A Comprehensive View of Automatic Speech Recognition System - A Systematic Literature Review

Yogesh Kumar
Ph.D. Research Scholar
Punjabi University, Patiala
Punjab, India
yogesh.arora10744@gmail.com

Dr. Navdeep Singh
Associate Professor
Mata Gujri College
Sri Fatehgarh Sahib, India
navdeep_jaggi@yahoo.com

***Abstract*— Humans have always attempted to correspond with objects in a natural language. Communications have been the essential feature of human life, a powerful tool for sharing and building the information that is passed from generation to generation. Among speech processing problems, automatic speech recognition mechanisms of converting the recorded speech signals into the text are one of the most challenging tasks. The signals are typically processed in a digital representation, so speech processing can be observed as a particular case of digital signal processing. The overall performance of an automatic speech recognition system greatly depends upon the acoustic modeling. Hence, building a precise and robust acoustic model holds the key to a suitable recognition performance. People have used different methods for automated speech recognition system. For recognizing the speech people always choose the English language in the majority of the research and implementation but very less work is done in other languages. Our analysis presents the study of the different speech recognition systems present in Indian and foreign languages in the systematic review of speech recognition paper. This paper gives the review of different aspects related to Automatic Speech recognition. We have elaborated the recent advancement in the speech recognition system, robust method for the development of an automatic speech recognition system and application of automatic speech recognition system in different fields.**

Keywords—Acoustic Model; MFCC; Sphinx; Hidden Markov Model; Word Error Rate

I. Introduction

Speech communication has become a dominant model of information exchange and human social bonding. In human machine interaction reflection is find out by spoken language communication is preferred by human along with person to person interaction. In particular capability of speaking naturally and responding properly to spoken languages has intrigued century's scientists and engineers and human behavior is mimics using designing of machine. In this field acoustic and electronic engineer is pioneered by Homer W. Dudley. First electronic voice synthesizer is created in 1930s by Bell Labs and during world war II secure voice transmission is sent by method development. Entire system development from scratch problem has been faced by people or researchers as difficulty of core speech recognition research. In [2] authors have given HTK as a open source speech recognition systems.

ISIP [3], AVCSR [4] and previous versions of the Sphinx systems [5, 6]. The existing systems are optimized for single approach to speech system design that creates a barrier to future research which is not original motive of system. Pluggable and modular framework is first Sphinx-4 that includes existing systems design patterns and provides flexibility to support emerging research areas. In case of modular framework dedicated to specific tasks there is separate components and at run time it is easily replaced by pluggable module. There are lots of other modules includes by Sphinx-4 that provides working system to researchers and implements state of the art speech recognition techniques. From last decades ASR filed is explored and working on providing an eye free and hands-free interfaces to devices. The objective of ASR is to capture an acoustic signal of speech and determine the words that were spoken by pattern matching. To do this, a set of acoustic and language models have to be stored in a computer database that represents the actual patterns.

These language models are then compared with captured signals. Speech is a vocalized form of human communication and there is need of screen, mouse and keyboard as interface to communicate with machine using software. In [7] authors have given Automatic Speech Recognition (ASR) system as a software interface form of an alternative form of hardware interfaces. Capture by telephone, microphone capture a input to ASR task is utterance of speech signal. Authors [8] have converted it into text sequence close to spoken data. Due to environmental disturbances and human beings speaking styles are main difficulties of ASR system implementation. The speech signal transformation from device independent text message is the main aim of ASR system in an efficient and accurate manner. Telephone or microphone captured acoustic signals is converted into set of words using speech recognition machine. Anyone speech can be recognized and there is need of large training data to make voice machine independent. The natural communication between machine and man is done using ASR.

This complete paper is divided into different sections. The first sections brief introduction to spontaneous speech recognition and why there is need of it. The architecture of speech recognition system is given in second section. The third section gives brief about ASR system for various languages. Fourth section contains review on different recent advancements done in speech recognition system that includes different subsections for different approaches and different aspects related to it. Fifth section includes Study and analysis of ASR Systems. Sixth section covers review on Feature extraction mechanism of Speech Recognition System.

II. Architecture of Speech Recognition System

Acoustic Modeling: An acoustic model is a systemized that is enclosed by arithmetical demonstrations for single discrete significance that helps in making an utterance. Phoneme related tag is assigned by each of arithmetical demonstrations and sub-word units guarded phones like set of sounds are used to collect speech from lexis sounds. In verbal communication every phoneme arithmetical depictions generation and wide corpus of words are captivating for constructing a sound model. Then front end supplied incoming features against HMM and unit of speech is depicted by acoustic model module.

Linguistic Models: A language model (LM) is a class of past statistics regarding the speech. This statistic is autonomous of a speech to be acquired. Information concerning about a speech can be uttered in terms of sequences that are probable or how commonly they emerge. Word-level language structure is provided by the linguist of the Language Model module, which can be depicted by any number of pluggable executions. These executions normally fall into one of two categories which are stochastic N-Gram models and graph-driven grammars. In previous n-1 word observation given words probabilities is imparted by stochastic N-Gram models. On other hand directed word graph is depicted by graph driven grammar in which each arc personify word transition probability and each node personifies a single word. A variety of formats are supported by the Sphinx-4 Language Model, including the following.

Simple Word List Grammar: It describes a grammar built over the list of words. An optional parameter is there which helps in explaining whether the grammar “loops” or not. For isolated word grammar can be constructed if its not looping on other hand it will be used to build trivial connected word recognition if it loops.

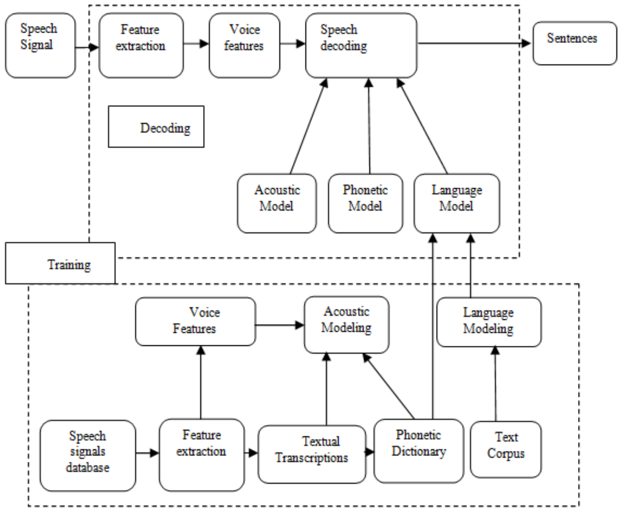

Fig. 1. Architecture of Speech Recognition System [8]

JSGF Grammar: It maintains the Java Speech API Grammar Format (JSGF) [9], which explains a vendor-independent Unicode representation of grammars, BNF-style and platform-independent.

LM Grammar: A statistical language based model is used to define a grammar. Approximately 1000 words bi-gram and smaller unigram grammars are used by LM grammars and makes one grammar node per word.

FSTGrammar: It provides FST (finite-state transducer) in the ARPA FST grammar format [10, 11].

SimpleN-GramModel: In the ARPA format a ASCII N-Gram models is enabled using it. Ni attempt is made it for memory usage optimization.

Large Trigram Model: It is developed by the CMU Cambridge Statistical Language Modeling Toolkit [11] that provides true N-Gram models. The Large Trigram Model includes memory storage, enabling it to work with huge files of 100MB or more.

III. ASR System for Various Languages

Various researchers have used different approaches for their work. Each approach has their own advantages and disadvantages. This work helps in getting the idea about different approaches that can be used and which will be the best one in getting better results. Various authors have worked on Marathi, Malayalam, Kannada, Bangla, Punjabi, Gujurati, etc languages. Some has proposed neural network technique. In 2015, some authors have used a new algorithm which is having different features like format frequencies, energy measure and zero crossing rates. Some have used Hidden Markov model and Mel frequency cepstral coefficients and for gujurati language Hidden Markov Model Toolkit is used for measuring the performance and various error parameters by Tailor, J.H. et al. [33] (2016). For Punjabi language used acoustic models-based tri-phone that is designed for Punjabi language continuous speech has been used.

Along with Indian languages various researchers have worked on foreign languages like Chinese, Romanian, Hungarian, Japanese, English, Thai, etc. Some have worked on designing speech recognition system model for various languages by gathering texts from Web. Burileanu, C. et al. (2010) [34] have described speech recognition architecture for this language (based on components in the “Hidden Markov Modeling Toolkit” – HTK) with two aspects i.e. training and decoding. Mihajlik, P. et al. (2010) [35] have represented various techniques for acoustic modeling and morphological for spontaneous Hungarian large-vocabulary continuous speech recognition (LVCSR) task. Data sparseness gets increase using effective Hungarian language that come by small training databases. One author has described a Paraphrase spontaneous speech to written style sentences integrated framework. Other have used Hidden Markov Model (HMM) that help in providing a most appropriate and reliable way of recognizing speech.

IV. Recent Advancements in Speech Recognition System

Speech is transformed to text using speech recognition systems and some major advances are made by voice recognition technology. In past it was a problem to use speech technology but now Dragon assistant, Google voice and Siri like speech technology has become a very popular tool. A lot of advancement and research has been done by Nuance, Intel, Apple, Microsoft and Google like big companies. In speech recognition major task is performed by spontaneous speech recognition. An accuracy of 90% has been achieved in news broadcast by GMM-HMM based system. But when the same system is applied to recognize spontaneous speech than achieving that much accuracy has become a problem. Various advancements in speech recognition system are discussed in this section.

A. Deep Learning Approach for Automatic Speech Recognition System

In 2012, authors have introduced a model called context-dependent (CD) for (LVSR) large-vocabulary speech recognition provided advances for phone recognition using the deep belief network. To train the Deep Neural Network to provide assignment over the senones, deep neural network hidden Markov model (DNN-HMM) are pre-trained. The deep belief network pre-training algorithm was one of the robust methods to reduce generalization error and to produce an efficient result in the deep neural networks. George, E., et.al, also discussed some key points of the model and explained procedure to apply CD-DNN-HMMs to LVSR. It has been seen that DNN modules were well suited to matched training and testing data and to improve performance in challenging environments. Then in 2015, Narayanan, A. et al. [12] introduced a concept that combines the acoustic modeling and separation by using joint adaptive training. DNN was used to implement the modules for acoustic modeling and speech separation. Unification was possible by using additional hidden layers with fixed weights and correct network architecture it was possible. By using CHiME-2 medium-large vocabulary ASR task and Log Mel-spectral features the improved error rates of an independently trained ratio masking frontend was 10.9%. The error rates of jointly trained system were 14.4%. New experiments were also done by the Narayanan, A. et al [12]. to increase the standard Log Mel-features such as noise and speech estimates from the separation module and the standard feature set used for IRM estimation. After that, best system improvement was 15.4%, there was 4.6 % of improvement over the next best result in this corpus. After that Fohr, D. et al. (2017) [13] introduced different architectures for DNN-based models that are very useful to decrease the word error rate as compared to a classical system.

B. Emotion Recognition using ASR System

Emotion recognition is an emerging research field that is finding lots of applications in recent days. A brief review based on Emotion Recognition using ASR System is discussed below.

In 2013, Hendy, N. et al [14]. focused on recognition of emotions using speech signal and formants, energy contours as well as spectral, statistics of pitch, perceptual and temporal features, jitter, and shimmer were the main extracted features. (ANN) The Artificial Neural Networks were selected, and the main objective was to select the accurate, fast and robust ANN classifier for real life applications and many experiments were done to check the success rate of ANN. It was shown that 85% of success rate or even more could be obtained by using 7 unique emotions in the database without increasing the system collusion and the computational time.

Then in 2014, Han, k. et al. [15] has also developed DNN to check the emotion level for each speech segment in an utterance. The main objective of the work was to build up an utterance level feature from segment level computations. To construct an ELM (extreme learning machine) for the utterance to understand the emotions was also the main focus. Their first focused was on making an emotion state probability distribution using DNNs and then secondly by using emotion state probability distribution utterance-level features were extracted. ELM was used for utterance-level features. ELM is a simple and effective single-hidden-layer neural network that is used to extract the utterance-level emotions. The results showed that by using ELM there were an improvement and 20% of accuracy were achieved as compared to the state of-the-art approaches.

Before Han, k. et al. [15], in 2009, He, L. et al. [16] described an automatic stress recognition based on the analysis of acoustic speech. Then non-linear Teager energy operator (TEO) based novel feature extraction approach has been used that is computed within discrete wavelet transform bands, wavelet packet bands and critical bands. They have used probabilistic and multilayer perceptron neural network for segmentation process. SUSAS database of 15 speakers actual stress is used for checking the efficiency of segmentation. The condition used was neutral, low stress and high stress and results show that TEO parameters used within perceptual wavelet packet bands gives best performance. The speech recognition system was explained in 2016 by Patadia, J. et al. [17]. They have also given review on existing features of the system and analyzed that any device can be used to detect the emotion identification. Emotion identification is very useful for human life as it helps in making intelligent system that supports human life in a very effective way.

C. Robust Method for the Development of ASR System

The various robust methods of automatic speech recognition system has been proposed by different authors each have their own pros and cons. The brief about different system is given in this section. The robust architecture and modeling technique is proposed in 2003 by Furui, S. [18] for understanding automatic speech. For language and acoustic methods unsupervised adjustment language and robust acoustic models were used. Then in 2008 other techniques are described by Furtuna, T.F. [19] for developing an algorithm to understand the isolated words in a speech recognition containing words. There is need of comparison between dictionary and entry signal words for recognition purpose and by dynamic algorithm comparison problems gets resolved. These algorithms were called as Dynamic Time Warping.

Then in 2005 WKER is proposed by Nanjo, H. et al. [20] WKER stands for the weighted keyword error rate which provides total errors weight from the received information. The work first described that this measure was effective for predicting the key sentence indexing of oral presentations performance and then made a decoding method to decrease WKER based on Minimum Bayes-Risk (MBR) framework and mentioned the decoding technique that worked for maximizing WKER and key sentence indexing.

D. Development of High-Performance Speech Recognition Systems

In communication between people, speech is considered as most important source. Real-time speech-to-text-conversion means transferring spoken words into written transcript (almost) concurrently. Various researchers have worked on it. From both horizontal and vertical point of view spoken language method growth is discussed by 2000 Juang, B.H. et al [21]. They have given a introduction of different language related problems solution using statistical techniques that are implemented to learn directly from speech signal structural regularities and information. Then Cui, X. et al. (2008) [22] have given a growth on IBM Iraqi/English speech to speech translation system type for DARPA transact project. In their work they have given the details of language and acoustic modeling that gives reduction in noise robustness and high recognition rate.

Silber, V. et al. (2014) [23] introduced the alternative way to handle the challenges of an ASR system in which main focus was on keyword recognition instead of decreasing word error rates (WER). The work focused on measuring the performance of two Hebrew ASR systems. For these systems academic lectures and audio books 40 min recording set is used then video lectures stenographer recording is used for comparison purpose. Over key phrases results keyword recognition gives advantage in keyness tests and both engines get exceeded by stenographer records. ASR can have satisfying accuracy level using proposed approach that gives 78% of keyword recognition which makes it suitable for searching web audio/video content.

V. Study and Analysis of ASR Systems

This section will give review of work done by various researchers in the field of ASR systems analysis.

Table 1: Study and analysis of ASR Systems [24, 25, 26, 27, 28, 29]

Sr. No.	*Citation*	*Research Contributions*
1	Saini, P. et al. (2013)	The author introduced better accuracy techniques for speech recognition in which the center of interest was to find out the best solution for the ASR.
2	Kaladharan, N. et al. (2016)	The author explored a method of speech recognition that involved different recognition rate with improved classification activities. This method deals with feature extraction, different classes of speech recognition and speech classification methods.
3	Vijayalakshmi, A. et al. (2015)	The author introduced an automated speech recognition method where the main focus was to identify speech and transform into text.
4	Kewatkar, S. et al. (2015)	The author developed an effective speech recognition technique that can recognize the speech faster, accurately, effectively. Attentive observations were required to overcome the issues in speech recognition method. These issues were different speech classes and their representation, feature extraction methods, database and performance analysis.
5	Karpagavalli S. et al. (2016)	The author described an automatic speech recognition in which main focus is to convert speech into a sequence of words by using a computer program. This technique deals with speech issues, tools, speech methodologies, parameterization and applications.
6	Kumar, A. et al. (2016)	The author described different methods used for making ASR models. The speech method was studied from audio recordings and transcripts were designed by taking recording as audio and their text transcription. Software’s were used to design a statistical representation of the sounds that deals with the creation of every word.
7	Badyal, I. et al. (2015)	The author discussed new techniques for speech recognition systems. Research discussed the various speech recognition techniques that represent the enhancement in the field to help provide a technological perspective of the progress made in the field. Moreover, it also discussed the fundamental principles and techniques of Speech Recognition to understand the basic design needed to develop a technique.
8	Gupta, S. et al. (2014)	The author proposed major methods with their ability of Feature extraction and Feature matching. A brief review showed that MFCC was mostly used for feature Extraction and VQ was best over DTW.
9	Ghai, W. et al. (2012)	The author proposed ASRs advancement of different languages and the technological perspective of automatic speech recognition in countries like China, Russian, Portuguese, Spain, Saudi Arab, Vietnam, Japan, UK, Sri-Lanka, Philippines, Algeria and India. Using ANN, mathematical models of the low-level circuits in the human brainto enhance speech-recognition through a model known as the ANN-Hidden Markov Model (ANN-HMM) played an important role for large-vocabulary speech recognition systems.

VI. Feature Extraction Mechanism of Speech Recognition System

A. Supervised Training Method and Extraction Process

Building ASR systems require the availability of speech databases with accurate orthographic transcriptions. A brief review is discussed below.

A method having two parts, i.e. sentence extraction and sentence compaction was introduced. On the basis of words confidence measures and amount of data important sentences was analyzed and sentence computation method was also used for compressing set of extracting sentences. Word set was selected for executing a sentence compaction that result in summarization score improvement. It also includes word concatenation probability and word strings linguistic likelihood and summary is made by combining the selected words. Spontaneous presentation is summarized for declaring proposed technique effectiveness. For lightly supervised training of acoustic model a novel data selection technique was described by Li, S., et al. (2015) [30]. It is not a faithful transcript in exploiting massive data with closed caption texts. Firslty they have used baseline system for creating ASR hypothesis and closed caption text sequence. Then trained and implemented a set of dedicated classifier that helps in choosing correct one and reject other. The result shows that classifier is able to effectively filter usable information for acoustic model training. They have not tuned classifier to threshold parameters. Use of baseline system helps in achieving improvement in accuracy of ASR. They have compared a proposed techniques with conventional method of confidence measure scores and simple matching based lightly supervised training. ITS and ASR phonetic variation approaches are described by Schotz, S. (2002) [31] that classify speech paralinguistic and linguistic phenomena. They have also given attempted paralinguistic phonetic attempted and variation solutions issues. Speech researchers are more attracted towards paralinguistic nature prosodic and phonetic variation that is considered as possible solution in text to speech synthesis and ASR systems. In order to learn sub word lexicon optimized for given task is done by new technique introduced by Parada, C. et al. (2011) [32]. Hybrid model output dependent vocabulary word detection work is introduced. The proposed approach is able to achieve 6.3% of sub word lexicon minimized error and at false alarming rate of 5% gives 7.6% of absolute error of MIT lectures and English broadcast news respectively.

VII. CONCLUSIONS

The underlying command for current research on speech technology and science is to understand and model an individual variation in spoken language. Each individual has their own way of speaking, which depends upon various factors that may include the dialect and accent of the speaker as well as the socioeconomic background of the speaker. The most important way for humans to communicate with each other and acquire information is with the help of speech. Since the invention of the telephone in the late 19th century using machines to extend people's ability to process speech has been a hot topic for various researchers. Among various speech processing problems, automatic speech recognition (ASR) for converting recorded speech automatically to text is one of the most challenging tasks. In this paper firstly we have given the brief introduction to speech recognition system and architecture that gives idea about it. Speech recognition systems are the techniques that transform speech to text are generally the result of computer learning. Voice recognition technology has made some major advances. Using speech technology ten years ago was more of a headache than the value that it was supposed to bring. So, different recent advancements done in speech recognition system that includes different subsections for different approaches and different aspects related to it also covered. Automatic speech recognition (ASR) is the recognition of the information inserted in a speech signal and its transcription in expression to a different set of characters. The ASR labels the issue of acoustic signal to the patterns of words. The review on ASR system for Indian and foreign languages is given in one of the section of this paper that gives idea of different work done on different languages and what approaches currently been used for it. In the last section of this paper we have given review on Feature extraction mechanism used by various researchers of Speech Recognition System. Overall this paper covers review on all aspects related to speech recognition system and in future with the help of this review we can work on any of the language using different improved approaches that gives better results than existing work done on it.

REFERENCES

[1] Y. Kumar, N. Singh, “An automatic speech recognition system for spontaneous Punjabi speech corpus”, Int J. Speech Technol, Springer, vol. 20(2), pp. 297-303, 2017.

[2] W. Ghai, N. Singh, “Analysis Of Automatic Speech Recognition Systems For Indo Aryan Languages : Pu”, 2012.

[3] S. Gupta, A. Pathak, A. Saraf, “A Study on Speech Recognition System: A Literature Review. International Journal of Science, Engineering and Technology Research, vol. 3(8), pp. 2192–2196, 2014.

[4] I. Badyal, M. D. Gupta, “The State of the Art of Automatic Speech Recognition : An Overview”, International Journal of Computer Science and Mobile Computing, vol. 4(2), pp. 359–368, 2015.

[5] S. Karpagavalli, E. Chandra, “A Review on Automatic Speech Recognition Architecture and Approaches”, International Journal of Signal Processing, Image Processing and Pattern Recagnition, vol. 9(4), pp. 393–404, 2016.

[6] A. Kumar, M. Dua, T. Choudhary, “Continuous Hindi Speech Recognition Using Monophone Based Acoustic Modeling”, International Conference on Advances in Computer Engineering & Applications, vol. 1, pp. 163–167, 2015.

[7] A. Vijayalakshmi, M. Jimmy, M. Nair, “A Study on Automated Speech Recognition Technique”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 4(3), pp. 614–617, 2015.

[8] H. Hofmann, S. Sakti, R. Isotani, H. Kawai, “Improving Spontaneous English ASR Using A Joint-Sequence Pronunciation Model”, 2010 4th International Universal Communication Symposium, IUCS 2010 - Proceedings, vol. 72, pp. 58–61, 2010.

[9] I. Kipyatkova, A. Karpov, V. Verkhodanova, “Modeling of Pronunciation, Language and Nonverbal Units at Conversational Russian Speech Recognition”, International Journal of Computer Science and Applications, vol. 10(1), pp. 11–30, 2013.

[10] S. Furui, “Recent Progress in Spontaneous Speech Recognition and Understanding”, IEEE Workshop on Multimedia Signal Processing, vol. 21, pp. 253–258.

[11] H. Atassi, Z. Smekal, A. Esposito, “Emotion Recognition from Spontaneous Slavic Speech”, 3rd IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2012 - Proceedings, vol. 25(5), pp. 389–394, 2012.

[12] A. Narayanan, D. Wang, “Improving Robustness of Deep Neural Network Acoustic Models Via Speech Separation and Joint Adaptive Training”, IEEE Trans on Audio, Speech, and Language Processing, vol. 23(1), pp. 92–101, 2015.

[13] D. Fohr, O. Mella, I. Illina, “New Paradigm in Speech Recognition : Deep Neural Networks”, IEEE International Conference on Information Systems and Economic Intelligence, Apr 2017, Marrakech, Morocco, vol. 1, pp. 870-879, 2017.

[14] N. A. Hendy, H. Farag, “Emotion Recognition Using Neural Network: A Comparative Study”, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 7(3), 1149–1155, 2013.

[15] K. Han, D. Yu, I. Tashev, “Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine”, Fifteenth Annual Conference, (September), vol. 25(8), pp. 223–227, 2014.

[16] L. He, M. Lech, N. Maddage, N. Allen, Neural Networks and TEO Features for an Automatic Recognition of Stress in Spontaneous

Speech. 5th International Conference on Natural Computation, ICNC 2009, vol. 2, pp. 227–231, 2009.

[17] J. Patadia, A. Reshamwala, “Feature Extraction Approach in Emotional Speech Recognition System”, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6(5), pp. 706–710, 2016.

[18] S. Furui, “Robust Methods in Automatic Speech Recognition and Understanding”, Proceedings EUROSPEECH, vol. 3, pp. 1993–1998, 2003.

[19] T. F. Furtuna, “Dynamic Programming Algorithms in Speech Recognition”, Word Journal of the International Linguistic Association, vol. 2, pp. 94–99, 2008.

[20] H. Nanjo, T. Kawahara, “New ASR Evaluation Measure and Minimum Bayes-Risk Decoding for Open-Domain Speech Understanding”, (ICASSP) IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 1, pp. 1053–1056, 2005.

[21] B. H. Juang, S. Furui, “Automatic Recognition and Understanding of Spoken Language - A First Step Toward Natural Human-Machine Communication”, Proceedings of the IEEE, vol. 8, pp. 1142–1165, 2000.

[22] X. Cui, L. Gu, B. Xiang, W. Zhang, Y. Gao, “Developing High Performance ASR in the IBM Multilingual Speech-To-Speech Translation System”, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 27(2), pp. 5121–5124, 2008.

[23] V. Silber-varod, N. Geri, “Can Automatic Speech Recognition be Satisficing for Audio/Video Search? Keyword-Focused Analysis of Hebrew Automatic and Manual Transcription”, Online Journal of Applied Knowledge Management, vol. 2(1), pp. 104–121, 2014.

[24] P. Saini, P. Kaur, “Automatic Speech Recognition: A Review”, International Journal of Engineering Trends and Technology, vol. 4(3), pp. 132–136, 2013.

[25] A. Vijayalakshmi, M. Jimmy, M. Nair, “A Study on Automated Speech Recognition Technique”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 4(3), pp. 614–617, 2015.

[26] S. Karpagavalli, E. Chandra, “A Review on Automatic Speech Recognition Architecture and Approaches”, International Journal of Signal Processing, Image Processing and Pattern Recagnition, vol. 9(4), pp. 393–404, 2016.

[27] A. Kumar, M. Dua, T. Choudhary, “Continuous Hindi Speech Recognition Using Monophone Based Acoustic Modeling”, International Conference on Advances in Computer Engineering & Applications, vol. 1, pp. 163–167, 2014.

[28] I. Badyal, M.D. Gupta, “The State of the Art of Automatic Speech Recognition : An Overview”, International Journal of Computer Science and Mobile Computing, vol. 4(2), pp. 359–368, 2015.

[29] W. Ghai, N. Singh, “Analysis Of Automatic Speech Recognition Systems For Indo-Aryan Languages : Punjabi A Case Study”, International Journal of Soft Computing and Engineering (IJSCE), vol. 2, pp. 379–385, 2012.

[30] S. Li, Y. Akita, T. Kawahara, “Discriminative Data Selection For Lightly Supervised Training Of Acoustic Model Using Closed Caption Texts”, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2015–January, vol. 4, pp. 3526–3530, 2015.

[31] S. Schötz, “Linguistic & Paralinguistic Phonetic Variation in Speaker Recognition & Text-To-Speech Synthesis”, GSLT: Speech Technology, vol. 9, pp. 1–10, 2002.

[32] C. Parada, M. Dredze, “Learning Sub-Word Units for Open Vocabulary Speech Recognition”, Proc. ACL, vol. 4, pp. 712–721, 2011.

[33] J. H. Tailor, “Speech Recognition System Architecture for Gujarati Language”, International Journal of Computer Applications, vol. 12, pp. 28–31, 2016.

[34] C. Burileanu, V. Popescu, A. Buzo, “Spontaneous Speech Recognition for Romanian in Spoken Dialogue Systems”, Proceedings of the Romanian Academy, vol. 1, pp. 83–91, 2010.

[35] P. Mihajlik, Z. Tuske, B. Tarjan, B. Nemeth, T. Fegyo, “Improved Recognition of Spontaneous Hungarian Speech-Morphological and Acoustic Modeling Techniques for A Less Resourced Task”, IEEE Transactions on Audio, Speech and Language Processing, vol. 6, pp. 1588–1600, 2010.