International Journal of Engineering Sciences Paradigms and Researches (IJESPR)
Vol. 48, Issue 02, Quarter 02 (April-May-June 2019)
(An Indexed, Referred and Impact Factor Journal)
ISSN (Online): 2319-6564
www.ijesonline.com

Automatic Speech Recognition: A Review

Dr Mukesh Singla

Professor, SPGOI, Rohtak
mukesh27singla@yahoo.co.in

Publishing Date: April 15, 2019

Abstract

With the advancement in innovation and the inherent benefit of voice based correspondence because of its fluctuation, speed and security has driven consideration towards automated acknowledgment of speech. (ASR) Automatic speech recognition has been broadly study amid the previous couple of decades. Today, a large portion of the ASR framework in view of measurable demonstrating, and HMM is the most well-known one among them. Present days, execution of ASR wind up one of the real bottleneck for its handy utilizes. While Deep Neural Networks have made gigantic progress for substantial vocabulary constant discourse acknowledgment tasks, preparing these systems is moderate. Deep neural network design upgrades the partition execution as far as various target measures under the semi-directed mode where the preparation information of the objective speaker is given while the inconspicuous interferer in the detachment arrange is anticipated by utilizing numerous meddling speakers blended with the mixed with the target speaker in the training stage. Consolidated worldly and otherworldly handling strategy is utilized as a preprocessing method for improving the debased discourse. Language discriminative data in high resonance areas of discourse is utilized for the assignment of dialect language identification.

Keywords: *Speech Recognition, ASR Architecture, Continuous Speech Recognizer, Acoustic Models.*

1. Introduction

1.1 Speech Recognition: Speech to text conversion is the ability of a machine to recognize speech sound and convert in to text sequence as close as possible. During the past few decades the speech is used as a communication medium. There are many commercial products are also available for more than 20 years, at first isolated word recognition, and then connected word recognition andcontinuous speech recognition. Most of these systems arebased on statistical modeling. The first ASR system came into existence in 1952, which was developed by Devis at Bell laboratory. This is speaker dependent, isolated digit recognition system. Basically, ASR systems are based on two acoustic models such as: (1) Word model, (2) Phone model. Word model isused when the size of vocabulary is small. In this model, the words are modeled as complete. In phone model instead of modeling the whole word, we model only parts of words generally phones. Continuous speech recognition is achieved by phone modeling. Phone model can be further classified into two parts: (1) context- independent phone mode, (2) context– dependent phone model. In context- independent phone model individual phones are modeled. ASR system based onmono-phones are comes under this category. While modeling phones in context dependent phone model their neighbors are also measured. Tri-phone based continuous speech recognition systems are comes under this category[7].

1.2ASR Architecture: ASR system contains four basic modules which are made up of feature extraction, acoustic model, language model and the recognizer for word and sentence level matching. A framework of an integrated technique to continuous speech recognition is given infigure.1. The element extraction module computes the acoustic feature vectors used to portray the spectral properties of time fluctuating speech signal. The acoustic match module looks at the comparability among input include vector grouping and an arrangement of acoustic models generated for all word in the assignment vocabulary.

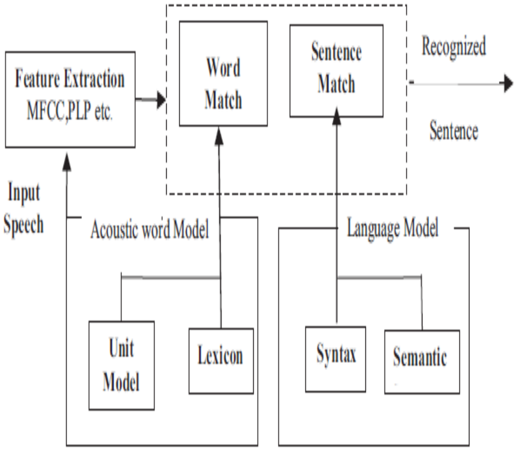

Figure 1: Block diagram of continuous speech recognizer [1]

2. RELATED WORK

ShwetaSinhaet al. assess number of Gaussian mixture part for Hindi database in light of the extent of vocabulary.MFCC and PLP includes alongside its expanded variant has been utilized as speech feature. HLDA is applied for feature decrease while utilizing broadened highlights.

Sabato Marco Siniscalchi et al. Two principle commitments have activated such another pattern: 1) a noteworthy progress has been made in preparing the weights in profound neural systems (DNNs), and a pre-prepared profound neural system shrouded Markov display (DNN-HMM) half and half engineering has outflanked a traditional Gaussian blend demonstrate concealed Markov display (GMM-HMM) programmed discourse acknowledgment (ASR) framework on a testing business seek dataset, and 2) it has been demonstrated that phoneme grouping can be helped by utilizing a various leveled structure of multi-layer perceptron's (MLPs) prepared to show long-traverse worldly examples with gainful consequences for dialect acknowledgment assignments. In this work, author join these two lines of research and exhibit that word acknowledgment exactness can be essentially improved by masterminding DNNs in a various leveled structure to show long haul vitality directions.**Tara N. Sainath et al.**investigate a wide range of enhancement procedures to enhance DNN preparing speed. This incorporates parallelization of the slope calculation amid cross-entropy and arrangement preparing, and in addition diminishing the quantity of parameters in the system utilizing a low-rank network factorization. Applying the proposed improvement procedures, creator demonstrates that DNN preparing can be accelerated by a factor of 3 on a 50-hour English Broadcast News (BN) atask with no loss in accuracy.

Yulan Liu etal. introduces an examination of far field discourse acknowledgment utilizing beam forming and divert connection with regards to Deep Neural Network (DNN) based element extraction. While discourse upgrade with beamforming is appealing, the calculations are ordinarily flag based with no data about the extraordinary properties of discourse. A straightforward alternative option to beamforming is linking different channel highlights. Results exhibited in this paper show that channel link gives comparative or better outcomes. All things considered the DNN front-end yields a 25% relative diminishment in Word Error Rate (WER).**Tu**

Yanhui et al. In this paper, a novel deep neural network (DNN) design is proposed to produce the speech features of both the objective speaker and interferer for speech partition without utilizing any earlier data about the interfering speaker. DNN is received here to straightforwardly display the profoundly nonlinear connection between speech highlights of the merged signs and the two competing speakers. Trial comes about on a monaural speech detachment and acknowledgment challenge assignment demonstrate that the proposed DNN system upgrades the division execution as far as various target measures under the semi-regulated mode where the preparation information of the objective speaker is given while the inconspicuous interferer in the partition organize is anticipated by using different meddling speakers blended with the objective speaker in the preparation arrange. Besides, as a preprocessing advance in the testing stage for powerful discourse acknowledgment, speech partition approach can accomplish critical enhancements of the acknowledgment precision over the benchmark framework with no source detachment.

Darryl Stewart et al. shows the (MWSP) maximum weighted stream posterior display as a strong and productive stream combination technique for varying media speech acknowledgment in conditions, where the sound or video streams might be subjected to obscure and time-shifting defilement. A huge preferred standpoint of MWSP is that it doesn't require a particular estimation of the flag in either stream to ascertain proper stream weights amid acknowledgment, and in that capacity it is methodology autonomous.

AnkitKuamr et al. The point of this paper is to examine the ideal number of Gaussian blend that shows most extreme precision with regards to Hindi discourse acknowledgment. Well toolbox HTK 3.4.1 is utilized for the usage of this framework, in which Mel recurrence cepstral coefficient (MFCC) is utilized as an element extraction system. The test comes about demonstrate that the most extreme execution of the proposed framework is accomplished when creator utilize four segment Gaussian blends HMM show.

Anil kumarVuppala et al.In this paper execution of the LID framework is considered in different background situations like clean room, auto manufacturing plant, high recurrence, and pink and background noise. In this work, Indian Institute of Technology Kharagpur - Multi Lingual Indian Language Speech Corpus (IITKGP-MLILSC) is utilized for building dialect recognizable proof framework. Clamor discourse tests from the NOISEX database are utilized in the present examination. The execution of the proposed technique is very acceptable contrasted with existing methodologies.

Xiong Xiao et al. This paper researches (DNN)deep neural networks in light of nonlinear component mapping and measurable direct element adjustment approaches for diminishing resonation in discourse signals. In the nonlinear element mapping approach, DNN is prepared from parallel clean/twisted discourse corpus to outline and loud discourse coefficients, (for example, log size range) to the fundamental clean speech coefficients.

Jun Ren et al. this paper embraces Deep Belief Neural Networks (DBNs) to demonstrate the circulation of dysarthric discourse flag. A ceaseless dysarthric discourse acknowledgment framework is delivered, in which the DBNs are utilized to anticipate the back probabilities of the states in Hidden Markov Models (HMM) and the Weighted Finite State Transducers structure was used to manufacture the discourse decoder. Test comes about demonstrate that the proposed technique gives better forecast of the likelihood appropriation of the unearthly portrayal of dysarthric discourse that outflanks the current strategies, e.g., GMM-HMM based dysarthric discourse recognition approaches. To the best of our insight, this work is the first run through to manufacture a consistent discourse acknowledgment framework for dysarthric discourse with profound neural system procedure, which is a promising methodology for enhancing the correspondence between those people with discourse obstacles and typical speakers.

Table 1: Summary of Speech Recognition Techniques

S No	Year	Model/Technique	Outcome
1	2013	Novel approach to develop a speech recognition system for Hindi using	Results show that for a medium sized vocabulary of

		MFCC, PLP and their extensions.	500 to 600 words 8 gaussian components gives optimal result.
2	2013	By arranging DNNs in a hierarchical structure to model long-term energy trajectories	The word recognition accuracy can be significantly enhanced,
3	2013	Variety of different optimization techniques 1. hybrid pre-training 2.Hessian-freeoptimization, 3.low-rank matrix factorization	1. Successfully parallelize the gradient computation, achieving a 3× speedup for cross-entropy fine tuning time on a 50 hr English BN task. 2.Sequence training can be sped up by a factor of 3 3. Reduce the number of parameters by 33% with no loss in accuracy.
4	2014	The DNN based front-end on the AMI meeting corpus was tested in the context of far field speech recognition.	On both overlapping and non-overlapping speech, BN features gave an average 25% relative WER reduction over using PLP features, regardless of the number and type of microphones.
5	2014	Novel architecture of DNN for separating speech of both the target and the interfering speaker	Approach also shows the effectiveness for robust speech recognition as a preprocessing step.
6	2014	A novelapproach called the MWSP model that does not requireany specific measurement of the noise level in the signalof either modality.	MWSP has been shown in experiments to offera smooth integration of the two modalities, making bestuse of the available reliable information on a frame-byframebasis and remaining simple to implement without*a priori* knowledge of the environmental conditions in whichit will bedeployed or tested.

3. Conclusion

Recognition of human speech by machine is a very challenging task from the last few decades and still there is no accurate system which acts as an interface between man and machine. Even though research being carried out in this area for last so many decades no accurate system has yet been developed. There are still many open problems that need quick and specific solution in speech recognition process. Variety of diverse optimization techniques are used to get better DNN training speed. The DNN based front-end on the AMI meeting corpus was tested in the context of far field speech recognition. On both overlapping and non-overlapping speech results of experiments show that, BN features have an average 25% relative WER decrease over when using PLP features, in spite of of the number and type of microphones. A novel approach called the MWSP model that does not require any specific measurement of the noise level in the signal of either modality. As (Maximum Weighted Stream Posterior Mod) MWSP is independent of modality- it complements the past research here and could be utilized as an option or maybe alongside by different methodologies. DNN mapping makes twisting upgraded discourse waveforms particularly when the resonation is solid. This is in part because of the reverberant stage spectrogram being utilized with the DNN-upgraded size spectrogram to re-orchestrate the discourse waveforms.

References

[1] Shweta Sinha, S S Agrawal and Aruna Jain, ”Continuous Density Hidden Markov Model for Context Dependent Hindi speech Recognition”, International Conference on Advances in Computing, Communications and Informatics (ICACCI),IEEE,2013,PP.1953-1958

[2] Sabato Marco Siniscalchi, DongYu, LiDeng, and Chin-Hui, ”Speech Recognition Using Long-Span Temporal Patterns in a Deep Network Model”, IEEE SIGNAL PROCESSING LETTERS, VOL. 20, NO. 3, MARCH 2013,PP.201-204

[3] Tara N. Sainath, Brian Kingsbury, Hagen Soltau, and Bhuvana Ramabhadran,”Optimization Techniques to Improve Training Speed of Deep Neural Networks for Large Speech Tasks”,IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 21, NO. 11, NOVEMBER 2013,PP.2267-2276.

[4] Yulan Liu, Pengyuan Zhang and Thomas Hain,”USING NEURAL NETWORK FRONT-ENDS ON FAR FIELD MULTIPLE MICROPHONES BASED SPEECH RECOGNITION”, IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP),PP.5542-5546.

[5] TuYanhui, Du Jun, Xu Yong, Dai Lirong, Lee Chin-Hui,”Deep Neural Network Based Speech Separation for Robust Speech Recognition”, IEEE,2014,PP.532-536.

[6] Darryl Stewart, Rowan Seymour, Adrian Pass, and Ji Ming,”Robust Audio-Visual Speech Recognition under Noisy Audio-Video Conditions”, IEEE TRANSACTIONS ON CYBERNETICS, VOL. 44, NO. 2, FEBRUARY 2014,PP.175-184.

[7] AnkitKuamr, MohitDua, Tripti Choudhary, ”Continuous Hindi Speech Recognition Using Gaussian Mixture HMM”,2014 IEEE Students’ Conference on Electrical, Electronics and Computer Science,PP.1-4

[8] Anil kumar Vuppala Mounika K.V Hari Krishna Vydana, ”Significance of Speech Enhancement and Sonorant Regions of Speech for Robust Language Identification”, IEEE,2015,PP1-5

[9] Xiong Xiao, Shengkui Zhao, Duc Hoang Ha Nguyen, Xionghu Zhong, Douglas L. Jones, EngSiong Chng and Haizhou Li, ”Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation”, EURASIP Journal on Advances in Signal Processing, 2016,PP.1-18.

[10] Jun Ren, Mingzhe Liu, ”An Automatic Dysarthric Speech Recognition Approach using Deep Neural Networks”, International Journal of Advanced Computer Science and Applications, Vol. 8, No. 12, 2017,PP.48-52