A Review on Domain Based Automatic Speech Recognition Technology

Nobel Jacob Varghese[1] , Dr. Cini Kurian[2]

[1] *Associate Engineer,Virtusa Consulting Services Private Limited*

[2] *Associate Professor, Al-Ameen College, Edathala, Aluva, Kerala*

Abstract *- Automatic Speech Recognition Technology is a multidisciplinary research area having tremendous potential. It has become a integral part of future intelligent system in which speech recognition and speech synthesis are used as the basic mode of communicating with humans. In this paper, a survey on the technological development of domain based automatic speech recognition is presented.*

Keyword — *Automatic Speech Recognition, Malayalam.*

I. INTRODUCTION

Designing a machine that converse with human, particularly responding properly to spoken language has intrigued engineers and scientists for centuries. Today speech technology enabled applications are commercially available for a limited but interesting range of tasks. Very useful and valuable services are provided by these technology enabled machines, by responding correctly and reliably to human voices. In order to bring us closer to the "Holy Gr ail" of machines that recognize and understand fluently spoken speech, many important scientific and technological advances have been took place, but still we are far from having a machine that mimics human behaviour. Speech recognition technology has become a topic of great interest to general population, through many blockbuster movies of 1960's and 1970's [1]. The anthropomorphism of "HAL", a famous character in Stanley Kubrick's movie "2001: A Space Odyssey", made the public aware of the potential of intelligent machines. In this movie, an intelligent computer named "HAL" spoke in a natural sounding voice and w as able to recognize and understand fluently spoken speech, and respond accordingly. George Lucas, in the famous Star Wars saga, extended the abilities of intelligent machines by making them intelligent and mobile Droids like R2D2 and C3PO were able to speak naturally, recognize and understand fluent speech, move around and interact with their environment, with other droids, and with the human population. Apple Computers in the year of 1988, created a vision of speech technology and computers for the year 2011, titled "Knowledge Navigator", which defined the concepts of a Speech User Interface (SUI) and a Multimodal User Interface (MUI) along with the theme of intelligent voice-enabled agents. This video had a dramatic effect in the technical community and focused technology efforts, especially in the area of visual talking agents [1].

Languages, on which so far automatic speech recognition systems have been developed are just a fraction of the total around 7300 languages. Chinese, English, Russian, Portuguese, Vietnamese, Japan, Spanish, Filipino, Arabic, Bengali, Tamil, Malayalam, Sinhala and Hindi are prominent among them [2]

.

II. EARLIER TECHNOLOGIES FOR AUTOMATIC SPEECH RECOGNITION

Many attempts have been started in the 2nd half of the 18th century to develop machines to mimic a human's speech communication capability. The early interest was not on recognizing and understanding speech but instead on creating a speaking machine [3, 4]. Speech pioneers like Harvery Fletcher and Homer Dudley firmly established the importance of the signal spectrum for reliable identification of the phonetic nature of a speech sound. Following the convention established by these two outstanding scientists, most modern systems and algorithms for speech recognition are based on the concept of measurement of the speech power spectrum (or its variants such as the cepstrum), due to the fact that measurement of the power spectrum from a signal is relatively easy to accomplish with modern digital signal processing techniques. Theory of acoustic-phonetics, which describes the phonetic elements of speech (the basic sounds of the language), was the guided factor in the design of early automatic speech recognition systems. Non-uniformity of time scales in speech events was one of the difficult problems of speech recognition. Martin at RCA Laboratories developed some time normalization methods [5]. Vintsyuk in the soviet Union proposed the use of Dynamic programming methods for time aligning a pair of speech utterances. (Generally known as dynamic time wrapping (DTW), including algorithms for connected word recognition [6]). At the same time, in an independent effort in Japan, Sakoe and Chiba at NEC Laboratories also started to use dynamic programming technique to solve the non-uniformity problem [7]. Publication by Sakoe and Chiba in the field of dynamic programming and its numerous variant forms (including Viterbi Algorithm which came from the communication theory community), has become an indispensable technique in automatic speech recognition [8]. Reddy at Carnegie Mellon University conducted pioneering research in the field of continuous speech recognition by dynamic tracking of phonemes [9]. Pattern recognition methods based on LPC was a significant technological advancement in the speech recognition research. Velichko and Zagoruyko in Russia advanced the use of pattern – recognition in speech recognition [10]. Atal and Itakura independently formulated the fundamental concepts of Linear Predictive coding (LPC), which greatly simplified the estimation of the vocal tract response from speech wave forms [11,12]. The basic idea of applying fundamental pattern technology to speech recognition based on LPC methods were proposed by itakura, Rabiner et.al and others [13, 14].

III. SURVEY BASED ON TASK /DOMAIN

Isolated digit recognition, Connected digit recognition, Continuous Speech recognition and spontaneous speech recognition are the common type of domains on which speech recognition works/operates. A survey on each of these tasks is outlined below.

A. Isolated Digit Recognition Task

Davis et.al from Bell laboratories built the isolated digit recognizer for a single speaker [67]. Another effort was from Japan where the digit recognizer hardware was built by Nagata and co-workers at NEC Laboratories [68]. In Malay language, a speaker independent recognizer was built by Al-Haddad et al in 2007 using Discrete Time Wrapping methods [69]. In 1985 L.Rabiner et.al have developed isolated digit recognizer based on Hidden Markov Model with continuous mixture density [70]. In 1990, Neural Prediction models being used by Iso to develop speaker independent digit recognizer [71]. Sakoe et.al has used Dynamic programming Neural Networks for digit recognition and the result was compared with that of HMM [72].

B. Connected Word Recognition Task

Connected word speech recognition is the system where the words are separated by pauses. It is s a class of fluent speech strings where the set of strings are derived from small-to-moderate size vocabulary such as digit strings, spelled letter sequences, combination of alphanumeric etc. Rabiner et al. analyzed three algorithms designed for connected word recognition: Two level DP approach, Level Building approach and One Pass approach [73]. These algorithms differ in computational efficiency, storage requirement and ease of realization in real time hardware. Garg et al have developed a speaker dependent connected digits recognition system by applying unconstrained Dynamic time warping technique in which they recognized each digit by calculating distance with respect to matching of input spoken digit with stored template [74].

C. Continuous Speech Recognition (CSR)

Mohammad A. M. et al. worked for the development of continuous speech recognition system on Arabic Language using Sphinx as well as HTK tools [75]. Five-state Hidden Markov Models (HMM) having 3 emitting states for triphone acoustic modelling were used. Their statistical Language model contained unigrams, bigrams, and trigram. This system was tested for different combinations of speakers and sentences. To ensure and validate the pronunciation correctness of the speech data, a manual human classification and validation of the correct speech data was conducted. A round robin technique was applied for fair testing and evaluation of this system and to make this system speaker independent. The word recognition accuracy was best for different speakers with similar sentences and was least for different speakers and different sentences. These problems have a solution in the additional morpheme level of speech signal representation. Ronzhin and Karpov kept these factors into consideration while developing a large vocabulary continuous speech recognition system [76]. On incorporation of morpheme level, the size of needed vocabulary was reduced. HMM with mixture Gaussian, probability density function was used as an acoustic model. They studied the peculiarities of Russian language to a great depth such as longer size of Russian words, set of accents and dialects, strict grammatical constructions, diverse phonetic structure. They applied dynamic warping of sentences to create a search of optimal matching of two sentences. Recognition accuracy of morpheme-based recognizer came about 95% which was found to be 1.7 times faster than word based recognizer. Another work was on Vietnamese language. It is a syllabic tonal language with six tones where each syllable has only one tone. The meaning of the word depends on the tone. Keeping this factor into consideration, Thang Tat Vu et al. developed a LVCSR for Vietnamese language and applied the combination of Mel Frequency Cepstral Coefficient (MFCC) and F0 features and bigram language model to improve the accuracy of their ASR. Incorporation of F0 gave a significant increase of around 10% in recognition accuracy [77]. For better speech recognition on small size of trained speech data, another smaller component has been used i.e. sub-syllable. For example, Chinese is a mono-syllable and tonal language in which each syllable of a character is composed of an initial and a final tone. Huang feng-long used sub-syllable for generating features while developing an independent speech recognition system using HMM for small vocabulary [78]. To improve the performance, they applied keyword-spotting criterion. This criterion has a basis that in spite of the ASRs being guided by grammatical constraints, speaking natural sentences and noise lowers the performance of an ASR. Sinhala is one of the less-resourced non-Latin language for which speaker dependent continuous speech recognizer have been developed using HTK by Nadungodage and Weerasinghe [79]. A considerable increase in the size of vocabulary for continuous speech recognizer due to the differences between written and spoken Sinhala, pushed them to take only written Sinhala vocabulary.

D. Spontaneous Speech Recognition

In order to increase recognition performance for spontaneous speech, several projects have been conducted. In Japan, a five year national project “Spontaneous Speech: corpus a nd processing technology” was conducted [80]. A world- largest spontaneous speech corpus, “Corpus of Spontaneous Japanese(CSJ) “ consisting of approximately 7 M words, corresponding to 650 hours of speech, was built, and various new technique were investigated. A spontaneous speech is a speech, which is natural sounding and not rehearsed. An ASR for such a speech handles a variety of natural speech features i.e. words being run together, “ums”, ”ahs” and slight utter. Usually recognition accuracy drastically decreases for spontaneous speech [80, 81]. One of the major reasons for this decrease is acoustic and language models used up until now have generally been built using written language or speech read from a text. In spontaneous speech, pronunciation variation is so diverse that multiple surface form entities are needed for many lexical items. Kawahara et al. have found that statistical modelling of pronunciation variations integrated with language modelling is effective in suppressing false matching of less frequent entries [82]. Another difficulty of spontaneous speech recognition is that generally no explicit sentence boundary is given. Therefore, it is impossible to recognize spontaneous speech sentence by sentence. Lousier et al. investigated combinations of unsupervised language model adaptation methods for CSJ utterances. Shinozaki et al. have proposed a combination of cluster-based language models and acoustic models in the framework of a Massively Parallel Decoder (MPD) to cope with the problem of acoustic as well as linguistic variations of utterances [83, 84].

IV. CONCLUSION

Speech recognition technology is a fast developing area, with the onus currently on user friendly gadgets designed to serve the general public. The advancement of this technology from machines that can partially mimic human speech capabilities to the designing of a machine that can function like an intelligent human is a foregone conclusion. However the barriers that challenge and confound this surge to a fully fledged success are yet to be breached. Presumably many years will pass before natural conversation between human beings and machines becomes a reality.

REFERENCES

[1] B. H. Juang and L. R. Rabiner (2005), 'Automatic speech recognition–a brief history of the technology', in Elsevier Encyclopaedia of Language and Linguistics, Second Edition, Elsevier.

[2] Wiqas Ghai , Navdeep Singh "Literature Review on Automatic Speech recognition" International Journal of Computer Applications Volume 41– March 2012.

[3] H. Dudley and T. H. Tarnoczy, The Speaking Machine of Wolfgang von Kempelen, J. Acoust.Soc. Am., Vol. 22, pp. 151-166, 1950.

[4] H. Dudley, R. R. Riesz, and S. A. Watkins, A Synthetic Speaker, J. Franklin Institute, Vol.227, pp. 739-764, 1939.

[5] T. B. Martin, A. L. Nelson, and H. J. Zadell, Speech Recognition by Feature abstraction Techniques, Tech. Report AL-TDR-64-176, Air Force Avionics Lab, 1964.

[6] T. K. Vintsyuk, Speech Discrimination by Dynamic Programming, Kibernetika, Vol. 4, No. 2, pp. 81-88, Jan.-Feb. 1968.

[7] H. Sakoe and S. Chiba, Dynamic Programming Algorithm Quantization for Spoken Word Recognition, IEEE Trans. Acoustics, Speech and Signal Proc., Vol. ASSP-26, No. 1, pp. 43-49, Feb. 1978.

[8] A. J. Viterbi, Error Bounds for Convolution Codes and an Asymptotically Optimal Decoding Algorithm, IEEE Trans. Information Theory, Vol. IT-13, pp. 260-269, April 1967.

[9] D.R Reddy , " An approach to computer speech recogn ition by direct analysis of the speech wave", Tech. Report No.C549 , computer Science Dept. , Stanford Univ., 1966.

[10] V.M.Velichko and N.G.Zagoruyko, Automatic Recognition of 200 words, Int.J.Man-Machine Studies, 2:223, June 1970.

[11] B. S. Atal and S. L. Hanauer, Speech Analysis and Synthesis by Linear Prediction of the Speech Wave, J. Acoust. Soc. Am. Vol. 50, No. 2, pp. 637-655, Aug. 1971.

[12] F. Itakura and S. Saito, A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies, Electronics and Communications in Japan, Vol. 53A, pp. 36-43, 1970.

[13] F. Itakura, Minimum Prediction Residual Principle Applied to Speech Recognition, EEE Trans. Acoustics, Speech and Signal Proc., Vol. ASSP-23, pp. 57-72, Feb. 1975

[14] L. R. Rabiner, S. E. Levinson, A. E. Rosenberg and J. G. Wilpon, Speaker Independent Recognition of Isolated Words Using Clustering Techniques, IEEE Trans. Acoustics, Speech and Signal Proc., Vol. Assp-27, pp. 336-349, Aug. 1979.

[15] Jean Francois, Automatic Word Recognition Based on Second Order Hidden Markov Models , IEEE Transactions on Audio, Speech and Language processing Vol.5,No.1, Jan.1997.

[16] Mark J. F. Gales, Katherine M. Knill, et.al., State-Based Gaussian Selection in Large Vocabulary Continuous Speech Recognition Using HMM s ,IEEE Transactions On Speech And Audio Processing, Vol. 7,o. 2, March 1999.

[17] Qiang Huo et.al, Bayesian Adaptive Learning of the parameters of Hidden Markov model for speech recognition ,IEEE Transactions on Audio, Speech and Language processing Vol.3,No.5, Sept..1995.

[18] R. P. Lippmann, Review of Neural Networks for Speech Recognition, Readings in Speech Recognition, A. Waibel and K. F. Lee, Editors, Morgan Kaufmann Publishers, pp. 374-392,1990.

[19] B.H. Juang, C.H. Lee and Wu Chou, Minimum classification error rate methods for speech recognition, IEEE Trans. Speech & Audio Processing, T-SA, vo.5, No.3, pp.257-265, May 1997.

[20] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. Phoneme recognition using time-delay neural networks. IEEE Transactions on Acoustics, Speech and Signal Processing, 37:328–339, 1989.

[21] T. Robinson and F. Fallside. A recurrent error propagation network speech recognition system. Computer, Speech and Language, 5:259–274, 1991.

[22] H. Bourlard and N. Morgan. Continuous speech recognition by connectionist statistical methods. IEEE Transactions on Neural Networks, 4:893–909, 1993.

[23] H. Bourlard and N. Morgan. Connectionist speech recognition: a hybrid approach. Boston: Kluwer Academic, Norwell, MA (USA), 1994.

[24] T. Robinson, M. Hochberg, and S. Renals. The Use of Recurrent Neural Networks in Continuous Speech Recognition (Chapter 19), pages 159–184. Kluwer Academic Publishers, Norwell, MA (USA), 1995.

[25] W. Reichl and G. Ruske. A hybrid rbf-hmm system for continuous speech recognition. In Proceedings of the International Conference on Acoustics,Speech and Signal Processing (ICASSP), pages 3335–3338, Detroit, MI (USA), 1995.

[26] K. Iso and T. Watanabe. Speaker-Independent Word Recognition using a Neural Prediction Model. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 441–444, Alburquerque, New Mexico (USA), 1990 .

[27] J. Tebelskis, A. Waibel, B. Petek, and O. Schmidbauer. Continuous Speech Recognition using Predictive Neural Networks. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP),pages 61–64, Toronto, Canada, 1 991

[28] D. Ellis, R. Singh, and S. Sivadas. Tandem-acoustic modeling in largevocabulary recognition. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 517–520, Salt Lake City, Utah (USA), 2001.

[29] B.H. Juang, C.H. Lee and Wu Chou, Minimum classification error rate methods for speech recognition, IEEE Trans. Speech & Audio Processing, T-SA, vo.5, No.3, pp.257-265, May 1997.

[30] L. R. Bahl, P. F. Brown, P. V. deSouza and L. R. Mercer, Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition, Proc. ICASSP 86,Tokyo, Japan, pp. 49-52, April 1986.

[31] 6. V. N. Vapnik, Statistical Learning Theory, John Wiley and Sons, 1998.

[32] A. Ganapathiraju, J.E. Hamaker, and J. Picone. Applications of support vector machines to speech recognition. IEEE Transactions on Signal Processing, 52:2348–2355, 2004.

[33] N. Thubthong and B. Kijsirikul. Support vector machines for Thai phoneme recognition. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 9:803–13, 2001.

[34] J.M. Garc´ıa-Cabellos, C. Pel´aez-Moreno, A. Gallar do-Antol´ın, F. P´erez- Cruz, and F. D´ıaz-de-Mar´ıa. SVM Classifie rs for ASR: A Discussion about Parameterization. In Proceedings of EUSIPCO 2004, pages 2067–2070, Wien, Austria, 2004.

[35] A. Ech-Cherif, M. Kohili, A. Benyettou, and M. Benyettou. Lagrangian support vector machines for phoneme classification. In Proceedings of the 9th International Conference on Neural Information Processing (ICONIP'02), volume 5, pages 2507–2511, Singapore, 2002.

[36] D. Mart´ın-Iglesias, J. Bernal-Chaves, C. Pel´aez-M oreno, A. Gallardo-Antol´ın, and F. D´ıaz-de-Mar´ıa. A Speech Recognizer based on Multiclass SVMs with HMM Guided Segmentation, pages 256–266. Springer, 2005.

[37] R. Solera-Ure˜na, D. Mart´ın-Iglesias, A. Gallardo- Antol´ın, C. Pel´aez-Moreno, and F. D´ıaz-de-Mar´ıa. Robust ASR using Support Vector Machines.Speech Communication, Elsevier, 2006.

[38] S.V. Gangashetty, C. Sekhar, and B. Yegnanarayana. Combining evidence from multiple classifiers for recognition of consonant-vowel units of speech in multiple languages. In Proceedings of the International Conference on Intelligent Sensing and Information Processing, pages 387–391, Chennai, India, 2005.

[39] H. Shimodaira, K.I. Noma, M. Nakai, and S. Sagayama. Support vector machine with dynamic time-alignment kernel for speech recognition. In Proceedings of Eurospeech, pages 1841–1844, Aalb org, Denmark, 2001.

[40] H. Shimodaira, K. Noma, and M. Nakai. Advances in Neural Information Processing Systems 14, volume 2, chapter Dynamic Time-Alignment Kernel in Support Vector Machine, pages 921–928. MIT Press, Cambridge, MA (USA), 2002.

[41] K.-F. Lee, Large-vocabulary speaker-independent continuous speech recognition: The Sphinx system, Ph.D. Thesis, Carnegie Mellon University, 1988.

[42] R. Schwartz and C. Barry and Y.-L. Chow and A. Derr and M.-W. Feng and O. Kimball and F. Kubala and J. Makhoul and J. Vandegrift, The BBN BYBLOS Continuous Speech Recognition System, in Proc. of the Speech and Natural Language Workshop, p. 94-99, Philadelphia, PA, 1989.

[43] H. Murveit and M. Cohen and P. Price and G. Baldwin and M. Weintraub and J. Bernstein,SRI's DECIPHER System, in proceedings of the Speech and Natural Language Workshop,p.238-242, Philadelphia, PA, 1989.
[44] S. Young, et. al., the HTK Book, http://htk.eng.cam.ac.uk/.
[45] Adoram Erell et.al., Energy conditioned spectral estimation for Recognition of noisy speed , IEEE Transactions on Audio, Speech and Language processing, Vol.1,No.1, Jan 1993.
[46] Adoram Erell et.al., Filter bank energy estimation using mixture and Markov models for Recognition of Nosiy Speech , IEEE Transactions on Audio, Speech and Language processing Vol.1,No.1, Jan.1993.
[47] Javier Hernando and Climent Nadeu, Linear Prediction of the One-Sided autocorrelation Sequence for Noisy Speech Recognition **,**IEEE Transactions On Speech And Audio Processing, Vol. 5, No. 1, January 1997.
[48] C. J. Leggetter and P. C. Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, Computer Speech and Language, 9, 171-185, 1995.
[49] A. P. Varga and R. K. Moore, Hidden Markov model decomposition of speech and noise, Proc. ICASSP, pp.845-848, 1990.
[50] M. J. F. Gales and S. J. Young, Parallel model combination for speech recognition in noise, Technical Report, CUED/FINFENG/ TR135, 1993.
[51] K. Shinoda and C. H. Lee, A structural Bayes approach to speaker adaptation, IEEE Trans. Speech and Audio Proc., 9, 3, pp. 276-287, 2001.
[52] Mazin G.Rahim et.al., Signal Bias Removal by maximum Likelihood Estimation for Robust Telephone Speech Recognition ,IEEE Transactions on Audio, Speech and Language processing Vol.4,No.1, Jan.1996.
[53] Ananth Sankar, A maximum likelihood approach to stochastic matching for robust speech recognition, IEEE Transactions on Audio, Speech and Language processing Vol.4, No.3, May.1996.
[54] Doh-Suk Kim, Auditory Processing of Speech Signals for Robust Speech Recognition in Real-World Noisy Environments ,IEEE Transactions On Speech And Audio Processing, Vol. 7, No. 1, January 1999.
[55] Mark J. F. Gales, Katherine M. Knill, et.al., State-Based Gaussian Selection in Large Vocabulary Continuous Speech Recognition Using HMM s ,IEEE Transactions On Speech And Audio Processing, Vol. 7,No. 2, March 1999.
[56] Jen-Tzung Chien, Online Hierarchical Transformation Of Hidden Markov Models for Speech Recognition, IEEE Transactions On Speech And Audio Processing, Vol.7, No. 6, November 1999.
[57] K.H.Davis, R.Biddulph, and S.Balashek, Automatic Recognition of spoken Digits, J.Acoust.Soc.Am., 24(6):637-642,1952.
[58] K. Nagata, Y. Kato, and S. Chiba, Spoken Digit Recognizer for Japanese Language, NEC Res. Develop., No. 6, 1963.
[59] Md Sah Bin Hj Salam, Dzulkifli Mohamad, Sheikh Hussain Shaikh Salleh: Malay isolated speech recognition using neural network: a work in finding number of hidden nodes and learning parameters. Int. Arab J. Inf. Technol. 8(4): 364-371 (2011).