Global Journal of Computer Science and Technology: D
Neural & Artificial Intelligence
Volume 22 Issue 1 Version 1.0 Year 2022
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Acoustic Features based Accent Classification of Kashmiri Language using Deep Learning

By Shehzen Sidiq Malla

Abstract- Automatic identification of accents is important in today's world, where we are souranded by ASR systems. Accent classification is the problem of knowing the native place of a person from the way He/She speaks the language into consideration. Accents are present in almost all the languages and it forms an important part of the language. Accents are produced from prosodic and articulation characteristics; in this research the aim is to classify accents of Kashmir Language. We have considered using the MFCC and Mel spectrograms for our research. A lot of research has been done for languages like English and is being done in this field and many models of machine learning and deep learning have shown state of the art results, but this problem is new for Kashmiri Language. The accents in Kashmir, vary from area to area and we have chosen 6 areas as our classes. We extracted the features from the audio data, converted those features into Images and then used the CNN architectures as our model. This research can be taken as base research for further researches in this language. Our custom models achieved the loss of 0.12 and accuracy of 98.66% on test data using Mel spectrograms, which is our best for our features.

Keywords: *accent classification, CNN, RELU, mel-spectrograms, MFCC.*

GJCST-D Classification: *I.2.7*

Strictly as per the compliance and regulations of:

Acoustic Features based Accent Classification of Kashmiri Language using Deep Learning

Shehzen Sidiq Malla

Abstract- Automatic identification of accents is important in today's world, where we are souranded by ASR systems. Accent classification is the problem of knowing the native place of a person from the way He/She speaks the language into consideration. Accents are present in almost all the languages and it forms an important part of the language. Accents are produced from prosodic and articulation characteristics; in this research the aim is to classify accents of Kashmir Language. We have considered using the MFCC and Mel spectrograms for our research. A lot of research has been done for languages like English and is being done in this field and many models of machine learning and deep learning have shown state of the art results, but this problem is new for Kashmiri Language. The accents in Kashmir, vary from area to area and we have chosen 6 areas as our classes. We extracted the features from the audio data, converted those features into Images and then used the CNN architectures as our model. This research can be taken as base research for further researches in this language. Our custom models achieved the loss of 0.12 and accuracy of 98.66% on test data using Mel spectrograms, which is our best for our features.

Keywords: accent classification, CNN, RELU, mel-spectrograms, MFCC.

I. Introduction

Kashmiri or Koshur is a Dardic language subgroup from Indo-Aryan, spoken by over seven million Kashmiris [Wikipedia]. There are many accents Spoken in Kashmir. There are some major accents and some minor accents in this language. This leads to diversity in the language and adds to its beautiful sounds and variations. The aim of this research is to classify these different accents. Although many accents are being spoken in this language, for this research, we have classified the prominent accents belonging to Kupwara, Srinagar, Islamabad, Shopian, and Bandipora. The proposed approach is on the basis of using Convolution Neural Networks (CNN) and training Neural networks on the images of features extracted from the audio files. The features are Mel- spectrogram and MFCCs. Our approach uses CNN as the classifier and MFCCs, Mel spectrograms as Features. Three types of MFCCs are extracted, 13, 24 and 36. We got excellent results on our dataset.

Accent classfication refers to the problem of inferring the native language of a speaker from his or her foreign accented speech. Identifying idiosyncratic differences in speech production is important for improving the robustness of existing speech analysis systems. For example, automatic speech recognition (ASR) systems exhibit lower performance when evaluated on foreign accented speech. By developing pre-processing algorithms that identify the accent, these systems can be modified to customize the recognition algorithm to the particular accent [1] [2]. In addition to ASR applications, accent identification is also useful for forensic speaker profiling by identifying the speaker's regional origin and ethnicity in applications involving targeted marketing [3] [4]. In this paper we propose a method for classification of 11 accents directly from the speech acoustics.

For example, Deshpande et al. used GMMs based on formant frequency features to discriminate between standard American English and Indian accented English [6]. Chen et al. explored the effect of the number of components in GMMs on classification performance [7]. Tang and Ghorbani compared the performance of HMMs with Support Vector Machine (SVM) for accent classification [8]. Kumpf and King proposed to use linear discriminant analysis (LDA) for identification of three accents in Australian English [9].

Artificial neural networks, especially Deep Neural Networks (DNNs) and Recurrent Neural Networks (RNNs) and CNNs have been widely used in state-of-the-art speech systems and Image Processing Systems [10] [11] [12] [13]; however,in the area of accent identification, there are only a few studies evaluating the performance of neural networks [14] [15]. Nonetheless, in a related area, language identification (LID), neural networks have been investigated exhaustively [16] [17] [18].In a recent paper [19], where they used spectrograms for accent classification and speaker recognition and achieved an accuracy of 92%. Inspired by their work, we also propose to use Mel-Spectrograms and MFCCs for our research

The rest of the paper is organized as follows: In section 2, we discuss the collection and making of dataset. In section 3, we discuss proposed system and discuss in detail the features that we have used for our research. In section 4, we discuss the experiments and show our results and finally in section 5, we conclude our research.

Author: e-mail: mallashehzen786@gmail.com

II. Dataset

a) Collection of Dataset

Data is very important in every machine learning and deep learning project or deep learning research. The data, we required for our research was, the audio files of people speaking some sentences that we choose. These sentences, to some extent captured a wide range of accent changes in the spoken Kashmiri Language. In total, 20 sentences were chosen for research purposes and people were recorded speaking these sentences in their native accents of Kashmir language. The data was collected from 5 districts or areas of Kashmir and all these files were saved with the extension of 'ogg', which in preprocessing, were changed to 'wav' format. In total, we got almost100 voice samples from each area and thus we had, 500 total voice samples of these sentences spoken by different people.

b) Making of dataset

The data we had, were audio files and we decided on getting the MFCC and Mel-Spectrograms from these audio files. So, our final dataset consisted of images of MFCC and Mel-Spectrograms. Since deep learning models require huge amount of data, we had to augment the data to increase the size of our dataset. There are many great techniques of augmenting the data, when it comes to images and audio. Since the images were of features, we could not use the normal augmentation techniques like distortion, rotation and many more[reference]. A special kind of augmentation known as specAugment [20], which produces augmented images on spectrograms was used. This augmentation performed following operations on the Images of Mel-Spectrograms. 1) Frequency masking is where certain part of the frequency is masked out, and 2) Time masking, where certain part of time is masked out. Even though, we performed augmentation on Mel-Spectrogram images, the data was not enough. so, we had to perform the augmentation on the audio files also. The audio files were augmented by increasing the speed, pitch and amplitude of the audio files, thus giving somewhat variability in the initial dataset of audio files.

After performing, such augmentations we had large sufficient dataset for deep learning.

III. Proposed System

a) Architecture

We used CNN based architecture with ReLU activation function for internal nodes and SoftMax function to output the probability distribution of output classes. CNN [19] models show state of the art performance with image data. Since our motive was to extract the features from the audio data and plot them as images and then those images were input to the model, so we chose the model based on CNN architecture. Our model has, six convolution layers and six Maxpooling layers, with 5 dense layers and a flatten layer.

[image of model] Accuracy and loss varied based on the feature used and learning rate of the model. We choose different learning rates based on different features that were input to the model. Our models were trained on learning rates between 0.001 to 0.0001

b) Features Used

Many features have been used in researches of audio processing. We decided to keep our research simple so we decided on two features, MFCC and Mel Spectrograms. MFCC have been found to perform well in case audio classification [21] purposes and Mel spectrograms and Spectrograms have also shown such performance in many cases [19]. Different number of coefficients can be used, mostly 13 are taken. The selection of such number of constants, depends on the problem in hand. We experimented on various number of coefficients and finally decided on 13, 24 and 36 coefficients. These features were extracted and plotted as images and then such images were input to our model.

Mel Spectrograms - A Mel spectrogram is a spectrogram that converts the frequencies to the Mel scale. When the spectrogram from the audio file is plotted using Mel scale, we get the Mel-Spectrogram. These spectrograms were plotted as images, same as the MFCCs and given input to the model.

All these operations of feature extraction were done using librosa library [22], which makes working with audio very easy.

IV. Experimental Setup And Results

Different experiments were performed on different features and different learning rates were set during the training of the models.

The features, were stored in two ways -

Images of the features were generated, and in other features were extracted and a dimension was added, no image was generated and the features were stored in JSON data format. Below, we show the results of various experiments:

a) Experiment 1

This Experiment was done using images of Mel spectrograms and MFCC for the CNN having 3 color channels. In this experiment, the images were generated from the audio files. Those images were saved and later were loaded back into the model. The models were trained on those images and evaluated on the validation and testing sets.

i. *Mel Spectrograms*

The below figure shows the metrices graphically and we can conclude from the graph that the model is showing state of art results on our data.

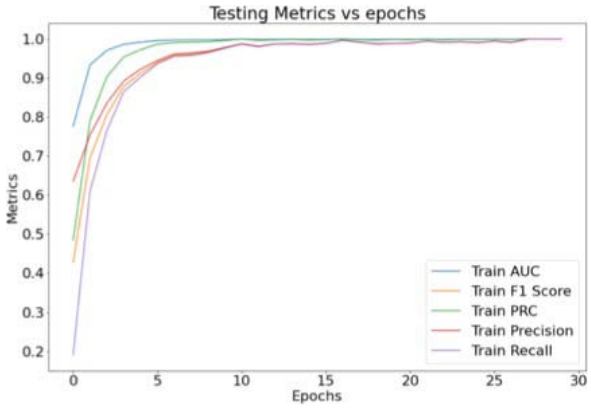

Fig. 1: Plot of Metrics of trainig data

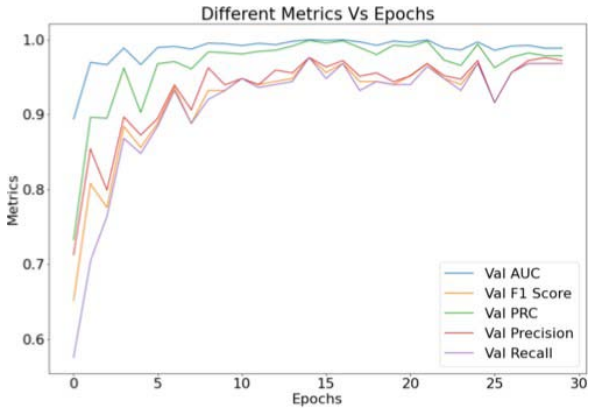

Fig. 2: Plot of metrics of validation data.

The above figure shows the metrices for the validation set. All these metrices were evaluated for the model which uses spectrogram images as the input.

```
Loss - 0.1259424239397049
Accuracy - 0.9866666793823242
F1 Score - 0.9866666793823242
Precision - 0.9866666793823242
Recall - 0.9866666793823242
AUC - 0.9916666746139526
PRC - 0.9788309931755066
```

Fig. 3: Metric Scores for testing data.

These above results were calculated on the testing data and we can conclude that our model performed much better than expected and showed state of the art performance on our data.

ii. *MFCCs*

The MFCC features were extracted from the audio files and plotted as images and these Images were saved and loaded at the time of model training. The following figures show the accuracies and losses with respect to the epochs. Three types of constants were extracted and same model was trained on these images generated from the audio files. The training was done using the training data, and validated on validated. Following table shows our results.

Table 1: Validation losses and Validation accuracies for Different features [images]

Feature	Validation Loss	Validation Accuracy
Mel Spectrograms	0.0392	0.9848
MFCC 13	0.04	0.98
MFCC 24	0.031	0.99
MFCC 36	0.06	0.97

From the above table, we can see that the 24 constants performed slightly better than the others on validation data.

Using the images as input to the model, Mel-Spectrograms and MFCC 24 constant features performed better than the other features.

b) *Experiment 2*

This experiment was done using Json files of extracted features and giving to CNN having 1 color channel.

In this experiment, the features were extracted and were saved in JSON files. No Images were generated for this data. Then the features were loaded back and the model was trained on these features. The below table show the Testing accuracies and testing Losses for various features extracted from audio files.

Table 2: Testing Loss and Testing Accuracy for Various Features [JSON]

Feature	Testing Loss	Testing Accuracy
MFCC 13	0.086	0.87
MFCC 24	0.110	0.87
MFCC 36	0.507	0.865

c) *Experiment 3*

This experiment was done by splitting the audio files in the chunks of 2 seconds.

In this experiment the audio was splitted into two second chunks a The Mel spectrograms features were extracted from the splitted audio files and then saved as images and as well as JSON files. Following accuracies and Losses were calculated on validation data.

Table 3: Validation Loss and Accuracy for 2 Sec Splitted Data

Type of Feature	Validation Loss	Validation Accuracy
Images	0.0331	0.98
JSON files	0.069	0.97

V. Conclusion

This research paper proposes a solution to the accent classification for Kashmiri Language using Convolutional Neural networks. The solution is based on deep learning techniques using CNNs that adapt to the multi-dimensional data. CNNs provide solution to this

problem using the supervised approach where they first undergo training session during which they are fed with labelled data from which they learn the relationships in the data and attain the learning capability. In later stage, they are presented with unseen data of same domain and are able to make remarkable inferences from this unseen data by utilizing the attained learning capability. In addition to reporting the state-of-art classification results, its accuracy is also remarkable. In our research, we can conclude that the models and the data that we used, the Mel Spectrograms performed better and showed better performance than the MFCCs, also we saw that the images with 3 color channels performed better than the features that were saved with extended dimension. Overall we can conclude, our model showed state of the art performance for the Accent classification of Kashmiri Language with five output classes.

VI. Future Improvements

There is a lot of improvement to be done in this field of research. Since, this is the first research in this language, as we could not find any other research related to Kashmiri language, so the area of improvement is vast. We propose following enhancements for this research-

- Collection of more data for efficient model training
  a. The dataset can be increased in size
  b. The dataset can be in such a way, that it captures the maximum of the features and variations present in the language.
- The model can be made more complex and more sophisticated that would be able to handle more data and not underfit.
- Making efficient model for being able to capture most of the features of accent classification
- Improving the classification error and thus being able to classify wide range of the language.
- Making use of different architectures and techniques available for making the overall application most fruitful.
- The classification classes can be increased to more than 5 accents or regions.

References Références Referencias

1. L. Kat and P. Fung, "Fast accent identification and accented speech recognition," in Acoustics, Speech, and Signal Processing (ICASSP), IEEE International Conference on, vol. 1. Phoenix, AZ, USA: IEEE, 1999, pp. 221–224.
2. C. Huang, T. Chen, and E. Chang, "Accent issues in large vocabulary continuous speech recognition," International Journal of Speech Technology, vol. 7, no. 2-3, pp. 141–153, 2004.
3. D. C. Tanner and M. E. Tanner, Forensic aspects of speech patterns: voice prints, speaker profiling, lie and intoxication detection. Lawyers & Judges Publishing Company, 2004.
4. F. Biadsy, J. B. Hirschberg, and D. P. Ellis, "Dialect and accent recognition using phonetic-segmentation supervectors," 2011.
5. S. Deshpande, S. Chikkerur, and V. Govindaraju, "Accent classification in speech," in Automatic Identification Advanced Technologies, Fourth IEEE Workshop on. Buffalo, NY, USA: IEEE, 2005, pp. 139–143.
6. T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using gaussian mixture models," in Automatic Speech Recognition and Understanding, IEEE Workshop on. Madonna di Campiglio, Italy: IEEE, 2001, pp. 343–346.
7. H. Tang and A. A. Ghorbani, "Accent classification using support vector machine and hidden markov model," in Advances in Artificial Intelligence. Springer, 2003, pp. 629–631.
8. K. Kumpf and R. W. King, "Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks," in Proc. EuroSpeech, vol. 4, pp. 2323–2326, 1997.
9. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," Signal Processing Magazine, IEEE, vol. 29, no. 6, pp. 82–97, 2012.
10. H. Zen and H. Sak, "Unidirectional long short- term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. Brisbane, Australia: IEEE, 2015, pp. 4470–4474.
11. Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," Signal Processing Letters, IEEE, vol. 21, no. 1, pp. 65–68, 2014.
12. Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Online speaking rate estimation using recurrent neural netwroks," in Acoustics, Speech and Signal Processing, IEEE International Conference on. Shanghai, China: IEEE, 2016.
13. M. V. Chan, X. Feng, J. A. Heinen, and R. J. Niederjohn, "Classfication of speech accents with neural networks," in Neural Networks, IEEE World Congress on Computational Intelligence., IEEE International Conference on, vol. 7. IEEE, 1994, pp. 44834486.
14. Rabiee and S. Setayeshi, "Persian accents identification using an adaptive neural network," in Second International Workshop on Education Technology and Computer Science. Wuhan, China: IEEE, 2010, pp. 7–10.

15. G. Montavon, "Deep learning for spoken language identification," in NIPS Workshop on deep learning for speech recognition and related applications, Whistler, BC, Canada, 2009, pp. 1–4.
16. R. A. Cole, J. W. Inouye, Y. K. Muthusamy, and M. Gopalakrishnan, "Language identification with neural networks: a feasibility study," in Communications, Computers and Signal Processing, IEEE Pacific Rim Conference on. IEEE, 1989, pp. 525– 529.
17. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," in Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. Florence, Italy: IEEE, 2014, pp. 5337–5341.
18. Y. Zeng, H. Mao, D. Peng, and Z. Yi, 'Spectrogram based multi-task audio classification', *Multimed Tools Appl*, vol. 78, no. 3, pp. 3705–3722, Feb. 2019, doi: 10/gnkrrs
19. D. S. Park *et al.*, 'SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition', *Interspeech 2019*, pp. 2613–2617, Sep. 2019, doi: 10/ghbzt4
20. Zhao, Huimin & Xianglin, Huang & Wei, Liu & Yang, Lifang. (2018). Environmental sound classification based on feature fusion.
21. B. McFee *et al.*, 'librosa: Audio and Music Signal Analysis in Python', Austin, Texas, 2015, pp. 18–24. doi: 10/gf4wxc. MATEC Web of Conferences. 173. 03059. 10.1051/matecconf/201817303059.