International Journal For Technological Research In Engineering
Volume 7, Issue 11, July-2020 *ISSN (Online): 2347 - 4718*

A COMPREHENSIVE STUDY ON AUTOMATIC SPEECH RECOGNITION SYSTEM BASED ON DEEP LEARNING

Shefali Gupta[1], Er. Vikas Kumar[2]
[1]M.Tech Scholar, [2]Asst. Professor
Electronics & Communication Engg. Department, Galaxy Global Group of Institutions, Ambala

Abstract: The improvement of speech intelligibility is a traditional problem which still remains open and unsolved. The recent boom of applications such as hands-free communications or automatic speech recognition systems and the ever-increasing demands of the hearing-impaired community have given a definitive impulse to the research in this area. The aim of the noise reduction algorithms is to estimate the clean speech signal from the noisy recordings in order to improve the quality and intelligibility of the enhanced signal. This paper provides a comprehensive study on speech enhancement and further processing. This one contains a preliminary study of the problem and a thorough review of the state of the art in this field. Some author's work is also explained.
Keywords: Speech Processing, Speech Enhancement, Deep Learning, Hearing Aid, Feature Extraction etc .

I. INTRODUCTION

In perfect situation there ought to be no corruption in quality or potentially clarity of unique discourse as well as human subjects have ordinary discourse generation and discernment frameworks. In down to earth situation there is corruption in quality or potentially comprehensibility and additionally human subjects have debilitated discourse creation and observation frameworks. So the objective of discourse upgrade is to improve quality and coherence. But when contributions from various mouthpieces are accessible (in some uncommonly masterminded cases), it has been extremely hard for discourse upgrade frameworks to improve clarity. In this way most discourse upgrade techniques raise quality, while limiting any misfortune in clarity. As watched, certain parts of discourse are more perceptually significant than others. The sound-related framework is more delicate to the nearness than nonattendance of vitality, and will in general overlook numerous parts of stage. In this manner discourse improvement calculations frequently center around precise demonstrating of tops in the discourse adequacy range, as opposed to on stage connections or on vitality at more fragile frequencies [1]. Speech processing or noise reduction has been one of the main investigated problems in the speech community for a long time. The problem arises whenever a desired speech signal is degraded by some disturbing noise. The noise can be additive or convolutive. In practice, a convolutive noise should be rather considered due to the reverberation. However, it is usually assumed that the noise is additive since it makes the problem simpler and also the developed algorithms based on this assumption lead to satisfactory results in practice. Even this additive noise can reduce the quality and intelligibility of the speech signal considerably. Therefore, the aim of the noise reduction algorithms is to estimate the clean speech signal from the noisy recordings in order to improve the quality and intelligibility of the enhanced signal. There are various applications of speech enhancement in our daily life. A speech signal consists of three classes of sounds. They are voiced, fricative and plosive sounds. Voiced sounds are caused by excitation of the vocal tract with quasi-periodic pulses of airflow [2].

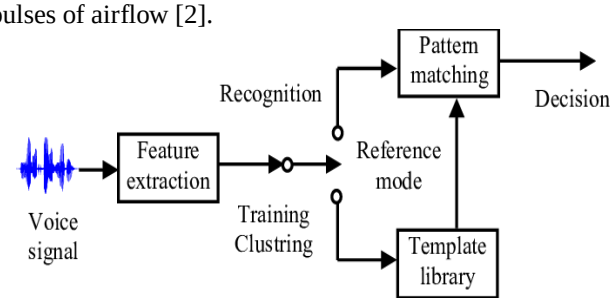

Fig 1: Speech Recognition System Process [1]

Fig1 described the process of speech recognition with clear block diagram. It explained the concept of speech signal processing with clear representation. Multichannel or multi amplifier commotion decrease frameworks, use the transient and phantom data just as the spatial data to appraise an ideal discourse signal from the given boisterous chronicles Consonants, be that as it may, can be begun by a voiced or an unvoiced sound and are delegated:

- Stops: which happen when the wind stream is blocked and all of a sudden discharged.
- Nasals: created when the air is halted in the oral pit yet not through the nasal cavity.
- Approximants: created when there is a tightening yet not limited enough to bring about choppiness.
- Fricatives: a thin narrowing in the vocal tract bringing about a tempestuous wind stream.

Characteristics and Estimation of Noise Signals

Commotion, rather than discourse, can start from any sort of source and have any phantom and transient qualities. There are, in any case, some basic suppositions made about the commotion when moving toward the discourse upgrade issue [2]:

- The control range of clamor is more stationary than that of discourse, and
- Speech and commotion are measurably free.

Numerous discourse improvement strategies require an estimation of the clamor control range, or, proportionately, the SNR at each time-recurrence container. The exactness of

the commotion estimation method majorly affects both the quality and understand ability execution of the prepared discourse. The principal commotion estimation approaches utilized Voice Activity Detector (VAD) estimators to distinguish clamor just interims. The commotion could be then determined by a fleeting normal during the discourse nonappearances utilizing an averaging time-steady that relies upon the expected stationary of the clamor. The premise of this methodology is that over a given time-interim there will be delays in the discourse in each recurrence band and thus the base estimation of the uproarious discourse range inside a recurrence band will relate to the commotion control. The clamor control range can likewise be determined by utilizing a Minimum Mean Squared Error (MMSE) estimator. A MMSE estimator was utilized to limit the intensity of the distinction work between the assessed and the genuine clamor control range.

Applications of Speech Enhancement in Daily Life

Some applications of speech enhancement that can be found in the daily life are:

1. Hearing Aids

Hearing misfortune an influences a significant level of individuals, and this figure is expanding because of the developing presentation to over the top commotion in their day by day lives. One of the primary issues for hearing-impeded individuals is the decrease of discourse understand-ability in boisterous situations, which is for the most part brought about by the loss of worldly and phantom goals in the sound-related arrangement of the hindered ear. The utilization of portable amplifier gadgets that just give intensification doesn't take care of the issue, because of the way that they enhance both discourse and commotion.

2. Hands-Free Communication Systems

Lately, the interest of sans hands correspondences for vehicles or video chat frameworks has radically expanded the innovative work of this sort of gadgets. The accomplishment of these frameworks depends on the nature of the gained discourse, which is sullied by various kinds of commotion and impedances. Thusly, the signs gained by the mouthpieces of the framework are normally improved before being transmitted through the correspondence channel.

3. Automatic Speech Recognition (ASR)

Much advancement has been made in ASR in the most recent years. PDAs, PCs or savvy TVs are just a few instances of current innovations that incorporate ASR. The likelihood of achievement in the acknowledgment firmly relies upon the nature of the obtained signal, and the presentation of ASR frameworks quickly debases within the sight of clamor. This reality makes a past phase of discourse improvement important for ASR frameworks.

4. Recording Systems

Sound chronicles have numerous applications, for example, security, programmed music interpretation, sound data recovery or electronic observation. One alluring activity to perform with these accounts is to recuperate the first sources with high calibre, isolating the diverse sound sources and evacuating foundation commotion.

The paper is ordered as follows. In section II, it provided concept of speech enhancement with hearing aid. In Section III, It defines the author researches related to speech processing & enhancement. In section IV, it describes the major gaps related to work. Finally, conclusion is explained in Section V.

II. SPEECH ENHANCEMENT IN DIGITAL HEARING AIDS

Portable hearing assistants are electronic gadgets worn by hearing-impeded individuals in a perfect world to improve the diminished understand ability brought about by hearing misfortune. Straightforward gadgets frequently produce enhanced commotions when the client is in a multi-source condition (for example a packed bar). Present day gadgets incorporate some kind of upgrade framework to defeat this impediment, for example, directional mouthpieces or discourse improvement calculations. Be that as it may, notwithstanding the issues found by discourse improvement calculations while improving coherence, their application in portable amplifiers involves three principle extra issues: hearing-disabled audience members have more noteworthy vulnerability to the twists presented by signal preparing calculations, the little size of hearing gadgets restrains the quantity of receivers gathered in the gadget, and the diminished existence of the present batteries compels the computational expense of the executed calculations.

Hearing Impairment

The quantity of individuals with hearing misfortune is expanding at a disturbing rate not just on account of the maturing of the total populace, yet in addition due to the developing introduction to commotion in everyday life. Hearing misfortune is usually spoken to by an audiogram, which shows the sound-related edge in logarithmic units (dB) for normalized frequencies estimated by an audiometer. Hearing impedance suggests bigger edges than typical hearing however the degree of misfortune among frequencies isn't uniform and relies upon every individual. The level of hearing misfortune is normally characterized as the normal hearing misfortune estimated at a specific octave-band, and the degree of misfortune is typically ordered into gentle (up to 40 dB), moderate (from 40 to 60 dB) and extreme (more than 60 dB). For hearing-disabled individuals experiencing gentle to direct hearing misfortune, a listening device is useful, yet on account of extreme hearing misfortune, the utilization of portable amplifiers is of little advantage, and some different arrangements might be thought of. Hearing-debilitated individuals face a wide range of sound-related issues that diminish their capacity of comprehension. These issues are depicted beneath.

1. Decreased Level of Audibility

Contingent upon the degree of hearing misfortune, an individual will hear a few sounds yet miss some different sounds. All in all, the high-recurrence segments of discourse

are more vulnerable than the low-recurrence segments, and hearing loss of older individuals is higher at high frequencies. Thus, hearing-impeded individuals will in general miss high-recurrence data, essentially consonants. This reality prompts miss fundamental pieces of certain phonemes lessening the clarity.

2. Reduced Dynamic Range

The dynamic scope of the sound-related framework is characterized as the level distinction between the sound-related limit and the uneasiness edge (for example limit of torment). For hearing disabled individuals, the sound-related limit is expanded in contrast with ordinary hearing individuals, subsequently the dynamic range is diminished. So as to abstain from surpassing the inconvenience limit, listening devices must enhance feeble sounds more than extraordinary sounds.

3. Reduced Frequency Resolution

Recurrence goals continuously diminishes as the level of hearing misfortune increments, and hearing-disabled individuals discover hard to recognize hints of various frequencies at the same time. This is because of the loss of affectability of the hair cells of the cochlea, which diminishes the capacity of separating frequencies.

4. Decreased Temporal Resolution

As a rule, more fragile sounds are now and again conceal by serious sounds that promptly go before or tail them, which diminishes the odds of understand ability. Moreover, the capacity to hear feeble sounds during brief timeframe openings slowly diminishes as the level of hearing misfortune increments, and hearing-debilitated individuals as a rule experience diminished worldly goals, which includes that the discourse understand ability saw by them is additionally diminished.

III. LITERATURE REVIEW

Yousheng X. et. al. [2014] [2] proposed a novel multi-channel discourse improvement strategy by consolidating the wiener separating and subspace sifting with a raised combinational coefficient. It proposed multi-channel discourse upgrade technique had a superior presentation in powerfully expelling hued clamor from loud discourse signals. Re-enactment models affirmed that under various hued commotion, the proposed multi-channel discourse upgrade strategy can get preferred discourse recuperation results over the conventional subspace multi-channel discourse improvement technique. Jie Z. et. al. [2014] [3] examined the reasonableness of discourse quality assessment gauges under different commotion conditions in the utilization of ghostly subtraction discourse improvement. At that point fitting assessment calculations were picked for discourse upgrade dependent on ghastly subtraction. The recreation results demonstrated that in the utilization of discourse upgrade, the reasonableness of discourse quality assessment calculations is constrained to the SNR of boisterous discourse, recording individuals, recording substance and foundation commotion condition. Ogawa A. et. al. [2014] [4] proposed a quick section look technique for corpus based discourse upgrade. It was for the most part dependent on two systems got from discourse acknowledgment innovation. The main was a quest like portion assessment work for precisely finding the longest coordinating fragments. The second was a tree and direct associated quest space for effectively sharing the section probability estimations. In the examinations for non-stationary uproarious perceptions utilizing the 26 multi-condition TIMIT parallel discourse corpus, the proposed inquiry strategy found the fragments nearly progressively without corrupting the nature of the upgraded discourse.

Prasanna A. et. al. [2014] [5] introduced a Codebook-based discourse upgrade (CBSE) utilizing prepared discourse and clamor codebooks for taking care of non-stationary commotion. Notwithstanding, the high register serious nature of this procedure rendered it inapplicable progressively discourse improvement situations by presenting a huge deferral in discourse transmission. In this work, this issue was tended to by giving a proficient, parallel CBSE calculation. The proposed parallel CBSE calculation was then utilized as a premise to give a novel cloud based structure to accomplish constant discourse improvement in portable correspondence as a proof-of idea. Shah Z. et. al. [2014] [6] introduced the assurance of ideal estimations of TXOP and FA that amplify VoIP limit. It originally decided the ideal estimation of FA that expands VoIP limit. The reproduction results indicated that ideal estimation of FA that amplifies VoIP limit is 14. At 10 ms packetization interim this estimation of FA gives an increase of 26% in VoIP limit when contrasted with the VoIP limit with no FA. Besides, it found that the ideal estimation of TXOP that expands VoIP limit is 13. At 10ms packetization interim this estimation of TXOP gives an increase of 32% in VoIP limit when contrasted with VoIP limit with default estimation of TXOP. We at that point decide the VoIP limit when ideal estimations of TXOP and FA are at the same time utilized.

Tan L. et. al. [2014] [7] proposed an element improvement procedure for commotion strong discourse acknowledgment. Existing inadequate model based element improvement techniques utilized clean discourse and unadulterated clamor Mel-ghostly models, or perfect and loud discourse log-Mel-phantom model sets, in their word references. The meager direct blend of SMest word reference models that best spoke to the test expression's SMest was acquired by taking care of a L1-minimization issue. This meager straight mix was applied to the SMref model word reference to create an improved delicate veil for denoising the articulation's Mel-spectra before MFCC extraction. Yun S. et. al. [2014] [8] set up a monstrous language and discourse database nearest to nature where discourse to-discourse interpretation gadget really was being utilized in the wake of assembling a lot of individuals dependent on the overview on clients 'requests. Besides, with the discourse to-discourse interpretation UI, an easy to understand UI had been structured; and simultaneously, mistakes were diminished during the procedure of interpretation the same number of measures to improve client fulfilment were utilized. Araki S. et. al.

[2015] [9] examined a multi-channel de-noising auto-encoder (DAE)- based discourse upgrade approach. As of late, profound neural system (DNN)- based monaural discourse improvement and hearty programmed discourse acknowledgment (ASR) approaches have pulled in a lot of consideration because of their elite. Despite the fact that multichannel discourse upgrade as a rule outflanks single channel draws near, there has been little research on the utilization of multi-divert preparing with regards to DAE. In this paper, they investigated the utilization of a few multi-channel includes as DAE contribution to affirm whether multi-channel data can improve execution.

Gong Z. et al. [2015] [10] created two implanted amplifier frameworks with clamor decrease, separately utilizing Kalman channel and Wiener channel procedures. The relative outcome demonstrated that the portable hearing assistant framework dependent on the Kalman channel based discourse upgrade can build the pace of discourse acknowledgment and the conference solace of hearing disabled people in a loud domain, contrasted and the listening device framework dependent on the Wiener channel based discourse improvement. Feng Deng et. al. [2015] [11] proposed an inadequate shrouded Markov model (HMM) based single-channel discourse improvement technique that models the discourse and commotion gains precisely in non-stationary clamor conditions. The probability standard for finding the model parameters is expanded with a regularization term bringing about a scanty autoregressive HMM (SARHMM) framework that empowers sparsity in the discourse and clamor displaying. In the SARHMM just few HMM states contribute altogether to the model of every specific watched discourse section.

Mowlaee P. et al. [2015] [12] exhibited a consonant stage estimation technique depending on basic recurrence and sign to-clamor proportion (SNR) data evaluated from boisterous discourse. The proposed strategy depends on SNR-based time-recurrence smoothing of the unwrapped stage got from the deterioration of the loud stage. The viability of the proposed stage estimation strategy is assessed for both stage just improvements of uproarious discourse and in mix with an adequacy just upgrade conspire. Pawar S. et. al. [2015] [13] introduced a calculation for improving discourse comprehensibility. This proposed calculation can be utilized for down to earth hearing prosthetic gadgets. Execution of the paired concealing calculation utilizes a bank of band-pass channels to perform mapping of sign. Additionally, grouping is performed with a sign to-commotion (SNR) gauge and a comparator. This incorporates spatial sifting strategy, characterization of sign, for example, unique and boisterous sign. After this dependent on SNR edge level sign are recombined to get decreased commotion level in discourse signal. Sun M. et al. [2016] [14] displayed a profound auto encoder (DAE) for precisely demonstrating the spotless discourse range. In the resulting phase of discourse upgrade, an extra DAE was acquainted with speak to the lingering part acquired by subtracting the assessed clean discourse range (by utilizing the pre-prepared DAE) from the uproarious discourse range. The improved discourse signal was hence acquired by changing the assessed clean discourse range once again into time space.

Renjith S et al. [2017] [15] planned for building up a component based feeling acknowledgment framework. The discourse chronicles with the feelings – outrage, joy and bitterness in Tamil and Telugu dialects are utilized for this work. Commotion and different unsettling influences from the discourse waveforms were isolated from the crude discourse signals utilizing pre-handling. Two highlights Linear Predictive Cepstral Coefficients (LPCC) and Hurst Parameter were separated. In view of the measurable parameters acquired from these highlights, the characterization was finished. Classifiers – K-Nearest Neighbor (KNN) and Artificial Neural Network (ANN) were utilized to distinguish separate feelings. The presentation of the classifiers were analyzed as far as exactness, accuracy and review.

Wang H. et al. [2018] [16] presented a front-end speech enhancement approach to robust speech recognition in automotive environments. It combined hybrid voice activity detection (VAD), relative transfer function (RTF) based generalized side lobe cancelation, and single-channel post filtering to enhance the speech signal of interest, thereby improving the robustness of speech recognition. Experiments were conducted in real automotive environments. The results showed that the developed method can significantly improve the performance of both VAD and automatic speech recognition. Wood S. et al. [2019] [17] exhibited a widespread codebook-based discourse improvement structure that depends on a solitary codebook to encode both discourse and commotion parts. The nuclear discourse nearness likelihood (ASPP) is characterized as the likelihood that a given codebook molecule encodes discourse at a given point in time. show that the proposed ITF-based ASPP approach accomplishes a decent parity of the exchange off between binaural clamor decrease and binaural prompt safeguarding.

IV. GAPS IN STUDY

Speech enhancement or noise reduction has been one of the main investigated problems in the speech community for a long time. The problem arises whenever a desired speech signal is degraded by some disturbing noise. The noise can be additive or convolutive. Even this additive noise can reduce the quality and intelligibility of the speech signal considerably. Therefore, the aim of the noise reduction algorithms is to estimate the clean speech signal from the noisy recordings in order to improve the quality and intelligibility of the enhanced signal. Due to this, it proposes a method for speech enhancement using mask estimation iteratively. The main objectives of the work are recording of Real time speech signal and to design a modified speech processing & enhancement using deep learning.

V. CONCLUSION

This paper provides a comprehensive study on speech enhancement and further processing. This one contains a preliminary study of the problem and a thorough review of the state of the art in this field. Signal degradation, however, is most commonly caused by noise from unwanted acoustic

sources in the environment, which may affect the speech quality and/or intelligibility of the wanted signal. There are several approaches which enhance the signal using time-frequency gain modification, such as spectral subtraction or MMSE-based algorithms. Although the most approaches aim to estimate the clean speech by applying a continuous gain. Therefore, the aim of the noise reduction algorithms is to estimate the clean speech signal from the noisy recordings in order to improve the quality and intelligibility of the enhanced signal. The review includes existing solutions to both the single-channel and multichannel speech enhancement problem, considering the noise reduction and the source separation approaches.

REFFERENCES

[1] H. Veisi H. Sameti, (2012) "Hidden-Markov-model-based voice activity detector with high speech detection rate for speech enhancement", IET Signal Processing, pp. 01-06.
[2] Xia Yousheng, Huang Jianwen, (2014) "Speech Enhancement Based on Combination of Wiener Filter and Subspace Filter", IEEE, pp. 81-86.
[3] Zhang Jie, Xiaoqun Zhao, Jingyun Xu, (2014) "Suitability of Speech Quality Evaluation Measures in Speech Enhancement", IEEE pp. 102-108.
[4] Atsunori Ogawa, Keisuke Kinoshita, Takaaki Hori, (2014) "Fast Segment Search For Corpus-Based Speech Enhancement Based On Speech Recognition Technology", IEEE International Conference on Acoustic, Speech and Signal Processing.
[5] AN.SaiPrasanna, Iyer Chandrashekaran, (2014) "Real Time Codebook Based Speech Enhancement with GPUs", International Conference on Parallel, Distributed and Grid Computing, IEEE, pp. 1042-1048.
[6] Zawar Shah, Ather Suleman, Imdad Ullah, (2014) "Effect of Transmission Opportunity and Frame Aggregation on VoIP Capacity over IEEE 802.11n WLANs", IEEE pp. 256-262.
[7] Lee Ngee Tan, Abeer Alwan, (2014) "Feature Enhancement Using Sparse Reference And Estimated Soft-Mask Exemplar-Pairs For Noisy Speech Recognition", IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 22-28.
[8] Seung Yun, Young-Jik Lee, and Sang-Hun Kim, (2014) "Multilingual Speech-to-Speech Translation System for Mobile Consumer Devices", IEEE Transactions on Consumer Electronics, Vol. 60, No. 3, pp. 232-238.
[9] Shoko Arakit, Tomoki Hayashi, (2015) "Exploring Multi-Channel Features for Denoising-Auto-encoder-Based Speech Enhancement", IEEE pp. 424-430.
[10] Zheng Gong and Youshen Xia, (2015) "Two Speech Enhancement-Based Hearing Aid Systems and Comparative Study", IEEE International Conference on Information Science and Technology, April 24-26 pp. 122-128.
[11] Feng Deng, Changchun Bao, (2015) "Sparse Hidden Markov Models for Speech Enhancement in Non-Stationary Noise Environments", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 23, No. 11.
[12] Pejman Mowlaee and Josef Kulmer, (2015) "Harmonic Phase Estimation in Single-Channel Speech Enhancement Using Phase Decomposition and SNR Information", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 23.
[13] Swati R. Pawar, Hemant kumar B. Mali, (2015) "Implementation of Binary Masking Technique for Hearing Aid Application", IEEE International Conference on Pervasive Computing, 2015.
[14] Meng Sun, Xiongwei Zhang, Hugo Van hamme, (2016) "Unseen Noise Estimation Using Separable Deep Auto Encoder for Speech Enhancement", IEEE Transactions On Audio, Speech, And Language Processing, Vol. 24, No. 1.
[15] Renjith S, Manju K G, (2017) " Speech Based Emotion Recognition in Tamil and Telugu using LPCC and Hurst Parameters", International Conference on circuits Power and Computing Technologies, pp. 4967-4973.
[16] Wang H,, Ye Z., (2018), " A Speech Enhancement System For Automotive Speech Recognition With A Hybrid Voice Activity Detection Method", International Workshop on Acoustic Signal Enhancement, pp. 456-460.
[17] Sean U. N. Wood, Johannes K. W. Stahl, (2019) "Binaural Codebook-based Speech Enhancement with Atomic Speech Presence Probability", IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 01-12.

