Extended Abstract: Verifying Urarina Language Phonemes With TensorFlow (A work in progress)

Michael Dorin
Graduate Programs in Software
University of St. Thomas
St. Paul, Minnesota
mike.dorin@stthomas.edu

Judith Dorin
Director of Prestigious Scholarships
University of St. Thomas
St. Paul, Minnesota
judith.dorin@stthomas.edu

Abstract

Between 1985 and 1987 a collection of audio cassettes and handwritten notes were created as part of study of the Urarina language. Attempts were made to augment the Urarina language documentation with information from this study, and the recordings and note cards were digitized. A challenging part of the documentation effort is validating the phonemes indicated for the recorded words. Human hearing and interpretation, especially as a person ages, can be subjective. This means it is necessary to perform this task via automated means. The availability of free and open machine learning software has substantially increased over the past few years and many solutions for classification are available. These solutions are easily implemented, and a background in artificial intelligence is no longer necessary. By converting the recorded words to image spectrographs, then using TensorFlow from Google combined with DeepImageSearch from Nilesh Verma, it is possible to indicate the start and end of phonemes within recorded words. This project supports Urarina language documentation and demonstrates the feasibility of inexpensive machine learning for linguistic research. Although this is a work in progress, the preliminary results are favorable.

1 Introduction

In 1987 students from the University of San Marcos completed a research project studying the Urarina language. These students were working to produce a thesis as part of the requirements for earning the title of Licenciado in Linguistics. The result was the creation of the thesis *Kacá eĵe (lengua urarina)/Aspectos de la fonología* (Cajas and Gualdieri, 1987).

The Urarina people live in the Chambira River Basin, and their population ranges between two and three thousand. There is evidence to classify this language as linguistically isolated from other regional languages. Urarina's typical and syntactic structure is OVS (OVS/VS)* (Olawsky, 2011).

Several aspects of the original study were left open for future research. One such study was completed by Beatriz Gualdieri who reviewed the original phonological analysis using updated theories more than twenty years after the original work (Gutiérrez et al., 2012).

As part of the newly found data review, it became apparent that an assessment of the documented phonemes would be useful. In an effort to minimize complexity and speed up the process, common off-the-shelf software (COTS) was selected to perform the analysis using machine learning.

There is abundant research describing machine learning speech recognition concepts, which are useful for this investigation. For example, an important article, ”Understanding Audio data, Fourier Transform, FFT and Spectrogram features for a Speech Recognition System” by Kartik Chaudhary, describes various techniques applicable to this project (Kartik Chaudhary, 2020).

This project restored and digitized the recorded audio and analyzed it using the COTS TensorFlow (Abadi et al., 2016). That the software works on the Urarina language likely means it is also applicable to other languages.

2 Materials and Methods

2.1 Preparation of the Data

All the cassettes' audio was digitized into waveform audio file format (WAV) files. Except for certain exceptions, such as numbers, each recorded word was spoken once in Spanish by a Spanish speaker, followed by a Urarina speaker giving the word in Urarina twice. Audacity, a free software tool, was used to extract individual words into separate WAV files (The Audacity Team, 1999).

2.2 Generation of Spectrographs

Spectrograms were created for each spoken word using the library Matplotlib employing the "Accent" style (The Matplotlib development team, 2022). According to Chaudhary, a human cannot speak more than one phoneme in a time window of 20 to 30 milliseconds. Chaudhary also suggests a window overlap of 25% to 75% when analyzing speech (Kartik Chaudhary, 2020). This project used a 20 millisecond window, with 10 milliseconds allotted on each side for overlap, giving a total window size of 40 milliseconds. Each full spectrograph was then partitioned into smaller images of the specified window size.

2.3 Spectrograph Analysis

A Python program was created using the open-source library DeepImageSearch (Nilesh Verma, 2021) coupled with TensorFlow (Abadi et al., 2016) to compare image segments to each other. DeepImageSearch is built upon Tensorflow and requires very little extra code to sort images by how closely they resemble one another (Nilesh Verma, 2021).

3 Results and Discussion

The created Python program separates the common audio segments found in each recorded word. These segments were then organized into groups. These groups are correlated back to the words, allowing for the identification of each word's phonemes. In addition, the timing of the start and end of each phoneme is possible, signifying the additional possibility of verification of phonotactics.

3.1 Conclusion

Though the initial results of this updated work are encouraging, more data needs to be analyzed to move forward correctly. The data from the 1987 study can be used for more than phoneme validation; it can also be used to support dialectological analysis and research of phonological changes in the language.

The 1987 study did not have computer technology readily available to assist or validate their work. Though the participants were aware of the important prosodic features involved in the language, they were not practical to explore at the time of that study. Using COTS for phoneme identification and other aspects of language analysis helps support data overall for Urarina documentation. This project also shows a path for COTS use in supporting research of other endangered languages.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. {TensorFlow}: a system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283.

Judith Cajas and Cecilia Gualdieri. 1987. *Kacá eȇje (lengua urarina)/Aspectos de la fonología*. Ph.D. thesis, Thesis Licenciatura, UNMSM, Lima, Peru.

Analía Gutiérrez, Hebe A González, and Beatriz Gualdieri. 2012. La metátesis como un fenómeno fonológico: El caso nivacle. *Volúmenes temáticos de la Sociedad Argentina de Lingüística Serie 2012*, page 75.

Kartik Chaudhary. 2020. Understanding audio data, fourier transform, fft and spectrogram features for a speech. `https://towardsdatascience.com/understanding-audio-data-fourier-transform-fft-spectrogram-and-speech-recognition-a4072d228520`, Last accessed on 2022-10-9,.

Nilesh Verma. 2021. Deep image search. `https://github.com/TechyNilesh/DeepImageSearch`, Last accessed on 2022-10-9,.

Knut J Olawsky. 2011. *A grammar of Urarina*, volume 37. Walter de Gruyter.

The Audacity Team. 1999. Audacity software. `https://https://www.audacityteam.org/`, Last accessed on 2022-10-9,.

The Matplotlib development team. 2022. Matplotlib: Visualization with python. `https://https://matplotlib.org/`, Last accessed on 2022-10-9,.