Hidden Markov Models for various speech to text recognition techniques

Er. Parnnet Kaur
Assistant Professor
Department of CSE
Chandigarh University
Punjab, INDIA
Parneet1207@gmail.com

Er. Arshpreet Kaur
Assistant Professor
Department of CSE
Chandigarh University
Punjab, INDIA
arshpreetofficial @gmail.com

***Abstract*- Punjabi language is popular Indo-Aryan language. Its phoneme sounds are tonal in nature which dissent in almost Indian side of Punjab. Recent research works reveal less significant work done towards developing a speech recognition system in Punjabi. For the liability of local people of Punjab and to build an Automatic Speech Recognition system, the work done is intended to feature the variability in the correctness and accuracy of various feature extraction techniques.**

Hidden Markov Models

The model given in Figure 1.2 could be drawn-out to a HMM as characterized in Figure 1.3. In the second model all observations generated can be emitted through a finite state probability. This serves as a better platform than the previously one discussed .The main usage of HMM is to verify that the recognised states for observation sequence top-top-bottom are actually exact; hence the state sequence is 'hidden'. Therefore with the guidance of this terminology we can calculate the final probability of the results produced and also sum with generating most likely state observations.

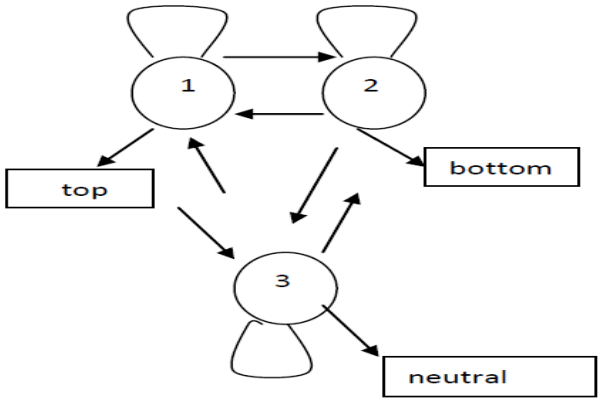

Figure 1.2: Markov Process (Example)

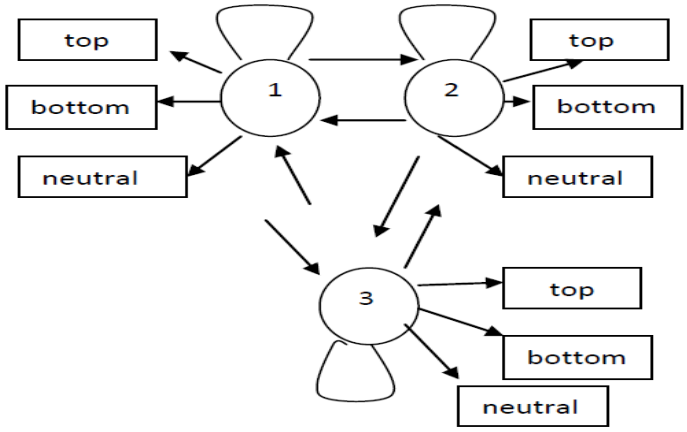

Figure 1.3: Hidden Markov Model (Example)

Applications of ASR

1. Blind people find difficult to read from the screen and writing on paper every time is inefficient in today's time, therefore Speech Recognition is an eventual interface for physically challenged people.

2. To interact with a computer interface has been inconvenient for the people, as it should be user friendly. Hence, speech recognition systems help a lot as it reduces the use of hands and eyes in operating the system.

3. It has been used eventually over telephone (automatic dial systems), telephone directory assistance, NLIDB (Natural Language Interface for Databases), voice dictation systems like expert systems, automatic voice translation for any language etc.

4. Speech applications installed publically in railways, multiplexes, airport communication and tourist information where tourists are answered for their instant queries. It would play an important role in levelling the gap between people having language communication problem.

Gap Analysis

1. The ASR systems for isolated words have been implemented using MFCC, PLP feature extraction techniques based on HMM and the maximum database was taken for 115 words [5] [7] with an achieved accurate level of 95.63%. Hence, the database must be expanded to analyze the behavior of the proposed system. Moreover ASR systems are yet to be developed for input that is not defined in pre-defined vocablury taking to a spontaneous recognition approach.

2. The ASR systems for continuous words have been implemented with a maximum for 100 sentences [8] and got accurate results up to 82.18%. Therefore the database requires more no of sentences for the analysis of continuous stream approach.

3. Speakers used for Automatic Speech Signal testing and recording purpose [5] [8] are on a maximum count of 9. More speakers may give tremendous variations to our Speech Recognition purpose. Hence, new ASR systems should be built taking more number of speakers to develop a robust and more redundant system.

4. Features extraction techniques like MFCC and LPC [8] have been implemented using HMM and DTW and achieved 94% and 96% accuracy. LPCC and other new techniques are required to be implemented through HMM, DTW and MATLAB to compare the outputs and get a super-oriented technique for ASR systems.

5. For Punjabi language almost all Automatic Speech Recognition systems were built using feature extraction techniques like MFCC, LPCC, LPC, and PLP [5] [7] [9] [10] for isolated and continuous Speech Signal. Hence, research area needs some other feature extraction techniques to be tested on the subjected language so as to have a wider approach in the testing performance environment.

Generation of Acoustic Model

To enable a system so as to perform analysis on a speech signal, it has to be defined in a static text form which is known as acoustic analysis. At the end of results the outcome is compared with some references previously defined in the acoustic model analysis. Two types of forms exist for acoustic model i.e. word model and phoneme model. Earlier the HMM framing is used as a prototype for declaring each word in the dictionary. Then this topology is used for HMM's up to 4 states.

In the next step *HCompV* tool is used instead of the combinational tool platform of *HInit* and *HRest* to estimate optimal values for transition probability, mean and variance vectors for all observation functions in a loop fashion to train each HMM known as re-estimation step. To get to at most value where the value of convergence vector does not decrease further, the loop process is further continues by *HERest* tool. In our system it has been repeated for 5 times.

Generation of Language Model and Decoding through Viterbi Search

Language Model

Language Model involves the creation of grammar that defines the set of rules to define word so that the word can be easily recognized by the system. For the purpose, task grammar uses some special symbols like braces, square brackets and others. Language Model defines the validity of word.

Grammar model contains chosen words with a pipeline symbol in between to give the probability for the generation of words.

Viterbi Algorithm

Viterbi algorithm [6] is used to decide the output sequence from acoustic and language models created in previous steps. For that purpose the parametric value of the HMM's need to be surfaced to predict out most likely state sequence out of all stare sequences.

Dynamic programming based algorithm Forward (or Backward) is used to predict the final path from state sequences. Moreover the state sequences either coding or non coding are generated using HMM model which defines the overall probability of the output sequence.

Improvement in current systems

To improve the above system following issues need to be tackled:

Spontaneous speech recognition system needs enlarged set of vocabulary set for recognition of every uttered word through the speaker.

Accuracy for noisy corpus needs various noise compression techniques and filters to gain higher form of accurate results.

Tonal aspect in Punjab is a major concern as various speakers speak according to their regional aspect laying a greater effect to the recognition accuracy.

As many inter regional population live in Punjab and use a mixture of Hindi sounds with Punjabi pronounciation, so there is a greater need of developing an hybrid system for Punjabi language.

Large database for connected approach is still in a great concern to be taken care for developing a robust speech recognition system.

Sphinx, Auditory toolbox, MATLAB and KALDI can attain higher accuracy than HTK tool.

REFERENCES

[1]. Antwarg, L. Rokach, L. & Shapira, B, "Attribute-Driven Hidden Markov Model Trees for Intention Prediction Part C (Applications and Reviews)", IEEE Transactions on Systems, Man, and Cybernetics, 2012, 42, PP. 1103-1119.

[2]. Kim, C. & Stern, R., "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition", IEEE/ACM Transactions on Audio and Language Processing, Speech, 2016, PP. 1

[3]. Kim, C. & Stern, R. M., "Power-Normalized Cepstral Coefficients (PNCC) for robust speech recognition", Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, 2012, PP. 4101-4104.

[4]. Dan, Z. M. & Monica, F. S, "A study about MFCC relevance in emotion classification for SRoL database", Electrical and Electronics Engineering (ISEEE), 2013 4th International Symposium on, 2013, PP. 1-4

[5]. Yücesoy, E. & Nabiyev, V. V., "Comparison of MFCC, LPCC and PLP features for the determination of a speaker's gender", Signal Processing and Communications Applications Conference (SIU), 2014 , 2014, PP. 321-324.

[6]. W. Zhu, D.O.Shaughnessy, "Using noise reduction and spectral emphasis techniques to improve ASR performance in noisy conditions", Proc. ASRU 2003.

[7]. W. Zhu, D. O. Shaughnessy, "Incorporating frequency masking filtering in a standard MFCC feature extraction algorithm," Proc. ICSP, 2004, PP. 617-620.

[8].Turan, M. A. T. & Erzin, E, "Source and Filter Estimation for Throat Microphone Speech Enhancement", IEEE/ACM Transactions on Audio, Speech and Language Processing, 2016, PP.265-275.

[9]. Alam, J., Ouellet, P., Kenny, P., O. Shaughnessy, D., "Comparative Evaluation of Feature Normalization Techniques for Speaker Verification," Proc NOLISP, LNAI 7015, 2011, PP. 246-253.

[10]. F. H. Liu, R. M. Stern, X. Huang, and A. Acero, "Efficient cepstral normalization for robust speech recognition", in Proc. ARPA Human Language Technology Workshop, 1993, PP.69-74.