RETRAINING THE PHONET LIBRARY USING US ENGLISH

by

HARSHA VEENA TADAVARTHY

(Under the Direction of Margaret E. L. Renwick)

ABSTRACT

This research presents re-training of Phonet Library, a speech technology that calculates posterior probabilities for phonological classes by leveraging distinctive features, on an American corpus. We re-trained the model on 49 phonemes classified by 24 distinctive features + pause (silence). We call the resulting model *Phonet_English.* It considers both the acoustic features and the phonetic features to estimate the posterior probabilities for a given audio signal. This statistical approach helps us understand patterned variability in speech. Phonet_English exhibits an impressive range of accuracies for phonological class recognition, with the lowest accuracy value of 80.7% and the highest accuracy value of 96.3%. This thesis also delves into the model's phoneme recognition accuracy and examines how its distinctive feature probabilities align with linguistic expectations for selected vowels and consonants. Our results showcase that Phonet_English is successful in capturing fundamental relationships between theoretical natural classes of sounds and their realization in English, making it highly useful in speech analysis and phonetic research.

INDEX WORDS: phoneme recognition, phonological posteriors, phonological classes distinctive features, phonology, recurrent neural network, natural classes.

PHONOLOGICAL FEATURE DETECTION FOR US ENGLISH USING THE PHONET LIBRARY

by

HARSHA VEENA TADAVARTHY

B.Tech., SASTRA University, India, 2018

A Dissertation Submitted to the Graduate Faculty of

The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2024

PHONOLOGICAL FEATURE DETECTION FOR US ENGLISH

USING THE PHONET LIBRARY

by

HARSHA VEENA TADAVARTHY

Major Professor:	Margaret E. L. Renwick
Committee:	Khaled Rasheed
	Shannon Quinn

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2024

ACKNOWLEDGEMENTS

I am extremely grateful to Dr. Margaret Renwick for her patience, continuous support, and guidance throughout this journey. Her expertise and insightful guidance have been invaluable. I would like to express my deepest appreciation to my committee members, Dr. Rasheed and Dr. Quinn, for agreeing to serve on my committee and for their contributions to improving this dissertation. I also extend my thanks to Austin Jones for his assistance in enhancing the research. I would like to acknowledge Aakash Gunda, without whose presence I would not be where I am today. Finally, I would like to thank my parents for their unwavering support and encouragement.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Background

In the world of language study, understanding how we make sounds (phonetics) and how these sounds work in different languages (phonology) is crucial. There are certain key elements, called distinctive features, that make up these speech sounds. These features are immensely beneficial for linguists in analyzing and understanding the diverse phonetic elements utilized in speech. They not only facilitate a deeper comprehension of our communicative methods and the construction of languages but are also regarded by phonologists as a formal system. This system undergoes rigorous scientific validation across various languages across the globe, highlighting its foundational role in the study of phonology.

For linguistic applications of speech technology, researchers often study systematic variations in pronunciation that may indicate language change, dialect differences among speakers, or context-dependent (allophonic) alternations. For example, in American English varieties, /t/ can change depending on its lexical positioning and adjacent phonemes. For instance, in words like *water* and *butter*, /t/ is typically realized as /ɾ/. Many tools, such as ASR systems or forced alignment systems, include a lexicon, and their goal is to assign specific word or phoneme labels to portions of the acoustic signal. These technologies can struggle with

capturing subtle speech variations, understanding context-dependent pronunciation, and recognizing diverse dialects, while also constraining researchers to analyze variation only at the level of individual segments.

In the field of pathological speech processing, it is difficult to obtain the speech features of the patient, which helps us to understand the patient's speech condition, by incorporating the traditional speech processing features like Mel frequency cepstral coefficients (MFCCs) and perceptual linear predictive coefficients (PLPs), which are widely used in the field of automatic speech recognition (ASR), speaker identification etc. It is difficult to extract the clinically interpretable features from the pathological context due to their complexity. Addressing this issue, the Phonet library (Vásquez-Correa et al. 2019) was developed, a recent tool that utilizes bidirectional Recurrent Neural Networks (RNNs) with Gated Recurrent Units (GRUs) to extract posteriors from speech signals. These phonological posteriors contain explainable data on the place and manner of the articulation that can be well understood by clinicians. The Spanish corpus CIEMPIESS, that consists of 17 hours of FM podcasts in Mexican Spanish with a sampling frequency of 16kHz and16-bit resolution, was used to train the original model. The model was trained on chunks of speech that were 0.5 seconds. The input features are extracted from the 33 Mel-filter banks. Using the Adam optimizer, the networks were trained with a weighted categorical cross-entropy loss function. To improve generalization, dropout and batch normalization layers were included. The model achieved an Unweighted Average Recall (UAR) ranges from 80.4% to 93.3% for phonological classes and recall values ranging from 0.50 to 0.80 for phoneme recognition.

This thesis presents a version of the Phonet system (Vásquez-Correa et al. 2019) trained on US English. We call it Phonet_English. Phonet is a classifier that uses recurrent neural networks to assign *phonological posterior probabilities* to portions of an input. Posteriors

correspond to natural classes, which in this case are assigned according to the sound system of US English. Natural classes are groups of speech sounds that share one or more phonetic or articulatory characteristics and behave similarly within the phonological system of a language (Mielke 2004). Phonet uses the full acoustic signal for analysis (just like speaker-hearers do during speech perception), rather than focusing e.g., on a small number of acoustic correlates, as in traditional phonetic analysis. We re-train the Phonet library using the Common Voice 14.0 dataset from the Mozilla Common Voice Project's US-based English corpus. Our basic method for evaluating Phonet is to compare its phonological posterior probabilities and phoneme recognition results against segmental labels from forced alignment. We focus on major natural class features for a subset of US English vowels and consonants and demonstrate that Phonet performs well.

1.2 Experiments and Results

This research focuses on retraining the model Phonet on a US English corpus and its subsequent performance. The re-trained model, which we call Phonet_English, exhibited an impressive range of accuracies with the lowest accuracy value of 80.7% and the highest accuracy value of 96.3% and is able to identify the phonemes with recall values ranging from 0.146 to 0.760.

To evaluate the model's performance, we plotted the AUC-ROC curve for each phonological class, and we obtained AUC values ranging from 0.56 to 0.92. Therefore, we can say that the model is successfully able to classify the phonemes. Our research proceeded to a more granular level of analysis focusing on specific phonological categories such as tense vowels, lax vowels, and coronal obstruents. We plotted violin graphs to analyze the distribution

of the posteriors. These graphs provided an intuitive way to visualize the data that allowed us to look into the distribution of the posteriors and in-depth patterns that might not be immediately seen from just looking at numerical data. This approach helped us understanding the similarities and distinctions between the phonological classes and their respective phonemes.

1.3 Contributions

This thesis contributes to the field of computational linguistics and speech analysis, especially in spoken language processing, by refining the Phonet (Vásquez-Correa et al. 2019) tool such that it can be used to analyze US English speech. This work has mostly focused on improving the performance of the model on analyzing the English speech. The results of the experiments shows that Phonet_English can be used to obtain phonological posteriors and it can be used in various linguistic explications. This comprehensive method has the potential to transform how phonological features are utilized and analysed in a wide range of linguistic applications, advancing both theoretical and applied linguistics. This work facilitates linguistic study and allows researchers to perform in-depth, nuanced studies of English speech. This works aims to significantly enhance linguistic analysis and phonetic transcription applications across a spectrum of fields including speech recognition and transcription, language learning and teaching, dialectology and sociolinguistics, phonetic research, speech therapy and rehabilitation, forensic linguistics, speech synthesis, and linguistic documentation of endangered languages.

1.4 Outline of the thesis

The outline of the thesis is as follows: Chapter 2 discusses more background information on human vocal tract, articulatory phonetics, natural classes, distinctive features, various state-

of-the-art machine learning models and their training processes used for the extraction of the signal features. Chapter 3 discusses the traditional and adopted distinctive feature theory and also defines the distinctive features that are relevant to this thesis. Chapter 4 provides in-depth information on the steps involved in training Phonet_English by describing the dataset used, explaining the necessary data preprocessing steps that need to be implemented before training the model and the expected output. Chapter 5 provides the results of the trained model. Chapter 6 gives in-depth details of evaluation on the performance of the model. Finally, Chapter 7 provides the conclusion of this thesis and possible future developments.

CHAPTER 2

DISTINCTIVE FEATURES IN PHONOLOGY

2.1 Introduction

The journey into understanding the complexity of phonological analysis through computational means, requires us to delve deeper into the anatomy of speech production, building blocks of linguistics and the cutting-edge computational methodologies that were used for speech recognition. In this section, we begin with exploring the vocal tract which is responsible for the production of speech (Zsiga 2013). This subsection focuses on the physiological mechanisms and articulatory mechanisms that are essential to create human speech sounds. To identify and understand the underlying pattern of the speech sounds and how these sounds are categorized, we next delve into the concepts of natural classes and distinctive features; we can better comprehend the complex structure of the language by grasping how sounds are grouped together according to shared phonetic characteristics. This investigation helps us understand the importance of elements from linguistic theory in developing computational models for speech analysis. Finally, we explore the evolution of sophisticated machine learning algorithms from traditional methods to interpret the nuances of spoken language. This part will elucidate how these computational techniques are employed to extract

and analyze phonological features from audio data. Moreover, we will also discuss various optimization techniques that were used to improve the performance of the model.

2.2 Anatomy and function of the Human Vocal Tract

The human vocal tract is responsible for speech in the human body. The vocal tract consists of 3 parts: sub-laryngeal, laryngeal or larynx, and supra-laryngeal. Speech is produced when air is driven into and out of the lungs. The larynx regulates pitch and voice vibration. The upper part of the larynx is said to be supra-laryngeal which is responsible for producing different speech sounds. The structure of the mouth is divided into passive articulators and active articulators. The lower lip, tongue tip, tongue blade, tongue body, and tongue root fall under the active articulators. The passive articulators include upper lip, upper teeth, alveolar ridge, post-alveolar region, hard palate, soft palate which is also known as velum, and pharyngeal wall. These articulators are used to control the flow of the airstream. Figure 1 shows the supralaryngeal vocal tract.

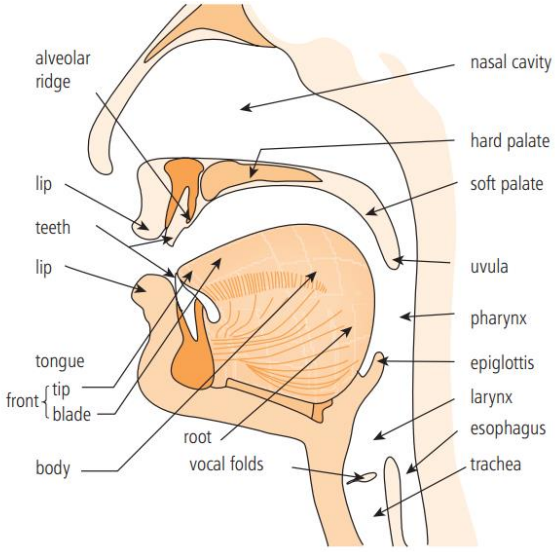


Figure 1: The supralaryngeal vocal tract, essential for understanding speech sound production and analysis. (Figure 1.8 in (Zsiga 2013)*)*

2.3 A quick overview of Phonetics and Phonology

Phonetics and Phonology are both subfields of linguistics that deal with how speech sounds are made and how we organize these sounds (Zsiga 2013). Within the context of *general phonetic theory* (Laver 2017), phonetics tells us how speech is produced by using the human organs and how these sounds are transmitted and perceived. Phonetics is mainly divided into three sub-categories articulatory phonetics, acoustic phonetics, and auditory phonetics (Skandera 2011).

2.4 Articulatory Phonetics: Manner and Place of Articulation

Building on the foundational understanding of the vocal tract's anatomy, we now focus on the specifics of how speech sounds are formed. Articulatory phonetics details the usage of the vocal organs to produce speech sounds, which are also called articulators. According to the movements and the position of different articulators, various speech sounds are created. Based on the level of constriction made in the vocal tract to produce the sound the "Manner of Articulation" is assigned. The place where this constriction is happening defines the "Place of Articulation." (We will discuss only the ones that are relevant to this research.).

The English consonants are categorized as follows (Zsiga 2013). The sounds that are formed by blocking the airflow are said to be *obstruents.* The sounds that resonate are *sonorants.* Vowels fall under sonorants. The sounds that are made where there is a complete stoppage of the air flow are said to be *Plosives.* These are also called *Oral stops.* In this research, the plosives are [p, b, t, d, k, g, ʔ]. We also considered the aspirated stops [p^h, t^h, k^h]. The sounds made by the vibration of the vocal folds (shown in Figure 1) are said to be *voiced* and the sounds that are

made without the vibration of the vocal fold are said to be *voiceless*. The action of opening the glottis, the space between the vocal folds, at the moment of the release of the oral closure of a stop is known as *aspiration* (Kim 1970). The sounds that are formed by pressing the two articulators together without closing them, resulting in a forced airflow, are *fricatives*. The sounds /f, v, θ, ð, s, z, ʃ, ʒ, h/ are considered for our research purposes. The sounds that are the combination of the plosive and fricative are said to be *affricates.* The sounds /tʃ, dʒ/ are considered. When the vocal tract is narrowed down by the active articulator, *approximants* are formed, without producing any friction. Additionally, this constriction facilitates resonance within vocal tract. The approximants included are /ɹ, j, w, l/. The l-sounds /l ɬ/ are the *lateral approximants.* When the active articulator strikes the passive articulator *flap* sounds are formed. The flap [ɾ] is considered.

The sounds are also distinguished into various categories depending on which place they are produced at. The sounds that are produced with both lips are *bilabials*. The bilabial stops and nasals are [p, b, m, p^h]. If the sounds are made with bottom lip and upper teeth, then they are *labiodentals.* The fricatives /f, v/ are labiodentals analyzed here. The sounds that are formed with the front tongue part are *coronals*. The sounds that are made with upper teeth and front part of the tongue are *dentals.* The considered dental fricatives are / θ, ð /. *Alveolars,* [t, d, n, s, z, l, ɹ, ɾ, t^h] are formed by using the front of the tongue and the alveolar ridge. *Post-alveolar* sounds /ʃ, ʒ, dʒ, tʃ/are made with blade of the tongue and post-alveolar region. When the middle part of the tongue is raised to meet the hard palate *palatals* /j/ are formed. When the back of the tongue is raised to meet the velum *velars* [k, g, ŋ, k^h ɬ] are formed. The sounds [h, ʔ] that are made at the glottis are called *glottals.* For a comprehensive overview of the consonants discussed, please refer to Table 1.

Manner of Articulation	Place of Articulation								
		Bilabial	Labiodental	Dental	Alveolar	Post-alveolar	Palatal	Velar	Glottal
	Plosive	p, p^h, b			t, t^h, d			k, k^h, g	ʔ
	Nasal	m			n			ŋ	
	Flap				ɾ				
	Fricative		f, v	θ, ð	s, z	ʃ, ʒ			h
	Approximant				ɹ		j		
	Lateral Approximant				l			ɬ	
	Affricate					dʒ, tʃ			

Table 1: English Consonants: An IPA Chart Overview.

The English vowels are categorized (Zsiga 2013) according to the position of the tongue (by height, backness, tenseness) and shape of the lips (rounded or unrounded). The front vowels, central vowels and back vowels that are considered are /i, ɪ, e, æ, ɛ/, /ə, ɚ/, and /ʊ, ɑ, ɔ/ respectively. When the sound is formed by combining two vowels, *diphthongs* are formed. The considered diphthongs are /aj, aw, ɔj, ej, ow/. Among these, /ɑ/ is an open vowel, /æ/ is a near-open vowel. The vowel /i/ are considered as high vowels and near-close vowels are /ɪ, ʊ/, and mid vowels include /ə, ɚ, ɛ, o/. The vowels discussed are shown in Table 2.

	Front	Central	Back
Close (High)	i	ɪ, ʉ	ʉː
Close-mid	ej	ə, ɚ	ʊ
Open-mid	ɛ		ɔj, ow
Open (Low)	æ		ɑ, aw

Table 2: English Vowels: An IPA Chart Overview.

2.5 Natural Classes

Natural classes are significant in phonology because these groups predict which sounds may participate in phonological processes across different languages. When a set of sounds demonstrates similar characteristics, it is typically likely that they are phonetically alike (Mielke 2008). Natural classes share certain phonetic properties and behave similarly in phonological processes. These classes are essential to phonological theory as they offer insights into how languages organize sounds and predict phonological patterns. According to Sylak-Glassman (2014), a natural class frequently appears across various unrelated languages, suggesting a universal linguistic principle that accounts for the arrangement of phonemes within that class. Typically, a natural class represents a phonological grouping of phonemes that can be justified by phonetic properties which distinguish these phonemes from all others. Conversely, an unnatural class consists of a phonological grouping that does not adhere to this phonetic rationale and is generally unique to a specific language. For instance, in Evenki (Tungusic), the phonemes /v, s, g/ transform into /m, n, ŋ/, respectively, when they follow a nasal consonant. Also, in Dravidian languages[1], a long radical long vowel shortens why they follow certain suffixes. For example, in the Telugu language, the long vowels are shortened when followed *-ku*, *-gu*, *-cu*, *-tu*, and *-du*. For instance, the root word "*pāl-*" changes to "*paluku*" (meaning "to say" or "to speak") when the suffix "*-uku*" is added, shortening the long vowel "*ā*" to "*a*" (Krishnamurti 1955).

According to Sagey (1990), natural classes are thought to be defined by specific combinations of phonological features. This theory offers precise and robust predictions but falls short in encompassing phonological classes that are natural due to similar patterns in phonetic features. For example, Noam Chomsky and Morris Halle (1968) wrote that “if a theory of

[1] Generally, South India Languages like Telugu, Tamil, Kannada, Tulu, and Malayalam are Dravidian languages.

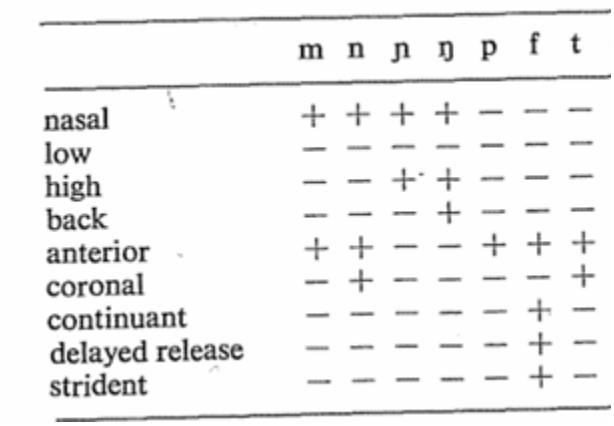
language failed to provide a mechanism for making distinctions between more or less natural classes of segments, this failure would be sufficient reason for rejecting the theory as being incapable of attaining the level of explanatory adequacy". According to Flemming (2005), natural classes align formally with phonological classes since they are defined as any group of phonemes affected identically during a derivation. Consequently, there is no formal differentiation between natural and unnatural classes in terms of having a phonetic foundation and potential occurrence across various languages. This similarity implies that this concept lacks the capability to predict which phonological classes are likely to appear in multiple languages.

2.6 Distinctive features

In phonological theory, features are the units that constitute individual segments (e.g., vowels and consonants); they are defined both by their articulatory and acoustic properties. The idea of distinctive features has long been a cornerstone of phonology. The concept of distinctive features is based on the premise that the identification of a sound relies on its lexical contrast with other sounds within the system (Trubetzkoy 1939). Later, in 1952, Jakobson, Fant and Halle published *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates.* Distinctive features are used in theoretical phonology to place phonemes into natural classes based on shared phonetic properties. Traditional approaches hold that features are binary, to capture mutually exclusive "oppositions" in characteristics like [+voice] (for /b/) vs. [–voice] (for /p/) (Jakobson, Fant & Halle 1963). Distinctive features help us understand the phonetic and phonological properties of the sound, as well as the difference(s) between two sounds. They capture the phonological properties of a segmental inventory, using a closed class of descriptors

that link the mental representation of speech to its specific articulatory or acoustic properties, such as place or manner of articulation for consonants.

According to Chomsky & Halle (1968), distinctive features in phonological theory are binary attributes that provide a systematic representation of the phonetic and phonological properties of the speech sounds. These features describe how individual phonemes are articulated and perceived across languages. They break down phonemes, the smallest units of sound that can differentiate meaning, into a series of binary oppositions such as [+voice] or [–voice], [+nasal] or [–nasal]. This binary system allows linguists to categorize phonemes not just based on their individual characteristics, but also in relation to one another, enabling a deeper understanding of the internal structure of phonological systems in any given language. One prominent way to represent these binary attributes is through the traditional feature matrix. In this approach, phonemes are viewed as bundles of distinctive features that are organized in a two-dimensional grid: one dimension is for the distinctive feature set, and the other is for the phonemes. For example, in a feature matrix phoneme as shown in Figure 2, /n/ is represented as [+nasal] and /p/ is represented as [–nasal].



	m	n	ɲ	ŋ	p	f	t
nasal	+	+	+	+	−	−	−
low	−	−	−	−	−	−	−
high	−	−	+	+	−	−	−
back	−	−	−	+	−	−	−
anterior	+	+	−	−	+	+	+
coronal	−	+	−	−	−	−	+
continuant	−	−	−	−	−	+	−
delayed release	−	−	−	−	−	+	−
strident	−	−	−	−	−	+	−

Figure 2: Example of a Feature Matrix - shows the initialization of phonemes in a traditional binary system. (Chomsky & Halle 1968)

This structured approach greatly facilitates the description and classification of the sounds of any language; linguists can analyze and compare phonemes in the matrix according to shared features. These features simplify the description and classification of sounds across languages. Therefore, these are important in phonological theory. With the help of a set of universal phonetic categories, linguists can effectively describe and predict phonemic behavior in phonological contexts, such as sound changes in different environments or the organization of phoneme systems in unfamiliar languages. Hence, this universal applicability of these features suggests their fundamental role in what Chomsky describes as Universal Grammar (UG) – a set of innate linguistic principles and structures shared by all human languages. When placed under such a light, the distinctive features as part of UG are almost the same as a cognitive blueprint guiding the process of language acquisition and processing, claiming that knowledge of language perception and production is not completely acquired but is also highly predetermined by genetic factors. Furthermore, the integration of distinctive features into UG supports the argument that the human brain is pre-wired with a specific set of linguistic capabilities. These capabilities facilitate the idea of rapid acquisition of language in early childhood, guiding the development of phonological systems by providing a framework within which all languages operate. For example, the presence of features like place, and manner of articulation helps children distinguish and produce the sounds necessary for effective communication within their linguistic environment without explicit instruction. Moreover, the research also discusses the application of the distinctive features in phonological rules, which are guidelines of how sounds interact and change in specific linguistic contexts. These rules are universally applicable due to their foundation in distinctive features. Hence, this further exemplifies the innate and systematic nature of human language as proposed by theories of UG.

In (1985), Clements addresses a weakness in the traditional feature matrix and proposes a multi-tiered hierarchical structure for the description of phonological features. He argues that the two-dimensional grid approach does not capture the complexity and hierarchical nature of the phonological systems. It treats features just as if they were independent and unstructured, not being able to reflect the real interactions and dependencies of features in natural languages. He concludes that the proposed structure offers a way in which phonological features and their interactions can be represented in a dynamic manner. This new model, also known as feature geometry, arranges the features into natural classes and considers insights from auto segmental and metrical phonology. This approach provides a more precise and comprehensive framework for analyzing phonological phenomena. His research is a significant advancement in phonological theory. It helps linguists to understand the behavior of the phonological features across various languages.

Hall (2001) explores two main topics. Firstly, how features are represented in phonology, like the structure of features and when certain features are left unspecified. Secondly, how these phonological features are related to their actual use and interpretation in phonetics. Hall (2007), stands as a comprehensive resource in the field of phonology, offering a detailed exploration of phonological theories and concepts that are fundamental to the study of language sounds. This chapter delves into a variety of phonological features, including segmental features that form the basic units of sound, major class features which distinguish between different types of sounds like sonorants and obstruents, laryngeal features that deal with voice and voicelessness, manner features which describe how air flows through the mouth during speech, and place features that indicate where in the mouth sounds are articulated.

The research in the Mielke (2008), challenges the traditional view described by Chomsky & Halle (1968), which states that distinctive features are part of UG. A finite, innate set of

features can describe all natural classes in any language, reflecting an inherent aspect of human linguistic capability. Mielke contends that distinctive features are not inherent but develop with the use of language. His study involved a large-scale cross-linguistic sample that involved 561 languages and 6,077 different sound classes. The results of the survey showed that many natural classes are not predicted by such theories. More than 24% of these classes remain unaccounted for by any existing theory, which points out that the occurrence of phonological patterns is wider and more variable than innate theories can account for. He introduces Emergent Feature Theory. This theory states that distinctive features arise from the phonological patterns observed in languages rather than being pre-specified. In this perspective, features are abstract generalizations that are developed while learning, under the influence of phonetic properties, cognitive processes, and social interaction. Therefore, according to the Emergent Feature Theory, the features themselves are learned constructs, reflecting the language-specific experiences of the speaker, while phonological patterns are constrained by phonetic factors. This theory accounts for variability and adaptability across the phonological systems of different languages; in that sense, it goes against the universality claimed by the innate feature theories.

Cohn (2011) includes a detailed review of the multifaceted role of distinctive features as defined by Chomsky & Halle (1968), focusing on their ability to capture contrast, phonological patterns, and phonetic realizations. The research argues that while the characterization of segments as combinations of universally defined distinctive features is approximately correct, it is not entirely accurate. This study looks at different views that could have the primitives of phonology be either segments, features, neither, or both, as well as the transition from the concept of segments to the understanding of distinctive features being the primitives in generative phonology. The study also discusses the concept of distinctive features as innate and universal elements which are necessary to describe the possible speech sounds and to explain the

concept of natural classes. Moreover, it details language-specific phonetics, showing that phonetic implementation varies across languages, with the help of examples, such as vowel lengthening and intonation patterns. It illustrates that the same phonological features can manifest differently depending on the language. The paper also examines how distinctive features help characterize phonological inventories and alternations. Mielke (2008) demonstrates that while distinctive features account for many phonological patterns, they also show limitations. The study also details the relationship of adult phonological systems and language acquisition and points out that the way in which those phonological systems are learned might be different from the end state knowledge of adult systems.

The phonological representations which are also called the representational aspect of phonology are defined by syntagmatic and paradigmatic dimensions (van der Hulst 2016). The organizational structure of breaking down phonemes into syllables, then those syllables into words, and beyond is said to be the syntagmatic dimension. This approach highlights how speech sounds are arranged in a hierarchical structure and in a sequential order within language. On the other hand, the concept of considering the idea that phonemes are not the smallest units of sound structure, rather, they can be broken down into even smaller, fundamental elements is said to be the paradigmatic dimension. Van der Hulst's paper examines the unary elements with the help of using the three frameworks: Dependency Phonology (DP), Government Phonology (GP), and Radical CV Phonology (RCVP). Each framework helps us to understand the unary features and how they are used to depict the underlying phonological structures across different languages. Dependency Phonology (introduced by Anderson 1987) explains the importance of unary features and how these help us to learn about the *dependency relationships* seen in the phonological structures. Government Phonology (Kaye, Lowenstamm & Vergnaud 1985; Kaye, Lowenstamm & Vergnaud 1990) introduces a set of unary components that are simple, and the

phonological information is processed cognitively. Radical CV Phonology further simplifies the idea and proposes that all phonological structures can be expressed by only two basic unary elements, simplifying the complexity of phonological representation. The analysis concludes that the unary features are more beneficial when compared to binary features to capture the phonological characteristics. Another alternative is a theoretically motivated mixture of feature types (Zsiga 2013).

According to Mitkov et al. (2014), phonetic similarity is judged primarily by place and manner of articulation, as organized in the International Phonetic Alphabet (IPA) consonant chart. This chart categorizes sounds based on their articulatory features. The IPA chart follows the binary distinctive feature system. The IPA chart for the phonemes in this study are represented in Tables 1 and 2.

2.7 Phonetic correlates of distinctive features

There are regular linguistic patterns such as *all tense vowels* or *all coronal obstruents*, within specific segmental contexts. For example, tense vowels are produced with more peripheral articulatory movements than lax vowels. This results in more extreme formant frequencies, leading to a larger vowel space (Hillenbrand et al. 1995). Therefore, the vowel space between the first (F1) and second formant (F2) frequencies is generally large when compared to lax vowels. Tense vowels show greater lengthening (temporal modifications) whereas lax vowels exhibit greater changes in their dynamic spectral properties. For example, the duration of the sound /i/ in “see” or /u/ in “boot” is longer in clear speech. On the other hand, lax vowels, such as /ɪ/ in “sit” or /ʊ/ in “put”, display more significant changes in their acoustic qualities over time, with noticeable shifts in formant frequencies (F1, F2, F3) (Leung 2016). Coronal

obstruents undergo place assimilation (Zuraiq & Zhang 2006). This phonological process involves altering the place of articulation to align with the place of articulation of the following consonant.

From these patterns we can interpret the underlying structures of languages and their systematic variations. Phonological features often do not correspond directly to segments on a one-to-one basis. In American English, for example, vowels typically become nasalized before nasal consonants, as in the word *pan* [pæ̃n], where the [nasal] feature extends over two segments (mapping one feature to two phones). Also, consider the affricates /tʃ/ and /dʒ/, which start with a stop closure, represented by [–continuant] (lacking the [continuant] feature), and end with fricative noise, represented by [+continuant] (two features mapping to one phoneme). Since these phonological elements often span multiple segments and can extend across word boundaries, it is crucial to have tools capable of identifying features without being restricted by segmental limits (further explored in subsequent sections). Phonetics researchers seek to identify precise acoustic correlates for distinctive features, independent of symbolic transcription.

According to Johnson (2005), listeners can often perceive two acoustically different signals as the "same sound". However, identifying consistent acoustic cues that correspond to commonly recognized phonological features is quite a challenge (Stevens & Blumstein 1981). On the other hand, early acoustic phonetic research showed that the relationships between phonetic features and acoustic features were typically centered on a limited number of human-measurable and interpretable correlates of features. The study, Peterson & Barney (1952), analyzed the acoustic characteristics of vowels to investigate the relation between the vowel sounds that speakers targeted, and the vowel sounds that listeners perceptually experienced. The values of these formants represent the resonant frequencies of the vocal tract. For this paper, the researchers recorded and measured the pronunciation of ten monosyllabic words – each with

different vowel sounds for 76 speakers and 70 listeners – and analyzed vowel quality. The results conclude that identification of the vowels dependent on the speaker's dialectal background and the formant value was an important acoustic cue in vowel quality. Delattre, Liberman & Cooper (1955) studied the role of second-formant transitions of the stop and nasal consonants. The results conclude that that each consonant has a fixed frequency position, or locus, of the second formant, and this locus can be associated with a consonant's place of articulation. The research also investigated the invariant acoustic cues for the place of articulation in stop consonants within the consonant-vowel syllables. The study (Stevens & Blumstein 1978) was carried out by pairing synthetic stop consonants with different vowels and manipulating acoustic attributes such as noise bursts at onset and formant transitions following consonantal release. The study concludes that the stimuli containing formant transitions, with or without noise bursts, were consistently identified according to place of articulation, while those are amplified with noise bursts. Lisker & Abramson (1964) measured the voice onset time (VOT) for initial stop consonants in their cross-language work. The duration of the release of the stop closure and the onset of voicing is measured in several different languages. The research concludes that VOT is one important cue to the distinction between voiced and voiceless stops, or to the distinction between aspirated and unaspirated stops. For example, in English, voicing-onset time for the voiced stops /b/ and /d/ is low, while for the voiceless aspirated stops /p/ and /t/ is high[2].

[2] Please refer to (Olive, et al., 1993) for more examples.

2.8 Speech technologies

In phonetic analysis, speech technologies are necessary to extract distinctive features within phonetic analysis because they are more precise, efficient, and can work with voluminous data. Due to subjective perceptions, there are bound to be errors with manual transcriptions, meaning that they are not reliable. Technological tools make objective and consistent measurements of acoustic characteristics like frequency, amplitude, and duration. Such tools are essential for the visualization of fine phonetic details that usually are lost – fine detail in the position of formants, structure of their transitions, or fine structure of the pitch contours. Besides, speech technologies allow large-scale analysis that permits researchers to process great amounts of speech data. This, in turn, is of utmost importance for cross-linguistic comparisons and universal phonetic features. It allows complex analysis of the speech signal because it uses advanced signal processing techniques such as Fourier analysis and Mel-Frequency Cepstral Coefficients (MFCCs). Practical examples and case studies (which are discussed later further in detail) represent the necessity and reason for technological tools in modern phonetic analysis; accordingly, detailed, objective, and comprehensive studies of speech sounds have been made possible in this area.

Lee (1989) utilizes the Hidden Markov models (HMM) for speaker independent phone recognition. The analysis was carried out on the subset of the TIMIT database (Lamel & Kassel 1986; Fisher, Zue & Bernstein 1987). The data considered consists of 2830 sentences from 357 speakers, used for training, and 160 sentences from 20 speakers used for testing. The model was trained under two scenarios: *Context - Independent* and *Context – Dependent*. In the case of the latter, the phone was dependent on the neighboring phones. This paper also proposes the concept of co-occurrence smoothing, which determines the similarity of each pair of phones, and then

modifies the distributions of each phone according to it. The model achieved a phone recognition accuracy of 73.8% speaker-independent phone recognition. This was the best accuracy reported at that time.

The study Lee et al. (2000) suggests a brand-new technique for extracting speech features from human audio data using Independent Component Analysis (ICA). The authors show that by extracting Gabor-like characteristics that are confined in both time and frequency, ICA can effectively encode speech signals. These characteristics, which together constitute a novel filter bank, outperformed conventional Mel-Frequency Cepstral Coefficients (MFCCs) in speech recognition tests. The structure of the ICA network is composed of inputs and outputs that have equivalent sizes of speech segments and the feature vectors are represented by inverse of the trained weight matrix. The ICA model was trained by using samples of human voice signals from 75 phonetically balanced Korean words said by 59 speakers. The ICA network was trained with randomly produced speech segments to extract basis function. To compare the effectiveness of ICA-based features against MFCCs, experiments were conducted to train the ICA network by determining dominant feature vectors and performing isolated-word recognition tasks. The study demonstrates that the proposed method achieved an error reduction of 47.4% when compared to MFCCs. The paper concludes that the extracting features employed by the ICA technique outperformed a traditional MFCCs approach and is an effective method for speech feature extraction.

In Lee et al. (2009) the authors use convolutional deep belief networks (CDBNs) to learn features of unlabeled data (speech and music), for tasks like speaker identification, gender classification, phone classification, and music genre classification. The authors demonstrate that CDBNs can learn hierarchical representations from the given signal, and these hierarchical representations increased the performance of the deep learning models, when compared with the

models trained on traditional features MFCCs, for audio classification tasks especially in case of limited labelled data. The architecture of the CDBNs is refined from the architecture of the convolutional restricted Boltzmann machine (CRBMs) for unsupervised learning of the hierarchical audio features. The CDBN is a CRBM with probabilistic max-pooling. For this research purpose the unlabeled TIMIT dataset (Fisher, Doddington & Goudie-Marshall 1986) is used. The first and second layers of the CDBN are trained via unsupervised learning method with the spectrogram as the input to the layers. The spectrogram is extracted from the utterances of the training data and has a window size of 20ms with 10ms overlaps. The model undergoes greedy layer wise training; the hidden layers are trained in a bottom-up fashion. The contrastive divergence technique is used to approximate the gradient effectively. The results of the experiments demonstrate that the proposed method outperformed the MFCC features by achieving an over 90% accuracy for speaker identification, around 95% for gender classification and around 80% for phone classification.

Graves, Mohamed & Hinton (2013) explore the application of Deep Recurrent Neural Networks (RNNs), with a focus on Long Short-term Memory (LSTM) architectures, for speech recognition. The analysis was carried out on the TIMIT corpus (DARPA-ISTO 1990), utilizing audio recordings from 74 speakers. The audio signals were analyzed using a Fourier-transform-based filter-bank. The signal data was normalized to make sure that all inputs have zero mean and unit variance. The models that were used are LSTM, Deep LSTM, Bidirectional LSTM (Bi-LSTM) and Deep Bidirectional LSTM. The architecture of the models ranged from simple configurations CTC-1L-250H, a single-layer LSTM network trained with Connectionist Temporal Classification (CTC), to more complex structures like the CTC-5L-250H. The paper mainly focuses on the impact of the deepening of the number of multiple layers of recurrent hidden units and explores the efficiency of these architectures in capturing the long-term

dependencies to enhance the accuracy of the speech recognition. The models were trained to recognize 61 phoneme labels which were mapped to 39 phonological classes. The stochastic gradient descent optimization technique was used with a learning rate of 0.0001. The proposed LSTM RNN model achieved a minimum test error of 17.7%.

In Arora, Lahiri & Reetz (2015), the research explores the advancement of ASR systems. The study proposes a novel framework that integrates phonological insights by transforming acoustic signals into a phonological feature space using Artificial Neural Networks (ANNs), to improve digit recognition. This approach is tested on the TIDIGITS database. The aim of the experiment is to overcome traditional ASR limitations in handling phonological variations such as assimilation and coarticulation. The demographic and dialectal diversity in the database gives the system robustness. Here, instead of the conventional statistical methods, the experiment incorporated the ANNs for the extraction of phonological features and a ternary matching scheme for phoneme estimation. It showed an overall accuracy of 62% in the recognition of digits. The study concludes that integration of phonological knowledge in ASR systems provides a promising way ahead, opening the gates to much more adaptive and robust speech recognition technologies.

The article Arora, Lahiri & Reetz (2018) introduces an ASR system dedicated to the specific task of improving pronunciation training for the non-native English learners (L2 learners). The proposed method uses Deep Neural Networks (DNNs) and Hidden Markov Models (HMMs) to analyze the speech of L2 learners at a sub-phonemic level. This approach allows the feedback produced regarding the pronunciation error to be more precise. The performance of the system was tested using an experiment with participants that are German and Italians learning English. The results show that the proposed approach resulted in high accuracy in detecting mispronunciations and diagnosing specific phonological errors. In real-world

application situations, the Deep Neural Networks (DNNs) for feature extraction to the HMMs for their interpretation, have been demonstrated to high performance. The study concludes that phonological feature-based systems can give detailed, constructive feedback in a way that helps learners substantially improve their pronunciation skills in a new language.

Sailor & Patil (2016) used Convolutional Restricted Boltzmann Machine (ConvRBM) model to extract the features from the given signal based on the unsupervised learning. In this model, initially, the speech signal is fed into a convolution layer, followed by the application of Rectified Linear Unit (ReLU) activation function. Subsequently, the pooling operation is performed with a window length of 25 milliseconds and a window shift of 10 milliseconds. Finally, the logarithmic transformation is applied to the output. The trained model achieved a performance improvement of 5% on the TIMIT corpus (Garofolo et al. 1993) and the Wall Street Journal WSJ0 database (Paul & Baker 1992) when compared to MFCC and Mel-filterbank. The data from TIMIT includes utterances from around 500 speakers and the from WSJ0 includes data from 84 speakers which consists of 7138 utterances. The model was fed normalized speech samples and was trained with a learning rate of 0.005 for the first 10 epochs. Following that, the learning rate was decreased.

Phonvoc (Cernak & Garner 2016) is a toolkit equipped with fully connected parallel networks that could identify 15 different phonological patterns in the English language, including nasal, strident, and vocalic classes, with more than 96% accuracy. Feature-based tools have been used in clinical applications; Jiao, Berisha & Liss (2017) used recurrent neural networks with long short-term memory, trained on the TIMIT database, to identify 15 sound patterns and assess the pronunciation abilities of individuals with speech difficulties, achieving over 90% accuracy in detecting these patterns.

2.9 Previous work using Phonet

In this thesis, we aim to re-train the Phonet Library (Vásquez-Correa et al. 2019) to adapt it for English speech. It provides deeper insights on the dynamics of spoken language, which will aid theoretical linguistic research and practical applications such as speech recognition and language-teaching tools. There are previous tools that utilized Phonet to understand the phonetic phenomena.

For instance, Wayland et al. (2023), offer in-depth research specifically focused on the lenition of voiced and voiceless stops in intervocalic positions of Argentinian Spanish. That study compares classic acoustic measurements with state-of-the-art deep learning techniques to enhance the quantification of lenition, an extremely common phonological process in which consonants become less obstructive in their articulation, hence affecting their acoustic quality. More particularly, the study deals with the lenition process of the Spanish voiced stops /b, d, g/ that surface as fricatives [β, ð, ɣ] in intervocalic positions, conditioned by phonetic factors such as stress, place of articulation, surrounding vowel quality, and speaking rate. In this experiment the researchers adopted a multi-method approach for the analysis of a corpus of Argentinian Spanish. They utilized both classic acoustic metrics (intensity, duration, spectral properties) and Phonet to analyze the Argentinian Spanish corpus. The output posterior probabilities of sonorant and continuant phonological features that are obtained by Phonet are compared against traditional measurements to assess its efficacy in capturing the nuances of lenition. This comparison was critical to evaluate the model's generalization capacity over lenition patterns that fulfil phonetic restrictions. Results from the study showed that the traditional acoustic measurements and Phonet gave good guidance as to the process of lenition. Specifically, the predictions of the Phonet model compared quite well with traditional acoustic measurements,

such as intensity and duration, in its ability to simulate expected phonological changes. The posterior probability predictions of sonorant and continuant classes made by Phonet took on a similar pattern to those taken from the relative acoustic intensity measures, which illustrated its effectiveness in reflecting the effort-based view of lenition and previous phonetic findings. Thereby, it allowed Phonet to be a valuable and a reliable tool alternative or supplementary source for a detailed phonological analysis.

The research (Tang et al. 2023) presents a study of lenition in speech using Phonet. The model is trained to recognize posterior probabilities of sonorant and continuant phonological features in Argentinian Spanish. The research focuses on voiced and voiceless stops, uncovering lenition patterns that align with previous studies and revealing additional patterns. Results of the study showed that Phonet can simulate the lenition patterns that were otherwise detected using traditional quantitative acoustic methodologies. More lenition of voiceless stops was found than for voiced stops, while lenition was also more prevalent in an unstressed compared to a stressed syllable. This would seem to accord with linguistic theory that lenition is conditionally due to the phonetic environment and to articulatory effort and would confirm the model's ability to interpret these subtle phonetic changes. The study confirms the model's effectiveness as an alternative or complement to traditional quantitative acoustic measures of lenition. It also concludes that Phonet will be the new application of automated lenition measurement, through the translation of complex acoustic data into interpretable phonological features. This model frees researchers from manual labeling and helps to analyze a great quantity of data. The authors argue that such a tool will greatly help in allowing scalable and efficient research on phonetic and phonological patterns.

Both aforementioned studies used a corpus of crowd-sourced recordings from 44 native Argentinian Spanish speakers, encompassing over 8.0 hours of speech with 7449 unique words,

to analyze word tokens with voiced and voiceless stops in varying vocalic environments. Both studies conclude that Phonet can automate and refine the analysis of complex phonological features and provides a fine-grained understanding of lenition when compared to traditional methods. Therefore, Phonet is a very useful research tool that enables in-depth analysis of the dynamics of language sound systems. It expands the frontier of linguistic research with methodologies for acoustic speech analysis, which opens a scope for further exploration of the complex interrelationship of phonetics and phonology in natural languages.

CHAPTER 3

DISTINCTIVE FEATURE ASSIGNMENTS

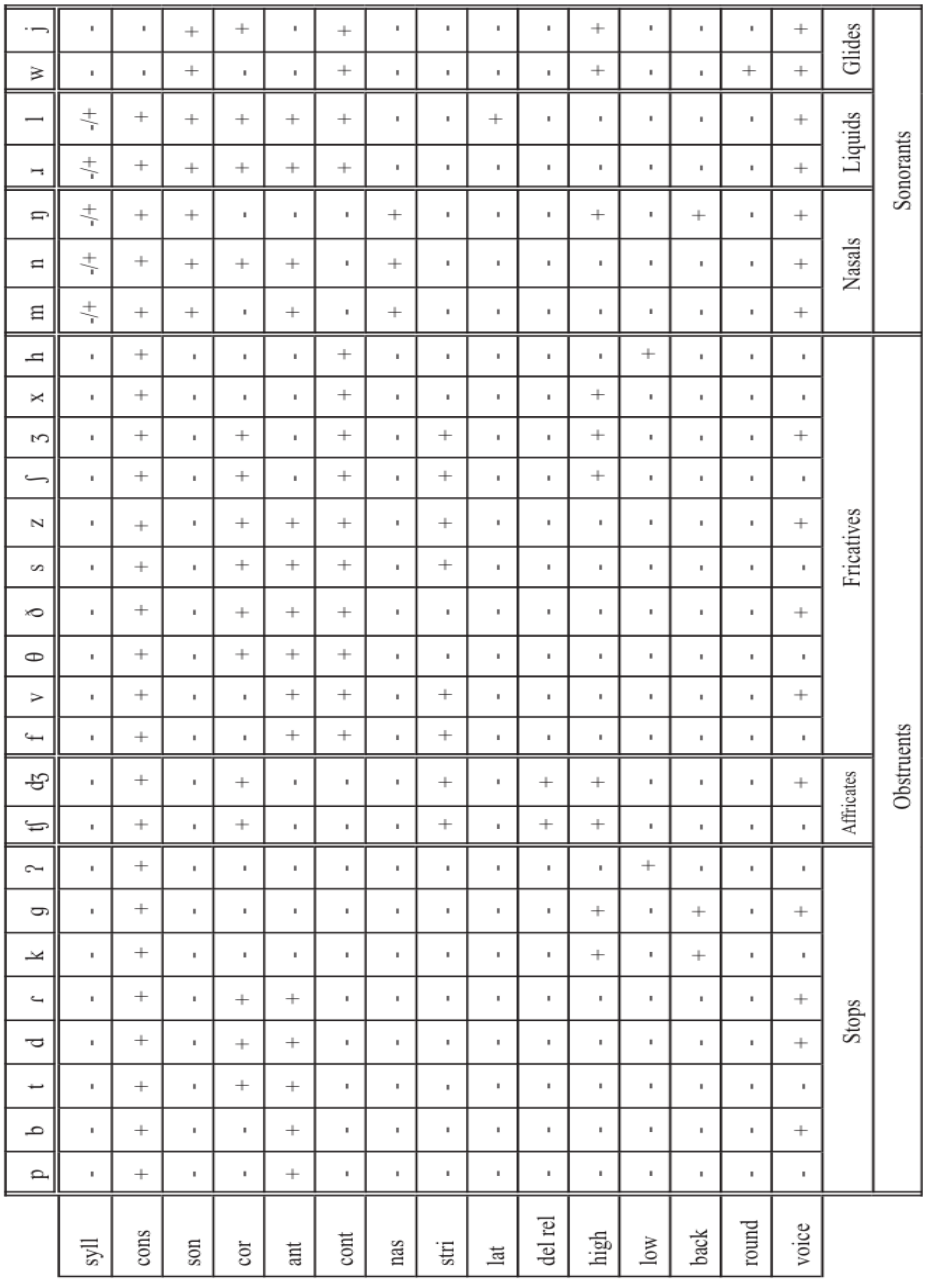
3.1 Introduction

In this chapter we discuss the distinctive feature assignments that were adapted in Phonet_English. We start with Section 3.2, that elaborates binary and unary distinctive feature system in Phonet_English. Next, Section 3.3 defines the Manner Features, Section 3.4 defines the Laryngeal Features, Section 3.4 defines the Place Features, Section 3.5 defines the Vowel Features and Section 3.6 describes the other features that are considered for training Phonet_English.

3.2 Binary and Unary Distinctive Feature System

The traditional binary distinctive feature systems for consonants and vowels are shown in Figures 3 and 4, respectively. On the other hand, the Phonet architecture uses a unary distinctive feature system. For example, phones like /i e a o u/ are typically characterized as [+syllabic]. In a binary feature theory, consonants outside of this natural class like /d t z/ would be designated as [-syllabic]; but for Phonet, [syllabic] is specified for all vowels, and simply not specified otherwise. This system reduces the complexity of the input data to the neural network during the

training phase of the model, leading to quicker model convergence and lower computational demands.



	p	b	t	d	ɾ	k	g	ʔ	ʧ	ʤ	f	v	θ	ð	s	z	ʃ	ʒ	x	h	m	n	ŋ	ɹ	l	w	j		
syll	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-/+	-/+	-/+	-/+	-/+	-	-		
cons	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-		
son	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+		
cor	-	-	+	+	+	-	-	-	+	+	-	-	+	+	+	+	+	+	-	-	-	+	-	+	+	-	+		
ant	+	+	+	+	+	-	-	-	-	-	+	+	+	+	+	+	-	-	-	-	+	+	-	+	+	-	-		
cont	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	+	+		
nas	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-		
stri	-	-	-	-	-	-	-	-	+	+	+	+	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-		
lat	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-		
del rel	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
high	-	-	-	-	-	+	+	-	+	+	-	-	-	-	-	-	+	+	+	-	-	-	+	-	-	+	+		
low	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-		
back	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-		
round	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-		
voice	-	+	-	+	+	-	+	-	-	+	-	+	-	+	-	+	-	+	-	-	+	+	+	+	+	+	+		
	Stops								Affricates		Fricatives										Nasals			Liquids		Glides		Sonorants	Obstruents

Figure 3: Distinctive feature specifications for British English Consonants.
Table 7.1 in (Davenport & Hannahs 2010)

	iː	ɪ	uː	ʊ	ɔ	oː	ɒ	ɑː	ʌ	æ	eː	ɛ	ə	ɜː
high	+	+	+	+	−	−	−	−	−	−	−	−	−	−
low	−	−	−	−	−	−	+	+	+	+	−	−	−	−
back	−	−	+	+	+	+	+	+	−	−	−	−	−	−
front	+	+	−	−	−	−	−	−	−	+	+	+	−	−
round	−	−	+	+	+	+	+	−	−	−	−	−	−	−
tense	+	−	+	−	−	+	−	+	−	−	+	−	−	+

Figure 4: Distinctive feature specifications for British English Vowels Table 7.2 in (Davenport & Hannahs 2010)

Now, let us dive into different distinctive features namely: Manner features, Place features, Laryngeal features, Vowel features. These features help us to understand the phonetic characteristics that distinguish different speech sounds. This detailed analysis aids in enhancing speech recognition systems and improving the accuracy of phonetic transcriptions.

3.3 Manner Features

[syllabic] sounds form the nucleus of a syllable, and mainly include vowels and syllabic consonants. [consonantal] sounds are the ones that are formed due to significant constriction in the vocal tract. [sonorant] sounds are produced by the accumulation of pressure behind an oral constriction and encompass vowels, nasals, approximants, and laterals. If the sound is produced when the oral cavity is not restricted, then it is a [continuant]. Vowels, fricatives, and approximants are Continuant. [nasal] sounds emerge when the velum is open, allowing air to escape through the nose, as in the sounds /m/, /n/, and /ŋ/. The [lateral] feature indicates a sound that is formed when the vocal tract is open at sides and closed at the center, allowing the airstream to flow over the sides of the tongue.

3.4 Laryngeal Features

[spread glottis] features indicate that there is a significant glottal opening gesture. The feature [voice] indicates the vibration of the vocal folds.

3.5 Place Features

The [distributed] feature tells if the sounds are produced using the blade of the tongue with a long constriction. A sound is [anterior] if it is made with the tongue front at or in front of the alveolar ridge. Dentals and alveolars are [anterior], while postalveolars, retroflexes and palatals crucially are not. [strident] sounds have high-amplitude and high-pitched frication. The [dorsal] feature tells if the sound is formed by moving the back of the tongue against or toward the velum. [labial] tells if the sounds are formed by using the lower lip. The [coronal] feature tells if the sounds produced use the tip or blade of the tongue.

3.6 Vowel Features

The [high] feature tells if vowels are produced with the tongue positioned close to the top of the mouth, the [low] feature tells if vowels have the tongue positioned at the bottom of the mouth, the [back] feature tells if vowels are articulated when the tongue is moved towards the back of the mouth, and the [front] feature tells if the vowels are articulated when the tongue advances towards the front of the mouth. Diphthongs, as defined before, start with one vowel and glide into another vowel within the same syllable. [round] tells if the sounds are articulated with rounded lips. The [tense] feature tells if the vowels are formed by stiffening the tongue root.

3.7 Other Features

In our research, we also use the [flap] feature that tells if the sound is formed when the tongue and alveolar ridge are in contact. Incorporating this feature is necessary, as it helps the model to identify the alveolar stops /t/ and /d/. The [rhotic] feature tells if the sounds possess a r-like quality.

Table 3 shows the distribution of phones into phonological classes for US English as they were specified for Phonet.

Phonological Feature	**Phonemes**
Syllabic	/a, æ, aj, aw, ɔj, ə, ɚ, ej, ɛ, i, ɪ, ow, ʉ, ʉː[3], ʊ/
Consonantal	/b, d, ð, dʒ, f, g, h, k, kʰ, l, ɬ, m, n, ŋ, p, pʰ, ɹ, ɾ, s, ʃ, t, tʰ, tʃ, v, z, ʒ, ʔ, θ/
Sonorant	/a, æ, aj, aw, ɔj, ə, ɚ, ej, ɛ, i, ɪ, j, l, ɬ, m, n, ŋ, ow, ɹ, ɾ, ʊ, ʉ, ʉː, w/
Continuant	/a, æ, aj, aw, ɔj, ð, ə, ɚ, ej, ɛ, f, h, i, ɪ, j, l, ɬ, o, ow, ɹ, ɾ, s, ʃ, tʃ, ʉ, ʉː, ʊ, v, w, z, ʒ, θ/
Nasal	/m, n, ŋ/
Voice	/a, æ, aj, aw, b, ɔj, d, ð, dʒ, ə, ɚ, ej, ɛ, g, i, ɪ, j, l, ɬ, m, n, ŋ, ow, ɹ, ɾ, u, ʉ, ʉː, ʊ, v, w, z, ʒ/
Labial	/b, f, m, p, pʰ, v, w/
Round	/aw, ɔj, ow, ʊ/
Coronal	/d, ð, dʒ, j, l, n, ɹ, ɾ, s, ʃ, t, tʰ, tʃ, z, ʒ, θ/
Distributed	/ð, dʒ, ɹ, ʃ, ʒ, θ/
Anterior	/d, ð, l, n, ɹ, ɾ, s, t, tʰ, z, θ/
Strident	/dʒ, s, ʃ, tʃ, z, ʒ/
Spread glottis	/h, kʰ, pʰ, tʰ/
Lateral	/l, ɬ/
Dorsal	/g, j, k, kʰ, ɬ, ŋ, w/

[3] The symbol '**ː**' , indicates it is a long vowel.

Rhotic	/ɚ, ɹ, ɾ/
Flap	/ɾ/
Long	/ ʉː/
High	/i, ɪ, j, ʉ, ʉː, ʊ, w/
Low	/a, æ, aj, aw/
Back	/a, aj, aw, ɔj, ow, u, ʉ, ʉː, ʊ, w/
Front	/æ, ɐj, ɛ, i, ɪ/
Tense	/ej, i, ow, ʉ, ʉː/
Diphthong	/aj, aw, ɔj, ej, ow/

Table 3: Assignment of English phones to phonological classes. Symbolic notation is based on MFA's (McAuliffe 2017) *US English lexicon.*

CHAPTER 4

PHONET FOR US ENGLISH

4.1 Introduction

In this chapter we outline the architecture and the retraining process of Phonet for English speech. Firstly, Section 4.2 details the architecture of the model. This section describes the utilized neural networks and how acoustic features, and phonetic features are incorporated in the training process which helps the resulting model to consider both acoustic and phonetic features while estimating the posterior probabilities of the given audio file. This aspect is especially important for the model to estimate the posterior probabilities of the input audio files and it strongly enhances the model's performance. Next, Section 4.3 provides the details of the training corpus like number of words, speaker diversity, and overall hours of audio used for training. Next, Section 4.4 goes more into detail on the preprocessing that needs to be done on the data prior to training and the procedure of retraining. Finally, Section 4.5 discusses the expected output of the model.

4.2 Overview of Phonet's architecture

The Phonet library (Vásquez-Correa et al. 2019) is an advanced phonetic analysis tool utilizing bidirectional recurrent neural networks (RNNs) with gated-recurrent units (GRUs) (Cho

et al. 2014). Its primary function is the estimation of posterior probabilities of phonological features, which can assist linguistic data analysis. The model is trained with both acoustic and phonetic features extracted from the audio files.

Initially, the raw audio signals which are in the .mp3 format[4] are converted into .wav format with a sampling frequency of 16 Hz. Next, each audio signal undergoes preprocessing which helps to enhance uniformity and quality of the signal. In this step, we perform both mean normalization and amplitude normalization. Mean normalization involves subtracting the average value from the signal to keep the signal centered around zero and amplitude normalization is carried out to ensure signal's highest value is below a consistent value. These steps are important as they help us to handle the varying recording volumes and background noise. To capture the dynamic nature of the audio signal over time, the signal is segmented into overlapping frames, specified by a frame size of 25 ms and a time shift (or hop length) of 10 ms. Next, by using the pyfeat.fbank function, we carry out a filter bank analysis on each section of the audio. This is carried out with the help of 33 triangular Mel filters and a setting of the Fourier Transform size (nfft=512) ensuring to capture the frequency-based features. These features provide a granular and detailed acoustic representation. Subsequently, we calculate the energy of each frame, which helps us control the loudness of the signal changes over time. Therefore, along with these spectral characteristics we also capture the dynamic changes in signal intensity. Then, we concatenate both filter bank coefficients and energy values to create a feature vector for each frame. Finally, logarithmic scaling is applied to these features. This step improves the performance of the machine learning models as we normalize the distribution of the data. These

[4] The raw signals do not need to be in .mp3 format.

extracted acoustic features are then saved in a .pickle file for each audio. Figure 5 shows the steps involved in extracting the acoustic features from the signals.

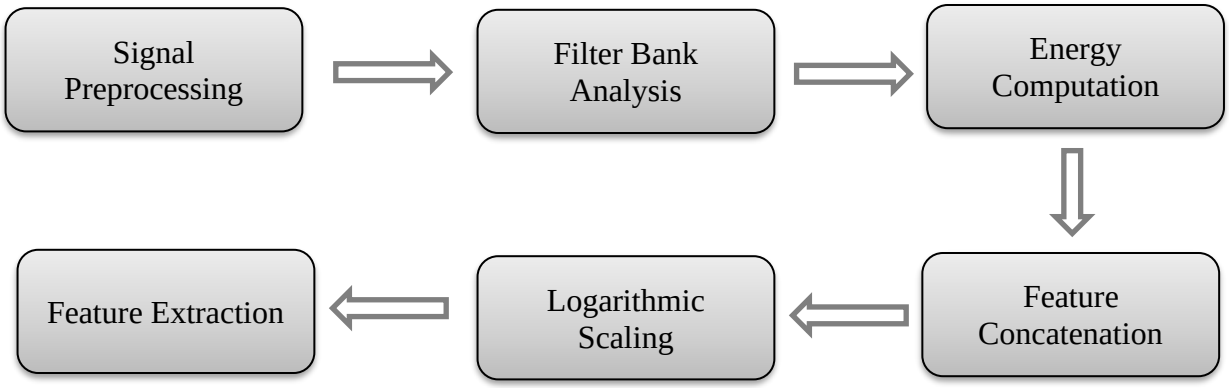


Figure 5: Acoustic feature extraction roadmap for Phonet training – details essential steps for model training.

Along with the acoustic features, we also extract the phonetic features, which are crucial for understanding the dynamics of speech. The phonetic features include the start and end time of a phoneme, and the phonological class of each phoneme. To extract these features accurately, we employed forced alignment. The process of synchronizing the audio signal with respect to its corresponding orthographic transcription is called forced alignment. This technique helps us to align the text, each word, and each phoneme, precisely with the corresponding segment in the audio signal. We used Montreal Forced Aligner (MFA; McAuliffe 2017). The MFA was applied to audio recordings and their associated text files to generate Praat TextGrid files (Boersma & Weenink 2023), containing word- and segment-level time alignments. The TextGrids identify the phones present in the recordings, generated by grapheme-to-phoneme conversion and with assistance from pronunciation variants in MFA's lexicon. Figure 6 illustrates the alignment of each phoneme and word to the audio signal. This figure shows the alignment of the waveform, spectrogram, and Textgrid file tiers for the utterance 'a tray of lighted embers /ə tɹej əv lajtjɪd ɛmbəz/' as produced by MFA. The top panel displays the audio waveform, the middle panel

shows the spectrogram, and the bottom panels present the word-level and phone-level annotations (tiers present in TextGrid file).

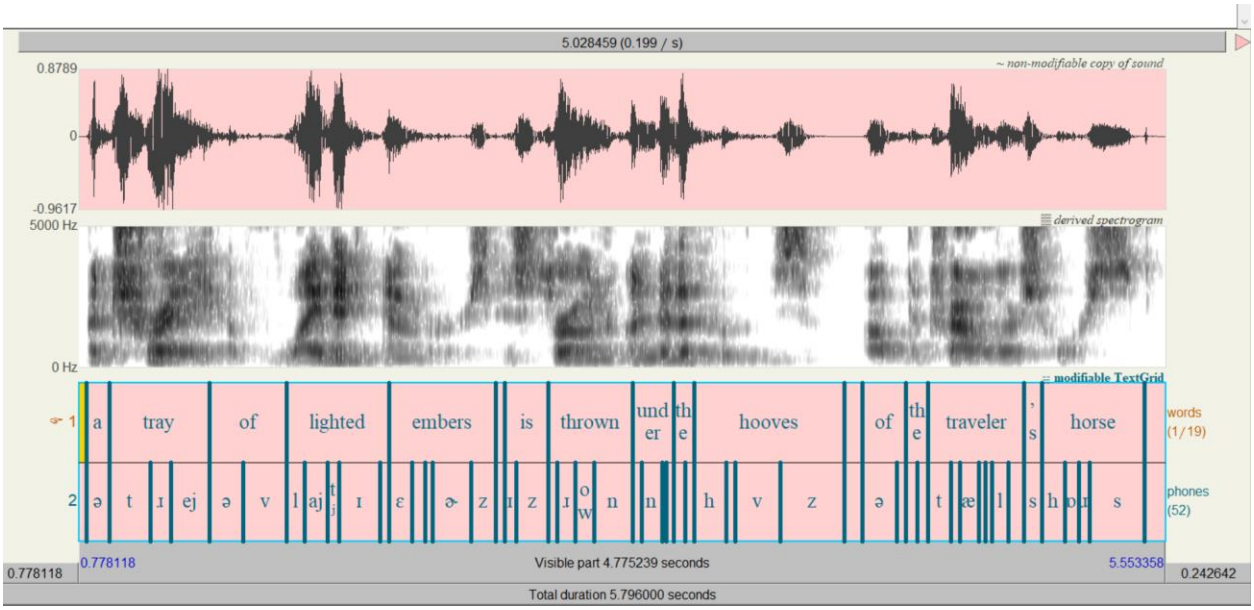


Figure 6: Display of audio file and textgrid file in Praat. The Textgrid file is obtained by MFA for utterance: a tray of lighted embers /ə tɹej əv lajt̪ɹɪd ɛmbəz/ - helps us to understand a textgrid file in which the phonemes and words are aligned with audio.

Using the pre-defined phoneme-to-phonological class mappings (Table 3), phonemes are first encoded to phonological features in the phonetic feature extraction segment. This encodes phonemes in a binary vector according to their phonological classes. The next step is to convert these phonemes into numerical indices so that every phoneme is recorded in a way that can be analyzed computationally. The last step involves processing TextGrid files to extract the maximum and minimum timing and identification of each phoneme. This data is then arranged into structured dictionaries for analysis at the frame and phoneme levels. Figure 7 shows the steps involved in extracting the phonetic features from the signals.

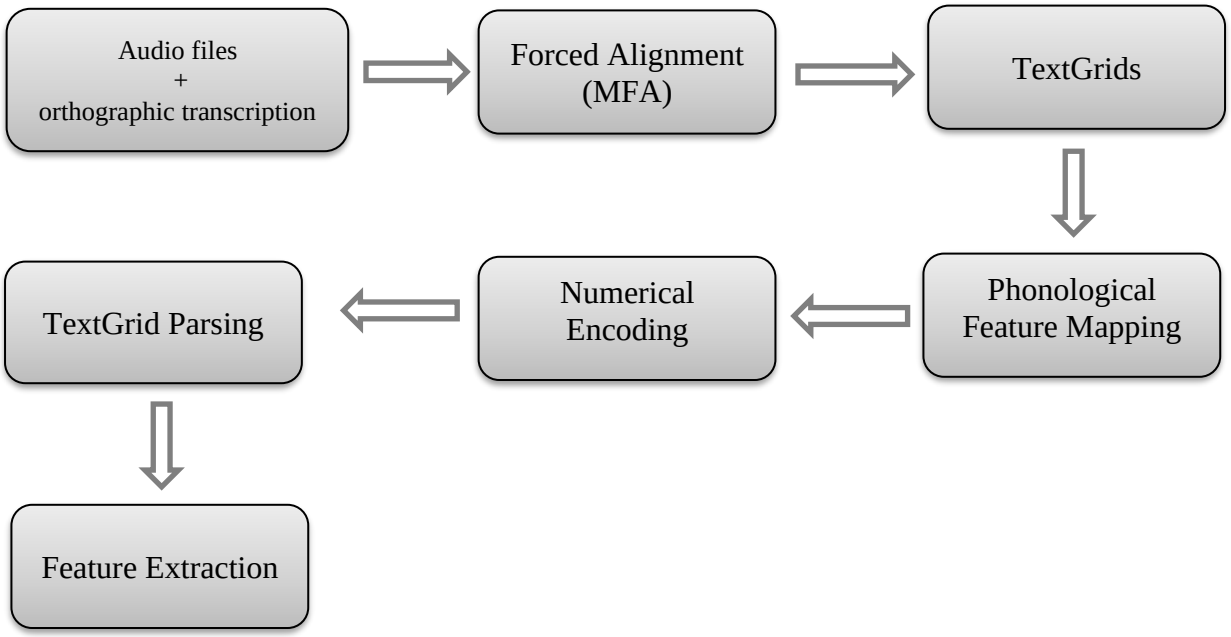


Figure 7: Phonetic features Extraction roadmap – details essential steps for model training.

Both input feature sequences, acoustic features and phonetic features are processed by two layers of bidirectional GRUs. These layers enable the system to integrate temporal context by processing data from both previous (backward) and upcoming (forward) temporal states. This dual-directional analysis enhances the system's ability to capture dynamic temporal dependencies within speech. After the GRU layers, the data passes through a time-distributed, densely connected neural layer. This layer operates as a hidden dense layer, further refining the feature representation and maintaining temporal sequence integrity. The final output generation involves a time-distributed layer equipped with a softmax activation function. This layer categorizes the processed features into distinct phonological classes, effectively translating the complex acoustic patterns into linguistically relevant categories. Additionally, Phonet contains a model for phoneme recognition. Source code of Phonet is accessible through its open-source repository at: https://github.com/jcvasquezc/phonet

We have adapted the Phonet's architecture with significant changes for the retraining procedure. The architecture is used to train a total of 26 neural networks: 25 phonological classes and one layer for the recognition of 49 phonemes. The original Phonet model that was trained on Spanish has 19 neural networks (18 for phonological classes and one layer for 21 phonemes). We utilized weighted categorical cross-entropy for the loss function and adopted the Adam optimizer (Kingma & Ba 2017) for efficient model training, ensuring precise weight adjustments and optimal convergence. The usage of weighted categorical cross-entropy loss function helps to tackle the class imbalance problem. The weight factors for each class are determined from the training set based on the proportion of samples that belong to each class.

4.3 Dataset

Our dataset is a subset of Common Voice 14.0 from the Mozilla Common Voice project (Ardila 2020) which is an open-free source voice database. Our focus was exclusively on the data labeled as "United States English", consisting of 6448 audio files, which consists of around 68,000 words and is approximately 10 hours of spoken English, involving 237 speakers. The data set consists of audio files in .mp3 format and their orthographic transcriptions in .tsv files.

4.4 Data Preprocessing & Model Training

We converted the audio files from their original .mp3 format to .wav format with a sampling frequency of 16kHz, as required for the Phonet's processing needs. Alongside this, we extracted the corresponding orthographic texts from the dataset's provided .tsv files, creating an individual text file for each audio piece. We then divided this data into two sets: training and testing. The training set, comprising 80% of the data, was used to train the model, while the

remaining 20% of the data was used as a validation dataset to test and evaluate the performance of the model (validation dataset = test dataset). We utilized the 'fit_generator' method in Keras and defined our validation dataset by using 'validation_data' parameter. For both sets, we extracted acoustic features and phonetic features from the audio files, saving them into .pickle files. As described in section 4.2, we combined the acoustic features with the time-aligned phone labels (obtained via MFA) to train the Phonet_English architecture.

We trained the model for 50 epochs utilizing an early stopping strategy to avoid overfitting. Early stopping[5] is an optimization technique that prevents the model from learning only the details that are specific to the training data. It monitors the performance of the model on the validation set (we used the test data for validation) and saves the best model if the validation loss is not improved. This approach ensures that the model generalizes well. Implementing an early stopper improves the performance of the model on unseen data. Additionally, it also reduces the computational cost by stopping the training process once the validation accuracy stops improving. Our model converged at the 26th epoch.

Data preprocessing and model training employed Python scripts running on Python version 3.12.1. Our computational setup consisted of a system equipped with an Intel i7 6850K processor, an Nvidia Titan V graphics card, 128GB of DDR4 RAM, and storage comprising a 1TB SSD and a 4TB HDD.

[5] https://www.tensorflow.org/api_docs/python/tf/keras/callbacks/EarlyStopping

4.5 Output of the Model

In phonological analysis, Posterior Probabilities, also known as the phonological probabilities, are log probabilities that show the possibility of phonological classes existing at each time step. These probabilities are the direct outputs form a trained model, like Bidirectional Recurrent Neural Networks in this study, that help us understand the presence of different phonological classes in a given speech signal.

The Phonet_English model is designed to do two things: Phonological Class recognition and Phoneme identification. It is a multi-label classification problem. In the case of Phoneme identification, the model outputs the predicted phoneme at each time stamp. In the case of phonological class recognition, at each time step, this model outputs log-likelihood ratios This value indicates the likelihood of that sound belonging to a phonological class.

Firstly, the acoustic features are extracted from the input audio. Then these features are segmented and normalized, before passing through the model whereby posterior probabilities are obtained for all phonological classes. These posteriors provide a temporal map of the phonological structure present in the audio and represent the likelihood of specific phonological features at each step in time. From these posterior probabilities, the log-likelihood probabilities are computed using (1), Phonological Log-Likelihood Ratio. This transformation converts the posterior probabilities into a feature space that emphasizes the relative likelihoods of phonological classes, enhancing their ability to distinguish between classes. Additionally, to avoid the bounding effect and improve robustness, the PLLR features can be projected into an orthogonal space. By transforming the posterior probabilities into a feature space that emphasizes the relative likelihoods of phonological classes, the ability of the phonological class distinction of the model is enhanced. Posterior probabilities provide direct insights into the

phonological composition of speech. On the other hand, PLLR features further enable differentiation between phonological classes based on relative likelihoods.

$$PLLR = \log_{10}(\frac{P}{1-P}) \quad (1)$$

where,

- PLLR: Phonological Log-Likelihood Ratio.
- P: Posterior probability of a specific phonological class being present at a given time step.
- $1 - P$: Probability of the specific phonological class not being present at the given time step.

CHAPTER 5

RESULTS

5.1 Introduction

This chapter describes the results obtained by running the Phonet_English model. In Section 5.2, we delve into the performance accuracy of the model for classification of phonological classes. Next, Section 5.3 explores the performance of the model with respect to the phoneme recognition task. Finally, Section 5.4 provides in-depth analysis of the posterior probabilities obtained from the Phonet_English model applied to an English audio file. This chapter highlights the efficacy of the trained model and potential areas of improvement in the phonological class identification and phoneme recognition tasks.

5.2 Recognition of Phonological classes

After training Phonet, we obtained evaluation metrics for every phonological class, shown in Table 4. As mentioned in section 4.4, the data is divided into training and test sets. After training, the test set is used to evaluate the model's performance. Since it is a classification problem, we have considered recall, f1-score, precision, and validation accuracy as the evaluation metrics. These metrics were derived using a confusion matrix, discussed in the later sections, that compared the predicted labels against the true labels in our test dataset. The

model's recall values ranged from 80.8% to 98.5%, highest for [flap], [long], [lateral], and [strident] features.

Recall indicates the model's ability to identify true positives of the phonological classes accurately. Recall is the ratio of the number of true instances of a phonological class predicted to be that phonological class divided by the total predicted true instances of the phonological class. It indicates how well the model predicts the true positive instances of a phonological class. Precision is calculated by dividing the number of true instances of a phonological class predicted to be that phonological class and the total number of instances predicted as that phonological class. Precision shows how well the model was performing in terms of correctness for its positive predictions. The F1-score, being the harmonic mean of precision and recall, can be balanced in giving an overview of the performance by the model. Validation accuracy measures the correctly predicted instances to the ratio of the total instances in the validation set in gauging the overall correctness of the model's predictions.

Phonological Class	**Recall**	**F-score**	**Precision**	**Validation Accuracy**
Syllabic	83.10%	83.80%	85.90%	83.40%
Diphthong	88.90%	91.20%	95.20%	89.60%
Consonantal	82.30%	82.60%	83.80%	82.80%
Sonorant	87.50%	87.60%	88.10%	87.40%
Continuant	85.10%	85.10%	85.50%	85.10%
Nasal	89.40%	91.20%	95.00%	89.30%
Voice	86.20%	86.20%	86.30%	86.10%
Labial	83.30%	86.60%	93.40%	84.10%
Round	88.80%	91.80%	96.40%	89.30%
Coronal	81.20%	82.10%	85.00%	81.30%
Distributed	84.90%	88.10%	94.20%	85.10%

Anterior	80.30%	81.50%	85.30%	80.50%
Strident	91.60%	92.40%	94.50%	91.90%
Spread Glottis	92.40%	94.30%	97.40%	92.30%
Lateral	92.20%	94.30%	97.70%	92.90%
Long	96.30%	97.40%	98.90%	96.30%
Dorsal	85.10%	88.40%	94.40%	85.50%
High	80.50%	82.70%	88.00%	80.70%
Low	89.00%	91.30%	95.70%	89.50%
Back	86.80%	89.30%	93.90%	87.20%
Front	82.30%	84.20%	89.10%	82.30%
Tense	87.00%	89.40%	93.90%	87.60%
Rhotic	87.30%	90.00%	95.10%	87.10%
Flap	98.20%	99.00%	99.80%	98.20%
Pause	92.20%	92.10%	92.20%	91.80%

Table 4: Phonological Class Performance Metrics

5.3 Recognition of Phonemes

To evaluate results of the phoneme recognition task, output from Phonet's trained phoneme recognition model was compared against the original MFA-generated input transcriptions (see section 4.2 and Fig. 6), and results are summarized in Table 5. The sounds /ʃ/, /s/, and /ej/ were recognized more accurately than others, showing higher recall values; /ɪ/, with low recall, was often recognized as /i/ instead. The Recall values tell us about the accuracy of the model for recognizing specific phonemes. High recall values indicate the model's effectiveness in differentiating phonemes' spoken forms.

Phoneme	Precision	Recall	F1-score
ɑ	0.330	0.619	0.431
æ	0.492	0.495	0.493
aj	0.521	0.724	0.606
aw	0.335	0.628	0.437
b	0.319	0.517	0.394
ɔj	0.239	0.376	0.292
d	0.212	0.279	0.241
ð	0.077	0.292	0.122
dʒ	0.371	0.573	0.451
ə	0.570	0.185	0.280
ɚ	0.351	0.582	0.438
ej	0.442	0.643	0.524
ɛ	0.356	0.431	0.390
f	0.540	0.622	0.578
g	0.155	0.381	0.220
h	0.389	0.654	0.488
i	0.366	0.575	0.447
ɪ	0.577	0.170	0.263
j	0.286	0.461	0.353
k	0.360	0.570	0.441
k^h	0.422	0.760	0.542
l	0.369	0.572	0.449
ɬ	0.438	0.666	0.528
m	0.407	0.510	0.452
n	0.585	0.449	0.508
ŋ	0.245	0.517	0.332
ow	0.322	0.469	0.382
p	0.341	0.514	0.410
p^h	0.424	0.616	0.502
ɹ	0.650	0.433	0.520
ɾ	0.105	0.262	0.150
s	0.680	0.652	0.666
ʃ	0.573	0.739	0.646
t	0.344	0.280	0.309
t^h	0.429	0.656	0.519
tʃ	0.356	0.470	0.405
ʉ	0.115	0.146	0.129
ʊ	0.110	0.299	0.161
ʉː	0.271	0.344	0.303
v	0.398	0.468	0.430
w	0.461	0.664	0.544
z	0.559	0.540	0.549

ʒ	0.209	0.229	0.219
ʔ	0.076	0.360	0.126
θ	0.156	0.364	0.218

Table 5: Phoneme Recognition Performance for Phonemes

5.4 Sample output

Figure 8 shows a sample output from the Phonet_English model.

time	phoneme	syllabic	diphthong	consonantal	sonorant	continuant	nasal	voice	labial	...	long	dorsal	high	low
1.04	f	0.425841	-0.752595	1.109577	0.444475	1.063063	0.114637	0.587234	1.314847	...	-1.034360	-0.133025	0.500955	-0.803052
1.05	f	0.377311	-0.667004	0.910175	0.413821	0.937248	-0.301298	0.492412	1.205326	...	-1.121672	0.151775	0.236561	-0.402658
1.06	f	0.672355	-0.170637	0.438656	0.930617	0.821242	-0.010050	0.859083	0.746712	...	-2.054916	0.120603	0.277434	0.485777
1.07	ɛ	1.325583	-0.022707	0.178282	1.514029	1.260311	-0.122505	1.472127	0.434847	...	-2.357608	-0.548841	0.446349	1.066883
1.08	ɛ	1.917959	0.189903	-0.100032	1.848210	1.684640	-0.240912	1.749724	0.145926	...	-3.390336	-0.369667	0.393315	1.603711
1.09	ɛ	1.985357	0.280517	-0.202721	1.885612	1.730273	-0.236886	1.859957	0.150720	...	-3.535275	-0.056524	0.381493	1.727518

Figure 8: Sample output of the Phonet_English model outputting predicted phonemes and log-likelihood probabilities at each time stamp of 10ms. This figure demonstrates the model's ability to predict phonemes and their associated phonological posteriors over time.

As discussed in section 3.3, in phonetic theory, [syllabic] and [consonantal] are theoretically opposed: [syllabic] refer to the sounds that are formed from the nucleus of a syllable, which are typically vowels, while [consonantal] refers to sounds that are formed due to the constriction in the vocal tract, which are typically consonants. Figure 9 shows the waveform of an audio file with its corresponding [syllabic] and [consonantal] probability values. From the figure we can say that the model is successful in predicting the phonological classes for the phonemes at each time stamp. The phoneme labels in Figure 10 are the true phonemes obtained by MFA: the blue line indicating [syllabic] is typically high probability during portions of the audio labeled with vowel symbols and having high amplitude, while the orange line indicating [consonantal] has the opposite pattern. Where the two lines have similar values, such as during the sequence /v l/, we note that while these sounds are [consonantal], they are voiced and /l/ is a sonorant, meaning they share phonetic characteristics with [syllabic] vowels.

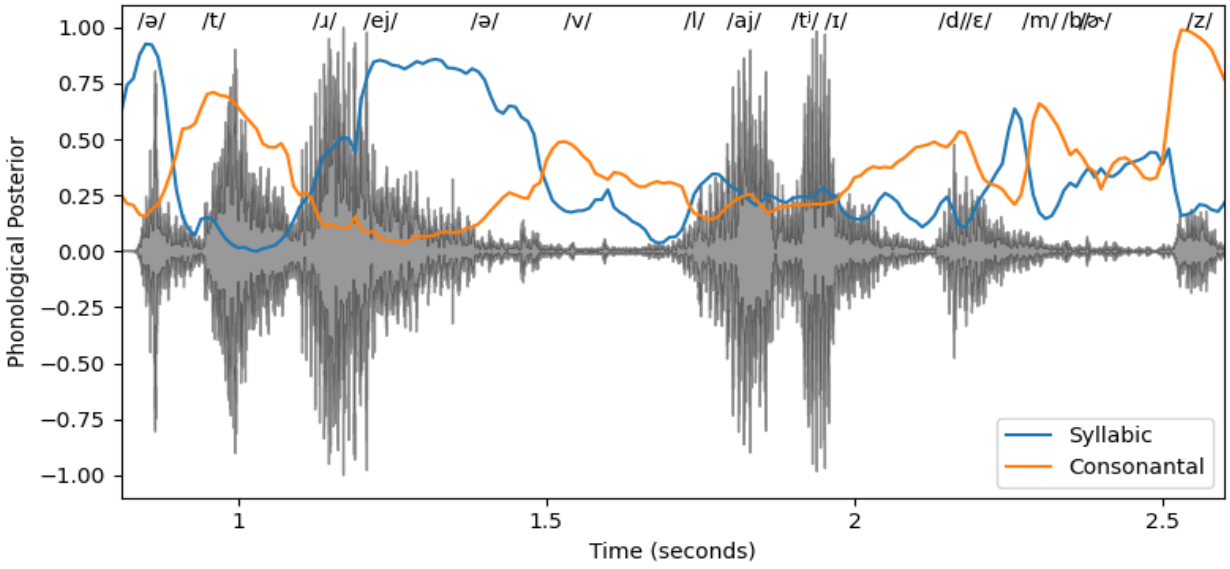


*Figure 9: A waveform of the English utterance "a tray of lighted embers" /ə tɹej əv lajtʲɪd ɛmbəz/, with phonological posterior values for the theoretically opposed distinctive features [**syllabic**] and [**consonantal**]. This figure displays model's ability to distinguish phonological features.*

The figure 10 shows the heat map of the posteriors obtained for the English audio speech "He also taught at the Art Institute of Chicago" /*çiː ɒɬsow tʰaʔ œʔ d̪iː aɻt ɪnstʲɪtʲɯːt əv ʃikʰaːgow*/. The greater presence of blue on the heatmap indicates a lower frequency of that phonological class at the corresponding timestamp.

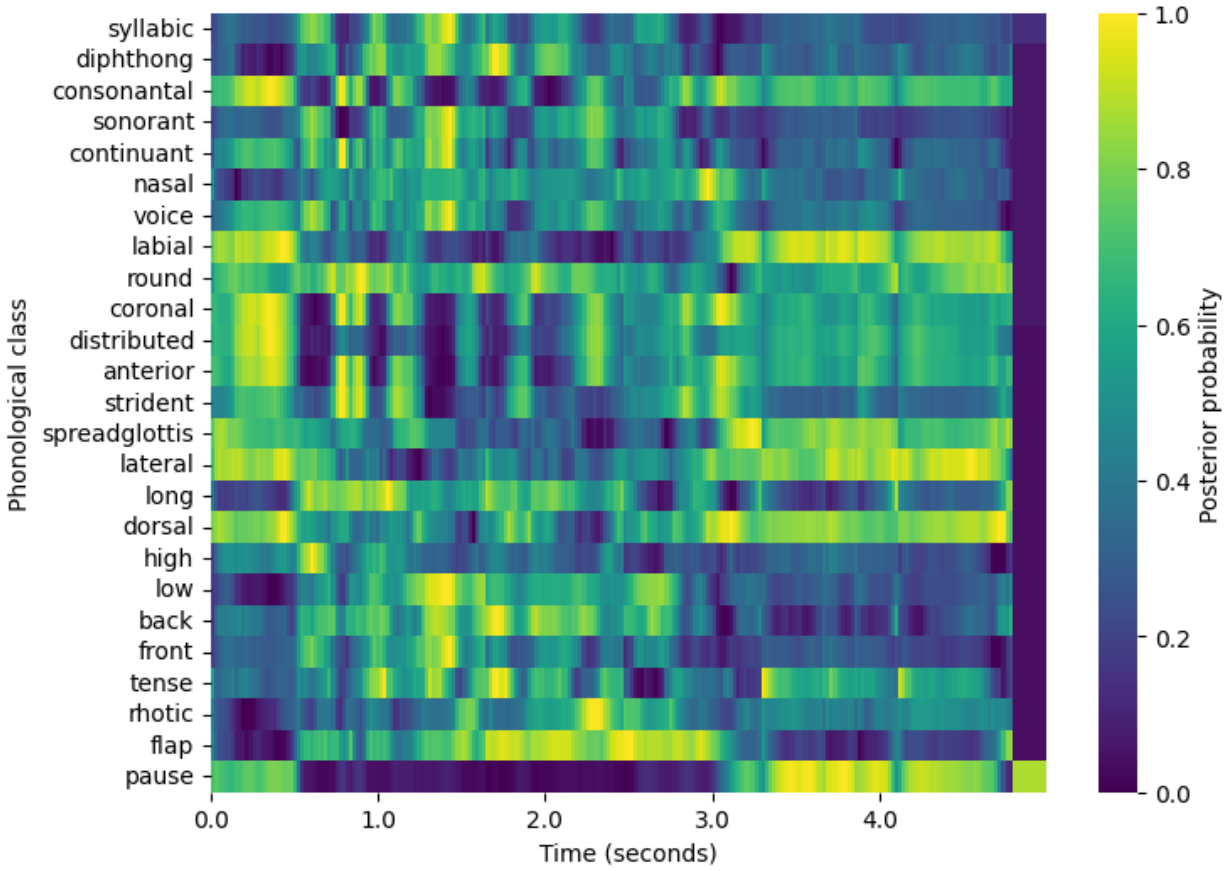


Figure 10: shows the Posteriors obtained for the English sentence "He also taught at the Art Institute of Chicago," demonstrating model's ability to capture fine-grained phonetic details.

CHAPTER 6

EVALUATION & DISCUSSION

6.1 Introduction

In this chapter, we analyze the outcomes of the Phonet_English model. The obtained Phonet_English model is used to obtain the posteriors for the test data audio files. Subsequently, we utilized Praat (Boersma & Weenink 2023) to obtain the MFA (McAuliffe 2017) labelled phonemes, named as phoneme_vox, in 10ms intervals. Then, we combined the results of all audio files into a single file which contains the MFA-labelled phonemes (representing a "ground truth" set of labels), the corresponding phonemes predicted by Phonet_English, and respective phonological posteriors for each distinctive feature. We took the average values of posteriors per individual segment for each audio file, rather than treating each sample separately, to control for the fact that some sounds have longer duration than others and could thus be overrepresented in the data. The averaged results of all audio files are merged into a single file, used for the evaluation of Phonet_English.

In Section 6.2, we analyze the results by violin plots. Section 6.3 provides the AUC-ROC curves for each phonological class and the entire model, that tells the training accuracy of the model. For the plots we have used the true phonemes, obtained via MFA. Finally, in Section 6.4,

we analyze a confusion matrix to obtain further insights into the underlying problems and the errors of the phoneme recognition task.

6.2 Violin Plots

To evaluate Phonet_English's performance, we considered 14 phonemes: the natural class of US English front vowels (Table 6) and the natural class of coronal obstruents (Table 7). The front vowels are noteworthy because they undergo systematic changes in their pronunciation across different US regions (Clopper, Pisoni & de Jong 2005), including variations in how tense vowels /i ej/ and lax vowels /ɪ ɛ æ/ are pronounced. From Renwick & Cassidy (2015), (Dunagan & Renwick 2021) we can say that coronal obstruents consist of a broad category of consonants encompassing stops, fricatives, and affricates and these are further divided by the attributes [voice], [continuant], and [anterior]. These consonants commonly exhibit extensive variation. For example, as heard in the pronunciation of *miss you* as [mɪʃ ju], the anterior series [t d s z] may transform into [tʃ dʒ ʃ ʒ] when it occurs before /j/. In Miller, Brailey-Jones & Renwick (2022) the methods for the automatic detection of palatalization by /t, d/ are described. Just like automatic speech recognition, digital symbolic representation of the speech sounds is also affected by variation. Moreover, in future research we can study dialect variation with the Phonet_English model as it produces the continuously varying phonological posteriors.

	Tense	**Lax**
High	/ i /	/ ɪ /
Mid	/ ej /	/ ɛ /
Low		/ æ /

Table 6: English front vowels.

	Alveolar	**Post-Alveolar**
Plosive	/ t /, / d /	
Fricative	/ s /, / z /	/ ʃ /, / ʒ /
Affricate		/ tʃ /, / dʒ /

Table 7: English coronal obstruents

From Figure 11, that displays a violin plot of posteriors for the feature [tense], we can say that the model is partially successful at differentiating the vowels. We can infer that the model correctly provided the highest [tense] probability for vowels /i ej/, and lowest [tense] probability for vowels /ɛ æ/. However, for lax vowel /ɪ/ the model provides the indication that Phonet may be confusing vowel height with tenseness. On contrast, from the distribution patterns in Figure 12 we can infer that the model is able to identify the [low] feature among the front vowels.

In case of coronal obstruents, we compared the posterior probabilities of [continuant] and [anterior] among the phonemes. In Figure 13, we can infer that the [continuant] feature value for stops /t, d/ is low when compared to the fricatives, which have high probabilities for this feature. For affricates /tʃ dʒ/ the [continuant] feature values are intermediate values since they are a stop-fricative sequence. From Figure 14, the subtle place of articulation differences that separate fronter /t d s z/ from the remaining coronals are similarly displayed via [anterior].

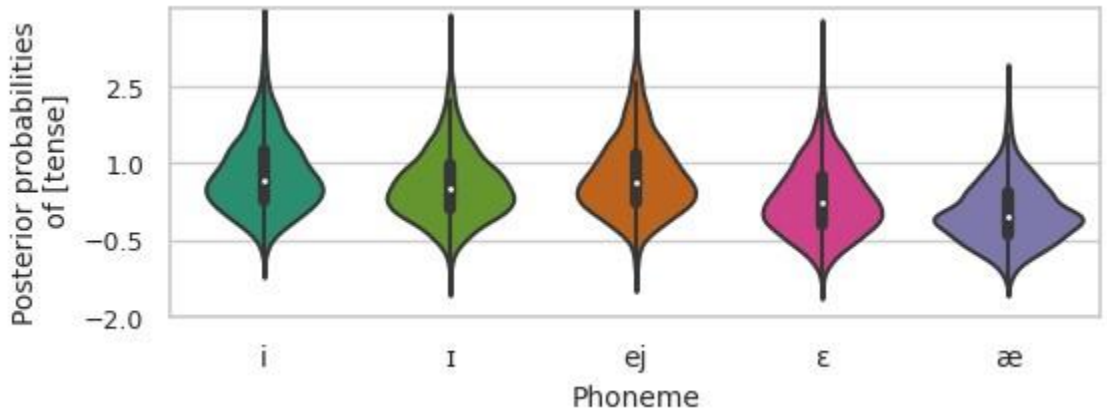


Figure 11: Phonological Log-Likelihood Ratios of [tense] for front vowels, demonstrates the model's effectiveness in identifying [tense] vowels.

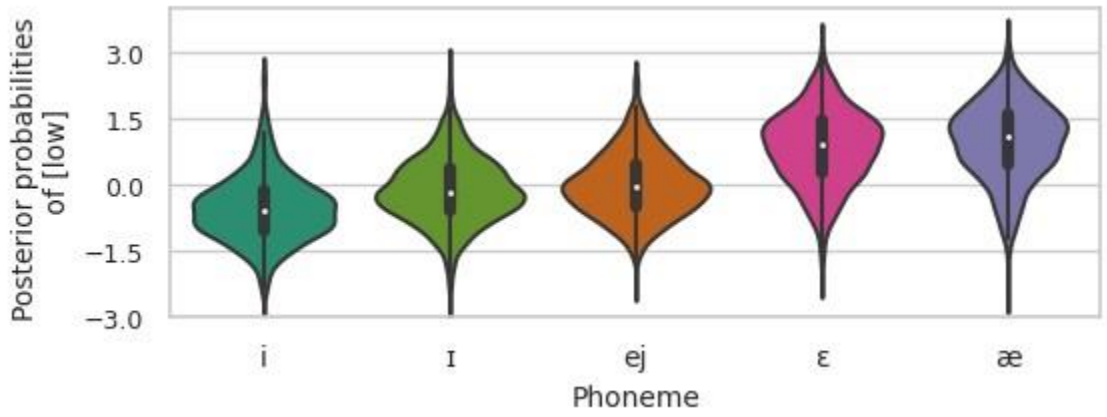


Figure 12: Phonological Log-Likelihood Ratios of [low] for front vowels, showcases the model's ability to identify [low] vowels.

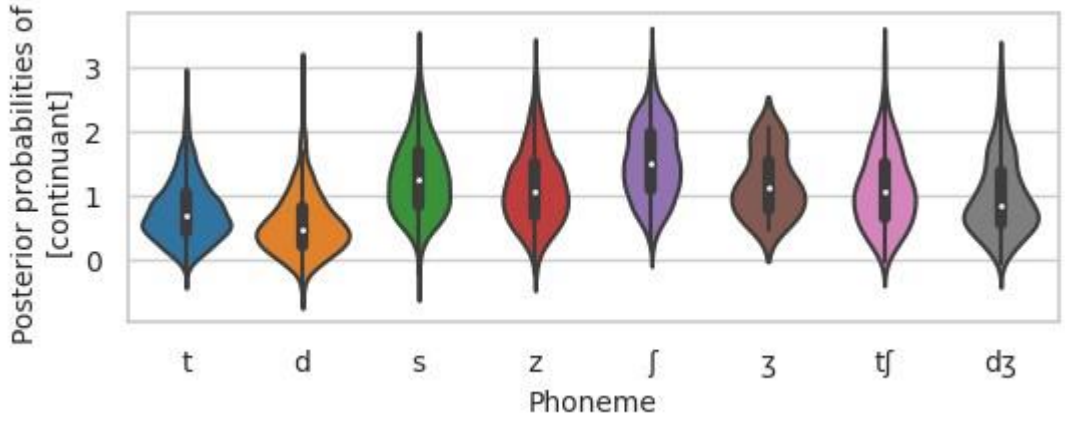


Figure 13: Phonological Log-Likelihood Ratios of [continuant] for coronal obstruents, demonstrates the model's ability to distinguish [continuant] phones.

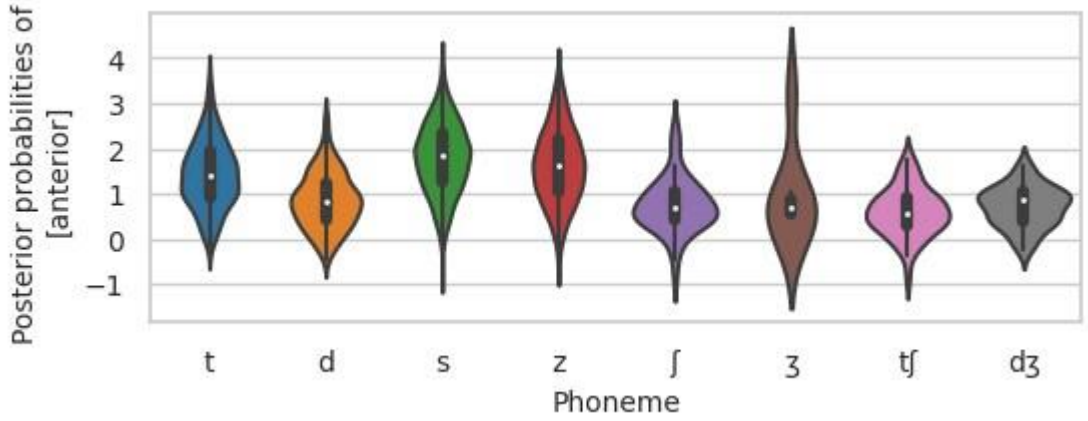


Figure 14: Phonological Log-Likelihood Ratios of [anterior] for coronal obstruents, showcases the model's ability to identify [anterior] phones.

6.3 AUC-ROC Curves

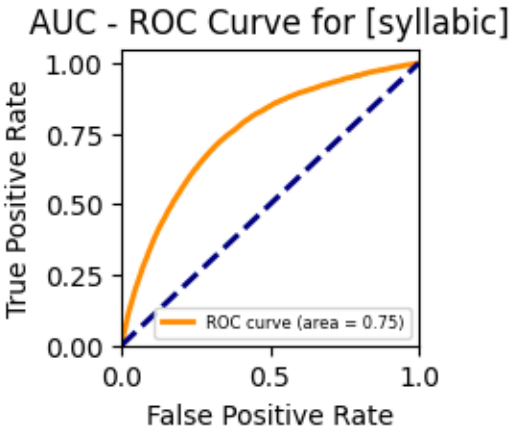
For further analysis we plotted AUC - ROC (Area Under the Curve of the Receiver Operating Characteristic) curves for each phonological class to analyze the performance of the model. For this purpose we normalized the data using Min-Max scaler[6], a technique used to scale/normalize the data so that all values fall between 0 and 1. Later we created the ideal posteriors for the phonemes with respect to phonological class; in case of [strident], for example, the posteriors will be 1 if the phoneme is among /dʒ, s, ʃ, tʃ, z, ʒ/ else the value will be 0. Then, we plotted the AUC-ROC curve for each class between posteriors obtained by Phonet_English and the ideal posteriors for that respective class.

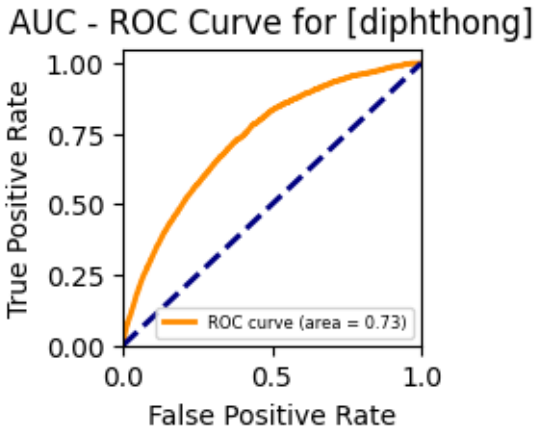
The AUC-ROC curves in Figure 15 show the performance of the Phonet_English model for each phonological class. All phonemes in Table 3 are considered. The ROC curve is plotted with True Positive Rate (TPR, also known as recall or sensitivity) on the y-axis against the False Positive Rate (FPR, or 1 – specificity). The TPR shows the percentage of positive data points that are correctly interpreted as positive with respect to positive data points and the FPR shows the percentage of negative data points that are falsely interpreted as positive with respect to positive data points. In our context, considering [syllabic], TPR illustrates an answer to the question, “of all the syllabic phonemes present, how many did the model correctly identify as [syllabic]?”, FPR provides information regarding “Of all the non-syllabic phonemes, how many did the model incorrectly label as [syllabic]?” and the area under curve (AUC) indicates the percantage of the phonemes that are correctly identified for each phonological class. In the AUC-ROC curve for [syllabic] the AUC is 0.75 which indicates that 75% of the time model is capable of identifying the syllabic and non-syllabic phonemes. The five phonological classes with

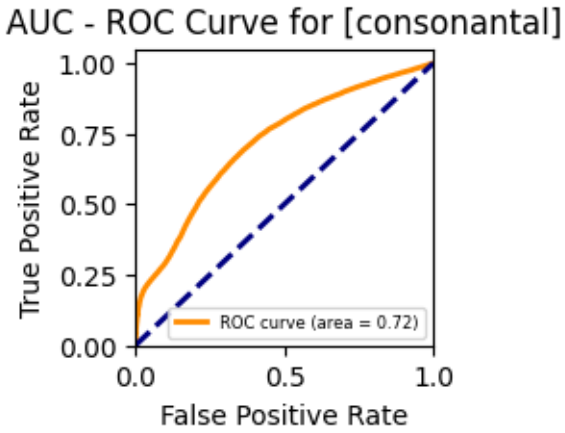
[6] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

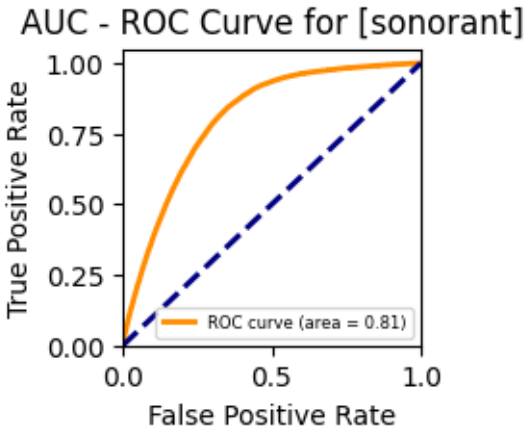
highest AUC values are as follows: [strident], [spreadglottis], [low], [nasal], [sonorant]. Therefore, we can say that the model was able to categorize the phonemes correctly for these classes for most of the cases.

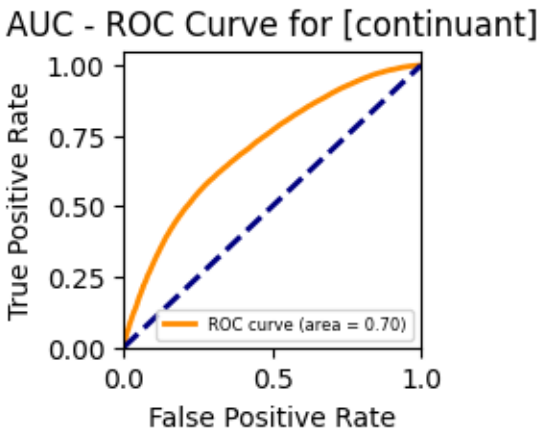
Note that from Table 4 we can infer that the highest Recall values are for [flap], [long] and [lateral], whereas the AUC values are 0.75, 0.74, 0.67. This indicates that the model is good at identifying phonemes of these classes but there is also a tendency of the model to identify phonemes under these classes which, in reality, do not belong these classes.

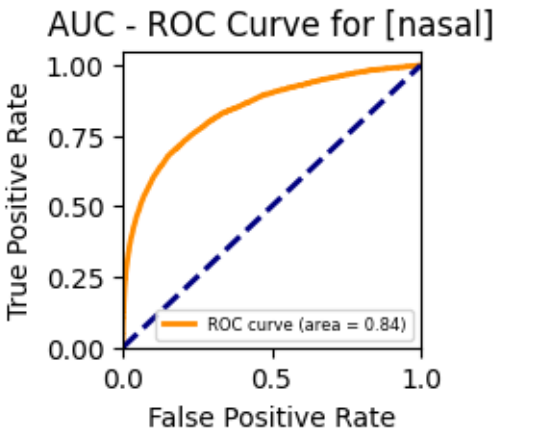


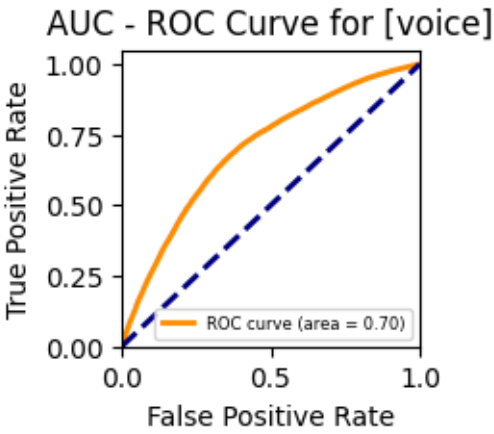


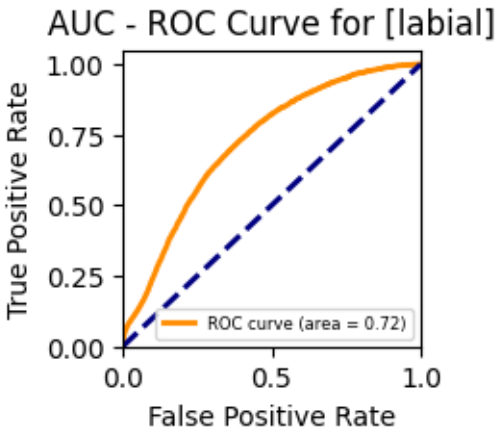


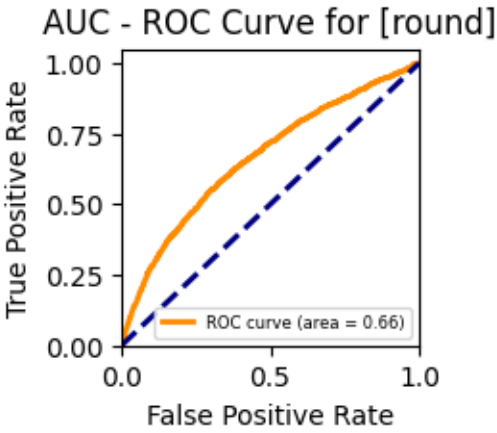


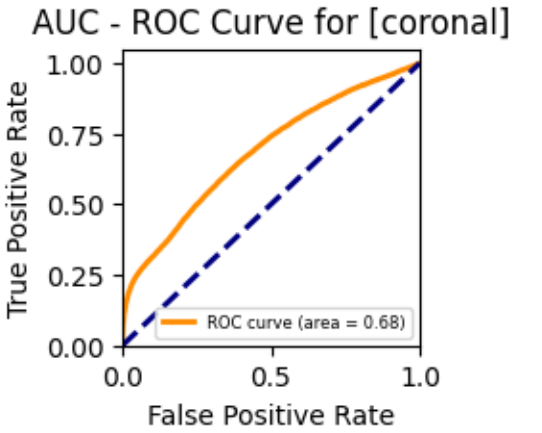

AUC - ROC Curve for [continuant]
True Positive Rate
False Positive Rate
ROC curve (area = 0.70)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0

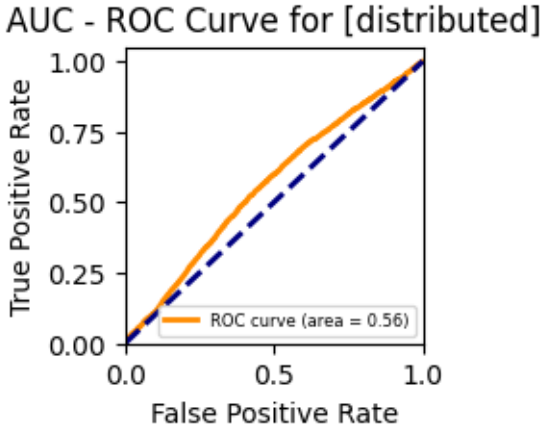

AUC - ROC Curve for [nasal]
True Positive Rate
False Positive Rate
ROC curve (area = 0.84)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0

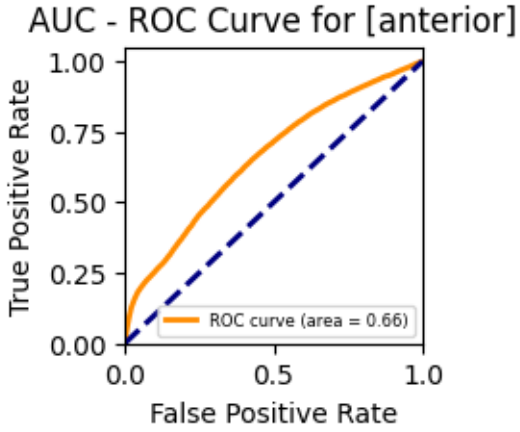

AUC - ROC Curve for [voice]
True Positive Rate
False Positive Rate
ROC curve (area = 0.70)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0

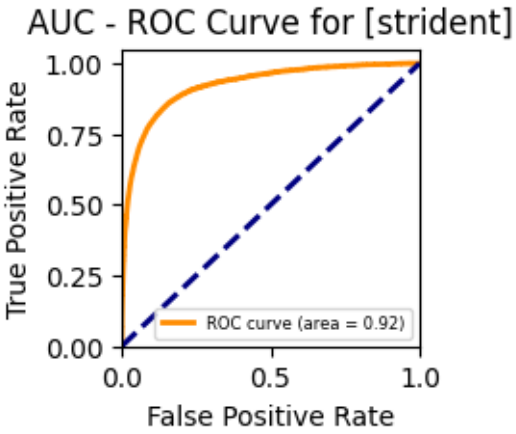

AUC - ROC Curve for [labial]
True Positive Rate
False Positive Rate
ROC curve (area = 0.72)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0

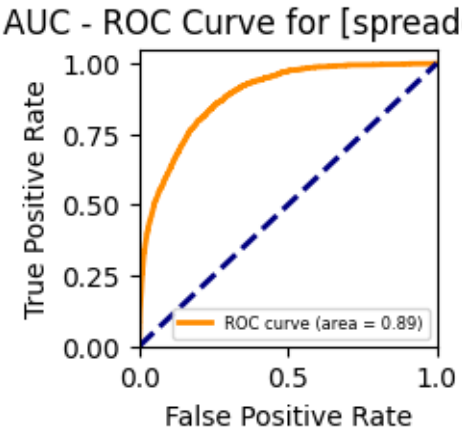

AUC - ROC Curve for [round]
True Positive Rate
False Positive Rate
ROC curve (area = 0.66)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0

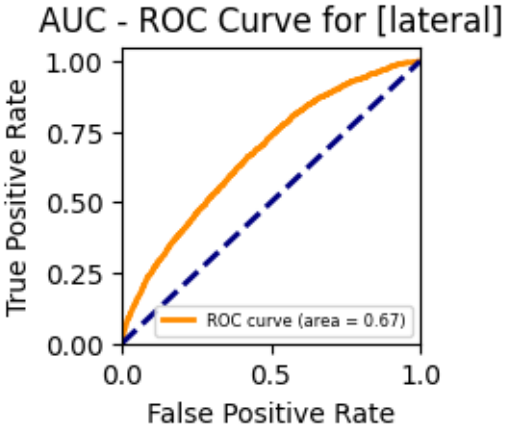

AUC - ROC Curve for [coronal]
True Positive Rate
False Positive Rate
ROC curve (area = 0.68)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0

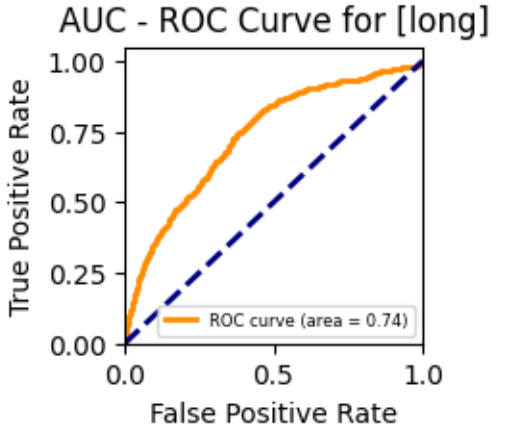

AUC - ROC Curve for [distributed]
True Positive Rate
False Positive Rate
ROC curve (area = 0.56)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0

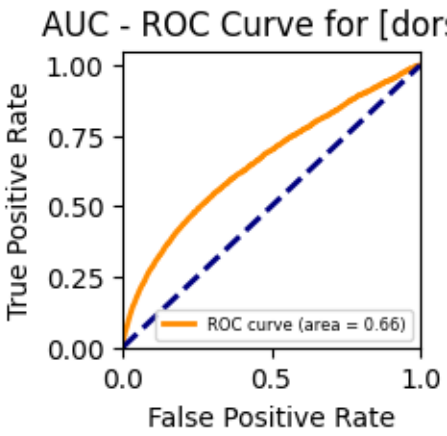

AUC - ROC Curve for [anterior]
True Positive Rate
False Positive Rate
ROC curve (area = 0.66)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0

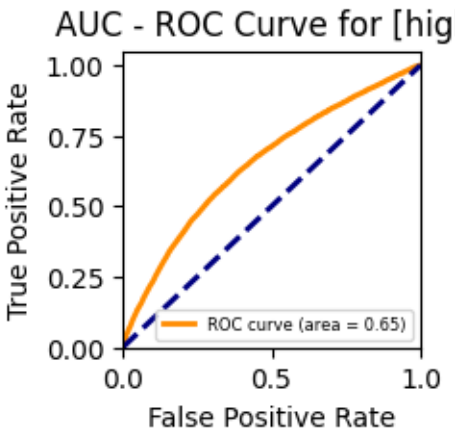

AUC - ROC Curve for [strident]
True Positive Rate
False Positive Rate
ROC curve (area = 0.92)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0

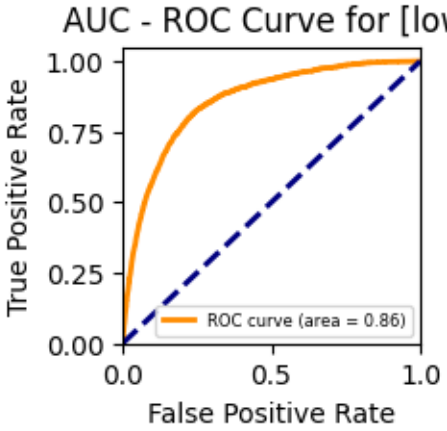

AUC - ROC Curve for [spreadglottis]
True Positive Rate
False Positive Rate
ROC curve (area = 0.89)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0

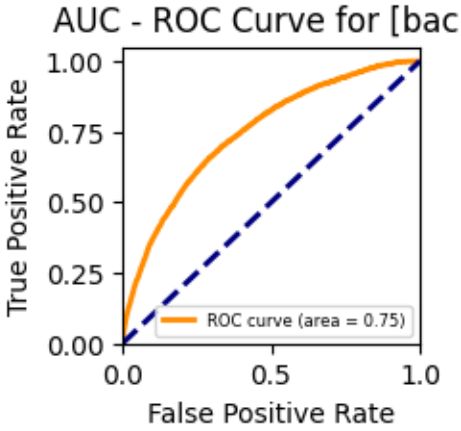

AUC - ROC Curve for [lateral]
True Positive Rate
False Positive Rate
ROC curve (area = 0.67)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0

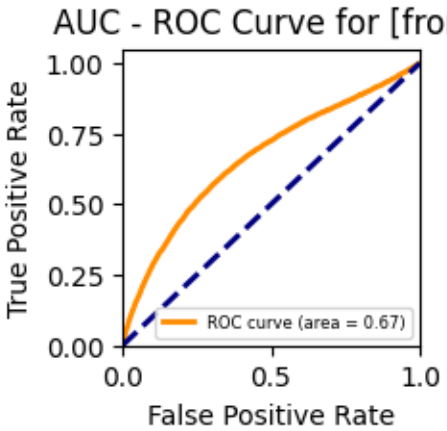

AUC - ROC Curve for [long]
True Positive Rate
False Positive Rate
ROC curve (area = 0.74)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0

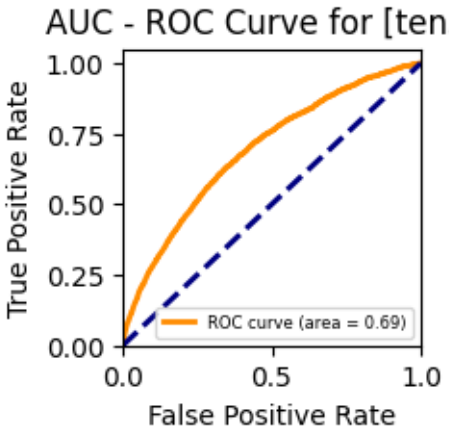

AUC - ROC Curve for [dorsal]
True Positive Rate
False Positive Rate
ROC curve (area = 0.66)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0


AUC - ROC Curve for [high]
True Positive Rate
False Positive Rate
ROC curve (area = 0.65)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0


AUC - ROC Curve for [low]
True Positive Rate
False Positive Rate
ROC curve (area = 0.86)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0


AUC - ROC Curve for [back]
True Positive Rate
False Positive Rate
ROC curve (area = 0.75)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0


AUC - ROC Curve for [front]
True Positive Rate
False Positive Rate
ROC curve (area = 0.67)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0


AUC - ROC Curve for [tense]
True Positive Rate
False Positive Rate
ROC curve (area = 0.69)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0

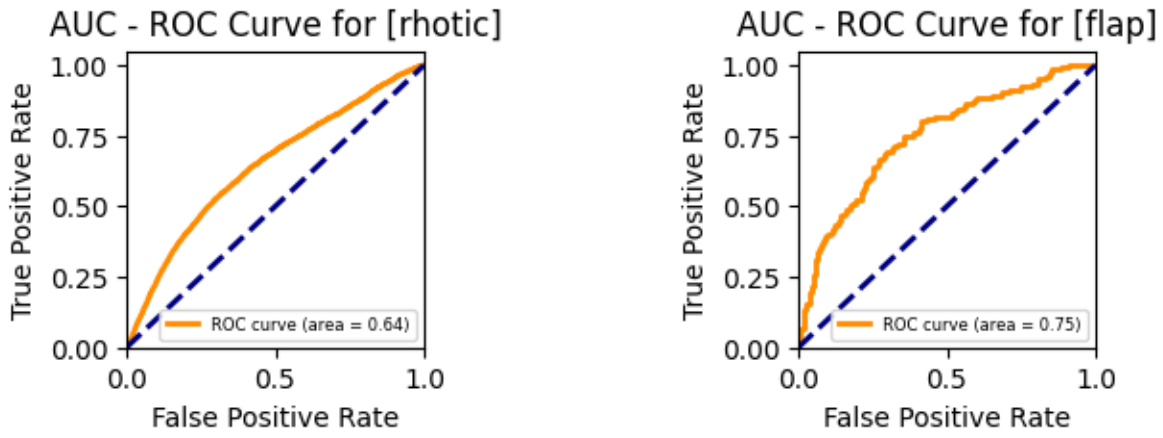


Figure 15: AUC-ROC curves for each phonological class, indicating model performance and classification accuracy for each class.

Figure 16 illustrates the ROC curve of the entire model, obtained using the macro-average technique[7] to provide a comprehensive assessment of the model's performance. For a multi-label classification problem, calculating the ROC curve for the entire model involves aggregating the performance across all labels. There are two main techniques, namely macro-average, and micro-average. The macro-average technique calculates the ROC metrics (TPR and FPR) for each label. On the other hand, the micro-average technique aggregates the True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) across all labels. In the case of the macro-average technique each label equally contributes to the overall metric. This is important especially in the case of imbalanced datasets such as this one. In the case of micro-average technique, more weights are given to the label that has more samples. The macro-average technique is preferred in multi-label classification problems, especially when dealing with imbalanced datasets, as it ensures a more balanced and fair evaluation across all classes. Therefore, for our study we have considered utilizing macro-average technique.

[7] https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

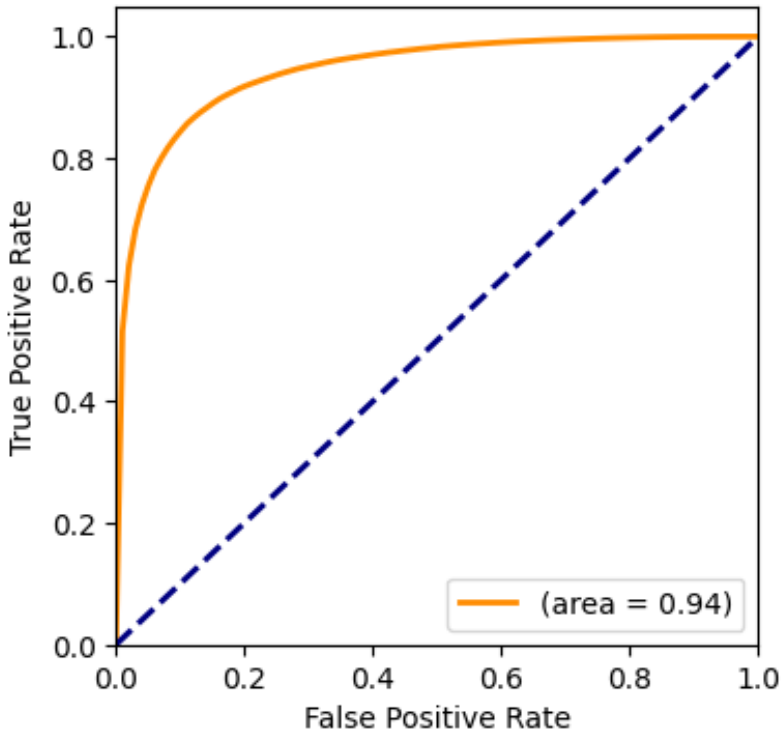


Figure 16: AUC-ROC curves for entire model using macro-average technique, demostrates the overall performance of the model for phonological classification.

These results confirm a high effectiveness of the library Phonet_English library in prediction of the phonological features and the accuracy of differentiation of complex phonological classes. The model's ability to model and predict both segmental and suprasegmental features makes the library an important utility for theoretical and applied studies in phonology. Such results evidence the usefulness of Phonet_English in supporting the conducting of linguistic research in the field of dialectics, the accuracy of phonetic transcriptions, and the improvement in developing automatic speech recognition systems.

6.4 Confusion Matrix Results

Figures 17 and 18 show the normalized confusion matrices for Consonants and Vowels, respectively. The confusion matrices were generated using the complete test dataset, with MFA-labelled phonemes as the true labels and the phonemes predicted by Phonet_English as the predicted labels. We have considered the normalized confusion matrix as it reduces the impact of the imbalances in the dataset. From the confusion matrices we can say that there is a notable confusion between certain phonemes that have similar acoustic properties. For example, the phoneme /ɹ/ is often misclassified as /ɬ/, /s/ is more classified as /ʒ/ and /ʃ/. Similarly, for /i/ and /ʉ/. Therefore, we can say that phonemes that share similar places of articulation or manner of articulation tend to have higher misclassification rates. Phonemes with lower diagonal values indicate areas where the model performs poorly. While certain phonemes are recognized with high accuracy, there is considerable room for improvement, particularly in distinguishing between similar-sounding phonemes.

However , phonemes with high confusions offer tantalizing opportunities for investigation of context-specific phonological variation and dialectal variation. For example, if /ɪ/ is often mistaken for /i/, which suggests that the lax vowel is being pronounced more like the tense vowel. Does this happen in all words, stress patterns, and consonantal contexts, or just in specific cases? or could it be a dialectal difference? To investigate, we could look at instances where /ɪ/ is misclassified and check for patterns in the words and their acoustic properties (like F1, F2, duration) compared to /i/ to find out why this overlap happens. Similar investigations are possible for consonants, such as /ʃ/ vs. /s/ or /ɬ/ vs. /ɹ/.

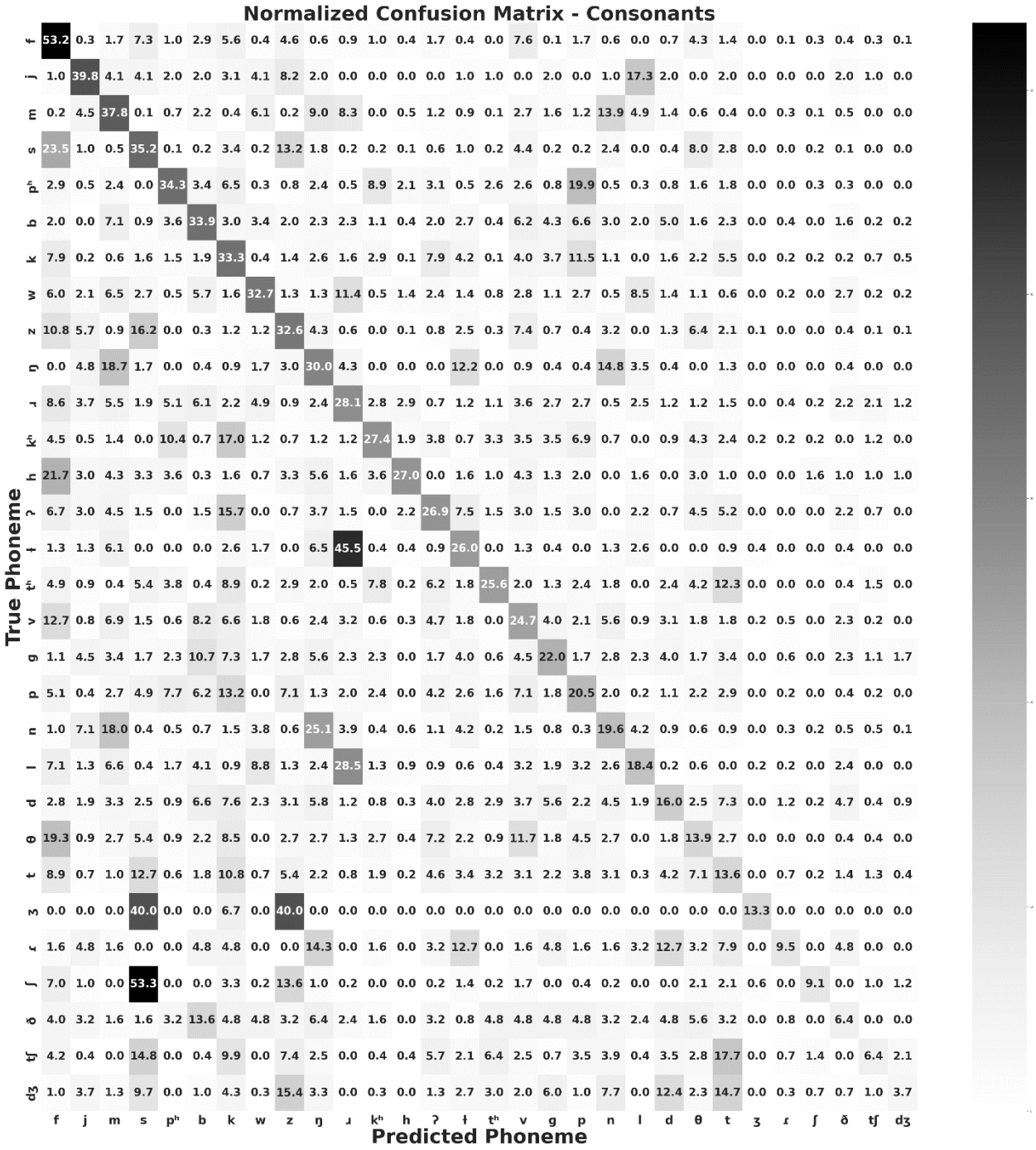


Figure 17: Normalized Confusion matrix – Consonants for phoneme recognition task. The color bar indicates the percentage of predictions per actual Phoneme.

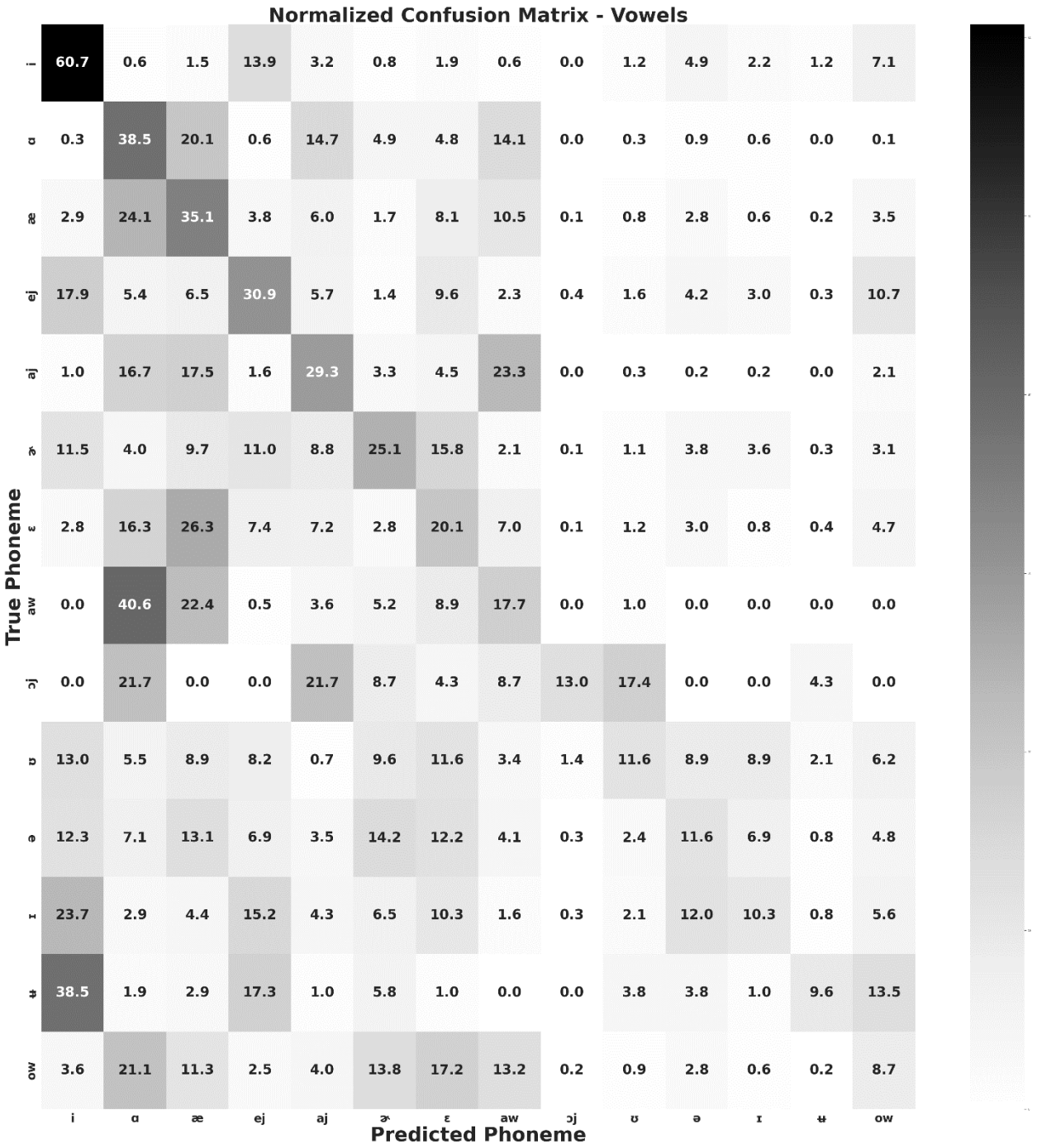


Figure 18: Normalized confusion matrix (in %) – Vowels for phoneme recognition task. The color bar indicates the percentage of predictions per actual Phoneme.

CHAPTER 7

CONCLUSION

In this study, we have re-trained Phonet, a tool which incorporates the architecture of RNN and bidirectional GRU units, on the US English Common Voice 14.0 dataset, enabling it to identify that language's phonological classes in speech samples. This model has good capability to recognize 49 sounds, more than the 24 sounds Phonet was developed with (Vásquez-Correa et al. 2019). We have mainly investigated two things: to see if the distinctive feature values obtained by our model match our expectations for the sounds and testing whether the model is able to predict the phoneme at the given timestamp. The first element was analyzed by violin plots, and AUC-ROC curves and the second was analyzed by confusion matrix. Based on these results we conclude that the model is well-trained and can provide appropriate, accurate posteriors for the given audio. The trained *Phonet_English* model is freely available at https://github.com/dhvl1754/Phonet_English

7.1 Limitations

A closer look at Phonet shows some problems. The major one is the absence of negative natural class specifications, which are essential in phonological analysis. For example, the class [-sonorant] includes stops /p, t, k/ and fricatives /f, s, ʃ/. The class [-voice] includes all voiceless

sounds, and the class [-continuant] distinguishes plosives from fricatives. However, since the framework of Phonet is constructed based on a unary specification system, those classes cannot be explicitly specified with negative specifications. Instead, they should be indirectly specified with either low posterior probabilities or with additional unary features such as [obstruent].

7.2 Future developments

One immediate avenue for future research is testing the generalization of this model to other English dialects and accents or retraining it using other accent-specific corpora. Additionally, research will explore the integration of more nuanced phonological features to investigate spoken-language variation across English varieties. While analyzing the performance of model on predicting the posteriors along with AUC-ROC curve we can also plot Area Under the Curve for Precision-Recall (AUC-PR). This approach will enable a more comprehensive analysis of the model's accuracy and reliability.

References

Anderson, John Mathieson & Colin J. Ewen. 1987. *Principles of Dependency Phonology* (Cambridge Studies in Linguistics). Cambridge University Press.

Ardila, R. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *s.l., s.n*, 4211–4215.

Arora, Vipul, Aditi Lahiri & Henning Reetz. 2015. Digit recognition with phonological features. *The Journal of the Acoustical Society of America* 138(3_Supplement). 1811–1811. https://doi.org/10.1121/1.4933750.

Arora, Vipul, Aditi Lahiri & Henning Reetz. 2018. Phonological feature-based speech recognition system for pronunciation training in non-native language learning. *The Journal of the Acoustical Society of America* 143(1). 98–108. https://doi.org/10.1121/1.5017834.

Boersma, P. & D. Weenink. 2023. Praat: Doing Phonetics by Computer [Computer Program. s.l.:s.n.

Cernak, M. & P.N. Garner. 2016. PhonVoc: A Phonetic and Phonological Vocoding Toolkit. In *s.l., s.n*, 988–992.

Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk & Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Alessandro Moschitti, Bo Pang & Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. Doha, Qatar: Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1179.

Chomsky, N. & M. Halle. 1968. *The Sound Pattern of English* (Studies in English). Harper & Row. https://books.google.com/books?id=y6liAAAAMAAJ.

Clements, G.N. 1985. The geometry of phonological features. *Phonology* 2. 225–252.

Clopper, Cynthia G., David B. Pisoni & Kenneth de Jong. 2005. Acoustic characteristics of the vowel systems of six regional varieties of American English. *The Journal of the Acoustical Society of America*. United States 118(3 Pt 1). 1661–1676. https://doi.org/10.1121/1.2000774.

Cohn, A. 2011. Features, segments, and the sources of phonological primitives. In *Where Do Phonological Features Come From?: Cognitive, physical and developmental bases of distinctive speech categories. s.l.:s.n*, 15–41.

DARPA-ISTO. 1990. The DARPA TIMIT Acoustic-Phonetic.

Davenport, M. & S.J. Hannahs. 2010. *Introducing Phonetics and Phonology*. Hodder Education. https://books.google.com/books?id=G4K1Es0bXuAC.

Delattre, P.C., A.M. Liberman & F.S. Cooper. 1955. Acoustic Loci and Transitional Cues for Consonants. *The Journal of the Acoustical Society of America* 27. 769–773.

Dunagan, D.G. & M.E.L. Renwick. 2021. Word-Boundary Palatalization and Production Planning in UK English. In *Proceedings of Meetings on Acoustics*, vol. 42, 060005.

Fisher, W M, V J Zue & D Bernstein. 1987. An acoustic phonetic data base," presented at the 113th Meet. *Acoust. Soc. Amer*.

Fisher, William M., George R. Doddington & Kathleen M. Goudie-Marshall. 1986. The DARPA Speech Recognition Research Database: Specifications and Status". In *Proceedings of DARPA Workshop on Speech Recognition*, 93–99.

Flemming, E. 2005. Deriving natural classes in phonology. *Lingua* 115. 287–309.

Garofolo, J. S., L. F. Lamel, W. M. Fisher, J. G. Fiscus & D. S. Pallett. 1993. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1.

Graves, A., A.-r Mohamed & G. Hinton. 2013. Speech recognition with deep recurrent neural networks. In *s.l., s.n*, 6645–6649.

Hall, T.A. (ed.). 2001. *Distinctive Feature Theory*. Berlin(Boston: De Gruyter Mouton.

Hall, T.A. 2007. Segmentalfeatures. In P. Lacy (ed.), *The Cambridge Handbook of Phonology*, 311–334. s.l.:Cambridge University Press.

Hillenbrand, James, Laura A. Getty, Michael J. Clark & Kimberlee Wheeler. 1995. Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America* 97(5). 3099–3111. https://doi.org/10.1121/1.411872.

Hulst, H. 2016. Monovalent `Features' in Phonology. *Language and Linguistics Compass* 10. 83–102.

Jakobson, R., G. Fant & M. Halle. 1963. *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. 3rd ed. Cambridge(MA: M.I.T. Press.

Jiao, Y., V. Berisha & J. Liss. 2017. *Interpretable phonological features for clinical applications*. New Orleans, LA, USA: IEEE Press.

Johnson, K. 2005. Speaker Normalization in Speech Perception. In D.B. Pisoni & R.E. Remez (eds.), *The Handbook of Speech Perception*, 364–389. Malden(MA: Blackwell.

Kaye, Jonathan, Jean Lowenstamm & Jean-Roger Vergnaud. 1985. The internal structure of phonological elements: a theory of charm and government. *Phonology Yearbook* 2(1). 305–328. https://doi.org/10.1017/S0952675700000476.

Kaye, Jonathan, Jean Lowenstamm & Jean-Roger Vergnaud. 1990. Constituent structure and government in phonology. *Phonology* 7(1). 193–231. https://doi.org/10.1017/S0952675700001184.

Kim, C.-W. 1970. A Theory of Aspiration. *Phonetica* 107–116.

Kingma, D.P. & J. Ba. 2017. Adam: A Method for Stochastic Optimization. s.l.:s.n.

Krishnamurti, B. 1955. The History of Vowel-Length in Telugu Verbal Bases. *Journal of the American Oriental Society* 237–252.

Lamel, L F R H & S Kassel. 1986. Speech database development: Design and analysis of the acoustic-phonetic corpus. In L S Workshop (ed.), *Proc. DARPA Speech*, 100–109.

Laver, J. 2017. Linguistic Phonetics. In *The Handbook of Linguistics*, 159–184. s.l.:John Wiley & Sons, Ltd.

Lee, Honglak, Peter Pham, Yan Largman & Andrew Ng. 2009. Unsupervised feature learning for audio classification using convolutional deep belief networks. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams & A. Culotta (eds.), *Advances in Neural Information Processing Systems*, vol. 22. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2009/file/a113c1ecd3cace2237256f4c712f61b5-Paper.pdf.

Lee, Jong-Hwan, Ho-Young Jung, Te-Won Lee & Soo-Young Lee. 2000. Speech feature extraction using independent component analysis. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, vol. 3, 1631–1634 vol.3. https://doi.org/10.1109/ICASSP.2000.862023.

Lee, K.-F.a H.H.-W. 1989. Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 1641–1648.

Leung, K.K.W.a J.A.a W.Y.a S.J.A. 2016. Acoustic characteristics of clearly spoken English tense and lax vowelsa. *The Journal of the Acoustical Society of America* 07, Volume 140. 45–58.

Lisker, L. & A.S. Abramson. 1964. A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. *Word* 20. 384–422.

McAuliffe, M. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *s.l., s.n*, 498–502.

Mielke, J. 2004. The Emergence of Distinctive Features. s.l.:s.n.

Mielke, J. 2008. *The Emergence of Distinctive Features*. s.l.:Oxford: Oxford University Press.

Miller, S.E., A. Brailey-Jones & M.E.L. Renwick. 2022. Postlexical Palatalization of /d/ across Word Boundaries in UK English. In *Proceedings of Meetings on Acoustics*, vol. 50, 060005.

Mitkov, Ruslan, Steven Bird, Jeffrey Heinz, Kemal Oflazer, Patrick Hanks, Ronald Kaplan, David Beaver, et al. 2014. The Oxford Handbook of Computational Linguistics. https://doi.org/10.1093/oxfordhb/9780199573691.001.0001.

Paul, D.B. & J.M. Baker. 1992. The design for the Wall Street Journal-based CSR corpus. In *Workshop on Speech and Natural Language*, 357–362. Stroudsburg, PA, USA.

Peterson, G.E. & H.L. Barney. 1952. Control Methods in a Study of the Vowels. *Journal of the Acoustical Society of America* 24. 175–184.

Renwick, M.E.L. & C.N. Cassidy. 2015. *Detecting Palatalization in Spontaneous Spoken English*. Pittsburgh: Acoustical Society of America.

Sagey, E.H. 1990. The Representation of Features in Non-linear Phonology: The Articulator Node. s.l.:Garland.

Sailor, H.B. & H.A. Patil. 2016. *Filterbank learning using Convolutional Restricted Boltzmann Machine for speech recognition*. Shanghai: IEEE Press.

Skandera, P.a B.P. 2011. A Manual of English Phonetics and Phonology: Twelve Lessons with an Integrated Course in Phonetic Transcription. s.l.:Narr.

Stevens, K.N. & S.E. Blumstein. 1978. Invariant Cues for Place of Articulation in Stop Consonants. *The Journal of the Acoustical Society of America* 64. 1358–1368.

Stevens, K.N. & S.E. Blumstein. 1981. The Search for Invariant Acoustic Correlates of Phonetic Features. In P.D. Eimas & J.L. Miller (eds.), *Perspectives on the Study of Speech*, 1–38. Hillsdale(NJ: Lawrence Erlbaum Associates.

Sylak-Glassman, J. 2014. Deriving Natural Classes: The Phonology and Typology of Post-Velar Consonants. *ProQuest Dissertations and Theses* 241.

Tang, Kevin, Ratree Wayland, Fenqi Wang, Sophia Vellozzi, Rahul Sengupta & Lori Altmann. 2023. From sonority hierarchy to posterior probability as a measure of lenition: The case of Spanish stops. *The Journal of the Acoustical Society of America* 153(2). 1191–1203. https://doi.org/10.1121/10.0017247.

Trubetzkoy, N.S. 1939. *Principles of Phonology*. Berkely: University of California Press.

Vásquez-Correa, J.C., P. Klumpp, J.R. Orozco-Arroyave & E. Nöth. 2019. Phonet: A Tool Based on Gated Recurrent Neural Networks to Extract Phonological Posteriors from Speech. In *s.l.*, *s.n*, 549–553.

Wayland, Ratree, Kevin Tang, Fenqi Wang, Sophia Vellozzi & Rahul Sengupta. 2023. Quantitative Acoustic versus Deep Learning Metrics of Lenition. *Languages* 8(2). https://doi.org/10.3390/languages8020098.

Zsiga, E.C. 2013. The Sounds of Language: An Introduction to Phonetics and Phonology. s.l.:Wiley-Blackwell.

Zuraiq, Wael & Jie Zhang. 2006. Phonological Assimilation in Urban Jordanian Arabic. *Kansas Working Papers in Linguistics* 28. 33–64. https://doi.org/10.17161/KWPL.1808.1229.