WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Lyrebird: Voice-To-Text Note Making Automated Software with Speech Recognition

***Rohan Agrawal [1], Rohan Rajesh Purandhar[2], Siddharth Mehta[3], Yash Bhalla[4], Dr. C. Nandini[5]***

[1][2][3][4]BE Students, Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bangalore, Karnataka, India

[5] HOD, Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bangalore, Karnataka, India

A B S T R A C T

Our project LyreBird, is a Mobile Application of Voice-to-Text Automated Software with Voice Classification, which is an innovation which would nullify the need for paper and ink in the corporate world and otherwise. The aim of this paper is to incorporate Deep Learning techniques in the field of Natural Language Processing and Image Processing to produce the transcripts of a meeting. Apart from the transcripts, the finished application would have an in-built Voice Classification Software, capable of identifying and distinguishing between each participant, thus personalizing the transcripts with respect to each participant.

## INTRODUCTION

In our day to day lives, there is a need to transcribe the events that happen, for future needs with respect to legislation, Courts, Business meetings, etc.

The application incorporates the use of a Deep Learning Model. In order for the application to be able to identify speakers, it would require the data of each speaker's voice in the form of a 60 second (approximately) audio clip and the speaker's name. This data will be used to formulate the dataset, which in turn, will be used to train the deep learning model. This will be part of the first-time application setup and does not need to be repeated at the start of each meeting. The data of each speaker along with the trained model will be stored, so as to provide hassle-free meetings in the future.

Since the dataset will be saved, upon the joining of a new speaker, the entire dataset need not be formulated again. Only the data of the new speaker is required, which will be added to the existing database. However, this would require the model to be re-trained.

The transcript of the meeting would be generated at the very end as a reference for all the participants of the meeting. The transcript of the meeting would be generated at the very end as a reference for all the participants of the meeting.

The main purpose of the project is to reduce the amount of manual work of note making required in a meeting, thus enabling the participants to indulge more into the discussion, rather than focusing on creating notes for future reference. And what's more, it saves paper thus becoming environment friendly.

## 2 PRIOR AND RELATEDWORK

OrkenMamyrbayev. et al. performed a comparative analysis of five classification algorithms for identification and classification of people who speak Finnish, Kazakh, and Turkish.

A. Rahul. et al. performed comparative analysis using five classification algorithms, utilizing a dataset of over 3000 records to determine if the speaker was male or female. The main parameter used for classification was pitch and frequency.

Nishtha H. Tandelet. et al. created a fully automatic deep learning-based system about speaker recognition by speaker identification and speaker verification.

Gaurav Aggarwal. et al. zeroed in on extraction and order of the discourse highlights utilizing Mel-Frequency Cepstral Coefficient (MFCC) and Support Vector Machine (SVM) to recognize youngster and grown-up voice.

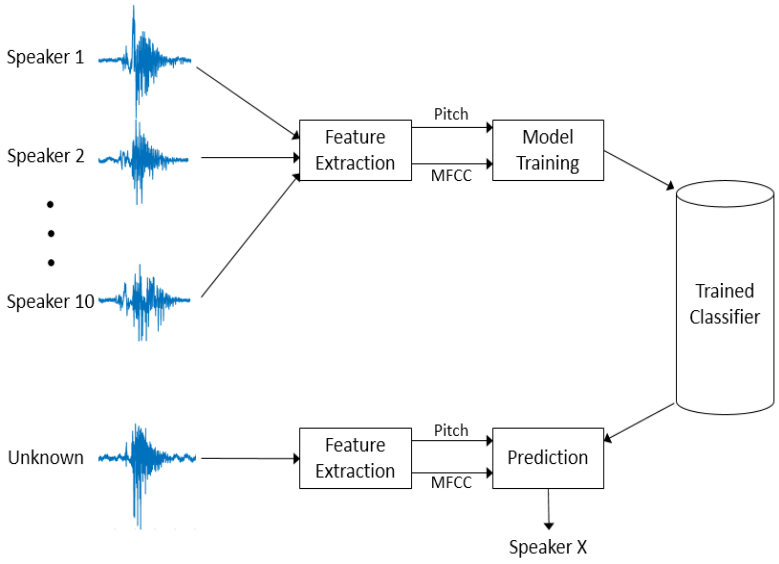
## 3 METHODOLOGY

**Feed Forward Neural Network (FNN)**

The data used to make the neural association contains 5435 named sounds from 10 remarkable classes. The classes are caution, street music, infiltrating, engine waiting, constrained air framework, vehicle horn, canine bark, drilling, gun shot and drill. Most classes are adjusted yet there are two that have low portrayal. Most speak to 11% of the information however one just speaks to 5% and one just 4%. The classes were lopsided as it was a decent test for the designer to fabricate a decent model with to some degree unequal classes.

We stacked the csv record that accompanied the preparation information into an information outline with all the names of the sound documents and its relating mark. We separated the highlights through a capacity that repeats through each column of the information outline getting to the record in the PC by perusing the document's way. We utilized Chromagram, Tonal Centroid Features (tonnetz), Mel-scaled Spectrogram,Mel-recurrence Cepstral Coefficients (MFCCs) and Spectral Contrast. We got a variety of 193 highlights with their separate name. We set them to be our X and y. We break them into preparing, approval and test information, watched that we saved similar extents for the classes as our all-out information and scaled the information. We picked 3435 sounds for our train information, 1000 for our approval information and 1000 for our test information.

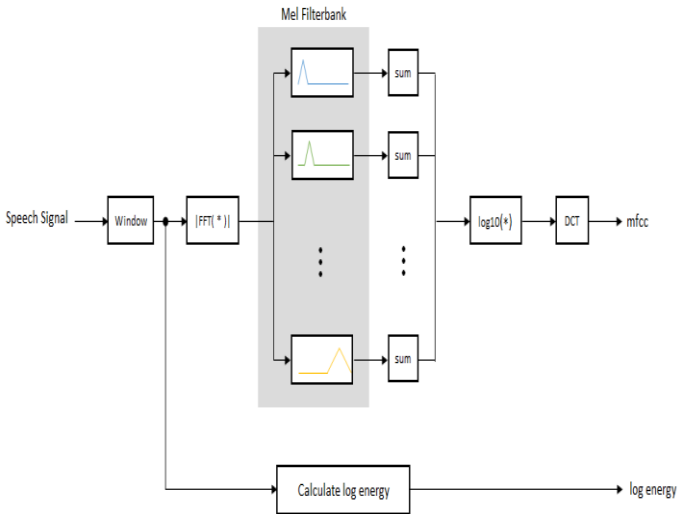
We assembled a feed forward neural organization with thick layers with two concealed layers utilizing relu and softmax for the 10 yields. We assembled the model utilizing the adam enhancer and downright cross-entropy for misfortune. We structure checked the best limits for the amount of neurons and the dropout degrees for our layers and thought about a reasonable model that foreseen our test at no other time seen (or heard) data with an exactness of 93%.



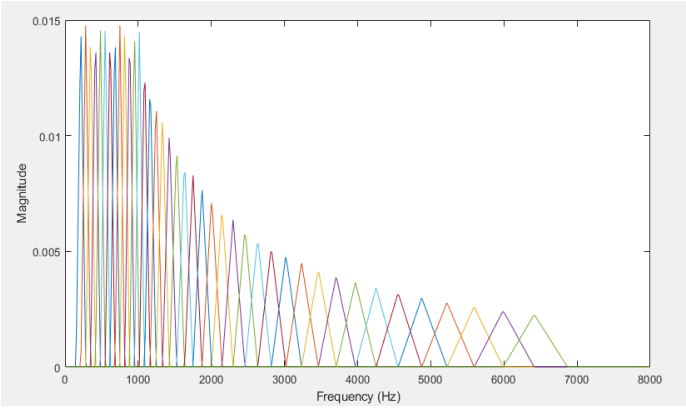

**Convolutional Neural Network (CNN)**

Utilizing similar information and similar strides as above, we produced our dataframe for our sound documents. Presently we expected to utilize a capacity to make pictures for each sound record. As in the past, our capacity iterated through each column of the dataframe and made a picture utilizing librosa and saved it to a neighborhood envelope.

Subsequent to making the pictures, we again split our information introduction preparing, approval and test (we utilized similar extents likewise with the neural organization from previously). We watched that we have similar equilibrium on classes and changed our dataframe record names from .wav to .jpg so we could utilize the equivalent dataframe to get to the pictures from our neighbourhood organizer.




We fabricated a convolutional neural organization with a Conv2D and MaxPooling2D information and five concealed layers: three Conv2D with their separate MaxPooling2D, at that point level and two Dense layers (all with relu). At last, we had the Dense yield layer for the 10 classes with softmax initiation. We once more, accumulated the model utilizing the adam analyzer and straight out crossentropy for misfortune. We didn't utilize gridsearchas it engaged for excessively long time (around two hours) hence we prepared to train it on 250 epochs. We acquired 92% precision on our testing data which was not never seen before the testing procedure.




## Voting Classifier

We inferred that considering we acquired two models that accomplish something very much like, we ought to use them together. We obtained the assumption probabilities for each class from our Neural Network and the assumption probabilities for our Convolutional Neural Network. After adding them together, we obtained maximum for each of the models. With everything taken into account, if our Neural Network was 65% sure the sound was that of a couple of young people playing and my Convolutional Neural Networkwas 95% sure it was fairly street music, by then street music would have a higher probability thusly our assumption would should be street music. We did this and thump our estimates to be 95% definite on at no other time seen test data. We found this to be a dazzling strategy to consolidate different models to have better figures.

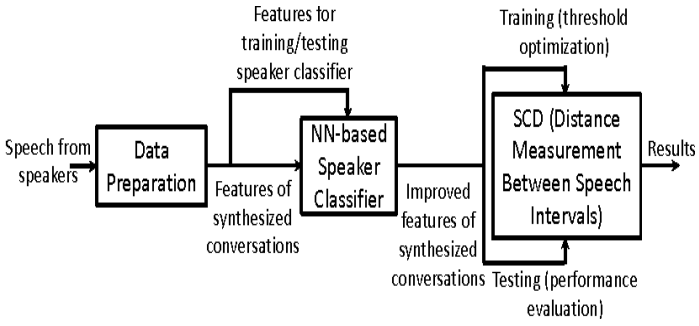
## Speaker Classifier

Presently we have the devices to handle our unique issue which was grouping speakers. First issue was to get acceptable sound information. After much examination, we run into a totally astounding information base for sound bites from book recordings chronicles at: http://www.openslr.org/12/. The downloaded dataset contains numerous gigabytes of clean information in ".flac" records that work incredible with macintoshes. We utilized a subset of the train-clean-100.tar.gz. The information comes very efficient in organizers by speakers with speaker ids, books, parts and record number. We additionally get a ".txt" document with data from the speakers telling us their name, sex and how long are their accounts. The subset we utilized ranges from 12 to 18 seconds and has over 13,000+speech cuts.

For our Urban Challenge issue, we attempted the use of feed forward neural organization as it is a lot quicker and it gave finerexactness. We obeyed similar strides as before, now managing significantly additional information (around 13k rather than 5k) and extended voice cuts (using normal voice cuts of 15 seconds rather than 5 seconds). Extricating the features took around 3 hours yet it should be done only a solitary time and we can save that group as a NumPy display in a close by coordinator and weight it at whatever point we need to use it or change anything.

We used 115 remarkable speakers both genders, male and female, where the base number of voice cuts per speaker was 56 and the most extraordinary was 166 (heedlessly picked). The standard deviation of the quantity of claps per speaker,as recorded, was around 16. We can trust them to be changed classes.

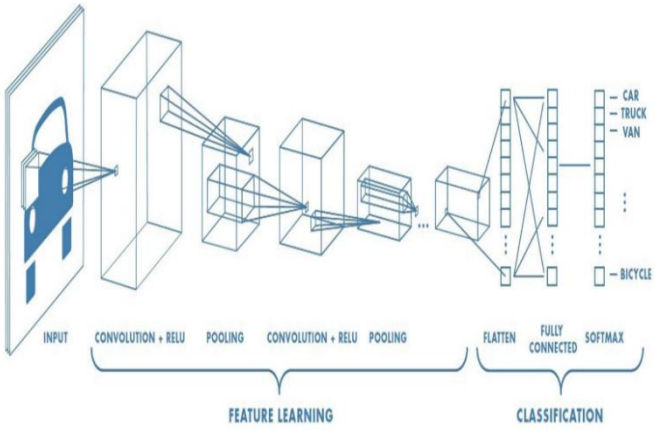
We fitted the information into a neural organization model with a similar arrangement of our gridsearched model from the Urban Challenge and got an astounding 99.8% exactness. We foreseeing on 1312 sound examples and arranged them into the 115 speakers and just got two sound examples wrong. The model just required 20 seconds to fit and it was practically great so we concluded it was not important to do the Convolutional Neural Network model.




## 4 EXISTING SYSTEM

We have executed this framework by characterizing 10 diverse metropolitan sounds. We utilized 2 unique ways to deal with arrive at our ultimate objective. The primary methodology was to remove mathematical highlights from the brief snippets utilizing the librosa library from python and utilizing those highlights to prepare a neural organization model and the subsequent methodology was to change over the sound bites to pictures and utilize those pictures to prepare a convolutional neural organization model.

## 5 PROPOSEDSYSTEM




We again utilized a similar Convolutional Neural Networkarrangement as in the past and fitted the model (which required a few hours). We got 97.7% precision on our test information. Recall that we got 99.8% exactness on our test information with our basic thick feed forward neural organization so this was a bit of baffling. We again, created forecasts on 100 new at no other time heard speakers and got 95% exactness.

We could consolidate these two models with a democratic classifier and improve exactness or train the models with more information to make them more precise however we have a period limitation and need to execute this model to have the option to utilize it for an intuitive exhibition and the Convolutional Neural Networkmodel's cycle takes excessively long. It requires some investment to make the pictures and furthermore to fit the model so we will utilize the NN with 97% exactness since it is quick and precise.

## 6 ACKNOWLEDGEMENT

Rohan Agrawal, Rohan Rajesh Purandhar, Siddharth Mehta, Yash Bhalla would like to thank Dr. C. Nandini, HOD, Department of CSE, DSATM for her constant encouragement, guidance and support throughout the course of the research paper, without whose efforts this would not have been possible.

## 7 CONCLUSION

We got great outcomes with the NN (93% on test information) and with the Convolutional Neural Network(92% on test information). We consolidated those two models together in a democratic classifier by joining the likelihood of the expectations and got a 95% precision when utilizing the Neural Network and Convolutional Neural Networktogether. We utilized a Neural Network model for foreseeing a grouping among 115 speakers and got 99.8% exactness. We didn't do the Convolutional Neural Networkfor this in light of the high exactness (practically awesome) of our Neural Network model. We utilized a Neural Network model for foreseeing sex and got 99.8% exactness while ordering the sex of speakers that the model had tuned in to previously. We got new information from speakers that the model had never heard and got a 95% exactness.

## SOURCES AND REFERNCES

[1] Gender differences in vowel duration in read Swedish: Ericsdotter C, Ericsson AM

[2] Temporal-Based Acoustic-Phonetic Patterns in Read Speech - Some Evidence forSpeaker Sex Differences: J International Phonetic Association

[3] Fact and fiction in the description of male and female pitch: Henton C G

[4] Voice Identification Using Classification Algorithms: OrkenMamyrbayev, NurbapaMekebayev, MussaTurdalyuly, NurzhamalOshanova,Tolga Ihsan Medeni, AigerimYessentay

[5] Voice Recognition and Voice Comparison using Machine Learning Techniques:Nishtha H. Tandel, Harshadkumar B. Prajapati, Vipul K. Dabhi

[6] Aida-zade K, Xocayev A, Rustamov S. Speech recognition using support vector machines. In: AICT'16. 10th IEEE International Conference on Application of Information and Communication Technologies; 2016