INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 4 - V7I4-1212)
Available online at: https://www.ijariit.com

Review paper on HMM Model for speech recognition systems

Abha Anand B.
abhaanandb.is17@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

Atul Ranjan Shipu
atulranjanshipu.is17@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

S. G. Raghavendra Prasad
raghavendrap@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

Sharadadevi Kaganurmath
sharadadeviks@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

Abstract- In a practical sense, speech reputation via machine has matured. Several speech-recognition algorithms are currently in use in a variety of programs, from a smartphone voice dialer to a voice response tool that costs stock values based on spoken input. One substantial development is that the software of statistical approaches, one of that is that the hidden Markov model (HMM). A speech-reputation process is generally divided into taxonomies primarily based totally on whether or not it needs to cope with particular or nonspecific talkers (speaker-based vs. speaker-impartial) and whether or not absolutely remoted utterances or fluent speech are acceptable (remoted phrase vs. related phrase). Present-day generation can also additionally effortlessly achieve near-ideal accuracy in speaker-impartial remoted-digit reputation, with the handiest 2-three percentage digit-string mistakes while the digit collection is uttered in a certain associated style with the aid of using the regular speaker.

Keywords— *Speech Recognition System, HMM, Speaker based, speaker impartial*

1. INTRODUCTION

Speech recognition could be a useful method for exchanging information via acoustic signals. As a result, it's no surprise that the voice signal has been the subject of study for decades. Speech recognition is a technology that uses a microphone to allow a computer to capture the words said by a person. Following that, the speech recognizer recognizes these words, and the system produces the recognized words. Speech recognition is the science of talking to a computer and having it recognize what you're saying. Speech recognition is the process of determining the meaning of a speech so that one can reply appropriately regardless of whether or not all of the words have been correctly recognized.

Database development and recognition operations are both involved in speech recognition. The process of creating a database entails acquiring speech samples from the speaker and extracting attributes for certain words and recognition could be a method of detecting the vocable by comparing current voice features to previously stored voice features. In real-time, the popularity algorithm compares the likelihood of an unknown vocable to a database of known words and then chooses the word with the highest likelihood.

The two types of speech recognition are text-dependent and text-independent voice recognition. Text-dependent voice recognition matches the vocable to the words he knew at the time the database was compiled. The text in the recognition phase is identical to the text in the training phase in this case. The vocable is identified regardless of the words in text-independent voice recognition. Speaker-dependent and speaker-independent speech recognition are two types of speech recognition. The speech of the speakers is recognized in speaker-dependent form as long as their speech samples are taken during training. Independent of the speakers, speaker-independent speech recognition recognizes the vocable.

2. OBJECTIVE

By reading research papers on the subject, this review aims to grasp the notion of Speech Recognition and its applications, as well as the issues that come with coping with the field's diversity and vastness. Also, suitable cost-effective efficient replacements to existing systems may be suggested.

3. LITERATURE SURVEY

The paper titled "A Systematic Review of Hidden Markov Models and Their Applications"[1]. Hidden mathematician models have applied mathematics models that have been used in a variety of real-world applications and communities. Mathematical models have become increasingly popular in recent decades, as evidenced by a large number of published works. The paper is a concise but thorough overview of analysis on the hidden mathematician model and its modifications for a variety of applications. The study of hidden mathematician

model variants and their applications reveals several interesting themes.

The fundamental theory of the Markov chain was glorious to mathematicians for roughly eighty years before HMMs were developed in the late 1960s. The study gives an overview of HMM variants and their potential applications.

The paper titles "Advances in subword-based HMM-DNN speech recognition across languages"[2]. They set off to construct and assess the utilization of subword units in reformist discourse acknowledgment, just as cutting edge neural organization based acoustic and etymological models, in their article. To achieve along these lines, they took a gander at word and subword structures in four unique dialects from different language families. The ideal size of the subword dictionary changed per language, going from several thousands in Finnish to a little more than a hundred thousand in English, yet for no situation did subword models outflank a comparative model utilizing word units. However this impact is now useful when utilizing n-gram language models, it is considerably more so when utilizing RNN-based language models, which have a superior capacity to catch lengthier settings.

With right RNN language models, exploitation of single characters as subword units to boot yields incredibly great outcomes despite the fact that it completely was at that point powerful to consolidate acoustic models from three very surprising structures, exploitation models with altogether entirely unexpected language displaying units were to boot prosperous, with mix frameworks diminishing the WER of the sole best framework by more than 100% family member. At last, the examination of more modest language demonstrating informational collections uncovered another strength of subwords, where debasement was fundamentally lower than that of word units. This backings the hypothesis that exploitation subword units reduce data shortage while expanding model heartiness.

The paper titled "Murmured Speech Recognition Using Hidden Markov Model"[3].
When criminals or militants are given a dose in order to obtain the truth, they mumble it. It's difficult for the human ear to comprehend. In the field of battle, when a commando wants to give confidential directions to his warriors in far-flung areas. This study provides a method for recording, transmitting, and converting non loud murmur speech to regular speech. The backside of a mumbled human's ear is connected to a NAM mike, which is becoming more linked to a wi-fi phone. This type of setup can even be employed to act decisively. As a consequence of the NAM mic consecutively connected with the Wi-Fi handset directly transmitting the signal to the conversion and identification system, the output voice is powerful against ambient sounds. To improve recognition accuracy, the speech recognition system in this study employs the murmured voices' lexicon. The configuration makes use of transceivers. The mumbled voice is collected and sent to the computer code built during this effort, despite the setup's maltreatment. The s/w formula transforms the input mumbled speech into a revised spectrum.

The paper is titled "End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition"[4]. Displaying the applied numerical connection between the acoustic discourse input and the HMM states that reflect lingually determined subword units like phonemes could be an indispensable advance in a secret numerical model-based programmed discourse acknowledgment framework. This exploration contemplates a start to finish acoustic demonstrating methodology using convolutional neural organizations, during which the CNN acknowledges a crude discourse signal as information and predicts the HMM states class contingent prospects at the yield. Through ASR studies and investigations on various dialects and numerous errands, partner inclination} to call attention to that the projected methodology yields reliably a further developed framework with less boundaries contrasted with the quality methodology of cepstral include extraction followed by ANN business, in differentiation to the quality system of the discourse strategy, inside the projected methodology the applicable component portrayals region unit learned by the primary technique the info crude discourse at the sub-segmental level.

The paper is titled "An Overview of End-to-End Automatic Speech Recognition"[5]. Programmed discourse acknowledgment, especially goliath jargon persistent discourse acknowledgment, is a vital issue inside the field of AI. For a significant time frame, the secret Andrei Markov model - Gaussian blended model has been the idea discourse acknowledgment structure. notwithstanding as of late, the HMM-profound neural organization model and furthermore the start to finish model abuse profound learning has accomplished execution on the far side HMM-GMM every misuse profound learning procedures, these 2 models have tantamount exhibitions.They took a gander at the occasion of the start to finish model in this examination. This examination initially talks about the essential ideas, advantages, and disadvantages of the HMM-based model and start to finish models, stressing that the start to finish model is the discourse acknowledgment advancement bearing. Programmed discourse acknowledgment might be an example acknowledgment task inside the field of applied science, that might be a subject of Symmetry.

4. UNDERSTANDING OF CONCEPTS

4.1 Speech Recognition System:

The voice signal is primarily responsible for conveying the words or message being delivered. The goal of speech recognition is to figure out what the underlying meaning of the utterance is. Speech recognition success hinges on the extraction and modelling of speech-dependent features that can effectively distinguish one word from another.

The speech recognition system may be viewed as working in four stages:
1.Feature Extraction
2. Pattern Training
3. Pattern Matching
4. Decision Logic

4.2 The Hidden Markov Model:

A hidden Markov model (HMM) is a statistical model that treats the represented system as a Markov chain. The task is to deduce the hidden parameters from the visible data in a process with unknown parameters. The state of a hidden Markov model is not explicitly visible during operation, but variables influenced by the state are. Each state includes a probability distribution for all possible output tokens. As a result, the token sequence generated by an HMM contains some state sequence information.

A hidden Markov model is a generalisation of a mixed model in which the hidden variables that impact the mixture component to be chosen for each observation are linked through a Markov process rather than being independent. By constructing stochastic models from known utterances, HMM compares the

likelihood that each model generates the unknown speech. Our feature vectors are reorganised into a Markov matrix (chains) that contains state transition probabilities using statistics theory. That is, if each of our code words represented a state, the HMM would keep track of state transitions and build a model that took into account the likelihood of each state moving on to the next.

HMMs are increasingly widely used since they can be automatically trained and are simple and computationally practical. For short periods of time, HMM considers the speech signal to be quasi-static and models these frames for recognition. It divides the signal's feature vector into several states and calculates the likelihood of a sign transitioning from one state to another. HMMs are basic networks that generate speech (cepstral vector sequences) by utilising a number of states for each model and modelling the short-term spectra associated with each state. Multivariate Gaussian distributions are generally heterogeneous. The model's parameters are the state transition probabilities, as well as the means, variances, and mixing weights that characterise the state output distributions. This (kind of) arranges our feature vectors into a Markov matrix (chains) that holds the probability of state transitions using statistics theory. That is, if each of our code words represented a state, the HMM would track the sequence of state changes and construct a model that included the probability of each state proceeding to the next. Mel Frequency Cepstral Coefficients (MFCC) are used in the feature extraction method, which extracts speech features for all of the speech samples. The pattern trainer is then given all of these features to train, and HMM uses them to generate an HMM model for each word. Then Viterbi decoding will be used to choose the one that has the highest possibility of being a recognised word.

4.3 Acoustic Model

The HMM-based speech recognizer employs the acoustic model to convert the spoken signal into a sequence of acoustic units, which is then translated into a phoneme sequence, and lastly, the desired text is generated by translating the phoneme sequence into text. The HMM-based speech recognizer employs the acoustic model to convert the spoken signal into a sequence of acoustic units, which is then translated into a phoneme sequence, and lastly, the desired text is generated by translating the phoneme sequence into text. If we choose a small unit similar to a phone, we will have an HMM for each potential phone in the language; however, the problem with this option is that the phone does not model its context. A model like this is known as a context-independent model. These models are often used in speech segmentation systems. Other acoustic units that consider context include the diphone, which represents the transition between two phones, the triphone, which represents the transition between three phones, subwords, and words. Context-dependent models are what they're termed.

4.4 Speech Recognition Process

The above-mentioned paradigm serves as the foundation for voice recognition technologies. Allow S to be the recognition voice signal. Recognizing entails determining the syntactic network's most likely path. The sequence of observation O is obtained by first transforming into a sequence of acoustic vectors using the same feature extraction method used for training. The model P (O|), which maximises the likelihood of witnessing O, is the most likely path. This probability can be calculated using the forward technique or the Viterbi algorithm.

4.5 Advantage of HMM Model

The mathematical framework and implementation structure of the HMM method is divided into two parts. In terms of the mathematical framework, the method's consistent statistical procedure and the manner in which it gives unambiguous solutions to linked problems may be identified. In terms of the implementation structure, we address the method's inherent flexibility in dealing with a variety of complex speech-recognition tasks, as well as its ease of implementation, which is a critical aspect in many actual engineering systems.

5. IMPLEMENTATION USED FOR REVIEW.

This execution isn't the best of discourse recognizer framework innovation to date; however it will give understanding into how HMMs can be utilized for discourse acknowledgment and different assignments. It will characterize what Hidden Markov Models are, tell the best way to carry out one structure: Gaussian Mixture Model HMM, GMM-HMM, and how to utilize this calculation for single speaker discourse acknowledgment. To exhibit this calculation, we have utilized distinctive datasets to discover the exactness of the model for various datasets. We utilized code as a source of perspective while making our own GMM-HMM model. This helped in testing the execution just as giving an edge of reference for execution. The records will be joined into a solitary information network (zero cushioning documents to uniform length), and a name vector with the right mark for every information record will be made.

When the information has been downloaded and turns into an information network, the resulting step is to remove highlights from the information, as is finished in numerous other AI pipelines. Most "purchaser grade" speaker ID frameworks utilize broad handling to extricate an assortment of attributes that characterize the sound across time and recurrence, and "custom highlights" were one of the keys to delivering a superb distinguishing proof framework as of not long ago. The present status of the workmanship has as of late moved to utilize profound neural organizations for include extraction, which is expected to appear in a future post. For now, keeping things to exceptionally fundamental highlights to give the most straightforward working model.

Maybe than the plenty of master includes regularly used in an advanced discourse acknowledgment pipeline, straightforward recurrence top location was utilized in this model (MFCCs, or all the more as of late, a pretrained multi-facet neural organization). This highlights an immediate impact on execution yet considers an all-encompassing execution that matches during a solitary post. The STFT is a basic segment of most DSP pipelines, and incredibly effective schedules for registering it is accessible (see FFTW, which NumPy wraps).

Following that, each FFT casing of every information document is exposed to top identification. All things being equal, we'll search for tops utilizing a moving window. Coming up next are the means of this calculation:

Make an information window of length X. For instance, X=9, however any window size is frequently utilized. This window ought to be partitioned into three segments: left, focus, and right. This might be LLLCCCRRR for the 9 example time span.Apply a capacity to each segment of the window (mean, middle, max, min, and so forth).Proceed to the following check if the most extreme worth of the capacity across the middle segment is more prominent than the outcome for left or right. GOTO next if not something else. In the event that the most extreme incentive for f(CCC) is inside the actual focal point of the window, you have discovered a pinnacle! Imprint it and proceed. In the event that this isn't the situation, continue to the following stage. Shift the information document by one example and rehash the technique.

Some distinguished pinnacles will be seen once the entirety of the information has been handled. Sort them by adequacy in plunging request, then, at that point produce the most elevated N tops.

The HMM is prepared utilizing the Baum-Welch calculation. There are various materials accessible on this calculation, which won't rehash here. Carrying out this HMM was genuinely interesting, and that we enthusiastically suggest utilizing a library.[12]

6. RESULTS

We used a GMM-HMM model to determine the accuracy rates for several single speaker word recognition tasks. We test the model using data and try to recognise speech after it has been trained. However, the model's accuracy rates with different datasets vary.

Different Datasets	**Accuracy Rates**
Dataset 1 - [‘apple’,’eye’,’book’,'dog', 'human', 'cat', 'fast', 'god']	50%
Dataset 2 - ['apple', 'banana', 'peach', 'lime', 'orange', 'kiwi', 'pineapple']	63.64%
Dataset 3 - ['five', 'seven', 'one', 'nine', 'six', 'three', 'zero', 'four', 'eight', 'two']	42%

7. CONCLUSION

While the hidden Markov model has made significant progress, it also provides a flexible but rigorous stochastic model with which to develop our systems. In addition, the framework includes training algorithms for estimating model parameters that are computationally efficient. It can be highly advantageous to incorporate belief functions theory into the voice recognition process. When there is noise present in the speech data, the model performs poorly, and accuracy and efficiency suffer when the speech type is varied. HMM can only be utilised when ideal conditions are met, otherwise, the efficiency is insufficient. Speech recognition deals with a wide range of datasets.

8. REFERENCES

[1] Bhavya Mor, Sunita Gharval, and Ajay Kumar, ‘A Systematic Review of Hidden Markov Models and Their Application’, Mar 2020.

[2] Peter Smit, Sami Virpioja, and Mikko Kurimo, ‘Advances in subword-based HMM-DNN Speech Recognition Across Languages’, Computer Speech & Language, Volume 66,2021,101158, ISSN 0885-2308.

[3] R. Kumar T., L. S. Videla, S. Sivakumar, A. G. Gupta, and D. Haritha, ‘Murmured Speech Recognition Using Hidden Markov Model’, 2020, 7th International Conference on Smart Structures and Systems (ICSSS), 2020, pp. 1-5, DOI: 10.1109/ICSSS49621.2020.9202163.

[4] Dimitri Palaz, Mathew Magimai-Doss, Ronan Collobert, ‘End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition, Speech Communication, Volume 108,2019, Pages 15-32, ISSN 0167-6393

[5] Wang, Dong; Wang, Xiaodong; Lv, Shaohe. ‘An Overview of End-to-End Automatic Speech Recognition’ Symmetry 11, no. 8: 1018, 2019. [6] Mankala, S.S.R., Bojja, S.R., Ramaiah, V.S.: Automatic speech processing using HTK for Telugu language. Int. J. Adv. Eng. Technol. 6(6), 2572–2578 (2014)

[7]Dua, M., Aggarwal, R.K., Kadyan, V., Dua, S.: Punjabi automatic speech recognition using HTK. Int. J. Comput. Sci. Issues (IJCSI) 9(4), 0814–1694 (2012)

[8] Kumar, K., Aggarwal, R.K., Jain, A.: A Hindi speech recognition system for connected words using HTK. Int. J. Comput. Syst. Eng. 1(1), 25–32 (2012) [9] O’Shaughnessy, D.: Acoustic analysis for automatic speech recognition. Proc. IEEE 101(5), 1038–1053 (2013)

[10]Zhaojuan Song, “English speech recognition based on deep learning with multiple features”, Springer-Verlag GmbH Austria, part of Springer Nature 2019

[11]Hassan Satori, Ouissam Zealouk,· Khalid Satori,· Fatima ElHaoussi, “Voice comparison between smokers and non-smokers using HMM speech recognition system”, Springer Science+Business Media, LLC 2017.

[12]https://colab.research.google.com/github/kastnerkyle/kastnerkyle.github.io/blob/master/posts/single-speaker-word-recognition-with-hidden-markov-models/single-speaker-word-recognition-with-hidden-markov-models.ipynb#scrollTo=FRpyQ0VyI4Tk,(accessed 4 July 2021)

