A Speaker Accent Recognition System for Filipino Language

Batman Odulio*, Karl Adrian Cruz, Justin Raphael Ariaso, Mico Ian Orjalo
Angelica Dela Cruz, Ramon Rodriguez and Manolito Octaviano Jr.
College of Computing and Information Technologies
Manila, Philippines
National University
*oduliobc@students.national-u.edu.ph

Abstract

This paper presents the development of an accent recognition system for the native speakers of Bikol and Tagalog using deep learning. The results of the work serve as baseline for the advancement of recognizing speakers with Tagalog and Bikol accents in Filipino language. A monologue written in Filipino is prepared as script for the development of the speech corpus. The script is used to capture the Bikol accent and Tagalog accent in the recordings. The corpus was validated, cleaned and divided into 80:20 ratios for training and testing. Afterwards, Praat is utilized to analyze and extract prosodic features such as F1 and energy of speech. The model was tested and yields 79.28% and 78.33% accuracy for Tagalog and Bikol accent, respectively.

1 Introduction

Accent can be defined as the pronunciation style in a language. In the Philippine setup, the accent of a speaker can be highly influenced by other speakers in an approximate geographical location. For this reason, people can speak the same language but with a different accent resulting to a language (e.g. Filipino) used with multiple accents. This can provide information about the status, age, gender, dialect and ethnicity of a speaker when analyzed intensively (Tjalve, 2007). For the past years, the value of recognizing an accent of a speaker has begun to receive attention in the field of computing. Its influence was acknowledged as foundation in developing different large-scale speech applications such as automatic speech recognition (Petkar, 2016), (Zheng et al., 2005). However, automatic accent recognition is a challenging research task since a language can have multiple style pronunciation.

Automatic accent recognition, also known as accent identification, is based on the consistency of acoustic patterns that can be identified in speaking style that leads to the distinction of pronunciation on the same accent cluster. The work of (Lazaridis et al., 2014) categorized accent recognition system into foreign and regional. Foreign accent classification is characterized by the distinct difference in utterances of a foreign language as spoken by a non-native speaker. On the other hand, regional accent classification is characterized by the changes in pronunciation mainly in speaking styles among native speakers of the language.

There have been numerous research efforts to develop accent recognition system in different languages. However, as of the writing of this paper, the work of (Danao et al., 2017) is the only existing study that explored accent recognition in Philippine languages.

In order to expand the research of accent recognition, this paper focused on the development of accent recognition between Bikol and Tagalog speakers using Filipino monologue. The work shows a baseline work for recognizing speakers with Tagalog and Bikol accents in Filipino language.

2 Related Works

There are different efforts made to develop accent recognition in different regions using various approaches. Various models created from acoustic features, but also deep neural networks were explored. Gaikwad et al. (2013) focused on the English pronunciation of native speakers of Marathi and Arabic language. Acoustic features such as energy, pitch, and formant frequency were

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 511-515, Hakodate, Japan, September 13-15, 2019

extracted and used in the experiment. It was noted that formant frequency feature gives a promising result for accent of Marathi speakers while energy feature for Arabic speakers. In another study, Pham et al. (2016) explored the combination of Mel Frequency Cepstrum and F0 for Gaussian Mixture Model in recognizing Vietnamese dialects. Combining formants and bandwidths with normalized F0 boost the baseline dialect identification of the language from 58.6% up to 72.2%. While the study of Biadsy (2011) described a variety of approaches that make use of different acoustic features of a speech signal such as frame-based acoustic, phonetic, phonotactics features and high-level prosodic features in building a system that recognizes the regional dialect and accent of a speaker. The best approach of the study was tested in four broad Arabic dialects, ten Arabic subdialects, American English vs. Indian English accents, American English Southern vs. Non-Southern, American dialects at the state level plus Canada and three Portuguese dialects. The approach introduced by the study was able to achieve an Equal Error Rate (EER) of 4% for four broad Arabic dialects, an EER of 6.3% for American vs. Indian English accents, 14.6% for American English Southern vs. Non-Southern dialects, and 7.9% for three Portuguese dialects. To further test the approach, it was applied to an automatic speech recognition system that was significantly improved by 4.6%.

On the other hand, approaches based on deep neural networks were also explored. Astrid et al. (2017) analyzed speech accents on videogames using deep learning. The intuition is that characters from a videogame have traits, such as appearance and speech accent which can determine their characteristics. A model was trained using AlexNet to differentiate American, British, and Spanish accents in a videogame. Reported result shows a low accuracy of 61% that opens for various questions for further research extension. Similarly, Jiao et al. (2016) experimented on the fusion of deep neural networks and recurrent neural networks. The fusion of network performed better as compared to individual networks tested on a speech corpus with 45-second utterances.

3 Methodology

3.1 Data Collection

A 3-page monologue written in Filipino is prepared for the speakers that was examined by a Filipino linguist to ensure that the accent and emotion when read will be emphasized (Hatzidaki et al., 2015). Speech collection is done in Bikol University, Legazpi to collect Bikol accent and National University, Manila to collect Tagalog accent. The participants selected lived in the region for at least 10 years to ensure their Tagalog and Bikol accents. The speech was recorded in a closed room using Audacity[1], a free, open source, cross-platform audio software and headset with digital stereo sound and noise canceling microphone. A total of 106 Bikol and 51 Tagalog speech data were collected from each university.

3.2 Data Processing

A native speaker of the collected language is sought to validate the speech data. During the validation, it was noted that there exist five (5) variations of Bikol accent as shown in Table 1.

Accent Variations	Male	Female
Bikol-Buhi	1	4
Bikol-Daraga	14	10
Bikol-Legazpi	28	25
Bikol-Masbate	2	5
Bikol-Sorsogon	6	11
Tagalog	25	26

Table 1. Collected speech data

In order make the Bikol accent comparable to the Tagalog speech data, the dominant accent (Bikol-Legazpi) in the corpus is used. The speech data from the two languages were cleaned by removing the unnecessary background noises using noise profiling using Audacity. Speech signals were divided into frames to help in determining the uniqueness of a certain accent. Each sentence was slice into 42 parts. Praat, a free computer software package for the scientific analysis of speech, was used to extract prosodic features necessary for

[1] https://www.audacityteam.org/

training using 25ms frame length and 10ms window. The extracted features were F0, Mean Energy, Duration, Minimum Pitch, Maximum Pitch, Minimum Energy and Maximum Energy. Afterwards, extracted features from the speech data were split into two: 80% for training data and 20% for testing data.

3.3 Modelling

The features were extracted from the speech signals and fed in a 1-Dimensional Convolutional Neural Networks (1D-CNN) using Keras[2]. Below is the architecture of the model (see Figure 1 for the visualized neural network architecture):

- hidden layers: convolutional layer (4)
- activation function (hidden layers): ReLU
- dropout layer: 0.5
- dense layer (fully-connected layer): softmax activation function
- output layer: Sigmoid activation function

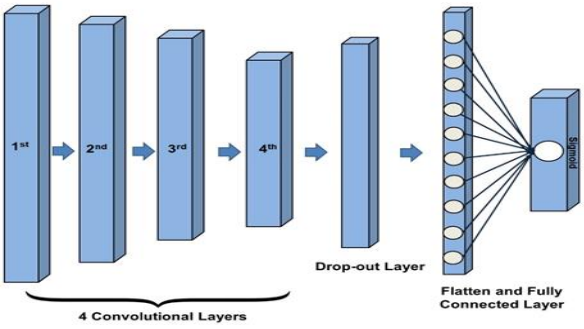

Figure 1. Convolutional Neural Network Architecture

Different parameters were explored by changing the values of output filters, batch size and epoch which yields different training and testing accuracy.

3.4 Feature Selection

The different features were experimented in varying combinations to find the suitable features that will be used to train the model of the accent recognition. These experiment setups were fed in the 1D-CNN. See Table 2 for the experiment setups used in the study.

Upon experimentation, the features that yielded highest accuracy is experiment #6 that has an accuracy of 82.89%, while the features that yielded the lowest accuracy is experiment #4 that has an accuracy of 63.27%. Experiment numbers 1-5 used the duration feature that affected the accuracy since the duration per data is different from one to another due to the uneven length of slice per data. Based on the previous studies and this experimentation, Energy and F0 were used as features for the accent recognition.

Experiment #	Features	Acc. (in %)
1	F0, Mean Energy, Duration, Minimum Pitch, Maximum Pitch, Minimum Energy and Maximum Energy.	76.75
2	F0, Mean Energy, Duration, Minimum Pitch, Maximum Pitch and Minimum Energy.	77.78
3	F0, Mean Energy, Duration, Minimum Pitch and Maximum Pitch.	77.31
4	F0, Mean Energy, Duration and Minimum Pitch.	63.27
5	F0, Mean Energy and Duration.	77.99
6	F0 and Mean Energy.	82.89

Table 2. Set of features for different experiments

3.5 Evaluation

The evaluation metrics used are accuracy and F1 score based from previous works in accent recognition. Accuracy is one of the evaluation metrics for assessing classification models appropriate to this study.

[2] https://keras.io/

4 Results and Discussion

Exp #	Output Filters	Batch Size	Epoch	Train Acc. (in %)	Testing Acc. (in %)
1	16/32	42	200	80	77
2	32/64	42	200	82	78
3	64/128	42	200	83	79
4	64/128	32	100	81	77
5	64/128	22	100	81	80
6	64/128	12	100	80	65
7	64/128	32	50	82	76
8	64/128	22	50	81	69
9	64/128	12	50	82	77

Table 3. Results of the experiments for parameter tuning

The convolutional neural network was trained using the determined hyperparameters (shown in table 3) to generate the accent classifier model. The generated model was then evaluated using the evaluation metrics. Table 4 shows the results of evaluation of the model.

Accent	Accuracy	F1 score	Recall	Precision
Tagalog	78.33	79.00	79.00	79.00
Bikol	79.28	78.50	78.00	79.00

Table 4. Results of evaluation

In addition, the researchers implemented the model by developing a user-friendly prototype in python that follows Input-Process-Output (IPO) scheme. The prototype is named "PARS: Philippine Accent Recognition System". Philippine Accent Recognition System (PARS) aims to distinguish the Accent of Bikol and Tagalog languages through utilizing the prosodic features of speech using the developed model developed. The researchers tested the prototype by having 840 testing data set and utilized the developed model and the result is as shown in the confusion matrix on Table 5.

Out of 420 Bikol speech data, the model correctly recognized 329 Bikol accent and out of 420 Tagalog speech data, the model correctly recognized 333 Tagalog accents. The results have shown the performance of the developed model and it reflects that the amount of correctly recognized input is more than the misrecognized input.

There are different factors that affect the performance of the model. The amount of data in this study is way less than the amount of data used by another study that also used deep learning in recognizing accents in speech.

It was also observed that during data cleaning or noise removal stage, words in low volume were removed in noise reduction activity.

Accent	Bikol	Tagalog
Bikol	329	91
Tagalog	87	333

Table 5. Confusion matrix

5 Conclusion and Future Work

A speaker accent recognition model for Filipino was developed using a 1D-CNN architecture. The study focused on two accents, Bikol and Tagalog. The model was tested and yields 79.28% and 78.33% accuracy for Tagalog and Bikol accent, respectively. The model was also implemented by developing a prototype named PARS. For future work, an increase in speech data and the inclusion of a wider scope of regions from the selected area is highly recommendable. The balance distribution of the participants with regards to age-group, gender and language can also be considered in order to improve the speaker accent recognition model. The researchers suggest to robust the model by considering the noise and making the model noise resistant. The researchers also suggest that in order to maximize the speech patterns and optimize the algorithm, adding more prosodic and other speech features can be explored.

References

Astrid Ensslin, Tejasvi Goorimoorthee, Shelby Carleton, Vadim Bulitko, Sergio Poo Hernandez 2017 Deep Learning for Speech Accent Detection in Videogames. In Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference.

Biadsy, F. 2011. Automatic dialect and accent recognition and its application to speech recognition Doctoral dissertation. Columbia University.

Glorianne Danao, Jolea Torres, Jamila Vi Tubio, and Larry Vea 2017. Tagalog Regional Accent

Classification in the Philippines. 2017 IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM).

Harshalata Petkar. 2016 A Review of Challenges in Automatic Speech Recognition, 151-No.3. International Journal of Computer Applications.

Hatzidaki, A., Baus, C., & Costa, A. (2015). The way you say it, the way I feel it: emotional word processing in accented speech. *Frontiers in psychology*, *6*, 351. doi:10.3389/fpsyg.2015.00351.

Michael Tjalve 2007. Accent features and idiodictionaries: on improving accuracy for accented speakers in ASR. Dissertation. University College London.

Lazaridis, Alexandros and Khoury, Elie and Goldman, Jean-Philippe and Avanzi, Mathieu and Marcel, S´ebastien and Garner, Philip N 2016 Swiss French regional accent identification. In Proceedings of odyssey 2014: The speaker and language recognition workshop.

PhamNgocHung,TrinhVanLoan,NguyenHongQuang 2016 Automatic identification of vietnamese dialects, V.32, N.1 (2016). Journal of Computer Science and Cybernetics.

Santosh Gaikwad, Bharti Gawali, Kale, K.V. 2013 Accent Recognition for Indian English using Acoustic Feature Approach International Journal of Computer Applications.

Yishan Jiao, Ming Tu, Visar Berisha, and Julie Liss 2016 Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features Interspeech 2016.

Zheng, Yanli and Sproat, Richard and Gu, Liang and Shafran, Izhak and Zhou, Haolang and Su, Yi and Jurafsky, Daniel and Starr, Rebecca and Yoon, Su-Youn. 2005 Accent detection and speech recognition for shanghai-accented mandarin. Ninth European Conference on Speech Communication and Technology.