Performance Evaluation of Different ASR Classifiers on Mobile Device

Gulbakshee J. Dharmale [1]
and
Dipti D. Patil [2]

[1] Computer Science and Engineering Department, PIET, Parul University, Vadodara, Gujarat, INDIA
[2] Information Technology Department, MKSSS's Cummins College of Engineering for Women, Pune, Maharashtra, INDIA *gul12dharmale@gmail.com

Automatic speech recognition is an option in contrast to composing on cell phones. Recently, it is usual and increasingly popular trend in communication. Classifier is used to classify the fragmented phonemes or words after the fragmentation of the speech signal. Several techniques are used for the classification of phoneme or word such as Neural Network, Support Vector Machine, Hidden Markov Model and Gaussian Mixture Model (GMM). This paper presents detailed study and performance analysis of above classification techniques. The performance evaluation is done to prove that GMM is better at the classification of signal data, and can be effectively used for improving the classification accuracy of the existing system. Our results show that accuracy of GMM is more than 20 % better than other three classifiers. The performance of ASR classifier is evaluated on android phones, and evaluated for normal conversations in Hindi language used in day to day human to machine communications, using high-quality recording equipment.

Keywords: Automatic Speech Recognition (ASR), Mel Frequency Cepstral Coefficient (MFCC),
Speech Enhancement, Hidden Markov Model (HMM), Gaussian Mixture Model (GMM).

1. INTRODUCTION

Speech Recognition is an innovation which grants machine to take out oral contained from a discourse signal and produce an instant message by utilizing highlight extraction and classification methods. The ASR innovation utilizes artificial knowledge and computational technique of signal handling. The Speech signal is created through various parts of the mouth with the assistance of changing pneumatic stress outside the mouth. At that point, these progressions can be tested intermittently and recorded in an advanced waveform. This recorded wave structure conveys all the data about the verbally expressed word. At that point, highlights of the discourse signal removed to accomplish discourse acknowledgment.

The precision of discourse acknowledgment relies on the possibility of a signal which will, in general, be traded off due to the environment wherein it is recorded or talked. Upgrading the quality and intelligibility of uproarious discourse is the challenge of implementing a robust system.

In recent times, globally increasing in trends of using a Smartphone as it is convenient for communication as well as data transmission. The main advantages of Mobile devices are Ubiquity, reachability, and Convenient compare to personal computers. As per today's requirement in improved applications that can employ automatic speech recognition on mobile devices.

1.1 Feature Extraction Part using MFCC

Feature extraction is used to derive expressive features from the improved and windowed speech signal to enable the classification of sounds. Mel Frequency Cepstral Coefficient (MFCC) is useful feature extraction techniques used in speech recognition system based on the frequency plot using the Mel scale.

MFCC is straightly appropriated inside the Mels or initial 1000 Hz and is then logarithmically conveyed above 1000Hz [Kamm et al. [1997]], B and M [2016]. This technique depends on the Mel scale. Essentially, the MFCC is registered by taking a straightly spaced frequency scale signal and afterward increasing it with a bunch of triangular bandpass channels. To more readily reflect the dynamic changes of the discourse, its first and once in a while second subsidiary of the information include vectors of the discourse signal [Dharmale and Patil [2019]] [Lähdesmäki and Shmulevich [2008]]. The recurrence part changing over to their Mel scale identical is given by the following equation;

$$Mel(K) = 1127\log(1+\frac{k}{700}) \tag{1}$$

where k is the frequency factor to be changed over to its comparable Mel esteem;

In this article, the introduction section presents ASR on mobile devices. The related work of speech recognition on mobile devices for multilingual is elaborated and explained, NN, HMM, SVM, and GMM classifiers. An idea regarding the execution of the above classifiers for the Hindi language gives in section 3. At last, this paper closes with the outcome and results of the execution assessment of NN, HMM, SVM, and GMM classifiers in section 4. The conclusion is stated in section five.

2. RELATED WORK

ASR for 10 digits trained for standard Arabian (SA). 98.62 %accuracy has been calculated using the connected digits corpus of the SA and 94.02 % using continuous SA corpus of speech [Walha et al. [2012]].

Kumar and Singh worked on Punjabi speech recognition having mainly based application creation on desktop computers [Dua et al. [2012]] [Kumar and Singh [2017]].

The output of audio games was analyzed by Luvcic Built for pupils with visual Impairments to consider the effects of Applications based on speech. These applications are very beneficial for pupils to understand the environment and to learn how to do daily tasks. They observed that the applications should be built with various levels of impairment and age groups.

The initial success of introducing visually impaired pupils to the ASR and TTS technologies resulted in the inspiration for further research in this area [Lučić et al. [2015]].

Isolated ASR was developed for Kannada by Thalengala and Shama. Two types of dictionaries were constructed: (a) Phone level and (b) syllable level. The Kanada news database is used to create dictionaries for pronunciation. For monophone and triphone acoustic versions, they obtained general word recognition accuracy of 60.2 % and 74.35 % respectively. They concluded that by selecting a suitable acoustic model based on the vocabulary size, the performance of an ASR system can be improved [Thalengala and Shama [2016]] [Schmitt et al. [2008]].

Schmitt emphasized the constraints and limitations ASR applications are confronted with, under different architectures [Nasereddin and Omari [2017]].

The morphological approach to the development of a comprehensive and widely representative Northern Sotho pronunciation dictionary for ASR was defined by Nkosi et al.Using the, which provides word pronunciation, they built the dictionary. If the pronunciation is not correct, the dictionary that has been developed specifies the correct pronunciation. Using HTK, they developed an acoustic model with word accuracy rate of 63.9 % [Ruan et al. [2016]].

Ruan et al. compared text entry and speech based dictation on mobile phones. They discovered that dictation based on speech is 3 times faster than text entry [Beulen et al. [1997]].

State tying for context dependent phoneme models was proposed by Beulen et. al.They suggested a state tying decision tree based on the VERBMOBIL corpus. They concluded that the gain is smaller than on the WSJ task due to the state tying is not as big as the contextual dependence of telephones in the German language as in the English language. This highlights that context based models cannot be suitable for every language. In order to generate an optimal

No.	ASR Technique	Features of ASR classification techniques
1	Neural Network (NN)	Suitable for design acknowledgment Simply conform to the powerful and new environmental factors More Training of information is required Self-learning and self-putting together [Essa et al. [2008]]
2	Hidden Markov Model (HMM)	Supports enormous vocabulary size Training complex Continuous and confined word acknowledgment. [Garg et al. [2000]] [Kurzekar et al. [2014]]
3	Support Vector Machine (SVM)	Ability to manage the powerful and high dimensional information vector. [Aida-zade et al. [2016]][Shanthi Therese and Lingam [2013]] Computational cost increments with gain in a few classes Needs fixed length input
4	Gaussian Mixture Model (GMM)	Training is composite Supports enormous vocabulary size. [Nainan and Kulkarni [2016]] [Deshmukh and Alasadi [2018]] Independent speaker and nonstop word acknowledgment

Table I: Explanation of various ASR classifiers

acoustic phonetic model, each language has distinctive features that need to be carefully modeled [Kaur and Mittal [2017]].

2.1 ASR Classifiers

Speech recognition is one of the main application areas of advanced signal training. Speech recognition frameworks have been isolated into two levels, the main degree of the ASR framework is the component extraction level utilizing Linear Predictive Cepstral Coding (LPC) and MFCC. The characterization method is the subsequent level utilizing HMM, GMM, NN, and SVM. Classifiers are utilized to order the separate fragmented words or phonemes after an efficient fragmentation of the speech signal. The most prevalent methods used for this classification are explained in detail in table1.

3. IMPLEMENTATION AND RESULT

In this section of the paper, the illustration of the analysis based on the n classification techniques studied is done. We implemented a backpropagation NN, HMM, SVM, and GMM-based calculation utilizing the android speech engine. The speech engine enables the input sounds to be processed in real-time and operates by splitting the input speech signal into segments, where each segment is a user-spoken word. Each of the words is then given to an extraction unit of the MFCC, which extracts the sound samples. These features are then contrasted with the Hindi Word network of words of the standard IIT Bombay. A complete list of Hindi words used as a database for this method is found in the IIT Hindi Word system. The process of matching the feature is performed with the help of the appropriate classification technique, which is implemented in its standard form. The progression of the proposed work is as appeared in the given figure 1.

The presentation of ASR classifier is assessed on Android telephone, utilizing excellent recording equipment, while the outcomes for the upgrade of existing frameworks is done on a regular human to machine correspondence, and assessed for ordinary discussions in the Hindi language. Receiver operating characteristics (ROC) curve is plotted to clarify the demonstration of ASR classifiers. It is utilized in speech recognition planning between false positive rate and true positive rate

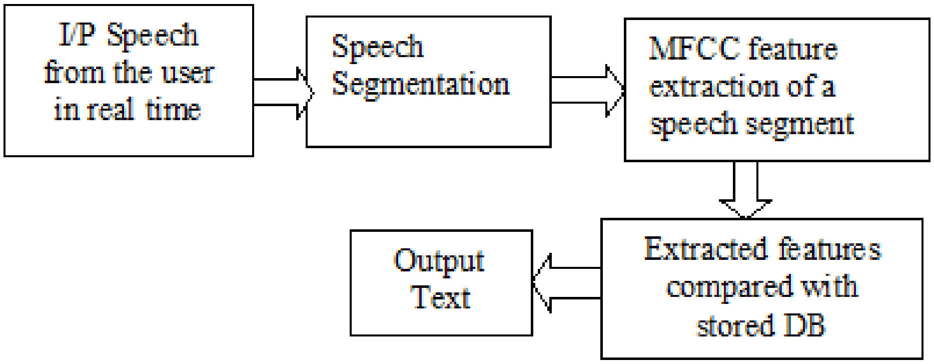

I/P Speech from the user in real time
Speech Segmentation
MFCC feature extraction of a speech segment
Extracted features compared with stored DB
Output Text

Figure 1: Process for performance evaluation of ASR classification techniques.

of ASR classifiers [Fawcett [2006]]. The ROC curve of four ASR classifiers, for example, NN, HMM, SVM and GMM depicts by ROC curve as appeared in figure 2. ROC curve is plotted between false positive rate and true positive rate. There are four probable parameters are true positive methods effectively perceived words (TP), Incorrectly recognized words are labelled as true negative (TN), a word which isn't expressed, however recognized is marked as a false positive (FP) and expressed words yet not recognized are considered false negative (FN). These four cases are utilized to compute tp rate, fp rate, precision, affectability, and particularity. Various focuses in ROC space show positive and negative acknowledgment. One upper left point (0, 0.9) speaks to consummate speech recognition, which demonstrates GMM performs in a way that is better than other three classifiers.

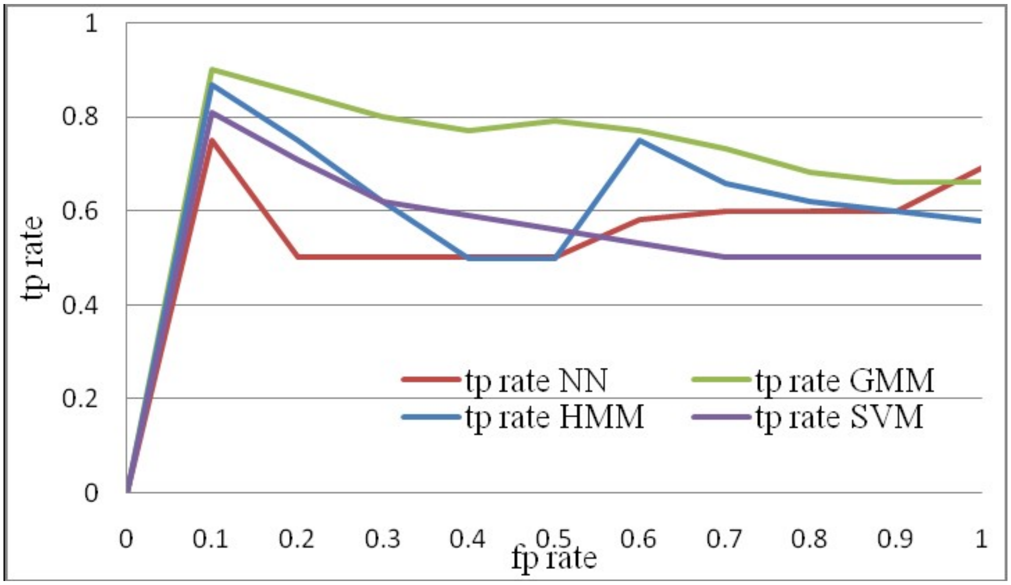


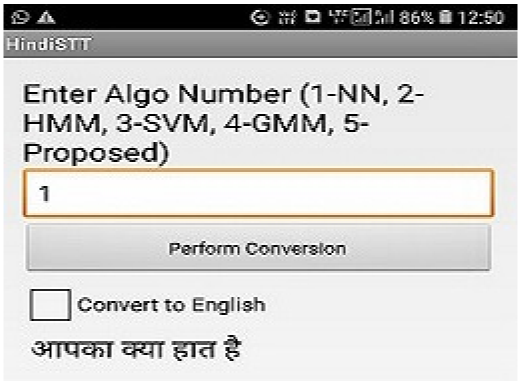
Figure 2: ROC curve plotted between four ASR classifiers

For short and long sentences used in day-to-day communication, the accuracy of each classifier is assessed and delays in speech recognition are measured. The results of each algorithms can be shown from the following table 2.

Sr. No.	ASR classification techniques	Recognition delay (ms)	Accuracy (%)
1	SVM	0.023	71
2	NN	0.024	64
3	GMM	0.020	83
4	HMM	0.022	80

Table II: Delay and accuracy of speech recognition of various ASR classification techniques

The outputs of the system implemented on the android device can be seen from figure 3 and figure 4. as follows;



HindiSTT

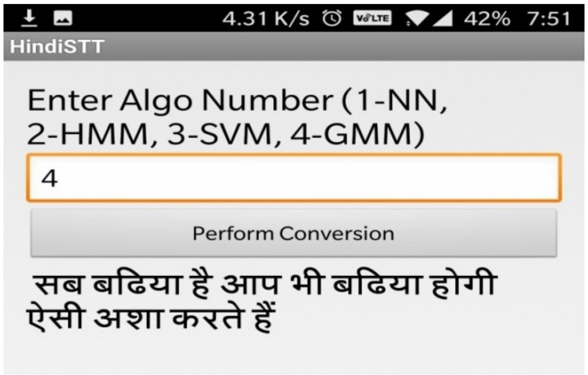
Enter Algo Number (1-NN, 2-HMM, 3-SVM, 4-GMM, 5-Proposed)

1

□ Convert to English

आपका क्या हात है

Figure 3: Short text conversion with the NN algorithm



HindiSTT

Enter Algo Number (1-NN, 2-HMM, 3-SVM, 4-GMM)

4

सब बढिया है आप भी बढिया होगी ऐसी अशा करते हैं

Figure 4: Long text conversion with the GMM algorithm

Our next comparison is between the recognition rate of HMM and SVM models for the English language. The main objective of the work is to establish an improved HMM-SVM strategy for the recognition of English speech. The task is to extract the different statistical characteristics based on the HMM method and use the SVM approach to conduct the classification of the speech signal.

The database of speech of words was created using Matlab, and each of the speech samples was taken to be 1 second long with a sampling rate of 8 kHz. The trained real-time samples were tested in different environments with minimal noise to perform the true evaluation of the classifiers either HMM or SVM. During the training phase, the system asks users to enter speech, then the original speech signal waveform and waveform after feature extraction appears. After this, the system inquires for entering text to train and create the database by adding the number of words. In the testing phase, one needs to enter words to test along with it and select a classification technique either HMM or SVM. If a selected classifier is HMM, then HMM directly recognizes the spoken word, while with SVM, the output shows all classes in a database with related count and provides recognized word as output with the highest count.

The following formula is used to calculate accuracy;

$$A = \frac{Correctlyrecognisedword}{Totalnumberofword} * 100\% \tag{2}$$

Extraction and choice for best parametric portrayal of acoustic signs play a very significant role in planning the discourse acknowledgment system on mobile device. For acquiring the results, the speech signal is recorded. Performance of HMM and SVM classification techniques for the English language is evaluated by the training and testing the word “Hello”. The speech signal, which is recorded for the word “Hello”, is shown in figure 5, and signal waveform after features extraction of the word “Hello” is given in figure 6.

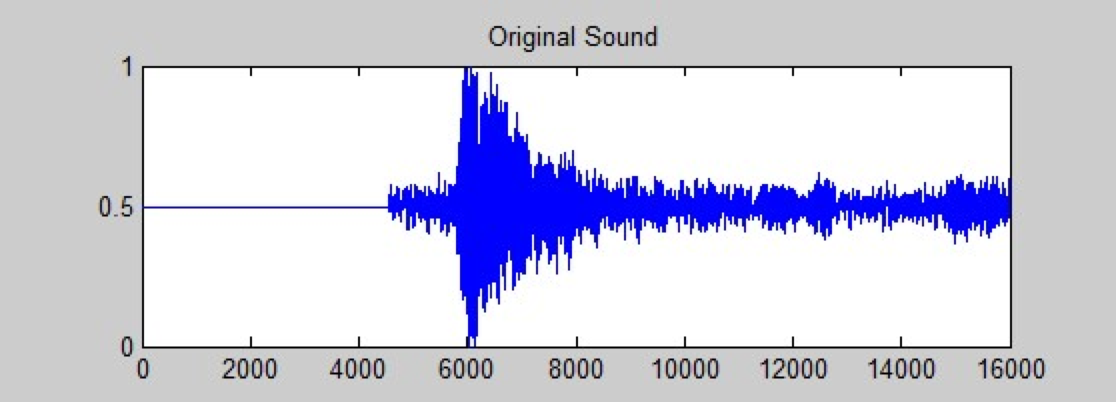


Figure 5: Original signals for word "Hello"

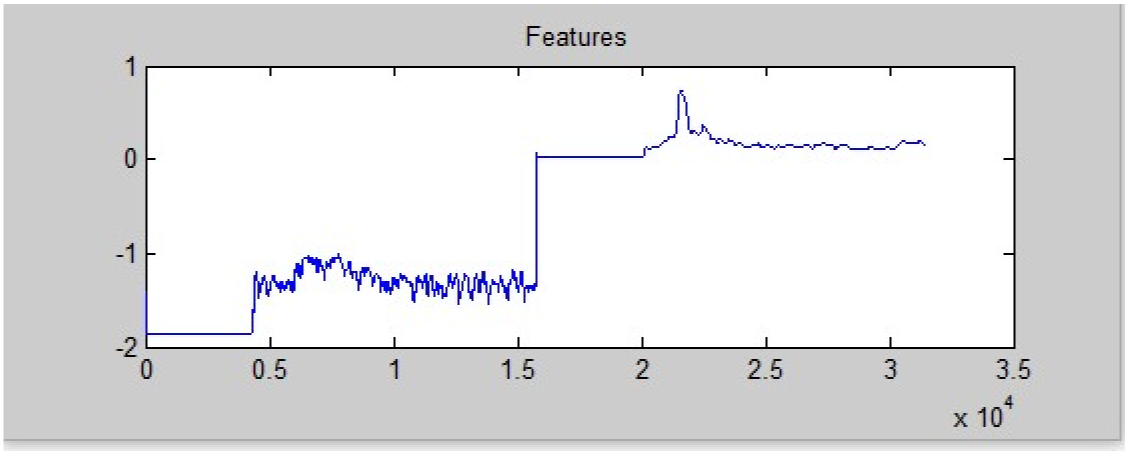


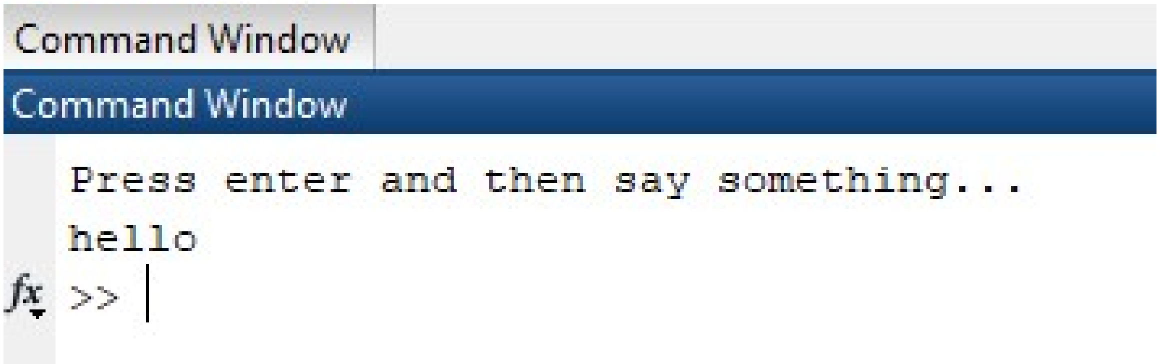
Figure 6: Speech signal for "Hello" word with features

Number of train words	Correctly tested words by SVM	Correctly tested words by HMM	Accuracy improvement (SVM over HMM) %
10	7	9	20
20	16	18	20
50	38	46	16
75	61	69	11
100	77	86	9

Table III: Speech recognition accuracy between SVM HMM

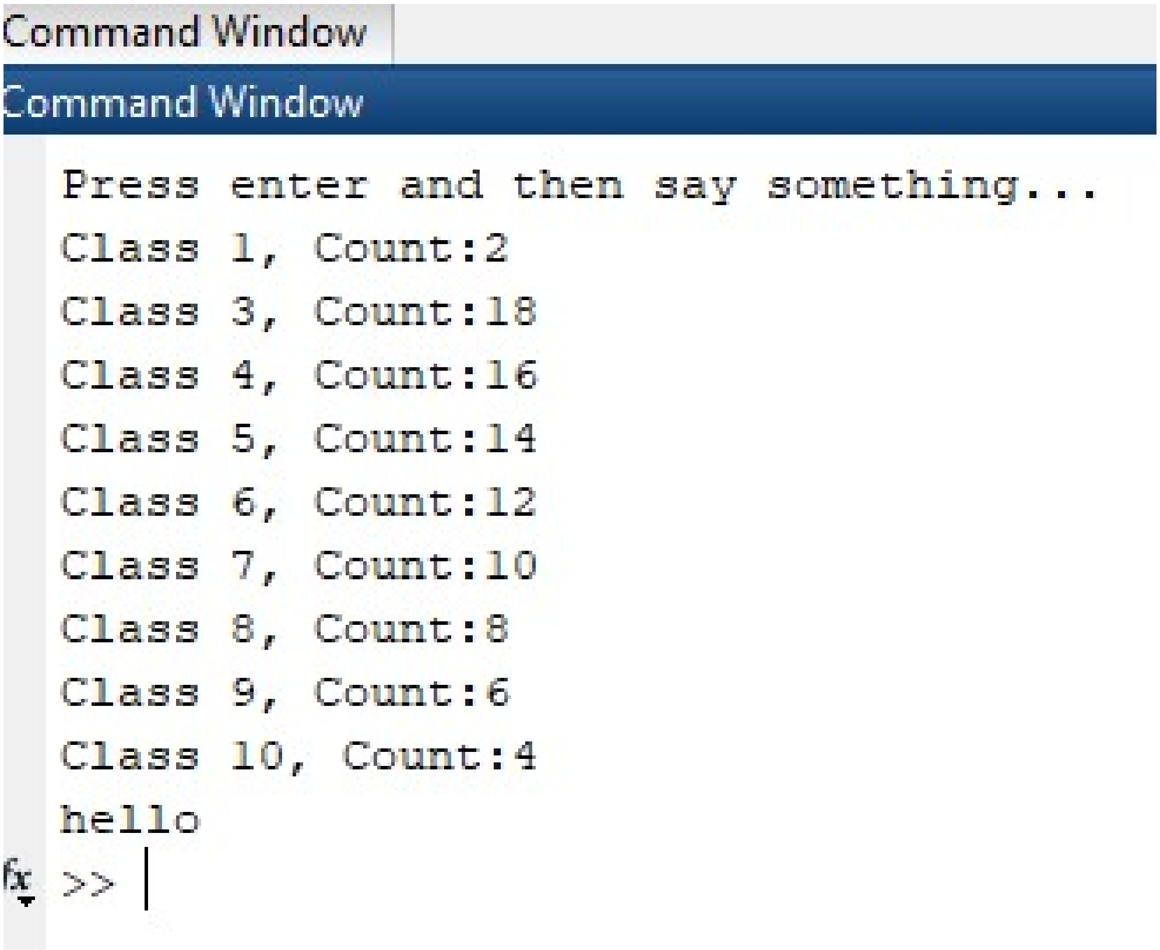
The system is trained for multiple words such as 'Hello, 'How are you', 'Good Morning' etc. Accuracy of HMM and SVM classification techniques are described in table 3. Accuracy of speech recognition is calculated by training 100 words and computing word fault rate at the time of testing. Result of the testing shows that HMM outperform the SVM as shown in table 3.

In case of HMM, the speech input is converted to text and after matching with the created database, the corrected output text is shown directly. Output with HMM is shown in figure 7. Testing of the word "Hello" with SVM classifier is shown in figure 8 in which matching of converted text with the created database is displayed on the output window, as correct text is with class having highest count.



```
Press enter and then say something...
hello
>>
```

Figure 7: Data testing with HMM classifier



```
Press enter and then say something...
Class 1, Count:2
Class 3, Count:18
Class 4, Count:16
Class 5, Count:14
Class 6, Count:12
Class 7, Count:10
Class 8, Count:8
Class 9, Count:6
Class 10, Count:4
hello
>>
```

Figure 8: Dataset testing with SVM classifier

4. CONCLUSION

From our study, we observe that Mel recurrence segments alongside HMM-based strategies give preferable order exactness over different techniques. NN is also a good method for classification but has limited accuracy due to its linear classification nature. Feature vectors of MFCC provide a good description of the signal under test and can be used for real-time implementations for speech recognition and translation into text. We plan to implement and improve the accuracy of HMM combined with MFCC using intelligent feature selection and hybrid classification techniques. Progressively executed, GMM beats different calculations and gives high classification precision. The accuracy of GMM is the best because it is evaluated under real-time outdoor and indoor noisy conditions.

References

Aida-zade, K., Xocayev, A., and Rustamov, S. 2016. Speech recognition using support vector machines. In *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*. IEEE, 1–4.

B, R. J. and M, J. S. 2016. A survey on speaker recognition with various feature extraction and classification techniques. *Int. Res. J. Eng. Technol.(IRJET) 3,* 4, 709–712.

Beulen, K., Bransch, E., and Ney, H. 1997. State tying for context dependent phoneme models. In *Fifth European Conference on Speech Communication and Technology.*

Deshmukh, R. and Alasadi, A. 2018. Automatic speech recognition techniques: A review.

Dharmale, G. J. and Patil, D. D. 2019. Evaluation of phonetic system for speech recognition on smartphone. *International Journal of Innovative Technology and Exploring Engineering (IJITEE) 8,* 10.

Dua, M., Aggarwal, R., Kadyan, V., and Dua, S. 2012. Punjabi speech to text system for connected words.

Essa, E., Tolba, A., and Elmougy, S. 2008. Combined classifier based arabic speech recognition. *INFOS2008, March,* 27–29.

Fawcett, T. 2006. An introduction to roc analysis. *Pattern recognition letters 27,* 8, 861–874.

Garg, A., Pavlovic, V., and Rehg, J. M. 2000. Audio-visual speaker detection using dynamic bayesian networks. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. IEEE, 384–390.

Kamm, T., Hermansky, H., and Andreou, A. G. 1997. Learning the mel-scale and optimal vtn mapping. In *Center for Language and Speech Processing, Workshop (WS 1997). Johns Hopkins University.*

Kaur, M. J. and Mittal, P. 2017. On developing an automatic speech recognition system for commonly used english words in indian english. *International Journal on Recent and Innovation Trends in Computing and Communication 5,* 7, 87–92.

Kumar, Y. and Singh, N. 2017. An automatic speech recognition system for spontaneous punjabi speech corpus. *International Journal of Speech Technology 20,* 2, 297–303.

Kurzekar, P. K., Deshmukh, R. R., Waghmare, V. B., and Shrishrimal, P. P. 2014. A comparative study of feature extraction techniques for speech recognition system. *International Journal of Innovative Research in Science, Engineering and Technology 3,* 12, 18006–18016.

Lähdesmäki, H. and Shmulevich, I. 2008. Learning the structure of dynamic bayesian networks from time series and steady state measurements. *Machine Learning 71,* 2-3, 185–217.

Lučić, B., Ostrogonac, S., Vujnović Sedlar, N., and Sečujski, M. 2015. Educational applications for blind and partially sighted pupils based on speech technologies for serbian. *The Scientific World Journal 2015.*

Nainan, S. and Kulkarni, V. 2016. A comparison of performance evaluation of asr for noisy and enhanced signal using gmm. In *2016 International Conference on Computing, Analytics and Security Trends (CAST)*. IEEE, 489–494.

Nasereddin, H. H. and Omari, A. A. R. 2017. Classification techniques for automatic speech recognition (asr) algorithms used with real time speech translation. In *2017 Computing Conference*. IEEE, 200–207.

Ruan, S., Wobbrock, J. O., Liou, K., Ng, A., and Landay, J. 2016. Speech is 3x faster than typing for english and mandarin text entry on mobile devices. *arXiv preprint* *arXiv:1608.07323*.

Schmitt, A., Zaykovskiy, D., and Minker, W. 2008. Speech recognition for mobile devices. *International Journal of Speech Technology 11,* 2, 63–72.

Shanthi Therese, S. and Lingam, C. 2013. Review of feature extraction techniques in automatic speech recognition. *International Journal of Scientific Engineering and Technology 2,* 6, 479–484.

Thalengala, A. and Shama, K. 2016. Study of sub-word acoustical models for kannada isolated word recognition system. *International Journal of Speech Technology 19,* 4, 817–826.

Walha, R., Drira, F., El-Abed, H., and Alimi, A. 2012. On developing an automatic speech recognition system for standard arabic language. *International Journal of Electrical and Computer Engineering 6,* 10, 1138–1143.

Dr. Gulbakshee J. Dharmale completed Ph.D. in Computer Science and Engg. from SGB Amravati University, Amravati. She is a life member of ISTE since 2011. She has published 10 research papers in reputed journals including in Scopus, SCI, etc., and conferences including in IEEE and it's also available online. Her main research work focuses on Artificial Intelligence and Machine learning.

Dr. Dipti D. Patil pursued M.E. Computer Engineering from TSEC Mumbai University, Mumbai in 2007 and Ph.D. Computer Engineering from SGB Amravati University, Amravati in 2014. She is currently working as Associate Professor in MKSSS's Cummins College of Engineering for Women, Pune since 2014. She is a member of BoS-Information Technology, SPPU, LMISTE, and LMCSI. She has published more than 50 research papers in reputed journals including in Scopus, SCI, ACM, etc., and conferences including in IEEE and it's also available online. Her research work focuses on Machine Learning, Pattern Recognition, Classification, Neural Networks, and Artificial Intelligence.