CONTINUOUS SPEECH RECOGNITION SYSTEM: A REVIEW

Article · June 2014

DOI: 10.15520/ajcsit.v4i6.3

Pratik Kurzekar
Samanavay Pratisthan's Institute of Knowledge
4 PUBLICATIONS 37 CITATIONS
SEE PROFILE

Ratnadeep R. Deshmukh
Dr. Babasaheb Ambedkar Marathwada University
214 PUBLICATIONS 440 CITATIONS
SEE PROFILE

Dr. Vishal Waghmare
Shri Swami Vivekanand Shikshan Sanstha, Kolhapur
25 PUBLICATIONS 110 CITATIONS
SEE PROFILE

Contents lists available at www.innovativejournal.in

Asian Journal of Computer Science and Information Technology

Journal homepage: http://innovativejournal.in/ajcsit/index.php/ajcsit

ISSN 2249 - 5126

Review Article

CONTINUOUS SPEECH RECOGNITION SYSTEM: A REVIEW

Pratik K. Kurzekar*, Ratnadeep R. Deshmukh, Vishal B. Waghmare, Pukhraj P. Shrishrimal

Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad – 431004 (MS), India.

ARTICLE INFO

Corresponding Author:
Pratik K. Kurzekar
Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad – 431004 (MS), India

DOI: http://dx.doi.org/10.15520/ajcsit.v4i6.3

ABSTRACT

Speech is most common mode of communication between human. Human are trying to develop systems which can understand and accept the command via speech. This paper gives an overview of continuous speech recognition systems developed in different languages. This paper would be helpful for the researchers to find the brief overview of continuous speech recognition systems developed in different languages around the world and the recognition rate achieved by these system.

Keywords: Speech Recognition, Continuous Speech Recognition, Speech Database, MFCC, HMM, HTK, Sphinx.

1. INTRODUCTION

Speech is the most commonly and widely used form of communication between humans. There are various spoken languages which are used throughout the world. The communication among the human being is mostly done by vocally, therefore it is natural for people to expect speech interfaces with computer [1].

Since early 1960's researchers are trying to develop system which can record, interpret and understand human speech. The use of speech for interacting with the computer may help the developing nations as the language technologies can be implemented for the e-governance system.

Speech recognition (SR) means translation of spoken words to the text or commands. Development of Speech recognition systems has attained new heights but robustness and noise tolerant recognition systems are few of the problems which make speech recognition systems inconvenient to use [2]. Many Research projects have been completed and currently in progress around the world for the development of robust speech recognition systems.

The paper presents the review of the continuous speech recognition systems.

2. CLASSIFICATION OF SPEECH RECOGNITION SYSTEMS:

The speech recognition systems can be classified in different type depending upon different classes. The speech recognition system can be classified on the basis of type of utterances, vocabulary size and speaker dependency.

2.1 Classification on the basis of utterances:

2.1.1 Isolated Words:

Isolated word recognizers usually require each utterance to have quite on both sides of the sample window. It accepts single words or single utterance at a time. These systems have "Listen/Not-Listen states", where they require the speaker to wait between utterances.

2.1.2 Connected Words:

Connected word systems are similar to isolated words, but it allows separate utterances to be 'run-together' with a minimal pause between them.

2.1.3 Continuous Speech:

Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. Recognizers with continuous speech capabilities are some of the most difficult to create because they utilize special methods to determine the utterance boundaries.

2.1.4 Spontaneous Speech:

At a basic level, it can be thought of as a speech that is natural sounding and not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, "ums" and "ahs", and even slight stutters [3].

2.2 Classification on the basis of Vocabulary size:

2.2.1 Small Vocabulary:

The speech recognition systems which can recognize limited and given set of vocabulary (i.e. few hundred words or sentences) are known as limited vocabulary speech recognition system.

2.2.2 Medium Vocabulary:

The speech recognition system which can recognize a considerable number of vocabularies (i.e. few from few hundred up to few thousands of words or sentences) such systems are known as medium vocabulary speech recognition system.

2.2.3 Large Vocabulary:

The speech recognition system which can recognize a large number of vocabularies (i.e. more than few thousands of words or sentences) such systems are known as large vocabulary speech recognition system [4].

2.3 Classification on the basis of Speaker mode:

2.3.1 Speaker Dependent:

Speaker dependent speech recognition systems learn the unique characteristics of a single person's voice, in a way similar to voice recognition. The system is trained on the basis of the training dataset and it may use templates.

2.3.2 Speaker Independent:

In speaker-independent speech recognition systems there is no training of the system to recognize a particular speaker and so the stored word patterns must be representative of the collection of speakers expected to use the system. The word templates are derived by first obtaining a large number of sample patterns from a cross-section of talkers of different sex, age-group and dialect, and then clustering these to form a representative pattern for each word.

2.3.3 Speaker Adaptive:

In speaker adaptive speech recognition systems the uses the speaker dependent data and adapt to the best suited speaker to recognize the speech and decrease the error rate by adaption [5].

3. EXISTING CONTINUOUS SPEECH RECOGNITION SYSTEM

3.1 Continuous speech recognition system for phonetic transcription:

A continuous speech recognition system for phonetic transcription was developed at AT&T Bell laboratories. Hidden Markov Model in conjunction with an appropriate dynamic programming algorithm was used to do the acoustic to phonetic mapping. The test was performed on DARPA data that has been filtered and down sampled to a 4 kHz bandwidth. The training set consisted of 3267 sentences spoken by 109 different speakers. Two test sets each consisting of 300 sentences which were spoken by 10 speakers. The third set consisted of 54 sentences spoken by one of speaker which were recorded using similar equipment that was used for DARPA data. The researchers attained 88% correct word recognition with 3% insertion yielding a word accuracy of 85%. The researchers resynthesized the phonetic transcription accuracy by directly accessing the phonetic transcription. In a few informal listing tests, the word intelligibility rate was judged to be approximately 75% [6].

3.2 The Dragon Continuous Speech Recognition System:

A 1000 word Continuous speech recognition system was developed at Dragon systems. The system was designed for large vocabulary natural language task. The developed system was a real time continuous speech recognition system. The input speech signal was sampled at 12 KHz and low pass filtered at 6 KHz. The researchers used HMM based speech recognizer. The task of the application consisted o recognizing mammography reports. The system was implemented on intel 486 PC. The research got a word error rate of 3.4 % and sentence error rate of 19.5% while calculating the recognition performance [7].

3.3 Continuous speech recognition system for Korean language:

A continuous speech recognition system for Korean language was developed using phone-based semi-continuous hidden Markov model (SCHMM) method at University of Korea. The system comprised of three features, first an embedded bootstrapping training method, second was the HMM parameter estimation and third a between-word modelling techniques in word boundaries. The task domain of the system contains 244 words including digits, English alphabet. Speech database for simulation consisted of two parts, one was the word data which consisted of the utterances pronounced by 51 speakers and other was the set of 5610 different sentences pronounced by the 51 speakers. For the phone models, DHMM and SCHMM were used and a model topology with 3 states and 8 transitions including skip transition was defined. For this the researchers used four feature vectors; LPC cepstrum with a bandpass lifter, delta cepstrum, delta-delta cepstrum and energy. For HMM training, two training stages were applied to the word and sentence data. In speaker independent experiment, the discrete HMM (DHMM) method was applied and it showed 89.7% word accuracy and the SCHMM method showed 89.0% [8].

3.4 The HMM based Continuous Speech Recognition System:

A two stage continuous speech recognition system using HMM based on phonological features was developed at University of Edinburg, United Kingdom. The researchers worked on three phonological features system :(1) the Sound Pattern of English (SPE) system which used binary features, (2) a multi-valued (MV) feature system which used traditional phonetic categories like manner, place, etc., and (3) a Government Phonology (GP) which used a set of structured primes. All the experiments were carried on the TIMIT speaker-independent database with the accuracy for a single feature ranging from 86% and 93%. The developed system gave higher phone recognition accuracy of 63.5% [9].

3.5 Continuous speech recognition system using phonologically-constrained morphological analysis:

A Continuous speech recognition system for 100 million British English words was developed using phonologically-constrained morphological analysis at Department of Phonetic and Linguistic, University College London. The corpus contained 4124 files out of that 3209 are written and 815 are transcribed speech. The researchers used 80 million words as training material and 20 million words comprised of written and spoken text. The researchers selected test sentences from BNC test set and LOB corpus. From BNC, 100 sentences of 1786 words were selected which were referred as BNC1 and 100 sentences of 2002 words from LOB corpus were referred as LOB1. These sentences were recorded by the male speaker of southern British English in an anechoic environment with the sampling frequency 16 kHz. Word vocabularies of 20000, 40000 and 65000 were chosen from the list of most frequent word in 10 million word sample of the training data set. These three vocabularies tested with 200 test sentences from LOB and BNC. The word pronunciation for 20, 40 and 65k vocabularies are mapped from a dictionary of British English pronunciation. Language model were build for both word and morph vocabularies for 20, 40 and 65k and 30 models were build in total. The best result showed 16% relative reduction in word error rate [10].

3.6 Across-word model continuous speech recognition system:

An across-word model continuous speech recognition system was developed at University of Technology, Aachen, Germany in which they three different experimental setup in which three different corpora were used for performing the set of experiments. All the experiments were carried out on Pentium III 600MHz PC.This system was used for the experiment on VERBMOBIL II corpus and which characterized of 16 cepstral coefficient with first and second derivative of the energy, 10 millisecond frame shift, a 3-state HMM triphone models with skip, forward and loop transition; gender independent Gaussian mixture and speaker track normalisation by vocal tract normalisation. By using developed model the word error rate was reduced from 23.1% to 21% and RTF was increased by the factor of 2.2 [11].

3.7 Continuous speech recognition system for polysyllabic word:

A continuous speech recognition system for polysyllabic word was developed at Radboud University Nijmegen, Netherlands. For the developed speech recognition system 1463 polysyllabic words were selected for experiment. For 885 utterances the system showed an accuracy of 64% for the polysyllabic words at the end of the utterance. Automatic speech recognition and SpeM system was used for the recognition of the words. It used two types of predictors: first type of predictor was related with absolute and relative values of the word activation and second type of predictor was related to the number of phones that were present till the end of the word. The system gave 81.1% accurate result when local words activation was used to identify word before its last phone was available and the system gave 64.1% of those words which were already recognised as one phone after the uniqueness point [12].

3.8 Gini SVM continuous speech recognition system:

A Ginisupport vector machines continuous speech recognition system was developed by incorporating segmental minimum Bayes risk decoding at Fair Isaac Corporation, San Diego, USA. It used lattice cutting to convert the Automatic Speech recognition search space into the sequence of smaller recognition problem. It was found that, on a small vocabulary recognition task, the use of GiniSVm can improve the performance of a well trained system based on hidden markov model under the Maximum Mutual Information (MMI) criterion. The developed system used GiniSVM toolkit to train SVMs for the 50 dominant confusion pairs extracted from the lattice generated by the MMI system. The word error rate was reduced with respect to MMI baseline from 9.07% to 8.00% [13].

3.9 Continuous speech recognition system for Indian Languages:

A continuous speech recognition system for Indian Languages was developed at IIIT Hyderabad. A speech database of 560 speakers in three different Indian languages i.e. Tamil, Telugu and Marathi was developed by using HMM technique. An acoustic model was developed on the developed databases by using Sphinx 2 speech tool kit. For database development, they categorized the speakers into different categories according to age and gender. While collecting the samples care was taken for the minimal noise and pronunciation mistakes by the speaker. The performance of the system was evaluated by taking three iterations. The developed system achieved word error rate (WER) of 23.2%, 20.2%, 28% for Marathi, Tamil and Telugu languages respectively [14].

3.10 Large vocabulary continuous speech recognition system for spectral representation:

A large vocabulary continuous speech recognition system was developed at Centre for Speech Technology Research, School of Informatics, and University of Edinburgh, United Kingdom. IT was developed to investigate the combination of complementary acoustic feature. The acoustic features were obtained by using a pitch-synchronous analysis in combination with the conventional features. The baseline acoustic models were trained on conventional MFCC. The MFCC used 25 milliseconds hamming window with a shift of 10 milliseconds. The developed system gave a relative reduction in word error rate (WER) of 3.2% [15].

3.11 Sparse code continuous speech recognition system:

A sparse code continuous speech recognition system was developed at University of Pretoria, South Africa. The developed system used TIGIDITS database which consisted of 8623 utterances. The system transformed the waveform into spectrogram and a sparse code for the spectrogram was found by the means of a linear generative model. The data was recorded at a sampling frequency of 24 kHz. The system used iterative subset selection algorithm with quadratic programming to find a sparse code in reasonable time. At the used parameters, a word error rate of 19% was achived which when compared with a system based on HMM have a word error rate of 15% using the same parameters [16].

3.12 Continuous speech recognition system for Arabian languages:

A speaker-independent continuous automatic speech recognition system based on a phonetically rich and balanced speech corpus was developed at University of Malaya, Malaysia. The speech corpus contained total 415 sentences recorded by 40 native Arabic speakers (20 male and 20 female) from 11 different Arab countries representing the three major regions (Levant, Gulf, and Africa) in the Arab world. Sphinx tools, and the Cambridge HTK tools were used to develop the system. The speech engine used 3-emitting state Hidden Markov Models (HMM) for tri-phone based acoustic models. The language model consisted of bi-grams and tri-grams. For similar speakers with different sentences, the system obtained a word recognition accuracy of 92.67% and 93.88% and a word error rate (WER) of 11.27% and 10.07% with and without diacritical marks, respectively. For different speakers with similar sentences, the system obtained a word recognition accuracy of 95.92% and 96.29%, and a WER of 5.78%, and 5.45% with and without diacritical marks, respectively [17].

4. COMPARISON BETWEEN THE EXISTING SPEECH RECOGNITION SYSTEM:

The overall paper describes the few of the continuous speech recognition system developed around the world. In section 3 we have described briefly the various continuous speech recognition system developed at different places using different techniques and the accuracy rate achieved. The speech recognition systems described in previous section are compared in this section on the basis of what features/parameters were used, language, speech database used, size of the database, techniques used along with used tool kits/systems and the accuracy of the system which in some cases are word error rate. Table 1 shows comparison of the different speech recognition systems studied for the paper. When we compare all the 12 speech recognition systems for the study we observed that most of the research is done for English language. Out of the studied 12 speech recognition systems 8 are for English language and remaining four are developed for different languages.

For developing the continuous speech recognition system for different languages 7 systems were based on previously developed standard speech database and for 5 systems the speech database was developed during the time span of the work. For most of the developed speech recognition system HMM technique was used with HTK or Sphinx toolkit. Few of the systems were developed using other techniques like MFCC, PCMA, GiniSVM and NICO. The speech recognition systems developed by implementing HMM with combination of other techniques showed higher accuracy rate and considerable reduction was observed in word error rate.

The comparative study shows that a lot of work has been carried out for English language the work for languages spoken in the developing nations is far less. The above study reveals that the researchers in the developing nations should try to work for their languages; which may help in development of language technologies which can ultimately benefit their society.

Table 1: Comparison between the existing speech recognition systems for continuous speech

Sr. No.	Speech Recognition System	Language	Database Used	Database Size	Techniques and Toolkits/ systems	Recognition Rate
1	Phonetic transcription	English	DARPA	3267 Sentences	HMM	85%
2	Speech Recognition	English	Developed during the research	1000 Words	HMM on Intel 486 PC	(WER) 3.4%
3	Speech Recognition for Korean Language	Korean	Developed during the research	244 Words and 5610 Sentences	HMM	89%
4	HMM based Speech Recognition	Northern American	TIMIT	3648 training utterances and 1344 test utterances	MFCC with NICO	63.5%
5	Speech Recognition	British English	BNC2A-J LOB1	1000 Sentences or 16522 Words	PCMA	WER reduced by 2.5%
6	Across word Model Speech Recognition	English	VERBMOBIL II		HMM	WER reduced by 2.1%
7	Polysyllabic Word Speech Recognition	Dutch	VIOS	1463 Polysyllabic words in 885 utterances	HMM with HTK	81.1%
8	GiniSVM Speech Recognition	English	Developed during the research	36 Words (26 letters and 10 digits)46730 training and 3112 testing utterances	GiniSVM	WER reduced by 5%
9	Speech Recognition for Indian Language	Tamil, Telugu and Marathi	Developed during the research	Marathi-155541 Sentences Tamil-303537 Sentences Telugu-444292 Sentences	HMM with Sphinx - 2	WER for Marathi-23.2% Tamil-20.2% Telugu-28%
10	Large Vocabulary Speech Recognition	English	WSJCAM0	50000 Words	MFCC	WER reduced by 3.2%
11	Sparse Code Speech Recognition	English	TIDIGITS	8623 utterances and contains 28,329 words	HMM with HTK	WER achieved up to 15%
12	Speech Recognition for Arabic Language	Arabian	Developed during the research	415 Sentence	HMM with Sphinx and HTK	WER is 5.78% and 5.45% with and without diacritical marks

CONCLUSION

In this paper we have studied few continuous speech recognition systems developed in different languages around the world. In the study we observed that most of the work is done for English language and little work is done for the languages of developing nations. It was interesting to find that most of the system developed for continuous speech used HMM with combination of other toolkits and the accuracy of developed systems showed higher recognition rate with reduction in word error rate up to 5%. It was also observed that very few attempts has been tried to develop the system with large vocabulary.

To use speech as an interface for computer there is need to develop continuous speech systems based on large vocabulary and increase the accuracy rate of such systems. The system developed should be able to handle background noise efficiently if we want to implement these systems in real life application. In this study we have concentrated to study few continuous speech recognition systems. This study will motivate the people working in the field of speech recognition system in developing countries to concentrate on development of large vocabulary speech database and recognition system for the developed speech databases which would be robust and that can handle background noise efficiently.

ACKNOWLEDGEMENT

This work is supported by University Grants Commission under the scheme Major Research Project entitled as "Development of Database and Automatic Recognition System for Continuous Marathi Spoken Language for agriculture purpose in Marathwada Region". The authors would also like to thank the University Authorities for providing the infrastructure to carry out the research.

REFERENCES

[1] Pukhraj Shrishrimal, R.R. Deshmukh, and Vishal Waghmare “Indian Language Speech Database: A Review”. *International Journal of Computer Application (IJCA)*, Vol 47, No. 5, pp. 17-21, 2012.

[2] Chao Huang, Eric Chang, Tao Chen “Accent Issues in Large Vocabulary Continuous Speech Recognition (LVCSR)”, *Microsoft Research China*, MSR-TR-2001-69, pp.1-27

[3] Santosh K. Gaikwad, Bharti Gawli, Pravin Yannawar, “A Review of Speech Recognition Technique”, *International Journal of Computer Applications (0975–8887)* Volume 10, No.3, November 2010.

[4] [4] M. A. Anusuya, S. K. Katti, “Speech Recognition by Machine: A Review”, *International Journal of Computer Science and Information Security (IJCSIS)*, Vol. 6, No. 3, pp. 181-205, 2009.

[5] X. D. Huang, "A Study on Speaker - Adaptive Speech Recognition", *Proc. DARPA Workshop on Speech and Natural Language*, pp. 278-283, February 1991.

[6] S. E. Leninson, A. Ljolie, L.G. Miller “Continuous speech Recognition from a Phonetic Transcription” *Acoustics, Speech, and Signal Processing*, vol.1 pp. 93 – 96, Apr 1990.

[7] Paul Bamberg, Yen-lu Chow, Laurence Gillick, Robert Roth and Dean Sturtevant, "The Dragon Continuous Speech Recognition System: A Real-Time Implementation", *Proceedings of DARPA Speech and Natural Language Workshop*, Hidden Valley, Pennsylvania, pp. 78-81, June 1990.

[8] H. R. Kim , K. W. Hwang , N. Y. Han and Y. M. Ahn "Korean Continuous Speech Recognition System Using Context-Dependent Phone SCHMMs", *Proceedings of the Fifth Australian International Conference on Speech Science and Technology*, vol. II, pp.694 -699,1994

[9] Simon King, Paul Taylor “Detection of phonological features in continuous speech using neural networks” *Computer Speech & Language*, Volume 14, Issue 4, October 2000, Pages 333–353.

[10] Mark Huckvale and Alex Chengyu Fang "Using Phonologically-Constrained Morphological Analysis in Continuous Speech Recognition", *Computer Speech and Language*, vol. 16, pp.165-181, 2002.

[11] Achim Sixtus and Hermann Ney, "From within-word model search to across-word model search in large vocabulary continuous speech recognition", *Computer Speech and Language*, Vol 16, 2002, pp.245–27.

[12] Odette Scharenborg, Louis ten Bosch, Lou Boves "Early recognition of polysyllabic words in continuous speech" *Computer Speech and Language*, Vol 21, pp. 54–71, 2007.

[13] Veera Venkataramani, Shantanu Chakrabartty, William Byrne "Ginisupport vector machines for segmental minimum Bayes risk decoding of continuous speech", *Computer Speech and Language*, Vol 21, pp. 423–442, 2007.

[14] Gopalakrishna Anumanchipalli, Rahul Chitturi, Sachin Joshi, Rohit Kumar, Satinder Pal Singh R.N.V. Sitaram, S P Kishore “Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems” *International Institute of Information Technology*, Hyderabad, India July 2007.

[15] Giulia Garau, Steve Renals "Combining Spectral Representations for Large-Vocabulary Continuous Speech Recognition", *IEEE transactions on Audio, Speech, and Language Processing*, Vol. 16, no. 3, March 2008

[16] W.J. Smit, E. Barnard "Continuous speech recognition with sparse coding", *Computer Speech and Language*, Vol 23, pp. 200–219, 2009.

[17] Mohammad Abushariah, Raja Ainon, Roziati Zainuddin, Moustafa Elshafei, and Othman Khalifa “Arabic Speaker-Independent Continuous Automatic Speech Recognition Based on a Phonetically Rich and Balanced Speech Corpus” *The International Arab Journal of Information Technology*, Vol. 9, No. 1, January 2012.

How to cite this article: Kurzekar, P., Deshmukh, R., & Pukhraj P. Shrishrimal, V. (2014). Continuous Speech Recognition System A Review. *Asian Journal Of Computer Science And Information Technology, 4*(6). doi:10.15520/ajcsit.v4i6.3