An Efficient Machine Learning Framework for Speaker Authentication using Voice Input

Jagadish N, PranaySaha, Sunny Singh

Abstract: ***With the advancements in the hardware industry, an increase in the computation power and development in Artificial Methods, we can think of working on cognitive tasks. We have worked on various speech recognition methods using Natural language processing and Hidden Markov Models. We have done the classification of the users on the basis of their utterances. In this paper, we propose a discrete probability approach. The result which we have got gives us high accuracy results in recognizing the speakers. This helps in concluding that Learning and Uncertain reasoning are important components of Artificial Intelligence that could help in the development of solutions to problems which are interesting and complex.***

Index Terms**:** ***Machine learning, VUI, Cognitive technology, NLP..***

I. INTRODUCTION

The man has various criteria to recognize various sounds and songs. But now the challenge is how we will teach a computer to acquire enough knowledge to classify and categorize sounds and voices. Once this is accomplished, we can leverage a lot from this. We can use computers to identify the speaker and tell who he or she is. This is done on the various utterances and many other parameters. In order to develop this, two things are very essential:

- Mathematical model should be created which define the system.
- And, a large set of sample training sounds which will help the computer to learn and train itself.

A very important aspect of AI is that it does better with learning. A system can recognize sounds better when it gets to hear new sound by some other speaker. In fact, this helps us to understand that learning is very much essential for any AI system. The input can be in Microsoft Wave Format and given as input to train the whole system before using it. We also use Discrete Hidden Markov Models as they are quite useful in data modeling. We use CMU Sphinx as it works even in platforms with low resources. Some important basic terms which will help us are as follows:

1. Lattice: It is defined as a directed graph which is used to represents recognition variants.
2. N-best list: This is a list of variants which are just like lattices, but they are not as dense as lattice when it comes to representations.
3. Word confusion networks: These are the lattices where lattice edges become the source of the strict order of the nodes.
4. Speech database: This is a typical set of recordings which are from the task database. All the dialogs are recorded if we develop dialog system.
5. Text database: All the samples of the texts are collected. Usually, databases of texts are accumulated in sample form of the text. The issue associated with this is that all the documents are to be stored in spoken format.

As we went through various models, we decided that the Hidden Markov Model will be suitable for this. It says the current state is responsible for the transition probability and not on any previous history.

We use CMU Sphinx to approach this problem. It is a toolkit which is quite portable for building the HMM. It is mainly used in speech synthesis and speech research. Many open source projects use it too.

II. LITERATURE SURVEY

A. Existing System

[1] Demonstrates a framework which uses a phonetic division for upgrading the achievement rate of speaker recognizable proof over an all-inclusive timeframe dependent on Hidden Markov Model (HMM). It was seen that the section limit data got from HMMs gave a methods for normalizing the formant examples acquired from a computerized cochlear channel The advanced cochlear channel and HMMs was utilized for two well-presented issues in discourse acknowledgment for the most part, i.e., phonetic beat fluctuation and inconstancy over fleeting units of a given length, commonly days. The analysis showed in this paper analyzes the achievement rates of speaker distinguishing proof for 18 distinct speakers more than four days. During this trial, 18 unique speakers were recorded saying the word login multiple times every day. These articulations were then used to enlist the clients' voices for later recognizable proof. Three techniques were examined: conventional, normalized, and normalized and weighted during this investigation.

Revised Manuscript Received on May 20, 2019

Jagadish N. Assistant Professor, Department of Computer Science and Engineering, S J C Institute of Technology. Chickaballapur, Karnataka,

Sunny Singh Student, Department of Computer Science and Engineering, S J C Institute of Technology. Chickaballapur, Karnataka,

Pranay Saha, Student, Department of Computer Science and Engineering, SJ C Institute of Technology. Chickaballapur, Karnataka,

Conventional speaker ID utilizes just the Cochlear Filter Bank to process the discourse signal. Standardization utilizes the phonetic division gotten by methods for the HMM notwithstanding the Cochlear Filter Bank. Standardization with weighting utilizes the F-proportions of each phonetic fragment. An improvement of 14.5% in identification rate was recorded by Day 4 utilizing their technique compared to conventional method.

[2] aimed at determining how Speech identification (SID) precision decays with the expansion in the quantity of speakers. A lot of 963 female speakers was considered during the test. The trial utilized spoken digit strings for distinguishing proof. The scores were utilized as the proportion of comparability of voices. For the shut arrangement of 963 ladies, brilliant speaker recognizable proof exactness (97.8%) was acquired when just carbon microphone speech was utilized. At the point when carbon amplifier discourse was utilized for training and electret microphone for testing, the exactness dropped to 91.1% for a set of of 122 women and further dropped to 82.7% for a set of 308 women. The discourse was gathered over the long distance telephone network of 22 locales in the United States. From each site, there were 100 speakers (50 male and 50 female). The subjects were in the age gathering of 18 to 70. The associated digit speech samples consist of 66 spoken digit strings of lengths 2 to 7 collected in a solitary session. The digit strings are part to such an extent that half were spoken with a carbon receiver and the other half with an electret microphone. The digit strings of the speech database were separated into generally a balance of for preparing and testing. For every female speaker, two sets of speaker-dependent (SD) HMM's were worked from the training digit strings. One set utilized carbon microphone speech and the other set utilized electret speech. Each set contained 11 HMMs modelling the digits 0 to 9 and 'oh.'

[3] speaker recognition method new text-independent which combined a speaker-specific Gaussian Mixture Model (GMM) with syllable based HMM. This paper assessed the vigor of speaker recognition when the style of talking changes. This analysis utilized the NTT database. This database comprised of sentences information expressed at three distinctive speed modes: typical(normal), quick(fast) and moderate(slow) by 35 Japanese speakers (22 guys and 13 females) on five sessions more than ten months. 5 training expressions of around 20 seconds altogether was recorded by every speaker. An exactness of 98.8% was acquired during this procedure for the three talking style modes: typical (normal), quick (fast), moderate (slow). Short test expression of around 4 seconds was utilized. The error decrease rate of about half for each expression and 43% per 2 seconds, was observed accordingly. It was affirmed that the technique proposed was better than traditional text-independent speaker identification techniques.

III. SYSTEM DESIGN

Viterbi algorithm along with Hidden Markov Model becomes the mathematical approach which we are incorporating. The HMM is similar to Finite State Machines as it gives transition probability instead of discrete inputs. There is a transition probability matrix which gives the transition probability to any other state. Another characteristic of HMM is the output it produces. This is specially encoded in another matrix known as Observation Matrix. This matrix makes processing easy for a computer. We have one more matrix called Observation Matrix. It’s a 1xn matrix where n represents states in a model.

The second basic part is the Viterbi calculation. It keeps running on a lot of the succession. It can assess the ideal likelihood network and the perception lattice. It depends on Dynamic programming to decide the ideal Markov model. It is utilized in view of the effortlessness and simple execution on a computerized PC.

The development of this system is as follows:

1. Configuration files are set up. These regulate how the CMU Sphinx will process the code. (eg. How many samples of each speaker should be chosen?)
2. After this, we proceed to form the grammar for our voice system. Since the model is only a simple case of classification, we need to have only names in the grammar which we create.
3. Followed by extracting all the parameters of each wave file into one parameterized file and then form a codebook from a single file for further quantization. Codebook formation includes clustering of data that are approximate centroid in the vector space.
4. Quantization f the wave files so they that can be used by the Viterbi algorithm.
5. Training the HMM model on the basis of the settings.
6. The testing is done, and it is then compared against the HMM models.
7. Likelihood of each utterance of the speaker as per the training set.
8. The result is then made in a matrix formation with the classification of the speaker.

IV. ARCHITECHTURE

The above diagram represents an architecture of how the system works. There are several phases: Signal Capture and Preprocessing, Feature Extraction, Speaker Recognition Model Training and Verification (using GMM), Speech Recognition (using HMM training and verification) and Taking the final decision.

The first step of capturing the audio signal and pre-processing involves operation like converting the analog voice input to the digital data that can be stored and processed by the algorithms. The data is captured in the form of signals having specific wavelength, pulse, frequency and amplitude. The data gathered is then processed in order to extract useful data. Also, the noise cancellation is carried out during data pre-processing. After all these a specific window or frame of voice model is obtained.

The second phase which is the feature extraction phase involves the extraction of useful vectors (or features) from the data gathered in the previous step. This step generally involves converting the data into several vectors. The outcome is the vector with certain values and direction. These vectors are of individual voice frame or windows.

The next two phases of model training and speaker verification (or classification) includes the Gaussian Mixture Model. The GMM adopted here helps to train all the feature vector gathered for each individual frame. Apart from training we can also recognize the speaker or verification of new test models is done using GMM.

The next two phases of speech recognition is done using Hidden Markov Model. This algorithm uses finite state automata along with conditional probabilities. The model is used in order to train the data gathered for predicting the probability of the speech input given later. HMM is used along with Viterbi algorithm. Viterbi algorithm takes input the new test case and the FSM created using HMM and gives output the best classified word.

Based on speaker profile identified and the word identified the matching process takes place. The last phase occurs after the matching. A concrete decision is taken by the system whether the speaker is authenticated or not. If the speaker is unauthenticated then a suitable error message is displayed

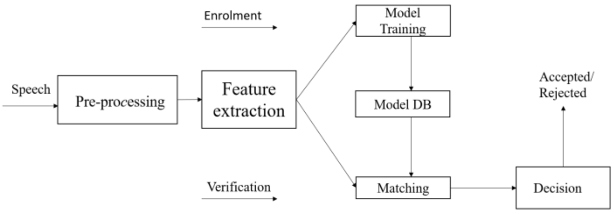
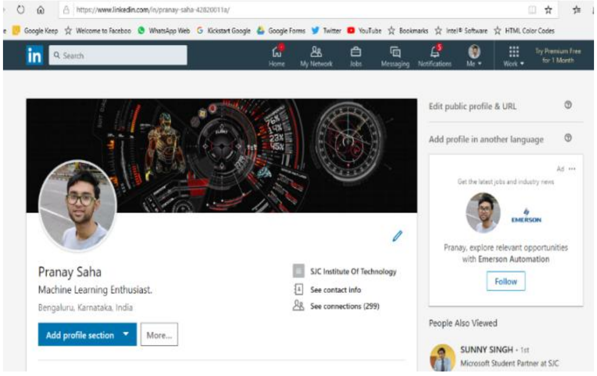


Fig1: Architectural Block diagram of Speaker/Speech Recognition System

V. RESULT

The method mentioned in the above section was implemented in a web-based framework. A UI was implemented for the experiment. During this experiment, voice profiles were created using the algorithms mentioned for two speakers. Each speaker was assigned a key word (or tag word) which could provide another layer security. There were three test cases: user 1 authenticated, user2 authenticated and unauthenticated user.


https://www.linkedin.com/in/pranay-saha-42820011a/

Pranay Saha
Machine Learning Enthusiast.
Bengaluru, Karnataka, India

SJC Institute Of Technology
See contact info
See connections (299)

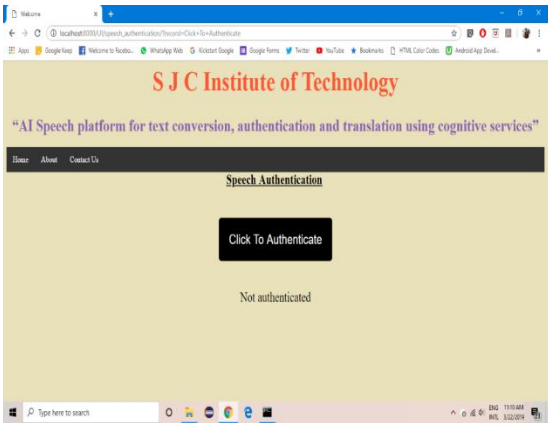
Get the latest jobs and industry news
Pranay, explore relevant opportunities with Emerson Automation

People Also Viewed
SUNNY SINGH · 1st
Microsoft Student Partner at SJC

Fig2: LinkedIn page of user 1

Figure 2 represents the user 1 being authenticated. Once the user 1 is authenticated we are opening a link (LinkedIn profile) of an individual user. In this test case the key word provided was “unlock me”, which was said by user 1 and the link was opened. The same methodology was adopted for the second user and suitable result was obtained.

Figure 3 represents an anonymous user trying to authenticate himself. Once the user tries to authenticate himself by saying any random word which doesn’t match to the keyword, the system takes input that specific word and then runs it through the algorithm and deny the authentication of the user and displays appropriate message to the user.


localhost:8000/UI/speech_authentication/?record=Click+To+Authenticate

S J C Institute of Technology

“AI Speech platform for text conversion, authentication and translation using cognitive services”

Speech Authentication

Not authenticated

Fig3: Not Authenticated User

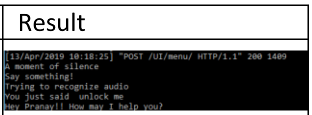
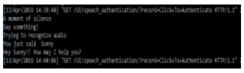
	Input Given	Result
1	Unlock me	 [13/Apr/2019 10:18:25] "POST /UI/menu/ HTTP/1.1" 200 1409 A moment of silence Say something! Trying to recognize audio You just said unlock me Hey Pranay!! How may I help you?
2	Sunny	 [13/Apr/2019 14:38:48] "GET /UI/speech_authentication/?record=Click+To+Authenticate HTTP/1.1" A moment of silence Say something! Trying to recognize audio You just said Sunny Hey Sunny!! How may I help you? [13/Apr/2019 14:40:06] "GET /UI/speech_authentication/?record=Click+To+Authenticate HTTP/1.1"
3	Something Else (Any random word)	 S J C Institute of Technology “AI Speech platform for text conversion, authentication and translation using cognitive services” Speech Authentication Click To Authenticate Not authenticated

Table 1: Overall Result Analysis

VI. CONCLUSION AND FUTURE ENHANCEMENT

This paper has helped us explore the domain of speaker authentication. We have got the opportunity to learn more about algorithms like Hidden Markov Model (HMM), Classifiers, Viterbi, etc. After analyzing the result this can be concluded that the system depicted in the paper is classifying the users and authenticating them successfully. Although the accuracy of this system can be further increased using more complex models. The work can obviously continue further to improve the result. We can also consider further classifications based on various parameters like amplitude, frequency, the timber of the sound. The idea is to use voice to authenticate the users and improve the human-machine interaction.

REFERENCES

1. Speaker identification experiments using HMMs Webb,J.J.; Rissanen, E.L.; Acoustics, Speech, and Signal Processing, 1993. ICASSP-93,1993 IEEE International Conference on, Volume: 2 ,27-30 April 1993 Page@): 387 -390 v01.2
2. Speaker identification using bidden Markov models, Inman, M.; Danforth, D.; Hangai, S.; Sato, K.; Signal Processing hceedmgs. 1998. ICSP'98.1998 Fo@ Intemational Conference on, 12-16 Oct. 1998 Page@): 609 -612 V0l.l
3. Speaker identification using Hidden Conditional Random Field-based speaker models, Wei-Tyng Hong, 2010 International Conference on Machine Learning and Cybernetics, Year: 2010 , Volume: 6, Pages: 2811 - 2816
4. H. C. Andrews, Mathematical Techniques in Pattern Recognition
5. E. Bunge, "AUROS—Automatic recognition of speakers by computers principles of the speaker recognition system", 9th International Congress on Acoustics, 1977.
6. S. Furui, "An analysis of long-term variation of feature parameters of speech and its application to talker recognition", Electr.and Comm. in Japan, vol. 57-A, pp. 34-42, 1974.
7. H. Ney, "Telephone-line speaker recognition using clipped autocorrelation analysis", Acoustics Speech and Signal Processing IEEE International Conference on ICASSP '81., vol. 6, pp. 188-192, 1981.
8. M. Sidorov, A. Schmitt, S. Zablotskiy, W. Minker, “Survey of automated speaker identification systems,” in Proc. 9th Int. Conf. Intell. Environ. , 236-239, 2013
9. Fuzzy-Clustering-Based Decision Tree Approach for Large Population Speaker Identification, Yakun Hu ; Dapeng Wu ; Antonio Nucci
10. SocialSense: A Collaborative Mobile Platform for Speaker and Mood Identification, Mohsin Y. AhmedSeanKenkeremathJohnStankovic
11. Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram, PawanK.Ajmera, DattatrayV.Jadhav,RaghunathS.Holambe

AUTHORS PROFILE

Jagadish N. Assistant Professor, Department of Computer Science and Engineering, S J C Institute of Technology.Chickaballapur, Karnataka,

Sunny Singh Student, Department of Computer Science and Engineering, S J C Institute of Technology. Chickaballapur, Karnataka,

PranaySaha, Student, Department of Computer Science and Engineering, S J C Institute of Technology.Chickaballapur, Karnataka,
India.