A Review on Technological Development of Automatic Speech Recognition

Cini Kurian

Abstract*— Speech recognition has been a challenging and multidisciplinary research area since decades. Speech Recognition technology has tremendous potential as it is an integral part of future intelligent devices, in which speech recognition and speech synthesis are used as the basic means for communicating with humans. In this paper, a survey of major landmarks in the research and development of automatic speech recognition is presented to provide a review of technological perspective and an appreciation of the fundamental progress that has been made in this area.

Keywords — Automatic Speech Recognition

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 Grail" 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 behavior. Speech recognition technology has become a topic of great interest to general population, through many block buster 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 general public aware of the potential of intelligent machines. In this movie, an intelligent computer named "HAL" spoke in a natural sounding voice and was 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].

Manuscript Received on September 2014.

Cini Kurian, Department of Computer Science , Al-Ameen College, Edathala, Aluva, Kerala, India.

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, Bangali, Tamil, Malayalam, Sinhala and Hindi are prominent among them [2]

II. AN INSIGHT INTO EARLIER TECHNOLOGIES FOR ASR

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].

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

III. TECHNOLOGIES FOR SPEAKER INDEPENDENT SYSTEMS

A speaker independent system that could deal with the acoustic variability intrinsic in the speech signals coming from many different talkers, often with notably different regional accents was the next goal for speech recognition researchers. This led to the creation of a range of speech clustering algorithms. Furthermore, researches to understand and to control the acoustic variability of various speech representations across talkers led to the study of a range of spectral distance measures (e.g., the Itakura distance and statistical modeling techniques that produced sufficiently rich representations of the utterances from a vast population[13,15]. Bell Laboratories' came up with a new concept of keyword spotting as a primitive form of speech understanding [16]. The technique of keyword spotting aimed at detecting a keyword or a key-phrase of some particular significance that was embedded in a longer utterance where there was no semantic significance to the other words in the utterance. The need for such keyword spotting was to accommodate talkers who preferred to speak in natural sentences rather than using rigid command sequences when requesting services (i.e., as if they were speaking to a human operator).

IV. SURVEY OF HIDDEN MARKOV MODEL (HMM) FOR SPEECH RECOGNITION

Speech recognition research in last three decades was characterized by a shift in methodology from the more intuitive template-based approach (a straightforward pattern recognition paradigm) towards a more rigorous statistical modeling framework. Although the basic idea of the hidden markov model (HMM) was known and understood early on in a few laboratories (e.g.IBM (International Business Machines) and the Institute for Defense Analyses (IDA)[17]), the methodology was not complete until the mid-1980's and it wasn't until after widespread publication of the theory that the Hidden Markov Model became the preferred method for speech recognition[17,18]. The popularity and use of the HMM as the main foundation for automatic speech recognition and understanding systems has remained constant over the past three decades, especially because of the steady stream of improvements and refinements of the technology. A weighted Hidden Markov Model algorithm and a subspace projection algorithm are proposed by Su and Lee , to address the discrimination and robustness issues for HMM based speech recognition[19]. A new hybrid algorithm based on combination of HMM and learning vector were proposed by L.R.Bahl et.al [20]. Learning Vector Quantization (LVQ) method showed an important contribution in producing highly discriminative reference vectors for classifying static patterns [21]. The Maximum Likelihood (ML) estimation of the parameters via Feedback (FB) algorithm was an inefficient method for estimating the parameters of HMM. To overcome this problem, Adoram Erell et.al. in their paper proposed a corrective training method that minimized the number of errors of parameter estimation [22]. A novel approach for a hybrid connectionist HMM speech recognition system based on the use of Neural Network (NN) showed the important innovations in training the Neural Network. Nam Soo Kim et.al. have presented various methods for estimating a robust output probability distribution (PD) in speech recognition based on the Discrete Hidden Markov Model (DHMM) [23,24]. An extension of the Viterbi algorithm made the second order HMM computationally efficient when compared with the existing Viterbi algorithm. Mark J. F. Gales et.al [25]. Investigate the use of Gaussian selection (GS) to increase the speed of a large vocabulary speech recognition system [26]. The theoretical frame work for Bayesian adaptive training of the parameters of Discrete Hidden Markov model (DHMM) and Semi Continuous HMM (SCHMM) with Gaussian mixture state observation densities were proposed by Qiang Huo et.al [27]. The proposed *maximum a- posteriori* (MAP) algorithms discussed are shown to be effective especially in the cases in which the training or adaptation data are limited.

V. SURVEY OF NEURAL NETWORKS FOR SPEECH RECOGNITION

Another technology that was (re)introduced in the late 1980's was the idea of artificial neural networks (ANN). Neural networks were first introduced in the 1950's, but failed to produce notable results initially [28]. The advent, in the 1980's, of a parallel distributed processing (PDP) model, which was a dense interconnection of simple computational elements, and a corresponding "training" method, called error back-propagation, revived interest around the old idea of mimicking the human neural processing mechanism. A particular form of PDP, the multilayer perceptron, received perhaps the most intense attention then, not because of its analog to neural processing but due to its capability in approximating any function (of the input) to an arbitrary precision. Early attempts at using neural networks for speech recognition centered on simple tasks like recognizing a few phonemes or a few words (e.g.,isolated digits), with good success [29]. However, as the problem of speech recognition inevitably requires handling of temporal variation, neural networks in their original form have not proven to be extensible to this task. In this context, two different classes of neural networks which consider the correlation between the temporal structures in the speech patterns were proposed: Time-Delay Neural Networks (TDNNs) and Recurrent Neural Networks (RNNs) [30,31]. TDNNs can be considered as a special type of the well-known Multilayer Perceptron (MLP) in which input nodes integrate shift registers (or time delays). Although these systems have shown to achieve good results on phoneme or isolated word recognition tasks, ANNs have not been successful on more complex tasks as continuous speech recognition. The main reason for this lack of success has been their inability to model the time variability of the speech signal even when recurrent structures are used.

VI. SURVEY OF HYBRID ANN/HMM FOR SPEECH RECOGNITION

To overcome these difficulties, several researchers have proposed the so-called Hybrid ANN/HMM-based ASR systems. The basic idea underlying these schemes is to combine HMMs and ANNs into a single system to get benefits from the best properties of both approaches: the ability of HMMs to model the time variability of the speech signal and the discrimination ability provided by ANNs. In the initially proposed approach of hybrid systems, ANN is

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

used to estimate jointly all the HMM state emission probabilities [32,33]. Several types of neural networks have been used for this purpose: MLPs [33-48], RNNs [34] and Radial Basis Function (RBF) networks [35-50]. Other approaches for speech recognition use Predictive Neural Networks [36, 37] which capture the temporal correlations between acoustic vectors. Finally, in the hybrid ANN/HMM system proposed [38], ANNs are used to estimate phone posterior probabilities and these probabilities are used as feature vectors for a conventional GMM (Gaussian Mixture Model) - HMM recognizer. This approach is called Tandem Acoustic Modeling and it achieves good results in context-independent systems. Numerous studies show that hybrid systems achieve comparable recognition results than equivalent (with a similar number of parameters) HMM-based systems or even better in some tasks and conditions. Even though, they present a better behavior when a little amount of training data is available, this, hybrid ANN/HMM have not been yet widely applied to speech recognition, because of the some of the problems which still remain open; such as the design of optimal network architectures and the difficulty of designing a joint training scheme for both, ANNs and HMMs.

VII. Survey of SVM and Hybrid SVM/HMM for Speech Recognition

A number of innovations took place in the field of pattern recognition. The problem of pattern recognition, which traditionally followed the framework of Bayes' which required estimation of distributions for the data, was transformed into an optimization problem involving minimization of the empirical recognition error [39]. This fundamental change of paradigm was caused by the recognition of the fact that the distribution functions for the speech signal could not be accurately chosen or defined, and that Bayes' decision theory would become inapplicable under these circumstances [40]. After all, the objective of the design of a recognizer should be to achieve the least recognition error rather than the best fitting of a distribution function to the given (known) data set as advocated by the Bayes criterion. The concept of minimum classification or empirical error subsequently spawned a number of techniques, among which discriminative training and kernel-based methods such as the support vector machines (SVM) have become popular subjects of study [40,41]. The improved discrimination ability of SVMs has attracted the attention of many speech technologists. However, its application to ASR is by no means straight forward. Variable time duration of the acoustic speech units has prevented ASR from being approached as a simple static classification problem. In fact, this has been for many decades one of the fundamental problems faced by the speech processing community and the main reason for the success of the HMMs. The main problem stems from the fact that conventional kernels can only deal with vectors of fixed length. Standard parameterization techniques, on the other hand, generate variable length sequences of feature vectors depending on the time duration of each speech utterance. Different approaches have been proposed to deal with the variable time duration of the acoustic speech units [42, 43]. The first one is Thai tone recognition in which they try to classify the five different lexical tones in that language: mid, low, falling, high and rising. They have proposed different feature length normalization procedures for each of them. Several ways of preprocessing the speech sequence to obtain a fixed dimension vector are analyzed for a noisy digit recognition task [44]. To solve the problem of the computational complexity of the SVM classical formulation by using an alternative Lagrangian one on the TIMIT database has been attempted [45]. Some hybrid approaches have also been proposed by the speech community [46,47]. Here the non-uniform distribution of analysis instants is provided by the internal states transitions of an HMM with a fixed number of states and a Viterbi decoder is used for dimensional normalization. In another work, it is acknowledged the fact that the classification error patterns from SVM and HMM classifiers can be different and thus their combination could result in a gain in performance [48]. Some new methods were introduced that belong to the family of methods based on sequence kernels, which try to solve the problem of different length sequences by adapting the kernel of the SVM to one that can work with samples of variable dimensionality [49,50]. The new approach, called Dynamic Time Alignment Kernel (DTAK) method has been proposed which uses as a kernel, where the score is obtained by means of a Dynamic Time Warping (DTW) algorithm. This seems simple and effective, which rescues an old technique of speech recognition: Dynamic Time Warping (DTW). They applied this technique to isolated words only [49].

VIII. Software Tools for Speech Recognition

The success of statistical methods caused the development of several new speech recognition systems including the Sphinx system from CMU , the BYBLOS system from BBN and the DECIPHER system from SRI [51,52,53]. CMU's Sphinx system successfully integrated the statistical method of hidden Markov models with the network search strength of the earlier Harpy system. Hence, it was able to train and embed context-dependent phone models in a sophisticated lexical decoding network, achieving remarkable results for large-vocabulary continuous speech recognition. The system that was made available by the Cambridge University team (led by Steve Young), called the Hidden Markov Model Tool Kit (HTK), was one of the most widely adopted software tools for automatic speech recognition research [54].

IX. Technologies for Robust Speech Recognizer

Following are the efforts made by few researchers to deal with different technical problems related to speech. Such as the sources of variability, noise etc. Not much work has been done on noisy speech recognition in the last two decades. One of the important methods introduced is the minimum mean square error (MMSE) method which is an estimate of the filter log energies. This has introduced a significant improvement over existing algorithms and is proposed by Adoram Erell and et.al [55]. A model based spectral estimation algorithm is derived that improves the robustness of ASR system to additive noise [56]. This algorithm is tailored for filter bank based systems where the estimation should seek to minimize the distortions as measured by the recognizers distance metric. Another model based spectral estimation algorithm is

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

derived to improve the robustness of speech recognition in noisy environment [57]. Various techniques were investigated to increase the robustness of speech recognition systems against the mismatch between training and testing conditions, caused by background noises, voice individuality, microphones, transmission channels, room reverberation etc. Major techniques include the maximum likelihood linear regression (MLLR), the model decomposition, parallel model composition (PMC), and the structural maximum aposteriori (SMAP) method for robust speech recognition. Mazin G.Rahim et.al present a signal bias removal (SBR) method based on maximum likelihood estimation for the minimization of the undesirable effects which occur in telephone speech recognition system such as ambient noise, channel distortions etc [58-62]. A maximum likelihood (ML) stochastic matching approach to decrease the acoustic mismatch between test utterances, and a given set of speech models was proposed to reduce the recognition performance degradation caused by distortions in the test utterances and/or the model set [63]. A new approach to an auditory model for robust speech recognition for noisy environments was proposed [64]. The proposed model consists of cochlear band pass filters and nonlinear operations in which frequency information of the signal is obtained by zero-crossing intervals. Compared with other auditory models, the proposed auditory model is computationally efficient, free from many unknown parameters, and able to serve as a robust front-end for speech recognition in noisy environments. Uniform distribution, is adopted to characterize the uncertainty of the mean vectors of the CDHMM's [65]. This works proposed two methods, namely, a model compensation technique based on Bayesian predictive density and a robust decision strategy called Viterbi. The proposed methods are compared with the conventional Viterbi decoding algorithm in speaker independent recognition experiments on isolated digits and connected digit strings , where the mismatches between training and testing conditions are caused by: 1) additive Gaussian white noise, 2) each of 25 types of actual additive ambient noises, and 3) gender difference. In an another attempt, a novel implementation of mini-max decision rule for continuous density hidden Markov model based robust speech recognition was proposed [66].

X. 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 have 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 modeling 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 got 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

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

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 and 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 stutters. Usually recognition accuracy drastically decreases for spontaneous speech [80, 81]. One of the major reasons for this decrease is that 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 modeling of pronunciation variations integrated with language modeling 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. Lussier 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].

XI. 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 recognition 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] B. H. Juang, S. E. Levinson and M. M. Sondhi, Maximum Likelihood Estimation forMultivariate Mixture Observations of Markov Chains, IEEE Trans. Information Theory, Vol.It-32, No. 2, pp. 307-309, March 1986.
[16] J. G. Wilpon, L. R. Rabiner, C. H. Lee and E. R. Goldman, Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models, IEEE Trans. On Acoustics, Speech and Signal Processing, Vol. 38, No. 11, pp. 1870-1878, November 1990.
[17] J. D. Ferguson, Hidden Markov Analysis: An Introduction, in Hidden Markov Models for Speech, Institute for Defence Analyses, Princeton, NJ 1980.
[18] S. E. Levinson , L. R. Rabiner and M. M. Sondhi "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition", Bell Syst. Tech. J., vol. 62, no. 4, pp.1035 -1074 1983.
[19] Su, K. Y., and C. H. Lee, "Speech Recognition Using Weighted HMM and Subspace Projection Approaches," IEEE Trans. on speech and audio processing,Vol. 2, No. 1, pp. 69-79, Jan. 1994.
[20] L.R.Bahl et.al, A method for the construction of Acoustic Markov Models for Words , IEEE Transactions on Audio, Speech and Language processing, Vol.1,No.4, Oct.1993.
[21] Shigeru Katagiri et.al., A New hybrid algorithm for speech recognition based on HMM segmentation and learning Vector quantization , IEEE Transactions on Audio, Speech and Language processing Vol.1,No.4, Jan 1995.
[22] Adoram Erell et.al., Filter bank energy estimation using mixture and Markov models for Recognition of Noisy Speech , IEEE Transactions on Audio, Speech and Language processing Vol.1,No.1, Jan.1993.
[23] Lalit R.Bahl et.al, Estimating Hidden Markov Model Parameters So as to maximize speech recognition Accuracy , IEEE Transactions on Audio, Speech and Language processing Vol.1,No.1, Jan.1993.
[24] Nam Soo Kim et.al., On estimating Robust probability Distribution in HMM in HMM based speech recognition , IEEE Transactions on Audio, Speech and Language processing Vol.3,No.4, July1995.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

[25] 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.
[26] 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.
[27] 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.
[28] 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.
[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] 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.
[31] T. Robinson and F. Fallside. A recurrent error propagation network speech recognition system. Computer, Speech and Language, 5:259–274, 1991.
[32] H. Bourlard and N. Morgan. Continuous speech recognition by connectionist statistical methods. IEEE Transactions on Neural Networks, 4:893–909, 1993.
[33] H. Bourlard and N. Morgan. Connectionist speech recognition: a hybrid approach. Boston: Kluwer Academic, Norwell, MA (USA), 1994.
[34] 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.
[35] 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.
[36] 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.
[37] 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, 1991.
[38] 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.
[39] 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.
[40] 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.
[41] 6. V. N. Vapnik, Statistical Learning Theory, John Wiley and Sons, 1998.
[42] 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.
[43] 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.
[44] J.M. Garc´ıa-Cabellos, C. Pel´aez-Moreno, A. Gallardo-Antol´ın, F. P´erez- Cruz, and F. D´ıaz-de-Mar´ıa. SVM Classifiers for ASR: A Discussion about Parameterization. In Proceedings of EUSIPCO 2004, pages 2067–2070, Wien, Austria, 2004.
[45] 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.
[46] D. Mart´ın-Iglesias, J. Bernal-Chaves, C. Pel´aez-Moreno, 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.
[47] 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.
[48] 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.
[49] 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, Aalborg, Denmark, 2001.
[50] 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.
[51] K.-F. Lee, Large-vocabulary speaker-independent continuous speech recognition: The Sphinx system, Ph.D. Thesis, Carnegie Mellon University, 1988.
[52] 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.
[53] 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.
[54] S. Young, et. al., the HTK Book, http://htk.eng.cam.ac.uk/.
[55] 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.
[56] 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.
[57] 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.
[58] 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.
[59] A. P. Varga and R. K. Moore, Hidden Markov model decomposition of speech and noise, Proc. ICASSP, pp.845-848, 1990.
[60] M. J. F. Gales and S. J. Young, Parallel model combination for speech recognition in noise, Technical Report, CUED/FINFENG/ TR135, 1993.
[61] 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.
[62] 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.
[63] 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.
[64] 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.
[65] 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.
[66] 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.
[67] K.H.Davis, R.Biddulph, and S.Balashek, Automatic Recognition of spoken Digits, J.Acoust.Soc.Am., 24(6):637-642,1952.
[68] K. Nagata, Y. Kato, and S. Chiba, Spoken Digit Recognizer for Japanese Language, NEC Res. Develop., No. 6, 1963.
[69] 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).

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

[70] Rabiner, B. Juang, S. Levinson, and M. Sondhi, "Recognition of isolated digits using hidden markov models with continuous mixture densities", AT&T Tech. Journal, 64(6), 1985.
[71] Iso, K. and Watanabe, T., "Speaker-Independent Word Recognition using a Neural prediction Model", Proc. ICASSP-90, pp. 441-444; Alburquerque, New México, USA, 1990.
[72] Sakoe, H., Isotani, R., Yoshida, K., Iso, K., Watanabe, T., "Speaker-Independent Word Recognition using Dynamic Programming Neural Networks"; Proc.ICASSP-89, pp. 29-32; Glasgow, Scotland; 1989.
[73] Rabiner, L. Juang, B. H., Yegnanarayana, B., "Fundamentals of Speech Recognition", Pearson Publishers, 2010.
[74] Garg, A., Nikita, Poonam, "Connected digits recognition using Distance calculation at each digit", IJCEM International Journal of Computational Engineering & Management, Vol. 14, October 2011, ISSN (Online): 2230-7893.
[75] Mohammad A. M. Abushariah, Moustafa Elshafei, Othman O. Khalifa, "Natural Speaker-Independent Arabic Speech Recognition System Based on Hidden Markov Models Using Sphinx Tools", May 2010.
[76] A.L. Ronzhin, A.A. Karpov. Large Vocabulary Automatic Speech Recognition for Russian Language. In Proc. of Second Baltic Conference on Human Language Technologies, Tallinn,Estonia, 2005, pp. 329-334
[77] Thang Tat Vu, Dung Tien Nguyen, Mai Chi Luong, John-Paul Hosom, "Vietnamese Large Vocabulary continuous speech recognition", INTERSPEECH 2005:1689-1692.
[78] Huang Feng-Long, "An Effective approach for Chinese speech recognition on small size of vocabulary", Signal & Image Processing: An International Journal (SIPIJ) Vol.2, No.2, June 2011.
[79] Nadungodage, T. and Weerasinghe, R., "Continuous Sinhala Speech Recognizer", Conference on Human Language Technology for Development, Alexandria, Egypt, 2-5 May 2011.
[80] Furui, S., 2003a. Recent advances in spontaneous speech recognition and understanding. In: Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, pp. 1-6.
[81] Furui, S., 2003b. Toward spontaneous speech recognition and understanding. In: Pattern Recognition in Speech and Language Processing. Chou, W., Juang, B.-H. (Eds.), CRC Press, New York, pp. 191-227.
[82] Kawahara, T., Nanjo, H. and Furui, S. 2001. Automatic transcription of spontaneous lecture speech, In: Proc. IEEE Workshop on Automatic Speech Recognition and understanding, Madonna di Campiglio, Italy.
[83] Lussier, L., Whittaker, E. W. D. and Furui, S., 2004. Combinations of language model adaptation methods applied to spontaneous speech. In: Proc. Third Spontaneous Speech Science & Technology Workshop, Tokyo, pp. 73-78.
[84] Shinozaki, T. and Furui, S., 2004. Spontaneous speech recognition using a massively parallel decoder. In: Proc. Inter speech-ICSLP, Jeju, Korea, 3, pp. 1705-1708.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.