I-Sueat

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

CARPIO, Joecel Eman
GARCIA, Michael John
RABE, Jett Adriel

Francis Dimzon
Adviser

December 14, 2021

Abstract

This paper describes the integration of automatic speech recognition toolkits namely "Kaldi" on languages with little to no existing entries in widely available language databases. The goal of this research is to transcribe naturally spoken language data in the "Aklanon" language automatically.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Insert Title here

With the rapid advancement of computer science and computational linguistics, numerous technologies are now being applied in various real-world situations. Among them is the interdisciplinary subfield known as speech recognition. Also known as automatic/computer speech recognition (ASR) or speech-to-text (STT), it utilizes computers in order to recognize and translate natural spoken language into text.

One of the key applications of automatic speech recognition is to transcribe speech documents such as talks, presentations, lectures, and broadcast news (Furui, Kikuchi, Shinnaka, & Hori, 2004). A known challenge in speech transcription is that it can be quite taxing to retrieve and reuse speech documents if they are only recorded as audio. Although high recognition accuracy can be easily obtained for speech read from a text, such as anchor speakers' broadcast news utterances, technological ability for recognizing spontaneous speech is still limited (Furui et al., 2004).

About 4500 languages exist in the world, but the majority of languages are spoken by less than 100,000 speakers; only about 150 languages (3%) have more than 1 Million speakers (Schultz, 2002).

Aklanon, which is often spelled as "Akeanon" by its local writers, is a dialect spoken by people located in the province of Aklan on the island of Panay in the Philippines. It somewhat varies with the dialects of neighboring provinces and islands and it belongs to a family of dialects whose ancestor might be proto-

West Visayan, which in turn was a member of the Malayo-Polynesian family of languages, to which such languages as Tagalog and Cebuano belong (De la Cruz & Zorc, 1968).

Aklanon/Akeanon is a specific language that is mostly exclusive to people who have lived in Aklan therefore most Filipino citizens wouldn't have familiarity with it. This would prove to be problematic since this would limit possible communications between locals and others. Additionally, there are barely any applications that are able to promote the language and get it out to the public. A speech-to-text recognition system would pave the way for the language to be recognized and appreciated.

This paper presents the development of a speech-to-text system that would be able to recognize Akeanon words with a decent accuracy rating. The system would be made using the open-source speech recognition toolkit "Kaldi". Kaldi is a speech recognition toolkit written in C++ and licensed under the Apache License v2.0. The system would be developed in a way such that the audio, acoustic and language data would be catered to the Akeanon dialect.

1.2 Problem Statement

The creation of an automatic speech-to-text system for the Aklanon language, a language that is not widely used will demonstrate the capabilities of automatic speech recognition toolkits such as "Kaldi". Since there is no currently available research on speech-to-text for the Aklanon language, this initial research will serve as an aid or foundation for other speech recognition systems targeting other possible local languages. The existence of speech recognition systems for the Aklanon language will also aid in transcription of language data and therefore easier translation of information and cultural knowledge

1.3 Research Objectives

1.3.1 General Objective

The general objectives of this research include creating a fully functional speech-to-text system with an adequate level of accuracy when given audio or language data as input.

1.3.2 Specific Objectives

A. To review related literature on existing speech-to-text or automatic speech recognition toolkits and compare them with each other to determine which toolkit will be appropriate for the research problem.

B. To gather audio or language data from different speakers and convert them to appropriate bit rates and frequencies wherein the toolkit will recognize.

C. To develop a system that will utilize the chosen toolkit to recognize simple words from the Aklanon language.

1.4 Scope and Limitations of the Research

The usage of "Kaldi" as opposed to other speech recognition toolkits limits the prototype to only the features that the toolkit has to offer. However, "Kaldi" is more accessible and more supported among any other which makes it ideal as the one to be utilized.

The audio samples used for the speech recognition toolkit were taken from just ten different people. The participants who were recorded for the audio samples used were mostly Aklanon native speakers. This was limited to just speakers of this particular language in order to keep the accuracy of the pronunciations of the words as precise as possible. Moreover, rules on pronunciation for this language could pose difficulties for non-native speakers such as the consonant "e".

With regards to the words that the prototype can recognized, it has only been limited to a few simple words. These include numbers, basic directions and common objects.

1.5 Significance of the Research

Currently, there is no speech recognition software available for the Aklanon language. This study would pave the way for future projects involving Aklanon speech-to-text. The prototype could also serve as a foundation for these future studies as they would expand further on the findings here.

Considering that Aklanon is similar to its neighboring languages or dialects with only pronunciation and diction being one of its key challenging differences,

this opens the door for its respective neighbors to be able to adapt a speech-to-text like this for their own language. This enables access for people to speech recognition technology as more languages would possibly be incorporated in the future.

Chapter 2

Review of Related Literature

2.1 Automatic Speech technology

Automatic speech recognition (ASR) by machine has been a field of research for more than five decades. The industry has developed a broad range of commercial products where ASR as user interface has become ever more useful and pervasive (Li, Deng, Gong, & Haeb-Umbach, 2014). Automatic Speech Technologies are used daily in several applications and services. However, ASR technology has not reached a point where computers understand all speech in any acoustic environment or by any person (Rabiner, 1993).

The basic architecture of ASR involves four main components: signal processing and feature extraction, acoustic model (AM), language model (LM), and hypothesis search. Audio signals which are taken as input for signal processing and feature extraction components are enhanced and converted so it becomes suitable for acoustic models. The acoustic models are responsible for integrating knowledge about acoustics and phonetics. The language model processes the probability of a hypothesized word sequence from what it has learned from training. The hypothesis search component will combine the previous scores from both AM and LM and outputs a word sequence (Yu & Deng, 2016).

2.2 Speech Recognition Toolkits

Over the last few decades, commercial proprietary speech recognition systems have emerged such as AT&T Watson .(Goffin et al., 2005), Microsoft Speech

Server (Dunn, 2007), Google Speech API (Adorf, 2013) and Nuance Recognizer. However, because these systems are difficult to incorporate into other software and provide little control over the recognizer's capabilities, open-source automatic speech recognition systems have emerged.

The training of linguistic and acoustic models for open-source ASR toolkits such as HDecode, Julius, pocketsphinx, Sphinx-4, and Kaldi was done in comparative research in 2014. Their trials revealed a ranking of the toolkits based on the effort-to-performance ratio. When given out-of-the-box training and decoding pipelines, the results showed that Kaldi outperformed all other recognition toolkits. The Sphinx also demonstrated that it could give training pipelines with a high chance of producing good results in a short period of time (Gaida et al., 2014).

2.3 About Kaldi

This study focuses on Kaldi as the toolkit of choice given the performance of Kaldi in comparison to other toolkits. Kaldi is an open-source voice recognition toolkit. It's developed in C++ and released under the Apache License Version 2.0. Kaldi's purpose is to create code that is both current and versatile, simple to comprehend, alter, and extend.

The following is a list of the important features of Kaldi:

- Integration with Finite State Transducers (FSTs): uses OpenFst toolkit as a library.
- Extensible design: algorithms are designed to be in generic form
- Open License: The Apache License belongs to one of the least restrictive licenses available.
- Complete Recipes: provides recipes for building speech recognition systems using widely available databases.
- Thorough testing: built with the goal of having corresponding test routines for nearly all code.

2.4 Etymology of the Aklanon Language

The Aklanon language is spoken and understood by roughly around 360,000 people who are residents of the province of Aklan, as well as those who reside near its borders. While it is almost impossible to track the actual stages of development of the language, in a book published by De La Cruz and Zorc on Aklanon Grammar (De la Cruz & Zorc, 1968), the language comes from a long history of evolution of proto-languages which now turned out to be the present day Aklanon.

The language itself has freely adopted English words. These words were simplified and have taken their own form of spellings. Aside from English words, there are also words borrowed from China, Spain and Japan that have been derived from to form Aklanon words (Salas Reyes et al., 1969).

2.5 Glottal Stop

The glottal stop presents a potential issue with speech to text recognition as native speakers of the Aklanon language, like most of its neighboring dialects, generally have a system wherein the words do not actually sound like they are spelled. These glottal stops have different rules concerning their positions such as initially before a vowel, medially between vowels, consonants or when a double glottal appears, and lastly the glottal at the final position (Salas Reyes et al., 1969). Depending on how an individual might pronounce certain words, the results from speech to text recognition could possibly vary as different stress intonations on several words could result in a mismatch in recognition.

2.6 The Consonant "e"

As mentioned by Reyes and his team in their book The Aklanon Dialect, the letter "e" can have two different roles depending on its position in the word. It could either be treated as the standard vowel or a consonantal sound. If the letter "e" appears in an environment with a vowel, it is treated as if it's a consonant and therefore no longer pronounced as the vowel "e". This is because diphthongs do not exist in the Aklanon language which makes the consonantal "e" quite distinguishable (Salas Reyes et al., 1969).

This version of the "e" is often described as voiced velar fricative denoted be

the latinized variant of the symbol of the Greek letter gamma. In various words, it tends to replace the letter “l” coming from words of other languages that have the same word format.

Chapter 3

Research Methodology

The project was mainly based on the Kaldi tutorial page located in the Kaldi ASR documentation site (`https://kaldi-asr.org/doc/kaldi_for_dummies.html`). The page provides a brief introduction to Kaldi as well as the preferred instructions and environment necessary for Kaldi installation and development.

The prototype was set up in a Fedora Linux environment using baseplate files obtained from the Kaldi ASR repository in GitHub `https://github.com/kaldi-asr/kaldi.git`. The Aklanon speech-to-text recognition system was then placed and developed alongside other example scripts within the egs folder (kaldi/egs).

3.1 Audio Data

Currently, the words that are being implemented and tested in the project have been limited to digits. The overall corpus would eventually increase over time in order to include basic directions, common objects and regular sentences.

A set of 100 .wav audio files (formatted into 44.1 khz) were prepared wherein each file contains three spoken digits recorded in the Aklanon language, one by one. Each of these audio files were then placed and categorized representing a speaker. 10 different speakers were chosen, with each speaker saying 10 sentences/files which would then make up to 100 different audio files. One speaker was chosen to be tested whereas the remaining nine were placed in a separate category in order to train the ASR system.

Digits	Aklanon Representation
1	Isaea
2	Daywa
3	Tatlo
4	Ap-at
5	Li-ma
6	An-om
7	Pito
8	Waeo
9	Siyam
10	Pueo

Table 3.1: Numbers 1-10 along with their Aklanon representations

3.2 Acoustic Data

Various files were made according to the instructions provided in the Kaldi tutorial. These include the files spk2gender, wav.scp, text, utt2spk and corpus.txt. The file spk2gender specifies the genders of the speakers, wav.scp connects every utterance to an audio file on a given path, text contains every utterance matched with its corresponding text transcription, utt2spk specifies which utterance belongs to which speaker, and corpus.txt contains every single possible utterance transcription among the 100 files. These data files were then prepared and sorted.

3.3 Language Data

The language data files are necessary for the modeling of a language in Kaldi. These include the files lexicon.txt, nonsilence_phones.txt, silence_phones.txt and optional_silence.txt. The file lexicon.txt contains every word available in corpus.txt along with their phonemic representations, nonsilence_phones contain nonsilent phonemes, silence_phones contain silent phonemes, and optional_silence contains optional silent phonemes.

After preparing the various kinds of data, scripts were then executed in order to showcase the results of the testing and training of the speech-to-text system.

References

Adorf, J. (2013). Web speech api. *KTH Royal Institute of Technology*.

De la Cruz, B. A., & Zorc, R. (1968). A study of the aklanon dialect. volume one: Grammar.

Dunn, M. D. (2007). *Pro microsoft speech server 2007: Developing speech enabled applications with. net.* Springer.

Furui, S., Kikuchi, T., Shinnaka, Y., & Hori, C. (2004). Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Transactions on Speech and Audio Processing*, *12*(4), 401–408.

Gaida, C., Lange, P., Petrick, R., Proba, P., Malatawy, A., & Suendermann-Oeft, D. (2014). Comparing open-source speech recognition toolkits. In *11th international workshop on natural language processing and cognitive science.*

Goffin, V., Allauzen, C., Bocchieri, E., Hakkani-Tur, D., Ljolje, A., Parthasarathy, S., ... Saraclar, M. (2005). The at&t watson speech recognizer. In *Proceedings.(icassp'05). ieee international conference on acoustics, speech, and signal processing, 2005.* (Vol. 1, pp. I–1033).

Li, J., Deng, L., Gong, Y., & Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(4), 745–777.

Rabiner, L. (1993). Fundamentals of speech recognition. *Fundamentals of speech recognition*.

Salas Reyes, V., et al. (1969). A study of the aklanon dialect, volume two: Dictionary (of root words and derivations), aklanon to english.

Schultz, T. (2002). Globalphone: a multilingual speech and text database developed at karlsruhe university. In *Seventh international conference on spoken language processing.*

Yu, D., & Deng, L. (2016). *Automatic speech recognition.* Springer.