Readable: A reading miscue detection for the Hiligyanon Language

A Special Problem

Presented to

the Faculty of the Division of Physical Sciences and Mathematics

College of Arts and Sciences

University of the Philippines Visayas

Miag-ao, Iloilo

In Partial Fulfillment

of the Requirements for the Degree of

Bachelor of Science in Computer Science by

GONZALES, Benjie Jr.

PANIZALES, John Patrick

PIORQUE, Lester

Francis F. DIMZON

Adviser

May 6, 2023

Abstract

Advancements in technology have allowed people to utilize machine learning techniques such as automatic speech recognition (ASR). ASR is a technology that allows computers to automatically recognize and transcribe spoken language , and it has made significant advances in recent years. Studies have been conducted and results show the potential of technology-based reading tutors in improving reading skills. However, it can still be challenging to achieve high levels of accuracy for some languages and accents, especially those that are underrepresented in ASR training data. This paper, therefore, aims to develop a system to detect the acceptability of an input reading speech relative to a reference speech in Hiligaynon. The system will use the open source speech recognition toolkit — Kaldi, to build the acoustic model and to perform the viterbi-forced alignment process to determine the input speech's similarity against the reference speech.

Keywords: Reading miscue detector, automated reading tutor, Hiligaynon language, viterbi-forced alignment

Contents

List of Tables

Chapter 1

Introduction

1.1 Overview of the Current State of Technology

The ability to read is a fundamental skill that is necessary for success in many areas of life. Unfortunately, there are many adults in the Philippines who have never learned to read or who have difficulty reading due to various reasons such as illiteracy, limited education, or learning disabilities. These individuals often face significant barriers to employment, education, and social participation, leading to a cycle of poverty and marginalization.

In 2019, the Philippines achieved a literacy rate of 96.5 % for the segment of the population aged 10 and over according to the PSA's Functional Literacy, Education and Mass Media Survey (FLEMMS), as reported in an article from Business World entitled, "Literacy rate estimated at 93.8% among 5 year olds or older — PSA." Literacy was defined as the ability to read and write "with

understanding of simple messages in any language or dialect." However, the same article notes that this was the same rate observed in 2013, a matter described as alarming by University of Asia and the Pacific Senior Economist Cid L. Terosa, stating that even minimal improvements should be expected especially after six years.

More recently, according to a report published by UNICEF in collaboration with UNESCO and the World Bank, the percentage of 10-year-olds in low- and middle-income countries who are unable to read is as high as 70%. This figure has likely been affected by school closures brought about by the COVID-19 pandemic. The same report also stated that only 10% of children in the Philippines were able to read simple text as of March 2022. Alarmingly, a separate report published by the World Bank in 2021 found that the rate of learning poverty - defined as the inability to read simple text by age 10 - in the Philippines was at 90%. These statistics highlight the urgent need to address the education crisis in the Philippines and the need to further augment the country's current literacy situation.

To address this problem, we propose the development of an automatic reading miscue detection system called Readable, specifically targeting the Hiligaynon language. Hiligaynon, also known as Ilonggo, is an Austronesian language spoken in the Western Visayas region of the Philippines, particularly in the provinces of Iloilo, Guimaras, Negros Occidental, and Capiz. It is one of the major languages of the Philippines, spoken by millions of people as a first or second language. Our reading miscue detection system will utilize machine learning techniques including automatic speech recognition (ASR). ASR is a technology that allows computers to automatically recognize and transcribe spoken language, and it has made significant advances in recent years. However, it can still be challenging to achieve

high levels of accuracy for some languages and accents, especially those that are underrepresented in ASR training data. By targeting local languages like Hiligaynon and designing our ASR system to work well for these languages, we can help ensure that our reading tutor is accessible and effective for non-reading adults in the Philippines.

While there are some similar applications like Google Read Along available for reading instruction, they may not be accessible or relevant for many non-reading adults in the Philippines due to language barriers or lack of internet connectivity. By targeting local languages like Hiligaynon and utilizing the benefits of natural language processing and machine learning techniques, our automatic reading tutor can provide personalized and effective reading instruction that is accessible and relevant for non-reading adults in the Philippines. By providing accessible, effective, and scalable reading education in Hiligaynon, we hope to improve the lives and prospects of non-reading adults in the Philippines and break the cycle of poverty and illiteracy. Children with strong literacy skills grow more consistently and confidently in their studies, and reading literacy is a crucial gateway to other learning areas such as the humanities, mathematics, and the sciences. By addressing learning poverty and promoting reading literacy, we can help ensure that children in the Philippines have the opportunity to reach their full potential and succeed in their studies.

1.2 Problem Statement

Given the current educational crisis our country is facing (UNICEF, UNESCO, & Bank, 2022) and with the aim to further improve the current state of literacy rate of our country (Hernandez, 2020), the development of automatic reading tutor systems which entails building reading miscue detection systems and other related programs becomes relevant. Furthermore, the limited resources available for Hiligaynon in the context of speech processing technologies opens a good opportunity to attempt to make a contribution for the said domain of interest.

1.3 Research Objectives

1.3.1 General Objective

The aim of this project is to develop a reading miscue detection system that would determine the acceptability of an input speech relative to a reference speech pattern.

1.3.2 Specific Objectives

Specifically, the project targets to:

1. Train and model a DNN-based acoustic model for Hiligaynon.
2. Evaluate the model in terms of the Word Error Rate via 6-fold cross validation.

3. Use the developed acoustic model to derive phonemic transcriptions of the input speeches/audio.

4. Develop a system that will determine the acceptability of a user's speech input, given a set of predetermined words to read, in terms of its deviation from reference transcriptions via forced alignment.

1.4 Scope and Limitations of the Research

The system is specific to the Hiligaynon language. The words used in the audio data are limited to 2-3 syllable Hiligaynon words from the book "Hiligaynon Lessons" by Cecille L. Motus. Deviations are only measured word-by-word. In terms of the toolkit, the system is limited by the features offered by Kaldi - an open source speech recognition toolkit for speech recognition and signal processing.

1.5 Significance of the Research

This project aims to take one step towards developing a solution for improving the reading skill of Filipinos, particularly non reading adults. The authors also aim to contribute to the growing efforts of including Philippine local languages, specifically, Hiligaynon in literatures related to speech processing, particularly those relevant to the development of automated reading tutors.

Chapter 2

Review of Related Literature

2.1 Philippine Education Crisis

The Philippines has faced a persistent and pervasive educational crisis, as evidenced by low literacy rates and learning outcomes, particularly among disadvantaged and marginalized groups. As mentioned earlier in the introduction, UNICEF, UNESCO, and the World Bank found that only 10% of children could read simple text as of March 2022 in the Philippines. The World Bank also found that 90% of children in the Philippines cannot read simple text by age 10. These statistics highlight the urgent need to address the education crisis in the Philippines and ensure that children have access to quality reading instruction.

Pascual and Guevara (2017) conducted a study evaluating the performance of a reading miscue detector and automated reading tutor for Filipino, a language spoken in the Philippines. The study was conducted with a group of elementary

school students in the Philippines, and the results showed that the reading miscue detector and automated reading tutor were effective in improving reading skills. The students who used the technology demonstrated significant improvements in reading fluency, accuracy, and comprehension, compared to a control group. These results highlight the potential of technology-based reading tutors to improve reading skills among children in the Philippines.

2.2 Speech Recognition and Reading Miscue Detection

Reading miscue detection tasks inevitably borrow concepts from the development of speech processing technology, particularly speech recognition. Automatic Reading Tutors, such as the one developed by Pascual and Guevara (2017), are machine-aided systems designed to help its users improve their skill in reading by offering help or guidance when it detects reading miscues or disfluencies in user input reading speech. Part of the approach in their system involved deriving the phone symbol sequence corresponding to the input speech, which they force aligned with a reference speech to determine deviations based on a computed likelihood score. This process can benefit from speech recognition, where various types of acoustic and/or language models, especially machine learning models are used to make sense of acoustic signals by extracting features from the said signals. These features can then be used as inputs for analysis or to whatever tasks the authors deem to be appropriate. A study by Rasmussen, Tan, Lindberg, and Jensen (2009) also used an ASR component in their system for detecting miscues

in dyslexic read speech.

In their study "Listen, Attend and Spell," Chan, Jaitly, Le, and Vinyals (2015) proposed a neural network architecture for automatic speech recognition (ASR) that they dubbed the "Listen, Attend and Spell" (LAS) model. The LAS model was designed to be a more efficient and accurate ASR system, particularly for languages with limited data availability. The LAS model utilizes an attention mechanism, which allows the model to focus on specific parts of the input audio, rather than processing the entire audio signal at once. This allows the model to better handle the variability and noise present in real-world audio, and it enables the model to learn more efficiently and accurately. The LAS model was tested on several datasets and outperformed existing ASR systems, demonstrating its effectiveness for automatic speech recognition tasks.

Indeed, speech recognition is one key component of computer-assisted language learning systems.

2.3 The Language Dilemma

While there is a wealth of English-oriented ASR systems, other languages, especially lesser known languages tend to struggle in these situations. For instance, most leading tech companies tend to focus on developing speech recognition technologies for the English language such as DeepSpeech, Mozilla's speech to text engine and OpenAI's Whisper. In the context of Philippine languages, research in speech processing technologies for the Filipino language is by no means a desert, however developing efficient ASR systems for the said language have yet to be

seen, as noted by Dimzon and Pascual (2020) . For instance the previously mentioned authors — guided by the motivation to fill in knowledge gaps in Filipino phoneme recognition — were able to develop an "Automatic Phoneme Recognizer for Children's Filipino Read Speech". Additionally, Aquino, Tsang, Lucas, and de Leon (2019) was able to develop a system using a grapheme to phoneme (G2P) approach, in conjunction with selected ASR models which have been found out to be just as effective as human transcribers. Other local languages, however, are challenged by limited resources but efforts are underway. Billones and Dadios (2014) conducted a study, where they created a 5-word vocabulary speech recognition system for Hiligaynon terms used as motion commands implemented for a breast self-examination (BSE) multimedia training system.

2.4 Kaldi ASR Toolkit

Povey et al. (2011) described Kaldi as a modern toolkit for speech recognition. It is designed to be extensible and has one of the least restrictive licenses making it more accessible. Several studies have incorporated Kaldi into their implementations.

For instance, Upadhyaya, Farooq, Abidi, and Varshney (2017) , developed a continuous Hindi speech recognition model using Kaldi, citing the toolkit for its ability to create high quality lattices and sufficient speed for real time recognition. It also said that the mentioned toolkit is actively maintained and accessible.

Additionally, not only is Kaldi able to support conventional models such as gaussian mixture models (GMMs) but is also able to implement deep neural network

based structures. For example, Kipyatkova and Karpov (2016) developed a "DNN-Based Acoustic Modeling for Russian Speech Recognition Using Kaldi." The paper mentioned using DNN implementations in Kaldi, ultimately choosing Dan's implementation because of its support for parallel training on multiple CPUs.

2.5 The Hiligaynon Language

Hiligaynon, also known as Ilonggo, is an Austronesian language spoken in the Western Visayas region of the Philippines, particularly in the provinces of Iloilo, Guimaras, Negros Occidental, and Capiz. It is one of the major languages of the Philippines, spoken by millions of people as a first or second language.

Hiligaynon has a rich and varied vocabulary, with many loanwords from Spanish, English, and other languages. According to Hiligaynon Reference Grammar by Wolfenden (2019), Hiligaynon has a complex verb conjugation and tense system, with a range of tense markers including markers for past, present, and future tense, as well as markers for perfective and imperfective aspect. The book also notes that Hiligaynon has a number of mood markers, including markers for indicative, imperative, and subjunctive mood.

Additionally, Hiligaynon Reference Grammar by Wolfenden (2019) describes the phonemic alphabet of Hiligaynon as consisting of 28 letters: A, B, C, D, E, F, G, H, I, J, K, L, M, N, Ñ, O, P, Q, R, S, T, U, V, W, X, Y, and Z. The book notes that the letters C, F, J, Q, V, X, and Z are not used as frequently in Hiligaynon as in other Philippine languages, and that the letter Ñ is used to represent the Spanish sound "ny."

Table 2.1: Table of Hiligaynon-specific phonemes used in training the system's acoustic model (Gavieta, et al., 2022, p. 20)

Phone Class	Phones/Diphone
Bilabial stops	/p/, /b/
Dental stops	/t/, /d/
Velar stops	/k/, /g/
Africate	/j/
Fricatives	/s/, /sh/, /v/, /z/, /f/
Nasals	/m/, /n/, /ng/
Liquids	/l/, /r/
Semivowels/Glides	/w/, /y/
Vowels	/i/, /e/, /a/, /o/, /u/
Diphones	/ha/, /he/, /hi/, /ho/, /hu/, /at/, /aw/, /ay/, /oy/

Chapter 3

Research Methodology

This chapter lists and discusses the specific steps and activities that will be performed to accomplish the project.

3.1 Research Activities

3.1.1 Data Gathering

This chapter presents the research methodology employed by the researchers in conducting their study on developing a reading miscue detector using Hiligaynon words. The methodology includes the selection of the words, the creation of a dictionary, the selection of the speakers, the equipment used, and the recording setup.

Word Selection and Dictionary Creation

The researchers selected one thousand words from a corpus given by their thesis adviser. These words were limited to two to three syllables and were grade-appropriate. To ensure consistency in the pronunciation of the words, a dictionary was created for the Hiligaynon words. The phonemes used in the dictionary were based on the study conducted by Gavieta et al entitled Hilispeech: A Hiligaynon Speech Recognition System.

Selection of Speakers and Equipment Used

To gather the audio files needed for the study, six speakers were selected. Three of the speakers were the researchers themselves, while the other three were chosen based on availability. All of the speakers were native Hiligaynon speakers to ensure that the pronunciations were accurate. The researchers used noise canceling microphones to record the speakers.

Script Creation and Recording Setup

For each speaker, a script was created by randomly selecting five hundred words. The recording was done in a closed quiet room to ensure minimal background noise. Each audio file contained only 25 words, and the duration of the audio files was not limited. To ensure easy organization and identification of the audio files, the researchers came up with a naming system that includes the gender of the speaker, the speaker number, and the audio file number.

3.1.2 Preprocessing

The open-source digital audio editor Audacity was used for preprocessing the audio data. The data was first compressed using Audacity's compression effect, and then normalized using Audacity's Normalize effect.

Leading and trailing parts of the audio were then cut to remove silent parts at the beginning and ending of the recording.

The audio files were then saved to WAV format.

3.1.3 Acoustic Modelling

The Kaldi ASR toolkit was used for building the acoustic model.

Files for the project was placed inside a folder of the same name under Kaldi's 'egs' directory. All of the recordings were placed under the directory data/audio (assuming readable as root directory).

Following a 6-fold training and testing scheme, six folders were created for each fold. Each 'fold' folder contained metadata for the files corresponding to each fold. The following metadata files was created:

- **wav.scp:** This file contains information about each file's file id and where the file is located. It contains data int the format of <file _id ><path _to _file >
- **text:** This file contains information about the file and the corresponding words uttered in that particular file's audio recording. It is writte in the

format <utterance _id ><series _of _words >

- **utt2spk:** This file contains information about the mapping of a specific file to it's corresponding speaker. It is written in the format <utterance _id ><speaker _id >

- **spk2gender:** This file contains information indicating a specific speaker's gender. It is written in the format <speaker _id ><gender >

The following metadata which are not associated with a specific fold is also created.

- **corpus.txt** This file contains all the words uttered in all of the recordings. Each line represents the words uttered in a specific file. This was placed under the data/local directory.

- **lexicon.txt** This file contains information about all the words considered in the project's dictionary together with their phonemic transcriptions. Also included are the silence phones.

- **nonsilence _phones** This file contains all of the non silent phones considered by the project

- **(silence _phones.txt and optional _silence.txt** These files contains the silence phones included in the project.

Training scripts were sourced from Kaldi's builtin scripts for different training algorithims namely: monophone, triphone, LDA+MLLT, LDA+MLLT+SAT and DNN.

Training was done for each of the six folds, with each training algorithm applied to each fold. The best results were then noted for each training algorithm for each fold.

3.1.4 Evaluation

Evaluation was done by comparing the word error rate (WER) and sentence error rate (SER) for each model for each fold, across a 6-fold cross validation test.

Chapter 4

Results and Discussions/Analyses

This chapter provides a summary and analysis of the results obtained by decoding different training models, such as monophones, triphones (delta, delta + delta-delta), LDA + MLLT, SAT, and DNN. The information presented includes data extracted from each decoding model's WER files output, including the number of errors made (insertions, deletions, and substitutions), as well as the word and sentence error rates (WER and SER), and the average percentage of errors for each model.

In order to have a reliable result, this project was tested through 6-fold cross-validation of every training model. There is a total of six speakers for this project. Each speaker is subjected to testing for every fold while the remaining speakers that are not chosen are set to be on the training part. Succeeding tables will present the results gathered through this process.

4.1 Mean results of five models

Table 4.1: Mean WER scores of the different models used for acoustic modelling

Model	Mean WER
Monophone	4.69
Triphone	14.46
LDA + MLLT	17.34
LDA + MLLT + SAT	2.13
DNN	0.69

Table 4.1 shows the different mean word error rate of every acoustic model through six folds. The LDA + MLLT got the highest word error rate which is 17.34% , followed by triphone model with mean WER of 14.46%, monophone model got 4.69% mean WER, LDA + MLLT + SAT model with 2.13 % mean WER, lastly, the DNN model got the lowest mean WER which is 0.69% among the five models. The numbers have proved that the DNN model is the best choice for this project.

4.2 DNN results through 6-fold cross-validation

	Mean WER	Mean SER
DNN Fold 1	0.78	19.55
DNN Fold 2	1.0	20.00
DNN Fold 3	0.95	21.36
DNN Fold 4	0.00	0.00
DNN Fold 5	0.24	5.91
DNN Fold 6	1.15	16.82

Table 4.2: DNN model result across 6-folds

Data from Table 4.2 presents the different word error rates and sentence error rates of the DNN model through each fold. Fold 6 DNN model have the highest

mean WER which is 1.15%, followed by fold 2 with mean WER of 1.00%, next is fold 3 with 0.95% mean WER, fold 1 with mean WER of 0.78%, fold 5 DNN model with 0.24% mean WER, and lastly, fold 4 got the lowest mean word error rate which is 0.00%. Fold 4 comprises the fourth speaker on testing and the rest of the speaker is on the training part.

4.3 DNN result in fourth fold

	WER	SER	Insert	Delete	Substitution
wer_7	0.00	0.00	0	0	0
wer_8	0.00	0.00	0	0	0
wer_9	0.00	0.00	0	0	0
wer_10	0.00	0.00	0	0	0
wer_11	0.00	0.00	0	0	0
wer_12	0.00	0.00	0	0	0
wer_13	0.00	0.00	0	0	0
wer_14	0.00	0.00	0	0	0
wer_15	0.00	0.00	0	0	0
wer_16	0.00	0.00	0	0	0
wer_17	0.00	0.00	0	0	0
MEAN	0.00	0.00			

Table 4.3: DNN model result in fold 4

Table 4.3 has the data gathered for the fourth fold DNN model after training the acoustic model. Each value for this fold is zero, including the WER and SER, and the insertion, delete, and substitution columns. The result implies that the fourth fold of the DNN model has no word or sentence errors and it did not detect either insertion, deletion, or substitution. Thus, among the different folds of the DNN model, the fourth fold is the most fitting model for this project.

Chapter 5

References

Aquino, A., Tsang, J. L., Lucas, C. R., & de Leon, F. (2019, 8). G2P and ASR techniques for low-resource phonetic transcription of Tagalog, Cebuano, and Hiligaynon. *2019 International Symposium on Multimedia and Communication Technology (ISMAC)*. Retrieved from `http://dx.doi.org/10.1109/ismac.2019.8836168` doi: 10.1109/ismac.2019.8836168

Billones, R. K. C., & Dadios, E. P. (2014, 11). Hiligaynon language 5-word vocabulary speech recognition using Mel frequency cepstrum coefficients and genetic algorithm. *2014 International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*. Retrieved from `http://dx.doi.org/10.1109/hnicem.2014.7016247` doi: 10.1109/hnicem.2014.7016247

Chan, W., Jaitly, N., Le, Q. V., & Vinyals, O. (2015, 8). Listen, Attend and Spell. *arXiv: Computation and Language*.

Dimzon, F. D., & Pascual, R. M. (2020, 12). An Automatic Phoneme Recognizer for Children's Filipino Read Speech. *2020 IEEE International Confer-*

ence on Teaching, Assessment, and Learning for Engineering (TALE). Retrieved from http://dx.doi.org/10.1109/tale48869.2020.9368399 doi: 10.1109/tale48869.2020.9368399

Hernandez, J. (2020, 10). *Literacy rate estimated at 93.8PSA.* Retrieved from https://www.bworldonline.com/economy/2020/10/29/325932/literacy-rate-estimated-at-93-8-among-5-year-olds-or-older-psa/

Kipyatkova, I., & Karpov, A. (2016). DNN-Based Acoustic Modeling for Russian Speech Recognition Using Kaldi. *Speech and Computer*, 246–253. Retrieved from http://dx.doi.org/10.1007/978-3-319-43958-7_29 doi: 10.1007/978-3-319-43958-7\{_}29

Pascual, R., & Guevara, R. (2017). Experiments and Pilot Study Evaluating the Performance of Reading Miscue Detector and Automated Reading Tutor for Filipino: A Children's Speech Technology for Improving Literacy. *Science Diliman*, *29*(1), 5–36. Retrieved from https://journals.upd.edu.ph/index.php/sciencediliman/article/view/5622

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N. K., ... Vesely, K. (2011, 1). The Kaldi Speech Recognition Toolkit. *IEEE Automatic Speech Recognition and Understanding Workshop*. Retrieved from https://publications.idiap.ch/downloads/papers/2012/Povey_ASRU2011_2011.pdf

Rasmussen, M. H., Tan, Z.-H., Lindberg, B., & Jensen, S. H. (2009, 9). A system for detecting miscues in dyslexic read speech. *Interspeech 2009*. Retrieved from http://dx.doi.org/10.21437/interspeech.2009-448 doi: 10.21437/interspeech.2009-448

UNICEF, UNESCO, & Bank, W. (2022, 3). *Where are we on Education Recovery?*

(Tech. Rep.). Retrieved from https://www.unicef.org/reports/where-are-we-education-recovery

Upadhyaya, P., Farooq, O., Abidi, M. R., & Varshney, Y. V. (2017, 3). Continuous hindi speech recognition model based on Kaldi ASR toolkit. *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. Retrieved from http://dx.doi.org/10.1109/wispnet.2017.8299868 doi: 10.1109/wispnet.2017.8299868

Wolfenden, E. (2019). *Hiligaynon Reference Grammar* (Open Access ed.). University of Hawaii Press. Retrieved from https://core.ac.uk/display/211329359