International Journal of Research Publication and Reviews

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

Implementation Lyrebird: Voice-To-Text Note Making Automated Software With Speech Recognition

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

[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

ABSTRACT

Our project LyreBird, is a Voice-to-Text Automated Software with Voice Classification, which is an innovation that 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 Speech Processing to produce the transcripts of a meeting. Apart from the transcripts, the finished application would also have an in-built Voice Classification Software, capable of identifying and distinguishing between each participant, thus personalizing the transcripts with respect to each participant.

Keywords: Voice Classification, Voice-to-Text, Speech Recognition, Deep Learning, AWS.

1 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 150 seconds (approximately) audio clip and the speaker's name.

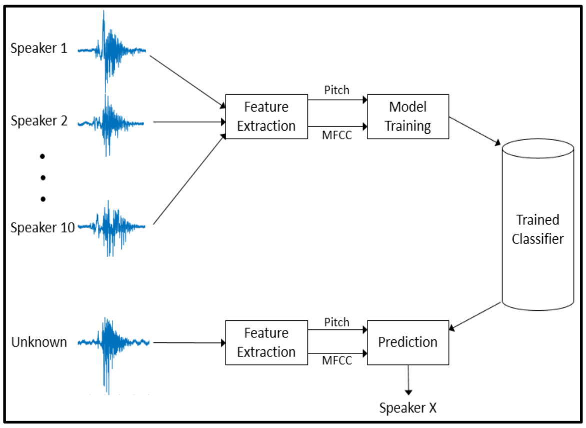


Fig. 1: Voice Classification Block Model/Flowchart

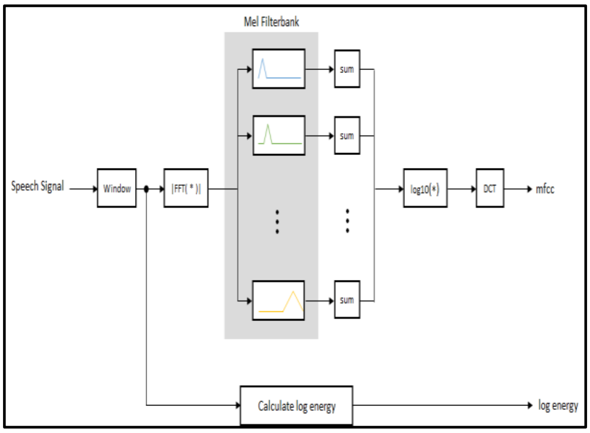


Fig. 2: Extraction of Features from Voice Clips

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 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 PROBLEM STATEMENT

A person usually attends a meeting daily, ranging from 30 to 45 minutes, where they have to give their inputs, make a note of what other representatives at the meeting are saying and also understand everyone. In this procedure, the note-making process is tedious and cumbersome because of various factors such as speed of writing, paying attention, making sense of notes after the lecture, and deciding what information to record in notes and hence in turn, reduces the overall efficiency of a person during the time period of the meeting.

We know that an integral part of any meeting be it court proceedings, business meetings, military meetings, police investigations, classrooms, air traffic control towers and even for content creation, is note-making. To automate the process of note-making during important meetings in this fast moving world, LyreBird has been created.

3 RELATED WORKS

The aim of our project is automated transcript generation in which 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 150 seconds (approximately) audio clip and the speaker's name.

At the end of each meeting, the model would generate the transcript of the meeting 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.

A. Voice Identification Using Classification Algorithms

The paper focused primarily on the identification and classification of people who speak Finnish, Kazakh, and Turkish. A comparative analysis of five classification algorithms were carried out to find the best one for personality identification by voice using machine learning methods.

In the first experiment, the support vector method was determined to show the best results. Two popular sets of features, often used in the analysis of the speech signal, were Mel Frequency Cepstral Coefficients (MFCC) and the Linear Prediction Cepstral Coefficients (LPCC). It's limitation is that the results produced were not comparable in performance to the models based on English. The reason for this is not only the difficulty of modeling the language, but also the lack of suitable resources for speech and text learning.

B. Voice Based Gender Classification Using Machine Learning

This paper performed a comparative analysis of the five classification algorithms, Linear Discriminant Analysis (LDA), K-Nearest Neighbour (KNN), Classification and Regression Trees (CART), Random Forest (RF), Support Vector Machine (SVM)). It used a vast dataset of 3169 records.

It contains 1584 male and 1985 female records. Each record has different acoustic properties like mean, frequency, standard deviation , median, etc. The main parameters for gender classification were pitch and frequency. It's limitation was that adults are capable of spontaneous vocal length adjustments which allows them to sound more masculine or feminine. So, it becomes difficult to classify the gender.

C. Voice Recognition and Voice Comparison using Machine Learning Techniques: A Survey

The paper focused on an elaborated literature survey on both traditional and deep learning-based methods of speaker recognition and voice comparison. It discusses speaker identification, speaker verification and finally describes the forensic approach of voice comparison which are, auditory, spectrographic, acoustic and automatic approach.

Deep learning based systems (CNN, Siamese NN) are fully automatic. CNN is good for classification problems while Siamese NN is good for comparison problems and even if the training data is less, Siamese NN can estimate well as compared to CNN.

D. Characterization between Child and Adult voice using Machine Learning Algorithm

The paper focused on extraction and classification of the speech features using Mel-Frequency Cepstral Coefficient (MFCC) and Support Vector Machine (SVM) to distinguish between child and adult voice. MFCC is used for feature extraction from voices and SVM is used to classify the datasets in a child's and an adult's speech. In the feature extraction phase, the parameters of sound waves are extracted and transformed from time domain to frequency domain.

The Mel-scale values are calculated for child and adult voices and then finally the Discrete Cosine Transform (DCT) of the list of Mel values is taken. The paper also conveyed that MFCC is not very robust in the presence of additive noise and normalises their values in speech recognition systems, thus introducing inaccuracy.

4. METHODOLOGY

The implementation of this system is done by splitting the system into various modules. The system mainly consists of six modules, i.e. Data Collection and Dataset Creation, Feature Extraction and Data Manipulation, Data Splitting and Statistical Methods, Model Definition, Training, Testing and Validation, Meeting Audio Processing and Transcript Creation.

A. Data Collection and Dataset Creation

We have used our own voices to create a dataset. The steps taken are: we collect the voice clips from the speakers who will be in the meeting. We will ask each attendee of the meeting to speak for at least 150 seconds. The clips are processed to remove the silences. This is achieved by splitting the clip whenever silence is encountered. Finally, we process the silence-free clips we received in the previous step and split them into clips of one second each.

Fig. 3: Data Collection and Dataset Creation

B. Feature Extraction and Data Manipulation

For feature extraction, we made use of “Librosa”, which is a python package for music and audio analysis. From Librosa, we made use of the functions that would extract around 5 features as follows:

- Chroma energy normalized statistics (CENS): It is best used for tasks such as audio matching and similarity.
- Short-time Fourier transform (STFT): It is a Fourier-related transform used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time.
- Tonal Centroids (or Tonnetz): It contains harmonic content of a given audio signal.
- Mel-frequency cepstral coefficients (MFCC): It is a scale that relates the perceived frequency of a tone to the actual measured frequency. It scales the frequency in order to match more closely what the human ear can hear.
- Spectral Contrast: It extracts the spectral peaks, valleys, and their differences in each sub-band, it represents the relative spectral characteristics. Spectral Contrast includes more spectral information than MFCC.

Once extracted, we need to convert it into an NumPy array, where each field contains another array of the respective features of that clip.

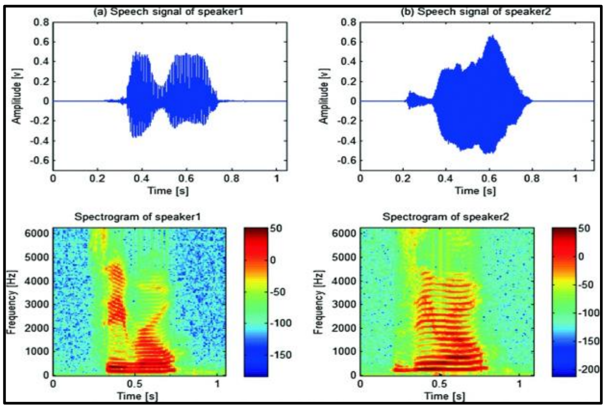


Fig. 4: Feature extraction and visualisation

C. Data Splitting and Statistical Methods

Once the dataset is created and features are extracted, we divide it into training data (70%), testing data (10%), and validation data(20%). Using ‘Label Encoder’, we one-hot encoded the data, so as to create a more matrix view of the data, as needed by the model later on. We then applied Standardization, to bring the mean value to 0, and the standard deviation to 1, which would bring the data in a proper range. The reason we applied Standardization and not normalization is because the data is not in a gaussian distribution.

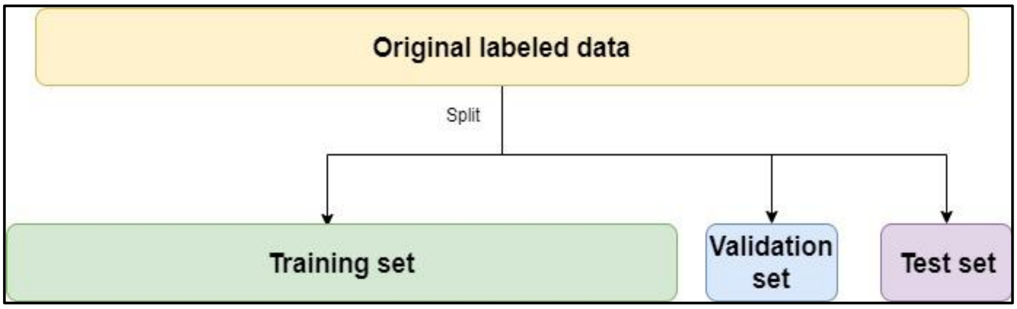
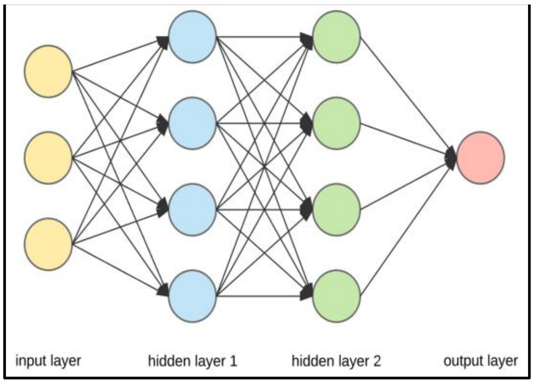


Fig. 5: Splitting data into 3 separate sets

D. Model Definition, Training, Testing and Validation

We selected a sequential Neural Network Model to fit this data to. This model contains an input layer, 2 hidden layers, and 1 output layer. At each stage, the activation function given was the relu function. We gave the Adam optimizer to the model. After training the model, we got an accuracy of 99%. After applying the model on the validation set, we got an accuracy of 97%. With more data utilised, that may increase as well.On applying the model to the test data, we observed 99% accuracy again. We were able to successfully identify which speaker ID belongs to which speaker.



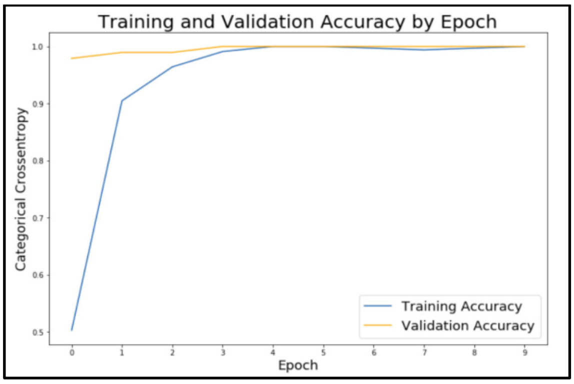


Fig. 6: Neural Network Model with 2 hidden layers

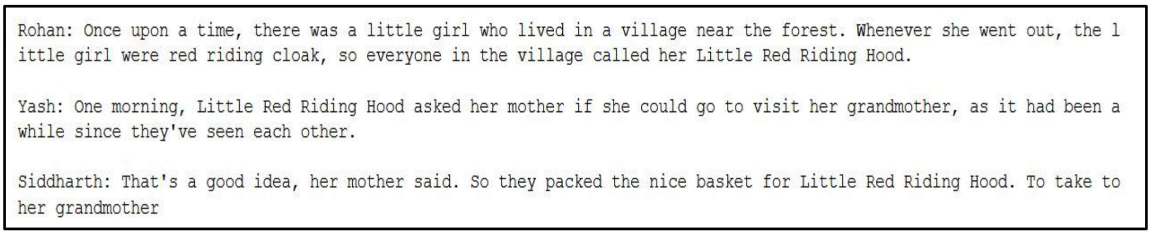
Fig. 7: Training vs Validation Accuracy

E. Meeting Audio Processing

We have implemented 2 libraries for processing of audio files, namely pydub and soundfile. The meeting audio is processed to remove the silences. This is achieved by splitting the clip whenever silence is encountered. Then feature extraction will be applied to the clips and all the features that we previously discussed above will be extracted. These extracted features will be used for speaker identification using the Voice Classification model we created earlier.

F. Transcript Creation

We have implemented 2 libraries for Speech-to-text conversion, namely boto3 and urllib. We upload the clips from our local file system to Amazon S3 bucket. For each file uploaded in the bucket, we use Amazon's Speech-To-Text service to get the punctuated text. The punctuated text and the speaker prediction results we received earlier are combined simultaneously to generate the final transcript. This transcript is then written to a file for future use.



```
Rohan: Once upon a time, there was a little girl who lived in a village near the forest. Whenever she went out, the little girl were red riding cloak, so everyone in the village called her Little Red Riding Hood.

Yash: One morning, Little Red Riding Hood asked her mother if she could go to visit her grandmother, as it had been a while since they've seen each other.

Siddharth: That's a good idea, her mother said. So they packed the nice basket for Little Red Riding Hood. To take to her grandmother
```

Fig. 8: Sample transcript generated

5. CONCLUSION

The project aims at putting away the need for stationary at meetings for the process of making notes or minutes of the meeting by automating the whole process through the developed application, hence saving time in the fast-moving corporate world.

This software is capable of generating transcripts of long duration meetings via Speech-to-Text conversion technique. The speakers are identified by their voice and the transcripts will be personalized by adding the speaker's name before their dialogues. This application will also play a huge role in decreasing the carbon footprint by reducing deforestation and global warming.

REFERENCES

[1] Gender Differences in Vowel Duration in Read Swedish: Ericsdotter C, Ericsson AM

[2] Temporal-Based Acoustic-Phonetic Patterns in Read Speech - Some Evidence for Speaker 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: Orken Mamyrbayev, Nurbapa Mekebayev, Mussa Turdalyuly, Nurzhamal Oshanova,Tolga Ihsan Medeni, Aigerim Yessentay

[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