Proceedings of the Sixth International Conference on Electronics, Communication and Aerospace Technology (ICECA 2022)
IEEE Xplore Part Number: CFP22J88-ART; ISBN: 978-1-6654-8271-4

2022 6th International Conference on Electronics, Communication and Aerospace Technology (ICECA) | 978-1-6654-8271-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICECA55336.2022.10009602

IMPLEMENTATION OF SPEECH TO TEXT CONVERSION USING HIDDEN MARKOV MODEL

A.Elakkiya, K.Jaya Surya, Konduru Venkatesh, S.Aakash
Department of ECE, Saveetha Engineering College, Saveetha Nagar, Sriperumbadur Taluk, Chennai, 602105.
elakkiyaa@saveetha.ac.in

***Abstract* - Deep learning is revolutionary when used to transcribe spoken language into text that computers can read with the same intent as human readers. The fundamental idea is to give intelligent systems with human language as data that may be utilized in various domains. A speech-to-text synthesizer is a piece of software that can convert an audio file into text using Digital Signal Processing (DSP) algorithms that analyze and process the speech signal in the audio file. The objective of Speech To Text (STT) is to convert audio input from a user or computer into readable text. The STT is proposed to be transformed using the Hidden Markov Model (HMM) method. The development of a speech-to-text synthesizer will be a tremendous advantage for the visually handicapped and will make reading lengthy texts much easier.**

Keywords: Speech to Text (STT), HMM, DSP, Deep learning

I Introduction

For STT, the system listens to audio input from a person or machine and detects and converts individual words and phrases into text. This streamlines the communication between humans, machines, and machines. Its greatest utility is when speakers of diverse languages and dialects must collaborate or communicate. Without a STT conversion mechanism, it might be challenging for speakers of different languages and dialects to converse. A STT converter [1] might be beneficial in this circumstance, as it can transform the words spoken by a person with an unknown accent or dialect into standard written English. The aforementioned qualities were accomplished by employing a multitude of techniques [2]. After collecting pertinent features from the input audio, word and phrase matching based on acoustic word models and the established syntax and meaning of the sentences [3-4] are performed. Each stage of this technique may be carried out independently of the others.

The paper structure is as follows: The second chapter reviews algorithms for speech to text conversion models. Section 3 explains the proposed methodology and HMM. Section 4 analyses experimental outcomes. Section 5's model conclusions.

II Literature work

The author advises in [5] that for STT conversion, the audio message should be recorded and then converted to text, but for TTS conversion, the text should be translated to audio and the audio message should then be played to the user. The proposed speech-based email system employs three modules: STT conversion, TTS conversion, and interactive voice response (IVR). The proposed system [6] focuses on providing its clients with an intuitive interface. The system makes use of IVR (Interactive Voice Response) technology. A taped voice will teach the user on how to access specific features in this system. The article suggests the construction of a system that enables visually impaired, blind, and other users to use email with the same proficiency as a regular user. The system is essentially independent of the mouse and keyboard and supports STT and TTS operations. Face Recognition is also used to verify the identity of a person. The author suggests replacing regular MFCC with HMM in STT systems in [7]. HMM was offered as an alternative to the conventional MFCC approach for extracting features from speech signals, which proved ineffective. In contrast to the MFCC approach, the HMM network's input features increased the identification of input audio characteristics. HMM revealed a significant improvement in the quality of feature extraction from audio, resulting in a Speech-To-Text conversion system with enhanced computational speed and accuracy. The authors of [8] offered many speech representation and classification strategies. In addition to database evaluation and efficiency, they applied a number of feature extraction algorithms. They assessed the several problems involved with Automatic Speech Recognition

and proposed solutions. Speech recognition is discussed in terms of the AI Approach, pattern recognition, and acoustic phonetic techniques [9]. The authors [10] Using a variety of tactics, it is possible to raise the proposed STT conversion rate and create text of greater quality. The objective is to develop a continuous STT system with a much bigger vocabulary and speaker independence that can correctly differentiate the voices of various speakers. For the creation of such a system, considerable use of ANN and HMM will be made [11]. The authors [12] illustrate that when a system's performance and reliability are impaired, certain constraints are imposed. Additionally, the data shows that the bulk of STT work is completed in English. The evolution of Indian and other regional languages has received less attention than that of other languages. English has the highest rate of speech recognition compared to all other languages, according to the study. Due to their phonetic nature, Indian languages have a low identification rate.

III Proposed Methodology

After analyzing the performance of each STT approach, we determined that HMM is the most successful overall. Neural networks provide the maximum level of STT efficiency achievable. As a result, we have suggested a hybrid technique for STT conversion employing HMM and neural networks. As shown in Fig. 1, HMM and Neural Network are utilized to perform the model for STT conversion because they give the highest accuracy for STT.

Hidden Markov Model

HMM [13-14] are a subtype of dynamic Bayesian models due to their distinct structure, which allows them to be utilized in a number of scenarios. Due to the fact that the system being modeled is a Markov process with hidden states, it is referred to as a hidden-markov model (HMM). A hidden Markov model (HMM) may be considered in its simplest form as a probabilistic model having a state variable (S), an observation variable (O), and a transition (T) between the states (with a probability attached to T). HMMs are a sort of graphical model [15] that may be used to make predictions about hidden variables using just known inputs.

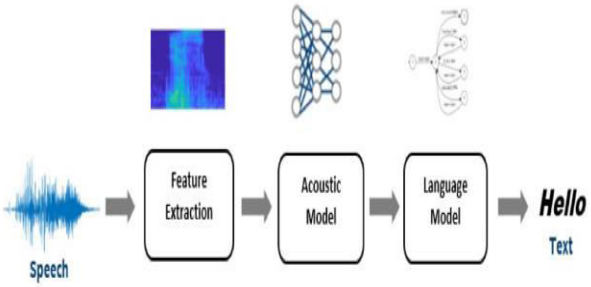

Fig 1 Proposed Speech to Text conversion model

Practice of forecasting the weather in a certain region based on the apparel of the people is an example of a straightforward use of HMM. The Markov assumption asserts, "Occurrence of the future is completely independent of the occurrence of the past and exclusively reliant on the present." This is the primary advantage of adopting the HMM. In other words, it is sufficient to know the current state to predict the future state without any training data. HMM is utilized extensively in speech and text recognition, robot localization, biological sequence analysis, handwriting recognition, pattern recognition, and other reinforcement learning applications. The methods necessary to convert STT, which is an HMM-based problem, are depicted in fig 2.

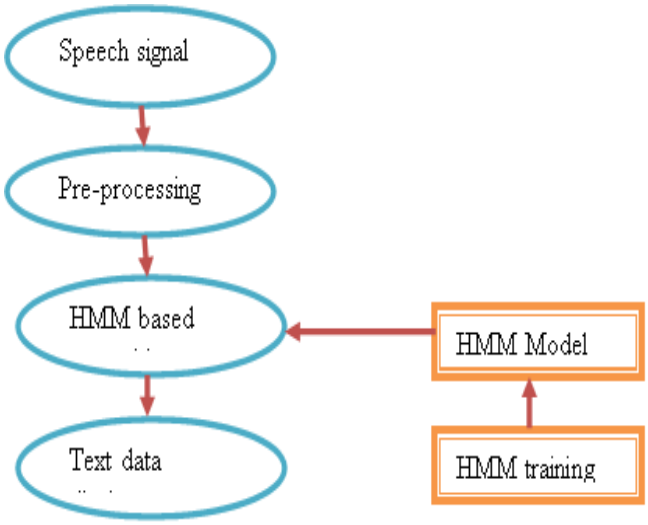

Fig 2. Steps for STT conversion using HMM

After being picked up by a microphone, the words are stored into the system's memory. The second phase, referred to as "speech preprocessing," involves cleaning up the speech signal so that it may be transformed. Using vocal activity detection, these gaps in the spoken input are eliminated. Third, during the training phase, the system creates a hidden Markov model (HMM) for each vocabulary term and trains these models. All the way from voice preprocessing through HMM model building, training is performed and the resulting HMM is loaded. Textual output is produced by saving the matched text and then reassembling it.

IV Results and Discussion

Automatic speech recognition (ASR) is the process of translating spoken words into written text using a machine or computer. ASR software recognises words and sub-words within the signal and analyses auditory characteristics of speech, including inspection and analysis of the physical properties of speech such as frequency, intensity, and duration. This method is advantageous for both speech recognition and text conversion. By painstakingly analysing the data, it is feasible to deduce behaviour from the uttered phrase. Deep learning architectures encompass a variety of variations on the same underlying concept, and it is simple to study a large amount of speech data. Each has accomplished noteworthy success in their respective disciplines. MFCC, DWT, GMM-HMM, and DNN-HMM were utilised throughout the speech recognition training, testing, and assessment stages.

Comparing DNN-HMM's accuracy to that of the other models in Table 1 is encouraging, as it reveals that DNN-HMM has achieved a high level of precision.

Table 1: Accuracy measurement of proposed model

Algorithms	Training Accuracy	Testing Accuracy	Validation Accuracy
MFCC [16]	80	81	83
DWT [17]	81.5	85.6	81.5
GMM-HMM [18]	86.2	91.2	93.1
Deep Neural Network - HMM	92.4	93.6	95.2

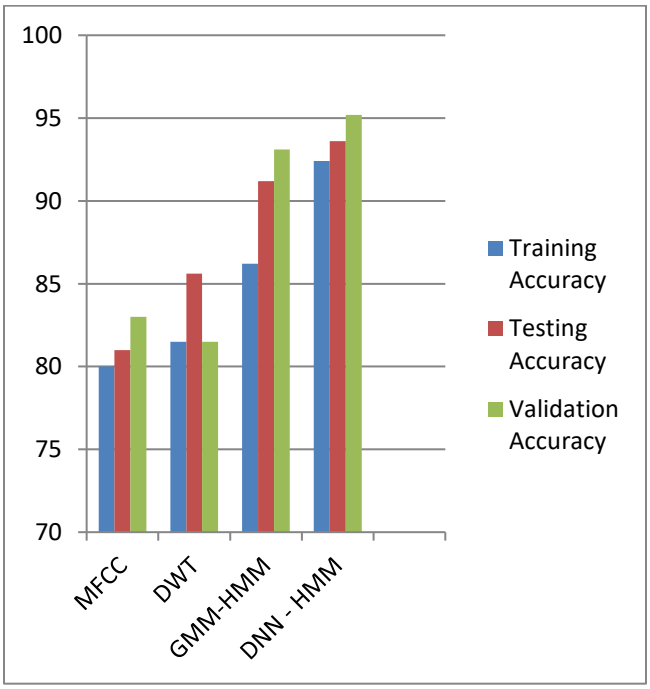

Fig 3 Performance comparison of different algorithms with DNN-HMM

Fig 3 shows that the DNN-HMM model gives highest accuracy than other models.

V CONCLUSION

Speech-to-text synthesis is a rapidly emerging area of computer science that plays an increasingly important role in the interfaces and systems we use every day. The Hidden Markov Model approach may be utilized to boost the efficacy of STT operations and to generate speech and text of greater quality. Using a Deep Neural Network, which can be developed in Python with the Speech Recognition API package from Google, is the optimal way for STT conversion. Consideration of punctuation throughout the speech-to-text translation process would be a significant improvement over the existing method. Our program connects with a speech-to-text engine optimized for American English. In order to increase the availability of speech-to-text technology in Nigeria, we plan to spend future funds in the development of locally optimized engines for the Nigerian language. Several indigenous languages, including Swahili, Konkani, the Vietnamese synthesis system, and Telugu, already employ a system similar to this. Additional research might focus on transferring speech-to-text systems to non-computer situations, such as telephones, ATMs, and computer games, or any other medium where such technology would be valuable.

VI REFERENCES

1. Ogunfunmi, T.; Ramachandran, R.P.; Togneri, R.; Zhao, Y.; Xia, X. A Primer on Deep Learning Architectures and applications in Speech Processing. Circuits Syst. Signal Process. 2019, 38, 3406–3432.

2. Passricha, V.; Aggarwal, R.K. PSO-based optimized CNN for Hindi ASR. Int. J. Speech Technol. 2019, 22, 1123–1133.

3. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. Pattern Recognit. 2019, 22, 354–377.

4. Dua, M.; Aggarwal, R.K.; Biswas, M. Optimizing Integrated Features for Hindi Automatic Speech Recognition System. J. Intell. Syst. 2019, 29, 959–976.

5. Palaz, D.; Doss, M.M.; Collobert, R. End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition. Speech Commun. 2019, 108, 15–32

6. Nagajyothi, D.; Siddaiah, P. Speech Recognition Using Convolutional Neural Networks. Int. J. Eng. Technol. 2018, 7, 133–137.

7. Poliyev, A.V.; Korsun, O.N. Speech Recognition Using Convolutional Neural Networks on Small Training Sets. Workshop on Materials and Engineering in Aeronautics. In IOP Conference Series: Materials Science and Engineering; IOP Publishing: Moscow, Russia, 2019; p. 714.

[8] Ingle, P., Kanade, H. and Lanke, A., 2016. Voice based e-mail System for Blinds. International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 3(01), pp.25-30.

[9] Jain, V., AK, K., Shenoy, R.N. and Ahmed, M., 2021. Voice Based Email for the Visually Impaired. International Journal of Engineering Research & Technology (IJERT) Vol. 11 Issue 07, July-2022

[10] Biruntha, S., Priya, M.G., Kiruthika, R., Indupriya, N. and Ashwini, R., 2021. Voice Based Email for Blind People Using Speech Recognition through Artificial Intelligence. International Journal of All Research Education and Scientific Methods (IJARESM), 9(04).

[11] Parkhi Bhardwaj, Gunjan Sethi Voice Based E-mail System for Visually Impaired: A Review published in International Research Journal of Engineering and Technology (IRJET) on December 12th, 2020.

[12] Pathan, N., Bhoyar, N., Lakra, U. and Lilhare, D., 2019. V-Mail (Voice Based E-Mail Application). International Research Journal of Engineering and Technology (IRJET), 6(03).

[13] Tharani K K, Shalini R, Jeyanthi I, Dr.Deepalakshmi R (2017), Voice Based Mail Attachment For Visually Challenged People, in International Journal of Scientific and Engineering Research, Vol.8, Issue.5, pp. 126-130.

[14] D Kiran kumar, User Interface for Visually Impaired People, IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) (Jan.-Feb. 2017), PP 65-71

[15] Shoba, G., Anusha, G., Jeevitha, V. and Shanmathi, R., 2014. An interactive email for visually impaired. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), (5089-5092).

[16] Bhowmick, A. and Hazarika, S.M., 2017. An insight into assistive technology for the visually impaired and blind people: state-of-the-art and future trends. Journal on Multimodal User Interfaces, 11(2), pp.149- 172.

[17] Tiwari, P.A., Zodawan, P., Nimkar, H.P., Rotke, T., Wanjari, P.G. and Samarth, U., 2020, February. A Review on Voice based E-Mail System for Blind. In 2020 International Conference on Inventive Computation Technologies (ICICT) (pp. 435-438). IEEE.

[18] Kulkarni, O., Alhat, A., Tejankar, N. and Patil, M., 2019. Voice based e- mail system for blind people. Open access international journal ofscience and engineering, 4(01).