CHAPTER 1

Introduction

"To invent, you need a good imagination and a pile of junk." - Thomas A. Edison

Human uses speech to communicate messages. The listener's ability to comprehend speech and the amount of fatigue during listening depend on the intelligibility and quality of speech. It is not always possible to ensure a noise free background during talking. Electronic speech communication is no longer confined to quieter, home or office environment. People today communicate speech from places as diverse as train, airport, market place, factory, restaurant etc. Each of this has different kind of noisy background and can include environments such as a very noisy cockpit of a fighter aircraft. Electronic communication of speech through a transmission media is influenced by the character of the media. The transmission media introduces distortions and generate noisy speech signals. Speech enhancement algorithms help to reduce background noise and improve the perceptual quality or intelligibility of speech signals. In a moderately adverse environment though speech is perceptible, but applications such as automatic speaker recognition, distributed speech recognition, etc., drive the effort to build adaptive noise suppression algorithms. The goal of speech enhancement varies according

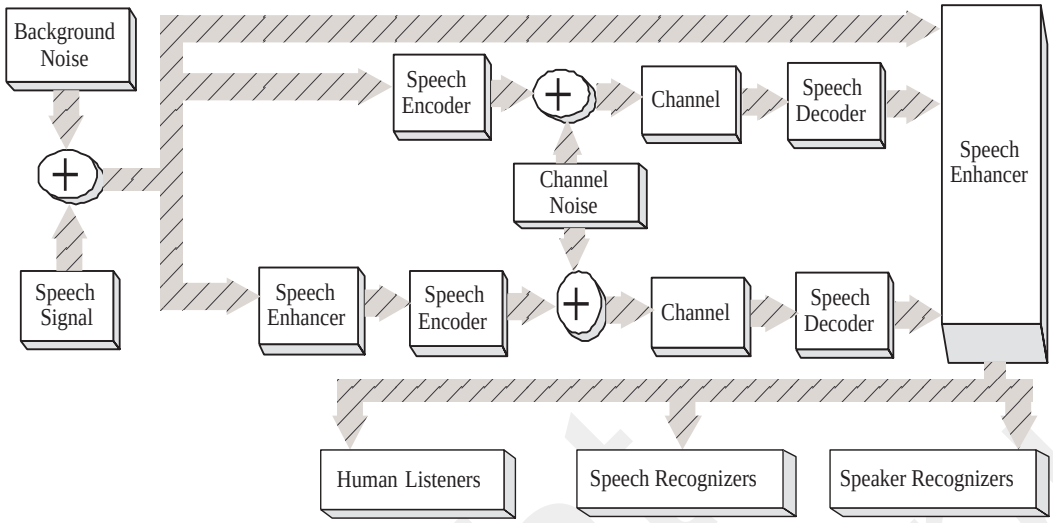

Figure 1.1: Application areas for Speech Enhancement

to specific applications, such as to boost the overall speech quality, to increase intelligibility, and to improve the performance of voice communication devices [1], [2]. Computational complexity and the trade off between the amount of noise suppression and speech degradations of the adaptive algorithms are also of concern for real world speech based applications. The insertion of speech enhancer into various application areas are shown in Fig. 1.1.

1.1 Literature Review

This section presents reviews on the production of speech in humans, various noise signals and different speech enhancement strategies.

1.1.1 Fundamentals of Speech Production

Speech is an acoustic waveform and is also known as a dynamic information-bearing signal. The speech is generated in the mouth of the speaker due to sound pressure. But the way of production and perception still remain a mystery to the researchers [3]. Researchers need good knowledge of these processes to develop efficient methods to represent and transform the acoustic waveforms to achieve

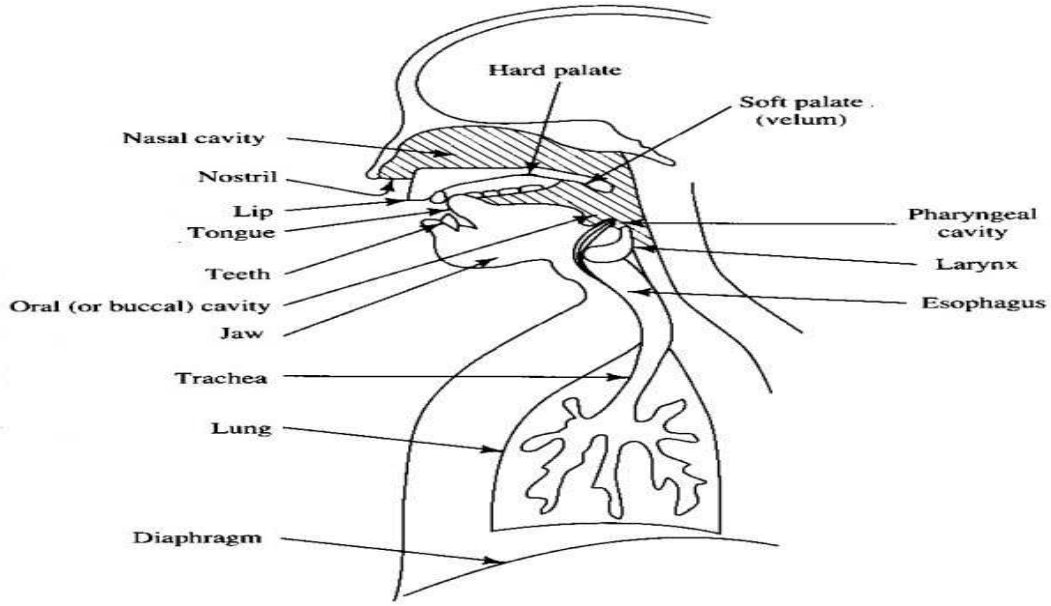

Figure 1.2: Speech Production System (from [4])

the desired accuracy. The basic human speech production mechanism consists of, lungs, trachea, larynx, pharyngeal cavity, buccal cavity, nasal cavity, velum, tongue, jaw, teeth and lips [3]. The respiratory subsystem of the mechanism build up with the help of lungs and trachea. When the air is expelled from the lungs into the trachea, this subsystem acts as the energy source for speech waveform. The resulting airflow passes through the larynx, which provides periodic excitation to the system to produce the voiced, unvoiced, mixed, plosive, whisper and silence waveforms [3]. The acoustic filter then shapes the generated waveform. The articulators i.e. velum, tongue, jaw, teeth and lips, give the finer adjustments to generate speech waveform [3]. The basic speech production system is shown in Fig. 1.2.

1.1.2 Fundamentals of Noise Signal

In speech enhancement, the behavior of the noise is very crucial. Therefore, an accurate noise model is important for the enhancers [5]. It is also important how well an enhancer works with different types of noise having different statistical and spectral properties. Based on the behavior and properties of the noise sources,

noise can be classified as in Table 1.1. Examples of different noise types based on source of noise is listed next.

Table 1.1: Classification of Noise based on various properties (from [5])

Structure	Continuous / Impulsive / Periodic
Type of Interaction	Additive / Multiplicative / Convolutive
Temporal behavior	Stationary / Non-stationary
Frequency range	Broadband / Narrowband
Signal dependency	Correlated / Uncorrelated
Statistical properties	Dependent / Independent
Spatial properties	Coherent / Incoherent

- Background: additive in nature, usually uncorrelated with the signal and present in various environment scenarios [5].
- Multi-talker: additive in nature, composed of single or multiple "competing" speakers. The "cocktail party effect" is multi-talker babble noise with very similar characteristics and frequency range to the speech signal of interest [5].
- Impulse: slamming of doors, noise present in archived gramophone recordings [5].
- Non-additive: due to non-linearities of microphones, speakers and channel distortion [5].
- Non-additive due to speaker stress: due to Lombard effect [6]. This results in speech having different spectral properties as compared to clean speech, a detailed summary of the changes in speech characteristics due to this effect are given in [7].
- Correlated with the signal: reverberations and echoes [5].
- Convolutive: corresponds to convolution in time domain. Also, changes in speech signal due to changes in room acoustics or changes in microphones etc [5].

- Multiplicative: due to fading in cellular channels [5].

1.1.3 Classification of Speech Enhancement Techniques

Speech enhancement techniques can be classified [5], [8] as shown in Table 1.2.

Table 1.2: Speech enhancement processing strategies (from [5])

Number of input channels	One / Two / Multiple
Domain of processing	Time / Frequency / Time-Frequency
Type of algorithm	Non-adaptive / Adaptive
Additional constraints	Speech production / Perception

The speech signal usually captured from single or multi-channel microphones and various types of noise can make speech enhancement algorithms difficult to deal with. The single channel microphone is a cost effective solution whereas the dual microphone approach is a costly solution to this speech enhancement problem. Spatial analysis can give useful information in multi-channel input but further increases the cost. The noise source is assumed to be statistically independent and additive [3]. The base of such assumption is the fact that most environmental noise is typically additive in nature [3]. The proposed methods in the present thesis work will be based on single channel enhancement techniques.

1.2 Previous Approaches : An Outline

Over the past three decades, variety of speech enhancement methods have been used to suppress background noise signals. Boll proposed spectral subtraction [9], Virag tackled the problem by utilizing masking properties of the human auditory system [5], Knecht *et al.* proposed noise reduction by artificial neural networks [10] and Dahl *et al.* used microphone array to suppress the car noise [11]. From the point of view of the number of microphones considered, the first two algorithms are single channel the other two are multi-channel approaches.

For single channel approaches, the [9]-[13] are shown to be very effective for the noise suppression. The Gaussian modeling of both speech and noise spectral

components have been reported and was successfully combined with the Minimum Mean Square Error (MMSE) estimator in speech enhancement systems [12], [13]. But this approach is limited to the large Discrete Fourier Transform (DFT) frames [12], [13] where the span of correlation of the signal under test is much shorter than the DFT frame length. The pdf of speech in the time domain and in the frequency domain is much better modeled by a Laplacian or a Gamma distribution rather than a Gaussian [14]-[18]. In past decade, extensive research on non-Gaussian modeling of speech samples has been reported [14]-[18]. The work in [15], [16], are MMSE based estimation using Laplacian and the Gamma modeling of speech and noise. However the method in [16], uses the fixed distribution parameters of the Gamma modeling. This indeed limits the application in general cases. The main reason is that the derivations of the MMSE estimate of the magnitude spectrums are computationally expensive. Alternative solutions were also explored in [17]-[20]. In [17], the pdf of the amplitude and phase of the DFT coefficients was approximated with a parametric function to derive joint MAP estimator. Martin *et al.* [18] use the super-Gaussian speech priors based MMSE estimation to estimate magnitude-squared DFT Coefficients. In [19], statistical speech feature enhancement in the cepstral domain is presented. The algorithm exploits joint prior distributions in the clean speech model. A noncausal estimator for the a priori SNR is proposed in [20]. A multi-band spectral subtraction with adjusting subtraction factor is given in [21]. E. Zavarehei *et al.* enhances the speech using Kalman Filtering for restoration of short time DFT trajectories [22]. In [23], the corrupted cepstrum is assumed to follow the Gaussian mixture as is the case with the clean spectrum. The non-linearity imposed on the clean spectrum pdf is approximated by a Taylor series. Extension of this work to model space is presented in [24]. The efficient MAP [25] and Maximum Log-Likelihood ratio [26] techniques require rather long adaptation data in order to obtain good estimates for the modification of the clean speech model set. In [27], work on multiple statistical models for soft decision in noisy speech enhancement is presented.

The Discrete Cosine Transform (DCT) has been found to perform better than the DFT [28] in speech enhancement because of its higher speech energy compaction and faster implementation than DFT [28], [29]. The amplitude estimation is then obtained by the Gaussian distribution assumption for both the noise and

clean speech signal. However in [29], for DCT spectra, the Laplacian distribution is shown to be more effective than the Gaussian distribution. In [30], Gazor *et al.* also investigated the clean speech signal's distribution by Karhunen-Loeve Transform and DCT. The statistics of DCT coefficients of the clean speech signal are shown to be similar to Laplacian pdf excluding silence intervals [30]. This is analogous to the pdf of the DCT coefficients of image signals in [31]. But the DCT has some inherent drawbacks such as it is not shift invariant, not noise robust, compared to DFT, it is also not cyclic and the transformation matrix is not Vandermonde. Independent Component Analysis (ICA) have also been used in separating original independent source signals from the observed mixed signals [32]. Jong *et al.* have extracted speech features by using ICA and Gabor filter [32] and used those features with MAP estimation for speech enhancement. In [33], the environmental noise corrupted speech signals were taken using array of microphones. Rosca *et al.* tackled the problem where parts of the time-frequency content of the speech signal are missing [34]. Speech enhancement method using ICA and delay-sum beamforming is presented in [35]. In [36], a single channel algorithm is proposed with the Wiener filter and the ICA technique. Various other approaches such as hidden markov modeling [37], signal subspace methods [38] and wavelet based methods [39]-[41] have also been proposed.

Wavelets are used to analyze different types of signals [39]-[41]. The Wavelet Transform are used to separate the speech signal and noise components efficiently. The speech wavelet coefficients are localized, but the noise coefficients are distributed in nature. Therefore the energy of the speech signal will be limited to few number of coefficients than the noise. The magnitude of the noise wavelet coefficients are then suppressed by different thresholding schemes and the remaining coefficients will be representing the speech signal data. The transform can then be inverted to reconstruct the speech signal [42]-[44]. The main problem with the Discrete Wavelet Transform (DWT) is that it is not translation invariant. Therefore, if the test signal is displaced by one sample then the corresponding wavelet coefficients do not displace by the same. The shape of the reconstructed signal will depend on the amount of signal translation. The orthonormal wavelets has been devoted to analyze their suitability on speech signals in very few literature due to their inherent complexities [42], [43]. The wavelet packet expansions [44] have

also become very popular in the signal processing domain. The original paper on best basis search [45] exploited the fact that 1-D wavelet packet bases [44] and local cosine bases [46] can be organized on a single dyadic tree. This work [45] is proposed using an entropy based cost function. Since then, a number of papers have proposed different criteria, e.g., Minimum Description Length principle [47], [48], Bayesian estimation [49], and rate-distortion framework [50]. To address the issues of the entropy based basis search [51], new class of algorithms have been studied in [52] and also in [53].

Several subspace filtering approaches have also been studied that uses Singular Value Decomposition (SVD) [54]-[56]. This approach requires the estimation of the dimensions of the clean speech signal and the information of the noise. Like spectral subtraction method, this technique also assumes that the noise is additive and uncorrelated with the clean speech signal. The SVD technique enhances a noisy signal by keeping a few of the singular values and ignores those which are associated with the noise signal. The enhanced signal is reconstructed from the reduced rank matrix. There are three frequently cited features to justify the success of the SVD [54]: (i) the information on the data generating mechanism is completely contained in certain subspaces of the data matrix, (ii) the complexity is given by the approximate rank of the data matrix and (iii) since the data are corrupted by additive noise it is expected that the SVD has a certain noise filtering effect due to low rank speech model. The minimum variance estimation gives the best linear estimate of the clean data [54], [57]. The superior noise reduction capabilities of subspace filtering are confirmed by several studies by least squares estimate [58] and with the principle optimization criteria [55].

1.3 Objectives and Scope

The investigation carried under this dissertation attempts:

- to design novel speech enhancement algorithms for miscellaneous single channel recording of speech signals and noise types.
- to exploit special properties of LGW in speech enhancement through an analysis - synthesis frame work.

- to use Bayesian Marginal Statistics, Bayesian Joint Statistics and Speech Presence Uncertainty in LGW frame work for speech enhancement task.
- to study the performance of the proposed methods against various objective, subjective and composite measures and compare with existing methods.
- to use the proposed method in ASR application and find its suitability through comparison of performances.

1.4 Contributions Made

This dissertation considers the problem of speech enhancement when only a single microphone is used and when the statistics of the interfering noise and speech are not available a priori. Thus it seeks to address a deficiency of many current enhancement techniques and looks toward a system which would have application in the real world. This dissertation focuses on LGW based LT-SSA estimators using the MAP criterion. The contributions are listed below:

1. Combination of LGW, SIRPs and SPU have been introduced for background noise reduction.
2. LGW and LT-CMS has also been used to reduce the effects of both telephone channel and handsets.
3. The pdf of the filtered speech coefficients in each scale of decomposition is modeled with GLD and corresponding shrinkage rule is proposed by MAP estimator from the statistics of the decomposed noisy speech signals.
4. Circularly Symmetric Probability Density Function (CSpdf) related to the family of SIRPs has been introduced to exploit the inter-scale dependency between the coefficients and their parents and corresponding joint shrinkage estimators are derived by MAP estimation theory.
5. The inter-scale noise variance of the coefficients is kept constant which gives closed form solution.

6. Consideration of SPU estimator is another contribution to the proposed estimator. This idea refines the estimate of the magnitudes by scaling them by the SPU probability.

1.5 NOIZEUS : Noisy Speech Corpus for Evaluation of Speech Enhancement Algorithms

Over the past three decades, different noise suppression algorithms have been used to improve the speech enhancement performance. But, it still remains a challenging task to find a reliable speech enhancement algorithm, where the noise level and noise characteristics are constantly changing [59], [60]. Comparison between various algorithms has also been not possible due to lack of common noisy speech database, different noise types used and different testing methodology [59], [60]. Therefore, it is impossible for researchers to compare the objective and subjective performance of different noise suppression algorithms [59], [60].

NOIZEUS database [61] was developed to facilitate comparison of different speech enhancement algorithms. The database contains 30 IEEE sentences [62], corrupted by eight different noises at four different SNRs (as in Table 1.3). The noise signals were taken from the AURORA database that includes suburban train, babble, car, exhibition hall, restaurant, street, airport and train-station noise. The sentences were produced by three male and three female speakers. The way thirty sentences were selected from the IEEE database ensures inclusion of all phonemes in the American English language [61]. The sentences were originally sampled at 25kHz and then down sampled to 8kHz [61]. Noise signals are artificially added to the speech signal as in [61]: The Intermediate Reference System (IRS) [61] filter is independently applied to the clean and noise signals. The IRS filter is actually a band pass filter whose frequency response corresponds to that for analog part of the transmitter of telephone equipment. The active speech level of the filtered clean speech signal is first determined using the method B of ITU-T P.56 [63]. A noise segment of the same length as the speech signal is randomly cut out of the noise recordings, appropriately scaled to reach the desired SNR level and finally added to the filtered clean speech signal. The long-term spectra of the above

Table 1.3: List of sentences used in NOIZEUS Database(from [61])

Filename	Speaker	Gender	Sentence text
sp01.wav	CH	M	The birch canoe slid on the smooth planks.
sp02.wav	CH	M	He knew the skill of the great young actress.
sp03.wav	CH	M	Her purse was full of useless trash.
sp04.wav	CH	M	Read verse out loud for pleasure.
sp05.wav	CH	M	Wipe the grease off his dirty face.
sp06.wav	DE	M	Men strive but seldom get rich.
sp07.wav	DE	M	We find joy in the simplest things.
sp08.wav	DE	M	Hedge apples may stain your hands green.
sp09.wav	DE	M	Hurdle the pit with the aid of a long pole.
sp10.wav	DE	M	The sky that morning was clear and bright blue.
sp11.wav	JE	F	He wrote down a long list of items.
sp12.wav	JE	F	The drip of the rain made a pleasant sound.
sp13.wav	JE	F	Smoke poured out of every crack.
sp14.wav	JE	F	Hats are worn to tea and not to dinner.
sp15.wav	JE	F	The clothes dried on a thin wooden rack.
sp16.wav	KI	F	The stray cat gave birth to kittens.
sp17.wav	KI	F	The lazy cow lay in the cool grass.
sp18.wav	KI	F	The friendly gang left the drug store.
sp19.wav	KI	F	We talked of the sideshow in the circus.
sp20.wav	KI	F	The set of china hit the floor with a crash.
sp21.wav	SI	M	Clams are small, round, soft and tasty.
sp22.wav	SI	M	The line where the edges join was clean.
sp23.wav	SI	M	Stop whistling and watch the boys march.
sp24.wav	SI	M	A cruise in warm waters in a sleek yacht is fun.
sp25.wav	SI	M	A good book informs of what we ought to know.
sp26.wav	TI	F	She has a smart way of wearing clothes.
sp27.wav	TI	F	Bring your best compass to the third class.
sp28.wav	TI	F	The club rented the rink for the fifth night.
sp29.wav	TI	F	The flint sputtered and lit a pine torch.
sp30.wav	TI	F	Let's all join as we sing the last chorus.

noises are given in [64]. The noise signals were added to the speech signals at SNRs of 0dB, 5dB, 10dB, and 15dB.

1.6 Objective Quality Assessment

The accurate way to evaluate speech quality is through subjective listening tests [60]. But subjective evaluation tests are performed under stringiest conditions which are costly and time consuming [60]. For that reason, the objective measures with high correlation are used to predict speech quality. Many objective speech quality measures have been proposed in the past [65]. As pointed out in [60], these measures were developed for the distortion evaluation purpose introduced by speech codecs and/or communication channels. Therefore, it is ill-defined whether the objective measures are suitable for quality evaluation of the enhanced speech or not [60]. The distortion types can be broadly divided into two categories [60]: (i) the speech distortion and (ii) noise distortion. The listeners are mostly influenced by the speech distortion when making judgments of overall quality [60]. But no such objective measure is available to correlate with either distortion type or with the overall quality of enhanced speech [60]. The objective quality measures represent distortions between the clean and enhanced speech signals and are classified into time and frequency domain measures. In the time domain, SNR and segmental SNR measure are used to calculate the distortions between the clean and enhanced speech signals. In the frequency domain, spectral distortion and spectral envelope distortion are used to calculate the distortions between the clean and enhanced speech signals. After comparison of the different objective measures in [66], the conclusion was (i) frequency domain measures are better than time domain measures, (ii) spectral envelope measures are better than spectral measures in the frequency domain, and (iii) Linear Predictive Cepstral (LPC) distance measures are best among the several LPC based spectral envelope calculating methods. In this thesis, three such quality measures are used which are commonly used to evaluate the speech quality after enhancement [60].

1.6.1 Segmental SNR (SSNR) Measure

The correlation of SNR with subjective quality is very poor. We can not use SNR as a general objective measure of speech quality [67]. We therefore consider the frame based SNR [67]. The measure (d_{SEGSNR}) is defined as in [67]:

$$d_{SEGSNR} = \frac{10}{M}\sum_{i=0}^{M-1} log\frac{\sum_{n=N_i}^{N_i+N-1} s_\phi^2(n)}{\sum_{n=N_i}^{N_i+N-1}[s_d(n) - s_\phi(n)]^2} \tag{1.1}$$

where, M is the total no. of frames, s_ϕ is clean signal, s_d is enhanced signal, N_i is the data length of of a particular frame and N is the total data length of the given signal. As the frames containing SNRs above 35dB do not reflect large perceptual differences, then those SNRs can be replaced with 35dB [67]. For silence periods, the SNR values can be negative since signal energies are small. These frames do not contribute perceptually. Therefore, lower threshold can be set to provide a bound on frame based SNR. We have selected -10dB as suggested in [67]. This measure is easy to implement, have low computational complexity and can provide indications of perceived speech quality for a specific waveform preserving speech system. Unfortunately, when used to evaluate transmission systems SSNR shows little correlation to perceived speech quality. This measure is also sensitive to a time shift, and therefore require precise signal alignment. The higher the value of d_{SEGSNR}, the better is the performance of the speech enhancer.

1.6.2 Weighted Spectral Slope (WSS) Measure

The WSS measure by Klatt is based on an auditory model in which 36 overlapping filters of progressively larger bandwidth are used to estimate the smoothed short-time speech spectrum. The measure finds a weighted difference between the spectral slopes in each band. The magnitude of each weight (w_a) reflects whether the band is near a spectral peak or valley, and whether the peak is the largest in the spectrum. A per-frame measure in decibels is found as [68]:

$$d_{WSS}(j) = K_{spl}(K - \hat{K}) + \sum_{k=1}^{36} w_a(k)(s_\phi(k) - s_d(k))^2 \tag{1.2}$$

where, where K and $\hat{K}$ are related to overall sound pressure level of the original and enhanced utterances, and K_{spl} is a parameter which can be varied to increase overall performance. Noise affects the speech time contour and the corresponding frequency content. Weighted Spectral Slope is better measures of perceptual speech quality as they represent deviation in the spectrum. The lower the value

of d_{WSS}, the better is the performance of the speech enhancer.

1.6.3 Log Area Ratio (LAR) Measure

LAR measure is calculated from the differences of p^{th} order LP reflection coefficients for the clean $r_\phi(j)$ and enhanced $r_d(j)$ signals for a given frame j [67]. The measure (d_{LAR}) is defined as follows [67]:

$$d_{LAR} = \left| \frac{1}{M} \sum_{i=1}^{M} \left[log\frac{1+r_\phi(j)}{1-r_\phi(j)} - log\frac{1+r_d(j)}{1-r_d(j)} \right]^2 \right|^{\frac{1}{2}} \tag{1.3}$$

where, M is the order of the LPC analysis. LAR measure is known to be significantly better correlated with human perception but still relatively simple to implement. One of the critical advantage is that they are less sensitive to signal misalignment. The lower the value of d_{LAR}, the better is the performance of the speech enhancer.

1.7 Subjective Quality Assessment

1.7.1 Perceptual Evaluation of Speech Quality (PESQ) Measure

PESQ measure is the most complex objective measure [69], [70] to compute and is recommended by ITU-T for speech quality assessment of 3.2 kHz handset telephony and narrow-band speech codecs [71], [72]. PESQ score is computed as a linear combination of the average disturbance value D_{ind} and the average asymmetrical disturbance values A_{ind} as mentioned in [69], [70]:

$$PESQ = a_0 + a_1 D_{ind} + a_2 A_{ind} \tag{1.4}$$

where, $a_0 = 4.5$, $a_1 = -0.1$ and $a_2 = -0.0309$ [69], [70]. The parameters were optimized for speech processed through networks and not for speech enhanced by noise suppression algorithms [69], [70]. To correlate the PESQ measure with all three objective quality measures, the undertaken in this investigation also consid-

ered optimized PESQ measure for each of the three rating scales by choosing a different set of parameters for each rating scale as in [69], [70]. Three different modified PESQ measures are used to predict signal distortion, noise distortion and overall speech quality. They are: (i) the speech signal alone using a five-point scale of signal distortion (SIG), (ii) the background noise alone using a five-point scale of background intrusiveness (BAK), and (iii) the overall effect using the scale of the Mean Opinion Score (MOS) (OVRL) [1 = bad, 2 = poor, 3 = fair, 4 = good, 5 = excellent] [69], [70]. The simulation codes are taken from [73]. All perceptual measures report outstanding quality correlations over large range of degradations. PESQ, in particular, is standardized as ITU-T P.862 in 2003 and is widely acknowledged as the state-of-the-art with reported quality correlation at 0.95. In fact, perhaps at the absence of reliable intelligibility measure, ITU-T has formed a study group (period : 2005-2008) to extend PESQ for intelligibility assessment.

1.7.2 Composite Measure

These measures were obtained by combining the objective measures. Composite measures are used to correlate with distortions related to speech/noise and overall speech quality [70]. It is derived by utilizing multiple linear regression analysis in [70]. We have applied the same concept for signal distortion (CorrSIG), for noise distortion (CorrBAK), and for overall speech quality (CorrOVRL). The simulation codes are taken from [73]. Two figures of merit are computed for the objective measures. The first one is the correlation coefficient (Pearson's correlation) between the subjective quality ratings S_d and the objective measure O_d, and is given by [70]:

$$\rho = \frac{\sum_d (S_d - \overline{S_d})(O_d - \overline{O_d})}{[\sum_d (S_d - \overline{S_d})]^{1/2} [\sum_d (O_d - \overline{O_d})]^{1/2}} \tag{1.5}$$

where, $\overline{S_d}$ and $\overline{O_d}$ are the mean values of S_d and O_d respectively. The second figure of merit is an estimate of the standard deviation of the error, and is given by [70]:

$$\hat{\sigma_e} = \hat{\sigma_d}\sqrt{1-\rho^2} \tag{1.6}$$

where, $\hat{\sigma_d}$ and $\hat{\sigma_e}$ are the standard deviation of S_d and computed standard deviation of error respectively. A smaller value of $\hat{\sigma_e}$ indicates that the objective measure is better at predicting subjective quality [70].

Two parametric (linear regression) and nonparametric regression analysis techniques were used [70] as third merit. The nonparametric regression was based on Multivariate Adaptive Regression Splines (MARS) analysis [70]. The MARS modeling technique is data driven and derives the best fitting function from the data. The MARS modeling is used to locally fit the data in a region by spline functions. Using the basis functions, it generates a global model after combining the data regions. The simulation codes are taken from [70]. The same five point scale is taken under consideration and to evaluate the measure. They are [1 = bad, 2 = poor, 3 = fair, 4 = good, 5 = excellent].

1.8 Organization of the Thesis

The thesis is organized into six chapters.

Chapter 2 contains the proposed framework and formulation by describing importance of LGW Filter, their frequency domain representation followed by design specifications in the current context, telephone channel and handset effect reduction, followed by ambient noise reduction, and estimator design using the proposed technique. The implementation details are also given in terms of block diagram and pseudo code. The objective, subjective and composite measures for eight different noise conditions with different SNR levels against seven different existing algorithms are discussed in details.

Chapter 3 contains the proposed framework and formulation that use joint subband statistics, signal analysis by LGW filters, Bayesian JSC with constant inter-scale variance. The implementation details are also given in block diagram and pseudo code. The objective, subjective and composite measures are shown against different existing algorithms, proposed marginal and Bivariate estimator.

Chapter 4 contains the proposed framework and formulation by describing

performance analysis of LGW with Bayesian JSC with SPU in constant inter-scale variance model estimator. The implementation details are given in block diagram and pseudo code. The performance evaluations are discussed in details in terms of objective, subjective and composite measures.

Chapter 5 contains the proposed framework and formulation in Automatic Speech Recognition contexts. The existing and proposed marginal as well Bivariate estimators are evaluated against four different noise conditions with different SNR levels.

Chapter 6 contains the conclusion and provides some directions for future work.

References

[1] H. Levitt, "Noise reduction in hearing aids: An overview", *Journal of Rehabilitation Research and Development*, vol. 38, no. 1, Jan./Feb. 2001. (Cited in section 1.)

[2] Y. Ephraim and I. Cohen, "Recent Advancements in Speech Enhancement", *The Electrical Engineering Handbook*, CRC Press, 2006. (Cited in section 1.)

[3] S. D. Kamath, "A Multiband Spectral Subtraction Method for Speech Enhancement", Master Thesis, University of Dallas, 2001. (Cited in sections 1.1.1 and 1.1.3.)

[4] https://www.msu.edu/course/asc/232/Charts/Speech%20Production.html. (Cited in sections (document) and 1.2.)

[5] N. Virag, "Speech enhancement based on masking properties of the human auditory system", Master thesis, Swiss Federal Institute of Technology, 1996. (Cited in sections (document), 1.1.2, 1.1, 1.1.2, 1.1.3, 1.2 and 1.2.)

[6] J. C. Junqua, "The influence of acoustics on speech production: a noise induced stress phenomenon known as the Lombard reflex", *ESCA-NATO Workshop on Speech under Stress*, Lisbon, pp. 83-90, Sep. 1995. (Cited in section 1.1.2.)

[7] J. C. Junqua and J. P. Haton, "Robustness in automatic speech recognition: Fundamentals and applications", *Kluwer A. Publishers*, 1996. (Cited in section 1.1.2.)

[8] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech", *Proc. IEEE*, vol. 67, no. 12, pp. 221-239, Dec. 1979. (Cited in section 1.1.3.)

[9] S. F. Boll, "Suppression of Acousitc Noise in Speech using Spectral Subtration", *IEEE ASSP*, vol. 27, no. 2, pp. 113-120, 1979. (Cited in section 1.2.)

[10] W. G. Knecht, M. E. Schenkel, and G. S. Moschytz, "Neural network filters for speech enhancement", *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 6, pp. 433-438, Nov. 1995. (Cited in section 1.2.)

[11] M. Dahl and I. Claesson, "Acoustic noise and echo cancelling with microphone array", *IEEE Trans. Vehicular Technology*, vol. 48, no. 5, pp. 1518-1526, Sep. 1999. (Cited in section 1.2.)

[12] Y. Ephraim and D. Malah, "Speech Enhancement using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109-1121, Dec. 1984. (Cited in section 1.2.)

[13] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator", *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443-445, Apr. 1985. (Cited in section 1.2.)

[14] T. H. Dat, K. Takeda, and F. Itakura, "Generalized Gamma Modeling of Speech and its Online Estimation for Speech Enhancement", *Proceedings of ICASSP-2005*, 2005. (Cited in section 1.2.)

[15] R. Martin and C. Breithaupt, "Speech Enhancement in the DFT Domain using Laplacian Speech Priors", *in Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC 03)*, Kyoto, Japan, pp. 87-90, Sep. 2003. (Cited in section 1.2.)

[16] R. Martin, "Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors", *IEEE ICASSP'02*, Orlando, Florida, May 2002. (Cited in section 1.2.)

[17] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model", *EURASIP Journal on Applied Signal Processing*, vol. 2005, issue 7, pp. 1110-1126, 2005. (Cited in section 1.2.)

[18] C. Breithaupt and R. Martin, "MMSE Estimation of Magnitude-Squared DFT Coefficients with Super-Gaussian Priors", *IEEE Proc. Intern. Conf. on Acoustics, Speech and Signal Processing*, vol. I, pp. 896-899, Apr. 2003. (Cited in section 1.2.)

[19] J. D. Deng and A. Acero. "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features", *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 3, pp. 218-233, May 2004. (Cited in section 1.2.)

[20] I. Cohen, "Speech Enhancement Using a Noncausal A Priori SNR Estimator", *IEEE Signal Processing Letters*, vol. 11, no. 9, pp. 725-728, Sep. 2004. (Cited in section 1.2.)

[21] S. Kamath and P. Loizou, "A Multi-Band Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise", *In Proceedings International Conference on Acoustics, Speech and Signal Processing*, 2002. (Cited in section 1.2.)

[22] E. Zavarehei, S. Vaseghi, and Q. Yan, "Speech Enhancement using Kalman Filters for Restoration of Short-Time DFT Trajectories", *Automatic Speech Recognition and Understanding (ASRU), 2005 IEEE Workshop*, pp. 219-224, Nov. 27, 2005. (Cited in section 1.2.)

[23] P. Moreno, B. Raj, and R. Stern, "A vector Taylor series approach for environment-independent speech recognition", *Proc. ICASSP*, pp. 733-736, 1996. (Cited in section 1.2.)

[24] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adapation using vector Taylor series for noisy speech recognition", *ICSLP Bejing*, pp. 869-872, 2000. (Cited in section 1.2.)

[25] J. Gauvain and C. Lee, "MAP estimation for multivariate Gaussian mixture observation of Markov Chains", *IEEE Trans. Speech & Audio Processing*, vol. 2, pp. 291-298, 1994. (Cited in section 1.2.)

[26] C. Leggetter and P. Woodland, "Maximum Likelihood Linear Regression for speaker adaptation of continuous density HMMs", *Comp. Sp. & Lang.*, pp. 171-185, 1995. (Cited in section 1.2.)

[27] J. -H. Changa, S. Gazor, N. S. Kim, and S. K. Mitra, "Multiple statistical models for soft decision in noisy speech enhancement", *Pattern Recognition*, vol. 40, issue 2007, pp. 1123-1134, 2007. (Cited in section 1.2.)

[28] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Noisy speech enhancement using discrete cosine transform", *Speech Commun.*, vol. 24, no. 3, pp. 249-257, 1998. (Cited in section 1.2.)

[29] J. -H. Chang and N. S. Kim, "Speech enhancement using warped discrete cosine transform", *in Proceedings of the IEEE Speech Coding Workshop*, Tsukuba, Japan, pp. 175-177, Oct. 2002. (Cited in section 1.2.)

[30] S. Gazor and W. Zhang, "Speech probability distribution", *IEEE Signal Process. Lett.*, vol. 10, no. 7, pp. 204-207, 2003. (Cited in section 1.2.)

[31] R. C. Reininger and J. D. Gibson, "Distributions of the two dimensional DCT coefficients for images", *IEEE Trans. Commun.*, vol. Com-31, no. 6, pp. 835-839, 1983. (Cited in section 1.2.)

[32] J. H. Lee, H. Y. Jung, T. W. Lee, and S. Y. Lee, "Speech coding and Noise reduction using ICA based speech Features", *Electronics Letters*, vol. 36, no. 17, pp. 1506-1507, 2000. (Cited in section 1.2.)

[33] T. Rutkowski, A. Cichocki, and A. K. Barros, "Speech Enhancement Using Adaptive Filters and Independent Component Analysis Approach", 1999. (Cited in section 1.2.)

[34] D. -C. Balcan and J. Rosca, "Independent Component Analysis for Speech Enhancement with Missing TF Content", *LNCS:ICA 2006*, LNCS 3889, pp. 552-560, 2006. (Cited in section 1.2.)

[35] X. Ma, Y. Wang, W. Liu, and F. Yin, "A New Speech Enhancement Method for Adverse Noise Environment", ISNN 2005, *LNCS*, LNCS 3497, pp. 586-591, 2005. (Cited in section 1.2.)

[36] L. Hong, J. Rosca, and R. Balan, "Independent Component Analysis Based Single Channel Speech Enhancement Using Wiener Filter", *in Proceedings of ISSPIT*, 2003. (Cited in section 1.2.)

[37] H. Sameti, H. Sheikhzadeh, L. Dend, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise", *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 445-455, 1998. (Cited in section 1.2.)

[38] M. Klein and P. Kabal, "Signal subspace speech enhancement with perceptual post-filtering", *Proc. IEEE Internet. Conf . Acoust. Speech Signal Process. (ICASSP)*, vol. 1, pp. 537-540, 2002. (Cited in section 1.2.)

[39] D. L. Donoho, "De-noising by soft thresholding", *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613-627, 1995. (Cited in section 1.2.)

[40] J. Seok and K. Bae, "Speech enhancement with reduction of noise components in the wavelet domain", *Proc. IEEE Internet. Conf. Acoust. Speech Signal Process (ICASSP)*, vol. 2, pp. 1323-1326, 1997. (Not cited.)

[41] A. Grossman, R. Kronland-Martinet, and J. Morlet, "Analysis of Sound patterns through wavelet transforms", I*nt . J. Patt. Roeog. Artificial Intell.*, vol. 1, pp. 97-126, Aug. 1987. (Cited in section 1.2.)

[42] I. Daubechies, "Orthonormal bases of compactly supported wavelets", *Commun. Pure Appl. Math.*, vol. 41, pp. 909-996, 1988. (Cited in section 1.2.)

[43] I. Daubechies, "Orthonormal bases of wavelets with compact support-connection with discrete filters", *In wavelets: Time-Frequency Methods and Phase Space, Berlin:Springer, IPTI*, pp. 38-66, 1989. (Cited in section 1.2.)

[44] R. R. Coifman, Y. Meyer, and M. V. Wickerhauser, "Wavelet analysis and signal processing", *in Wavelets and Their Applications, M.B. Ruskai et al., Eds.*, pp. 153-178. Jones and Bartlett, Boston, 1992. (Cited in section 1.2.)

[45] R. R. Coifman and M. V. Wickerhauser, "Entropy based algorithms for best basis selection", *IEEE Trans. Inf. Th.*, vol. 38, no. 2, pp. 713-718, Mar. 1992. (Cited in section 1.2.)

[46] H. Malvar, "Signal Processing with Lapped Transforms", *Artech House*, 1992. (Cited in section 1.2.)

[47] H. Krim and J. -C. Pesquet, "On the statistics of best bases criteria", *in Wavelets and Statistics, Lecture Notes in Statistics, A. Antoniadis, Ed.*, Springer-Verlag, pp. 193-207, 1995. (Cited in section 1.2.)

[48] P. Moulin, "Signal estimation using adapted tree-structured bases and the MDL principle", *in Proc. IEEE-SP Int. Symp. TFTS*, Paris, pp. 141-143, Jun. 1996. (Cited in section 1.2.)

[49] H. Krim, D. Tucker, S. Mallat, and D. Donoho, "On denoising and best signal representation", *IEEE Trans. Inf. Th.*, vol. 45, no. 7, pp. 2225-2238, Nov. 1999. (Cited in section 1.2.)

[50] K. Ramchandran and M. Vetterli, "Best wavelet packet bases in a rate distortion sense", *IEEE Trans. Im. Proc.*, vol. 2, no. 2, pp. 160-175, Apr. 1993. (Cited in section 1.2.)

[51] M. V. Wickerhauser, INRIA lectures on wavelet packet algorithms. In Ondelettes et paquets d'ondelettes, pages 31-99, Roquencourt, Jun. 17-21, 1991. (Cited in section 1.2.)

[52] D. Donoho and I. Johnstone, "Ideal denoising in an orthogonal basis chosen from a library of bases", preprint Dept. Stat., Stanford Univ., Oct. 1994. (Cited in section 1.2.)

[53] H. Krim and J. -C, Pasquet, "Robust multiscale representation of processes and optimal signal reconstruction", *In Time Freq./Time Scale Symp.*, Philadelphia, PA, 1994. (Cited in section 1.2.)

[54] B. D. Moor, "The singular value decomposition and long and short spaces of noisy matrices", *IEEE Trans. on Sig. Proc.*, vol. 41, no. 9, pp. 2826-2838, Sept. 1993. (Cited in section 1.2.)

[55] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sørensen, "Reduction of Broad-Band noise in speech by Truncated QSVD", *IEEE Trans. on Speech and Audio Process.*, vol. 3, no. 6, pp. 439-448, Nov. 1995. (Cited in section 1.2.)

[56] P. C. Hansen and S. H. Jensen, "FIR filter representation of reduced-rank noise reduction", *IEEE Trans. on Sig. Proc.*, vol. 46, no. 6, pp. 1737-1741, Jun. 1998. (Cited in section 1.2.)

[57] S. V. Huffel, "Enhanced resolution based on minimum variance estimation and exponential data modeling", *Signal Proc.*, vol. 33, no. 3, pp. 333-355, 1993. (Cited in section 1.2.)

[58] M. Dendrinos, S. Bakamidis, and G. Caraynnis, "Speech Enhancement from noise : a regenerative process", *Speech Comm.*, vol. 10, no. 1, pp. 45-57, 1991. (Cited in section 1.2.)

[59] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms", *Speech Communication*, vol. 49, pp. 588-601, 2007. (Cited in section 1.5.)

[60] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement", *IEEE Trans. on Audio, Speech and Language Process.*, vol. 16, no. 1, pp. 229-238, Jan. 2008. (Cited in sections 1.5 and 1.6.)

[61] Available at: http://www.utdallas.edu/∼loizou/speech/noizeus. (Cited in sections (document), 1.5 and 1.3.)

[62] IEEE Subcommittee, "IEEE Recommended Practice for Speech Quality Measurements", *IEEE Trans. Audio and Electroacoustics*, vol. AU-17, no. 3, pp. 225-246, 1969. (Cited in section 1.5.)

[63] ITU-T P.56, "Objective measurement of active speech level", ITU-T Recommendation P.56, 1993. (Cited in section 1.5.)

[64] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", *in ISCA ITRW ASR2000*, Paris, France, 2000. (Cited in section 1.5.)

[65] S. Quackenbush, T. Barnwell, and M. Clements, "Objective Measures of Speech Quality", *Englewood Cliffs, NJ: Prentice-Hall*, 1988. (Cited in section 1.6.)

[66] N. Kitawaki, K. Itoh, M. Honda, and K. Kakehi, "Comparison of objective speech quality measures for voice band coders", *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 1000-1003. 1982. (Cited in section 1.6.)

[67] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms", *Int. Conf. on Spoken Language Processing*, vol. 7, pp. 2819-2822, 1998. (Cited in sections 1.6.1, 1.6.1 and 1.6.3.)

[68] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low bit-rate speech coding systems", *IEEE Jour. on Select. Areas in Comm.*, vol. 6, no. 2, pp. 262-273, Feb. 1988. (Cited in section 1.6.2.)

[69] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms", *Speech Communication*, vol. 49, pp. 588-601, 2007. (Cited in sections 1.7.1 and 1.7.1.)

[70] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement", *IEEE Trans. on Audio, Speech and Language Process.*, vol. 16, no. 1, pp. 229-238, Jan. 2008. (Cited in sections 1.7.1, 1.7.1, 1.7.2, 1.7.2 and 1.7.2.)

[71] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-A new method for speech quality assessment of telephone networks and codecs", *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, pp. 749-752, 2001. (Cited in section 1.7.1.)

[72] "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs", ITU, ITU-T Rec. P. 862, 2000. (Cited in section 1.7.1.)

[73] P. Loizou, "Speech Enhancement: Theory and Practice", *CRC Press, Taylor and Francis*, Boca Raton, FL, 2007. (Cited in sections 1.7.1 and 1.7.2.)