IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 25, Issue 2, Ser. III (Mar. – Apr. 2023), PP 11-16
www.iosrjournals.org

English Accent Classification and Conversion Using Machine Learning

J. Vamsinath[1], Belapure Ashish[2] , Gangula Aravind Reddy[2] , Kammari Sai Aditya Chary[2] ,Nallani Chakravarthula Vedith[2]

[1] Asst. Professor, Dept. of Computer Science and Engineering, VNR VJIET, Hyderabad - 500090, TS, India
[2] Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Vignana Jyothi Nagar,Pragathi Nagar, Nizampet, Hyderabad, India

Abstract
Considered to be the main means of human communication language. The language of worldwide communication, English is widely spoken around the world. Accent recognition is a significant issue in modern technology. Because of differences in their accents, even a speaker who is unfamiliar with the accent as familiar as the local speaker occasionally has trouble understanding. For the purpose of bridging these communication barriers, we have proposed a system to classify and transform speech between people caused by the individuality of their accents. We think machine learning can help with accent recognition. This research aims to increase spoken English accent recognition's precision. An innovative method of increasing the number of tactics by using Convolutional Neural Networks (CNN), which improves accuracy of accent classification. Additionally, English accent conversion refers to the process of changing one English accent into another so that the listener can comprehend the speaker's accent. Speech recognition and text-to-speech modules can aid with this.

Keywords: *CNN, MFCC, Accent recognition, English accent conversion.*

Date of Submission: 07-04-2023 Date of Acceptance: 20-04-2023

I. Introduction

Speech recognition systems can recognise accents of non-native speakers of a language. As we know, accented speech often contains phones or sounds which are not common in standard pronunciation. They have been trained to understand the accents and accent classification can enhance speech recognition systems by classifying the speaker’s ethnicity and classifying the accents using the trained data. Accent recognition or classification, which provides identification of a speaker’s origins or ethnicity, is important in applications like crime investigation. In real life applications, it becomes crucial to recognize accents from short audio clips or audio snippets. Furthermore, only understanding or recognizing a person’s accent is not sufficient in most cases; it is highly desirable to even be able to convert speech recognized from one accent to another. For both accent detection and conversion, technologies such as Machine Learning can offer robust solutions. Our research findings indicate that combinations of various neural networks could be used to process audio parameters to both detect as well as convert the accent portion of speech. Main challenge here is trying to use non-parallel data, that is, speech data where source and target data utterances are not the same, which is the case in real world systems. CNN is what we propose to use predominantly for accent detection modules whereas for accent conversion modules we propose to use gTTS. Taking into account that the focus of this project is the tourism industry, we have only limited ourselves to English Language and its accents as it is a language which is very widely used throughout the world. With the availability of enough relevant data, taking our model as a basis, similar solutions can be provided for other regional languages as well.

II. Literature Survey

A Neural Network Approach to Accent Classification

Through the use of MFCC taken from the speaker's audio, this study aims to classify distinct accents. and to forecast the accent using a Convolutional Neural Network model. The audio files undergo data preprocessing before being fed into the model. The MFCCs are extracted from the audio data to do this. The dataset is ready to be fed into the neural network after the extraction has been completed. For this project, they extracted 13 MFCC features from the audio files. After testing the model on the testing data, an accuracy of 62.81% has been achieved.

Speech Accent Detection in Video games

They described a study on how to train a deep neural network (Alexnet) to automatically categorize audio recordings with accents in this article. They used audio recordings taken from a video game to train AlexNet using the Speech Accent Archive data. accuracy of the network. For instance, a video game producer may find it helpful to employ a network that can identify the veracity of rage exhibited by a voice actor in the initial screening of character audition submissions. A test is run to build a confusion matrix after the parameters have been identified. 60% of the files were identified correctly by the network.

Foreign accent classification using deep neural nets

The acoustic patterns that a person's voice exhibits are what the authors of this study are particularly interested in. There are some differences between individuals in these auditory patterns. The rationale for the prediction model is that these variances are more pronounced if they are country-specific. In this work, CNN and CRNN, two DNN architectures, were employed. They also discovered that the CRNN is more efficient than the CNN architecture.

English Language Accent Classification and Conversion using Machine Learning

In this study, the authors make the suggestion that combinations of different neural networks may be utilized to process audio parameters in order to both detect and convert the accented speech. By utilizing non-parallel data, in which the source and destination data's utterances diverge from one another and from those of the real world, presented their biggest hurdle. For the accent conversion module, they recommend using GANs rather than CNNs and RNNs, which are employed in the accent detection module.

Accent classification in human speech biometrics for native and non-native English speakers

Through the classification of place, this exploratory study investigates various techniques for accent recognition in people. The speech recognition dataset included seven distinct phonetic sounds that were each spoken 10 times by test subjects from Mexico and the United Kingdom. Each dataset contained 26 MFCC logs that were retrieved at a varying time period of 200 ms, and each piece of information contained the 26 MFCC features that were matched to the accent of the speaker. While Chihuahua and Mexico City provided the Mexican accents, London and the West Midlands provided the British accents. The weights of all four classes were balanced to replicate a dataset with an equal distribution. Training of dataset using HMM and prediction, the dataset was represented as a time series (with relational properties). The classifying capability of Information Gain.

Accent Conversion Using Phonetic Posteriorgrams

This paper proposes a method for matching frames between two speakers based on audio similarity. This technique uses post-speech grams to link audio he frames from source and target speakers. Our argument is simple: if a speech recognition engine trained on native data concludes that an L2 speech segment is very similar to a native segment that produces a particular phoneme, then we should consider it to produce the same phoneme. It makes sense to match against native segments. specify. In addition, our method uses an articulatory synthesizer based on unit selection to replace mispronounced L2 diphons with those from the L2 corpus that match the reference articulatory configuration.

Voice Impersonation using Generative Adversarial Networks

In this research, an unique method for conversion of voice using data that has been trained is proposed. An SI-ASR (Speak-Independent Automatic Speech Recognition) system that does not depend on the speaker which produces phonetic posteriorgrams (PPGs). PPGs can depict how speech sounds are articulated in a speaker-normalized space. One kind of generative model that may be taught to produce samples that closely resemble pulls from the real distribution of the data is the Generative Adversarial Network (GAN). GANs are trained to be discriminative, so samples produced by the model are indistinguishable from samples taken from the original data. It can be seen that the VoiceGAN model can change a speaker's style. Other stylistic traits that might be recognised can easily be included in the process. In theory, though. Longer-term prosodic-level stylistic characteristics could potentially.

Voice Conversion Using Artificial Neural Networks

The spectral properties of a source speaker can be translated to a target speaker using artificial neural networks (ANN). Modern Gaussian Mixture Model (GMM) and ANN are contrasted in terms of voice conversion. The outcomes demonstrate that ANNs produce outputs that are understandable and can modify speech more successfully than GMMs. A continuous speech stream can be spectrally modified using an ANN in the voice conversion framework. Many speaker pairs have been used in the demonstrations of ANN's efficiency.

ANN-based spectral transformation surpasses GMM in terms of both objective and evaluatively-based results, as can be shown when comparing the two methods.

Non-native speech conversion with consistency-aware recursive network and generative adversarial network

This paper proposes a method based on similarity of the audio between the two speakers, frame matching. This technique uses post-speech grams to link audio frames from both the speakers. i.e source speaker and destination speaker. Our argument is simple. An L2 speech segment that produces a certain phoneme is found to be remarkably close to a native segment by a speech recognition engine trained on native data; it should be assumed that it produces the same phoneme. It makes sense to match the native segment. show. In addition, our method uses a unit-selection-based articulatory synthesizer in place of mispronounced L2 diphons compared to that of L2 corpus that conform to the articulatory layout of the reference.

Articulatory-based conversion of foreign accents with deep neural networks

They have provided an articulatory technique for the modulation of non-native accents in this study. DNN and MFCCs are components of the architecture. The GMM approach and the DNN accent-conversion method have both been compared. Accent conversions made using the DNN were more understandable and seemed more natural than those made using the GMM. The difference in how accents are perceived using the two ways could be due to how the acoustic quality impacts how non-native accents are perceived.Compared to the GMM, the DNN typically synthesized speech with better acoustic quality and despite the fact that both methods involve articulatory synthesis, according to a recent study.

Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network

In this study, spectrograms and DCNN are used to demonstrate a technique for speech emotion identification (CNN) The deep CNN is fed spectrograms produced from the voice inputs. According to preliminary findings, the suggested strategy based on a recently trained model performs better than a finely calibrated model. With the use of a feature learning strategy based on Deep CNNs, we attempt to address the SER problem in this research. Deep CNNs receive input in the form of spectrograms, which represent speech signals. To enhance SER performance even further, we intend to employ more data along with moderately complex models.

Accent Conversion Using Artificial Neural Networks

In this study, we present a technique for accent conversion that produces a series of transformation matrices that can be used to Mel Frequency Cepstral Coefficients and discovers the distinctions between two accents. This is accomplished using a feedforward artificial neural network, alignment preprocessing, MCD, and a softmax classifier for validation. The feedforward architecture successfully modifies the MCDs of an accent sample, but it fails to take into account other speech characteristics that aren't represented by MFCCs. The focus of future study should be on expanding the model for reconstruction. It's also crucial to look into how well RNNs can capture temporary data.

Phonetic Posteriorgrams For Many-to-one Voice Conversion Without Parallel Data Training

They recommended conducting this study's voice conversion utilising without data for parallel training. The posterior probability of each phonetic class for each distinct time frame of one utterance is shown in a time-versus-class matrix, or PPG. The suggested method is initially used to acquire the PPGs of the target speech. A DBLSTM structure can then be used to represent the connections between the PPGs and the target speech's acoustic features. This method allows for many-to-one conversion without the need for ongoing training data.

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

The technique of identifying a speaker's original language from speech that has a foreign accent is known as accent classification. This study's authors propose a system for categorizing 11 dialects based only on speech acoustics. The suggested accent recognition approach consists of using DNNs and RNNs together that have been trained on data across the long and short terms, respectively. The outcome demonstrates that the suggested system outperforms the conventional SVM-based system.

Voice Conversion Using Deep Bidirectional Long Short-term Memory Based Recurrent Neural Networks

Voice-to-voice (VC) is a technology that modifies your voice so that it sounds like your target speaker. A novel speech-to-speech technique using DBLSTM and RNN can replicate both the frame-by-frame interaction between source and target languages and the long-term contextual dependence of auditory patterns. Experimental findings demonstrate that both unbiased and arbitrary metrics recommended methods for

DBLSTM-RNN significantly improves the consistency and naturalness of transformed speech by boosting the MOS from the range of 2.3 - 3.2. It shows that we can improve.

Accent Classification Using Support Vector Machines

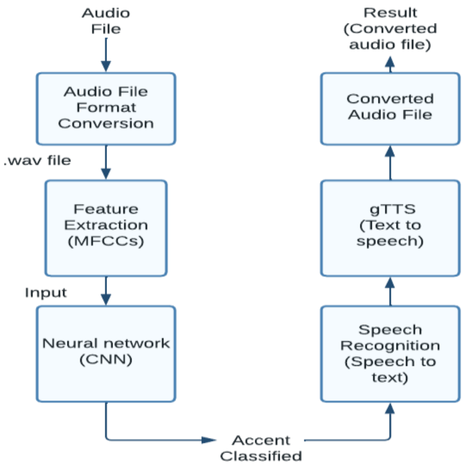
This study, We study an accent classification system using an input SVM with MFCC features on time-based segments. The audio sample size and time segment length required for efficient performance are considered for three audio samples with different accents. The mean sequence of MFCCs for a given sample served as a trait vector for a given subject. We used leave-one (speaker)-out cross-validation (LOO) to evaluate performance with an emphasis on accuracy, precision, recall metrics, and ROC curve analysis.

III. Challenges faced by Existing System

The current approach for categorising and converting English accents suffers a number of problems, including a lack of data and being limited to a small number of accents. The algorithms employed in current systems include RNN, DNN, GMM, and DBLSTM. These algorithms are all inefficient in terms of computation. To overcome these difficulties, the accuracy and generalisation skills of the current system must be improved.

IV. Methodology

The proposed system is to classify as well as convert accents. It consists of the following two parts : Accent classification and Accent Conversion. Accent classification is to predict the native or non-native region of the speaker. Accent conversion is to convert an accent to a selected target accent.

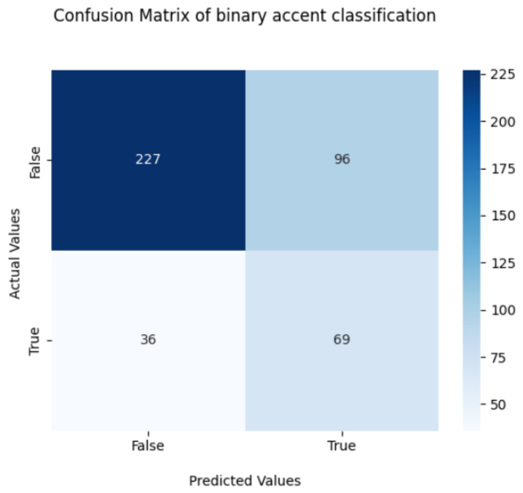

Audio File
Audio File Format Conversion
.wav file
Feature Extraction (MFCCs)
Input
Neural network (CNN)
Accent Classified
Speech Recognition (Speech to text)
gTTS (Text to speech)
Converted Audio File
Result (Converted audio file)

V. Results

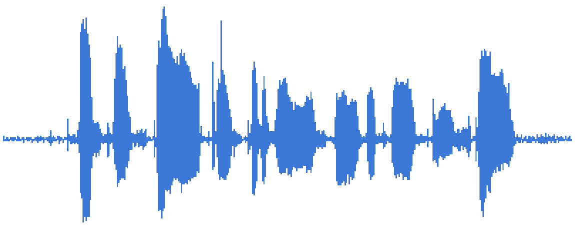
There were 2138 samples obtained in all. The 0.2 test size was chosen. The number of native english accents is around 579 and the number of non- native english accents is about 1559.

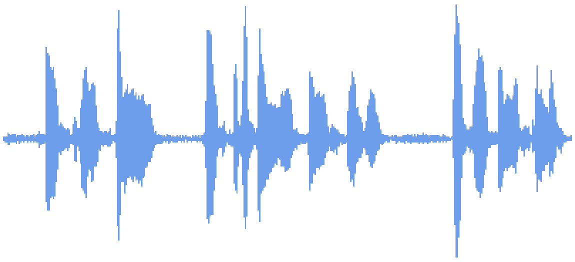
Accent Classification:

The number of training epochs used was 20 and 30. The two models' accuracy values are 0.6752 and 0.6916. We choose the second model in this instance since it provides the maximum accuracy. Here is the confusion matrix.



Accent Conversion:
These are the segments of the spectrograms of a sample audio file before and after conversion.





VI. Conclusion

We have worked to develop a system that makes accent classification easier and more accurate by utilising CNN. In comparison to more conventional methods, the gTTS module for accent conversion yields favourable and accurate results. We have established a well-optimized accent categorization and conversion system that users may easily use based on our findings and analysis.

References

[1]. Duduka, S., Jain, H., Jain, H.P.V. and Chawan, P.M., 2021. A Neural Network Approach to Accent Classification. International Research Journal of Engineering and Technology (IRJET), 8(03), pp.1175-1177.

[2]. Ensslin, A., Goorimoorthee, T., Carleton, S., Bulitko, V. and Hernandez, S.P., 2017, September. Deep Learning for Speech Accent Detection in Video Games. In Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference.

[3]. Singh, U., Gupta, A., Bisharad, D. and Arif, W., 2020. Foreign accent classification using deep neural nets. Journal of Intelligent & Fuzzy Systems, 38(5), pp.6347-6352.

[4]. Parikh, P., Velhal, K., Potdar, S., Sikligar, A. and Karani, R., 2020, May. English language accent classification and conversion using machine learning. In Proceedings of the International Conference on Innovative Computing & Communications (ICICC).

[5]. Bird, J.J., Wanner, E., Ekárt, A. and Faria, D.R., 2019, June. Accent classification in human speech biometrics for native and non-native english speakers. In Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to

Assistive Environments (pp. 554-560).

[6]. Zhao, G., Sonsaat, S., Levis, J., Chukharev-Hudilainen, E. and Gutierrez-Osuna, R., 2018, April. Accent conversion using phonetic posteriorgrams. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5314-5318). IEEE.

[7]. Gao, Y., Singh, R. and Raj, B., 2018, April. Voice impersonation using generative adversarial networks. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2506-2510). IEEE.

[8]. Desai, S., Raghavendra, E.V., Yegnanarayana, B., Black, A.W. and Prahallad, K., 2009, April. Voice conversion using artificial neural networks. In 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 3893-3896). IEEE.

[9]. Aryal, S. and Gutierrez-Osuna, R., 2015. Articulatory-based conversion of foreign accents with deep neural networks. In the Sixteenth Annual Conference of the International Speech Communication Association.

[10]. Oyamada, K., Kameoka, H., Kaneko, T., Ando, H., Hiramatsu, K. and Kashino, K., 2017, December. Non-native speech conversion with consistency-aware recursive network and generative adversarial network. In 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 182-188). IEEE.

[11]. Badshah, A.M., Ahmad, J., Rahim, N. and Baik, S.W., 2017, February. Speech emotion recognition from spectrograms with deep convolutional neural networks. In 2017 international conference on platform technology and service (PlatCon) (pp. 1-5). IEEE.

[12]. Bearman, A., Josund, K. and Fiore, G., 2017. Accent conversion using artificial neural networks. Stanford University, Tech. Rep, Tech. Rep..

[13]. Sun, L., Li, K., Wang, H., Kang, S. and Meng, H., 2016, July. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In 2016 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

[14]. Jiao, Y., Tu, M., Berisha, V. and Liss, J.M., 2016, September. Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features. In Interspeech (pp. 2388-2392).

[15]. Sun, L., Kang, S., Li, K. and Meng, H., 2015, April. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4869-4873). IEEE

J. Vamsinath, et. al. “English Accent Classification and Conversion Using Machine Learning.” *IOSR Journal of Computer Engineering (IOSR-JCE)*, 25(2), 2023, pp. 11-16.