HMM-based Automatic Speech Recognition Systems: A survey

Author's Details:
[1]Rajota Kharwal, [1]Dil Nawaz Hakro, [2]Intzar Lashari, [1]Tuba Qureshi, [1]Maryam Hameed
[1]Institute of Information and Communication Technology, University of Sindh, Jamshoro,Pakistan
[2]Institute of Business Administration, University of Sindh, Jamshoro,Pakistan
Corresponding emails: dill.nawaz@gmail.com,

Abstract:

Natural language processing enables computer and machines to understand and speak human languages. Speech recognition is a process in which computer understand the human language and processes further instructions as per recognition of the human language. The human language varies so the machine or computer needs entirely different algorithms as the human languages differ in various aspects, such as sounds, phonemes, words, meanings and much more. Understanding human language is a challenging job and for this purpose Hidden Markov Models are used commonly as they possess promising results in understanding human language. A survey of various researches employing Hidden Markov models is presented to highlight the importance of HMM in the process of speech recognition.

INTRODUCTION

A computer is a more powerful and useful device because it gives the number of benefits including to provide the ability to complete a task with high quality in seconds, it also facilitates you to learn new skills that are too important in current technological years. In current years of the technological world, computers are stiffly used to communicate with another via text and speech. There are many tools are freely available for several languages such as electronic dictionaries, text processing tools and leading speech processing system likewise text-to-speech (Speech generation) as well as speech-to-text (speech recognition) (Laurent Besacier et al., 2013). The development of speech recognition was started since 1950's at AT&T in Bell Labs, too many efforts and investments were carried for the advancements of speech recognition research. In late1960s, the conducted research was focused on the recognition of phonemes and isolated words. From 1970s to late 1980s, the major tendency of speech recognition research was attracted by the connected words and continues speech recognition (Sameti, H. et al., 2011) Automatic speech recognition (ASR) introduced the responsibility of recognizing a person based on her/ his sound with the guidance of machine (Patil, H. A., et al 2008). The intelligence of speech recognition system empowered the machine to understand the voice instructions (Watile, Y., et al 2015). Speech recognition system differ in complexities. For example To recognise the simple and bounded isolated words vocabulary of a Speaker dependent recognition under controlled environment but it can be too complicated to recognise the large vocabulary of a speaker dependent under the noisy environment(Ali, H., et al, 2014). The size of small vocabulary systems contributes the recognition capability about 100 words, the capability to recognize the medium size vocabulary contributes up to 100 to 1000 words, and the capability to recognize the large vocabulary system contributes more than 1000 words (Gosavi, S. D.,Khot, U. P., & Shah, S. (2013)The collection of audio recordings called speech corpus which is basic fundamental to build automatic speech recognition system to perform its functions(Rauf, S., et al, 2015)

Markov Chain:

Abdullah and kasaboy (1999) presented an algorithm classified by HMM which is purely following Markov chain. (Abdullah and Kasab OV) proposed an example in their research paper about the weather. A model has been constructed to conclude tomorrow's weather depends on present day condition. The prototype based on three strategies which show all the days weather of the town under-investigated. This could be Rainy (R), Sunny (S), and Cloudy (C) from the weather's history. Here is a table (Table-1) of following probabilities of today's condition based weather for tomorrow

		Tomorrow		
		Sunny(S)	Cloudy(C)	Rainy(R)
Today	Sunny(S)	.7	.2	.1
	Cloudy(C)	.05	.8	.15
	Rainy(R)	.15	.25	.6

Table 1: probabilities of weather expectation

State w is referred to the condition of weather that is fragment at instant t and to find the tomorrow's weather condition being in present day condition $P(w_t+1/w_t)$.

History of an unobjectionable similarity for n instant is:

$P(wt + 1/wt , wt-1 , wt-2 , ... , wt-n) \approx P(wt+1 /wt)$.

This is considered the earliest structure of Markov chain in history has been premeditated for only one instant. Figure no (1) presents the finite sate diagram of the weather probabilistic. If today's weather is sunny (S) then what are the probabilities of upcoming R, C and S considered as five days and S mentioned in above model?

P(w1 = S, w2=S, w3=C, q4=C, w5=R, w6=S) = P(S).P(w2=S/w1=S). P(w3=C/w2=S). P(w4=C/w3=C). P(w5=R/w4=C). P(w6=S/w5=R)

Initial probability of day S denoted by P(S) is 100%, at the time of considered current day weather is sunny.

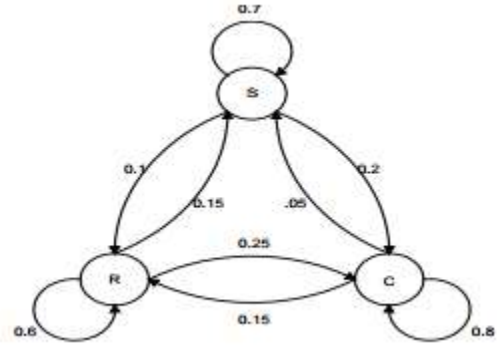

Fig. (1) Finite state representation of weather forecast problem (Abdullah and kasaboy ,1999))

Hidden Markov Model

As understood most prosperous and fruitful modelling access to ASR is the use of HMM in acoustic models apply to recognise the sub word or whole word unit speech unit (Hui Jiang et al., 2006). One of the major advantages of the HMM is the implication of statistical method used in speech recognition (Juang and Rabiner., 1991). Hidden Markov Models along with speech recognition combination has proved an attractive tool in the past decades. While the number of recorded speech recognition system depends on HMM is too expensive to examine detail in this spot. It is very helpful to indicate some most effective and more significant as well as successful for speech recognition system. Voice dictation system is another well-established effort by IBM (Averbuch,et al., 1987; Jelinek 1976; Jelinek, and Mercer 1983;) The Defense Advanced Research Projects Agency Resource Management Task for the consecutive speech recognition system (Chow et al. 1987; Lee 1989). According to the traditional pattern recognition, the development of the statistical pattern recognition is the perfect key to the automatic speech recognition puzzle (R.O. Duda et al., 1973; Fukunaga, K. 2013) Hidden Markov Model is also applied to recognise the natural language model (Kupiec, J.1992). The popularity of the HMM framework is designed straightforward and simple structure to implement and to give its clear performance. Hidden Markov Model be also permitted to serve the sequence of the sounds within a sector of speech. Each unit of speech known as phoneme and a phoneme can be modelled by an individual Hidden Markov Model (Rabiner, L. R.1989; Veera valli2005;).

Modules Of ASR:

There are five identified modules for automatic speech recognition systems given below are fig (1):

- Speech signal preprocessing
- Feature extraction
- Acoustic modeling
- Lexical and language modelling
- Recognition word

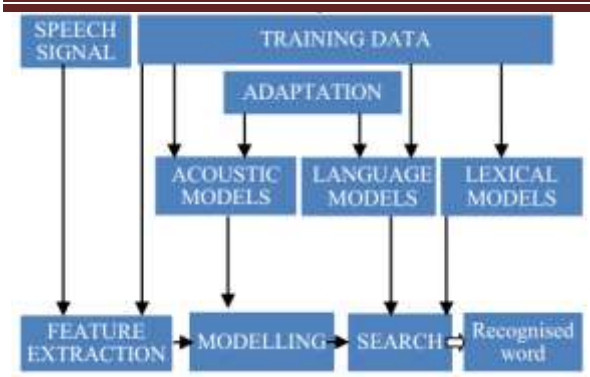

Fig no: (1) modules of ASR

Feature Extraction:

The very first conversion applied in the input speech signal of the automatic speech recognition is known as feature extraction (Sivadas, S., 2004). Feature extraction can be understood as the crucial segment of steps in any speech recognition system. It is assumed that the feature extraction is the heart of the speech recognition (Narang, S., & Gupta, M. D., 2015).Theoretically, it should be also possible to recognise the given input speech signal directly through digitised waveform, although, the reason to do feature extraction is to reduce the variability of large vocabulary speech. There are so many techniques are available for feature extraction including: *ƒ* Mel-frequency cepstral coefficients (MFCC), *ƒ* Power spectral analysis (FFT), Linear predictive cepstral coefficients (LPCC) *ƒ* First order derivative (DELTA), Linear predictive analysis (LPC), *ƒ* Relative spectral filtering of log domain coefficients (RASTA) and others (Shrawankar , 2013).

1	Principal Component analysis (PCA)	Non linear feature extraction method, Linear map, fast, eigenvector-based	Traditional, eigenvector base method, also known as karhunen-Loeve expansion; good for Gaussian data
2	Linear Discriminate Analysis(LDA)	Non linear feature extraction method, Supervised linear map, fast, eigenvector-based	Better than PCA for classification[9]
3	Independent Component Analysis (ICA)	Non linear feature extraction method, Linear map, iterative non- Gaussian	Blind source separation, used for de-mixing non- Gaussian distributed sources(features)
4	Linear Predictive coding	Static feature extraction method,10 to 16 lower order coefficient,	It is used for feature Extraction at lower order
5	Cepstral Analysis	Static feature extraction method, Power spectrum	Used to represent spectral envelope[9]
6	Mel-frequency scale analysis	Static feature extraction method, Spectral analysis	Spectral analysis is done with a fixed resolution along a Subjective frequency scale i.e. Mel-frequency Scale.
7	Filter bank analysis	Filters tuned required frequencies	
8	Mel-frequency cepstrum (MFCCs)	Power spectrum is computed by performing Fourier Analysis	This method is used for find our features
9	Kernel based feature extraction method	Non linear transformations	Dimensionality reduction leads to better classification and it is used to redundant features, and improvement in classification error.[11]
10	Wavelet	Better time resolution than Fourier Transform	It replaces the fixed bandwidth of Fourier transform with one proportional to frequency, which allow better time resolution at high frequencies than Fourier Transform
11	Dynamic feature extractions i)LPC ii)MFCCs	Acceleration and delta coefficients i.e. II and III order derivatives of normal LPC and MFCCs coefficients	It is used by dynamic or runtime feature
12	Spectral subtraction	Robust Feature extraction method	It is used basis on Spectrogram[4]
13	Cepstral mean subtraction	Robust Feature extraction	It is same as MFCC but working on Mean statically parameter
14	RASTA filtering	For Noisy speech	It is find out Feature in Noisy data
15	Integrated Phoneme subspace method (Compound Method)	A transformation based on PCA+LDA+ICA	Higher Accuracy than the existing Methods[4]

Techniques of feature extraction with their properties (Gaikwad, S. K, et al., 2010)

Acoustic Modelling:

An acoustic model is the most paramount constituent of an automatic speech recognition which interprets for almost all of the computational load and performance of the system. The main important role in developing of acoustic modelling is to detecting the phonemes spoken by users. Their creation includes the use of phonic recordings of user speech and then their word scripts and then systemizes all them into the mathematical representation of speech which makes up words.(Ghai, W., & Singh, N. (2012)

Lexical Modelling:

Lexicon model is constructed to provide users with the syllabification of individual words in a given language. Along lexical model, there are multifarious kinds of combinations of phonemes are interpret to recognise and show valid words for the

recognition. The neural network is accompanied to construct Non-native speech recognition for the lexical model.

Language modelling:

The language model is usually used in smaller, the language models vary in categories, for example, limited vocabulary tasks are commonly described manually in the terminology of deterministic a definite state representations, For the huge speech vocabulary recognition project namely N-grams based stochastics (trigrams, bigrams and much more) has been used widely. As the results presented by sparse problem of data training, N-gram smoothing contingency is conventionally compulsory for cases with $N \geq 2$. Class dependent trigrams and bigrams have also been imported. To store for more much huge language constraints, hierarchal language models have also been presented. (Jelinek, F, 1985)The usage of context-free language in r automatic speech recognition ASR is still bound usually cause of the increasing in computation recommended to implement such type of grammars. (Ney, H. 1991)

Conclusion:

By using various approaches the problems of Sindhi phonemes and pronunciation can be handled at various levels specially by using Hidden Markov Models or applying some other classifiers such as artificial neural network and support vector machines. A working speech recognition system can be built understanding building blocks of Sindhi language followed by a complex system for Sindhi language which can understand and generate a complete Sindhi language as Sindhi is considered one of the most difficult spoken language of the available scripts. Various approaches employed in various speech recognition systems can be applied as well as some new approaches can be proposed as there is no any speech recognition system for Sindhi is available.

References:

1. Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, *56*, 85-100.
2. Sameti, H., Veisi, H., Bahrani, M., Babaali, B., &Hosseinzadeh, K. (2011). A large vocabulary continuous speech recognition system for the Persian language. *EURASIP Journal on Audio, Speech, and Music Processing*,*2011*(1), 6.
3. Patil, H. A., &Basu, T. K. (2008). Development of speech corpora for speaker recognition research and evaluation in Indian languages.*International Journal of Speech Technology*, *11*(1), 17-32.
4. Watile, Y., Ghotkar, P., &Rohankar, B. (2015). COMPUTER CONTROL WITH VOICE COMMAND USING MATLAB. *COMPUTER*, *3*(6).
5. Ali, H., Ahmad, N., Zhou, X., Iqbal, K., & Ali, S. M. (2014). DWT features performance analysis for automatic speech recognition of Urdu.*SpringerPlus*, *3*(1), 204.
6. Rauf, S., Hameed, A., Habib, T., & Hussain, S. (2015, October). District names speech corpus for Pakistani Languages. In *Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLER), 2015 International Conference*(pp. 207-211). IEEE.
7. Gosavi, S. D., Khot, U. P., & Shah, S. (2013). Speech recognition for robotic control. *IJERA*, 408-413.
8. Jiang, H., Li, X., & Liu, C. (2006). The Large margin is hidden Markov models for speech recognition. *IEEE transactions on audio, speech, and language processing*, *14*(5), 1584-1595.
9. Juang, B. H., &Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, *33*(3), 251-272.
10. Baker, J. (1975). The DRAGON system--An overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *23*(1), 24-29.
11. Averbuch, A., Bahl, L., Baker, R., Brown, P., Daggett, G., Das, S., ...&Fraleigh, D. (1987, April). Experiments with the TANGORA 20,000 word speech recognizer. In *Acoustics, Speech, and Signal Processing,*

IEEE International Conference on ICASSP'87. (Vol. 12, pp. 701-704). IEEE.

12. Jelinek, F. (1976). Continuous speech recognition by statistical methods.*Proceedings of the IEEE*, *64*(4), 532-556.
13. Bahl, L. R., Jelinek, F., & Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, (2), 179-190.
14. Chow, Y., Dunham, M., Kimball, O., Krasner, M., Kubala, G., Makhoul, J., ...& Schwartz, R. (1987, April). BYBLOS: The BBN continuous speech recognition system. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87*. (Vol. 12, pp. 89-92). IEEE.
15. R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973
16. Fukunaga, K. (2013). Introduction to statistical pattern recognition. Academic press.
17. Kupiec, J. (1992). Robust part-of-speech tagging using a hidden Markov model. *Computer Speech & Language*, *6*(3), 225-242.
18. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286.
19. Veeravalli, A. G., Pan, W. D., Adhami, R., & Cox, P. G. (2005, March). A tutorial on using hidden Markov models for phoneme recognition. In*System Theory, 2005. SSST'05. Proceedings of the Thirty-Seventh Southeastern Symposium on* (pp. 154-157). IEEE.
20. Abdulla, W., &Kasabov, N. K. (1999). *The concepts of hidden Markov* model in speech recognition. Department of Information Science, University of Otago.
21. Sivadas, S. (2004). Tandem feature extraction for automatic speech recognition.
22. Narang, S., & Gupta, M. D. (2015). Speech Feature Extraction Techniques: A Review. International Journal of Computer Science and Mobile Computing, 4(3), 107-114.
23. Shrawankar, U., &Thakare, V. M. (2013). Techniques for feature extraction in speech recognition system: A comparative study. arXiv preprint arXiv:1305.1145.
24. Gaikwad, S. K., Gawali, B. W., &Yannawar, P. (2010). A review on speech recognition technique. International Journal of Computer Applications, 10(3), 16-24.
25. Ghai, W., & Singh, N. (2012). Literature review on automatic speech recognition. International Journal of Computer Applications, 41(8).
26. Jelinek, F. (1985). The development of an experimental discrete dictation recognizer. Proceedings of the IEEE, 73(11), 1616-1624.
27. Ney, H. (1991). Dynamic programming parsing for context-free grammars in continuous speech recognition. IEEE Transactions on Signal Processing, 39(2), 336-340.