A Neural Network Approach to Accent Classification

Saiprasad Duduka[1], Henil Jain[1], Virik Jain[1], Harsh Prabhu[1], Prof. Pramila M. Chawan[2]

[1]B.Tech Student, Dept of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India
[2]Associate Professor, Dept of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India

---***---

Abstract - *Accent classification forms an important part of many Automatic Speech Recognition (ASR) systems. Neural networks are being used increasingly in many ASR tasks including recognizing and classifying accents. This paper aims to classify different accents by extracting Mel Frequency Cepstral Coefficients from the audio of the speaker and predict the accent using a Convolutional Neural Network model.*

Key Words: *Mel Frequency Cepstral Coefficients (MFCC), Convolutional Neural Network (CNN)*

1. INTRODUCTION

Recognizing and distinguishing accents form an integral part of Automatic Speech Recognition (ASR) systems. Considering the current popularity of smart voice assistants on both smart speakers as well as mobile phones, improving ASR systems has become an important focus for research. One of the aspects in this is the accent of the speaker. For example, the interpretation of spoken English depends on whether the speaker is speaking in say an American accent or British accent. Based on that, the system can get more information about the user's background and use that information in other use cases. In this paper, we predict accents of three different native speakers namely English, Arabic and Mandarin using a 2-D CNN model which uses 13 MFCC features of each sample audio file. The audio files used were taken from Speech Accent Archive by George Mason University.

2. LITERATURE REVIEW

Bryant et al. achieved 42% accuracy on George Mason University dataset using Gaussian Discriminant Analysis (GDA) for five male accents: English, Spanish, Arabic, French and Mandarin. [1]

Chan et al. achieved 90% accuracy using Artificial Neural Network (ANN). However, the dataset was limited as it consisted of audio samples from only 22 speakers and the model was only able to classify native and non-native English Speakers. [2]

Parikh et al. achieved 68.67% accuracy on George Mason University dataset using Convolutional Neural Network (CNN) for three accents: Spanish, Indian and American. [3]

Blackburn et al. classified accents in 4 stages using Artificial Neural Network containing categories: voiced, unvoiced, stopped and energy dip. However, they also work on limited dataset only 50 people and three accents: Arabic, Chinese and Australian. [4]

Jiao et al. achieved 51.92% accuracy on the dataset for the INTERSPEECH 16 Native Language Sub-Challenge contains a training, a development, and a test set using Deep Neural Network (DNN) with ReLU at output of each layer. [5]

Duduka et al. have done a comprehensive survey on this topic of accent classification incorporating the above as well as other papers in their survey.[6]

3. PROPOSED SYSTEM

3.1 PROBLEM STATEMENT

"To classify accent of a speaker among Arabic, English and Mandarin using a Neural Network Classifier."

3.2 DATASET DESCRIPTION

As suggested by More, Nikhil T. et al., the success of any software project depends on the quality of the requirements.[7]

For this project, we have taken audio files from the Speech Accent Archive hosted by George Mason University. For the purpose of this project, we have limited the classification to the three most spoken accents: English, Mandarin and Arabic. In each of the audio files, the speaker recites the following English paragraph:

'Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.'

The frequency of each spoken accent in our dataset is as follows:

Accent	Frequency	Size (in MB)
Arabic	194	563
English	646	1228
Mandarin	151	394
Total	991	2185 (2.1 GB)

We reserved 20% of this data for testing purposes. The dataset was split into training and testing data by random sampling.

Along with the audio files, the dataset also include certain metadata about the speakers like sex, birth place, other spoken languages, age, age of English onset, method of learning English, native language, etc.

3.3 DATA PREPROCESSING

We first preprocess the audio files before feeding it to our model by extracting the Mel Frequency Cepstral Coefficients (MFCCs) from the audio files.

These form the features of our dataset that are fed to the neural network. For this project, we extract 13 MFCCs from the audio files to be used as features.

The following diagram represents the flow of entire data preprocessing pipeline.

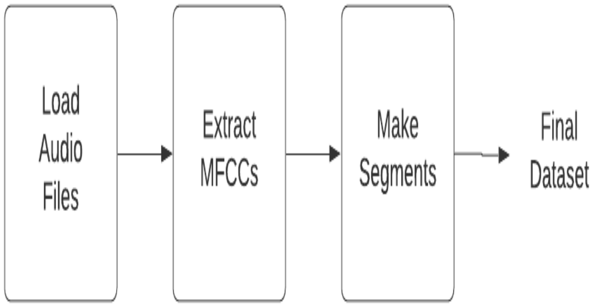


Fig -1: Data Preprocessing

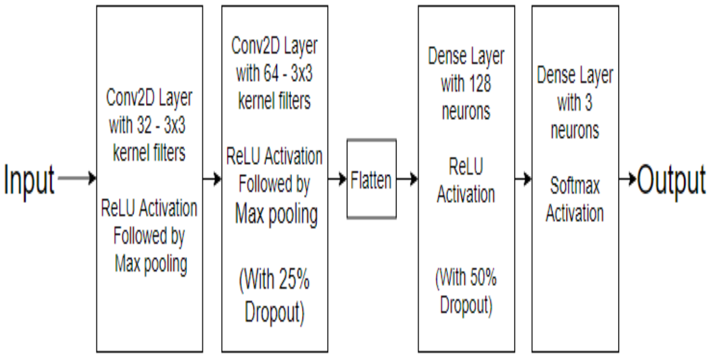
The following are the data preprocessing steps:

1. Load Audio Files:
   We use the librosa python library to load the .wav files.

2. Extract MFCCs:
   We then use the librosa library to extract 13 MFCCs from each audio file.

3. Make Segments:
   We then break down the MFCCs from each training example into smaller chunks.

3.4 MODEL DESCRIPTION


Input
Conv2D Layer with 32 - 3x3 kernel filters
ReLU Activation Followed by Max pooling
Conv2D Layer with 64 - 3x3 kernel filters
ReLU Activation Followed by Max pooling
(With 25% Dropout)
Flatten
Dense Layer with 128 neurons
ReLU Activation
(With 50% Dropout)
Dense Layer with 3 neurons
Softmax Activation
Output

The model consists of four layers – one input layer, two hidden layers and one output layer.

Layer 1: The input layer is 2D convolutional layer that consists of 32 filters with each filter having a dimension of 3x3.The activation function used is a ReLU (Rectified Linear Unit) activation function. Output of the activation function is given as input to the 2D Max Pooling with pool size of 2x2.

Layer 2: The first hidden layer is also 2D convolutional layer that consists of 64 filters with each filter having a dimension of 3x3. The activation function used is a ReLU (Rectified Linear Unit) activation function. Output of the activation function is given as input to the 2D Max Pooling with pool size of 2x2. Dropout rate is kept at 25%. The multidimensional output generated by the previous layer is transformed into a one-dimensional array which is fed as input to the next hidden layer.

Layer 3: The second hidden layer is a dense layer with 128 neurons with each neuron having a ReLU activation function. Dropout rate in this layer is kept at 50%.

Layer 4: The last layer which is the output layers is a dense layer which consists of 3 neurons with SoftMax as the activation function.

4. IMPLEMENTATION:

Our primary programming language for implementation of our project is Python. We take the help of various Python libraries to implement our functionalities. The major libraries used are:

a. Librosa: To load and process audio files and to extract MFCCs from the .wav files.

b. Scikit-learn: To create train and test datasets from the database.

c. Keras: To build the neural network.

5. RESULTS:

Out of the total available dataset, we used 80% of the data for training and reserved the remaining 20% for testing. The exact size of our training and testing data for each accent is as follows:

Training samples:

Accent	Frequency
Arabic	155
English	520
Mandarin	118

Testing samples:

Accent	Frequency
Arabic	40
English	126
Mandarin	33

After testing our model on the testing data, we achieved an accuracy of 62.81%.

6. CONCLUSION AND FUTURE SCOPE:

In conclusion, our model gives an accuracy of 62.81% for these three accents: English, Arabic and Mandarin using Convolutional Neural Network (CNN). As we can see from some of the previous related works that were done on other accents, our results are quite satisfactory.

For future scope, we can try other deep neural network models such as Recurrent Neural Network. Also, we can try training the model on larger dataset along with some more hyperparameter tuning.

7. REFERENCES

[1] Morgan Bryant, Amanda Chow, Sydney Li, “Classification of Accents of English Speakers by Native Language”.

[2] M. V. Chan, Xin Feng, J. A. Heinen and R. J. Niederjohn, "Classification of speech accents with neural networks," Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94), Orlando, FL, USA, 1994, pp. 4483-4486 vol.7, doi: 10.1109/ICNN.1994.374994. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[3] K Parikh, Pratik and Velhal, Ketaki and Potdar, Sanika and Sikligar, Aayushi and Karani, Ruhina, English Language Accent Classification and Conversion using Machine Learning (May 14, 2020). Proceedings of the International Conference on Innovative Computing & Communications (ICICC) 2020, Available at SSRN: https://ssrn.com/abstract=3600748 or http://dx. doi.org/10.2139/ssrn.3600748.

[4] Blackburn, C., J. Vonwiller and R. King. “Automatic accent classification using artificial neural networks.” EUROSPEECH (1993).

[5] Jiao, Yishan & Tu, Ming & Berisha, Visar & Liss, Julie. (2016). Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long- and Short-Term Features. 2388-2392. 10.21437/Interspeech.2016-1148.

[6] Duduka, Saiprasad & Jain, Henil & Jain, Virik & Prabhu, Harsh & Chawan, Pramila. (2021). Accent Classification using Machine Learning.

[7] More, Nikhil T.; Sapre, Bhushan S.; and Chawan, Pramila M. (2017) "An Insight into the Importance of Requirements Engineering," *International Journal of Computer and Communication Technology*: Vol. 8 : Iss. 1 , Article 5. DOI: 10.47893/IJCCT.2017.1394

[8] Singh, Utkarsh et al. ‘Foreign Accent Classification Using Deep Neural Nets’. 1 Jan. 2020: 6347 – 6352

[9] Y. Ma, M. Paulraj, S. Yaacob, A. Shahriman and S. K. Nataraj, "Speaker accent recognition through statistical descriptors of Mel-bands spectral energy and neural network model," 2012 IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (STUDENT), Kuala Lumpur, 2012, pp. 262-267, doi: 10.1109/STUDENT.2012.6408416.