iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VII **Month of publication:** July 2021

DOI: https://doi.org/10.22214/ijraset.2021.37234

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

Improving Pronunciation for Non-Native Speakers Using Neural Networks

R. Santhoshi[1], Sreekaran Srinath[2], A. Joseph Sathiadhas Esra[3]
[1]*Dept. of ECE, Meenakshi Sundararajan Engineering College*
[2]*Dept. of ECE, Meenakshi Sundararajan Engineering College*
[3]*Assistant Professor, Dept. of ECE, Meenakshi Sundararajan Engineering College, Tamil Nadu*

Abstract: ***While learning a new language through the internet or applications, a lot of them focus on teaching words and sentences and do not concentrate on the pronouncing ability of the users. Even though many speakers are proficient in a language, their pronunciation might be influenced by their native language. For people who are interested in improving their pronunciation capabilities this proposed system was introduced. This system is primarily focused on improving the pronunciation of English words and sentences for non-native speakers i.e., for whom English is a second language. For a given audio clip, we scale the audio and extract the features, input the features to the model developed and the output of the model gives the phonemes that are spoken in the clip. Many models detect phonemes and various methods have been proposed but the main reason for choosing deep learning is that the learning and the features that we tend to oversee or overlook are picked up by the model provided the dataset is balanced and the model is built properly. The features to be considered varies for every speech processing project and through previous research work and through trial and error we choose the features that work best for us. Comparing the phonemes with the actual phonemes present, we can give the speaker which part of their speech they need to work on. Based on the phoneme, feedback is given on how to improve their pronunciation.***
Keywords: Mispronunciation detection, Feedback analysis, Neural Network, Speech processing, Phoneme

I. INTRODUCTION

While trying to learn a new language, a lot of applications online help with the teaching. Not many applications guide the users and provide them feedback as they are learning the language. This paper addresses this concern and aims to improve the pronunciation of non-native speakers trying to learn English with the help of neural networks. Computer-aided pronunciation tools that use multiple methods have been developed over the years, this paper focuses on using neural networks to build models to calculate the pronunciation and give feedback based on it. For many of us, English is a second language and while trying to learn a language we learn the structure and grammar of it. Pronunciation is not something that we work on just by listening but by getting active feedback. Also, not everyone has the time and resources to attend classes or have a tutor, to build a basic prototype of what we think would be a good solution to this problem is our motivation. While non-native learners of languages can easily upgrade their reading comprehension and writing skills, there is a lack of tools to help them improve pronunciation and speaking skills. While trying to learn a new language, we bring the accent, the method of pronouncing our mother tongue, our native dialect into the new language that we are trying to learn. Usually, the mistakes made while learning a language from another language are quite similar. Once we understand where exactly we are going wrong, it's much easier to learn the nuance of the new language. While the range of words is high the number of phonemes is just 44. Words are the unique combinations and meanings of these 44 sounds. Analysing this would thus enable us to create a model such that we can enable them to voice out the words in a better way. The feedback given to the users is based on linguist suggestions on how the mouth position should be for each phoneme. This is standardized and for 44 phonemes, we have 44 methods on how to pronounce each one respectively.

II. LITERATURE REVIEW

A. For Modelling and Feature Selection

This paper[2] forms the base for the features that are to be considered for the model that we propose. The paper is focused on Arabic phonemes but our focus is based on English phonemes. The paper has classified phonemes using a Deep convolutional neural network as well as transfer learning models. They also tried preprocessing the data and feature selection before inputting it to the model as well as using the model to select features on its own. This was compared and contrasted as well.

B. For Analyzing the Learning Pattern of non-native Speakers

This paper[1] focuses on the people trying to learn Hindi whose native language is Tamil. The contrast between the two languages in terms of speech is that the latter doesn't have a lot of aspirated unvoiced stops. Thus the learners substitute these phones with the unaspirated unvoiced stops. They sample the audio at the rate of 16KHz for this process. They compute the likelihood ratio computed for the speakers helps in giving a good separation of distribution. The paper also emphasizes that acoustic phoneme features are better than cepstral features in terms of the statistical likelihood pronunciation model.

III.PROPOSED MODEL

A. Dataset

For this paper, we used the DARPA TIMIT acoustic-phonetic Continuous Speech Corpus (TIMIT) dataset. This is a common dataset used for speech analysis and evaluation. It consists of 6300 sentences. Ten speakers spoke 630 sentences each. The table provides the gender and dialect distribution of the sentences. This dataset was considered because it gave a detailed analysis for the phonemes, words, sentences pronounced along with their timestamps which can be used for further modelling. The sentences are also phonetically rich to give the contrasts and help understand the features better.

TABLE I
Dialect distribution for complete test set

Dialect	Male	Female	Total
1	7	4	11
2	18	8	26
3	23	3	26
4	16	16	12

B. Features Extracted

1) *Mel-Frequency Cepstral Coefficients:* Mel frequency cepstral coefficients are the coefficients that together make the MFC. The coefficients are extracted using filter banks. Since the MFC is non-linear the frequency bands are equally spaced on the mel scale and thus give a better representation of the sounds as perceived by the human ears and provide a better depiction of vocal tract resonances. Thus they give the overall shape of the spectral envelope. The signal is windowed, DFT is applied. The log of the magnitude is taken and warped in the mel scale to which DCT is applied to get the MFCC features.

$$mel(f)=2595xlog10(1+f/700)\ E1$$

2) *The Short-Time Fourier Transform:* Short-time Fourier transform is the Fourier transform taken over a short interval of time. It helps depict the changes over a short interval of time easily. It gives the representation of the amplitude versus time and frequency for a given signal. It approximates the time-frequency analysis performed by the ear. It also models on a short spectrum. STFT is computed as consecutive FFT for a windowed frame that continuously slides and hops in the time axis. It is a very useful audio processing tool.
3) *The Mel Scale:* It is the logarithmic scale of the given signal frequency. It helps represent the human voice better. Humans cannot hear high-frequency signals but can find the difference in the low-frequency range. The mel scale provided a better resolution to find the cepstral features of the signal in the lower frequency range to help distinguish the sound better. It is modelled like that of the human ears. The transformation from the Hertz scale to the Mel Scale is the following:

$$M = 1127.log\ (1+f/700)$$

4) *Spectral Contrast:* It is the decibel difference between the peak and valley of a spectrum. Each phoneme has a unique spectrum and can be represented as a vector. If the difference is zero degrees then it's the same but, if it's 90 degrees difference then they are different. All the above features are extracted with the help of a python library known as Librosa.

C. Data Preprocessing:

Each audio file is sampled and then the time at which each phoneme occurs has been calculated with the help of the TIMIT file. Using the sampled audio file and the time, the audio signal is broken into segments. Features were extracted for each audio frame and the values were stored in a data frame. A total of 167 features were extracted for all the phoneme segments.

D. Model

The number of features extracted from the given audio clip is 167 for each phoneme present in the audio clips of the dataset. These features include the Mel Frequency Cepstral Coefficients, Mel Frequency Spectral Values, Chroma Features, and Short Time Fourier Transform. These 167 features are given as input to the model. Thus the number of nodes or the input layer contains 168 nodes (167 features + 1 for bias).

A hidden layer configuration uses just two rules: (i) the number of hidden layers equals one; and (ii) the number of neurons in that layer is the mean of the neurons in the input and output layers. This constraint along with trial and error resulted in creating two hidden layers of 128 nodes each with 0.5 dropout class and ReLU optimizer. There are a total of 61 unique phonemes present in the dataset. There are 44 phonemes in the English language. The TIMIT dataset uses the CMU dictionary phoneme. Since the model is classification-based, we have 61 output nodes with a sigmoid activation function.

E. Classification

The preprocessed features are fed into the model and the classification is predicted. Upon prediction, the classified phonemes are matched to their respective labels. The sentence is then compared to these labels and is stored as a dictionary.

F. Feedback Model

The actual pronunciation is compared with the model predictions and the feedback is generated. To improve the pronunciation we help the users by telling which part of the sentence they have mispronounced along with how to pronounce by giving them feedback on the mouth positions to pronounce the sound. The output of the model is saved and compared with the actual phonemes of the audio file. Thus, inaccurate phonemes are stored. The TIMIT dataset has 61 classes, this and the IPA are mapped together to give feedback. Most languages, including English, can be described in terms of a set of distinctive sounds, or phonemes. In particular, for American English, there are about 42 phonemes including vowels, diphthongs, semi-vowels, and consonants. The internationally standard method to represent phonemes is International Phonetic Alphabet (IPA). For a better representation of the computers, we code them as ASCII characters and several schemes such as TIMIT, CMU, WSJ, ARPABET have been proposed. The phoneme is mapped with the respective TIMIT to IPA and the feedback is given based on Table II[13].

TABLE II
Feedback Based On Phoneme

Phoneme	Mouth Feel
/a/	Jaw and tongue are down.
/b/	Lips start out together, then they open and a puff of air comes out. Voice box on.
/d/	Tip of tongue touches above your top teeth. Voice box on.
/e/	Mouth open, tongue behind bottom teeth.
/f/	Top teeth touch your bottom lip.
/g/	Mouth is open, tongue humped at back of mouth. Voice box on.
/h/	Air comes out of open mouth.
/i/	Mouth is open, tongue is slightly lowered.
/j/	Tongue is up, lips are open.
/k/ (letter c)	Tongue is humped in the back of your mouth.
/l/	Tip of tongue touches above your top teeth and stays there.
/m/	Lips come together.
/n/	Tongue behind top teeth, air comes out your nose.
/o/	Mouth open, jaw dropped.
/p/	Lips open and a puff of air comes out.
/kw/ (letters qu)	Back of tongue is humped in back of mouth, lips make a circle.
/r/	Tongue curls up to roof of mouth. Voice box on.
/s/	Tip of tongue touches above your top teeth. Then blow a hiss of air.
/t/	Tip of tongue touches above your top teeth.
/u/	Mouth is open, tongue is down.
/v/	Top teeth touch bottom lip. Voice box on.
/w/	Lips make a circle.
/ks/ (letter x)	Begins with back of tongue humped in back of your mouth. Then hiss.
/y/	Tongue is touching side of teeth, mouth is open.
/z/	Tip of tongue touches above your top teeth. Voice box on.

IV. RESULTS

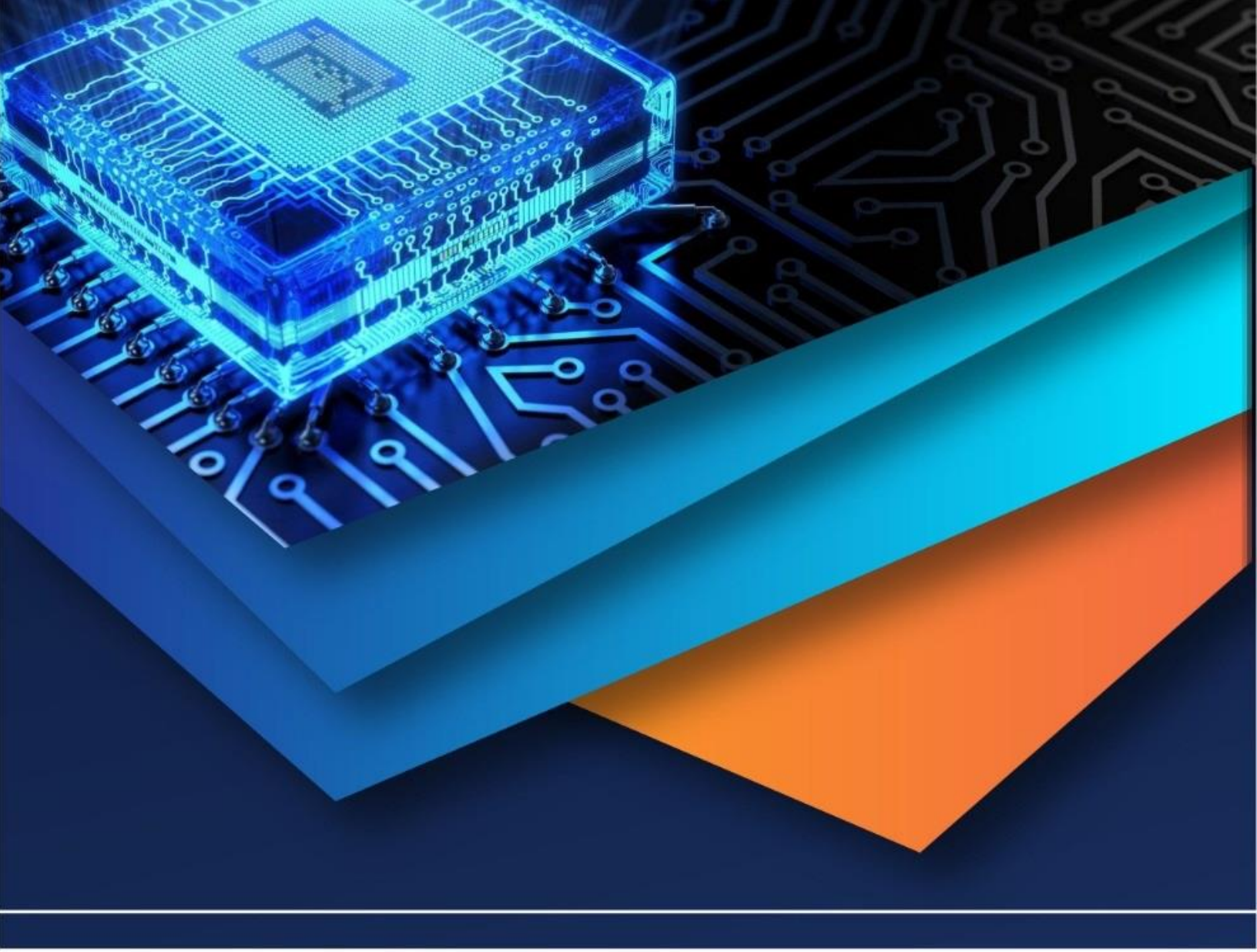
The model performed with an accuracy of approximately 45% and the feedback for the users were given accordingly. The proposed model thus extracted the features from the audio file and gave the phonemes detected in the same. Comparing this with the reference phonemes pronounced, we found the mispronounced phonemes. Based on this, we gave feedback on how to improve pronunciation. The phoneme detection was done by neural networks which is a promising field for speech and audio processing.

V. CONCLUSIONS AND FUTURE SCOPE

To process audio files in real-time, we would need various noise removal, voice enhancement mechanisms to remove silences in parts of speech. There are varying dialects in different regions, hence audio needs to be gathered and utilized to build better models. The model also needs to automatically extract features based on the given audio signal and classify the phonemes without preprocessing. The feedback model can be developed from a linguistic perspective and can be customized to the native speaker's language and dialect. These are the aspects that need to be worked on in the future scope.

REFERENCES

[1] Vaishali Patil, Preeti Rao, “Automatic pronunciation assessment for language learners with acoustic-phonetic features” in Proceedings of the Workshop on Speech and Language Processing Tools in Education, pages 17–24, COLING 2012, Mumbai, December 2012 .
[2] F. Nazir, M. N. Majeed, M. A. Ghazanfar and M. Maqsood, "Mispronunciation Detection Using Deep Convolutional Neural Network Features and Transfer Learning-Based Model for Arabic Phonemes," in IEEE Access, vol. 7, pp. 52589-52608,2019,doi:10.1109/ACCESS.2019.2912648.
[3] Agarwal, C., Chakraborty, P. A review of tools and techniques for computer aided pronunciation training (CAPT) in English. Educ Inf Technol 24, 3731–3743 (2019). https://doi.org/10.1007/s10639-019-09955-7.
[4] https://towardsdatascience.com/calculating-audio-song-similarity-using-siamese-neural-networks 62730e8f3e3d
[5] https://www.lingualift.com/blog/best-language-learning-apps/
[6] https://jonathan-hui.medium.com/speech-recognition-gmm-hmm-8bb5eff8b196
[7] https://jonathan-hui.medium.com/speech-recognition-series-71fd6784551a 8. https://youtu.be/wA9--WJSPws
[8] https://jonathan-hui.medium.com/speech-recognition-phonetics-d761ea1710c0
[9] https://librosa.org/doc/latest/index.html
[10] https://keras.io/api/
[11] https://www.tensorflow.org/api_docs/python/tf
[12] https://www.kaggle.com/nltkdata/timitcorpus
[13] http://wp.auburn.edu/rdggenie/home/teaching-ideas/mouthmoves/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY