



UNIVERSITY OF THE PHILIPPINES

Grapheme-to-phoneme and automatic speech recognition techniques
for low-resource phonetic transcription of Tagalog, Cebuano, and Hiligaynon

Angelina Aquino

Bachelor of Science in Electronics and Communications Engineering

Joshua Lijandro Tsang

Bachelor of Science in Electronics and Communications Engineering

Undergraduate Project Advisers:

Crisron Rudolf Lucas, M.S. EE

Franz de Leon, Ph.D.

Electrical and Electronics Engineering Institute

University of the Philippines Diliman

Undergraduate Project Examiner:

Neil Irwin Bernardo, M.S. EE

Electrical and Electronics Engineering Institute

University of the Philippines Diliman

Date of Submission

24 May 2019

Permission is given for the following people to have access to this thesis:

Circle one or more concerns:	<input checked="" type="checkbox"/> I	<input checked="" type="checkbox"/> P	<input checked="" type="checkbox"/> C	
Available to the general public				Yes <input checked="" type="checkbox"/> No
Available only after consultation with author/thesis adviser				<input checked="" type="checkbox"/> Yes/No
Available only to those bound by confidentiality agreement				<input checked="" type="checkbox"/> Yes/No

Students' signature/s:

aa Aquino *Joshua Lijandro Tsang*

Signature/s of undergraduate project advisers:

Crisron Rudolf Lucas *Franz de Leon*

University Permission Page

We hereby grant the University of the Philippines non-exclusive worldwide, royalty-free license to reproduce, publish, and public distribute copies of this work in any form subject to the provisions of applicable laws, the provisions of the UP IPR policy and any contractual obligations, as well as more specific permission marking on the Title Page.

Specifically we grant the following rights to the University:

- to upload a copy of the work in the theses database of the college/school/institute/department and in any other databases available on the public internet;
- to publish the work in the college/school/institute/department journal, both in print and electronic or digital format and online; and
- to give open access to above-mentioned work, thus allowing “fair-use” of the work in accordance with the provisions of the Intellectual Property Code of the Philippines (Republic Act No. 8293), especially for teaching, scholarly, and research purposes.


Angelina Aquino

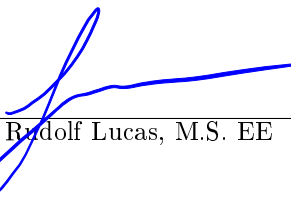
24 May 2019
Date


Joshua Lijandro Tsang

24 May 2019
Date

Approval Sheet

In partial fulfillment of the requirements for the degree of Bachelor of Science in Electronics and Communications Engineering, this project entitled "Grapheme-to-phoneme and automatic speech recognition techniques for low-resource phonetic transcription of Tagalog, Cebuano, and Hiligaynon", prepared and submitted by Angelina Aquino and Joshua Lijandro Tsang, is hereby recommended for approval.



Crisron Rudolf Lucas, M.S. EE
Adviser

24 May 2019

Date

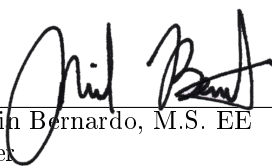


Franz de Leon, Ph.D.
Adviser

24 May 2019

Date

Accepted in partial fulfillment of the requirements for the degree of Bachelor of Science in Electronics and Communications Engineering.



Neil Irwin Bernardo, M.S. EE
Examiner

24 May 2019

Date

John Richard E. Hizon, Ph.D.
Director, Electrical and Electronics Engineering Institute

Date

Grapheme-to-phoneme and automatic speech recognition techniques
for low-resource phonetic transcription of Tagalog, Cebuano, and Hiligaynon

Undergraduate Project

by

Angelina Aspra Aquino

2014-06313

B.S. Electronics and Communications Engineering

Joshua Lijandro Lugay Tsang

2014-06489

B.S. Electronics and Communications Engineering

Advisers:

Crisron Rudolf G. Lucas

Franz A. de Leon

University of the Philippines, Diliman

May 2019

Abstract

Grapheme-to-phoneme and automatic speech recognition techniques
for low-resource phonetic transcription of Tagalog, Cebuano, and Hiligaynon

Philippine linguists are tasked with documenting over 170 indigenous languages. A key part of this documentation is the phonetic transcription of recorded speech, which is typically done by hand, and is often expensive and time-consuming. Automated phonetic transcription (APT) systems provide a faster and cheaper alternative to manual transcription, but no such system has yet been developed for most Philippine languages. In this project, we implement and evaluate three APT methods—grapheme-to-phoneme (G2P) conversion, automatic speech recognition (ASR), and adaptive alignment—for transcription of small speech corpora in Tagalog, Cebuano, and Hiligaynon. We show that the G2P, adaptive, and select ASR models perform at par with human transcribers while reducing total time and costs by at least 75%. These systems serve as a competent baseline for future developments in APT for Philippine languages, which are expected to facilitate further research in Philippine linguistics and speech technology.

Contents

List of Figures	iii
List of Tables	iv
1 Introduction	1
1.1 Overview	1
1.2 Flow and Organization	2
2 Related Work	3
2.1 Automated phone recognition	3
2.1.1 Generative modeling	4
2.1.2 Discriminative classification (neural networks)	5
2.2 Grapheme-to-phoneme conversion	6
2.2.1 G2P using weighted finite state transducers	7
2.2.2 G2P using recurrent neural networks	8
2.3 Adaptive phonetic transcription	8
2.3.1 Rule-based G2P with ASR for phonetic transcription of Basque	9
2.3.2 Phonetic transcription of Austrian German using forced alignment and burst detection	10
2.4 Automated speech transcription for Philippine languages	11
3 Problem Statement and Objectives	13
3.1 Problem Statement	13
3.2 Objectives	13
3.3 Scope and Limitation	14
4 Methodology	15
4.1 Transcription and preprocessing of speech corpora	16
4.2 Data allocation for the train and test sets	18
4.3 Training of a grapheme-to-phoneme conversion model	18
4.4 Training of the automatic speech recognition system	19
4.5 Alignment of G2P and ASR transcriptions	20
4.6 Substitution of allophones	21
4.7 Evaluation metrics	22
5 Results and Discussion	23

5.1	ASR performance optimization	23
5.2	Transcription accuracy	24
5.3	Runtime comparison	26
6	Conclusion and Recommendations	27
6.1	Conclusion	27
6.2	Future Work	28
6.2.1	Increasing speech duration and language model order for ASR	28
6.2.2	Cross-lingual effectiveness	28
6.2.3	Alternative phoneme categorization for Visayan languages	28
	Bibliography	29

List of Figures

2.1	Segmented waveform, phonetic transcription, and spectrogram for the word <i>triangle</i>	4
2.2	Visual representation of a hidden Markov model	5
2.3	Perceptron node (from skymind.ai)	5
2.4	An example DNN with an input layer, three hidden layers, and an output layer [1]	6
2.5	WFST alignment lattice for the word <i>RIGHT</i> :/ɪ aɪ t/. Adapted from [2]	8
2.6	Flowchart of an adaptive phonetic transcriber for Basque [3]	10
4.1	Block diagram of the automated phonetic transcription system	15
4.2	Consonants in the phone inventory of Tagalog, Cebuano, and Hiligaynon	16
4.3	Vowels and diphthongs in the phone inventory of Tagalog, Cebuano, and Hiligaynon	16
4.4	G2P training and testing instance using Multilingual Set 1	19
4.5	Sequence alignment with Needleman-Wunsch for $s = \pm 1, W_1 = 1$	21
5.1	Percent accuracies of APT systems versus manual transcriptions	25

List of Tables

2.1	Phoneme recognition systems for the Filipino Speech Corpus	11
4.1	Expanded phone inventory used for transcription of speech corpora	17
4.2	Summary of transcribed data	17
4.3	Data allocation for training and testing	18
4.4	Phonological Rules in Target Languages	21
4.5	Hardware specifications for ASR runtime evaluation	22
5.1	ASR performance using Set 1 data allocation	23
5.2	ASR performance using Set 2 data allocation	24
5.3	G2P performance for single-language and multilingual datasets	24
5.4	ASR performance for single-language and multilingual datasets	24
5.5	Adaptive system performance for single-language and multilingual datasets	25
5.6	System runtimes for single-language and multilingual datasets	26
5.7	Average runtimes across varying dataset durations	26

Chapter 1

Introduction

1.1 Overview

Language is an important aspect of everyday life. It is a means of communication and enables people to express their thoughts and feelings. It is a tool which conveys traditions and values related to group identity and holds the key in understanding the basis of human interaction.

Linguists analyze the function and formation of language. This includes multiple forms such as written, verbal, and signed languages [4]. Proper documentation of languages allows research in linguistics to advance, and a fundamental part of this documentation is the phonetic transcription, where the distinct sounds that make up a spoken utterance, be it a word, phrase or a sentence, are written down. This data enables sociolinguists to understand how pronunciations vary across different groups and enables historical linguists to understand how pronunciations evolve over time. Additionally, to create modern systems for speech synthesis and recognition, understanding the intricacies of speech is the first step.

Phonetic transcriptions of recorded speech data are necessary in Philippine linguistics for documenting, characterizing, and conducting further research on native languages. The development of a system capable of phonetic transcription would facilitate further advances in linguistic analysis and speech processing for Philippine languages.

In the Philippines, the UP Department of Linguistics and DLSU Center for Language Technologies are among the few academic research centers which are actively documenting and analyzing Philippine languages [5]. Traditionally, transcriptions are performed manually by trained phoneticians, but this process requires considerable time and expense. With an expert, 30 minutes of clean audio would take 1.5 to 2 hours to transcribe. Considering a rate of Php 1000 per hour, this undertaking is quite expensive. Shifting this workload into computers, this tedious process can

be completed, with the same accuracy, in significantly less time, effectively saving money.

Automated phonetic transcribers have been widely developed in recent years for foreign languages [3] [6]. However, automated transcription for Philippine languages has thus far only been designed for word-level and phoneme-level recognition, and cannot resolve speech at the phone-level. Hence, there is a need to fill this gap and develop an automated phone transcription system for Philippine languages.

1.2 Flow and Organization

Chapter 1 discusses the background and motivation behind the study and gives an overview of what this project aims to accomplish. Chapter 2 contains the review of related literature and introduces some concepts to better understand the project. Chapter 3 contains the problem statement, objectives, scope and limitations of the project. Details on the methodology can be found in Chapter 4 followed by the proposed schedule and current progress in Chapter 6.

Chapter 2

Related Work

Phonetic transcription has been a recurrent problem in speech and language processing research for over three decades [7]. This chapter presents an overview of prevalent historical and contemporary techniques for automated phonetic transcription of acoustic and text-based speech data. It also contains a discussion on prior development of automated speech transcription and segmentation systems for Philippine languages.

2.1 Automated phone recognition

Every spoken language is composed of *phones*, or speech sounds. Individual phones are produced by positioning the human vocal tract in a certain way—[k] is pronounced by bringing the back of the tongue close to the soft palate then releasing a buildup of air pressure in a short burst; [v] is pronounced by touching the bottom lip to the front row of teeth and forcing air through this point of contact while the vocal folds are vibrating. The possible points of articulation in the human vocal tract are limited; as such, there exists only a finite set of phones that humans can produce, and each spoken language uses a subset thereof [8].

When a phone is uttered, its distinct articulation results in a specific set of peak frequencies, called *formants*. In Figure 2.1, we can see how the formants in the word *triangle*, denoted by the dark horizontal bands in its spectrogram, vary across phones. These formant frequencies have roughly the same value for a given phone, regardless of the speaker. Computers can therefore analyze the formants of a speech signal, and use them to identify the phones being said. This process of automated phone recognition is the foundation for speech processing, allowing higher levels of speech recognition (e.g. identifying words in an utterance) and synthesis (i.e. artificially summing frequencies over time to produce a sequence of phones).

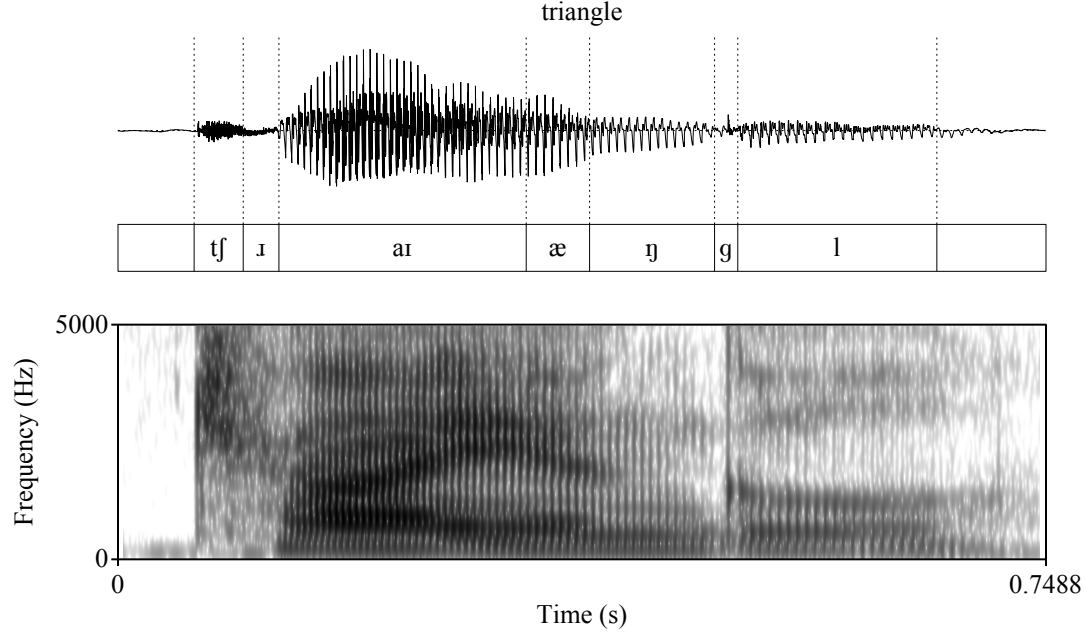


Figure 2.1: Segmented waveform, phonetic transcription, and spectrogram for the word *triangle*

Phone recognition techniques can be broadly classified into two groups: generative modeling, and discriminative classification [7]. In the following subsections, we discuss these two approaches, and give examples of notable speech recognition systems under each category.

2.1.1 Generative modeling

When we process a speech signal to determine its formants and other features, we are making a series of acoustic observations $\mathbf{A} = \{a_1, a_2, \dots, a_n\}$ about the signal. We then want to determine the most likely sequence of phones $\hat{\mathbf{P}}\mathbf{h} = \{ph_1, ph_2, \dots, ph_n\}$ that fit these observations (with $\mathbf{P}\mathbf{h}$ being the actual sequence of phones that were uttered). This can be calculated using Bayes' theorem:

$$\hat{\mathbf{P}}\mathbf{h} = \operatorname{argmax}_{Ph} P(\mathbf{P}\mathbf{h}|\mathbf{A}) = \operatorname{argmax}_{Ph} \frac{P(\mathbf{P}\mathbf{h}) \cdot P(\mathbf{A}|\mathbf{P}\mathbf{h})}{P(\mathbf{A})}$$

For a given signal, our acoustic observations, and consequently the probability $P(\mathbf{A})$, are constant. We can therefore simplify the above equation as:

$$\hat{\mathbf{P}}\mathbf{h} = \operatorname{argmax}_{Ph} P(\mathbf{P}\mathbf{h}) \cdot P(\mathbf{A}|\mathbf{P}\mathbf{h})$$

The optimal phone sequence $\hat{\mathbf{P}}\mathbf{h}$ can thus be determined by learning or *generating* an acoustic model $P(\mathbf{A}|\mathbf{P}\mathbf{h})$, and a language model $P(\mathbf{P}\mathbf{h})$ [9].

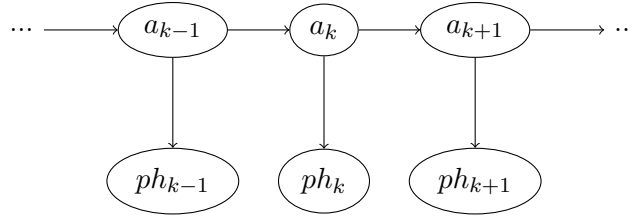


Figure 2.2: Visual representation of a hidden Markov model

The hidden Markov model (HMM), visualized in Figure 2.2, is one such generative model that is widely used and highly successful in speech recognition [7] [9]. Since recorded speech (i.e. our set of acoustic observations \mathbf{A}) takes discrete samples of continuous audio data over time, it is inherently a sequence of distinct states, with each current state a_k possibly switching to the next state a_{k+1} with a certain transition probability. However, in generating an acoustic model $P(\mathbf{A}|\mathbf{Ph})$, these states a_k are treated as “hidden” or unknown. Instead, we know the correct sequence of phones \mathbf{Ph} , and we can say that each physical sound a_k is interpreted as a phone ph_k with a certain emission probability (represented in Figure 2.2 by the arrows from a_k to ph_k). The acoustic model can therefore be learned by estimating the transition and emission probabilities from a training dataset with known phonetic transcriptions.

2.1.2 Discriminative classification (neural networks)

Over the last few years, advances in both machine learning algorithms and computer hardware have led to more efficient methods for training *deep neural networks* (DNNs) [10]. A DNN is a conventional *multilayer perceptron* (MLP) with multiple hidden layers [1]. These layers are made up of *nodes* which consist of four parts: input values, *weights* and *bias*, a *weighted sum* and an *activation function*. These are shown in Figure 2.3.

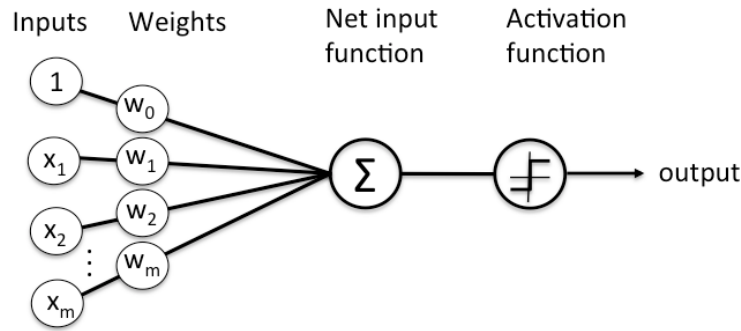


Figure 2.3: Perceptron node (from skymind.ai)

The inputs are multiplied by their given weights and bias. These values are then added together to form a weighted sum. The weighted sum is applied to the correct activation function, to determine the progress of the signal through the network, and ultimately, a classification is made. Following a *discriminative system*, an input produces a single output. A sample DNN is depicted in Figure 2.4.

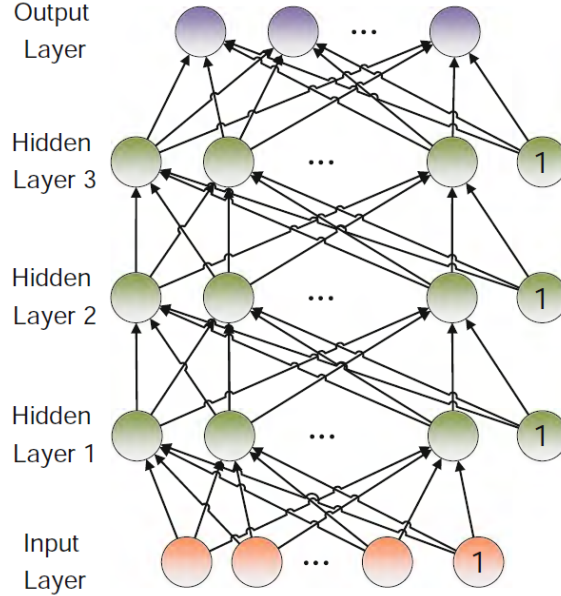


Figure 2.4: An example DNN with an input layer, three hidden layers, and an output layer [1]

DNNs are the latest hot topic in speech recognition. Kaldi is an open-source toolkit for speech recognition with codebases for deep neural nets [11].

As opposed to other open-source systems and kits, Kaldi is accessible and wide spread in the community. Other open-source recognition toolkits that meet these criteria include HTK and Sphinx. In evaluating these three systems, Kaldi outperforms the two. The time spent to set up, prepare, run and optimize the toolkits was the most for HTK and the least for Kaldi. At the same time, Kaldi provides training and decoding pipelines including the most advanced techniques out of the box [12].

2.2 Grapheme-to-phoneme conversion

Many of the world’s languages employ writing systems, which are visual representations of spoken language. In these writing systems, a *grapheme* is the smallest significant unit of text. In

English, for example, the set of graphemes includes twenty-six alphabetical letters, ten numerical digits, and several punctuation marks. These graphemes can then be used to form larger written structures, such as words, phrases, or sentences, as dictated by the *orthography*, or rules of standard writing and spelling, of a language.

An alphabet is a type of writing system wherein each grapheme represents a *phoneme* (a sound, or set of sounds, considered to be distinctive in a given language). However, because a language’s pronunciation evolves much faster than its orthography, graphemes and phonemes in an alphabet tend to lose their one-to-one correspondence over time. This can be observed in Philippine languages: written text is largely expressed using the Latin alphabet, which originally evolved to represent Latin sounds, and lacks graphemes for common Philippine speech sounds such as the velar nasal [ŋ] (the sound of *ng* in *panga*) or the glottal stop [ʔ] (the pause pronounced between the first two syllables in *paalam*).

Due to these discrepancies between text and speech, transcribing the sounds in a given text is not as straightforward as a one-to-one conversion. *Grapheme-to-phoneme (G2P) conversion* techniques have thus been developed to address this problem, the simplest of which include dictionary lookup and rule-based transcription [13]. A dictionary lookup system stores a database of words and their corresponding pronunciations, then refers to this database for conversion; however, a comprehensive dictionary is tedious to encode and store, and may be ineffective when a language has multiple pronunciations for similarly-spelled words (such as Tagalog *paso*—“scald, burn” and *paso*—“earthen pot”). On the other hand, in a rule-based transcription system, a set of conversion rules are formulated by analyzing the language in question, but this requires proficiency in linguistics; moreover, languages are highly complex and dynamic, and cannot be readily characterized by a finite set of rules.

In recent years, data-driven approaches to G2P have gained increasing popularity and accuracy. These take a limited number of example words and pronunciations, and use them to generate probable pronunciations for novel words. Two of the most prominent data-driven models are discussed in the next subsections.

2.2.1 G2P using weighted finite state transducers

When a G2P system is trained using a dataset of word-pronunciation pairs, it aims to learn the optimum alignment of one or more graphemes to one or more phonemes. This alignment may be modeled using a *weighted finite state transducer*—a lattice of states and transitions with a defined starting and ending state, wherein each transition represents the mapping of a grapheme subsequence to a phoneme subsequence. In training the WFST model, each transition is assigned

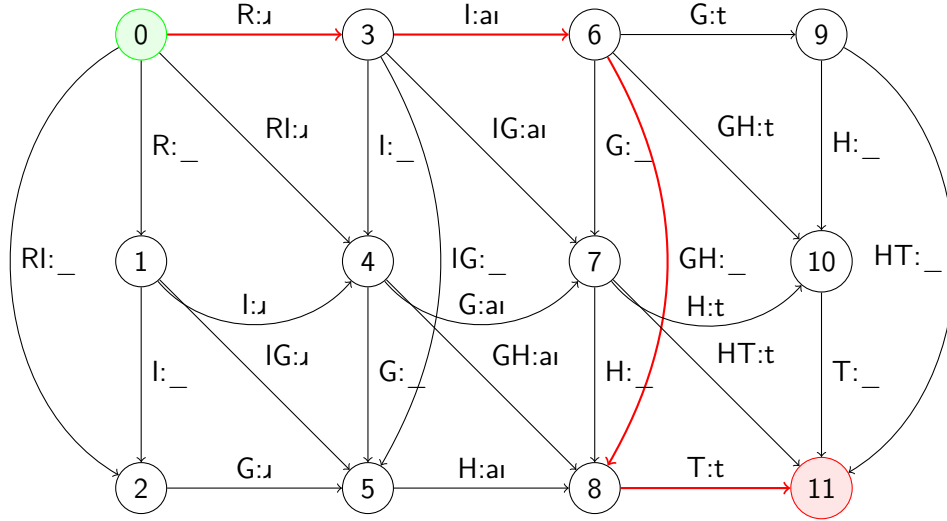


Figure 2.5: WFST alignment lattice for the word *RIGHT*:/ɹ aɪ t/. Adapted from [2]

a certain weight, such that the shortest path from the starting state to the ending state gives the optimal G2P alignment. An example of a WFST alignment lattice for the word *RIGHT* and its grapheme sequence /ɹ aɪ t/ is shown in Figure 2.5, with the correct alignment denoted by the red arrows.

Phonetisaurus is a system which uses WFSTs for G2P conversion [14]. It is an open-source toolkit which is considered state-of-the-art [15] and has been shown to outperform other G2P models in large-vocabulary conversion tasks [16].

2.2.2 G2P using recurrent neural networks

Increasing developments have also been made towards the use of recurrent neural networks (RNNs) in G2P. These are a class of neural network which are said to have temporal memory, since they take into consideration not only the inputs fed into the network, but also the network’s prior outputs. One such RNN-based G2P converter has been shown to match and even outperform state-of-the-art systems like Phonetisaurus when applied to low-resource languages including Tagalog [15].

2.3 Adaptive phonetic transcription

In the previous sections, we discussed how a language has a finite set of phones (speech sounds) and phonemes (groups of distinctive sounds). A sound, or group thereof, is said to be *phonemic* in a language if it distinguishes one word from another. On the other hand, two sounds

are said to be *allophones* of a single phoneme if they can be used interchangeably in a word without changing its meaning [17]. Sounds which are phonemic in one language may be allophones in another—for example, the phones [r] and [l] are phonemic in Tagalog (as seen by the distinctive meanings of *turo*—“to teach” and *tulo*—“drip”), but are both understood as the single /r/ phoneme in Japanese.

In recent years, *adaptive phonetic transcription* systems have been developed which use the concepts of phonemes and allophones to their advantage. These systems first process the text equivalent of a given utterance using G2P conversion to transcribe the sequence of phonemes in the utterance. Phone-level ASR is then used to determine what specific sound is being said when variations in pronunciation are permissible in the language (i.e. when a phoneme has several allophones). This adaptive process aims to remedy the shortcomings of either method if used alone—G2P techniques are often unable to capture the specific phonetic realizations of a given word, while phone recognition systems are prone to insertions, deletions, or substitutions which can be easily corrected by referring to an utterance’s text transcription.

The following subsections outline several recent projects on adaptive phonetic transcription, and discuss the relevance of each project’s context and methodology in relation to the goal of phonetic transcription for Philippine languages.

2.3.1 Rule-based G2P with ASR for phonetic transcription of Basque

In 2015, an adaptive phonetic transcriber was developed for the Basque language [3]. Basque is a minority language spoken by inhabitants of the Basque Country on the Spanish-French border. Though a standard pronunciation of Basque exists, the language’s geographical situation has given rise to several dialects and non-native speech variants with a wide range of phonetic realizations. Accurate phonetic transcription of Basque therefore requires taking these varying pronunciations into account; hence, a twofold adaptive procedure was implemented as seen in Figure 2.6.

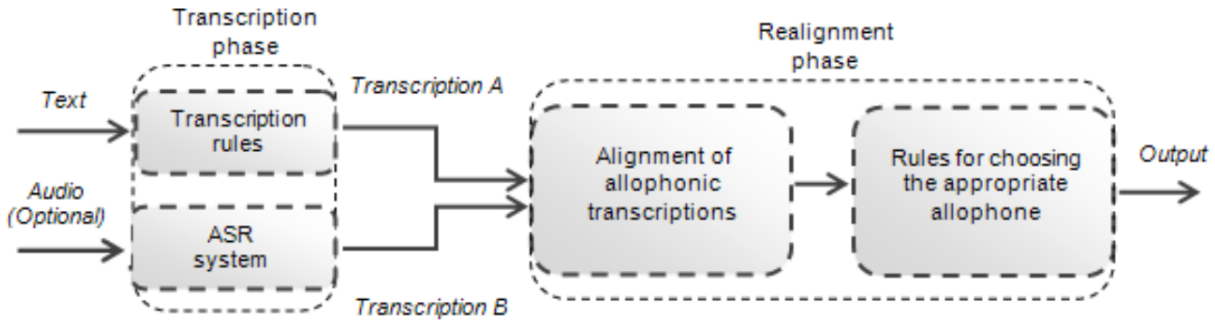


Figure 2.6: Flowchart of an adaptive phonetic transcriber for Basque [3]

A rule-based grapheme-to-phoneme transcriber was trained using data from multiple news media to produce text-based transcriptions in standard Basque pronunciation. An automatic speech recognition (ASR) system was also trained using audio data to produce phone-level transcriptions of speech recordings. If both text and speech data were available for a given utterance, the system was designed to align the transcriptions of each, then make appropriate substitutions in the text transcription based on a list of possible allophones.

This adaptive system yielded a correctness of 79.92% upon testing, a 12.20% improvement over the G2P-only transcription of the same dataset at 67.72% correctness. Since several dialects also exist for Philippine languages, using an adaptive process to transcribe them may produce a similar increase in transcription accuracy over text-only or speech-only methods.

2.3.2 Phonetic transcription of Austrian German using forced alignment and burst detection

The first large-scale corpus for read and spontaneous speech in Austrian German was created in 2014 [6]. For it to be usable in speech technology and linguistic analysis, this corpus needed to be transcribed and segmented at the phone level. Although segmentation tools did exist for standard German, these were unable to capture several phonetic variations found in conversational Austrian German. The creators of the corpus therefore found it necessary to develop a new system which could accurately transcribe and segment spontaneous Austrian German speech.

First, a pronunciation dictionary was created by passing all words in the corpus into a G2P converter to generate pronunciations in standard German. Next, a set of transformation rules was applied to each word in the dictionary, to account for phonetic variation in spontaneous Germany-based and Austrian German; additional pronunciation variants for the 150 most frequent words were also annotated. HTK, an HMM-based speech recognition toolkit, was then used along

with the dictionary to transcribe the corpus and segment each utterance (i.e. specify the duration of every phone through timestamps). Finally, an algorithm was developed to differentiate between plosives (specifically, the phones [p], [b], [t], [d], [k], and [g]) which are pronounced with bursts versus those without bursts, as such observations were deemed relevant in further phonetic analysis of Austrian German.

With spontaneous German speech, phoneticians experienced an average discrepancy of 21.2%. The system reports a PER of 18.5% which falls within the range. This shows that adaptive method is effective even for spontaneous speech, which is faster and inherently harder to transcribe due to its unpredictability. Additionally, the systems could be made to classify more specific phone features, highlighting the need for phonetic transcription.

2.4 Automated speech transcription for Philippine languages

As of 2018, no automated phonetic transcriber has yet been created for Philippine languages. However, several phoneme-level transcription and segmentation systems have previously been developed for Tagalog, referred to in these projects as Filipino. These systems were created primarily for speech recognition purposes, as transcription at the broader phoneme-level is often sufficient for recognizing more complex structures such as words and phrases. A summary of Filipino phoneme recognition systems is provided in Table 2.1.

Table 2.1: Phoneme recognition systems for the Filipino Speech Corpus

Year	Speech Technology	Accuracy
2003	TIMIT + Filipino bootstrapping	42.12%
2003	MLP	62.64% (par/sen) 63.93% (words) 72.60% (syllables)
2006	HMM-NN + STC	67.0%

The first Filipino Speech Corpus (FSC) was created in 2003 by the UP Digital Signal Processing (DSP) Laboratory [18]. With its creation came a need to label several hundred hours of speech data at the phoneme level, so that the data could be used to develop speech applications for the Tagalog language. Three Filipino phoneme recognition systems were subsequently developed to address this need:

1. A phoneme recognizer initially trained on the TIMIT database (a phonetically annotated corpus of American English speech [7]) was then bootstrapped with a Filipino phonotactic

model to recognize phonemes in the FSC [18]. This system obtained an average phoneme recognition rate (or accuracy) of 42.12%.

2. A Filipino phoneme transcription and segmentation system was trained using a multi-layer perceptron (i.e. a deep neural network) [19]. The performance of this system was evaluated using three subcorpora of the FSC: 62.64% accuracy was obtained for the paragraphs/sentences subcorpus, 63.93% accuracy for the words subcorpus, and 72.60% accuracy for the syllables subcorpus.
3. A phoneme-level segmentation and transcription system was developed for the FSC using a hybrid HMM-neural network architecture with split temporal context (i.e. temporal analysis of left- and right-adjacent acoustic features) [20]. A peak accuracy of 67.0% was obtained using this system.

Chapter 3

Problem Statement and Objectives

3.1 Problem Statement

Phonetic transcriptions of recorded speech data are necessary in linguistics research for characterizing the phonology of Philippine languages. They are also fundamental in the development of modern speech recognition systems. Traditionally, such transcriptions are performed manually by trained phoneticians. A modern alternative is the use of automated phonetic transcribers, which provide a faster and cheaper alternative to manual transcription, and have been widely developed for foreign languages. However, automated transcription for Philippine languages has thus far only been designed for word-level and phoneme-level recognition, and cannot resolve speech at the phone-level. The development of a system capable of phonetic transcription would facilitate further advances in phonological analysis and speech recognition for Philippine languages.

3.2 Objectives

The aim of this project was to develop and evaluate the performance of an automated phonetic transcription system for recorded speech data in select Philippine languages. The system consists of three phases:

1. a grapheme-to-phoneme (G2P) converter, which parses the orthographic text transcription of a given speech utterance and returns the likeliest phonetic equivalent of each word;
2. an automatic speech recognition (ASR) tool, which processes the audio recording of a speech utterance, and returns the sequence of recognized phones; and
3. an adaptive substitution phase, which aligns the two prior transcriptions using an optimal

matching algorithm, modifies the G2P transcription if a permissible allophone is detected using the ASR system, and returns the adapted phonetic transcription.

The system was evaluated based on its phone error rate (PER) versus manual transcriptions, and its performance was compared with the PER of the individual G2P and ASR phases. The speed of the system was recorded based its total running time for a given duration of input speech data.

3.3 Scope and Limitation

The system in this project was trained and tested using 7.5 hours of speech data in three languages—Tagalog, Cebuano, and Hiligaynon—acquired from previous projects of the UP Digital Signal Processing Laboratory. The system produces text-based phonetic transcriptions of input speech utterances. However, it does not segment (i.e. mark start and end timestamps of) phones on the input audio files, as temporal segmentation of speech is not a typical requirement of transcription for linguistic analysis.

Chapter 4

Methodology

This chapter provides a discussion on the development of the automated phonetic transcription system. A block diagram of the system is presented in Figure 4.1. The Phonetisaurus toolkit [14] was used for grapheme-to-phoneme conversion, while the Kaldi toolkit [11] was used for automated phone recognition.

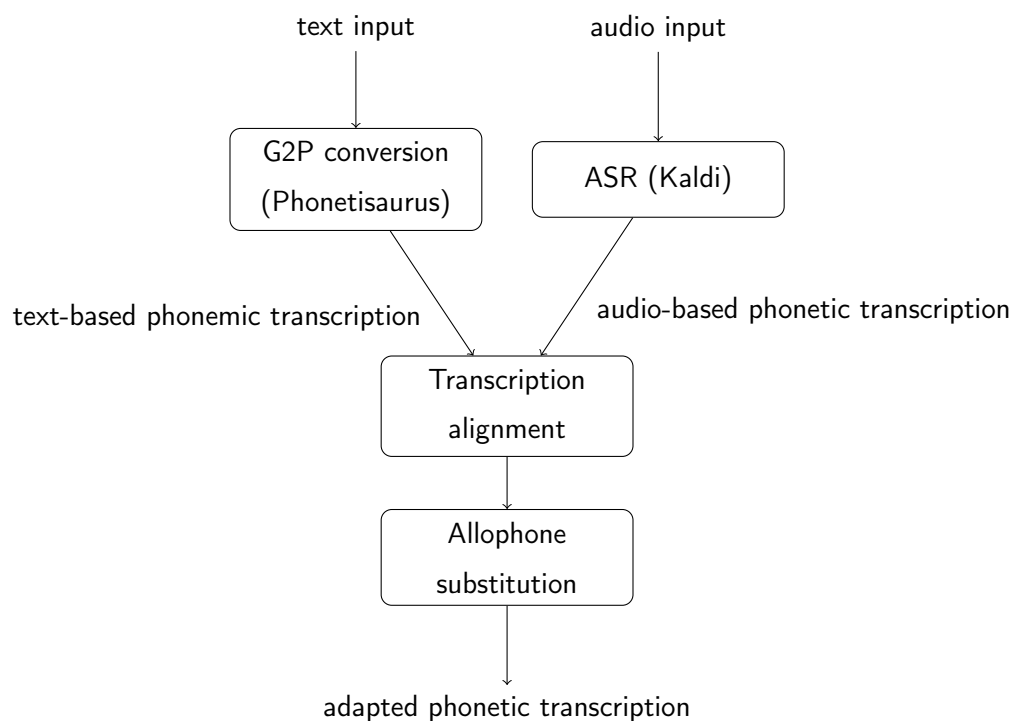


Figure 4.1: Block diagram of the automated phonetic transcription system

4.1 Transcription and preprocessing of speech corpora

Speech corpora for Tagalog, Cebuano, and Hiligaynon were obtained from the UP DSP Laboratory. The recordings and corresponding orthographic transcriptions in these corpora were produced as part of the Interdisciplinary Signal Processing for Pinoys (ISIP) Project 6: *Philippine Languages Database for Mother Tongue-based Multilingual Education and Applications* from 2011 to 2013 [21].

An inventory of the phones in the three target languages, based on several reference grammars and phonological studies [22]–[23], has been compiled as shown in Figures 4.2 and 4.3. This phone inventory consists of 22 consonants, 5 vowels, and 7 diphthongs.

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Palatal	Velar	Glottal
Plosive	p b		t d			k g	ʔ	
Nasal	m		n			ŋ		
Tap or Flap			ɾ					
Fricative		f v		s z	ʃ			h
Affricate			tʃ dʒ					
Approximant	W					j	w	
Lateral approximant			l					

Figure 4.2: Consonants in the phone inventory of Tagalog, Cebuano, and Hiligaynon

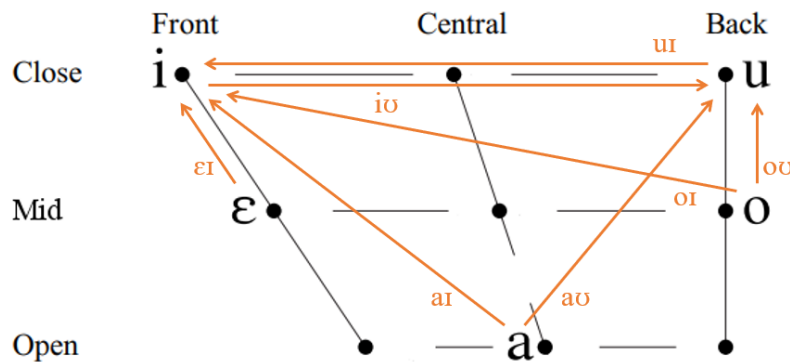


Figure 4.3: Vowels and diphthongs in the phone inventory of Tagalog, Cebuano, and Hiligaynon

We initially limited our transcription to the phone inventory above, to reflect only the speech sounds found in native vocabulary. However, we found that the speech corpora contain a

significant percentage of English words and phrases as well as Spanish loan words, whose pronunciation cannot be accurately captured using native phones alone. We therefore expanded the phone inventory for transcription to include select English and Spanish phones, a summary of which is provided in Table 4.1 with additional phones in bold font.

Table 4.1: Expanded phone inventory used for transcription of speech corpora

Vowels			Consonants						
a	ε	i	p	b	t	θ	d	ð	
o	u	aɪ	k	g	ʔ	m	n	ɲ	
aʊ	ɛɪ	iʊ	ŋ	r	ɹ	f	v	s	
oɪ	oʊ	uɪ	z	ʃ	h	tʃ	dʒ	w	
ə	æ	ɔ	j	l	ʎ				

The speech corpora for Tagalog, Cebuano, and Hiligaynon consist of speech audio files and their corresponding orthographic transcriptions. Since verified phonetic transcriptions were also required to train both the grapheme-to-phoneme converter and automated speech recognition system, phonetic transcriptions of each speech file were produced by hand. The accuracy of these manual transcriptions have been verified by representatives of the UP Department of Linguistics who are native speakers of each language. The speech data used is summarized in Table 4.2. Each ID corresponds to a speaker and a unique dataset.

Table 4.2: Summary of transcribed data

Cebuano		Hiligaynon		Tagalog	
ID	Length	ID	Length	ID	Length
1691	34m 52s	0738	26m 18s	0156	17m 2s
2295	38m 10s	0739	25m 12s	0812	18m 35s
2998	25m 46s	0740	25m 55s	2359	17m 42s
4703	8m 15s	0741	14m 40s	4281	19m 20s
8973	27m 46s	0742	14m 45s	5093	17m 47s
		0744	29m 5s	5546	20m 40s
				6093	16m 37s
				7859	20m 54s
				8125	17m 58s
				8453	16m 43s
Total	2h 15m	Total	2h 16m	Total	3h 5m

4.2 Data allocation for the train and test sets

Roughly eighty percent (80%) of the speech data was allocated as the training set, while the remaining twenty percent (20%) was used as the test set. To compare system performance with respect to available data, two train and test sets were formed. One hour of speech data for each language were used for set 1 while two hours of Cebuano, two hours of Hiligaynon, and three hours of Tagalog speech data were used for Set 2.

Table 4.3: Data allocation for training and testing

	Set 1		Set 2	
	Train	Test	Train	Test
Cebuano	2998, 8973 (54m)	4703 (8m)	1691, 2295, 2998, 8973 (127m)	4703 (8m)
Hiligaynon	0739, 0740 (51m)	0738 (26m)	0739, 0740, 0741, 0742, 0744 (1h 50m)	0738 (26m)
Tagalog	0812, 2359 (37m)	0156 (17m)	2359, 4281, 5093, 5546, 6093, 7859, 8125, 8453 (2h 29m)	0156, 0812 (36m)
Multilingual	2998, 8973, 0739, 0740, 0812, 2359 (2h 22m)	4703, 0738, 0156 (51m)	1691, 2295, 2998, 8973, 0739, 0740, 0741, 0742, 0744, 2359, 4281, 5093, 5546, 6093, 7859 8125, 8453 (6h 26m)	4703, 0738, 0156, 0812 (1h 10m)

Due to the inconsistency between the lengths of speech data, the original distribution of 80% training and 20% testing set was not strictly followed as moving datasets between the train and test sets would inevitably skew the distribution.

4.3 Training of a grapheme-to-phoneme conversion model

The orthographic transcriptions of utterances in the speech corpora were stored in formatted log files, while the verified manual transcriptions were encoded in spreadsheets and converted to CSV files. Scripts were created in Python to parse these files and create dictionaries for each training instance, containing a list mapping all words in the training subset to their corresponding pronunciation variants.

The Phonetisaurus toolkit was then trained using these dictionaries to generate WFST pronunciation models, which are capable of producing the likeliest phonetic transcription of both known and unknown words. Phonetic transcriptions for the test subsets were then generated

by applying the pronunciation model to each word in the test log files, and concatenating the pronunciations to produce a single transcription per utterance, with a summary of all transcriptions encoded into CSV files.

Shell scripts were created to automate the G2P transcription process, and were used to generate standalone formatted transcriptions from the orthographic input files. One instance of the G2P subsystem is shown in Figure 4.4, with the training files, model generation process, testing files, and evaluation metrics displayed in the command line.

```
dsp@ASUS-Ubuntu1804:/mnt/windows/Users/Admin/Desktop/AAA1/198/scripts$ time ./run-g2p-subsystem.sh
../data/_small/train/CEB/2998.111024.084752.csv
../data/_small/train/CEB/8973.111024.032108.csv
../data/_small/train/HIL/0739.130130.063449.csv
../data/_small/train/HIL/0740.130201.010543.csv
../data/_small/train/TGL/0812.110816.021250.csv
../data/_small/train/TGL/2359.110818.085852.csv
INFO:phonetisaurus-train:2019-04-29 02:00:49: Checking command configuration...
INFO:phonetisaurus-train:2019-04-29 02:00:49: Checking lexicon for reserved characters: '}', '|', '_'...
INFO:phonetisaurus-train:2019-04-29 02:00:49: Aligning lexicon...
INFO:phonetisaurus-train:2019-04-29 02:00:53: Training joint ngram model...
INFO:phonetisaurus-train:2019-04-29 02:00:53: Converting ARPA format joint n-gram model to WFST format...
INFO:phonetisaurus-train:2019-04-29 02:00:54: G2P training succeeded: train/model.fst
../data/_small/test/CEB/4703.111020.092714.csv
../data/_small/test/HIL/0738.130117.075101.csv
../data/_small/test/TGL/0156.110816.035620.csv
../data/_small/testref/CEB/4703.111020.092714.csv
../data/_small/testref/HIL/0738.130117.075101.csv
../data/_small/testref/TGL/0156.110816.035620.csv
total phones in reference: 30439
total edit distance: 2381
phone error rate: 0.07822201780610401

real    0m7.689s
user    0m6.860s
sys     0m0.117s
```

Figure 4.4: G2P training and testing instance using Multilingual Set 1

4.4 Training of the automatic speech recognition system

An ASR system was created using the Kaldi toolkit [11]. Data for training the acoustic and language models were first prepared. Data prepared for the acoustic model included the audio files and their corresponding transcriptions, speaker information, and a lexicon of all phones to be recognized. The SRI Language Modeling Toolkit (SRILM) was used to build the language model. As the system was expected to distinguish individual phones instead of words, the lexicon was encoded to contain only the inventory of phones as listed in Table 4.1. Likewise, the utterances within the corpus were transcribed in phones instead of words.

A Gaussian mixture model–hidden Markov model (GMM-HMM) framework was used for acoustic modeling while a bigram model was chosen for language modeling after its performance

was compared with a unigram model. Using a bigram model, the probability of detecting a certain phone was dependent on the previously detected phone, as opposed to a unigram model with phone probabilities independent of previously detected phones. The acoustic modeling process was better optimized by cycling through training and alignment phases.

The ASR was trained using monophone and triphone models, as well as Mel-frequency cepstral coefficients with delta feature computation (MFCC+ Δ + $\Delta\cdot\Delta$), and linear transform estimation with maximum-likelihood linear transform estimation (LDA+MLLT) models. The resulting transcriptions with the highest accuracy were then selected for use in the adaptive phase.

4.5 Alignment of G2P and ASR transcriptions

After the speech data was processed by the trained G2P and ASR phases, two transcriptions (a text-based and an audio-based transcription) were generated for each utterance. In order to make allophonic corrections to each phonemic transcription, these two sequences were first aligned. This involved finding matching pairs of phones, as well as identifying subsequences of the transcription where substitution, deletion, or insertion of phones is possible.

In this project, the Needleman-Wunsch algorithm was used to align the two transcriptions. The Needleman-Wunsch is a dynamic programming algorithm which determines the optimal global alignment of two sequences $A = a_1, a_2, \dots, a_n$ and $B = b_1, b_2, \dots, b_m$ by incrementally scoring the similarity of their subsequences [24]. In this algorithm, a two-dimensional scoring matrix is generated, with either axis marked by elements of one sequence. A score is given to a pair of elements to denote the similarity of a subsequence ending in that pair, with a more positive score indicating higher similarity. This score is calculated by:

$$H_{i,j} = \max \begin{cases} H_{i-1,j-1} + s(a_i, b_j), \\ H_{i-1,j} - W_1, \\ H_{i,j-1} - W_1 \end{cases} \quad 0 \leq i \leq n, 0 \leq j \leq m$$

where $H_{i,j}$ is the score for a pair of elements a_i and b_j , $s(a_i, b_j)$ is the substitution score for the two elements (positive if a_i and b_j are identical, negative otherwise), and W_1 is the penalty for a deletion or insertion (otherwise known as the gap penalty). For this project, a substitution score of ± 1 and a gap penalty of 1 were applied. The optimal alignment of each pair of transcriptions was thus found by beginning at the highest-indexed element and tracing back the path of adjacent elements with maximum scores.

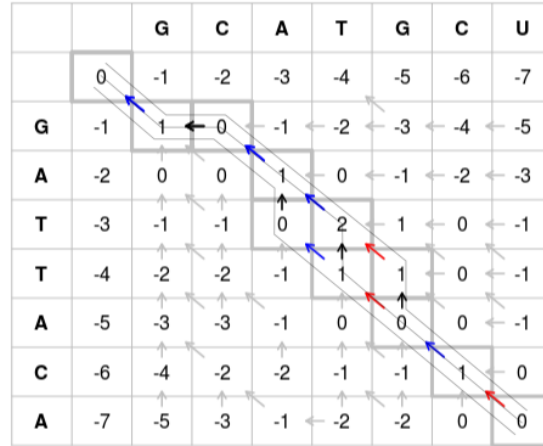


Figure 4.5: Sequence alignment with Needleman-Wunsch for $s = \pm 1$, $W_1 = 1$

4.6 Substitution of allophones

Once the G2P and ASR transcriptions were aligned, the phonemic G2P transcription was then adapted to account for possible allophonic variations detected by the ASR. A list of observed phonological rules in the target languages was compiled as shown in Table 4.4. Aligned pairs of phones were then checked. If a mismatch (i.e. a pair of two non-identical, aligned phones) was detected, and a substitution between the G2P and ASR phones was permissible based on the rules in Table 4.4, the ASR phone was encoded; otherwise, the G2P transcription was retained. In this manner, we were able to generate a third, allophone-adapted phonetic transcription of each utterance.

Table 4.4: Phonological Rules in Target Languages

Phonological Process	Environment
$\varepsilon \rightleftharpoons i$	all
$o \rightleftharpoons u$	all
$oi \rightleftharpoons ui$	all
$r \rightleftharpoons \text{ɹ}$	all
$\emptyset \rightleftharpoons ?$	$\#_V, V_V, V_ \#$
$\text{tʃ} \rightleftharpoons \text{ti(j)}$	all
$\text{dʒ} \rightleftharpoons \text{di(j)}$	all
$\text{ʃ} \rightleftharpoons \text{si(j)}$	all
$\text{n} \rightleftharpoons \text{ni(j)}$	all

4.7 Evaluation metrics

The accuracy of the system was evaluated using the phone error rate (PER) metric, which is defined by the following expression:

$$\text{PER} = \frac{S + D + I}{N_T} \times 100\%$$

where N_T is the total number of phones in the manual transcription, and the sum $S + D + I$ is the Levenshtein edit distance, or the total number of substitutions, deletions, and insertions required to transform the system-created transcription into the original manual transcription [7]. Note that $PER = 100\% - accuracy$, and a lower PER indicates better system performance.

Average PER for the end-to-end system, as well as the individual G2P and ASR phases, were computed based on manual transcription of the test dataset. These results were tabulated and compared to known inter-transcriber error rates (i.e. the discrepancy in manual phone transcriptions among professional phoneticians).

The speed of the adaptive system, as well as its prior subsystems, was measured and compared to known manual transcription speeds. To this end, scripts were created to automate the entire system, and the total runtime of the system for a given duration of speech data was recorded. The ASR runtime was evaluated using a Lenovo Ideapad 530s with no concurrent background processes, and the following technical specifications as listed in Table 4.5:

Table 4.5: Hardware specifications for ASR runtime evaluation

OS	Ubuntu 18.04.2 LTS
CPU	Intel Core i5-8250U 1.6GHz
RAM	8GB DDR4
GPU	Intel UHD Graphics 620

Chapter 5

Results and Discussion

This chapter presents the results of automated phonetic transcription performed by the adaptive system and its prior subsystems, and compares their accuracy and finishing time to those of transcription performed by hand.

5.1 ASR performance optimization

The ASR was trained using monophone, triphone, MFCC+ Δ + $\Delta\cdot\Delta$, and LDA+MLLT acoustic models, with each acoustic model being further tested for performance of both unigram and bigram language models. The results are summarized in Tables 5.1 and 5.2.

Table 5.1: ASR performance using Set 1 data allocation

	monophone		triphone		MFCC+ Δ + $\Delta\cdot\Delta$		LDA+MLLT	
	unigram	bigram	unigram	bigram	unigram	bigram	unigram	bigram
Cebuano								
Phone Error Rate	55.18%	50.17%	64.37%	61.33%	66.63%	62.97%	67.51%	63.05%
Accuracy	44.82%	49.83%	35.63%	38.67%	33.37%	37.03%	32.49%	36.95%
Hiligaynon								
Phone Error Rate	59.77%	53.89%	65.31%	61.57%	65.29%	62.05%	68.06%	65.32%
Accuracy	40.23%	46.11%	34.69%	38.43%	34.71%	37.95%	31.94%	34.68%
Tagalog								
Phone Error Rate	35.08%	31.88%	34.00%	30.48%	35.48%	31.49%	34.80%	30.90%
Accuracy	64.92%	68.12%	66.00%	69.52%	64.52%	68.51%	65.20%	69.10%
Multilingual								
Phone Error Rate	40.26%	37.17%	38.13%	35.325%	37.31%	35.39%	35.80%	34.00%
Accuracy	59.74%	62.83%	61.87%	64.75%	62.69%	64.61%	64.20%	66.00%

Table 5.2: ASR performance using Set 2 data allocation

	monophone		triphone		MFCC+ Δ + Δ · Δ		LDA+MLLT	
	unigram	bigram	unigram	bigram	unigram	bigram	unigram	bigram
Cebuano								
Phone Error Rate	53.57%	49.81%	55.52%	51.67%	60.40%	56.82%	49.83%	45.34%
Accuracy	46.43%	50.19%	44.48%	48.33%	39.60%	43.18%	50.17%	54.66%
Hiligaynon								
Phone Error Rate	39.93%	35.25%	36.40%	33.87%	37.25%	34.63%	38.59%	35.37%
Accuracy	60.07%	64.75%	63.60%	66.13%	62.75%	65.37%	61.41%	64.63%
Tagalog								
Phone Error Rate	29.10%	25.67%	18.46%	17.07%	18.08%	16.66%	16.33%	15.14%
Accuracy	70.90%	74.33%	81.54%	82.93%	81.92%	83.34%	83.67%	84.86%
Multilingual								
Phone Error Rate	34.05%	31.05%	24.29%	22.54%	23.79%	21.95%	21.81%	20.32%
Accuracy	65.95%	68.95%	75.71%	77.46%	76.21%	78.05%	78.19%	79.68%

5.2 Transcription accuracy

Tables 5.3 to 5.5 show the results obtained by applying single-language and multilingual data to each of the three APT systems. The accuracies of the systems for each dataset are also visualized in Figure 5.1.

Table 5.3: G2P performance for single-language and multilingual datasets

	Cebuano		Hiligaynon		Tagalog		Multilingual	
	1h	2h	1h	2h	1h	3h	3h	7h
Phones in reference	7827	7827	11523	11523	11059	22630	30439	41980
Edit distance	485	514	1026	992	651	1339	2381	3095
Phone Error Rate	6.20%	6.57%	8.90%	8.61%	5.87%	5.92%	7.82%	7.37%
Accuracy	93.80%	93.43%	91.10%	91.39%	94.13%	94.08%	92.18%	92.63%

Table 5.4: ASR performance for single-language and multilingual datasets

	Cebuano		Hiligaynon		Tagalog		Multilingual	
	1h	2h	1h	2h	1h	3h	3h	7h
Phones in reference	3851	3851	11636	11636	11082	22611	26569	38098
Edit distance	1932	1746	6271	3941	3378	3423	9033	7740
Phone Error Rate	50.17%	45.34%	53.89%	33.87%	30.48%	15.14%	34.00%	20.32%
Accuracy	49.83%	54.66%	46.11%	66.13%	69.52%	84.86%	66.00%	79.68%

Table 5.5: Adaptive system performance for single-language and multilingual datasets

	Cebuano		Hiligaynon		Tagalog		Multilingual	
	1h	2h	1h	2h	1h	3h	3h	7h
Phones in reference	3851	3851	11513	11513	11079	22610	26443	37974
Edit distance	263	297	1169	1131	778	1495	2400	3049
Phone Error Rate	6.83%	7.71%	10.15%	9.82%	7.02%	6.61%	9.08%	8.03%
Accuracy	93.17%	92.29%	89.85%	90.18%	92.98%	93.39%	90.92%	91.97%

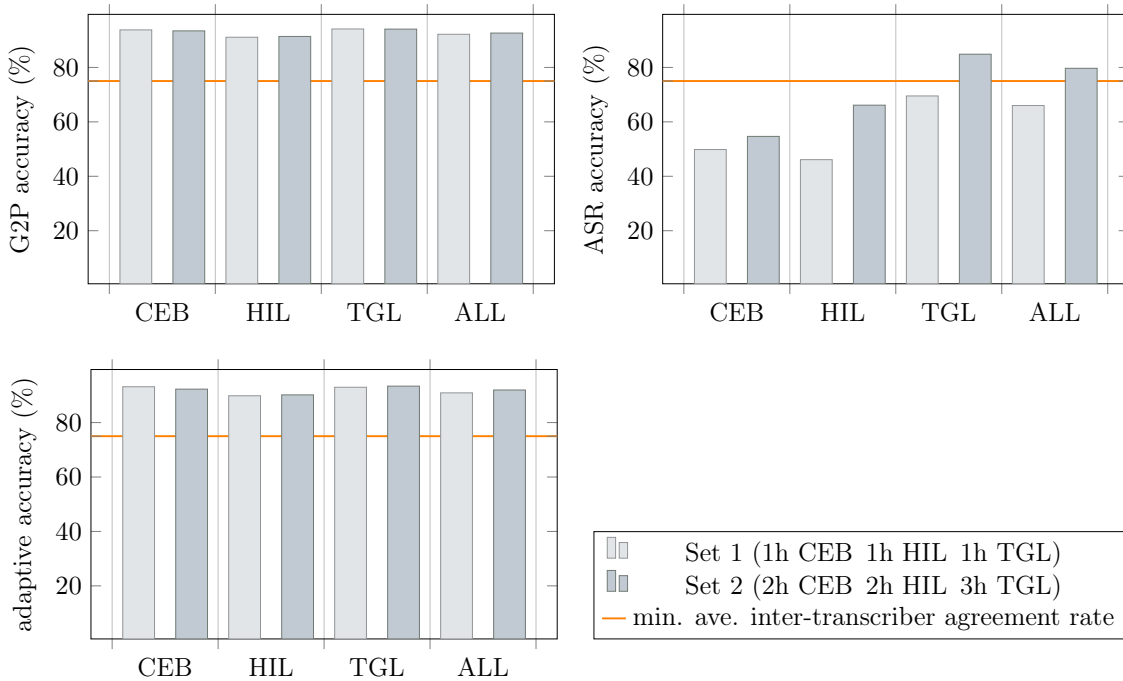


Figure 5.1: Percent accuracies of APT systems versus manual transcriptions

The G2P system yielded the best performance among the three APT systems, showing improvements of 16.10%–19.13% over the minimum 75% human inter-transcriber accuracy. Meanwhile, the ASR system performed the poorest of the three, with only the 3-hour Tagalog and 7.5-hour multilingual models operating at par with human transcribers. The adaptive system also returned accuracies of at least 15% above the minimum inter-transcriber rate across all datasets; however, these results were consistently lower than the G2P accuracy rates by 0.63%–1.26%. As the ASR system had subpar accuracy compared to the G2P system in all eight test cases, it may have introduced more errors than it corrected during the allophone substitution phase, and therefore lowered the overall accuracy of the adaptive system instead of improving it as desired.

Notably, while the G2P and adaptive systems did not show any significant differences in performance with the increase in data from Set 1 to Set 2, the ASR system consistently yielded higher accuracies (ranging from 4.83% to 20.02% improvement) using the expanded data set. The ASR system may therefore potentially reach human accuracy levels for Cebuano and Hiligaynon transcription by increasing the amount of training data provided. Preliminary testing of the ASR system also showed improvements in accuracy by increasing the order of the n-gram language model, with accuracies of over 90% achieved at $n \geq 5$ for the 3-hour Tagalog dataset and at $n \geq 6$ 7.5-hour multilingual dataset.

5.3 Runtime comparison

The running times of each system for different durations of input data were recorded and averaged as shown in Tables 5.6 and 5.7. Of the three APT methods, the G2P system was observed to have the fastest transcription speeds (1.936 to 5.395s per hour of data), while the adaptive system was the slowest (5m 47s to 19m 34s per hour of data), having been implemented as a combination of both prior systems. Across the eight tests performed, all three systems exhibited a decrease in required transcription time of over 75% versus reported typical rates of human transcribers (180m per hour of data).

Table 5.6: System runtimes for single-language and multilingual datasets

	Cebuano		Hiligaynon		Tagalog		Multilingual	
	1h	2h	1h	2h	1h	3h	3h	7h
G2P	2.596s	4.382s	3.110s	17.21s	2.874s	22.09s	7.617s	13.55s
ASR	13m 14s	9m 44s	34m 50s	22m 3s	10m 26s	13m 51s	33m 46s	40m 13s
Alignment	0.296s	0.310s	1.659s	1.694s	1.224s	2.580s	3.075s	4.464s
Adaptive	13m 17s	9m 49s	34m 55s	22m 22s	10m 30s	14m 16s	33m 57s	40m 31s

Table 5.7: Average runtimes across varying dataset durations

Dataset Duration	System Running Time		
	G2P	ASR	Adaptive
1h	2.86s	19m 30s	19m 34s
2h	10.79s	15m 54s	16m 5s
3h	14.85s	23m 49s	24m 6s
7h	13.55s	40m 13s	40m 31s

Chapter 6

Conclusion and Recommendations

6.1 Conclusion

In this project, the implementations of three automated phonetic transcription methods—grapheme-to-phoneme conversion, automatic speech recognition, and adaptive substitution—are presented as alternatives to manual transcription of speech data in Tagalog, Cebuano, and Hiligaynon. As available speech corpora for Philippine languages are limited in number and scale, we trained and tested each system on small amounts of data in each of the target languages, ranging from 1 to 7.5 hours of recorded speech, and measured the speed and accuracy of the systems across varying lengths of input data.

Based on our results, the G2P system is evaluated as the most suitable alternative to manual phonetic transcription, having been shown to exceed human accuracy while reducing overall transcription time by over 98%, and is recommended for use whenever orthographic transcriptions of the speech data to be processed are available. In cases where only recorded audio of speech utterances is provided, the ASR models for Tagalog and mixed speech are also shown to perform at par with human accuracy while yielding faster transcription speeds. These systems serve as a competent baseline for future developments in APT for Philippine languages, and are expected to significantly reduce the costs of transcription, facilitating further research and advancements in Philippine linguistics and speech technology.

6.2 Future Work

6.2.1 Increasing speech duration and language model order for ASR

The ASR system exhibited consistent improvements in accuracy by providing greater durations of speech data. Preliminary tests also showed potential improvements in system performance with increasing order of n-gram language models. Further study on the relationships between dataset duration, language model order, and ASR performance is recommended.

6.2.2 Cross-lingual effectiveness

Verified phonetic transcriptions are scarce for many Philippine languages, which hinders the generation of new APT models. Further tests may therefore be done to evaluate the performance of models trained with known languages in generating transcriptions for unknown languages. (For example, how well would the Tagalog, Cebuano, or Hiligaynon models perform when transcribing data in Ilokano, Bikolano, or Waray?)

6.2.3 Alternative phoneme categorization for Visayan languages

ASR phone recognition in both Cebuano and Hiligaynon was over 20% less accurate than Tagalog recognition. This may have been due to the distinctive phonology of Visayan languages [25], [23], wherein the [ɛ]-[i] and [o]-[u] pairs of phones are characterized as single phonemes, and are typically articulated in roughly the same position of the mouth. An alternative phone inventory, which identifies these pairs as only a single sound each, could thus be investigated for potential improvements in Visayan phone recognition.

Bibliography

- [1] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Switzerland: Springer Publishing Company, Incorporated, 2014.
- [2] J. R. Novak, “WFST-based G2P conversion with Phonetisaurus.” Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing, July 2012. Tutorial presentation.
- [3] N. Barroso, K. L. de Ipina, and P. M. Calvo, “An automatic and adaptive phonetic transcriber for the Basque language,” in *4th International Work Conference on Bio-inspired Intelligence*, IEEE, June 2015.
- [4] F. de Saussure, *Course in General Linguistics*. London, England: Bloomsbury Academic, 2013.
- [5] D. T. Dayag and S. N. Dita, “Linguistic research in the Philippines: Trends, prospects, & challenges,” in *Philippine social sciences: Capacities, directions, and challenges* (V. A. Miralao and J. Agbisit, eds.), pp. 110–126, Quezon City: Philippine Social Science Council, January 2012.
- [6] B. Schuppler, S. Grill, A. Menrath, and J. A. Morales-Cordovilla, “Automatic phonetic transcription in two steps: Forced alignment and burst detection,” in *Statistical Language and Speech Processing: Second International Conference*, October 2014.
- [7] C. Lopes and F. Perdigão, “Phone recognition on the TIMIT database,” in *Speech Technologies* (I. Ipsic, ed.), InTech, 2011.
- [8] M. Dobrovolsky, “Phonetics: the sounds of language,” in *Contemporary Linguistic Analysis: An Introduction* (W. O’Grady and J. Archibald, eds.), ch. 2, Ontario: Pearson Canada, Inc., 2016.

- [9] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, New Jersey: Prentice Hall, 2001.
- [10] G. Hinton, L. Deng, D. Yu, G. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi speech recognition toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.
- [12] C. Gaida, P. Lange, R. Petrick, P. Proba, A. Malatawy, and D. Suendermann-Oeft, “Comparing open-source speech recognition toolkits,” tech. rep., DHBW Stuttgart, Germany, 2014.
- [13] M. Bisani and H. Ney, “Joint-sequence models for grapheme-to-phoneme conversion,” *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.
- [14] J. R. Novak, N. Minematsu, and K. Hirose, “WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding,” in *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, pp. 45–49, Association for Computational Linguistics, July 2012.
- [15] P. Jyothi and M. Hasegawa-Johnson, “Low-resource grapheme-to-phoneme conversion using recurrent neural networks,” in *International Conference on Acoustics, Speech and Signal Processing*, IEEE, March 2017.
- [16] S. Hahn, P. Vozila, and M. Bisani, “Comparison of grapheme-to-phoneme methods on large pronunciation dictionaries and LVCSR tasks,” 2012.
- [17] W. O’Grady, C. Dyck, Y. Rose, E. Czaykowska-Higgins, and M. Dobrovolsky, “Phonology: contrasts and patterns,” in *Contemporary Linguistic Analysis: An Introduction* (W. O’Grady and J. Archibald, eds.), ch. 3, Ontario: Pearson Canada, Inc., 2016.
- [18] R. C. Guevara, M. Co, E. Espina, I. D. Garcia, E. Tan, R. Ensomo, and R. Sagum, “Development of a Filipino speech corpus,” tech. rep., DSP Laboratory, Department of Electrical and Electronics Engineering, University of the Philippines, Diliman, Quezon City, Philippines, 2003.

- [19] R. G. Sagum, R. A. Ensomo, E. M. Tan, and R. C. L. Guevara, "Phoneme alignment of Filipino speech corpus," in *TENCON 2003 Conference on Convergent Technologies for Asia-Pacific Region*, IEEE, October 2003.
- [20] I. J. Chua and N. Eustaquio, "Automatic isolated word recognition system and phoneme-level segmentation of the Filipino speech corpus using HTK." Undergraduate project, Department of Electrical and Electronics Engineering, University of the Philippines, 2006.
- [21] A. F. B. Laguna and R. C. L. Guevara, "Development, implementation and testing of language identification system for seven Philippine languages," *Philippine Journal of Science*, vol. 144, no. 1, pp. 81–89, 2015.
- [22] P. Schachter and F. T. Otones, *Tagalog Reference Grammar*. Berkeley and Los Angeles, California: University of California Press, 1972.
- [23] A. A. Bolas, "Comparative analysis on the phonology of Tagalog, Cebuano, and Itawis," tech. rep., University of the Philippines–Diliman, August 2013. Available: <http://www.academia.edu/4427395>.
- [24] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.
- [25] E. P. Wolfenden, *Hiligaynon Reference Grammar*. Honolulu: University of Hawaii Press, 1971.