Audio Classifier for Endangered Language Analysis and Education

Meghna Reddy and Min Chen [0000-0002-2370-6255]

University of Washington Bothell, Bothell WA 98011, USA
{meghnard, minchen2}@uw.edu

Abstract. Around 42% of the world languages are considered endangered due to the decline in the number of speakers. MeTILDA (Melodic Transcription in Language Documentation and Application) is a collaborative platform created for researchers, teachers, and students to interact, teach, and learn endangered languages. It is currently being developed and tested on the Blackfoot language, an endangered language primarily spoken in Northwest Montana, USA and Southern Alberta, Canada. This study extends MeTILDA functionality by incorporating machine learning framework in documenting, analyzing, and educating endangered languages. Specifically, this application focuses on two main components, namely audio classifier and language learning. Here, the audio classifier component allows users to automatically obtain instances of vowels and consonants in Blackfoot audio files. The language learning component enables users to visually study the pitch patterns of these instances and improve their pronunciation by comparing with that of native speakers using a perceptual scale. This application reduces manual efforts and time-intensive tasks in locating important segments of Blackfoot language for research and educational purpose.

Keywords: Audio Classification, Endangered Language Education, MeTILDA.

1 Introduction

42% of the world languages [1] are currently endangered. Language endangerment is considered one of the most urgent problems facing humanities [2] because endangered languages represent a vast repository of human knowledge about the natural world and cultural traditions which is irreplaceable. Linguists have responded to this issue through a renewed commitment towards language documentation and education [3].

This paper presents a collaborative research study that aims to document and educate Blackfoot, an endangered language with around 3,250 active speakers [4]. Blackfoot is a pitch accent language where words with the same sound sequence can convey different meanings when changing in pitches, e.g., ápssiw means 'it is an arrow' while apssíw means 'it is a fig; snowberry.' Studies have shown the important role of rhythm and melody in language acquisition, especially with pitch-accent endangered languages like Blackfoot [5]. In our work, we developed an audio classifier that au-

tomatically identifies vowels and consonants instances, and a cloud-based platform called MeTILDA (Melodic Transcription in Language Documentation and Application) to help Blackfoot language education.

2 Related Work

Many studies have been done to classify audio for various purposes, such as speech classification [6], event detection [7], and music recommendation [8]. Some work focuses on extracting useful audio features including low-level spectral and temporal features [6], mid-level featured such as Mel-Frequency Cepstral Cooefficients (MFCCs), spectrograms and mel-spectrograms [8]. Other work aims to develop effective machine learning or deep learning models. For example, Artificial Neural Networks (ANN) [7], Convolution Neural Networks (CNN) and Deep Belief Networks [9] have demonstrated promising results in automating the audio classification process. However, there has been limited research in audio classification for language education. In addition, as a pitch accent language, Blackfoot audio classification and language education faced several unique challenges including limited audio dataset for training, varied lengths of vowels and consonants, and impact of pitch on word meanings. Therefore, models developed for other purposes fail to support effective Blackfoot vowel/consonant classification.

In addition, existing software systems offered limited support to address the urgent need of endangered language education. Most existing linguistics tools such as FLEx[1], ELAN[2], and Praat[3] provide essential support for linguistic research but are less effective for language education, while language learning tools such as Babbel[4], Rosetta Stone[5] largely focus on commonly spoken languages. The support is even more inadequate for pitch accent languages. Research has shown that pronunciation learning technique is significantly understudied in Indigenous languages [10]. Pitch movements are often not explicitly represented in instructions and remain unclear to learners.

3 Audio Classifier and MeTILDA Framework

In our study, we developed an audio classifier to automatically extract vowels and consonants from Blackfoot speech recordings, which can be input into the MeTILDA platform for users to visually study pitch patterns and improve pronunciations. MeTILDA uses a perceptual scale developed in our earlier study [11] to help researchers and learners “see what they hear.” It consolidates and automates the workflow of generating perceived pitch changes of words in visual aids, called Pitch Art.

[1] fieldworks.sil.org
[2] https://archive.mpi.nl/tla/elan
[3] https://www.fon.hum.uva.nl/praat/
[4] https://www.babbel.com/
[5] https://www.rosettastone.com/

This visual diagram would allow Blackfoot teachers and learners to understand how their pronunciation compares to that of native speakers.

3.1 Audio classifier

The audio classifier is trained and tested on Blackfoot speech recordings and their associated TextGrid annotation files provided by our collaborators[6].

Data Processing and Feature Extraction. The first step is to process the data and extract audio features for machine learning model development. To automate this process, we incorporate "praatIO"[7] in our system to split training data into segments of vowels and consonants based on timestamps specified in the TextGrid annotation files. For testing data, a challenge is to split speech recordings into small segments with appropriate lengths whereas the lengths of vowels and consonants can be varied in Blackfoot speech. In the literature, attempts were made to identify the lengths of vowels and consonants in languages spoken in Australia and Philippines [12]. Most fall within a band of 50 milliseconds (ms) to 250 ms. This information is useful and matches with our observations on Blackfoot vowels and consonants based on speech recordings. Specifically, most vowels vary between 100 - 150 ms in length and most of the consonants between the 100 – 180 ms. Therefore, currently 100 ms is used as the window size for the testing data.

Various features are then extracted from audio segments including MFCCs, spectrograms and low-level features. Here, MFCCs are a set of coefficients that make up the mel-frequency spectrum to represent an audio signal's overall shape and frequency-based features [13]. A spectrogram is a visual representation of an audio unit with its most important frequency-based features taken into consideration. It works best as an input for deep learning models as image data. In addition, we also explored a set of low-level features including chroma STFT, root-mean-square (RMS) value, spectral centroid, spectral bandwidth, spectral rolloff and zero crossing rate [14]. We developed tools that allow users to select different combinations of audio features from the above-mentioned set for further processing and re-training models.

Model Training. After feature extraction, we explore different deep learning models including ANN, CNN, deep belief network for vowel/consonant classification. ANNs have proven to perform well for regression and classification applications ranging from banking to time-series forecasting and medicine. Sigmoid activation function is used in our study as it works best with binary classifiers [7] such as ours with 2 possible output classes (vowel or consonants). CNNs use images as inputs for training. In our application, spectrograms of audio units are a good visual representation to be used to train a CNN. In terms of deep belief network, in our previous study, it has been used successfully in identifying important Blackfoot sounds of 'h' and 'okii' as part of the DeepAudioFind project [15]. It is therefore used in this study to

[6] Dr. Mizuki Miyashita, a professor of linguistics at University of Montana (UM) and Mr. Naatosi Fish, a Blackfoot community linguist.

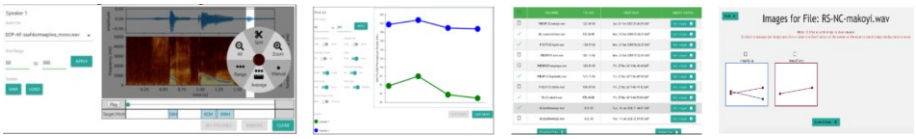
[7] https://github.com/timmahrt/praatIO

serve as a baseline. The classification results can be passed into MeTILDA for language learning.

3.2 MeTILDA

In MeTILDA, with the goal to help researchers and learners "see what they hear," we incorporated a perceptual scale developed by our previous study [11]. This scale provides a common reference for comparing pitch across recordings, regardless of the speakers' natural pitch. MeTILDA is hosted on the cloud[8] and the current features are organized into three components: Create, Learn, and Login/Team Project.

Create: The *Create* page allows users to upload recordings to the cloud database and select sound file(s) from the database for further processing. As shown in Fig. 1 (a), when a sound file is uploaded or selected, its waveform and spectrogram are shown on the screen. Users can then identify vowel or consonant instances based on the input from audio classifier or adjust the duration of the instances using the tools provided. F0 measurements are then computed using the perceptual scale. User can select the Average button, shown on a pie menu or enter orthographic symbols for the syllables, based on which, Pitch Art is automatically generated in a drawing window. MeTILDA allows the measurement of multiple recordings on the same page (as in Fig. 1 (b)) and prints word melodies of all selected recordings. This feature is useful for comparing the pitch movements produced by native speakers and learners. Other features offered by the *Create* page include saving Pitch Art images, listening to only the tones of the word melody, and toggling a variety of appearance options (e.g., displaying syllable text, showing pitch in an F0 contour instead of averaged, showing pitch in hertz instead of the psychoacoustic scale). Fig. 1 (c-d) show the My Files and History pages: Once the Pitch Art is drawn, users can save the uploaded sound files as well as measurement data (c) and Pitch Art images (d) for future access, all in the cloud database.



(a) Measuring and labeling (b) Pitch Art (c) Data (d) Saved Pitch Art

Fig. 1. *Create* page images

(a) Target word & Pitch Art (b) Practice & Recording

Fig. 2. *Learn* page images

[8] https://metilda.herokuapp.com/

Learn: The *Learn* page supports practice on word pronunciations. As shown in Fig. 2 (a), users may choose a syllable pattern, each containing words for users to select and practice. Users can listen to a native speaker's pronunciation or pitch tones for reference, and they can record their own pronunciations. As in Fig. 2 (b), the user's pitch track appears in a dotted line, enabling visual comparison with the model Pitch Art. Each previously recorded sound is presented, with an option to play or pause. The *Learn* page also has a View Students option so the tool can be used as a course supplement, allowing teachers to view their students' pitch tracks.

Login/Team Project: Users can obtain access to MeTILDA by logging in based on their user category: researcher, teacher, student, or other. Different users have varied access privileges to data in the system. For example, researchers and teachers can access all features: measurement tools, Pitch Art creation, saving, and stored data. Students have access to play, record, and submit in the *Learn* page only. Teachers can view their students' submitted work, while students cannot see each other's work.

4 Experimental results

Performance comparison is done on different audio classifiers using various machine learning models with a combination of features. The dataset provided by the collaborators includes 4,852 audio segments with 2,211 vowels and 2,641 consonants. Using 10-fold cross validation scheme, two top performers are listed in Table 1. As a comparison, using the baseline model from our previous study [15] only achieved 77% in Precision and Recall. We want to acknowledge the dataset is relatively small to thoroughly reflect the model performance, but the preliminary results show great potential. The collaborators are working to provide more data for us to further improve and verify the models.

Table 1. Summary of trained models with top performance

Features + Model Type	Precision	Recall	Features + Model Type	Precision	Recall
MFCCs + ANN	89%	89%	Spectrograms + CNN	88%	90%

MeTILDA is currently used by linguistics researchers in finding patterns among Blackfoot speakers and in documenting the language. It has also been used by teachers and students in evaluating a learner's pronunciation as compared to that of native speakers. A recent survey was conducted to gather feedback on usefulness, ease of learning, and satisfaction from representative user groups. Totally 14 users participated with 3 self-identified as linguists, 10 students, and 1 teacher. By average, the rating for each question is above 4.0 out of 5 and over half of the questions received more than 4.5 ratings.

5 Conclusions

With the goal to promote research and education in endangered languages, this project develops an audio classification application to automate the process of obtaining instances of vowels and consonants in Blackfoot. We also present a working prototype called MeTILDA that has shown promising results towards analyzing and learning Blackfoot speech. As a result, linguistics researchers are provided with multiple tools to document and analyze language recordings. Teachers have access to collections of words and recordings from native speakers to teach students. Students are given the tool to compare their own pronunciation of words to that of native speakers. In the future work, we plan to further extend our work to classify and educate other sounds in Blackfoot and other endangered languages.

6 ACKNOWLEDGMENTS

This work is supported by National Science Foundation (NSF BCS-2109654). We also appreciate the late Mr. Earl Old Person for his audio recording as a native speaker and the learners of the Blackfoot language.

References

1. Eberhard, D, M., Simons, G. F., Fennig, C. D. (eds.): Ethnologue: languages of the world. SIL International (2022).
2. Rogers, C., Campbell, L.: Endangered Languages. *obo* in Linguistics (2011).
3. State of the art of Indigenous languages in research: A collection of selected research papers. ISBN: 978-92-3-100521-3 (2022).
4. Kaneko, I.: A metrical analysis of Blackfoot nominal accent in optimality theory. Doctoral dissertation, University of British Columbia (2000).
5. Bird, S., Miyashita, M.: Teaching phonetics in the context of Indigenous language revitalization. In: International Symposium on Applied Phonetics, pp. 39-44 (2018).
6. Bhattacharjee, M., Prasanna, S. M., Guha, P.: Time-frequency audio features for speech-music classification. arXiv:1811.01222 (2018).
7. Eutizi, C., Benedetto, F.: On the performance improvements of deep learning methods for audio event detection and classification. In: International Conference on Telecommunications and Signal Processing, pp. 141-145 (2021).
8. Chen, K., Liang, B. Ma, X., Gu, M.: Learning audio embeddings with user listening data for content-based music recommendation. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3015-3019 (2021).
9. Scarpiniti, M., et al.: Deep Belief Network based audio classification for construction sites monitoring. Expert Systems with Applications, vol. 177 (2021).
10. McIvor, O.: Adult Indigenous language learning in western Canada: what is holding us back? In: Michel, K., Walton, P., Bourassa, E., Miller, J. (Eds.) Living Our Languages: Papers from the 19th Stabilizing Indigenous Languages Symposium, pp. 37-49 (2015).
11. Chen, M., Borad, J., Miyashita, M., Randall, J.: Integrated cloud-based system for endangered language documentation and application. In: IEEE 4th International Conference on Multimedia Information Processing and Retrieval, pp. 235-238 (2021).

12. Aoyama, K., Reid, L. A.: Cross-linguistic tendencies and durational contrasts in geminate consonants: An examination of Guinaang Bontok geminates. Journal of the International Phonetic Association, 36(2), 145-157 (2006).
13. Prabakaran, D., Sriuppili, S.: Speech processing: MFCC based feature extraction techniques-an investigation. Journal of Physics: Conference Series, 1717(1), 012009 (2021).
14. McFee, B., et al.: librosa: Audio and music signal analysis in python. In: 14th python in science conference, vol. 8, pp. 18-25 (2015).
15. Sandeep R.: Unsupervised feature extraction for data mining of endangered language audio data. University of Washington Bothell (2016).