IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

MISPRONOUNCIATION RECOGNITION SYSTEM USING CNN AND RNN ALGORITHM

NIVETHA.V
STUDENT
DEPT OF COMPUTER SCIENCE AND ENGINEERING
MANAKULA VINAYAGAR ENGINEERING COLLEGE,
PUDUCHERRY

Dr.T.MEGALA
ASSISTANT PROFESSOR
DEPT OF COMPUTER SCIENCE AND ENGINEERING SRI
SRI MANAKULA VINAYAGAR ENGINEERING COLLEGE
PUDUCHERRY

ABSTRACT

English is not taught in the traditional method of teaching speaking skills but it is taught through writing and grammar. However, pronunciation difficulties must be addressed although some of these may be insurmountable for today's AI models. Deep learning has been successful in various tasks as it can recognize complex patterns and representations than any other model. This application attempts at using a deep learning algorithm to predict how well someone pronounces words. It uses deep learning to provide better results faster. The current algorithms have limitations when dealing with local pronunciation problems so this approach also aims to enhance English language training especially on oral communication skills using deep learning methods

Keywords: machine learning models, embracing deep learning, deep learning.

1.INTRODUCTION

The phenomenon of increasing globalization has made people less likely to value their speaking skills when learning a language. The process of learning how to speak a language can usually be divided into two parts: learning grammar, structure, idioms etcetera; and acquiring good pronunciation. Deep learning refers to those kinds of machine learning which belong to neural networks consisting of no fewer than three layers. These networks aim at mimicking the functionality in human brains that allows them learn from massive volumes of data. A single-layer network may yield approximate predictions even though an additional hidden layer could greatly enhance the model's accuracy in making predictions by optimizing and refining it further. In deep learning, machines gain independence from humans and can comprehend as well as perform tasks without any intervention. Visual content, written language or audio input

can all be used for classification tasks. It is often the case that deep learning models make exceedingly accurate predictions surpassing even human performance levels frequently but not always becoming educated about

something only when given enough information about it. The significance of English as the world's most widely spoken language is highlighted by using a large amount of annotated data and different kinds of neural network architectures. Consequently, students who are learning English pay much attention to speaking skills development. Speaking practice is believed to be important in improving one's ability to communicate well in English As such, it requires timely and relevant feedback for correction. In teaching Chinese students English, teachers usually adopt the traditional way which involves listening to a recording, reading the text aloud and repeating sentences after a language repeater. However,this routine lacks immediacyof effect.

Feedback may fail to establish an explicit relationship between students' pronunciation and machine speech. Also, no input during practice makes it difficult for learners to connect spoken machine language with their reading abilities . Prompting exact feedback for classroom students poses challenges for teachers because it does not allow them realize their mispronunciations quickly enough for corrective action.. With the development and availability of computers, computer-assisted instruction has become an important part of educational technology today. Most of them aim at helping learners use and understand pronunciation while learning a language software neglects oral communication skills. Previously, computer-assisted language learning was mainly concerned with grammar, sentence structure, idiomatic expressions ad pronunciation among others. However, it should be known that accurate pronunciation is very important because it affects greatly on how well someone can express themselves in speaking or writing English for example. In this case many people tend to forget about improving their spoken skills when they are studying another language. Generally, the process of acquiring spoken language involves two major parts: understanding grammar, structure, idioms etcetera and being able to pronounce words correctly or accurately if you will. Deep learning is one type of machine learning which falls under the neural network family characterized by having more than three layers. These neural networks aim to replicate the conditioning of the mortal brain, allowing it to gather new information from large volumes of data. of fresh layers in a neural network can greatly ameliorate its capability to directly prognosticate issues. The fresh layers that aren't readily apparent have the eventuality to make delicacy more by optimizing and enriching the model. Our design, the Automatic Pronunciation Mistake Detection, is a dependable system that efficiently corrects pronunciation crimes in English. The purpose of

this tool is to help scholars or druggies enhance their capacities in pronouncing words rightly. The thing of the design is to ameliorate the delicacy of error discovery by exercising Speech recognition, pyaudio, and pyttsx3 to drop the error rate effectively. In the field of deep literacy, a machine algorithm gains the capability to understand and perform bracket tasks using visual accoutrements , written words, or audio cues. Deep literacy models retain the capability to achieve extremely precise prognostications, constantly surpassing mortal performance. Models are trained using a comprehensive gathering of data and neural network infrastructures that cover a wide range of complexity situations.

2.RELATED WORK

A model called arbitrary timber is being appertained to. In the history, English education in China has concentrated primarily on written chops and alphabet, disregarding the significance of enhancing oral language capacities. This emphasis has led to a situation in which multitudinous Chinese scholars perform exceptionally well in written English examinations but face difficulties in effectively expressing themselves in English in their everyday lives. As transnational communication becomes further pivotal, there's a growing understanding that the focus of language literacy should be on developing oral language chops. still, the process of tutoring spoken English has its difficulties, especially when it comes to giving scholars helpful feedback during one- on-

one oral practice sessions. This understanding has urged the disquisition of computer-grounded results to attack these difficulties, specifically targeting models that automatically correct pronunciation.

A Support Vector Machine(SVM) is employed in a system designed to identify crimes.This system, offered for relating pronunciation crimes in English as a alternate language, is a fresh approach that merges confidence scoring at the phone position with corner- grounded Support Vector Machines(SVMs). This new approach focuses on specific phonemes that L2 learners, specifically Korean learners in this exploration, constantly struggle with and make miscalculations in. The findings showed that when dealing with data containing multitudinousnon-phonemic crimes, the SVM system yielded a vastly advanced F- score(0.67) in comparison to counting solely on confidence scoring(0.60). The exploration points out how pivotal it's to acclimatize styles for detecting pronunciation crimes to the difficulties encountered by learners of a particular language. It demonstrates the effectiveness of using both confidence scoring and SVMs grounded on milestones to enhance delicacy, especially when dealing with problematic sounds.

Discovery of pronunciation crimes and generation of feedback for call operations.

The composition presents a new system for automatically relating crimes in Computer supported Language Learning systems, combining both verbal knowledge and ultramodern speech technology. This point of having two functions makes it unique, as the classifier is excellent at both relating miscalculations in phone operation in alternate language(L2) speech, similar as negotiations and deformations of phonemes. The CAPT system becomes further effective when it detects crimes with a high position of perfection. This perfection allows for a more thorough analysis of pronunciation performance. By combining verbal knowledge, advanced speech technology, and multimedia support, this system provides a complete result for CAPT systems. It not only directly detects crimes but also offers instructional feedback for language learner

The study aims to descry incorrect pronunciation innon-native speech through the use of aural model and convolutional intermittent neural networks. The composition explores the use of a CRNN model, which integrates convolutional neural networks and long short- term memory networks, combined with connectivity time series bracket. This model aims to convert aural signals into pinyin marker sequences. This system is specifically designed for Detecting mispronunciations in Mandarin within the frame of Computer- backed Language Learning. The development of language chops is personalized and people learn at their own speed. a process of thorough disquisition and examination, one can acquire a deeper understanding and sapience into a subject. The study of language and how it's used orders, the exploration uncovers four orders. Pronunciation miscalculations are connected to the habitual way native Spanish speakers gasp words, which provides sapience into particular patterns.

Using neural spectrogram recognition, a system is designed to descry phonological crimes in pronunciation for training purposes.

The paper presents an innovative system that uses neural networks to descry miscalculations in speech pronunciation bynon-native speakers. This system is specifically designed for training programs that help with pronunciation. The neural- based system allows for a detailed analysis ofnon-native speech by rooting unique phonological groups from visible spectrogram patterns. This approach enhances stoner feedback in pronunciation training by fastening on specific speech features. The paper highlights the significance of using neural networks to classifynon-native speech parts grounded on their unique phonological orders. The use of neural-grounded technology offers stopgap in giving druggies more specific and helpful feedback, perfecting the effectiveness of computer-

supported pronunciation training, and playing a part in the development of technology- driven language literacy tools.

3.PROBLEM IDENTIFICATION

In the being traditional English education system in China, the predominant focus has been on written chops and alphabet, leading to a insufficiency in the development of oral language proficiency among scholars. This work addresses the limitations of being models by proposing an enhanced arbitrary timber(RF) model was created with the purpose of relating and amending automated pronunciation miscalculations in English assignments. Using the enhanced arbitrary timber algorithm, the model is able of determining and relating if learners have accurate pronunciation. point birth is conducted using Mel cepstral portions(MFCC) to capture essential audio signal aspects, while top element analysis is a statistical fashion used to reduce the dimension of a dataset. The testing results show table advantage. The essential capability of deep literacy algorithms to discern complex connections and excerpt hierarchical features positions them as precious tools for addressing challenges related to pronunciation delicacy. This operation of deep literacy holds pledge in advancing the capabilities of pronunciation vaticination models, communication that the integration of MFCC, PCA, and RF within a bracket frame presents a promising result for resolving pronunciation difficulties.

DISADVANATAGE IN THE EXISTING SYSTEM

The literature check on being algorithms highlights challenges in studying the original optima wavelength, particularly in the environment of algorithms where the difficulty extends to understanding original optima wavelengths. also, the limitations include the incapability to effectively study long- term dependences due to dataset constraints. The being algorithms face difficulties when applied to larger datasets, and the need for expansive labeled data for optimal performance poses a significant challenge. also, the lack of a standardized pronunciation in some cases further complicates the accurate labeling of crimes, presenting fresh hurdles in the development of effective models for tasks like automatic pronunciation correction.

4. PROPOSED SYSTEM

The challenge of pronouncing multitudinous terms in being algorithms can significantly stymie the study of original optima wavelength. In response to this issue, the focus of the work is to develop an operation that leverages the capabilities of deep literacy to prognosticate pronunciation delicacy with a high position of perfection. Deep literacy has surfaced as a important paradigm, demonstrating remarkable delicacy across colorful tasks owing to its capacity to grasp intricate patterns and representations within data. Its proficiency in managing large datasets and learning hierarchical features stands out, enabling it to generalize effectively to new and different data. In the specific environment of studying original optima, deep literacy models parade a no table advantage. This operation of deep literacy holds pledge in advancing the capabilities of pronunciation vaticination models, potentially offering further robust and environment- apprehensive results in the realm of language literacy and communication.

ADVANTAGE OF THE PROPOSED SYSTEM

The proposed system offers a distinctive advantage in its application of a deep literacy algorithm for prognosticating pronunciation delicacy. Unlike traditional approaches, deep literacy excels in landing intricate patterns and representations within data, leading to emotional vaticination delicacy. This rigidity positions the proposed system as a potent tool for addressing challenges in pronunciation vaticination, furnishing further nuanced and environment-apprehensive feedback ffor language learners.

5 .ARCHITECTURE DIAGRAM

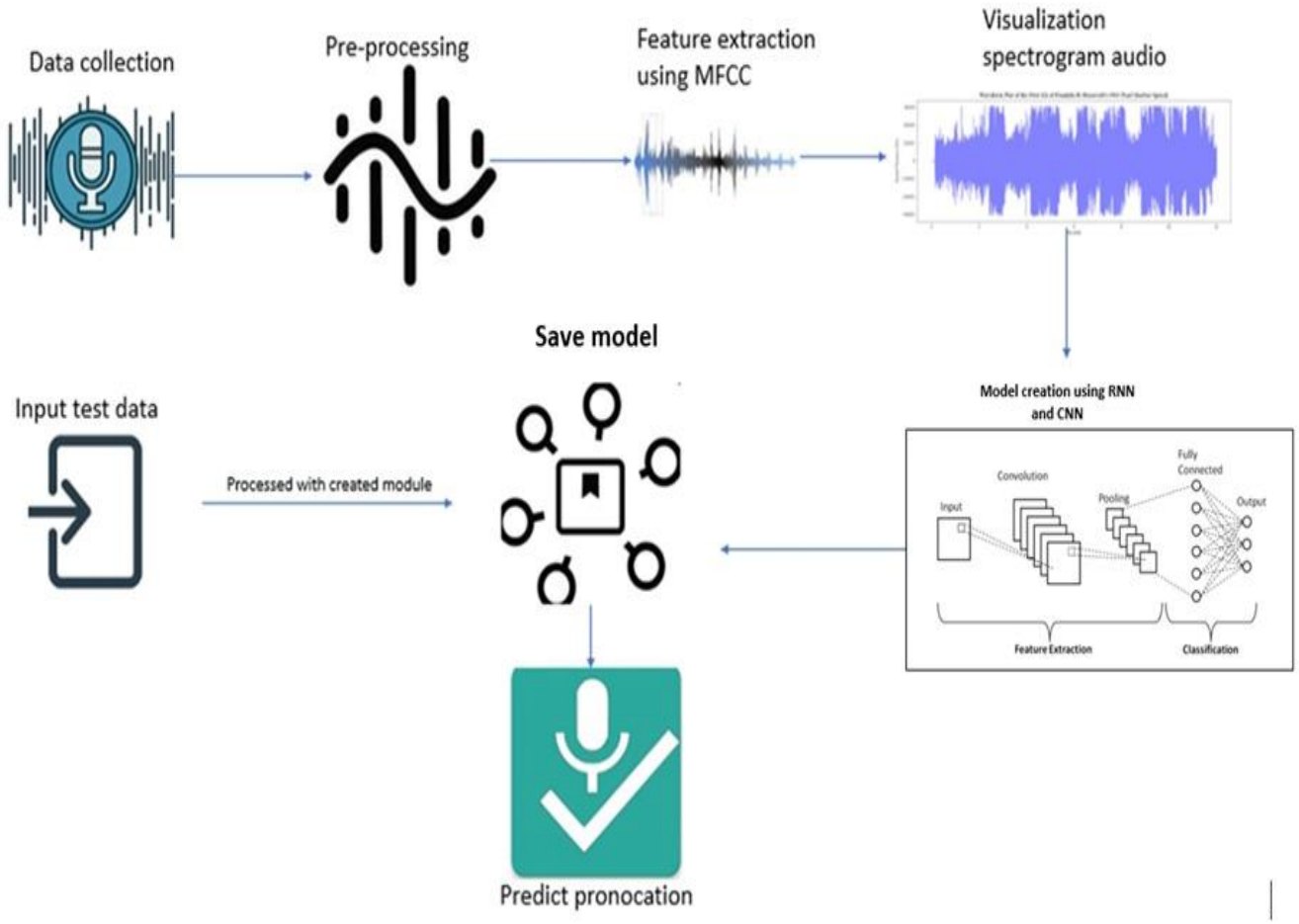


Figure 1 Speech recognition framework

An armature illustration for pronunciation using LSTM(Long Short- Term Memory) illustrates a neural network model designed to understand and induce accurate pronunciations of words or expressions. At its core, the armature consists of layers of connected memory cells, each able of retaining and recycling successional information over time. In the environment of pronunciation, the input to the LSTM network would generally be represented as sequences of phonemes, graphemes, or other verbal units, depending on the specific task or language. The cells within the network learn to capture and render the temporal dependences between these verbal units, enabling the model to prognosticate the pronunciation of words or sequences of words grounded on their contextualinformation.The input data is fed into the network, where it passes through multiple layers of cells. Each cell has gates that control the inflow of information, allowing the model to widely flash back or forget information as demanded. As the input data

propagates through the network, the LSTM cells learn to prize meaningful features and patterns

from the successional input, gradationally erecting an understanding of pronunciation rules and patterns essential in the language. During training, the model is optimized to minimize the difference between its prognosticated pronunciations and the ground verity pronunciations handed in
the training data. This is generally achieved through ways like back propagation and grade descent, where the model's parameters are acclimated iteratively to ameliorate its pronunciation accuracy Once trained, the LSTM pronunciation model can be used to induce pronunciations for unseen words or sequences of words, furnishing precious backing in tasks like textbook- to- speech conflation, automatic pronunciation correction, or language literacy operations. Overall, the armature illustration for pronunciation using LSTM showcases the intricate network of connected memory cells that enable the model to understand and induce accurate pronunciations grounded on contextual verbal information.

6. BLOCK DIAGRAM

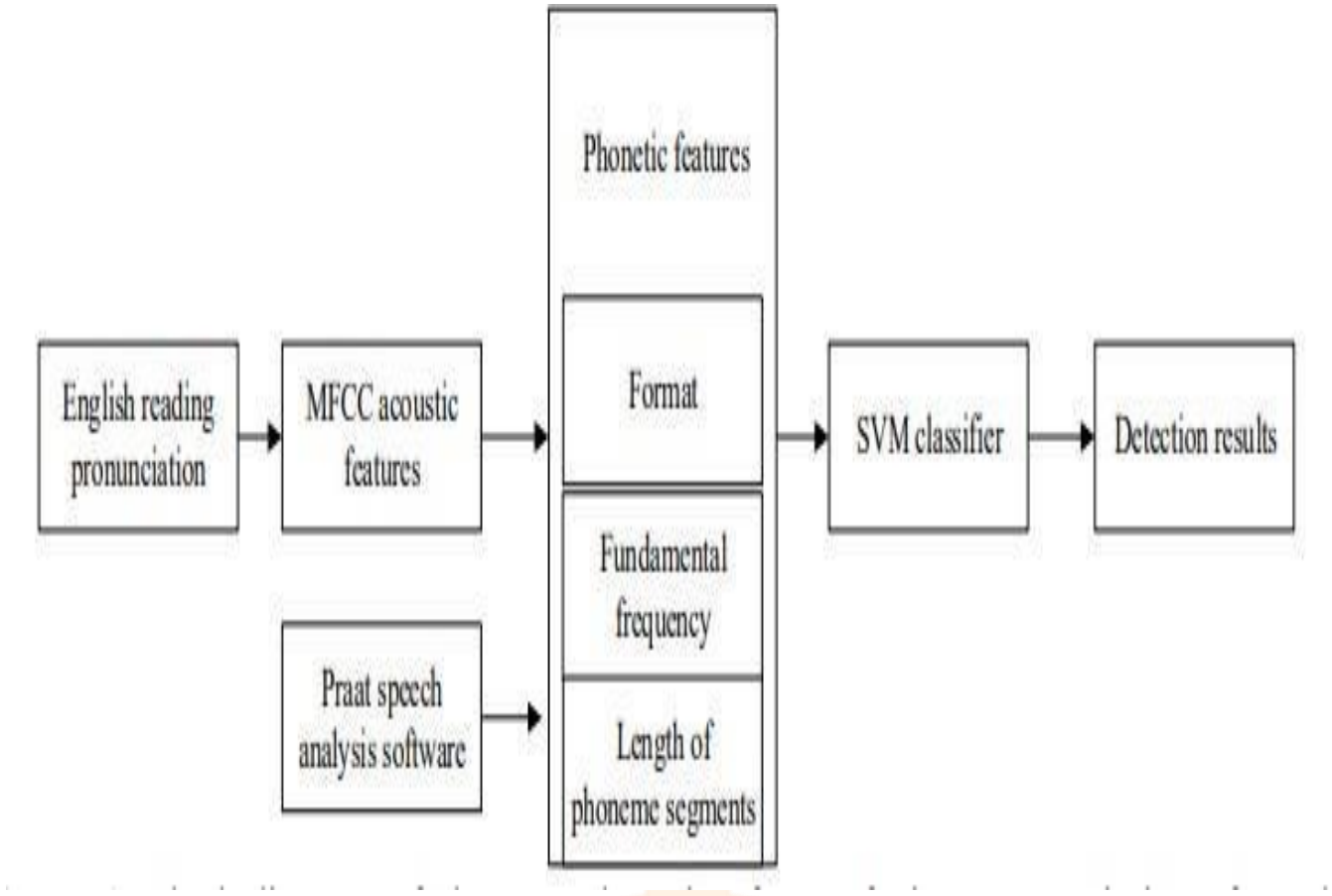


Figure 2 block diagram of mispronounciaton recognition system using cnn and rnn algorithm

7.METHODOLOGY

FEATURE EXTRACTION:

This law grain excerpts several audio features from an audio train using the librosa library in Python. After loading the audio train, it processes the audio signal to gain its Short- Time Fourier Transform(STFT) representation, which is also used to cipher colorful features similar as Mel- frequency Cepstral Portions(MFCCs), Root Mean Square Energy(RMSE), spectral flux, and zero- crossing rate(ZCR). These features prisoner important characteristics of the

audio signal related to its timbre, dynamics, and spectral content. The uprooted features are generally

employed in tasks similar as audio bracket, speech recognition, and sound analysis.

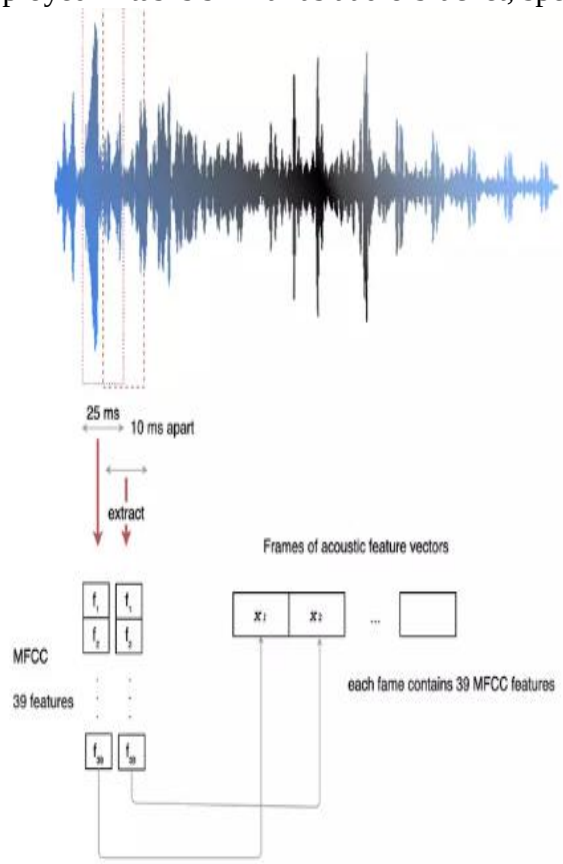


Figure 4 Mel frequency cepstral coefficient

DISPLAYING THE AUDIO:

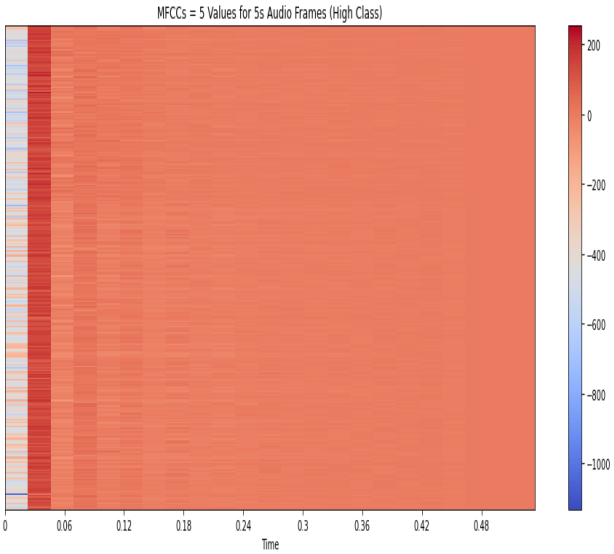


Figure 3 Audio frequency

VISULASISATION OF SPECTROGRAM:

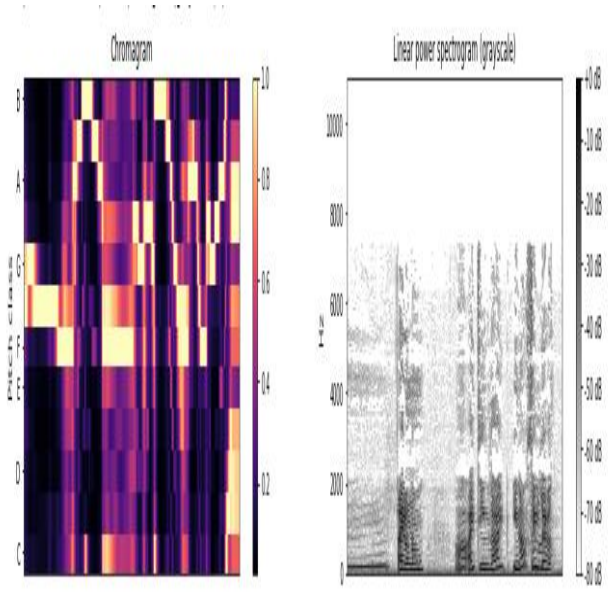


Figure 5 Sonographs

```
from pydub import AudioSegment
from flask import *
from flask_cors import CORS, cross_origin
import sqlite3 as sq
import speech_recognition as sr
from gtts import gTTS
```

LIBRARIES USED IN FRONT END

This Python script utilizes a combination of libraries for to construct a Beaker web grounded software tool featuring with audio processing capabilities. By importing the script can manipulate audio lines, performing tasks like format conversion and member slicing. The beaker library facilitates the creation of a web frame, while steins handles cross- origin resource sharing, pivotal for allowing web runners to interact with the operation. With sqlite3, the script can communicate with SQ databases, probably used for storing audio-affiliated data. speech, recognition enables speech- to- textbook functionality, allowing the operation to transcribe audio content. Together, these libraries form the backbone of a protean web operation able of processing, transcribing, and generating audio content, potentially feeding to tasks like automated recap services or voice- enabled relations.

REGISTER

```
def register():
    conn = sq.connect("register.db")
    conn.execute("create table if not exists info(id integer primary key

    r = request.json
    conn.execute("insert into info(name,password,email,phone) values(?,?,
                (r["name"], r["password"], r["email"], r["phone"]))
    conn.commit()
    conn.close()
    return 's'
@app.route('/login', methods=["POST"], strict_slashes=False)
```

The `register()` function is responsible for handling user registration within a Flaskweb application. It likely operates in response to a form

information such as name, email, and password throughthe request object. The function establishes a connection to an SQL database named"register. db",where user data is stored. It then inserts the provided user information into the "info" table using a SQL INSERT query, effectively creating a new user record in the database. After successfully registeringthe user, the function may return a response indicating the success of the registration process, such as a JSON object with a success message or a status code. This function serves as the backend logicfor adding new users to the application, ensuring their information is securely stored for future authentication and interaction with the system.

LOG IN

```
def login():
    r = request.json
    print(r)
    conn = sq.connect("register.db")
    n = conn.execute("select * from info where name=? and password=?",
                     (r["email"], r["password"])).fetchone()
    return json.dumps(n)
```

The `login()` function is designed to handle user authentication within a Flask web application. It begins by retrieving JSON data from the request object, likely containing user credentials such asemail and password. The function then establishes a connection to an SQLite database named "register.db", where user information is stored. Using a SQL query,it searches for a matching record in the "info" table based on the provided email and password. If a matching record is found, it returns the user information in JSON format; otherwise, it returns null. This function essentially serves as the backend logic for authenticating users during the login process, verifying their credentials against stored records in the database.

VOICE TEXT

```
def voicetotext(file):
    r = sr.Recognizer()
    with sr.AudioFile(file) as source:
        audio_text = r.listen(source)
        text = r.recognize_google(audio_text)
    return text
@app.route('/audio', methods=['POST'])
```

Speech recognition, generally refered to as voice to textbook conversion, is the act of rephrasing spoken wods into written language. It enables computers to understand and interpret mortal speech, converting it into a format that can be reused and anatomized programmatically. In practice, voice to textbook systems generally involve several stages, including landing audio input through a microphone, preprocessing the audio signal to enhance clarity, and using algorithms to fete and restate speech patterns into textbook.

8.PERFORMANCE EVALUTION

TRAINING AND VALIDATION

TRAINING ACCURACY

Training Delicacy is a measure of how well the model performs on the training dataset. It's calculated by comparing the model's prognostications on the training data to the factual markers and calculating the proportion of rightly classified samples.

$$\text{Training Accuracy} = \frac{\text{Number of Correctly Classified Training Examples}}{\text{Total Number of Training Examples}} \times 100\%$$

A high training delicacy indicates that the model has learned to fit the training data well and can directly prognosticate the markers of exemplifications it has seen during training. still, high training delicacy alone doesn't guarantee good performance on unseen data, as the model may have simply learned the training exemplifications without learning the underpinning patterns.

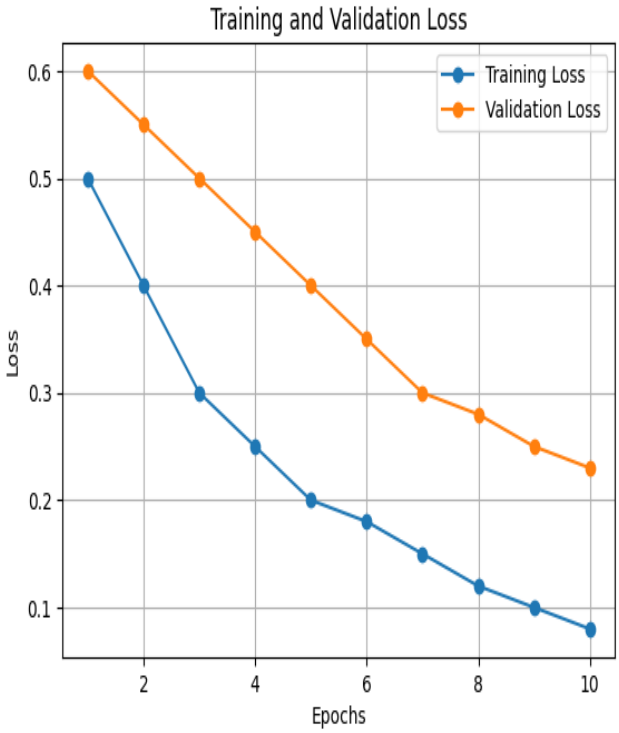
VALIDATION ACCURACY

Confirmation delicacy, on the other hand, is a measure of how well the model generalizes to new, unseen data. It's reckoned by assessing the model on a separate confirmation dataset that isn't used during training. The confirmation dataset serves as a deputy for real- world data and allows us to assess how well the model performs on exemplifications it has not encountered ahead.

$$\text{Validation Accuracy} = \frac{\text{Number of Correctly Classified Validation Examples}}{\text{Total Number of Validation Examples}} \times 100\%$$

A high confirmation delicacy indicates that the model has learned to generalize well from the training data to new cases, making it more likely to perform well in practice. Monitoring confirmation delicacy helps help overfitting, where the model learns to study the training data's noise and fails to generalize**.**

VALIDATION AND TRAINING LOSS:



TRAINING LOSS:

Training loss measures the error between the model's prognostications and the factual target values on the training data. It indicates how well the model is learning from the training data and conforming its parameters(weights) to minimize vaticination crimes. Lower training loss signifies better model performance on the training data**.**

$$L_{\text{train}} = \frac{1}{N}\sum_{i=1}^{N} \text{Loss}(\hat{y}_i, y_i)$$

Here, Loss(y^i,yi) represents the loss incurred for the *i*ith example, comparing the model's prediction y^i with the true target value *yi*.

VALIDATION LOSS

Confirmation loss measures the error between the model's prognostications and the factual target values on a separate confirmation dataset that the model has not seen during training. It serves as an estimate of how well the model generalizes to unseen data. Monitoring confirmation loss helps descry over befitting, where the model performs well on the training data but inadequately on new data.

9.CONCLUSION

In conclusion, the development of an operation employing the power of deep literacy to prognosticate pronunciation delicacy represents a significant step forward in addressing the challenges associated with studying original optima wavelength. By using deep literacy's remarkable capacity to decrypt intricate patterns and representations within data, the proposed operation seeks to offer an accurate also effective result the pronunciation chain. also, in the specific environment of studying original optima, deep literacy models showcase a distinct advantage over traditional styles. The operation of deep literacy in this sphere holds great pledge, offering further robust and environment- apprehensive results that can significantly contribute to language literacy and communication. Eventually, the integration of deep literacy into the study of pronunciation delicacy not only addresses being challenges but also opens new avenues for

invention in language- related technologies.

REFERENCES

1] F. Marty, “Reflections on the use of computers insecond language acquisition-II,” System, vol. 10, no.1, pp. 1–11, 1982.
2] M. C. Intelligent, “Computer assisted language learning as cognitive science: the choice of syntactic frameworks for language tutoring,” Journal of Artificial Intelligence in Education, vol. 5, no. 4, pp. 533–556, 1994.
3] Q. Chen, “Auto adapted English pronunciation evaluation: a fuzzy integral approach,” International Journal of Pattern Recognition and Artificial Intelligence, vol. 22, no. 1, pp. 153–168, 2008.
4] A. O. Husby, “Dealing withL1 background and L2 dialects in Norwegian CAPT,” in Proceedings of the Speech and LanguageTechnology in Education Venice, Italy, August 2011.
5] C. T. Ha, “Common pronunciation problems of Vietnamese learners of English,” Journal of Science,pp. 2135–2146, 2005.
6] B. Dong, “Automatic scoring of flat tongue and raised tougue in computer- assisted Mandarin learning,” in Proceedings of the ISCSLP. IEEE, Singapore, December 2006.
7] L. Neumeyer, “Automatic scoring of pronunciation quality,” Speech Communication, vol. 30, pp. 83–93, 2000.
8] S. M. Witt and S. J. Young, “Phone- level pronunciation scoring and assessment for interactive language learning,” Speech Communication, vol. 30, no. 2, pp. 95–108, 2000.
9] H. Franco, “Automatic detection of mispronunciation for language instruction,” in Proceedings of the Eur. Conf. Speech communication Greece, September 1997.
10] F. K. Soong, “Capturing L2 segmental mispronunciations with joint- sequence models in Computer Aided pronunciationtraining (CAPT),” in Proceedings of the ISCSLP, Taiwan, November 2010 .

[11] A. R. Mohamed, “Deep belief networks using discriminative features for phone recognition,” in Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Process, pp. 5060–5063, Prague, Czech Republic, May 2011.
[12] K. Li, “Mispronunciation detection and diagnosis in L2 English speech using m deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 1, pp. 193–207, 2017.
[13] J. Tao “Exploring deep learning architectures for automatically grading non-native spontaneous speech,” in Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6140–6144, Shanghai, China, March 2016.
[14] W. Li “Improving non-native mispronunciation detection and enriching diagnostic feedback with dnn-based speech attribute modeling,” in Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6135– 6139, Shanghai, China, March 2016.
[15] H. Strik, “Comparing different approaches for automatic pronunciation error detection,” Speech Communication, vol. 51, no. 10, pp. 845–852, 2009.
[16] R. Duan and T. Kawahara, “Multi-lingual and multi-task DNN learning for Articulatory error detection,” in Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 1–5 Korea, December 2016.
[17]N. Altman, “Markov models-training and evaluation of hidden Markov models,” Nature Methods, vol. 17, no. 2, pp. 121-122, 2020.
[18] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[19] M. Tuba, “Performance analysis of the fireworks algorithm versions,” Lecture Notes in Computer Science, vol. 12689, pp. 415–422, 2021.
[20] B.ilagavathi, “Evaluating the Ada Boost algorithm for biometric-based face recognition,” Data Engineering and Communication Technology, vol. 63, pp. 669–678, 2021.