Robust Speech Filter And Voice Encoder Parameter Estimation using the Phase-Phase Correlator

Abul K. Azad

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering

Lamine Mili, Chair

T. Charles Clancy

Allan MacKanzie

Amir Zaghloul

Narenderan Ramakrishbnan

September 20th, 2019

Falls Church, Virginia

Keywords: Robust statistics, speech processing, auto-regressive parameter estimation, estimator bias curve, estimator variance curve, mixed-excitation linear prediction, phase-phase correlator, impulsive noise.

Robust Speech Filter And Voice Encoder Parameter Estimation using the Phase-Phase Correlator

Abul K Azad

(ABSTRACT)

In recent years, linear prediction voice encoders have become very efficient in terms of computing execution time and channel bandwidth usage while providing, in the absence of impulsive noise, natural sounding synthetic speech signals. This good performance has been achieved via the use of a maximum likelihood parameter estimation of an auto-regressive model of order ten that best fits the speech signal under the assumption that the signal and the noise are Gaussian stochastic processes. However, this method breaks down in the presence of impulse noise, which is common in practice, resulting in harsh or non-intelligible audio signals. In this paper, we propose a robust estimator of correlation, the Phase-Phase correlator that is able to cope with impulsive noise. Utilizing this correlator, we develop a Robust Mixed Excitation Linear Prediction encoder that provides improved audio quality for voiced, unvoiced, and transition speech segments. This is achieved by applying a statistical test to robust Mahalanobis distances for identifying the outliers in the corrupted speech signal, which are then replaced with filtered signals. Simulation results reveal that the proposed method outperforms in variance, bias, and breakdown point three other robust approaches based on the arcsin law, the polarity coincidence correlator, and the median-of-ratio estimator without sacrificing the encoder bandwidth efficiency and the compression gain while remaining compatible with real-time applications. Furthermore, in the presence of impulsive noise, the proposed speech encoder speech perceptual quality also outperforms the state of the art in terms of mean opinion score.

Robust Speech Filter And Voice Encoder Parameter Estimation using the Phase-Phase Correlator

Abul K Azad

(GENERAL AUDIENCE ABSTRACT)

Impulsive noise is a natural phenomenon in everyday experience. Impulsive noise can be analogous to discontinuities or a drastic change in natural progressions of events. Specifically in this research the disrupting events can occur in signals such as speech, power transmission, stock market, communication systems, etc. Sudden power outage due to lighting, maintenance or other catastrophic events are some of the reasons why we may experience performance degradation in our electronic devices. Another example of impulsive noise is when we play an old damaged vinyl records, which results in annoying clicking sounds. At the time instance of each click, the true music or speech or simply the audible waveform is completely destroyed. Other examples of impulse noise is a sudden crash in the stock market; a sudden dive in the market can destroy the regression and future predictions. Unfortunately, in the presence of impulsive noise, classical methods methods are unable to filter out the impulse corruptions.

The intended filtering objective of this dissertation is specific, but not limited, to speech signal processing. Specifically, research different filter model to determine the optimum method of eliminating impulsive noise in speech. Note, that the optimal filter model is different for time series signal model such as speech, stock market, power systems, etc. In our studies

we have shown that our speech filter method outperforms the state of the art algorithms. Another major contribution of our research is in speech compression algorithm that is robust to impulse noise in speech. In digital signal processing, a compression method entails in representing the same signal with less data and yet convey the the same same message as the original signal. For example, human auditory system can produce sounds in the range of approximately 60 Hz and 3500 Hz, another word speech can occupy approximately 4000 Hz in frequency space. So the challenge is, can we compress speech in one of half of that space, or even less. This is a very attractive proposition because frequency space is limited but the wireless service providers desires to service as many users as possible without sacrificing quality and ultimately maximize the bottom line. Encoding impulse corrupted speech produces harsh quality of synthesized audio. We have shown if the encoding is done with the proposed method, synthesized audio quality is far superior to the sate of the art.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Bandwidth in wireless voice communication systems is limited and very expensive. Therefore, extensive research has been carried out for developing methods able to achieve high level of speech compression while maintaining good audio signal quality. Well-known methods include the transmission of the estimates of the parameters of an Auto-Regressive (AR) model, the Fourier magnitudes of the error signal, the pitch periods, the signal gain, and the transition frames. The AR model parameters are estimated at the receiver end to reproduce a synthesized version of the actual speech signal. Note that AR model of a short time speech segment is synonymous with a linear prediction model, which is encoded via the Line Spectral Pairs (LSP), which are resilient to quantization errors. Linear prediction, AR model and LSP represent the same parameter in different form. However, there are two challenges that have not been yet satisfactorily addressed; these are: (i) achieving good quality of the speech encoder model estimation using low data rate [1], and that despite a significant progress made during the last decades [2], [3], [4], [5], [6], and (ii) copying with speech impulsive noise [7, 8, 9, 10]. If the model parameters are not estimated with acceptable accuracy from the bias and variance stand point, the synthesized audio may sound harsh, or tonal, or non-human like, or non intelligible [2].

Let us now review the most popular prefiltering techniques for noise suppression in speech. One of them is the median filter [8], [11], which computes the median of a sliding window, one sample at a time, and then tags the sample as an outlier via a thresholding mechanism. A corrupted sample is replaced with an estimated value obtained using a linear prediction model. Unfortunately, the median filter is not practical for real-time application because it filters the speech signal one sample at a time, which is time consuming. Furthermore, it unduly over-tags as outliers good signal segments, making the filtered speech sound distorted. A second method is the binary mask filter proposed by Ruhland *et al.* [7], which processes the speech signal in the frequency domain. Assuming a disjoint speech and noise spectra, a Signal-to-Noise Ratio (SNR) thresholding algorithm is utilized to identify the noise bins, which are then zeroed. Finally, an inverse Fourier transform is applied to recover the filtered speech signal in the time domain. Unfortunately, this approach provides a strongly biased spectrum estimate, resulting in a poorly synthesized audio signal. A third method is proposed by Veselinovic and Graupe [9]; it makes use of a wavelet filter bank that decomposes the noisy speech in the wavelet domain, and then calculates highpass and lowpass filter output wavelet coefficients, whose noise components, which have lower values, are filtered out via a thresholding method. Then, the noise is either replaced with zeroes or interpolated. Finally a filtered speech is obtained by applying the inverse wavelet transform. Unfortunately, this method produces sounds of poor quality in the presence of impulsive noise because the latter have a power spectrum that leaks over the filter bandwidths. A fourth method is proposed by Wen and Tao [12], which aims at suppressing the voice signal from the noise by utilizing an inverse AR filter and applying a thresholding mechanism. Unfortunately, this method produces harsh sounds when the original speech signal is corrupted by impulsive noise. A fifth method proposed by Hassen and Clements [10] consists in applying the Wiener filter that minimizes the mean-square-error between a target signal and the estimated signal, yielding distorted audio signal in the presence of impulsive noise.

A brief overview of our proposed contributions consists of a) a robust estimator of correlation, applicable to speech, based on the Phase-Phase Correlator (PPC) [13], b) a robust method to estimate the power spectral density, c) a robust approach to estimate the Auto-Regressive model of order p, AR(p). d) a novel speech filter algorithm and e) a robust speech encoder algorithm. Although speech filter can be applied on its own or as the pre-filter to speech encoder process, all of the contributions are applicable to designing a robust speech encoder that can cope with impulsive noise condition. A robust estimator of correlation may be applied in a speech encoder to estimate the fundamental frequency, to make voice/unvoiced framing decision, estimate band-pass voicing strength, etc. The robust estimator of correlation can be used to define auto-correlation function of the error signal and subsequently take the Fourier transform to estimate the PSD, which can be utilized to estimate the robust excitation signal. The robust estimator of correlation, along with the Burg's algorithm can be applied to estimate the minimum phase $AR(p)$, which can be transformed into Line Spectral Pair (LSP) for encoder model.

In this paper, we develop a new pre-filtering algorithm that is robust to impulsive noise, which has the following novel features. As shown in Fig. 2.4, it consists in identifying the outliers using a new statistical test applied to Robust Mahalanobis Distances (RMD) based on the PPC. A new method for generating the exciting error signal of the AR(10) model is developed. First, this model is robustly estimated using a robust version of the Burg's algorithm based on the PPC. Then, the PPC is executed again to robustly estimate the autocorrelation function at various time lags. Next, by applying the Fourier transform to this function, the Power Spectrum Density (PSD) of the error signal is estimated and the Fourier magnitudes are calculated. Finally, the missing samples associated with the outliers are replaced with the outputs of the AR(10) model that are excited with the Inverse Fast Fourier Transform (IFFT) of the Fourier magnitudes. Our simulation results show significant improvements on the filtered output speech provided by our pre-filtering method over the methods advocated in the literature mentioned earlier.

After the execution of the prefiltering step just described, we apply a new robust speech compression algorithm based on the Mixed-Excitation Linear Prediction (MELP) coding. Initiated by McCree and Barnwell [2] in 1995, it has been the top candidate for the U.S. federal standards and since then, has been adopted by NATO as the standard for the voice encoder. Indeed, in the absence of impulsive noise, the MELP is able to synthesize superior quality and natural sounding speech. The MELP encoder algorithm is based on the Linear

Predictive Coding (LPC) with several additional analysis functions as described next. The speech signal is first segmented into 22.5-ms frames, which are each processed to estimate the pitch periodicity, the parameters of an AR(10) model, the time correlations, the Fourier transform magnitudes, and five bandpass voicing strengths. The model parameters are then encoded and sent to the receiver for decompression. Besides the MELP, there are a few other algorithms proposed in the literature. These include the Vector Sum Excited Linear Prediction (VSELP) coding [6], the Advanced Multi-Band Excitation (AMBE) coding [14], and the Code Excited Linear Prediction (CELP) [5] coding. Unfortunately, all these methods are vulnerable to impulsive noise.

In this paper, we propose a new robust version of the MELP, termed the RMELP. We choose the MELP algorithm because it outperforms all the other methods in terms of speech quality in the absence of impulsive noise. By contrast, our RMELP robustly estimates four critical parameters, namely the fundamental frequency or the pitch period, the band-pass voicing strengths, the AR(10) model parameters, and the Fourier magnitudes of the residual signal, which are all estimated using a correlation estimator. In this paper, we investigate four robust correlation estimation methods proposed in the literature to identify which one is the most suitable for speech processing; these are the Arcsin Law (ASL), the Complex Polarity Coincidence Correlator (CPCC), the PPC, and the Median-of-Ratios Estimator (MRE), which is the maximum likelihood estimator at the Laplace probability distribution. Specifically, we develop the Complex-valued PPC (CPPC), which has been adapted from the real-valued hard limitor estimator, namely the ASL. Here, the CPCC is the intermediate step for deriving the PPC with infinite phase quantization. Furthermore, we show that a hard limited correlation estimator in the complex domain is more efficient than in the real domain. Since speech signals are real-valued processes, we propose to first take the Hilbert transform to map them in the complex domain, and then to estimate the complex-valued correlation coefficients using the CPCC to improve its variance. To evaluate the performance of each of these four estimators at various contamination rate, we carry out extensive simulations on synthetic signals and evaluate their variance, bias, and breakdown point at various correlation and contamination rate level. We conclude that the PPC is the robust estimator of choice for speech signal analysis when subject to impulsive noise. The paper is organized as follows. Section 2 analyzes the characteristics of the corrupted speech signal. Section 3 describes the ASL, the CPCC, the PPC, and the MRE as robust correlation estimators for speech signals. Section 4 proposes robust pre-filtering methods based on the PPC. Section 5 develops a robust PPC-based MELP encoder algorithm, the RMELP. Section 6 provides some simulation results of the proposed methods and compares their performances to the ASL, the CPCC, and the MRE in terms of bias, variance, and breakdown point. Finally, Section 7 concludes the paper.

Chapter 2

Corrupted Speech Signal Characteristics

Over the last few decades, a number of different algorithms for speech encoding have been proposed in the literature. However, most of them utilize spectral and excitation estimation methods under the assumption that short-time speech frames are stationary and Gaussian. However, this assumption is strongly violated in the presence of impulsive noise, making the speech signal and the noise to be intermixed in both the frequency and the time domain. Generally, an impulsive noise amplitude is large compared to that of the true signal and may span up to 2 or 5 ms. For instance, the MELP vocoder algorithms involve the estimation of the pitch period, the voiced/unvoiced flag, the aperiodic flag, the voicing strength, the AR(10) parameters, the Fourier magnitudes, and the AR model excitation in 22.5 ms frames. Evidently, if all these parameters are not robustly estimated, which is the case for all the

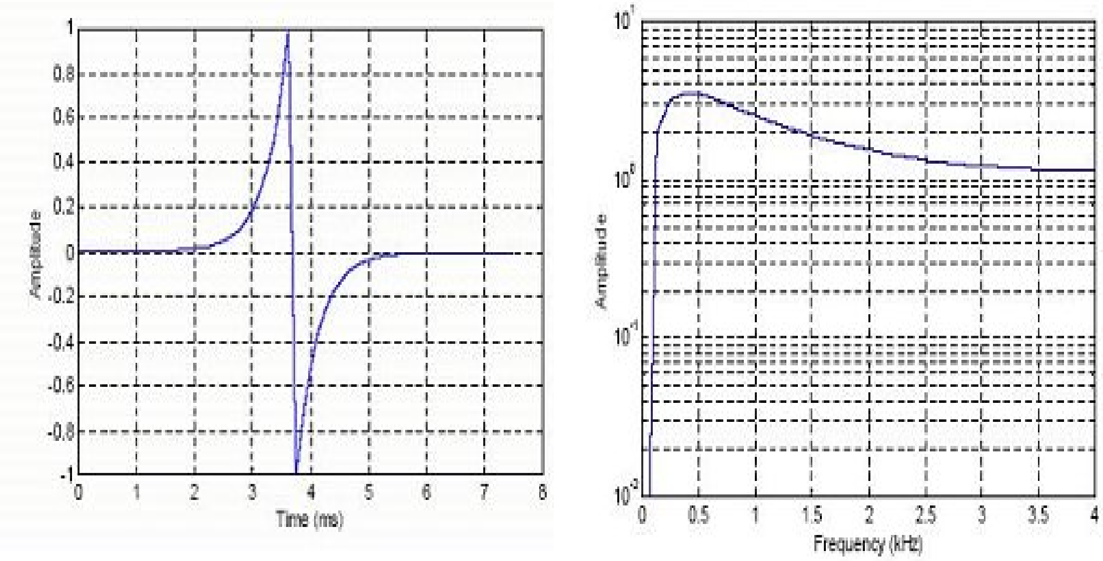

Figure 2.1: Impulse model in time and frequency domain at 8Ksps; observing the frequency response, clearly it destroys the entire spectrum.

conventional methods, the synthesized audio signal will be biased in the presence of impulsive noise.

Impulse corruption in time domain can be modelled as a double sided exponential function; increase in amplitude, reaching infinite slope and hence discontinuous, while immediately decays to mean value following the discontinuity point. Clean and impulse corrupted time series of a speech segment is depicted in 2.2. In general, impulse noise variance is greater than the signal variance and hence noise samples may be clearly visible. However, this is not always true, meaning impulse noise variance may be at the same level as the signal variance, where noise detection becomes very challenging. It is further evident from the spectrogram in 2.3 that when the impulses occurs, the spectrum characteristics are destroyed across all frequency bands.

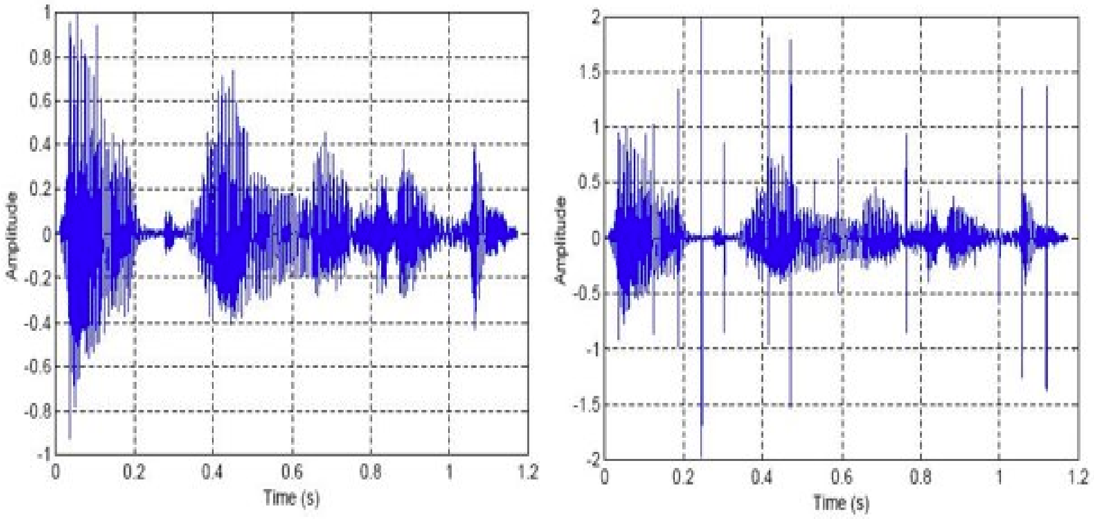

Figure 2.2: Speech sampled at 8ksps, corrupted with impulse noise. It may appear the impulse corruptions is relatively higher in amplitude but in reality, they can occur at the same level as the true signal.

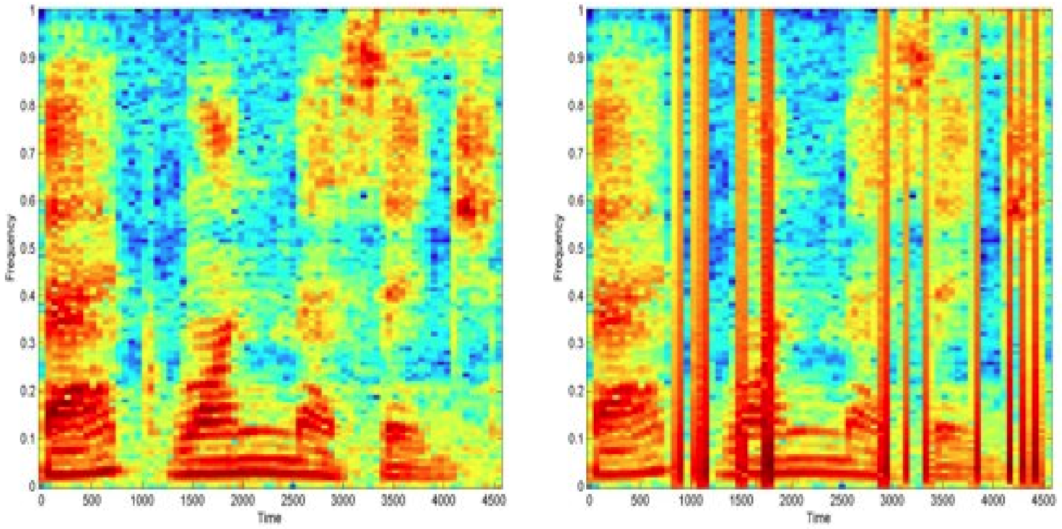

Figure 2.3: Impulse corrupted speech spectrogram.

To address this problem, we develop a robust estimator of correlation based on the PPC that is implemented within the RMELP voice encoder algorithm, which is executed at the rate of 2400 bits/sec. As shown in Fig. 2.4, a parametric model of the speech signal is generated at the output of the RMELP encoder with the same characteristics as the input speech waveform. Since a speech signal is non stationary and a-periodic due to the characteristics of the glottal excitation, the underlying process used to produce speech, it should be in principle modeled as a non-stationary stochastic process. However, this modeling is mathematically intractable for real-time applications. To address this problem, a speech signal is first broken into frames with time intervals of 20 to 30 ms, during which the signal can be reasonably assumed to be stationary, at least in the wide-sense. This is a very important assumption that makes the signal modeling, filtering, and compression algorithms realistic and execute. Hence, a frame is modeled as a stochastic process $X(t)$ with a constant mean, $E[X(t)] = \mu$, a constant variance, $E[X(t)^2] = \sigma^2$, and a time invariant autocorrelation function, $R_{xx}(t, t+\tau) = R_{xx}(\tau)$. Note it is very costly and nearly impossible to segment speech into frames that are near stationary. And hence stationary assumption is often violated, specially in transition frames, resulting in sub-optimal parameter estimation and poor speech synthesis.

Research has shown that speech synthesis quality is just as sensitive to excitation signal estimation as AR model parameters. This is specially true when signal in question is non-stationary. However, in the presence of impulsive noise, our research has shown that generating the excitation signal is more challenging than estimating the PSD of a given frame. In this paper, we propose to robustly estimate the power spectral density, by taking the

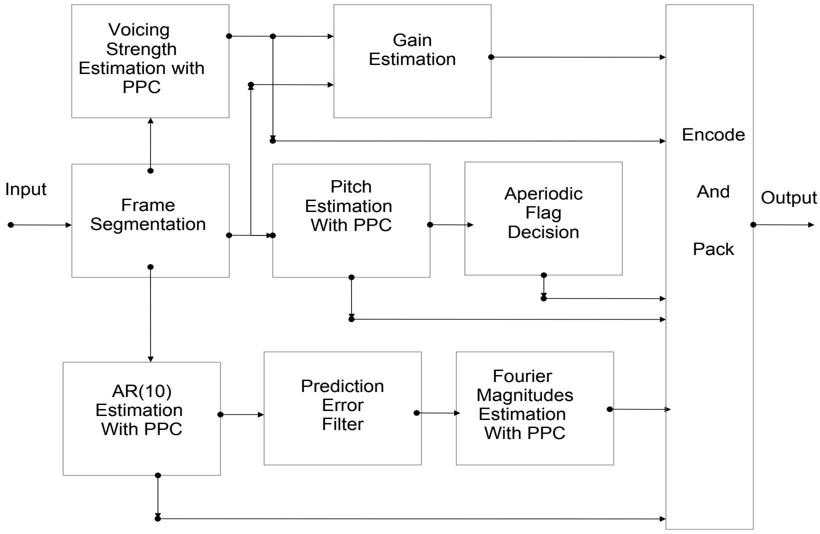
Voicing Strength Estimation with PPC
Gain Estimation
Encode And Pack
Input
Frame Segmentation
Pitch Estimation With PPC
Aperiodic Flag Decision
Output
AR(10) Estimation With PPC
Prediction Error Filter
Fourier Magnitudes Estimation With PPC

Figure 2.4: Flowchart of the RMELP encoder.

Fourier transform of the PPC-based autocorrelation function, which is encoded to estimate the spectral shape of the error signal. Unfortunately, this approach still produces some noisy artifacts in the synthesized audio. A method will be proposed next to deal with this problem.

Robust statistics deal with the development of estimators with stable bias and variance under violations of the assumptions. In signal processing, the ϵ-replacement contaminated process is considered; it is defined as $X_t = X_t(1 - Z_t) + Z_tW_t$, where Z_t is the zero-one process such that $P(Z_t = 1) = \epsilon$, which determines the contamination rate. The breakdown point of an estimator is the largest contamination rate associated with an acceptable bias, where the asymptotic bias is defined as $b = |\hat{\theta}(F) - \hat{\theta}(G)|$, where $\hat{\theta}$ is the estimator, G is the contaminated model distribution of true model F. In this paper, we analyze the bias, the variance, and the breakdown point of various correlation estimators at various levels of correlation and signal contamination. Asymptotic variance is defined as

Chapter 3

Speech Filter Algorithms

Speech filtering is a popular topic in modern speech signal processing. Consequently there are way too many algorithms proposed in the IEEE papers that can withstand the scholars criticism of their technical approach and performances. Earlier, we briefly mentioned four of the best performing speech filter algorithms, specifically, the Median Filter (MF), the Binary Mask Filter (BMF), the Wavelet Filter (WF) and the Inverse Filter. In this chapter we go a little deeper into their implementations.

3.1 The Median Filter

The Sample median is the ML estimator of location at the Laplacian distribution, which sounds very attractive because speech conforms to Laplacian PDF over long time. The class of Median Filters actively smooths out the corrupted signal of generic distribution. Median

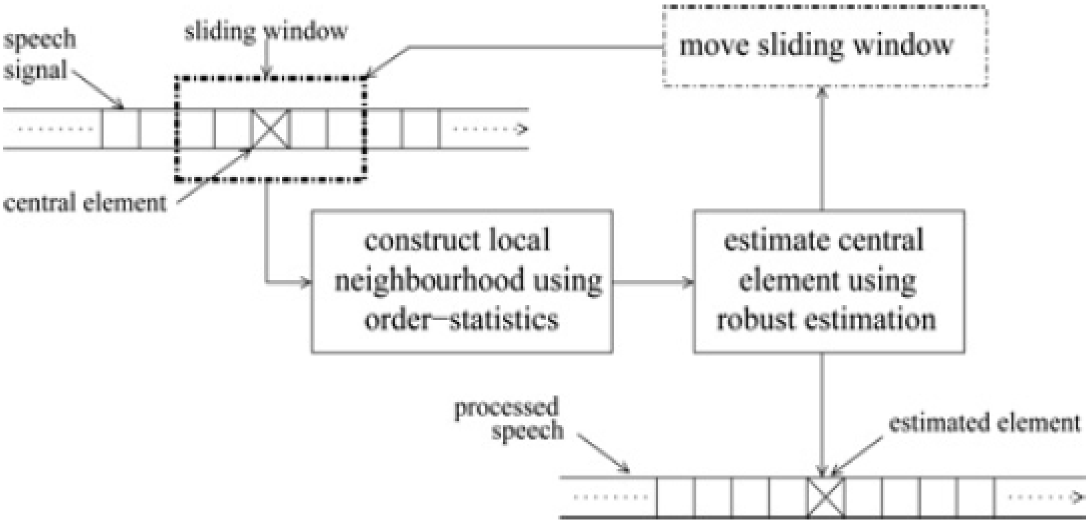
speech signal
sliding window
move sliding window
central element
construct local neighbourhood using order−statistics
estimate central element using robust estimation
processed speech
estimated element

Figure 3.1: The basic Median Filter Process

filter, is a non-linear filter, which operates on the running frame and replaces the outliers with the moving median. It should be apparent that a the MF assumes the observations are Independent and Identically distributed. Since Speech signal is highly correlated and closely follows Gaussian PDF withing short time frames, it is often the case that MF false detects the outliers. As mentioned earlier, there is no well defined threshold method, where it can suppress the undesired samples, while retaining the desired information. This make sense because both the desired and the noise signal occupy the same space in frequency. The basic processing blocks of the Median filter is demonstrated in 3.1, where the central element is the median of the sliding window.

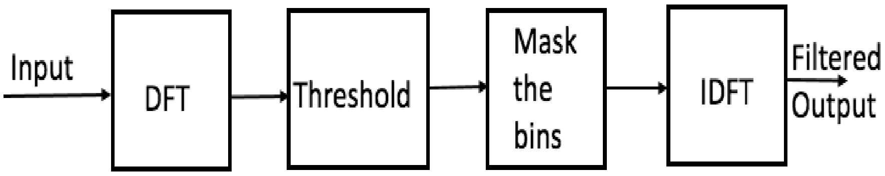

Figure 3.2: The Simple Binary Mask Filter

3.2 The Binary Mask Filter

As its name suggests, binary masking is the process of masking the frequency bins of a signal in frequency domain. The simplest of methods is to replace the outlier bins with zeros and take the inverse Fourier transform to clean up the corrupted signal. As shown in 3.2, the process begins by taking Fourier transform of the input frame and utilize a threshold mechanism in the frequency domain to mask out the noise and then convert back into time domain. This filter. This filter performs very well for applications where the true signal is known and the noise bins are easily estimated from the corrupted signal. As seen in 3.3, the improvement is remarkable. It should be noted that the clean speech was corrupted with Gaussian white noise, which is a big problem in real applications where the noise model is unknown. Furthermore, determining the optimum threshold to mask out the noise bins is very difficult.

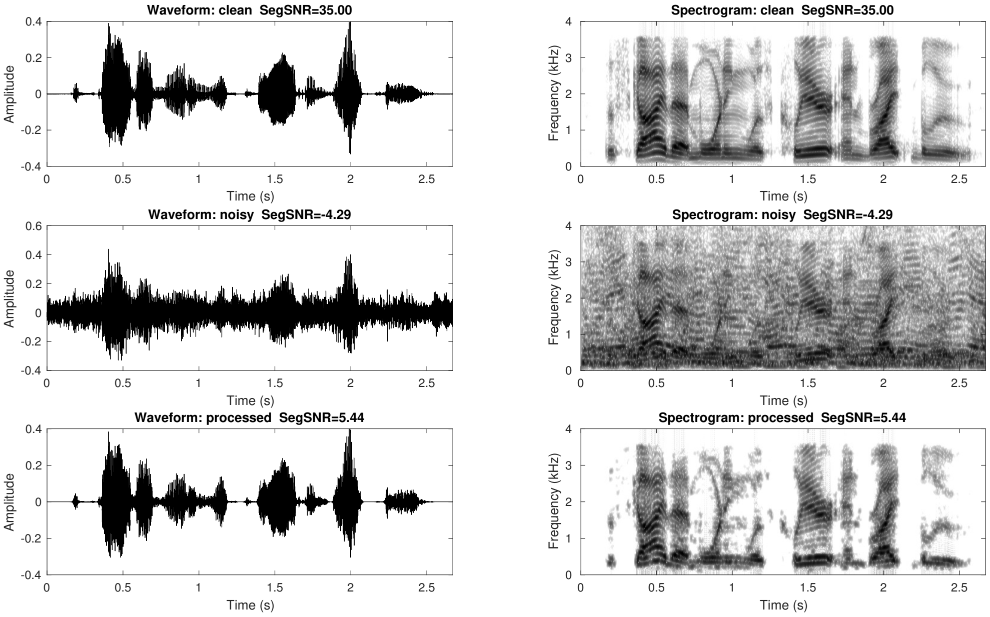

Figure 3.3: The BMF performance under the ideal condition

3.3 The Wavelet Filter

It is well documented that the Wavelet Filter is able to decompose a signal in both time and frequency domain, apply a threshold mechanism to identify the outliers, replace them and reconstruct an estimated, filtered signal. Simultaneous time and frequency domain filtering is claimed because the corrupted signal is filtered with a high-pass filter and a low-pass filter, which results in two time series' which are separated in frequencies. In theory, his process can be iterated until the lowest time resolution is achieved; however, after so many iterations there will be no information left and hence no need in continuing. In wavelet analysis, the low-pass filter output is referred to as the approximation signal, while the high-pass filter output is is considered as the details. Indeed, the classifications of the wavelet transform

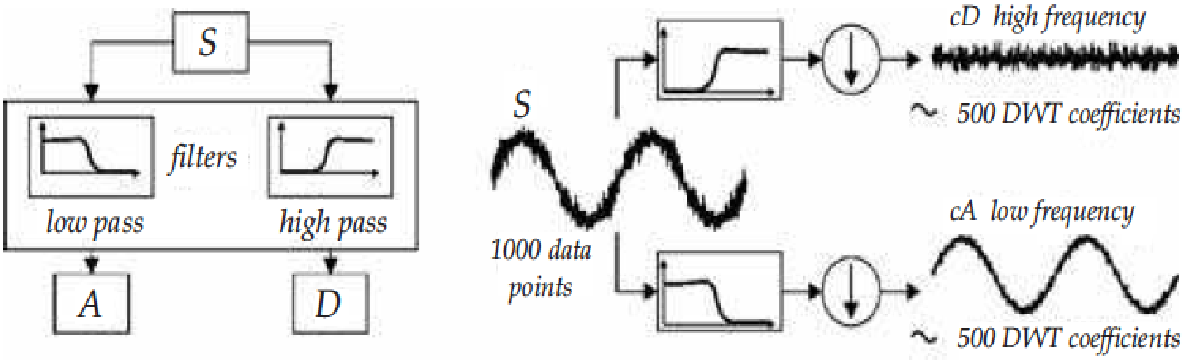

Figure 3.4: The Wavelet Transform Decomposition

outputs, specifically for speech, makes good sense. Most of he information if a speech signal is contained int he lower end of the spectrum. It is easy to show that if we filter the speech with a low-pass filter with $1kHz$ cutoff frequency, we would still be able to understand a conversation, while loosing some of the finer individual speaker characteristics. However, if we filter the speech with a high-pass filter with $1kHs$ cutoff frequency, the words in the conversation will be much harder to make any cognitive sense. It should be noted that in the presence of impulsive noise, classical wavelet transform does not work as the impulse corruption spectrum is flat and it shows up on both the approximation and the details signals. Wavelet transform decomposition and reconstruction methods are depicted in 3.4 and 3.5, where a down means decimation and an up arrow means interpolation.

3.4 The Inverse Filter

The Inverse Filter might sound elusive, however, the concept makes sense for outlier identification. The concept is to estimate the $AR(p)$ model for a given frame, and if you apply

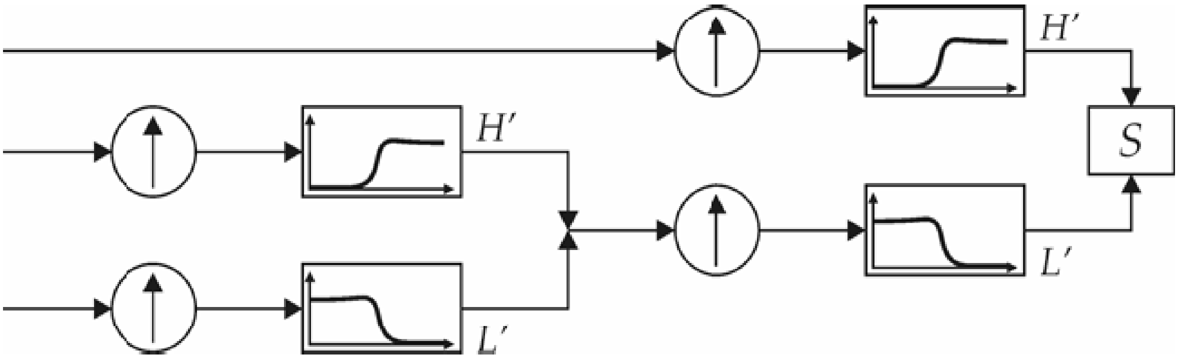

Figure 3.5: The Wavelet Transform Reconstruction

the inverse filter of the AR model to the input speech, the output will be white. If there are noise corruptions, specifically impulse corruptions, the inverse filter will have a significant effect, in terms of amplitude, on the inverse filter output. The objective then is to determine a threshold that is significantly larger for impulse corruptions than the mean error signal variance. This is plausible because when you apply the inverse filter, the output of correlated speech samples will be suppressed and become noise like, while impulsive noise samples will be smeared and transformed to a scaled version of the impulse response of the LPC inverse filter. Another word, the scale of speech signal is reduced to white noise, while the scale of the noise remains unchanged or, in most cases, increases. Assuming the inverse filter output is a white Gaussian random process, a filter matched to the LPC inverse filter can further improve smeared impulsive noise detection ability. However the key is the robust estimation of the AR model; otherwise the inverse filter performs very poor in maximizing the corruptions in the error signal and hence identifying the outliers. Indeed a robust method in identifying the outliers is the essential first step. However, the problem of replacing them is just as challenging. One such method is to replace the outliers in the error signal with

Gaussian noise and synthesize the estimated, filtered signal, while replacing only the outliers in the input frame.

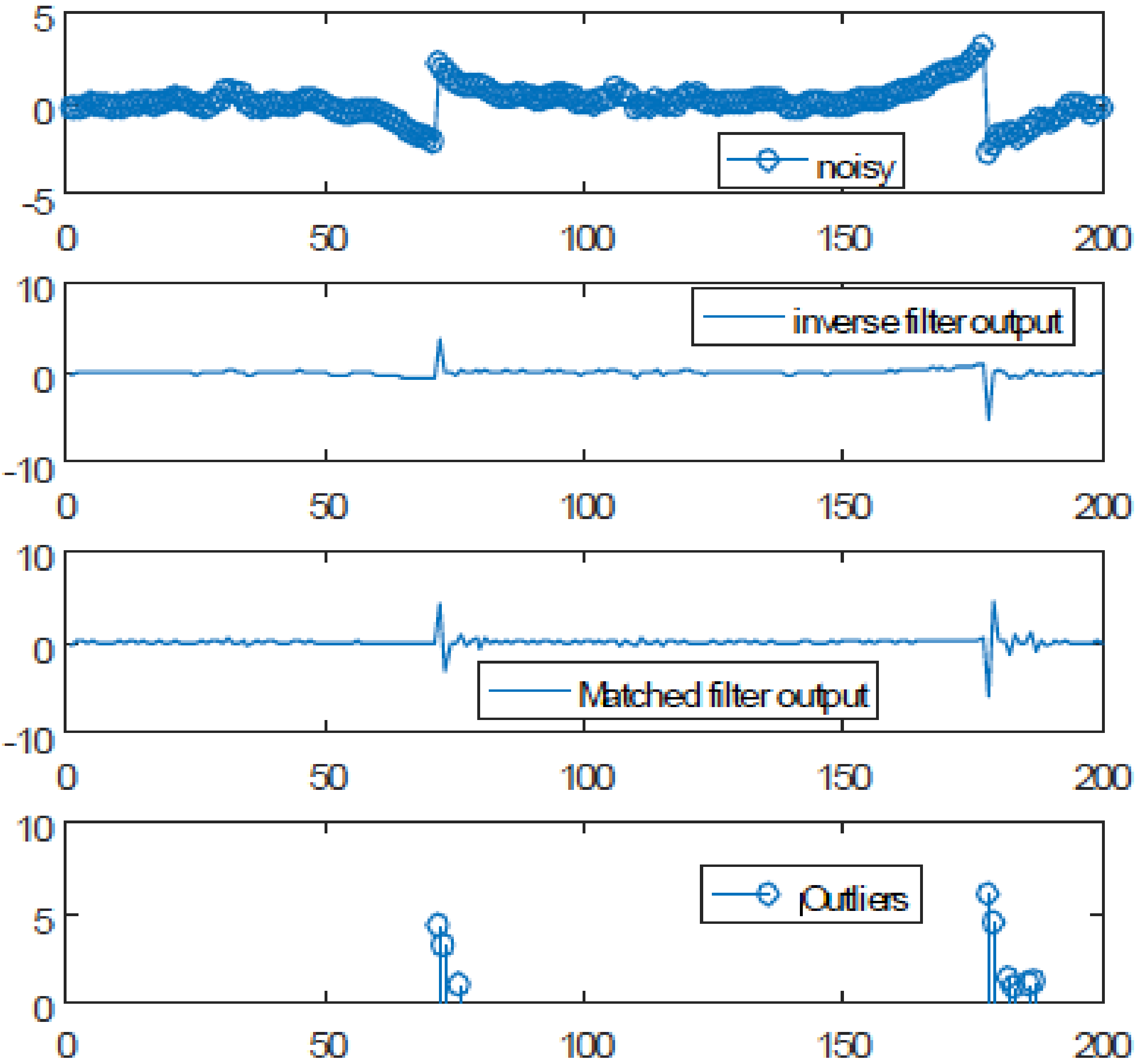

Figure 3.6: The Inverse Filter outlier identification

Chapter 4

Speech Encoder Algorithms

Encoding of the speech signal is how the compression gain is achieved. Classical digital telephony samples voice stream at 8ksps with 8 bits per sample linear quantizer, resulting in 64kbps, effective bandwidth. This may not be a reason for attention for wired communication such as Public Switched Telephone Network. However, when it comes to wireless communication systems, spectrum is very much limited and enterprises like AT&T, Intelsat, Iridium made it their mission to pack more user in a narrow-band voice channel, while maintaining speech quality as if you are in a 64kbps pipe. From the acoustic signal processing prospective, this translates into the highest possible compression ratio in speech signal encoding without compromising the fidelity. Hence voice encoder algorithms such as the MELP, the AMBE, the CELP and others alike that only requires 2.4kbps or lower pipe and still demonstrate acceptable speech synthesis quality are highly sought after. Our proposed speech compression algorithm is based on the MELP encoder, which is discussed throughout

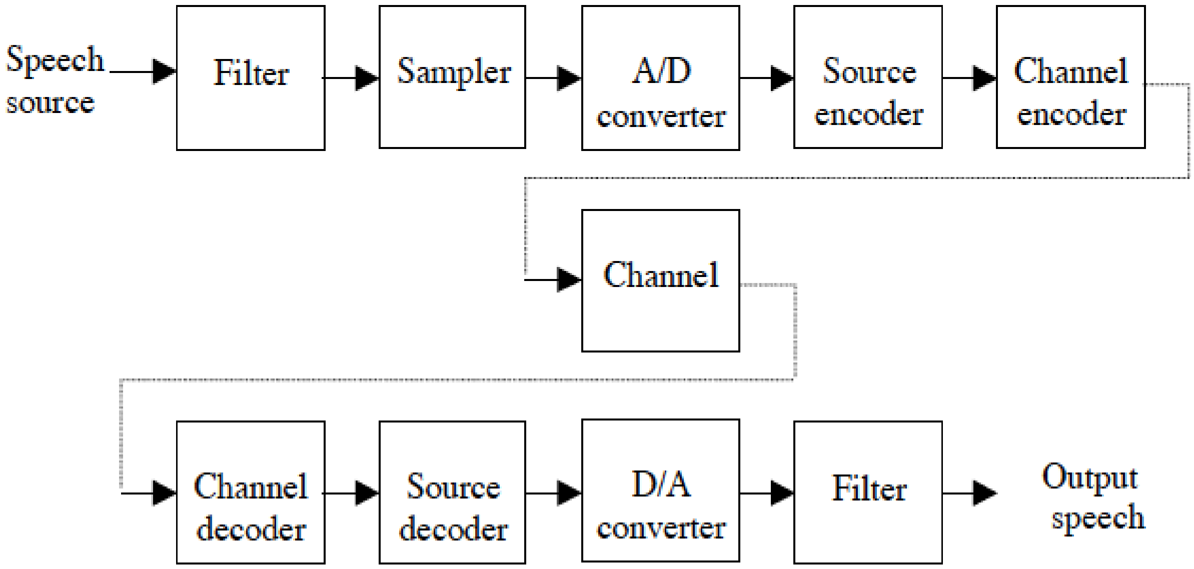

Figure 4.1: Speech Encoder System Model

this paper and in great details in a later section. Here we briefly discuss the three popular algorithms, namely, the VSELP, the CELP and the AMBE coding algorithms mentioned in the introduction.

4.1 The Code Excited Linear Prediction Encoder

In speech encoder algorithm, there are two main objectives, retain fidelity, while compressing the encoder information as much as possible. Another word, we hope to retain the fidelity of a 64kbps speech signal, as an example, into as low amount of bits/sec as possible. CELP encoder is another attempt in improving compression and quality performance of the LPC encoder. It was perceived that if we can define a finite code-book that can represent arbitrary excitation signal in the form of an AR model, then we only need to transmit the index of the

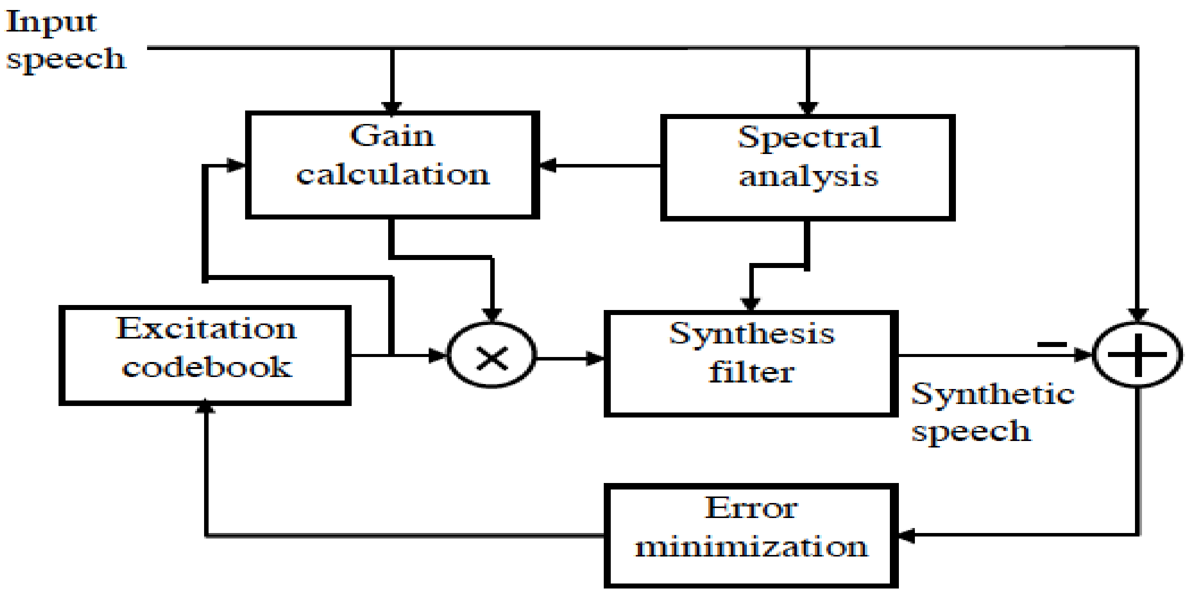

Figure 4.2: CELP Encoder Model

excitation code index, which in theory would result in huge compression gain. Because now you can encode the index of the LPC code, which is known to the synthesizer as opposed to encode all of the LPC filter coefficients. The basic CELP encoder process is also referred to as the Analysis-by-Synthesis (AbS) encoder. As shown in 4.2, the CELP encoder proposes to minimize the error between the input spectral shape and a code-book of quantized spectral shapes. The index of the code book spectral shape is chosen that minimizes error in the synthesized speech. Clearly, there is one fundamental problem to this method, that it is impossible to quantize the excitation signal so the error is negligible. This approach is promising, only if one can define a finite code book that can represent all possible excitation signals.

4.2 The Vector Sum Excited Linear Prediction Encoder

Originally developed by Motorola Corporation, the VSELP encoder algorithm is and evolution to the CELP encoder, with an attempt to lower code book size and hence improving computational cost. Excitation vectors from the stochastic codebook, however, are obtained through linear combination of a number of fixed basis vectors—hence the name of vector sum excitation. The VSELP coder was designed to achieve the highest possible quality with reasonable computational complexity while providing robustness to channel errors, essential requirements for cellular telephony applications. Figure 4.3, shows the core encoding structure. For this coder, a frame consists of 160 samples, and a subframe contains 40 samples. A total of $2^7 = 128$ codevectors are included in each stochastic codebook, with each codevector having 40 elements.

4.3 The Advanced Multi-band Excitation Encoder

The AMBE encoder is a variation of the Multiband Excitation Encoder. The excitation signal and the spectral envelopes are estimated simultaneously so that the synthesized spectrum is closest in the least squares sense to the spectrum of the original speech. This encoder also falls under the general class of analysis by synthesis encoders. Analysis consists of prefiltering, parameter estimation, quantization, and coding. Parameter decoding and frame-by-frame

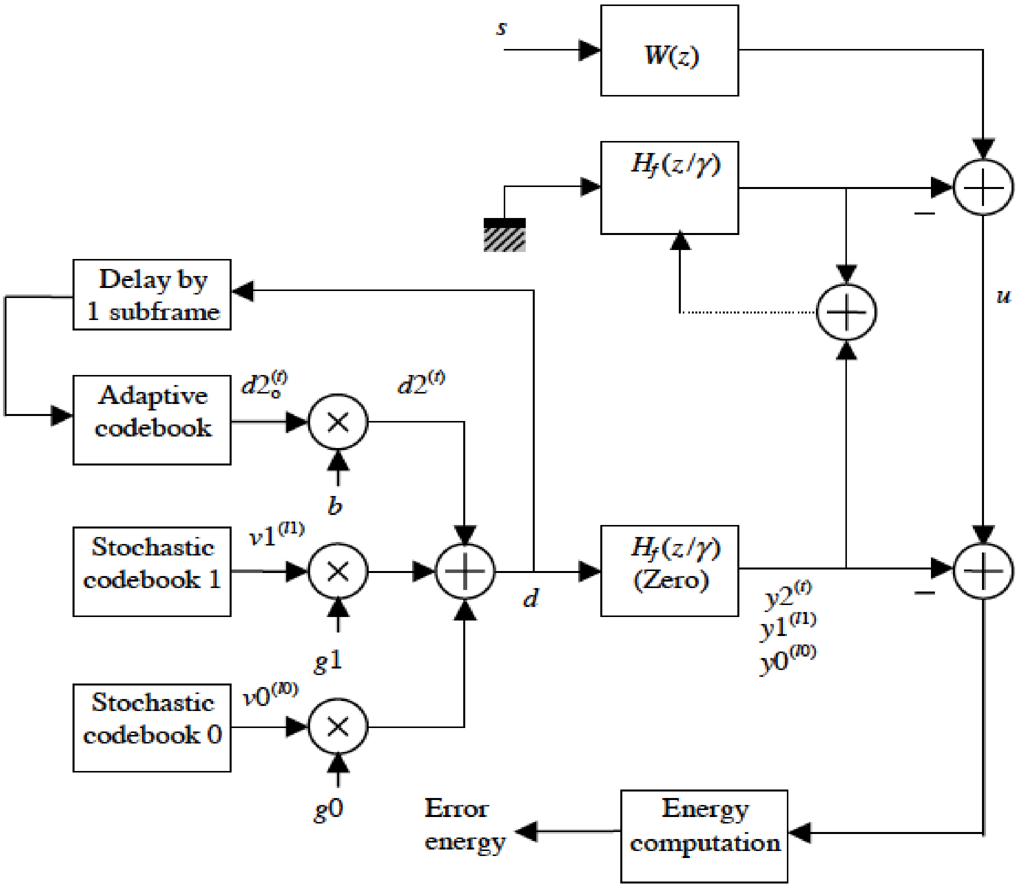

Figure 4.3: VSELP Encoder Model

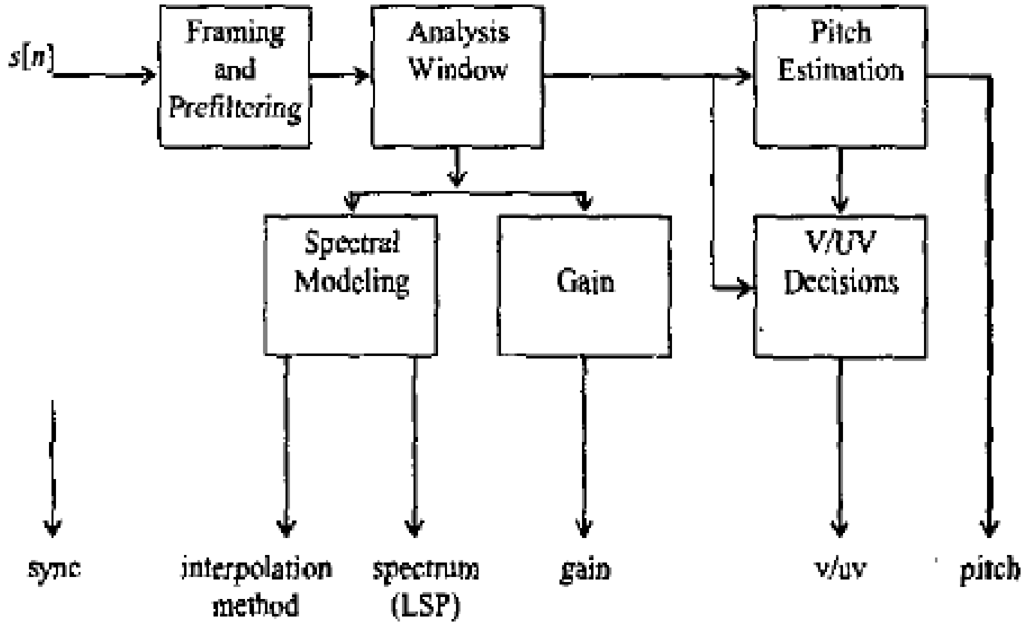

Figure 4.4: AMBE Encoder Model

reconstruction of the coded speech form the synthesis stage. The relevant parameters which are used to represent the input speech waveform are fundamental frequency (pitch), vocal tract spectral estimate, voicing decisions, and frame gain. An example block diagram of the AMBE analyzer is shown in Figure 4.4.

Chapter 5

Robust Correlation Estimation Methods

The most widely used estimator of correlation is the Gaussian Maximum Likelihood Estimator (GMLE) of the Pearson's ρ based on the sample mean and the sample covariance. Unfortunately, in the presence of impulsive noise, this estimator breaks down. This motivates us to explore the application of four robust alternative correlation estimators to speech processing, which are defined next. Specifically we evaluate the asymptotic bias and variance properties [15] to isolate the optimum robust estimator for speech processing.

Moronna [15] defined asymptotic bias as $b = |T(F) - T(G)|$, where $G = (1 - \epsilon)F + \epsilon H$, is the gross error model, F is the true distribution and H is the contamination distribution. It is clear that for a given distribution, the bias of the estimator is simply the difference between the true value and estimate. It is also clear that maximum bias is achieved when $T(G) = 0$, at some arbitrary contamination rate ϵ. This contamination rate is known as the breakdown point. Part of our research is to reveal which one of the estimators can withstand highest breakdown point ϵ, making it robust compared to others.

Given a probability distribution F, minimum asymptotic variance is attained with the maximum likelihood estimator and in short speech frame situation, it is the GMLE. However, the robust estimators we are evaluating in this paper are clearly not the GMLE. As a result there is a cost in terms of increased asymptotic variance of the estimator, in contrast to having better bias performances. Marona defined efficiency as $\eta = VAR(T_{MLE}(F)/VAR(T(F)$ [15], as the ratio of the variances of the MLE and the variance of the estimator in question. In addition to quantitative bias and variance performance of the four robust estimators, we will evaluate auditory perception at different bias and variance thresholds utilizing the Perceptual Evaluation Quality Estimator (PESQ)[16].

5.1 The Arcsin Law

If the noise power is equal or greater than the desired signal power, large distortions of the signal amplitude and power spectrum will result. Therefore, it is very important to analyze the noise power and its spectrum characteristics when developing robust correlation estimators. In the early 1940's, Van Vleck [17] developed a mathematical theory of the clipped signal spectrum. He showed that if the clipping is not down to more than the Root-

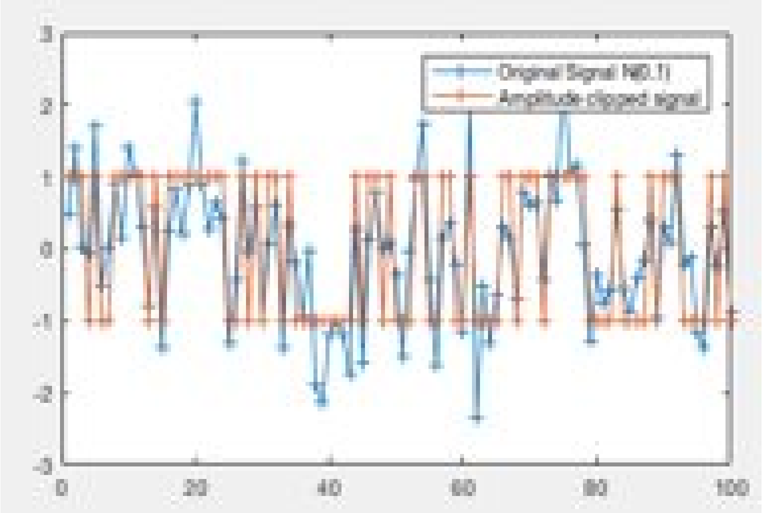

Figure 5.1: Hard limited speech signal.

Mean-Square (RMS) level before limiting, which is equivalent to clipping at about 1.4 times the RMS level, there is practically no distortion of the power spectrum. Even in the case of extreme clipping or hard limiting signal amplitude to +1 or −1, the distortion of the spectrum due to the presence of harmonics in the signal is small.

Given a real, zero mean Gaussian stationary process $\{z(t); -\infty < t < \infty\}$ with autocorrelation function with time lag τ, $R_{zz}(\tau)$ and Normalized Auto-Correlation Function (NACF), $\rho_{zz(\tau)}$, hard limiting $z(t)$ presents a unit variance process $\{y(t) = sign[z(t)]\}$ and its NACF, $\rho_{yy} = \frac{1}{\sigma_{yy}^2} E\{y(t)y(t+\tau)\} = \frac{2}{\pi} arcsin[\rho_{zz}(\tau)]$. In other words, if $z(t)$ is a stationary ergodic process, $\rho_{zz}(\tau)$ can be calculated from $\rho_{yy}(\tau)$ using the relationship given by

$$\rho_{zz}(\tau) = sin[\frac{\pi}{2}\rho_{yy}(\tau)]. \tag{5.1}$$

5.2 The Complex Polarity Coincidence Correlator

McGraw and Wagner [18] have shown that Van Vleck's work [17] on the ASL is applicable to any elliptically symmetric distribution. As a special case, we consider a stationary zero-mean circular complex Gaussian process, $\{z(t) = \{z_r(t) + jz_i(t)\}; -\infty < t < \infty\}$, with a bivariate probability density function,

$$f_Z(\underline{z}) = \frac{1}{\sqrt{\pi^2 det(\mathbf{R})}} e^{-\underline{z}^T \mathbf{R}^{-1} \underline{z}}, \tag{5.2}$$

where $\mathbf{R}$ is the Hermitian symmetric, positive-definite covariance matrix. Its complex-valued NACF is defined as $\rho_{zz}(\tau) = \frac{R_{zz}(\tau)}{\sigma_z^2} = \rho^r_{zz}(\tau) + \rho^i_{zz}(\tau)$, where $R_{zz}(\tau) = E[z(t)z^*(t+\tau)]$ is the autocovariance function and σ_z^2 is the variance of $z(t)$. The NACF $\rho_{zz}(\tau)$ can be calculated using (5.1), with the exception that the time shifted version of the signal is complex conjugate. Jacovitti [13] proposed a "hard limited" process, $y(t) = sign[z_r(t)] + jsign[z_i(t)]$, obtained through a constant magnitude, memoryless complex nonlinear transformation. In the complex phase plane, the simplest hard limiting transformation results in four discrete phases, where both the real and the imaginary parts can be +1 or −1. This can be thought of as four-level phase quantization in the phase plane, which intuitively may increase the phase error in the complex-valued NACF estimate. This is known as the CPCC, and for the ideal infinite-level phase quantization is known as the PPC, which will be discussed in the next section.

By extending (5.1) in the complex domain, the CPCC can be written as

$$\rho_{yy}(\tau) = E\{csign[z(t)]csign^*[z(t+\tau]\}, \tag{5.3}$$

where $csign[z(t)] \triangleq \frac{1}{\sqrt{2}}\{sign[z_r(t)]+jsign[z_i(t)]\}$. For a discrete signal of length N, the above expectation can be computed numerically as $\rho_{yy}(\tau) = \frac{2}{\pi}\{arcsin[\rho^r_{zz}(\tau)] + jarcsin[\rho^i_{zz}(\tau)]\}$. Ultimately, the CPCC can be seen as the extension of the ASL in the complex domain for the special case where the phase plane is quantized to four levels on the unit circle. This condition will force the *csign* operator to produce complex numbers in the set $[1+j, 1-j, -1+j, -1-j]$. Similarly to the ASL, we can write the NACF of the CPCC as $\rho^{CPCC}_{zz} = sin[\frac{\pi}{2}\rho^r_{yy}(\tau)] + jsin[\frac{\pi}{2}\rho^i_{yy}(\tau)]$.

In order to evaluate the variance of the CPCC estimator, we realize that $\rho^{CPCC}_{yy} \leq 1$, yielding $\rho^{CPCC}_{yy} = \frac{1}{2}[\rho_{y_r y_r}(\tau) + \rho_{y_i y_i}(\tau)] + j\frac{1}{2}[\rho_{y_i y_r}(\tau) - \rho_{y_r y_i}(\tau)]$. It is apparent that the variances of the real and the imaginary components of the CPCC are respectively given by

$$E[Re(\rho^{CPCC}_{yy}(\tau))^2] = \frac{1}{4}E[\rho^2_{y_r y_r}(\tau)] + \frac{1}{4}E[\rho^2_{y_i y_i}(\tau)] + \frac{1}{2}E[\rho_{y_r y_r}\rho_{y_i y_i}], \tag{5.4}$$

$$E[Im(\rho^{CPCC}_{yy}(\tau))^2] = \frac{1}{4}E[\rho^2_{y_i y_r}(\tau)] + \frac{1}{4}E[\rho^2_{y_r y_i^2}(\tau)] - \frac{1}{2}E[\rho_{y_i y_r}\rho_{y_r y_i}]. \tag{5.5}$$

Since y_i and y_r have identical statistics, from the Schwart's inequality, we can claim that the variance of the complex-valued correlation estimate is more efficient than the real-valued

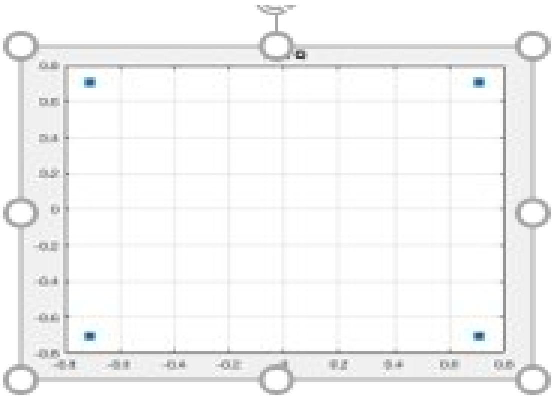

Figure 5.2: The CPCC signal space.

one, that is, $E[Re(\rho_{yy}^{CPCC}(\tau))^2] \leq E[\rho_{y_r y_r}^2(\tau)]$ and $E[Im(\rho_{yy}^{CPCC}(\tau))^2] \leq E[\rho_{y_i y_i}^2(\tau)]$. For sufficiently large sample size, Jacovitti and Neri [13] derived the variances of the real and the imaginary parts of the CPCC as

$$E[Re(\rho_{zz}^{CPCC}(\tau))^2] \cong \frac{\pi^2}{8N}(1 - (\rho_{zz}^r)^2)[1 - \frac{4}{\pi^2}arcsin^2(\rho_{zz}^r) - \frac{4}{\pi^2}arcsin^2(\rho_{zz}^i)], \quad (5.6)$$

$$E[Im(\rho_{zz}^{CPCC}(\tau))^2] \cong \frac{\pi^2}{8N}(1 - (\rho_{zz}^i)^2)[1 - \frac{4}{\pi^2}arcsin^2(\rho_{zz}^i) - \frac{4}{\pi^2}arcsin^2(\rho_{zz}^r)]. \quad (5.7)$$

Note that the approximations of the Gaussian variates by the linear terms of the Taylor's series expansion of the associated characteristic function are used to derive the variance of the CPCC.

5.3 The Phase-Phase Correlator

Jacovitti and Neri [13] proposed the PPC estimator based on Reeds' derivation of the cross-correlation function of two general, envelope-distorting filters [19], where the joint probability density of the envelope is a Gaussian process. The PPC can be viewed as a generalization of the CPPC since the complex-valued signal is also hard limited while retaining all the phase information without quantization. By contrast, the CPCC estimator assumes four discrete points in the complex phase plane, which results in the loss of the phase information due to quantization. Jacovitti and Neri [13] classified the CPPC as a fine phase quantized estimator. For both the PPC and the CPCC, we consider the process $\{z(t) = \{z_r(t) + jz_i(t)\}; -\infty < t < \infty\}$ obeying a joint bi-variate circularly complex Gaussian probability distribution, the phase preserving hard-limiting process, $y(t) = \frac{z(t)}{|z(t)|} = e^{jArg[z(t)]}$, where $Arg[z(t)]$ is the phase of complex process $z(t)$. The PPC can be applied to calculate the NACF of $y(t)$ at lag τ as $\rho_{yy}(\tau) = E[e^{jArg[z(t)]}e^{jArg[z(t+\tau)]}]$, which is related to $\rho_{zz}(\tau)$ via $\rho_{yy}(\tau) = \frac{\pi}{4}\rho_{zz}(\tau)_2F_1(\frac{1}{2}, \frac{1}{2}, 2; |\rho_{zz}(\tau)|^2)$, where ${}_2F_1(a, b; c; z)$ is the Gaussian hypergeometric function defined as

$$ {}_2F_1(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \sum_{n=0}^{\infty} \frac{\Gamma(a+n)\Gamma(b+n)}{\Gamma(c+n)} \frac{z^n}{n!}. \tag{5.8} $$

It can be seen from (5.8) that ${}_2F_1(a, b; c; z)$ is a real function and hence, it follows $|\rho_{yy}(\tau)| = \frac{\pi}{4}|\rho_{zz}(\tau)|_2F_1(\frac{1}{2}, \frac{1}{2}, 2; |\rho_{zz}(\tau)|^2)$, and since the PPC is phase preserving, we may write $Arg[\rho_{yy}(\tau)] = Arg[\rho_{zz}(\tau)]$. For complex-valued signals, if $z(t)$ is a stationary ergodic process, an estimate

$\hat{\rho_{zz}}(\tau)$ can be calculated from an estimate of $\hat{\rho_{yy}}(\tau)$, with the inverse function given by $\hat{\rho}_{zz}(\tau) = e^{jArg[\rho_{yy}(\tau)]}g^{-1}[|\rho_{yy}(\tau)|]$. For N-samples of a discrete-time hard-limited signal, an estimate $\hat{\rho}_{yy}(\tau)$ can be calculated as

$$\hat{\rho}_{yy}(\tau) = \frac{1}{N-\tau}\sum_{i=1}^{N-\tau} e^{\{jArg[z(t_i)]-Arg[z(t_i+\tau)]\}}. \tag{5.9}$$

It can be seen that $|\hat{\rho}_{yy}(\tau)|$ is less than one, which forces $|\hat{\rho}_{zz}(\tau)|$ also to be less than one. This property is particularly attractive in estimating the AR parameters for speech processing since it ensures that the synthesis filter is stable.

The variances of the real and the imaginary parts and the covariance of the PPC estimator can be expressed in terms of the hypergeometric function as

$$\begin{aligned} E[Re(\rho_{yy}^{PPC}(\tau))^2] \cong \frac{1}{N}\Big\{\frac{1}{2} + \frac{1}{4}[[\rho_{zz}^r]^2] - [[\rho_{zz}^i]^2]_2F_1(1,1,3;|\rho_{zz}(\tau)|^2) \\ - \frac{\pi^2}{16}[\rho zz^r]^2\big[{}_2F_1(\frac{1}{2},\frac{1}{2},2;|\rho_{zz}(\tau)|^2)\big]^2\Big\}, \end{aligned} \tag{5.10}$$

$$\begin{aligned} E[Im(\rho_{yy}^{PPC}(\tau))^2] \cong \frac{1}{N}\Big\{\frac{1}{2} - \frac{1}{4}[[\rho_{zz}^r]^2] - [[\rho_{zz}^i]^2]_2F_1(1,1,3;|\rho_{zz}(\tau)|^2) \\ - \frac{\pi^2}{16}[\rho zz^i]^2\big[{}_2F_1(\frac{1}{2},\frac{1}{2},2;|\rho_{zz}(\tau)|^2)\big]^2\Big\}, \end{aligned} \tag{5.11}$$

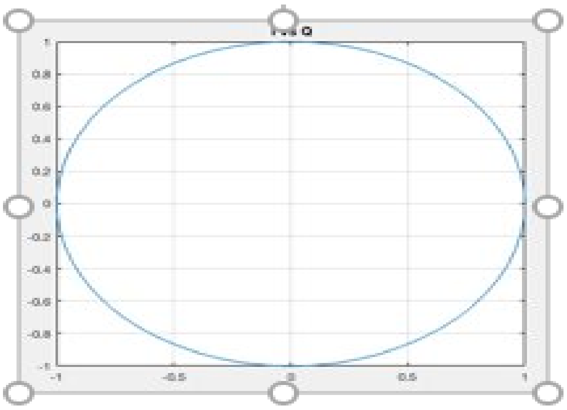

Figure 5.3: The PPC signal space.

$$Cov[Re(\rho_{yy}^{PPC})Im(\rho_{yy}^{PPC})] \cong \frac{1}{N}\Big\{\frac{1}{2}\rho_{zz}^{r}\rho_{zz}^{i}{}_2F_1(1,1,3;|\rho_{zz}(\tau)|^2) \\ + \frac{\pi^2}{16}\rho_{zz}^{r}\rho_{zz}^{i}\big[{}_2F_1(\frac{1}{2},\frac{1}{2},2;|\rho_{zz}(\tau)|^2)\big]^2\Big\}. \quad (5.12)$$

However, a closed form representation of the inverse of the variances of the real and the imaginary part and their covariance matrices does not exist; hence, look up tables are used to calculate the variance of $\hat{\rho}_{zz}^{PPC}(\tau)$.

5.4 The Median-of-Ratios Estimator

An estimator of the correlation ρ for the bi-variate Gaussian distribution using the sample median assumes that the two sets of observations obey a zero-mean bi-variate probability density function written as

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}}e^{-\frac{1}{2\sigma^2(1-\rho^2)}(x^2+y^2-2\rho xy)}. \quad (5.13)$$

As shown in [20], a random variable V equal to the ratio X/Y follows a Cauchy distribution with density function, $f_V(v) = \frac{\sqrt{1-\rho^2}}{\pi(1-2\rho v+v^2)}$ and cumulative probability density function, $F_V(v) = \frac{1}{2} + \frac{1}{\pi} arctan\left(\frac{v-\rho}{\sqrt{1-\rho^2}}\right)$. It follows that when $v = \rho$, $F_V(v) = \frac{1}{2}$, which is the median. For a zero mean Gaussian stationary time series MRE of correlation is defined as $\hat{\rho}_x(\tau) = \hat{\xi} F_V(v)$, where $\hat{\xi}$ is the sample median and $F_V(v)$ is the cumulative probability density function of the ratio $\frac{X_{t+\tau}}{X_t}$. For a complex-valued signal, Tamburello and Mili [21] showed that the independent application of the coordinate-wise median of the ratio to the real and the imaginary components is justified. Given a zero mean Gaussian random process X_t, if $Y_t = \frac{X_{t+\tau}}{X_t}$ is defined as a coordinate-wise ratio at some lag τ, then the MRE is expressed as $\hat{\rho}_x(\tau) = \hat{\xi}(Re\{Y_t(\tau)\}) + \hat{\xi}(Im\{Y_t(\tau)\}))$.

Chapter 6

Robust Pre-Filtering Methods

Impulsive noise is a naturally occurring phenomenon in communication systems, which can be thought of as sharp keystrokes, discontinuities in transmission, fast fading, interference, etc. Gandhi, Ledoux and Mili [22] proposed an impulsive noise model using two exponential functions of opposite signs, where a discontinuity occurs at the changing of the signs. While an impulse can last between 1ms to 5ms, it behaves like the Dirac impulse in the time domain, resulting in a flat power spectrum, which makes it impossible to filter out in frequency domain. Study shows that in the presence of impulsive noise, the robust PPC estimator of the encoder parameters still produces some noisy artifacts in the synthesized audio. Indeed, speech encoder requires the estimation of the Fourier magnitudes of the error signal; consequently, if the original signal is corrupted, the error signal, even when robustly estimated with a robust AR model, will also be corrupted. Now, because the parameters encoded in the speech coder model are the Fourier magnitudes, which are needed to generate the excitation of the AR model for speech synthesis, the synthesized speech will exhibit noisy artifacts if the Fourier magnitudes are corrupted. One of our contributions in speech encoder parameter estimation is the development of a pre-filtering algorithm based on the Robust Mahalonobis Distance (RMD) method, which is entirely implemented in the time domain. It consists of two main steps: (1) a robust MD is first calculated using the PPC to identify the outliers and (2) the flagged outliers are then replaced by estimated values obtained from a robust AR model, also estimated with the PPC.

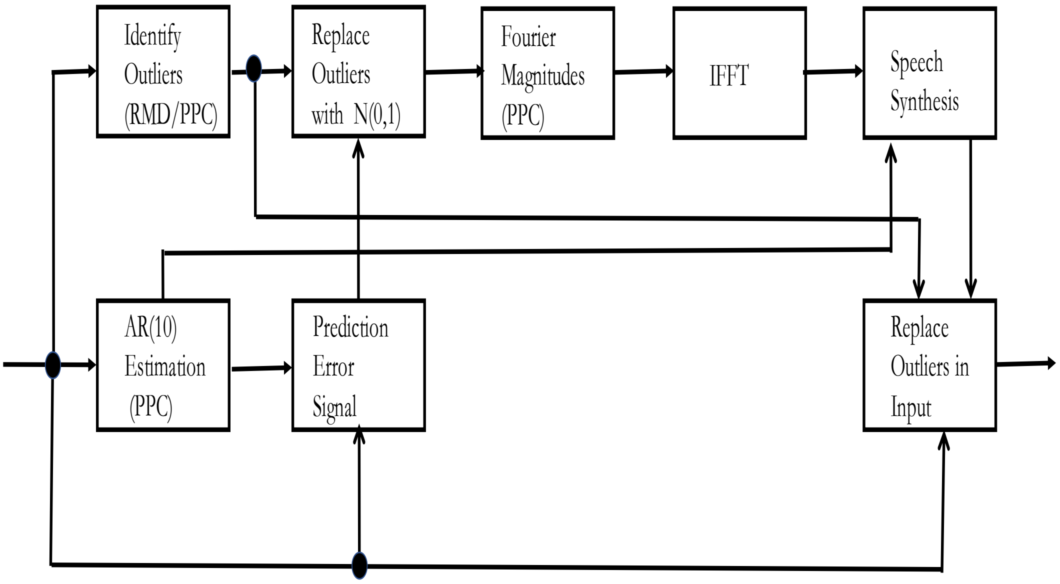

Figure 6.1: Flowchart of the proposed robust pre-filtering algorithm.

6.1 A Robust Mahalanobis Distance

An outlier may be induced by a lighting on a transmission line as an impulsive noise, which is easy to identify. However, it is also possible, i.e. in fast fading, that the noise amplitude is at the same level as the true signal amplitude. In that case, the outlier identification becomes much more difficult and requires sophisticated algorithms to be reliability fulfilled. A reliable method consists in applying a statistical test to robust MD values as described next. The MD of an n-dimensional vector, $\mathbf{h}_i$, which is the i-th column vector of the observation matrix $\mathbf{H}^T$ and has a sample mean given by $\overline{\mathbf{h}}$, is defined as $MD_i = \sqrt{(\mathbf{h}_i - \overline{\mathbf{h}})^T \mathbf{C}^{-1} (\mathbf{h}_i - \overline{\mathbf{h}})}$; it is a measure of distance of the associated poin with respect to the bulk of the point cloud. If we assume that the $\mathbf{h}_i$'s are drawn from a normal distribution, $N(\mu, \mathbf{C})$, then MD_i^2 will approximately obey a chi-squared distribution with n-degrees of freedom, that is, χ^2_n. Classical outlier identification method flags all the data points having $MD_i > \sqrt{\chi^2_{n,0.975}}$. However, this method is not robust because it is prone to the masking effect; indeed, a sufficiently large outlier can bias the sample mean and inflates the sample covariance matrix to the point where a second outlier closer to the bulk will stand below the confidence threshold $\sqrt{\chi^2_{n,0.975}}$, resulting in the breakdown of the method. Our robust version of MD replaces $\overline{\mathbf{h}}$ with the median of $\mathbf{h_i}$ and the ij-th element of the covariance matrix, $\mathbf{C}$, with $C_{ij} = \sqrt{E[\{\mathbf{z}_i - med(\mathbf{z}_i)\}^2] E[\{\mathbf{z}_j - med(\mathbf{z}_j)\}^2]}$, whose normalized version is calculated with the PPC.

6.2 Robust AR Model Estimation

Now that we have described a robust way of identifying outliers in the pre-filtering process step, the next step is to robustly estimate the desired signal segment that has been corrupted with the impulsive noise. Here, outlier replacement is considered, that is, a set of signal samples have been replaced with outliers, which is the worst case scenario as compared to outlier addition. After being identified, the outliers are replaced with reliable estimated values from a robust AR model. As mentioned earlier, the speech signal is non-stationary, which makes the AR parameter estimation of a large ensemble unrealistic. To get around this problem, the speech signal is first segmented into smaller frames, which may be considered to be piece-wise stationary. In low data rate speech compression, specifically the signal is sampled at 8 kHz with 8 sps, and research has shown 22.5 ms, yielding 180 samples, to be an optimum frame length. It is important to note that a frame length plays a significant role in model parameter estimation, affecting the quality of speech synthesis. Further discussions on frame length will be provided on later sections.

Speech signal spectral response is very well represented by an AR(p) model, which is also the fundamental building block for any LPC. This time series model assumes that a sample value at time k can be written as a linear combination of previous sample values up to time $k-p$ plus a white noise. Formally, we have

$$z_k = \sum_{i=1}^{p} a_i z_{k-i} + e_k, \tag{6.1}$$

where $\{a_1, ..., a_p\}$ are the parameters of the model and e_k is the error value. Assuming that N samples, z_1,..., z_N, are available, we can write $(N-p)$ equations (12) for $k = p+1, ..., N$, which can be put in a matrix form as $\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{e}$, where $\mathbf{z}$, $\mathbf{x}$, $\mathbf{e}$ are column vectors and

H is the observation matrix. The two standard methods for estimating the AR(p) model parameters are the least square estimator and the Levinson-Durban recursion, which is derived from the Yule-Walker equations. The least squares parameter estimation minimizes the sum of the squared residuals, where **e** is assumed to be a vector of independent Gaussian random variables, with zero mean. The Levinson-Durban recursion provides a fast solution involving a Toeplitz correlation matrix. Under ideal conditions, both of these methods can result in stable parameter estimation. When estimating the auto-correlation function, they implicitly assume that the beginning and the end of the time series samples are zero. This is what Burg refers to as ”end effect” [23]. Therefore, when we are dealing with near periodic signals, they can lead to unstable AR model estimation, mainly due to an ill-conditioned covariance matrix.

Burg [23] proposes instead to operate on the time series data in iterations and estimate the reflection coefficients that minimizes forward and backward prediction error vectors $\mathbf{f}_i$ and $\mathbf{b}_i$. Specifically, given a discrete, zero mean, Gaussian stationary random process, $\{z(n); 0 \leq n < N\}$, the reflection coefficient Γ_i, that minimizes the prediction error is defined as

$$\Gamma_i = -\frac{2\mathbf{b}_i^H \mathbf{f}_i}{\mathbf{f}_i^H \mathbf{f}_i + \mathbf{b}_i^H \mathbf{b}_i}, \tag{6.2}$$

where $\mathbf{f}_i = x(1 : N - i - 1)$ and $\mathbf{b}_i = x(0 : N - i - 2)$. The prediction error vectors are then updated as $\mathbf{f}_{i+1} = \mathbf{f}_i + \Gamma_i \mathbf{b}_i$ and $\mathbf{b}_{i+1} = \mathbf{b}_i + \Gamma_i^* \mathbf{f}_i$ to calculate Γ_{i+1} using (6.2) until iteration i reaches order p. It can be shown that indeed, the prediction error minimizing reflection coefficients satisfy the condition, $\{0 \leq \Gamma_i < 1; \forall i\}$. Although the Burg's algorithm is more stable than the least square method and the Levinson-Durban recursion, they are all formulated given the assumption that the noise obeys a normal distribution. In the presence of impulsive noise, all three methods break down. It can be easily seen from (6.2) that large amplitude outliers can cause partial correlation estimate to breakdown, resulting in an unstable AR(p) model. In order to robustify the Γ estimation, any of the four previously discussed methods, the ASL, the CPPC, the PPC or the MRE can be utilized. However, considering the bias and variance trade-offs, we proposed to use the PPC with the Burg's method to estimate the AR(p) parameters from noise corrupted data.

For a large data set, study shows that a speech signal closely obeys the Laplace distribution. However, when the speech frame is short, such as 22.5 ms in our low data rate encoder model, piece-wise frames more closely follow a Gaussian distribution. Therefore, we propose to replace the outliers with prediction estimates obtained from an AR(p) model excited with zero-mean, unit variance normal process.

6.3 Outlier Replacement Method

To circumvent the difficulty in robustly generating the excitation signal directly from corrupted speech that will produce good quality synthesized speech, we pre-process the input speech with a non-linear, time domain, filter algorithm before carrying out the encoder parameters estimation. A convenient way to filter the signal is to replace the signal value, one sample at a time, by the output estimate of the AR model. However, this is computationally intensive, and therefore, may be too slow for real time applications. Instead, we propose to do the following. For each region of clustered outliers, an analysis window of 180 samples that is centered at the midpoint of the outlier cluster is formed from the 360 sample running buffer. Next, an AR(20) model is estimated via the robust PPC-based Burg's algorithm. The speech frame is then filtered with the inverse of the AR(20) model to generate the error signal. Now, because the latter is estimated using the corrupted signal, the associated residuals retain the impulses. Furthermore, because the inverse filter has an opposite effect on the error signal, it tends to amplify the error signal samples corresponding to the impulses instead of forcing them to behave like a white noise. To address this probem, we propose to replace the corrupted error signal samples with samples drawn from a standard Gaussian distribution, $N(0,1)$. Then, we estimate a robust PSD from the PPC and take its inverse Fourier transform to generate the robust excitation signal. Finally, we apply this excitation signal to the input of the robust AR(20) model to produce a synthesized audio signal. Here, only the samples tagged as outliers are replaced at the output of the prefilter. Great care must be taken with respect to the gain of the replacement samples. Inappropriate gain calculation can cause erratic transitions from replacement to adjacent samples, which can induce audio distortion.

Chapter 7

The Proposed, Robust Speech Encoder Algorithm

We now describe the MELP vocoder algorithm developed by McCree and Barnwell [2]. As portrayed in Fig. 2.4, the RMELP is a robust implementation of the MELP encoder algorithm with four fundamental differences. Specifically in RMELP, pitch period, voicing strength, AR(10) model and Fourier magnitudes are estimated with the robust estimator of correlation from (7.2). The RMELP analyzes the speech frame duration of 22.5 ms, sampled at 8000 samp/sec, which is equivalent to 180 samples that are 16-bit quantized. It defines a parametric model with merely 54 bits [2] and hence, it is able to achieve compression ratio of 53.33 over a traditional digital speech signal at 64 kbps. It should be reemphasized that we propose to utilize our robust estimator of correlation to estimate pitch period, voicing strengths, AR(10) model and Fourier magnitudes for frames tagged with outliers; otherwise

GMLE will be applied. In the following sections, we propose several improvements to the fundamental encoder blocks while making the RMELP encoder robust against impulsive noise.

7.1 Pitch Period Estimation and Voiced/Unvoiced Frame Identification

The first step of the MELP encoder analysis process consists in estimating the pitch period on each 22.5 ms frame, which is processed with a 1-kHz lowpass filter. For real-valued signal, $z_n(t)$, the pitch period is defined by a lag τ where the GMLE estimator of the Pearson's ρ given by

$$\hat{\rho}_{zz}(\tau) = \frac{\sum\limits_{n=-\lfloor\tau/2\rfloor-80}^{-\lfloor\tau/2\rfloor+79} z_n z_{n+\tau}}{\sqrt{\sum\limits_{n=-\lfloor\tau/2\rfloor-80}^{-\lfloor\tau/2\rfloor+79} z_n z_n \sum\limits_{n=-\lfloor\tau/2\rfloor-80}^{-\lfloor\tau/2\rfloor+79} z_{n+\tau} z_{n+\tau}}} \tag{7.1}$$

reaches its maximum. In this paper, it is calculated for lag $\tau = 40, 41, ..., 160$, where the lag values have been carefully chosen. For speech signal samples at 8000 ksps, the pitch period τ translates into fundamental frequencies between 50 and 200 Hz. Naturally, female and male speech signals tend to fall within and on the higher end of the spectrum, respectively.

Furthermore, voiced and unvoiced decisions are also determined based on the values taken by $\hat{\rho}_{zz}(\tau)$, since the voiced frames are highly periodic and have much higher correlation coefficients than the unvoiced frames. Robust estimation in pitch period and voiced/unvoiced frame classification is critical to preserving the fidelity of the speech signal. Impulsive noise in either voiced or unvoiced frames can cause the encoder to falsly identify the frames and incorrectly estimating the fundamental frequency. Our research shows that these miss classifications can severely degrade the quality of the synthesized speech signals. In order to robustify the pitch period estimation and the voiced/unvoiced frame classification, we propose to use the PPC.

As portrayed in Fig. 2.4, the MELP is an extension of the classical LPC model with two fundamental additions, an aperiodic pulse excitation for the transition frames and an periodic and noise mixed excitation for the voiced and unvoiced frames. Other improvements are an adaptive spectral enhancement, a pulse dispersion filter and a Fourier-magnitude-based excitation generation. Speech signals exhibit strong time correlations (larger than 0.6) for time lags up to the 10-th order. The MELP analyzes the speech frame duration of 22.5 ms, sampled at 8000 samp/sec, which is equivalent to 180 samples that are 16-bit quantized. It defines a parametric model with merely 54 bits [2] and hence, it is able to achieve compression ratio of 53.33 over a traditional digital speech signal at 64 kbps.

The encoder maintains 360 samples in a circular buffer, that is, 180 samples of the current frame, 90 samples of the previous frame, and 90 samples of the next frame. The speech signal analysis window varies depending on the specific parameter to be estimated. The estimation of the pitch period, which corresponds to the fundamental frequency of the signal, is the first and critical step of the RMELP encoder algorithm. All subsequent parameter estimations are dependent upon the accuracy of the fundamental frequency estimate and the correct classification of the original speech signal into voiced, or unvoiced, or transition frames. Hence, the widest analysis window is used to estimate the pitch period. It is clear from (7.2) that for a given lag τ, the NACF is a function of 160 samples. However, as τ takes on values between 40 and 160, the center of the correlation window can also shift by as much as 160 samples. As a result, a total of 320 samples analysis window is required to calculate the pitch period, where 160 samples belong to the current frame and 160 samples to the next frame. Prior to the pitch period estimation, the input speech is processed with a high-pass filter to remove any low frequency hum below 60 Hz, and then is passed through a low-pass filter with cut-off frequency of 1 kHz. Furthermore, in order to increase the accuracy of the pitch period estimation, a fractional pitch period refinement is accomplished utilizing the PPC. For more details on fractional pitch refinement, the reader is referred to [1].

The fundamental frequency estimation is instrumental in accurately analyzing the band-pass voice strength of the upper four bands as described in Section 7.2. The band-pass voicing strength analysis is a two-step process. First, the NACF, $\hat{\rho}_{zz}(\tau)$, is calculated at the pitch lag. However, because the speech signal is non-stationary, it is possible that the pitch period may transition within a single frame. As a result, at higher frequencies this method can often provide poor correlation estimates. In order to circumvent this problem, we propose to use a second method initiated by McCree and Barnwell [2]. Here, the output of the band-pass filter is full-wave rectified to eliminate the high frequency transitions, which makes the generated signal to behave as if its envelopes rise and fall, in line with each pitch pulse. Then, the rectifier output is smoothed with a one-pole low-pass filter followed by a notch filter to remove DC component. The NACF is once again calculated from the full-wave rectified signal utilizing the PPC. Finally, the higher estimated value of the $\rho(\tau)$ of the above two methods is utilized to encode the pitch periodicity in each of the upper four bands. For the robust AR(10) model estimation, we use the Burgs' algorithm from (6.2). Bandwidth expansion is performed on the AR(10) model spectral response by multiplying each linear prediction coefficient by a factor of 0.994. Simulation results show that our proposed method of AR model estimation based on the Burg's algorithm is quite robust; consequently, an a priori removal of the outliers may not be necessary.

In the standard modeling procedure, which is different from the MELP, a speech signal is fragmented as binary, voiced or unvoiced frames. With a binary voicing decision, it is common to falsely identify a frame as voiced or unvoiced. In the former case, it will sound tonal while in the latter case, it will sound harsh. This problem is specially pronounced for female speakers with a higher fundamental frequency. By contrast, the MELP identifies a third type of frames, the transition frames, where the speech signal is neither highly periodic nor noise-like. For the transition frames, McCree and Barnwell [2] propose to remove the periodicity in the voiced excitation pulses by randomly varying pitch period by ±25%, which is proved to be a better fit in characterizing erratic glottal pulses in transition frames. In practice, the voiced frames are highly periodic and identified with high normalized correlation coefficients at the fundamental frequency while the transition frames are those with marginal correlation coefficients.

As for the LPC, it generates either periodic or white noise pulses that are used to excite an all-pole filter to synthesize voiced or unvoiced frames, respectively. Since the natural sound of a human voice is periodic with some level of added white noise, the LPC synthesizes output signals that sound as tonal or harsh. On the other hand, the MELP generates dynamically mixed excitation by combining periodic and white noise pulses in different proportion in each frame, determined by the periodicity intensity in different voicing frequency bands. The speech frames are filtered using five frequency bands and the periodicity at the fundamental frequency is calculated in each band. The periodic pulse excitation and the noise excitation intensity is determined in each band based on the normalized autocorrelation coefficients for each frame. The relative voice and noise power in each band characterizes the pulse shaping filter. The periodic and noise excitations are first filtered using the pulse shaping and the noise shaping filter, respectively. Here, the filters' outputs are added together to form the total excitation, known as the mixed excitation [24], since some portions of the noise and the pulse train are mixed together. Basically, mixed excitation along with the transition frame identification are the keys for improving the tonal or the buzzy quality sounds, hence making the speech signal to sound natural.

For voiced or transition frames, the periodic pulse excitation is still insufficient to synthesize a human-like voice. If the order of the AR model is sufficiently large, the error signal will be approximately white. However, for a low data rate speech encoder model, a higher order of the AR model costs more bandwidth; consequently, an AR(10) is proposed for the MELP encoder. If an inverse filter from the AR(10) is applied, the error signal does indeed remain colored, that is, the error signal still has important information that is critical in estimating optimum excitation impulses. In order to capture that information, the Fourier magnitudes are also encoded. On the synthesizer, the Fourier magnitudes are used to shape the periodic excitation sequence to closely estimate the encoder error signal.

The intended purpose of the spectral enhancement filter is to enhance the quality of the synthesized speech by closely matching the natural speech waveform in the formant regions [25]. According to [2], formants between pitch pulses do not decay as rapidly as they do at the all-pole filter output. If the poles are close to or greater than the unit circle, the LPC synthesis filter output may sound chirpy or even make the filter unstable. To address this problem, a bandwidth expanded pole-zero filter, estimated directly from the synthesis filter, has been proposed in [2]. The purpose of the pulse dispersion filter is to improve the quality of the synthesized speech signal in the frequency regions with low formant resonance. This filter is a 65-tap Finite Impulse Response (FIR) filter based on a spectrally flattened triangular pulse to create a time domain stretch to the synthetic speech, which makes it more natural sounding. By implementing the above signal processing blocks, the MELP is able to synthesize and mimic natural sounding human voice without buzzy, synthetic artifacts. In the following sections, we propose several improvements to the fundamental encoder blocks while making the MELP encoder robust against impulsive noise.

The first step of the MELP encoder analysis process consists in estimating the pitch period on each 22.5 ms frame, which is processed with a 1-kHz lowpass filter. For real-valued signal, $z_n(t)$, the pitch period is defined by a lag τ where the robust estimator of correlation is given by

$$\hat{\rho}_{zz}(\tau) = e^{jArg[\rho_{yy}(\tau)]} g^{-1}[|\rho_{yy}(\tau)|], \tag{7.2}$$

where $\rho_{yy}(\tau)$ is given by (5.9), reaches its maximum. In this paper, the pitch is calculated for $\tau = 40, 41, ..., 160$, where the lag values have been carefully chosen to cover male and female speaker fundamental frequency range. For speech signal sampled at 8000 ksps, the pitch period τ translates into fundamental frequencies between 50 and 200 Hz. Naturally, male and female speech signals tend to fall within and on the higher end of the spectrum, respectively.

Furthermore, robust voiced and unvoiced decisions are also determined based on the values taken by $\hat{\rho}_{zz}(\tau)$, since the voiced frames are highly periodic and have much higher correlation coefficients than the unvoiced frames. Robust estimation in pitch period and voiced/unvoiced frame classification is critical to preserving the fidelity of the speech signal. Impulsive noise in either voiced or unvoiced frames can cause the encoder to falsely identify the frames and incorrectly estimating the fundamental frequency. Our research shows that these miss classifications can severely degrade the quality of the synthesized speech.

The encoder maintains 360 samples in a circular buffer, that is, 180 samples of the current frame, 90 samples of the previous frame, and 90 samples of the next frame. The speech signal analysis window varies depending on the specific parameter to be estimated. The estimation of the pitch period, which corresponds to the fundamental frequency of the signal, is the first and critical step of the RMELP encoder algorithm. All subsequent parameter estimations are dependent upon the accuracy of the fundamental frequency estimate and the correct classification of the original speech signal into voiced, or unvoiced, or transition frames. Hence, the widest analysis window is used to estimate the pitch period. It is clear from (7.2) that for a given lag τ, the NACF is a function of 160 samples. However, as τ takes on values between 40 and 160, the center of the correlation window can also shift by as much as 160 samples. As a result, a total of 320 samples analysis window is required to calculate the pitch period, where 160 samples belong to the current frame and 160 samples to the next frame. Prior to the pitch period estimation, the input speech is processed with a high-pass filter to remove any low frequency hum below 60 Hz, and then is passed through a low-pass filter with cut-off frequency of 1 kHz. Furthermore, in order to increase the accuracy of the pitch period estimation, a fractional pitch period refinement is accomplished utilizing the robust estimator of correlation from (7.2). For more details on fractional pitch refinement, the reader is referred to [1].

n addition to voiced and unvoiced frames, the RMELP identifies a third type of frames, the transition frames, where the speech signal is neither highly periodic nor noise-like. For the transition frames, McCree and Barnwell [2] propose to remove the periodicity in the voiced excitation pulses by randomly varying pitch period by ±25%, which is proved to be a better fit in characterizing erratic glottal pulses in transition frames. In practice, the voiced frames are highly periodic and identified with high normalized correlation coefficients at the pitch while the transition frames are those with marginal correlation coefficients. Hence robust pitch estimation plays a significant role in transition frame identification.

7.2 Multiband Voicing Strength Estimation

As mentioned earlier, a key factor in determining the optimum mixture of periodic and noise pulses in mixed excitation process is pulse shaping filter. So it is important to understand the spetral characteristics of a given frame. For this purpose, speech signal in each frame is processed with five filter banks, with passbands of 0-500, 500-1000, 1000-2000, 2000-3000, and 3000-4000 Hz. We need to precondition the excitation signal, which is a combination of both pitch and noise. To separate the two types of signals, we calculate normalized autocorrelation estimates for each five bands, which are evaluated at the pitch period lag τ. Then, if a coefficient estimate exceeds a threshold on a particular band, the signal is flagged as voice-like, otherwise it is flagged as only noise. To make the test immune to impulsive noise, we propose to use the PPC for robustly estimating the autocorrelation function.

The fundamental frequency estimation is instrumental in accurately analyzing the band-pass voicing strengths of the upper four bands. The band-pass voicing strength analysis is a two-step process. First, the NACF, $\hat{\rho}_{zz}(\tau)$, is robustly estimated via (7.2) at the pitch lag. However, because the speech signal is non-stationary, it is possible that the pitch period may transition within a single frame. As a result, at higher frequencies this method can often provide poor correlation estimates. In order to circumvent this problem, we propose to use a second method initiated by McCree and Barnwell [2]. Here, the output of the band-pass filter is full-wave rectified to eliminate the high frequency transitions, which makes the generated signal to behave as if its envelopes rise and fall, in line with each pitch pulse.

Then, the rectifier output is smoothed with a one-pole low-pass filter followed by a notch filter to remove DC component. The NACF is once again robustly estimated with (7.2)from the full-wave rectified signal utilizing the PPC. Finally, the higher estimated value of the $\rho(\tau)$ of the above two methods is utilized to encode the pitch periodicity in each of the upper four bands.

The relative voice and noise power in each band characterizes the pulse shaping filter. The periodic and noise excitations are first filtered using the pulse shaping and the noise shaping filter, respectively. Here, the filters' outputs are added together to form the total excitation, known as the mixed excitation [24], since some portions of the noise and the pulse train are mixed together. Basically, mixed excitation along with the transition frame identification are the keys for improving the tonal or the buzzy quality sounds, hence making the speech signal to sound natural. In the presence of impulsive noise, it is clear that robust estimation of voicing strengths to generate the pulse shaping filter is critical to generating the optimum excitation signal.

7.3 AR Model Estimation

For the robust AR(10) model estimation, we use the Burgs' algorithm from (6.2). AR(10) model is robustly estimated on the input speech signal by applying the Burg's algorithm based on the PPC while using a 200-sample (i.e., a 25-ms signal segment) Hamming window centered on the last sample in the current frame. Bandwidth expansion is performed on the AR(10) model spectral response by multiplying each linear prediction coefficient by a factor of 0.994. Simulation results show that our proposed method of AR model estimation based on the Burg's algorithm is robust against impulsive noise; consequently, an a-priori removal of the outliers may not be necessary. Note that our robust PPC based AR(10) is guaranteed to be minimum phase and stable, where MMSE minimization method such as Levinson-Durban recursion is not. However, if the speech input is clean then variance of the robustly estimated AR(10) coefficients may be greater than one, negatively impacting synthesized speech quality.

The intended purpose of the spectral enhancement filter is to enhance the quality of the synthesized speech by closely matching the natural speech waveform in the formant regions [25]. According to [2], formants between pitch pulses do not decay as rapidly as they do at the all-pole filter output. If the poles are close to or greater than the unit circle, the LPC synthesis filter output may sound chirpy or even make the filter unstable. To address this problem, a bandwidth expanded pole-zero filter, estimated directly from the synthesis filter, has been proposed in [2]. Since spectral enhancement filter is directly related to AR(10) model, the need for robust estimation of the sysnthesis filter is further validated.

7.4 Fourier Magnitudes Estimation of The Error Signal

In theory, if the order of the AR model is sufficiently large, the error signal will be approximately white. However, for a low data rate speech encoder model, a higher order of the AR model is cost prohibitive; consequently, an AR(10) is proposed for the RMELP encoder. If an inverse filter from the AR(10) is applied, the error signal does indeed remain colored, that is, the error signal still has important information that is critical in estimating optimum excitation impulses. In order to capture that information, the low order Fourier magnitudes are encoded. On the synthesizer, the Fourier magnitudes are used to shape the mixed excitation sequence to closely estimate the encoder error signal.

In order to accurately estimate and reproduce the information in the error signal, McCree and Barnwell [2] propose to use a Fourier series expansion of the signal and encode the ten most dominant magnitudes around integer multiple of the fundamental pitch harmonics. This procedure makes sense since a digital impulse can be represented in the frequency domain by means of the Discrete Fourier Transform (DFT) if all the bins are represented with unit magnitude and zero phase. At the decoder one pitch period speech is synthesized at a time. So, instead of estimating one-pitch period impulse train to generate the excitation, the inverse Fourier transform of the Fourier magnitudes is taken at the pitch harmonics to obtain the time-domain excitation signal. By using sufficient points for the DFT, the magnitudes and phases of the excitation of the transmitted error signal can be systematically configured to reproduce near ideal error signal. The Inverse Discrete Fourier Transform (IDFT) output is then processed with pulse shaping and spectral enhancement filters. Finally, the output is excited with the AR(10) filter to synthesize the output speech signal.

Prior to computing the FFT, a Hamming window is applied to 200 samples of the transmitted error signal, centered at the last samples of the current frame. The output is then zero padded to compute a 512-point complex Fourier transform. A peak search algorithm is performed on the normalized magnitudes with an RMS value of one and the first ten peaks at integer multiples of the fundamental frequency are estimated and encoded with a vector quantizer. However, it is well known that the Fourier transform performs poorly in presence of outliers. To overcome this weakness, we propose a robust method to estimate the power spectral density and Fourier magnitudes of the error signal using the PPC. Specifically, we propose to estimate, two-sided, symmetric autocorrelation vector, directly from the error signal and then take its Fourier transform to obtain the PSD. Formally, it is given by $S_z(f) = \int_{-\infty}^{\infty} R_{zz}(\tau)e^{-2\pi i f\tau}d\tau$, where $R_{zz}(\tau)$ is robustly estimated using the PPC.

Chapter 8

Simulation Results

The performance of the robust correlation estimators is investigated using Monte-Carlo simulations using synthetic signals with different levels of contamination rate and correlation. Specifically, we generate their bias and variance curves for different contamination rates with a fixed correlation coefficient and study their robustness against outliers. Furthermore, the robust pre-filter performance is evaluated by comparing the variance of the error signal obtained before and after the pre-filtering process. Finally, the conventional MELP algorithm is compared with our proposed RMELP algorithm at a contamination rate of $\epsilon = 0.05$ using a a recorded real speech signal corrupted by impulsive noise.

Our research extends over a broad range of the speech signal processing aspects and over time it became apparent that there is no single accepted method to quantify improvements in different stages of the speech processing. For example, we first attempt to identify robust estimators of correlation for speech in terms of bias and variance, then propose a filter algorithm that improves impulse corrupted speech and finally robustly implement the RMELP to synthesize the improved version of the corrupted speech. At different stages of these process, the Signal to Noise Ratio (SNR) improvement is a good metric for a filter performance while the Perceptual Evaluation of Speech Quality (PESQ) or the Mean Opinion Score (MOS) metrics make more sense for speech quality enhancement evaluation.

Signal to noise ratio method is the most basic method out of the three aforementioned metrics. Strictly for research and bench-marking, if the known true signal is corrupted with a known noise model, then we can apply some arbitrary signal processing on the input speech such as our proposed robust filter and measure the gain in terms of $G = \frac{SNR(filtered)}{SNR(corrupted)}$. SNR measurements is simple for the ϵ contamination model if the noise is precisely known.

Perceptual Evaluation of Speech Quality (PESQ) is a weighted measure of the perceived distortion in reproduced speech compared to the reference signal. The PESQ measurement begins by level aligning both signals to a standard listening level. They are filtered (using an FFT) with an input filter to model a standard telephone handset. The signals are aligned in time and then processed through an auditory transform. The transformation also involves equalising for linear filtering in the system and for gain variation. Two distortion parameters are extracted from the disturbance (the difference between the transforms of the signals), and are aggregated in frequency and time and mapped to a prediction of subjective mean opinion score (MOS). A block diagram of the PESQ algorithm is shown in Figure

The Mean Opinion Score (MOS), is the widely accepted subjective evaluation where the

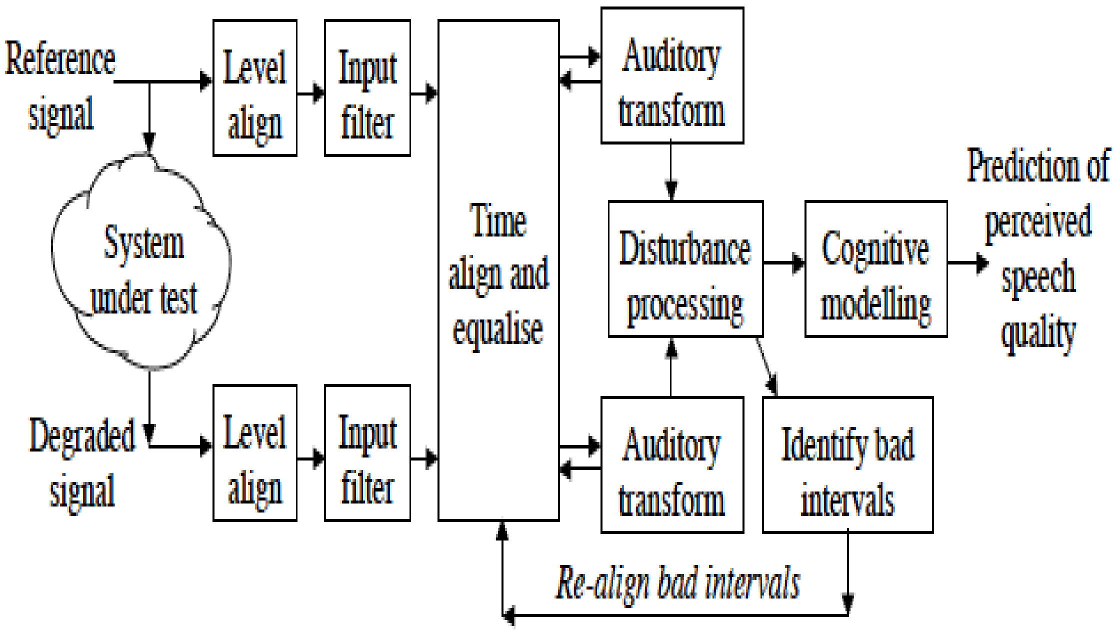

Figure 8.1: The PESQ Architecture

measure is an average of many subjective evaluations of speech signal under test. Proper set up of the MOS test is quite elaborate and address some very important details. For example, the test lab environment has to be properly calibrated so as not not create echos or other artificial effects. All of the sample speech have to be properly conditioned with the reference gain control so the listen tests are not biased. The test evaluation subject would have to be carefully chosen at random. Speech samples have to be chosen carefully to maintain balance between male vs female speakers. Furthermore, an encoder performance may not behave the same way, for example, German Vs. American English and so on. With our limited resources, we have attempted to conduct MOS evaluation with ten subjects chosen at random. Essentially, the speech quality published is ten opinions spread over four male and four female speech samples.

8.1 Performance Analysis of the Robust Correlation Estimators

In order to generate bias and variance curves, N = 2000 samples are drawn from a zero-mean, unit variance, complex-valued Gaussian distribution. White noise samples are then colored with lag-one or single-pole filter, where the filter parameter is the complex correlation coefficient. The ϵ contaminating replacement outliers are selected at random so they are uniformly distributed for the duration of N samples. Note that replacement outliers are also drawn from a zero-mean Gaussian distribution with $\sigma^2_{outlier} >> \sigma^2_{inlier}$. For bias curve, one thousand iterations of white noise samples are drawn and then colored with complex correlation coefficients given by $0.5e^{j\frac{\pi}{4}}$ and $0.9e^{j\frac{\pi}{4}}$. The lag-one correlation coefficient is estimated while varying contamination rate, ϵ, between zero and fifty percent. Finally, the sample mean of all the iteration estimates' bias is plotted against the contamination rate. We observe from Fig.8.2 that for correlation magnitude of 0.50, the GMLE bias is the largest. That estimator breaks down at very low contamination rate, as expected. The CPCC exhibits the best bias performance, followed by the ASL, the PPC and the MRE, in that order. For highly correlated data, the ASL performs the best, followed by the CPCC, the PPC, and the MRE, in that order. Therefore from the bias point of view, the ASL and the CPCC are the two estimators of choice. However, as we will now see, these two estimators may not be as efficient as the others. Let us examine the variances of these four estimators. The Gaussian white noise is correlated with AR(1) model, where lag-one correlation coefficient is varied

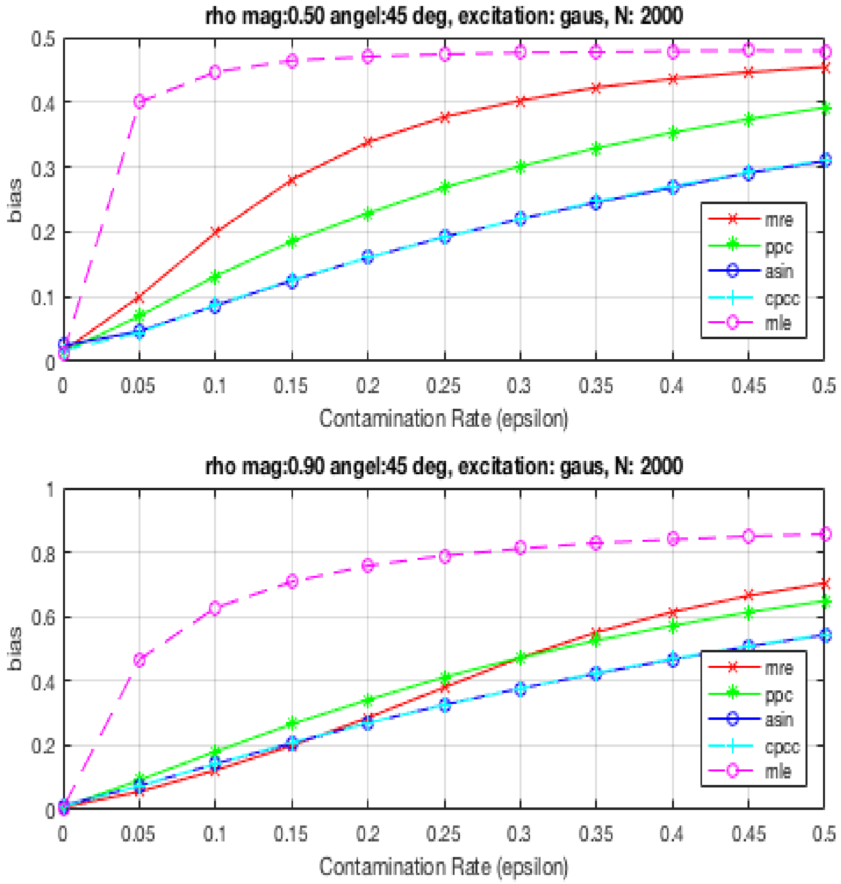

Figure 8.2: Bias curves of the ASL, the CPCC, the PPC, the MRE, and the GMLE versus the contamination rate ϵ at $\rho = 0.5$ (top) and $\rho = 0.9$ (bottom).

between zero and one. Colored signal is contaminated with a fixed ϵ. By means of the ASL, the CPCC, the PPC, the MRE, and the GMLE the correlation coefficient is estimated one thousand times and the variance of the estimates is calculated and plotted against an array of ρ values. As observed from Fig. 8.3, when there is no contamination, the GMLE clearly outperforms the others, while the PPC, the CPCC and the MRE exhibit slightly higher but approximately the same variances, and the ASL performs the worst. At a contamination rate of 10%, the GMLE starts to break down for $\rho > 0.7$ while the PPC, the CPCC, and the MRE

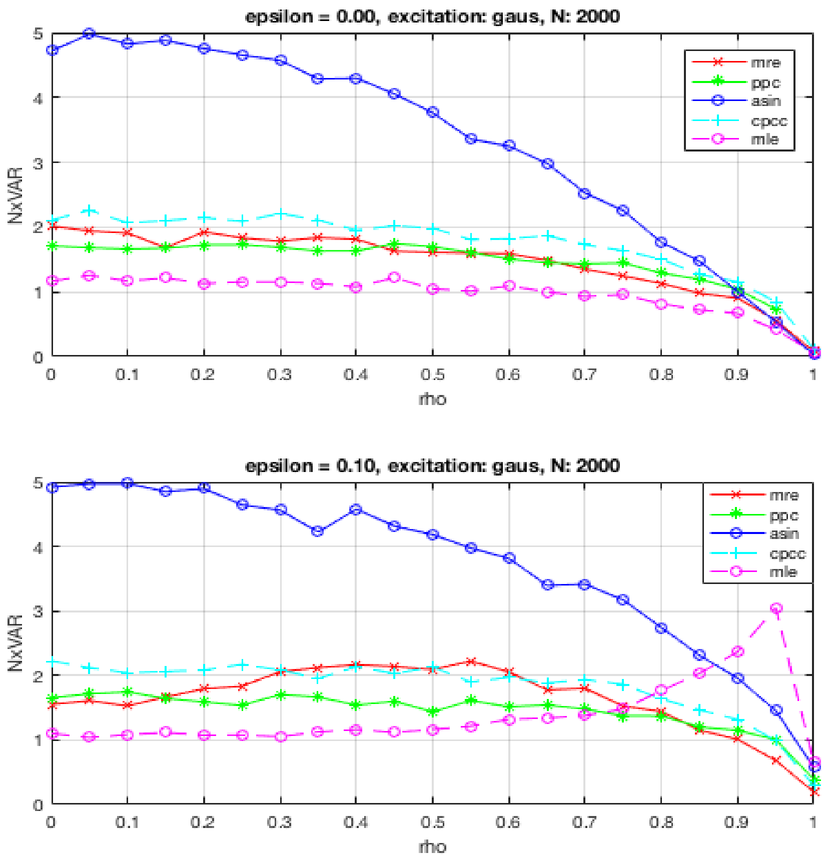

Figure 8.3: Variance curves at $\epsilon = 0.0$ (top) and $\epsilon = 0.10$ (bottom) of the PPC, the MRE, the CPCC, and the ASL.

do not break down and perform similarly. It is important to note that although variance of the ASL is quite high, it still outperforms the GMLE. At higher contamination rates of 20 and 35 percent, we see from Fig. 8.4 that the MRE begins to break down for $\rho > 0.6$. This observation will prove to be very important in selecting an estimator of correlation for signal such as speech. It is known that speech signal is highly correlated and in fact, frames that are classified as voice type, exhibit normalized correlation $\rho >> 0.6$. Considering bias and variance performance, we have chosen the PPC for our speech encoder algorithm. Note that while the PPC is less efficient than the GMLE; however, in the presence of impulsive noise, it is able to estimate stable speech model parameters. We have also evaluated all of the five estimators' performance for an excitation obeying a complex-valued Laplace distribution. Our simulation results show that the MRE breaks down at low contamination rate while the ASL, the CPCC, and the PPC perform similarly when complex-valued Laplacian excitation is used.

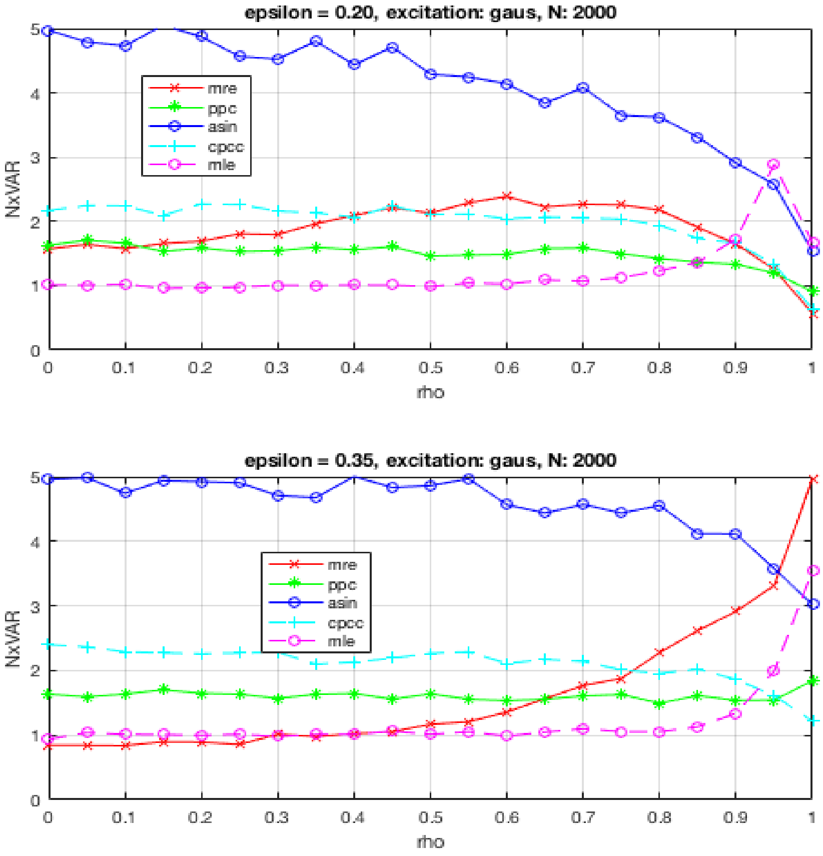

Figure 8.4: Variance curves versus the correlation ρ at $\epsilon = 0.20$ (top) and $\epsilon = 0.35$ (bottom) of the PPC, the CPCC, and the MRE.

Through analysis and simulations, we have shown robust correlation estimators statistical properties in terms of bias and variance performances. However, it may not be clear how speech perceptual quality is impacted with increased bias and variance levels. To this end we performed a very control set of experiments by artificially injecting different levels of bias or variance errors to the maximum likelihood, AR(10) estimates. We then synthesized the speech with the known, ideal excitation and error injected AR(10) estimates. Finally, the PESQ tool is used to determine how MOS scores are affected with different bias and variance thresholds. Note that a clean speech segment, sampled at $8ksps$, is fragmented into $22.5ms$ frames for this experiment. Table 8.1 shows PESQ test results of bias thresholds between 0 and 0.5 in 0.1 increments and variance thresholds between 1 and 2 in 0.2 increments. Note that when bias is varied, variance remained constant at one and likewise when variance is increased, bias is held at zero. The choice of the optimum estimator of correlation for speech is a careful balance between bias and variance evaluation. One may deduce from table 8.1 that speech quality is more sensitive to bias than variance of the estimates.

Table 8.1: PESQ of different bias or variance thresholds

Bias	Variance	PESQ	Bias	Variance	PESQ
0	1	4.5	0	1	4.5
0.1	1	2.2	0	1.2	2.445
0.2	1	1.456	0	1.4	2.373
0.3	1	1.555	0	1.6	1.788
0.4	1	1.190	0	1.8	1.728
0.5	1	1.064	0	2	1.567

8.2 Robust Filter Performance Analysis

For consistency with the RMELP encoder speech frame length of 22.5 ms or 180 samples, the analysis window for the impulsive outlier identification and replacement comprises 180 samples. For the outlier identification, we apply the RMD and for outlier replacement, we develop a new algorithm that utilizes the robust Burg's algorithm and the robust PSD, both derived from the PPC.

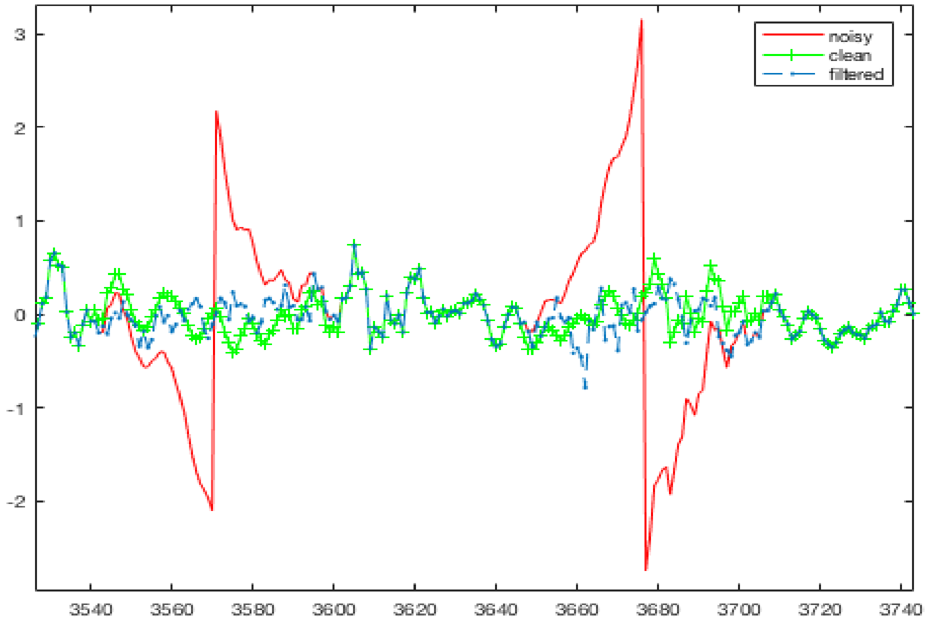

Figure 8.5: The actual clean speech signal (green) and the RMELP pre-filter input (red) and output signal (blue).

In preparation for calculating the RMD, we start by forming an (180x4) dimensional observation matrix **H**, each row vector of which is used to calculate an RMD. The outliers are then flagged as those with $RMD_i > \sqrt{\chi^2_{4,0.975}}$. Then, the outliers of each frame are replaced with robust signal estimates. Note that the outliers located at the edge of a frame are treated in the same way as those located inside because, by design, the RMELP analysis frame is centered at the last sample of the processed frame while maintaining a running buffer of 360 samples. However, with the RMD threshold criterion, it is often difficult to identify the beginning and the end of an impulse. Inaccurate identification of the impulse edges can cause sharp transitions in the filtered signal, which may degrade perceptual quality. This is a well-known and a difficult problem. To alleviate it, we propose to replace 10 samples preceding and following an impulse. It can be observed from Fig. 8.5, which displays a real speech frame corrupted by simulated impulses, the prefilter output signal exhibits tremendous improvements. It is important to note that the latter will never match the clean signal. More importantly, it will mimic the underlying characteristics, i.e. the spectral shape and the perceptual quality.

In order to evaluate the proposed filter performance in comparison to other filter methods discussed in section **??**, we resort to signal to noise ratio metric, where the true signal is known. Corrupted or filtered speech signal noise level can simply be calculated as $SNR = s^2/(s-x)^2$, where s is the true signal and x is the corrupted or filtered signal. True speech was corrupted with 2% impulse outliers to produce SNR of $-4.4dB$. We then filtered the corrupted speech with our proposed method, along with the binary mask filter, the inverse filter and the Wavelet filter. Our proposed filter outperforms the second best BMF fitler by more than $3dB$. Note that BMF would perform better if true signal is known and noise frequency bins are correctly identified. This is not possible in real applications; instead local threshold mechanism is used to identify the noisy frequency bins. In contrast, the inverse filter and and the wavelet filter perform worse because either they are unable to identify majority of the outliers or they falsely tag portion of the true signal as outliers.

Table 8.2: Filter SNR Improvement Comparison

Filter Method	Input SNR	Output SNR	SNR Gain
Proposed robust filter	-4.4	7	11.4
Binary mask filter	-4.4	3.5	7.9
Inverse filter	-4.4	4.3	8.7
Wavelet filter	-4.4	2	6.4

8.3 Speech Synthesis Performance Analysis

Let us now analyze the simulation results of the RMLEP algorithm for estimating the pitch period, the voice strength, the AR(10) model, and the error signal PSD estimation.

To assess the performance of our proposed PPC-based speech synthesis method, we do the following. First, a 200-sample, highly correlated, a clean voiced frame is processed with the GMLE to benchmark the desired AR(10) model. Then, we induce a replacement of the outliers using the clean signal with the ϵ contamination model, and then we re-estimate the AR(10) parameters. The outlier replacement is made at random on the time-axis with zero-mean and a variance of $\sigma^2_{outlier} = 10\ \sigma^2_{clean}$. As it can be seen from Fig. 8.6, the speech frame contaminated with a fraction of $\epsilon = 0.02$, the GMLE estimator of the AR(10) from the Levinson-Durbin recursion breaks down, while our proposed estimator follows closely the desired, clean model. Furthermore, our study shows that for highly correlated voiced frames, the robust PPC estimator does not break down until $\epsilon > 0.20$. In order to estimate

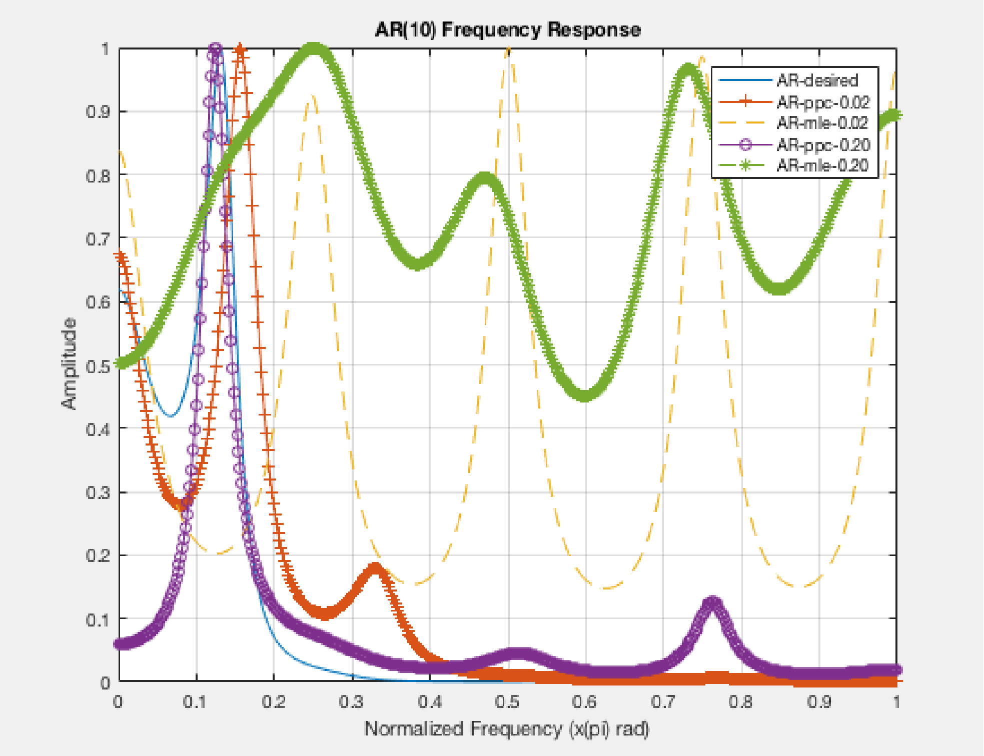

Figure 8.6: PSD amplitude curves calculated from the AR(10) model; the latter is estimated with the PPC at $\epsilon = 0.02$ (red) and $\epsilon = 0.20$ (violate) and with the GMLE at $\epsilon = 0.0$ (blue), $\epsilon = 0.02$ (yellow) and $\epsilon = 0.20$ (green). As observed, the PPC-based PSD curves closely follow the uncontaminated GMLE-based PSD curve.

and encode the error signal, the RMELP encoder algorithm performs a 512-point Fourier magnitude estimation from the output of the inverse filter, whose parameters are those of the AR(10) model, which are estimated using the Burg's algorithm. The goal is to encode the error signal information as correctly as possible so that the receiver may generate the optimum excitation with the aid of the pitch period, the band-pass voicing strengths, and the LPC model. Identical signal model is used to examine the PSD estimation algorithm based on the PPC. Note that we are attempting to capture information contained in the error signal and if the AR model is estimated correctly, the output of the inverse filter will indeed be a white noise. However, if the AR model is not of sufficient order, then there remains some information in the error signal. Clearly, contamination rate ϵ would have to be much lower for PSD to exhibit any discernible characteristic. Error signal PSD is estimated with $\epsilon = 0.05$, contaminated input frame. It can be seen from 8.7, PSD estimated from the robust PPC is far superior than the FFT method. The PPC performs much better in all frequency regions. Last but not least, Fig. 8.3 shows synthesized speech from our proposed encoder algorithm results in tremendous improvements over the state of the art.

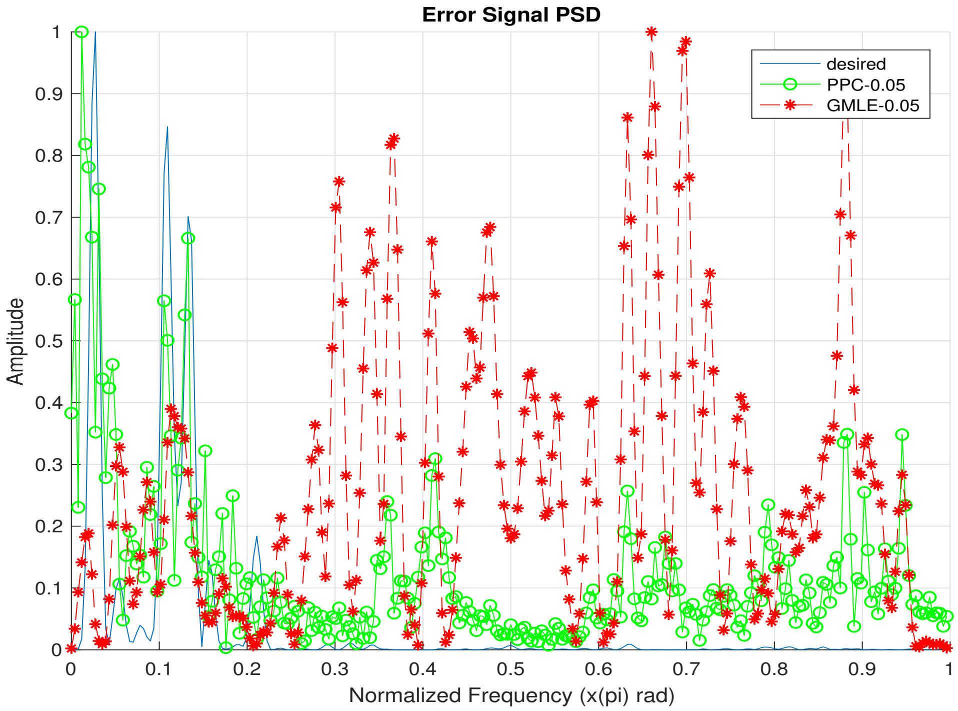

Figure 8.7: PSD amplitudes of the error signal estimated with the PPC at $\epsilon = 0.05$ (green) and the GMLE at $\epsilon = 0.0$ (blue) and $\epsilon = 0.05$ (red). As observed, the robust PSD curves remain close to the uncontaminated GMLE-based PSD curve while the GMLE-based PSD curve under contamination strongly deviate from it.

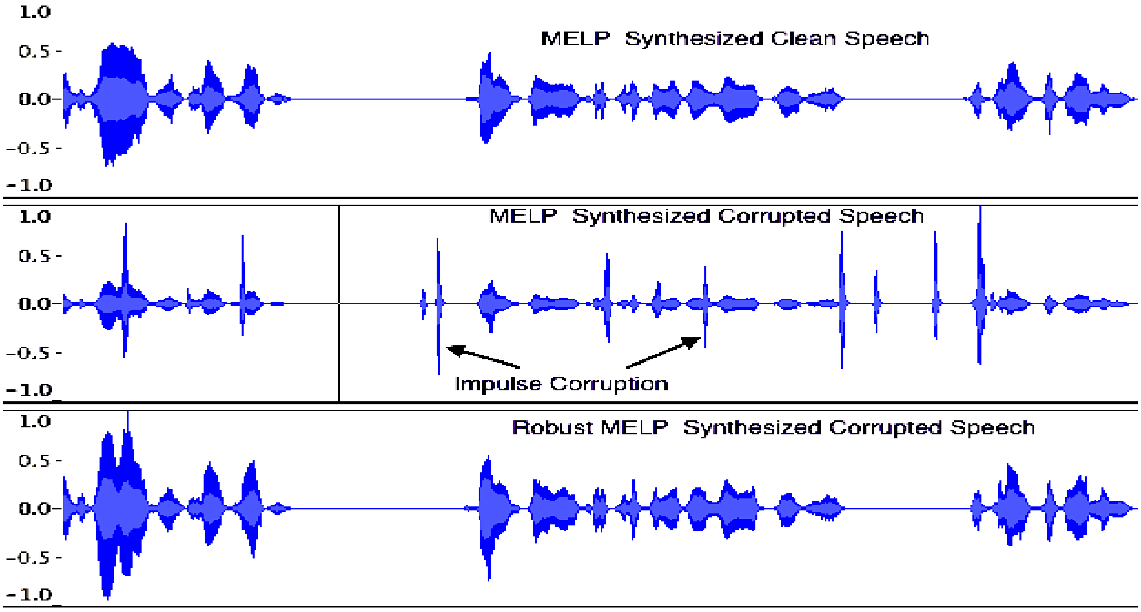

Figure 8.8: MELP Synthesis with the proposed robust encoder, in the presence of impulsive noise. Synthesized speech with MELP encoded method clearly shows (middle) spiky residuals, while our RMELP encoder produces speech (bottom) with impulses, completely removed.

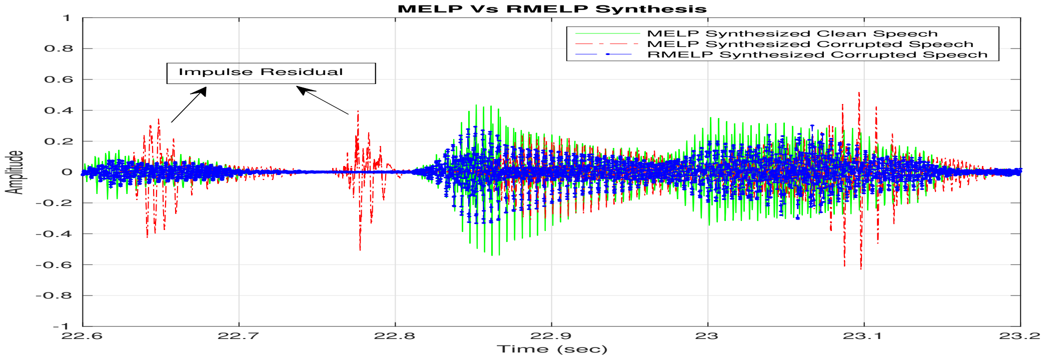

Figure 8.9: MELP Synthesis with the proposed robust encoder, in the presence of impulsive noise. Synthesized speech with MELP encoded method clearly shows (in red) spiky residuals, while our RMELP encoder produces speech (in blue) with impulses removed.

However, there may be some residuals of the impulses, buried within true signal, which may be hard to visualize. Nonetheless, a true evaluation of the speech perceptual quality can only be performed by subjective tests resulting in MOS scores. For MOS scores, we resort to Degradation Category [?] method, accepted and adopted by the International Telecommunications Union (ITU). Note that clean speech stimulus used for our simulations is a set of eight speech segments, four male and four female speakers, sampled at $8ksps$. Each speaker segment is an ensemble of four short sentences, approximately three seconds long. Ten subjects were chosen at random to evaluate clean and corrupted speech quality performances. As noted earlier, corrupted speech was generated with 1% outliers and it can be seen from table 8.3 that performance is dramatically affected for such small contamination rate. Given a robust estimator will produce sub-optimal encoder parameters when speech stimulus is uncorrupted, RMELP is applied only when outliers are detected in a given frame; otherwise GMLE is applied. Naturally, RMELP performs similar to MELP when speech is clean. However, in the presence of impulsive noise, RMELP MOS score is improved from 2.3 to 4.0 for female speapkers and for male speakers it improved from 2.6 to 4.5. Furthermore our research shows that clean, female, voiced frames are more likely to be falsely tagged with outliers than the clean, male, voiced frames. Which may explain why MELP performs better compared to RMELP.

Table 8.3: MOS performance of MELP Vs. RMELP

Input	RMELP	MELP
Female clean	4.3	4.2
Male clean	4.4	4.4
Female corrupted	4.01	2.3
Male corrupted	4.5	2.6

8.3.1 Execution Time Analysis

Analytically it is difficult to conduct a fair comparison of execution time between the GMLE and the PPC. The main reason is because the PPC utilizes a lookup table to estimate the correlation coefficients, which can vary significantly depending on the hardware architecture. For a 180 sample signal window, the GMLE of normalized auto-correlation at lag one can be computed with approximately 360 multipliers, 360 additions, one square root and one division. In contrast the PPC can be computed with 721 multiplications, 720 additions, 181 divisions and one lookup operation. Clearly, GMLE requires less computation resources than the PPC. In order to make sense of how execution time compares between the GMLE and the PPC, we estimated the auto-regressive model of order ten from the 180 sample window. Through Monte-Carlo simulations, we have found that processing time is increased by approximately a factor of 4.8 when using the robust PPC to estimate the AR(10) model relative to that of the GMLE.

Chapter 9

Future Work

Speech recognition technologies for applications in devices such as smart speakers, mobile phones, home appliance controllers have become integral to our daily lives. For future work as a continuation of robust speech processing research, we propose to develop a robust pattern recognition algorithm based on correlation analysis utilizing the Mel Frequency Cepstral Coefficient (MFCC) [ref]. Pattern recognition is the fundamental building block to speech recognition where acquired speech is anaylyzed by follwoing the human auditory system model. Analysis patterns must then be correlated against a database of a comprehensive dictionary of patterns and translate them into text or other formats. Futher processing of pattern recognition is required to arrange the words so they form meaningful sentences in the right context. As it may be obvious speech recognition is extemely expensive in terms of computation resources. State of the art algorithms have made enormous progress and yet often make false pattern recognition even in the ideal environment. This is well known in applications such text translation from voice when speaker pronounces one word and the application translates it to a similar sounding but completely different text word. In moderate noise environments such as inside of a moving car, present algorithms performance are poor and they perform extremely poor in the presence of impulsive noise. Unfortunately, a full speech recognition application is very complex and it is out of scope for our future work. Instead, in the presence of impulsive noise, we focus on a robust algorithm for feature extraction in individual spoken words and accurately correlate them against a known pattern dictionary.

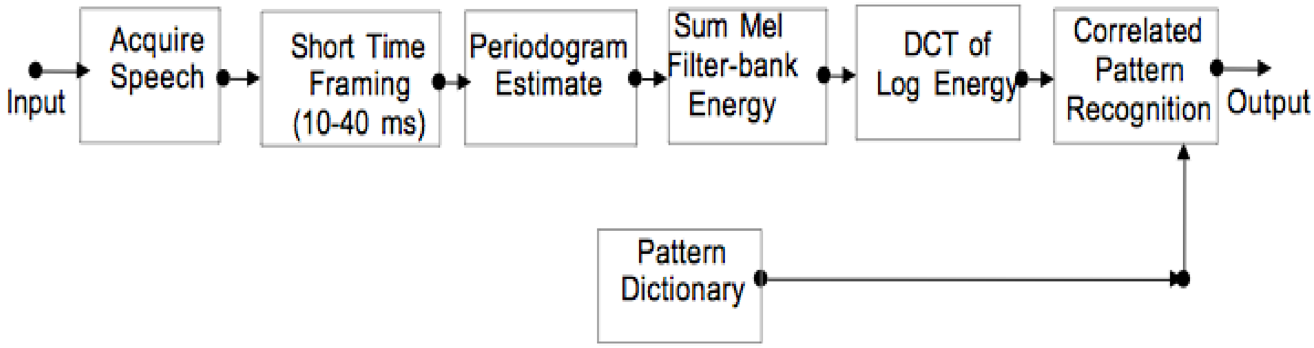
Input
Acquire Speech
Short Time Framing (10-40 ms)
Periodogram Estimate
Sum Mel Filter-bank Energy
DCT of Log Energy
Correlated Pattern Recognition
Output
Pattern Dictionary

Figure 9.1: Pattern recognition Algorithm

In computer science and electrical engineering, Automatic Speech Recognition (ASR) is the translation of spoken words into text. Some SR systems use "speaker-independent speech recognition while others use "training" where an individual speaker reads sections of text into the SR system. These systems analyse the person's specific voice and use it to fine-tune the recognition of that person's speech, resulting in more accurate transcription. Systems that do not use training are called "speaker-independent" systems. Systems that use training are called ”speaker-dependent” systems [ref]. Speech recognition is a very complex problem because vocalizations vary in terms of accent, pronunciation, articulation, roughness, nasality, pitch, volume, speed etc. Three main algorithms used in state of the art speech recognition systems are Hidden Markov Model(HMM), Dynamic Time Warping(DTW) and Artificial Neural Networks(ANN). When an HMM is applied to speech recognition, the states are interpreted as acoustic models, indicating what sounds are likely to be heard during their corresponding segments of speech; while the transitions provide temporal constraints, indicating how the states may follow each other in sequence. Because speech always goes forward in time, transitions in a speech application always go forward [ref].Dynamic time warping is a well-known technique to find an optimal alignment between two given (time-dependent) sequences under certain restrictions and the sequences are warped in a nonlinear fashion to match each other. The reasoning behind DTW is that the rate of speech may not be constant throughout the word; in other words, the optimal alignment between a template and the speech sample may be nonlinear. A neural network can be defined as a model of reasoning based on the human brain where learning can be supervised or unsupervised. For all three of the speech recognition algorithms, they require one common fundamental process, which is pattern recognition of short time speech frames.

A system model of our proposed speech pattern recognition algorithm is depicted in Fig. 9.1. Acquired speech is first segmented into short time frames to ensure each frame is stationary, Power Spectral Density (PSD) is then calculated on each frame followed by taking the log of Mel filter bank energies and Discrete Cosine Transform (DCT). For correlation based pattern recognition methodology only a set of the DCT coefficients, specifically thirteen, retained for comparison against a pattern database. In order to robustify this algorithm, we propose a two-step process. First we propose to replace PSD calculation in Fig. 9.1 with autocorrelation function derived from the PPC in section 3.3 and evaluate its performance. Next apply our robust pre-filter from chapter 4, in addition to PPC based robust PSD estimation and evaluate pattern recognition performance.

Machine learning is a trending area of research, with applications in communication systems, specifically in signal processing. However, most machine learning algorithms propose to find the optimum result by minimizing the mean square error method. It should be noted that the sample mean is not robust to outliers and hence a robust decoder algorithm is warranted. Furthermore, if the optimum noise model is known, in theory, it would be possible to train the encoder to learn similarities in the characteristics of the corrupted Vs. the clean training codes and make the correct decoding decision. For low data rate speech encoder application, we circle back to the same issue of how to define a finite code-book that captures all possible noise characteristics. If we were to assume a large enough, the optimum, code-book, it would still be challenging to minimize the right error at the receiver for real-time applications. Nonetheless, if computing power is of no concern, it would be of great value to pursue further research in low data rate speech encoding that might result in acceptable performances.

Chapter 10

Conclusions

In the presence of impulsive noise, all the methods proposed in the literature produce synthesized audio signals of very poor quality that are either harsh or non-intelligible. This paper describes the first robust method able to overcome this problem without sacrificing the encoder bandwidth efficiency, the compression gain, and the computation time while remaining compatible with real-time applications. Simulation results show that our method achieves much needed improvements in synthesized speech, both in terms of perceptual quality and error variance magnitudes. This has been achieved thanks to the development of a new robust version of the state-of-the-art MELP speech compression method utilizing a new robust outlier identification method based on the RMD and a novel algorithm for replacing the missing data by synthesizing a robustly estimated AR model. For the MELP encoder improvements, robust estimation methods based on the PCC have been developed for four sets of encoder parameters, namely the pitch period, the multiband voicing strengths, the

AR(10) model, and the Fourier magnitudes.

As a future work, we will initiate an outlier replacement method based on a robust pattern recognition algorithm. Furthermore, we will develop a fixed-point implementation of the proposed RMELP algorithm for real-time applications to improve its computation time. We will also evaluate its performance for speaker authentication and voice recognition in impulsive noise environments.

Bibliography

[1] W. C. Chu, "Foundation and evolution of standardized coders," in *Speech Coding Algorithms*, July 2003.

[2] A. V. McCree and T. P. Barnwell, "A mixed excitation lpc vocoder model for low bit rate speech coding," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 242–250, Jul 1995.

[3] A. McCree, K. Truong, E. B. George, T. P. Barnwell, and V. Viswanathan, "A 2.4 kbit/s melp coder candidate for the new u.s. federal standard," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1, May 1996, pp. 200–203 vol. 1.

[4] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 8, pp. 1223–1235, Aug 1988.

[5] E. Pryadi, K. Gandi, and H. Y. Kanalebe, "Speech compression using celp speech coding technique in gsm amr," in *2008 5th IFIP International Conference on Wireless and Optical Communications Networks (WOCN '08)*, May 2008, pp. 1–4.

[6] I. A. Gerson and M. A. Jasiuk, "Vector sum excited linear prediction (vselp) speech coding at 8 kbps," in *International Conference on Acoustics, Speech, and Signal Processing*, Apr 1990, pp. 461–464 vol.1.

[7] M. Ruhland, J. Bitzer, M. Brandt, and S. Goetze, "Reduction of gaussian, supergaussian, and impulsive noise by interpolation of the binary mask residual," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1680–1691, Oct 2015.

[8] N. Gallagher and G. Wise, "A theoretical analysis of the properties of median filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1136–1141, December 1981.

[9] D. Veselinovic and D. Graupe, "A wavelet transform-based blind adaptive filter of unknown noise from speech," in *Proceedings of the 43rd IEEE Midwest Symposium on Circuits and Systems (Cat.No.CH37144)*, vol. 3, 2000, pp. 1362–1365 vol.3.

[10] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to automatic speech recognition," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, Apr 1988, pp. 561–564 vol.1.

[11] V. H. Diaz-Ramirez and V. Kober, "Robust speech processing using local adaptive non-linear filtering," *IET Signal Processing*, vol. 7, no. 5, pp. 345–359, July 2013.

[12] Z. Wen and J. Tao, "An excitation model based on inverse filtering for speech analysis and synthesis," in *2011 IEEE International Workshop on Machine Learning for Signal Processing*, Sept 2011, pp. 1–5.

[13] G. Jacovitti and A. Neri, "Estimation of the autocorrelation function of complex gaussian stationary processes by amplitude clipped signals," *IEEE Transactions on Information Theory*, vol. 40, no. 1, pp. 239–245, Jan 1994.

[14] S. F. C. Neto, F. L. Corcoran, J. Phipps, and S. Dimolitsas, "Performance assessment of 4.8 kbit/s ambe coding under aeronautical environmental conditions," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1, May 1996, pp. 499–502 vol. 1.

[15] R. Maronna, R. Martin, and V. Yohai, *Robust Statistics: Theory and Methods.* Wiley, New York, NY, USA, 2006.

[16] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, May 2001, pp. 749–752 vol.2.

[17] J. H. V. Vleck and D. Middleton, "The spectrum of clipped noise," *Proceedings of the IEEE*, vol. 54, no. 1, pp. 2–19, Jan 1966.

[18] D. McGraw and J. Wagner, "Elliptically symmetric distributions," *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 110–120, Jan 1968.

[19] I. Reed, "On the use of Laguerre polynomials in treating the envelope and phase components of narrow-band Gaussian noise," *IRE Transactions on Information Theory*, vol. 5, no. 3, pp. 102–105, September 1959.

[20] Y. Chakhchoukh, P. Panciatici, and P. Bondon, "Robust estimation of sarima models: Application to short-term load forecasting," in *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, Aug 2009, pp. 77–80.

[21] P. Tamburello and L. Mili, "Robustness analysis of the phase-phase correlator to white impulsive noise with applications to autoregressive modeling," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 6053–6058, Nov 2012.

[22] M. A. Gandhi, C. Ledoux, and L. Mili, "Robust estimation methods for impulsive noise suppression in speech," in *Proceedings of the Fifth IEEE International Symposium on Signal Processing and Information Technology, 2005.*, Dec 2005, pp. 755–760.

[23] J. Burg, "Maximum entropy spectral analysis." in *Proceedings of 37th Meeting, Society of Exploration Geophysics, Oklahoma City.*, 1967.

[24] J. Makhoul, R. Viswanathan, R. Schwartz, and A. Huggins, "A mixed-source model for speech compression and synthesis," in *ICASSP '78. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, Apr 1978, pp. 163–166.

[25] J. Holmes, "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE Transactions on Audio and Electroacoustics*, vol. 21, no. 3, pp. 298–305, Jun 1973.