HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Electrical and Communications Engineering
Laboratory of Acoustics and Audio Signal Processing

Carlo Magi

All-Pole Modelling of Speech: Mathematical Analysis Combined with Objective and Subjective Evaluation of Seven Selected Methods

Master's Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

Espoo, 5th December 2005

Supervisor: Prof. Paavo Alku
Instructor: D.Sc. Tom Bäckström

HELSINKI UNIVERSITY OF TECHNOLOGY

ABSTRACT OF THE MASTER'S THESIS

Author:	Carlo Magi
Name of the thesis:	All-Pole Modelling of Speech: Mathematical Analysis Combined with Objective and Subjective Evaluation of Seven Selected Methods
Date:	5th December 2005 **Number of pages:** 63
Department:	Electrical and Communications Engineering
Professorship:	S-89
Supervisor:	Prof. Paavo Alku
Instructors:	D.Sc. Tom Bäckström

In this work, we study spectral modelling of speech using all-pole models. With those mathematical speech production models, our objective is to find the essential information in natural speech communication. The underlying assumption is that speech can be modelled with the so-called source-filter model. The all-pole model is an implementation of such source-filter models and it model the spectral envelope of the short-time spectrum of speech.

Seven different methods for obtaining the parameters of all-pole models were presented. All methods were formulated using the same notation, in order to present a uniform thery covering the all-pole methods in question. The stability regions of the all-pole models optimised in the time domain were analysed and derived thoroughly. Moreover, a new stability region for the *weighted linear prediction* (WLPC) model was derived.

The spectral modelling properties of these all-pole models were compared using both objective and subjective testing. This was done be comparing their behaviour in the presence of uncorrelated Gaussian and Laplacian background noise. A certain objective measure used was the *logarithmic spectral differences* and the subjective test was carried out as listening tests where the *Degradation Category Rating* testing procedure was used. In both tests, the WLPC model, where the weighting function was the short time energy of the speech signal, gave the best results. The correlation between the objective and subjective results was found to be remarkable strong.

Keywords: All-pole model, speech analysis, linear prediction, prediction polynomial

TEKNILLINEN KORKEAKOULU DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Carlo Magi
Työn nimi:	Puheen AR- mallinnus: Seitsemän valitun menetelmän matemaattinen analyysi sekä niiden objektiivinen ja subjektiivinen evaluointi
Päivämäärä:	5.12.2005 **Sivuja:** 63
Osasto:	Sähkö- ja tietoliikennetekniikka
Professuuri:	S-89
Työn valvoja:	Prof. Paavo Alku
Työn ohjaajat:	TkT Tom Bäckström

Tämä työ käsittelee puheen spektraalista mallinnusta, autoregressiivisiä (AR) malleja apuna käyttäen. Lineaariset puheentuottomallit pyrkivät etsimään ihmisen puheentuotosta kommunikaation kannalta tärkeimmät tekijät. Tämä tehdään yleisesti jakamalla lineaarinen puheentuottomalli lähteeksi ja ääntöväylän suotimeksi. Tällainen jako voidaan toteuttaa AR-mallinnuksella, missä puheen lyhytaikaisen spektrin verhokäyrä saadaan mallinnettua tehokkaasti.

Seitsemän AR-mallia määriteltiin ja formuloitiin yhtenäisiä merkintätapoja käyttäen, minkä seurauksena menetelmiä voitiin vertailla keskenään teoreettisella tasolla. Aika-alueessa optimisoitujen AR-mallien stabiilisuus ominaisuudet formuloitiin rakentavalla ja osittain uudella tavalla. Tämän seurauksena painotetulle lineaariselle ennustusmenetelmälle (WLPC) johdettiin uusi stabiilisuusalue käytettävän painofunktion suhteen.

Kyseisten seitsemän AR-menetelmän ominaisuuksia, kohinaisen puhesignaalin spektriä mallinnettaessa, vertailtiin objektiivisten ja subjektiivisten mittojen valossa. Molemmissa tapauksissa kohinana käytettiin korreloimattomia Gaussin ja Laplacen jakautuneita satunaislukuja. Objektiivisena mittana käytettiin logaritmista spektrin eroavaisuustunnuslukua (SD) ja subjektiivisena mittana kuuntelukokeita. Kuuntelukokeissa käytettiin diskreettiä näytteen huonontuma skaalaa (DCR). WLPC menetelmä, missä painofunktiona käytettiin puhesignaalin lyhytaikaista energiaa, toimi selvästi parhaana menetelmänä molemmissa testeissä. Kyseiset mitat (SD ja DCR) osoitettiinkin korreloivan huomattavan hyvin keskenään.

Avainsanat: AR-malli, puheanalyysi, lineaarinen ennustus, ennustuspolynomi

Acknowledgements

This Master's thesis has been done for the Laboratory of Acoustics and Audio Signal Processing.

I want to thank my supervisor Prof. Paavo Alku, as well as, my instructor and friend Tom Bäckström for their guidance and teaching. I wish to thank all people working in the AKU lab, especially: Prof. Unto K. Laine, Petri Korhonen, Toni Hirvonen and Jouni Pohjalainen. Finally, I would like to thank my precious butterfly and the rest of my family for support.

Otaniemi, 5th December 2005

Carlo Magi

Contents

Abbreviations

ACR	Absolute Category Rating
ASSP	Acoustics, Speech, and Signal Processing
DAM	Diagnostic Acceptability Measure
DAP	Discrete All-Pole (model)
DAT	Digital Audio Tape
DCR	Degradation Category Rating
DMOS	Degradation Mean Opinion Score
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
HUT	Helsinki University of Technology
ICASSP	International Conference on Acoustics, Speech, and Signal Processing
ICSLP	International Conference on Spoken Language Processing
IEEE	Institute of Electrical and Electronics Engineers
IIR	Infinite Impulse Response
IRLS	Iterative Re-weighted Least Squares (algorithm)
ISP	Immittance Spectrum Pair
ITU	International Telecommunication Union
ITU-T	ITU Telecommunication Standardization Section
LMS	Least Mean Square (algorithm)
LP	Linear Prediction
SD	Spectral Distortion
MLE	Maximum Likelihood Estimate
MVDR	Minimum Variance Distortionless Response (model)
RMS	Root Mean Square
SNR	Signal to Noise Ratio
STE	Short Time Energy
WLPC	Weighted Linear Prediction Coefficients (model)
WLSP	Weighted-sum Line Spectrum Pair (model)
WSS	Wide Sense Stationary

List of Figures

List of Tables

List of Symbols

$A(z)$	z-transform of the predictor polynomial
$\mathscr{E}(\cdot)$	cost function
σ^2	error energy
$\rho(\cdot)$	loss function
$\mu(\cdot)$	Itakura-Saito error measure
$\Xi(\cdot)$	flatness measure
$E[\cdot]$	expectation operator
$\nabla_{\mathbf{x}}$	gradient operator with respect to $\mathbf{x}$
$\int_{\gamma}$	definite integral
$\mathcal{F}(\mathbf{A})$	numerical range of the matrix $\mathbf{A}$
$det(\mathbf{A})$	determinant of matrix $\mathbf{A}$
$span\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$	n dimension vector space spanned by linearly independent vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$
$\mathbf{x} \perp \mathbf{y}$	$\mathbf{x}$ and $\mathbf{y}$ are orthogonal
$\mathbb{1}$	vector whose first component is equally to 1 and other to zero
$\mathbf{I}$	identity matrix
$\mathbf{J}$	row reversal operator
$\|\cdot\|$	Hilbert space norm
$\lvert\cdot\rvert$	absolute value
$\mathbf{x}^T$	transpose of vector $\mathbf{x}$
$\mathbf{x}^*$	transpose and complex conjugate of vector $\mathbf{x}$
$\mathbb{R}^n$	n dimension real space
$\mathbb{C}^n$	n dimension complex space
$\mathbb{Z}_+$	set of positive integers
$\mathscr{H}$	complex Hilbert space
$B(\mathscr{H})$	algebra of bounded linear operators on the Hilbert space
$\mathcal{Z}$	z-transform
$\bigcap$	intersection
G_N^{ϕ}	N-polygon with centre at the origin

Chapter 1

Introduction

Speech is the most essential method of human communication. It is the most natural way to transmit variety of information such as feelings and thoughts. If the speech signal is considered at the acoustic level it is composed of rapid fluctuations in air pressure. This is obvious because of the nature of the primary speech transmission channel, atmosphere [14]. When the air molecules are vibrating and colliding the speech information is transmitted through the atmosphere to the ear and comprehended by the brain.

The requirement for the mathematical speech production models is to find the essential inducement concerning the information relevant to the communication. This is the reason why it is important to understand the basic physiological principle of speech production where the lungs act as the source of air for exciting the vocal mechanism. The muscle force pushes air out of lungs and through the entire vocal tract. When the vocal folds are tensed, the air flow causes them to vibrate. This is how *voiced* speech sounds are produced. *Unvoiced* sounds are produced when the vocal folds are relaxed and constrictions in the vocal tract cause turbulent noise, in order to produce a sound.

The underlying assumption in this work is that speech can be modelled with the source-filter model. One of the most widely used speech production models of this kind is the *linear speech production model* developed by Fant [13]. This model assumes that the speech production model or the speech signal can be separated as the *glottal flow* $G(z)$, *vocal tract* $V(z)$, *lip radiation* $L(z)$, and *source* $E(z)$ such that $S(z) = E(z)G(z)V(z)L(z)$, where $S(z)$ is the speech signal.

The all-pole model $\frac{1}{A(z)}$ is an implementation of such source-filter models and yields for good estimates of the factor $G(z)V(z)L(z)$ especially in the case of voiced speech signals ($\frac{1}{A(z)} \simeq G(z)V(z)L(z)$). The reason behind the excellence of all-pole model is the fact that the *vocal tract* can be approximated by the rather simple tube model working as an resonator. The all-pole model is an mathematical implementation of such resonator [13].

All-pole models model the spectral envelope of the short-term spectrum of speech. The short-time spectrum has been one of the most used representations of speech signals. It is widely used in various fields of speech processing, for instance in speech recognition and speech synthesis.

In this work we will present several different methods for obtaining the parameters of all-pole models. We will concentrate on the stability properties of the all-pole models optimised in the time domain, because of the necessity of stability properties in the practical adaptation. It is also relatively easy to derive new all-pole models which do not preserve the stability properties (see [4]). The most popular method for obtaining the parameters of all-pole models is linear prediction (LP) [23]. The stability of the LP filter based on the autocorrelation method is guaranteed, however yet even so it is suffering from several limitations. It is well-known fact that in the presence of the background noise the LP method suffers from many problems for example robustness against the uncorrelated background noise is poor [12, 26]. In this thesis, different all-pole methods are compared in the presence of Gaussian and Laplacian uncorrelated background noise. The quality of speech processing in the presence of additive noise is of interest in a various speech technology applications, such as in speech transmission.

This thesis has been organised as follows. In Chapter 2 the seven all-pole models used in this thesis are formulated. The all-pole methods in question are: linear prediction (LP) [23], weighted linear prediction (WLPC) [10], maximum-likelihood-type estimates (M-estimates) such as HUBER and ℓ_1 [21], weighted sum of line spectrum pair polynomial (WLSP) [9], discrete all-pole model (DAP) [12], and minimum variance distortion-less response (MVDR) [26]. We will present a coherent way to formulate and optimise the seven selected all-pole models. The theoretical aspects explaining modelling errors given by different methods, optimisation criterion, and the spectral envelope properties, are presented in a detailed manner. In the next section, we characterise the stability properties of the all-pole models optimised in the time domain. Different stability regions are being presented, for the weighted LP methods and the stability properties of the LP method and WLSP method is derived in more detailed and accurate way than usually found in literature. Chapters 4, 5, and 6 are concerned with objective and subjective measures for evaluation of the spectral differences of the spectrum envelopes, calculated from clean and contaminated speech samples. The contamination is done by adding Gaussian and Laplacian uncorrelated noise in the clean speech samples. Chapter 7 deals with the results of the comparisons between the different all-pole models with respect to these measures. Finally, in Chapter 8 we will present the conclusions for this work and give suggestions for future work.

Chapter 2

Model Formulations and Basic Properties

2.1 Linear Prediction

2.1.1 Optimisation in the Time Domain

The idea behind the linear prediction (LP) is to estimate a future sample x_n by linear combination of the p past samples [23], [8]. This estimate can be formulated as

$$\hat{x}_n = -\sum_{i=1}^{p} a_i x_{n-i}, \tag{2.1}$$

where weights $a_i \in \mathbb{R}, \forall i = 1, \ldots, p$. The prediction error $\varepsilon_n(\mathbf{a})$ is defined as

$$\varepsilon_n(\mathbf{a}) = x_n - \hat{x}_n = x_n + \sum_{i=1}^{p} a_i x_{n-i} = \mathbf{a}^T \mathbf{x}_n, \tag{2.2}$$

where $\mathbf{a} = (a_0\ a_1\ \cdots\ a_p)^T$ where $a_0 = 1$ and $\mathbf{x}_n = (x_n \cdots x_{n-p})^T$. The goal is to find the coefficient vector $\mathbf{a}$ which minimise the cost function $\mathscr{E}_{LP}(\mathbf{a})$ which is also known as the error energy. This problem can be formulated as the constrained minimisation problem:

$$\begin{aligned} &\text{minimise } \mathscr{E}_{LP}(\mathbf{a}) \\ &\text{subject to } \mathbf{a}^T \mathbb{1} = 1, \end{aligned} \tag{2.3}$$

where the unit vector $\mathbb{1}$ is defined as $\mathbb{1} = (1\ 0\ \cdots\ 0)^T$. The purpose of the constraint is to guarantee that the first element of the optimal solution vector is equal to one. This minimisation depends on the nature of the cost function $\mathscr{E}_{LP}(\mathbf{a})$. Traditionally the cost

function is been defined as $\mathscr{E}_{LP}(\mathbf{a}) = E[|\varepsilon_n(\mathbf{a})|^2]$ where the operator $E[\cdot]$ is defined as the expectation operator. By simple calculation we get

$$\begin{aligned}\mathscr{E}_{LP}(\mathbf{a}) &= E[\varepsilon_n(\mathbf{a})^2] = E[\sum_{k=0}^{p}\sum_{h=0}^{p} a_k a_h x_{n-k} x_{n-h}] \\ &= \sum_{k=0}^{p}\sum_{h=0}^{p} a_k a_h E[x_{n-k} x_{n-h}].\end{aligned} \tag{2.4}$$

Autocorrelation Method

Next we will consider the factor $E[x_n x_{n-k}]$ from Eq. 2.4. The autocorrelation method assumes the signal to be wide sense stationary (WSS), that is $E[x_{n-k}x_{m-k}] = E[x_n x_m]$, $\forall\, k \in \mathbb{Z}$. Let us assume that the signal $\mathbf{x}$ is windowed such that it is zero outside the interval $[0, N]$

$$\tilde{x}_n = x_n w_n,\ \forall n \in \mathbb{Z} \tag{2.5}$$

where the w_n is a window function with $w_n = 0,\ \forall n \in \mathbb{Z}\backslash[0, N]$. An asymptotically unbiased estimator[1] for the expectation operator is

$$E[x_n x_{n-k}] \approx \frac{1}{N+1}\sum_{i=k}^{N} \tilde{x}_i \tilde{x}_{i-k}. \tag{2.6}$$

Let us approximate Eq. 2.4 by using Eq. 2.6, then

$$\begin{aligned}\mathscr{E}_{LP}(\mathbf{a}) &= \sum_{k=0}^{p}\sum_{h=0}^{p} a_k a_h E[x_{n-k}x_{n-h}] = \sum_{k=0}^{p}\sum_{h=0}^{p} a_k a_h E[x_n x_{n-|h-k|}] \\ &\approx \frac{1}{N+1}\sum_{k=0}^{p}\sum_{h=0}^{p}\sum_{i=|h-k|}^{N} a_k a_h \tilde{x}_i \tilde{x}_{i-|h-k|} = \mathbf{a}^T \mathbf{R}_I \mathbf{a},\end{aligned} \tag{2.7}$$

where the matrix $\mathbf{R}_I$ is defined as the autocorrelation matrix whose elements have the property: $\mathbf{R}_I[i, j] = \mathbf{R}_I[i+k, j+k],\ \forall k \in \mathbb{Z}$. $\mathbf{R}_I$ is also a symmetric matrix, that is $\mathbf{R}_I^T = \mathbf{R}_I$. From these properties we conclude that $\mathbf{R}_I$ is symmetric Toeplitz matrix and from Eq. 2.7 it can be alternatively written as

$$\mathbf{R}_I = \frac{1}{N+1}\sum_{n\in I} \tilde{\mathbf{x}}_n \tilde{\mathbf{x}}_n^T = \frac{1}{N+1}\sum_{n\in I} \mathbf{W}_n \mathbf{x}_n \mathbf{x}_n^T \mathbf{W}_n, \tag{2.8}$$

[1] For definition, see Appendix A.

where $\tilde{\mathbf{x}}_n = \mathbf{W}_n\mathbf{x}_n$ and the windowing operator $\mathbf{W}_n$ is the diagonal matrix such that $\mathbf{W}_n = diag(w_n \ \cdots \ w_{n-p})$ and the index set is defined as $I := \{0, \ldots, N+p\}$. Finally, from Eq. 2.7 we find that the cost function to be minimised in Eq. 2.3 is

$$\mathscr{E}(\mathbf{a}) = \mathbf{a}^T\mathbf{R}_I\mathbf{a}, \tag{2.9}$$

where the matrix $\mathbf{R}_I$ is defined in Eq. 2.8. Furthermore the matrix $\mathbf{R}_I$ is known to be symmetric positive definite Toeplitz matrix [15] and this means that quadratic function $\mathscr{E}$ in Eq. 2.9 is convex.

Covariance Method

In the covariance method, the speech signal $\mathbf{x}$ is not assumed to be WSS. Furthermore in this case we use the unbiased estimator for the expectation operator from Eq. 2.4

$$E[x_{n-k}x_{n-h}] \approx \frac{1}{N-p+1}\sum_{i=p}^{N} x_{i-k}x_{i-h}. \tag{2.10}$$

In this case the the Eq. 2.4 can be approximated as

$$\begin{aligned}\mathscr{E}_{LP}(\mathbf{a}) &= \sum_{k=0}^{p}\sum_{h=0}^{p} a_k a_h E[x_{n-k}x_{n-h}] \approx \frac{1}{N-p+1}\sum_{k=0}^{p}\sum_{h=0}^{p}\sum_{i=p}^{N} a_k a_h x_{i-k}x_{i-h} \\ &= \mathbf{a}^T\mathbf{C}_I\mathbf{a},\end{aligned} \tag{2.11}$$

where the covariance matrix $\mathbf{C}_I$ can be written as

$$\mathbf{C}_I = \frac{1}{N-p+1}\sum_{n\in I}\mathbf{x}_n\mathbf{x}_n^T, \tag{2.12}$$

where the index set is defined as $I := \{p+1, \ldots, N\}$. In this case the cost function to be minimised in Eq. 2.3 is

$$\widetilde{\mathscr{E}}(\mathbf{a}) = \mathbf{a}^T\mathbf{C}_I\mathbf{a}. \tag{2.13}$$

The matrix $\mathbf{C}_I$ does not have Toeplitz structure but it is positive definite, which implies that the quadratic function $\widetilde{\mathscr{E}}$ is convex.

Constrained Minimisation Problem

In both (autocorrelation and covariance) methods the quadratic function to be minimised, in the constrained minimisation problem, is convex. Let us use the Lagrange multiplier

method [6] in order to define the new objective function in the case of convex quadratic cost function such as Eq. 2.9

$$\eta(\mathbf{a}, \lambda) = \frac{1}{2}\mathbf{a}^T\mathbf{R}_I\mathbf{a} - \lambda(\mathbf{a}^T\mathbb{1} - 1), \tag{2.14}$$

where $\lambda \neq 0$ is Lagrange multiplier. It is well known that $\mathbf{a}$ minimises η iff it satisfies the linear equation

$$\nabla_{\mathbf{a}}\eta(\mathbf{a}, \lambda) = \mathbf{R}_I\mathbf{a} - \lambda\mathbb{1} = 0, \tag{2.15}$$

where $\nabla_{\mathbf{a}}$ is gradient operator with respect to $\mathbf{a}$. This yields to the equation

$$\mathbf{R}_I\mathbf{a} = \sigma^2\mathbb{1}, \tag{2.16}$$

where Eq. 2.16 is known as the normal equation and $\sigma^2 = \lambda$ denotes error energy. This can be seen by substituting Eq. 2.16 to Eq. 2.9. In the case of the covariance method the normal equation can be obtained in the similar way and finally the normal equation is

$$\mathbf{C}_I\mathbf{a} = \varsigma^2\mathbb{1}. \tag{2.17}$$

Note that the cost function to be minimised in Eq. 2.3 can be defined in a more abstract and general way. That is

$$\mathscr{E}_{LP}(\mathbf{a}) = E[\rho(\varepsilon_n(\mathbf{a}))] \approx \sum_{n\in I}\rho(\varepsilon_n(\mathbf{a})). \tag{2.18}$$

where ρ is the loss function and $\varepsilon_n(\mathbf{a})$ is the prediction error. Finally the general cost function can be written in terms of the prediction error

$$\mathscr{E}(\mathbf{a}) = \sum_{n\in I}\rho(\varepsilon_n(\mathbf{a})). \tag{2.19}$$

Note that choosing $\rho(x) = x^2$ and $I := \{p+1, \ldots, N\} \subset \mathbb{Z}$ and if the prediction error $\varepsilon_n(\mathbf{a})$ is defined like in Eq. 2.2 we get the cost function defined in Eq. 2.13.

2.1.2 LP in the Frequency Domain

In order to understand the LP method it requires the use of frequency domain approach [23]. Let us consider Eq. 2.2. By applying the $\mathcal{Z}$-transform to the error $\varepsilon_n(\mathbf{a})$ one gets

$$E(z) = A(z)X(z) \tag{2.20}$$

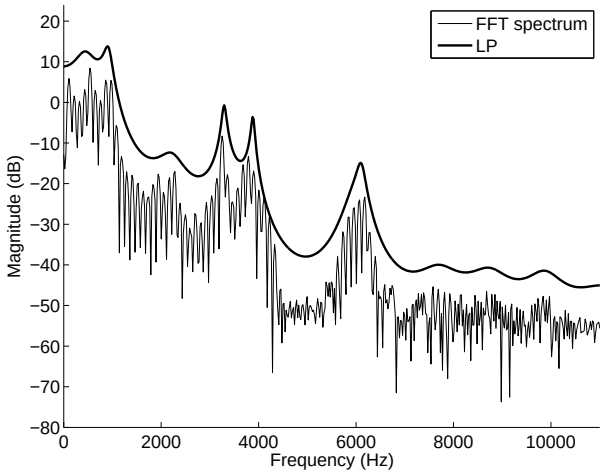

Figure 2.1: FFT spectrum of a male vowel /a/ (thin line) and the all-pole spectrum of the LP method (thick line). The LP order was 22 and the sampling frequency was 22050 Hz. For the sake of clarity the magnitude level of the prediction model have been lifted 10 dB.

where $A(z) = a_0 + a_1 z^{-1} + \ldots + a_p z^{-p}$ is the model $\mathcal{Z}$-transform and it is noted as inverse filter. Recall that $\frac{1}{A(z)}$ was the all-pole filter. $E(z)$ was the $\mathcal{Z}$-transform of the prediction error $\varepsilon_n(\mathbf{a})$ and $X(z)$ was the $\mathcal{Z}$-transform of the speech signal x_n respectively. The total error in Eq. 2.9 can be considered as the infinite sum

$$\mathscr{E}(\mathbf{a}) = \sum_{n=-\infty}^{\infty} \varepsilon_n(\mathbf{a})^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(e^{j\omega})|^2 d\omega \tag{2.21}$$

where the last equality is due to Parseval's theorem and the fact that x_n is assumed to be a deterministic signal. The power spectrum $P(\omega)$ of the signal x_n is defined as

$$P(\omega) = |X(e^{j\omega})|^2 = \frac{|E(e^{j\omega})|^2}{|A(e^{j\omega})|^2}, \tag{2.22}$$

where the last equality is obtained from Eq. 2.20. The signal spectrum $P(\omega)$ is approximated by the all-pole model spectrum $\widetilde{P}(\omega)$. If we assume that the noise $E(z)$ is white, then $|E(z)|^2 = \sigma^2$ and from [23] we have

$$\widetilde{P}(\omega) = \frac{\sigma^2}{|A(e^{j\omega})|^2}, \tag{2.23}$$

where σ^2 is the error energy. If we compare Eqs. 2.22 and 2.23 we see that the more "flat" the residual power spectrum is the better approximation is accomplished ($|E(e^{j\omega})|^2 \approx \sigma^2$).

Now combining Eqs. 2.20-2.23 we obtain

$$\begin{aligned}\mathscr{E}(\mathbf{a}) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}|E(e^{j\omega})|^2 d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi}|A(e^{j\omega})|^2|X(e^{j\omega})|^2 d\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}|A(e^{j\omega})|^2 P(\omega) d\omega = \frac{\sigma^2}{2\pi}\int_{-\pi}^{\pi}\frac{P(\omega)}{\widetilde{P}(\omega)}d\omega.\end{aligned} \tag{2.24}$$

If we minimise $\mathscr{E}(\mathbf{a})$ as in Eq. 2.3 and get the vector $\tilde{\mathbf{a}}$ solving Eq. 2.16, the resulting minimum energy $\mathscr{E}_{min} = \mathscr{E}(\tilde{\mathbf{a}})$ is equal to the gain factor σ^2 from Eqs. 2.16 and 2.23. This means that Eq. 2.24 can be written in the form

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{P(\omega)}{\widetilde{P}_{min}(\omega)}d\omega = 1, \tag{2.25}$$

where $\widetilde{P}_{min}(\omega)$ is the all-pole model spectrum corresponding to the vector $\tilde{\mathbf{a}}$ such that $\mathscr{E}(\tilde{\mathbf{a}}) = \sigma^2$.

Using Eqs. 2.24 and 2.25 we can define two major properties of the LP error measure $\mathscr{E}(\mathbf{a})$. These properties are called *global property* and *local property* [23]. The *global property* means that the spectral match at frequencies with high energy is not better than the match at the frequencies with little energy. This is because the error energy $\mathscr{E}(\mathbf{a})$ is determined by the ratio of the two spectra seen in Eq. 2.24 and therefore the spectral matching process is performed uniformly over the entire frequency range.

If a small region of the spectrum is considered, one observes that in order to minimise Eq. 2.24, a better fit is obtained when $\widetilde{P}(\omega) > P(\omega)$ (i.e. $P(\omega)/\widetilde{P}(\omega)$ is small), on average, than vice versa. This means that the resulting estimate $\widetilde{P}(\omega)$ is above the original spectrum. This property is called the *local property* and that is why the resulting model spectrum $\widetilde{P}(\omega)$ is a good estimate of the spectral envelope of the signal spectrum $P(\omega)$ see Fig. 2.1. From Eq. 2.25 one can notice one of the major disadvantages of LP modelling. This disadvantage is called the cancellation of error and it means that the contributions to the error when $\widetilde{P}(\omega) > P(\omega)$ cancel those when $\widetilde{P}(\omega) < P(\omega)$.

2.2 Weighted Linear Prediction Analysis

The weighted LP (WLPC) analysis [10] uses the loss function defined as

$$\rho(\varepsilon_n(\mathbf{a})) = (\varepsilon_n(\mathbf{a}))^2 w_n, \tag{2.26}$$

where w_n is defined as a discrete weight function and the prediction error $\varepsilon_n(\mathbf{a})$ is defined as in Eq.2.2.. By substituting Eq. 2.26 to Eq. 2.19 the weighted-residual energy (cost function)

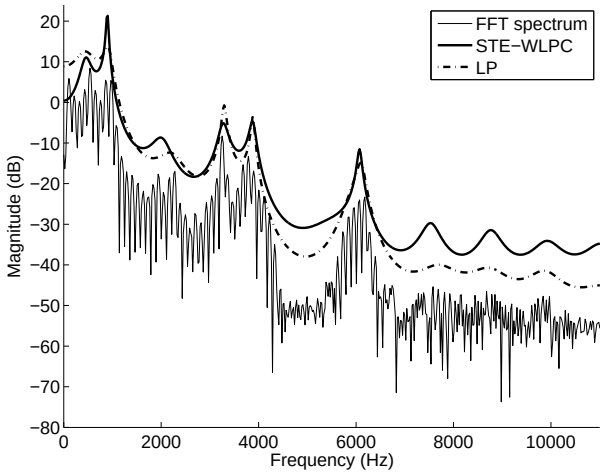

Figure 2.2: FFT spectrum of a male vowel /a/ (thin line) and the all-pole spectrum of the STE-WLPC method (thick line) together with the LP method ($M = 12,\ k = 1$). The order of the all-pole models was 22 and the sampling frequency was 22050 Hz. For the sake of clarity the magnitude levels of the prediction models have been lifted 10 dB.

$\mathscr{E}$ becomes

$$\mathscr{E}(\mathbf{a}) = \sum_{n \in I} (\varepsilon_n(\mathbf{a}))^2 w_n. \tag{2.27}$$

In order to minimise $\mathscr{E}(\mathbf{a})$ one results in similar kind of normal equation as in LP method.

$$\mathscr{E}(\mathbf{a}) = \sum_{n \in I} (\varepsilon_n(\mathbf{a}))^2 w_n = \mathbf{a}^T \left(\sum_{n \in I} w_n \mathbf{x}_n \mathbf{x}_n^T \right) \mathbf{a} = \mathbf{a}^T \widetilde{\mathbf{R}}_I \mathbf{a}, \tag{2.28}$$

where $\widetilde{\mathbf{R}}_I = \sum_{n \in I} w_n \mathbf{x}_n \mathbf{x}_n^T$. By using the same minimisation method as in Eq. 2.3 it can be seen that $\mathbf{a}$, which minimises $\mathscr{E}$ in Eq. 2.28, satisfies the linear equation

$$\widetilde{\mathbf{R}}_I \mathbf{a} = \tilde{\sigma}^2 \mathbb{1}. \tag{2.29}$$

The idea behind the weight function is to over-weight or select [32] the speech samples that fit the LP model well. This means that those temporary excitation free speech samples produce small LP residual are over-weighted. On the other hand, those speech samples during the waveform changes rapidly due to, for example closure of the vocal folds are more difficult to predict and hence they results in a larger residual. Those samples should be down-weighted. Generally, if the change in the wave form is too rapid, linear models are not able to follow such changes. It has been observed in [10], that the pre-emphasised vowel sounds show clear peaks just after, and clear valleys just before, the moments of excitations.

These peaks corresponds also to peaks in the LP residual. This observation suggests that we could use the short time energy (STE) of the signal in weighting Fig. 2.2. Altogether, there are many ways of doing the proper weighting for the sum of squared prediction errors. In literature, it has been proposed to use the STE of the signal either as a selection criterion or as a weighting function. In this thesis, the STE of the signal is employed as a weighting function, because the sample-selective methods suffers from several shortcomings such as high computational complexity (see [10]). Finally, the the short-time energy weighting function can be formulated as

$$w_n = \sum_{i=0}^{M-1} x_{n-i-k}^2, \tag{2.30}$$

where the k is the delay and M is the length of the STE window. From Eq. 2.30 we can readily see that indeed the speech samples which follow the main excitation are over-weighted and those samples which contain excitations are down-weighted (see Fig. 2.3).

2.3 Maximum-Likelihood-Type Estimates

Let us consider more precisely the concept of maximum-likelihood-type estimates (M-estimates). Our objective is to minimise the cost function $\mathscr{E}$ defined in the same manner as in Eq. 2.19.

$$\min_{\mathbf{a}} \mathscr{E}(\mathbf{a}) = \min_{\mathbf{a}} \sum_{n \in I} \rho(\varepsilon_n(\mathbf{a})). \tag{2.31}$$

The Huber has shown that the loss function ρ should be symmetric $\rho(-x) = -\rho(x)$ and it should have a bounded derivative $|\frac{\partial \rho(x)}{\partial x}| = |\psi(x)| < M|x|$ [18, 21], where the $\psi(x)$ is defined as

$$\psi(x) = \frac{\partial \rho}{\partial x}(x). \tag{2.32}$$

In general the solution to Eq. 2.31 is not scale-invariant. This means that if the data is multiplied by a constant, the new estimate differs from the original estimate. It is possible to define the scale-invariant solution by $\hat{\psi}(x) = \hat{r}\psi(x/\hat{r})$. Where the scalar $\hat{r}$ is defined as a robust scale estimate [18]. The scale-invariant solution $\mathbf{a}$ to Eq. 2.31 satisfies

$$\sum_{n \in I} x_{n-j} \hat{\psi}(\varepsilon_n(\mathbf{a})) = 0 \quad j = 1, \ldots, p. \tag{2.33}$$

This group of equations is, in general, nonlinear and iterative methods are required in order to solve the vector $\mathbf{a}$. The Newton algorithm and the iterative re-weighted least squares

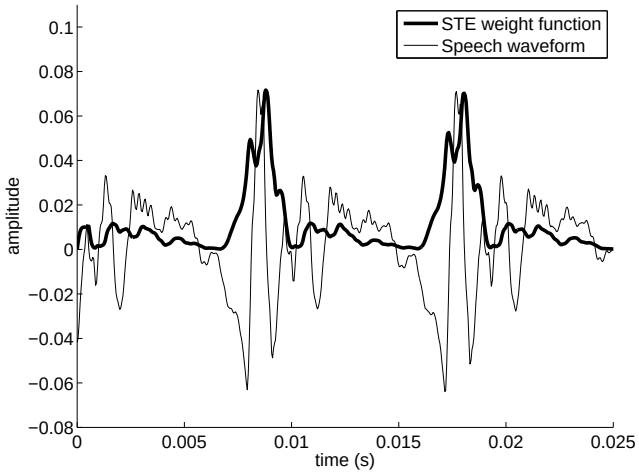

Figure 2.3: Time waveform of a male vowel /a/ (thin line) and STE-weight function (thick line) where from Eq. 2.30: $M = 12,\ k = 1$. The STE-weight function is scaled to the same level as the speech waveform.

algorithm (IRLS) have been proposed in order to solve this system of equations [21]. Let us consider in more detail the IRLS solution. If we assume that x is very small (this is a valid assumption because of in applications x is the prediction error) the derivative $\frac{\partial\psi(x)}{\partial x}$ can be approximated as $\frac{\partial\psi(x)}{\partial x} \approx \frac{\psi(x)}{x} = W(x)$. By substituting this estimate in Eq. 2.33 we obtain the equation

$$\sum_{n\in I} x_{n-j}\varepsilon_n(\mathbf{a}^{k+1})W(\varepsilon_n(\mathbf{a}^k)) = 0, \quad j = 1,\ldots,p \tag{2.34}$$

The $\mathbf{a}^k$ is the kth iteration of the solution. If the definition of the prediction error from Eq. 2.2 is applied to Eq. 2.34 and we set $w_n = W(\varepsilon_n(\mathbf{a}^k))$, then we have

$$\sum_{n\in I}\sum_{i=1}^{p} x_{n-j}x_{n-i}w_n a_i^{k+1} = -\sum_{n\in I} x_{n-j}x_n w_n, \quad j = 1,\ldots,p. \tag{2.35}$$

This group of equations can be written in the matrix notation with $a_0 = 1$, as

$$\widetilde{\mathbf{R}}_I\mathbf{a} = \tilde{\sigma}^2\mathbb{1}, \tag{2.36}$$

Where $\widetilde{\mathbf{R}}_I = \sum_{n\in I} w_n\mathbf{x}_n\mathbf{x}_n^T$. (Note that the choice of the index set I defines whether matrix $\mathbf{R}_I$ is a symmetric Toeplitz matrix or not.) Finally, from Eq. 2.36, the approximated M-estimate method yields to the WLPC method. Therefore one can straightforwardly derive the requirements for the weight function $W(\varepsilon_n(\mathbf{a}^k))$ that guarantee the corresponding all-pole filter to be in the minimum phase as we see later in Section 3.2.

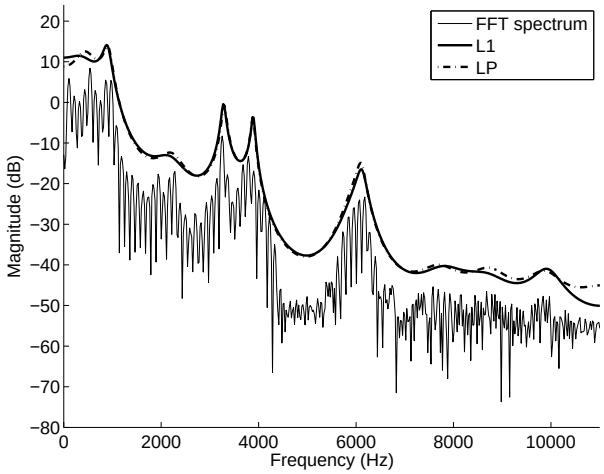

Figure 2.4: FFT spectrum of a male vowel /a/ (thin line) and all-pole spectrum of M-estimate method with ℓ_1 loss function (thick line) together with LP method. The order of the all-pole models was 22 and the sampling frequency was 22050 Hz. For the sake of clarity the magnitude levels of the prediction models have been lifted 10 dB.

2.3.1 Properties of the loss function

So far we have assumed that the error density of the error signal, or equivalently the speech signal in Eq. 2.2, to be Gaussian (linear mapping is invariant with respect to the distribution of the random number). It is well known from [21, 18] that if the error density f is known, then the loss function in Eq. 2.31 can be chosen as

$$\rho(x) = -\ln\left[f(x)\right] \tag{2.37}$$

and the estimate for the weights $\mathbf{a}$ obtained from Eq. 2.31, is the maximum likelihood estimate (MLE). It is easy to see that if we choose the error density to be Gaussian $f(x) = e^{-\frac{1}{2}x^2}$ we obtain the loss function of the form $\rho(x) = -\ln(e^{-\frac{1}{2}x^2}) = \frac{1}{2}x^2$, which is exactly the same loss function that we have used in LP analysis. Moreover, if the derivative is calculated as $\psi(x) = \frac{\partial}{\partial x}(\frac{1}{2}x^2) = x$ and if we calculate the weight function w_n from Eq. 2.35, we get $w_n = \frac{\psi(x)}{x} = 1$, which is the original LP method.

There are a more general error criteria called the $\ell_\wp$ error measures. The loss function in the $\ell_\wp$ method is defined as

$$\rho_\wp(x) = \frac{1}{\wp}|x|^\wp, \quad 1 \leqslant \wp \leqslant 2 \tag{2.38}$$

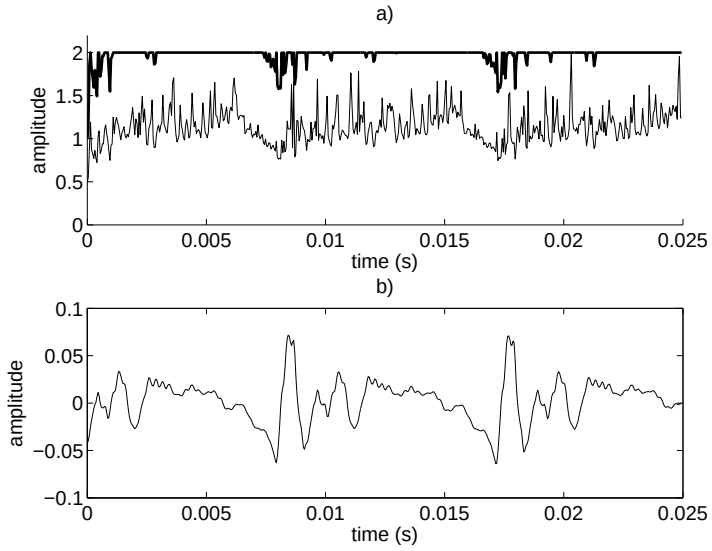

Figure 2.5: **a)** Normalised Huber logarithmic loss function (thick line), where the tuning constant c is the mean value of the signal amplitude from panel b). The thin line represent the normalised ℓ_1 logarithmic loss function. Both functions have been scaled in order to draw them in the same picture and both have been calculated from the excitation signal from the covariance method calculated from speech signal from lower panel.**b)** Time waveform of a male vowel /a/ where the loss functions have been calculated.

and

$$\psi_{\wp}(x) = sgn(x)|x|^{\wp-1}, \quad 1 \leqslant \wp \leqslant 2. \tag{2.39}$$

Notice that $\psi_{\wp}(x)$ is bounded only when $\wp = 1$. That is why, in this work, the ℓ_1 error measure ($\rho_1(x) = |x|$) is used and the corresponding estimate is called the sample median estimate see Figs. 2.4 and 2.5. The resulting estimate for the least absolute deviation estimate is optimal if the error and distribution are Laplacian. The sample median estimate is sensitive to the behaviour of the error distribution at its median. On the other hand, if the ℓ_2 error measure is used, which is the classical LP error measure, the obtained estimate, called the sample mean estimate, is very sensitive to the tail behaviour of the error distribution.

It is well known that the non-Gaussian nature of the model excitation for voiced speech should be taken into account when one is choosing the proper loss function. For all-pole modelling for natural speech, the error density is not exactly known. This is because there are always some outliers affecting on the signal. The distribution is assumed to be a mixed distribution F of the form [18]

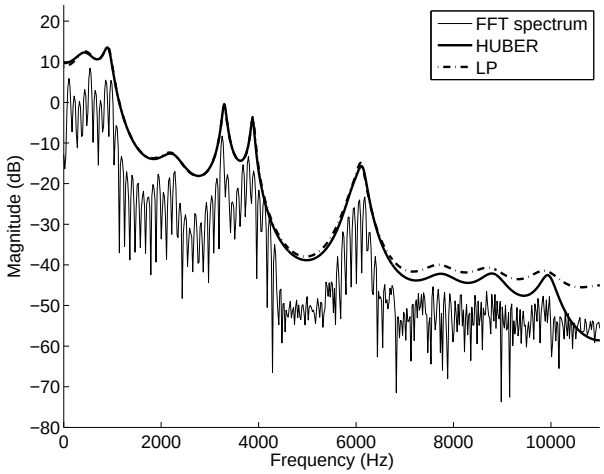

Figure 2.6: FFT spectrum of a male vowel /a/ (thin line) and the all-pole spectrum of the M-estimate method with the Huber loss function (thick line) together with the LP method. The order of the all-pole models was 22 and the sampling frequency was 22050 Hz. For the sake of clarity the magnitude levels of the prediction models have been lifted $10\ dB$.

$$\mathrm{F} = (1-\epsilon)\Phi + \epsilon \mathrm{L}, \tag{2.40}$$

where Φ is the standard normal cumulative, L is an unknown contaminating distribution and $\epsilon \in [0, 1)$. The error density corresponding to F is formulated as in [18, 21]

$$f_H(x) = \frac{(1-\epsilon)}{\sqrt{2\pi}} e^{-\epsilon \mathrm{L}}. \tag{2.41}$$

This density is Gaussian in the middle and Laplacian at the tails. The corresponding loss function, known as the Huber's loss function, is defined as

$$\rho_H(x) = \begin{cases} \frac{1}{2}x^2 & \text{if} \quad |x| \leqslant c \\ c|x| - \frac{1}{2}c^2 & \text{if} \quad |x| > c \end{cases} \tag{2.42}$$

where c is an efficiency tuning constant as in [21], which is a function of the corrupted percentage ϵ. The tuning constant c should be chosen to achieve high efficiency both for the nominal Gaussian distribution and for most mixture distributions. If we take the derivate of Eq. 2.42 in order to calculate the psi-function we obtain

$$\psi_H(x) = \min\{c, \max\{x, -c\}\}. \tag{2.43}$$

The ψ_H belongs to the class of *minmax* estimators that lie between the sample mean (ℓ_2) and the sample median (ℓ_1) see Fig. 2.5. From Eq. 2.43 we can find that ψ_H is indeed

bounded, monotonically nondecreasing (this assures uniqueness of the estimate solutions), and continuous.

2.4 Weighted Sum of Line Spectrum Pair Polynomials

Let us introduce the zero ended coefficient vector $\tilde{\mathbf{a}} = (\mathbf{a}^T \ 0)^T$, where $\mathbf{a} = (1 \ a_1 \ \cdots \ a_p)^T$ is the vector that solves Eq. 2.16. The coefficient vectors for the LSP polynomials can be defined as

$$\begin{aligned} \mathbf{p} &= \tilde{\mathbf{a}} + \mathbf{J}\tilde{\mathbf{a}} \\ \mathbf{q} &= \tilde{\mathbf{a}} - \mathbf{J}\tilde{\mathbf{a}} \end{aligned} \tag{2.44}$$

where $\mathbf{J}$ denotes the row reversal operator, which can be implemented as

$$\mathbf{J} = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ \vdots & & 1 & 0 \\ 0 & & & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix}. \tag{2.45}$$

This implies that the vector $\mathbf{p}$ is symmetric and $\mathbf{q}$ antisymmetric. For the weighted sum of line spectrum pair polynomials (WLSP) method [9] one defines coefficient vector $\mathbf{d}$ of the corresponding polynomial $D(z)$ as

$$\mathbf{d} = \lambda\mathbf{p} + (1-\lambda)\mathbf{q}, \tag{2.46}$$

or equivalently, taking the $\mathcal{Z}$-transform from Eq. 2.46

$$D(z) = \lambda P(z) + (1-\lambda)Q(z), \tag{2.47}$$

where $\lambda \in (0, 1)$.

2.4.1 Properties of the Error Energy as a Function of λ

We are ready to perform the same error analysis as in [9]. Consider Eq. 2.16. If one takes the extended positive definite symmetric Toeplitz matrix $\mathbf{R}$ and multiply it from the right by the zero extended coefficient vector $\tilde{\mathbf{a}}$ then from Eq. 2.44 we have

$$\mathbf{R}\tilde{\mathbf{a}} = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 & \gamma \end{pmatrix}^T, \tag{2.48}$$

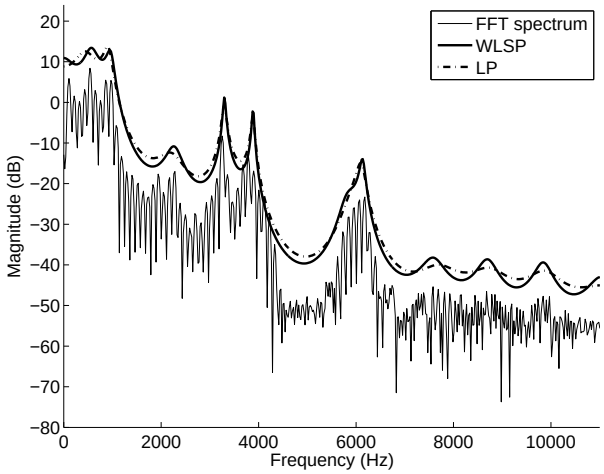

Figure 2.7: FFT spectrum of a male vowel /a/ (thin line) and the all-pole spectrum of the WLSP method (thick line) together with the LP method. The order of the all-pole models was 22 and the sampling frequency was 22050 Hz. For the sake of clarity the magnitude levels of the prediction models have been lifted 10 dB.

where $\gamma = \sum_{i=0}^{p} a_i R(p-i+1)$. In view of Eq. 2.48 one can write

$$\begin{aligned} \mathbf{Rp} &= \mathbf{R\tilde{a}} + \mathbf{RJ\tilde{a}} = \begin{pmatrix} \sigma^2 + \gamma & 0 & \cdots & 0 & \sigma^2 + \gamma \end{pmatrix}^T \\ \mathbf{Rq} &= \mathbf{R\tilde{a}} - \mathbf{RJ\tilde{a}} = \begin{pmatrix} \sigma^2 - \gamma & 0 & \cdots & 0 & \gamma - \sigma^2 \end{pmatrix}^T \end{aligned} \tag{2.49}$$

Combining Eqs. 2.46 and 2.49 we can write

$$\mathbf{Rd} = \lambda\mathbf{Rp} + (1-\lambda)\mathbf{Rq} = \begin{pmatrix} \sigma^2 + (2\lambda-1)\gamma & 0 & \cdots & 0 & (2\lambda-1)\sigma^2 + \gamma \end{pmatrix}^T. \tag{2.50}$$

It is interesting to note that choosing λ such that the last element is equally to zero which implies

$$\lambda = -\gamma/(2\sigma^2) + 1/2, \tag{2.51}$$

we obtain a similar equation as Eq. 2.16

$$\mathbf{Rd} = \tilde{\sigma}^2 \mathbb{1}, \tag{2.52}$$

where the error energy $\tilde{\sigma}^2 = \sigma^2 - (\gamma/\sigma)^2 \leqslant \sigma^2$. This means that the residual energy of the $p+1$ order all-pole model corresponding to vector $\mathbf{d}$ is smaller than the residual energy

given by the LP model of order p. In fact, the vector $\mathbf{d}$ from Eq. 2.50 yields to classic LP model of order $p+1$, if λ is chosen as in Eq. 2.51. This implies, interestingly, that the residual energy of the classical LP model of order p is smaller than the LP model of order $p-1$. This is the classical result for the LP analysis [23]. If we choose $\lambda = 1/2$ we obtain from Eqs. 2.44 and 2.46 that $\mathbf{d} = \tilde{\mathbf{a}}$. Then the $\mathbf{d}$ is the original LP model of order p from Eq. 2.16. Finally, it is worth noticing that the WLSP method can be interpreted as interpolation between the LP models of order p and $p+1$.

2.5 Discrete All-Pole Modelling

The discrete all-pole model (DAP) uses the discrete Itakura-Saito (IS) error measure [30, 12, 25, 2] and the optimisation criterion is derived in the frequency domain. The reason behind the idea is to overcome the well-known limitations of LP. That is for example the ambition of the all-pole spectral envelopes to bias towards the pitch harmonics and the error cancellation property [12, 24, 23]. The one of the major limitation of the LP method can be seen by computing the all-pole envelope of the discrete spectrum, which is always the case when spectra is computed using FFT. The all-pole envelope tend to bias towards the pitch harmonics especially for the voiced speech. This is the case especially when the F_0 is very high. From Eq. 2.25 and in view of the *local properties* of the LP error measure in subsection 2.1.2 the reason behind this behaviour is obvious, see [12], [24] for more details. In the case of conventional LP method the minimum error from Eq. 2.25 is obtained without the identical match between the spectra $P(\omega)$ and $\widetilde{P}(\omega)$. In the DAP modelling, the error function reaches the minimum ($\mathscr{E}(\mathbf{a}) = 0$) only when the model spectrum coincide on all discrete points. Moreover, the DAP method tries to maximise the error flatness. Next we will consider the concept of spectral flatness in the case of original LP model in order to motivate new minimisation method in DAP based on these observations.

2.5.1 Maximising the Spectral Flatness of the Error Spectrum

Let us look at the concept of maximising spectral flatness of continuous spectra for the original LP model. Let $E(z)$ and $\mathscr{E}(\mathbf{a})$ be defined as in Eqs. 2.21 and 2.22. Let us define the normalised log spectrum of the error as

$$V(\omega) = ln\left[|E(e^{j\omega})|^2/\mathscr{E}(\mathbf{a})\right]. \tag{2.53}$$

The Itakura-Saito error measure can be written as [25]

$$\mu(E) = \int_{-\pi}^{\pi}\left[e^{V(\omega)} - V(\omega) - 1\right]\frac{d\omega}{2\pi}. \tag{2.54}$$

Note that from Eqs. 2.21 and 2.53 we get $\int_{-\pi}^{\pi} e^{V(\omega)} \frac{d\omega}{2\pi} = 1$. We can define the flatness measure $\Xi(E)$ based on Itakura-Saito error measure as follows

$$\Xi(E) = \exp\left[\int_{-\pi}^{\pi} V(\omega)\frac{d\omega}{2\pi}\right] = \frac{\exp\left[\int_{-\pi}^{\pi} ln|E(e^{j\omega})|^2 \frac{d\omega}{2\pi}\right]}{\mathscr{E}(\mathbf{a})}, \tag{2.55}$$

In view of Eqs. 2.53 and 2.55 it is easy to see that the spectral flatness measure $\Xi(E)$ lie between zero and one, and it is equal to one for a perfectly flat spectrum (since $\mathscr{E}(\mathbf{a})$ does not depend on ω).

Recall the Eq. 2.20. Then

$$\begin{aligned}\int_{-\pi}^{\pi} ln|E(e^{j\omega})|^2 \frac{d\omega}{2\pi} &= \int_{-\pi}^{\pi} ln\left(|A(e^{j\omega})|^2 |X(e^{j\omega})\,|^2\right)\frac{d\omega}{2\pi} \\ &= \int_{-\pi}^{\pi} ln|X(e^{j\omega})|^2 \frac{d\omega}{2\pi} + \int_{-\pi}^{\pi} ln|A(e^{j\omega})|^2 \frac{d\omega}{2\pi} \\ &= \int_{-\pi}^{\pi} ln|X(e^{j\omega})|^2 \frac{d\omega}{2\pi},\end{aligned} \tag{2.56}$$

where the last equality is due to the fact that if $A(z)$ is restricted to have zeros inside the unit circle then its log spectrum has zero average[2] value (see [25] for more detail). Finally Eq. 2.56 can be substituted in to Eq. 2.55 and rewritten as

$$\Xi(E) = \frac{\exp\left[\int_{-\pi}^{\pi} ln|X(e^{j\omega})|^2 \frac{d\omega}{2\pi}\right]}{\mathscr{E}(\mathbf{a})} \tag{2.57}$$

If the input to filter $A(z)$ is fixed, then

$$\Xi(E) = \frac{c}{\mathscr{E}(\mathbf{a})}, \tag{2.58}$$

where $c = \exp\left[\int_{-\pi}^{\pi} ln|X(e^{j\omega})|^2 \frac{d\omega}{2\pi}\right]$ is a constant. From Eq. 2.58 we see that minimising the total error $\mathscr{E}(\mathbf{a})$ is equivalent to choosing the inverse filter $A(z)$ that maximises the spectral flatness in Eq. 2.55 at its output.

2.5.2 Minimisation criterion in DAP

The Eq. 2.58 motivates us to maximise the residual spectral flatness instead of minimising the error energy. This is performed in DAP using discrete spectra. That is why the Itakura-Saito error measure $\mu(E)$ in Eq. 2.54 must be used in the discrete form $\mu_D(E)$. It has been

[2]For details, see Appendix A

done in [12] in order to derive the minimisation criterion for DAP

$$\begin{aligned}\mu_D(E) &= \frac{1}{N}\sum_{m=1}^{N}\left[e^{V(\omega_m)} - V(\omega_m) - 1\right] \\ &= \frac{1}{N}\sum_{m=1}^{N}\left[\frac{|E(e^{j\omega_m})|^2}{\sigma^2} - ln\left[\frac{|E(e^{j\omega_m})|^2}{\sigma^2}\right] - 1\right] \\ &= \frac{1}{N}\sum_{m=1}^{N}\left[\frac{P(\omega_m)}{\widetilde{P}(\omega_m)} - ln\frac{P(\omega_m)}{\widetilde{P}(\omega_m)} - 1\right]\end{aligned} \tag{2.59}$$

where the σ^2 is the error energy and $P(\omega_m)$ is the given discrete spectrum defined at N frequency points ω_m and $\widetilde{P}(\omega_m)$ is the all-pole model spectrum, defined in Eq. 2.23. Note that the discrete spectral points can be chosen freely but, in practise, in the DAP algorithm sampling is performed at the harmonics. The vector $\mathbf{a}$ from Eq. 2.2 is obtained by setting $\partial\mu_D(E)/\partial\mathbf{a} = 0$. We do not go into the details (see [12] for more detail) but the vector obtained solves the equation

$$\mathbf{R}_I\mathbf{a} = \tilde{\mathbf{h}}, \tag{2.60}$$

where $\mathbf{R}_I$ is symmetric Toeplitz matrix defined in Eq. 2.8 and vector $\tilde{\mathbf{h}}$ is the impulse response of the model and for further definition the reader is referred to see [12]. The example of a spectral envelope given by DAP is seen in Fig. 2.8. After minimisation of Eq. 2.59, we find that [12]

$$\mu_{Dmin} = ln\frac{\left[\prod_{m=1}^{N}\widetilde{P}(\omega_m)\right]^{1/N}}{\left[\prod_{m=1}^{N}P(\omega_m)\right]^{1/N}} \tag{2.61}$$

and

$$\frac{1}{N}\sum_{m=1}^{N}\frac{P(\omega_m)}{\widetilde{P}(\omega_m)} = 1. \tag{2.62}$$

The Eq. 2.62 is the discrete form of the Eq. 2.25. From these equations one can conclude that the minimum error $\mathscr{E}_{min}$ is equal to the logarithm of the ratio of the geometric means of the model spectrum and the original spectrum.

2.6 Minimum Variance Distortionless Response Modelling

So far we have presented the different disadvantages of the original LP method and tried to overcome those limitations by defining the set of methods such as WLPC, WLSP, DAP, and

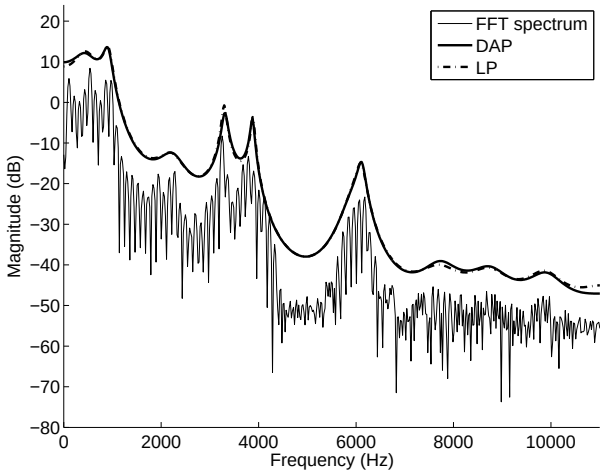

Figure 2.8: FFT spectrum of a male vowel /a/ (thin line) and the all-pole spectrum of the DAP method (thick line) together with the LP method. The order of the all-pole models was 22 and the sampling frequency was 22050 Hz. For the sake of clarity the magnitude levels of the prediction models have been lifted 10 dB.

M-estimates in general. A further problem is that if the order of the original LP method is increased, then the corresponding envelope overestimates the original voiced speech power spectrum. This means that the LP envelope is resolving the harmonics and not the spectral envelope. The minimum variance distortionless response (MVDR) method [26] provides a smooth spectral envelope even when the model order is increased. In particular, if one chooses the proper order for the MVDR method the all-pole envelope obtained, models a set of spectral samples exactly.

Let us recall the constrained minimisation problem defined in Sec. 2.1 Eq. 2.3. In MVDR methodology the constrained minimisation problem can be stated as

$$\begin{aligned} &\text{minimise } \mathscr{E}(\mathbf{h}_\ell) = \mathbf{h}_\ell^* \mathbf{R}_I \mathbf{h}_\ell \\ &\text{subject to } \mathbf{h}_\ell^* \mathbf{v}(\omega_\ell) = 1 \end{aligned} \tag{2.63}$$

where $\mathbf{R}_I$ is the positive definite symmetric Toeplitz matrix defined as in Eq. 2.8, $\mathbf{h}_\ell = (h_0 \ h_1 \ \cdots \ h_p)^T$ is the distortionless filter to be optimised and $\mathbf{v}(\omega) = (1 \ e^{j\omega} \ \cdots \ e^{jp\omega})^T$. The distortionless constraint $H_\ell(e^{j\omega_\ell}) = \mathbf{v}^*(\omega_\ell)\mathbf{h}_\ell = 1$, where H_ℓ is noted as the frequency response, ensures that the input signal components with frequency ω_ℓ will pass through undistorted. Let us use the Lagrange multiplier method to solve the underlying filter from

Eq. 2.63 as in Sec. 2.1 then

$$\eta(\mathbf{h}_\ell, \lambda) = \frac{1}{2}\mathbf{h}_\ell^* \mathbf{R}_I \mathbf{h}_\ell - \lambda(\mathbf{h}_\ell^* \mathbf{v}(\omega_\ell) - 1), \tag{2.64}$$

where $\lambda \neq 0$. If the derivative with respect to the vector $\mathbf{h}_\ell$ is set to zero we obtain the equation

$$\mathbf{R}_I \mathbf{h}_{\ell,\text{opt}} = \lambda \mathbf{v}(\omega_\ell) \quad \Leftrightarrow \quad \mathbf{h}_{\ell,\text{opt}} = \lambda \mathbf{R}_I^{-1} \mathbf{v}(\omega_\ell), \tag{2.65}$$

where the correlation matrix $\mathbf{R}_I$ is assumed to be nonsingular (this is the case whenever $\mathbf{R}_I$ is the symmetric positive definite Toeplitz matrix as autocorrelation matrix is). Let us multiply the Eq. 2.65 by the vector $\mathbf{v}(\omega_\ell)$ and apply the criterion $\mathbf{v}^*(\omega_\ell)\mathbf{h}_{\ell,\text{opt}} = 1$ then we get

$$\mathbf{v}^*(\omega_\ell)\mathbf{h}_{\ell,\text{opt}} = \lambda \mathbf{v}^*(\omega_\ell)\mathbf{R}_I^{-1}\mathbf{v}(\omega_\ell) = 1 \quad \Leftrightarrow \quad \lambda = \frac{1}{\mathbf{v}^*(\omega_\ell)\mathbf{R}_I^{-1}\mathbf{v}(\omega_\ell)} \tag{2.66}$$

If the λ is substituted in Eq. 2.65 we obtain that

$$\mathbf{h}_{\ell,\text{opt}} = \frac{\mathbf{R}_I^{-1}\mathbf{v}(\omega_\ell)}{\mathbf{v}^*(\omega_\ell)\mathbf{R}_I^{-1}\mathbf{v}(\omega_\ell)}. \tag{2.67}$$

The optimum FIR filter $\mathbf{h}_{\ell,\text{opt}}$ is obtained in similar way in [16]. Let us calculate the corresponding minimum error energy $\mathscr{E}_{\ell,\text{min}}$ by use of Eq. 2.67

$$\begin{aligned} \mathscr{E}_{\ell,\text{min}} = \mathbf{h}_{\ell,\text{opt}}^* \mathbf{R}_I \mathbf{h}_{\ell,\text{opt}} &= \frac{\mathbf{v}^*(\omega_\ell)\mathbf{R}_I^{-1}}{\mathbf{v}^*(\omega_\ell)\mathbf{R}_I^{-1}\mathbf{v}(\omega_\ell)} \mathbf{R}_I \frac{\mathbf{R}_I^{-1}\mathbf{v}(\omega_\ell)}{\mathbf{v}^*(\omega_\ell)\mathbf{R}_I^{-1}\mathbf{v}(\omega_\ell)} \\ &= \frac{1}{\mathbf{v}^*(\omega_\ell)\mathbf{R}_I^{-1}\mathbf{v}(\omega_\ell)}. \end{aligned} \tag{2.68}$$

Next we will introduce a essential property of MVDR analysis. We claim that if the error energy in Eq. 2.63 is minimised in this way, then the minimum error energy $\mathscr{E}_{\ell,\text{min}}$ is a good estimate for the original signal power spectrum at a frequency point ω_ℓ that is $P_{\text{MV}}(\omega_\ell) = \mathscr{E}_{\ell,\text{min}} \approx P(\omega_\ell)$, where P is defined like in Eq. 2.22. If we use the same notation for error energy as in Eq. 2.24 we obtain

$$P_{\text{MV}}(\omega_\ell) = \mathscr{E}_{\ell,\text{min}} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H_{\ell,\text{opt}}(e^{j\omega})|^2 P(e^{j\omega}) d\omega, \tag{2.69}$$

where $H_{\ell,\text{opt}}$ is the frequency response corresponding to the optimal filter $h_{\ell,\text{opt}}$. In the next section we will explain why this is a good estimate especially when we are studying periodic signals such as voiced speech. This analysis is based on [26].

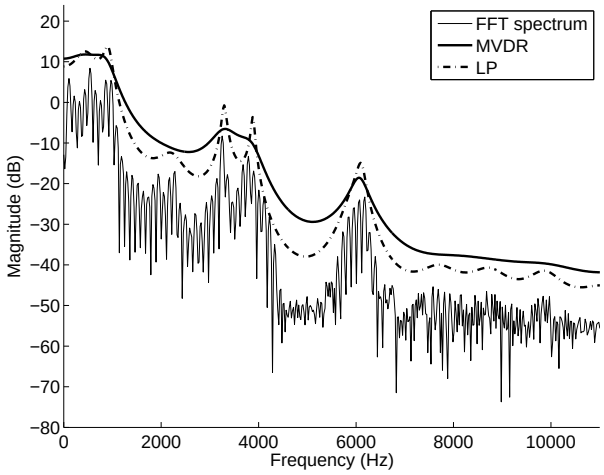

Figure 2.9: FFT spectrum of a male vowel /a/ (thin line) and the all-pole spectrum of the MVDR method (thick line) together with the LP method. The order of the all-pole models was 22 and the sampling frequency was 22050 Hz. For the sake of clarity the magnitude levels of the prediction models have been lifted 10 dB.

2.6.1 MVDR Envelope

Let $P(\omega)$ be the power spectrum of a periodic signal with L harmonics and fundamental frequency equal to ω_0, then

$$P(\omega) = 2\pi \sum_{k=1}^{L} \frac{|c_k|^2}{4} \left[\delta(\omega + k\omega_0) + \delta(\omega - k\omega_0)\right], \tag{2.70}$$

where c_k's are the amplitudes at the harmonics, and $\delta(\omega)$ is the Dirac delta function. Next we will calculate the estimate $P_{\text{MV}}(\omega_\ell)$ at the frequency point $\omega_\ell = \omega_0 \ell$. From Eqs. 2.69 and 2.70 we have

$$\begin{aligned}
P_{\text{MV}}(\omega_0\ell) &= \int_{-\pi}^{\pi} |H_{\ell,\text{opt}}(e^{j\omega})|^2 \sum_{k=1}^{L} \frac{|c_k|^2}{4} \left[\delta(\omega + k\omega_0) + \delta(\omega - k\omega_0)\right] d\omega \\
&= \sum_{k=1}^{L} \left\{ \int_{-\pi}^{\pi} |H_{\ell,\text{opt}}(e^{j\omega})|^2 \frac{|c_k|^2}{4} \delta(\omega + k\omega_0) d\omega + \int_{-\pi}^{\pi} |H_{\ell,\text{opt}}(e^{j\omega})|^2 \frac{|c_k|^2}{4} \delta(\omega - k\omega_0) d\omega \right\} \\
&= \sum_{k=1}^{L} \left\{ |H_{\ell,\text{opt}}(e^{-j\omega_0 k})|^2 \frac{|c_k|^2}{4} + |H_{\ell,\text{opt}}(e^{j\omega_0 k})|^2 \frac{|c_k|^2}{4} \right\}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{|c_\ell|^2}{4} + \frac{|c_\ell|^2}{4}|H_{\ell,\text{opt}}(e^{-j\omega_0\ell})|^2 + \sum_{k=1,k\neq\ell}^{L} \frac{|c_k|^2}{4}\left\{|H_{\ell,\text{opt}}(e^{-j\omega_0 k})|^2 + |H_{\ell,\text{opt}}(e^{j\omega_0 k})|^2\right\} \\
&= \frac{|c_\ell|^2}{4} + \frac{|c_\ell|^2}{4}|H_{\ell,\text{opt}}(e^{-j\omega_0\ell})|^2 + \sum_{k=1,k\neq\ell}^{L} \frac{|c_k|^2}{2}|H_{\ell,\text{opt}}(e^{j\omega_0 k})|^2,
\end{aligned}
\tag{2.71}
$$

where the second to last equality is due to the fact that $|H_{\ell,\text{opt}}(e^{j\omega_0\ell})|^2 = 1$ and the last equality is due to the fact that $H_{\ell,\text{opt}}(z) = H^*_{\ell,\text{opt}}(z^*), \ \forall z \in \mathbb{C}$. If we set

$$
K(\mathbf{h}_\ell) = \frac{|c_\ell|^2}{4}|H_{\ell,\text{opt}}(e^{-j\omega_0\ell})|^2 + \sum_{k=1,k\neq\ell}^{L} \frac{|c_k|^2}{2}|H_{\ell,\text{opt}}(e^{j\omega_0 k})|^2 \geq 0 \tag{2.72}
$$

and remember that $\frac{|c_\ell|^2}{4} = P(\omega_0\ell)/2\pi$ then we can rewrite Eq. 2.71 as

$$
P_{\text{MV}}(\omega_0\ell) = P(\omega_0\ell)/2\pi + K(\mathbf{h}_\ell) \geq P(\omega_0\ell)/2\pi. \tag{2.73}
$$

Note that when the error energy $\mathscr{E}_{\ell,\text{min}}$ is minimised, then $K(\mathbf{h}_\ell)$ should be very small and we obtain a good estimate.

So far we have derived the power spectrum estimate at a one frequency point and showed that it is indeed a good estimate. Let us rewrite this estimate in the matrix notations

$$
P_{\text{MV}}(\omega_\ell) = \frac{1}{\mathbf{v}^*(\omega_\ell)\mathbf{R}_I^{-1}\mathbf{v}(\omega_\ell)}. \tag{2.74}
$$

Should we derive a new filter $\mathbf{h}_e$ in order to produce the signal power spectrum estimate at different frequency point ω_e? The answer is no because the filter $\mathbf{h}_\ell$ is only conceptual and is being used only in the theoretical matter in order to clarify the method. Note that the $P_{\text{MV}}(\omega_\ell)$ does not depend on $\mathbf{h}_\ell$. So if we derive the new filter $\mathbf{h}_e$ and do all computations the power spectrum estimate is obtained

$$
P_{\text{MV}}(\omega_e) = \frac{1}{\mathbf{v}^*(\omega_e)\mathbf{R}_I^{-1}\mathbf{v}(\omega_e)}. \tag{2.75}
$$

Finally we are ready to write the pth order MVDR power spectrum for all frequencies as

$$
P_{\text{MV}}(\omega) = \frac{1}{\mathbf{v}^*(\omega)\mathbf{R}_I^{-1}\mathbf{v}(\omega)} = \frac{1}{|B(e^{j\omega})|^2}, \tag{2.76}
$$

where the $\mathbf{R}_I \in \mathbb{R}^{(p+1)\times(p+1)}$ is symmetric positive definite Toeplitz matrix. The coefficients for the MVDR prediction polynomial $B(e^{j\omega})$ can be obtained from the original LP coefficients see [26] for more details.

Chapter 3

Stability Analysis for the Non-Iterative All-Pole Models Optimised in the Time Domain

3.1 LP Method

In the autocorrelation case, where $\mathbf{R}_I$ is a symmetric Toeplitz matrix the $\mathcal{Z}$-transform $A(z)$ of the vector $\mathbf{a}$ is known to be minimum phase. This means that this inverse filter $A(z) = 1 + a_1 z^{-1} + \cdots + a_p z^{-p}$ has all its roots inside the unit circle. This is not the case in the covariance method. The minimum-phase property or equivalently the stability of the all-pole filter $A^{-1}(z)$ ensures that the corresponding impulse response is convergent. The stability property has been proved in among others [11, 29, 15]. In the following, we will prove this in a detailed manner.

Let us rewrite the Eq. 2.16 in the case where we assume only that the matrix $\mathbf{R}_I$ can be factorised as $\mathbf{R}_I = \mathbf{Y}^*\mathbf{Y}$, where the matrix $\mathbf{Y} = (\mathbf{y}_0\ \mathbf{y}_1\ \cdots\ \mathbf{y}_p) \in \mathbb{C}^{(N+p+1)\times(p+1)}$.

$$\begin{pmatrix} \mathbf{y}_0^*\mathbf{y}_0 & \mathbf{y}_0^*\mathbf{y}_1 & \cdots & \mathbf{y}_0^*\mathbf{y}_p \\ \mathbf{y}_1^*\mathbf{y}_0 & \mathbf{y}_1^*\mathbf{y}_1 & \cdots & \mathbf{y}_1^*\mathbf{y}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}_p^*\mathbf{y}_0 & \mathbf{y}_p^*\mathbf{y}_1 & \cdots & \mathbf{y}_p^*\mathbf{y}_p \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.1}$$

In other words, the coefficients $a_1, \ldots, a_p$ are unique solutions to the group of equations

$$\mathbf{y}_k \perp (\mathbf{y}_0 + a_1\mathbf{y}_1 + \ldots a_p\mathbf{y}_p), \quad \forall k = 1, \ldots, p \tag{3.2}$$

where the $\mathbf{x}\perp\mathbf{y}$ means that the vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal ($\mathbf{x}^*\mathbf{y} = 0$).
Before deriving any results concerning the locations of the roots of the prediction polynomial we define the vector spaces $\mathscr{K} := span\{\mathbf{y}_1, \ldots, \mathbf{y}_p\}$ and $\widetilde{\mathscr{K}} := span\{\mathbf{y}_0, \ldots, \mathbf{y}_{p-1}\}$ and linear projection operator $\mathcal{P}_\mathbf{g}$ on $\mathscr{K}$ such that if vector $\mathbf{g} = (g_1 \ \ldots \ g_p)$ then

$$\begin{aligned} \mathcal{P}_\mathbf{g}\mathbf{y}_0 &= -(g_1\mathbf{y}_1 + \cdots + g_p\mathbf{y}_p) \\ \mathcal{P}_\mathbf{g}\mathbf{y}_k &= \mathbf{y}_k, \ \ k = 1, \ldots, p, \end{aligned} \tag{3.3}$$

There is a very important property concerning the linear projection operator and before going further let us consider more closely this property.

Lemma 3.1.1 *If the coefficients $g_1, \ldots, g_p$ are solutions to the group of equations defined in Eq. 3.2 then the linear projection operator $\mathcal{P}_\mathbf{g}$, where $\mathbf{g} = (g_1, \ldots, g_p)$, has the property $\mathbf{v}^*\mathcal{P}_\mathbf{g}\mathbf{u} = \mathbf{v}^*\mathbf{u} \ \ \forall\mathbf{v} \in \mathscr{K}, \ \mathbf{u} \in \widetilde{\mathscr{K}}$.*

Proof Assume that $g_1, \ldots, g_p$ are solutions to the group of equations defined in Eq. 3.2. Then

$$\begin{aligned} &\mathbf{y}_k\perp(\mathbf{y}_0 + g_1\mathbf{y}_1 + \ldots g_p\mathbf{y}_p), \ \ \forall k = 1, \ldots, p \ \ \Leftrightarrow \\ &\mathbf{y}_k^*\mathbf{y}_0 + \mathbf{y}_k^*(g_1\mathbf{y}_1 + \ldots + g_p\mathbf{y}_p) = 0, \ \ \forall k = 1, \ldots, p \ \ \Leftrightarrow \\ &\mathbf{y}_k^*\mathbf{y}_0 = -\mathbf{y}_k^*(g_1\mathbf{y}_1 + \ldots + g_p\mathbf{y}_p) = \mathbf{y}_k^*\mathcal{P}_\mathbf{g}\mathbf{y}_0, \ \ \forall k = 1, \ldots, p \end{aligned} \tag{3.4}$$

Take $\mathbf{v} \in \mathscr{K} \ \Leftrightarrow \mathbf{v} = \sum_{i=1}^p d_i\mathbf{y}_i$ and $\mathbf{u} \in \widetilde{\mathscr{K}} \ \Leftrightarrow \mathbf{u} = \sum_{i=1}^p c_i\mathbf{y}_{i-1}$ then we can write

$$\mathbf{v}^*\mathcal{P}_\mathbf{g}\mathbf{u} = \left(\sum_{i=1}^p d_i^*\mathbf{y}_i^*\right)\mathcal{P}_\mathbf{g}\left(\sum_{i=1}^p c_i\mathbf{y}_{i-1}\right) = \mathbf{v}^*\mathbf{u}.$$

The last equality is due to the property derived in Eq. 3.4 and the basic property of the projection operator $\mathbf{y}_k^*\mathcal{P}_\mathbf{g}\mathbf{y}_j = \mathbf{y}_k^*\mathbf{y}_j \ \ \forall k, j = 1, \ldots, p$.
❑

We are ready to derive the theorem concerning the locations of the roots of the inverse filter $A(z)$.

Theorem 3.1.2 *Let the vector $\mathbf{a} = (1 \ a_1 \ \cdots \ a_p)^T$ be the solution to Eq. 3.1. Further, assume that $\exists\mathbf{M} \in \mathbb{C}^{(N+p)\times(N+p)}$ such that the vectors $\mathbf{y}_k$ from Eq. 3.1 have the relation $\mathbf{y}_k = \mathbf{M}\mathbf{y}_{k+1}, \ k = 0, \ldots, p-1$. Then the zeros of the $\mathcal{Z}$-transform $A(z)$ of the vector $\mathbf{a}$ belong to the numerical range[1] of the matrix $\mathbf{M}$ denoted as $\mathcal{F}(\mathbf{M})$.*

[1] For definition, see Appendix A

Proof Let us define the linear operator $\mathscr{A} : \mathscr{K} \to \mathscr{K}$ as

$$\mathscr{A}\mathbf{x} = \mathcal{P}_{\mathbf{g}}\mathbf{M}\mathbf{x}, \qquad \forall \mathbf{x} \in \mathscr{K}. \tag{3.5}$$

Where $\mathcal{P}_{\mathbf{g}}$ is the linear projection operator defined as in Eq. 3.3, where $\mathbf{g} = (a_1 \ \cdots \ a_p)$. Notice that the choosing $\mathbf{g}$ in this way assures operator $\mathcal{P}_{\mathbf{g}}$ has the property of Lemma 3.1.1. Next, take $\mathbf{v} \in \mathscr{K} \Leftrightarrow \mathbf{v} = \sum_{k=1}^{p} \xi_k \mathbf{y}_k$. we have

$$\begin{aligned} \mathscr{A}\mathbf{v} &= \mathcal{P}_{\mathbf{g}}\mathbf{M}(\sum_{k=1}^{p} \xi_k \mathbf{y}_k) = \mathcal{P}_{\mathbf{g}}(\xi_1 \mathbf{y}_0 + \sum_{k=2}^{p} \xi_k \mathbf{y}_{k-1}) \\ &= -\xi_1(a_1\mathbf{y}_1 + \cdots + a_p\mathbf{y}_p) + \sum_{k=2}^{p} \xi_k \mathbf{y}_{k-1} = \sum_{k=1}^{p} \pi_k \mathbf{y}_k, \end{aligned} \tag{3.6}$$

where the property $\mathbf{y}_k = \mathbf{M}\mathbf{y}_{k+1},\ k = 0, \ldots, p-1$ and Eq. 3.3 have been used. From Eq. 3.6 the coordinate vector $\boldsymbol{\pi} = (\pi_1 \cdots \pi_p)^T$ can be calculated as

$$\boldsymbol{\pi} = \begin{pmatrix} -a_1 & & & \\ -a_2 & & \mathbf{I}_{(p-1)\times(p-1)} & \\ \vdots & & & \\ -a_p & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_p \end{pmatrix} = \mathbf{C}\boldsymbol{\xi}, \tag{3.7}$$

The matrix $\mathbf{C}$ is the companion matrix of $A(z)$, that is, the zeros of $A(z)$ are the eigenvalues of $\mathbf{C}$. Matrix $\mathbf{C}$ is also the representation of the operator $\mathscr{A}$ with respect to the basis of space $\mathscr{K}$ which can be see from Eq. 3.7. This means that $\mathbf{C}$ and $\mathscr{A}$ have the same eigenvalues. It remains to show that eigenvalues of operator $\mathscr{A}$ belongs to the numerical range of matrix $\mathbf{M}$.
Take a normalised eigenvector $\mathbf{v}$ of operator $\mathscr{A}$, that is $\mathscr{A}\mathbf{v} = \lambda\mathbf{v}$, where $\|\mathbf{v}\|^2 = 1$ and λ denotes the corresponding eigenvalue. By simple calculation we get

$$\lambda = \lambda\|\mathbf{v}\|^2 = \mathbf{v}^*\lambda\mathbf{v} = \mathbf{v}^*\mathscr{A}\mathbf{v} = \mathbf{v}^*\mathcal{P}_{\mathbf{g}}\mathbf{M}\mathbf{v} = \mathbf{v}^*\mathbf{M}\mathbf{v} \in \mathcal{F}(\mathbf{M}) \tag{3.8}$$

where the last equality is due to the fact that because $\mathbf{v} \in \mathscr{K}$ and $\mathbf{M} : \mathscr{K} \to \widetilde{\mathscr{K}}$ we can apply the property of the operator $\mathcal{P}_{\mathbf{g}}$ from Lemma 3.1.1.
❑

Next we return to the stability properties of the LP model, where the matrix $\mathbf{R}_I$ from Eq. 2.16 is a real symmetric Toeplitz matrix. From [29], the symmetric Toeplitz matrix

$\mathbf{R}_I$ can be factored in the following way $\mathbf{R}_I = \mathbf{Y}^T\mathbf{Y}$, where

$$\mathbf{Y} = \begin{pmatrix} y_0 & 0 & \cdots & 0 \\ y_1 & y_0 & \ddots & \vdots \\ \vdots & \vdots & & 0 \\ y_{p+1} & y_p & \cdots & y_0 \\ \vdots & \vdots & & \\ y_N & y_{N-1} & \cdots & y_{N-p} \\ 0 & y_N & & y_{N-p+1} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & y_N \end{pmatrix} = \begin{pmatrix} \mathbf{y}_0 & \mathbf{y}_1 & \cdots & \mathbf{y}_p \end{pmatrix} \in \mathbb{R}^{(N+p+1)\times(p+1)}. \tag{3.9}$$

The columns $\mathbf{y}_k$ of the matrix $\mathbf{Y}$ can be generated via the formula $\mathbf{y}_k = \mathbf{M}\mathbf{y}_{k+1}$, $k = 0, \ldots, p-1$, where

$$\mathbf{M} = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \omega & \mathbf{0}^T \end{pmatrix} \in \mathbb{C}^{(N+p+1)\times(N+p+1)}, \tag{3.10}$$

$\mathbf{I} \in \mathbb{R}^{(N+p)\times(N+p)}$ is an identity matrix and $\mathbf{0} = (0 \cdots 0)^T \in \mathbb{R}^{N+p}$. The ω is arbitrary so it can be chosen freely.

Theorem 3.1.3 *Let the vector* $\mathbf{a}$ *be the solution to the Eq. 2.16 where* $\mathbf{R}_I$ *is a symmetric Toeplitz matrix and let us choose the corresponding matrix* $\mathbf{M}$ *in Eq. 3.10 such that* $\omega = e^{i\phi}$ *where* $\phi \in [0, 2\pi]$. *Then* $A(z)$ *has all its roots inside the circle with centre at the origin and radius equal to* $\cos(\frac{\pi}{N+p+1})$.

Proof In view of Theorem 3.1.2 it is clear that the roots of the $A(z)$ belong to the numerical range of the matrix $\mathbf{M}$. The numerical range of matrix $\mathbf{M}$ coincides with the convex hull[2] of its eigenvalues because with $\omega = e^{i\phi}$ the matrix $\mathbf{M}$ become unitary, that is $\mathbf{M}^*\mathbf{M} = \mathbf{M}\mathbf{M}^* = \mathbf{I}$. The characteristic polynomial[3] $p_{\mathbf{M}}(x)$ of matrix $\mathbf{M}$ is $p_{\mathbf{M}}(x) = x^{N+p+1} - e^{i\phi}$. This can be seen by computing $det(x\mathbf{I} - \mathbf{M})$ by a Laplace cofactor expansion [17] along the first column. That is why the eigenvalues ψ_k (the roots of characteristic polynomial) are

$$\psi_k = e^{(2\pi k+\phi)i/(N+p+1)} \quad k = 0, \ldots, N+p. \tag{3.11}$$

When the convex hull is composed from the eigenvalues ψ_k, one observes that it has the geometrical shape of polygon. The roots of $A(z)$ are located inside the $(N+p+1)$-polygon

[2,3] For definition, see Appendix A

G^{ϕ}_{N+p+1} with centre at the origin. Let us remember that $\phi \in [0, 2\pi]$ was chosen arbitrary. If a different ω is chosen such as $\omega = e^{i(\phi+\epsilon)}$ we get the same results from Theorem 3.1.2 but a different polygon $G^{\phi+\epsilon}_{N+p+1}$. This polygon is the polygon G^{ϕ}_{N+p+1} rotated an amount of ϵ in the complex plane. The roots of polynomial $A(z)$ are thus located inside the area G_{N+p+1} defined as

$$G_{N+p+1} = \bigcap_{\phi\in[0,2\pi]} G^{\phi}_{N+p+1}, \tag{3.12}$$

see Fig. 3.1. By calculating the intersection points for rotated polygons we obtain that the

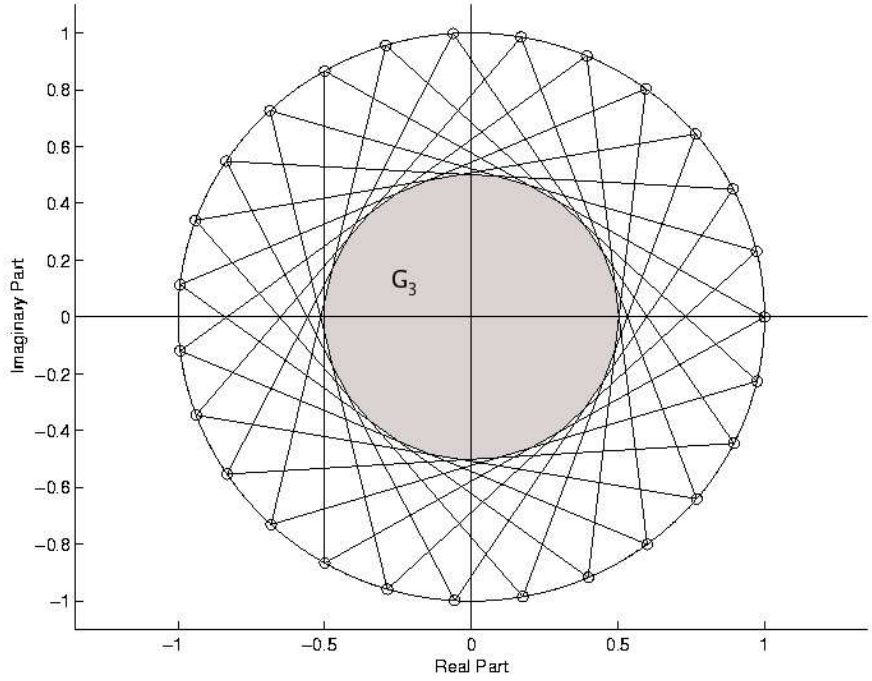

Figure 3.1: Example of the intersection area for rotated G^{ϕ}_{3} polygons.

radius ρ of G_{N+p+1} is equal to $\rho = \cos(\frac{\pi}{N+p+1})$. This concludes the proof.
❑

3.2 WLPC Method

The stability criterion of the WLPC method, in the case of autocorrelation, is similar to the LP method. In the case of autocorrelation method ($I := \{0, \ldots, N+p\}$) the left-hand side

in Eq. 2.29 can be separated in a similar way as in Sec. 3.1

$$\widetilde{\mathbf{R}}_I = (\mathbf{WY})^T(\mathbf{WY}) = \widetilde{\mathbf{Y}}^T\widetilde{\mathbf{Y}} \tag{3.13}$$

where $\mathbf{Y}$ is defined like in Eq. 3.9 and the matrix $\mathbf{W}$ is a diagonal matrix

$$\mathbf{W} = \begin{pmatrix} \sqrt{w_0} & 0 & \cdots & 0 \\ 0 & \sqrt{w_1} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \sqrt{w_{N+p}} \end{pmatrix}. \tag{3.14}$$

In this case the matrix $\widetilde{\mathbf{R}}_I$ become a symmetric matrix without Toeplitz structure. Let us consider the stability properties of the WLPC method in a similar way as in [10].

From Eq. 3.13 one observes that the columns $\tilde{\mathbf{y}}_k$ of the matrix $\widetilde{\mathbf{Y}} = \mathbf{WY}$ can be generated via the formula

$$\tilde{\mathbf{y}}_k = \widetilde{\mathbf{M}}\tilde{\mathbf{y}}_{k+1} \quad k = 0, 1, \ldots, p, \tag{3.15}$$

where

$$\widetilde{\mathbf{M}} = \begin{pmatrix} 0 & \sqrt{w_0/w_1} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{w_1/w_2} & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{w_{N+p-1}/w_{N+p}} \\ \omega & 0 & \cdots & 0 & 0 \end{pmatrix} \in \mathbb{C}^{(N+p+1)\times(N+p+1)}, \tag{3.16}$$

For Eq. 3.15, we have from theorem 3.1.2, that the zeros of the $\mathcal{Z}$-transform of the vector $\mathbf{a}$ solving Eq. 2.29 belong to the numerical range of the matrix $\widetilde{\mathbf{M}}$. It remains to derive the numerical range of the matrix $\widetilde{\mathbf{M}}$.

It has been shown [10] that the numerical range $\mathcal{F}(\widetilde{\mathbf{M}})$, where $\omega = 0$, is located inside a circle with centre at the origin and with radius $\rho = \frac{1}{2}\max_n\{\sqrt{w_n/w_{n+1}} + \sqrt{w_{n+1}/w_{n+2}}\}$, for $n = 0, 1, \ldots, N+p-2$. Let us introduce a new stability region with respect to weights w_n and a length of the data sequence. Denote the algebra of bounded linear operators[4] on the Hilbert space $\mathscr{H}$ by $B(\mathscr{H})$.

[4]For definition, see Appendix A.

Theorem 3.2.1 *Let $\mathscr{H}$ be a complex Hilbert space and let $\widetilde{\mathbf{M}} \in B(\mathscr{H})$ be a nilpotent operator*[5] *with power of nilpotency n. The numerical range $\mathcal{F}(\widetilde{\mathbf{M}})$ is a circle (open or closed) with centre at the origin and radius ρ not exceeding $\|\widetilde{\mathbf{M}}\| \cos(\frac{\pi}{n+1})$.*

Proof The proof can be found in [20].
❑

Let us apply Theorem 3.2.1 to the case where $\widetilde{\mathbf{M}}$ is defined as in Eq. 3.16 where $\omega = 0$. Then the $\widetilde{\mathbf{M}}$ is a nilpotent operator with power of nilpotency $n = N + p + 1$. The norm of the Hilbert space for the matrix $\widetilde{\mathbf{M}}$ is clearly equal to $\|\widetilde{\mathbf{M}}\| = \max_n\{\sqrt{w_n/w_{n+1}}\}$, for $n = 0, \ldots, N + p - 1$. In view of Theorems 3.1.2 and 3.2.1 we get

Theorem 3.2.2 *The zeros of the inverse filter of the weighted linear prediction model are located inside a circle with centre at the origin and with radius*

$$\rho = \max_n\{\sqrt{w_n/w_{n+1}}\} \cos(\frac{\pi}{N+p+1}), \quad n = 0, \ldots, N + p - 1.$$

Note that by choosing the weights $w_n = 1,\ \forall n = 0, \ldots, N + p$, one gets the same stability region as the LP method has ($\rho = \cos(\frac{\pi}{N+p+1})$).

3.3 WLSP Method

It has been proved that the WLSP polynomial $D(z)$ is minimum phase [8] but for the sake of completeness let us introduce a different kind of proof. In Sec. 2.4, the vector $\mathbf{d}$ in Eq. 2.44 was defined using the zero extended vector $(\mathbf{a}\ \ 0)^T$. Let us define the vector $\mathbf{d}$ in the different manner without using the extended vector (this kind of definition is similar to the concept of the immittance spectral pairs (ISP) from [7])

$$\begin{aligned} \mathbf{p} &= \mathbf{a} + \mathbf{J}\mathbf{a} \\ \mathbf{q} &= \mathbf{a} - \mathbf{J}\mathbf{a} \end{aligned} \tag{3.17}$$

Then the vector $\mathbf{d}$ is defined as in Eq. 2.46 by using the vectors $\mathbf{p}$, $\mathbf{q}$ from Eq. 3.17. The polynomial $D(z)$ denotes the $\mathcal{Z}$-transform of the vector $\mathbf{d}$. The matrix $\mathbf{J}$ is defined as in Eq. 2.45. It is easy to see that proving the stability in this situation is equivalent to the stability of the original inverse filter in Sec. 2.4.

[5] Matrix $\mathbf{A}$ is nilpotent with power of nilpotency n if it is the smallest integer such as $\mathbf{A}^n = 0$.

If the vector $\mathbf{d}$ is multiplied from left by the positive definite symmetric Toeplitz matrix $\mathbf{R}_I$ from Eq. 2.16 we get

$$\begin{aligned}\mathbf{R}_I\mathbf{d} &= \mathbf{R}_I[\lambda\mathbf{p} + (1-\lambda)\mathbf{q}] = \mathbf{R}_I[\lambda(\mathbf{a}+\mathbf{J}\mathbf{a}) + (1-\lambda)(\mathbf{a}-\mathbf{J}\mathbf{a})] \\ &= \mathbf{R}_I[\mathbf{a} + (2\lambda-1)\mathbf{J}\mathbf{a}] = \mathbf{R}_I[\mathbf{a} + \alpha\mathbf{J}\mathbf{a}],\end{aligned} \tag{3.18}$$

where $\alpha = 2\lambda - 1$. Remember that vector $\mathbf{a}$ was the solution to the Eq. 2.16. By using the fact that $\mathbf{R}_I\mathbf{J} = \mathbf{J}\mathbf{R}_I$ if $\mathbf{R}_I$ is centrosymmetric matrix (symmetric Toeplitz matrix is centrosymmetric) we can write

$$\mathbf{R}_I\mathbf{d} = \sigma^2(\mathbf{I} + \alpha\mathbf{J})\mathbb{1}, \tag{3.19}$$

or equivalently

$$\mathbf{R}_I\mathbf{d} = \sigma^2\mathbf{T}\mathbb{1}, \tag{3.20}$$

where $\mathbf{T} = \mathbf{I} + \alpha\mathbf{J}$ and the unit vector $\mathbb{1} = (1\ 0\cdots 0)^T$.

Before going further we have the property concerning the positive definite symmetric Toeplitz matrix $\mathbf{R}_I$

Theorem 3.3.1 *The $\mathcal{Z}$-transform of the vector* $\mathbf{a}$ *solving Eq. 2.16 is in minimum phase iff* $\mathbf{R}_I$ *is positive definite symmetric Toeplitz matrix.*

Proof The proof can be found in [15].
❑

We are ready to prove the stability of the WLSP method.

Theorem 3.3.2 *The WLSP polynomial* $D(z) = \lambda P(z) + (1-\lambda)Q(z)$ *is minimum phase iff* $\lambda \in (0, 1)$.

Proof ('⇐') After some straightforward calculations the inverse $\mathbf{T}^{-1}$ of a matrix $\mathbf{T}$ in Eq. 3.20 is

$$\mathbf{T}^{-1} = \frac{1}{1-\alpha^2}(\mathbf{I} - \alpha\mathbf{J}). \tag{3.21}$$

Now Eq. 3.20 can be written as

$$\frac{1}{1-\alpha^2}(\mathbf{I} - \alpha\mathbf{J})\mathbf{R}_I\mathbf{d} = \frac{1}{1-\alpha^2}\mathbf{R}_I(\mathbf{I} - \alpha\mathbf{J})\mathbf{d} = \mathbf{R}_I\tilde{\mathbf{d}} = \sigma^2\mathbb{1}, \tag{3.22}$$

where

$$\tilde{\mathbf{d}} = \frac{1}{1-\alpha^2}(\mathbf{I} - \alpha\mathbf{J})\mathbf{d} \tag{3.23}$$

and the first equality is due to the fact that $\mathbf{R}_I$ and $(\mathbf{I} - \alpha\mathbf{J})$ commutate. From Theorem 3.3.1 we find that $\mathcal{Z}$-transform of the vector $\tilde{\mathbf{d}}$ is minimum phase. It remains to prove that the $\mathcal{Z}$-transform of the vector $\mathbf{d}$ is also in the minimum phase. If the $\mathcal{Z}$-transform is applied to Eq. 3.23 we obtain

$$(1-\alpha^2)\widetilde{D}(z) = D(z) - \alpha z^{-p} D^*(\frac{1}{z^*}), \tag{3.24}$$

where the polynomial $D(z)$ is of order p. Let us divide Eq. 3.24 by $D(z)$ and define a new rational function G of z as

$$G(z) = (1-\alpha^2)\frac{\widetilde{D}(z)}{D(z)} = 1 - \alpha\frac{z^{-p}D^*(\frac{1}{z^*})}{D(z)}. \tag{3.25}$$

We are ready to use the well known *principle of the argument*[6] from complex analysis [31], [15]. This principle is explained in the same manner as in [15] that is if one is given a rational function f of z and let ζ be a simple closed curve in the z plane. As the path ζ is traversed in a counterclockwise direction, a closed curve is generated in the $f(z)$ plane that encircles the origin $N_z - N_p$ times in a counterclockwise direction where N_z is the number of zeros inside ζ and N_p is the number of poles inside ζ.

Let us choose the simple closed curve ζ to be the unit circle. Then the polynomial $\widetilde{D}(z)$ has p roots inside the ζ and p poles at $z = 0$. The polynomial $D(z)$ has also p poles at $z = 0$ and let us assume that it has k roots outside the unit circle. Then it is obvious that the polynomial $G(z)$ in Eq. 3.25 has $N_z = p$ roots and $N_p = p - k$ poles and this means that the curve in $G(z)$ plain encircles the origin $p - (p - k) = k$ times in a counterclockwise direction. Now since the contour ζ is the unit circle ($z = e^{j\omega}$) from Eq. 3.25 one gets

$$|G(z) - 1|_\zeta = |\alpha| \left| \frac{z^{-p}D^*(\frac{1}{z^*})}{D(z)} \right|_\zeta = |\alpha| \left| \frac{e^{-jp\omega}D^*(e^{j\omega})}{D(e^{j\omega})} \right|_\zeta = |\alpha|. \tag{3.26}$$

The assumption was that $\lambda \in (0,1) \Leftrightarrow \alpha \in (-1,1)$. This means that the closed curve generated in the $G(z)$ plane does not encircles the origin see Fig. 3.2. This implies that k must be equal to zero and hence, $D(z)$ has all its roots inside the unit circle.

('$\Rightarrow$') If the roots of the (inverse filter) polynomial $D(z)$ are inside the unit circle then

[6] For definition, see Appendix A.

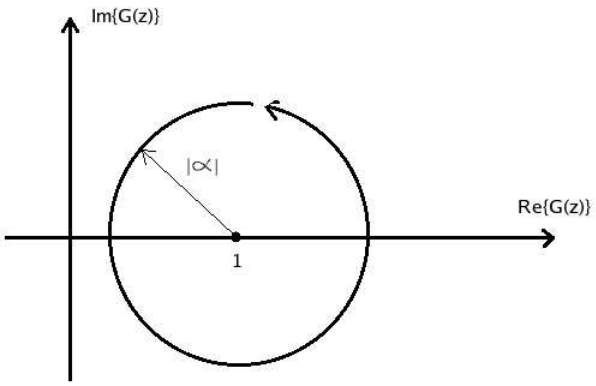

Figure 3.2: $G(z)$ traces out a curve which is a circle of radius $|\alpha|$ that is centred at $z = 1$.

k is equal to zero. This implies that the closed curve generated in the $G(z)$ plane, from Fig. 3.2, does not encircles the origin. From this property we obtain that $\alpha \in (-1, 1)$ which concludes the proof.

❑

Chapter 4

Objective Assessment in All-Pole Modelling of Speech

The quality of the spectral modelling can be understood in many different ways [18, 21, 25]. In this thesis, we are especially interested in behaviour of all-pole modelling in the presence of the uncorrelated background noise. In practise, if two different all-pole spectra are compared, the most convenient way of doing this is to use spectral distortion measures. There are several objective spectral-distortion measures in literature such as *RMS log spectral measure*, *cepstral distance measure*, *cosh measure*, *likelihood ratios* etc. (see [30, 3]). In this thesis the log spectral distance $\mathrm{SD}_{\wp}$ for the normalised all-pole spectrum is used in order to evaluate the effectiveness of the all-pole models to model the voiced speech spectrum in presence of uncorrelated noise. Moreover, if we consider the perceptually relevant measures in detail, the shifting of the first two formants as a function of the signal to noise ration should be examined. The reason behind this lies in the fact that the locations of the first two formants determines which vocal is in question. In this work, we will consider this measure and therefore explain the concept of signal to noise ratio and the concept of vocal tract resonances called formants.

4.1 Log Spectral Distance

The SD_2 measure is based on the well known Itakura-Saito distortion measures as we shall see later. Let us consider the general $\mathrm{SD}_{\wp}$ measure in detailed manner.
Take two power spectra $P_1(\omega)$ and $P_2(\omega)$, and let the energies of the original signals (p_1 and p_2) be equal. Define the difference function of these two logarithmic spectra to be

$$V(\omega) = \log_{10} P_1(\omega) - \log_{10} P_2(\omega). \tag{4.1}$$

Continuous Case

In the continuous case the $\mathrm{SD}_{\wp}$ measure is defined as

$$\mathrm{SD}_{\wp} = \sqrt[\wp]{\frac{1}{2\pi}\int_{-\pi}^{\pi} |V(\omega)|^{\wp} d\omega}. \tag{4.2}$$

In mathematics this is called the $L_{\wp}$ norm (see [31] for more detail). What is the connection between $L_{\wp}$ norms for a different $\wp$? One observes that when $\wp$ is increased the effects of the large errors are more heavily weighted and the lower errors effects on the $\mathrm{SD}_{\wp}$ measure is decreased. If $\wp$ approaches infinity, only the maximum value of $|V(\omega)|$ defines the $\mathrm{SD}_{\wp}$ measure. In this thesis, the discrete version of SD_2 measure is used as we see in the next section. The connection between the Itakura-Saito distortion measures from Eq. 2.54 and SD_2 can be seen by rewriting Eq. 2.54 and using the Taylor series [1] of the exponential function about the point equal to zero (which is known as a Maclaurin series of the exponential function).

$$\begin{aligned}\mu(E) &= \int_{-\pi}^{\pi} \left[e^{V(\omega)} - V(\omega) - 1\right] \frac{d\omega}{2\pi} = \int_{-\pi}^{\pi} \left[\sum_{k=0}^{\infty} \frac{V(\omega)^k}{k!} - V(\omega) - 1\right] \frac{d\omega}{2\pi} \\ &\approx \int_{-\pi}^{\pi} \left[1 + V(\omega) + \frac{V(\omega)^2}{2} - V(\omega) - 1\right] \frac{d\omega}{2\pi} = \frac{1}{2}\mathrm{SD}_2^2,\end{aligned} \tag{4.3}$$

where the approximation $\sum_{k=0}^{\infty} \frac{V(\omega)^k}{k!} \approx 1 + V(\omega) + V(\omega)^2/2$ is fairly good if $|V(\omega)| \ll 1, \ \forall\, \omega \in [-\pi, \pi]$.

Discrete Case

The $\mathrm{SD}_{\wp}$ measure can be defined also in the discrete case. In this work we are interested in measuring the difference between two all-pole spectra. Take two different all-pole models $A_1(z)$ and $A_2(z)$. In order to compare these models using the $\mathrm{SD}_{\wp}$ measure, one must normalise the gains of the all-pole filters to be equal. The power spectra for these models is defined as in Eq. 2.23

$$P_i(\omega) = \frac{\sigma^2}{|A_i(e^{j\omega})|^2} \qquad i = 1, 2. \tag{4.4}$$

Next, the difference function can be calculated as in Eq. 4.1. Notice that one typically computes the power spectra using the FFT algorithm and thus the spectra are discrete.

[1] Taylor series is an expansion of a real function f about a point x_0 : $f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$.

Therefore, the discrete $\mathrm{SD}_{\wp}$ measure must be used. That is

$$\mathrm{SD}_{\wp} = \sqrt[\wp]{\frac{2}{f_s}\sum_{f=0}^{f_s/2}|V(2\pi f)|^{\wp}}, \tag{4.5}$$

where f_s is the sampling frequency.

The greatest advantage of the SD_2 measure is the fact that the contributions to the total difference in Eq. 4.1 are equally important whether $P_1(\omega) < P_2(\omega)$ or vice versa and that there is no error cancellation property because of the square in Eq. 4.5 and 4.2 when the $\wp$ is chosen to be equal to 2. SD_2 is also a perceptually relevant distortion measure. That is because the loudness of a signal is approximately logarithmic. It is also very appealing that when one compares two slightly different all-pole spectra, the largest values of $|V(\omega)|$ occurs when there is lot of variations in the formant frequencies of the all-pole models in question.

4.2 Formant Shifting as a Function of Signal to Noise Ratio

In this thesis, we will measure formant shifting as a function of signal to noise ratio (SNR). This kind of a measure gives us straightforward information about the quality of the present all pole model if it is used in order to synthesise or re-synthesise speech. This property will be explained later and it is based on the concept of the formant map. Let us first explain the concept of the SNR and the vocal tract resonance called formant. Later on we will show the way to obtain the formants from the spectrum in practise.

4.2.1 Signal to Noise Ratio

Consider a clean signal sequence $\{s(n)\}$ and a sequence of uncorrelated noise $\{e(n)\}$, where the length of the both sequences is N. Let us assume that we have a additive model for corruption

$$x(n) = s(n) + e(n) \qquad n = 0, \ldots, N \tag{4.6}$$

Then the SNR value of the signal sequence $\{x(n)\}$ is defined as

$$\mathrm{SNR}(x) = 10\log_{10}\left[\frac{\sum_{n=1}^{N} s(n)^2}{\sum_{n=1}^{N} e(n)^2}\right]. \tag{4.7}$$

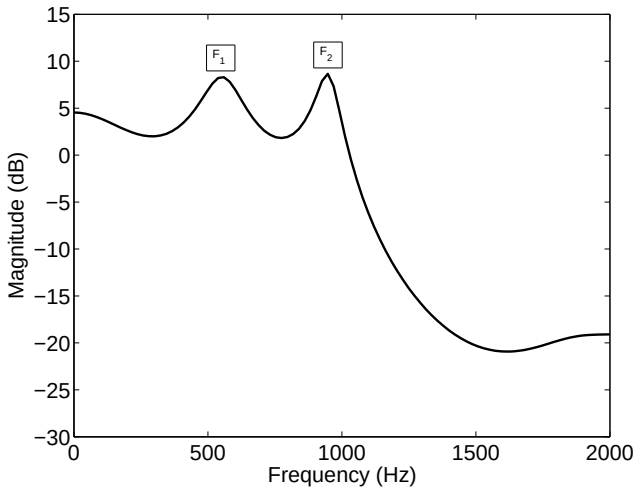

Figure 4.1: The two clear peaks of the power spectrum of the synthesise filter shows the first two formants, F_1 and F_2. All-pole model was computed from the Finnish vowel /a/ pronounced by male 3, using sampling frequency 22050 Hz and a large LP-order.

4.2.2 Vocal Tract Resonance and the Influence of Formant Shifting

In this section we consider the vocal tract resonances called the formants and then explain the effect of the formant shifting. Let us first consider the basic speech production model by rewriting the Eq. 2.20:

$$E(z) = A(z)X(z), \tag{4.8}$$

Eq. 4.8 can be written as

$$E(z)\frac{1}{A(z)} = X(z). \tag{4.9}$$

The all-pole filter $\frac{1}{A(z)}$ can be interpreted as a vocal tract filter and the clear characteristics of the filter should be an approximation of the entire vocal tract, serving as a acoustically resonant system. Eq. 4.9 can be understood as the approximation to the source-tract model by Fant [13]. In this model the all-pole model $\frac{1}{A(z)}$ models the combined effect of the *glottal flow*, *vocal tract*, and the *lip radiation*. The vocal tract can be considered a time-varying filter that prohibit the passage of sound energy at certain frequencies while allowing its progress at other frequencies. The formants are the resonant frequencies at which local energy maxima are sustained by the vocal tract and are determined, in part, by the overall physical dimensions of the vocal tract (length,volume and overall shape of the vocal tract). When different phones are produced, the physical dimensions of the vocal tract are changed

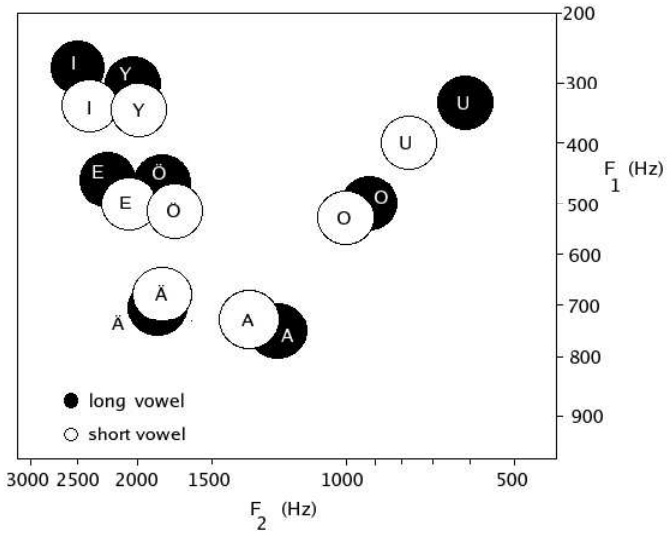

Figure 4.2: Vocal map for Finnish vowels. The information behind this figure can be found in [19].

and that is why the formants appear in different frequency positions for different vowels (see Fig. 4.2). The formant frequencies thereby correspond to the peaks in the power spectrum of the synthesis filter. This can be seen from Fig. 4.1. (For further information, the reader is referred to see [27].)

Chapter 5

Subjective Testing of All-Pole Models

There are many standards concerning subjective testing of speech. In psychoacoustics, it is very common to study and evaluate theoretical results using different kind of listening test setups see ([28, 1]). Next we will consider some aspects of subjective testing concerning the listening tests.

5.1 Listening Tests

In subjective tests, different speech samples are played to a group of listeners. The test subjects are asked to rate the quality of the speech samples. The tests are performed for a appropriate group of listeners and at the end the results are processed using tools of mathematical statistics. It is clear that the larger the number of listeners, the more reliable the results will be. This is the case because the confidence interval decreases for a large numbers of listeners. Some subjective tests use so called expert listeners. This means that before the test is being performed the group of listeners is trained to listen to certain properties of the speech samples. Usually non-expert listeners are used because, if we are considering the real applications, the system is eventually used by non-expert listeners (common people).

There are many issues affecting the results of the subjective testing. For example, if the test samples are too loud or they are clipping, the effects on the results could be remarkable. This is why the test setup should be planned very carefully and the speech samples should be normalised to the same level and the reference signals should be processed in the right manner. There are different kind of subjective tests belonging to the category of listening-only subjective tests such as *absolute category rating* (ACR), *degradation category rating* (DCR), and *diagnostic acceptability measure* (DAM). In this work, we are considering the DCR testing because it is widely used test for background noise conditions. Next we will

give a closer look in DCR testing (see [28]).

Degradation Category Rating

The DCR test uses an annoyance scale and it is suitable for evaluating good quality speech. It is also a good test for evaluating the effects of background noise conditions in speech samples. The idea is to play the uncorrupted (clean) speech sample first as a reference sample and then play the contained speech sample. The listener has to rate the degradation at the discrete DCR scale (see Appendix B Table B.1). The reader should notice that the order of speech samples (reference and corrupted speech sample) played might bias the scores. This is a basic difficulty when designing listening tests in general and it should be taking into account. In this work, the DCR test is used in the manner that listener is able to listen the sample pair as many times as he or she wants and in any order. It is well known that this kind of setting also suffers from some difficulties.

Chapter 6

Speech material and Tests Setups

In this chapter, we will consider shortly the processing of speech signals as well as the test setups of objective and subjective tests. Let us first consider the processing of speech signals.

6.1 Processing of Speech Signals

The speech materials was recorded in an anechoic chamber using a high-quality condenser microphone (Brüel&Kjær), and the data was saved onto a DAT. Next the speech samples were down-sampled at $f_s = 22050\ Hz$, using 16 bits per sample in order to transfer the vowels into a computer. The recorded speech material consists of Finnish vowel /a/ produced by five male speakers. Finally, the speech samples were processed with computer in the following way. Sustained segment with duration of $25\ ms$ located in the middle of speech sample (551 number of samples at each speech sample) are selected and used in this work. We did not use any pre-emphasise filtering techniques and from Tab. B.6 the reader can see in which all-pole methods the speech samples were windowed by the Hamming window before calculating any model parameters.

6.2 Tests Setups

6.2.1 Objective Tests

As mentioned earlier, the discrete logarithmic SD_2 measure will be calculated at the different circumstances. In order to complement these results, the first two formants, namely F_1 and F_2, are calculated at the same situations. The clean speech samples from five male were processed in the manner explained in Sec. 6.1.

First the corrupted speech samples were created. We generated for each speech sample,

Table 6.1: Table shows for which all pole models the speech samples were windowed by a Hamming window, before calculating the model parameters.

All pole model	Hamming window
LP	yes
WLPC	no
M-estimates	no
WLSP	yes
DAP	yes
MVDR	yes

seven corrupted samples with SNR $0, 5, \ldots, 30$. The corruption was done by adding Gaussian noise or the Laplacian noise as described in Eq. 4.6. In the end, we have 15 speech samples for one male speaker (totally 75 speech samples). In order to calculate SD_2 and the formants (F_1, F_2) as a function of the SNR, we calculate an all-pole model for each corrupted signal and calculate the spectral difference for each corrupted all-pole spectral envelope from the all-pole model envelope created to the clean speech sample. This is done for every male speech sample (vowel /a/) in the both cases, namely the Gaussian corrupted samples and the Laplacian corrupted samples.

In order to calculate SD_2 in every situation we use the discrete form on SD_2 from Eq. 4.5 and the calculation is performed in two cases, on the entire frequency range ($0Hz \rightarrow 11025Hz = f_s/2$) and in half range ($0Hz \rightarrow 5525Hz$).
When the formant shifting as a function of the SNR is being analysed, we have to be able to define where the first two formants are located. This has been done by calculating the 45 order LP model for the five clean speech samples and taking the first two maximum peaks from the LP spectra. The location of the peaks can be found by calculating the zero points of the derivative of the LP envelope. The same method has been used for every method at all different SNRs in order to calculate the formant shifting as a function of SNR. The calculations are performed in two different situations, namely the presence of the Gaussian noise and Laplacian noise. Reader should notice that the respective spectra are discrete so the formant location is the nearest spectral peak to the actual peak in the speech spectrum. That is why one observes from the Figs. 7.4-7.5 that for example the F_1 location for two different method can be the exactly same.

The order p of the all-pole models was 22. The LP, WLPC, WLSP, MVDR, and DAP methods were derived using the autocorrelation method. The length of the STE window

M in WLPC method was set to be 12. M-estimates (Huber and ℓ_1) uses the covariance method with robust scale estimate $\hat{r} = 1$ such that in Huber method the tuning constant c was chosen to be 1.5. We calculated only one iteration round using the covariance matrix, where the initial vector was chosen to be equal to the LP coefficient vector calculated by the covariance method.

6.2.2 Subjective Tests

The subjective tests is carried out as an listening test as mentioned earlier. In the listening tests, the speech samples (vowel /a/) of male 1 and male 3 have been used in the following manner. First the samples were processed in the same manner as in Sec. 6.2. Then we calculated the all-pole models (filters) for the same seven selected methods, using the same parameters and orders as in objective tests in Sec. 6.2. The all-pole models for both male were calculated from the original ("clean") speech samples and the corrupted versions of the same two samples. The corruption was performed using Gaussian noise at seven different signal to noise ratios $0, 5, 10, \ldots, 30$. The excitations for male 1 and male 3, were composed by means of the temporal structures of the respective clean speech samples. This means that the excitations were impulse vectors of approximate length equally to 300 ms (6000 samples). The impulses location of the impulse train were set to be the same as the peaks in the original time waveform of the speech sample in question. The excitations were filtered through the all-pole filters (reverse of the original filter) in order to synthesise the speech. The synthesised speech samples were windowed by Hanning window of length 10 ms (551 samples) and the energies were normalised to one.

Listening procedure

The listeners in this degradation category rating test were non-expert listeners but most of them were experts in many fields of speech technology. There were 10 Finnish listeners, 8 male and 2 female. Listening test were performed in a noiseless room designed for the listening purpose. Listening was done over headphones using a mono signal for a both ears. The test was performed such that only one listeners was doing the test at a time and the test took about 30 min. The test was organised in the following way:

First the listeners were instructed verbally, and then they carried through a short training session consisting on 9 sample pairs (18 speech samples in total) which were in random order. During the training session the listeners were advised to adjust the volume manually to a pleasant level. After the level was chosen the listeners were not allowed to change it during the actual test. There were 112 samples in the test and the comparison between

the reference and the corrupted sample (56 sample pair) was performed in random order such that the listeners were able to listen to both samples as many time as they want and in an arbitrary order. The reference speech sample was in every situation the synthesised speech sample calculated from the clean speech sample. The validity of each listeners were checked in the following manner. The comparison between the two same speech samples were included for both tests, the training stage and the actual test. We also included cases where the comparison of some sample pairs were repeated twice. Next, the listeners were given a discrete grade $1-5$ (see Appendix B Tab. B.2). The voting was accomplished using the mouse in order to move the scroll bar.

Chapter 7

Results

In this chapter, we will present the results of objective and subjective testing. At the end, the correlation between the subjective and objective testing is described and calculated. In both cases, the signals used were processed in the same manner explained in Chapter. 6.

7.1 Objective Results

As we can see from Figs. 7.2 and 7.3, the STE-WLPC method worked markedly better in the presence of the Gaussian and Laplacian noise than the others all-pole methods. The only methods which is almost as good as STE-WLPC method in this kind of test setups is the MVDR method. We may ask the underlying reason behind the superior of the STE-WLPC method? The answer for this is evident if we consider Fig. 7.1. From this figure it is clear that if we calculate the SNR at a time interval 10 ms $\rightarrow$ 15 ms it is worse than SNR calculated from interval 15 ms $\rightarrow$ 20 ms. The STE weight function is the one which down-weight those intervals for low SNR. From Fig. 2.5 it is easy to see that the weight functions from the L_1 and HUBER method do not have this property. The reader should notice that this holds in the particular case, when the noise is uncorrelated Gaussian and Laplacian. Note that the results for both types of noise are fairly similar. The SD_2 calculated from the entire and half frequency range gives also similar results in the both cases. The reason why the MVDR method obtains such a good results if the SD_2 measure is being considered. The reason could be because of the smoothness of the method. This means that small changes in the speech waveform do not affect the estimation of the smooth spectral envelope (see Fig. 2.9). The reason why DAP and WLSP are inferior to STE-WLPC in terms of SD_2 is explained by the fact that they do not posses the weighting property described in Sec. 2.2.

The calculation of formant shifting as a function of SNR is performed over the frequency range from 0 Hz to 2000 Hz. The calculation is stopped in the situation where only one

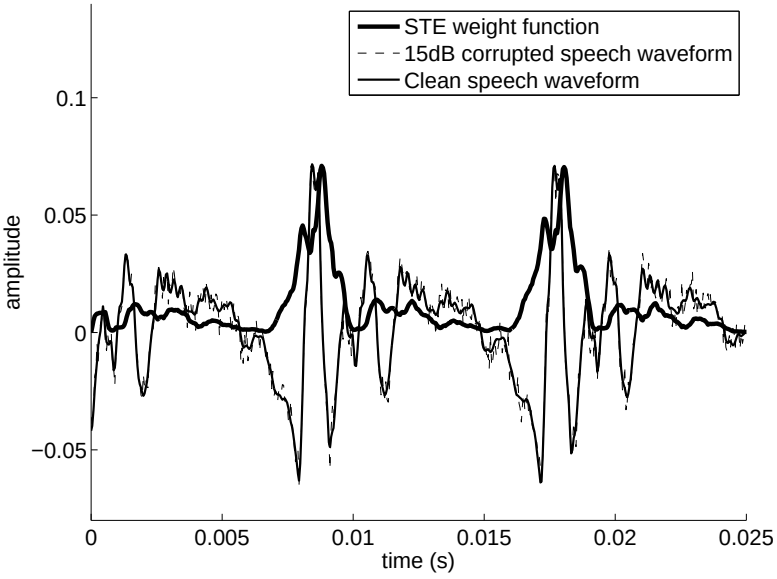

Figure 7.1: A clean time waveform of a male vowel /a/ and the same waveform corrupted at a SNR=15 dB Gaussian white noise. The STE-weight function is calculated from the corrupted speech sample and scaled to the same level as speech waveforms ($M = 12$, $k = 1$).

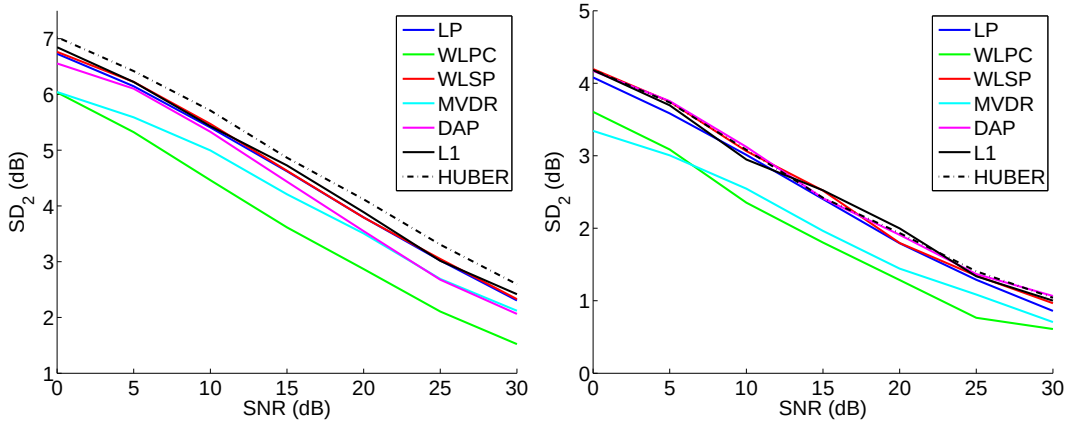

Figure 7.2: Spectral distance values between clean and corrupted all pole envelopes calculated for a vowel /a/ using SNR values from $0dB$ to $30dB$. The speech sample was corrupted by Gaussian uncorrelated noise. The values were computed as an average over five male speakers. Left panel: SD_2 was calculated using frequency range from 0Hz to 11025Hz. Right panel: SD_2 was calculated using frequency range from 0Hz to 5525Hz.

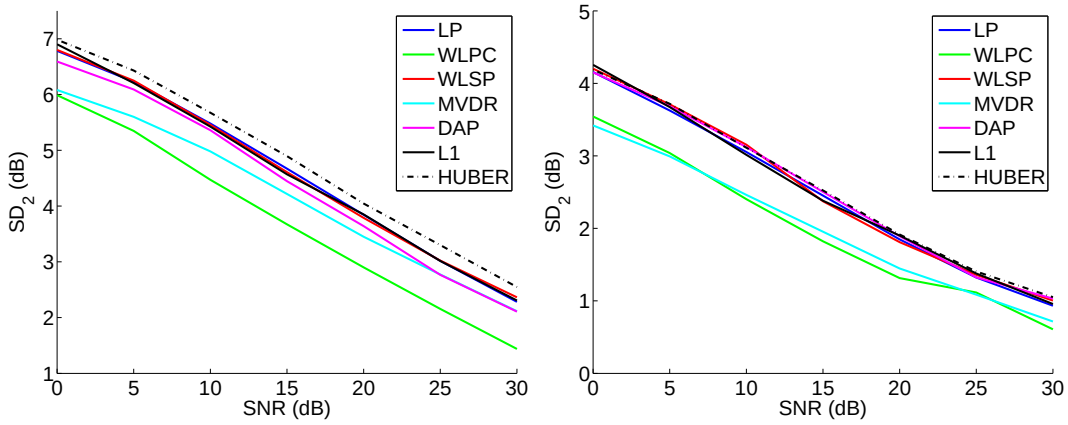

Figure 7.3: Spectral distance values between clean and corrupted all pole envelopes calculated for a vowel /a/ using SNR values from $0dB$ to $30dB$. The speech sample was corrupted by Laplacian uncorrelated noise. The values were computed as an average over five male speakers. Left panel: SD_2 was calculated using frequency range from 0Hz to 11025Hz. Right panel: SD_2 was calculated using frequency range from 0Hz to 5525Hz.

formant is located in the frequency range in question. This means that the first two formants are combined to one or they have moved out of the frequency range. This can be seen from Figs. 7.4 and 7.5 where the discrete function (at a specific colour corresponding to the all-pole method in question) as a function of the SNR stops suddenly when the SNR is near to $0\ dB$. The reader should notice that the MVDR method can not be found from the pictures. The reason behind this is obvious when studying Fig. 2.9. It is clear that the spectral envelope produced by the MVDR method is so smooth that even at a clear speech only the one peak value can be located at a frequency range in question. It is also interesting to note that in both Figs. 7.4 and 7.5, both formants are travelling toward zero. The reason behind this is not clear and should be studied in future.

7.2 Subjective Results

From Figs. 7.6 and 7.7 it is clear that the WLPC method is considerably better than other all-pole methods. The MVDR method do not cope as well as in the SD_2 tests because of the spectral smoothness. It is difficult to rate the sound quality at a DMOS scale if the vowel /a/ does not sound good even at a clean synthesised speech. We can conclude that only the WLPC method is clearly differing from other methods. Altogether the results derived from the subjective test are fairly similar to the objective tests, the SD_2 measure particularly.

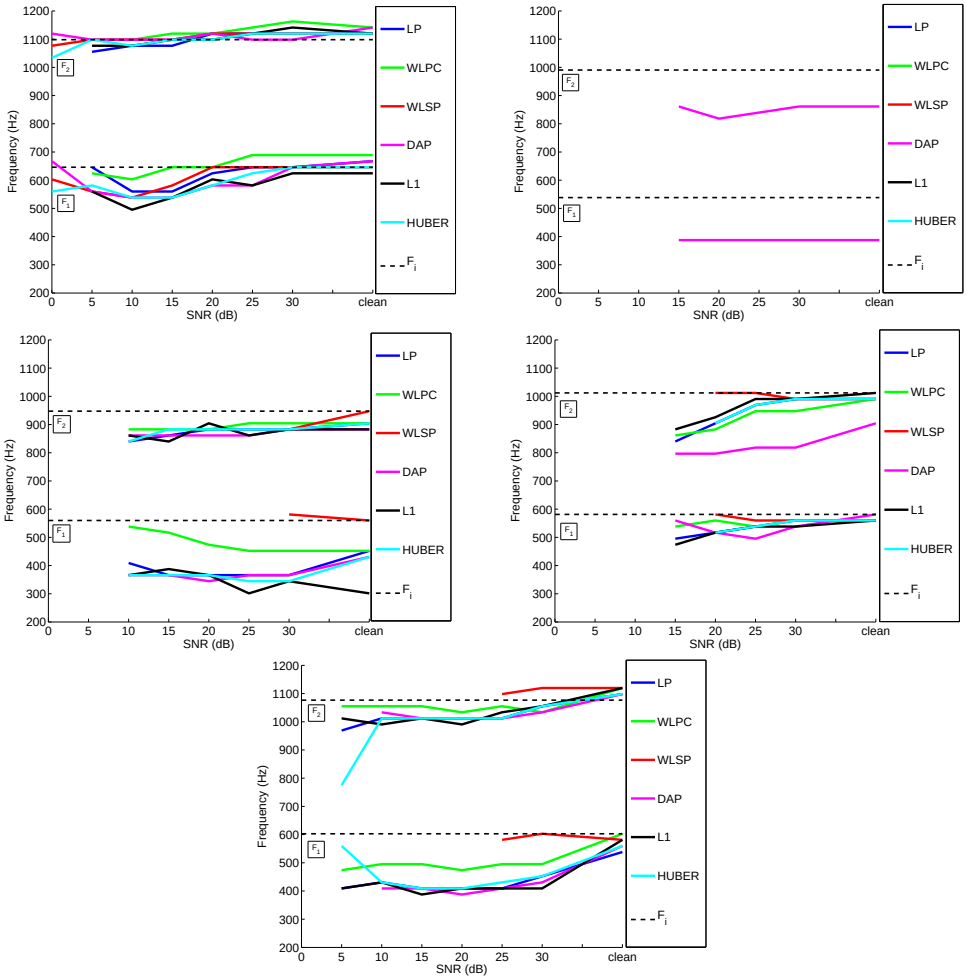

Figure 7.4: First and second formant ($F_1,\ F_2$) locations (Hz) as a function of SNR (dB). The analyses were using all seven selected all-pole methods. Speech was corrupted by Gaussian additive noise. The "correct" formants (dashed lines) were derived by using 45 order LP model matched to the underlying vowel. Results were given separately for all five male speakers.

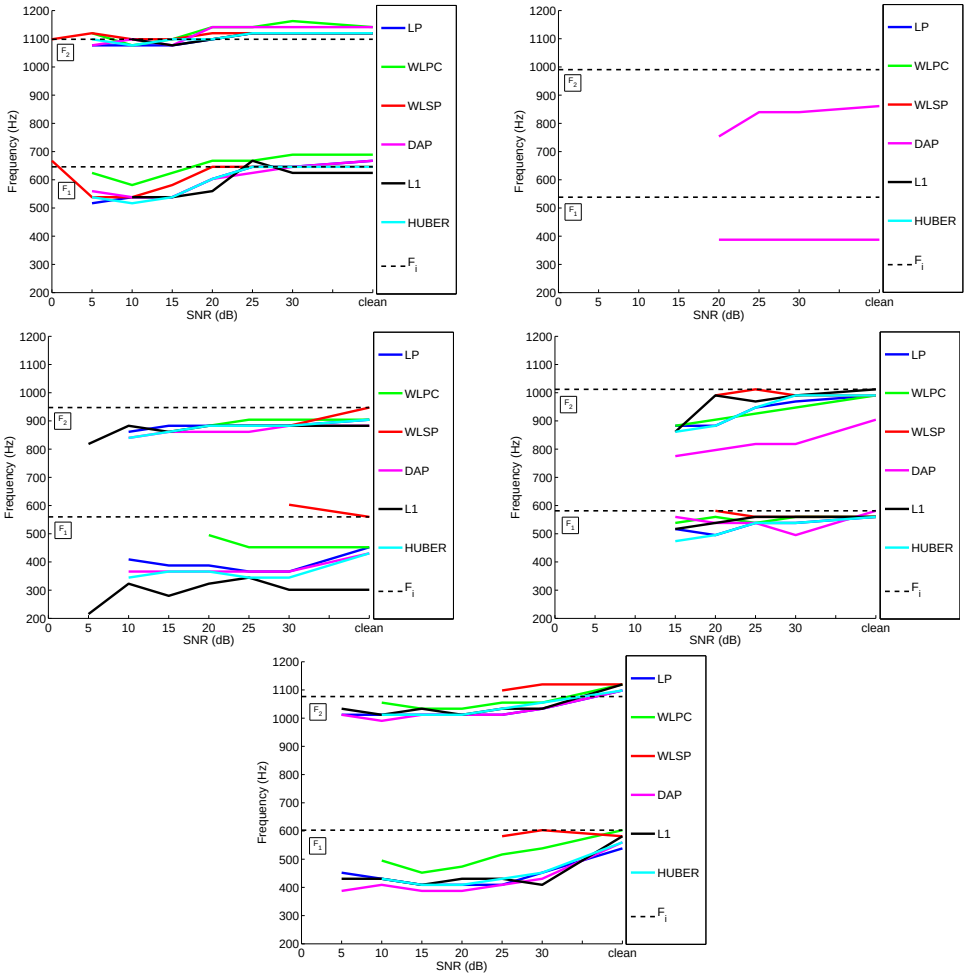

Figure 7.5: First and second formant (F_1, F_2) locations (Hz) as a function of SNR (dB). The analyses were using all seven selected all-pole methods. Speech was corrupted by Laplacian additive noise. The "correct" formants (dashed lines) were derived by using 45 order LP model matched to the underlying vowel. Results were given separately for all five male speakers.

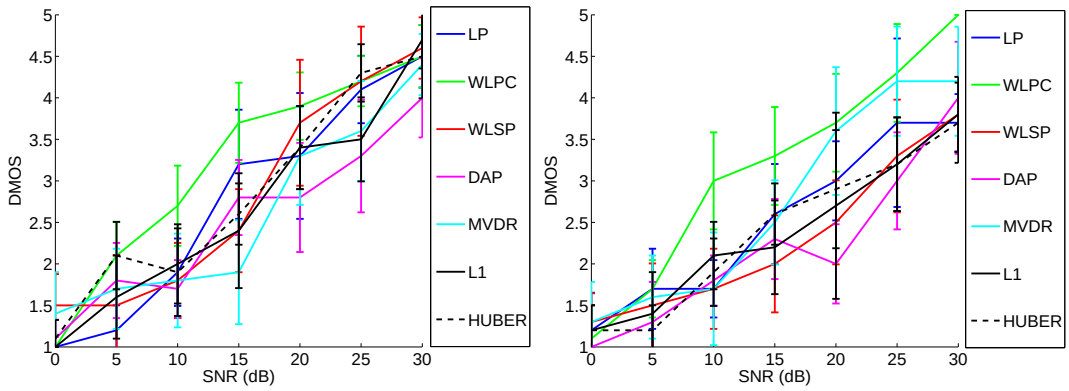

Figure 7.6: Mean DMOS values for all seven selected all-pole models. The values were computed as an average over the data obtained from 10 listeners. The synthesised speech sample was calculated from Finnish vowel /a/. Left panel: synthesise computed using data pronounced male 1. Right panel: synthesise computed using data pronounced male 3.

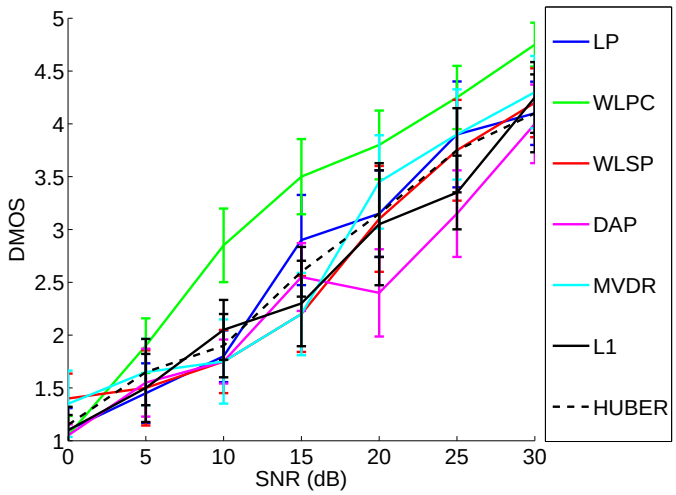

Figure 7.7: Mean DMOS value for all seven selected all-pole models. The values were computed as an average over the data obtained from 10 listeners for both synthesised speech samples pronounced by male 1 and 3. The synthesised speech sample was calculated from Finnish vowel /a/.

7.3 Correlation Between Subjective and Objective Results

In this section we compare the results of the subjective and objective testing. This is done simply by studying SD_2 as a function of the DMOS value and by fitting the straight line for the data by minimising the mean square error. The correlation coefficient r was also

calculated using the definition from [5, 22]

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_j (y_j - \bar{y})^2}}, \tag{7.1}$$

where the $\bar{x}$ is the mean value calculated from the data vector $\mathbf{x}$. Fig. 7.8 shows clearly that there is simple relationship between DMOS and SD_2 scores. The correlation coefficient between the SD_2 (calculated for the entire frequency range) and the DMOS, where, in both situations, the contamination was performed using Gaussian noise, was $r = -0.9670$ see Fig. 7.8 left panel. The correlation coefficient was calculated also for the SD_2 where the calculation was performed over the half frequency range. In this situation the correlation coefficients was $r = -0.9654$ see Fig. 7.8 right panel. One observes that the correlation coefficients are fairly high. The reason behind this is the linear nature of the both, subjective and objective results as a function of the SNR. Other reason is that there were only 49 points included in the correlation coefficients calculations in both cases, which is rather low.

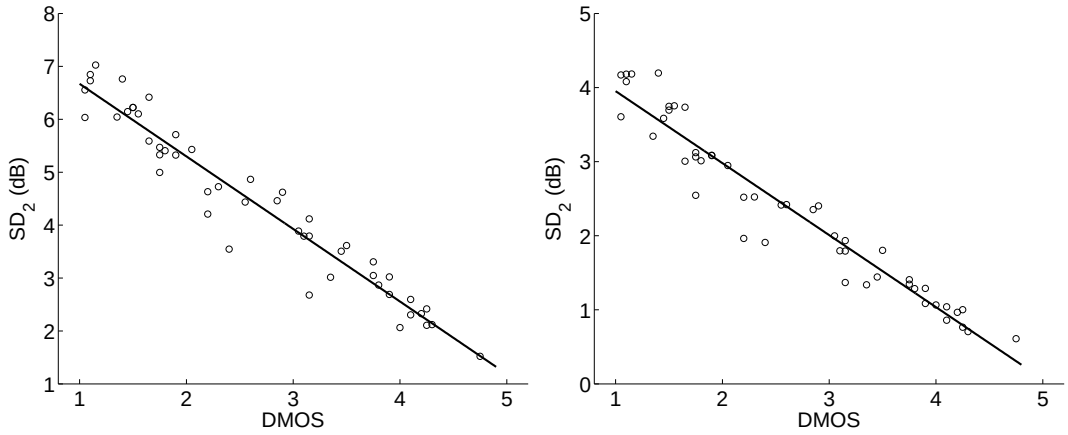

Figure 7.8: Objective SD_2 values in dB as a function of subjective DMOS values. The straight line is fitted to the data with the least square method. Left panel: SD_2 calculated from the entire frequency range and the correlation coefficient was $r = -0.9670$. Right panel: SD_2 calculated from the half frequency range and the correlation coefficient was $r = -0.9654$.

Chapter 8

Conclusions

In this work, the spectral modelling properties and the model errors of seven all-pole models were first compared on a theoretical level. Subsequently, the spectral modelling properties were compared using both objective and subjective testing. The all-pole models were formulated, and their modelling properties as well as modelling error behaviour were examined in the time and frequency domain. The different limitations of LP model were presented in order to derive and justify the existence of the other all-pole methods: WLPC, WLSP, DAP, MVDR, and the M-estimates such as HUBER and ℓ_1 methods. All methods were formulated using the same notation, in order to present a uniform theory covering the all-pole methods in question. Moreover, the connection between the M-estimates and the WLPC method was introduced. In fact, using the approximated IRLS method for the M-estimates, the solutions were showed to be identical. Particularly in the M-estimates and WLPC methods the properties of the different weighting functions were discussed.

Especially emphasise in this work was devoted in the stability of the all-pole models optimised in the time domain. Stability of the LP method was presented using a new approach based on new property which applies the linear projection operator defined in Eq. 3.3. Using this result we were able to present the connection between the stability region and the normal equation in a profound manner. A new stability region for the weighted LP method (WLPC) with respect to the weighting function was derived. This new stability region is tending towards the stability region derived for the LP method as the weights approach to the unity. This is clearly a considerable improvement to the old stability region for the WLPC method. The stability properties of the WLSP method were derived in new accurate way by using a well known *principle of the argument* from complex analysis. The properties of the objective $\mathrm{SD}_{\wp}$ measures were described in a mathematical way and the connection between the Itakura-Saito error measure and the SD_2 measure was introduced.

In addition, the behaviour of the all-pole models in the presence of uncorrelated Gaussian

and Laplacian background noise were examined with objective and subjective tests. The objective measures used were the logarithmic spectral differences and the shifting of the first two formants as a function of signal to noise ratio. The subjective test was carried out as listening tests where the DCR testing procedure was used. The correlation between the subjective and objective results were calculated using the correlation coefficient and the correlation was found to be remarkably strong.

As a conclusion, the WLPC model, for which the weighting function was selected to be the short time energy, gave the best results both in the objective and subjective tests. A new stability region for the WLPC model with respect to the weighting function was derived. Finally, we conclude that it is recommendable to use the WLPC method with appropriate weighting function in the presence of uncorrelated noise. The stability of this method can be guaranteed by choosing the weighting function in the right manner using our new stability region with respect to the weighting function. We also conclude that because of the stability and robustness of the WLPC method, it could be extremely feasible in many fields of speech technology, such as speech enhancement and inverse filtering.

In future work, the different weighting functions of the WLPC method should be studied in more detailed way in the presence of different kind of background noise, and stability analysis for adaptive methods in general, should be derived in order to present more general results concerning the stability regions.

Bibliography

[1] Methods for subjective determination of transmission quality. *ITU-T Recommendation P.800*, August 1996.

[2] Jr. A.H. Gray and J.D. Markel. A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-22(3):207–217, June 1974.

[3] Jr. A.H. Gray and J.D. Markel. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-24(5):380–391, October 1979.

[4] P. Alku and S. Varho. A new linear predictive method for compression of speech signals. In *Proc. International Conference on Spoken Language Processin ICSLP*, volume 6, pages 2563–2566, Sydney, Australia, 1998.

[5] A. Bayya and M. Vis. Objective measures for speech quality assessment in wireless communications. In *Proc. IEEE Acoustics, Speech, and Signal Proc. ICASSP'96*, volume 1, pages 495–498, 1996.

[6] M.S. Bazaraa, H.D. Sherali, and C.M. Shetty. *Nonlinear programming: Theory and algorithms*. J. Wiley & Sons, Inc., New York, 2nd edition, 1993.

[7] Y. Bistritz and S. Peller. Imittance spectral pairs (ISP) for speech encoding. In *Proc. IEEE Acoustics, Speech, and Signal Proc. ICASSP'93*, volume 2, pages 9–12, 1993.

[8] T. Bäckström. *Linear Predictive Modelling of Speech – Constraints and Line Spectrum Pair Decomposition*. PhD thesis, Helsinki University of Technology (TKK), Espoo, Finland, 2004. http://lib.tkk.fi/Diss/2004/isbn9512269473/.

[9] T. Bäckström and P. Alku. All-pole modeling technique based on weighted sum of lsp polynomials. *IEEE Signal Processing Letters*, 10(6):180–183, June 2003.

[10] Y. Kamp C. Ma and L.F. Willems. Robust signal selection for linear prediction analysis of voiced speech. *Speech Communication*, 12(1):69–81, March 1982.

[11] G. Cybenko. Locations of zeros of predictor polynomials. *IEEE Transactions on Automatic Control*, AC-27(1):235–237, February 1982.

[12] A. El-Jaroudi and J. Makhoul. Discrete all-pole modelling. *IEEE Transactions on Signal Processing*, 39(2):411–423, February 1991.

[13] G. Fant. *Acoustic Theory of Speech Production*. Mouton, The Hague, 1960.

[14] J.L. Flanagan. *Speech Analysis Synthesis and Perception*. Springer-Verlag, New York, 1972.

[15] M.H. Hayes. *Statistical Digital Signal Processing and Modelling*. John Wiley & Sons, Inc, 1996.

[16] S. Haykin. *Adaptive Filter Theory*. NJ: Prentice Hall, 1991.

[17] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[18] P.J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, March 1964.

[19] A. Iivonen. Homepage of university of helsinki: Department of speech science, May 2000. http://www.helsinki.fi/hum/hyfl/projektit/vokaalikartat.html#finswedish-long_vowels.

[20] M.T. Karaev. The numerical range of a nilpotent operator on a Hilbert space. *Proceedings of the American Mathematical Society*, 132(8):2321–2326, February 2004.

[21] C-H. Lee. On robust linear prediction of speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(5):642–650, May 1988.

[22] Malden Electronics Ltd. Speech quality assessment, 2000. Background Information For DSLA Users.

[23] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, April 1975.

[24] J. Makhoul. Spectral linear prediction: Properties and applications. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-23(3):283–296, June 1975.

[25] J.D. Markel and A. H. Gray. *Linear prediction of speech*. Berlin : Springer, 1976.

[26] M.N. Murthi and B.D. Rao. All-pole modelling of speech based on the minimum variance distortionless response spectrum. *IEEE Transactions on Speech and Audio Processing*, 8(3):221–239, May 2000.

[27] D. O'Shaughnessy. *Speech Communication Human and Machine*. Addison-Wesley, Canada, 1987.

[28] D. O'Shaughnessy. *Speech Coding and Synthesis*. Elseviever Science B.V., 1995.

[29] Y. Genin P. Delsarte and Y. Kamp. Stability of linear predictors and numerical range of a linear operator. *IEEE Transactions on Information Theory*, IT-33(3):412–415, May 1982.

[30] L. Rabiner and B-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall PTR, 1993.

[31] W. Rudin. *Real and Complex Analysis*. McGraw-Hill International Editions, 1987.

[32] Y. Miyoshi K. Yamato R. Mizoguchi M. Yanagida and O. Kakusho. Analysis of speech signals of short pitch period by a sample-selective linear prediction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-35(9):1233–1240, September 1987.

Appendix A

Definitions

Asymptotically unbiased estimator: Estimator $\hat{\theta}$ is called the asymptotically unbiased estimator of θ if the $E[\hat{\theta}]$ approaches θ as sample size n is increased.

$$\lim_{n\to\infty} E[\hat{\theta}] = \theta.$$

Bounded linear operator: A bounded linear operator Λ is a linear transformation between normed vector spaces X and Y such as there exists some $M > 0$ such that $\forall x \in X$,

$$\|\Lambda(x)\|_Y \leqslant M\|x\|_X.$$

Characteristic polynomial: The characteristic polynomial of $\mathbf{A}$ is defined by

$$p_{\mathbf{A}}(x) := det(x\mathbf{I} - \mathbf{A}).$$

Convex hull: Let Ω be arbitrary set. The convex hull of Ω, denoted as $conv(\Omega)$, is the collection of all convex combinations of Ω. That is

$$conv(\Omega) := \left\{ \sum_{i=1}^{n} \alpha_i x_i \ : \ x_1, \ldots, x_n \in \Omega, \ \sum_{i=1}^{n} \alpha_i = 1, \ \ \alpha_i \leqslant 0 \ \forall i, \ n \in \mathbb{Z}_+ \right\}.$$

$conv(\Omega)$ is also the minimal convex set that contains Ω.

Numerical range of the matrix: The numerical range of an $n \times n$ complex matrix $\mathbf{A}$, also known as its field of values, is defined as

$$\mathcal{F}(\mathbf{A}) := \{\mathbf{x}^*\mathbf{A}\mathbf{x} \ : \ \|\mathbf{x}\| = 1, \ \ \mathbf{x} \in \mathbb{C}^n\}.$$

Principle of the argument: Let $f(z)$ be a single-valued function that is analytic a region Ω enclosed by a contour ζ and let $f(z)$ be of the form $f(z) = \frac{x(z)}{y(z)}$. Let N_z be the number of complex roots of $f(z)$ in ζ, and N_p be the number of poles in ζ, then

$$N_z - N_p = \frac{1}{2\pi j} \int_{\zeta} \frac{f\,'(z)}{f(z)} dz.$$

Zero mean property of the all-pole model This property is based on the proof presented in [25]. Consider the prediction polynomial $A(z)$ that has all its roots inside the unit circle. Then if the $\Re\{\cdot\}$ denotes the real part operator of the complex number and $\text{Res}(\cdot)$ denotes the residue, we then have [31]

$$\begin{aligned}
&\int_{-\pi}^{\pi} ln|A(e^{j\omega})|^2 \frac{d\omega}{2\pi} = \int_{-\pi}^{\pi} ln|A(e^{-j\omega})|^2 \frac{d\omega}{2\pi} \\
&= 2\Re\left\{\int_{-\pi}^{\pi} ln\left[A(e^{-j\omega})\right] \frac{d\omega}{2\pi}\right\} = 2\Re\left\{\oint_{\Gamma} ln\left[A(1/z)\right] \frac{dz}{2\pi jz}\right\} \\
&= 2\Re\left\{\text{Res}(0)\right\} = 2\Re\left\{\lim_{z\to 0} ln\left[A(1/z)\right]\right\} = 0
\end{aligned}$$

where we have used the *residue theorems* from complex analysis [31] and the fact that the first coefficient of the prediction polynomial is equally to one.

Appendix B

Tables

Table B.1: Quality rating scale for a degradation category rating test.

Description	Rating
Degradation not perceived	5
Degradation perceived but not annoying	4
Degradation slightly annoying	3
Degradation annoying	2
Degradation very annoying	1

Table B.2: Finnish quality rating scale for a degradation category rating test. The question to be asked was: Näytteen A huonontuma referenssiääneen on:

Description	Rating
Ei kuultavissa oleva	5
Kuultavissa, mutta ei häiritsevä	4
Hieman häiritsevä	3
Häiritsevä	2
Erittäin häiritsevä	1

Table B.3: Spectral distance measure between clean and corrupted all pole envelope as a function of SNR, calculated for a vowel /a/ for MALE 1-5. The speech sample was corrupted by Gaussian white noise and the SD_2 (in dB) was calculated in frequency range $0Hz \rightarrow 11025Hz$.

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	5.6779	5.1267	4.4165	3.6417	2.76286	2.07068	1.46873
WLPC	4.8904	4.2415	3.39457	2.56506	1.94256	1.58799	1.04158
WLSP	5.6383	5.0912	4.37655	3.6161	2.68897	2.23745	1.46874
MVDR	5.1262	4.5252	4.03488	3.34229	2.59931	1.95351	1.39385
DAP	5.4072	4.8977	4.1236	3.33132	2.54245	1.94107	1.30772
ℓ_1	5.516	4.8634	4.12393	3.28757	2.62873	1.69613	1.35935
HUBER	5.8853	5.1964	4.52992	3.74786	3.02854	2.24529	1.59304

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	7.1721	6.5681	5.80673	5.08459	4.26693	3.4804	2.73869
WLPC	6.2706	5.7229	4.83635	3.99002	3.1526	2.42196	1.73996
WLSP	7.3471	6.6946	5.91559	5.07481	4.23245	3.4062	2.71431
MVDR	6.4516	6.0749	5.42314	4.66475	3.91317	3.00442	2.48334
DAP	7.3514	6.7119	5.89794	4.63571	3.64835	2.35429	2.17893
ℓ_1	7.609	6.9687	6.13986	5.4345	4.63121	3.41422	3.03133
HUBER	7.4468	6.7932	6.00483	5.22682	4.38127	3.65121	2.98418

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	6.3408	5.7458	4.9923	4.1663	3.3708	2.6008	1.8692
WLPC	5.2107	4.5487	3.425	2.6336	1.9939	1.1207	0.54938
WLSP	6.5295	5.9358	5.1018	4.2949	3.4859	2.684	1.9652
MVDR	5.6245	5.1438	4.5949	3.776	3.1222	2.2474	1.7149
DAP	6.115	5.6525	4.7361	4.2139	3.307	2.6346	1.7843
ℓ_1	6.5124	5.9423	5.0818	4.4165	3.6196	2.9769	2.1855
HUBER	6.9111	6.3172	5.6751	4.8289	4.0827	3.2288	2.5523

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	7.1451	6.524	5.7411	4.9173	4.073	3.2725	2.5565
WLPC	6.8201	6.2693	5.3746	4.4248	3.691	2.6636	2.215
WLSP	7.0005	6.5733	5.8155	4.836	4.0757	3.2236	2.536
MVDR	6.3492	5.9252	5.3258	4.401	3.7003	2.8771	2.3267
DAP	7.0762	6.4896	5.7512	4.7072	3.727	2.6744	2.2529
ℓ_1	7.4325	6.7587	5.7445	5.1676	4.1576	3.4724	2.5719
HUBER	7.6244	6.8852	6.2061	5.2835	4.5137	3.6522	2.9276

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	7.3039	6.7668	6.0783	5.2968	4.4936	3.6815	2.8929
WLPC	6.9805	5.835	5.2745	4.461	3.5604	2.7417	2.0618
WLSP	7.2961	6.8254	6.1437	5.3333	4.4791	3.6933	2.9579
MVDR	6.6634	6.2792	5.5993	4.8658	4.2001	3.37	2.6818
DAP	6.8259	6.7635	6.1387	5.2898	4.5119	3.7854	2.7969
ℓ_1	7.1617	6.5807	6.0615	5.3202	4.4037	3.5184	2.9365
HUBER	7.2574	6.8978	6.1382	5.2351	4.5846	3.7574	2.9223

Table B.4: Spectral distance measure between clean and corrupted all pole envelope as a function of SNR, calculated for a vowel /a/ for MALE 1-5. The speech sample was corrupted by Gaussian white noise and the SD_2 (in dB) was calculated in frequency range $0Hz \rightarrow 5525Hz$.

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	3.4352	2.9351	2.3301	1.7116	0.92402	0.4721	0.20148
WLPC	2.6941	2.1934	1.4399	0.76707	0.45285	0.13961	0.16561
WLSP	3.3256	2.9892	2.4216	1.5941	0.9217	0.54501	0.31387
MVDR	2.8698	2.3516	1.9488	1.4229	0.71798	0.51207	0.2171
DAP	3.3011	2.6833	2.0003	1.471	0.97873	0.78755	0.37746
ℓ_1	3.3967	2.8401	2.0721	1.4828	1.0235	0.44561	0.21649
HUBER	3.2161	2.9582	2.2248	1.5635	1.0444	0.57424	0.3096

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	4.3594	3.8147	3.1499	2.4205	1.8198	1.2983	0.86858
WLPC	4.0067	3.5859	2.6995	2.1295	1.5454	0.96783	0.64883
WLSP	4.5517	3.8666	3.2009	2.5079	1.7302	1.304	0.89701
MVDR	3.5199	3.4372	2.9039	2.1844	1.7276	1.1698	0.77302
DAP	5.2908	4.7069	3.9919	2.9357	2.2495	1.5455	1.3751
ℓ_1	4.7162	4.2119	3.3903	2.9892	2.3502	1.4597	1.4012
HUBER	4.5396	3.8364	3.1462	2.4424	1.9252	1.4608	0.96849

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	4.0035	3.5131	2.9208	2.443	1.8991	1.4	0.95098
WLPC	3.0888	2.675	1.6828	1.5146	0.7191	0.44024	0.13072
WLSP	4.2296	3.722	3.0425	2.6444	1.9995	1.5705	0.90086
MVDR	3.1007	2.671	2.048	1.6244	1.256	0.80784	0.52739
DAP	3.8375	3.5678	2.8082	2.2961	1.8027	1.1315	0.91106
ℓ_1	4.0298	3.5325	2.7942	2.4483	2.02	1.28	0.91805
HUBER	4.0403	3.5772	3.0104	2.4147	1.8408	1.3144	0.94933

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	5.2485	4.6991	4.0233	3.2736	2.6276	2.0686	1.4389
WLPC	4.9322	4.403	3.819	3.0082	2.689	1.7303	1.4751
WLSP	5.1783	4.8108	3.9303	3.3291	2.6584	2.2283	1.6051
MVDR	4.2948	3.9363	3.5297	2.7744	2.1295	1.8163	1.2867
DAP	5.0316	4.4705	4.0245	2.9768	2.5412	2.0343	1.7139
ℓ_1	5.3589	4.8143	3.903	3.5031	2.8051	2.2318	1.6559
HUBER	5.5538	5.0291	4.4142	3.6232	3.037	2.5076	2.0882

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	3.3541	2.9625	2.6353	2.1623	1.6958	1.2056	0.8395
WLPC	3.306	2.5745	2.1224	1.5908	1.0216	0.54941	0.62876
WLSP	3.6972	3.3435	2.7245	2.5235	1.6725	1.0799	1.1145
MVDR	2.9288	2.63	2.2995	1.8085	1.3841	1.1217	0.7221
DAP	3.39	3.3366	2.7825	2.3889	1.9749	1.3452	0.94811
ℓ_1	3.404	3.0789	2.5781	2.1935	1.7918	1.2728	0.81393
HUBER	3.5631	3.2728	2.6137	2.0557	1.8228	1.1758	0.88594

Table B.5: Logarithmic spectral distance measure between clean and corrupted all pole envelope as a function of SNR, calculated for a vowel /a/ for MALE 1-5. The speech sample was corrupted by Laplacian noise and the SD_2 (in dB) was calculated in frequency range $0Hz \rightarrow 11050Hz$.

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	5.7783	5.2217	4.3504	3.5586	2.7798	2.071	1.361
WLPC	4.9465	4.2939	3.2178	2.7368	2.0536	1.5269	1.0218
WLSP	5.7376	5.0485	4.4102	3.7942	2.6659	2.1764	1.5344
MVDR	5.1474	4.5954	4.1223	3.3949	2.5321	1.7376	1.4377
DAP	5.2058	4.9772	4.2022	3.2168	2.6334	2.1224	1.4055
ℓ_1	5.5042	4.7894	3.9532	3.2244	2.543	1.6969	1.2577
HUBER	5.862	5.256	4.5462	3.7674	2.8349	2.236	1.5981

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	7.2203	6.6232	5.9951	5.1529	4.314	3.5011	2.7435
WLPC	6.262	5.668	4.7519	3.9918	3.2887	2.4049	1.6558
WLSP	7.2291	6.8117	5.9398	4.8641	4.2572	3.519	2.6733
MVDR	6.5093	6.0305	5.3824	4.7019	3.8713	3.2825	2.4875
DAP	7.4321	6.7688	6.1322	4.9465	3.9883	2.547	2.0355
ℓ_1	7.7194	7.0707	6.1271	5.3389	4.6831	3.5232	2.6894
HUBER	7.3504	6.7901	5.9741	5.2636	4.3932	3.644	2.8169

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	6.312	5.8753	5.1359	4.3551	3.5703	2.5752	1.879
WLPC	5.2048	4.4625	3.5962	2.8012	1.8837	1.0182	0.34265
WLSP	6.4589	5.8284	5.029	4.2802	3.5415	2.5368	1.9854
MVDR	5.6849	5.1815	4.5761	3.7913	2.9506	2.3433	1.6606
DAP	6.0871	5.5884	4.9117	4.1812	3.4072	2.6658	1.8517
ℓ_1	6.5176	5.8987	5.1884	4.3884	3.5848	2.8058	2.1899
HUBER	6.8693	6.3616	5.5398	4.8253	4.024	3.2429	2.4864

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	7.2111	6.5928	5.8078	4.9665	4.0976	3.282	2.5247
WLPC	6.6551	6.0648	5.3905	4.4649	3.6408	2.8111	2.2385
WLSP	7.1624	6.7017	5.7505	4.7946	4.0072	3.2888	2.6242
MVDR	6.3141	5.9277	5.1916	4.4114	3.6542	2.989	2.3194
DAP	7.0366	6.4246	5.5245	4.6682	3.653	2.8277	2.2226
ℓ_1	7.4974	6.761	6.0054	4.8677	4.1472	3.4911	2.6612
HUBER	7.5828	6.9393	6.1635	5.3327	4.4551	3.6572	2.8714

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	7.3897	6.8341	6.1101	5.318	4.4508	3.6556	2.8998
WLPC	6.8742	6.249	5.3997	4.3702	3.6345	3.0169	1.9301
WLSP	7.4058	6.8552	6.1527	5.3083	4.4673	3.5853	3.0023
MVDR	6.7469	6.2635	5.6291	4.7855	4.2377	3.5145	2.6674
DAP	7.1956	6.6906	6.0281	5.2213	4.5204	3.6663	3.0206
ℓ_1	7.2662	6.5057	5.8513	5.023	4.2851	3.5605	2.7367
HUBER	7.2492	6.8047	6.1418	5.2885	4.5128	3.7212	2.9715

Table B.6: Spectral distance measure between clean and corrupted all pole envelope as a function of SNR, calculated for a vowel /a/ for MALE 1-5. The speech sample was corrupted by Laplacian noise and the SD_2 (in dB) was calculated in frequency range $0Hz \rightarrow 5525Hz$.

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	3.4432	2.9418	2.2994	1.6638	1.047	0.55384	0.31909
WLPC	2.68	2.0455	1.3652	0.88177	0.53841	0.21644	0.14284
WLSP	3.311	2.8149	2.3236	1.7003	0.77913	0.53114	0.28099
MVDR	2.947	2.4763	1.8952	1.4386	0.82761	0.28468	0.33366
DAP	3.1002	2.7369	2.228	1.568	1.1234	0.52417	0.48667
ℓ_1	3.2586	2.7877	1.9978	1.507	0.87529	0.55824	0.24607
HUBER	3.3918	2.7439	2.1993	1.5794	0.87195	0.55079	0.25223

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	4.444	3.904	3.369	2.6921	2.0747	1.5498	1.0858
WLPC	4.0033	3.3965	2.7171	2.0727	1.5991	1.019	0.68256
WLSP	4.4747	3.8788	3.37	2.2642	1.6505	1.2879	0.81248
MVDR	3.8398	3.3437	2.7538	2.2152	1.6529	1.3719	0.7439
DAP	5.2739	4.8251	4.1705	3.183	2.4972	1.5139	1.0721
ℓ_1	4.9257	4.2304	3.5216	2.8297	2.3766	1.6192	1.0234
HUBER	4.2759	3.9215	3.1389	2.5073	2.007	1.3551	0.92766

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	4.078	3.344	2.7899	2.2671	1.7453	1.2138	0.82533
WLPC	3.2239	2.5215	1.9842	1.4005	0.80898	0.43321	0.25591
WLSP	4.1435	3.5978	3.1738	2.3058	2.0623	1.4902	0.99976
MVDR	3.0612	2.5139	2.1692	1.6557	1.142	0.83299	0.45221
DAP	3.9302	3.4054	2.8304	2.2586	1.704	1.2123	0.84282
ℓ_1	3.9893	3.5788	2.8868	2.2095	1.9825	1.2605	0.80791
HUBER	4.1195	3.6281	2.9745	2.4555	1.9643	1.4503	1.1851

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	5.2564	4.7586	4.0838	3.4012	2.6911	2.1079	1.6012
WLPC	4.7893	4.2692	3.8234	2.9975	2.4021	1.8451	1.2726
WLSP	5.3552	4.8399	3.9629	3.3064	2.6812	2.2162	1.8575
MVDR	4.3202	3.9559	3.247	2.6952	2.1254	1.6851	1.3121
DAP	5.0027	4.3268	3.5728	3.1894	2.3969	1.9949	1.7776
ℓ_1	5.4787	4.7783	4.1338	3.3257	2.5448	2.3335	1.9373
HUBER	5.5281	5.0012	4.3752	3.7104	2.949	2.3907	1.955

METHOD	SNR=0	SNR=5	SNR=10	SNR=15	SNR=20	SNR=25	SNR=30
LP	3.5601	3.1796	2.7214	2.2354	1.6766	1.187	0.83112
WLPC	3.0189	2.9613	2.1232	1.7574	1.2202	2.0597	0.68262
WLSP	3.7268	3.4252	2.9362	2.2953	1.8771	1.2703	1.0706
MVDR	2.9307	2.676	2.2416	1.7634	1.4832	1.2551	0.733
DAP	3.4479	3.1313	2.7984	2.3128	1.7369	1.3701	0.97409
ℓ_1	3.6244	2.9883	2.531	2.0365	1.7252	1.1161	0.76529
HUBER	3.6306	3.2959	2.8759	2.3636	1.7836	1.2776	0.93394