Hambaeon: Developing an Akeanon Speech and Text Corpus with Extensive Data Collection and Preliminary Speech Recognition Results

A Special Problem Proposal

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

VILLANUEVA, Joshua C.

VILLANUEVA, Mariefher Grace Z.

Francis D. DIMZON

Adviser

November 9, 2024

Abstract

From 150 to 200 words of short, direct and complete sentences, the abstract should be informative enough to serve as a substitute for reading the entire SP document itself. It states the rationale and the objectives of the research. In the final Special Problem document (i.e., the document you'll submit for your final defense), the abstract should also contain a description of your research results, findings, and contribution(s).

Suggested keywords based on ACM Computing Classification system can be found at `https://dl.acm.org/ccs/ccs_flat.cfm`

Keywords: Keyword 1, keyword 2, keyword 3, keyword 4, etc.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Speech-to-Text (STT) technology has rapidly evolved in recent years, especially with the emergence of deep learning algorithms, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs) that greatly enhanced the accuracy of STT systems (Televic, 2024). Open-source toolkits such as Kaldi have further accelerated research and development in this field by providing a flexible framework for building and training custom automatic speech recognition (ASR) models. These ASR systems are used to convert speech into text [17]. However, despite these advancements, only few Philippine languages have been explored and are integrated with this technology. For this special problem, the focus will be on one of the understudied Central Philippine languages, Akeanon (Wellstood, 2022).

Akeanon is an Austronesian language belonging to the Visayan subgroup (Biray, 2023). With more than 130,000 households (Authority, 2023) speaking the language, Akeanon is primarily spoken in the province of Aklan, located in northwestern Panay. Biray (2023) explains that the language has several dialects, each typically named after the town where it is spoken. These include Akeanon Buruangganon, Akeanon Nabasnon, Akeanon Bukidnon, and the standard Akeanon, which is spoken in Kalibo, the provincial capital of Aklan. For this special problem, the researchers will focus on developing the speech and text corpus specifically for standard Akeanon.

Up to this date, no studies have been conducted that is directly related to Akeanon and speech recognition altogether. However, there exist similar studies in the context of speech recognition on other Visayan languages such as Bisaya in the study of Cerna et al. (2023) and Hiligaynon, studied by Billones and Dadios (2014). This special problem aims to bridge the gap in speech recognition for Akeanon starting with establishing a foundational speech corpus for the language, which can lay the groundwork for future research and applications. The corpus development will draw on methodologies from similar studies conducted for Bisaya and Hiligaynon, adapting them to meet the specific needs of Akeanon. In doing so, the project aims to bring Akeanon closer to digital integration, promoting inclusivity in speech recognition technology for Philippine languages. By bridging this gap, this special problem aspires to create a resource that can benefit future ASR developments, language preservation efforts, and the broader field of computational linguistics.

Developing an STT system for Akeanon not only addresses the gap of lack of representation of the regional language but also contributes to the preservation

and digital inclusion of the language. This special problem also aims to create an ASR model that can accurately recognize Akeanon speech patterns and nuances, given the language's unique phonetic and linguistic characteristics. With the resources collated for this study, the researchers will use Kaldi toolkit to develop an ASR system that will provide initial speech recognition results for the Akeanon language. Lastly, the study seeks to explore the challenges involved in building speech models for low-resource languages, contributing insights to the broader field of speech technology development.

1.2 Problem Statement

Akeanon remains underrepresented in modern speech technologies. According to Khan et al. (2023), in machine learning, natural language can be categorized into two categories: low-resource languages (LRLs) and high-resource languages (HRLs). Among these resources are (a) collections of text in different formats, such as research papers, journal articles, social media content, etc.; (b) lexical, syntactic, and semantic resources, such as dictionaries, bag of words, semantic databases, etc.; and (c) task-specific resources, such as annotated text, machine translation corpora, part-of-speech tags, etc.. HRLs e.g. English, French, Japanese, etc., are languages that are highly accessible and have many data resources that can be used for natural language processing (NLP). LRLs, on the other hand, are understudied and have few data resources that can be utilized for NLP. Most regional languages in the Philippines are considered to be LRL, including the Akeanon language. Alejan et al. (2021) raised concerns on the Philippines' inclusion on a global list of the top ten "language hotspots", which

means that many of its languages are disappearing faster than they are being completely documented. Their study noted the global rate of language extinction, which is one in every two weeks. They also projected that around half of the 6,000 languages will become extinct by the end of the century, to which most of them are indigenous languages. According to Magueresse et al. (2020) [29], a language supported by NLP techniques can help preserve it from extinction. It will also make the language more available and accessible in digital format, which offers significant commercial value, societal purpose, and applications in a variety of domains (Tsvetkov, 2017).

This special problem aims to address the lack of resources, availability, and accessibility of the Akeanon language in, but not limited to, modern speech technologies by building and establishing a speech and text and corpus for the language. Additionally, by developing an ASR model specific for Akeanon, it would lay the foundation for future research in speech-to-text, and other modern speech technologies for the language. Lastly, this special problem seeks to inspire innovation and drive similar efforts to preserve and develop accessible language technologies for other regional languages in the Philippines.

1.3   Research Objectives

1.3.1   General Objective

The general objective of this study is to construct and establish a text and speech corpus the Akeanon language, and design and implement an automatic speech

recognition (ASR) system for the language using the Kaldi toolkit.

1.3.2 Specific Objectives

Specifically, the study targets to:

1. Construct Akeanon text and speech corpus.

2. Train and fine-tune ASR models on collected data on Akeanon language.

3. Evaluate model performance and optimize the parameters in terms of the Word Error Rate (WER).

1.4 Scope and Limitations of the Research

The system is specific to the standard Akeanon language that is spoken in Kalibo, Aklan, the provincial capital. It is limited to the standard Akeanon even though the language has several dialects, due to resource constraints and the need for specificity. The study is centered around gathering audio samples from native speakers of standard Akeanon to guarantee precision and uniformity in the training of the model. This does not include other variations or dialects that are based on standard Akeanon language such as Buruanganon, Akeanon Nabasnon, and Akeanon Bukidnon, which can have different and unique phonetic and lexical traits. The automatic speech recognition model will be constructed using the Kaldi toolkit, selected for its flexibility in low-resource language environments.

Nevertheless, the model's effectiveness might be influenced by the scarce availability of Akeanon data, potentially affecting its wide-ranging applicability.

1.5   Significance of the Research

Akeanon language, like many indigenous languages in the Philippines, lacks representation in digital technologies. Creating an automatic speech recognition (ASR) model for Akeanon language will help contribute to the preservation of the language in digital format, establishing a resource that will support documentation and education initiatives in the future. The dataset and model produced in the study of Akeanon language can act as a basis for further and additional linguistic research.

Akeanon and its incorporation in speech recognition technology fosters digital inclusivity. This enables Akeanon speakers to engage with technology in their mother tongue highlighting the areas in education, communication, and public service where language barriers are almost present when accessing the said areas. Mobile applications, virtual assistants and other tools can embed speech to text of Akeanon language to help enhance accessibility and boost engagement with technology.

The challenge faced and lessons learned from this study will help contribute to addressing the lack of representation of low-resource language in AI technology, aligning with the need for inclusivity in language processing (Poupard, 2024). This initiative will help in promoting linguistic diversity as well as safeguard cultural heritage through Akeanon speech recognition in technological advancement.

Poupard (2024) highlights that even minimal focus on languages with fewer resources can significantly influence their viability in an increasingly digital world where larger languages prevail.

References

Alejan, J. A., Ayop, J. I. E., Allojado, J. B., Abatayo, D. P. B., Abacahin, S. K. N., & Bonifacio, R. (2021, May). *Heritage language maintenance and revitalization: Evaluating the language endangerment among the indigenous languages in bukidnon, philippines.* Retrieved from `https://eric.ed.gov/?id=ED617996` (ERIC - Online Submission)

Authority, P. S. (2023). *Tagalog is the most widely spoken language at home (2020 census of population and housing).* Retrieved from `https://psa.gov.ph/content/tagalog-most-widely-spoken-language-home-2020-census-population-and-housing`

Billones, R. K. C., & Dadios, E. P. (2014). Hiligaynon language 5-word vocabulary speech recognition using mel frequency cepstrum coefficients and genetic algorithm. In *2014 international conference on humanoid, nanotechnology, information technology, communication and control, environment and management (hnicem)* (p. 1-6). doi: 10.1109/HNICEM.2014.7016247

Biray, E. (2023, 12). Derivational morphology features in common akeanon dialects. *International Journal of Language and Literary Studies*, *5*, 222-234. doi: 10.36892/ijlls.v5i4.1441

Cerna, P. D., Cascaro, R. J., Juan, K. O. S., Montes, B. J. C., & Caballero,

A. O. (2023). Bisayan dialect short-time fourier transform audio recognition system using convolutional and recurrent neural network. *International Journal of Advanced Computer Science and Applications*, *14*(3). Retrieved from `http://dx.doi.org/10.14569/IJACSA.2023.01403111` doi: 10.14569/IJACSA.2023.01403111

Khan, M., Ullah, K., Alharbi, Y., Alferaidi, A., Alharbi, T. S., Yadav, K., . . . Ahmad, A. (2023). Understanding the research challenges in low-resource language and linking bilingual news articles in multilingual news archive. *Applied Sciences*, *13*(15). Retrieved from `https://www.mdpi.com/2076-3417/13/15/8566` doi: 10.3390/app13158566

Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource languages: A review of past work and future challenges. *CoRR*, *abs/2006.07264*. Retrieved from `https://arxiv.org/abs/2006.07264`

Poupard, D. (2024). Attention is all low-resource languages need. *Translation Studies*, *17*(2), 424–427. Retrieved from `https://doi.org/10.1080/14781700.2024.2336000` doi: 10.1080/14781700.2024.2336000

Televic. (2024, 1). *The evolution of speech-to-text technology.* Retrieved from `https://www.televic.com/en/televicgsp/news/the-evolution-of-speechtotext-technology`

Tsvetkov, Y. (2017). *Opportunities and challenges in working with low-resource languages.* Retrieved from `https://www.cs.cmu.edu/~ytsvetko/jsalt-part1.pdf` (PDF)

Wellstood, Z. (2022). A relative clause analysis of event existential constructions in aklanon. *GLOSSA*, *7*(1). Retrieved from `https://www.glossa-journal.org/article/id/5866/` doi: 10.16995/glossa.5866