Review And Challenges In Speech Recognition (ICCAS 2005)

M.Masroor Ahmed*, Abdul Manan Bin Ahmed**

* Faculty of Computer Science & Information Systems, Universiti Teknologi Malaysia, Skudai, Malaysia
(Tel: +60(07)-5532070; Email: masroorahmed@gmail.com)
* * Faculty of Computer Science & Information Systems, Universiti Teknologi Malaysia, Skudai, Malaysia
(Tel: +60(07)-5532070; Email: manan@fsksm.utm.my)

Abstract: This paper covers review and challenges in the area of speech recognition by taking into account different classes of recognition mode. The recognition mode can be either speaker independent or speaker dependant. Size of the vocabulary and the input mode are two crucial factors for a speech recognizer. The input mode refers to continuous or isolated speech recognition system and the vocabulary size can be small less than hundred words or large less than few thousands words. This varies according to system design and objectives.[2]. The organization of the paper is: first it covers various fundamental methods of speech recognition, then it takes into account various deficiencies in the existing systems and finally it discloses the various probable application areas.

Keywords: Dynamic Time Warping, Speech Recognition, Hidden Morkov Models, Isolated Word Recognition System, Speaker Dependent / Independent Recognition System

1. INTRODUCTION

Research in automatic speech recognition (ASR) aims to develop methods and techniques that enable computer systems to accept speech input and to transcribe the recognized utterances into normal orthographic writing. Four basic approaches to attain this goal have been followed and tested over the years[3]:

I. template-based approaches, where the incoming speech is compared with stored units in an effort to find the best match)

II. knowledge-based approaches that attempt to emulate the human expert ability to recognize speech

III. stochastic approaches, which exploit the inherent statistical properties of the occurrence and co-occurrence of individual speech sounds

IV. connectionist approaches which use networks of a large number of simple, interconnected nodes which are trained to recognize speech.

Speech recognition is a complex process due the parameters that influence speech including gender, age, race and cultural characteristics, such as dialects and accents. Speakers vary greatly in the clarity and speed of their speech, and words are not spoken individually, but slurred into a stream of sounds. A speech recognition system must identify the beginning and end of each word by 'listening' to the stream of phonemes. An effective speech recognition system must also deal with homophones- two words sound identical but have different meanings, and often different spellings.

2. APPROACHES TO SPEECH RECOGNITION

At present there are various approaches to recognize speech. These important techniques are briefly discussed in the succeeding paragraphs.

2.1 User Dependant and User Independent Recognition System

When the recognition system is trained exclusively by one person or by few persons. It is called speaker dependant recognition system. On the contrary, if a system is designed so that anyone can take charge of the system and the systems responds with equivalent efficiency, such a system is termed as speaker independent system. Since people from different regions of the world have different accents and more over everybody can not speak with the same speed, so building an efficient real time speaker independent speech recognition system is a big challenge [2]

2.2 Dynamic Time Warping

Dynamic Time Warping (DTW) is one of the common approaches to isolated word speech recognition. In this approach the template of each word is stored in the vocabulary. And the template of incoming speech is compared with each of the already stored templates. The closest match among the two templates is found and that is declared as the recognized utterance. This presents two problems: what form do the templates take and how are they compared to incoming signals.

The simplest form for a template is a sequence of feature vectors -- that is the same form as the incoming speech. The template is a single utterance of the word selected to be typical by some process; for example, by choosing the template which best matches a cohort of training utterances. The matching process needs to compensate for length differences and take account of the non-linear nature of the length differences within the words.

The Dynamic Time Warping algorithm achieves this goal; it finds an optimal match between two sequences of feature vectors, which allows for stretched and compressed sections of the sequence. [6][13][35][36][37].

2.3. Isolated Speech Recognition

Isolated speech recognition enables the system to take word by word input. This means that before the start and end of the word their will be silence. This phenomenon has made the system very simple, because we just have to detect two silence zones and will extract the features whatever is falling between these zones. The information extracted through these features is then compared with the already stored information in the system and the closest match is taken as the uttered word. But this is the case of ideal conditions i.e for example, when we are using the system without any other noise in some sound proof room. [34][35][36] [37] [38].

2.4 Connected Word Recognition System

Connect word systems (or more correctly 'connected utterances') are similar to Isolated words, but allow separate utterances to be 'run-together' with a minimal pause between them. [12][13][14][15][31][32].

2.5 Hidden Markov Models (HMM)

HMM is a statistical modeling of the various pronunciations possible, or acoustic references of a word (or phoneme or expression depending on the case). The Markov models technique is used in most automatic speech recognition systems. Figure 1 shows a HMM structure usually applied in speech recognition systems [30]

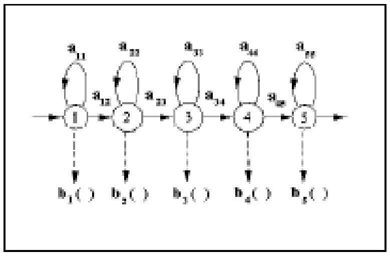

Fig. 1 A generic HMM.

As seen in Fiq. 1, we can define a HMM as follows:
$\lambda = (A,B,\pi)$ with this set of parameters:
A : { a_{ij} } = P { q_j / q_i} , Probability Transition Matrix, with dimension N^2 , and N is the number of states. This matrix describes a probability transition from state q_i to q_j .
B : Matrix $b_j(k)$ = P {V_k / q_j} , the probability to get the symbol V_k in the state q_j . and for DDHMM b_j (k) = {b_{jk}}; 1≤ j≤ N e 1≤ k≤ M
for all N model states and M symbols used on VQ.
π : Initial probability vector π(i) **.** Concerning HMM from figure 1 , this vector will always be defined as [1 0 0 0 ..],
{V_1,…V_k,…V_M}**:** set of M symbols
O = {O_1,…,O_T}: observation sequence in the interval [1,T]
Q = {Q_1,…,Q_T}: state sequence through the HMM in the interval [1,T].
N – Number of states
M – Number of symbols (number of centroids or, also, number of label-codes) For a more didactical approach in Hidden Markov Models with applications in speech recognition, we would recommend the following references
[10] [26] [27] [28] [29] [33].

2.6 Continuous Speech Recognition System

Continuous speech recognizers allow users to speak almost naturally. So finding the boundaries of utterance is a real challenging job due to the following important problems:

1) The quality of the speech signal may be affected by environmental noise or the transfer function of the trans-mission system, e.g., microphone and telephone.

2) In the acoustic signal, there is no clear indication or no indication at all of the boundaries between words or phonemes. Thus, not only the spoken words but also the phoneme boundaries and the word boundaries are unknown.

3) There is a large variation of the speaking rates in continuous speech.

4) The words and especially the word endings are pronounced less carefully in fluent speech than in an isolated speaking mode.

5) There is a great deal of interspeaker as well as intraspeaker variability caused by a number of factors, such as sex and physiological and psychological conditions.

6) For unrestricted natural-language speech input, the task-inherent syntactic-semantic constraints of the language should be exploited by the recognition system, in a way similar to human-to-human communication. [4][5][6][16][17][18][19][20][21][22][25]

2.7 Word Spotting

Key word spotting (KWS) is a recognition branch consisting of detecting a small set of keywords from a speech stream. Several methods for keyword spotting have been proposed a common approach is the use of filler models which represent the out of vocabulary words. In this method the recognition is made with a continuous speech recognition system, using a grammar formed by a filler and the keyword models other methods are based upon confidence measure to verify whether or not a given keyword exists within a segment of speech
[9][2]

3. FLAWS IN CURRENT SPEECH RECOGNITION SYSTEMS.

The present speech recognition systems carries the following shortcomings: These are approximately sorted, starting with concrete flaws in the lowest-level signal processing, and ending with problems in the entire abstract problem formulation as recovery of word sequences.

3.1 Blurring

Almost the first thing done with the signal data in current systems, after an initial spectral decomposition, is to blur or smooth the spectrum across frequency, either directly or by fitting a low-order spectral model [PLP]. Subsequent smoothing along the time axis is also becoming more common [RASTA]. Such blurring is demonstrably of great benefit to generalizing the phonemic templates, but the data removed by this step must carry some information of value to speech recognition.

3.2 Signal assumed speech

For obvious reasons, speech researchers have focused on the problem of distinguishing speech sounds from one another rather than the apparently simpler problem of distinguishing speech from non speech sounds. Unfortunately, working exclusively with valid speech signals has evolved efficient feature spaces that provide good phonetic discrimination but no basis by which to reject non speech insertions. While applications can be constructed that approach the speech-only assumption (such as close-talking microphones), misinterpretation of nonspeech is a serious problem in many scenarios, and it may also have led to an overly-reduced view of the speech signal that is not adequate to handle spontaneous and other marginal speech forms.

3.3 Vocal character ignored

Another simplifying assumption deeply embedded in speech recognizers is the idea that the vocal-tract shape, as reflected in smoothed, power-normalized spectral slices, holds all the information in speech. Every effort is made to remove the evidence of fundamental voice pitch (or voicing state), spectral cues to speaker identity, and amplitude modulation above the phoneme timescale. Although these cues may not be the primary determinants of phonetic class, it is likely that the overall problem of understanding speech cannot be solved unless machines, like their human antecedents, take them into account.

3.4 Absolute templates

Speech recognition proceeds by making probabilistic classifications of feature vectors, then finding the most likely utterance given the possible label sequences. The labels are usually phonemes, and the probabilistic classification generally amounts to measuring the distance between the observed feature vector and a set of prototypes in some suitably-weighted space. Perhaps the most awful problem with this is that all the different realizations of a phoneme are boiled down into a single exemplar, and apart from channel normalization and some limited context dependence (e.g. in triphone systems), all the variation due to different speakers and speaking styles is handled by broadening the predicted variance around this single ideal. It seems miraculous that absolute templates are able to work at all in classifying features which are as context-relative and adaptive as speech inflections, and that the approach of averaging all different speakers into a single template, rather than trying to find the appropriate speaker adaptation works as well as it does. But surely an approach that exploits the predictably consistent idiosyncrasies of a particular speaker across frames, phonemes and phrases, will work much better.

3.5 Independent frames discount duration

Many discussions of hidden-Markov model speech recognition systems start with the observation that the underlying probabilistic theory relies upon the assumption that successive feature-vectors are independent - meaning that their joint probability is the product of their individual probabilities - and this is clearly untrue in the case of speech, which displays a high degree of correlation along time in certain segments. The main ramification of this is that timing information - phonetic distinctions based on duration rather than spectrum - cannot be incorporated, despite their great significance in linguistics. Various approaches to `segmental modeling' have used criteria based on duration distributions measured in training to rescore the phonetic segmentation generated by the initial Markov decoding, but timing seems sufficiently important to demand a more directly integrated, and more context-sensitive, role within recognition.

3.6 Hard boundaries between labels

The speech recognition problem is currently posed as recovering a sequence of discrete state labels for a Markov process. The goal (as expressed in the training material) is traditional phonetic labeling, where the speech signal is partitioned into exclusive, adjacent regions labeled with distinct phoneme labels. Quite apart from the problems that may arise from treating all frames within a single labeled region as belonging to the same distribution, the task of placing an instantaneous boundary between different phonemes ranges from the tricky to the impossible, as reported by human transcribers. Discrete regions are not a good description of the speech signal, particularly at the phoneme level, where co-articulation and continuous transitions are the rule. A more satisfying foundation would involve labeling the speech signal with real-valued weights reflecting the varying amounts that each phoneme influences the sound at that moment; the phoneme sequence would then be realized as a set of overlapped curves reflecting the spread in time of a phoneme's acoustic evidence as imposed by the articulators.

3.7 Single class of low level elements

Although the phoneme has theoretical attractions, and has proved a successful foundation for existing speech recognition, looking at the different kinds of phonemes that exist shows that they are far from a homogeneous class. The hidden Markov model speech-recognition system is oriented towards vowels and sibilant, with their variable-length, pseudo-static spectra. However, equally important in speech are sounds defined by their dynamic properties: the transitions in initial and final consonants and the transients of stop-releases are not good candidates for spectral template matching, and typically have less durational variation. They might be better matched with a fixed time-frequency template. The wider point is that once we recognize that the form of the best model for different speech sounds varies depending on the particular sound in question, we can see the weakness in using a single signal model, the succession of repeatable spectral frames, as the foundation of speech recognition system.

3.8 Too much grammar

Initially, speech recognition was approached as a context-independent pattern recognition problem, but a more careful analysis of human speech perception reveals that the human ability to recognize speech gains considerably from the predictability of actual utterances; rather than having a free choice among 10,000 words, there are perhaps 5-10 that are most likely. Speech recognition systems that incorporate language models embodying the statistics of the grammar of the training corpus derive tremendous advantage from that information, yet the flaw here is that the powerful constraint of the language model is disguising weaknesses in the lower level features; like a person at a noisy party, a speech recognition system might be able to recover an utterance nearly perfectly as long as it conforms to expectations; when something less predictable is heard, the recognizer collapses in conditions that would be no challenge to a human listener. In the long term, it might be better at this time to work on speech recognizers that model human ability to recover spoken words in the absence of context, since this will give us a better idea of how we are approaching the solution of this preliminary problem.

3.9 Too modular

Reduction of a difficult tasks into more tractable pieces is one of the most powerful logical tools of problem solving. Yet any division of a problem limits the scope within which each part may be solved, and in abstraction problems like speech recognition, wider context is critical. Just as the particular meaning of an ambiguous word may be impossible to specify without knowing the whole sentence ("fruit flies like a banana"), so may the correct interpretation of a segment of speech vary considerably in different high-level contexts. For computational convenience, we construct our systems as sequences of modules - feature extraction, phoneme classification, word decoding - such that each operates independently of the others. Human auditory perception, free from constraints of logical analysis or systematic testing, very probably alters its lower levels of processing in response to the results of higher level analysis. By working with compartmentalized computer systems incapable of such top-down adaptation, we are at best performing much more computation than necessary (in order to anticipate all possible subsequent outcomes); at worst, we run the risk of performing completely inappropriate early processing for some of the time.

3.10 Punctuation not extracted

The speech recognition paradigm has, somewhat arbitrarily, become defined as the process of recovering word sequences from sound. This is not quite adequate, because when we transcribe language, we use a small amount of additional marking, most of which falls into the category of punctuation, in order to disambiguate the meaning of the word sequence. In many cases the phrase boundaries and emphases carried by punctuation can be inferred with some accuracy from the words alone, and thus for restricted or formalized tasks like TIMIT and the Wall Street Journal corpus, punctuation may have been a tolerable omission. However, in the case of spontaneous speech (e.g. the Switchboard corpus), the word sequences are much less neatly constructed, and the phrasing information, encoded in prosodic cues and transcribable as punctuation, are indispensable for an adequate representation of the meaning of the speech - and at the same time, adequate grammar-based language modeling for this kind of material may rely on these kinds of markers. Unfortunately, word-centered metrics such as the Word Error Rate are so central to the current business of speech recognition that the prospects for widespread acceptance of the importance of punctuation-style information appear discouraging [7][39]

4. APPLICATIONS OF SPEECH RECOGNITION

The speech recognition technology has seen a considerable maturity by now. This can be applied into various fields of engineering and medical sciences. On engineering side, the technology can be used for all types of automation. Whereas, in medical science, it can be used for rehabilitation of physically impaired people. [16]

REFERENCES

[1] http://www.infj/ulst.ac.uk/nlp/philips.html.

[2] Spanias, A.S.; Wu, F.H. “Speech coding and speech recognition technologies: a review.” *Circuits and Systems, 1991., IEEE International Sympoisum on , 11-14 June 1991*.

[3] http://www.htlcentral.org/page-827.0.shtml

[4] Ney, H.; Ortmanns, S..” Dynamic programming search for continuous speech recognition.” *Signal Processing Magazine, IEEE , Volume: 16 , Issue: 5 , Sept. 1999*

[5] Ney, H.; Ortmanns, S.. “Progress in dynamic programming search for LVCSR.”*Proceedings of the IEEE , Volume: 88 , Issue: 8 , Aug. 2000*

[6] Deshmukh, N.; Picone, J. “Methodologies for language modeling and search in continuous speech recognition.” *Southeastcon '95. 'Visualize the Future'., Proceedings., IEEE , 26-29 March 1995*

[7] http://www.icsi.berkely.edu/~dpwe/icsiprivate/asr-10-probs-1997jan.html

[8] Cuntai Guan; Ce Zhu; Yongbin Chen; Zhenya He. “Performance comparison of several speech recognition methods.” *Speech, Image Processing and Neural Networks, 1994. Proceedings, ISSIPNN '94., 1994 International Symposium on , 13-16 April 1994*

[9] Benayed, Y.; Fohr, D.; Haton, J.P.; Chollet, G.” A new keyword spotting approach based on reward function. “*Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on , Volume: 1 , 1-4 July 2003*

[10] Fabian Luis Vargas, Rubem Dutra Ribeiro Fagundes, Daniel Barros Junior.. “An FPGA-Based Viterbi Algorithm Implementation For Speech Recognition System**.”** *IEEE Signal Processing Society International Conference on Acoustics, Speech, and Signal Processing 2001*

[11] Roe, D.B.; Wilpon, J.G.” Whither speech recognition: the next 25 years.” *Communications Magazine, IEEE , Volume: 31 , Issue: 11 , Nov. 1993*

[12] Scharenborg, O.; ten Bosch, L.; Boves, L.”Early recognition of words in continuous speech” *Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE workshop on,30 Nov-3Dec 2003.*

[13] Myers, C.; Rabiner, L.; Rosenberg, A.” An investigation of the use of dynamic time warping for word spotting and connected speech recognition.” *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '80. , Volume: 5 Apr 1980.*

[14] Ishikawa, Y.Nakajima, K.”A real time connected word recognition system Pattern Recognition, 1990.” *Proceedings., 10th International Conference on , Volume: ii , 16-21 June 1990*

[15] Jouvet, D.; Monne, J.; Dubois, D.” A new network-based speaker-independent connected-word recognition system.” *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86. , Volume: 11 , Apr 1986*

[16] van der Walt, C.; Mortimer, B.“The practical application of a continuous speech recognition system.” *Communications and Signal Processing, 1993., Proceedings of the 1993 IEEE South African Symposium on 6 Aug 1993*

[17] Segawa, O.; Takeda, K.; Itakura, F. “Continuous speech recognition without end-point detection.”*Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on , Volume: 1 , 7-1May 2001*

[18] Morgan, N.; Bourlard, H.” Continuous speech recognition.” *Signal Processing Magazine, IEEE, Volume: 12 , Issue: 3 May 1995*

[19] Picone, J. “Continuous speech recognition using hidden Markov models.” *ASSP Magazine, IEEE [see also IEEE Signal Processing Magazine] , Volume: 7 , Issue: 3 , July 1990*

[20] Gopalakrishnan, P.S.; Nahamoo, D.” Models and algorithms for continuous speech recognition: a brief tutorial. “ *Circuits and Systems, 1993., Proceedings of the 36th Midwest Symposium on, 16-18 Aug. 1993*

[21] Murveit, H.; Mankoski, J.; Rabaey, J.; Brodersen, R.; Stoezle, T.; Chen, D.; Narayanaswamy, S.; Yu, R.; Schrupp, P.; Schwartz, R.; Santos, A.”**A** large-vocabulary real-time continuous-speech recognition system.” *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on , 23-26 May 1989*

[22] O'Shaughnessy, D.; Zhishun Li; Farhat, A.; El Meliani, R.; Vergin, R.; Heon, M. “Recent progress in automatic recognition of continuous speech.” *Electrical and Computer Engineering, 1997. IEEE 1997 Canadian Conference on , Volume: 1 , 25-28 May 1997*

[23] Hori, C.; Furui, S, “A new approach to automatic speech summarization.**”** *Multimedia, IEEE Transactions on , Volume: 5 , Issue: 3 , Sept. 2003*

[24] Rabiner, L.R. “Applications of speech recognition in the area of telecommunications.” *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on , 14-17 Dec. 1997*

[25] Liao, L.; Gregory, M.A. “Algorithms for speech classification**,** Signal Processing and Its Applications, “*ISSPA '99. Proceedings of the Fifth International Symposium on , Volume: 2 , 22-25 Aug. 1999*

[26] Huang, X.D.; Jack, M.A. “Performance comparison between semi continuous and discrete hidden Markov models of speech. “*Electronics Letters , Volume: 24 , Issue: 3 , 4 Feb. 1988*

[27] gpdsHMM: A Hidden Markov Model Toolbox In The Matlab Environment: http://www.gpds.ulpgc.es

[28] Comparative Study of Continuous Hidden Markov Models (CHMM) and Artificial Neural Network (ANN) on Speaker Identification System http://www.worldscinet.com

[29] *Schuller, B.; Rigoll, G.; Lang, M.,*”Hidden Markov model-based speech emotion recognition.” *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on , Volume: 1 , 6-9 July 2003*

[30] Rabiner, L.; Juang, B.,”An introduction to hidden Markov models.” *ASSP Magazine, IEEE [see also IEEE Signal Processing Magazine] , Volume: 3 , Issue: 1 , Jan 1986*

[31] Rabiner, L.R.; Wilpon, J.G.; Soong, F.K. “High performance connected digit recognition using hidden Markov models.” *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on , Volume*: 37 , Issue: *8 , Aug. 1989*

[32] Rabiner, L.R.; Wilpon, J.G.; Soong, F.K.”High performance connected digit recognition, using hidden Markov models.” In, *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on , 11-14 April 1988*

[33] Rabiner, L.R., “A tutorial on hidden Markov models and selected applications in speech recognition.” *Proceedings of the IEEE , Volume: 77 , Issue: 2 , Feb. 1989*

[34] Bok-Gue Park; Koon-shik Cho; Jun-Dong Cho, “Low power VLSI architecture of Viterbi scorer for HMM-based isolated word recognition.” *Quality Electronic Design, 2002. Proceedings. International Symposium on , 18-21 March 2002*

[35] Chauhan, S.; Sharma, P.; Singh, H.R.; Mobin, A.; Agrawal, S.S. “Design and development of voice-cum-auto steered robotic wheelchair incorporating reactive fuzzy scheme for anti-collision and auto routing.” *TENCON 2000. Proceedings , Volume: 1 , 24-27 Sept. 2000*

[36] Singh, H.R.; Mobin, A.; Kumar, S.; Chauhan, S.; Agrawal, S.S. “Design and development of voice/joystick operated microcontroller based intelligent motorised wheelchair.” *TENCON 99. Proceedings of the IEEE Region 10 Conference , Volume: 2 , 15-17 Sept. 1999*

[37] Singh, H.R.; Chauhan, S.; Mobin, A.; Agrawal, S.S.”Design and development of voice/tele operated intelligent mobile robot.” *TENCON '97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications'., Proceedings of IEEE , Volume 1 , 2-4 Dec. 1997*

[38] Rockland, R.H.; Reisman, S.“ Voice activated wheelchair controller. “*Bioengineering Conference, 1998. Proceedings of the IEEE 24th Annual Northeast , 9-10 April 1998*

[39] Phil Woodland.” Speech Recognition” Speech and Language Engineering state of the Art” *(Ref No: 1998 / 499) IEEE Colloquium, 19 Nov, 1998.*