A Review on Improvement in Speech Quality by Codebook Based ANN Technique

Lalita Devi
M.Tech Scholar
Computer Science & Engg. Department
R. P. Inderaprastha Institute of Technology, Kurukshetra

Er. Parmeet Kaur
Asst. Professor
Computer Science & Engg. Department
R. P. Inderaprastha Institute of Technology, Kurukshetra

Abstract*-. *Speech enhancement is an important stage to improve the perceptual quality of a noisy speech signal. The core problem in speech enhancement is the separation of speech and noise, for which a commonly deployed technique is estimating and removing the noise spectrum from the input noisy speech spectrum. Unseen noise estimation is a key yet challenging step to make a speech enhancement algorithm work in adverse environments. At worst, the only prior knowledge we know about the encountered noise is that it is different from the involved speech. 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 codebook based ANN iteratively. In this work, it provides speech enhancement under different noisy conditions. All simulations will be done in MATLAB tool.

***Keywords*-Speech Processing, Speech Enhancement, etc .**

I. INTRODUCTION

In vehicles, it is common to use a telephone hands free accessory. The main motivation for this is to facilitate both hands on the steering wheel when driving. The drawback is that the telephone microphone is situated at a farther distance (50 cm) from the speaker's mouth as compared to handheld telephony, in a high noise level environment. To increase the SNR and to allow for the listener to grasp the speech clearly, a speech enhancement method should be applied. The use of systems involving speech-based communication technology is now ubiquitous; such systems include mobile phones, hearing aids and video-conferencing technology. The perceived quality, and in more severe cases the intelligibility, of the speech signal in these systems is reduced when they are used under the adverse noise conditions encountered in real environments such as offices, crowded public spaces, or railway stations [1].

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. 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. Fricative sounds are formed by constricting the vocal tract and passing air through it, causing turbulence those results in a noise-like sound. Plosive sounds are created by closing up the vocal tract, building up air behind it then suddenly releasing it. A speech signal can be considered as a linear composite of the above three classes of sound, each of these sounds are stationary and remain fairly constant over intervals of the order of 30 to 40 ms [2].

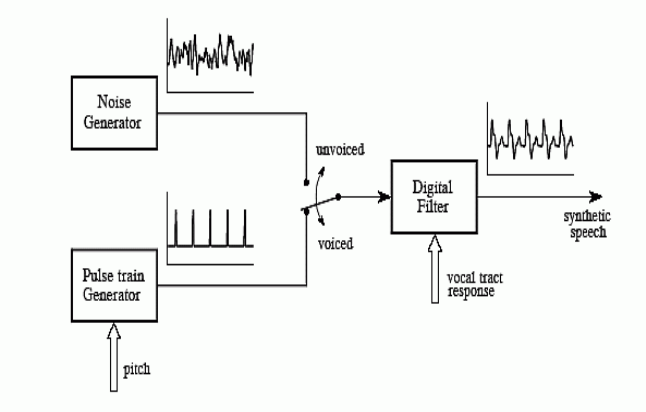

Fig 1: Process of Speech Recognition [1]

Speech sounds can be broadly divided into two categories: voiced and unvoiced. Voiced sounds are produced when the vocal folds are vibrating, producing a quasi periodic signal, while unvoiced sounds are articulated without vibration of the vocal folds. Speech consists of a sequence of vowels and consonants together with brief silences between phonemes and words. Vowels are created by a voiced sound without any constriction in the vocal tract. Estimation of a clean speech signal from a noisy recording is a typical signal estimation task. But due to the non-stationary of the speech and most of the practical noise signals, and also due to the importance of the problem, significant amount of research has been devoted to this challenging task. Single-channel speech enhancement algorithms e.g. use the temporal and spectral information of speech and noise to design the estimator. In this case, only the noisy recording obtained from a single microphone is given while the noise type, speaker identity or speaker gender is usually not known.

The paper is ordered as follows. In section II, we discuss related work of speech processing. In Section III, It defines. In section IV, it describes the problem related to work. Finally, conclusion is explained in Section V.

II. LITERATURE REVIEW

Fukane A. et. al. (2011) [5] revealed a presentation assessment of Spectral subtraction calculation and its adjusted renditions for Hearing guides in various situations, for example, café, train and Car conditions. Clean discourse signals were tainted by foundation commotion individually multi-talker chatter clamour, train commotion, and motor clamour at four distinctive sign to-clamour proportion levels - 2dB, 0 dB, 5 dB,10 dB. Abstract and target type assessment of upgraded discourse signals were done. The assessment of understand ability and nature of improved discourse was accounted for portable amplifiers.

Plourde E. et. al. (2011) [6] examined a multidimensional Bayesian STSA estimator that expect corresponded otherworldly parts. Since the shut structure arrangement of this ideal estimator isn't promptly accessible, it on the other hand inferred shut structure articulations for an upper and a lower bound on the ideal estimator. Assessment results from both goal and abstract discourse quality measures demonstrated that at moderate to high SNR values, where ghastly relationship of discourse was generally recognizable, the proposed estimators can accomplish critical upgrades over the Wiener channel estimators.

Christian D. et. al. (2012) [7] introduced that the objective of monaural discourse improvement was to isolate a solitary blend into its fundamental clean discourse and interferer parts. This under-decided issue was illuminated by fusing earlier information as scholarly discourse and interferer lexicons. Upgrade execution was estimated utilizing target gauges and is restricted by two impacts. A too scanty coding of the blend makes the discourse part be clarified with too not many discourse lexicon molecules, which prompts an estimate mistake we signify source mutilation.

Veisi H. et. al. (2012) [8] grew new voice action discovery (VAD) calculation with delicate choice yield in Mel-recurrence area dependent on shrouded Markov model (HMM) and was fused in a HMM-based discourse improvement framework. As the principle motivation behind the proposed VAD was to recognize discourse portions precisely, a hang-over instrument is proposed and is applied on the yield of the VAD to improve the discourse identification rate. The VAD was incorporated in the HMM-based discourse improvement framework in Mel-recurrence phantom (MFS) and cepstral (MFC) spaces.

Yousheng X. et. al. (2014) [9] 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 clamour 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) [10] 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) [11] 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) [12] introduced a Codebook-based discourse upgrade (CBSE) utilizing prepared discourse and clamour 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) [13] 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) [14] 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) [15] 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.

Wood S. et al. (2019) [16] presented a universal codebook-based speech enhancement framework that relies on a single codebook to encode both speech and noise components. The atomic speech presence probability (ASPP) is defined as the probability that a given codebook atom encodes speech at a given point in time. show that the proposed ITF-based ASPP approach achieves a good balance of the trade-off between binaural noise reduction and binaural cue preservation.

III. SINGLE CHANNEL SPEECH ENHANCEMENT

In general, speech enhancement methods can be categorized into two broad classes: unsupervised and supervised. In unsupervised methods such as Wiener and Kalman filters and estimators of the speech DFT coefficients using super-Gaussian priors, a statistical model is assumed for each of the speech and noise signals, and the clean speech is estimated from the noisy observations without any prior information on the noise type or speaker identity. Hence, no supervision and labeling of signals as speech or a specific noise type are required in these algorithms. For the supervised methods, e.g., on the other hand, a model is considered for both the speech and noise signals and the model parameters are learned using the training samples of that signal. Then, an interaction model is defined by combining speech and noise models and the noise reduction task is carried out [3].

The speech signal degradations may be attributed to various factors; viz. disorders in production organs, different sensors (microphones) and their placement (hands free), acoustic non-speech and speech background, channel and reverberation effect and disorders in perception organs. Considerable research recently has examined ways to enhance speech, mostly related to speech distorted by background noise (occurring at the source or in transmission)-both wideband noise and narrowband noise, clicks, and other non-stationary interferences. Most cases assume noise whose pertinent features change slowly (i.e., locally stationary over analysis frames of interest), so that it can be characterized in terms of mean and variance (i.e., second-order statistics), either during non-speech intervals of the input signals or via a second microphone (called reference microphone) receiving little speech input.

1. Weiner Filter

Wiener filtering is one of the oldest approaches that is used for noise reduction. In the following, we review the Wiener filter in the discrete Fourier transform (DFT) domain, in order to introduce the notation. Let us denote the quantized, time domain noisy speech, clean speech, and noise signals by y, s, and n, respectively. Also, denote the sample index by m. For an additive noise, the signal model is written as eq. (1):

$Ym = Sm + Nm$ ..... (1)

To transfer the noisy signal into the frequency domain, data is first segmented into overlapped frames, and then each frame is multiplied by a tapered window (such as Hamming window) to reduce the spectral leakage, and then DFT is applied to the windowed data. The signal is then processed in the DFT domain and the enhanced signal is reconstructed by using the overlap-add framework. The frame length is usually between 10 and 30 ms, and the speech signal within each frame is assumed to be stationary. Let k and t represent the frequency bin and short-time frame indices, respectively. Wiener filtering is a linear minimum mean squared error (LMMSE) estimator that is a special case of the Bayesian Gauss–Markov theorem. Using the Wiener filter, the clean speech DFT coefficients are estimated by an element-wise product of the noisy signal yt and a weight vector ht is represented by eq. (2):

$St = Ht * Yt$ ..... (2)

To obtain the weight vector ht, the mean square error (MSE) between the clean and estimated speech signals is minimized. Assuming that different frequency bins are independent, it can minimize the MSE for each individual frequency bin k separately.

2. Kalman Filter

Although the time-varying Wiener filter is optimal in the sense of mean square error for a given short-time frame, it does not use the prior knowledge about the speech production. For example, the temporal dependencies are not optimally used in the Wiener filtering. Therefore, Kalman filtering and smoothing have been presented in the literature to improve the performance of the noise reduction algorithms.

A. Speech Enhancement Techniques

The approach to speech enhancement varies considerably depending upon type of degradation. The speech enhancement techniques can be divided into two basic categories: (i) Single channel and (ii) Multiple channels (array processing) based on speech acquired from single microphone or multiple microphone sources respectively. In most cases the background random noise is added with the desired speech signal and forms an additive mixture which is picked up by microphone. It can be stationary or non-stationary, white [3] or colored and having no correlation with desired speech signal.

1. DFT Based Methods

They are most popular as they have less computational complexity and easy implementation. They use short time DFT and have been intensively investigated; also known as spectral processing methods. They are based on the fact that human speech perception is not sensitive to spectral phase but the clean spectral amplitude must be properly extracted from the noisy speech to have acceptable quality speech at output and hence they are called short time spectral amplitude based methods.

2. Wavelet Based Methods

The DFT based methods use short time spectral measurements and hence are suffered by time-frequency resolution trade-offs. Wavelet based methods are developed which provides more flexibility in time-frequency representation of speech. The Wavelet denoising algorithm is most commonly used and based on soft thresholding of the Wavelet coefficients. However uniform thresholding results in suppression of noise as well as unvoiced components of desired speech.

3. KLT Based Methods

The frequency domain methods are nowhere close to offering fully satisfactory solutions to their inherent problems: the musical noise artifact and the inevitable trade-off between signal distortion and the level of residual noise. It uses the singular value decomposition (SVD) of a data matrix to remove the noise subspace and then reconstruct the desired speech signal from the remaining subspace.

4. Adaptive Filtering

The adaptive filters which are mostly used in adaptive control applications can also be useful for speech enhancement. Mostly LMS and its variants are useful in multi microphone additive noise and echo cancellation problems. But for single channel speech enhancement Kalman and H∞ adaptive filters are found suitable. They can also address the problem of colored noise removal as the noise is not always white is real environments

IV. PROBLEM FORMULATION

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 codebook based speech enhancement signal to improve noise reduction.

V. CONCLUSION

Speech signals can be degraded in many ways during their acquisition in noisy environments and they can also be further degraded in the electronic domain. Serious 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. Due to this, it proposes a method for speech enhancement using mask estimation iteratively. In this work, it provides speech enhancement under different noisy conditions.

REFFERENCES

[1] J. S. Lim and A. V. Oppenheim, (1979) "*Enhancement and bandwidth compression of noisy speech*," Proc. IEEE, vol. 67, no. 12, pp. 1586–1604.

[2] P. Vary and R. Martin, (2005) "*Digital Speech Transmission: Enhancement, Coding and Error Concealment*", Chichester, U.K.: Wiley.

[3] P. Loizou, (2005), *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC.

[4] S. Doclo and M. Moonen, (2005) ''*On the output SNR of the speech-distortion weighted multichannel Wiener filter*,'' IEEE Signal Processing Letters, vol. 12, no. 12, pp. 809-811.

[5] Anuradha R. Fukane, L. Sahare, (2011) "*Enhancement of Noisy Speech Signals for Hearing Aids*", International Conference on Communication Systems and Network Technologies, 2011.

[6] E. Plourde, and B. Champagne, (2011) "*Multidimensional STSA Estimators for Speech Enhancement With Correlated Spectral Components*", IEEE Transactions On Signal Processing, Vol. 59, No. 7.

[7] Christian D. Sigg, Tomas Dikk, (2012) "*Speech Enhancement Using Generative Dictionary Learning*", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 6.

[8] 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.

[9] Xia Yousheng, Huang Jianwen, (2014) "*Speech Enhancement Based on Combination of Wiener Filter and Subspace Filter*", IEEE, pp. 81-86.

[10] Zhang Jie, Xiaoqun Zhao, Jingyun Xu, (2014) "*Suitability of Speech Quality Evaluation Measures in Speech Enhancement*", IEEE pp. 102-108.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] Shoko Arakit, Tomoki Hayashi, (2015) "*Exploring Multi-Channel Features for Denoising-Auto-encoder-Based Speech Enhancement*", IEEE pp. 424-430.

[18] 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.

[19] 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.

[20] 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.

[21] Swati R. Pawar, Hemant kumar B. Mali, (2015) "*Implementation of Binary Masking Technique for Hearing Aid Application*", IEEE International Conference on Pervasive Computing, 2015.

[22] 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.