Accent Classification using Machine Learning

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 - *This paper aims to perform a comprehensive survey on different machine learning and deep learning techniques used in literature for the task of accent classification. 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. Thus, we try to summarize the various models and techniques used for the task of detecting and classifying accents along with the datasets used by them and the corresponding accuracies for different models.*

Key Words: Machine learning, Deep learning, Accent Classification, Automatic Speech Recognition (ASR)

1. INTRODUCTION

Identifying and classifying accents form an important part of Automatic Speech Recognition (ASR) systems. With the increasing popularity and usage 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 summarize the work of 7 research attempts for the task which use many wide-ranging machine learning algorithms as well as various architectures of neural networks. In the next section, we briefly summarize these 7 research papers discussing the datasets used by them, the algorithms and models used and implemented as well as the results that they achieve on the corresponding datasets, wherever reported.

2. LITERATURE REVIEW

2.1 Classification of Accents of English Speakers by Native Language

The goal is to classify accents by the native language of the speaker, i.e. predict the speaker's native language. Dataset used: George Mason University Department of English Speech Accent Archive. It contains recordings of people speaking the same sentence. For each clip, the archive contains information about the speaker's background, like age, gender, birthplace, native language, other languages spoken, age of English onset, English residence and length of English onset. Five male accents are chosen for classification: English, Spanish, Arabic, French and Mandarin.

They designed their classification algorithm to capitalize on this difference by comparing different speakers' enunciations of each syllable in the recording and using this information to model how speakers of each language enunciate each syllable in the script. Pre-processing by splitting the recording and aligned using Munich Automatic Segmentation System (MAUS). Length of clip for the spoken word and volume are adjusted and normalized. Mel-frequency cepstral coefficients (MFCC's) are used as features. PCA is used to get the 16 most important features for each word. The models used were of two kinds: 1. Supervised – SVM, Naïve Bayes, SoftMax Logistic Regression and GDA, 2. Unsupervised – Gaussian Mixture Model (GMM) and k-Means Clustering. Best Result: 42% with GDA which is on par with previous attempts on this task.

2.2 Classification of Speech Accents with Neural Networks

They use Artificial Neural Networks (ANNs) for the task of identifying different accents, specifically Competitive Learning, Back-propagation and Counter Propagation models. Unsupervised competitive learning model architecture is relatively simpler compared to that of supervised back-propagation model and training is less time consuming, but results are usually more stable for the supervised back-propagation models. Finally, there is counter propagation model which combines competitive learning and back-propagation models.

Data from 22 speakers is used (10 native English speakers and 12 non-native English speakers) and 7 features were Age at which English was first used, Percentage of day when English is used, No. of years English has been used, Pitch frequencies averaged over time and First 3 formant frequencies. Architecture of all ANNs consists of 7 input neurons in the first layer, 3 neurons in the hidden layer, and 2 neurons in the output layer. The hidden layer uses a hyperbolic tangent function. The stop criterion is set to 0.0001 for weight changes in all ANNs.

Table -1: Comparison of results of different models

Classifier	Training	Testing
Competitive Learning	81.8%	45%
Counter Propagation	100%	64%
Back Propagation	100%	90.9%

2.3 English Language Accent Classification and Conversion using Machine Learning

The goal is to prevent communication barriers arising because of accents distinctness by creating a system to detect accent and convert it to the required form. Dataset Used: The dataset used is The Speech Accent Archive. It contains audio files of people saying the same lines. The demographic information of the people is also present viz., birthplace, native language, sex, other languages known, English learning method (academic or naturalist), age of English onset, location and duration of the speakers. However, only the native language is used for accent detection. Three accents are chosen for classification: Spanish, Indian and American.

For accent detection, the data is first pre-processed in which the pauses between the words are removed. Next, MFCC features are extracted and fed into an CNN. LSTM extraction is also done. The long-term features are fed into a DNN and short-term features are fed into an RNN. These two are combined with a probabilistic fusion algorithm, the results of which are again combined with the RNN to give the best possible accuracy.

For accent conversion, a cycle AN network is used with two generator and two discriminators. The first generator converts the source to target and the second discriminator does the comparison between generated target and true target files. The files for the first discriminator are generated by the second generator which takes the generated target file as an input. The first discriminator is used to compare the source file and generated source file. The output of the second generator is used as an input for the first. Best Result: An accuracy of 68.67% is achieved on accent detection with 95 epochs.

2.4 Foreign accent classification using deep neural networks

To classify the accent and get insights about the person's interests, culture, psychology and similar higher-level attributes. The study uses Common Voice corpus of speech data read by users on the Common Voice website. The corpus consists of speech data form 18 languages and consists of 1087 validated hours of recording in MP3 audio format (sampled at 22.05 kHz). The dataset also consists of demographic metadata like age, sex, and accent. The training and testing set has been made to be speaker independent. Audio samples from five different countries US, England, India, Canada, Australia have been used.

The audio dataset contains different length audio. To make all audio files of equal length following processing was done: 1. Calculate the median audio length for conversion from the dataset (3.62 seconds) 2. For longer clips, trim the audio file to have a length equal to median length 3. For shorter clips, pad the audio file with zeros since it adds no new information to the audio file and make its length equal to the median length.

The architecture consists of a cascade of Convolutional Neural Network (CNN) and Convolutional Recurrent Neural Network (CRNN) which are as follows: 1. CNN: Consists of 4 layers with ReLu activation layer and the last layer is a fully connected layer with SoftMax output function. 2. CRNN: Consists of 4 layers wherein each layer is followed by batch normalization and max-pooling and the last 2 are GRU (Gated Recurrent Units) Layers. CNN and CRNN exploit spatially local correlations across input data to improve the performance of audio processing tasks, such as speech recognition, accent identification and acoustic scene classification. The proposed model has an accuracy of 78.48% using CNN and 83.21% using CRNN.

2.5 Speaker Accent Recognition through Statistical Descriptors of Mel-bands Spectral Energy and Neural Network Model

The goal is to classify different accent types. Three English accents were classified namely Malay, Chinese and Indian using artificial neural network model. Experimental Setup: For the analysis purpose of this work, we took speech corpus recorded from 42 female volunteers of three main ethnics. It is composed of 15 Malay, 15 Chinese and 12 Indian female speakers. 630 speech samples were collected in total. The speakers are originated from various north, south, west and east regions of the country and as such they were also influenced by their regional accents. The background noise in that room was recorded approximately 22 dB which is considered very quiet. Laptop computer sound card was used for recording the speech and the sampling rate and bit resolution was set to 16kHz and 16 bps respectively in the MATLAB program. Feature Descriptors: Mel- filter bands Spectral Energy applying Mel-filter banks in frequency domain simply amounts to multiplying triangular-shaped windows to each region of spectrum of interest. Statistical Descriptors: Mean, Standard Deviation and Variance.

Accent Modelling using Neural Networks: We used two-layer FF-MLP to classify our features into one of three accent classes and the network was trained using Levenberg-Marquardt learning algorithm that is known for its fast convergence. We adopted Mean-squared Error (MSE) as an objective criterion for successful learning of the task. Best Result: Analysing the best performance, we observed that within the range, the min CR=93.00% required 27 epochs and the max CR=99.01% required 34 epochs. The average value of CR was 96.79% with standard deviation of slightly less than 2%.

2.6 Automatic Accent Classification Using Artificial Neural Networks

The goal is to develop an English accent classification system to discriminate between the speech of subjects whose first language is Arabic, Chinese and Australian English. Dataset Used: The speech data used in these studies were collected from 50 subjects, at least 5 male and 1 female speakers of each language. To obtain a reasonable degree of accent uniformity all the Chinese speakers were from Beijing; the Arabic speakers were from Lebanon; the Australian speakers were from Sydney. Three accents chosen for classification: Arabic, Chinese and Australian.

The system operates on continuous speech sample of arbitrary duration. The classification is performed in stages. A broad phonetic class segmented is used to divide the incoming speech into 4 categories, namely, voiced, unvoiced, stopped and energy dip. For each of these segment types, an artificial neural network is trained to classify the accent. A cumulative measure of the accent classification is obtained from sequence of accent labels from these four networks.

2.7 Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long- and Short-Term Features

The goal is to propose a combination of long-term and short-term training for automatic identification of foreign accents and analyse how elemental components of speech change with accent. The provided dataset for the INTERSPEECH 16 Native Language Sub-Challenge contains a training, a development, and a test set. The corpus contains one speech sample from 5132 speakers, labelled with one of the 11 native languages. The training set was assigned 3300 samples and development set was assigned 965 samples. The remaining 867 samples are assigned to the test set. The length of each sample is 45 seconds.

Voice activity detection (VAD) to remove the silence periods. The noise level was matched with the VAD threshold. Framing and Feature Extraction: The remaining speech samples were then trimmed into multiple segments with equal intervals of 4 seconds. Deep Neural Network: A DNN was constructed to make a prediction regarding the accent type from the long-term features. Rectifier linear units ("ReLU") were used at the output of each layer and we use the dropout method to prevent overfitting. Recurrent Neural Network: The short-term features extracted from 25ms frames of speech were used to train the RNN. Categorical labels were assigned to each frame of the segment. The results for each sample were calculated by averaging the predictions on all frames in all segments. The UAR is 52.24% and the overall accuracy is 51.92%,

3. PROPOSED SYSTEM

3.1 Problem statement

"To classify English Accents using neural networks."

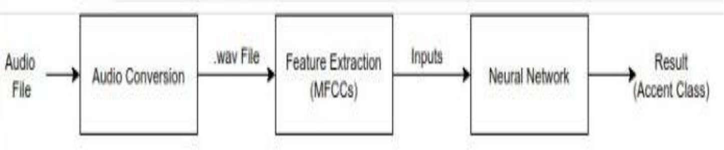
3.2 Problem Elaboration

Accent recognition can be used to develop efficient Automatic Speech Recognition Systems which are found in many voice assistants – to interpret their speech. Also, geographical location of a person can de deduced from their spoken accent. Deep learning methods are used for audio and image processing tasks. Various neural network architectures are implemented for these cases. Thus, we decided to use neural networks for the task of accent classification.

3.2 Proposed Methodology

Audio files will first be converted to .wav files because they are more universal and suitable for audio processing. To extract features from audio files we used Mel Frequency Cepstral Coefficients (MFCC). Once the features have been generated, they will be fed as inputs to the neural networks. Two possible neural network architectures that can be used for this task are:

1. Convolutional Neural Network (CNN)- We can use 1-D CNNs as they are suitable for time-series data, text and audio data. We can use available python frameworks such as Keras for building the model.
2. Recurrent Neural Network (RNN)- RNNs are used frequently for sequential data like time-series data. We can use Long-Short Term Memory (LSTM) cells to build this RNN. This model can also be implemented in Keras.



4. CONCLUSION

This paper thus summarizes the results of our survey on different approaches to accent classification used in literature for the task of accent identification and classification. As we see, most of these systems use deep learning approaches utilizing different architectures of neural networks, as is the case in many ASR tasks.

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

[5] 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.

[6] Blackburn, C., J. Vonwiller and R. King. "Automatic accent classification using artificial neural networks." *EUROSPEECH* (1993).

[7] 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.