CHAPTER 3

LITERATURE REVIEW
3.1 Introduction

In general, there exists a need for voice based communications, human-machine/ machine-machine interfaces, and automatic speech recognition systems to increase the reliability of these systems in noisy environments. In many cases, these systems work well in nearly noise-free conditions, but their performance deteriorates rapidly in noisy conditions. Therefore, improvement in existing pre-processing algorithms or introducing entire new class for algorithm for speech enhancement is always the objective of research community. The main requirement for speech enhancement systems varies according to specific applications, such as to boost the overall speech quality, to increase intelligibility, and to improve the performance of voice communication devices.

3.2 Estimation based Filtering Techniques

One of the early papers [1] in speech enhancement considers the problem of estimation of speech parameters from the speech, which has been degraded by additive background noise. In this work they propose the two suboptimal procedures which have linear iterative implementations in order to suppress the non-linear effect on the speech parameters due to background noise. In another similar problem [2] of enhancing the speech in presence of additive acoustic noise, spectral decomposition of frame of noisy speech was adopted. The attenuation of particular spectral component was determined based on how much the measured speech plus noise power exceeds an estimation of background noise leading an importance of proper choice of the suppression or subtraction factors. The short-time spectral amplitude (STSA) was used to model the speech and noise spectral components in [3]. The parametric estimation techniques, where parameters of underlying model, consist of small set of parameters, is determined and then numerical process is used to modify the parameters, can be contrasted by the non-parametric method which can be used as in [4] where no model is assumed and uses non-parametric spectrum estimation techniques.

In application point of view, there is work described in [5], where noisy speech enhancement algorithm has been discussed and implemented to compare its performance against the various levels of LPC (Linear Predictive coefficient) perturbation. Various speech enhancement techniques have been considered here such as spectral subtraction, spectral over subtraction with use of a spectral floor, spectral subtraction with residual noise removal and time and frequency domain adaptive MMSE filtering. The speech signal sued here for recognition experimentation was a typical sentence with additive normally distributed white noise distortion.
The single channel speech enhancement algorithm at very low SNR has been presented in [6], which uses masking properties of human auditory system. This algorithm is the subtractive type in its nature and subtraction parameter is adapted as per the levels of rough estimate of the background noise and the added musical residual noise and thus making this algorithm adaptable to noise present in every frame of speech. In another interesting research [7], speech was enhanced from noise along with coding using discrete wavelet packet transform decomposition. Two stages of subtractive-type algorithm used, once estimating noise and subtracting it from noisy speech to have rough estimate of speech later, this estimate is further used to determine the time-frequency masking threshold assuming high-energy frames of speech will partially mask the input noise and hence reducing the need for a strong enhancement process. The both of these work used Noisex-92 database to evaluate the performance of their proposed algorithms. In yet another similar work [8], the noise autocorrelation function is estimated during non-speech activity periods and it is used in deciding the masking threshold for the speech enhancement. Here, author also uses frequency to Eigen-domain transformation to provide the upper bound estimate of residual noise to be introduced in the speech.

It is believed that the time distribution of speech samples is much better modelled by a Laplacian or a Gamma density functions rather than a Gaussian density function. The same is valid for short time DFT domain, typically, frame size less than 100ms [9]. Optimal estimators for speech enhancement in the Discrete Fourier Transform (DFT) domain is used for estimating complex DFT coefficients in the MMSE sense when the clean speech DFT coefficients are Gamma distributed and the DFT coefficients of the noise are Gaussian or Laplace distributed. When the noise model is a Laplacian density, this estimator outperforms other estimators in the sense it show less annoying random fluctuations in the residual noise than for a Gaussian density noise. In [10] and [11], adaptive estimation of non-stationary noise present in the speech has been presented.

3.3 Beamforming based Speech Enhancement

Frost [12] has suggested constrained minimum power adaptive beamforming, which deals with the problem of a broadband signal received by an array, where pure delay relates each pair of source and sensor. Each sensor signal is processed by a tap delay line filter after applying a proper time delay compensation to form delay-and-sum beamformer. The algorithm is capable of satisfying some desired frequency response in the look direction while minimizing the output noise power by using constrained minimization of the total output power. This minimization is realized by adjusting
the taps of the filters under the desired constraint using constrained LMS-type algorithm. Griffiths and Jim [13] reconsidered Frost’s algorithm and introduced the generalized sidelobe canceller (GSC) solution. The GSC algorithm is comprised of three building blocks. The first is a fixed beamformer, which satisfies the desired constraint. The second is a blocking matrix, which produces noise-only reference signals by blocking the desired signal (e.g., by subtracting pairs of time-aligned signals). The third is an unconstrained LMS-type algorithm that attempts to cancel the noise in the fixed beamformer output. In [13], it is shown that Frost algorithm can be viewed as a special case of the GSC. The main drawback of the GSC algorithm is its delay-only propagation assumption.

In another work [14], switching adaptive filters were used to form the beamformer. This beamformer has two sections and interconnected with switch. The first section determines the adaptive look direction and cues in on the desired speech and is adapted only when speech is present. Second section which adapted during silence-only periods is implemented as multichannel adaptive noise canceller. In [15], authors have proposed the solution to GSC algorithm by estimating ratio of transfer functions (TFs), otherwise it is based on TFs which relates source signal and the sensors. The TF ratios are estimated by exploiting the non-stationarity characteristic of the desired signal. This algorithm can be used normally in reverberating room having acoustic environment. One interesting paper [16], describes how optimal finite-impulse response subband beamforming can be used by including coherent multipath propagation into optimality criterion for speech enhancement in multipath environment.

In application point of view, a constrained switched adaptive beamforming (CSA-BF) [17] was used for speech enhancement and recognition in real moving car environment. This algorithm consists of a speech/noise constraint section, a speech adaptive beamformer and noise adaptive beamformer. The performance obtained with this algorithm was compared with classic delay-and-sum beamforming (DASB) using CU-Move corpus and found decrease in word-error-rate (WER) by 31% in speech recognition. The computational complexity of DASB is very low and can be easily implemented for real-time requirement. It is also effective when direction of desired source is known and can be applied in the car as driver’s head position is restricted based on seat position. However, as there is possibility of change in drivers head direction, DASB algorithm could be inconsistent and this inconsistency can be solved by employing CSA-BF algorithm which can improve the SNR by up to 5.5 dB on the average. For the application of hands-free speech recognition, one of the works [18] uses sequence of features to be used for speech recognition itself, to optimize a filter-and-sum beamformer instead of separating the beamformer, to be used for speech enhancement, from speech recognition system. In this work, they used Mel Frequency Cepstral Coefficient (MFCC) and applied to the HMM based classifier for speech recognition.
Optimizing beamformer without knowledge of source or acoustic characteristic of environment is termed as "blind beamforming". One of the papers [19] proposes blind speech enhancement using beamformer which consist of subband soft-constrained adaptive filter using recursive least square (RLS) algorithm, combined with subband weighted time-delay estimator (TDE). Estimation of propagation time difference of arrival of a dominate speech source received by sensor array is based the steered response power with phase transform (SRP-PHAT) algorithm, which was modified to work in subband structure. One recent paper [20] presents phase-based dual-microphone speech enhancement technique based on prior speech model. In this work, it is claimed that around 23% improvement achieved using this algorithm as compared to the delay-and-sum beamformer, where experiments were conducted on the CARVUI database.

In application point of view, the study presented in [21] addresses the problem of distant speech acquisition in multiparty meeting s using multiple cameras and microphones. The camera, used as a multi-person tracker, was used to give the more precise location of each person to the microphone array beamformer. They evaluated the performance of speech recognition using data recorded in a real meeting room for stationary speaker, moving speaker and overlapping speech scenarios. The result obtained with audio-video speech enhancement was better than that with only audio. In one of the recent work [22], adaptive beamformer based on estimation of power spectral density (PSD) and noise statistics update was proposed. An inactive-source detector based on minimum statistics is developed to detect the speech presence and to acquire the noise statistics. The performances of this beamformers were tested in a real hands-free in-car environment. One of the most recent papers [23] uses GSC based speech enhancement using the location of speaker obtained via localization module. This algorithm relies on time delay compensation, DFT computations, fixed channel compensator, adaptive channel compensator.

3.4 Active Noise Cancellation

It is believed that if error sensor output is measured properly and mapped to the control speaker even with some propagation delay, then essential problem for the active noise cancellation structure is to predict the future values and/or components of noise. The work presented in [24], deals this problem with single sensor and predicts the noise model parameters with Kalman filter using non-gradient algorithm and gradient search algorithm. These two algorithms were applied on noise generated by a propeller aircraft, a helicopter and jet aircraft. The noise was reduced significantly, though; it was most in propeller case and least in jet noise. The gradient algorithm has
also less computational complexity over the non-gradient algorithms. Another way of having active noise cancellation is to employ adaptive filter to characterize the transfer function between error mic to control noise in frequency domain and it is described in the [25]. The least mean square (LMS) algorithm was applied in each band of frequency decomposition using DFT to model the control signal. Further, a frequency-domain periodic active noise equalization (ANE) system, which reshapes the residual noise by controlling the output of the adaptive comb filter at each frequency bin, is also presented in this work.

In real-life application point of view, one of the interesting papers [26] presents the integrated feedback based approach for noise reduction headset for the audio and communication purposes. This system uses single microphone per ear cup which makes it simple in its applicability with existing audio and communication devices. In another work [27] of designing ANC headset, filtered-x least means square algorithm (FXLMS), which introduces secondary path for synthesizing the reference signal was implemented. One of the study [28] based on FXLMS analysed the algorithm and extended the new algorithm to reduce the multi-tonal noise. Similar work was developed in [29] and evaluated with narrowband and broadband noise against the instability in convergence of adaptation algorithm. The modified FXLMS algorithm based on efficient affine projection technique is compared with conventional FXLMS algorithm in [30]. It is claimed that modified filtered-x structure provides better convergence speed over conventional filtered-x algorithm but need more computations comparatively as it requires additional filtering channel.

### 3.5 Noise Types and Sources

The noise reduction is any active area of speech processing where noise is being reduced to achieve the noise free speech. This becomes critical in so many applications including speech recognition where main objective is to reduce the word error rate (WER) and at the same time providing flexibility to use system in anywhere irrespective of which kind of noise present in acquired speech. To design the efficient and generalized algorithm for speech enhancement, it is critical to have knowledge about the noise that could be present in speech. Mathematically, noise can be sometimes modelled by its probability distribution of values and represented in some work [31] by Gaussian distribution or Laplacian types or Gama types distributed speech in case speech modeling. The level of SNR can be important indicator to the strategy to be employed in enhancement algorithms. In the presence of high level SNR and low level SNR, algorithm can be switched from one enhancement scheme to another after estimating the SNR present in speech.
Depending of the stationarity characteristics of noise, the technique is to be designed to handle stationary and non-stationary noise separately.

There are some noises, whose source are known and are easy to be modelled and analyzed. Despite of known sources of noise, it becomes difficult to reduce the noise when more than one type of noises is present. In applications like noise free headset, it may become easy to overcome the problem of noise as ears are isolated from external region in large extent. In car application, noises generated by engine, wind and rain wipers are to be handled intelligently. Hands free devices have echo voice as most dominant noise. In close room meeting application, acoustic characteristics of walls can generate multiple echoes of voice. In crowd or outdoor meetings, lot many people may talk simultaneously and can overlap with principle speech. In traffic area, noise may be raised from other vehicles and impulsive honking. In car or at home sometimes background music can become source of noise for the speech of interest. In hands-free operation of cellular phones in vehicles, the speech signal to be transmitted may be contaminated by reverberation and background noise.

3.6 Real Time Speech Processing

Some of the definitions of real time speech processing [32] found from the different dictionaries and researchers are given below,

- In real-time speech processing system the correctness of the computations not only depends upon the logical correctness of the computation but also upon the time at which the reconstructed speech is produced. There is occurrence of system failure if the timing constraints of the system are not met. (gillies @ ee.ubc.ca).
- It is the time in which the occurrence and the reporting or recording of an event are almost simultaneous.
- The actual time used by a computer to solve and control speech processing problems effectively at the same time (Webster's dictionary) (Agnes and Guralnik, 2000)
- Real time speech processing requires a fast enough data processing to maintain with an outside process, it is a form of transaction processing in which each transaction is executed as soon as complete data becomes available for the transaction. (http://www.wordwebonline.com).
There are basically two types of real-time systems 'soft' and 'hard' as shown in figure 3.1. Soft real-time system means a system which has reduced constraints on 'tardiness' but still must operate very quickly. In hard real-time system the type of a typical real-time system which requires a stringent deadline (http://media.wiley.com).

![Real Time Systems Diagram](image)

Fig. 3.1 Types of real time systems

Considering several definitions mentioned above, we can say a common basis of a real-time speech processing system requires explicit bounded response time constraints without system failure, with the logical correctness based on both the correctness of the reconstructed speech outputs and their timeliness. The response time is called the time between the presentation of a set of speech inputs and the appearance of all the associated reconstructed speech outputs. In one of the paper real time speech processing strategies are explores for different applications.

### 3.7 Performance measurement of Real Time Systems

A real time system performance is a measure of the percentage of non-idle processing often measured by CPU utilization. A system is said to be time overloaded if it is 100% or more time-loaded. Systems that are time-overloaded are unstable and exhibit missed deadlines and unpredictable response times.

---

Table 3.1 shows CPU utilization for real-time systems. Utilization factors in the 0%-69% range are generally considered as safe. Beyond 70% they have a high risk of missing deadlines, and above 100% are potentially terrible.

<table>
<thead>
<tr>
<th>Utilization Percentage (%)</th>
<th>Type of Zone</th>
<th>Applications Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 25</td>
<td>Excess</td>
<td>Different</td>
</tr>
<tr>
<td>26 - 50</td>
<td>More safer</td>
<td>Different</td>
</tr>
<tr>
<td>51 - 68</td>
<td>Safe</td>
<td>Different</td>
</tr>
<tr>
<td>69</td>
<td>Theoretical limit</td>
<td>Different</td>
</tr>
<tr>
<td>70 - 99</td>
<td>Dangerous</td>
<td>Embedded systems</td>
</tr>
<tr>
<td>100 and above</td>
<td>Overload</td>
<td>Strained systems</td>
</tr>
</tbody>
</table>

3.8 Conclusion

The speech, being a fundamental way of communication for the humans, has been embedded in various essential applications like speech recognition, voice-distance-talk and other forms of personal communications. There are so many applications of speech still to be far from reality just because of lack of efficient and reliable noise removal mechanism and preserving or improving the intelligibility for the speech signals. The broad categories of speech enhancement techniques can be listed as speech filtering techniques, beam forming techniques and active noise cancellation methods. In this chapter, an attempt has been stepped towards surveying the methodologies for speech improvement. It was also interesting to discuss, how these techniques affect the performance of various application systems like speech recognition and speech communication. Essentially, we also discussed about the types and sources of noise that can be considered in speech enhancements. In last phase we also discussed about definitions of real time speech processing and the performance measures for real time systems.