IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735.Volume 10, Issue 4, Ver. I (Jul - Aug .2015), PP 61-67
www.iosrjournals.org

Survey of Technical Progress in Speech Recognition by Machine over Few Years of Research

[1]Nnamdi Okomba S., [2]Adegboye Mutiu Adesina, and [3]Candidus O. Okwor.
Computer Engineering Department
[1,2,3]Federal University Oye-Ekiti, Ekiti State, Nigeria

Abstract: *This paper reviews some of various research carried out over the last decade in the area of Automatic Speech Recognition (ASR) and discusses the major themes and advance made in the last decade of research, in order to show the outlook of technology and an appreciation of the fundamental progress that has been achieved in this weighty area of speech communication. Over period of research and development, the accuracy of automatic speech recognition remains one of the important research challenges such as variation of the context, environmental condition, speaker's variation and poor-quality audio. The design of speech recognition requires careful attention to the following issue: Definition of various types of speech classes, speech representation, techniques, database and performance evaluation. The history, challenges of speech recognition system and various techniques to solve these challenges constructed by various research works have been presented in a chronological order. The objective of this paper is to compare and summarize well know approaches used in various steps of speech recognition system.*

Keywords: *Automatic Speech Recognition, Database, Feature extraction, Performance evaluation, Techniques*

I. Introduction

Speech is the primary means of communication between human. Speech Recognition (also known as Computer Speech Recognition or Automatic Speech Recognition (ASR)) is the process of converting a speech signal to a sequence of word or other linguistic unit,by means of an algorithm implemented as a computer program [1]. It is the most natural form of human communication and it has been one of the most existing areas of the signal processing. Generation of speech waveforms and speech recognition has been under development for several decades [2].

The main goal of speech recognition area is to develop techniques and systems for speech input to machine. This paper reviews key highpoints during the last few decades in the research and development of speech recognition, so as to provide a technological perspective. Although many technical progresses have been made, there are still remaining many research problems that need to be addressed.

A. Speech Recognition System

Speech recognition system is used as intelligence home in personal communication system, banking system and security system [3], [4]. Speech recognition technology was increasingly used within telephone networks to automate as well as to enhance the operation service. Fig 1 shows the schematic diagram of speech recognition system for human being.

It has four main building block speech analyses, feature extraction coding, language translation and message understanding. Speech analysis made up of noise removal, silence removal and end point detection which are necessary to improve the performance of speech recognition system. The speech analysis also deals with suitable frame size for further analysis using sub segmentation, segmentation and super segmental analysis techniques [5].

The feature extraction and coding unit reduce the dimensionality of the input vector and maintain discriminating power of the signal. The spectral signal output of speech analysis converted to activity signal on the auditory nerve using neural transducer method. Then, activity signal converted into a language code within the brain and message understanding is finally accomplished for speech recognition [6].

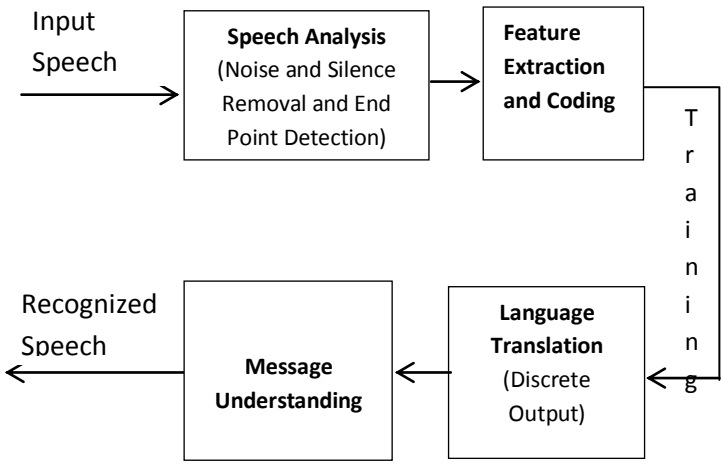

Fig. 1 Schematic diagram of Speech Recognition System

B. Types of Speech Recognition System

Automatic Speech Recognition System can be classified according to types of words they have the ability to recognize. The Wordsare classified as the following:

Isolated Words

Isolated word recognizers required each utterance to have quiet (lack of an audio signal) on both sides of the sample window. It accepts only single utterance/word at a time. It has Listen/ Non-Listen and requires the speaker to pause between utterances.

Connected Words

Connected word recognizers (more correctly as 'connected utterances') have almost the same characteristic with isolated words, but it is slightly difference in the sense that it allows separate utterance to be run-together with a least wait between them.

Spontaneous Speech

Recognizers with continuous speech proficiencies are sometimes most difficult to create because they utilize special technique to determine utterance boundaries. It allows users to speak nearly naturally, while system determines the content.

C. Automatic Speech Recognition(ASR)

Design Challenges

The task of Automatic Speech Recognition is to extract the primary linguistic message from a complex acoustic patter contains many sources of variability. The challenges on which recognition accuracy depends has been tabulated and presented in the Table 1.

Table 1: ASR System Design Challenges

Speakers Differences	Speaker dependence/ independence due to Sex, Age, Pronunciation etc.
Transducer	Telephone, Headset, Microphone
Environmental Variation	Noise Types, Surrounding noise level, Noise ratio and Working Condition
Vocabulary	Specific/ generic vocabulary, Characteristic of available training
Channel	Reflection of Sound, Band amplitude and distortion
Speech Style	Effect of stress, Speech, Loudness and Production

D. Application of Speech Recognition System

Speech recognition has been used in the various area of application. Table 2 present difference sector, input pattern, pattern classes and application area in which speech recognition systems have been used.

Table 2: Application of Speech Recognition

Sector	Input Pattern	Pattern Classes	Application Area
Education Sectors	Speech Wave Form	Spoken Words	Enable students who are physically handicapped and unable to use keyboard to enter text verbally
Non- Education Sectors	Speech Wave Form	Spoken Words	Computer games, Precision surgery
Translation	Speech Wave Form	Spoken Words	It is advanced application which translate from one language to another
Medical Sectors	Speech Wave Form	Spoken Words	Medical Transcriptions (digital speech to text), Health care
Military Sectors	Speech Wave Form	Spoken Words	Training air traffic controller, Fighter aircraft,
Artificial Intelligence (AI) Sector	Speech Wave Form	Spoken Words	Robotics
Generally	Speech Wave Form	Spoken Words	News reporting, Court reporting

II. Approach To Speech Recognition

Speech recognition research has been ongoing for more than 80 years. Over the period there have been three major approaches, each with various techniques as presented in the Table 3.

Table 3: Speech Recognition Technique

Approach	Technique	Representation	Recognition Function
Acoustic phonetic approach		Phonemes/ Segmentation and labeling	Probabilistic lexical access procedure
Pattern recognition approach	Template	Speech samples, Pixels and curves	Correlation, distance measure
	Dynamic Time Wave (DTW)	Set of a sequences of spectral vectors	Dynamic warping optimal algorithm
	Vector Quantization (QV)	Set of a spectral vectors	Clustering functions (Codebook)
	Statistical	Features	Discrimination
	Support vector Machine	Kernel based features	Maximal margin hyper plane
AI approach		Knowledge based	
	Neural Network	Rules/ Unit/ Procedures based	Network function

III. Feature Extraction

Feature extraction is a process of extracts data from the speech (Voice signal) that is unique for each speaker. Mel Frequency Ceptral Coefficient (MFCC) technique is frequently used to create the fingerprint of the sound files [7]. The Mel Frequency Ceptral Coefficient based on the know variation of the human ear's critical bandwidth frequencies with filters spaced linearly at low frequencies and logarithmically at high frequencies used to capture the important characteristics of speech [8], [9].

The main goal of the feature extraction step in speech recognition is to compute a parsimonious sequence of feature vectors providing a compact representation of the given input signal.

The three stage in which feature extraction usually performed are:

1st Stage: The first stage is called acoustic front end or speech analysis. It carries out temporal analysis of signal and produce raw feature that describing envelop of the power spectrum of short speech interval.

2nd Stage: The second stage composed of static and dynamic feature.

3rd Stage: The third stage transforms these extended feature vectors into more compact and robust vectors that are then delivered to the recognizer. The Table 4 presents various methods for feature extraction in speech recognition.

Table 4: Feature Extracted Method

S/N	Method	Property
1	Mel frequency cepstrum (MFFCs)	Power spectrum is computed by performing Fourier Analysis
2	Kernel based feature extraction method	Non- linear transformation
3	Cepstral Analysis	Static feature extraction method, Power Spectrum
4	Filter bank analysis	Filters tuned required frequencies
5	Principal component analysis (PCA)	Non-linear feature extraction method, Linear map/ fast/ eigenvector-based
6	Mel frequency scale analysis	Statics feature extraction method, Spectral analysis
7	Wavelet	Better time resolution than Fourier Transform
8	Dynamic feature extraction (LPC &MFCCs)	Acceleration and delta coefficients
9	Cepstralmeanssubstraction	Robust feature extraction
10	Linear Discriminant Analysis (LDA)	Non-linear feature extraction method, Linear map, iterative non Gaussian

IV. Overview Of Speech Recognition

A. Between Years 1920 To 1960s

In the early 1920s machine recognition came in to an existence. The first machine to recognize speech to any important step commercially named, Radio Rex (toy) was manufactured in 1920 [10]. Research into the concepts of speech technology started in the year 1936 at Bell Laboratory. In 1939, Bell Labs showed a speech synthesis machine at World Fair in New York before they later abandoned struggle to develop speech simulated listening and recognition based on an inappropriate conclusion that artificial intelligence would eventually be necessary for accomplishment.

In 1950s, earliest attempt to device system for automatics speech recognition by machine were made when several researchers tried to exploit the fundamental concepts of acoustic phonetics. During this period, most of the speech recognition system investigated special resonances during the vowel system of each utterance which were extracted from output region of each utterance which were extracted from output signals of an analogue of filter bank and logic circuits [11].

In 1952, at Bell Laboratories, David, Biddulph and Balashek developed a system for isolated digit recognition for a single speaker [12]. The system depends deeply on measuring spectral resonances during the vowel region of each digit. In 1956, at RCA Laboratory, Olson and Belar tried to recognize 10 distinct syllables of a single talker, as personified in 10 monosyllabic words [13]. In another effort of Fry and Denes at University College in England, in 1959, a phoneme recognizer that recognizes four vowels and nine consonants were developed [14]. They used a spectrum analyzer and a pattern matcher to make the recognition decision. Again during 1959 period vowel recognizer of forgie and forgie built at MIT Licoin Laboratory in which 10 vowels embedded in a /b/ vowel /t/ format were recognized in a speaker independent mode [15].

B. Between Years 1960 To 1980s

In the 1960s at Japanes Laboratory Suzuks and Nakata started their research in the speech recognition field and developed special purpose hardware as part of their system due to computation that were not fast enough then [16]. In 1962s, another hardware effort in Japan was the work of Sakari and Doshita of Kyoto University, who developed hardware phoneme recognizer [17]. The third Japanese work wasthe digit recognizer hardware of Negate and Coworkers at NEC Laboratory in 1963 [18].

In the separate effort of Japan Sakoe and Chiba at NEC Laboratories dynamic programming technique was used to solve the non-uniformity problems [19].The final achievement of annotation in the late 1960s was pioneering research of Reddy in the area of continuous speech recognition by dynamic tracking of phonemes [20].

The field of isolated word or discrete utterance recognition became a possible and functional technology in 1970s through the fundamental studies by Velichko and Zagoruyko in Russia [21], Itakura in United State, Cakoe and Chiba in Japan [22]. Durring this decade, striving speech understanding project was funded by Defence Advanced Research Projects Agencies (DARPA), which lead to various seminal system and technology [23]. One of the demonstration of speech understanding was achieved by CMU in 1973 and Heresay I system was able to used semantic data to significantly moderate the number of alternatives considered by the recognizer. CMU's Harphy system was displayed to be able to recognize speech using a vocabulary of 1, 011 words with the judicious accuracy [24].

Another success of research in the 1970s was the beginning of a longstanding, extremely successful effort of group in large vocabulary speech recognition at IBM in which researchers studied three different tasks over a period of almost two decades, namely the New Raleigh Language [25] for simple database queries, Laser potent text Language [26] for transcribing laser potent and lastly the oofice correspondent tasks called Tangora [27] for dictation of simple communications.

C. Between years 1980 to 200s

Research in field of speech recognition in 1980s was characterized by a shift in technology from template based approach to statistical modeling method, most especially the hidden Markov model (MM) approach [28], [29].

The approach of hidden markov model (HMM) was well known and understands in a few laboratories like Primary IBM, Institute for Defense Analysis (IDA) and Dargon Systems but it became extensively used in the middle of 1980s. Another innovative technology that came into existed in the late 1980s was the method of applying neural network to problem of speech recognition [30]. The approach was first introduced in the 1950s, but they did not prove useful initially because they had many practical problems [31] [32].

Era of 1980s was decade in which a major motivation was given to large vocabulary and continuous speech recognition system by the defences Advanced Research Project Agency (DARPA) community, sponsored a large research program aimed at accomplishing high word accuracy for 1000 word continuous speech recognition, database management task. Major research contributions resulted from effort at CMU [33], AT & T Bell Labs [34], Lincoln Labs [35] and SRI [36]. The CMU (also known as SPHINX System) successfully integrated the Statistical method of HMM with the network search strength of the earlier Harpy System.

D. Between Years 1990 To 2000

The year 1990s was a decade in which a number of innovations took place in the area of pattern recognition. The problem of pattern recognition which traditionally followed the framework of Bayes and required estimation of distributions for the data was changed into an optimization problem resulting to the reduction of the empirical recognition error [37].

During this decade, a key issue {38] in design and implementation of speech recognition system was how to appropriately select the speech material used to train the recognition algorithm. A number of human language technology projects funded by DARPA in the 1980s and 1990s further enhanced the progress, as showed by many papers published in the proceedings of the DARPA speech and Natural language/ Human language workshop. The research describes the development of accomplishments for speech recognition that were conducted in the 1990s [39], at Fajitsu Laboratories Limited.

E. Between Years 2000 Till Date

In year 2000, Variational Bayesian (VB) estimation and clustering technique were developed [40]. This VB method was based on a subsequent distribution of parameters. Giuseppe Richardi [41] developed this technique to solve the problem of adaptive learning in automatic speech recognition and also proposed active learning algorithm for automatic speech recognition in the year 2005. Some enhancements have been worked out on large vocabulary continuous speech recognition system on performance improvement [42].

Sadaoki and Furui investigated SR technique in 2005 that can adapt to solve speech variation using a large number of model trained based on clustering technique. De-waehter et al. [43], attempted to overcome the time dependencies probems in Speech Recognition(SR) by using straight forward template matching technique. In 2008, the authors [44] explained the application of corresponding framelevel of phone and phonological attributed classes. In the recent work carried out by Vrinda and Chander, suitable speech recognition was developed for hindilanguage, for the people thatare physically challenges and cannot able to operate the computer through keyboard and mouse,using hidden markov model (HMM) to recognize speech sample to give admirable result for isolated words [1].

V.Speech Databases

Speech databases have an extensive uses in Automatics Speech Recognition (ASR). It is also used in other key applications such as Automatic Speech Synthesis, Coding and Analysis, SpeakerLanguage Identification and Verification. These applications required large amounts of recoded database. Speech databases are most generally classified into multi-session and single-session databases.

Multi-session databases allow estimation of temporal intra-speaker variability. Based on acoustic environment, databases are recorded either in noise free environment, such as office or home. Also, according to the purpose of the databases, some corporations are designed for mounting and evaluating speech recognition.

V. Summary Of Thetechnology Progress

In the last 60 years, most especially in the last three decades, research in speech recognition has been highly carried out worldwide, encouraged on by advance in signal processing algorithms, architectures and hardwares. The technological progress in the 60 years is summarized and presented in the Table 4 [45].

Table4: Summary of the Technological Progress over 60 years

Past	New (Present)
Template matching	Corpus-based statistical modeling
Filter bank/ spectral resonance	Cepstral features, kernel based function, group delay function
Heuristic time normalization	DP/ DTW matching
Distance-based	Likelihood based method
Read speech recognition	Spontaneous speech recognition
Maximum likelihood	Discriminative
Hardware recognizer	Software recognizer
Monologue recognition	Dialogue/Conversation recognition
Clean speech recognition	Noisy/Telephone speech recognition
Small vocabulary	Large vocabulary
Isolated word recognition	Continuous speech recognition
Single speaker recognition	Speaker-independent/adaptive recognition
Single modality (audio signal only)	Multimodal (audio/visual) Speech recognition

VI. Performance Measure Of Speech Recognition

Speech recognition accuracy and speech recognition rate are two important to be consider in order measuring performances of speech recognition system. Speech recognition accuracy is measured in terms of Word Error Rate (WER), whereas speech recognition rate is measured in terms of computation rate.WER is a common metric of the performance of speech recognition.

VII. Conclusion

Speech is the essential, most effective, reliable and common medium to communicate in real time system. There are so many applications of Speech Recognition Systems that are still far from reality due to deficiency of resourceful and reliable noise removal machinery and technique for improving the quantity of recorded speech of every word being comprehensible. This paper attempts to provide a comprehensive survey of research on speech recognition and the progress made in some years back till date. Although significance improvement has been made in the last 50 years, still we believe that there are so many works to be doing under the full variation in Speech Style, Environmental Condition and Speech Variation.

References

[1] Vrinda and S. Chander, "Speech recognition system for English language," International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 1, pp. 919-922, 2013

[2] L. Rabiner and B. H. Juang, "Fundamental of Speech Recognition," AT & T, 1993.

[3] Burstein, A. Stolzle, and R. W. Broderson, "Using speech recognition in a personal communication system" IEEE in International Conference on Conference on Communication, Vol. 3, pp. 1717-1992.

[4] T. Isobe, M. Morishima, F. Yoshitani, N. Koizumi, and K. Murakani, "Voice-activated home banking system and it field trial," International Conference on Spoken Language, Vol. 3, pp. 1688-1691, 1996.

[5] H. SJayanna and S. R. Mahadeva, "Analysis, Feature Extraction, Modeling and Tech., Rev., 26: 181-190, 2009.

[6] L. Rabiner, B. H. Juang, and B. Yegnanarayana, "Fundamentals of speech recognition," Pearson Education. First edition, ISBN 978-81-7758-560-5, 2009.

[7] M. Lindasalwa, B. Mumtaj, and Elamvazuthi, "Voice recognition algorithms using Mel Frequency Cepstral Coefficient (MFCC) and DTW Technique" Journal of Computing, Volume 2, Issue 3, 2010.

[8] K. Ahsamul and M. A. Mohammad, "Vector quantization in text dependent automatic speech recognition using Mel-Frequency Cepstrum Coefficient," 6th WSEAS International Conference on Circuits System, Electronics, Control & Signal Processing, Cairo, Egypt, pp. 352-355, Dec 29-31, 2007.

[9] S. Mahd and T. Azizollah, "Voice command recognition system based on MFCC and VQ algorithm, "World Academy of Science, Engineering and Technology Journal, 2009.

[10] W. Stefan and H. Reinhold, "Approaches to iterative speech feature enhancement and recognition, "IEEE Transaction on Audio, Speech and Language Processing, Vol. No.5, July 2009.

[11] F. Sadaoki, "50 years of progress in speech and speaker recognition research," ECTI Transactions on Computer and Information Technology, Vol. 1. No.2, November 2005.

[12] K. H. Davis,R.Biddulp,and S.Balashek, "Automatic recognition of spoken digit," J. Acoust.Soc.Am., 24(6):637-642, 1952.

[13] H. F. Olson and H. Belar, " Phonetic Typewriter," J. Acoust.Soc.Am., 28(6): 1072-1081, 1956.

[14] P. Deves, "The design and Operation of the Mechanical speech recognizer at University College London," J. British Inst. Radio Engr., 19(4), 211-299, 1959.

[15] J. W. Forgie and C. D. Forgie, "Results obtained from a vowel recognition computer program," J.A.S.A., 31(11), pp. 1480-1489, 1959.

[16] J. Suzuki and K. Nakata, "Recognition of Japanese vowel-preliminary to the recognition of speech," J. Radio Res. Lab 37(8), pp. 193-212, 1961.

[17] T. Sakai and S. Doshita, "The phonetic typewriter, information processing 1962," Proc. ITIP congress 1962.

[18] K. Nagata, Y. Kato, and C. Chiba, "Spoken digit recognizer for Japanese Language," NEC Res. Develop., No.6, 1963.

[19] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoustic, Speech and Signal Processing ASSP. 26(1), pp. 43-49, 1978.

[20] 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., September 1966.

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

[22] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoustics, speech, signal proc., ASSP 26(1), pp. 43-49, Febreary 1978.

[23] D. Klatt, "Review of the ARPA speech understanding project," J.A.S.A. 62(6), pp. 1324-1366, 1977.

[24] Lowrre, The HARPY speech understanding project," Trends in Speech Recognition, W. Lea, Ed., Speech Science Pub., pp. 576-586, 1990.

[25] Tappert, N. R. Dixon, A. S. Rabinowitz, and W. D. Chapman, Automatic recognition of Continuous speech utilizing dynamic segmentation, Dual Classification, Sequential Decoding and Error Recover," Rome Air Dev. Cen, Rome, NY, Tech. Report TR-71-146, 1971.

[26] F. Jelink, L. R.Bahl, and R. L. Mercer, "design of a Linguistic Statistical Decoder for the recognition of continuous speech," IEEE Trans. Information Theory, IT- 21 pp. 250-256, 1975.

[27] Jelink, "The development of an experimental discrete dictation recognizer," Proc. IEEE, 73(11), pp.1616-624, 1985.

[28] J. Ferguson, "Ed., Hidden Markov Models for speech," IDA, Princeton, NJ, 1980.

[29] L. Rabiner, "A tutorial on hidden markov model and selected applications in speech recognition," Proc. IEEE, 77(2) pp.257-286, February 1989.

[30] W. Chou and B. H. Juang, "(Eds.) Pattern recognition in speech and language processing, CRC Press, pp. 115-147, 2003.

[31] R.P. Lippmann, "An introduction to computing with neural nets," IEEE Trans. ASSP Mag., 4(2), pp. 4-22, April 1987.

[32] A.Weibel, et at., "Phoneme recognition using time-delay neural networks," IEEE Trans. Acoustics, Speech, Signal Proc., 37, pp. 393-404, 1989.

[33] K. F. Lee, "An overview of the SPHINX speech recognition system," Proc ICASSP, 38, pp. 600-610, 1990.

[34] H. Lee, et al., "Acoustic modeling for large volcabulary speech recognition," Computer Speech and Language, 4 pp. 127-165, 1990.

[35] M. Weintraub et al., "Linguistic constraints in hidden markov model based speech recognition," Proc. ICASSP, pp. 699-702, 1989.

[37] Y. Chow, et al. "BYBLOS, the BBN continuous speech recognition system," Proc. ICASSP, pp. 89-92, 1987.

[38] S. Furui and T. Ichiba, "Cluster-based modeling for ubiquitous speech recognition,"Department of Computer Science Tokyo Institute of Technology, Interspeech 2005.

[39] M. K. Brown et al. "Advance in speech recognition technology, "IEEE Transaction on Signal Processing Vol.39, No.6, June 1991.

[40] M. Afify and O. Siohan, "Sequential estimation with optimal forgetting for robust speech recognition," IEEE Transaction on Speech and Audio Processing," Vol. 12, No.4, July 2004.

[41] M. Afify, F. Liu, and H. Jiang, "A new verification-based fast-match for large vocabulary continuous speech recognition," IEEE Transaction on Speech and Audio Processing, Vol. 133 No.4 July 2005.

[42] Riccardi, "Activate learning: Theory and application to Automatic speech recognition," IEEE Transaction on speech and Audio Processing, Vol. 13, No. 4, July 2005.

[43] A.Erell et al., "Filter bank energy estimation using mixture and morkov models for recognition of noisy speech," IEEE Transactions on Audio, Speech and Language Processing, Vol.1, No.1, 1993.

[44] J. weih, "Construction modulation frequency domain based feature for robust speech recognition," IEEE Transaction on Audio, Speech and Language Processing, Vol.16, No.1 Jan 2008.

[45] A.sloin et al. "Support vector Machine Training for improvement hidden markov model," IEEE Transactions on Signal Processing, Vol.56, No.1, Jan. 2008.