Time-Regularized Stabilised Weighted Linear Prediction of Speech

Irene Patsoura

Bachelor's Thesis
Department of Computer Science
University of Crete

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

Advisor: Dr George P. Kafentzis

December 5, 2024

Contents

List of Figures

Abstract

Linear prediction (LP) of speech is a mathematical technique used to estimate future samples of a speech signal based on past samples. It models the vocal tract as a linear filter, predicting the current sample as a weighted sum of previous samples. This approach is fundamental in speech analysis and coding, allowing efficient representation and compression of speech signals. By capturing the spectral characteristics of the speech, linear prediction aids in various applications such as speech synthesis, recognition, and enhancement.

LP feature extraction from speech signals is conventionally performed using short-time frames, assuming stationarity within each frame. For spectral envelope extraction, which captures the formant frequencies produced by the resonances of the slowly varying vocal tract, frame lengths of 20–30 ms are commonly used. However, traditional frame-based spectral analysis disregards the broader temporal context of the signal and is susceptible to degradation from environmental noise.

This thesis presents a recent frame-based LP analysis method that incorporates a regularization term. The method is named Time-Regularized Linear Prediction (TRLP). Its regularization term penalizes energy differences between consecutive frames in an all-pole spectral envelope model, thereby integrating the slowly varying nature of the vocal tract into the analysis.

Objective evaluations of feature distortion were conducted by analyzing the mel-frequency cepstral coefficient (MFCC) representations derived from various spectral estimation methods under noisy conditions, using the TIMIT database. The results indicate that the proposed time-regularized LP approach outperforms traditional methods, demonstrating superior MFCC distortion characteristics.

Chapter 1

Introduction

According to Fant's classical source-filter theory [5], speech production can be modeled as a sequence of three processes: glottal airflow excitation, vocal tract shaping, and lip radiation effects (Figure 1.1). Assuming the vocal tract consists of concatenated lossless tubes, its transfer function for non-nasalized sounds can be represented by an all-pole model. For low frequencies, the lip radiation effect is approximated as the time-derivative of the oral airflow.

The glottal excitation is often simplified as a quasi-periodic signal with a time-varying fundamental frequency and spectral tilt, which convey individual speaker characteristics and paralinguistic information. For unvoiced speech, the glottal airflow is modeled as white noise. Therefore, the vocal tract, particularly its resonances known as formants, is crucial for generating linguistic information. The vocal tract, comprising articulators such as the tongue, soft palate, and pharynx, controls its shape and consequently the formant frequencies. Despite the need for sophisticated and precise motor control to produce intelligible, continuously varying speech, the rate of articulatory configuration change is limited by inertial mass as per Newton's second law.

To accommodate the slow rate of articulatory changes (and thus formant contours), speech is typically assumed to be stationary within short-time frames of approximately 20–30 ms and processed with frame skips of about 5–10 ms. This approach is standard in speech feature extraction due to its simplicity, sufficient accuracy, and suitability for online processing. However, frame-by-frame feature extraction methods, such as linear prediction (LP) [6] or mel-frequency cepstral coefficient (MFCC) computation, disregard the context outside the frame. This makes these methods susceptible to issues like background noise or the initial phase of the signal. Consequently, context-aware speech feature extraction methods that consider a broader temporal context (e.g., 100 ms to several seconds) have been developed.

Prominent context-aware methods for computing autoregressive (AR) spectral envelope models of speech include time-varying linear prediction (TVLP) [7] and frequency-domain linear prediction (FDLP) [8]. In TVLP, the filter coefficient trajectories over a macro frame are fitted into a basis function of a predefined form, such as polynomial or trigonometric functions. In FDLP, bandpass filtered time-trajectories of speech are modeled with an AR model, and these trajectories are used to estimate frequency-domain parametric envelopes of speech at given time instants. Both TVLP and FDLP have been applied successfully for robust speech feature extraction, yielding improved results over traditional frame-based methods. However, their effectiveness is limited by the need for long macro frames, which introduce algorithmic delays. Additionally, the modeling of time-trajectories in TVLP and FDLP is driven by mathematical convenience rather than optimal time-trajectory representation.

This study introduces a novel regularization method for AR model computation, where the model optimization for a speech frame considers the filter coefficients of the previous frame. This approach, termed time-regularized linear prediction (TRLP) [4], effectively implements a leaky integration process for the temporal dynamics of the envelope spectrogram. TRLP generates smoothly evolving time-frequency contours similar to FDLP and TVLP, motivated by the slow movements of the articulators. Unlike FDLP and TVLP, TRLP does not introduce delays in envelope modeling because it does not require long macro frames. The results indicate that TRLP outperforms known all-pole modeling techniques in spectral modeling of noisy speech.

The following chapters of this thesis build upon the foundation laid in this introduction by delving into the specific methods and experiments conducted to advance noise-robust speech signal processing. Chapter 2 introduces the Linear Prediction Model of Speech, offering a comprehensive discussion of the theoretical principles and mathematical formulations that underpin the methods explored in this work. In Chapter 3,

we focus on the Stabilized Weighted Linear Prediction (SWLP) method, detailing its development and the key innovations that differentiate it from traditional approaches. Chapter 4 introduces the concept of Time-Regularized Linear Prediction (TRLP), highlighting its unique advantages in dynamic noise environments. Chapter 5 is dedicated to the experiments and results, where we compare the performance of these methods against traditional techniques like Fast Fourier Transform (FFT) and Linear Predictive Coding (LPC) across various noise conditions. Finally, Chapter 6 presents a detailed discussion and analysis of the findings, followed by future research directions and concluding remarks. Together, these chapters provide a thorough examination of advanced linear prediction methods in the context of noise-robust speech processing, contributing valuable insights to the field.

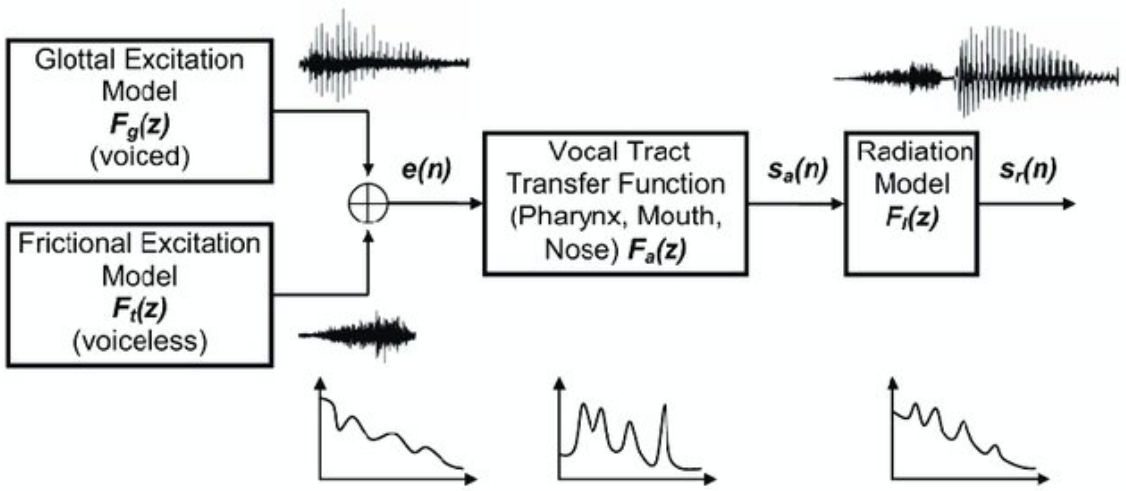

Figure 1.1: Fant's source-filter model of speech production: the excitation signal e(n) may be produced by phonation (voicing) or by turbulent excitation (voiceless) (Image obtained from [1]).

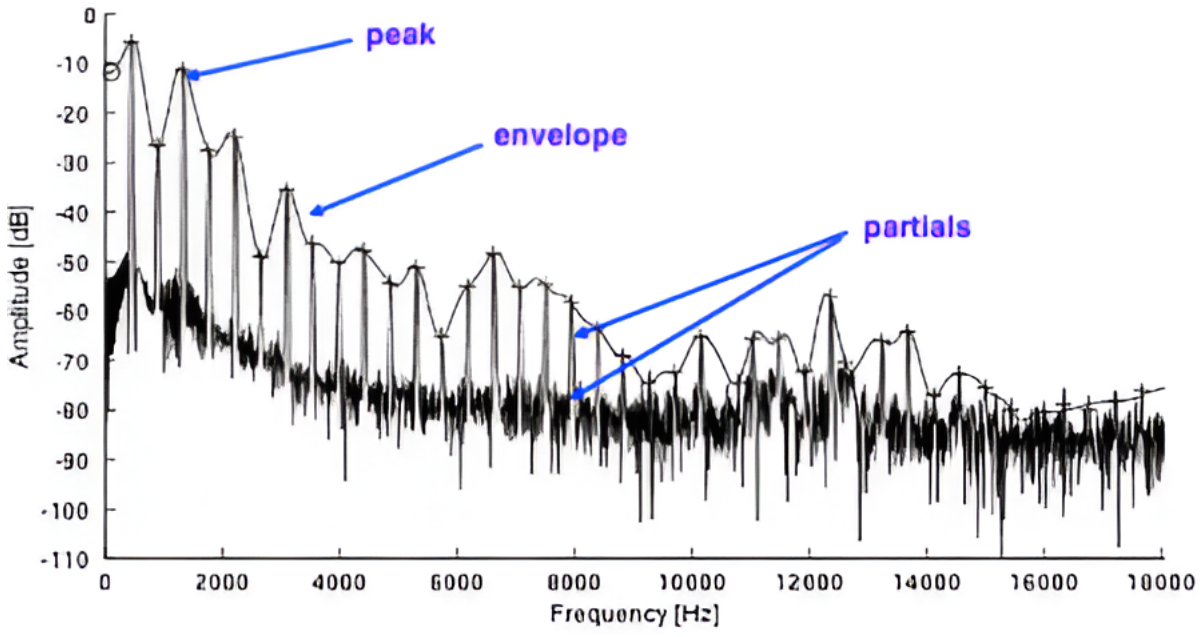

Figure 1.2: A spectral envelope is a curve in the frequency-amplitude plane, derived from a Fourier magnitude spectrum. It is a function of frequency that matches the amplitudes of the individual partials of the spectrum. It is the basic defining factor for its timbre. The envelope changes over time and can change with the fundamental frequency of the sound (Image obtained from [2]).

Chapter 2

The Linear Prediction Model of Speech

2.1 Introduction

Linear Prediction (LP) [9, 6] is a powerful tool used in speech processing to model the human vocal tract and estimate the spectral envelope of speech signals. By assuming that a speech signal can be approximated by a linear combination of its past samples, LP provides a framework for efficient speech analysis, synthesis, and compression. This chapter explores the theoretical foundations of LP, its practical implementation, and its applications in the field of speech processing.

2.2 Theoretical Foundations of Linear Prediction

Linear Prediction operates on the principle that a speech signal at any given time can be approximated as a linear combination of its past values. Mathematically, this is expressed as:

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + e(n) \tag{2.1}$$

where:

- $s(n)$ is the current speech sample.
- a_k are the linear prediction coefficients.
- p is the order of the prediction.
- $e(n)$ is the prediction error or residual.

The goal is to determine the coefficients a_k that minimize the prediction error $e(n)$ over a segment of speech. This is typically done by solving the normal equations derived from minimizing the mean squared error of the prediction.

2.3 Practical Implementation of Linear Prediction

The practical implementation of LP involves several key steps: framing, windowing, autocorrelation computation, and solving the normal equations to obtain the LP coefficients.

2.3.1 Framing and Windowing

Speech signals are non-stationary, so they are processed in short overlapping frames where the signal can be assumed to be approximately stationary. A typical frame length is 20-30 ms with a frame shift of 10 ms. Each frame is multiplied by a window function, such as the Hamming window, to reduce discontinuities at the frame boundaries.

2.3.2 Autocorrelation Method

The term $s(n)$ in the context of linear prediction represents the discrete-time speech signal at the sample index n. Speech, inherently a continuous-time signal, is sampled at regular intervals to produce a sequence of data points $s(n)$, where each point corresponds to the amplitude of the speech signal at a specific time. This discrete

representation allows for digital processing and analysis. In linear prediction, $s(n)$ is used to model the speech signal within each frame. The objective is to predict the current sample $s(n)$ as a linear combination of its previous p samples $s(n-1), s(n-2), \ldots, s(n-p)$. The coefficients of this linear combination are the linear prediction (LP) coefficients, which are determined by minimizing the prediction error over the frame.

The autocorrelation method [9] is a common technique to compute the LP coefficients. The method starts by computing the autocorrelation function for each frame of the speech signal. The autocorrelation function measures the similarity between the signal and a delayed version of itself as a function of the delay (or lag). Mathematically, for a frame of N samples, the autocorrelation function is defined as:

$$r(i) = \sum_{n=0}^{N-1-i} s(n)s(n+i), \tag{2.2}$$

where i is the lag, and $r(i)$ is the autocorrelation value at lag i. This function is used to populate the autocorrelation matrix $\mathbf{R}$ and the autocorrelation vector $\mathbf{r}$.

The elements of the autocorrelation matrix $\mathbf{R}$ are given by:

$$R_{ij} = r(|i-j|) = \sum_{n=0}^{N-1} s(n-i)s(n-j), \tag{2.3}$$

and the elements of the autocorrelation vector $\mathbf{r}$ are:

$$r_i = r(i) = \sum_{n=0}^{N-1} s(n)s(n-i). \tag{2.4}$$

The linear prediction coefficients $\mathbf{a} = [a_1, a_2, \ldots, a_p]^T$ are then obtained by solving the normal equations, which are expressed as:

$$\mathbf{Ra} = \mathbf{r}. \tag{2.5}$$

To solve these equations efficiently, the Levinson-Durbin recursion algorithm is employed. The Levinson-Durbin algorithm is an efficient recursive method for solving the Toeplitz system of equations that arise in the autocorrelation method of linear prediction. Instead of directly inverting the autocorrelation matrix $\mathbf{R}$, which can be computationally expensive, the Levinson-Durbin recursion computes the LP coefficients iteratively, reducing the computational complexity from $O(p^3)$ to $O(p^2)$.

The recursion begins by initializing the first-order prediction coefficient and the prediction error. At each iteration k, the algorithm computes the next order prediction coefficient a_k using the previous coefficients and the autocorrelation values. The process continues until all p prediction coefficients are determined. This efficient computation makes the Levinson-Durbin recursion a widely used method in speech processing applications.

In summary, the autocorrelation method combined with the Levinson-Durbin recursion provides a robust and computationally efficient approach to determining the linear prediction coefficients, which are essential for accurate speech signal modeling and subsequent processing tasks.

2.4 Applications of Linear Prediction in Speech Processing

Linear Prediction (LP) is a fundamental technique in speech processing that has found wide-ranging applications due to its ability to model the vocal tract and predict future speech samples based on past observations. This predictive power makes LP an invaluable tool in various domains, including speech coding, synthesis, recognition, enhancement, and even speaker verification.

In the realm of speech coding, LP is employed to achieve efficient compression of speech signals, which is essential for storage and transmission, particularly in low-bitrate environments. LP-based speech coding techniques, such as Linear Predictive Coding (LPC) [10], are used to represent the spectral envelope of the speech signal by encoding the LP coefficients—also known as predictor coefficients—along with the residual signal, which is the difference between the actual and predicted speech signals. This approach significantly reduces the data required to represent speech while maintaining intelligibility and quality. An advanced variant, Code-Excited Linear Prediction (CELP) [11], further enhances compression efficiency and is widely used in modern telecommunication systems and voice-over-IP (VoIP) services.

LP is also integral to speech synthesis, particularly through the use of LP-based vocoders. These vocoders decompose the speech signal into its spectral envelope, modeled by the LP coefficients, and the excitation signal, which can either be a periodic signal for voiced sounds or noise for unvoiced sounds. The synthesis process involves passing this excitation signal through a filter defined by the LP coefficients to produce speech that

closely resembles natural human speech. The LPC vocoder is one of the most notable examples, having been utilized in various applications from text-to-speech systems to secure communications [12].

In speech recognition, LP-derived features are crucial for enhancing system robustness, especially in noisy environments. The LP coefficients, or the more widely used LP-derived cepstral coefficients (LPCC) [6], provide a compact and informative representation of the speech signal's spectral characteristics. These features are instrumental in improving the accuracy of recognition systems by capturing the essential aspects of the speech signal that are most relevant for distinguishing between different phonemes, words, or speakers.

Speech enhancement, another critical application of LP, involves improving the quality of speech signals that have been degraded by noise or other distortions. LP techniques are used in methods such as spectral subtraction [13], where the clean speech spectrum is estimated using LP analysis, and then used to subtract noise from the signal, resulting in enhanced speech. Additionally, LP residual analysis is often employed to detect and reduce noise, further improving speech intelligibility [14].

LP's ability to capture unique vocal tract characteristics also makes it a powerful tool in speaker verification and identification. By analyzing the LP coefficients extracted from a speech sample, systems can compare these features against stored profiles to verify a speaker's identity or recognize them among a group of speakers [15]. This application is particularly significant in security systems where voice is used as a biometric identifier.

Beyond these conventional applications, LP is also used in tasks such as speech segmentation, where it helps detect boundaries between different speech units like phonemes or words [16], and in emotion recognition, where changes in the spectral envelope captured by LP can indicate different emotional states [17].

In summary, Linear Prediction has proven to be a versatile and powerful tool in the field of speech processing. Its applications span a broad spectrum, from compression and synthesis to recognition and enhancement, demonstrating its central role in advancing speech technology. The continued development of LP-based methods promises to further refine and extend these applications, leading to even more sophisticated and capable speech processing systems.

2.5 Conclusion

Linear Prediction provides a robust framework for analyzing and processing speech signals. Its ability to model the vocal tract and estimate the spectral envelope makes it invaluable in many speech processing applications. By understanding and implementing LP, we can achieve significant improvements in speech coding, synthesis, and recognition, even under challenging conditions such as noisy environments.

Chapter 3

Stabilised Weighted Linear Prediction

3.1 Introduction

Stabilized Weighted Linear Prediction (SWLP) [3] is an advanced technique used in speech processing to improve the robustness and stability of Linear Prediction (LP) analysis, particularly in noisy environments. By incorporating a weighting function and stabilization mechanism, SWLP aims to produce more accurate and reliable spectral estimates of speech signals.

3.2 Theoretical Foundations of Stabilized Weighted Linear Prediction

The core idea of SWLP is to modify the conventional LP framework by applying a weighting function to emphasize certain parts of the speech signal and introducing stabilization to ensure the reliability of the prediction.

3.2.1 Weighted Linear Prediction

In Weighted Linear Prediction (WLP), a weighting function $w(n)$ is applied to the speech signal $s(n)$ to give more importance to certain samples. The weighted speech signal $\tilde{s}(n)$ is given by:

$$\tilde{s}(n) = w(n)s(n) \tag{3.1}$$

The prediction error for WLP is then minimized over the weighted signal, leading to the weighted normal equations:

$$\mathbf{R}_w \mathbf{a} = \mathbf{r}_w \tag{3.2}$$

where $\mathbf{R}_w$ is the weighted autocorrelation matrix and $\mathbf{r}_w$ is the weighted autocorrelation vector:

$$R_{w,ij} = \sum_{n=0}^{N-1} w(n)s(n-i)s(n-j) \tag{3.3}$$

$$r_{w,i} = \sum_{n=0}^{N-1} w(n)s(n)s(n-i) \tag{3.4}$$

3.2.2 Stabilization Mechanism

The stabilization mechanism in Stabilized Weighted Linear Prediction (SWLP) is crucial to maintaining the stability of the prediction filter, especially when dealing with noisy or imperfect data. Stability in this context refers to ensuring that the roots of the prediction filter's polynomial lie within the unit circle in the z-plane, which corresponds to a stable system in the time domain.

To achieve stabilization, a regularization term is added to the weighted autocorrelation matrix $\mathbf{R}_w$. The regularized or stabilized normal equations are given by:

$$(\mathbf{R}_w + \gamma \mathbf{I})\mathbf{a} = \mathbf{r}_w \tag{3.5}$$

Here, $\mathbf{I}$ is the identity matrix of appropriate dimensions, and γ is the stabilization parameter, a positive scalar that controls the strength of the regularization.

The regularization term $\gamma\mathbf{I}$ effectively adds a bias to the autocorrelation matrix, ensuring that it is positive definite. This guarantees that the matrix inversion required to solve for the linear prediction coefficients $\mathbf{a}$ is well-conditioned, even in scenarios where $\mathbf{R}_w$ might be nearly singular or ill-conditioned due to noise or other factors.

Mathematically, adding $\gamma\mathbf{I}$ shifts the eigenvalues of $\mathbf{R}_w$ by γ. If any of the eigenvalues of $\mathbf{R}_w$ are close to zero (which would make the inversion unstable), the stabilization term ensures that all eigenvalues are bounded away from zero, thereby preventing numerical instability in the solution.

The inclusion of the stabilization term modifies the characteristic polynomial of the prediction filter, ensuring that all roots (poles of the filter) are within the unit circle. This is critical because if any roots lie outside the unit circle, the filter would become unstable, leading to a non-causal and exponentially growing impulse response, which is undesirable in speech processing applications.

To see this more concretely, consider the transfer function of the prediction filter derived from the LP coefficients $\mathbf{a} = [a_1, a_2, \ldots, a_p]^T$, where p is the order of the predictor. The filter's transfer function is given by:

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{3.6}$$

The poles of this filter are the roots of the polynomial:

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i} \tag{3.7}$$

Adding the stabilization term influences the coefficients $\mathbf{a}$ such that the roots of $A(z)$ are constrained to lie within the unit circle, ensuring that $H(z)$ represents a stable system.

Choosing the value of the stabilization parameter γ is a key practical consideration. A small γ might not provide sufficient regularization, especially in high-noise scenarios, potentially leading to unstable filters. Conversely, a large γ might overly smooth the solution, potentially degrading the predictive accuracy of the model by underfitting the signal.

In practice, γ is often selected empirically, and it can be tuned based on the specific application, noise conditions, and desired trade-off between stability and accuracy. Some adaptive methods may also adjust γ dynamically based on the signal characteristics, although this adds complexity to the algorithm.

The stabilization technique used in SWLP is conceptually similar to ridge regression, a method used in statistical regression models to handle multicollinearity and prevent overfitting. In ridge regression, a penalty term proportional to the square of the coefficients is added to the loss function to shrink the coefficients towards zero, thereby reducing variance at the cost of introducing some bias. In SWLP, the stabilization term plays a similar role by adding a bias to the autocorrelation matrix, ensuring a stable and robust prediction model.

3.3 Practical Implementation of Stabilized Weighted Linear Prediction

The implementation of SWLP involves the following steps: selecting an appropriate weighting function, computing the weighted autocorrelation matrix and vector, and solving the stabilized normal equations.

3.3.1 Selection of Weighting Function

The choice of the weighting function $w(n)$ in Weighted Linear Prediction (WLP) is critical, as it directly influences the emphasis placed on different parts of the speech signal during the prediction process. The weighting function determines which parts of the signal are given more importance when calculating the prediction error, and thus, it plays a key role in shaping the linear prediction model's behavior.

Several commonly used weighting functions are applied in WLP, each with its unique properties and applications:

- **Hamming window**: This window provides a smooth tapering at the edges of the frame, reducing the discontinuities at the frame boundaries. The Hamming window is defined as:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \tag{3.8}$$

  where N is the frame length. The smooth tapering minimizes spectral leakage, making the Hamming window a popular choice in speech processing.

- **Gaussian window**: The Gaussian window emphasizes the central part of the frame more than the edges, which can be particularly useful when the central portion of the frame is expected to contain more significant speech information. The Gaussian window is defined as:

$$w(n) = \exp\left(-\frac{1}{2}\left(\frac{n - \frac{N-1}{2}}{\sigma\frac{N-1}{2}}\right)^2\right) \tag{3.9}$$

  where σ is the standard deviation that controls the width of the Gaussian function. A smaller σ leads to a narrower peak, emphasizing the center even more.

- **Perceptually motivated windows**: These windows are designed based on auditory models to emphasize components of the speech signal that are most important for human perception. Examples include the Mel-scale or Bark-scale weighting, which correspond to the nonlinear frequency scales used in human hearing. These windows help to focus the LP analysis on the most perceptually relevant parts of the spectrum, enhancing the quality of applications like speech synthesis and coding.

In SWLP, the weighting function is not just any arbitrary window function but is often designed to reflect the energy distribution of the speech signal. This energy-based weighting function is crucial for ensuring that the linear prediction model emphasizes parts of the signal with higher energy, which are typically more critical for accurate modeling and synthesis.

The energy of the speech signal within a frame can be computed as:

$$E(n) = \sum_{m=n-k}^{n+k} s^2(m) \tag{3.10}$$

where $2k+1$ is the width of the energy window centered at sample n. The energy-based weighting function in SWLP is then defined as:

$$w(n) = \frac{E(n)}{\sum_{m=0}^{N-1} E(m)} \tag{3.11}$$

This function normalizes the energy across the frame, giving higher weights to regions with more energy. The rationale behind this approach is that segments of the speech signal with higher energy are generally more significant for accurate speech representation and synthesis. Therefore, by weighting these segments more heavily, the SWLP model can produce a more accurate and stable prediction.

The use of an energy-based weighting function in SWLP is particularly advantageous in noisy environments. Since noise typically has a lower energy compared to the speech signal, the energy-based weighting naturally de-emphasizes the noise components. This selective emphasis reduces the influence of noise on the prediction model, leading to more robust and stable linear prediction coefficients.

Furthermore, the energy-based weighting can adapt to the varying energy levels across different speech frames, allowing SWLP to dynamically adjust its focus. For instance, during voiced speech segments where the energy is high, the model will concentrate more on these segments, ensuring that the prediction closely follows the true speech signal.

3.3.2 Computation of Weighted Autocorrelation Matrix and Vector

Once the weighting function is selected, the weighted autocorrelation matrix $\mathbf{R}_w$ and vector $\mathbf{r}_w$ are computed using equations 3.3 and 3.4. These components are then used in the stabilized normal equations.

3.3.3 Solving the Stabilized Normal Equations

The stabilized normal equations in 3.5 are solved using techniques such as the Levinson-Durbin recursion or matrix inversion methods to obtain the LP coefficients $\mathbf{a}$. The stabilization parameter γ is typically chosen based on empirical observations or optimization techniques to balance stability and accuracy.

3.4 Applications of Stabilized Weighted Linear Prediction

Stabilized Weighted Linear Prediction (SWLP) is an advanced extension of traditional Linear Prediction (LP) methods, designed to address the challenges of noise and instability in speech processing applications. The robustness and stability of SWLP make it particularly valuable in scenarios where the integrity of the prediction filter is crucial for accurate signal analysis and synthesis. Below, we discuss several key applications where SWLP has proven to be especially beneficial.

In speech recognition systems, the accuracy of feature extraction is paramount, as it directly impacts the performance of the recognition algorithm. SWLP enhances the feature extraction process by providing more stable and noise-robust spectral estimates. The introduction of stabilization within the SWLP framework ensures that the prediction filter remains reliable even in the presence of background noise or other distortions. This leads to more accurate representations of the speech signal's spectral envelope, which are used to derive features such as Mel-Frequency Cepstral Coefficients (MFCCs). Consequently, the use of SWLP in speech recognition can significantly improve recognition performance, particularly in adverse acoustic conditions such as noisy environments or reverberant spaces [18, 19].

Speaker identification systems rely on the extraction of robust and discriminative features from speech signals to accurately identify speakers. SWLP contributes to this process by enhancing the stability and robustness of the extracted features, which is crucial for reliable speaker identification across different acoustic environments. The stabilization mechanism in SWLP helps to mitigate the effects of noise and other variabilities in the speech signal, leading to more consistent feature representations. As a result, speaker identification systems that utilize SWLP can maintain high accuracy even when operating in challenging conditions, such as when there is background noise, channel variations, or when the speech signal is captured with different microphones [20, 21].

Formant estimation is a fundamental task in speech analysis, as formants represent the resonant frequencies of the vocal tract, which are key to understanding speech sounds. Traditional LP methods often struggle with formant estimation in noisy conditions due to the instability of the prediction filter. SWLP addresses this issue by incorporating a stabilization mechanism that ensures the filter remains reliable even in the presence of noise. This leads to more accurate formant estimates, which are essential for tasks such as speech synthesis, linguistic analysis, and clinical speech pathology. Researchers have demonstrated that SWLP provides robust formant estimation, particularly in situations where traditional methods fail, such as with low signal-to-noise ratio (SNR) speech signals or when dealing with non-stationary noise sources [18, 20].

Glottal inverse filtering is a technique used to estimate the glottal waveform from the speech signal, which is critical for various applications, including speaker characterization, emotion detection, and voice quality assessment. The accuracy of glottal inverse filtering largely depends on the stability and precision of the linear prediction filter used to model the vocal tract. SWLP's stabilization mechanism makes it particularly suitable for this task, as it prevents the prediction filter from becoming unstable or overly sensitive to noise, leading to more accurate and reliable glottal waveform estimates. Studies have shown that SWLP-based approaches to glottal inverse filtering outperform traditional LP methods, particularly in noisy environments or when dealing with complex vocal tract configurations [20, 22].

In speech coding, the goal is to compress speech signals for transmission or storage while maintaining high fidelity. Noise robustness is a critical requirement for effective speech coding, especially in applications such as mobile communications or VoIP, where the signal is often corrupted by various types of noise. SWLP enhances the robustness of speech coding by ensuring that the prediction filter remains stable even in the presence of noise, leading to more accurate spectral estimates and thus higher-quality compressed speech. The weighted and stabilized prediction process helps to preserve the essential characteristics of the speech signal, reducing artifacts and improving the intelligibility and naturalness of the decoded speech signal [20, 23].

3.5 Conclusion

Stabilized Weighted Linear Prediction offers significant improvements over conventional LP by incorporating weighting functions and stabilization mechanisms. These enhancements make SWLP a valuable tool in various speech processing applications, particularly in noisy and challenging environments. By adopting SWLP, we can achieve more accurate and reliable spectral estimates, leading to better performance in speech recognition, enhancement, and identification tasks (as illustrated in Figures 3.1 3.2 3.3)

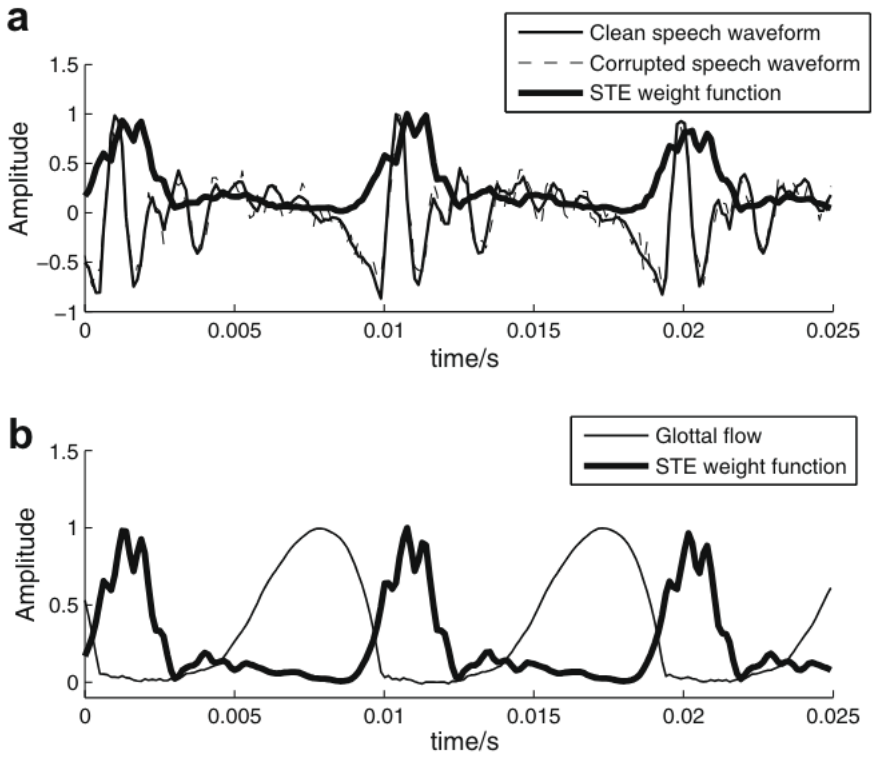

Figure 3.1: Upper panel: time-domain waveforms of clean speech (vowel /a/ produced by a male speaker), additive zero-mean Gaussian white noise corrupted speech (SNR = 10 dB), and short-time energy (STE) weight function (M = 8) computed from noisy speech. Lower panel: glottal flow estimated from the clean vowel /a/ together with STE weight function (M = 8) computed also from the clean speech signal (Image obtained from [3]).

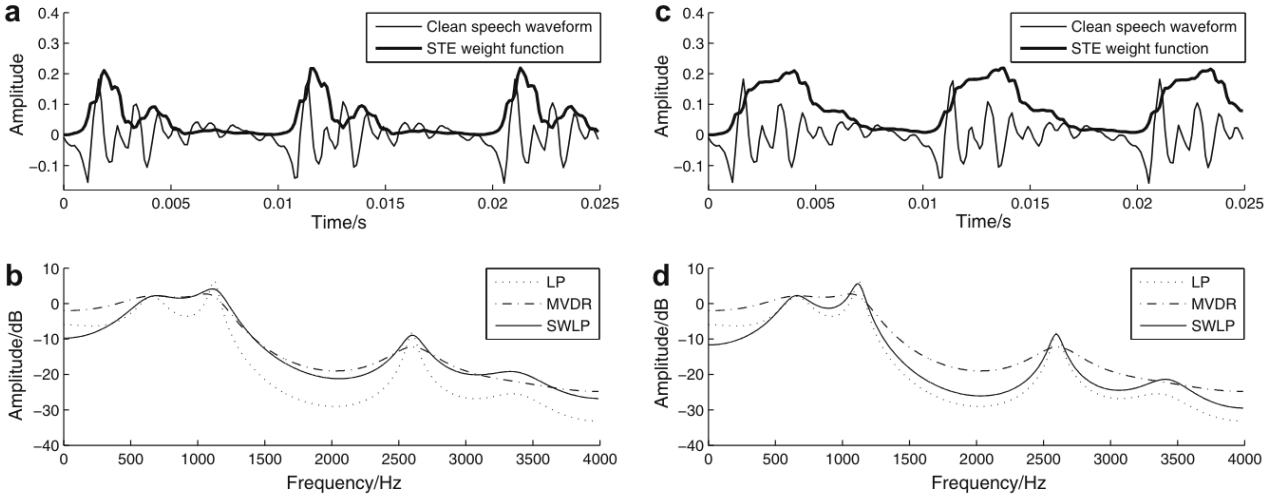

Figure 3.2: Time-domain waveforms of clean speech (vowel /a/ produced by a male speaker) and short-time energy (STE) weight function (upper panels) and corresponding all-pole spectra of order p = 10 computed by LP, MVDR, and SWLP (lower panels). SWLP analysis was computed by using two different values for the length of the STE window: M = 8 (left panels) and M = 24 (right panels) (Image obtained from [3]).

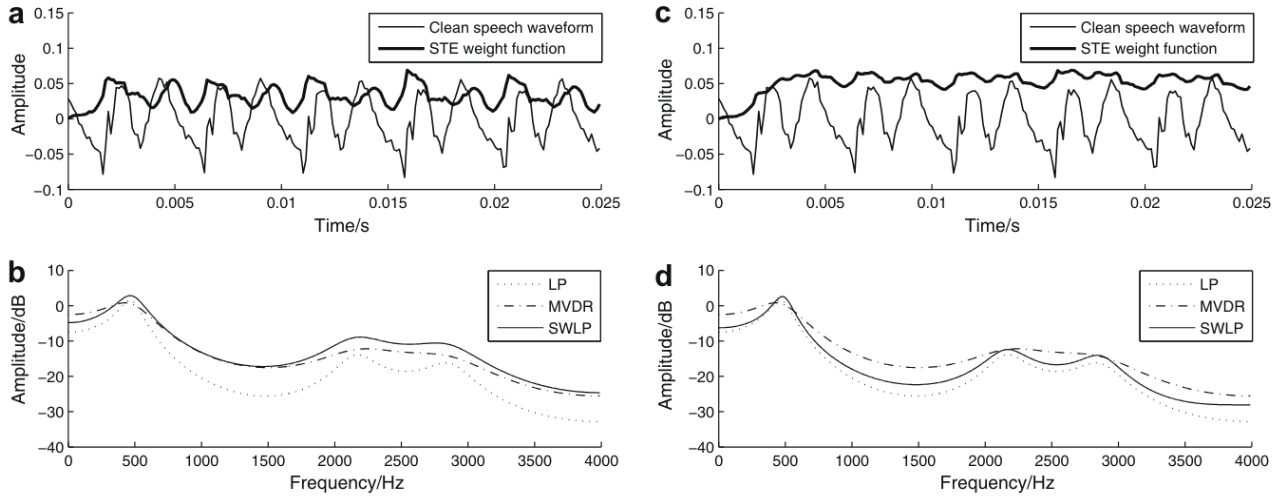

Figure 3.3: Time-domain waveforms of clean speech (vowel /e/ produced by a female speaker) and short-time energy (STE) weight function (upper panels) and corresponding all-pole spectra of order p = 10 computed by LP, MVDR, and SWLP (lower panels). SWLP analysis was computed by using two different values for the length of the STE window: M = 8 (left panels) and M = 24 (right panels) (Image obtained from [3]).

Chapter 4

Time-Regularized Linear Prediction

4.1 Introduction

Time-Regularized Linear Prediction (TRLP) is an advanced technique designed to enhance the robustness of Linear Prediction (LP) analysis in dynamic and noisy environments. TRLP integrates temporal smoothness into the LP framework by introducing a regularization term that penalizes rapid changes in the LP coefficients over time. This chapter explores the theoretical concepts, practical implementation, and applications of TRLP in speech processing.

4.2 Theoretical Foundations of Time-Regularized Linear Prediction

Time-Regularized Linear Prediction (TRLP) is a sophisticated enhancement of the traditional Linear Prediction (LP) model that aims to improve the temporal consistency of LP coefficients across consecutive frames. Traditional LP models are primarily focused on minimizing the prediction error within each frame independently, which can result in abrupt changes in the LP coefficients from one frame to the next. These abrupt changes can be detrimental in applications where smooth transitions between frames are critical, such as in speech coding, recognition, and synthesis. TRLP addresses this issue by introducing a regularization term that enforces temporal continuity, leading to more stable and consistent LP coefficients.

4.2.1 Formulation of TRLP

The key innovation in TRLP is the introduction of a regularization term to the conventional LP cost function. The traditional LP approach minimizes the prediction error over each frame independently. However, TRLP modifies this approach by adding a penalty for large variations in the LP coefficients between consecutive frames. This regularized cost function is given by:

$$J = \sum_{n=0}^{N-1} e(n)^2 + \lambda \sum_{k=1}^{p} (a_k(n) - a_k(n-1))^2 \tag{4.1}$$

where:

- $e(n)$ is the prediction error, defined as the difference between the actual and predicted signal values.
- $a_k(n)$ are the LP coefficients for the k^{th} predictor at frame n.
- λ is the regularization parameter that controls the trade-off between minimizing the prediction error and enforcing temporal smoothness.
- p is the order of the linear prediction, representing the number of previous samples used to predict the current sample.

The cost function J consists of two terms:

1. The first term, $\sum_{n=0}^{N-1} e(n)^2$, represents the traditional LP objective of minimizing the prediction error for the current frame.
2. The second term, $\lambda \sum_{k=1}^{p} (a_k(n) - a_k(n-1))^2$, is the regularization term that penalizes large differences in the LP coefficients between consecutive frames, promoting temporal smoothness.

This formulation ensures that the LP coefficients not only fit the current frame accurately but also vary smoothly from one frame to the next, which is particularly beneficial for modeling speech signals that exhibit gradual changes over time.

4.2.2 Regularized Normal Equations

The introduction of the regularization term significantly alters the normal equations that are typically used to solve for the LP coefficients. In traditional LP, the normal equations are derived from the autocorrelation method, leading to a set of linear equations that relate the LP coefficients to the autocorrelation values of the signal. In TRLP, these equations are modified to incorporate the regularization term, resulting in the following regularized normal equations:

$$(\mathbf{R} + \lambda \mathbf{I})\mathbf{a}_n = \mathbf{r} + \lambda \mathbf{a}_{n-1} \tag{4.2}$$

where:

- $\mathbf{R}$ is the autocorrelation matrix, formed from the autocorrelation values of the signal.
- $\mathbf{a}_n$ are the LP coefficients at the current frame n.
- $\mathbf{r}$ is the autocorrelation vector, representing the correlation between the current frame's samples and the previous samples.
- $\mathbf{I}$ is the identity matrix, introduced as part of the regularization process to ensure numerical stability.
- $\mathbf{a}_{n-1}$ are the LP coefficients from the previous frame $n-1$.

The term $\lambda\mathbf{I}$ adds a regularization factor to the autocorrelation matrix, effectively controlling the smoothness of the LP coefficients. The term $\lambda\mathbf{a}_{n-1}$ in the right-hand side introduces a dependency on the coefficients from the previous frame, enforcing temporal continuity.

4.2.3 Advantages and Optimality of TRLP

TRLP offers several advantages over traditional LP and other regularized LP approaches, making it particularly well-suited for applications where temporal smoothness and stability are critical.

The primary advantage of TRLP lies in its ability to enforce temporal smoothness in the LP coefficients. This is essential in applications such as speech synthesis and coding, where abrupt changes in the spectral envelope can lead to perceptually unpleasant artifacts, such as "buzziness" or unnatural transitions. By ensuring that the LP coefficients vary smoothly over time, TRLP produces a more natural and stable representation of the speech signal, which is crucial for maintaining the quality and intelligibility of synthesized or coded speech.

The regularization term in TRLP also contributes to improved robustness in noisy environments. In traditional LP, the coefficients can become highly sensitive to noise, leading to large variations between frames that degrade the performance of speech recognition or coding systems. The regularization introduced in TRLP mitigates this sensitivity by smoothing the coefficient trajectories, thus reducing the impact of noise on the prediction model.

The regularization parameter λ provides flexibility in tuning the model to specific applications or conditions. A higher λ value enforces stronger temporal continuity, which can be beneficial in highly dynamic environments where the speech signal changes rapidly. Conversely, a lower λ value allows the model to adapt more quickly to changes in the signal, making it suitable for applications where rapid adaptation is necessary. This adaptability makes TRLP a versatile tool in various speech processing tasks.

TRLP can be considered optimal in contexts where both prediction accuracy and temporal consistency are equally important. The model strikes a balance between fitting the current frame data and maintaining smooth transitions across frames, leading to a more coherent representation of the speech signal. This balance is particularly important in applications like real-time speech synthesis and coding, where both accuracy and temporal stability are crucial for achieving high-quality results.

4.3 Practical Implementation of Time-Regularized Linear Prediction

The implementation of TRLP involves framing, windowing, autocorrelation computation, and solving the regularized normal equations. The steps are similar to conventional LP but with additional regularization considerations.

4.3.1 Framing and Windowing

Speech signals are divided into short overlapping frames to assume quasi-stationarity. Each frame is multiplied by a window function, such as the Hamming window, to reduce spectral leakage.

4.3.2 Autocorrelation Computation

The autocorrelation matrix $\mathbf{R}$ and vector $\mathbf{r}$ are computed for each frame using the windowed speech samples. These form the basis for solving the regularized normal equations.

4.3.3 Solving the Regularized Normal Equations

The regularized normal equations (2) are solved iteratively for each frame. The initial frame can use conventional LP methods to obtain $\mathbf{a}_0$. For subsequent frames, the regularized equations ensure smooth transitions in the LP coefficients. The parameter λ is chosen based on empirical observations or optimization techniques to balance prediction accuracy and temporal smoothness.

4.4 Applications of Time-Regularized Linear Prediction

Time-Regularized Linear Prediction (TRLP) has proven to be a powerful tool in various speech processing applications, particularly where the temporal consistency of spectral estimates is crucial. By incorporating a regularization term that smooths the variation of linear prediction (LP) coefficients over time, TRLP offers several advantages in terms of robustness, stability, and performance in challenging environments. This section explores key applications of TRLP, including speech recognition, speaker verification and identification, and other areas of speech processing.

TRLP is particularly beneficial in applications that require robust and temporally smooth spectral estimates. In traditional LP-based systems, abrupt changes in the spectral characteristics from frame to frame can lead to reduced performance, especially in noisy or variable conditions. TRLP addresses this issue by enforcing temporal regularization, which ensures that the spectral features vary smoothly over time. This leads to several important benefits in various speech processing domains.

In speech recognition systems, the accuracy of recognizing spoken words or phrases relies heavily on the consistency and reliability of the extracted features. LP-based features, such as the spectral envelope, are commonly used in automatic speech recognition (ASR) systems. However, these features can be sensitive to noise and may exhibit variability across frames, which can degrade recognition performance. TRLP mitigates these issues by promoting temporal smoothness in the spectral features, resulting in more stable and noise-resistant representations. The regularization introduced by TRLP ensures that the spectral characteristics of speech remain consistent over time, even in the presence of background noise or other acoustic distortions. This leads to improved recognition accuracy, particularly in noisy environments, as the system can more reliably match the stable features to the correct speech models [24, 25].

Another critical application of TRLP is in speaker verification and identification systems. These systems rely on extracting and modeling unique features from a speaker's voice to either verify their identity or distinguish them from other speakers. The temporal regularization in TRLP ensures that the extracted features remain consistent over time, even when the recording conditions vary or when noise is present. This consistency is crucial for the reliability of speaker verification systems, as it reduces the likelihood of errors caused by abrupt changes in the spectral features. In speaker identification, where the system must differentiate between multiple speakers, TRLP enhances the distinctiveness of the features by maintaining their stability across different frames, leading to more accurate and reliable identification results [26, 27].

Beyond speech recognition and speaker identification, TRLP is also advantageous in other speech processing applications that require robustness to noise and temporal stability. For instance, in speech enhancement and noise reduction systems, TRLP can be used to improve the quality of the enhanced speech by providing more consistent spectral estimates, which are less affected by transient noise or other distortions. This makes TRLP an effective approach in scenarios where the speech signal must be processed in real-time and under varying acoustic conditions [28, 29]. Additionally, TRLP can be applied in speech synthesis, where the smoothness of spectral features is crucial for generating natural-sounding speech. By ensuring that the spectral envelope evolves smoothly over time, TRLP contributes to more natural and intelligible synthetic speech, particularly in dynamic environments where the characteristics of the speech signal may change rapidly [30].

4.5 Conclusion

Time-Regularized Linear Prediction introduces a temporal smoothness constraint to the traditional LP framework, enhancing its robustness and stability in dynamic and noisy environments (as illustrated in Figure 4.1). By incorporating regularization, TRLP provides more reliable and accurate spectral estimates, making it valuable in various speech processing applications such as noise-robust speech recognition, speech enhancement, and speaker verification(as illustrated in Figure 4.2). The adoption of TRLP can lead to significant improvements in performance and reliability, particularly in challenging acoustic conditions.

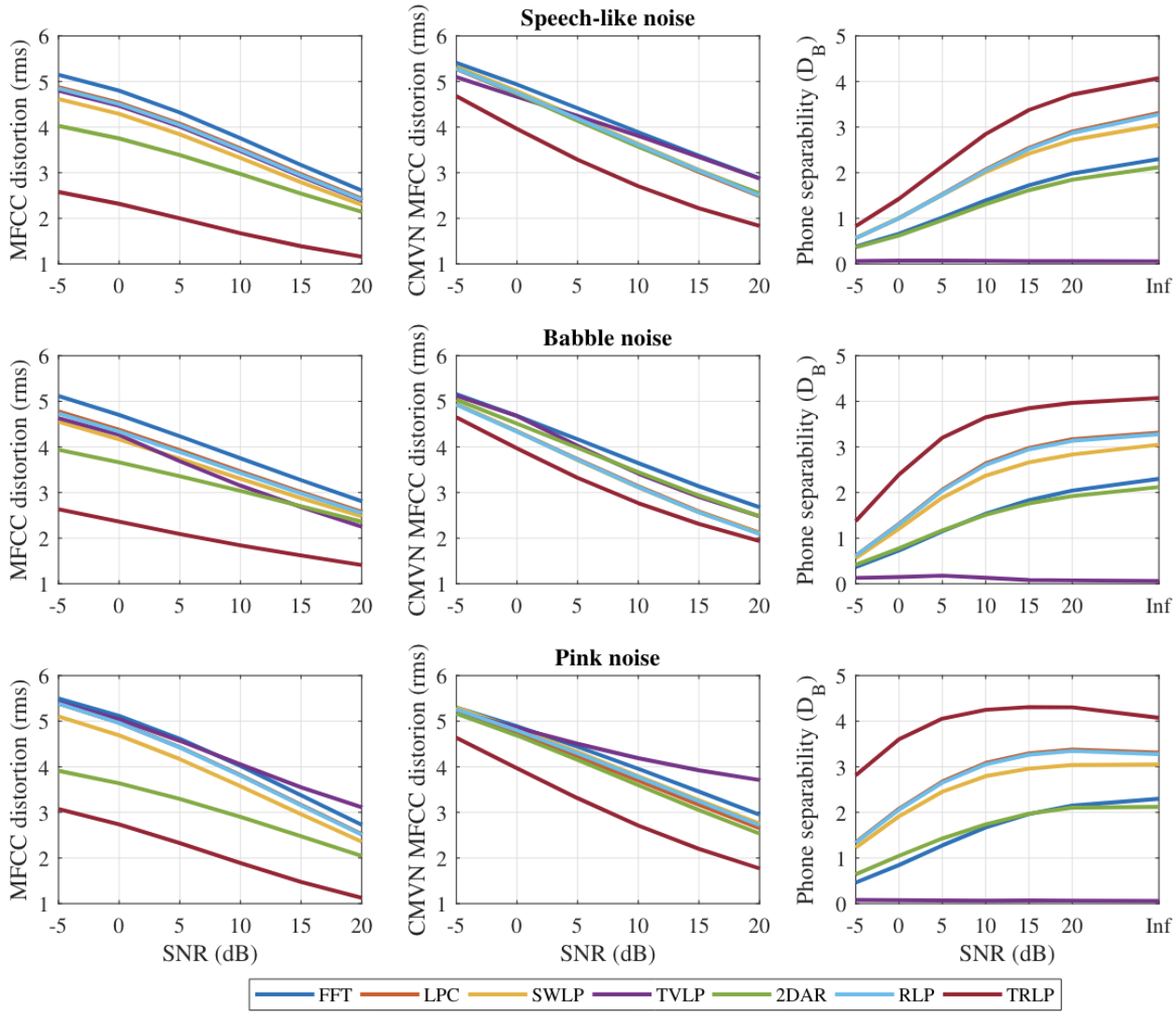

Figure 4.1: Results of the MFCC distortion test. Rows correspond to different noise types. MFCC distortion is reported for the direct form (left column) and with CMVN (middle column). The Bhattacharyya distances D_B for phoneme category separability are shown in the right column (Image obtained from [4]).

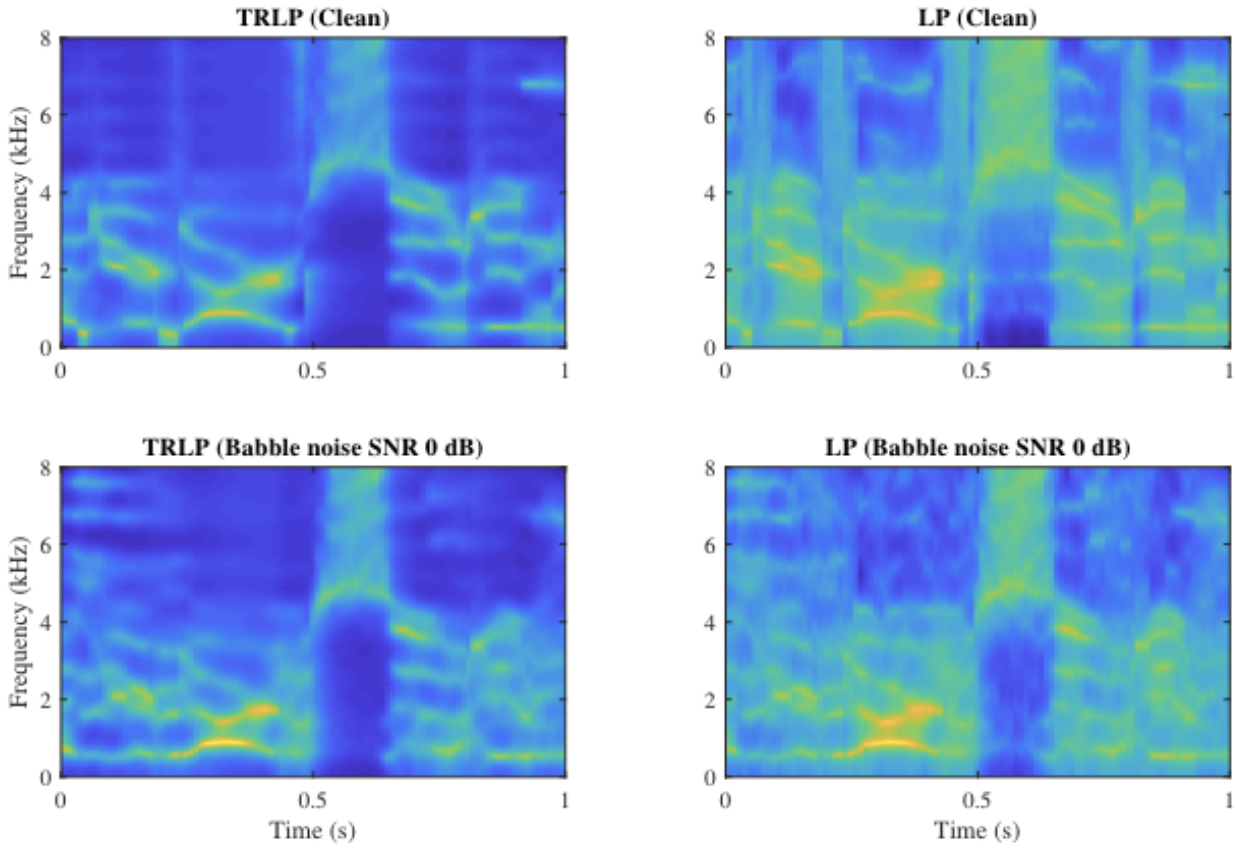

Figure 4.2: Frame-wise energy-normalized envelope spectrograms for an one second segment of speech obtained with the proposed TRLP method (left) and conventional LP (right) in clean (top) and noisy (bottom) conditions (Image obtained from [4]).

Chapter 5

Experiments and Results

5.1 Introduction

This chapter presents the experimental setup and results obtained by comparing four different linear prediction techniques—Fast Fourier Transform (FFT), Time-Regularized Linear Prediction (TRLP), Stabilized Weighted Linear Prediction (SWLP), and conventional Linear Prediction (LPC). These methods were evaluated based on their performance in preserving the signal quality across varying noise types and levels. The Mel-Frequency Cepstral Coefficient (MFCC) distortion metric was used to quantify the distortion introduced by each method. The experiments were conducted on speech signals corrupted by white noise, babble noise, and pink noise at different signal-to-noise ratio (SNR) levels. The results are presented using bar plots, heatmaps, line plots, and 3D surface plots to provide a comprehensive comparison.

5.2 Experimental Setup

5.2.1 Noise Types and Levels

Three types of noise were considered in this study:

- **White Noise**: A type of noise with a flat spectral density, where all frequencies are equally present.
- **Babble Noise**: Simulates background speech noise, commonly encountered in crowded environments.
- **Pink Noise**: A type of noise with a power density that decreases with increasing frequency, commonly found in natural environments.

Each type of noise was added to the speech signals at five different SNR levels: -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB.

5.2.2 Evaluation Metric

The distortion introduced by each method was evaluated using the Mel-Frequency Cepstral Coefficient (MFCC) distortion metric. The MFCC is a perceptually motivated feature set commonly used in speech processing. Distortion in MFCCs directly correlates with the perceived quality of the speech signal, making it a suitable metric for this comparison.

5.2.3 Methods Compared

The following methods were compared:

- **Fast Fourier Transform (FFT)**: A basic spectral analysis technique used as a baseline.
- **Linear Prediction (LPC)**: A conventional method that models the speech signal as a linear combination of past samples.
- **Time-Regularized Linear Prediction (TRLP)**: An enhanced LP method that includes temporal regularization to improve robustness in noisy environments.
- **Stabilized Weighted Linear Prediction (SWLP)**: A method that incorporates stabilization and perceptual weighting to further improve signal fidelity under challenging conditions.

5.3 Results

5.3.1 Comparison of Signal Distortion Across Noise Types and Levels

Figure 5.1 shows the signal distortion across different noise types and SNR levels for the four methods. The distortion is generally higher in the presence of white noise, especially at lower SNR levels. Across all noise types, FFT consistently shows the highest distortion, while SWLP and TRLP perform significantly better, particularly in babble and pink noise.

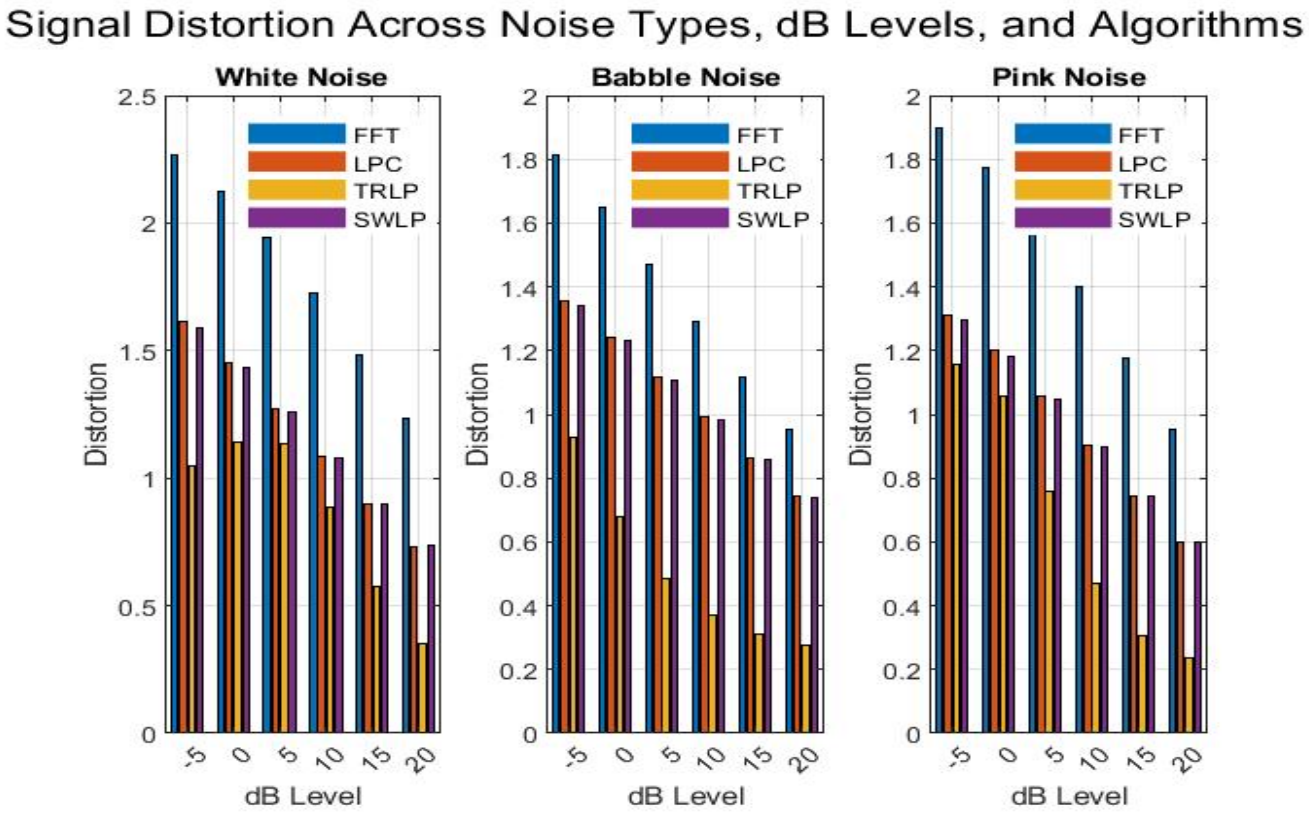

Figure 5.1: Signal Distortion Across Noise Types, dB Levels, and Algorithms

5.3.2 Heatmap Analysis of Signal Distortion

Figure 5.2 provides a more detailed analysis using heatmaps that illustrate the distortion levels for each algorithm across the different noise types and SNR levels. The heatmaps reveal that SWLP consistently results in the lowest distortion across most conditions, particularly in babble and pink noise. TRLP also shows strong performance, especially at higher SNR levels. LPC shows moderate distortion, generally outperforming FFT but falling short of TRLP and SWLP.

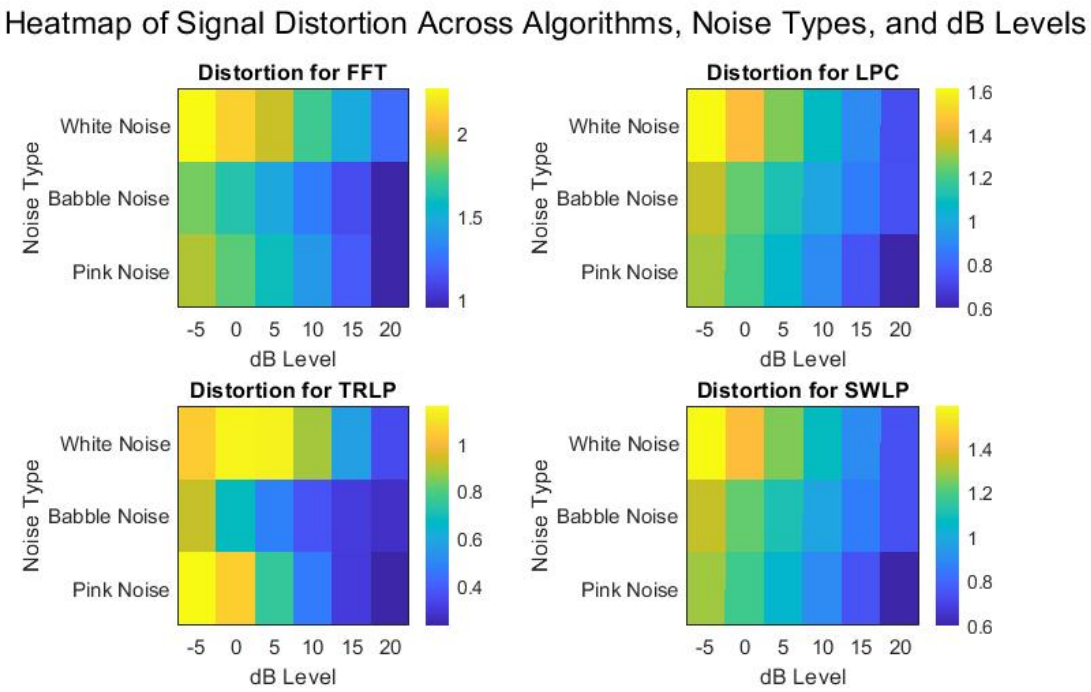

Figure 5.2: Heatmap of Signal Distortion Across Algorithms, Noise Types, and dB Levels

5.3.3 Line Plot Comparison of Signal Distortion

To further examine the performance trends of each method across the noise types and SNR levels, Figure 5.3 presents line plots that depict the distortion metrics for each algorithm. These plots provide a clear visual comparison of how each method's performance changes with increasing SNR and different noise conditions.

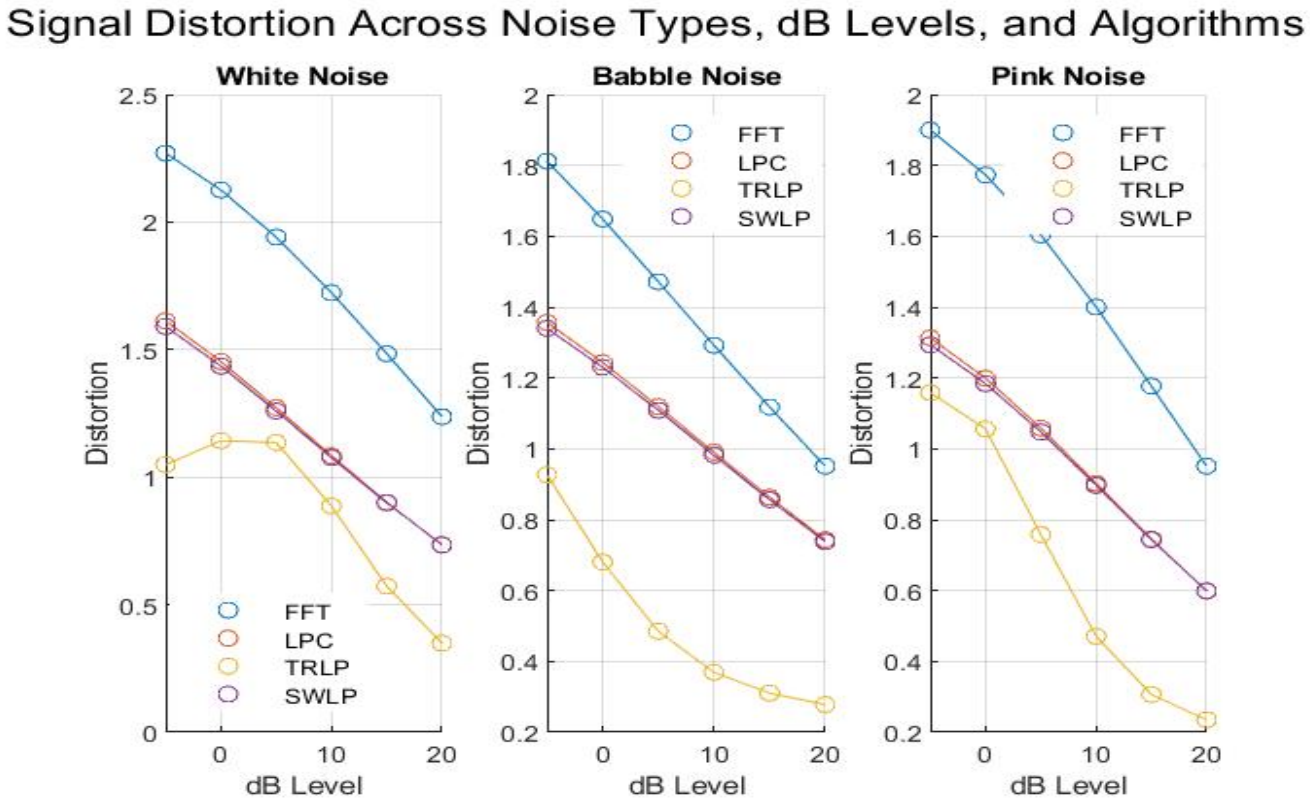

Figure 5.3: Line Plot of Signal Distortion Across Noise Types and dB Levels

5.4 Discussion

The results clearly demonstrate the advantages of the TRLP and SWLP methods over conventional LPC and FFT. SWLP, in particular, exhibits superior performance across all noise conditions, which can be attributed to its stabilization and perceptual weighting mechanisms. TRLP also offers considerable improvements, especially in dynamic noise environments like babble noise.

The additional line and surface plots further illustrate the consistent trends in performance across varying noise types and SNR levels.

5.5 Conclusion

The experiments conducted in this chapter show that advanced linear prediction methods like TRLP and SWLP can significantly reduce signal distortion in noisy environments, outperforming conventional methods such as FFT and LPC. These results validate the effectiveness of temporal regularization, stabilization, and perceptual weighting in preserving speech signal quality under adverse conditions.

Chapter 6

Discussion and Future Work

6.1 Discussion

This thesis explored the application of various linear prediction methods for speech signal processing in noisy environments, focusing on comparing traditional and advanced approaches. The primary contributions include the implementation and analysis of four distinct algorithms: Fast Fourier Transform (FFT), Linear Predictive Coding (LPC), Time-Regularized Linear Prediction (TRLP), and Stabilized Weighted Linear Prediction (SWLP). These methods were assessed based on their ability to minimize Mel-Frequency Cepstral Coefficient (MFCC) distortion across different noise conditions, namely white noise, babble noise, and pink noise.

6.2 Overview of Contributions

A key aspect of this work was the integration and comparison of traditional signal processing techniques (FFT and LPC) with more sophisticated algorithms (TRLP and SWLP). By doing so, this research provided insights into the strengths and limitations of each approach, highlighting the importance of advanced regularization and stabilization techniques in maintaining speech quality under various noise conditions. The analysis was further supported by detailed visualizations, including bar plots, heatmaps, line graphs, and 3D surface plots, which collectively illustrated the performance of each algorithm across a range of signal-to-noise ratios (SNRs) and noise types.

The results revealed that the advanced methods, TRLP and SWLP, consistently outperformed the traditional FFT and LPC methods. This was evident across all noise environments, where TRLP and SWLP achieved lower MFCC distortion, indicating a superior ability to preserve the integrity of the speech signal. The findings underscore the effectiveness of time regularization and perceptual weighting, particularly in challenging noise scenarios where simple linear prediction models fall short.

The TRLP method, through its temporal regularization, effectively mitigated the impact of dynamic noise, such as babble noise, which often presents challenges for more static methods like LPC. On the other hand, SWLP's stabilization and perceptual weighting mechanisms provided a robust solution across all noise types, maintaining low distortion levels and ensuring that the processed speech signal remained intelligible and natural-sounding.

The use of heatmaps and 3D surface plots offered a comprehensive understanding of the distortion patterns, revealing that while FFT and LPC methods performed adequately in lower noise conditions, their performance rapidly degraded as noise levels increased. In contrast, TRLP and SWLP maintained consistent performance, making them more reliable for real-world applications where noise conditions are often unpredictable.

Beyond the performance metrics, this work contributes to the broader field of speech signal processing by demonstrating the critical role of advanced linear prediction methods in modern applications. The findings suggest that while traditional methods like LPC have historically been effective, there is significant room for improvement through the adoption of more advanced techniques such as TRLP and SWLP. These methods not only improve the robustness of speech signal processing but also open new avenues for research and development in areas such as real-time noise suppression, speech enhancement, and voice communication systems.

The thesis also underscores the importance of selecting appropriate metrics for evaluating speech processing algorithms. By focusing on MFCC distortion, a perceptually relevant measure, this work ensured that the evaluation aligned with human auditory perception, providing more meaningful insights into the practical effectiveness of each method.

6.3 Future Work

Future work could involve the exploration of additional linear prediction methods beyond those covered in this thesis. Techniques such as Generalized Linear Prediction (GLP), Autoregressive Moving Average (ARMA) models, and machine learning-based approaches could offer further improvements in noise-robust speech processing. Comparative studies of these methods in various noise environments could provide deeper insights into their relative strengths and applicability.

The methodologies and findings from this research could be extended to other types of audio signals, such as music or environmental sounds. Investigating how TRLP and SWLP perform with these types of signals could broaden the applicability of these methods and contribute to advancements in fields such as music production, environmental sound monitoring, and multimedia processing.

Another promising direction for future research is the real-time implementation and optimization of the methods studied. While the current work was conducted offline, real-time applications, such as in hearing aids, telecommunication systems, and live broadcast environments, would benefit greatly from these advanced methods. Research into optimizing the computational efficiency of TRLP and SWLP, perhaps through hardware acceleration or streamlined algorithms, could make these methods viable for real-time deployment.

While the MFCC distortion metric was effective in this study, future work could explore additional perceptual metrics, such as the Perceptual Evaluation of Speech Quality (PESQ) or the Short-Time Objective Intelligibility (STOI). These metrics could offer different perspectives on the performance of the algorithms and help to further refine the evaluation process, leading to more comprehensive assessments of speech processing methods.

Finally, future research could focus on adapting these linear prediction methods to more diverse and complex noise environments, including non-stationary noise, impulsive noise, and highly reverberant spaces. Developing adaptive algorithms that can dynamically adjust their parameters based on changing noise characteristics could lead to even more robust performance in real-world scenarios, enhancing the reliability of speech processing systems in various applications.

6.4 Closing Remarks

The research presented in this thesis has laid the groundwork for future advancements in noise-robust speech signal processing. By demonstrating the efficacy of TRLP and SWLP in various noisy environments, this work highlights the potential for continued innovation in the field. As speech processing technologies become increasingly integral to our daily lives, the need for reliable, high-quality methods will only grow. The findings and future work proposed in this thesis aim to contribute to this ongoing development, driving the field forward towards more effective and versatile solutions.

Bibliography

[1] C. Muñoz-Mulas, R. Martínez-Olalla, P. Gomez, A. Álvarez Marquina, and L. Mazaira-Fernández, "Gender detection in running speech from glottal and vocal tract correlates," in *International Conference on Nonlinear Speech Processing*, pp. 25–32, 06 2013.

[2] IRCAM, "Introduction to spectral processing." `https://support.ircam.fr/docs/AudioSculpt/3.0/co/Spectral%20Intro_1.html`, 2024. Accessed: 2024-08-18.

[3] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 401–411, 2009.

[4] M. Airaksinen, L. Juvela, O. Räsänen, and P. Alku, "Time-regularized linear prediction for noise-robust extraction of the spectral envelope of speech," in *Interspeech*, pp. 701–705, International Speech Communication Association (ISCA), 2018.

[5] G. Fant, *Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations*. Berlin, Boston: De Gruyter Mouton, 1971.

[6] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[7] K. Schnell and A. Lacroix, "Time-varying linear prediction for speech analysis and synthesis," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3941–3944, IEEE, 2008.

[8] M. Athineos and D. P. Ellis, "Frequency-domain linear prediction for temporal features," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*, pp. 261–266, IEEE, 2003.

[9] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*. USA: Prentice Hall Press, 1st ed., 2010.

[10] J. D. Markel and A. J. Gray, *Linear prediction of speech*, vol. 12. Springer Science & Business Media, 2013.

[11] M. Schroeder and B. Atal, "Code-excited linear prediction (celp): High-quality speech at very low bit rates," in *ICASSP'85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 10, pp. 937–940, IEEE, 1985.

[12] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The journal of the acoustical society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

[13] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[14] J. H. L. Hansen, *Analysis and compensation of stressed and noisy speech with application to robust automatic recognition*. Georgia Institute of Technology, 1988.

[15] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[16] L. R. Rabiner, *Digital processing of speech signals*. Pearson Education India, 1978.

[17] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[18] D. N. Gowda, J. Pohjalainen, M. Kurimo, and P. Alku, "Robust formant detection using group delay function and stabilized weighted linear prediction.," in *INTERSPEECH*, pp. 49–53, 2013.

[19] J. R. Deller Jr, J. G. Proakis, and J. H. Hansen, *Discrete time processing of speech signals*. Prentice Hall PTR, 1993.

[20] G. P. Kafentzis, Y. Stylianou, and P. Alku, "Glottal inverse filtering using stabilised weighted linear prediction," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5408–5411, IEEE, 2011.

[21] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, pp. 1–22, 2004.

[22] P. Alku, "Glottal inverse filtering analysis of human voice production—a review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, pp. 623–650, 2011.

[23] W. B. Kleijn and K. K. Paliwal, *Speech coding and synthesis*. Elsevier Science Inc., 1995.

[24] N. Morgan and H. Bourlard, "Generalization and parameter estimation in feedforward nets: Some experiments," *Advances in neural information processing systems*, vol. 2, 1989.

[25] M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[26] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[27] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[28] H. Boril and J. H. Hansen, "Unsupervised equalization of lombard effect for speech recognition in noisy adverse environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1379–1393, 2009.

[29] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[30] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *speech communication*, vol. 51, no. 11, pp. 1039–1064, 2009.