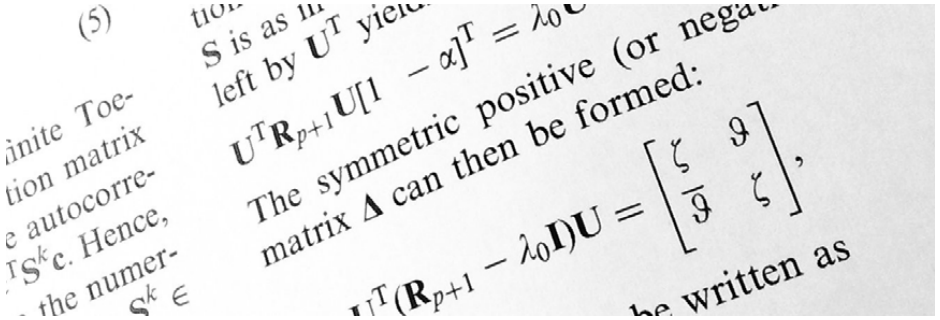
Helsinki University of Technology
Department of Signal Processing and Acoustics
Espoo 2009 Report 9

MATHEMATICAL METHODS FOR LINEAR PREDICTIVE SPECTRAL MODELLING OF SPEECH

Carlo Magi



Helsinki University of Technology
Faculty of Electronics, Communications and Automation
Department of Signal Processing and Acoustics

Teknillinen korkeakoulu
Elektroniikan, tietoliikenteen ja automaation tiedekunta
Signaalinkäsittelyn ja akustiikan laitos

Helsinki University of Technology
Department of Signal Processing and Acoustics
P.O. Box 3000
FI-02015 TKK
Tel. +358 9 4511
Fax +358 9 460224
E-mail Lea.Soderman@tkk.fi

ISBN 978-951-22-9963-8
ISSN 1797-4267

Multiprint Oy
Espoo, Finland 2009

Carlo Magi (1980-2008)

Preface

Brief is the flight of the brightest stars.

In February 2008, we were struck by the sudden departure of our colleague and friend, Carlo Magi. His demise, at the age of 27, was not only a great personal shock for those who knew him, but also a loss for science. During his brief career in science, he displayed rare talent in both the development of mathematical theory as well as application of mathematical results in practical problems of speech science. Apart from mere technical contributions, Carlo was also a valuable and highly appreciated member of the research team. Among colleagues, he was well known for venturing into exciting philosophical discussions and his positive as well as passionate attitude was motivating for everyone around him. His contributions will be fondly remembered.

Carlo's work was tragically interrupted just a few months before the defence of his doctoral thesis. Shortly after Carlo's passing, we realised that the thesis he was working on had to be finished. We, the colleagues of Carlo, did not have a choice, it was obvious to us that finishing his work was our obligation. It is our hope that this posthumous doctoral dissertation of Carlo Magi will honour the life and work of our dear colleague, as well as remind us how fortunate we were to have worked with such a talent.

Paavo Alku
Tom Bäckström
Jouni Pohjalainen

Abstract

Linear prediction (LP) is among the most widely used parametric spectral modelling techniques of discrete-time information. This method, also known as autoregressive (AR) spectral modelling, is particularly well-suited to processing of speech signals, and it has become a major technique that is currently used in almost all areas of speech science. This thesis deals with LP by proposing three novel, mathematically-oriented perspectives. First, spectral modelling of speech is studied with the help of symmetric polynomials, especially with those obtained from the line spectrum pair (LSP) decomposition. Properties of the LSP polynomials in parametric spectral modelling of speech are presented in the form of a review article and new properties of the roots of the LSP polynomials are derived. Second, the concept of weighted LP is reviewed and a novel, stabilized version of this technique is proposed and its behaviour is analysed in robust feature extraction of speech information. Third, this study proposes novel, constrained linear predictive methods that are targeted to improve the robustness of LP in the modelling of the vocal tract transfer function in glottal inverse filtering. The focus of this thesis is on the theoretical properties of linear predictive spectral models, with practical applications in areas such as in feature extraction of automatic speech recognition.

Contents

List of publications

I Carlo Magi, Tom Bäckström, Paavo Alku. "Simple proofs of root locations of two symmetric linear prediction models", Signal Processing, Vol. 88, Issue 7, pp. 1894-1897, 2008.

II Carlo Magi, Jouni Pohjalainen, Tom Bäckström, Paavo Alku. "Stabilised weighted linear prediction", Speech Communication, Vol. 51, pp. 401-411, 2009.

III Tom Bäckström, Carlo Magi. "Properties of line spectrum pair polynomials – A review", Signal Processing, Vol. 86, No. 11, pp. 3286-3298, 2006.

IV Tom Bäckström, Carlo Magi, Paavo Alku. "Minimum separation of line spectral frequencies", IEEE Signal Processing Letters, Vol. 14, No. 2, pp. 145-147, 2007.

V Tom Bäckström, Carlo Magi. "Effect of white-noise correction on linear predictive coding", IEEE Signal Processing Letters, Vol. 14, No. 2, pp. 148-151, 2007.

VI Paavo Alku, Carlo Magi, Santeri Yrttiaho, Tom Bäckström, Brad Story. "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering", Journal of the Acoustical Society of America, Vol. 125, No. 5, pp. 3289-3305, 2009.

VII Paavo Alku, Carlo Magi, Tom Bäckström. "Glottal inverse filtering with the closed-phase covariance analysis utilizing mathematical constraints in modelling of the vocal tract", Logopedics, Phoniatrics, Vocology (Special Issue for the Jan Gauffin Memorial Symposium). 2009 (In press).

Chapter 1

Introduction

Estimation of the power spectral density of discrete-time signals is a research topic that has been studied in several science disciplines, such as geophysical data processing, biomedicine, communications, and speech processing. In addition to conventional techniques utilising the Fourier transform, a widely used family of spectrum estimation methods is based on *linear predictive* signal models, also known as autoregressive (AR) models [Makhoul, 1975, Markel and Gray Jr., 1976, Kay and Marple, 1981]. These techniques aim to represent the spectral information of discrete-time signals in parametric forms using digital filters having an *all-pole* structure, that is their $\mathbb{Z}$-domain transfer functions comprise only poles. The term linear prediction (LP) refers to the time-domain formulation which serves as the basis in the optimisation of the all-pole filter coefficients: each time-domain signal sample is assumed to be predictable as a linear combination of a known number of previous samples. The optimal (typically in the "minimum mean square error" sense) set of filter coefficients is solved to obtain the all-pole spectral model.

In speech science, linear predictive methods have a particularly established role, due to their close connection to the source-filter theory of speech production and its underlying theory of the tube model of the vocal tract acoustics [Fant, 1970, Markel and Gray Jr., 1976]. The model provided

by LP is especially well-suited for voiced segments of speech, in which AR modelling allows a good digital approximation for the filtering effect of the instantaneous vocal tract configuration on the glottal excitation. Consequently, LP can be used as a straightforward technique to decompose a given (voiced) speech sound into two components, an excitation and a filter, a scheme that can be applied for various purposes. In addition, the most well-known form of linear predictive techniques, the conventional LP analysis utilising the autocorrelation criterion, has two important practical benefits: the all-pole model provided by the method is guaranteed to be stable and the optimisation of the model parameters can be performed with good computational efficiency. Since the early linear predictive studies by, for example, Atal and Hanauer [1970], these techniques have been addressed in numerous articles and many modifications of LP have been proposed. During the past three decades, LP has become one of the most widely used speech processing tools, applicable in almost all areas of speech science. In addition to its role as a general short-time spectral estimation method, LP has a crucial role in many ubiquitous speech technology areas, especially in low bit-rate speech coding. LP is used currently, for example, as the core element of speech compression in both the second and third generation of mobile communications, in the VoIP technology, as well as in the recent proposal for the next MPEG standard for unified speech and audio coding [Chu, 2003, Gibson, 2005, Neuendorf et al., 2009].

This thesis deals with novel, mathematically-oriented methods related to LP. The work presents three perspectives into linear predictive spectral modelling of speech. First, a well-known means to represent linear predictive information based on the line spectrum pair (LSP) polynomials is studied. In difference to its most widely used application as a method to quantise linear predictive spectral models, the LSP decomposition is employed in the present work as a general spectral modelling method. Certain new properties concerning the root locations of the LSP polynomials are described. Second, the thesis addresses weighted linear predictive models. These are AR models that are defined by utilising temporal weighting to the square of the prediction error, or residual, in computing the optimal filter coefficients. Third, the study proposes variants to the conventional LP by

imposing constraints in the formulation of linear predictive computation to obtain vocal tract models for glottal inverse filtering. The following three sections address the fundamental theory in the background of these three perspectives.

Chapter 2

Line spectrum pairs

Linear predictive methods provide accurate models of the short-time spectral envelope of speech that can be used in speech processing applications such as speech coding. However, to enable efficient transmission of the models, the model parameters have to be represented in a form robust to transmission errors. One such method is the line spectrum pair (LSP) representation [Itakura, 1975, Soong and Juang, 1984]. This representation is based on a decomposition of a polynomial, such as the linear predictor, into another domain, the line spectrum frequency (LSF) domain, where the parameters can be represented as angles of polynomial zeros on the unit circle.

The use of LSFs as a representation of the linear predictive model has three major advantages compared to antecedent representations. Firstly, when represented by LSFs, the stability of the predictive model can be readily retained even when corrupted by noise. Secondly, small errors in the LSFs produce small and well-localised errors in the model, whereby the representation responds predictably to transmission errors. Finally, the computational complexity of obtaining the LSFs is sufficiently low for practical applications. Due to these advantages, the LSFs appear in combination with the code excited linear prediction (CELP) method in most of the state-of-the-art speech coders [ETSI, a,b,c, ITU-T, Cox, 1995, Neuendorf

et al., 2009].

Prior to LSP, the most suitable representations of LP coefficients were log-area ratios (LAR) and inverse sine quantisation. However, in terms of spectral deviation due to quantisation errors, neither of these representations is optimal [Gray Jr. and Markel, 1976]. An article from the same time period claims that the reflection coefficients would be superior to other coding schemes [Viswanathan and Makhoul, 1975]. In any case, the LSP decomposition was soon found to be insurmountable to all of the above mentioned coding schemes. Quality-wise, coding with a 31-bit representation of the LSP polynomials is equivalent to, or better than, a 41-bit representation with reflection coefficients [Kang and Fransen, 1985].

The LSP polynomials are defined, for an order m predictor $A(z)$, with [Itakura, 1975, Soong and Juang, 1984]

$$\begin{aligned} P(z) &= A(z) + z^{-m-1}A(z^{-1}) \\ Q(z) &= A(z) - z^{-m-1}A(z^{-1}). \end{aligned} \tag{2.1}$$

We can readily see that, using $P(z)$ and $Q(z)$, polynomial $A(z)$ can be reconstructed as

$$A(z) = \frac{1}{2}\left[P(z) + Q(z)\right]. \tag{2.2}$$

The roots α_j and β_j of $P(z)$ and $Q(z)$, respectively, have a number of useful properties, namely, it holds that [Schüssler, 1976, Soong and Juang, 1984, Stoica and Nehorai, 1986]

1. α_j and β_j are on the unit circle $|\alpha_j| = |\beta_j| = 1$ and can be written as $\alpha_j = e^{i\pi\lambda_j}$ and $\beta_j = e^{i\pi\gamma_j}$.

2. λ_j and γ_j are simple and distinct $\lambda_j \neq \lambda_k$ and $\gamma_j \neq \gamma_k$ for $j \neq k$, and $\lambda_j \neq \gamma_k$ for all j.

3. λ_j and γ_j are interlaced, that is, $\gamma_j < \lambda_j < \gamma_{j+1}$ for all j.

Polynomials $P(z)$ and $Q(z)$ can be reconstructed from λ_j and γ_j, and since $A(z)$ can be reconstructed from $P(z)$ and $Q(z)$, the angles λ_j and γ_j can be

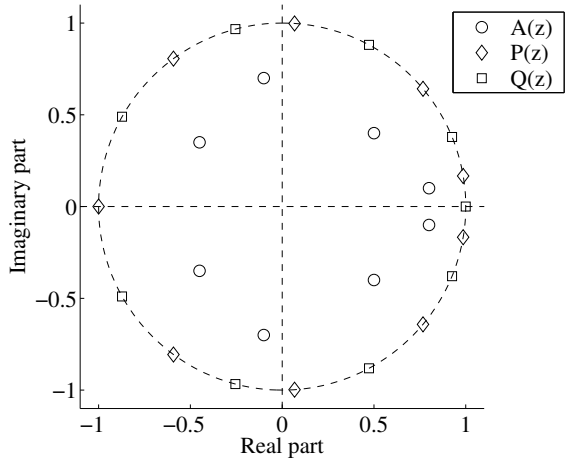


Figure 2.1: Illustration of root loci of LSP polynomials $P(z)$ and $Q(z)$ calculated from polynomial $A(z)$.

used to uniquely describe $A(z)$. This description is bounded since $\lambda_j, \gamma_j \in [0, 1]$ (the complex conjugate λ_j^* and γ_j^* are redundant and can be ignored).

Conversely, if two polynomials have zeros interlaced on the unit circle, their sum is minimum-phase [Soong and Juang, 1984]. Therefore, ensuring the interlacing property is retained, the description is robust in terms of stability of the all-pole model.

Properties 1 and 3 are called the *unit circle property* and the *intra-model interlacing property* of LSP polynomials. These properties are illustrated in Fig. 2.1.

Since the roots lie on the unit circle, the all-pole models $P^{-1}(z)$ and $Q^{-1}(z)$ will have infinite values at these locations. In terms of the spectrum, these roots can be seen as vertical lines at frequencies corresponding to the angle of each root. These lines are known as the line spectrum frequencies (LSFs) of the corresponding model.

It can be concluded that it is possible to describe the spectral envelope of a signal through the angles of the zeros of LSP polynomials $P(z)$ and

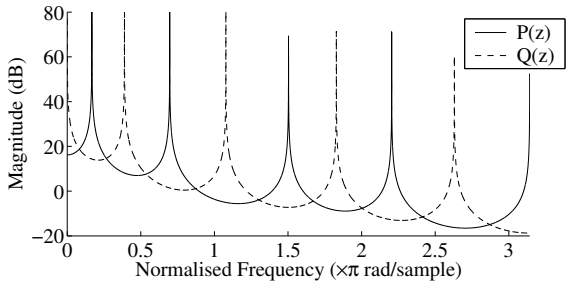


Figure 2.2: Illustration of the line spectrum frequencies in the spectra of polynomials $P(z)$ and $Q(z)$.

$Q(z)$ calculated from an LP polynomial $A(z)$. This yields a convenient representation of the LP coefficients, since the range of the angles is limited to $[0, \pi]$, and the stability of the all-pole model corresponding to $A(z)$ is guaranteed if the interlacing property is retained.

Chapter 3

Weighted linear prediction

Modern speech technology applications are increasingly being used in noisy environments such as in cars, on streets, and in restaurants. Ambient noise, which is typically assumed to be additive, corrupts voice samples. Consequently, extraction of the original spectral information of speech becomes more difficult in, for example, speech coding and speech recognition. There are plenty of previous studies indicating that AR-modelling, such as conventional LP, is sensitive to noise [e.g. Lee, 1988, Ma et al., 1993, Hu, 1998a]. Kay has demonstrated the degradation of AR modelling in several studies by using a simplified scheme based on the use of additive uncorrelated white noise [e.g. Kay, 1978, 1979, 1980, Kay and Marple, 1981]. The degradation of AR modelling is now described by assuming that the signal x_n obeys the following equation of a pth order autoregressive process:

$$x_n = \sum_{k=1}^{p} a_k x_{n-k} + e_n \tag{3.1}$$

, where e_n is a zero mean, white noise process with its autocorrelation function given by $R_e(k) = \sigma_e^2 \delta_k$. By assuming that noise-corruption is additive and that the noise is white and uncorrelated with the signal x_n, the observed process can be written as $y_n = x_n + w_n$, where w_n is noise, whose autocorrelation function is $R_W(k) = \sigma_W^2 \delta_k$. The power spectrum of

the noise-corrupted signal can now be written as

$$|Y(z)|^2 = |X(z)|^2 + |W(z)|^2 = \sigma_e^2 A^{-1}(z)A^{-1}(z^{-1}) + \sigma_W^2 \tag{3.2}$$

, where $A(z)$ is the $\mathbb{Z}$-transform of the AR coefficient sequence a_k.

Equation 3.2 above shows clearly that in the presence of noise the power spectrum of the signal can no longer be parametrised simply by all-pole modelling, because the spectrum now has both zeros and poles, that is the noise-corrupted spectrum becomes an ARMA (autoregressive moving average) process. In parametric modelling of the short-time speech spectrum, the degradation caused by environmental noise is typically seen as smoothing of the resulting AR spectrum (see Fig. 3.1); the formants indicated by the AR spectrum computed from noisy speech are of larger bandwidth than those shown by the AR spectrum estimated from clean speech. If the corruption is severe, the all-pole spectrum typically fails to indicate some of the formants. In addition, the overall dynamic range of the all-pole spectrum computed by LP is typically reduced when the speech is corrupted by noise.

There are plenty of studies aiming at improving the robustness of AR modelling with respect to noise. In speech science, this topic is called *robust LP* although robustness can be sometimes related also to other artifacts than noise, for example to the biasing of the resonances of the AR models by the spectral harmonics [El-Jaroudi and Makhoul, 1991]. In robust LP, the basic linear predictive model used in conventional LP is typically modified by using, for example, different error measures in model optimisation. An example thereof is the classical study by Lee [1988], who used special kinds of cost functions to give more emphasis to small residual values while attenuating residuals of large amplitude in order to improve modelling of formants. Hu [1998b], in turn, proposed a robust approach to LP based on the use of an orthogonal principle to facilitate the incorporation of various error minimisation criteria. Additionally, there are several studies on improved spectral models in which filter optimisation is based on the assumptions that the background noise is additive and white, and that its effects can be compensated by subtracting a suitable bias from

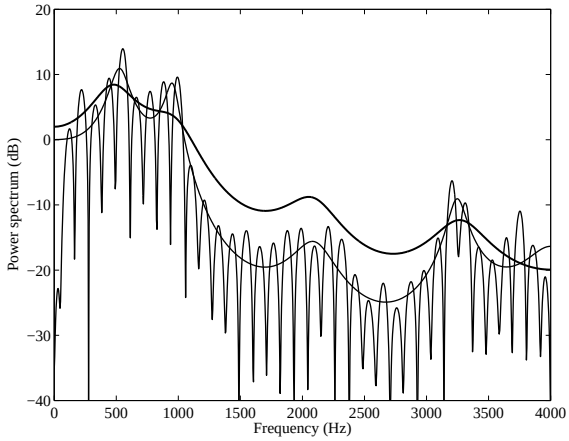


Figure 3.1: Examples of all-pole models computed by the conventional LP (of order $p = 10$) from the vowel /a/ produced by a male speaker. The FFT-spectrum of the clean vowel is shown together with the LP spectra computed from the clean (thin line) and noisy (thick line) vowel with signal-to-noise ratio equal to 3 dB.

the zero-lag autocorrelation in the normal equations [e.g. Kay, 1980, Hu, 1998a]. It is, however, worth noting that most of the existing techniques for computing robust linear predictive models iteratively update the filter parameters. Furthermore, some of the previously developed modifications of LP cannot guarantee the stability of the all-pole filter. Both of these issues impose serious limitations especially in applications such as speech coding and parametric speech synthesis, in which the stability of the parametric synthesis model is a prerequisite; in speech coding, the method also typically needs to be implemented to work in real time with modest hardware requirements. Moreover, if robust spectral models are developed based on simplified noise models, their performance typically deteriorates when processing speech corrupted by realistic distortion such as office noise, car noise, or babble noise.

In the present study, robust LP was studied based on the concept of *weighted linear prediction* (WLP). WLP is a method for computing all-pole models of speech by temporally weighting the square of the residual in model parameter optimisation. Following the notations used by Ma et al. [1993], the residual energy of the pth-order WLP-model is expressed as

$$E = \sum_{n=n_1}^{n_2} e_n^2 W_n = \sum_{n=n_1}^{n_2} \left(s_n - \sum_{k=1}^{p} a_k s_{n-k} \right)^2 W_n, \tag{3.3}$$

where e_n is the residual, W_n is the temporal weight function, and the residual energy is minimised in the time span between n_1 and n_2. In the case of the autocorrelation method, $n_1 = 1$ and $n_2 = N + p$, and the signal is assumed to be zero outside the interval $[1, N]$. The optimal filter coefficients can be determined by setting the partial derivatives of Eq. 3.3 with respect to each a_k to zero. This results in the WLP normal equations

$$\sum_{k=1}^{p} a_k \sum_{n=n_1}^{n_2} W_n s_{n-k} s_{n-i} = \sum_{n=n_1}^{n_2} W_n s_n s_{n-i}, \qquad 1 \leq i \leq p. \tag{3.4}$$

Note that conventional LP is obtained as a special case of WLP: if W_n is chosen as a finite nonzero constant for all n, it becomes a multiplier of both

sides of Eq. 3.4 and cancels out leaving the LP normal equations [Makhoul, 1975]. Eq. 3.4 can also be expressed in matrix form as

$$\left(\sum_{n=n_1}^{n_2} W_n \mathbf{s}_n \mathbf{s}_n^T \right) \mathbf{a} = \sum_{n=n_1}^{n_2} W_n s_n \mathbf{s}_n, \tag{3.5}$$

where $\mathbf{a} = [a_1, a_2, \ldots, a_p]^T$ and $\mathbf{s}_n = [s_{n-1}, s_{n-2}, \ldots, s_{n-p}]^T$.

In the study by Ma et al. [1993], temporal weighting was computed from the (noisy) speech signal by using the short-time energy (STE) function

$$W_n = \sum_{i=n-M+1-k}^{n-k} s_i^2, \tag{3.6}$$

where the length of the energy window is denoted by M and the delay of the energy envelope by k. With STE as the weighting function, WLP was proposed by Ma et al. [1993] as an improved linear predictive method based on emphasising those samples that fit the underlying speech production model well. The original WLP formulation, however, did not guarantee stability of the resulting all-pole models, hence restricting its use. This might have been the reason why the potential idea proposed in Ma et al. [1993] has remained largely unnoticed in the speech science community.

The idea of WLP was, however, revived recently in the work of Magi and his co-authors. This work is based on emphasising the effects of digital speech samples located in the glottal closed phase, known to be associated with speech samples of large amplitude, when computing linear predictive models. This temporal emphasis is justified by two rationales. Firstly, it can be argued that these large-amplitude segments of speech are less vulnerable to have been corrupted by stationary additive noise than segments consisting of samples of smaller amplitude. Secondly, there is plenty of evidence in speech science indicating that formants extracted during the closed phase of the glottal cycle are more prominent than those computed during the glottal open phase, due to the absence of sub-glottal coupling [e.g. O'Shaughnessy, 1987]. Hence, by emphasising the contribution of the samples during the glottal closed phase, one is expected to obtain spectral

models which show better modelling of the main phonemic cues of speech sounds, the formants, in noisy conditions.

The use of temporal weighting in the optimisation of linear predictive methods was studied in study II. This work proposes a new concept, *stabilised weighted linear prediction* (SWLP), which yields all-pole models whose general performance can be adjusted by properly choosing the length of the STE window. By choosing a large M value in Eq. 3.6, the SWLP spectra become similar to those obtained by the conventional LP analysis. A small value of M, on the other hand, results in SWLP filters, similar to those computed by the minimum variance distortionless response (MVDR), a method that has recently [e.g. Yapanel and Hansen, 2003, Dharanipragada et al., 2007] attracted increasing interests in the speech recognition community due to its favourable noise-robustness properties in the context of the most widely used feature extraction method, the mel-frequency cepstral coefficients (MFCCs) [Rabiner and Juang, 1993]. The new SWLP method is a promising spectral modelling algorithm because it produces stable all-pole filters whose spectral envelopes are either smooth or having large dynamics, depending on the choice of the length of the STE window.

SWLP has already been used in speech recognition applications to replace the standard FFT-based spectral estimation in the process of computing the MFCCs. Isolated word recognition experiments have indicated superior performance for the proposed SWLP method in comparison to current spectral modelling techniques, including the previously proposed MVDR, in several different types of noise corruption scenarios [Pohjalainen et al., 2008]. SWLP has recently been found to improve robustness in large vocabulary continuous speech recognition, as well [Kallasjoki et al., 2009]. These results encourage the application of SWLP in various speech signal classification and recognition tasks.

Chapter 4

Glottal inverse filtering

One of the applications for linear predictive techniques is glottal inverse filtering (GIF). This is a traditional area of speech science which aims to estimate the source of voiced speech, the glottal volume velocity waveform, either from the speech pressure signal recorded by a microphone in the free field or from the oral flow captured by a specially designed pneumotachograph mask, also known as the Rothenberg mask [Rothenberg, 1973]. Research on GIF methods has been motivated for the past four decades mainly by the needs of the basic research aiming to improve our understanding about the functioning of the human voice production mechanism. However, the GIF methods have also been used, especially in the past ten years, in a wide range of applications such as speech synthesis [Raitio et al., 2008], speaker recognition [Plumpe et al., 1999], classification of vocal emotions [Cummings and Clements, 1995, Airas and Alku, 2006], voice conversion [Childers, 1995], and analysis of occupational voices [Vilkman, 2004, Lehto et al., 2008].

The basic principle of GIF is straightforward and based on Fant's source-filter theory of speech production [Fant, 1970]. According to this fundamental theory, the production of (voiced) speech can be considered a linear system as $S(z) = G(z)V(z)L(z)$, where $S(z)$ is the (known) speech signal, $G(z)$ is the (unknown) glottal flow, and $V(z)$ and $L(z)$ correspond to the

transfer functions of the vocal tract and lip radiation effect, respectively. If the vocal tract and the lip radiation effects can be estimated from speech, the desired output, the excitation signal of speech, can be computed by cancelling the effects of the tract and lip radiation by filtering the speech sound through the corresponding inverse models: $G(z) = S(z)V^{-1}(z)L^{-1}(z)$. The lip radiation effect can be estimated with reasonable accuracy as a simple first order differentiator for low frequencies [Flanagan, 1972]. Therefore, the essential problem in GIF is the estimation of the vocal tract transfer function $V(z)$.

Since the introduction of the idea of inverse filtering (IF) by Miller [1959], many different IF methods have been developed. The methods developed are greatly different because, for example, some of them need user adjustments in defining the settings of the vocal tract resonances [e.g. Price, 1989, Sundberg et al., 2005] while others are completely automatic [e.g. Veeneman and BeMent, 1985]. From the methodological point of view, the techniques developed can be categorised based on how the effect of the glottal source is taken into account in the estimation of the vocal tract in the underlying IF method. From this perspective, there are, firstly, methods [e.g. Alku, 1992] that are based on the gross estimation of the glottal contribution during both the closed and open phase of the glottal pulse using all-pole modelling. Secondly, the use of a joint optimisation of the glottal flow and vocal tract is possible based on synthetic, pre-defined models of the glottal flow [e.g. Fröhlich et al., 2001, Fu and Murphy, 2006]. Thirdly, it is possible to estimate the glottal flow using closed-phase covariance analysis [Strube, 1974, Wong et al., 1979]. This is based on the assumption that there is no contribution from the glottal source to the vocal tract during the closed phase of the vocal fold vibration cycle. After identification of the closed phase, LP with the covariance criterion is computed to get a parametric, pth order inverse model for the vocal tract:

$$V(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}. \tag{4.1}$$

Closed-phase (CP) covariance analysis is among the most widely used glot-

tal inverse filtering techniques. Since the original presentation of the method by Strube (1974), the CP method has been used as a means to estimate the glottal flow, for instance, in the analysis of the phonation type [Childers and Ahn, 1995], prosodic features of connected speech [Strik and Boves, 1992], vocal emotions [Cummings and Clements, 1995], source-tract interaction [Childers and Wong, 1994], singing [Arroabarren and Carlosena, 2004], and speaker identification [Plumpe et al., 1999]. Despite its prevalence the CP analysis is known to suffer from methodological shortcomings. In particular, there are several studies indicating that the glottal flow estimates computed with the CP analysis vary greatly depending on the position of the covariance frame [e.g. Larar et al., 1985, Veeneman and BeMent, 1985, Yegnanarayana and Veldhuis, 1998, Riegelsberger and Krishnamurthy, 1993]. Given the fundamental assumption of the method, that is, the computation of the vocal tract model during an excitation-free time span, this undesirable feature of the CP analysis is understandable. The true length of the glottal closed phase is typically short, which implies that the amount of data used to define the parametric model of the vocal tract with the covariance analysis is sparse. If the position of this kind of a short data frame is misaligned, the resulting linear predictive filter typically fails to model the vocal tract resonances, which might result in severe distortion of the glottal flow estimates. The misalignment of the covariance frame results typically in distortion of the glottal pulses by a sharp component, called "jags" by Wong et al. [1979] (see Fig. 4.1). This distortion is explained by the occurrence of roots of the linear predictive vocal tract model either on the positive real axis of the $\mathbb{Z}$-domain or at low frequencies. While these root positions are correct in terms of the MSE-type optimisation criterion used in linear predictive analysis, they are difficult to be interpreted by the Fant's source-filter theory of vowel production.

In the present thesis (studies VI and VII), the idea of constrained linear prediction is proposed in order to alleviate problems caused by using conventional LP with short data frames in defining vocal tract models. The new approach implies imposing a pre-defined value for the gain of the linear predictive inverse filter at two frequencies: either at $\omega = 0$ or at $\omega = \pi$. By denoting the transfer function of a pth order linear predictive inverse filter

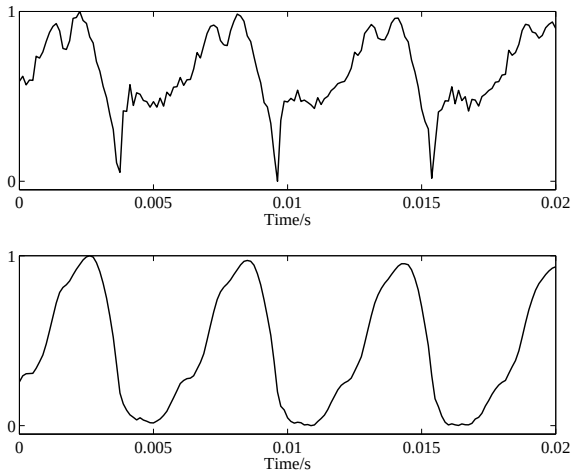


Figure 4.1: Examples of glottal flows estimated by the closed phase covariance analysis from the vowel /a/ produced by a male speaker. Vocal tract transfer function was estimated with the conventional LP (upper panel) and with the DC-constrained linear prediction (lower panel). The sharp edges, the "jags", are easily seen at glottal closure instants in the upper panel.

by $C(z) = \sum_{k=0}^{p} c_k z^{-k}$, where $c_0 = 1$, the following equations for the filter gain can be easily written at the two frequencies:

$$\begin{aligned} C(e^{i0}) = C(1) &= \sum_{k=0}^{p} c_k = l_{\mathrm{DC}} \\ C(e^{i\pi}) = C(-1) &= \sum_{k=0}^{p} c_k (-1)^k = l_{\pi}. \end{aligned} \tag{4.2}$$

With these constraints, the optimisation of the linear prediction results in the following normal equations:

$$\mathbf{c} = \mathbf{p}^{-1}\Gamma \left(\Gamma^T \mathbf{p}^{-1}\Gamma\right)^{-1} \mathbf{b} \tag{4.3}$$

, where the covariance matrix $\mathbf{p}$ is computed from the speech signal vector $\mathbf{x}_n$ during the closed phase consisting of N samples as $\mathbf{p} = \sum_{n=0}^{N-1} \mathbf{x}_n \mathbf{x}_n^T \in \mathbb{R}^{(p+1)\times(p+1)}$. In Eq. 4.3 above, Γ is a $(p+1) \times 2$ constraint matrix defined as

$$\Gamma = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 1 \end{bmatrix}^T \tag{4.4}$$

when the constraint is imposed at $\omega = 0$. In the case the gain of the optimal inverse filter is constrained at $\omega = \pi$ this matrix is defined as

$$\Gamma = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & -1 & 1 & -1 & \dots & \pm 1 \end{bmatrix}^T. \tag{4.5}$$

The positive real value defining the gain is given in matrix $\mathbf{b}$, which is defined as $\mathbf{b} = [1\, l_{\mathrm{DC}}]^T$ and $\mathbf{b} = [1\, l_{\pi}]^T$ when the constraint is imposed on $\omega = 0$ or at $\omega = \pi$, respectively. It is also possible to impose both of the constraints simultaneously. In this case, the matrices are defined as

$$\Gamma = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & -1 & 1 & -1 & \dots & \pm 1 \end{bmatrix}^T \tag{4.6}$$

and

$$\mathbf{b} = [1\, l_{\mathrm{DC}}\, l_{\pi}]^T. \tag{4.7}$$

It is shown in this thesis that the proposed idea of imposing constraints especially at $\omega = 0$ can be used to compute such AR models for the vocal tract that are less prone to include poles in positions of the $\mathbb{Z}$-domain that are difficult to interpret from the point of view of the classical source-filter theory of vowel production (e.g. vocal tract roots are located on the positive real axis). Hence, the proposed technique can be used to reduce the vulnerability of the CP analysis to the extraction of the covariance frame position. Examples of this phenomenon are shown in Figs. 4.1 and 4.2, which compare results of CP analyses when the vocal tract is determined with the conventional LP and with the proposed DC-constrained LP. It can be seen from the obtained glottal flow estimates in Fig. 4.1 that the false positions of the vocal tract roots determined by the conventional LP have resulted in extensive distortion of the flow pulses by "jags" that is, the waveform shows sharp edges especially at the instant of glottal closure. These artifacts, however, are clearly reduced in the glottal flow estimates computed by using the DC-constrained LP in modelling of the vocal tract. The spectra of the corresponding vocal tract models are shown in Fig. 4.2.

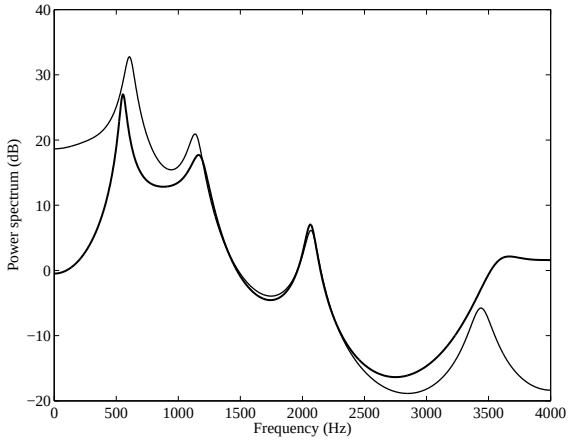


Figure 4.2: All-pole spectra of the vocal tract transfer functions used in the computation of the glottal flow estimates shown in Fig. 4.1: conventional LP(thin line) and DC-constrained LP (thick line).

Chapter 5

Summary of publications

5.1 Study I

This theoretical contribution presents new proofs to the root loci of two linear predictive models. Namely, it shows that the roots of symmetric linear predictive models and eigenfilters are located on the unit circle. As a side result, new information is provided on the separation of zeros expressed in angles.

Symmetric linear predictive models are filters that minimise modelling error under the constraint that the first and last coefficient of the filter are set to unity. It is well known that such filters always have zeros on the unit circle and a proof thereof is presented in [Stoica and Nehorai, 1988].

Eigenfilters represent the eigenvectors of the autocorrelation matrix. They appear when minimising the modelling error under the condition of unit-norm filters. As the eigenvectors of Toeplitz matrices, their properties are well researched with respect to matrix algebra [e.g. Grenander and Szegö, 1958], and the unit circle property of the eigenfilters corresponding to the minimum and maximum eigenvalue is also well known [Makhoul, 1981].

The proofs presented in this contribution are based on symmetries of the Toeplitz matrix that represent the autocorrelation of the signal as well

as the fact that the autocorrelation matrix is always positive definite. By extracting a zero, or a complex conjugate pair of zeros, from the linear predictive model or the eigenfilter, it is possible to show that these zeros must remain on the unit circle.

As stated above, autocorrelation matrices are positive definite, but, as it turns out, autocorrelation matrices are in fact positive definite by a positive margin [Delsarte et al., 1987]. In other words, this imposes a stronger condition than the positive definite property. This property is a result of the symmetries specific to autocorrelation matrices. Namely, the autocorrelation lag can be stated as a matrix operation using a matrix known as a shift matrix. The shift matrix is a nil-potent matrix and it is a property regarding the numerical range of nil-potent matrices [Karaev, 2004] that provides the positive margin to the positive definite property of autocorrelation matrices. Finally, this positive margin can be directly linked to the angle between the zeros of both symmetric linear predictive models and eigenfilters.

The presented proofs are straightforward and simpler than antecedent proofs. In addition to the simplification of proofs, analysis of zero separation is a completely new, unanticipated result.

5.2 Study II

This article deals with a new method, stabilised weighted linear prediction (SWLP), developed for computing robust AR models for spectral modelling of speech. The starting point of the study is a previously known method, weighted linear prediction (WLP), which introduces the idea of applying temporal weighting of the square of the residual signal in LP. The original WLP algorithm, proposed by Ma et al. [1993], used short-time energy (STE) as a weighting function. In its original form, however, WLP does not guarantee the stability of the all-pole models, thereby hindering the use of the method in such applications where speech sound must be synthesised using the all-pole model.

Study II revisits the concept of WLP by introducing a modified com-

putation of the correlation matrix, which always leads to stable all-pole models. This new SWLP method is shown to yield all-pole models whose general performance can be adjusted by tuning the length of the STE window. The study compares the performance of SWLP, minimum variance distortionless response (MVDR), and conventional LP in spectral modelling of speech corrupted by additive noise. The comparisons are performed by computing, for each method, the logarithmic spectral differences between the all-pole spectra extracted from clean and noisy speech, with different segmental signal-to-noise ratio (SNR) levels.

According to the evaluation results, the proposed SWLP algorithm is the most robust method against zero-mean Gaussian noise and the robustness is largest for the variant of SWLP utilising a short STE window length. These findings are corroborated by a small listening test. Finally, SWLP is compared to other short-time spectral estimation methods (FFT, LP, MVDR) in isolated word recognition experiments using MFCC features. Recognition accuracy obtained by SWLP, in comparison to other short-time spectral estimation methods, improves already at moderate segmental SNR levels for sounds corrupted by zero-mean Gaussian noise. For realistic factory noise having low pass characteristics, the SWLP method improves the recognition results at segmental SNR levels below 0 dB.

5.3 Study III

Line spectrum pair (LSP) decomposition is a method developed for robust quantisation of linear predictive models. It was introduced by Itakura [1975] and its beneficial properties for quantisation were demonstrated by Soong and Juang [1984]. While its usefulness as a practical tool was greatly appreciated in speech coding [Kleijn and Paliwal, 1995], the full extent of the mathematical structure remained for long forgotten in more or less obscure publications. Study III collects the known results under a common notation and methodology, as well as presents some extensions to these results.

The three most fundamental properties of the zeros of the LSP polyno-

mials are that 1) the zeros are on the unit circle, 2) the zeros are distinct, and 3) the zeros corresponding to the symmetric and antisymmetric polynomials are interlaced. In this contribution, a more general set of interlacing properties, relating consecutive order polynomials, is presented and proved.

The LSP polynomials also appear as a basis of the Levinson recursion, which is most often used for the solution of the coefficients of the linear predictive model from the normal equations [Hayes, 1996]. Study III demonstrates the commonalities of LSP and Levinson recursion, as well as extensions to the latter, the split Levinson [Delsarte and Genin, 1986] and Krishna's algorithm [Krishna, 1988].

A surprising connection between LSP and constrained linear predictive models exists and was first presented by Kleijn et al. [2003] and Bäckström et al. [2004]. It has been shown that the LSP polynomials correspond to filters estimated from certain low-pass and high-pass filtered signals. The current contribution collects prior results regarding the stability of constrained models and their relation to LSP polynomials, as well as presents corresponding interlacing properties of the constrained models.

Study III presents for the first time both the properties of LSP polynomials and the proofs thereof in a complete and rigorous manner. This work is a review paper in nature, but presents also new interlacing properties as natural extensions to the given proofs.

5.4 Study IV

The line spectrum pair (LSP) decomposition, ubiquitous in speech coding, has three celebrated properties, namely, its zeros are on the unit circle, distinct and interlaced. Study V presents a stronger condition than the properties of distinct and interlacing zeros, namely, that the zeros of the LSP polynomials, the line spectral frequencies (LSFs) are separated by a positive margin when estimated from the autocorrelation matrix of a finite signal.

LSP is used in speech coding since it is robust to quantisation errors and the stability of the prediction model can be readily guaranteed. When using

LSP in quantisation, the linear predictive model is uniquely represented by the angles of the LSP polynomial zeros. Solution of the LSFs is not possible by analytical methods and thus numerical search methods have to be employed. However, LSFs arbitrarily close to each other represent problems to the numerical search methods.

The current paper provides a lower limit to the distance between LSFs in terms of the maximal root radius of the linear predictive model. In other words, the positive margin between the roots and the unit circle provides a positive margin between LSFs. The proof is based on evaluating the group delay of the LSP polynomials, whereby a relation between the maximal root radius and maximal group delay is obtained. The final result is obtained by applying the mean value theorem of differentials.

The presented results are useful in designing new speech coding methods, whereby knowledge of the location and distribution of LSFs can assist in obtaining computationally effective algorithms. Alternatively, by modifying the LSFs, it is possible to constrain the root radius of the linear predictive model, thus smoothing its spectral envelope when desired.

5.5 Study V

White-noise correction is a method used in speech coding applications employing linear predictive modelling in order to ensure stability of the model in the presence of numerical round-off errors. Study V provides a mathematical analysis of the method and presents a relation between the location of the zeros of the linear predictive filter and level of white-noise correction.

While linear predictive modelling generally guarantees stable models, in practical applications the inevitable numerical round-off errors jeopardise this property. In effect, round-off errors might make the autocorrelation matrix ill-conditioned or indeed compromise its positive definitive property. White-noise correction is equivalent to a regularisation method that increases all the eigenvalues of the autocorrelation matrix by a constant regularisation coefficient, thus making the matrix positive definite by a positive margin. In terms of signal processing, white noise correction corresponds

to adding uncorrelated white noise to the input signal.

The current work provides a relation between the regularisation coefficient and the stability radius of the linear predictive model. Specifically, the larger the regularisation coefficient, that much larger is the margin between the unit circle and the zeros of the linear predictive filter. The proof is based on the fact that the autocorrelation values at different lags can be represented in terms of a shift matrix and a vector consisting of the original signal. The shift matrix is a nil-potent matrix and the numerical range of such matrices depends on the power of nil-potency of the shift matrix [Delsarte and Genin, 1986]. This property of nil-potent matrices is the basis of a proof of the convergence rate for autocorrelation coefficients depending on the lag. Using this result and by extracting a zero from the linear predictive model, a formula for the stability radius of the linear predictive model under white-noise correction can be readily derived.

Study IV presents properties of white noise correction, a method widely used in speech coding, whose theoretical properties had not previously been rigorously studied. The level of white noise correction (corresponding to the level of regularisation and the level of added white noise) is shown to correspond to the margin inside the unit circle within which the model zeros must remain. As a notable side result, the current work also presents a formula for the convergence rate of the autocorrelation of a signal.

5.6 Study VI

This study deals with the estimation of the glottal volume velocity waveforms with an inverse filtering approach, the closed-phase (CP) covariance analysis. CP analysis is one of the most widely used glottal inverse filtering methods, and its original form, proposed by studies of Strube [1974] and Wong et al. [1979], uses conventional LP with the covariance criterion in the modelling of the vocal tract. The data frame of the covariance analysis is located so as to cover those samples during which there is no excitation from the source into the tract, that is the closed phase of the glottal cycle. Since the length of the closed phase is typically short, the

resulting all-pole model is highly sensitive to the position of the extracted frame. Even a minor change of the frame position might greatly affect the $\mathbb{Z}$-domain locations of the roots of the all-pole model given by LP. This undesirable feature of the conventional CP analysis typically results in vocal tract models which have roots, both real and complex, at low frequencies or roots that are located outside the unit circle. These kinds of false root locations, in turn, result in distortion of the glottal flow estimates which is typically seen as unnatural peaks at the instant of glottal closure, the so-called "jags", or as increased formant ripple during the closed phase.

The study proposes an improved version of the CP analysis based on a combination of two algorithmic issues. Firstly, and most importantly, a constraint is imposed on the DC gain of the inverse filter prior to the optimisation of the coefficients. With this constraint, linear predictive analysis is more prone to give vocal tract models that can be justified from the point of view of the source-filter theory of vowel production, that is they show complex conjugate roots in the vicinity of formant regions rather than unrealistic resonances at low frequencies. The study shows how this idea of imposing a constraint on the DC gain of a linear predictive inverse filter can be mathematically presented and optimised. Secondly, the new CP method utilises an inverse filter that is minimum phase, a property that is not typically used in glottal inverse filtering. The method is evaluated using synthetic vowels produced by physical modelling, and natural speech. The results show that the algorithm improves the performance of the CP-type inverse filtering and its robustness with respect to the covariance frame position.

5.7 Study VII

This study is a sequel to study VI and was published in a special issue in honour of Swedish speech scientist Dr. Jan Gauffin. The idea of imposing constraints on linear prediction is revisited by formulating the problem for two frequencies: angular frequencies $\omega = 0$ (that is DC) and $\omega = \pi$ (that is half the sampling frequency). With these constraints, linear predictive

analysis is more prone to give vocal tract models in the CP analysis whose roots are located in the formant region rather than in unrealistic positions at low frequencies. Instead of using isolated vowels produced using sustained phonation as in study VI, this investigation uses more challenging utterances recorded from continuous speech to test the performance of CP analysis based on DC-constrained LP. The results show that the use of DC-constrained LP gives all-pole models of the vocal tract that are less vulnerable to errors that are made in the detection of the closed phase location.

Chapter 6

Conclusions

This thesis studies LP, one of the most prevalent approaches to parametric spectral modelling of speech, from a mathematical perspective by presenting novel modifications to the well-known classical LP analysis. The topic involves three approaches into the research problem: (1) spectral models utilizing the LSP decomposition of the LP model, (2) spectral modelling and feature extraction of speech based on weighted LP, and (3) constrained LP in vocal tract modelling of glottal inverse filtering. The main results obtained in these three areas are summarised as follows.

1. Line Spectrum Pair (LSP) decomposition is widely used in speech coding for the quantisation of the LP parameters representing the spectral envelope of a signal. In terms of quantisation, LSP has a number of beneficial properties such as robustness to quantisation noise and guaranteed model stability. However, although widely used, there were largely unresearched areas in the theory of LSP. The current work represents an effort in filling in the blanks. Previous results are collected from scattered sources into a review paper; interlacing properties are described in their full extent, the connection to Levinson recursion is explained, and the filtering interpretation of LSP is demonstrated. Two new results are also presented. Firstly, a relationship between the root radius of the linear predictive model and

the separation of line spectral frequencies (LSFs) is demonstrated. Secondly, the effect of white-noise correction is studied in relation to both the linear predictive model as well as the LSFs. These results demonstrate the rich structure of the LSP decomposition and provide a theoretical basis for the development of future speech coding methods.

2. The idea of using temporal weighting of the squared residual in optimising linear predictive models was proposed in the early 1990's, but during the past 15 years this idea seems to have been ignored by the speech research community. However, as shown in this thesis, this potent idea can be used to improve the noise robustness of LP analysis and is thus well-suited for applications where noise-corrupted speech must be modelled. This study revives the previously proposed idea of weighted LP and, importantly, proposes a novel variant, SWLP, which is guaranteed to result in stable all-pole synthesis filters. Unlike conventional LP, the behaviour of SWLP can be controlled by selecting the length of the short-time energy window used in the computation of the weighting function. Hence, a new all-pole modelling method has been developed in this thesis. In comparison to conventional LP, this method features improved robustness in the presence of additive noise. At the same time, it shares the important practical benefit of conventional autocorrelation LP, that is the method can be used in all linear predictive applications that use LP for synthesis and thus call for stability of the resulting all-pole filter.

3. The closed-phase (CP) analysis is a widely-used glottal inverse filtering algorithm that uses linear predictive analysis with the covariance criterion in order to estimate the vocal tract transfer function. It has been reported in various previous studies that the CP analysis is highly vulnerable to the correct positioning of the covariance frame. In order to alleviate this problem, this thesis proposes the idea of constrained linear prediction; the linear predictive filter is optimised by imposing certain *a priori* values to the filter gain at two frequencies

(DC and π). With this idea, the root locations of the linear predictive vocal tract model are less prone to occur in such positions that are unrealistic from the point of view of the classical source-filter theory of voice production. The glottal inverse filtering studies conducted in this thesis show that the use of constrained LP makes CP analysis less vulnerable to errors that are made in the detection of the closed phase location.

Bibliography

M. Airas and P. Alku. Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalized amplitude quotient. *Phonetica*, 63:26–46, 2006.

P. Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11:109–118, 1992.

I. Arroabarren and A. Carlosena. Vibrato in singing voice: the link between source-filter and sinusoidal models. *EURASIP Journal on Applied Signal Processing*, (7):1007–1020, 2004.

B. S. Atal and S. L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, 50:637–655, 1970.

T. Bäckström, P. Alku, T. Paatero, and W. B. Kleijn. A time-domain interpretation for the LSP-decomposition. *IEEE Transactions on Speech and Audio Processing*, 12(6):554–560, 2004.

D. G. Childers. Glottal source modeling for voice conversion. *Speech Communication*, 16:127–138, 1995.

D. G. Childers and C. Ahn. Modeling the glottal volume-velocity waveform for three voice types. *The Journal of the Acoustical Society of America*, 97:505–519, 1995.

D. G. Childers and C. F. Wong. Measuring and modeling vocal source-tract interaction. *IEEE Transactions on Biomedical Engineering*, 41(7): 663–671, 1994.

W. C. Chu. *Speech Coding Algorithms – Foundation and Evolution of Standardized Coders*. Wiley, 2003.

R. V. Cox. Speech coding standards. In W. B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis*, pages 49–78. Elsevier, 1995.

K. E. Cummings and M. A. Clements. Analysis of the glottal excitation of emotionally styled and stressed speech. *The Journal of the Acoustical Society of America*, 98:88–98, 1995.

P. Delsarte and Y. V. Genin. The split Levinson algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(3):470–478, 1986.

P. Delsarte, Y. Genin, and Y. Kamp. Stability of linear predictors and numerical range of a linear operator (corresp.). *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(3):412–415, 1987.

S. Dharanipragada, U. Yapanel, and B. Rao. Robust feature extraction for continuous speech recognition using the MVDR spectrum estimation method. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):224–234, 2007.

A. El-Jaroudi and J. Makhoul. Discrete all-pole modeling. *IEEE Transactions on Signal Processing*, 39:411–423, 1991.

ETSI. *Recommendation GSM 06.10; GSM full rate speech transcoding*. ETSI, 1992a.

ETSI. *Digital cellular telecommunications system (Phase 2); Enhanced full rate (EFR) speech transcoding, GSM 06.60 version 4.0.0*. ETSI, August 1997b.

ETSI. *3rd Generation Partnership Project: AMR Wideband Speech Codec, 3GPP TS 26.190 (Release 5)*. ETSI, March 2001c.

G. Fant. *Acoustic Theory of Speech Production*. Mouton, 1970.

J. L. Flanagan. *Speech Analysis Synthesis and Perception*. Springer Verlag, 1972.

M. Fröhlich, D. Michaelis, and H.W. Strube. SIM – simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals. *The Journal of the Acoustical Society of America*, 110:479–488, 2001.

Q. Fu and P. Murphy. Robust glottal source estimation based on joint source-filter model optimization. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):492–501, 2006.

J. Gibson. Speech coding methods, standards, and applications. *IEEE Circuits and Systems Magazine*, 5(4):30–49, 2005.

A. H. Gray Jr. and J. D. Markel. Quantization and bit allocation in speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(6):459–473, 1976.

U. Grenander and G. Szegö. *Toeplitz Forms and Their Applications*. American Mathematical Society, 1958.

M. H. Hayes. *Statistical Digital Signal Processing and Modeling*. J. Wiley & Sons, Inc., 1996.

H. Hu. Linear prediction analysis of speech signals in presence of white gaussian noise with unknown variance. *IEE Proceedings – Vision, Image and Signal Processing*, 145:303–308, 1998a.

H. Hu. Robust linear prediction of speech signals based on orthogonal framework. *Electronics Letters*, 34:1385–1386, 1998b.

F. Itakura. Line spectrum representation of linear predictive coefficients of speech signals. *The Journal of The Acoustical Society of America*, 57, suppl. 1:35, 1975.

ITU-T. *Recommendation G.729-Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP).* ITU-T, March 1996.

H. Kallasjoki, K. Palomäki, C. Magi, P. Alku, and M. Kurimo. Noise robust lvcsr feature extraction based on stabilized weighted linear prediction. In *Proceedings of the 13th International Conference on Speech and Computer (SPECOM 2009)*, St. Petersburg, Russia, 2009.

G. S. Kang and L. J. Fransen. Application of line-spectrum pairs to low-bit-rate speech encoders. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95)*, volume 1, pages 244–247, Dallas, Texas, USA, 1985.

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

S. Kay. Improvement of autoregressive spectral estimates in the presence of noise. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'78)*, pages 357–360, Tulsa, Oklahoma, USA, 1978.

S. Kay. The effects of noise on the autoregressive spectral estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(5):478–485, 1979.

S. Kay. Noise compensation for autoregressive spectral estimate. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28:292–303, 1980.

S. Kay and S. Marple. Spectrum analysis – a modern perspective. *Proceedings of the IEEE*, 69(11):1380–1419, 1981.

W. B. Kleijn, T. Bäckström, and P. Alku. On line spectral frequencies. *IEEE Signal Processing Letters*, 10(3):75–77, 2003.

W.B. Kleijn and K.K. Paliwal, editors. *Speech Coding and Synthesis*. Elsevier, 1995.

H. Krishna. New split Levinson, Schur and lattice algorithms for digital signal processing. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'88)*, volume 3, pages 1640–1642, New York, NY, USA, 1988.

J. Larar, Y. Alsaka, and D. Childers. Variability in closed phase analysis of speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'85)*, volume 10, pages 1089–1092, Tampa, Florida, USA, 1985.

C. Lee. On robust linear prediction of speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36:642–650, 1988.

L. Lehto, L. Laaksonen, E. Vilkman, and P. Alku. Changes in objective acoustic measurements and subjective voice complaints in call-center customer-service advisors during one working day. *Journal of Voice*, 22: 164–177, 2008.

C. Ma, Y. Kamp, and L. F. Willems. Robust signal selection for linear prediction analysis of voiced speech. *Speech Communication*, 12(1):69–81, 1993.

J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(5):561–580, 1975.

J. Makhoul. On the eigenvectors of symmetric Toeplitz matrices. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(4):868–872, 1981.

J.D. Markel and A.H. Gray Jr. *Linear Prediction of Speech*. Springer Verlag, 1976.

R. Miller. Nature of the vocal cord wave. *The Journal of the Acoustical Society of America*, 31:667–677, 1959.

M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach, F. Nagel, and J. Robilliard. A novel scheme for low bit rate unified speech and audio coding – MPEG RM0. In *Proceedings of the 126th AES Convention*, Munich, Germany, 2009.

D. O'Shaughnessy. *Speech Communication – Human and Machine*. Addison-Wesley, 1987.

M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Transactions on Speech and Audio Processing*, 7(5):569–586, 1999.

J. Pohjalainen, C. Magi, and P. Alku. Enhancing noise robustness in automatic speech recognition using stabilized weighted linear prediction (SWLP). In *Proceedings ISCA Tutorial and Research Workshop (ITRW) on "Speech Analysis and Processing for Knowledge Discovery*, Aalborg, Denmark, 2008.

P. J. Price. Male and female voice source characteristics: Inverse filtering results. *Speech Communication*, 8(3):261–277, 1989.

L. R. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku. HMM-based finnish text-to-speech system utilizing glottal inverse filtering. In *Proceedings Interspeech 2008*, Brisbane, Australia, 2008.

E. L. Riegelsberger and A. K. Krishnamurthy. Glottal source estimation: Methods of applying the LF-model to inverse filtering. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'93)*, volume 2, pages 542–545, Minneapolis, Minnesota, USA, 1993.

M. Rothenberg. A new inverse filtering technique for deriving the glottal airflow waveform during voicing. *The Journal of The Acoustical Society of America*, 53:1632–1645, 1973.

H. W. Schüssler. A stability theorem for discrete systems. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(1):87–89, 1976.

F. K. Soong and B.-H. Juang. Line spectrum pair (LSP) and speech data compression. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'84)*, volume 1, pages 1.10.1–1.10.4, San Diego, CA, USA, 1984.

P. Stoica and A. Nehorai. The poles of symmetric linear prediction models lie on the unit circle. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34:1344–1346, 1986.

P. Stoica and A. Nehorai. On linear prediction models constrained to have unit-modulus poles and their use for sinusoidal frequency estimation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(6): 940–942, 1988.

H. Strik and L. Boves. On the relation between voice source parameters and prosodic features in connected speech. *Speech Communication*, 11 (2-3):167–174, 1992.

H.W. Strube. Determination of the instant of glottal closure from the speech wave. *The Journal of the Acoustical Society of America*, 56:1625–1629, 1974.

J. Sundberg, E. Fahlstedt, and A. Morell. Effects on the glottal voice source of vocal loudness variation in untrained female and male voices. *The Journal of the Acoustical Society of America*, 117:879–885, 2005.

D. E. Veeneman and S. L BeMent. Automatic glottal inverse filtering from speech and electroglottographic signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):369–377, 1985.

E. Vilkman. Occupational safety and health aspects of voice and speech professions. *Folia Phoniatrica et Logopaedica*, 56:220–253, 2004.

R. Viswanathan and J. Makhoul. Quantization properties of transmission parameters in linear predictive systems. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(3):309–321, 1975.

D. Y. Wong, J. D. Markel, and A. H. Gray. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27:350–355, 1979.

U. Yapanel and J. Hansen. A new perspective on feature extraction for robust in-vehicle speech recognition. In *Proceedings Interspeech*, pages 1281–1284, 2003.

B. Yegnanarayana and R. N. J. Veldhuis. Extraction of vocal-tract system characteristics from speech signals. *IEEE Transactions on Speech and Audio Processing*, 6(4):313–327, 1998.

Study I

Carlo Magi, Tom Bäckström, Paavo Alku. "Simple proofs of root locations of two symmetric linear prediction models", Signal Processing, Vol. 88, Issue 7, pp. 1894-1897, 2008.

Available online at www.sciencedirect.com
ScienceDirect
Signal Processing 88 (2008) 1894–1897
SIGNAL PROCESSING
www.elsevier.com/locate/sigpro

Fast communication

Simple proofs of root locations of two symmetric linear prediction models

Carlo Magi[☒,1], Tom Bäckström[*], Paavo Alku

Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology (TKK), P.O. Box 3000, FI-02015 TKK, Finland

Received 4 December 2007; received in revised form 17 January 2008; accepted 19 January 2008
Available online 30 January 2008

Abstract

This paper gives simple proofs of the root locations of two linear predictive methods: the symmetric linear prediction model and the eigenfilter model corresponding to the minimal or maximal simple eigenvalues of an autocorrelation matrix. The roots of both symmetric models are proved to lie on the unit circle. Differently from previous proofs, the approach used in the present study also shows, based on the properties of the autocorrelation sequence, that the root angles of the symmetric linear prediction model are limited to occur within a certain interval. Moreover, eigenfilters corresponding to the minimum or maximum eigenvalue of an autocorrelation matrix that have multiplicity greater than unity are also studied. It turns out that it is possible to characterise the whole space spanned by the eigenvectors corresponding to the multiple eigenvalues by a single symmetric/antisymmetric eigenvector of the principal diagonal sub-block of the autocorrelation matrix having all the roots on the unit circle.

Keywords: Symmetric linear prediction; Eigenfilter; Toeplitz matrix

1. Introduction

The spectral envelope of speech signals originates from two major physiological parts of the human voice production mechanism, the glottal excitation and the vocal tract. The former is the source behind the over-all spectral envelope of speech, and it is characterised for voiced sounds as a low-pass process. The latter is the cause of the local resonances of the spectral envelope of speech sounds, the formants. It is well known that linear prediction (LP) constitutes an effective means to estimate the spectral envelope of speech signals by representing this information by a small number of parameters in the form of pth order all-pole filter [1]. Therefore, LP has become the most widely used spectral method to model speech and it has an established role in several speech technology applications, especially in low-bit rate voice coding [2]. Stability criteria of LP have been studied widely [1,3] and it is well known that the autocorrelation criterion guarantees the minimum-phase property of the LP inverse filter, a feature that is a pre-requisite for most speech technology applications of LP.

[*]Corresponding author. Tel.: +358 9 451 5843; fax: +358 9 460 224.
E-mail addresses: tom.backstrom@tkk.fi (T. Bäckström), paavo.alku@tkk.fi (P. Alku).
[☒]Passed away 11 February 2008.
[1]Supported by Academy of Finland (project no: 205962 and 111848) and TKK.

0165-1684/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2008.01.019

While the conventional LP analysis computed with the autocorrelation criterion always gives a minimum-phase inverse filter, there also exist *constrained* linear predictive models in which the impulse response of the inverse filter is *symmetric*. Two such variants of LP are represented by symmetric LP [4,5] and the eigenfilter model corresponding to the minimal and maximal simple eigenvalue of the autocorrelation matrix [5,6]. While symmetric models have been used in such applications as in the estimation of sinusoidals [5,7] their use in speech technology has remained limited. The rationale behind this is obvious: the over-all structure of speech sounds calls for using such filters that are both stable and capable of mimicing the spectral envelope of speech, which is typically of a low-pass nature and comprises local resonances of finite bandwidth. Due to known z-domain properties of symmetric impulse responses [8] this kind of spectral behaviour cannot be implemented with a symmetric LP model. Therefore, the use of symmetric linear predictive models are not feasible in spectral modelling of speech if the underlying spectrum is to be modelled in a similar manner as in the conventional LP, that is by a single all-pole filter computed directly from the speech signal. Contrary to this prevailing understanding, symmetric linear predictive methods *can* be successfully used also in spectral modelling of speech if they are implemented as a parallel structure, as indicated by a recent study [9]. This study is based on a previous investigation by Stoica and Nehorai [5], who showed that conventional autocorrelation LP can be interpreted with the help of two symmetric linear predictive polynomials. Unfortunately, the important relationship between the conventional and the symmetric LP presented by Stoica and Nehorai has remained largely unnoticed in speech science.

The present paper addresses mathematical properties of two symmetric linear predictive models optimised with the autocorrelation criterion, which, in principle, can be used in the implementation of stable all-pole spectral models of speech based on the similar parallel structure as used by Stoica and Nehorai [5] and Alku and Bäckström [9]. More specifically, a new proof is given to show that the roots of the symmetric LP are located on the unit circle. In contrast to the previous proofs [4,5,10], the root angles of the symmetric LP are shown here to be limited to occur within a certain interval. It will also be proved that the roots of the eigenfilter corresponding to the maximal or minimal simple eigenvalue of the symmetric positive definite autocorrelation matrix lie on the unit circle. This property has been proved without considering the case of multiple eigenvalues [6,11–13]. Here, we give a detailed treatment to the case where the eigenfilter corresponds to the maximal or minimal *multiple* eigenvalue of the symmetric positive definite autocorrelation matrix.

2. Results

Consider the root locations of a monic symmetric LP inverse filter

$$A(z) = 1 + a_1 z^{-1} + \cdots + a_{p/2-1} z^{-p/2+1} + a_{p/2} z^{-p/2} + a_{p/2-1} z^{-p/2-1} + \cdots + a_1 z^{-p+1} + z^{-p}. \quad (1)$$

In order to make the notation simple, only even-order polynomials are considered here. The case of odd-order polynomials can be treated similarly.

The coefficient vector $\mathbf{a} = [1 \;\; a_1 \;\; \cdots \;\; a_{p/2} \;\; \cdots \;\; a_1 \;\; 1]^{\mathrm{T}}$ of the inverse filter of symmetric LP can be solved from the equation [5]:

$$\mathbf{R}_{p+1}\mathbf{a} = \gamma[1 \;\; 0 \;\; \cdots \;\; 0 \;\; 1]^{\mathrm{T}}, \quad (2)$$

where $\mathbf{R}_{p+1} = \mathrm{toep}\{r_0, \ldots, r_p\} \in \mathbb{R}^{(p+1)\times(p+1)}$ is the autocorrelation matrix of signal x_n and $r_k = \sum_{i=0}^{N-k-1} x_i x_{i+k} = \mathbf{x}^{\mathrm{T}}\mathbf{S}^k\mathbf{x}$. Here, N is the length of the signal vector $\mathbf{x} = [x_0 \;\; \cdots \;\; x_{N-1}]^{\mathrm{T}}$ and $\mathbf{S}_{ij} = \delta_{j-i+1}$ is the $N \times N$ down-shift matrix, while γ is the filter gain factor.

Theorem 1. *The zeros of the pth order inverse filter $A(z)$ of a symmetric LP model lie on the unit circle.*

Proof. First, factor out a general factor of the symmetric polynomial: $A(z) = (1 + \alpha z^{-1} + z^{-2})B(z)$. Then, the coefficient vector $\mathbf{a}$ can be written in factored form as $\mathbf{a} = \mathbf{B}\,[1 \;\; \alpha \;\; 1]^{\mathrm{T}}$, where $\mathbf{B} \in \mathbb{R}^{(p+1)\times 3}$ is the convolution matrix $\mathbf{B} = [\mathbf{b} \;\; \mathbf{S}\mathbf{b} \;\; \mathbf{S}^2\mathbf{b}]$. Here $\mathbf{b} = [b_0 \;\; \cdots \;\; b_{p/2-1} \;\; \cdots \;\; b_0 \;\; 0 \;\; 0]^{\mathrm{T}}$, where b_i are the coefficients of the symmetric polynomial $B(z)$ and $\mathbf{S}_{ij} = \delta_{j-i+1}$ is the $p+1 \times p+1$ down-shift matrix. Multiplying Eq. (2) from the left by $\mathbf{B}^{\mathrm{T}}$ yields

$$\mathbf{B}^{\mathrm{T}}\mathbf{R}_{p+1}\mathbf{B}[1 \;\; \alpha \;\; 1]^{\mathrm{T}} = \gamma[b_0 \;\; 0 \;\; b_0]^{\mathrm{T}}. \quad (3)$$

Then, $\mathbf{B}^{\mathrm{T}}\mathbf{R}_{p+1}\mathbf{B} = \mathrm{toep}\{\hat{r}_0, \hat{r}_1, \hat{r}_2\}$ is the autocorrelation matrix of the convolved signal $x_n * b_n$. Furthermore, it is positive definite and due to the Sylvester criterion[2] $\hat{r}_0^2 - \hat{r}_1^2 > 0$ so $|\hat{r}_0| > |\hat{r}_1|$. From

[2]For symmetric positive definite matrices, all of the leading principal minors are positive.

Eq. (3), $2\hat{r}_1 + \alpha\hat{r}_0 = 0$, which implies

$$|\alpha| = 2\frac{|\hat{r}_1|}{|\hat{r}_0|} < 2 \tag{4}$$

then the general factor $1 + \alpha z^{-1} + z^{-2}$ of the symmetric polynomial $A(z)$ has its roots on the unit circle concluding the proof. □

Corollary 1. *The symmetric LP model of order p and analysis frame length N can have roots* $e^{\pm i\omega_k}$ *only inside the interval*

$$\omega_k \in \pi\left[\frac{1}{N+p-1}, 1 - \frac{1}{N+p-1}\right]. \tag{5}$$

Proof. Since the symmetric positive definite Toeplitz matrix $\mathbf{B}^{\mathrm{T}}\mathbf{R}_{p+1}\mathbf{B}$ is the autocorrelation matrix of the convolved signal $c_n = x_n * b_n$, the autocorrelation sequence can be written as $\hat{r}_k = \mathbf{c}^{\mathrm{T}}\mathbf{S}^k\mathbf{c}$. Hence, the autocorrelation sequence belongs to the numerical range of the nilpotent down-shift matrix $\mathbf{S}^k \in \mathbb{R}^{(N+p-2)\times(N+p-2)}$ with power of nilpotency $\lceil (N+p-1)/k \rceil$, where $\lceil\cdot\rceil$ indicates rounding upward to the closest integer. The numerical range of matrix $\mathbf{S}^k$ is a circle with its centre at the origin and radius not exceeding $\cos(\pi/(N+p-1))$ [14]. From [15] the following inequality is obtained:

$$|\hat{r}_1| \leqslant \hat{r}_0 \cos\left(\frac{\pi}{N+p-1}\right). \tag{6}$$

Eq. (4) gives

$$\alpha \in \left[-2\cos\left(\frac{\pi}{N+p-1}\right), 2\cos\left(\frac{\pi}{N+p-1}\right)\right].$$

Finally, polynomial $1 + \alpha z^{-1} + z^{-2}$ has its roots on the unit circle at $e^{\pm i\omega_0}$, where $\omega_0 = \arccos(-\alpha/2)$. □

Note that Corollary 2 implies that the estimated model cannot estimate sinusoidals at certain frequencies. However, since the input signal can have components at all frequencies, the estimate must be biased. The same conclusion was drawn in [5].

Next, consider the root locations of the eigenfilters of autocorrelation matrix $\mathbf{R}_{p+1}$ corresponding to the simple minimum or maximum eigenvalue

$$V(z) = v_0 + v_1 z^{-1} + \cdots + v_{p/2-1} z^{-p/2+1} + v_{p/2} z^{-p/2} \pm (v_{p/2-1} z^{-p/2-1} + \cdots + v_1 z^{-p+1} + v_0 z^{-p}). \tag{7}$$

The coefficients v_i can be solved from the equation

$$\mathbf{R}_{p+1}\mathbf{v} = \lambda_0\mathbf{v}, \tag{8}$$

where $\mathbf{R}_{p+1} = \mathrm{toep}\{r_0, \ldots, r_p\} \in \mathbb{R}^{(p+1)\times(p+1)}$, λ_0 is the minimum or maximum eigenvalue, and $\mathbf{v} = [v_0 \ \cdots \ v_{p/2} \ \cdots \ \pm v_0]^{\mathrm{T}}$ is the corresponding eigenvector.

Theorem 2. *The zeros of the pth order eigenfilter $V(z)$ corresponding to the minimum or maximum simple eigenvalue of a symmetric positive definite Toeplitz matrix $\mathbf{R}_{p+1}$ lie on the unit circle.*

Proof. Start by factoring the polynomial as $V(z) = (1 - \alpha z^{-1})U(z)$, where $\alpha \neq 0$. The coefficient vector $\mathbf{v}$ can be written in factored form as $\mathbf{v} = \mathbf{U}[1 \ \ -\alpha]^{\mathrm{T}}$, where $\mathbf{U} \in \mathbb{C}^{(p+1)\times 2}$ is the convolution matrix $\mathbf{U} = [\mathbf{u} \ \mathbf{Su}]$, $\mathbf{u} = [u_0 \ \cdots \ u_{p-1} \ 0]^{\mathrm{T}}$, and $\mathbf{S}$ is as in Theorem 1. Multiplying Eq. (8) from the left by $\mathbf{U}^{\mathrm{T}}$ yields

$$\mathbf{U}^{\mathrm{T}}\mathbf{R}_{p+1}\mathbf{U}[1 \ \ -\alpha]^{\mathrm{T}} = \lambda_0\mathbf{U}^{\mathrm{T}}\mathbf{U}[1 \ \ -\alpha]^{\mathrm{T}}. \tag{9}$$

The symmetric positive (or negative) semidefinite matrix $\mathbf{\Delta}$ can then be formed:

$$\mathbf{\Delta} = \mathbf{U}^{\mathrm{T}}(\mathbf{R}_{p+1} - \lambda_0\mathbf{I})\mathbf{U} = \begin{bmatrix} \zeta & \vartheta \\ \vartheta & \zeta \end{bmatrix}, \tag{10}$$

whereby Eq. (9) can be written as

$$\mathbf{\Delta}[1 \ \ -\alpha]^{\mathrm{T}} = \mathbf{0}. \tag{11}$$

The numerical range of a symmetric matrix $\mathbf{R}_{p+1} - \lambda_0\mathbf{I}$ is a closed interval on the real axis whose endpoints are formed by the extreme eigenvalues of $\mathbf{R}_{p+1} - \lambda_0\mathbf{I}$. Therefore, and since λ_0 is the smallest or greatest simple eigenvalue of the matrix $\mathbf{R}_{p+1}$, from $\zeta = \mathbf{u}^{\mathrm{T}}(\mathbf{R}_{p+1} - \lambda_0\mathbf{I})\mathbf{u} = 0$ it would follow that $(\mathbf{R}_{p+1} - \lambda_0\mathbf{I})\mathbf{u} = 0$, whereby $\mathbf{u} = \mathbf{v}$. This is a contradiction since $\alpha \neq 0$. Hence $\zeta \neq 0$. Moreover, because $\det(\mathbf{\Delta}) = 0$ then $\vartheta \neq 0$ and $\mathbf{\Delta} \neq \mathbf{0}$.

Finally, from Eq. (11)

$$\begin{cases} \zeta - \alpha\vartheta = 0, \\ \overline{\vartheta} + \alpha\zeta = 0. \end{cases} \tag{12}$$

Then $\overline{\vartheta} + \alpha^2\vartheta = 0 \Rightarrow |\alpha| = 1$. □

Next, consider the case of eigenvectors corresponding to the minimum/maximum eigenvalue of a symmetric positive definite Toeplitz matrix $\mathbf{R}_{p+1}$ of multiplicity m greater than 1.

Corollary 2. *Let the multiplicity m of the minimum or maximum eigenvalue λ_0 of a symmetric positive definite Toeplitz matrix $\mathbf{R}_{p+1}$ be $m \in \{1, \ldots, p+1\}$. Then there exists an eigenfilter $F(z)$ of order $p - m + 1$ that has all its roots on the unit circle. Moreover, the coefficient vector $\mathbf{f}$ of the polynomial $F(z)$*

characterises all eigenvectors corresponding to the eigenvalue λ_0.

Proof. Let the multiplicity of the maximum/minimum eigenvalue λ_0 be $m \in \{1,\ldots,p+1\}$. Then the rank of the positive/negative semidefinite matrix $\mathbf{R}_{p+1}-\lambda_0\mathbf{I}$ is equal to $p-m+1$. From [16] it follows that the matrix $\mathbf{R}_{p-m+2}-\lambda_0\mathbf{I}$, defined as the principal diagonal sub-block of matrix $\mathbf{R}_{p+1}-\lambda_0\mathbf{I}$, is of rank $p-m+1$. Then there exists a symmetric/antisymmetric vector $\mathbf{f}\in\mathbb{R}^{p-m+2}$ such that

$$(\mathbf{R}_{p-m+2}-\lambda_0\mathbf{I})\mathbf{f}=0, \tag{13}$$

where λ_0 is the minimum/maximum *simple* eigenvalue of the positive definite symmetric Toeplitz matrix $\mathbf{R}_{p-m+2}$ and $\mathbf{f}$ is the corresponding eigenvector. From Theorem 2 it follows that the roots of the eigenfilter $F(z)$ of order $p-m+1$ corresponding to the eigenvector $\mathbf{f}$ are on the unit circle.

Define the linearly independent zero padded vectors $\mathbf{f}_i$ as

$$\mathbf{f}_i=[\underbrace{0\ \cdots\ 0}_{i\ \text{zeros}}\ \ \mathbf{f}^{\mathrm{T}}\ \ \underbrace{0\ \cdots\ 0}_{m-i-1\ \text{zeros}}\ ]^{\mathrm{T}},\quad \forall i=0,\ldots,m-1. \tag{14}$$

Due to the symmetric Toeplitz structure,

$$\mathbf{f}_i^{\mathrm{T}}(\mathbf{R}_{p+1}-\lambda_0\mathbf{I})\mathbf{f}_i=\mathbf{f}^{\mathrm{T}}(\mathbf{R}_{p-m+2}-\lambda_0\mathbf{I})\mathbf{f}=0,\quad \forall i=0,\ldots,m-1. \tag{15}$$

Since $(\mathbf{R}_{p+1}-\lambda_0\mathbf{I})$ is positive/negative semidefinite, it must be that $(\mathbf{R}_{p+1}-\lambda_0\mathbf{I})\mathbf{f}_i=0$, $\forall i=0,\ldots,m-1$. Then

$$\mathbf{R}_{p+1}\mathbf{f}_i=\lambda_0\mathbf{f}_i,\quad \forall i=0,\ldots,m-1.\quad \square \tag{16}$$

3. Conclusion

Congruent proofs that the roots of the symmetric LP model and the eigenfilter model corresponding to the minimal or maximal simple eigenvalue of positive definite symmetric Toeplitz matrix lie on the unit circle have been presented. Moreover, a new limit is derived for the root angles of the symmetric LP model. In the case of multiple eigenvalues the theory characterising all the eigenvectors corresponding to that eigenvalue has been presented. Furthermore, in this case the eigenfilter obtained has all its roots on the unit circle.

References

[1] J.D. Markel, A.H. Gray, Linear Prediction of Speech, Springer, Berlin, 1976.

[2] W.B. Kleijn, K.K. Paliwal, Speech Coding and Synthesis, Elseviever Science B.V., Amsterdam, 1995.

[3] P. Delsarte, Y. Genin, Y. Kamp, Stability of linear predictors and numerical range of a linear operator, IEEE Trans. Inf. Theory IT-33 (3) (May 1982) 412–415.

[4] D. Goodman, E. Miller, A note on minimizing the prediction error when the zeros are restricted to the unit circle, IEEE Trans. Acoust. Speech Signal Process. ASSP 30 (3–4) (June 1982) 503–505.

[5] P. Stoica, A. Nehorai, On linear prediction models constrained to have unit-modulus poles and their use for sinusoidal frequency estimation, IEEE Trans. Acoust. Speech Signal Process. 36 (6) (June 1988) 940–942.

[6] J. Makhoul, On the eigenvectors of symmetric Toeplitz matrices, IEEE Trans. Acoust. Speech Signal Process. ASSP 29 (4) (August 1981) 868–872.

[7] P. Stoica, B. Friedlander, T. Söderström, Asymptotic bias of the high-order autoregressive estimates of sinusoidal frequencies, Circuits Syst. Signal Process. 6 (3) (September 1987) 287–298.

[8] A. Oppenheim, R. Schafer, Discrete-Time Signal Processing, Prentice-Hall, New Jersey, 1989.

[9] P. Alku, T. Bäckström, Linear predictive method for improved spectral modelling of lower frequencies of speech with small prediction orders, IEEE Trans. Speech Audio Process. 12 (2) (March 2004) 93–99.

[10] P. Stoica, A. Nehorai, The poles of symmetric linear prediction models lie on the unit circle, IEEE Trans. Acoust. Speech Signal Process. ASSP-34 (5) (October 1986) 1344–1346.

[11] U. Grenander, G. Szegö, Toeplitz Forms and Their Applications, University of California Press, Berkeley, California, 1958.

[12] C. Gueguen, Linear prediction in the singular case and the stability of eigen models. in: Proceedings of the IEEE International Conference Acoustics Speech and Signal Processing, Atlanta, GA, March, 1981, pp. 881–884.

[13] P. Stoica, A. Nehorai, On stability and root location of linear prediction models, IEEE Trans. Acoust. Speech Signal Process. 35 (4) (April 1987) 582–584.

[14] M.T. Karaev, The numerical range of a nilpotent operator on a Hilbert space, Proc. Am. Math. Soc. 132 (8) (February 2004) 2321–2326.

[15] T. Bäckström, C. Magi, Effect of white-noise correction on linear predictive coding, IEEE Signal Process. Lett. 14 (2) (February 2007) 148–151.

[16] G. Cybenko, Moment problems and low rank Toeplitz approximations, Circuits, Syst. Signal Process. 1 (3) (September 1982) 345–366.

Study II

Carlo Magi, Jouni Pohjalainen, Tom Bäckström, Paavo Alku. "Stabilised weighted linear prediction", Speech Communication, Vol. 51, pp. 401-411, 2009.

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 51 (2009) 401–411
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Stabilised weighted linear prediction

Carlo Magi [1], Jouni Pohjalainen [2], Tom Bäckström, Paavo Alku *

Helsinki University of Technology (TKK), Laboratory of Acoustics and Audio Signal Processing, P.O. Box 3000, FI-02015 TKK, Finland

Received 26 September 2007; accepted 1 December 2008

Abstract

Weighted linear prediction (WLP) is a method to compute all-pole models of speech by applying temporal weighting of the square of the residual signal. By using short-time energy (STE) as a weighting function, this algorithm was originally proposed as an improved linear predictive (LP) method based on emphasising those samples that fit the underlying speech production model well. The original formulation of WLP, however, did not guarantee stability of all-pole models. Therefore, the current work revisits the concept of WLP by introducing a modified short-time energy function leading always to stable all-pole models. This new method, stabilised weighted linear prediction (SWLP), is shown to yield all-pole models whose general performance can be adjusted by properly choosing the length of the STE window, a parameter denoted by M.

The study compares the performances of SWLP, minimum variance distortionless response (MVDR), and conventional LP in spectral modelling of speech corrupted by additive noise. The comparisons were performed by computing, for each method, the logarithmic spectral differences between the all-pole spectra extracted from clean and noisy speech in different segmental signal-to-noise ratio (SNR) categories. The results showed that the proposed SWLP algorithm was the most robust method against zero-mean Gaussian noise and the robustness was largest for SWLP with a small M-value. These findings were corroborated by a small listening test in which the majority of the listeners assessed the quality of impulse-train-excited SWLP filters, extracted from noisy speech, to be perceptually closer to original clean speech than the corresponding all-pole responses computed by MVDR. Finally, SWLP was compared to other short-time spectral estimation methods (FFT, LP, MVDR) in isolated word recognition experiments. Recognition accuracy obtained by SWLP, in comparison to other short-time spectral estimation methods, improved already at moderate segmental SNR values for sounds corrupted by zero-mean Gaussian noise. For realistic factory noise of low pass characteristics, the SWLP method improved the recognition results at segmental SNR levels below 0 dB.

Keywords: Linear prediction; All-pole modelling; Spectral estimation

1. Introduction

Linear prediction (LP) is the most widely used all-pole modelling method of speech (Makhoul, 1975). The prevalence of LP stems from its ability to estimate the spectral envelope of a voice signal and to represent this information by a small number of parameters. By modelling the spectral envelope, LP captures the most essential acoustical cues of speech originating from two major parts of the human voice production mechanism, the glottal flow (which is the physiological source behind the over-all spectral envelope structure) and the vocal tract (which is the cause of the local resonances of the spectral envelope, the formants). In addition to its ability to express the spectral envelope of speech with a compressed set of parameters, LP is known to guarantee the stability of the all-pole models, provided that the autocorrelation criterion is used. Moreover, implementation of the conventional LP can be done with a small computational complexity. LP analysis,

* Corresponding author. Tel.: +358 9 451 5680; fax: +358 9 460 224.
E-mail addresses: jpohjala@acoustics.hut.fi (J. Pohjalainen), tom.backstrom@tkk.fi (T. Bäckström), paavo.alku@tkk.fi (P. Alku).
[1] Deceased in February 2008. Supported by Academy of Finland (Project No. 111848).
[2] Supported by Academy of Finland (Project No. 107494).

0167-6393/$ - see front matter © 2009 Published by Elsevier B.V.
doi:10.1016/j.specom.2008.12.005

however, also suffers from various drawbacks, such as the biasing of the formant estimates by their neighbouring harmonics (El-Jaroudi and Makhoul, 1991). This is caused by aliasing that occurs in the autocorrelation domain and the phenomenon is, in general, most severe for high-pitch voiced speech. Additionally, it is well-known that the performance of LP deteriorates in the presence of noise (Sambur and Jayant, 1976). Therefore, several linear predictive methods with an improved robustness against noise have been developed (Lim et al., 1978; Zhao et al., 1997; Shimamura, 2004). However, it is worth noticing that most of these robust modifications of LP are based on the iterative update of the prediction parameters. Weighted linear prediction (WLP) uses time-domain weighting of the square of the prediction error signal (Ma et al., 1993). By emphasising those data segments that have a high signal-to-noise ratio (SNR), WLP has been recently shown to yield improved spectral envelopes of noisy speech in comparison to the conventional LP analysis (Magi et al., 2006). In contrast to many other robust methods of LP, the filter parameters of WLP can, importantly, be computed without any iterative update.

When the order of LP increases, the spectral envelopes given by LP might over-estimate the underlying speech spectrum (Murthi et al., 2000). This occurs especially in the analysis of voiced speech of sparse harmonic structure, in which case LP models not only the spectral envelope but also the multiples of the fundamental. The minimum variance distortionless response (MVDR) method tries to cope with this problem by providing a smooth spectral envelope even when the model order is increased. MVDR is popular in array processing but it has recently also attracted increasing interest in speech processing where it has been used, for example, in the feature extraction of speech recognition (Wölfel et al., 2003; Dharanipragada et al., 2007; Wölfel et al., 2005; Yapanel and Hansen, 2003).

This study addresses the computation of spectral envelopes of speech from noisy signals by comparing three all-pole modelling methods: the conventional LP, MVDR, and WLP. Because the original version of WLP presented in (Ma et al., 1993) does not guarantee stability of the all-pole model, the idea of WLP is revisited by developing weight functions which always result in a stable all-pole model. It will be shown that with a proper choice of parameters the proposed stabilised WLP method yields spectral envelopes similar to those given by low order MVDR model but with improved robustness against additive background noise.

2. Weighted linear prediction

The discussion is begun by briefly presenting the optimisation of the filter parameters in WLP. Both in conventional LP and in WLP, sample x_n is estimated by a linear combination of the p past samples. This estimate can be formulated as

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

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

$$\varepsilon_n(\boldsymbol{a}) = x_n - \hat{x}_n = x_n + \sum_{i=1}^{p} a_i x_{n-i} = \boldsymbol{a}^T \boldsymbol{x}_n, \tag{2}$$

where $\boldsymbol{a} = [a_0\ a_1 \cdots a_p]^T$ with $a_0 = 1$ and $\boldsymbol{x}_n = [x_n \cdots x_{n-p}]^T$. The goal is to find the coefficient vector $\boldsymbol{a}$, of a pth order FIR predictor, which minimises the cost function $\mathscr{E}(\boldsymbol{a})$, also known as the prediction error energy. This problem can be formulated as the constrained minimisation problem

$$\begin{aligned} &\text{minimise} \quad \mathscr{E}(\boldsymbol{a}), \\ &\text{subject to } \boldsymbol{a}^T\boldsymbol{u} = 1, \end{aligned} \tag{3}$$

where the unit vector $\boldsymbol{u}$ is defined as $\boldsymbol{u} = [1\ 0 \cdots 0]^T$. This minimisation depends on the nature of the cost function $\mathscr{E}(\boldsymbol{a})$. The cost function in the WLP method is defined as

$$\mathscr{E}(\boldsymbol{a}) = \sum_{n=1}^{N+p} (\varepsilon_n(\boldsymbol{a}))^2 w_n. \tag{4}$$

In matrix notation, Eq. (4) can be written as

$$\mathscr{E}(\boldsymbol{a}) = \boldsymbol{a}^T \left(\sum_{n=1}^{N+p} w_n \boldsymbol{x}_n \boldsymbol{x}_n^T \right) \boldsymbol{a} = \boldsymbol{a}^T \boldsymbol{R} \boldsymbol{a}, \tag{5}$$

where $\boldsymbol{R} = \sum_{n=1}^{N+p} w_n \boldsymbol{x}_n \boldsymbol{x}_n^T$. Here the signal x_n is assumed to be zero outside the interval $[1,N]$, $\boldsymbol{R}$ corresponds to the autocorrelation matrix if and only if $\forall n = 1,\ldots, N+p, w_n = 1$. According to Eq. (4), the formulation allows us to temporally emphasise the square of the residual signal. It should be noticed that in difference to conventional LP the autocorrelation matrix $\boldsymbol{R}$ is *weighted*.

Matrix $\boldsymbol{R}$, defined in Eq. (5), is symmetric but does not possess the Toeplitz structure. However, it is positive definite, thus making the minimisation problem in Eq. (3) convex. Using the Lagrange multiplier minimisation method (Bazaraa et al., 1993), it can be shown (Bäckström, 2004) that $\boldsymbol{a}$, which solves the minimisation problem in Eq. (3), satisfies the linear equation

$$\boldsymbol{R}\boldsymbol{a} = \sigma^2 \boldsymbol{u}, \tag{6}$$

where $\sigma^2 = \boldsymbol{a}^T \boldsymbol{R} \boldsymbol{a}$ is the error energy. The corresponding WLP all-pole filter is obtained as $H(z) = 1/A(z)$, where $A(z)$ is the z-transform of $\boldsymbol{a}$.

3. Model formulation

The key concept of WLP, introduced in Eq. (4), is the time-domain weight function w_n. By choosing an appropriate waveform for w_n, one can either temporally emphasise or attenuate the square of the residual signal prior to the optimisation of the filter parameters. In (Ma et al., 1993) the weight function was chosen based on the short-time energy (STE)

$$w_n = \sum_{i=0}^{M-1} x_{n-i-1}^2, \tag{7}$$

where M is the length of the STE window. The performance of WLP was analysed in the original study by Ma et al. (1993) by using only clean speech represented by a small set of both synthetic and natural vowels. In the current study, however, the idea of weighting is motivated from the point of view of computing linear predictive models of speech that are more robust against noise than the conventional LP. From this perspective, the use of the STE window can be justified by two arguments. Firstly, as illustrated in Fig. 1a, the STE function over-weights those sections of the speech waveform which consist of samples of large amplitude. It can be argued that these segments of speech are less vulnerable to additive, uniformly distributed noise in comparison to values of smaller amplitude. Hence, by emphasising the contribution of these strong data values in the computation of all-pole models one is expected to get spectral models which show better robustness in noisy conditions. Secondly, there is plenty of evidence in speech science indicating that formants extracted during the closed phase of a glottal cycle are more prominent than those computed during the glottal open phase due to the absence of sub-glottal coupling (Wong et al., 1979; Yegnanarayana et al., 1998; Childers and Wong, 1994; Krishnamurthy et al., 1986). Hence, emphasis of the contribution of the samples occurring during the glottal closed phase is likely to yield more robust acoustical cues for the formants. Especially in the case of wideband noise, this kind of emphasising should improve modelling of higher formants in comparison to spectral models such as the conventional LP, which treat all data samples equally.

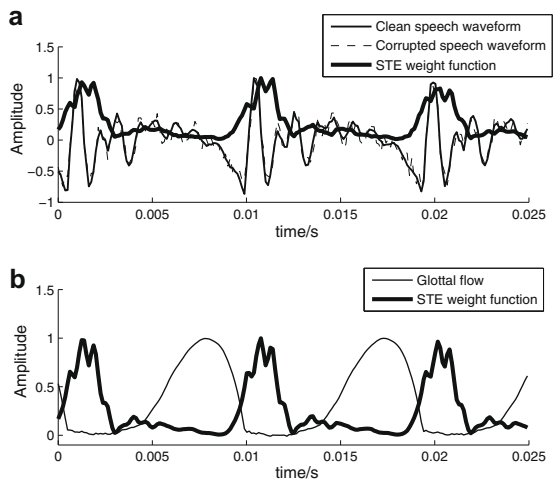


Fig. 1. Upper panel: time-domain waveforms of clean speech (vowel /a/ produced by a male speaker), additive zero-mean Gaussian white noise corrupted speech (SNR = 10 dB), and short-time energy (STE) weight function ($M = 8$) computed from noisy speech according to Eq. (7). Lower panel: glottal flow estimated from the clean vowel /a/ together with STE weight function ($M = 8$) computed also from the clean speech signal.

Fig. 1b illustrates how the STE weight function focuses on the glottal closed phase. In this example, the STE function was computed from the clean /a/ vowel shown in the upper panel of Fig. 1. The glottal flow was estimated from the same clean vowel using the inverse filtering algorithm presented in (Alku, 1992). Even though WLP enables emphasising the contributions of samples occurring during the closed phase, it is worth noticing that the goal of the method *is not* to try to define the vocal tract filter precisely during the closed phase, as is the case in the so-called closed phase covariance method of glottal inverse filtering (Wong et al., 1979; Huiqun et al., 2006).

The stability of the WLP method with the STE weight function, as proposed in (Ma et al., 1993), however, can not be guaranteed. Therefore, a formula for a generalised weight function to be used in WLP is developed here so that the stability of the resulting all-pole filter is always guaranteed. The autocorrelation matrix from Eq. (5) can be expressed as

$$\boldsymbol{R} = \boldsymbol{Y}^T \boldsymbol{Y}, \tag{8}$$

where $\boldsymbol{Y} = [\boldsymbol{y}_0\ \boldsymbol{y}_1 \cdots \boldsymbol{y}_p] \in \mathbb{R}^{(N+p)\times(p+1)}$ and $\boldsymbol{y}_0 = \left[\sqrt{w_1}x_1 \cdots \sqrt{w_N}x_N\ 0 \cdots 0\right]^T$. The columns $\boldsymbol{y}_k$ of the matrix $\boldsymbol{Y}$ can be generated via the formula

$$\boldsymbol{y}_{k+1} = \boldsymbol{B}\boldsymbol{y}_k, \quad k = 0, 1, \ldots, p-1, \tag{9}$$

where

$$\boldsymbol{B} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ \sqrt{w_2/w_1} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{w_3/w_2} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \sqrt{w_{N+p}/w_{N+p-1}} & 0 \end{bmatrix}. \tag{10}$$

The derivation of this stabilised WLP method can be expressed as follows. The weights are first calculated using Eq. (7) such that if $w_i = 0$ a small constant is added to the coefficient ($w_i = 10^{-6}$) and, before forming the matrix $\boldsymbol{Y}$ from Eq. (8), the elements of the secondary diagonal of the matrix $\boldsymbol{B}$ are defined (observe this difference in comparison to the original study by Ma et al. (1993)) for all $i = 1, \ldots, N+p-1$ as

$$\boldsymbol{B}_{i+1,i} = \begin{cases} \sqrt{w_{i+1}/w_i}, & \text{if } w_i \leqslant w_{i+1}, \\ 1, & \text{if } w_i > w_{i+1}. \end{cases} \tag{11}$$

Henceforth, the WLP method computed using matrix $\boldsymbol{B}$, defined above, is called the *stabilised weighted linear prediction* (SWLP) model, where the stability of the corresponding all-pole filter is guaranteed due to Eq. (11) (see Appendix A).

4. Results

The behaviour of SWLP in spectral modelling of speech is demonstrated in the two examples shown in Figs. 2 and 3. In these figures, the analysed speech sounds (vowels /a/ and /e/ in Figs. 2 and 3, respectively) are shown together with the STE weight functions in the upper panels. The lower panels show spectra of parametric all-pole models of order $p = 10$ computed with three techniques: conventional LP with the autocorrelation criterion, minimum variance distortionless response, and the proposed SWLP. In order to demonstrate the effect of the weight function length, the SWLP analysis was computed using M values equal to 8 (left panels) and 24 (right panels). The examples depicted demonstrate two characteristic features of SWLP. First, the weight function computed by the STE clearly emphasises those segments of speech where the data values are of large amplitude while segments of small amplitude values are given lesser weights. Second, the shape of the all-pole spectrum computed by SWLP is, in general, smooth. However, the behaviour of the SWLP spectrum depends on the length of the STE window: with $M = 8$, the SWLP shows a very smooth spectral behaviour reminiscent of low order ($p = 10$) MVDR, but for the larger M value the sharpness of the resonances in the SWLP spectrum increases and its general spectral behaviour approaches that of LP. The reason behind this is evident by referring to Eq. (10): the larger the value of M the more

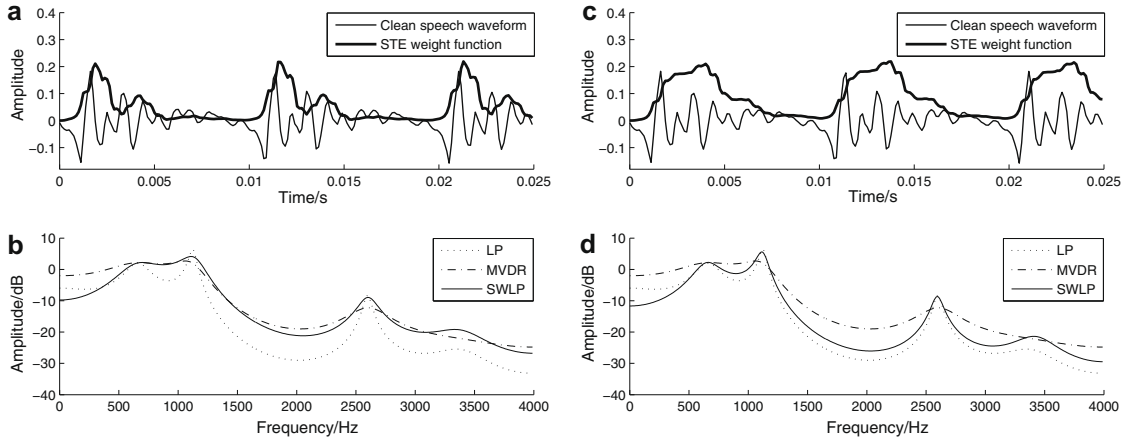


Fig. 2. Time-domain waveforms of clean speech (vowel /a/ produced by a male speaker) and short-time energy (STE) weight function (upper panels) and corresponding all-pole spectra of order $p = 10$ computed by LP, MVDR, and SWLP (lower panels). SWLP analysis was computed by using two different values for the length of the STE window: $M = 8$ (left panels) and $M = 24$ (right panels).

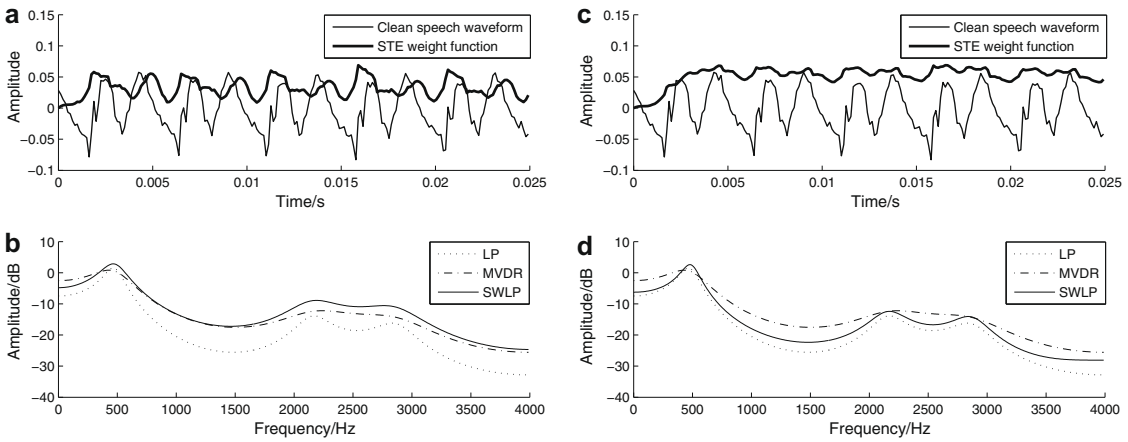


Fig. 3. Time-domain waveforms of clean speech (vowel /e/ produced by a female speaker) and short-time energy (STE) weight function (upper panels) and corresponding all-pole spectra of order $p = 10$ computed by LP, MVDR, and SWLP (lower panels). SWLP analysis was computed by using two different values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

elements of matrix $\boldsymbol{B}$ are equal to unity. In other words, the general spectral shape of the SWLP filter can be made similar to MVDR by selecting a small value of M and it can be adjusted to behave in a manner close to LP by using a larger value of M.

The following result section is divided into three major parts. First, objective spectral distortion measurements were computed for LP, MVDR, and SWLP by using the spectral distortion criterion, SD_2. Next, small scale subjective tests were organised in order to obtain subjective evidence for the performance of low order MVDR and SWLP. It is well known that the SD_2 measure favours smooth spectra. Therefore, automatic speech recognition tests were conducted as the third experiment to get evidence on the performance of the different short-time spectral estimation methods in the presence of noise.

The main focus in the experiments of this study was to measure how the proposed SWLP method works for speech corrupted by additive noise and, in particular, to compare the performance of SWLP to that of LP and MVDR in spectral modelling of noisy speech. All the experiments reported in this study were conducted using the sampling frequency of 8 kHz and the bandwidth of 4 kHz. The prediction order in all methods tested was set to $p = 10$, thereby fulfilling the known rule between the bandwidth and the prediction order (Markel and Gray, 1976). In addition, MVDR was also computed using a high model order ($p = 80$) which is a typical choice in studies in which MVDR has been used in automatic speech recognition (Wölfel et al., 2005; Dharanipragada et al., 2007). Corrupted signals with desired segmental signal to noise ratios (SNR) were generated by adding noise to clean speech sounds. Two types of noise were used: white zero mean Gaussian sequences produced by random number generator and factory noise recorded in realistic circumstances (Varga et al., 1992). Segmental SNR was computed as an average SNR over all 20 ms frames in the speech signal (Kleijn and Paliwal, 1995).

4.1. Objective spectral distortion measurements

Objective evaluation of the effect of noise on all-pole modelling was computed by adapting the widely used spectral distortion criterion, SD_2 (Rabiner and Juang, 1993; Gray et al., 1979). With this measure, the difference between all-pole spectra computed from clean and noisy speech is computed as follows:

$$V(\omega) = \log_{10} P_1(\omega) - \log_{10} P_2(\omega), \tag{12}$$

where P_1 and P_2 denote power spectra of the all-pole filters computed from clean and noisy speech, respectively

$$P_i(\omega) = \frac{\sigma_i^2}{|A_i(e^{j\omega})|^2} \quad i = 1, 2. \tag{13}$$

In Eq. (13), the gains σ_i of the all-pole filters are adjusted so that the impulse response energies of the filters become equal. Since power spectra are computed using FFT, the discrete version of SD_2 must be used

$$SD_2 = \sqrt{\frac{1}{N_s} \sum_{i=0}^{N_s - 1} |V(2\pi f_i)|^2}, \tag{14}$$

where N_s is the length of the discrete FFT spectra.

The experiments here were begun by running a test to analyse how much the performance of SWLP is affected by additive Gaussian noise for different values of M. Speech data, taken from the TIMIT database (Garofolo, 1993), consisted of 12 American English sentences from four different dialect regions produced by six female and six male speakers. The frame length was 25 ms (200 samples) and no pre-emphasis was used. The total number of speech frames analysed in this test was 654, comprising both voiced and unvoiced speech sounds. The difference in the SWLP spectral models computed from clean and noisy samples was quantified in five different segmental SNR categories by using SD_2. The experiments were conducted by using six different values (4, 8, 12, 16, 20, 24) of the STE window length M.

The results obtained from the first experiment are shown in Fig. 4. The data depicted show that the effect of noise on SWLP modelling depends greatly on the choice of the STE window length M: the smaller the value of M the larger the robustness of SWLP against noise. By referring to the examples shown in Figs. 2 and 3, this behaviour can be explained by the effect the value of M has on the shape of the STE function and, consequently, on the general shapes of the SWLP spectral models. In the case of a small M value, temporal fluctuations in the weighting function are greater than those computed with a larger value of M (see Figs. 2a and c & 3a and c). Consequently, the weighting in the case of a small M value emphasises samples of large amplitude more than the weight function defined with a larger M value. In the case of zero-mean Gaussian additive

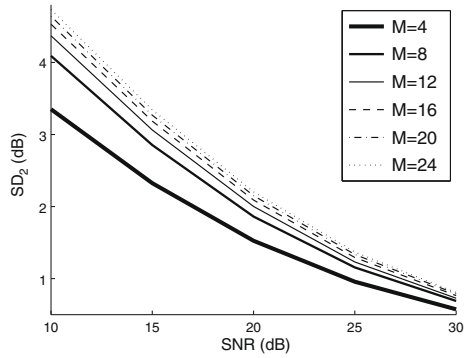


Fig. 4. Spectral distortion values (SD_2) between SWLP envelopes of order $p = 10$ computed from clean and noisy speech. The length of the STE window was varied in six steps from $M = 4$ to $M = 24$. Speech was corrupted by additive zero-mean Gaussian white noise in five segmental SNR categories. SD_2 values were computed as an average over all the analysed segments consisting of 654 frames from the TIMIT database.

noise, this implies that the all-pole models are computed by emphasising speech samples of larger local SNR over those with small local SNR. Hence, the resulting SWLP model computed with a small M value is less vulnerable to additive Gaussian noise. The results shown in Fig. 4 can also be understood from the point of view of the general shape of the SWLP filter (see Figs. 2b and d & 3b and d). In the case of a small M value the all-pole model indicates, also in the case of clean speech, a smoother spectral behaviour than the model computed with a larger M value. In other words, the poles of the SWLP filter computed from speech with large SNR tend to be closer to the origin of the z-plane when the STE function is computed with a small M value. It is understandable that an all-pole filter which has a smooth spectral envelope is less sensitive to noise than a model with sharp resonances, which also explains why Fig. 4 shows the best performance for the lowest value of M.

The second experiment was conducted to compare the performance of the proposed SWLP method to that of conventional LP and MVDR in spectral modelling of noisy speech. Since the behaviour of SWLP depends greatly on the value of the STE window length M, it was decided to compute the SWLP by using two different values for this parameter: a large value of $M = 24$ corresponding to SWLP which behaves similarly to the conventional LP, and a small value $M = 8$, yielding SWLP filters of smooth spectral shape similar to those computed by low order ($p = 10$) MVDR. The greatest M value used in previous experiments was 24, and hence it was selected to represent the SWLP with a large M value. The selection of the small M value was accomplished by running a special experiment in which the value of M that yielded the largest similarity between the all-pole spectra given by SWLP and MVDR ($p = 10$) was searched for. This was done by running an experiment where SD_2 was computed between the MVDR and SWLP all-pole envelopes by varying the STE window length M from 4 to 24. The SD_2 values were computed as an average over the entire (uncorrupted) training data consisting of 650 frames from TIMIT. The result of the experiment showed that the smallest spectral distortion value between SWLP and MVDR spectra was achieved with $M = 8$. Hence, all the further comparisons between SWLP and low order ($p = 10$) MVDR were computed by using the parameter value $M = 8$.

Performance of LP, MVDR, and SWLP was compared by measuring, for each method, how much the all-pole models computed from clean speech differ from those computed from noisy speech. SD_2 was used as an objective distance measure between the all-pole spectra extracted from clean and noisy signals. Again, noise corruption was done by adding zero-mean Gaussian noise to the clean utterances with five segmental SNR levels. Data consisted of 12 sentences, produced by 6 females and 6 males, taken from the TIMIT database. (These utterances were different from those used in the search of the M value yielding the largest similarity between SWLP and MVDR spectra). The total number of speech frames was 650. The SD_2 value for each method in each segmental SNR category was computed as an average over the SD_2 values obtained from individual frames.

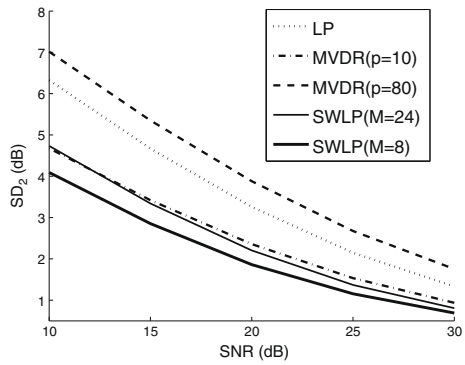


Fig. 5. Spectral distortion values (SD_2) between all-pole envelopes computed from clean and noisy speech with LP ($p = 10$), MVDR ($p = 10$ and $p = 80$) and SWLP ($p = 10$ with $M = 8$ and $M = 24$), where p is the model order and M is the length of the STE weight function. Speech was corrupted by additive zero-mean Gaussian white noise in five segmental SNR categories. SD_2 values were computed as an average over all the analysed segments consisting of 654 frames from the TIMIT database.

The results obtained in comparing the robustness of the three all-pole modelling techniques are shown in Fig. 5. As a general trend, all methods show an increase in SD_2 when segmental SNR decreases. This over-all trend implies, naturally, that the spectral difference between the clean all-pole model and the one computed from noisy speech increases for all the methods analysed when the amount of noise is raised. In comparing conventional LP and MVDR, the results here are in line with previous findings indicating that LP is sensitive to noise while MVDR shows a clearly better performance (Magi et al., 2006). The behaviour of SWLP, however, shows the best robustness against additive Gaussian noise. In particular, SWLP with a small M value is able to tackle the effect of additive Gaussian noise more effectively than any of the other methods tested.

4.2. Small scale subjective tests

Next, in order to get tentative subjective evidence for the performance of low order MVDR and SWLP in the modelling of both clean and noisy speech, a small listening test was organised. In this test, subjects ($n = 13$) listened to 200 ms sounds synthesised by exciting MVDR and SWLP filters of order $p = 10$ by impulse trains. The all-pole filters were computed with MVDR and SWLP both from clean and noisy utterances corrupted with additive zero-mean Gaussian noise with SNR = 10 dB. The utterances consisted of eight Finnish vowels produced by one male and one female subject. The test involved a perceptual comparison between three sounds (the reference sound, sounds A and B). The reference was always the original, clean vowel.

Sounds A and B were synthesised utterances produced, in random order, by impulse train excited MVDR and SWLP filters. In order to involve no pitch difference between the three sounds, the impulse train was always extracted from the reference signal. In addition, the loudness of the three sounds were normalised by adjusting the intensity levels of the sounds to be equal. The listener was asked to evaluate which one of the two alternatives (A or B) sounded more like the reference. In case the listener found that the quality difference between sound A and the reference was equal to that of sound B and the reference, she or he replied with *No preference*. The listener was allowed to listen to the three sounds as many times as she or he wished. The procedure was then repeated for all the vowels including both clean and noisy speech.

The results, shown in Table 1, indicated that for clean male vowels, the listeners preferred the quality of the all-pole filters computed by SWLP over that given by MVDR: in 71% of all comparisons, they rated the vowels synthesised by SWLP to be closer in quality to the original speech, while only in 17% of the cases the listeners were in favour of MVDR. However, there were differences between the vowels: for /a/, /e/, /o/, /ä/, and /ö/, all the listeners preferred the sound synthesised by SWLP while for /i/ and /u/ SWLP was preferred only in approximately 10% of the cases. For these two vowels, both SWLP and MVDR failed to model the second formant properly. MVDR, however, modelled the over-all spectral envelope of the original vowel sound slightly better which might have explained the higher preference of MVDR. When listening to the sounds synthesised from noisy speech, the responses were even more in favour of SWLP: in 73% of all the cases, the vowels produced by SWLP filters were preferred, while those synthesised by MVDR filters were preferred in only 1% of the responses. For clean female vowels, the listeners preferred SWLP in 46% of the cases while MVDR was assessed better in 19% of the comparisons. Again, when listening to the sounds synthesised from noisy speech, the listeners favoured the sounds synthesised by SWLP: in 45% of the cases, it was considered to yield quality closer to the original speech, while MVDR was preferred only in 5% of the cases.

4.3. Automatic speech recognition tests

As the third main part of the experiments, the performance of the proposed SWLP method was tested in feature extraction of a speech recogniser. In the field of automatic speech recognition (ASR), the mel-frequency cepstral coefficient (MFCC) representation is, by far, the most popular method of feature extraction. The stages of the MFCC computation for one speech frame can be outlined as follows (O'Shaughnessy, 2000): (1) estimation of the short-time magnitude spectrum; (2) computation of logarithmic mel-filterbank energies using triangular bandpass filters in the frequency domain; and (3) discrete cosine transformation of the logarithmic filtered energies. In the first stage, simple FFT (periodogram) spectrum estimation is typically used; however, it is not the best spectrum estimation method in terms of robustness when the signal is corrupted by noise. Indeed, it has been argued that both LP and MVDR spectrum estimation, when substituted as the first stage of the MFCC computation, improve noise robustness of the features in certain cases (de Wet et al., 2001; Dharanipragada et al., 2007). This raises the question of whether SWLP could also offer improvement to the robustness of ASR systems.

Table 1
Subjective evaluation between impulse train exited SWLP ($M = 8$) and MVDR filters of order $p = 10$. All-pole filters were computed from clean and noisy (SNR = 10 dB) male and female vowels.

Preferred method	Male vowels		Female vowels	
	Clean (%)	Noisy (%)	Clean (%)	Noisy (%)
MVDR	17	1	19	5
SWLP	71	73	46	45
No preference	12	26	35	50

The performance of six different spectrum estimation methods was evaluated in ASR: FFT, LP ($p = 10$), MVDR with $p = 10$ and $p = 80$, and SWLP ($p = 10$) with $M = 8$ and $M = 24$. This resulted in six slightly different 12-dimensional MFCC feature vectors, which were tested in isolated word recognition (IWR). The goal was to focus on the effect that the short-time spectrum in itself has on robustness. This means that the information given to the recogniser only involved the *shape* of the short-time spectrum. For this reason, neither the zeroth MFCC coefficient, which reflects frame energy, nor the inter-frame $\Delta/\Delta\Delta$-coefficients were included in the feature vector. It is well known that the inclusion of $\Delta/\Delta\Delta$-coefficients, which characterise the temporal changes of the spectrum, in the feature vector generally improves the performance of an ASR system (O'Shaughnessy, 2000). The $\Delta/\Delta\Delta$-coefficients are, however, based on short-time spectral estimation methods. Hence, it is reasonable to assume that whenever the spectrum estimation is distracted by noise, this will also have a negative effect on the obtained $\Delta/\Delta\Delta$-coefficients, resulting in lower recognition performance.

The use of IWR as the test problem can be justified by two reasons. First, state of the art continuous speech recognisers rely heavily on language models to improve their performance. Because language modelling compensates for shortcomings in the acoustic modelling, it may in an unpredictable fashion mask or distort the relative performance differences between the different features. Second, the acoustic modelling in both continuous and connected speech recognition benefits from long-time temporal structure. Instead, by focusing on IWR with vocabularies consisting of fairly short and common words, which might differ by just one phoneme, it can be argued that the feature evaluation concentrates more effectively on the importance of the correct identification of phonetic units based on the short-time spectrum.

The IWR system used in the present study is based on dynamic time warping (DTW) (O'Shaughnessy, 2000).

DTW has been widely replaced by HMM methods in continuous speech recognition, the main focus of current ASR research. However, DTW is still well suited for IWR tasks and provides a good test bench for the present purpose of feature evaluation.

The idea of DTW is to compute a meaningful time-aligned distance between two templates, a test template $T(n)$ consisting of N_T feature vectors and a reference template $R(n)$ consisting of N_R feature vectors, by warping their time axes in order to synchronise acoustically similar segments in the templates. The time alignment is accomplished by finding the minimum-cost path through a grid of $N_T \times N_R$ nodes, where each node (i,j) corresponds to a pair of feature vectors $(T(i), R(j))$ and has an associated cost $d(T(i), R(j))$. In the current implementation, $d(T(i), R(j))$ was chosen to be the squared Euclidean distance between the two MFCC feature vectors $T(i)$ and $R(j)$. The optimised DTW distance was given by the sum of the node costs along the best path. The current system uses the so-called constrained endpoints version of DTW, where the path is required to start from grid node $(1,1)$ and end at node (N_T, N_R) (O'Shaughnessy, 2000). The local continuity constraints of the present implementation dictate that along any permitted path any grid node (i,j) can be reached by one move only from one of the nodes $(i-1, j)$, $(i, j-1)$, or $(i-1, j-1)$. Exceptions naturally occur at grid boundaries where $i = 1$ or $j = 1$. In addition to these constraints, at most two consecutive moves from $(i, j-1)$ to (i,j) are permitted, except at the grid boundary where $i = N_T$.

The training templates were clustered (Rabiner and Juang, 1993) using complete link agglomerative clustering (Theodoridis and Koutroumbas, 2003). This involves computing pairwise DTW distances between all training templates corresponding to the same vocabulary word. For each word in the vocabulary, 10 clusters were generated and one reference template was chosen from each cluster. The representative template for each cluster was chosen as the one with the minimum average distance between it and every other template in the same cluster. During the recognition phase, each test template (test word) was recognised as follows. DTW distances were computed between the test template and each reference template (of which there are 10 for each word in the vocabulary). For each vocabulary word, the average of the three smallest DTW distances was computed. The recognition decision was then made based on the smallest such averaged distance. Similar averaging is suggested in (Rabiner and Juang, 1993).

The test material consisted of words extracted from continuous speech in the TIMIT database. The vocabulary of the recognition task was the 21 words in the two "SA" sentences spoken by every speaker in TIMIT. These sentences were "She had your dark suit in greasy wash water all year" and "Don't ask me to carry an oily rag like that". The training set consists of these words spoken by 136 randomly chosen male and 136 female speakers in the "train" subset of TIMIT (this number was chosen because it is the number of female speakers in the TIMIT "train" subset). The testing set has the words spoken by 50 randomly chosen male and 50 randomly chosen female speakers in the TIMIT "test" subset (which has completely different speakers from the "train" subset). Thus, the training and testing sets contained totals of 5712 and 2100 word tokens, respectively. A similar TIMIT-based corpus (albeit with slightly different sizes of the training and testing sets and non-balanced male–female speaker populations) was used in (Wu et al., 1993), where the best evaluated HMM-based recognisers using single-frame acoustic features achieved a word recognition performance of 91.0%.

The speech material was down-sampled to 8 kHz for the evaluation. All features were computed using a frame length of 20 ms and a frame shift of 10 ms. No pre-emphasis was used. Noise corruption was done by adding prerecorded, down-sampled noise from the Noisex-92 database (Varga et al., 1992) to the test data with seven different segmental SNR levels. Two types of noise were used: white noise and factory noise recorded in a carproduction hall. The averaged power spectra of these two noise signals are shown in Fig. 6. It can be seen that the two noise types have very different characteristics, as the spectrum of the factory noise has a steep downward slope.

The correct recognition rates for the two noise types are shown in Tables 2 and 3. For each noise type and segmental SNR level, the two best scores are shown in boldface. With clean speech, the two most conventional methods, FFT-MFCC and LP-MFCC, showed the best performance. The results for FFT-MFCC and LP-MFCC are in agreement with a previous study, which found LP-based MFCC features to be more robust than their FFT-based counterparts in moderate noise conditions (de Wet et al., 2001). MVDR-MFCC with $p = 10$ slightly outperformed FFT-MFCC in white noise with some segmental SNR levels, while MVDR-MFCC with $p = 80$ showed, in general, modest improvement over FFT-MFCC in factory noise. Considering that the factory noise used here is of a low-pass type, like most other real-world noises used in other

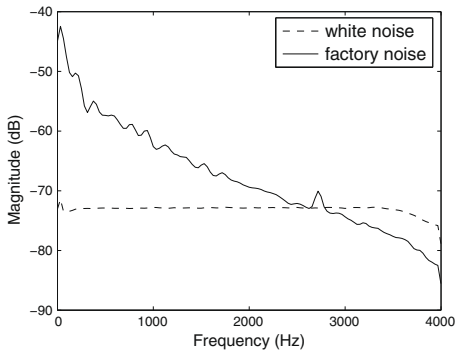


Fig. 6. The averaged power spectra of white noise and car factory noise from the Noisex-92 database (Varga et al., 1992), after re-sampling the signal to 8 kHz. The spectra were estimated using Welch's method with a 20 ms window.

Table 2
Correct recognition rates (%) with white noise. Two best scores are shown in boldface.

Feature vector	Signal to noise ratio (dB)							
	Clean	20	15	10	5	0	−5	−10
FFT-MFCC	**90.9**	86.5	78.3	61.7	42.1	24.6	13.3	**8.7**
LP-MFCC	**91.6**	**87.8**	80.0	65.9	49.9	**32.7**	**15.7**	7.2
MVDR-MFCC, $p = 10$	89.5	84.8	75.8	60.3	44.2	28.0	13.1	6.9
MVDR-MFCC, $p = 80$	89.7	85.2	76.6	61.7	45.0	25.8	12.3	6.2
SWLP-MFCC, $M = 8$	88.7	86.7	**82.5**	**73.7**	**58.0**	**38.5**	**19.2**	**9.5**
SWLP-MFCC, $M = 24$	90.3	**87.8**	**84.3**	**73.6**	**54.0**	32.0	15.4	7.2

Table 3
Correct recognition rates (%) with car factory noise. Two best scores are shown in boldface.

Feature vector	Signal to noise ratio (dB)							
	Clean	20	15	10	5	0	−5	−10
FFT-MFCC	**90.9**	89.3	88.0	86.0	78.8	65.5	43.2	22.9
LP-MFCC	**91.6**	**91.2**	**90.5**	**88.4**	**83.3**	**69.7**	49.3	26.8
MVDR-MFCC, $p = 10$	89.5	87.9	85.6	82.0	73.3	57.2	38.4	21.5
MVDR-MFCC, $p = 80$	89.7	**89.8**	**88.1**	**86.4**	**81.1**	**68.1**	45.9	24.4
SWLP-MFCC, $M = 8$	88.7	88.4	87.1	83.5	78.6	67.1	**50.9**	**30.4**
SWLP-MFCC, $M = 24$	90.3	89.2	87.3	85.3	79.4	67.9	**51.9**	**34.8**

studies, the latter observation appears to be well in line with the findings reported in the literature (e.g. Dharanipragada et al., 2007). The SWLP-MFCC features were superior to the other methods in white noise conditions, in particular when used with the parameter value $M = 8$. With factory noise, SWLP-MFCC became the best method when speech was severely corrupted by noise (that is SNR < 0 dB), and in these cases SWLP-MFCC was on an average 10% units better than the baseline FFT-MFCC.

The results indicate that SWLP-based feature extraction outperformed the other techniques in recognition of speech corrupted by white noise already at segmental SNR value of 20 dB. In the case of factory noise, the major improvements achieved by SWLP occurred at clearly smaller segmental SNR values of -5 dB and -10 dB. The difference in the performance of SWLP between the two noise types can be explained by the fact that in the case of white noise, upper frequencies of voiced speech are masked by noise already at reasonably high segmental SNR levels. This, in turn, implies that traditional spectral modelling techniques, such as LP, cannot model upper formants properly from speech corrupted by white noise. The proposed SWLP, however, emphasises the contribution of speech samples during the closed phase of the glottal cycle and thereby models formants during the time span inside the fundamental period when the resonances are more prominent (see Section 3). This implies that higher formants modelled by SWLP are less likely to be masked by additive noise as severely as those modelled by LP and, consequently, the acoustical cues embedded in them will be more effectively used in the feature extraction. The spectral envelope of factory noise, however, is of a low-pass nature and reminds that of voiced speech. Therefore, higher formants of speech corrupted by factory noise are not distorted severely until at the lowest segmental SNR categories below 0 dB. Hence, the improved recognition accuracy achieved by the proposed SWLP method takes place at the lowest values of the segmental SNR range in the case the additive noise is of low-pass nature.

5. Summary

LP was analysed in this study by using temporal weighting of the residual energy. The work is based on the previous study by Ma et al. (1993) where the concept of WLP was introduced by applying short-time energy waveform as the weighting function. In contrast to the original work by Ma et al., the present study established a modified STE weighting which guarantees the stability of the resulting all-pole filter. This new method, named stabilised weighed linear prediction, was then compared to two known all-pole modelling methods, conventional LP and minimum variance distortionless response, by analysing speech corrupted by additive noise. It was shown that the proposed SWLP method gave the best performance in robustness against noise when quantifying the difference between the clean and noisy spectral envelopes using the objective spectral distortion measure SD_2. This finding was also corroborated by a small subjective test in which the majority of the listeners assessed quality of impulse train excited SWLP all-pole filters extracted from noisy speech to be perceptually closer to original clean speech than the corresponding all-pole responses computed by MVDR. Finally, SWLP was compared to other short-time spectral estimation methods in isolated word recognition experiments. It was shown to improve recognition accuracy already at moderate segmental SNR values for sounds corrupted by white noise. For realistic factory noise of low pass characteristics, the proposed method improved the recognition results at segmental SNR levels below 0 dB.

In difference to the original work by Ma et al. (1993), the present study also focused on how the length of the STE window, the parameter M, affects the general shapes of the all-pole envelopes given by WLP. It was shown, importantly, that by choosing the value of M properly, the behaviour of SWLP can be adjusted to be similar to either LP (corresponding to large M values) or to MVDR (corresponding to small M values). This makes SWLP an attractive method for speech processing because it enables, with the same method, the computation of stable all-pole filters that yield spectral envelopes which are either smooth or of large dynamics. In particular, we believe that the proposed

SWLP method when combined with a properly chosen value of M might become a potential technique in the development of new feature detection methods for recognition of noisy speech. This argument is justified by the increasing interest shown recently in the speech recognition community towards the MVDR technique, due to its promising performance in producing cepstral features for the recognition of noisy speech (Wölfel et al., 2003; Dharanipragada et al., 2007). The current study, however, shows evidence that MVDR is outperformed in robustness by the proposed SWLP in cases when the level of noise corruption is moderate to high. Hence, there are promising areas of future study in examining how the concept of WLP affects the recognition of noisy speech, when used in a state-of-the-art HMM-based continuous speech recognition framework.

Appendix A. Stability of SWLP all-pole filter

In this section, a proof is presented for the minimum phase property of the SWLP inverse filter $A(z) = 1 + a_1 z^{-1} + \cdots + a_p z^{-p}$, where the coefficients a_i are solved from Eq. (6). The structure of the proof is similar to that given in (Delsarte et al., 1982), but for the sake of completeness a more detailed treatment is given in the following:

Rewrite Eq. (6) in the case when the autocorrelation matrix $\boldsymbol{R}$ is factorised as $\boldsymbol{R} = \boldsymbol{Y}^T\boldsymbol{Y}$ and $\boldsymbol{Y} = [\boldsymbol{y}_0\ \boldsymbol{y}_1 \cdots \boldsymbol{y}_p] \in \mathbb{R}^{(N+p)\times(p+1)}$

$$\begin{bmatrix} \boldsymbol{y}_0^T\boldsymbol{y}_0 & \boldsymbol{y}_0^T\boldsymbol{y}_1 & \cdots & \boldsymbol{y}_0^T\boldsymbol{y}_p \\ \boldsymbol{y}_1^T\boldsymbol{y}_0 & \boldsymbol{y}_1^T\boldsymbol{y}_1 & \cdots & \boldsymbol{y}_1^T\boldsymbol{y}_p \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{y}_p^T\boldsymbol{y}_0 & \boldsymbol{y}_p^T\boldsymbol{y}_1 & \cdots & \boldsymbol{y}_p^T\boldsymbol{y}_p \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{A.1}$$

In the SWLP model formulation, the columns $\boldsymbol{y}_i$ of matrix $\boldsymbol{Y}$ were generated via Eq. (9), using matrix $\boldsymbol{B}$ from Eq. (11). However, the column vectors $\boldsymbol{y}_i$ of matrix $\boldsymbol{Y}$ can be expressed by the following reverse equation:

$$\boldsymbol{y}_k = \boldsymbol{M}\boldsymbol{y}_{k+1} \quad k = 0, 1, \ldots, p-1, \tag{A.2}$$

where

$$\boldsymbol{M} := \begin{bmatrix} 0 & 1/\boldsymbol{B}_{2,1} & 0 & \cdots & 0 \\ 0 & 0 & 1/\boldsymbol{B}_{3,2} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & 1/\boldsymbol{B}_{N+p,N+p-1} \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \tag{A.3}$$

and $\boldsymbol{B}_{i+1,i}$ are the elements of matrix $\boldsymbol{B}$ from Eq. (11). Matrix $\boldsymbol{M}$, defined in Eq. (A.3), is a nilpotent[3] operator with power of nilpotency $n = N + p$. Moreover, the norm of the Hilbert space for the matrix $\boldsymbol{M}$ is clearly equal to

$$\|\boldsymbol{M}\|_2 = \max_n\{1/\boldsymbol{B}_{n+1,n}\} = \max_n\left\{\sqrt{w_n/w_{n+1}}\right\}. \tag{A.4}$$

Note that, according to Eq. (11), $1 \leqslant \boldsymbol{B}_{n+1,n} < \infty, \forall n$ which implies that $\|\boldsymbol{M}\|_2 \leqslant 1$.

Defining the matrices $\boldsymbol{Y}_0 := [\boldsymbol{y}_0\ \boldsymbol{y}_1 \cdots \boldsymbol{y}_{p-1}] \in \mathbb{R}^{(N+p)\times p}$ and $\boldsymbol{Y}_1 := [\boldsymbol{y}_1\boldsymbol{y}_2 \cdots \boldsymbol{y}_p] \in \mathbb{R}^{(N+p)\times p}$ and corresponding subspaces $\mathscr{Y}_0 := \operatorname{span}\{\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{p-1}\} \subset \mathbb{C}^{N+p}$ and $\mathscr{Y}_1 := \operatorname{span}\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p\} \subset \mathbb{C}^{N+p}$ (where the base field is $\mathbb{C}$), respectively. Note that the reverse Eq. (A.2) can be written in a more compact form

$$\boldsymbol{Y}_0 = \boldsymbol{M}\boldsymbol{Y}_1. \tag{A.5}$$

Next, define the symmetric linear projection operator $\boldsymbol{P} : \mathbb{C}^{N+p} \to \mathscr{Y}_1$ as

$$\boldsymbol{P} := \boldsymbol{Y}_1(\boldsymbol{Y}_1^T\boldsymbol{Y}_1)^{-1}\boldsymbol{Y}_1^T. \tag{A.6}$$

Thus, for all $\boldsymbol{v} \in \mathscr{Y}_1$ the projection operator has the property

$$\boldsymbol{P}\boldsymbol{v} = \boldsymbol{v}. \tag{A.7}$$

By rearranging Eq. (A.1), the coefficients $\boldsymbol{a} = [a_1 \cdots a_p]^T$ can be solved from the equation

$$\boldsymbol{a} = -(\boldsymbol{Y}_1^T\boldsymbol{Y}_1)^{-1}\boldsymbol{Y}_1^T\boldsymbol{y}_0. \tag{A.8}$$

From this equation yet another important property for the projection operator $\boldsymbol{P}$ is obtained

$$\boldsymbol{P}\boldsymbol{y}_0 = \boldsymbol{Y}_1(\boldsymbol{Y}_1^T\boldsymbol{Y}_1)^{-1}\boldsymbol{Y}_1^T\boldsymbol{y}_0 = -\boldsymbol{Y}_1\boldsymbol{a}. \tag{A.9}$$

Lemma 1. *The zeros of the inverse filter $A(z)$ of the SWLP model are the nonzero eigenvalues of linear operator $\boldsymbol{P}\boldsymbol{M} : \mathbb{C}^{N+p} \to \mathscr{Y}_1$.*

Proof. Take the eigenpair $(\boldsymbol{v}, \lambda)$ of the linear operator $\boldsymbol{P}\boldsymbol{M}$, where the eigenvector $\boldsymbol{v} \in \mathscr{Y}_1$ can be expressed as $\boldsymbol{v} = \boldsymbol{Y}_1\xi$, where $\boldsymbol{\xi} = [\xi_1 \cdots \xi_p]^T \in \mathbb{C}^p$ is the coordinate vector with respect to the basis of space $\mathscr{Y}_1$. Using Eqs. (A.5), (A.7), (A.9) gives

$$\begin{aligned} \lambda\boldsymbol{Y}_1\boldsymbol{\xi} &= \boldsymbol{P}\boldsymbol{M}\boldsymbol{Y}_1\boldsymbol{\xi} = \boldsymbol{P}\boldsymbol{Y}_0\boldsymbol{\xi} = \left[\boldsymbol{P}\boldsymbol{y}_0\ \boldsymbol{P}\boldsymbol{y}_1 \cdots \boldsymbol{P}\boldsymbol{y}_{p-1}\right]\boldsymbol{\xi} \\ &= \left[-\boldsymbol{Y}_1\boldsymbol{a}\ \boldsymbol{y}_1 \cdots \boldsymbol{y}_{p-1}\right]\boldsymbol{\xi} = \boldsymbol{Y}_1\boldsymbol{C}\boldsymbol{\xi}, \end{aligned} \tag{A.10}$$

where

$$\boldsymbol{C} = \begin{bmatrix} -a_1 & & \\ -a_2 & & \boldsymbol{I}_{(p-1)\times(p-1)} \\ \vdots & & \\ -a_p & 0 \quad \cdots & 0 \end{bmatrix} \tag{A.11}$$

is the companion matrix of the inverse filter $A(z)$, that is the zeros of $A(z)$ are the eigenvalues of C. According to Eq. (A.10)

[3] Matrix $\boldsymbol{A}$ is nilpotent with power of nilpotency n if n is the smallest integer such that $\boldsymbol{A}^n = 0$.

$$\begin{aligned} &\boldsymbol{Y}_1\boldsymbol{C}\boldsymbol{\xi} = \lambda\boldsymbol{Y}_1\boldsymbol{\xi}, \\ &\boldsymbol{Y}_1(\boldsymbol{C}\boldsymbol{\xi} - \lambda\boldsymbol{\xi}) = 0, \\ &\boldsymbol{C}\boldsymbol{\xi} = \lambda\boldsymbol{\xi}, \end{aligned} \tag{A.12}$$

where the last implication is due to the fact that $\{\boldsymbol{x} \in \mathbb{C}^p | \boldsymbol{Y}_1\boldsymbol{x} = \boldsymbol{0}\} = \emptyset$. □

Theorem 1. *The zeros of the inverse filter $A(z)$ of the SWLP model are located inside a circle with centre at the origin and radius*

$$\rho = \max_n \left\{\sqrt{w_n/w_{n+1}}\right\} \cos\left(\frac{\pi}{N+p+1}\right).$$

Proof. Take a normalised eigenvector $v \in \mathscr{Y}_1$ and the corresponding eigenvalue $\lambda \in \mathbb{C}$ of the linear operator $\boldsymbol{PM}$. Straightforward calculation gives

$$\begin{aligned} &\lambda = \lambda\|\boldsymbol{v}\|^2 = \boldsymbol{v}^T\lambda\boldsymbol{v} = \boldsymbol{v}^T\boldsymbol{P}\boldsymbol{M}\boldsymbol{v} = (\boldsymbol{P}\boldsymbol{v})^T\boldsymbol{M}\boldsymbol{v} = \boldsymbol{v}^T\boldsymbol{M}\boldsymbol{v} \\ &\in \mathscr{F}(\boldsymbol{M}). \end{aligned} \tag{A.13}$$

Hence, the zeros of the inverse filter $A(z)$ belong to the numerical range $\mathscr{F}(\boldsymbol{M})$ of nilpotent linear operator M. It has been proved in (Karaev, 2004) that the numerical range of the nilpotent operator M with power of nilpotency $N+p$ is a circle (open or closed) with centre at the origin and radius ρ not exceeding $\|\boldsymbol{M}\|_2 \cos\left(\frac{\pi}{N+p+1}\right)$. Hence, according to Eq. (A.4), the zeros of the inverse filter $A(z)$ of the SWLP model are located inside a circle with centre at the origin and with radius $\rho = \max_n\left\{\sqrt{w_n/w_{n+1}}\right\} \cos\left(\frac{\pi}{N+p+1}\right)$. Note that, in the SWLP method, $\max_n\left\{\sqrt{w_n/w_{n+1}}\right\} \leqslant 1$ according to Eq. (11), which guarantees the stability of the corresponding all-pole filter $1/A(z)$. □

References

Alku, P., 1992. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Comm. 11 (2), 109–118.

Bazaraa, M., Sherali, H., Shetty, C., 1993. Nonlinear Programming: Theory and Algorithms, 2nd ed. John Wiley & Sons Inc., New York.

Bäckström, T., 2004. Linear Predictive Modelling of Speech – Constraints and Line Spectrum Pair Decomposition. Ph.D. Thesis, Helsinki University of Technology (TKK), Espoo, Finland. <http://lib.tkk.fi/Diss/2004/isbn9512269473/>.

Childers, D., Wong, C.-F., 1994. Measuring and modeling vocal source-tract interaction. IEEE Trans. Biomed. Eng. 41 (7), 663–671.

de Wet, F., Cranen, B., de Veth, J., Boves, L., 2001. A comparison of LPC and FFT-based acoustic features for noise robust ASR. In: Proc. Eurospeech 2001, Aalborg, Denmark.

Delsarte, P., Genin, Y., Kamp, Y., 1982. Stability of linear predictors and numerical range of a linear operator. IEEE Trans. Inform. Theory IT-33 (3), 412–415.

Dharanipragada, S., Yapanel, U., Rao, B., 2007. Robust feature extraction for continuous speech recognition using the MVDR spectrum estimation method. IEEE Trans. Audio Speech Language Process. 15 (1), 224–234.

El-Jaroudi, A., Makhoul, J., 1991. Discrete all-pole modeling. IEEE Trans. Signal Process. 39 (2), 411–423.

Garofolo, J., 1993. DARPA TIMIT Acoustic-phonetic Continuous Speech Corpus. US Department of Commerce. <http://www.mpi.nl/world/tg/corpora/timit/timit.html>.

Gray, A., Markel, J., 1979. Distance measures for speech processing. IEEE Trans. Acoust. Speech Signal Process. ASSP-24 (5), 380–391.

Huiqun, D., Ward, R., Beddoes, M., Hodgson, M., 2006. A new method for obtaining accurate estimates of vocal-tract filters and glottal waves from vowel sounds. IEEE Trans. Audio Speech Language Process. 14 (2), 445–455.

Karaev, M., 2004. The numerical range of a nilpotent operator on a Hilbert space. Proc. Amer. Math. Soc. 132 (8), 2321–2326.

Kleijn, W., Paliwal, K., 1995. Speech Coding and Synthesis. Elsevier Science B.V..

Krishnamurthy, A., Childers, D., 1986. Two-channel speech analysis. IEEE Trans. Acoust. Speech Signal Process. ASSP-34 (4), 730–743.

Lim, J., Oppenheim, A., 1978. All-pole modelling of degraded speech. IEEE Trans. Acoust. Speech Signal Process. ASSP-26 (3), 197–210.

Ma, C., Kamp, Y., Willems, L., 1993. Robust signal selection for linear prediction analysis of voiced speech. Speech Comm. 12 (1), 69–81.

Magi, C., Bäckström, T., Alku, P., 2006. Objective and subjective evaluation of seven selected all-pole modelling methods in processing of noise corrupted speech. In: CD Proc. 7th Nordic Signal Processing Symposium NORSIG 2006, Reykjavik, Iceland, June 7–9.

Makhoul, J., 1975. Linear prediction: a tutorial review. Proc. IEEE 63 (4), 561–580.

Markel, J., Gray, A.H., 1976. Linear Prediction of Speech. Springer, Berlin.

Murthi, M., Rao, B., 2000. All-pole modeling of speech based on the minimum variance distortionless response spectrum. IEEE Trans. Speech Audio Process. 8 (3), 221–239.

O'Shaughnessy, D., 2000. Speech Communications: Human and Machine, second ed. IEEE Press.

Rabiner, L., Juang, B.-H., 1993. Fundamentals of Speech Recognition. Prentice Hall PTR.

Sambur, M., Jayant, N., 1976. LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise. IEEE Trans. Acoust. Speech Signal Process. ASSP-24 (6), 488–494.

Shimamura, T., 2004. Pitch synchronous addition and extension for linear predictive analysis of noisy speech. In: CD Proc. 6th Nordic Signal Processing Symposium NORSIG 2004, Espoo, Finland, June 9–11.

Theodoridis, S., Koutroumbas, K., 2003. Pattern Recognition, second ed. Academic Press.

Varga, A., Steenneken, H., Tomlinson, M., Jones, D., 1992. Noisex-92 Database. <http://spib.rice.edu/spib/select_noise.html>.

Wong, D., Markel, J., Gray, A., 1979. Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Trans. Acoust. Speech Signal Process. ASSP-27 (4), 350–355.

Wu, J., Chan, C., 1993. Isolated word recognition by neural network models with cross-correlation coefficients for speech dynamics. IEEE Trans. Pattern Anal. Machine Intell. 15, 1174–1185.

Wölfel, M., McDonough, J., 2005. Minimum variance distortionless response spectral estimation. IEEE Signal Proc. Mag. 22 (5), 117–126.

Wölfel, M., McDonough, J., Waibel, A., 2003. Warping and scaling of the minimum variance distortionless response. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop, Vol. 387–392, pp. 387–392.

Yapanel, U., Hansen, J., 2003. A new perspective on feature extraction for robust in-vehicle speech recognition. In: EUROSPEECH 2003, Geneva, Switzerland, September 1–4.

Yegnanarayana, B., Veldhuis, R., 1998. Extraction of vocal-tract system characteristics from speech signals. IEEE Trans. Speech Audio Process. 6 (4), 313–327.

Zhao, Q., Shimamura, T., Suzuk, J., 1997. Linear predictive analysis of noisy speech. In: Communications, Computers and Signal Processing. PACRIM'97, Victoria, Canada, August 20–22, Vol. 2, pp. 585–588.

Study III

Tom Bäckström, Carlo Magi. "Properties of line spectrum pair polynomials – A review", Signal Processing, Vol. 86, No. 11, pp. 3286-3298, 2006.

Available online at www.sciencedirect.com
ScienceDirect
Signal Processing 86 (2006) 3286–3298
SIGNAL PROCESSING
www.elsevier.com/locate/sigpro

Properties of line spectrum pair polynomials—A review

Tom Bäckström*, Carlo Magi[1]

Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology (TKK), P.O. Box 3000, FI-02015 TKK, Finland

Received 16 June 2005; received in revised form 17 January 2006; accepted 23 January 2006
Available online 21 February 2006

Abstract

This review presents mathematical properties of line spectrum pair polynomials, especially those related to the location of their roots, a.k.a. line spectral frequencies. The main results are three interlacing theorems for zeros on the unit circle, which we call the intra-model, inter-model and filtered-model interlacing theorems.

Keywords: Line spectral frequencies; Symmetric polynomials; Linear prediction; Levinson recursion

1. Introduction

Line spectrum pair (LSP) decomposition is a method developed for robust representation of the coefficients of linear predictive (LP) models [1]. Explicitly, the angles of LSP polynomial roots are called line spectrum frequencies (LSFs) and they provide an unambiguous representation of the LP model. Since linear prediction is the de facto standard (as well as, in most cases, a written standard) for spectral modelling of speech [2] and LSP its standard representation, the LSPs are ubiquitous in speech processing. The prevalence of LSP relies on its strong mathematical properties, namely that the stability of the LP model is guaranteed with simple criteria in the LSF domain and robustness of the LSF representation to quantisation noise [3]. Due to their simplicity of use, LSFs are also frequently used in other areas of speech processing, such as speech recognition, enhancement and analysis (e.g. [4–8]).

Line spectral polynomials were first introduced in speech processing by Itakura [1] and their most essential properties presented by Soong and Juang [3]. However, these properties had been independently proved earlier by Schüssler in a more general context [9].

The purpose of the current article is to provide an up-to-date review of the mathematical theory of LSP and give a consistent set of proofs for that theory. In order to keep the size of the article reasonable, we have purposely excluded all applications of LSP and restricted the scope of this article to the theory of LSPs only. Most theorems are given complete proofs, many of which are new or have significant extensions to prior proofs. The goal has been to make the theorems accessible to all while retaining mathematical rigour. Some parts of the

*Corresponding author. Tel.: +358 9 4515843; fax: +358 9 460224.
E-mail address: tom.backstrom@tkk.fi (T. Bäckström).
[1]Supported by the Academy of Finland (project number 205962).

0165-1684/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2006.01.010

current work are based on the doctoral thesis of the first author [10]. However, this review contains newer results, developed after the doctoral thesis as well.

Concerning vocabulary, the abbreviation LSP refers in some sources to Line Spectrum Pair polynomials and in others to line spectral polynomials. Moreover, the operation yielding LSP polynomials has been called LSP transformation, LSP decomposition or simply LSP. Since there is virtually no difference in meaning, we will use all forms interchangeably. The LSFs, which refer to the angle (frequency) of the zeros of LSP polynomials, are sometimes loosely used to refer to LSP methods in general, but we prefer to use LSFs to denote the actual angle or frequency only. The LSP decomposition is based on a transformation of a polynomial to its symmetric and antisymmetric parts, and many different names have appeared to denote these symmetries. According to our understanding the following terms are equivalent: symmetric, self-reciprocal, and palindromic polynomial. The same terms apply for the antisymmetric polynomial but with a prefix of either "anti", "skew", or "conjugate". A polynomial $A(z)$ that has all its zeros inside the unit circle is said to be minimum-phase, or equivalently, its inverse $A^{-1}(z)$ is said to be stable. If it has all zeros outside the unit circle, it is maximum-phase and if the zeros are on the unit circle it is sinusoidal.

2. Definitions and basic properties

The reciprocal polynomial of an mth order polynomial $A(z)=\sum_{l=0}^{m}\alpha_l z^{-l}$ is $A^{\#}(z)=z^{-m}A(z^{-1})$. A polynomial $A(z)$ is said to be *symmetric* if $A(z)=A^{\#}(z)$ and *antisymmetric* if $A(z)=-A^{\#}(z)$. Both types of polynomials are linear-phase FIR filters, when interpreted as transfer functions. It follows that if z_i is a zero of a symmetric or antisymmetric polynomial, then also $z=z_i^{-1}$ must be a zero.

Zeros of symmetric and antisymmetric real polynomials can thus be in the following constellations:

(1) root quadruples symmetric to the unit circle and real axis (z_i, z_i^*, z_i^{-1}, z_i^{*-1}),
(2) root pairs on the unit circle symmetric to the real axis (z_i, z_i^*),
(3) root pairs on the real axis symmetric to the unit circle (z_i, z_i^{-1}),
(4) trivial zeros at $z_i=\pm 1$.

Specifically, simple trivial zeros of symmetric and antisymmetric polynomials must appear in the following combinations [11]:

m	Symmetric	Antisymmetric
Even	None	$z=+1$, $z=-1$
Odd	$z=-1$	$z=+1$

The proof of this result is trivial.

Definition 1. The LSP decomposition (with displacement $k\geqslant 0$) of a polynomial $A(z)$ of order m is

$$\mathscr{P}_k\{A\}(z)=A(z)+z^{-m-k}A(z^{-1}),$$
$$\mathscr{Q}_k\{A\}(z)=A(z)-z^{-m-k}A(z^{-1}). \quad (1)$$

The $(m+k)$th order polynomials $\mathscr{P}_k\{A\}(z)$ and $\mathscr{Q}_k\{A\}(z)$ are called the LSP polynomials. We will often write this as $\mathscr{P}_k\{A\}$ and $\mathscr{Q}_k\{A\}$, when that short form does not cause confusion.

Immediately, we notice that $\mathscr{P}_k\{A\}$ is symmetric and $\mathscr{Q}_k\{A\}$ is antisymmetric and so we obtain (for any k)

$$\tfrac{1}{2}[\mathscr{P}_k\{A\}(z)+\mathscr{Q}_k\{A\}(z)]=A(z). \quad (2)$$

The LSP decomposition is therefore a bijective transformation.

In applications, the most commonly used displacements are $k=0$ and 1, but because most theoretical results generalise nicely over all $k\geqslant 0$, we will apply this more general notation.

Definition 2. We say that two real polynomials $F_1(z)$ and $F_2(z)$ (not necessarily of the same order) are interlaced on the unit circle if

(1) all zeros $z=z_l$ of $F_j(z)$ are on the unit circle, that is, $F_j(z_l)=0 \Leftrightarrow |z_l|=1$;
(2) zeros of $F_1(z)$ and $F_2(z)$ are simple and distinct, with the exception of possible simple trivial zeros at $z=\pm 1$;
(3) the N non-trivial zeros $z\neq\pm 1$ of $F_1(z)$ and $F_2(z)$ are interlaced on both the upper and the lower halves of the unit circle, that is, the zeros $z_l^{(j)}=\exp(\mathrm{i}\theta_l^{(j)})$ of $F_j(z)$ have

$$-\pi<\cdots<\theta_{N/2-1}^{(2)}<\theta_{N/2-1}^{(1)}<\theta_{N/2}^{(2)}<\theta_{N/2}^{(1)}<0$$
$$0<\theta_{N/2+1}^{(1)}<\theta_{N/2+1}^{(2)}<\theta_{N/2+2}^{(1)}<\theta_{N/2+2}^{(2)}<\cdots<\pi.$$

(Note that N is always even since we have omitted trivial zeros.)

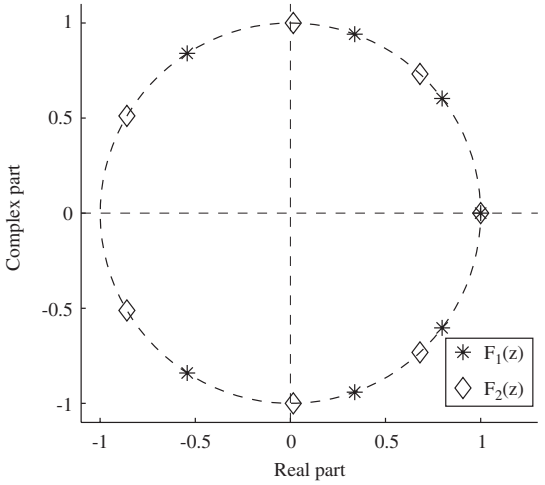


Fig. 1. Illustration of two polynomials $F_1(z)$ and $F_2(z)$ interlacing on the unit circle, denoted as $F_1(z) \doteqdot F_2(z)$.

Moreover, we will define the relation $\doteqdot$, such that $F_1(z) \doteqdot F_2(z)$ means that $F_1(z)$ and $F_2(z)$ interlace on the unit circle and $F_1(z)$ has a zero-pair closer to $z = 1$ (the angle zero).

The interlacing property is illustrated in Fig. 1.

We can now state the most famous LSP property, the *intra-model interlacing* property, in the following theorem.

Theorem 3. *Let $A(z)$ be a real polynomial with all zeros inside the unit circle. Then the zeros of the LSP polynomials are interlaced on the unit circle*

$$\mathscr{P}_k\{A\} \doteqdot \mathscr{Q}_k\{A\}$$

for all $k \geqslant 0$. Conversely, if the zeros of two real polynomials of the same degree, one symmetric and the other antisymmetric, are thus interlaced, then their sum always has all zeros within the unit circle.

This theorem has been proved in [3] (for $k = 1$) and in [9] (for all $k \geqslant 0$). Note that the theorem holds for $k \geqslant 0$ in both directions.

Fig. 2 illustrates this property.

Since the roots of LSP polynomials lie on the unit circle, they can, in principle, be readily found. Moreover, the zeros of the LSP polynomials define the polynomial unambiguously up to scaling and we can reconstruct the LSP polynomials from their zeros (and scaling coefficients) and thereby obtain the original $A(z)$ as well. The zeros can, in turn, be represented by their angles only, since they lie on the unit circle. Finally, the angles are bounded and if

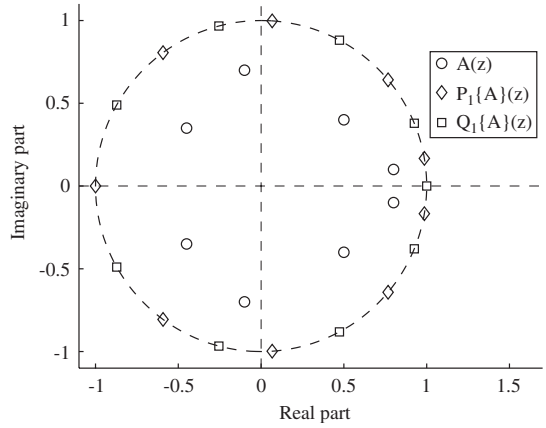


Fig. 2. Illustration of root loci of LSP polynomials. The zeros of $A(z)$ are represented by circles ◯, those of $\mathscr{P}_1\{A\}$ by diamonds ◇ and $\mathscr{Q}_1\{A\}$ by squares □.

the ordering property is ensured, the minimum-phase property of the reconstructed $A(z)$ is retained. It is therefore this theorem that justifies the use of LSP in speech coding.

We can also note that if the LSP transform is calculated with displacement $k \geqslant 1$, then the first coefficients of $\mathscr{P}_k\{A\}$ and $\mathscr{Q}_k\{A\}$ will be equal to that of $A(z)$. The scaling of the LSP polynomials for $k \geqslant 1$ is thus known by construction.

Behind the scenes, it is well known that some speech coding standards (e.g. AMR-WB [12]) choose to communicate the scaling coefficient instead of using displacements $k \geqslant 1$ in order to circumvent patented technology. Such methods are usually based on the concept of immittance spectral pairs (ISP) defined as [13]

$$I(z) = \frac{\mathscr{Q}_0\{A\}(z)}{\mathscr{P}_0\{A\}(z)}.$$

Due to the similarity of definitions, many theorems of LSP find a correspondence in ISP [13,14].

Another interlacing theorem can be derived using the reflection coefficients of a polynomial $A(z)$. Recall that a minimum-phase polynomial can be unambiguously constructed using the recursion

$$A_{n+1}(z) = A_n(z) + \Gamma_n z^{-n-1} A_n(z^{-1}). \tag{3}$$

The polynomial $A_{n+1}(z)$ is minimum-phase iff $|\Gamma_n| < 1$ and $A_n(z)$ is minimum-phase. This is also the basis for the Schur–Cohn and Marden–Jury stability tests [15]. Conversely, the reflection coefficient $\Gamma_n \in (-1, +1)$ guarantees that all zeros of $A_{n+1}(z)$ are within the unit circle given that the

previous order polynomial $A_n(z)$ has all zeros within the unit circle. Quite clearly, we can write equivalently

$$A_{n+1}(z) = \frac{1+\Gamma_n}{2}\mathscr{P}_1\{A_n\}(z) + \frac{1-\Gamma_n}{2}\mathscr{Q}_1\{A_n\}(z). \tag{4}$$

Often we define $\alpha_n = (1+\Gamma_n)/2$ whereby simply $A_{n+1}(z) = \alpha_n\mathscr{P}_1\{A_n\} + (1-\alpha_n)\mathscr{Q}_1\{A_n\}$ and the increase in polynomial order becomes interpolation between LSP polynomials with $\alpha_n \in (0,1)$. This recursive relation is employed in the Levinson recursion (see Section 3).

Theorem 4. *Let polynomials $A_n(z)$, $A_{n+1}(z)$ and $A_{n+2}(z)$, related through Eq. (4), be minimum-phase. Then the zeros of the LSP polynomials with displacements $k \in \{0,1\}$ are interlaced in the following combinations*:

$$\mathscr{P}_k\{A_{n+1}\} \doteqdot \mathscr{P}_k\{A_n\}, \tag{a}$$

$$\mathscr{Q}_k\{A_{n+1}\} \doteqdot \mathscr{Q}_k\{A_n\}, \tag{b}$$

$$\mathscr{P}_k\{A_n\} \doteqdot \mathscr{Q}_k\{A_{n+1}\}, \tag{c}$$

$$\mathscr{P}_k\{A_{n+1}\} \doteqdot \mathscr{Q}_k\{A_n\}, \tag{d}$$

$$\mathscr{P}_k\{A_{n+2}\} \doteqdot \mathscr{P}_k\{A_n\}, \tag{e}$$

$$\mathscr{Q}_k\{A_{n+2}\} \doteqdot \mathscr{Q}_k\{A_n\}. \tag{f}$$

This property is called the *inter-model interlacing* property and is illustrated in Fig. 3 (together with the intra-model interlacing property of Theorem 3).

Cases (a) and (b) have been previously proved in [16,17] and complete proof for cases (a)–(f) will follow.

Proof. From Eq. (4), we observe that $\mathscr{P}_0\{A_{n+1}\} = ((1+\Gamma_n)/2)\mathscr{P}_1\{A_n\}$ and it thus suffices to prove case $k=1$ whereby $k=0$ will directly follow.

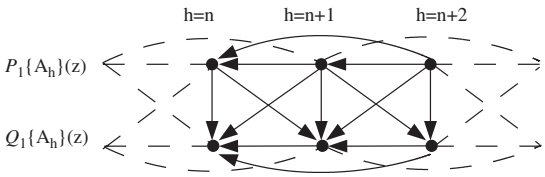


Fig. 3. Illustration of the intra- and inter-model interlacing properties of Theorems 3 and 4. Arrows of graph signify an interlacing-of-zeros between polynomials $\mathscr{P}_k\{A_{h_1}\}$ and $\mathscr{Q}_k\{A_{h_2}\}$ for given orders h_1 and h_2, such that $A \doteqdot B$ corresponds to $A \to B$. Vertical arrows represent intra-model, and all other lines inter-model, interlacing.

To simplify notations, let $P_n(z) = \mathscr{P}_1\{A_n\}(z)$ and $Q_n(z) = \mathscr{Q}_1\{A_n\}(z)$. By substituting Eq. (4) into Eq. (1) we obtain

$$\begin{aligned} P_{n+1}(z) &= \frac{1+\Gamma_n}{2}P_n(z) + \frac{1-\Gamma_n}{2}Q_n(z) \\ &\quad + z^{-n-2}\frac{1+\Gamma_n}{2}P_n(z^{-1}) \\ &\quad + z^{-n-2}\frac{1-\Gamma_n}{2}Q_n(z^{-1}) \\ &= \frac{1+\Gamma_n}{2}P_n(z)(1+z^{-1}) \\ &\quad + \frac{1-\Gamma_n}{2}Q_n(z)(1-z^{-1}), \end{aligned} \tag{5}$$

where we have used $P_n(z) = z^{-n-1}P_n(z^{-1})$ and $Q_n(z) = -z^{-n-1}Q_n(z^{-1})$.

Now, if $P_n(z_i) = 0$ for some $z_i \neq \pm 1$, then $P_{n+1}(z_i) = ((1-\Gamma_n)/2)Q_n(z_i)(1-z_i^{-1}) \neq 0$ for all $\Gamma_n \in (-1,+1)$, since the zeros of $P_n(z)$ and $Q_n(z)$ do not overlap. The zeros of $P_n(z)$ and $Q_n(z)$ do therefore not coincide and, by analogy, the zeros of $Q_n(z)$ and $P_{n+1}(z)$ do not coincide.

Letting Γ_n travel from -1 to $+1$ along the real axis, we find that the zeros of $P_{n+1}(z)$ will travel (on continuous tracks [18]) from the zeros of $P_n(z)(1+z^{-1})$ to the zeros of $Q_n(z)(1-z^{-1})$. Since (1) the zeros will remain on the unit circle for all Γ_n due to Theorem 3, (2) the zeros do not coincide with those of $P_n(z)$ or $Q_n(z)$, and (3) the zeros of $P_n(z)$ and $Q_n(z)$ are interlaced by Theorem 3, then the zeros of $P_{n+1}(z)$ must remain between the zeros of $P_n(z)$ or $Q_n(z)$. The zeros are thus interlaced as $P_{n+1}(z) \doteqdot P_n(z)$ and $P_{n+1}(z) \doteqdot Q_n(z)$.

This proves cases (a) and (c), while cases (b) and (d) follow by analogy.

Similarly as in Eq. (5), we can derive an equation for $P_{n+2}(z)$ using $P_n(z)$ and $Q_n(z)$ to obtain

$$\begin{aligned} P_{n+2}(z) &= \frac{1+\Gamma_n}{2}P_n(z)(1+\Gamma_{n+1}z^{-1}+z^{-2}) \\ &\quad + \frac{1-\Gamma_n}{2}Q_n(z)(1-z^{-2}). \end{aligned} \tag{6}$$

With an identical rationale as for cases (a)–(d) and using Eq. (6), we find that $P_{n+2}(z)$ is interlaced with $P_n(z)$, thus proving case (e). Note that $P_{n+2}(z)$ is not interlaced with $Q_n(z)$ because, in Eq. (6), the zero $Q_n(z_i) = 0$ does not imply that $P_{n+2}(z_i)$ would be non-zero.

Again, case (f) follows by analogy. □

Corollary 5. *The LSP polynomials on consecutive displacements of a polynomial $A(z)$ with all zeros inside the unit circle are interlaced in the following*

combinations ($k \geqslant 0$):

$$\mathscr{P}_{k+1}\{A\} \doteqdot \mathscr{P}_k\{A\}, \quad \text{(a)}$$

$$\mathscr{Q}_{k+1}\{A\} \doteqdot \mathscr{Q}_k\{A\}, \quad \text{(b)}$$

$$\mathscr{P}_k\{A\} \doteqdot \mathscr{Q}_{k+1}\{A\}, \quad \text{(c)}$$

$$\mathscr{P}_{k+1}\{A\} \doteqdot \mathscr{Q}_k\{A\}, \quad \text{(d)}$$

$$\mathscr{P}_{k+2}\{A\} \doteqdot \mathscr{P}_k\{A\}, \quad \text{(e)}$$

$$\mathscr{Q}_{k+2}\{A\} \doteqdot \mathscr{Q}_k\{A\}. \quad \text{(f)}$$

Proof. Let $A_n(z) = A(z)$, $\Gamma_{n+h} = 0$ ($h \geqslant 1$) and define $A_{n+h}(z)$ through Eq. (4). Then $\mathscr{P}_{k+h}\{A_n\} = \mathscr{P}_k\{A_{n+h}\}$ and $\mathscr{Q}_{k+h}\{A_n\} = \mathscr{Q}_k\{A_{n+h}\}$. The desired results follows directly from Theorem 4. □

We will call this interlacing property the *displacement interlacing* property.

3. Levinson recursion

The Levinson recursion is an efficient algorithm for the solution of Toeplitz systems [19,20]. Its connection to LSPs is well known [21]. In speech processing it is most often used to solve the coefficients of linear predictive systems [2,22]. This topic has, however, been covered in numerous other works (e.g. [22,23]) and we will only point out those details that are of relevance to LSP or those that are necessary preliminaries for Section 4.

In order to justify the Levinson recursion, we will first present the linear predictive model. Given a wide-sense stationary signal x_n, we define an mth order linear estimate $\hat{x}_n$ of future samples by $\hat{x}_n = -\sum_{k=1}^{m} \alpha_k x_{n-k}$. The estimation error is

$$e_n = x_n - \hat{x}_n = x_n + \sum_{k=1}^{m} \alpha_k x_{n-k}.$$

By minimising the expected value of the squared error $\mathscr{E}[e_n^2]$, we obtain the normal equations

$$Ra = \sigma^2 [1 \;\; 0 \;\; 0 \;\; \cdots \;\; 0]^{\mathrm{T}}, \quad (7)$$

where R is the real symmetric Toeplitz autocorrelation matrix, $a = [\alpha_0 \;\; \ldots \;\; \alpha_m]^{\mathrm{T}}$ the model parameters (with $\alpha_0 = 1$) and σ^2 the residual energy or the minimum prediction error variance.

The Levinson recursion, which can be used to solve Eq. (7), can then be stated as follows: let us assume that on iteration step l, we have the intermediate solution $a_l = [\alpha_{l,0} \; \alpha_{l,1} \; \ldots \; \alpha_{l,l}]^{\mathrm{T}}$ (with $\alpha_{l,0} = 1$) such that

$$R_l a_l = \begin{bmatrix} r_0 & r_1 & r_2 & \cdots & r_l \\ r_1 & r_0 & r_1 & \cdots & r_{l-1} \\ r_2 & r_1 & r_0 & \cdots & r_{l-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_l & r_{l-1} & r_{l-2} & \cdots & r_0 \end{bmatrix} \begin{bmatrix} \alpha_{l,0} \\ \alpha_{l,1} \\ \vdots \\ \alpha_{l,l} \end{bmatrix} = \sigma_l^2 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (8)$$

where R_l is the $(l+1) \times (l+1)$ symmetric Toeplitz autocorrelation matrix and σ_l^2 the minimum prediction error variance.

In vector notation, the LSP transform in Eq. (1) corresponds to

$$\mathscr{P}_k\{a_l\} = \begin{bmatrix} a_l \\ 0_k \end{bmatrix} + \begin{bmatrix} 0_k \\ Ja_l \end{bmatrix},$$

$$\mathscr{Q}_k\{a_l\} = \begin{bmatrix} a_l \\ 0_k \end{bmatrix} - \begin{bmatrix} 0_k \\ Ja_l \end{bmatrix},$$

where J is the row-reversal matrix, and 0_k is a $k \times 1$ vector of zeros. Clearly,

$$\begin{aligned} R_{l+1}\mathscr{P}_1\{a_l\} &= [\sigma_l^2 + \gamma_l, 0, \ldots, 0, \gamma_l + \sigma_l^2]^{\mathrm{T}} \\ &= \hat{\gamma}_l^{+}[1, 0, \ldots, 0, 1]^{\mathrm{T}} \\ R_{l+1}\mathscr{Q}_1\{a_l\} &= [\sigma_l^2 - \gamma_l, 0, \ldots, 0, \gamma_l - \sigma_l^2]^{\mathrm{T}} \\ &= \hat{\gamma}_l^{-}[1, 0, \ldots, 0, -1]^{\mathrm{T}}, \end{aligned} \quad (9)$$

where $\gamma_l = \sum_{k=0}^{l} \alpha_{l,k} r_{l-k}$.

The Levinson step is then obtained using Eq. (4) and we have

$$\begin{aligned} R_{l+1}a_{l+1} &= R_{l+1}\left[\frac{1+\Gamma_l}{2}\mathscr{P}_1\{a_l\} + \frac{1-\Gamma_l}{2}\mathscr{Q}_1\{a_l\}\right] \\ &= \left[(\sigma_l^2 + \Gamma_l\gamma_l), 0, \ldots, 0, (\gamma_l + \Gamma_l\sigma_l^2)\right]^{\mathrm{T}}. \end{aligned}$$

Choosing Γ_l such that the last term is zero, $\gamma_l + \Gamma_l\sigma_l^2 = 0$, we obtain a form identical to Eq. (8) for $l+1$ with $\sigma_{l+1}^2 = \sigma_l^2 + \Gamma_l\gamma_l$, and we have completed one recursion step. Numerous proofs are available to show that the final model $A(z) = \sum_{l=0}^{m} \alpha_{m,l} z^{-l}$ is minimum-phase if R is positive definite [22,24–27].

There are some improvements to the Levinson algorithm, namely the split Levinson algorithm [21] and Krishna's algorithm [28]. The split Levinson is based on a three-term recursion instead of the two-term recursion in the conventional Levinson, reducing the required multiplications to half. Numerically, the split Levinson is weakly stable [29], and since Krishna's algorithm employs a similar

recursion, one is led to believe that it is also weakly stable, even though the authors are not aware of any proof thereof.

The split Levinson algorithm is based on the intermediate solution

$$R_l\tilde{a}_l = \eta[1\ 0\ 0\ \ldots\ 0\ 1]^{\mathrm{T}}, \tag{10}$$

where η is a scalar. The iteration is based on the three-term equation

$$\tilde{a}_{l+1}^{\mathrm{T}} = \delta[0\ \tilde{a}_{l-1}^{\mathrm{T}}\ 0] + [\tilde{a}_l^{\mathrm{T}}\ 0] + [0\ \tilde{a}_l^{\mathrm{T}}],$$

where scalar δ is chosen so that the structure of Eq. (10) is retained. Observe how nicely the inter-model interlacing between $\mathscr{P}_1\{A_{l-1}\}$, $\mathscr{P}_1\{A_l\}$ and $\mathscr{P}_1\{A_{l+1}\}$ correlates with this three-term recursion. In fact, it is the inter-model interlacing theorem that ensures that the zeros of the LSP polynomials remain on the unit circle and interlaced.

Observe also that if a_l is the solution to Eq. (8), then

$$R_l\mathscr{P}_0\{a_l\} = 2\sigma_l^2[1\ 0\ 0\ \ldots\ 0\ 1]^{\mathrm{T}}$$

and Eq. (10) is thus the symmetric part of the conventional Levinson solution. A similar split algorithm could be developed for the antisymmetric part as well.

Krishna's algorithm is otherwise identical to the split Levinson, but it is based on the intermediate solution

$$R_l\hat{a}_l = \eta[+1, \pm1, +1, \pm1, +1, \pm1, +1, \ldots]^{\mathrm{T}}, \tag{11}$$

where we choose either "+" or "−" according to preference. (Interestingly, this equation appears also as a solution to an inverse problem for transmission lines [30].) It can be readily shown that this means that in Krishna's method, the trivial zeros are omitted from intermediate solutions. This fact can be easily observed by the following calculations. Let $\hat{a}_l = [\hat{\alpha}_{l,0} \ldots \hat{\alpha}_{l,l}]^{\mathrm{T}}$ be the solution to Eq. (11) (with '−' signs and l even) and $\hat{A}_l(z)$ its z-transform. Adding a trivial zero $z=-1$ to $\hat{A}_l(z)$, that is, forming $(1+z^{-1})\hat{A}_l(z)$, in vector notation, and multiplying by R_{l+1}, corresponds to

$$R_{l+1}\begin{bmatrix} \hat{\alpha}_{l,0} & 0 \\ \hat{\alpha}_{l,1} & \hat{\alpha}_{l,0} \\ \vdots & \vdots \\ \hat{\alpha}_{l,l} & \hat{\alpha}_{l,l-1} \\ 0 & \hat{\alpha}_{l,l} \end{bmatrix}\begin{bmatrix} +1 \\ +1 \end{bmatrix} = \eta\begin{bmatrix} +1 \\ -1 \\ +1 \\ \vdots \\ \xi \end{bmatrix} + \eta\begin{bmatrix} \xi \\ +1 \\ -1 \\ \vdots \\ +1 \end{bmatrix}$$

$$= \eta\begin{bmatrix} 1+\xi \\ 0 \\ \vdots \\ 0 \\ 1+\xi \end{bmatrix},$$

which is obviously equivalent to Eq. (10).

4. Constrained linear prediction

A surprising and intriguing connection exists between LSP and a modified linear predictive model called constrained linear prediction (CLP) [31,32]. It is a theoretical result interesting in its own right.

The constrained linear predictive models considered here are models where the frequency response of the model is constrained to attenuate certain frequency regions. This is achieved by shaping the spectrum of the input signal with fixed FIR filters and estimating the original signal from the filtered signal. Consequently, the predictive power of the model is concentrated to those frequency regions which are not filtered out by the FIR filter.

Specifically, the CLP model is constructed as follows. Let us define a linear predictive model where future samples of signal x_n are estimated from past filtered samples of x_n. The modelling error becomes

$$e_n = x_n + \sum_{h=0}^{m-1} a'_h\hat{x}_{n-h}, \tag{12}$$

where $\hat{x}_n = c_n * x_n$ is the filtered input signal and the a'_h's are the model parameters. Let $C(z) = \sum_{l=0}^{N} c_l z^{-l}$ be the transfer function of the filter and v_i^{-1} its zeros. In vector form this equation becomes $e_n = b^{\mathrm{T}}x + a'^{\mathrm{T}}C^{\mathrm{T}}x$, where C is the convolution matrix corresponding to $C(z)$ and $b = [1\ 0\ \ldots\ 0]^{\mathrm{T}}$. In the current context, only the case $N=1$ is relevant, and we will, in the following, thus consider only filters $C(z)$ with a single zero v^{-1}.

Our objective is to minimise the expected value of e_n^2 from Eq. (12). This can be easily achieved by defining $a = b + Ca'$, whereby we have $e_n = a^{\mathrm{T}}x$. In the optimisation problem, we can use the convex objective function $a^{\mathrm{T}}Ra$, where R is the symmetric Toeplitz autocorrelation matrix, similarly as in classical LP, but with the constraint that vector a must be of the form $a = b + Ca'$. The constraint can readily be included in the minimisation problem

using Lagrange coefficient γ and the null-space c_v^0 of convolution matrix C (whereby $C^{\mathrm{T}}c_v^0 = 0$), where $c_v^0 = [1\ v\ v^2 \dots v^m]^{\mathrm{T}}$.

Now, the constrained objective function for our optimisation function becomes $\zeta(a,\gamma) = a^{\mathrm{T}}Ra - (a-b)^{\mathrm{T}}c_v^0\gamma$. Through some straightforward calculations, the minimum of the expected value of the squared error becomes [32]

$$Ra = \gamma[1\ \ v\ \ v^2\ \ \dots\ \ v^m]^{\mathrm{T}} \tag{13}$$

and we can show that if $|v|<1$ with $v \in \mathbb{C}$, then a has all its zeros within the unit circle. For $|v| = 1$ it has zeros *on* the unit circle and $|v|>1$ zeros outside the unit circle. Moreover, if $v = 0$, then a becomes the conventional LP model, whose minimum-phase property is well known [22,24–26,33].

Theorem 6. *Let vector* $a = [\alpha_0 \dots \alpha_m]^{\mathrm{T}}$ *be a solution to Eq.* (13), *R a positive definite symmetric Toeplitz matrix and scalar* γ *chosen such that* $\alpha_0 = 1$. *Then the z-transform* $A(z)$ *of a has all its zeros*

- *inside the unit circle if* $|v|<1$,
- *on the unit circle if* $|v| = 1$,
- *outside the unit circle if* $|v|>1$.

The proof has appeared in [32].

If $v = \pm 1$, then Eq. (13) becomes identical to Eq. (11), thereby tying the constrained linear predictive model to Levinson recursion and LSP. In fact, if $v = \pm 1$ then the constrained linear predictive model becomes an estimate of future samples of x_n using averaged or differentiated past samples of x_n [31,34,35]. Consequently, the zeros of symmetric LSP polynomials are shifted in the FFT spectrum towards zero, since it is based on an averaged input signal, whose spectral components close to the Nyquist frequency are attenuated. Similarly, the zeros of antisymmetric LSP polynomials are shifted towards the Nyquist frequency, since it is based on a differentiated input signal, whose spectral components close to zero are attenuated. See Fig. 4 for illustration.

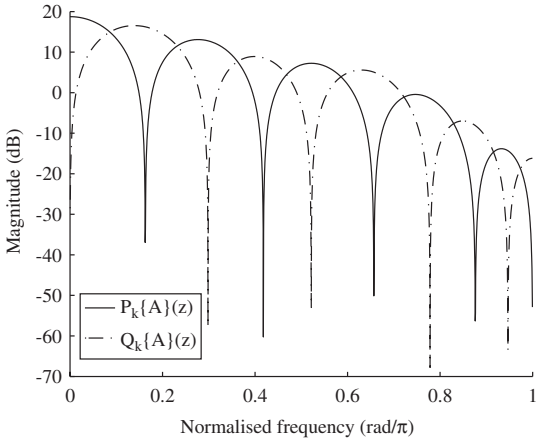


Fig. 4. Illustration of the shifting of zeros of the symmetric and antisymmetric LSP polynomials, $\mathscr{P}_k\{A\}$ (solid line) and $\mathscr{Q}_k\{A\}$ (dash-dotted line), to the left and the right, respectively.

On the other hand, since solutions to Eq. (13) have all zeros inside the unit circle for $|v|<1$ their symmetric and antisymmetric parts (the LSP polynomials) have zeros on the unit circle due to Theorem 3. Explicitly, solutions to the equations

$$\begin{aligned} R\mathscr{P}_0\{a_v\} &= \gamma \begin{bmatrix} 1+v^m \\ v+v^{m-1} \\ v^2+v^{m-2} \\ \vdots \\ v^m+1 \end{bmatrix}, \\ R\mathscr{Q}_0\{a_v\} &= \gamma \begin{bmatrix} 1-v^m \\ v-v^{m-1} \\ v^2-v^{m-2} \\ \vdots \\ v^m-1 \end{bmatrix} \end{aligned} \tag{14}$$

have all zeros on the unit circle due to Theorem 3. Note that the LSP polynomials of a_v are real if $v \in \mathbb{R}$ or $|v| = 1$. (For complex v with $|v| = 1$, a_v can be scaled to real by division with $(1 \pm v^m)$, since $(v^k \pm v^{m-k})/(1 \pm v^m)$ is real for all k.) With $v = 0$, Eq. (14) becomes exactly the solution of a step in the split Levinson recursion.

An intriguing special case is $v \to \pm 1$. For example, in Eq. (14), when $v \to +1$, we observe that $\mathscr{Q}_0\{a_v\} \to 0$. However, by dividing the equation by $1 - v$, we obtain a scaled model $a_v^- = a_v/(1-v)$ that does not converge to zero. By similar scalings, we obtain non-zero vectors $a_v^{\pm} = a_v/(1 \pm v)$ at the limit

all $m, v \to +1$,

$$R\mathscr{Q}_0\{a_v^+\} = \gamma\left[1, 1-\frac{2}{m}, 1-\frac{4}{m}, \dots, -1+\frac{2}{m}, -1\right]^{\mathrm{T}},$$

m even, $v \to -1$,

$$R\mathscr{Q}_0\{a_v^-\} = \gamma\left[1, -1+\frac{2}{m}, 1-\frac{4}{m}, \dots, 1-\frac{2}{m}, -1\right]^{\mathrm{T}},$$

m odd, $v \to -1$,

$$R\mathscr{P}_0\{a_v^-\} = \gamma\left[1, -1+\frac{2}{m}, 1-\frac{4}{m}, \ldots, -1+\frac{2}{m}, 1\right]^{\mathrm{T}}.$$

These polynomials having coefficients $a_v^{\pm}$ all have zeros on the unit circle due to continuity (polynomial zeros follow continuous tracks for continuous transformations of the polynomial coefficients [18]).

While it could seem that we now have come quite far from the central topics in LSP theory, the following theorem will bring us right back and demonstrate yet another interlacing property of LSP. Its proof has not been presented before, and we are therefore bound to present the full proof even if it is long and rather complex.

Theorem 7. *Let x_n be a wide-sense stationary process and c_n and d_n the impulse response coefficients of two first-order real FIR-filters with z-transforms $C(z)$ and $D(z)$, whose zeros are $z=\mu^{-1}$ and $z=\theta^{-1}$, respectively, with $-1<\mu<\theta<+1$. Calculate the LP models a_c and a_d, respectively, for x_n filtered with c_n and d_n. Then the LSP polynomials of $A_c(z)$ and $A_d(z)$ corresponding to a_c and a_d, respectively, are interlaced in the following pairs*:

$$\mathscr{P}_0\{A_c\} \doteqdot \mathscr{P}_0\{A_d\}, \quad \text{(a)}$$

$$\mathscr{Q}_0\{A_c\} \doteqdot \mathscr{Q}_0\{A_d\}, \quad \text{(b)}$$

$$\mathscr{P}_0\{A_c\} \doteqdot \mathscr{Q}_0\{A_d\}, \quad \text{(c)}$$

$$\mathscr{P}_0\{A_d\} \doteqdot \mathscr{Q}_0\{A_c\}. \quad \text{(d)}$$

Note that in contrast to the constrained LP model in Eq. (12), which estimates future values of the *original* signal x_n from filtered past values, in Theorem 7 we have used the conventional LP that estimates future *filtered* values from past filtered values.

Fig. 5 illustrates the generation of LSP polynomials in Theorem 7. We will denote these interlacing properties as the *filtered-model interlacing* property. In addition to the interlacing pairs listed in Theorem 7, naturally, the intra-model interlacing properties of Theorem 3 are still valid. The filtered-model interlacing property is illustrated in Fig. 6.

Since the proof is long and complex, it is presented in Appendix A.

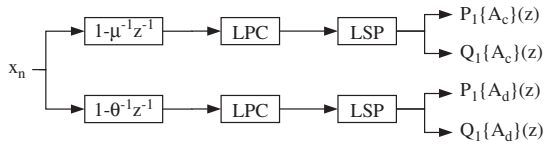


Fig. 5. Illustration of the filtered-model interlacing of LSP polynomials in Theorem 7.

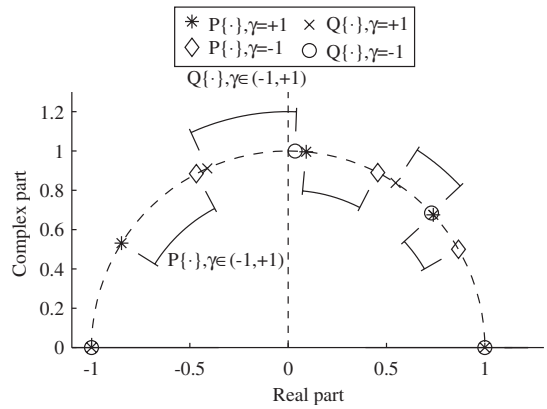


Fig. 6. Illustration of the filtered-model interlacing property. The intervals plotted on the inside of the unit circle signify the range of the zeros of $\mathscr{P}\cdot$ for $\gamma \in (-1,+1)$, and similarly, intervals outside the range of $\mathscr{Q}\cdot$ for $\gamma \in (-1,+1)$. Note that the ranges themselves are interlacing.

5. Other topics

In speech coding, traditionally, the LSP polynomials are not as much of interest as the angles of their roots, the LSFs. Since the roots lie on the unit circle, it is sufficient to express their angle only for a complete description of the root. In the spectral domain, unit circle zeros (or poles) appear as vertical lines, thereby warranting the name line spectral frequencies (see Fig. 4).

The problem thereby becomes finding one of the roots of given LSP polynomials. Since the polynomials are symmetric and antisymmetric, they have only $m/2$ or $(m-1)/2$ degrees of freedom, and we can expect that root finding should benefit from this symmetry. Indeed, for symmetric polynomials, we can use the Chebyshev transform and substitute $x=z+z^{-1}$ to obtain a real polynomial of x with all zeros on the real axis in the interval $x \in [-2,+2]$ [36]. By removal of trivial zeros, the same operation can be done for antisymmetric polynomials as well [11]. While the Chebyshev transform is widely used, it has received a rigorous

mathematical treatment only recently in [37]. The root-finding problem has, anyhow, received a lot of attention (e.g. [3,36,38–40]), but it still is the most computationally costly step in spectral modelling of speech coding applications. In order to improve root-finding algorithms, it is possible to develop statistical distributions for the location of zeros [41].

It has been proved that "the poles of symmetric linear prediction models lie on the unit circle" [33]. It is important to realise, however, that this is true only if we use explicit constraints to fix the first and last coefficients of a causal predictor to be equal to unity, but that it is not true for all symmetric LP models. In practical applications, the only value for the first coefficient that makes sense is naturally unity, but it is possible to use other constraints that enforce symmetry and only afterwards scale the predictor to make the first coefficient equal to unity. For example, the forward–backward predictive model (which has been reinvented several times) constrains the middle coefficient to be equal to unity and thus has the normal equations [42–46]

$$Ra = \gamma[0 \ \ldots \ 0 \ 1 \ 0 \ \ldots \ 0]^{\mathrm{T}},$$

where the single "1" is the middle coefficient. It can be easily verified through experiments that this model often has zeros off the unit circle even though it is symmetric. The presented proof for the unit-circle property of (anti)symmetric polynomials in Eq. (14) is thus relevant since it provides a general formula which ensures the unit circle property of symmetric polynomials. Using these equations, we can generate an arbitrary number of polynomials with zeros (interlaced) on the unit circle. The authors are, however, not aware of any similar solutions that would have all zeros off the unit circle.

It would be tempting to try to use LSFs as a basis for sinusoidal modelling; after all, the LSP decomposition generates two models with zeros on the unit circle, which are in the spectral domain equivalent to sinusoidal models. In other words, we could use minimum-phase polynomials for the generation of sinusoidal models with the aid of LSP decomposition. However, in the presence of noise, the LSFs are inconsistent and biased estimates of sinusoidal signals [47,48].

6. Conclusions

In this work, we have presented a review of the properties of line spectrum pair polynomials. The most significant results are three interlacing theorems, which can be summarised as follows:

Intra-model interlacing—the interlacing of the zeros of the symmetric and antisymmetric LSP polynomials on the unit circle.

Inter-model interlacing—the interlacing of the zeros of the LSP polynomials of consecutive order.

Filtered-model interlacing—the interlacing of the zeros of LSP polynomials of LP models calculated from signals filtered with first-order FIRs.

In addition, we presented an interlacing relation which is of lesser importance, the *displacement interlacing*, or the interlacing of LSP polynomials of consecutive displacement.

These interlacing properties provide a rich basis for studies in many subfields of speech processing, for example, spectral modelling of speech with stable all-pole models for speech coding and enhancement, and improved features for speech recognition. Moreover, we have to keep in mind that linear prediction is used in a wide range of applications other than speech processing, such as geology, economics, transmission line theory, and systems identification. The possible applications of the presented results therefore have a very wide range and large potential.

Appendix A. Proof for Theorem 7

Since the proof of Theorem 7 is rather complex, we have factored it into three preliminary lemmas, which we will present before presenting the main proof.

Lemma 8. *Let x_n be a wide-sense stationary process and c_n the impulse coefficients of a first-order FIR-filter that has a zero at $z = v^{-1}$. Calculate the conventional LP model a_c from the filtered signal $c_n * x_n$ and the constrained LP model a_v with* Eq. (13). *Then the LSP polynomials (displacement $k = 0$) of a_c and a_v are equal up to scaling.*

Proof. Defining a vector $c_v^0 = [1 \ v \ v^2 \ \ldots \ v^m]^{\mathrm{T}}$ and the row-reversal matrix J, we can write Eq. (14) as

$$R_m \mathscr{P}_0\{a_v\} = \gamma(c_v^0 + Jc_v^0).$$

Define the convolution matrix C_v for c_n as

$$C_v = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ v & 1 & \ddots & \vdots \\ 0 & v & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 \\ 0 & \ldots & 0 & v \end{bmatrix}.$$

Convolving a zero v^{-1} to $\mathscr{P}_0\{a_v\}$ is then equivalent to $C_v\mathscr{P}_0\{a_v\}$ and we obtain

$$\begin{aligned}
&R_{m+1}C_v\mathscr{P}_0\{a_v\}\\
&\quad= R_{m+1}\begin{bmatrix}\mathscr{P}_0\{a_v\} & 0\\ 0 & \mathscr{P}_0\{a_v\}\end{bmatrix}\begin{bmatrix}1\\ -v^{-1}\end{bmatrix}\\
&\quad= \gamma\begin{bmatrix}c_v^0+Jc_v^0 & \beta\\ \beta & c_v^0+Jc_v^0\end{bmatrix}\begin{bmatrix}1\\ -v^{-1}\end{bmatrix}\\
&\quad= \gamma(1-v^{-2})\hat{c}_v^0+\gamma(v^m+v^{-2}+v^{-1}\beta)[1\ 0\ 0\cdots 0\ 1]^{\mathrm{T}},
\end{aligned}$$

where the symmetric Toeplitz matrix R_{m+1} is a positive definite extension to R_m (such extensions are always readily available) and $\hat{c}_v^0=[1\ v\ \dots\ v^{m+1}]^{\mathrm{T}}$. Since $C_v^{\mathrm{T}}c_v^0=0$, we can multiply from the left by C_v^{T} to obtain

$$\tilde{R}_m\mathscr{P}_0\{a_v\}=\tilde{\gamma}[1\ 0\ 0\ \dots\ 0\ 1]^{\mathrm{T}}, \tag{A.1}$$

where $\tilde{\gamma}=\gamma(v^m+v^{-2}+v^{-1}\beta)$ and $\tilde{R}_m=C_v^{\mathrm{T}}R_{m+1}C_v$ is the autocorrelation matrix of the filtered signal. Comparing Eq. (A.1) to Eq. (9), we observe that $\mathscr{P}_0\{a_v\}$ is the symmetric part of a conventional LP model (with autocorrelation matrix $\tilde{R}_m$). In other words, the symmetric part of an LP model from a filtered signal is therefore equal to the constrained LP model with the same filtering. A similar proof for the antisymmetric LSP polynomial is, *mutatis mutandis*, readily available by the same approach. □

Lemma 9. *Let the symmetric LSP polynomials $\mathscr{P}_0\{a_\mu\}$ and $\mathscr{P}_0\{a_\theta\}$ be solutions to*

$$R_m\mathscr{P}_0\{a_\mu\}=\gamma(c_\mu^0+Jc_\mu^0)\quad \textit{and}$$
$$R_m\mathscr{P}_0\{a_\theta\}=\gamma(c_\theta^0+Jc_\theta^0),$$

where $c_v^0=[1\ v\ v^2\ \dots\ v^m]^{\mathrm{T}}$, and let $\mathscr{P}_0\{A_\mu\}(z)$ and $\mathscr{P}_0\{A_\theta\}(z)$ be their z-transforms.

Iff θ is a zero of $\mathscr{P}_0\{A_\mu\}(z)$ then μ is a zero of $\mathscr{P}_0\{A_\theta\}(z)$ and all the other zeros are equal.

The same holds for antisymmetric LSP polynomials.

Note that if μ is real, then the zeros of $\mathscr{P}_0\{A_\mu\}$ lie on the unit circle and consequently, θ must be on the unit circle for the lemma to apply.

Furthermore, as a corollary, observe that if the polynomials $\mathscr{P}_0\{A_\mu\}$ and $\mathscr{P}_0\{A_\theta\}$ have a joint root v, we then can form a new polynomial $\mathscr{P}_0\{A_v\}$ for which $R_m\mathscr{P}_0\{a_v\}=\gamma(c_v^0+Jc_v^0)$. Then $\mathscr{P}_0\{A_\mu\}$ and $\mathscr{P}_0\{A_v\}$ have joint roots except for v and μ, and likewise, $\mathscr{P}_0\{A_v\}$ and $\mathscr{P}_0\{A_\theta\}$ have joint roots except for v and θ. It follows that $\mathscr{P}_0\{A_\mu\}$ and $\mathscr{P}_0\{A_\theta\}$ have joint roots except for μ and θ. In other words, it is not possible that some roots would be joint and some not, but rather, it is all or nothing.

Proof. The lemma is symmetric with respect to μ and θ and it is thus sufficient to prove it in one direction "⇒" and we obtain the other direction "⇐" by analogy.

Define the convolution matrix C_μ with the second-order FIR-filter $[1,-(\mu+\mu^{-1}),1]$ and let $\mathscr{P}_0\{a_\mu\}$ be the solution to

$$R_m\mathscr{P}_0\{a_\mu\}=\gamma(c_\mu^0+Jc_\mu^0).$$

Convolve two zeros μ and μ^{-1} to $\mathscr{P}_0\{a_\mu\}$, that is, $C_\mu\mathscr{P}_0\{a_\mu\}$, and we obtain

$$\begin{aligned}
&R_{m+2}C_\mu\mathscr{P}_0\{a_\mu\}\\
&\quad= R_{m+2}\begin{bmatrix}\mathscr{P}_0\{a_\mu\} & 0 & 0\\ 0 & \mathscr{P}_0\{a_\mu\} & 0\\ 0 & 0 & \mathscr{P}_0\{a_\mu\}\end{bmatrix}\begin{bmatrix}1\\ -(\mu+\mu^{-1})\\ 1\end{bmatrix}\\
&\quad= \gamma\begin{bmatrix}c_\mu^0+Jc_\mu^0 & \beta_1 & \beta_2\\ \beta_1 & c_\mu^0+Jc_\mu^0 & \beta_1\\ \beta_2 & \beta_2 & c_\mu^0+Jc_\mu^0\end{bmatrix}\begin{bmatrix}1\\ -(\mu+\mu^{-1})\\ 1\end{bmatrix}\\
&\quad= \gamma\begin{bmatrix}1+\mu^m-(\mu+\mu^{-1})\beta_1+\beta_2\\ -\mu^{-1}-\mu^{m+1}+\beta_1\\ 0\\ \vdots\\ 0\\ -\mu^{-1}-\mu^{m+1}+\beta_1\\ 1+\mu^m-(\mu+\mu^{-1})\beta_1+\beta_2\end{bmatrix}\\
&\quad= \gamma[\upsilon_0\ \upsilon_1\ 0\ 0\ \dots\ 0\ \upsilon_1\ \upsilon_0]^{\mathrm{T}},
\end{aligned}$$

where $\upsilon_0=1+\mu^m-(\mu+\mu^{-1})\beta_1+\beta_2$ and $\upsilon_1=-\mu^{-1}-\mu^{m+1}+\beta_1$.

Now if θ is a zero of $\mathscr{P}_0\{a_\mu\}$, then also θ^{-1} is a zero since $\mathscr{P}_0\{a_\mu\}$ is symmetric. We can therefore deconvolve both zeros to obtain $\mathscr{P}_0\{a_\mu\}=C_\theta d$, where C_θ is defined as above. By defining $\hat{\beta}_1=\upsilon_1+\theta^{-1}+\theta^{m+1}$ and $\hat{\beta}_2=\upsilon_0+(\theta+\theta^{-1})\upsilon_1+\theta^{-2}+\theta^{m+2}$, and following the above calculation in reverse, we obtain

$$\begin{aligned}
&R_{m+2}C_\mu\mathscr{P}_0\{a_\mu\}\\
&\quad= R_{m+2}C_\mu C_\theta d=\gamma[\upsilon_0\ \upsilon_1\ 0\ 0\ \dots\ 0\ \upsilon_1\ \upsilon_0]^{\mathrm{T}}
\end{aligned}$$

$$
\begin{aligned}
&= \gamma \begin{bmatrix} 1+\theta^m-(\theta+\theta^{-1})\hat{\beta}_1+\hat{\beta}_2 \\ -\theta^{-1}-\theta^{m+1}+\hat{\beta}_1 \\ 0 \\ \vdots \\ 0 \\ -\theta^{-1}-\theta^{m+1}+\hat{\beta}_1 \\ 1+\theta^m-(\theta+\theta^{-1})\hat{\beta}_1+\hat{\beta}_2 \end{bmatrix} \\
&= \gamma \begin{bmatrix} c_\theta^0+Jc_\theta^0 & \hat{\beta}_1 & \hat{\beta}_2 \\ \hat{\beta}_1 & c_\theta^0+Jc_\theta^0 & \hat{\beta}_1 \\ \hat{\beta}_2 & \hat{\beta}_2 & c_\theta^0+Jc_\theta^0 \end{bmatrix} \begin{bmatrix} 1 \\ -(\theta+\theta^{-1}) \\ 1 \end{bmatrix} \\
&= R_{m+2} \begin{bmatrix} \mathscr{P}_0\{a_\theta\} & 0 & 0 \\ 0 & \mathscr{P}_0\{a_\theta\} & 0 \\ 0 & 0 & \mathscr{P}_0\{a_\theta\} \end{bmatrix} \begin{bmatrix} 1 \\ -(\theta+\theta^{-1}) \\ 1 \end{bmatrix} \\
&= R_{m+2}C_\theta \mathscr{P}_0\{a_\theta\}. \qquad \text{(A.2)}
\end{aligned}
$$

Now $R_{m+2}C_\mu\mathscr{P}_0\{a_\mu\} = R_{m+2}C_\theta\mathscr{P}_0\{a_\theta\}$ according to Eq. (A.2), then $C_\mu\mathscr{P}_0\{a_\mu\} = C_\theta\mathscr{P}_0\{a_\theta\}$ since R_{m+2} is full rank, and it follows that $\mathscr{P}_0\{a_\mu\}$ and $\mathscr{P}_0\{a_\theta\}$ have all the same zeros except for θ and μ, respectively. The proof for antisymmetric LSP polynomials is analogous. □

Lemma 10. *Let the set of vectors d_i be such that those a_i that solve $Ra_i = d_i$ are minimum-phase for all positive definite, symmetric, real, Toeplitz matrices R. Furthermore, let a be the solution to*

$$Ra = Dg,$$

where $D = [d_0\ d_1\ d_2\ \ldots]$ and $g = [\gamma_0\ \gamma_1\ \gamma_2 \ldots]^{\mathrm{T}}$. If $\gamma_i > 0\ \forall i$, then a is minimum-phase.

This lemma has been proved in [32]. Now, we are finally ready to proceed to the proof of Theorem 7.

Proof of Theorem 7. Let us study the symmetric LSP polynomial $\mathscr{P}_0\{a_\nu\}$ and the antisymmetric LSP will again follow by similar arguments.

We now observe that:

(1) According to Lemma 8, the LSP polynomial of a conventional LP model formed from the filtered signal $c_n * x_n$ has the form

$$R_m\mathscr{P}_0\{a_\mu\} = \gamma(c_\mu^0 + Jc_\mu^0), \qquad \text{(A.3)}$$

where μ^{-1} is the single zero of the FIR-filter c_n.

(2) According to Theorem 3, the zeros of $\mathscr{P}_0\{a_\mu\}$ solving Eq. (A.3) are separate and on the unit circle.
(3) Choosing $\mu = 0$ in Eq. (A.3), we arrive to a form identical to Eq. (10), and therefore, $\mathscr{P}_0\{a_{\mu=0}\}$ is the LSP polynomial of order m of the unfiltered signal.
(4) Choosing $\mu = \pm 1$ in Eq. (A.3), we arrive to a form identical to Eq. (11) and therefore, $\mathscr{P}_0\{a_{\mu=\pm1}\}$ is the LSP polynomial of order $m+1$ of the unfiltered signal, with one trivial zero removed.
(5) Since LSP polynomials of consecutive orders are interlaced according to Theorem 4, then also $\mathscr{P}_0\{a_{\mu=0}\}$ and $\mathscr{P}_0\{a_{\mu=\pm1}\}$ are interlaced.
(6) Due to Theorem 9, polynomials $\mathscr{P}_0\{a_\nu\}$ and $\mathscr{P}_0\{a_\theta\}$cannot have joint zeros unless all zeros are equal (which would imply $\nu = \theta$).
(7) The zeros of $\mathscr{P}_0\{a_\nu\}$ follow continuous tracks as a function of ν [18].

Let $\mathscr{P}_0\{a_\nu\}$ and $\mathscr{P}_0\{a_\theta\}$ solve Eq. (A.3) for some scalars ν and θ with $-1 \leqslant \nu < \theta \leqslant +1$. Now suppose that the zeros of $\mathscr{P}_0\{a_\nu\}$ and $\mathscr{P}_0\{a_\theta\}$ are *not* interlaced. If we let $\widehat{\nu}$ travel from -1 to ν and $\widehat{\theta}$ from $+1$ to θ, then at $\widehat{\nu} = -1$ and $\widehat{\theta} = +1$ the zeros of $\mathscr{P}_0\{a_{\widehat{\nu}}\}$ and $\mathscr{P}_0\{a_{\widehat{\theta}}\}$ are interlaced, but for $\widehat{\nu} = \nu$ and $\widehat{\theta} = \theta$ they are not interlaced. Therefore, there must exist a point $\widehat{\nu}' \in (-1, \nu)$ and $\widehat{\theta}' \in (\theta, +1)$ where the zeros pass each other, that is they overlap, since the transformation is continuous and the zeros are always on the unit circle. Since the two models cannot have joint zeros unless all zeros are equal, then our assumption must be false, the zeros of $\mathscr{P}_0\{a_\nu\}$ and $\mathscr{P}_0\{a_\theta\}$ must be interlaced. This proves cases (a) and (b).

Let a_ν solve $Ra_\nu = \gamma c_\nu^0$, where $c_\nu^0 = [1\ \nu\ \ldots\ \nu^m]^{\mathrm{T}}$ and a_ν is thus minimum-phase. Then, also $a_{\nu,\Gamma}$ is minimum-phase if it solves

$$Ra_{\nu,\Gamma} = \gamma[c_\nu^0 + \Gamma c_\nu^0] \qquad \text{(A.4)}$$

and $\Gamma \in (-1, +1)$, since adjusting $\Gamma \in (-1, +1)$ corresponds to adjusting the last reflection coefficient in the open interval $(-1, +1)$.

Let a_1 and a_2 be solutions to Eq. (A.4) with ν_1, Γ_1 and ν_2, Γ_2, respectively, and then their sum $a = a_1 + a_2$ is minimum-phase due to Lemma 10. Moreover, then $\mathscr{P}_0\{a\} = \mathscr{P}_0\{a_1\} + \mathscr{P}_0\{a_2\}$ and $\mathscr{Q}_0\{a\} = \mathscr{Q}_0\{a_1\} + \mathscr{Q}_0\{a_2\}$.

If we go arbitrarily close to the edges, $\Gamma_1 \to +1$ and $\Gamma_2 \to -1$ then $\mathscr{Q}_0\{a_1\} \to 0$ and $\mathscr{P}_0\{a_2\} \to 0$. It follows that $a \to \mathscr{P}_0\{a_1\} + \mathscr{Q}_0\{a_2\}$. Since a is

minimum-phase as long as we stay inside the limits (even when we go arbitrarily close to the limits), then $\mathscr{P}_0\{a_1\}$ and $\mathscr{Q}_0\{a_2\}$ must be interlaced for all inside points $\Gamma_i, v_i \in (-1,+1)$. Furthermore, since $\mathscr{P}_0\{a_1\}$ and $\mathscr{P}_0\{a_2\}$ are interlaced due to the first part of this proof, $\mathscr{P}_0\{a_1\}$ and $\mathscr{Q}_0\{a_2\}$ are interlaced also at the edges $\Gamma_1 = 1$ and $\Gamma_2 = -1$ as long as the v_is are strictly inside $v_i \in (-1,+1)$. This proves cases (c) and (d). □

References

[1] F. Itakura, Line spectrum representation of linear predictive coefficients of speech signals, J. Acoust. Soc. Amer. 57 (Suppl. 1) (1975) 35.

[2] J. Makhoul, Linear prediction: a tutorial review, Proc. IEEE 63 (5) (April 1975) 561–580.

[3] F.K. Soong, B.-H. Juang, Line spectrum pair (LSP) and speech data compression, in: Proceedings of the IEEE Acoustics, Speech, and Signal Processing ICASSP'84, vol. 1, San Diego, CA, March 1984, pp. 1.10.1–1.10.4.

[4] K.K. Paliwal, A study of line spectrum pair frequencies for vowel recognition, Speech Commun. 8 (1989) 27–33.

[5] C.S. Liu, M.T. Lin, W.J. Wang, H.C. Wang, Study of line spectrum pair frequencies for speaker recognition, in: Proceedings of the IEEE Acoustics, Speech, and Signal Processing ICASSP'90, 1990, pp. 277–280.

[6] H.K. Kim, K.C. Kim, H.S. Lee, Enhanced distance measure for LSP-based speech recognition, Electron. Lett. 29 (16) (1993) 1463–1465.

[7] M. Hasegawa-Johnson, Line spectral frequencies are poles and zeros of the glottal driving-point impedance of a discrete matched-impedance vocal tract model, J. Acoust. Soc. Amer. 108 (July 2000) 457–460.

[8] R.W. Morris, M.A. Clements, Modification of formants in the line spectrum domain, IEEE Signal Process. Lett. 9 (1) (January 2002) 19–21.

[9] H.W. Schüssler, A stability theorem for discrete systems, IEEE Trans. Acoust. Speech Signal Process. ASSP-24 (1) (February 1976) 87–89.

[10] T. Bäckström, Linear Predictive Modelling of Speech—constraints and line spectrum pair decomposition, Ph.D. Thesis, Helsinki University of Technology (TKK), Espoo, Finland, 2004, ⟨http://lib.tkk.fi/Diss/2004/isbn9512269473/⟩.

[11] J. Rothweiler, On polynomial reduction in the computation of LSP frequencies, IEEE Trans. Speech Audio Process. 7 (5) (1999) 592–594.

[12] B. Bessette, R. Salami, R. Lefebvre, M. Jelínek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, K. Järvinen, The adaptive multirate wideband speech codec (AMR-WB), IEEE Trans. Speech Audio Process. 10 (8) (November 2002) 620–636.

[13] Y. Bistritz, S. Peller, Imittance spectral pairs (ISP) for speech encoding, in: Proceedings of the IEEE Acoustics, Speech, and Signal Processing ICASSP'93, vol. 2, 1993, pp. 9–12.

[14] Y. Bistritz, H. Lev-Ari, T. Kailath, Immittance-domain Levinson algorithms, IEEE Trans. Inform. Theory 35 (May 1989) 675–682.

[15] Y. Bistritz, Reflections on Schur–Cohen matrices and Jury–Marden tables and classification of related unit-circle zero location criteria, Circuits Systems Signal Process. 15 (1996) 111–136.

[16] G.A. Mian, G. Riccardi, A localization property of line spectrum frequencies, IEEE Trans. Speech Audio Process. 2 (4) (1994) 536–539.

[17] H.K. Kim, H.S. Lee, Interlacing properties of line spectrum pair frequencies, IEEE Trans. Speech Audio Process. 7 (1) (1999) 87–91.

[18] R. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, Cambridge, UK, 1999.

[19] N. Levinson, The Wiener RMS error criterion in filter design and prediction, J. Math. Phys. 25 (1947) 261–278.

[20] J. Durbin, The fitting of time series models, Rev. Internat. Statist. Inst. 23 (1960) 233–244.

[21] P. Delsarte, Y.V. Genin, The split Levinson algorithm, IEEE Trans. Acoust. Speech Signal. Process. ASSP-34 (3) (1986) 470–478.

[22] M.H. Hayes, Statistical Digital Signal Processing and Modeling, Wiley, New York, 1996.

[23] G. Golub, C. van Loan, Matrix Computations, The Johns Hopkins University Press, London, 1996.

[24] S.A. Tretter, The all-pole model is theoretically stable, IEEE Trans. Audio and Electroacoust. AU-20 (4) (October 1972) 316.

[25] S.W. Lang, J.H. McClellan, A simple proof of stability for all-pole linear prediction models, Proc. IEEE 67 (5) (May 1979) 860–861.

[26] S. Treitel, T.J. Ulrych, A new proof of the minimum-phase property of the unit prediction error operator, IEEE Trans. Acoust. Speech Signal Process. ASSP-27 (1) (February 1979) 99–100.

[27] P. Stoica, A. Nehorai, On stability and root location of linear prediction models, IEEE Trans. Acoust. Speech Signal Process. ASSP-35 (4) (April 1987) 582–584.

[28] H. Krishna, New split Levinson, Schur and lattice algorithms for digital signal processing, in: Proceedings of the IEEE Acoustics Speech and Signal Processing ICASSP'88, vol. 3, 1988, pp. 1640–1642.

[29] Y. Wang, H. Krishna, B. Krishna, Split Levinson algorithm is weakly stable, in: Proceedings of the IEEE Acoustics, Speech, and Signal Processing ICASSP'89, Glasgow, UK, 1989, pp. 1215–1218.

[30] R.E. Caflisch, An inverse problem for Toeplitz matrices and the synthesis of discrete transmission lines, Linear Algebra Appl. 38 (1981) 207–225.

[31] W.B. Kleijn, T. Bäckström, P. Alku, On line spectral frequencies, IEEE Signal Process. Lett. 10 (3) (2003) 75–77.

[32] T. Bäckström, P. Alku, A constrained linear predictive model with the minimum-phase property, Signal Processing 83 (10) (October 2003) 2259–2264.

[33] P. Stoica, A. Nehorai, The poles of symmetric linear prediction models lie on the unit circle, IEEE Trans. Acoust. Speech Signal Process. ASSP-34 (October 1986) 1344–1346.

[34] T. Bäckström, P. Alku, W.B. Kleijn, A time-domain reformulation of linear prediction equivalent to the LSP decomposition, in: Proceedings of the IEEE Acoustics, Speech, and Signal Processing ICASSP'02, vol. 1, Orlando, FL, USA, 2002, pp. 661–664.

[35] T. Bäckström, P. Alku, T. Paatero, W.B. Kleijn, A time-domain interpretation for the LSP-decomposition, IEEE Trans. Speech Audio Process. (November 2004) 554–560.

[36] P. Kabal, R.P. Ramachandran, The computation of line spectral frequencies using Chebyshev polynomials, IEEE Trans. Acoust. Speech Signal Process. ASSP-34 (6) (1986) 1419–1426.
[37] P. Lakatos, On zeros of reciprocal polynomials, Publ. Math. Debrecen 61 (2002) 645–661.
[38] N. Farvardin, R. Laroia, Efficient encoding of speech LSP parameters using the discrete cosine Transformation, in: Proceedings of the IEEE Acoustics, Speech, and Signal Processing ICASSP'89, vol. 1, 1989, pp. 168–171.
[39] S. Grassi, A. Dufaux, M. Ansorge, F. Pellandini, Efficient algorithm to compute LSP parameters from 10th-order LPC coefficients, in: Proceedings of the IEEE Acoustics, Speech, and Signal Processing ICASSP'97, vol. 3, 1997, pp. 1547–1550.
[40] J. Rothweiler, A rootfinding algorithm for line spectral frequencies, in: Proceedings of the IEEE Acoustics, Speech, and Signal Processing ICASSP'99, vol. 2, 1999, pp. 661–664.
[41] J.-W. Tourneret, Statistical properties of line spectrum pairs, Signal Processing 65 (1998) 239–255.
[42] P. Alku, U. Laine, Bi-directional linear prediction—a new LPC-based method for signal processing, in: Proceedings of the AMSE International Conference Signals & Systems, Brighton, UK, vol. 1, 1989, pp. 101–114.
[43] A.-C. Lee, A new autoregressive method for high-performance spectrum analysis, J. Acoust. Soc. Amer. 86 (1) (1989) 150–157.
[44] S. David, B. Ramamurthi, Two-sided filters for frame-based prediction, IEEE Trans. Signal Process. 39 (4) (1991) 789–794.
[45] J.-J. Hsue, A.E. Yagle, Blind deconvolution of symmetric noncausal impulse responses using two-sided linear prediction, IEEE Trans. Signal Process. 42 (6) (June 1994) 1509–1518.
[46] S.H. Leung, H.C. Ng, K.F. Wong, Speech synthesis using two-sided linear prediction parameters, in: Proceedings of the Speech, Image Processing and Neural Networks, ISSIPNN'94, Hong Kong, April 1994.
[47] D.M. Goodman, E.K. Miller, A note on minimizing the prediction error when the zeros are restricted to the unit circle, IEEE Trans. Acoust. Speech Signal Process. ASSP-30 (3) (June 1982) 503–505.
[48] P. Stoica, A. Nehorai, On linear prediction models constrained to have unit-modulus poles and their use for sinusoidal frequency estimation, IEEE Trans. Acoust. Speech Signal Process. 36 (6) (June 1988) 940–942.

Study IV

Tom Bäckström, Carlo Magi, Paavo Alku. "Minimum separation of line spectral frequencies", IEEE Signal Processing Letters, Vol. 14, No. 2, pp. 145-147, 2007.

Minimum Separation of Line Spectral Frequencies

Tom Bäckström, *Member, IEEE*, Carlo Magi, and Paavo Alku

***Abstract*—We provide a theoretical lower limit on the distance of line spectral frequencies for both the line spectrum pair decomposition and the immittance spectrum pair decomposition. The result applies to line spectral frequencies computed from linear predictive polynomials with all roots within a zero-centered circle of radius $r < 1$.**

***Index Terms*—Immittance spectrum pair, linear prediction, line spectrum pair, root finding, speech coding.**

I. Introduction

LINE SPECTRUM pair (LSP) decomposition is a frequently used method in speech processing as a representation of linear predictive models [1]. In the LSP decomposition, a linear predictive (LP) polynomial $A(z)$ is decomposed into two polynomials, one symmetric and one antisymmetric, the zeros of which are separate, on the unit circle and interlaced [2]. Since the zeros are on the unit circle, they can be represented by their angles only, the line spectral frequencies (LSFs). Given an LP polynomial $A(z)$, the computation of LSFs calls for numerical root-finding algorithms, which generally employ finite step-sizes or search grids in their iteration process. These root-finding algorithms typically use Chebyshev-polynomials in expression of the LSP decomposition [3]. If the LSFs lie arbitrarily close to each other, it is impossible to determine a fixed step-size small enough to ensure that all zeros are found. Conversely, a lower limit on LSF separation enables us to choose a step-size small enough such that all zeros are recovered.

Our result applies to a polynomial $A(z)$ whose zeros are, in the complex plane, within a circle centered at origin with radius $0 < r < 1$. For example, predictive polynomials, calculated with the autocorrelation method, have zeros within a radius of $r = \cos[\pi/(N+m+1)]$, where $N+1$ is window length (in samples) and m model order [4].

II. Derivation of Lower Limit

Given a real LP polynomial $A(z) = \sum_{k=0}^{m} a_k z^{-k}$ of order m, the corresponding LSP polynomials are defined as [2]

$$\begin{aligned} P(z) &= A(z) + z^{-m-1}A(z^{-1}) \\ Q(z) &= A(z) - z^{-m-1}A(z^{-1}). \end{aligned} \tag{1}$$

Manuscript received January 18, 2006; revised June 5, 2006. This work was supported by the Ministry of Education and the Academy of Finland under Project Number 205962. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Patrick A. Naylor.

The authors are with the Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Espoo, Finland.

Digital Object Identifier 10.1109/LSP.2006.881514

Alternatively, and slightly more generally, we can decompose $A(z)$ into symmetric and antisymmetric polynomials $F_+(z)$ and $F_-(z)$, respectively, as

$$F_\pm(z) = A(z) \pm z^{-m-p}A(z^{-1}) \tag{2}$$

where $p \geq 0$. The notation thus encompasses both LSP ($p = 1$), and the immittance spectrum pair (ISP) forms ($p = 0$) [5]. Polynomials $F_\pm(z)$ can be rewritten as

$$\begin{aligned} F_\pm(z) &= A(z)\left[1 \pm \frac{z^{-m-p}A(z^{-1})}{A(z)}\right] \\ &= A(z)[1 \pm H(z)] \end{aligned}$$

where $H(z) = z^{-m-p}A(z^{-1})/A(z)$ is all-pass. Then zeros of $F_\pm(z)$ appear at points $\hat{z}_n = e^{i\hat{\omega}_n}$, where $H(\hat{z}_n) = \pm 1$.

The phase of $H(e^{i\omega}) = e^{i\phi(\omega)}$ can be written as [6, p. 235]

$$\phi(\omega) = -(m+p)\omega - \sum_{k=1}^{m} 2\arctan\frac{r_k \sin(\omega - \omega_k)}{1 - r_k\cos(\omega - \omega_k)}$$

and its negative derivative, the group delay as [6, p. 238]

$$\begin{aligned} \tau(\omega) &= -\frac{\partial\phi(\omega)}{\partial\omega} \\ &= p + \sum_{k=1}^{m} \frac{1 - r_k^2}{1 + r_k^2 - 2r_k\cos(\omega - \omega_k)} \end{aligned}$$

where $z_k = r_k e^{\omega_k i}$ are the zeros of $A(z)$. We also note that $\tau(\omega) > 0$ for all ω. Now, the zeros of $F_\pm(z)$ appear at points where $\phi(\hat{\omega}_n) = n\pi$.

The group delay is limited from above by

$$\begin{aligned} \tau(\omega) &\leq p + \sum_{k=1}^{m} \frac{1 - r_k^2}{(1 - r_k)^2} = p + \sum_{k=1}^{m} \frac{1 + r_k}{1 - r_k} \\ &\leq p + m\frac{1 + r_{\max}}{1 - r_{\max}} \end{aligned} \tag{3}$$

where $r_{\max} = \max_k r_k$. Note that the group delay can reach the limit only at angle $\omega = \omega_k$ and when all roots of $A(z)$ lie at the same point in the Z-domain $z_k = z_j$ for all j. For $A(z)$ of real coefficients, this implies that all the roots are located at the same point on the real axis.

Since $\tau(\omega)$ is continuous, we can use the Mean Value Theorem with $\omega_0 \in (\omega, \omega + \Delta\omega)$, and we have

$$\begin{aligned} \left|\frac{\Delta\phi(\omega)}{\Delta\omega}\right| &= \left|\frac{\partial\phi(\omega_0)}{\partial\omega}\right| = |\tau(\omega_0)| \\ &\leq p + m\frac{1 + r_{\max}}{1 - r_{\max}} \end{aligned}$$

since $\tau(\omega) > 0$. As zeros of $F_{\pm}(z)$ appear at $\phi(\hat{\omega}_n) = n\pi$, we set $\Delta\phi(\omega_n) \triangleq \pi$ and obtain a lower limit for the LSF separation as

$$|\Delta\omega_n| \geq \frac{\pi(1 - r_{\max})}{m + p + (m - p)r_{\max}}. \quad (4)$$

III. Chebyshev Transformation Domain

Generally, root-finding algorithms do not search roots directly from $F_{\pm}(z)$, but first, remove trivial zeros at $z = \pm 1$ and second, perform the Chebyshev transformation to reduce the complexity of root-finding [3]. Let us denote the Chebyshev-transformed polynomial with $G(x)$. Then $x = \cos\omega$, where $z = e^{i\omega}$. In the Chebyshev domain, the zeros $z_k = e^{i\omega_k}$ of $F_{\pm}(z)$ become $x_k = \cos\omega_k$, and $G(x_k) = 0$. Then, a lower limit for separation of LSFs in Chebyshev domain for $(\omega_n, \omega_n + \Delta\omega_n) \subset [0, \pi]$ is

$$\begin{aligned} &\Delta x(\omega_n) \\ &\quad = \cos\omega_n - \cos(\omega_n + \Delta\omega_n) \\ &\quad \geq \cos\omega_n - \cos\left(\omega_n + \frac{\pi(1 - r_{\max})}{m + p + (m - p)r_{\max}}\right). \end{aligned} \quad (5)$$

Note that $\Delta x(\omega_n) \geq 0$. This limit varies as a function of ω_n, and it is therefore useful, in root-finding in the Chebyshev domain, to use a step-size that varies with ω_n.

IV. Numerical Examples and Discussion

The acquired lower limits of root separation for ISP and LSP are illustrated in Fig. 1 as a function of maximum root radius $0 < r_{\max} < 1$ for $m = 8$. The continuous line represents the theoretical lower limit of (4). The worst-case polynomial, that is, the polynomial with all zeros at the same point on the real axis, is depicted with a dotted line. The fact that this polynomial is indeed the worst case in terms of LSF separation is evident from (3). Minimum root separation obtained from LP polynomials computed from natural speech signals, represented by eight Finnish sustained vowels /a/, /e/, /i/, /o/, /u/, /y/, /ä/, and /ö/, sampled at 8 kHz, are depicted with dashed lines. Varying maximum root radius $\tilde{r}_{\max}$ was acquired by substitution of variables $\tilde{z} \triangleq \tilde{r}_{\max} z / r_{\max}$. It is evident that the worst-case polynomial is well below the typical range of $\min\Delta\omega$ with vowel sounds, since the roots of $A(z)$ are, in general, distributed to cover the whole frequency range, and only some roots are close to $r_{\max}$ in amplitude. However, polynomials close to the worst-case polynomial *can occur*, even though they are undoubtedly extremely rare. Therefore, the worst case must be taken into account when designing practical applications.

We find also that the lower limit is a very accurate estimate, since the worst-case polynomials have a minimum root separation that is practically equal to the lower limit. Furthermore, we observe that ISP has always a larger minimum separation, since ISP polynomials have one root fewer than LSP polynomials. However, the difference between root separation of ISP and LSP is, for $r_{\max} \approx 1$, negligible. The LSP limit is thus a reasonable lower limit for both LSP and ISP.

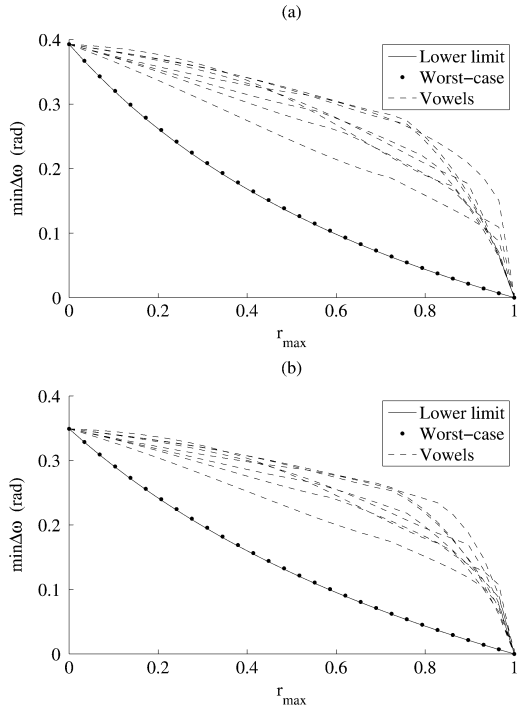


Fig. 1. Minimum root separation $\Delta\omega$ of (a) an ISP and (b) an LSP polynomial of degree $m = 8$ as a function of maximum root radius $r_{\max}$. Continuous line is the theoretical lower limit, dotted line is a simulation result for the worst-case polynomial, and dashed lines are examples of eight Finnish vowels.

TABLE I
Typical Combinations of Model Order m, Sampling Frequency F_s, and Frame Length l (Where $N = F_s l/1000$) for Linear Prediction With the Autocorrelation Method and Their Corresponding Minimum LSF Separation $\Delta\omega$. The Parameter Values Selected Are in Use in Speech Coding Applications [7]–[9]

Model order	Sampling frequency	Frame length	$\min\Delta\omega$ (rad)
8	8 kHz	20 ms	3.3×10^{-5}
10	8 kHz	20 ms	2.6×10^{-5}
16	12.8 kHz	20 ms	6.5×10^{-6}

A list of minimum LSF separation for commonly used model orders is given in Table I [7]–[9]. We observe that the theoretical minimum LSF separations are rather small. For example, AMR-NB uses a search grid with step-size $\pi/60$ (rad), which is several orders of magnitude larger than the values obtained in Table I [8]. Moreover, numerical round-off errors can increase the $r_{\max}$ limits used in Table I and thus decrease LSF separation limits. However, these values for LSF separation are not directly comparable since coders generally use white-noise correction and lag-windowing of the autocorrelation, both of which, effectively, reduce $r_{\max}$, thus also increasing the minimum separation of LSFs.

As a final example, we present one potential application of the derived lower limit of line spectral frequencies. Fig. 2 shows a speech spectrum of the Finnish vowel /i/, produced by a male speaker and sampled at 16 kHz (only the lower audio band of

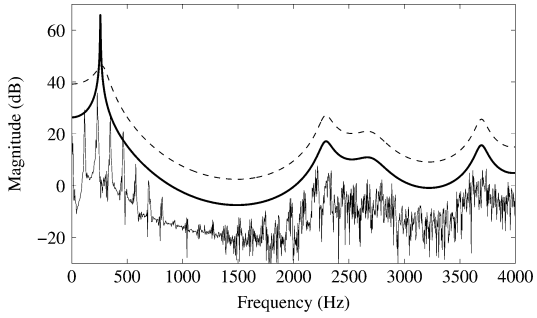


Fig. 2. Illustration of bandwidth expansion of a spectral peak by constraining the LSF separation for the vowel /i/. The thin continuous line represents the speech spectrum, the thick continuous line an all-pole model with an underestimated formant bandwidth, and the dashed line an all-pole model with constraint on the LSF separation. Spectra have been displaced by 10 dB for visual clarity.

4 kHz is displayed for visual clarity). The continuous line represents an all-pole model of order $m = 18$ estimated from the speech signal. In this spectral model, the original LP parameters have been distorted by quantization that has resulted in a severe underestimation of the bandwidth of the first formant located at approximately 260 Hz. The minimum LSF separation for this model is 8.4×10^{-4} corresponding to the maximum root radius of $r_{\max} = 0.9996$. Limiting the LSF separation to 0.32 yields a maximum root radius of $r_{\max} = 0.98$. Consequently, the sharp spectral peak at the first formant has been removed from the corresponding all-pole model, as shown by the dashed line in Fig. 2.

V. Conclusion

We have derived a lower limit for the distance of adjacent line spectral frequencies for both the LSP decomposition and the ISP decomposition. The acquired theoretical limit can make applications using line spectra more efficient. This result applies to polynomials whose roots lie in an origin-centered circle of radius $r < 1$.

The current results could be used, for example, in either of the following purposes. First, numerical root-finding algorithms for extraction of LSFs can use the lower limit as a resolution threshold below which it is unnecessary to search for roots. Second, LP models can be modified to remove excessively sharp peaks by altering LSFs. By moving LSFs further away from each other, we can guarantee a minimum bandwidth for spectral peaks.

References

[1] W. B. Kleijn and K. K. Paliwal, "Quantization of LPC parameters," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam, The Netherlands: Elsevier, 1995, pp. 433–466.

[2] F. K. Soong and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, San Diego, CA, Mar. 1984, vol. 1, pp. 1.10.1–1.10.4.

[3] P. Kabal and R. P. Ramachandran, "The computation of line spectral frequencies using Chebyshev polynomials," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 6, pp. 1419–1426, Dec. 1986.

[4] P. Delsarte, Y. Genin, and Y. Kamp, "Stability of linear predictors and numerical range of a linear operator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 3, pp. 470–478, May 1986.

[5] Y. Bistritz and S. Peller, "Imittance spectral pairs (ISP) for speech encoding," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1993, vol. 2, pp. 9–12.

[6] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[7] *GSM Full Rate Speech Transcoding*, Recommendation GSM 06.10, ETSI, 1992.

[8] *Digital Cellular Telecommunications System (Phase 2+); Adaptive Multi-Rate (AMR) Speech Transcoding*, GSM 06.90 version 7.2.1, ETSI, Apr. 2000.

[9] *3rd Generation Partnership Project: AMR Wideband Speech Codec*, 3GPP TS 26.190 V5.0.0, 3GPP, Mar. 2001.

Study V

Tom Bäckström, Carlo Magi. "Effect of white-noise correction on linear predictive coding", IEEE Signal Processing Letters, Vol. 14, No. 2, pp. 148-151, 2007.

Effect of White-Noise Correction on Linear Predictive Coding

Tom Bäckström, *Member, IEEE*, and Carlo Magi

***Abstract*—White-noise correction is a technique used in speech coders using linear predictive coding (LPC). This technique generates an artificial noise-floor in order to avoid stability problems caused by numerical round-off errors. In this letter, we study the effect of white-noise correction on the roots of the LPC model. The results demonstrate in analytic form the relation between the noise floor level and the stability radius of the LPC model.**

***Index Terms*—Autocorrelation function, linear predictive coding (LPC), numerical stability, speech coding, stability radius, white-noise correction.**

I. Introduction

In SPEECH processing, linear prediction is a classical and often used method for modeling of the spectral envelope of a speech signal [1]. The envelope is represented by an mth-order polynomial $A(z) = \sum_{k=0}^{m} a_k z^{-k}$, and the coefficient vector $\mathbf{a} = [a_0 \cdots a_m]^T$ is calculated from the normal equations

$$\mathbf{R_{m+1}a} = \sigma^2[1\,0\cdots 0]^T \tag{1}$$

where $\mathbf{R}_{m+1}$ is the autocorrelation matrix that is of size $(m+1) \times (m+1)$, symmetric, real, and has Toeplitz structure. If $\mathbf{R}_{m+1}$ is positive definite, then $A(z)$ is minimum-phase, i.e., it has all zeros inside the unit circle, and $A^{-1}(z)$ is stable. In some circumstances, which will be discussed in more detail in Section III, the stability criterion can be strengthened. Specifically, linear predictive models calculated using the autocorrelation criterion have their roots within an origin-centered circle of radius $r = \cos[\pi/(N+m)]$, where N is the frame length in samples [2].

However, when the autocorrelation is represented in the finite-length memory of a computer, the inevitable round-off errors can make the matrix $\mathbf{R}_{m+1}$ ill-conditioned or even jeopardize its positive definite property. In addition, problems in estimation of the spectrum can move zeros closer to the unit circle. It follows that the roots of $A(z)$ can be arbitrarily close to the unit circle or no longer remain within the unit circle, and the model could become unstable. Zeros close to the unit circle are not only troublesome in view of stability but also since the formants corresponding to such zeros have an unnaturally low bandwidth. Such formants cause severe reduction of perceived speech quality.

To fence these problems, it is possible to modify the autocorrelation matrix in such a way that the positive definite property is ensured by a positive margin. By adding a small positive constant μ to the autocorrelation r_k at lag $k = 0$ (where r_k is an element of the autocorrelation sequence at lag k), or equivalently, using the modified autocorrelation matrix

$$\widehat{\mathbf{R}}_{m+1} = \mathbf{R}_{m+1} + \mu\mathbf{I} \tag{2}$$

all eigenvalues of $\mathbf{R}_{m+1}$ are increased by μ, and for a sufficiently large μ, matrix $\widehat{\mathbf{R}}_{m+1}$ is positive definite and well-conditioned. This is a typical approach used in regularization of ill-conditioned matrices [3]. It has been used in speech coding for at least a quarter century [4], and it is being used in most of the major speech coding standards employing linear predictive coding, including, e.g., G.729, AMR, AMR-WB, and Speex [5]–[8]. Usually, we set $\mu = cr_0$, where c is a small constant. The choice of defining μ using r_0 is natural, since it can be interpreted as adding an artificial noise floor to the input signal, whose level depends only on the energy of the input signal. However, we find it surprising that, at least as far as the authors are aware, there is no conclusive evidence in the literature as to the behavior of the roots of $A(z)$ as a function of μ. Indeed, there are surprisingly few publications concerning white-noise correction overall.

In this letter, we are concerned with the roots of $A(z)$ as a function of $\mu > 0$ in the analytical case, where no round-off errors are present. We will show that the zeros lie in an origin-centered circle of radius $r_\mu < 1$, that is, the zeros are within the unit circle by a positive margin $1 - r_\mu$, where $r_\mu \in (0, 1)$ is a decreasing function of μ. This result is useful, for example, in the quantization of line spectral frequencies (LSFs), a representation of the linear predictive coefficients [9], since we can, with the aid of the stability radius, obtain a lower limit for the separation of LSFs and thus specify the required accuracy of numerical operations [10].

II. Autocorrelation Convergence Rate for Finite Sequences

The autocorrelation sequence r_k of a signal x_n, with $n \in (0, N-1)$, can be defined as

$$r_k = \sum_{j=0}^{N-|k|-1} x_j^* x_{j+k}.$$

Manuscript received March 14, 2006; revised June 23, 2006. This work was supported by the Ministry of Education and the Academy of Finland under Project Number 205962. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Alan McCree.

The authors are with the Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Helsinki, Finland.

Digital Object Identifier 10.1109/LSP.2006.881513

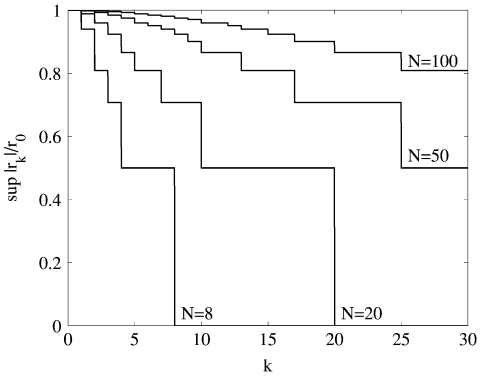


Fig. 1. Illustration of the upper limit for the autocorrelation sequence of signals of length $N = \{10, 20, 50, 100\}$.

Define the $N \times N$ down-shift matrix as $\mathbf{S}_{ij} = \delta_{i-j+1}$, that is,

$$\mathbf{S} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix}.$$

By setting $\mathbf{x} = [x_0 \cdots x_{N-1}]^T$, the autocorrelation can now be written as

$$r_k = \mathbf{x}^H \mathbf{S}^k \mathbf{x}.$$

The matrix $\mathbf{S}^k$ is nilpotent with power of nilpotency $\lceil N/k \rceil$, where $\lceil \cdot \rceil$ signifies rounding upward to the closest integer. The numerical radius, defined as $\rho(\mathbf{M}) = \sup_{\mathbf{y}} \{\mathbf{y}^H \mathbf{M} \mathbf{y} / \|\mathbf{y}\|^2\}$, for a nilpotent operator $\mathbf{F}$ is $\rho(\mathbf{F}) \leq \|\mathbf{F}\| \cos[\pi/(h+1)]$, where h is the power of nilpotency [11]. The autocorrelation sequence is therefore, for $k \in \{\pm 1 \cdots \pm N\}$, limited as

$$\begin{aligned} |r_k| = |\mathbf{x}^H \mathbf{S}^k \mathbf{x}| &\leq \|\mathbf{x}\|^2 \rho(\mathbf{S}^k) \\ &\leq r_0 \cos\left[\frac{\pi}{\lceil N/k \rceil + 1}\right]. \end{aligned} \tag{3}$$

This limit is illustrated in Fig. 1.

III. Root Loci With White-Noise Correction

The coefficients $\mathbf{a} = [a_0 \cdots a_m]^T$ of a linear predictive model $A(z) = \sum_{k=0}^{m} a_k z^{-k}$ are, in the autocorrelation method, calculated from the normal equations of (1). Applying white-noise correction to (1), we obtain

$$(\mathbf{R}_{m+1} + \mu \mathbf{I})\mathbf{a} = \sigma^2 [1\, 0 \cdots 0]^T. \tag{4}$$

Following the principal idea of [12, Ch. 5], let α be a zero of $A(z)$ and therefore $A(z) = B(z)(1 - \alpha z^{-1})$, with $B(z) = \sum_{k=0}^{m-1} b_k z^{-k}$. In matrix form, we have

$$\mathbf{a} = \begin{bmatrix} b_0 & 0 \\ b_1 & b_0 \\ \vdots & \vdots \\ b_{m-1} & b_{m-2} \\ 0 & b_{m-1} \end{bmatrix} \begin{bmatrix} 1 \\ -\alpha \end{bmatrix} = \mathbf{B} \begin{bmatrix} 1 \\ -\alpha \end{bmatrix}.$$

Multiplying (4) from the left by $\mathbf{B}^H$, we obtain

$$\left(\mathbf{B}^H \mathbf{R}_{m+1} \mathbf{B} + \mu \mathbf{B}^H \mathbf{B}\right) \begin{bmatrix} 1 \\ -\alpha \end{bmatrix} = \begin{bmatrix} \sigma^2 b_0^* \\ 0 \end{bmatrix}. \tag{5}$$

We can readily see that the matrix $\mathbf{U} = \mathbf{B}^H \mathbf{R}_{m+1} \mathbf{B}$ is the autocorrelation matrix of the convolved signal $x_n * b_n$. This signal is of length $N + m - 1$, and we thus have

$$\mathbf{U} = \begin{bmatrix} u_0 & u_1^* \\ u_1 & u_0 \end{bmatrix}$$

with $|u_1| \leq u_0 \cos[\pi/(N+m)]$ due to (3). By defining $\mathbf{b} = [b_0 \cdots b_{m-1}]^T$, we obtain $u_0 = \mathbf{b}^H \mathbf{R}_m \mathbf{b} \leq \|\mathbf{b}\|^2 \|\mathbf{R}_m\|$, where $\mathbf{R}_m$ is the $m \times m$ principal sub-matrix of $\mathbf{R}_{m+1}$.

Similarly, $\mathbf{V} = \mathbf{B}^H \mathbf{B}$ is the autocorrelation matrix of the sequence b_n. This sequence is of length m, and we thus have

$$\mathbf{V} = \begin{bmatrix} v_0 & v_1^* \\ v_1 & v_0 \end{bmatrix}$$

with $v_0 = \|\mathbf{b}\|^2$ and $|v_1| \leq \|\mathbf{b}\|^2 \cos[\pi/(m+1)]$.

Substituting $\mathbf{U}$ and $\mathbf{V}$ into (5), we obtain

$$\begin{cases} u_0 + \mu v_0 - \alpha(u_1 + \mu v_1)^* &= \sigma^2 b_0 \\ u_1 + \mu v_1 - \alpha(u_0 + \mu v_0) &= 0. \end{cases}$$

Solving for α and using the limits of u_1 and v_1, we find that

$$\begin{aligned} |\alpha| = \frac{|u_1 + \mu v_1|}{|u_0 + \mu v_0|} &\leq \frac{|u_1| + \mu |v_1|}{u_0 + \mu v_0} \\ &\leq \frac{\|\mathbf{b}\|^2 \|\mathbf{R}_m\| \cos[\frac{\pi}{N+m}] + \mu \|\mathbf{b}\|^2 \cos[\frac{\pi}{m+1}]}{\|\mathbf{b}\|^2 \|\mathbf{R}_m\| + \mu \|\mathbf{b}\|^2}. \end{aligned}$$

Finally, by choosing $\mu = \epsilon \|\mathbf{R}_m\|$, we conclude that

$$|\alpha| \leq \frac{\cos[\frac{\pi}{N+m}] + \epsilon \cos[\frac{\pi}{m+1}]}{1 + \epsilon}. \tag{6}$$

Note that when $\epsilon = 0$, the result simplifies to $|\alpha| \leq \cos[\pi/(N+m)]$, and it thus agrees with prior works [2].

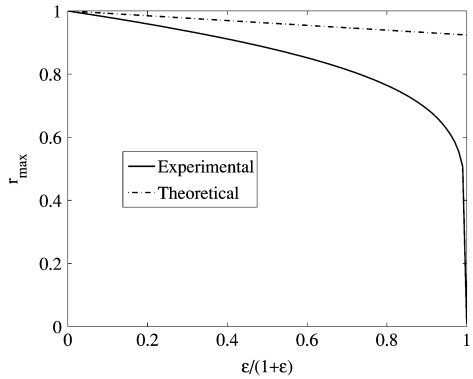


Fig. 2. Maximal root radius of the linear predictive model ($m = 8$) with white-noise correction as a function of ϵ scaled by $\|\mathbf{R}_m\|$. The dash-dotted line is the theoretical limit from (7), and the continuous line is the highest observed root radius for 1000 randomly generated models.

Moreover, when the window length is large $N \to \infty$, then $\cos[\pi/(N+m)] \to 1$ and (6) reduces to

$$|\alpha| \leq \frac{1+\epsilon\cos[\frac{\pi}{m+1}]}{1+\epsilon}. \tag{7}$$

As this limit is only slightly looser than that of (6), we can often use this simpler form.

Fig. 2 illustrates the limit of (7). In Fig. 2, we have used $\epsilon/(1+\epsilon)$ on the x-axis instead of ϵ in order to show the limit for all possible values of $\epsilon \geq 0$. We observe that for large ϵ, the limit is loose. The reason for this is that the limit used for $|v_1|$ overshoots, as it does not take into account the fact that all zeros of $B(z)$ travel simultaneously toward zero as a function of ϵ. However, since ϵ is small in practical applications, this is not a significant problem.

IV. Choice of Scaling Coefficient

Traditionally, the white-noise correction level μ has been defined using the lag 0 autocorrelation r_0, that is, $\mu = cr_0$. It is a choice motivated by convenience—the value of the scaling coefficient need not be calculated but can be extracted directly from the autocorrelation sequence. In addition, as r_0 is equal to the signal energy, the white-noise correction level is always set in relation to the signal energy, which is a well-warranted choice. The proof in Section III, however, used a different scaling $\mu = \epsilon\|\mathbf{R}_m\|$. In this section, we will describe the relation of these two coefficients.

Let $\mathrm{tr}(\cdot)$ denote the trace of a matrix, and λ_k its eigenvalues, with $\lambda_{\max}$ as the maximal eigenvalue; then due to the Toeplitz structure [13]

$$r_0 = \frac{\mathrm{tr}(\mathbf{R}_m)}{m} = \frac{1}{m}\sum_{k=0}^{m-1}\lambda_k \leq \lambda_{\max} = \|\mathbf{R}_m\|. \tag{8}$$

On the other hand, we can also limit r_0 from below by

$$r_0 = \frac{\mathrm{tr}(\mathbf{R}_m)}{m} = \frac{1}{m}\sum_{k=0}^{m-1}\lambda_k \geq \frac{\lambda_{\max}}{m} = \frac{\|\mathbf{R}_m\|}{m}. \tag{9}$$

Both limits are sharp in the sense that they can be reached for some combination of the eigenvalues $\lambda_k \geq 0$. In other words, r_0 lies in the interval $\|\mathbf{R}_m\|/m \leq r_0 \leq \|\mathbf{R}_m\|$, and we can estimate r_0 below by $\|\mathbf{R}_m\|/m$. This scaling, $\mu = \epsilon\|\mathbf{R}_m\|/m$, gives a slightly larger stability radius but agrees with convention. The limits corresponding to (6) and (7), using $\mu = \epsilon\|\mathbf{R}_m\|/m$, are thus

$$|\alpha| \leq \frac{\cos[\frac{\pi}{N+m}] + \frac{\epsilon}{m}\cos[\frac{\pi}{m+1}]}{1+\frac{\epsilon}{m}} \tag{10}$$

and

$$|\alpha| \leq \frac{1+\frac{\epsilon}{m}\cos[\frac{\pi}{m+1}]}{1+\frac{\epsilon}{m}}. \tag{11}$$

V. Discussion and Conclusions

In this letter, we have presented, in analytic form, an upper limit for the stability radius of the LP model as a function of white-noise correction level. Specifically, we showed that the maximum radius r_μ of the roots of the LP model is a decreasing function of the white-noise correction level μ, with $r_\mu < 1$ for all $\mu > 0$.

While we have discussed theory of white-noise correction in detail, most speech coders apply other modifications to the autocorrelation sequence as well. These modifications enhance, in effect, the condition number, and we should expect that the stability radius will decrease correspondingly. Detailed analysis of these modifications is left for further study.

Even though white-noise correction is applied to alleviate problems caused by numerical round-off errors, our analysis in this letter has been purely analytical. Practical applications of our results should therefore take into account issues of numerical accuracy as well.

The presented results are mainly of theoretical interest but also useful in determining the required accuracy of the root-finding process of the LSP representation [9]. The distance of adjacent LSFs is relative to the stability radius, and an upper limit for the latter gives us a lower limit on the former [10]. Specifically, for stability radius $r_{\max}$, the minimum separation of LSFs is

$$|\Delta\omega_n| \geq \frac{\pi(1-r_{\max})}{m+p+(m-p)r_{\max}} \tag{12}$$

where $p = 1$ for LSP, and $p = 0$ for immittance spectrum pair (ISP) [9], [14]. Using (6) and (12), we have calculated the minimum separation of LSFs for some common parameter combinations in Table I.

TABLE I
TYPICAL COMBINATIONS OF MODEL ORDER m, SAMPLING FREQUENCY F_s, FRAME LENGTH l (WHERE $N = F_s l/1000$), AND WHITE-NOISE CORRECTION LEVEL μ AND THEIR CORRESPONDING STABILITY RADIUS $r_{\max}$ AND MINIMUM LSF SEPARATION $\Delta\omega$ FROM (10) AND (12)

m	F_s	l	μ	$1 - r_{\max}$	$\min \Delta\omega$
10	8 kHz	20 ms	0.0001	1.7×10^{-4}	2.6×10^{-5}
10	8 kHz	30 ms	0.0001	7.9×10^{-5}	1.2×10^{-5}
16	12.8 kHz	20 ms	0.0001	6.6×10^{-5}	6.5×10^{-6}

ACKNOWLEDGMENT

The mathematical analysis in this letter was inspired by and a result of detailed analysis of the following publications: [2], [11], and [12, Ch. 5].

REFERENCES

[1] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. PROC-63, no. 5, pp. 561–580, Apr. 1975.

[2] P. Delsarte, Y. Genin, and Y. Kamp, "Stability of linear predictors and numerical range of a linear operator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 3, pp. 470–478, May 1986.

[3] P. Kabal, "Ill-conditioning and bandwidth expansion in linear prediction of speech," in *Proc. Int. Conf. IEEE Acoustics, Speech, Signal Processing3*, Apr. 6-10, 2003, vol. 1, pp. 824–827.

[4] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 3, pp. 247–254, Jun. 1979.

[5] *Recommendation G.729-Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)*, ITU-T, Mar. 1996.

[6] *Digital Cellular Telecommunications System (Phase 2+); Adaptive Multi-Rate (AMR) Speech Transcoding, (GSM 06.90 version 7.2.1)*, ETSI, Apr. 2000.

[7] *3rd Generation Partnership Project: AMR Wideband Speech Codec*, 3GPP TS 26.190 V5.0.0. 3GPP, Mar. 2001.

[8] J.-M. Valin, "Speex: A free codec for free speech," in *Proc. Linux Conf. Australia*, 2006 [Online]. Available: http://people.xiph.org/jm/papers/speex_lca2006.pdf.

[9] F. K. Soong and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, San Diego, CA, Mar. 1984, vol. 1, pp. 1.10.1–1.10.4.

[10] T. Bäckström, C. Magi, and P. Alku, "Minimum separation of line spectral frequencies," *IEEE Signal Process. Lett.*, vol. 14, no. 2, pp. 145–147, Feb. 2007.

[11] U. Haagerup and P. D. L. Harpe, "The numerical radius of a nilpotent operator on a Hilbert space," *Proc. Amer. Math. Soc.*, vol. 115, no. 2, pp. 371–379, Jun. 1992.

[12] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*. New York: Wiley, 1996.

[13] G. Golub and C. van Loan, *Matrix Computations*. Baltimore, MD: The Johns Hopkins Univ. Press, 1996.

[14] Y. Bistritz and S. Peller, "Imittance spectral pairs (ISP) for speech encoding,," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1993, vol. 2, pp. 9–12.

Study VI

Paavo Alku, Carlo Magi, Santeri Yrttiaho, Tom Bäckström, Brad Story. "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering", Journal of the Acoustical Society of America, Vol. 125, No. 5, pp. 3289-3305, 2009.

Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering

Paavo Alku[a] and Carlo Magi[b]
Department of Signal Processing and Acoustics, Helsinki University of Technology, P.O. Box 3000, Fi-02015 TKK, Finland

Santeri Yrttiaho
Department of Signal Processing and Acoustics and Department of Biomedical Engineering and Computational Science, Helsinki University of Technology, P.O. Box 3000, Fi-02015 TKK, Finland

Tom Bäckström
Department of Signal Processing and Acoustics, Helsinki University of Technology, P.O. Box 3000, Fi-02015 TKK, Finland

Brad Story
Speech Acoustics Laboratory, University of Arizona, Tuscon, Arizona 85721

(Received 23 November 2007; revised 5 February 2009; accepted 8 February 2009)

Closed phase (CP) covariance analysis is a widely used glottal inverse filtering method based on the estimation of the vocal tract during the glottal CP. Since the length of the CP is typically short, the vocal tract computation with linear prediction (LP) is vulnerable to the covariance frame position. The present study proposes modification of the CP algorithm based on two issues. First, and most importantly, the computation of the vocal tract model is changed from the one used in the conventional LP into a form where a constraint is imposed on the dc gain of the inverse filter in the filter optimization. With this constraint, LP analysis is more prone to give vocal tract models that are justified by the source-filter theory; that is, they show complex conjugate roots in the formant regions rather than unrealistic resonances at low frequencies. Second, the new CP method utilizes a minimum phase inverse filter. The method was evaluated using synthetic vowels produced by physical modeling and natural speech. The results show that the algorithm improves the performance of the CP-type inverse filtering and its robustness with respect to the covariance frame position. © *2009 Acoustical Society of America.* [DOI: 10.1121/1.3095801]

PACS number(s): 43.70.Gr, 43.70.Jt [CHS] Pages: 3289–3305

I. INTRODUCTION

All areas of speech science and technology rely, in one form or another, on understanding how speech is produced by the human voice production system. In the area of voice production research, glottal inverse filtering (IF) refers to methodologies that aim to estimate the source of voiced speech, the glottal volume velocity waveform. The basis for these techniques is provided by the classical source-filter theory, according to which the production of a voiced speech signal can be interpreted as a cascade of three separate processes: the excitation, that is, the glottal volume velocity waveform, the vocal tract filter, and the lip radiation effect (Fant, 1970). In order to compute the first of these processes, IF methodologies estimate the second and third processes typically in forms of linear, time-invariant digital systems and then cancel their contribution from the speech signal by filtering it through the inverse models of the vocal tract and lip radiation effect. Since the lip radiation effect can be estimated at low frequencies as a time-derivative of the flow (Flanagan, 1972), which is easily modeled digitally by a fixed first order finite impulse response (FIR) filter, the key problem in IF methods is the estimation of the vocal tract.

Among the main methodologies used to analyze human voice production, IF belongs to the category of acoustical methods. As alternatives to the acoustical methods, it is possible to investigate voice production with visual inspection of the vocal fold vibrations or with electrical (e.g., Lecluse *et al.*, 1975) or electromagnetic methods (Titze *et al.*, 2000). Visual analysis of the vibrating vocal folds is widely used especially in clinical investigation of voice production. Several techniques, such as video stroboscopy (e.g., Hirano, 1981), digital high-speed stroboscopy (e.g., Eysholdt *et al.*, 1996), and kymography (Švec and Schutte, 1996), have been developed, and many of them are currently used in daily practices in voice clinics. Acquiring visual information about voice production, however, always calls for invasive measurements in which the vocal folds are examined either with a solid endoscope inserted in the mouth or with a flexible fiberscope inserted in the nasal cavity. In contrast to these techniques, a benefit of glottal IF is that the analysis can be computed from the acoustic signal in a truly non-invasive manner. This feature is essential especially in such research areas in which vocal function needs to be investigated under as natural circumstances as possible, for instance, in under-

[a]Electronic mail: paavo.alku@tkk.fi
[b]Deceased in February 2008.

standing the role of the glottal source in the expression of vocal emotions (Cummings and Clements, 1995; Gobl and Ní Chasaide, 2003; Airas and Alku, 2006) or in studying occupational voice production (Vilkman, 2004; Lehto *et al.*, 2008). In addition to its non-invasive nature, glottal IF provides other favorable features. IF results in a temporal signal, the glottal volume velocity waveform, which is an estimate of a real acoustical waveform of the human voice production process. Due to its direct relationship to the acoustical production of speech, estimates of glottal excitations computed by IF can be modeled with their artificial counterparts to synthesize human voice in speech technology applications (Klatt and Klatt, 1990; Carlson *et al.*, 1991; Childers and Hu, 1994).

Since the introduction of the idea of IF by Miller (1959), many different IF methods have been developed. The methods can be categorized, for example, based on the input signal, which can be either the speech pressure waveform recorded in the free field outside the lips (e.g., Wong *et al.*, 1979; Alku, 1992) or the oral volume velocity captured by a specially designed pneumotachograph mask, also known as the Rothenberg mask (e.g., Rothenberg, 1973; Hertegård *et al.*, 1992). In addition, methods developed to do IF differ depending on whether they need user adjustments in defining the settings of the vocal tract resonances (e.g., Price, 1989; Sundberg *et al.*, 2005) or whether the analysis is completely automatic (e.g., Veeneman and BeMent, 1985). From the methodological point of view, the techniques developed can be categorized based on how the effect of the glottal source is taken into account in the estimation of the vocal tract in the underlying IF method. From this perspective, there are, firstly, methods (e.g., Alku, 1992) that are based on the gross estimation of the glottal contribution during both the closed and open phase of the glottal pulse using all-pole modeling. By canceling the glottal contribution from the speech signal, a model for the vocal tract is computed with linear prediction (LP) (Rabiner and Schafer, 1978) although other spectral envelope fitting techniques such as those based on the penalized likelihood approach (Campedel-Oudot *et al.*, 2001) or cepstrum analysis (Shiga and King, 2004) could, in principle, be used as well. Secondly, the use of a joint optimization of the glottal flow and vocal tract is possible based on synthetic, pre-defined models of the glottal flow (e.g., Milenkovic, 1986; Kasuya *et al.*, 1999; Fröhlich *et al.*, 2001; Fu and Murphy, 2006). Thirdly, it is possible to estimate the glottal flow using closed phase (CP) covariance analysis (Strube, 1974; Wong *et al.*, 1979). This is based on the assumption that there is no contribution from the glottal source to the vocal tract during the CP of the vocal fold vibration cycle. After identification of the CP, covariance analysis is used to compute a parametric all-pole model of the vocal tract using LP.

CP covariance analysis is among the most widely used glottal IF techniques. Since the original presentation of the method by Strube (1974), the CP method has been used as a means to estimate the glottal flow, for instance, in the analysis of the phonation type (Childers and Ahn, 1995), prosodic features of connected speech (Strik and Boves, 1992), vocal emotions (Cummings and Clements, 1995), source-tract interaction (Childers and Wong, 1994), singing (Arroabarren and Carlosena, 2004), and speaker identification (Plumpe *et al.*, 1999). In addition to these various applications, CP analysis has been a target of methodological development. The major focus of this methodological work has been the method of accurately determining the location of the covariance frame, the extraction of the CP of the glottal cycle. In order to determine this important time span from a speech waveform, an approach based on a series of sliding covariance analyses is typically used. In other words, the analysis frame is sequentially moved one sample at a time through the speech signal and the results of each covariance analysis are analyzed in order to determine the CP. Strube (1974) used this approach and identified the glottal closure as an instant when the frame was in a position which yielded the maximum determinant of the covariance matrix. Wong *et al.* (1979) instead defined the CP as the interval when the normalized squared prediction error was minimum, and this technique has been used by several authors since, although sometimes with slight modifications (e.g., Cummings and Clements, 1995). Plumpe *et al.* (1999), however, argued that the use of the prediction error energy in defining the frame position of the covariance analysis might be problematic for sounds which involve gradual closing or opening of the vocal folds. As a remedy, they proposed an idea in which sliding covariance analyses are computed and formant frequency modulations between the open and CP of the glottal cycle are used as a means to define the optimal frame position. Akande and Murphy (2005) suggested a new technique, adaptive estimation of the vocal tract transfer function. In their method, the estimation of the vocal tract is improved by first removing the influence of the glottal source by filtering the speech signal with a dynamic, multi-pole high-pass filter instead of the traditional single-pole pre-emphasis. The covariance analysis is then computed in an adaptive loop where the optimal filter order and frame position are searched for by using phase information of the filter candidates.

All the different CP methods referred to above are based on the identification of the glottal CP from a single source of information provided by the speech pressure waveform. Therefore, they typically involve an epoch detection block in which instants of glottal closure and opening are extracted based on algorithms such as DYPSA (Naylor *et al.*, 2007). Alternatively, if electroglottography (EGG) is available, it is possible to use two information channels so that the position and duration of the CP is estimated from EGG, and then the speech waveform is inverse filtered. This so-called two-channel analysis has been shown to yield reliable results in IF due to improved positioning of the covariance frame (Veeneman and BeMent, 1985; Krishnamurthy and Childers, 1986). In this technique, the CP analysis is typically computed by estimating the CP of the glottal cycle as the time interval between the minimum and maximum peaks of the first time-derivative of the EGG waveform (Childers and Ahn, 1995). It is important to notice that even though there have been many modifications to CP analysis since the work by Strube (1974), all the methods developed are based on the same principle in the mathematical modeling of the vocal

tract, namely, the use of conventional LP with the covariance criterion described in Rabiner and Schafer (1978).

Even though different variants of CP covariance analysis have been shown to yield successful estimates of the glottal flow by using simple synthesized vowels, this IF methodology has certain shortcomings. Several previous studies have in particular indicated that glottal flow estimates computed by the CP analysis vary greatly depending on the position of the covariance frame (e.g., Larar *et al.*, 1985; Veeneman and BeMent, 1985; Yegnanarayana and Veldhuis, 1998; Riegelsberger and Krishnamurthy, 1993). Given the fundamental assumption of the method, that is the computation of the vocal tract model during an excitation-free time span, this undesirable feature of the CP analysis is understandable. The true length of the glottal CP is typically short, which implies that the amount of data used to define the parametric model of the vocal tract with the covariance analysis is sparse. If the position of this kind of a short data frame is misaligned, the resulting linear predictive filter typically fails to model the vocal tract resonances, which might result in severe distortion of the glottal flow estimates. This problem is particularly severe in voices of high fundamental frequency (F0) because they are produced by using very short lengths in the glottal CP. In order to cope with this problem, previous CP methods typically exploit techniques to improve the extraction of the covariance frame position. In the present work, however, a different approach is suggested based on setting a mathematical constraint in the computation of the inverse model of the vocal tract with LP. The constraint imposes a predefined value for the direct current (dc) gain of the inverse filter as a part of the optimization of the filter coefficients. This results in vocal tract filters whose transfer functions, in comparison to those defined by the conventional covariance analysis, are less prone to include poles in positions in the z-domain that are difficult to interpret from the point of view of the classical source-filter theory of vowel production (e.g., on the positive real axis). This new dc-constrained vocal tract model is combined in the present study with an additional procedure, checking of the minimum phase property of the inverse filter, to yield a new CP algorithm.

In the following, typical artifacts caused by the CP analysis are first described using representative examples computed from natural vowels. These examples are then used to motivate the proposed new method to compute LP in vocal tract modeling of the CP analysis. The new method is then tested with both synthetic vowels produced by physical modeling of the human voice production mechanism and with natural speech of both female and male subjects.

II. METHODS

A. Sources of distortion in the conventional CP analysis

In this section, two major sources of error in the conventional CP analysis are described with the help of examples. The word "conventional" refers here to the CP analysis in which the vocal tract is modeled with a pth order all-pole filter computed by the basic form of the covariance analysis described by Rabiner and Schafer (1978), and the lip radiation effect is modeled with a fixed first order FIR filter. All the analyses described were computed using the sampling frequency of 8 kHz and the order of the vocal tract filter set to $p=12$. The length of the covariance frame was 30 samples (3.75 ms). The instant of glottal closure was extracted, when needed, as the instant of the negative peak of the EGG derivative.

First, the sensitivity of the glottal flow estimate about the position of the covariance frame is demonstrated. Figure 1 shows three glottal flow estimates, which were inverse-filtered from the same token of a male subject uttering the vowel [a] by using a minor change in the position of the covariance frame position: the beginning of the covariance frame in Figs. 1(b) and 1(c) was moved earlier in the signal by two and four samples, respectively, in comparison to the beginning of the covariance frame used in Fig. 1(a). The inverse filters obtained are shown in the z-domain in the left panels of Fig. 2, and the amplitude spectra of the corresponding vocal tract filters are depicted in the right panels of the same figure. The example indicates how a minor change in the position of the covariance frame has resulted in a substantial change in the estimated glottal flows. It is worth noticing that the covariance analyses illustrated in Figs. 2(a) and 2(b) have resulted in two inverse filters both of which have one root on the positive real axis in the z-domain. In Fig. 2(b), the position of this root is slightly closer to the unit circle than in Fig. 2(a). The CP analysis shown in Fig. 2(c) has, in turn, resulted in an inverse filter with a complex conjugate pair of roots at low frequencies. The effect of an inverse filter root which is located on the positive real axis approaches that of a first order differentiator [i.e., $H(z)=1-z^{-1}$] when the root approaches the unit circle, and a similar effect is also produced by a complex conjugate pair of roots at low frequencies. Consequently, the resulting glottal flow estimate, as shown in Figs. 1(b) and 1(c), becomes similar to a time-derivative of the flow candidate given by an inverse filter with no such roots or when these roots are located in a more neutral position close to the origin of the z-plane. This severe distortion of the glottal flow estimate caused by the occurrence of inverse filter roots, both real and complex conjugate pairs, at low frequencies is greatest at time instants when the flow changes most rapidly, that is, near glottal closure. As shown in Figs. 1(b) and 1(c), this distortion[1] is typically seen as sharp negative peaks, called "jags" by Wong *et al.* (1979), of the glottal flow pulses at the instants of closure.

The undesirable distortion of the glottal flow estimates by the occurrence of jags implies that the corresponding all-pole vocal tract model has roots on the positive real axis or at low frequencies, and, consequently, its amplitude spectrum shows boosting of low frequencies. This effect is clearly shown in the example by comparing the right panel of Fig. 2(a) to the corresponding panels in Figs. 2(b) and 2(c). It is worth emphasizing that the source-filter theory of voice production by Fant (1970) assumes that poles of the vocal tract for non-nasalized voiced sounds occur as complex conjugate pairs and the low-frequency emphasis of the vowel spectrum results from the glottal source. Therefore, it can be argued

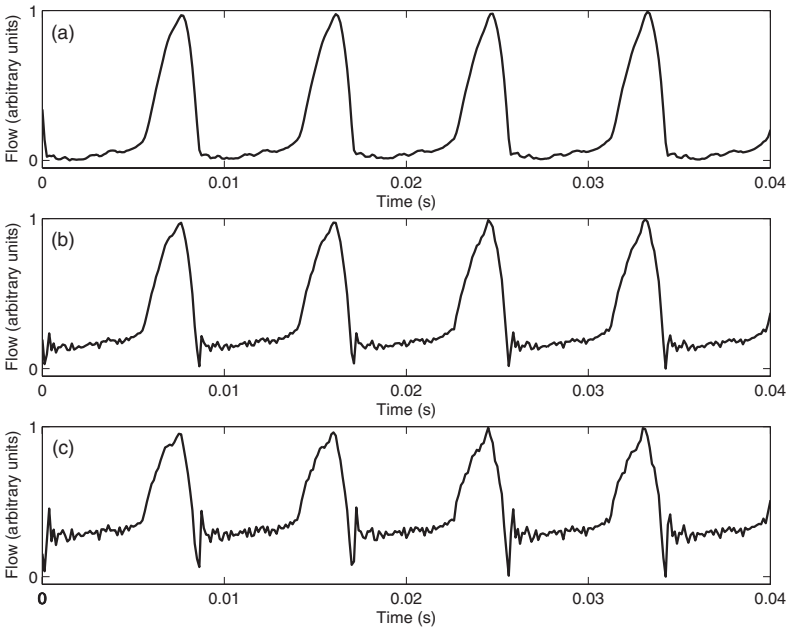


FIG. 1. Glottal flows estimated by IF the vowel [a] uttered by a male speaker by varying the position of the covariance frame in the CP analysis. The covariance frame was placed in the beginning of the CP using the differentiated EGG in panel (a), and its position was moved earlier by two samples in panel (b) and by four samples in panel (c).

that among the three vocal tract models computed by the CP analysis, the one depicted in Fig. 2(a) is the most plausible to represent an amplitude spectrum of an all-pole vocal tract of a vowel sound.

Quality of glottal flows computed by the CP analysis can be made less dependent on the position of the covariance frame by removing the roots of the vocal tract model located on the real axis (Wong *et al.*, 1979; Childers and Ahn, 1995). This is typically done by first solving the roots of the vocal tract model given by LP and then by removing those roots that are located on the positive real axis while preserving the roots on the negative real axis. This procedure was used for the example described in Figs. 1 and 2, and the results are shown in the time domain in Fig. 3 and in the frequency domain in Fig. 4. It can be seen that this standard procedure indeed decreased the distortion caused by the jags, as shown in Fig. 3(b). It is, however, worth noticing that this procedure is blind to complex roots located at low frequencies, which

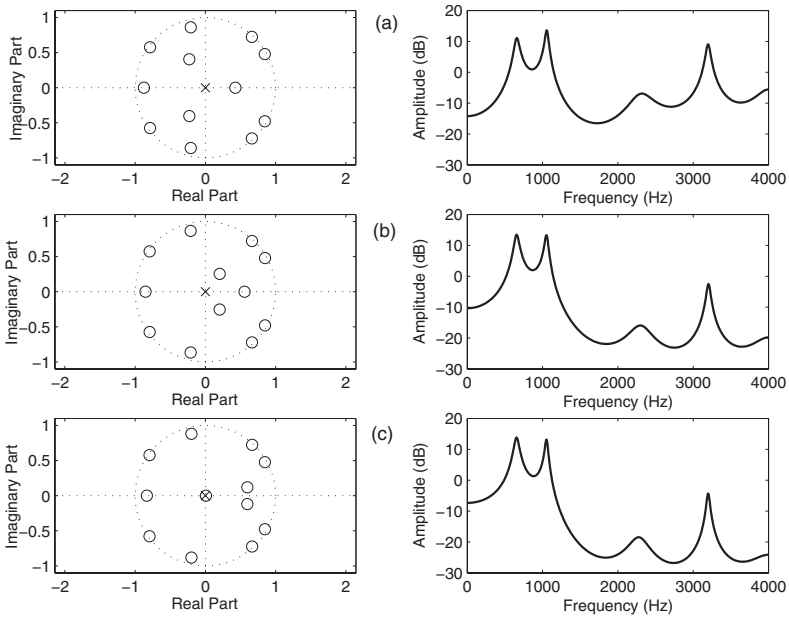


FIG. 2. Transfer functions of inverse filters in the z-domain (left panels) and the corresponding amplitude spectra of the all-pole vocal tract models (right panels) used in the CP analyses shown in Fig. 1.

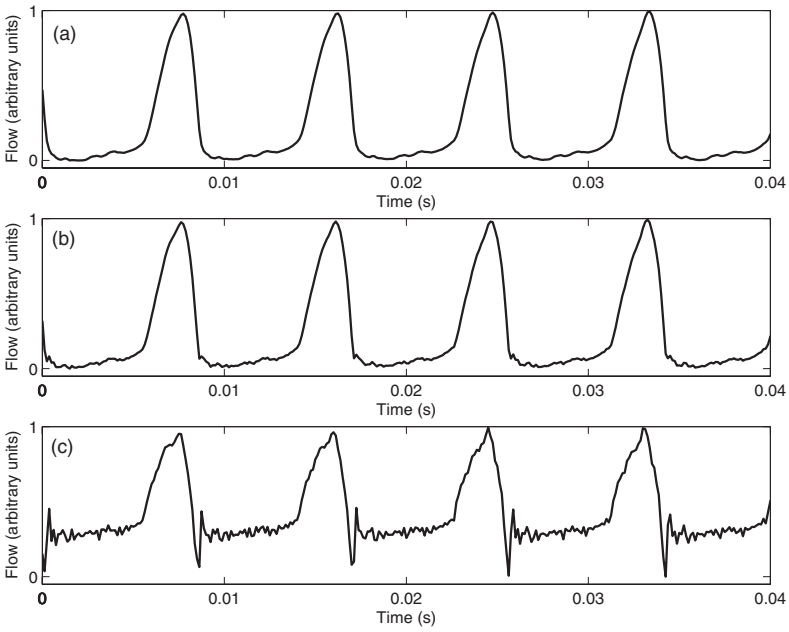


FIG. 3. Glottal flows estimated by IF the same [a] vowel used in Fig. 1. Roots located on the positive real axis were removed before IF. The covariance frame was placed in the beginning of the CP with the help of the differentiated EGG in panel (a), and its position was moved earlier by two samples in panel (b) and by four samples in panel (c).

cause distortion, described in Figs. 1(c) and 3(c), that might be even more severe than that resulting from the roots on the positive real axis.

In addition to the distortion caused by the occurrence of inverse filter real and complex roots at low frequencies as described above, the estimation of the glottal flow with the CP analysis might be affected by another issue. Namely, the computation of the linear predictive analysis with the covariance analysis might yield an inverse filter that is not minimum phase; that is, the filter has roots outside the unit circle in the z-domain. Although this property of the covariance analysis is well-known in the theory of LP (Rabiner and Schafer, 1978), it is, unfortunately, typically ignored in most glottal IF studies (exceptions are Akande and Murphy, 2005; Bozkurt *et al.*, 2005; Bäckström and Alku, 2006). A possible explanation of why the occurrence of non-minimum phase filters gets so little attention in glottal wave analysis is the fact that IF is always computed via FIR filtering. Hence,

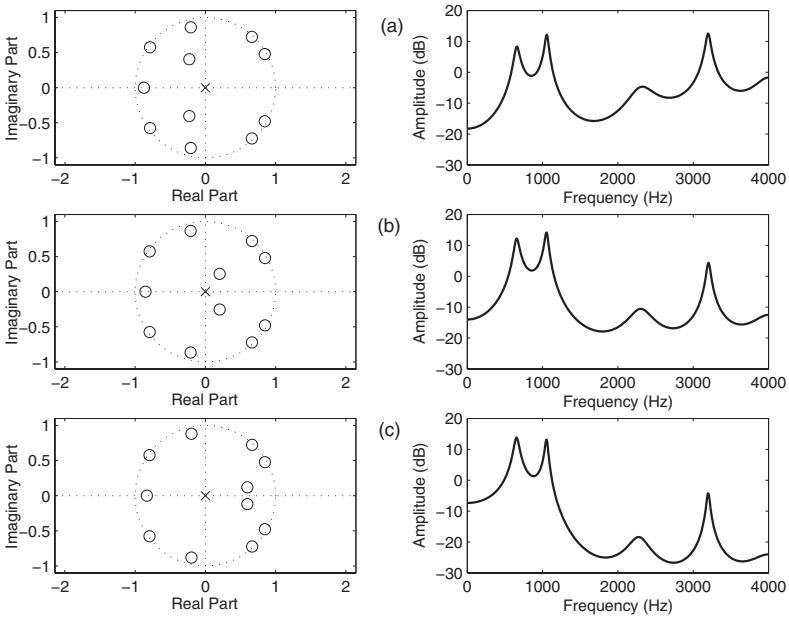


FIG. 4. Transfer functions of inverse filters in the z-domain (left panels) and the corresponding amplitude spectra of the all-pole vocal tract models (right panels) used in the CP analyses shown in Fig. 3.

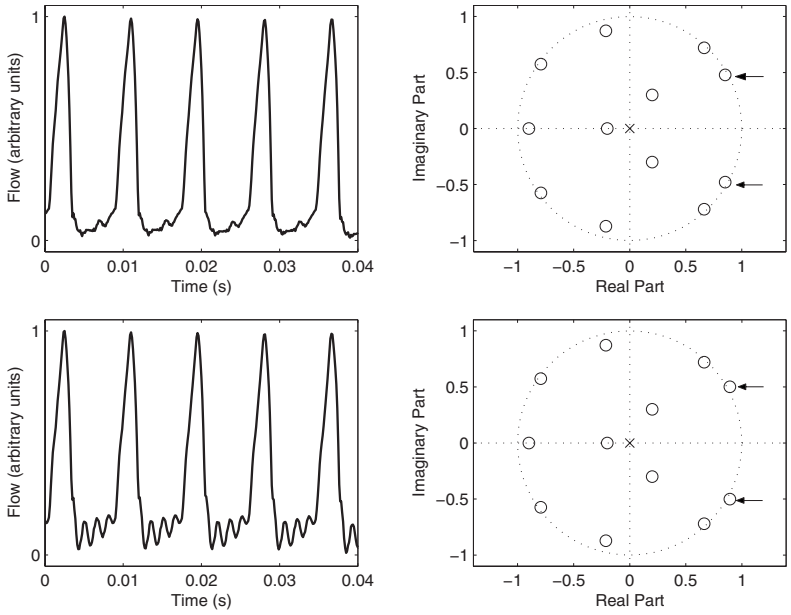


FIG. 5. Glottal flows estimated by the CP analysis (left panels) and inverse filter transfer functions in the z-domain (right panels) in the case of (a) minimum phase and (b) non-minimum phase IF. Radii of all roots in minimum phase filtering are less than unity. In non-minimum phase filtering, the complex conjugate root pair indicated by arrows in panel (a) is replaced by its mirror image pair outside the unit circle. The root radius of the indicated complex conjugate pair is 0.98 in panel (a) and 1.02 in panel (b).

non-minimum phase filters do not cause stability problems, which, of course, would be the case if non-minimum phase filters were used in all-pole synthesis, such as in speech coding or synthesis. Even though stability problems are not met in glottal in IF, the use of non-minimum phase inverse filters does cause other kinds of artifacts, as demonstrated below.

According to the source-filter theory of speech production, the glottal flow is filtered by a physiological filter, the vocal tract, which is regarded as a stable all-pole system for vowels and liquids. In the z-domain, this kind of system must have all its poles inside the unit circle (Oppenheim and Schafer, 1989). An optimal inverse filter cancels the effects of the vocal tract by matching each pole inside the unit circle with a zero of a FIR filter. However, it is well-known in the theory of digital signal processing that zeros of a FIR filter can be replaced by their mirror image partners; that is, a zero at $z=z_1$ is replaced by $z=1/z_1^*$, without changing the shape of the amplitude spectrum of the filter (Oppenheim and Schafer, 1989). In other words, an inverse filter that is minimum phase can be replaced with a non-minimum phase FIR by replacing any of its roots with a corresponding mirror image outside the unit circle without changing the shape of inverse filter's amplitude response. Therefore, from the point of view of canceling the amplitude response of the all-pole vocal tract, there are several inverse filters, of which one is minimum phase and others are non-minimum phase, that can be considered equal. These candidates are, however, different in terms of their phase characteristics, and canceling the effects of an all-pole vocal tract with a non-minimum phase inverse filter produces phase distortion, which might severely affect the shape of the glottal flow estimate. This distortion is especially strong in cases where zeros in the inverse filter located in the vicinity of the lowest two formants are moved from inside the unit circle to the outside. Figure 5 shows an example of this effect. In Fig. 5(a), a glottal flow estimated with a minimum phase inverse filter is shown in the left panel, and the z-plane representation of the corresponding inverse filter is shown on the right. This inverse filter was deliberately modified by replacing one complex conjugate root pair located inside the unit circle by its corresponding mirror image pair located outside the circle. The root pair selected corresponds to the inverse model of the first formant and is represented in the z-plane graph of Fig. 5(a) by the complex conjugate pair having the lowest angular frequency (indicated by arrows). Even though the modification caused only a minor change in the root radius (original radius: 0.98, modified radius: 1.02), the change from the minimum phase structure into the non-minimum phase form is manifested as increased ripple during the CP of the glottal cycle, as shown in the left panel of Fig. 5(b).

B. The improved CP analysis

A new approach is proposed in the present study to compute IF with the CP analysis. The proposed technique aims to reduce the effects of the two major artifacts, occurrence of low-frequency roots of the inverse filter and occurrence of inverse filter roots outside the unit circle, described in the previous section. The main part of the method, to be described in Sec. II B 1, is represented by a new mathematical algorithm to define a linear predictive inverse filter. The novel way to compute vocal tract inverse filters is then combined, as described in Sec. II B 2, with an additional processing stage to yield the new glottal IF algorithm.

1. Computation of the vocal tract inverse filter with constrained linear prediction

The conventional CP analysis involves modeling the vocal tract with an all-pole filter defined according to the clas-

sical LP based on the covariance criterion (Rabiner and Schafer, 1978). The filter coefficients of pth order inverse filter are searched for by using a straightforward optimization where the energy of the prediction error is minimized over the covariance frame. In principle, this kind of optimization based on the mean square error (MSE) criterion treats all the frequencies equally, and the filter coefficients are mathematically adjusted so that the resulting all-pole spectrum accurately matches the high-energy formant regions of the speech spectrum (Makhoul, 1975). However, it is worth emphasizing that the conventional covariance analysis does not use any additional information in the optimization process, for example, to bias the location of roots of the resulting all-pole filter. This inherent feature of the conventional covariance analysis implies that roots of the resulting all-pole model of the vocal tract might be located in such a position in the z-domain (e.g., on the positive real axis) that is correct from the point of view of MSE-based optimization but unrealistic from the point of view of the source-filter theory of vowel production and its underlying theory of tube modeling of the vocal tract acoustics. In his fundamental work, Fant (1970) related vocal tract shapes derived from x-rays to the acoustic theory of different tube shapes and developed the source-filter theory of speech production. According to this theory, the transfer function of voiced speech, defined as the ratio of the Laplace transforms of the sound pressure at the lips to the glottal volume velocity, includes only complex poles in the s-domain. According to the discrete time version of this theory (e.g., Markel and Gray, 1976), the z-domain transfer function of the vocal tract is expressed for vowel sounds as an all-pole filter of order $2K$, which models K formants as a cascade of K second order blocks, each representing an individual resonance of a certain center frequency and bandwidth. In other words, there might be a mismatch in root locations of vocal tract filters between those optimized by the conventional covariance analysis and those assumed both in the source-filter theory and its underlying acoustical theory of tube shapes. It is likely that this mismatch becomes prevalent especially in cases when the covariance frame consists of a small number of data samples. Hence, the phenomenon discussed is related to the sensitivity of the CP analysis about the position of the covariance frame, a drawback discussed in several previous studies (e.g., Larar *et al.*, 1985; Veeneman and BeMent, 1985; Yegnanarayana and Veldhuis, 1998; Riegelsberger and Krishnamurthy, 1993).

Based on the concept of *constrained* LP, the computation of the conventional covariance analysis, however, can be modified in order to reduce the distortion that originates from such vocal tract model roots that are located in unrealistic positions in the z-domain. The key idea is to impose such restrictions on the linear predictive polynomials *prior* to the optimization that can be justified by the source-filter theory of voice production. Intuitively, this means that instead of allowing the linear predictive model to locate its roots freely in the z-domain based solely on the MSE criterion, the optimization is given certain restrictions in the predictor structure, which then result in more realistic root locations. In order to implement restrictions that end up in equations which can be solved in closed form, one has to first find a method to express the constraint in a form of a concise mathematical equation and then use the selected equation in the minimization problem. One such convenient constraint can be expressed with the help of the dc gain of the linear predictive inverse filter. The rationales to apply this quantity are as follows. First, the dc gain of a digital FIR filter can be expressed in a very compressed and mathematically straightforward manner as a linear sum of the predictor coefficients [see Eq. (4) below]. Consequently, the optimization of the constrained linear predictive filter is mathematically straightforward, ending up with a matrix equation [see Eq. (9)] that can be solved noniteratively in a similar manner as the corresponding normal equations of the conventional LP. Second, it is known from the classical source-filter theory of voice production that the vocal tract transfer function of nonnasalized sounds approaches unity at zero frequency provided that the losses through vibration of the cavity walls are small (Fant, 1970, pp. 42–44). In conventional LP, the dc gain of the inverse filter is not constrained, and, consequently, it is possible that the amplitude response of the vocal tract model computed by the covariance analysis shows excessive boost at zero frequency. If the covariance frame is short and placed incorrectly, it might even happen that the amplitude response of the obtained vocal tract model shows larger gain at zero frequency than at formants, which violates the assumptions of the source-filter theory and its underlying acoustical theory of tube shapes. Hence, by imposing a predefined constraint on the dc gain of the linear predictive inverse filter, one might expect to get such linear predictive vocal tract models whose amplitude response shows better correspondence with Fant's source-filter theory; that is, the transfer function indicates peaks at formant frequencies, while the gain at zero frequency is clearly smaller and approaches unity. It must be emphasized, however, that even though the proposed idea to assign the dc gain of the inverse filter into a pre-defined value is undoubtedly mathematically straightforward, this technique does not involve imposing explicit constraints on the root positions *per se* prior to the optimization. In other words, the exact z-domain root locations of the vocal tract model are still determined by the MSE-type optimization, yet the likelihood for these roots to become located in such positions that they create an excessive boost at low frequency is less than in the case of the conventional LP. Mathematical derivations to optimize the proposed idea of the dc-constrained LP will be described below.

In the conventional LP, the error signal, known as the residual, can be expressed in matrix form as follows:

$$e_n = x_n + \sum_{k=1}^{p} a_k x_{n-k} = \sum_{k=0}^{p} a_k x_{n-k} = \mathbf{a}^T \mathbf{x}_n, \tag{1}$$

where $\mathbf{a}=[a_0,\ldots,a_p]^T$, with $a_0=1$, and the signal vector is $\mathbf{x}_n=[x_n \ldots x_{n-p}]^T$. The coefficient vector $\mathbf{a}$ is optimized according to the MSE criterion by searching for such parameters that minimize the square of the residual. In the covariance method, this minimization of the residual energy is computed over a finite time span (Rabiner and Schafer,

1978). By denoting this time span with $0 \leqslant n \leqslant N-1$, the prediction error energy $E(\mathbf{a})$ can be written as

$$E(\mathbf{a}) = \sum_{n=0}^{N-1} e_n^2 = \sum_{n=0}^{N-1} \mathbf{a}^T \mathbf{x}_n \mathbf{x}_n^T \mathbf{a} = \mathbf{a}^T \left[\sum_{n=0}^{N-1} \mathbf{x}_n \mathbf{x}_n^T \right] \mathbf{a} = \mathbf{a}^T \mathbf{\Phi} \mathbf{a}, \tag{2}$$

where matrix $\mathbf{\Phi}$ is the covariance matrix defined from speech samples as

$$\mathbf{\Phi} = \sum_{n=0}^{N-1} \mathbf{x}_n \mathbf{x}_n^T \in R^{(p+1)\times(p+1)}. \tag{3}$$

It is worth noticing that the computation of matrix $\mathbf{\Phi}$ requires speech samples located inside the energy minimization frame, that is, x_n, where $0 \leqslant n \leqslant N-1$, plus p samples occurring before this frame, that is, x_n, where $-p \leqslant n < 0$. The optimal filter coefficients can be computed easily by minimizing the prediction error energy $E(\mathbf{a})$ with respect to the coefficient vector $\mathbf{a}$. This yields $\mathbf{a} = \sigma^2 \mathbf{\Phi}^{-1} \mathbf{u}$, where $\sigma^2 = (\mathbf{u}^T \mathbf{\Phi}^{-1} \mathbf{u})^{-1}$ is the residual energy given by the optimized predictor and $\mathbf{u} = [1 0 \cdots 0]^T$.

The conventional LP can be modified by imposing constraints on the minimization problem presented above. A mathematically straightforward way to define one such constraint is to set a certain pre-defined value for the frequency response of the linear predictive inverse filter at zero frequency. By denoting the transfer function of a pth order constrained inverse filter $C(z)$, the following equation can be written:

$$C(z) = \sum_{k=0}^{p} c_k z^{-k} \Rightarrow C(e^{j0}) = C(1) = \sum_{k=0}^{p} c_k = l_{\mathrm{dc}}, \tag{4}$$

where c_k, $0 \leqslant k \leqslant p$, are the filter coefficients of the constrained inverse filter and l_{dc} is a pre-defined real value for the gain of the filter at dc. Using matrix notation, the dc-constrained minimization problem can now be formulated as follows: minimize $\mathbf{c}^T \mathbf{\Phi} \mathbf{c}$ subject to $\mathbf{\Gamma}^T \mathbf{c} = \mathbf{b}$, where $\mathbf{c} = [c_0 \cdots c_p]^T$ is the filter coefficient vector with $c_0 = 1$, $\mathbf{b} = [1 l_{\mathrm{dc}}]^T$, and $\mathbf{\Gamma}$ is a $(p+1) \times 2$ constraint matrix defined as

$$\Gamma = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 0 & 1 \end{bmatrix}. \tag{5}$$

The covariance matrix defined in Eq. (3) is positive definite. Therefore, the quadratic function to be minimized in the dc-constrained problem is convex. Thus, in order to solve the minimization problem, the Lagrange multiplier method (Bazaraa *et al.*, 1993) can be used. This procedure begins with the definition of a new objective function,

$$\eta(\mathbf{c}, \mathbf{g}) = \mathbf{c}^T \mathbf{\Phi} \mathbf{c} - 2\mathbf{g}^T (\mathbf{\Gamma}^T \mathbf{c} - \mathbf{b}), \tag{6}$$

where $\mathbf{g} = [g_1 g_2]^T > \mathbf{0}$ is the Lagrange multiplier vector. The objective function of Eq. (6) can be minimized by setting its derivative with respect to vector $\mathbf{c}$ to zero. By taking into account that matrix $\mathbf{\Phi}$ is symmetric (i.e., $\mathbf{\Phi} = \mathbf{\Phi}^T$), this results in the following equation:

$$\begin{aligned} \nabla_c \eta(\mathbf{c}, \mathbf{g}) &= \mathbf{c}^T (\mathbf{\Phi}^T + \mathbf{\Phi}) - 2\mathbf{g}^T \mathbf{\Gamma}^T = 2\mathbf{c}^T \mathbf{\Phi} - 2\mathbf{g}^T \mathbf{\Gamma}^T \\ &= 2(\mathbf{\Phi}\mathbf{c} - \mathbf{\Gamma}\mathbf{g}) = 0. \end{aligned} \tag{7}$$

By combining Eq. (7) with the equation of the constraint (i.e., $\mathbf{\Gamma}^T \mathbf{c} - \mathbf{b} = 0$), vector $\mathbf{c}$ can be solved from the group of equations

$$\begin{aligned} \mathbf{\Phi}\mathbf{c} - \mathbf{\Gamma}\mathbf{g} &= 0, \\ \mathbf{\Gamma}^T \mathbf{c} - \mathbf{b} &= 0, \end{aligned} \tag{8}$$

which yields the optimal coefficients of the constrained inverse filter:

$$\mathbf{c} = \mathbf{\Phi}^{-1} \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{\Phi}^{-1} \mathbf{\Gamma})^{-1} \mathbf{b}. \tag{9}$$

In summary, the optimal dc-constrained inverse filter, a FIR filter of order p given in Eq. (4) is obtained by solving for the vector $\mathbf{c}$ according to Eq. (9), in which the covariance matrix $\mathbf{\Phi}$ is defined by Eq. (3) from the speech signal x_n, matrix $\mathbf{\Gamma}$ is defined by Eq. (5), and matrix $\mathbf{b} = [1 l_{\mathrm{dc}}]^T$, where l_{dc} is the desired inverse filter gain at dc.

2. Checking the minimum phase property

In order to eliminate the occurrence of non-minimum phase filters, the roots of the inverse filter are solved, and if the filter is not minimum phase, those roots that are located outside the unit circle are replaced by their mirror image partners inside the circle. In principle, it is possible that the constrained LP computed according to Eq. (9) yields an inverse filter that has roots on the positive real axis. Due to the use of the dc constraint, the risk for this to happen is, however, clearly smaller than in the case of the conventional covariance analysis. Because the roots of $C(z)$ are solved for in order to eliminate the occurrence of non-minimum phase filters, it is trivial also to check simultaneously whether there are any roots on the positive real axis inside the unit circle. If so, these roots are simply removed, in a procedure similar to that used in the conventional CP analysis (Wong *et al.*, 1979).

3. Summary of the new algorithm

In summary, the new glottal IF algorithm can be presented by combining the procedures described in Secs. II B 1 and II B 2. The estimation of the glottal flow with this new CP-based IF algorithm consists of the following stages.

(1) Prior to the analysis, the speech pressure waveform is filtered through a linear-phase high-pass FIR with its cut-off frequency adjusted to 70 Hz. The purpose of this filter is to remove annoying low-frequency components picked up by the microphone during the recordings of the speech signals. The output of this stage, the high-pass filtered speech sound, is denoted by $S_{\mathrm{hp}}(n)$ below.
(2) The position of the covariance frame is computed using any of the previously developed methods based on, for

example, the maximum determinant of the covariance matrix (Wong *et al.*, 1979) or the EGG (Krishnamurthy and Childers, 1986).

(3) Vocal tract transfer function $C(z)$ is computed according to Eq. (9) by defining the elements of the covariance matrix in Eq. (3) from $S_{hp}(n)$ by using the covariance frame defined in stage (2).

(4) Roots of $C(z)$ defined in stage (3) are solved. Those roots of $C(z)$ that are located outside the unit circle are replaced by their corresponding mirror image partner inside the unit circle. Any real roots located on the positive real axis are removed.

(5) Finally, the estimate of the glottal volume velocity waveform is obtained by filtering $S_{hp}(n)$ through $C(z)$ defined in stage (4) and by canceling the lip radiation effect with a first order infinite impulse response filter, with its pole close to the unit circle (e.g., at $z=0.99$).

The algorithm runs in a frame-based manner, and the adjustable parameters are recommended to be set to values typically used in CP analysis: frame length: 50 ms; order of the vocal tract model: 12 (with sampling frequency of 8 kHz); the length of the covariance frame: 30 samples (a value that equals the order of the vocal tract model multiplied by 2.5). In the experiments conducted in the present study, the parameter l_{dc} used in the computation of the dc-constrained vocal tract inverse filters was adjusted so that the amplitude response of the vocal tract filter at dc was always equal to unity.[2]

III. MATERIALS AND EXPERIMENTS

In order to evaluate the performance of the new CP analysis technique, experiments were conducted using both natural and synthetic speech. The purpose of these experiments was to investigate whether the new modified covariance analysis based on the concept of constrained LP, when supplemented with the minimum phase requirement of the inverse filter, would make IF with the CP analysis less vulnerable to the position of the covariance frame.

A. Speech and EGG recordings

Simultaneous speech pressure waveform and EGG signals were recorded from 13 subjects (six females). The ages of the subjects varied between 29 and 43 (mean of 32), and none of them had experienced voice disorders. The speaking task was to produce the vowel [a] five times by using sustained phonation. Vowel [a] was used because it has a high first formant (F1).[3] Subjects were allowed to use the fundamental frequency of their own choice, but they were encouraged not to use a pitch that is noticeably higher than in their normal speech. The duration of each phonation was at least 1 s. The production was done by two types of phonation: normal and pressed. These two phonation types were selected because they are more likely to involve a CP in the vocal fold vibration, which would not be the case in, for example, breathy phonation (Alku and Vilkman, 1996). This, in turn, implies that the basic assumption of the CP analysis, that is, the existence of a distinct CP within the glottal cycle, should be valid. Consequently, using these two modes, one would expect to be able to demonstrate effectively the dependency of the CP analysis on the position of the covariance frame. The recordings were perceptually monitored by an experienced phonetician who trained the subjects to create the two registers properly. Phonations were repeated until the phonation type was satisfactory.

Speech pressure waves were captured by a condenser microphone (Brüel & Kjær 4188) that was attached to a sound level meter (Brüel & Kjær Mediator 2238) serving also as a microphone amplifier, and the EGG was recorded simultaneously (Glottal Enterprise MC2-1). The mouth-to-microphone distance was 40 cm. In order to avoid inconsistency in the synchronization of speech and EGG, the microphone distance was carefully monitored in the recordings, and its value was checked prior to each phonation. Speech and EGG waveforms were digitized using a (DAT) digital audio tape recorder (Sony DTC-690) by adopting the sampling rate of 48 kHz and the resolution of 16 bits.

The speech and EGG signals were digitally transferred from the DAT tape into a computer. Before conducting the IF analysis, the sampling frequency of both signals was downsampled to 8 kHz. The propagation delay of the acoustic signal from the glottis to the microphone was estimated by using the vocal tract length of 15 and 17 cm for females and males, respectively, the mouth-to-microphone distance of 40 cm, and the speed of sound value of 350 m/s. These values yielded the propagation delay of 1.57 and 1.63 ms for female and male speakers, respectively. The fundamental frequency of each vowel sound was computed by searching for the peak of the autocorrelation function from the differentiated EGG signal. For female speakers, the mean F0 was 195 Hz (min: 178 Hz, max: 211 Hz) and 199 Hz (min: 182 Hz, max: 216 Hz) in normal and pressed phonation, respectively. For males, the mean F0 was 104 Hz (min: 90 Hz, max: 119 Hz) and 114 Hz (min: 95 Hz, max: 148 Hz) in normal and pressed phonation, respectively.

B. Synthetic vowels

A fundamental problem present both in developing new IF algorithms and in comparing existing methods is the fact that assessing the performance of an IF technique is complicated. When IF is used to estimate the glottal flow of natural speech, it is actually never possible to assess in detail how closely the obtained waveform corresponds to the true glottal flow generated by the vibrating vocal folds. It is, however, possible to assess the accuracy of IF by using synthetic speech that has been created using artificial glottal waveform. This kind of evaluation, however, is not truly objective because speech synthesis and IF analysis are typically based on similar models of the human voice production apparatus, for example, the traditional linear source-filter model (Fant, 1970).

In the current study, a different strategy was used in order to evaluate the performance of different CP analysis methods in the estimation of the glottal flow. The idea is to use *physical modeling* of the vocal folds and the vocal tract in order to simulate time-varying waveforms of the glottal

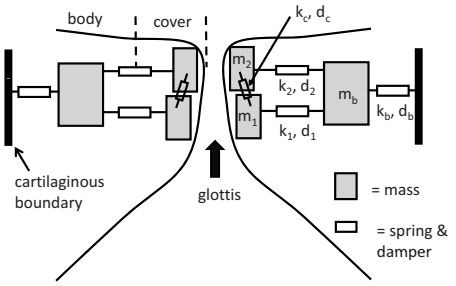


FIG. 6. Schematic diagram of the lumped-element vocal fold model. The cover-body structure of each vocal fold is represented by three masses that are coupled to each other by spring and damping elements. Bilateral symmetry was assumed for all simulations.

flow and radiated acoustic pressure. By using the simulated pressure waveform as an input to an IF method, it is possible to determine how closely the obtained estimate of the voice source matches the simulated glottal flow. This approach is different from using synthetic speech excited by an artificial form of the glottal excitation because the glottal flow waveform results from the interaction of the self-sustained oscillation of the vocal folds with subglottal and supraglottal pressures, as would occur during real speech production. Hence, the glottal flow waveform generated by this model is expected to provide a more stringent and realistic test of the IF method than would be permitted by a parametric flow waveform model where no source-tract interaction is incorporated.[4]

The sound pressure and glottal flow waveforms used to test the new IF technique were generated with a computational model of the vocal folds and acoustic wave propagation. Specifically, self-sustained vocal fold vibration was simulated with three masses coupled to one another through stiffness and damping elements (Story and Titze, 1995). A schematic diagram of the model is shown in Fig. 6, where the arrangement of the masses was designed to emulate the body-cover structure of the vocal folds (Hirano, 1974). The input parameters consisted of lung pressure, prephonatory glottal half-width (adduction), resting vocal fold length and thickness, and normalized activation levels of the cricothyroid (CT) and thyroarytenoid (TA) muscles. These values were transformed to mechanical parameters of the model, such as mass, stiffness, and damping, according to the "rules" proposed by Titze and Story (2002). The vocal fold model was coupled to the pressures and air flows in the trachea and vocal tract through aerodynamic and acoustic considerations as specified by Titze (2002), thus allowing for self-sustained oscillation. Bilateral symmetry was assumed for all simulations such that identical vibrations occur within both the left and right folds. Nine different fundamental frequency values (105, 115, 130, 145, 205, 210, 230, 255, and 310 Hz), which roughly approximate the ranges typical of adult male and female speech (e.g., Hollien *et al.*, 1971; Hollien and Shipp, 1972; Stoicheff, 1981), were generated by modifying the resting vocal fold length and activation levels of the CT and TA muscles; all other input parameters were held constant. The input parameters for all nine cases are shown in Table I. Those cases with the resting length (L_o) equal to 1.6 cm were intended to be representative of the male F0 range, whereas those with L_o=0.9 cm were intended to be in the female F0 range.

Acoustic wave propagation in both the trachea and vocal tract was computed in time synchrony with the vocal fold model. This was performed with a wave-reflection approach (e.g., Strube, 1982; Liljencrants, 1985) where the area functions of the vocal tract and trachea were discretized into short cylindrical sections or tubelets. Reflection and transmission coefficients were calculated at the junctions of consecutive tubelets, at each time sample. From these, pressure and volume velocity were then computed to propagate the acoustic waves through the system. The glottal flow was determined by the interaction of the glottal area with the time-varying pressures present just inferior and superior to the glottis as specified by Titze (2002). At the lip termination, the forward and backward traveling pressure wave components were subjected to a radiation load modeled as a resistance in parallel with an inductance (Flanagan, 1972), intended to approximate a piston in an infinite plane baffle. The output pressure is assumed to be representative of the pressure radiated at the lips. To the extent that the piston-in-a-baffle reasonably approximates the radiation load, the calculated output pressure can also be assumed to be representative of the pressure that would be transduced by a microphone in a non-reflective environment. The specific implementation of the vocal tract

TABLE I. Input parameters for the vocal fold model used to generate the nine different fundamental frequencies. Notation is identical to that used in Titze and Story (2002). The a_{CT} and a_{TA} are normalized activation levels (can range from 0 to 1) of the CT and TA muscles, respectively. L_o and T_o are the resting length and thickness of the vocal folds, respectively. ξ_{01} and ξ_{02} are the prephonatory glottal half-widths at the inferior and superior edges of vocal folds, respectively, and P_L is the respiratory pressure applied at the entrance of the trachea (see Fig. 7). The value of P_L shown in the table is equivalent to a pressure of 8 cm H_2O.

	Fundamental frequency (Hz)								
Parameter value	105	115	130	145	205	210	230	255	310
a_{CT}	0.1	0.4	0.1	0.4	0.2	0.3	0.3	0.4	0.7
a_{TA}	0.1	0.1	0.4	0.4	0.2	0.2	0.3	0.4	0.5
L_o (cm)	1.6	1.6	1.6	1.6	0.9	0.9	0.9	0.9	0.9
T_o (cm)	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
ξ_{01} (cm)	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
ξ_{02} (cm)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P_L (dyn/cm^2)	7840	7840	7840	7840	7840	7840	7840	7840	7840

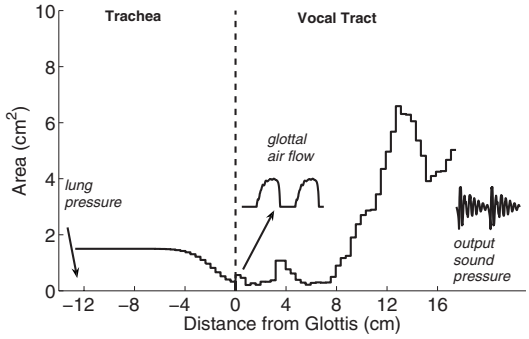


FIG. 7. Area function representation of the trachea and vocal tract used to simulate the male [a] vowel. The vocal fold model of Fig. 6 would be located at the 0 cm point indicated by the dashed vertical line. Examples of the glottal flow and output pressure waveforms are shown near the locations at which they would be generated.

model used for this study was presented in Story (1995) and included energy losses due to viscosity, yielding walls, heat conduction, as well as radiation at the lips.

In the model, a specific vocal tract shape is represented as an area function. For this study, glottal flow and output pressure waveforms were generated based on the area function for the [a] vowel reported by Story *et al.* (1996). For simulations of this vowel with the four lowest fundamental frequencies (105, 115, 130, and 145 Hz), the vocal tract length was set to 17.46 cm. For the five higher F0 speech simulations, exactly the same [a] vowel area function was used, but the length was non-uniformly scaled to 14.28 cm with scaling factors based on those reported by Fitch and Giedd (1999). The purpose of the shortened tract length was to provide an approximation of a possible female-like vocal tract to coincide with the higher F0 simulations. Although a measured female area function could have been used (e.g., Story, 2005), scaling the length of the male [a] vowel was done so that all cases resulted from fairly simple modifications of the same basic model.

A conceptualization of the complete model is given in Fig. 7, where the vocal fold model is shown to be located between the trachea and the vocal tract. The vocal tract is shown configured with the shape and length of the adult male [a] vowel, and the trachea is a uniform tube with a cross-sectional area of 1.5 cm^2 but tapered to 0.3 cm^2 near the glottis. An example glottal flow waveform is indicated near the middle of the figure. Note that the ripples in the waveform are largely due to interaction of the flow with the formant oscillations in the vocal tract. The coupling of the trachea to the vocal tract (via glottal area), however, will slightly alter the overall resonant structure of the system and, hence, will also contribute to glottal waveform shape. The sound pressure waveform radiated at the lips is also shown at the lip end of the area function and, as mentioned previously, can be considered analogous to a microphone signal recorded for a speaker.

In summary, the model is a simplified but physically-motivated representation of a speaker in which glottal airflow and output pressure waveforms result from self-sustained oscillation of the vocal folds and their interaction with propagating pressure waves within the trachea and vocal tract. The model generates both the signal on which IF is typically performed (microphone signal) and the signal that it seeks to determine (glottal flow), thus providing a reasonably realistic test case for IF algorithms.

C. Experiments

Four representative examples of glottal flow pulse forms computed by the proposed CP algorithm are shown in Fig. 8.

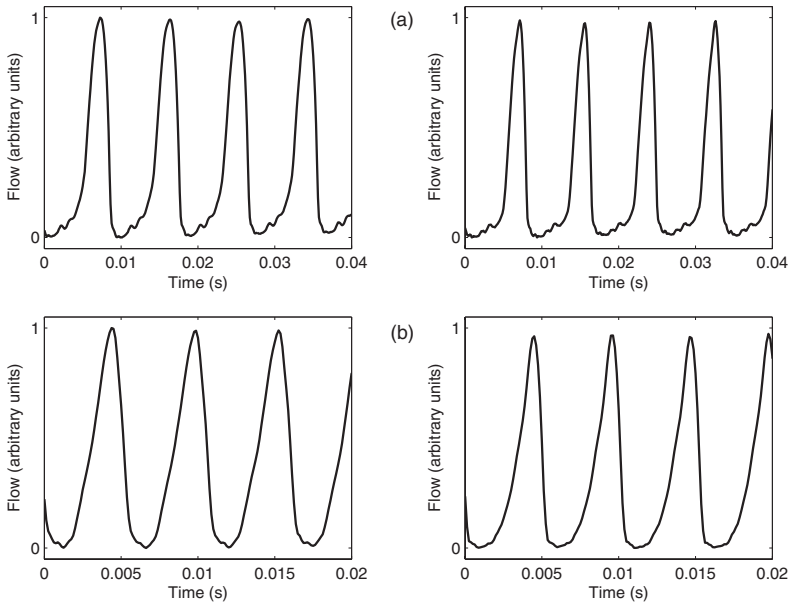


FIG. 8. Examples of glottal flows estimated by the proposed CP$_{con}$ method. IF was computed from [a] vowels produced by a male (panels a) and a female (panels b) speaker using normal (left panels) and pressed (right panels) phonation.

The examples shown in Figs. 8(a) and 8(b) were computed from a male and female speaker, respectively, by using both normal and pressed phonations of the vowel [a]. All these estimates of the glottal excitation were computed by using parameter values given at the end of Sec. II B 3. The beginning of the covariance frame was adjusted to a time instant three samples after the negative peak of the EGG derivative. It can be seen in Fig. 8 that none of the estimated glottal pulse forms show abrupt high amplitude peaks at the end of the closing phase, indicating that inverse filter roots are most likely located correctly in the formant region rather than in unrealistic positions at low frequency. CP can be identified rather easily from all the examples shown. However, the waveforms estimated from utterances spoken by the male speaker show a small ripple component. This ripple might be due to incomplete canceling of some of the higher formants by the inverse filter. Alternatively, this component might be explained by the existence of nonlinear coupling between the source and the tract, which cannot be taken into account in CP analysis because it is based on linear modeling of the voice production system.

The performance of the proposed CP analysis algorithm was tested by conducting two major experiments, one of which used synthetic vowels and the other natural speech. Both experiments involved estimating the glottal flow with three CP analysis types. The first one, denoted by CP_{bas} in the rest of the paper, is represented by the basic CP analysis in which the vocal tract model computed by the covariance analysis is used as such in IF. The second one, denoted by CP_{rem}, is the most widely used form of the CP analysis in which the roots of the inverse filter polynomial computed by the covariance analysis are solved, and those located on the positive real axis are removed before IF. The third type, denoted by CP_{con}, is the proposed method based on the constrained LP described in Sec. II B.

In both experiments, the robustness of each CP analysis to the position of the covariance frame was evaluated by varying the beginning of the frame position near its optimal value, n_{opt}, the instant of glottal closure. For synthetic vowels, n_{opt} was first adjusted by using the derivative of the flow pulse generated by the physical vocal fold model. In this procedure, the optimal beginning of the covariance frame was set to the time instant after the negative peak of the flow derivative when the waveform returns to the zero level. For each synthetic vowel, the beginning of the covariance frame was then varied in 11 steps by defining the start index as $n=n_{opt}+i$, where $i=-5$ to $+5$. (In other words, the optimal frame position corresponds to index value $i=0$.) For natural vowels, the position of the covariance frame was varied by first extracting the glottal closure as the time instant when the EGG derivative reached a negative peak within a glottal cycle. Again, 11 frame positions were analyzed around this instant of glottal closure.

For synthetic sounds, there is no variation between periods, and, therefore, only a single cycle was analyzed. The total number of CP analyses conducted for synthetic speech was 297 (3 CP methods $\times$ 9 F0 values $\times$ 11 frame positions per cycle). For natural vowels, the analysis was repeated for six consecutive glottal cycles. Hence, the total number of CP analyses conducted for natural speech was 5148 (3 CP methods $\times$ 2 phonation types $\times$ 13 speakers $\times$ 11 frame positions per cycle $\times$ 6 cycles). The estimated glottal flows were parametrized using two frequency-domain measures. The first of these, H1H2, is defined as the difference in decibel between the amplitudes of the fundamental and the second harmonic of the source spectrum (Titze and Sundberg, 1992). The second parameter, the harmonic richness factor (HRF), is defined from the spectrum of the glottal flow as the difference in decibel between the sum of the harmonic amplitudes above the fundamental and the amplitude of the fundamental (Childers and Lee, 1991). (Notice the difference in the computation of the spectral ratio between the two parameters: if only the second harmonic is included in HRF, then its value becomes equal to H1H2 multiplied by -1.) These parameters were selected for two reasons. First, both of them can be computed automatically without any user adjustments. In CP analysis with a varying frame position, this is highly justified because the glottal flow waveforms, especially those computed with CP_{bas}, are sometimes so severely distorted that their reliable parametrization with, for example, time-based glottal flow quotients is not possible. Second, both H1H2 and HRF are known to reflect the spectral decay of the glottal excitation: a slowly decaying source spectrum is reflected by a small H1H2 and a large HRF value. Hence, if the glottal flow estimate is severely distorted by artifacts seen as jags in the closing phase, as shown in Figs. 1(b), 1(c), and 3(c), one is expected to get a decreased H1H2 value and an increased HRF value because the spectrum of the distorted glottal flow approaches that of the impulse train, that is, a flat spectral envelope. Since HRF takes into account a larger number of spectral harmonics, one can argue that its value reflects more reliable changes in the glottal flow than H1H2. Therefore, HRF alone might represent a sufficient spectral parameter to be used from the point of view of the present study. H1H2 is, however, a more widely used parameter in voice production studies, which justifies its selection as an additional voice source parameter in the present investigation.

IV. RESULTS

A. Experiment 1: Synthetic vowels

Robustness of the different CP analyses to the covariance frame position is demonstrated for the synthetic vowels by the data given in Table II. H1H2 and HRF values were first computed in each covariance frame position with each of the three CP techniques. For both H1H2 and HRF, the difference between the parameters extracted from the original flow and the estimated flow was computed. The data in Table II show the absolute value of this difference computed as an average pooled over 11 frame positions. The obtained results indicate that the error in both H1H2 and HRF due to the variation of the CP frame position is smallest for all vowels with F0 less than 310 Hz when IF is computed with the proposed new method. The average value of H1H2, when pooled over all vowels with F0 less than 310 Hz, equaled to 2.6, 0.9, and 0.5 dB for CP_{bas}, CP_{rem}, and CP_{con}, respectively. For HRF, the average value equaled to 7.8, 3.8, and 2.4 dB for CP_{bas}, CP_{rem}, and CP_{con}, respectively. For the synthetic

TABLE II. Effect of the covariance frame position on H1H2 and HRF using vowels synthesized by physical modeling. Absolute value of the difference (in dB) was computed between parameter values extracted from the original flows and from the glottal flows estimated by IF. Inverse filtering was computed by three CP algorithms: CP_{bas}, CP_{rem}, and CP_{con}. Data were averaged over 11 different frame positions starting around the instant of glottal closure.

F0 (Hz)	Diff in H1H2 (dB)			Diff in HRF (dB)		
	CP_{bas}	CP_{rem}	CP_{con}	CP_{bas}	CP_{rem}	CP_{con}
105	1.36	0.06	0.03	5.08	2.30	1.87
115	2.93	0.14	0.08	9.15	2.37	1.75
130	1.81	0.13	0.06	6.14	2.23	1.63
145	3.42	0.10	0.07	10.36	1.74	1.59
205	2.98	1.66	0.83	8.80	6.13	2.72
210	2.67	1.40	0.82	8.06	5.91	2.91
230	3.17	1.28	0.90	8.31	4.00	3.21
255	2.40	2.13	1.24	6.35	5.44	3.40
310	0.69	0.69	3.38	6.24	6.24	5.02

vowel with the largest F0 value, the best result was also given by CP_{con} when the parametrization was performed with HRF. However, H1H2 indicated a surprisingly small error for this high-pitch vowel when IF was conducted with CP_{bas} and CP_{rem}. The waveforms, however, were greatly distorted, but the levels of the fundamental and the second harmonic, that is, those sole spectral components used in the computation of H1H2, were only marginally affected. It is, though, worth emphasizing that the length of the glottal CP for this high-pitch vowel with F0=310 Hz is only ten samples (1.25 ms). This implies that the underlying assumption underlying all the three assessed IF techniques, that is, the existence of sufficiently long CP, is greatly violated. Hence, the surprisingly small value of H1H2 difference for this signal is explained mainly by the shortcomings of the simple spectral parameter rather than by the successful voice source estimation. In summary, the experiments conducted with synthetic vowels indicate that the proposed CP algorithm was the least vulnerable to the covariance frame position among the three techniques when voices of different F0 were compared.

B. Experiment 2: Natural vowels

The standard deviations (std) and means of the H1H2 and HRF values extracted from the glottal flows computed from natural vowels of varying covariance frame positions were compared with repeated measures analyses of variance (ANOVAs). The data were analyzed with sex × method × phonation ANOVAs where "sex" included male and female sexes, factor "method" included three different CP algorithms, CP_{bas}, CP_{rem}, and CP_{con}, and factor "phonation" included phonation types normal and pressed. H1H2 and HRF data were analyzed with separate ANOVAs, and Newman–Keuls tests were used as a means of *post hoc* analysis for pairwise differences in the data. The standard deviations and mean values of H1H2 and HRF obtained from the 66 window positions (11 frame positions of 6 cycles) are shown in Fig. 9. The main and interaction effects of the corresponding ANOVA results are given in Table III.

The standard deviation of both H1H2 and HRF differed significantly between the IF methods. *Post hoc* analyses showed that the standard deviations of H1H2 and HRF were, on the average, smaller when the new CP method

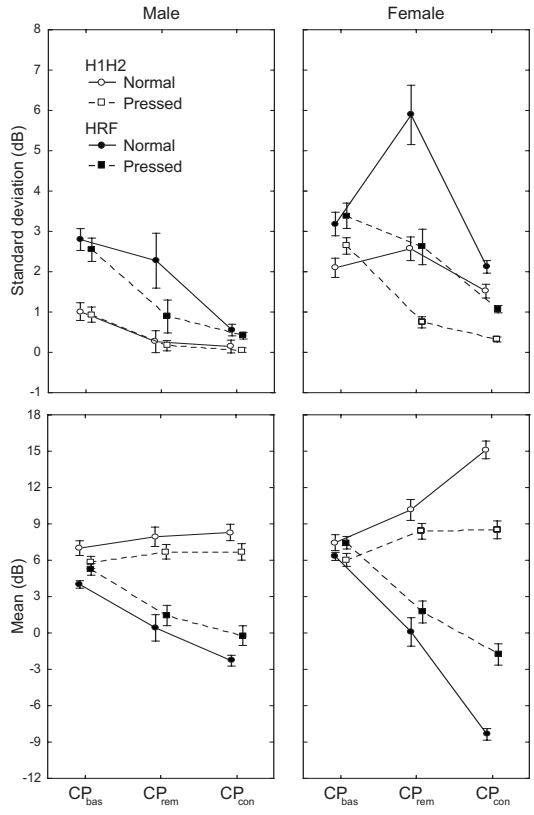


FIG. 9. Standard deviations (top panels) and means (bottom panels) of H1H2 and HRF according to the speaker sex and the type of phonation for CP analyses computed by CP_{bas}, CP_{rem}, and CP_{con}. Error bars represent standard error of the mean.

TABLE III. ANOVA results for standard deviations and means of H1H2 (upper table) and HRF (lower table). The degrees of freedom (DF), Greenhouse–Geisser epsilons (ε), F-values, and the associated probability (p) values are shown for each ANOVA effect. Analyses were conducted for utterances produced by 13 speakers using 11 covariance frame positions per glottal cycle and 6 successive periods.

		Standard deviations			Means		
H1H2 Effects and degrees of freedom (Df1, Df2)		ε	F	p	ε	F	p
Sex	1, 11		83.86	<0.001		9.38	<0.05
Method	2, 22	0.73	47.22	<0.001	0.69	88.85	<0.001
Method × sex	2, 22	0.73	4.12	<0.05	0.69	37.70	<0.001
Phonation	1, 11	1.00	10.42	<0.01	1.00	20.43	<0.001
Phonation × sex	1, 11	1.00	6.80	<0.05	1.00	3.58	ns
Method × phonation	2, 22	0.95	16.26	<0.001	0.68	23.22	<0.001
Method × phonation × sex	2, 22	0.95	15.57	<0.001	0.68	16.81	<0.001
		Standard deviations			Means		
HRF Effects and degrees of freedom (Df1, Df2)		ε	F	p	ε	F	p
Sex	1, 11		49.33	<0.001		0.83	ns
Method	2, 22	0.61	26.43	<0.001	0.87	262.27	<0.001
Method × sex	2, 22	0.61	6.26	<0.05	0.87	30.99	<0.001
Phonation	1, 11	1.00	18.12	<0.01	1.00	15.22	<0.01
Phonation × sex	1, 11	1.00	2.86	ns	1.00	2.06	ns
Method × phonation	2, 22	0.77	12.83	<0.001	0.67	9.29	<0.01
Method × phonation × sex	2, 22	0.77	3.10	ns	0.67	4.58	<0.05

ns=not significant

(H1H2-std=0.5, HRF-std=1.0) was used than when either CP_{bas} (H1H2-std=1.6, HRF-std=3.0) or CP_{rem} (H1H2-std =0.9, HRF-std=2.8) was used.

For H1H2, the difference between CP_{bas} and CP_{rem} was also significant. Additional effects on H1H2 and HRF variability were observed for sex and phonation type. The H1H2 and HRF standard deviations were larger for female (H1H2-std=1.6, HRF-std=3.0) than for male (H1H2-std =0.4, HRF-std=1.6) speakers. Further, the variability of H1H2 and HRF was larger for the normal (H1H2-std=1.2, HRF-std=2.7) than for the pressed (H1H2-std=0.8, HRF-std=1.8) type of phonation. Finally, significant method × sex, method × phonation, and method × phonation × sex interactions were found for both H1H2 and HRF, and a phonation × sex interaction was additionally significant for the H1H2.

The results indicated a statistically significant effect of CP method on the mean H1H2 and HRF values. The mean H1H2 and HRF values increased and decreased, respectively, when the IF algorithm CP_{bas} (H1H2=6.6 and HRF=5.7) was changed to CP_{rem} (H1H2=8.2 and HRF=0.9) and, then, further to the new CP_{con} algorithm (H1H2=9.5 and HRF= −3.0). While HRF mean values were similar for both sexes, the average H1H2 values were larger for female (9.3) than for male (7.1) speakers. Additionally, a smaller mean H1H2 and a larger mean HRF value was observed for the pressed phonation (H1H2=7.0 and HRF=2.3) than for the normal phonation (H1H2=9.2 and HRF=0.1). Finally, significant method × sex, method × phonation, and method × phonation × sex interactions were found for both H1H2 and HRF data.

V. CONCLUSIONS

CP covariance analysis, a widely used glottal IF method, computes a parametric model of the vocal tract by conducting linear predictive analysis over a frame that is located in the CP of the glottal cycle. Since the length of the CP is typically short, the resulting all-pole model is highly vulnerable with respect to the extraction of the frame position. Even a minor change in the frame position might greatly affect the z-domain locations of the roots of the all-pole model given by LP. This undesirable feature of the conventional CP analysis typically results in vocal tract models, which have roots, both real and complex, at low frequencies or roots that are located outside of the unit circle. These kinds of false root locations, in turn, result in distortion of the glottal flow estimates, which is typically seen as unnatural peaks at the instant of glottal closure, the so-called jags, or as increased formant ripple during the CP.

The present study proposed an improved version of the CP analysis based on a combination of two algorithmic issues. First, and most importantly, a constraint is imposed on the dc gain of the inverse filter prior to the optimization of the coefficients. With this constraint, linear predictive analysis is more prone to give vocal tract models that can be justified from the point of view of the source-filter theory of vowel production; that is, they show complex conjugate roots in the vicinity of formant regions rather than unrealistic resonances at low frequencies. Second, the new CP method utilizes an inverse filter that is minimum phase, a property that is not typically used in glottal IF.

The new glottal IF method, CP_{con}, was compared to two CP analysis techniques by using both synthetic vowels produced by physical modeling of the voice production apparatus and natural vowels produced by male and female speakers. In summary, the experiments conducted with synthetic vowels having F0 from 105 to 310 Hz indicate that the proposed CP method gave glottal flow estimates with better robustness to the covariance frame position than the conventional CP methods. The result suggests that the parametric model of the vocal tract computed with the dc-constrained linear predictive analysis is less prone to distortion by the problem typically met in the CP analysis, namely, the involvement of samples outside the CP in the computation of the vocal tract. This problem violates the basic assumption of the CP analysis that the estimation of the vocal tract transfer function is made during the excitation-free time span. It can be argued that this violation is larger for voices of high pitch because they typically show short CPs in the glottal excitation. Violation results in the occurrence of unjustified inverse filter roots at low frequencies, which, in turn, distorts the resulting glottal flow estimates. Based on the results achieved with synthetic speech, the involvement of the dc constraint in the optimization process of the vocal tract model, however, seems to reduce this distortion and hence improve the estimation robustness with respect to the CP frame position. It must be emphasized, though, that if the amount of data samples during the glottal CP becomes extremely small, which was the case in analyzing the vowel with F0=310 Hz in the present investigation, distortion of the glottal flow estimates becomes large with all CP techniques.

The experiments conducted with natural speech indicate that the deviation of H1H2 and HRF due to the varying of the covariance frame position inside the glottal cycle was larger for female speech than for male vowels and the deviation was also larger in normal than in pressed phonation. These results are in line with findings reported in previous studies (e.g., Veeneman and BeMent, 1985) as well as with experiments conducted in the present investigation with synthetic speech, indicating that the robustness of the CP analysis with respect to the frame position tends to decrease for shorter CP intervals, as in higher F0 speech or in normal as opposed to pressed phonation. The proposed new CP method, importantly, gave the smallest deviation of H1H2 and HRF, suggesting that the involvement of the dc constraint reduces the sensitivity of the CP analysis to the covariance frame position and that this holds true also for natural vowels. This finding is also supported by the fact that the mean levels of H1H2 and HRF were found to be largest and smallest, respectively, when IF was computed with CP_{con}. In other words, the average spectral decay of the glottal flow pulse forms computed by varying the frame position was steeper with CP_{con} than with the other two CP methods. This is explained by the frequency-domain effect produced by distortion represented by impulse-like jags: the larger their contribution, the flatter the spectrum.

In summary, the proposed IF method constitutes a potential means to compute the CP covariance analysis to estimate the glottal flow from speech pressure signals. It reduces distortion caused by one of the major drawbacks of the conventional CP analysis, the sensitivity of the analysis to the position of the covariance frame. The computational load of the new method is only slightly larger than that of the conventional CP method. In addition, the method can be implemented in a manner similar to the conventional one, that is, either based solely on the speech pressure signal or in a two-channel mode where an EGG signal is used to help extract the covariance frame position. Therefore, there are no obstacles in principle for the implementation of the proposed method in environments where the conventional analysis is used. One has to keep in mind, though, that the new method does not change the basic assumptions of the CP analysis, namely, that the voice source and vocal tract are linearly separable, and there is a CP of finite duration during which there is no excitation by the source of the tract.

ACKNOWLEDGMENTS

This work was supported by the Academy of Finland (Project No. 111848) and by the COST action 2103, "Advanced Voice Function Assessment."

[1]It is worth emphasizing that glottal pulses estimated from natural speech sometimes show fluctuation, typically referred to as "ripple," after the instant of glottal closure. This component might correspond to actual phenomena or it may result from incorrect inverse filter settings. If the pulse waveform is fluctuating after the instant of the glottal closure, it is, though, difficult, if not impossible, to define accurately which part of the fluctuation corresponds to real phenomena and which part results from incorrect IF. If, however, the flow waveform shows an abrupt peak at the end of the closing phase, such as in Fig. 1(c), and if this component is removed by, for example, a minor change in the position of the analysis frame, it is more likely that the component represents an artifact than a real phenomenon.

[2]By using Eq. (4), the gain of the vocal tract filter at dc, denoted by G_{dc}, is defined as the absolute value of the inverse of the frequency response of the constrained predictor at $\omega=0$: $G_{dc}=|1/C(e^{j0})|=|1/l_{dc}|$. In principle, the requirement $G_{dc}=1$ can be satisfied by assigning either $l_{dc}=1$ or $l_{dc}=-1$. Although both of these values result in vocal tract filters of equal gain at dc, they end up as different constrained transfer functions. In order to test the difference between the two values of l_{dc}, the glottal flows were estimated from the synthetic vowels described in Sec. III B by using H1H2 and HRF parameters described in Sec. III C and by conducting the constrained IF analysis by assigning both $l_{dc}=1$ and $l_{dc}=-1$. The results indicated clearly that the choice $l_{dc}=-1$ yielded glottal flow estimates that were closer to the original flows generated by the physical modeling approach.

[3]In the area of glottal IF, most studies analyze vowels with high first formant such as [a] or [ae]. The reason for this is the fact that the separation of the source and the tract becomes increasingly difficult from a mathematical point of view if the first formant is low. This is due to the fact that the strong harmonics at low frequencies bias the estimation of the first formant in all-pole modeling (El-Jaroudi and Makhoul, 1991). This, in turn, results in severe distortion of the glottal flow estimates.

[4]It should be noted that while synthetic vowels produced by the physical modeling approach mimic real speech production by involving source-tract interaction, this effect is not taken into account in CP analysis, which simply assumes that the source and tract are linearly separable (Strube, 1974; Wong *et al.*, 1979). The proposed dc-constrained LP is a new mathematical method to compute the vocal tract model of CP analysis, but it does not in any way change the underlying assumption of the linear coupling between the source and the tract. Therefore, the use of physically-motivated synthetic speech was justified by a need to have more realistic artificial vowels as test material, not by a goal to analyze how source-tract interaction affects different versions of the CP technique, all of which are based on the linear source-filter theory and are therefore unable to take into account the coupling between the source and the tract.

Airas, M., and Alku, P. (**2006**). "Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalized amplitude quotient," Phonetica **63**, 26–46.
Akande, O., and Murphy, P. (**2005**). "Estimation of the vocal tract transfer function with application to glottal wave analysis," Speech Commun. **46**, 15–36.
Alku, P. (**1992**). "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," Speech Commun. **11**, 109–118.
Alku, P., and Vilkman, E. (**1996**). "A comparison of glottal voice source quantification parameters in breathy, normal, and pressed phonation of female and male speakers," Folia Phoniatr Logop **48**, 240–254.
Arroabarren, I., and Carlosena, A. (**2004**). "Vibrato in singing voice: The link between source-filter and sinusoidal models," EURASIP J. Appl. Signal Process. **7**, 1007–1020.
Bäckström, T., and Alku, P. (**2006**). "Harmonic all-pole modelling for glottal inverse filtering," in CD Proceedings of the seventh Nordic Signal Processing Symposium, Reykjavik, Iceland.
Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (**1993**). *Nonlinear Programming: Theory and Algorithms* (Wiley, New York).
Bozkurt, B., Doval, B., D'Alessandro, C., and Dutoit, T. (**2005**). "Zeros of z-transform representation with application to source-filter separation of speech," IEEE Signal Process. Lett. **12**, 344–347.
Campedel-Oudot, M., Cappe, O., and Moulines, E. (**2001**). "Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach," IEEE Trans. Speech Audio Process. **9**, 469–481.
Carlson, R., Granström, B., and Karlsson, I. (**1991**). "Experiments with voice modelling in speech synthesis," Speech Commun. **10**, 481–489.
Childers, D., and Ahn, C. (**1995**). "Modeling the glottal volume-velocity waveform for three voice types," J. Acoust. Soc. Am. **97**, 505–519.
Childers, D., and Hu, H. (**1994**). "Speech synthesis by glottal excited linear prediction," J. Acoust. Soc. Am. **96**, 2026–2036.
Childers, D., and Lee, C. (**1991**). "Vocal quality factors: Analysis, synthesis, and perception," J. Acoust. Soc. Am. **90**, 2394–2410.
Childers, D., and Wong, C.-F. (**1994**). "Measuring and modeling vocal source-tract interaction," IEEE Trans. Biomed. Eng. **41**, 663–671.
Cummings, K. E., and Clements, M. A. (**1995**). "Analysis of the glottal excitation of emotionally styled and stressed speech," J. Acoust. Soc. Am. **98**, 88–98.
El-Jaroudi, A., and Makhoul, J. (**1991**). "Discrete all-pole modeling," IEEE Trans. Signal Process. **39**, 411–423.
Eysholdt, U., Tigges, M., Wittenberg, T., and Pröschel, U. (**1996**). "Direct evaluation of high-speed recordings of vocal fold vibrations," Folia Phoniatr Logop **48**, 163–170.
Fant, G. (**1970**). *Acoustic Theory of Speech Production* (Mouton, The Hague).
Fitch, T., and Giedd, J. (**1999**). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," J. Acoust. Soc. Am. **106**, 1511–1522.
Flanagan, J. (**1972**). *Speech Analysis, Synthesis and Perception* (Springer, New York).
Fröhlich, M., Michaelis, D., and Strube, H. (**2001**). "SIM—Simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals," J. Acoust. Soc. Am. **110**, 479–488.
Fu, Q., and Murphy, P. (**2006**). "Robust glottal source estimation based on joint source-filter model optimization," IEEE Trans. Audio, Speech, Lang. Process. **14**, 492–501.
Gobl, C., and Ní Chasaide, A. (**2003**). "The role of voice quality in communicating emotion, mood and attitude," Speech Commun. **40**, 189–212.
Hertegård, S., Gauffin, J., and Karlsson, I. (**1992**). "Physiological correlates of the inverse filtered flow waveform," J. Voice **6**, 224–234.
Hirano, M. (**1974**). "Morphological structure of the vocal cord as a vibrator and its variations," Folia Phoniatr Logop **26**, 89–94.
Hirano, M. (**1981**). *Clinical Examination of Voice* (Springer, New York).
Hollien, H., Dew, D., and Philips, P. (**1971**). "Phonational frequency ranges of adults," J. Speech Hear. Res. **14**, 755–760.
Hollien, H., and Shipp, T. (**1972**). "Speaking fundamental frequency and chronologic age in males," J. Speech Hear. Res. **15**, 155–159.
Kasuya, H., Maekawa, K., and Kiritani, S. (**1999**). "Joint estimation of voice source and vocal tract parameters as applied to the study of voice source dynamics," in Proceedings of the International Congress on Phonetic Sciences, San Francisco, CA, pp. 2505–2512.
Klatt, D., and Klatt, L. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.
Krishnamurthy, A., and Childers, D. (**1986**). "Two-channel speech analysis," IEEE Trans. Acoust., Speech, Signal Process. **34**, 730–743.
Larar, J., Alsaka, Y., and Childers, D. (**1985**). "Variability in closed phase analysis of speech," in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Tampa, FL, pp. 1089–1092.
Lecluse, F., Brocaar, M., and Verschuure, J. (**1975**). "The electroglottography and its relation to glottal activity," Folia Phoniatr. **17**, 215–224.
Lehto, L., Laaksonen, L., Vilkman, E., and Alku, P. (**2008**). "Changes in objective acoustic measurements and subjective voice complaints in call-center customer-service advisors during one working day," J. Voice **22**, 164–177.
Liljencrants, J. (**1985**). "Speech synthesis with a reflection-type line analog," DS dissertation, Department of Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm, Sweden.
Makhoul, J. (**1975**). "Linear prediction: A tutorial review," Proc. IEEE **63**, 561–580.
Markel, J., and Gray, A., Jr. (**1976**). *Linear Prediction of Speech* (Springer-Verlag, Berlin).
Milenkovic, P. (**1986**). "Glottal inverse filtering by joint estimation of an AR system with a linear input model," IEEE Trans. Acoust., Speech, Signal Process. **34**, 28–42.
Miller, R. (**1959**). "Nature of the vocal cord wave," J. Acoust. Soc. Am. **31**, 667–677.
Naylor, P., Kounoudes, A., Gudnason, J., and Brookes, M. (**2007**). "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," IEEE Trans. Audio, Speech, Lang. Process. **15**, 34–43.
Oppenheim, A., and Schafer, R. (**1989**). *Discrete-Time Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ).
Plumpe, M., Quatieri, T., and Reynolds, D. (**1999**). "Modeling of the glottal flow derivative waveform with application to speaker identification," IEEE Trans. Speech Audio Process. **7**, 569–586.
Price, P. (**1989**). "Male and female voice source characteristics: Inverse filtering results," Speech Commun. **8**, 261–277.
Rabiner, L., and Schafer, R. (**1978**). *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, NJ).
Riegelsberger, E., and Krishnamurthy, A. (**1993**). "Glottal source estimation: Methods of applying the LF-model to inverse filtering," in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Minneapolis, MN, Vol. **2**, pp. 542–545.
Rothenberg, M. (**1973**). "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," J. Acoust. Soc. Am. **53**, 1632–1645.
Shiga, Y., and King, S. (**2004**). "Accurate spectral envelope estimation for articulation-to-speech synthesis," in the CD Proceedings of the Fifth ISCA Speech Synthesis Workshop, Pittsburgh, PA.
Stoicheff, M. L. (**1981**). "Speaking fundamental frequency characteristics of nonsmoking female adults," J. Speech Hear. Res. **24**, 437–441.
Story, B. (**1995**). "Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract," Ph.D. dissertation, University of Iowa.
Story, B. (**2005**). "Synergistic modes of vocal tract articulation for American English vowels," J. Acoust. Soc. Am. **118**, 3834–3859.
Story, B., and Titze, I. (**1995**). "Voice simulation with a body-cover model of the vocal folds," J. Acoust. Soc. Am. **97**, 1249–1260.
Story, B., Titze, I., and Hoffman, E. (**1996**). "Vocal tract area functions from magnetic resonance imaging," J. Acoust. Soc. Am. **100**, 537–554.
Strik, H., and Boves, L. (**1992**). "On the relation between voice source parameters and prosodic features in connected speech," Speech Commun. **11**, 167–174.
Strube, H. (**1974**). "Determination of the instant of glottal closure from the speech wave," J. Acoust. Soc. Am. **56**, 1625–1629.
Strube, H. (**1982**). "Time-varying wave digital filters for modeling analog systems," IEEE Trans. Acoust., Speech, Signal Process. **30**, 864–868.
Sundberg, J., Fahlstedt, E., and Morell, A. (**2005**). "Effects on the glottal voice source of vocal loudness variation in untrained female and male voices," J. Acoust. Soc. Am. **117**, 879–885.
Švec, J., and Schutte, H. (**1996**). "Videokymography: High-speed line scanning of vocal fold vibration," J. Voice **10**, 201–205.
Titze, I. (**2002**). "Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model," J. Acoust. Soc. Am. **111**, 367–376.
Titze, I., and Story, B. (**2002**). "Rules for controlling low-dimensional vocal fold models with muscle activities," J. Acoust. Soc. Am. **112**, 1064–1076.

Titze, I., Story, B., Burnett, G., Holzrichter, J., Ng, L., and Lea, W. (**2000**). "Comparison between electroglottography and electromagnetic glottography," J. Acoust. Soc. Am. **107**, 581–588.
Titze, I., and Sundberg, J. (**1992**). "Vocal intensity in speakers and singers," J. Acoust. Soc. Am. **91**, 2936–2946.
Veeneman, D., and BeMent, S. (**1985**). "Automatic glottal inverse filtering from speech and electroglottographic signals," IEEE Trans. Acoust., Speech, Signal Process. **33**, 369–377.
Vilkman, E. (**2004**). "Occupational safety and health aspects of voice and speech professions," Folia Phoniatr Logop **56**, 220–253.
Wong, D., Markel, J., and Gray, A., Jr. (**1979**). "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE Trans. Acoust., Speech, Signal Process. **27**, 350–355.
Yegnanarayana, B., and Veldhuis, N. (**1998**). "Extraction of vocal-tract system characteristics from speech signals," IEEE Trans. Speech Audio Process. **6**, 313–327.

Study VII

Paavo Alku, Carlo Magi, Tom Bäckström. "Glottal inverse filtering with the closed-phase covariance analysis utilizing mathematical constraints in modelling of the vocal tract", Logopedics, Phoniatrics, Vocology (Special Issue for the Jan Gauffin Memorial Symposium). 2009 (In press).

Logopedics Phoniatrics Vocology
2009, 1–10, iFirst article

informa
healthcare

ORIGINAL ARTICLE

Glottal inverse filtering with the closed-phase covariance analysis utilizing mathematical constraints in modelling of the vocal tract

PAAVO ALKU, CARLO MAGI† & TOM BÄCKSTRÖM

Helsinki University of Technology, Department of Signal Processing and Acoustics, Otakaari 5A, PO Box 3000, FI-02015 TKK, Finland

Abstract
Closed-phase (CP) covariance analysis is a glottal inverse filtering method based on the estimation of the vocal tract with linear prediction (LP) during the closed phase of the vocal fold vibration cycle. Since the closed phase is typically short, the analysis is vulnerable with respect to the extraction of the covariance frame position. The present study proposes a modified CP algorithm based on imposing certain predefined values on the gains of the vocal tract inverse filter at angular frequencies of 0 and π in optimizing filter coefficients. With these constraints, vocal tract models are less prone to show false low-frequency roots. Experiments show that the algorithm improves the robustness of the CP analysis on the covariance frame position.

Key words: *Closed-phase analysis, glottal inverse filtering, linear prediction*

Introduction

Many different inverse filtering (IF) methods have been developed since the 1950s in order to estimate the source of voiced sounds in speech or singing, the glottal volume velocity waveform (e.g. (1–4)). Among the various proposed IF techniques, one of the most widely used is closed-phase (CP) covariance analysis pioneered by the works of Strube (5) and Wong et al. (6). The basis of the CP method is in the source tract theory of voice production, which assumes that human voice production can be modelled as a linear cascade of three processes: the glottal source, the vocal tract, and the lip radiation effect (7). In order to separate the first two parts, the CP analysis assumes that there is no contribution from the glottal source to the vocal tract during the closed phase of the vocal fold vibration cycle. By identifying this time-span of zero input, a parametric model of the vocal tract is computed using linear prediction (LP) with the covariance criterion. LP analysis yields an inverse model of the vocal tract in the form of a digital finite impulse response (FIR) filter which is then used to cancel the effects of the vocal tract resonances. Inverse filtering is typically conducted over several fundamental periods, that is the inverse model of the vocal tract estimated during the glottal closed phase is used pitch-asynchronously to cancel the effects of the vocal tract during both the open and closed phases of consecutive glottal periods.

The CP method has been used as a means to estimate the voice source in various studies since the mid-1970s. These investigations have addressed, for example, the characterization of different phonation types (e.g. (8,9)). In addition, Strik and Boves (10) used the CP analysis in elucidating how voice source features behave in conversational speech and how these effects can be modelled with synthetic models represented by the Liljencrants-Fant (LF) waveform (11). Cummings and Clements (12), in turn, took advantage of the CP technique in studying changes of the glottal flow waveform in emotionally styled and stressed speech. In addition to normal speech, the CP analysis has also been used in efforts to analyse and classify abnormal voice production

Correspondence: Professor Paavo Alku, Helsinki University of Technology, Department of Signal Processing and Acoustics, Otakaari 5A, PO Box 3000, FI-02015 TKK, Finland. Fax: +358-(0)9-460224. E-mail: Paavo.Alku@tkk.fi
†Deceased in February 2008.

(Received 27 November 2008; revised 20 February 2009; accepted 20 March 2009)

ISSN 1401-5439 print/ISSN 1651-2022 online
DOI: 10.1080/14015430902913519

(13–15). Furthermore, several studies have applied the CP method in the area of speech technology such as in speech synthesis (e.g. (16,17)) and in speaker identification (e.g. (18,19)). Finally, CP analysis of the glottal flow has been conducted in the area of singing research in analysing, for example, the production of vibrato (20).

The CP technique has also been a target of methodological development, and the major focus of this research activity has been devoted to the question how to insert accurately the covariance frame into the closed phase. In order to determine this important time-span from a speech waveform, a series of sliding covariance analyses is typically computed by moving the analysis frame sequentially one sample at a time through the speech signal, and the results of each covariance analysis are analysed in order to determine the closed phase. This approach was used by Strube (5), who identified the glottal closure as an instant when the frame was in a position which yielded the maximum determinant of the covariance matrix. Wong et al. (6) instead defined the closed phase as the interval when the normalized squared prediction error was minimum. Plumpe et al. (18) proposed an idea in which the formant frequency modulations between the open and closed phase are used as a means to define the optimal frame position. Akande and Murphy (21) suggested a technique which computes the covariance analysis in an adaptive loop where the optimal filter order and frame position are searched for by using phase information. Furthermore, there are several such studies that do not directly represent glottal inverse filtering but are closely related to the CP analysis, because they focus on the extraction of glottal closure and opening instants. These epoch extraction studies take advantage of, for example, the Hilbert envelope (22) and the group delay function (23,24) in estimating time instants of glottal closure and opening, and they can be applied, if desired, in the CP analysis.

In all CP methods referred to above, the optimal position of the covariance frame is computed from a single source of information represented by the speech signal. Alternatively, electroglottography (EGG) can be used to extract the position and duration of the closed phase. This so-called two-channel analysis has been shown to yield consistent glottal flow estimates due to improved positioning of the covariance frame (15,25). In the two-channel analysis, the CP analysis is typically computed by estimating the closed phase of the glottal cycle as the time interval between the minimum and maximum peaks of the first time derivative of the EGG waveform (9).

In spite of its prevalence, the CP covariance analysis has certain shortcomings. Several previous studies have in particular indicated that glottal flow estimates computed by the CP analysis vary greatly depending on the position of the covariance frame (e.g. (15,26–28)). Examples of this phenomenon are depicted in Figure 1, which shows how glottal flow estimates computed by the CP analysis vary extensively even though there is only a minor change in the position of the covariance frame. Given the fundamental assumption of the method, that is the computation of the vocal tract model during an excitation-free time-span, this undesirable feature of the CP analysis is understandable. The expected length of the closed phase is typically short, which implies that the amount of data used to define the parametric model of the vocal tract with the covariance analysis is sparse. This problem is particularly severe in voices of high fundamental frequency (F0) because they are produced by using very short lengths in the glottal closed phase. If the position of this kind of a short data frame is misaligned, the linear predictive filter model of the vocal tract, optimized in the mean square error (MSE) sense, might locate some of its roots at low frequencies rather than in the formant region. Consequently, the resulting glottal flow estimate, as demonstrated in Figure 1b, shows sharp peaks, called 'jags' by Wong et al. (6), and becomes similar to a time derivative of the flow candidate given by an inverse filter with no such false roots. This severe distortion of the glottal flow estimate caused by the occurrence of inverse filter roots, both real and complex conjugate pairs, at low frequencies, is greatest at time instants when the flow changes most rapidly, that is near glottal closure.

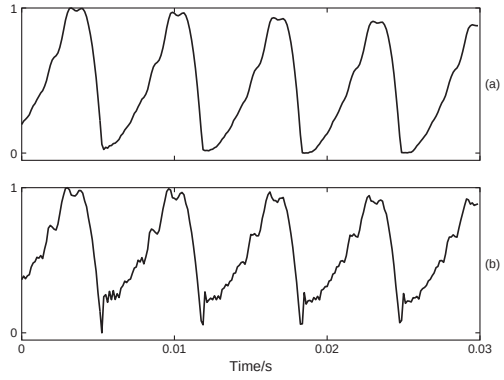


Figure 1. Examples of glottal flows estimated by the conventional CP analysis from a vowel produced by a male speaker. The beginning of the covariance frame in panel (a) is located three samples later than in panel (b).

In order to reduce artefacts caused by false inverse filtering roots, previous CP methods typically exploit techniques to improve the extraction of the closed phase. In addition, a standard procedure to reduce the effect caused by 'jags' is to check whether the vocal tract model computed by LP has any roots on the positive real axis in the z domain, and, if there are such roots, they are simply removed before conducting the inverse filtering (e.g. (6)). In the present work, however, a different approach is studied based on setting mathematical constraints in the computation of the inverse model of the vocal tract with LP. These constraints impose predefined values for the gain of the inverse filter either at zero frequency, that is at DC (direct current), or at the half of the sampling frequency, that is at angular frequency π. This idea results in vocal tract filters whose transfer functions, in comparison to those defined by the conventional covariance analysis, are less prone to comprise disturbing poles in the z domain. From the two possible constraints, the current study reports on experiments where the DC-constrained LP was used to compute the vocal tract filter in the estimation of the glottal source with the CP analysis.

Methods

The conventional CP analysis involves modelling the vocal tract with an all-pole filter defined according to the classical LP based on the covariance criterion (29). The filter coefficients of the pth order inverse filter are searched for by using a straightforward optimization, where the energy of the prediction error is minimized over the covariance frame. In principle, this kind of optimization based on the MSE criterion treats all the frequencies equally, and the filter coefficients are mathematically adjusted so that the resulting all-pole spectrum matches accurately the high-energy formant regions of the speech spectrum. However, it is worth emphasizing that the conventional covariance analysis does not pose any additional information to be used in the optimization process, for example, to bias the location of roots of the resulting all-pole filter. This inherent feature of the conventional covariance analysis implies that roots of the resulting all-pole model of the vocal tract might be located in such a position in the z domain (e.g. on the positive real axis) that is correct from the point of view of the MSE-based error criterion but not optimal from the point of view of glottal inverse filtering.

The computation of the conventional covariance analysis, however, can be modified by using the concept of *constrained* LP. Intuitively, this means that instead of allowing the linear predictive model to locate its roots freely in the z domain based solely on the MSE criterion, the optimization is given certain restrictions in the predictor structure which, then, results in more realistic root locations. In order to implement restrictions, one has to first find a method to express the constraint in a form of a mathematical equation and then to use the selected equation in the minimization problem. Two straightforward constraints can be postulated by assigning a predefined value for the gain of the linear predictive inverse filter either at zero frequency ($\varpi = 0$, where ϖ denotes angular frequency), at half the sampling frequency ($\varpi = \pi$), or at both. With these predefined gains, the optimal constrained linear predictive inverse filter can be optimized based on the mathematical derivations described below.

Let us start by presenting the known optimization of the conventional LP. In the conventional LP, the error signal, or the residual, can be expressed in matrix form as follows:

$$e_n = x_n + \sum_{k=1}^{p} a_k x_{n-k} = \sum_{k=0}^{p} a_k x_{n-k} = \mathbf{a}^{\mathrm{T}} \mathbf{x}_n, \qquad (1)$$

where $\mathbf{a} = [a_0 \ldots a_p]^{\mathrm{T}}$ with $a_0 = 1$, and the signal vector is $\mathbf{x}_n = [x_n \cdots x_{n-p}]^{\mathrm{T}}$. The coefficient vector $\mathbf{a}$ is optimized according to the MSE criterion by searching for such parameters that minimize the square of the residual. In the covariance method, this minimization of the residual energy is computed over a finite time-span (29). By denoting this time-span with $0 \leq n \leq N\text{-}1$, the prediction error energy $E(\mathbf{a})$ can be written as

$$E(\mathbf{a}) = \sum_{n=0}^{N-1} e_n^2 = \sum_{n=0}^{N-1} \mathbf{a}^{\mathrm{T}} \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \mathbf{a} = \mathbf{a}^{\mathrm{T}} \left[\sum_{n=0}^{N-1} \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \right] \mathbf{a}$$
$$= \mathbf{a}^{\mathrm{T}} \boldsymbol{\Phi} \mathbf{a}, \qquad (2)$$

where matrix $\boldsymbol{\Phi}$ is the covariance matrix defined from speech samples as

$$\boldsymbol{\Phi} = \sum_{n=0}^{N-1} \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \in R^{(p+1)\times(p+1)}. \qquad (3)$$

The optimal filter coefficients can be computed easily by minimizing the prediction error energy $E(\mathbf{a})$ with respect to the coefficient vector $\mathbf{a}$. This yields $\mathbf{a} = \sigma^2 \boldsymbol{\Phi}^{-1} \mathbf{u}$, where $\sigma^2 = (\mathbf{u}^{\mathrm{T}} \boldsymbol{\Phi}^{-1} \mathbf{u})^{-1}$ is the residual energy given by the optimized predictor and $\mathbf{u} = [1\ 0 \cdots 0]^{\mathrm{T}}$.

The conventional LP can be modified by imposing constraints on the minimization problem presented above. A mathematically straightforward way to define one such constraint is to set a certain

predefined value for the frequency response of the linear predictive inverse filter at zero frequency. By denoting the transfer function of a pth order *DC-constrained* inverse filter $C(z)$ the following equation can be written:

$$C(z) = \sum_{k=0}^{p} c_k z^{-k} \Rightarrow C(e^{j0}) = C(1) = \sum_{k=0}^{p} c_k = l_{DC}, \quad (4)$$

where c_k, $0 \leq k \leq p$, are the filter coefficients of the constrained inverse filter and l_{DC} is a predefined real value for the gain of the filter at DC. Using matrix notation, the DC-constrained minimization problem can now be formulated as follows: minimize $\mathbf{c}^T\mathbf{\Phi}\mathbf{c}$ subject to $\mathbf{\Gamma}^T\mathbf{c} = \mathbf{b}$, where $\mathbf{c} = [c_0 \cdots c_p]^T$ is the filter coefficient vector with $c_0 = 1$, $\mathbf{b} = [1\ l_{DC}]^T$, and $\mathbf{\Gamma}$ is a $(p+1) \times 2$ constraint matrix defined as

$$\mathbf{\Gamma} = \begin{bmatrix} 1 & 0 & . & . & . & 0 \\ 1 & 1 & . & . & . & 1 \end{bmatrix}^T \quad (5)$$

Similarly, a constraint can be assigned to the frequency response of the linear predictive inverse filter at angular frequency $\varpi = \pi$. By denoting the transfer function of a pth order *π-constrained* inverse filter $D(z)$ the following equation can be written:

$$D(z) = \sum_{k=0}^{p} d_k z^{-k} \Rightarrow D(e^{j\pi}) = D(-1) = \sum_{k=0}^{p} d_k (-1)^k = l_\pi. \quad (6)$$

The π-constrained minimization problem can now be formulated: minimize $\mathbf{d}^T\mathbf{\Phi}\mathbf{d}$ subject to $\mathbf{\Omega}^T\mathbf{d} = \mathbf{e}$, where $\mathbf{d} = [d_0 \cdots d_p]^T$ is the filter coefficient vector with $d_0 = 1$, $\mathbf{e} = [1\ l_\pi]^T$, and $\mathbf{\Omega}$ is a $(p+1) \times 2$ constraint matrix defined as

$$\mathbf{\Omega} = \begin{bmatrix} 1 & 0 & 0 & 0 & . & . & . & 0 \\ 1 & -1 & 1 & -1 & . & . & . & 1 \end{bmatrix}^T. \quad (7)$$

It is also possible to assign a third constraint by imposing simultaneously that the first inverse filter coefficient is equal to unity and that the filter gain at both $\varpi = 0$ and $\varpi = \pi$ are equal to l_{DC} and l_π, respectively. For the sake of brevity, the optimization is expressed in the following only for the DC-constrained LP. The remaining two constraints, however, can be derived using a similar approach.

The covariance matrix defined in Equation 3 is positive definite. Therefore, the quadratic function to be minimized in the DC-constrained problem is convex. Thus, in order to solve the minimization problem, the Lagrange multiplier method (30) can be used. This procedure begins with the definition of a new objective function

$$\eta(\mathbf{c}, \mathbf{g}) = \mathbf{c}^T\mathbf{\Phi}\mathbf{c} - 2\mathbf{g}^T(\mathbf{\Gamma}^T\mathbf{c} - \mathbf{b}), \quad (8)$$

where $\mathbf{g} = [g_1\ g_2]^T > \mathbf{0}$ is the Lagrange multiplier vector. The objective function of Equation 8 can be minimized by setting its derivative with respect to vector $\mathbf{c}$ to zero. By taking into account that matrix $\mathbf{\Phi}$ is symmetric (i.e. $\mathbf{\Phi} = \mathbf{\Phi}^T$) this results in the following equation:

$$\nabla_c \eta(\mathbf{c}, \mathbf{g}) = \mathbf{c}^T(\mathbf{\Phi}^T + \mathbf{\Phi}) - 2\mathbf{g}^T\mathbf{\Gamma}^T = 2\mathbf{c}^T\mathbf{\Phi} - 2\mathbf{g}^T\mathbf{\Gamma}^T = 2(\mathbf{\Phi}\mathbf{c} - \mathbf{\Gamma}\mathbf{g})^T = 0 \quad (9)$$

By combining Equation 9 with the equation of the constraint (i.e. $\mathbf{\Gamma}^T\mathbf{c} - \mathbf{b} = 0$), vector $\mathbf{c}$ can be solved from the group of equations

$$\begin{cases} \mathbf{\Phi}\mathbf{c} - \mathbf{\Gamma}\mathbf{g} = 0 \\ \mathbf{\Gamma}^T\mathbf{c} - \mathbf{b} = 0. \end{cases} \quad (10)$$

From above, the first equation yields $\mathbf{c} = \mathbf{\Phi}^{-1}\mathbf{\Gamma}\mathbf{g}$. Inserting this into the second equation in Equation 10 gives $\mathbf{\Gamma}^T\mathbf{c} = \mathbf{\Gamma}^T\mathbf{\Phi}^{-1}\mathbf{\Gamma}\mathbf{g} = \mathbf{b}$. By solving vector $\mathbf{g}$ from this equation and by inserting it into the first equation in Equation 10, one obtains the constrained inverse filter:

$$\mathbf{c} = \mathbf{\Phi}^{-1}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Phi}^{-1}\mathbf{\Gamma})^{-1}\mathbf{b}, \quad (11)$$

In summary, the optimal DC-constrained inverse filter, the FIR filter of order p given in Equation 4, is obtained by solving for the vector $\mathbf{c}$ according to Equation 11, in which the covariance matrix $\mathbf{\Phi}$ is defined by Equation 3 from the speech signal x_n, matrix $\mathbf{\Gamma}$ is defined by Equation 5, and matrix $\mathbf{b} = [1\ l_{DC}]^T$, where l_{DC} is the desired inverse filter gain at DC.

An example describing spectral models obtained by the proposed technique is depicted in Figure 2. The figure shows three all-pole models of order $p = 10$ obtained for a vowel /a/ produced a male speaker. The analyses were conducted pitch-asyn-

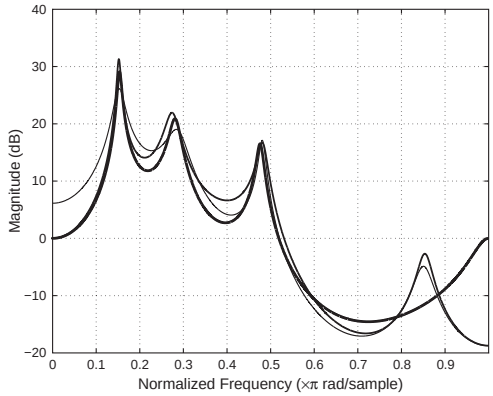


Figure 2. Examples of all-pole spectra of order $p = 10$ computed by linear predictive analyses with the covariance criterion: conventional LP (thin line), DC-constrained LP with $l_{DC} = 1.0$ (line with medium thickness), and constrained LP in which filter gains at $\varpi = 0$ and $\varpi = \pi$ are assigned simultaneously to $l_{DC} = l_\pi = 1.0$.

chronously using the covariance criterion with a frame length of 20 ms. The speech signal was pre-emphasized with a first-order FIR filter (zero at $z = 0.95$) prior to the linear predictive analyses. The spectra shown correspond to the conventional LP (thin line), the DC-constrained analysis with $l_{DC} = 1.0$ (line with medium thickness), and the analysis in which the inverse filter gain at $\varpi = 0$ was assigned to $l_{DC} = 1.0$ and the gain at $\varpi = \pi$ was assigned to $l_{\pi} = 1.0$ (thick line). It can be observed from the figure how filter gains at $\varpi = 0$ in both of the constrained cases are equal to 0 dB (that is, log(1.0)), and this level is shown by the thick line also at $\varpi = \pi$.

A recent study has demonstrated that the estimation of the glottal flow is distorted if the inverse model of the vocal tract has roots outside the unit circle, that is the FIR filter is non-minimum phase (31). In order to eliminate the occurrence of non-minimum phase filters, the roots of the inverse filter need to be solved, and if the filter is not minimum phase, those roots that are located outside the unit circle are replaced by their mirror image partners inside the circle. In principle, it is possible that the constrained LP computed according to Equation 11 yields an inverse filter that has roots on the positive real axis. Due to the use of the DC constraint, the risk for this to happen is, however, smaller than in the case of the conventional covariance analysis. Because the roots of $C(z)$ are solved for in order to eliminate the occurrence of non-minimum phase filters, it is trivial also to check simultaneously whether there are any roots on the positive real axis inside the unit circle. If so, these roots are simply removed in a procedure similar to that used in the conventional CP analysis (6).

In summary, the estimation of the glottal flow utilizing the concept of constrained LP comprises the following stages:

1. The position of the covariance frame is computed using any of the previously developed methods based on, for example, the maximum determinant of the covariance matrix (6) or the EGG (25).
2. The vocal tract transfer function $C(z)$ is computed according to Equation 11 by defining the elements of the covariance matrix in Equation 3 from the speech signal to be inverse-filtered.
3. The roots of $C(z)$ defined in stage 2 are solved. Those roots of $C(z)$ that are located outside the unit circle are replaced by their corresponding mirror image partner inside the unit circle. Any real roots located on the positive real axis are removed.
4. Finally, the estimate of the glottal volume velocity waveform is obtained by filtering the speech signal through $C(z)$ defined in stage 3 and by cancelling the lip radiation effect with a first order Infinite Impulse Response (IIR) filter with its pole close to the unit circle (e.g. at $z = 0.99$).

Materials and experiments

The performance of the proposed CP method was studied in experiments by using both synthetic and natural vowels. In both cases, the speech material comprised different representations of the vowel /a/. This vowel was selected due to its high first formant, which makes the linear predictive estimation of the vocal tract resonances more accurate in comparison to vowels whose first formant locates lower in frequency. In the following, the characteristics of the materials are explained separately for synthetic and natural utterances used in the experiments.

Synthetic speech

Synthetic glottal sources were created by utilizing the LF model (11). Glottal flow derivatives corresponding to three different phonation types (breathy, modal, and creaky) were synthesized by using the LF parameters published by Gobl (32). The fundamental frequency of the source waveforms was varied in order to create synthetic male and female vowels. F0 values of male voices (89, 100, and 92 Hz for breathy, modal, and creaky sources, respectively) were taken directly from the data provided by Gobl's study. In female vowels, the same LF parameters were used as in male sounds, but the F0 was two times larger. The vocal tract was modelled with 12th order all-pole filters. The coefficients of these filters were adjusted so as to create the four lowest formants of the vowel /a/. The centre frequencies and bandwidths of the formants were adjusted separately for male and female sounds by utilizing data from the study by Peterson and Barney (33).

Natural speech and EGG recordings

In the analysis of natural speech, glottal inverse filtering studies mostly use isolated vowels produced with sustained phonation. In the current study, however, a more realistic yet a more challenging approach was utilized by estimating glottal flows from continuous speech. The material was obtained by asking speakers to read a Finnish text consisting of three passages describing weather conditions. The text was repeated three times by each speaker. Each recitation took approximately one minute. The

middle recitation was chosen for further analyses. The weather forecast text was specifically tailored to comprise several words with a long vowel /a/ surrounded either by fricatives /s/ or unvoiced plosives /k/, /p/, and /t/ in order to obtain non-nasalized vowels of high first formant for inverse filtering analysis. From these words, the one starting the second passage, the word *Kaatosade* (Finnish for 'torrential rain'), was selected, and the long /a/ in the first syllable of this word was used in the inverse filtering analyses.

All subjects (six males, six females, age range from 18 to 48 years) were native speakers of Finnish with no history of any known speech or hearing deficit. Recordings took place in an anechoic chamber using a headset microphone (unidirectional Sennheiser electret capsule) and electroglottography (Glottal Enterprises two-channel EGG, model EG-4). Data were saved into a digital audio recorder (iRiver iHP-140). The distance between the speaker's lips and microphone was 12 cm.

Speech and EGG waveforms were transferred from the audio player onto a computer to be analysed. The signal was down-sampled to 8 kHz sampling rate, and both the voice sounds and the EGG waveforms were high-pass filtered with a linear phase FIR with a cut-off frequency of 70 Hz in order to remove any low-frequency noise picked up during the recordings. The propagation delay of the speech signals was compensated by using the lip-to-microphone distance of 12 cm, vocal tract length of 17 and 15 cm for males and females, respectively, and 350 m/s as the speed of sound. This resulted in a delay of 7 and 6 samples for males and females, respectively.

Experiments and parameter settings

The main goal of the experimental part of this study was to evaluate how the proposed new DC-constrained CP method behaves in comparison to the conventional CP technique when the covariance frame position is varied. In the conventional CP technique, the vocal tract roots on the positive real axis in the z-domain were removed prior to inverse filtering as suggested by Wong et al. (6). In order to compare the two methods, two experiments were conducted as follows.

Experiment 1

In the first experiment, synthetic vowels were inverse-filtered by both CP techniques, and the obtained glottal flow estimates were compared with the corresponding LF waveforms used in the sound synthesis. The instant of glottal closure was detected as the time when the amplitude of the LF waveform reduced below 2% of its peak value, denoted by EE in the LF terminology (11). Altogether nine covariance frame positions were tested for each synthetic vowel around this instant of glottal closure: the position that started at the detected glottal closure instant; four frame positions that started 1, 2, 3, and 4 samples before the instant of closure; and another four positions that started 1, 2, 3, and 4 samples after the closure instant. All in all, this first experiment resulted in 108 estimated glottal flows (2 CP methods, 2 speakers, 3 phonation types, 9 covariance frame positions per cycle).

All the obtained glottal flow estimates were quantified numerically by using a frequency domain glottal flow parameter, harmonic richness factor (HRF), which is defined from the spectrum of the glottal flow as the ratio, in dB, between the sum of the harmonic amplitudes above the fundamental and the amplitude of the fundamental (17). HRF was selected because it can be computed automatically without any subjective user adjustments.

Experiment 2

In the second experiment, natural speech and EGG were used to evaluate the effects of the varying covariance frame position on the estimation of the glottal flow with the two CP methods. The glottal closure was detected at the time instant when the differentiated EGG waveform reached its maximum negative peak (9). Again, nine different covariance frame positions were selected around this instant: the frame position that started at the instant of the negative peak; four positions that started 1, 2, 3, and 4 instants before the peak; and another four that started 1, 2, 3, and 4 time instants after the peak of the differentiated EGG wave. The analysis was repeated for four consecutive glottal cycles. All in all, this second experiment resulted in 864 estimated glottal flows (2 CP methods, 12 speakers, 4 glottal cycles, 9 covariance frame positions per cycle). All the estimated glottal flows were parameterized with HRF.

Parameter settings

All the inverse filtering analyses were conducted by using the order of the vocal tract filter, that is the prediction order of the covariance analysis $P=12$. Analyses with synthetic speech used a fixed length of 30 samples (3.75 ms) for the duration of the covariance frame. For natural utterances, the length of the covariance frame was set to the value that equalled the time distance between the negative and positive peak of the differentiated EGG (9).

However, if this distance was smaller than twice the order of LP (i.e. smaller than 24 samples), a fixed duration of 24 samples was used. The length of the inverse filtering analysis window was eight cycles. The value of the DC constraint parameter was adjusted by computing the DC gain of the synthetic vocal tract models of males and females and by rounding this value to the nearest integer. This yielded l_{DC} values equal to 2.0 and 4.0, which were used in the analysis of male and female speakers, respectively.

Results

Examples of estimated glottal flow waveforms of one male speaker are shown in Figures 3 and 4 when inverse filtering was conducted with the conventional and the proposed method, respectively. Similarly, Figures 5 and 6 show results obtained for one female speaker. The waveforms indicate how varying the beginning of the covariance frame around the negative peak of the differentiated EGG affects the time-domain waveforms of the glottal flow. When the frame is misaligned, the waveforms given by the conventional CP analysis clearly indicate sharp edges especially in the vicinity of the glottal closure (e.g. Figures 3b, 5a, 5c, 5d). One might argue that this kind of fluctuation can hardly be created by the vocal folds and therefore rapid changes in the waveform are most likely a consequence of the incorrect estimation of the antiresonances of the vocal tract by the conventional LP analysis. The use of the proposed DC constraint, however, results in glottal flow estimates for the same two speakers that are clearly less dependent on the position of the covariance frame.

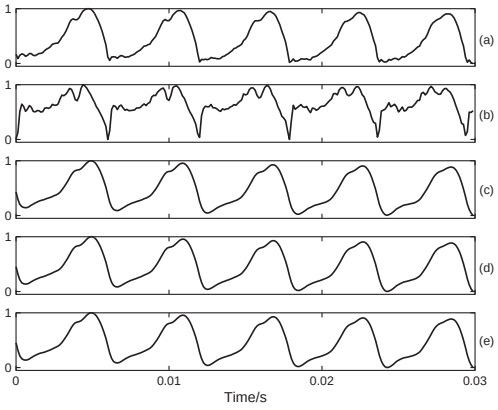


Figure 3. Examples of glottal flows estimated by the conventional CP analysis by varying the covariance frame position (male speaker). The beginning of the covariance frame was located at the instant of the negative peak of the differentiated EGG (electroglottography) in panel (c). The position of the frame was decremented by one and two time instants in panels (b) and (a), respectively. The position of the frame was incremented by one and two time instants in panels (d) and (e), respectively.

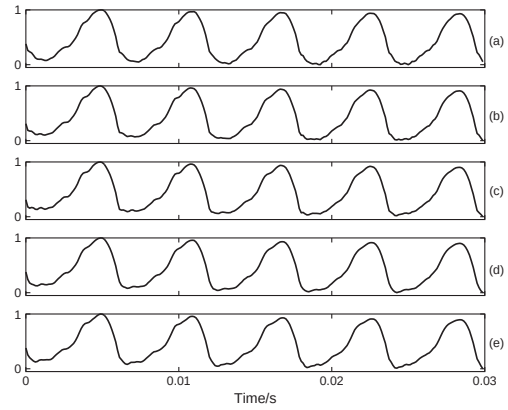


Figure 4. Examples of glottal flows estimated by the proposed new CP analysis by varying the covariance frame position (male speaker). The beginning of the covariance frame was located at the instant of the negative peak of the differentiated EGG (electroglottography) in panel (c). The position of the frame was decremented by one and two time instants in panels (b) and (a), respectively. The position of the frame was incremented by one and two time instants in panels (d) and (e), respectively.

Differences in the all-pole models of the vocal tract computed by the conventional CP analysis and by the DC-constrained analysis are demonstrated in Figure 7, which presents the spectra of the vocal tract models given by the conventional (thin line) and DC-constrained method (thick line) in the analyses whose resulting time-domain glottal flow estimates are shown in panels (b) in Figure 3 and 4. It is worth noting in Figure 7 how the vocal tract spectrum obtained by the conventional LP shows excessive boost at low frequencies whereas the level of the DC-constrained spectrum is clearly lower. The strong low-frequency components in the all-pole spectrum of the conventional LP analysis are explained by false roots, and they, in turn, have caused the severe distortion which is visible in the corresponding time-domain waveform of the glottal flow shown in Figure 3b. While the two all-pole spectra are relatively similar at higher frequencies, one can easily see how the amplified low-frequency components of the conventional LP spectrum have also 'pushed' the first formant peak higher in frequency in comparison to that in the DC-constrained spectrum.

The results of experiment 1 are given in Table I. These data were computed as follows: The HRF

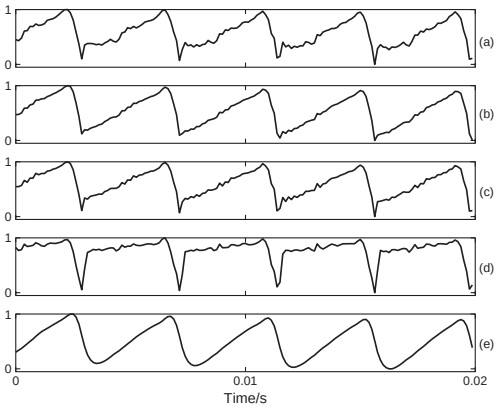


Figure 5. Examples of glottal flows estimated by the conventional CP analysis by varying the covariance frame position (female speaker). The beginning of the covariance frame was located at the instant of the negative peak of the differentiated EGG (electroglottography) in panel (c). The position of the frame was decremented by one and two time instants in panels (b) and (a), respectively. The position of the frame was incremented by one and two time instants in panels (d) and (e), respectively.

value of each glottal flow estimated by inverse filtering was subtracted from the HRF value computed from the corresponding original LF waveform. The absolute value of this difference was computed for each of the nine covariance frame positions and, finally, averaged over all nine positions. These data,

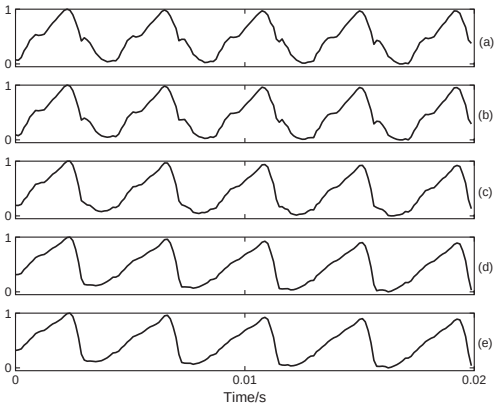


Figure 6. Examples of glottal flows estimated by the proposed new CP analysis by varying the covariance frame position (female speaker). The beginning of the covariance frame was located at the instant of the negative peak of the differentiated EGG (electroglottography) in panel (c). The position of the frame was decremented by one and two time instants in panels (b) and (a), respectively. The position of the frame was incremented by one and two time instants in panels (d) and (e), respectively.

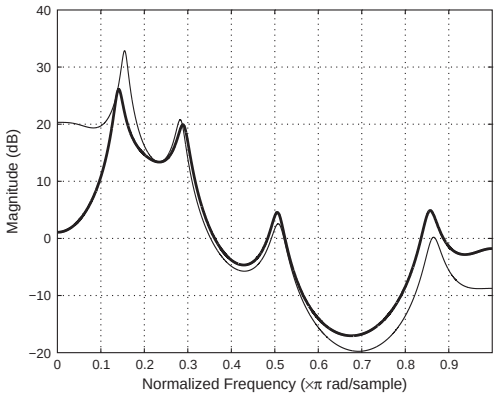


Figure 7. Examples of all-pole spectra computed in the closed-phase covariance analysis by the conventional LP (thin line) and by the DC-constrained LP (thick line). Spectra shown were taken from the analyses that resulted in the glottal flow estimates that are shown in Figure 3(b) and Figure 4(b).

as given in Table I, show that the glottal flow estimates computed from the synthetic speech by the proposed method were closer to the original LF waveforms than those obtained by the conventional CP analysis in all cases. For both inverse filtering methods, the error was larger in vowels synthesized with LF waveforms modelling female voice production than in those mimicking male voice production. This difference is explained by the shorter length of the closed phase in LF pulses of female speech; for glottal pulses of a short closed phase the underlying assumption of the closed-phase analysis (that is, the existence of a sufficiently long phase with zero flow) is violated more than for pulses of a longer closed phase, and, consequently, the absolute error becomes larger for high-pitch vowels when the frame position is varied.

The results of experiment 2 are given in Table II. These data were obtained by simply computing the standard deviation of the HRF values over all the different covariance frame positions (four cycles and nine positions per cycle). Again, the data show that the use of the constrained linear prediction has resulted in a clearly smaller fluctuation of the glottal flow parameters when the covariance frame position is varied: the standard deviation is smaller for the proposed CP technique in all the speakers analysed except for one male subject.

Conclusions

Closed-phase (CP) covariance analysis is a widely used glottal inverse filtering method. It is based on the extraction of the vocal tract model with the

Table I. Mean (in dB) of absolute error in the harmonic richness factor between the original synthetic glottal flow generated by the LF (Liljencrants-Fant) waveform and the estimated glottal flow given by the conventional closed-phase covariance analysis (CP_{con}) and by the closed-phase analysis utilizing the DC-constrained linear prediction (CP_{DC}). The analysis was conducted by incrementing the beginning of the covariance frame nine times in the vicinity of the glottal closure instant.

Gender	Phonation type	CP_{con}	CP_{DC}
Male	Breathy	10.65	0.09
	Modal	1.02	0.45
	Creaky	1.41	0.63
Female	Breathy	24.61	13.59
	Modal	11.95	5.48
	Creaky	9.57	4.83

conventional linear prediction using the covariance criterion during the time of the closed phase of the vocal fold cycle. Since the length of the closed phase is typically short, the resulting vocal tract model is highly vulnerable with respect to the extraction of the frame position. This undesirable feature of the conventional CP analysis typically results in vocal tract models which have roots, both real and complex, at low frequencies. These kinds of false root locations, in turn, result in distortion of the glottal flow estimates, which is typically seen as unnatural peaks at the instant of glottal closure, the so-called 'jags', or as increased formant ripple during the closed phase.

The present investigation studied an improved version of the CP analysis based on the concept of the constrained linear prediction. In this idea, predefined constraints can be imposed on the gain of the linear predictive inverse filter at DC or at angular frequency π. With these constraints, linear predictive analysis is more prone to give vocal tract models whose roots are located in the formant region rather than in unrealistic positions at low frequency. By using both synthetic vowels as well as natural utterances recorded from continuous speech, the present study showed that the use of the DC-constrained linear prediction resulted in closed-phase inverse filtering analysis that was less sensitive to the location of the covariance frame position than the conventional CP technique. In the present study, the value of the DC gain was adjusted separately for vowels produced by male and female subjects using data provided by synthetic vocal tract models. In our previous study, we used a fixed DC gain value for both genders (31). A possible topic of further studies would be to study how varying the value of the constraint inside a given range changes the results of the closed-phase analysis.

Table II. Standard deviation (in dB) of the harmonic richness factor given by the conventional closed-phase covariance analysis (CP_{con}) and by the closed-phase analysis utilizing the DC-constrained linear prediction (CP_{DC}). The analysis was conducted by incrementing the beginning of the covariance frame nine times in the vicinity of the glottal closure instant and by repeating the analysis for four consecutive fundamental periods.

Speaker	CP_{con}	CP_{DC}
Male 1	3.59	1.73
Male 2	5.64	1.08
Male 3	7.75	1.77
Male 4	2.13	0.72
Male 5	0.32	1.53
Male 6	2.21	1.12
Female 1	3.06	2.07
Female 2	13.15	1.81
Female 3	2.53	0.53
Female 4	6.24	1.29
Female 5	9.12	1.78
Female 6	4.41	2.66

Acknowledgements

This work was supported by the Academy of Finland (project no. 111848 and 127345) and by the COST action 2103, 'Advanced Voice Function Assessment'. After completing research related to the current article, Tom Bäckström has started working at The Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

1. Miller R. Nature of the vocal cord wave. J Acoust Soc Am. 1959;31:667–77.
2. Lindqvist-Gauffin J. Inverse filtering. Instrumentation and techniques. STL-QPSR. 1964;5:1–4.
3. Lindqvist-Gauffin J. Studies of the voice source by means of inverse filtering. STL-QPSR. 1965;6:8–13.
4. Gauffin J, Sundberg J. Spectral correlates of glottal voice source waveform characteristics. J Speech Hear Res. 1989;32: 556–65.
5. Strube H. Determination of the instant of glottal closure from the speech wave. J Acoust Soc Am. 1974;56:1625–9.
6. Wong D, Markel J, Gray A. Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Trans Acoust Speech Signal Process. 1979;27:350–5.
7. Fant G. Acoustic theory of speech production. The Hague: Mouton; 1970.
8. Price P. Male and female voice source characteristics: Inverse filtering results. Speech Commun. 1989;8:261–77.
9. Childers D, Ahn C. Modeling the glottal volume-velocity waveform for three voice types. J Acoust Soc Am. 1995;97: 505–19.
10. Strik H, Boves L. On the relation between voice source parameters and prosodic features in connected speech. Speech Commun. 1992;11:167–74.

11. Fant G, Liljencrants J, Lin Q. A four-parameter model of glottal flow. STL-QPSR. 1985;4:1–13.
12. Cummings K, Clements M. Analysis of the glottal excitation of emotionally styled and stressed speech. J Acoust Soc Am. 1995;98:88–98.
13. Berouti M, Childers D, Paige A. Glottal area versus glottal volume-velocity. In: IEEE International Conference on Acoustics, Speech and Signal Processing, May 9–11, 1977, Hartford, CT, USA. p. 33–6.
14. Deller J. Evaluation of laryngeal dysfunction based on features of an accurate estimate of the glottal waveform. In: IEEE International Conference on Acoustics, Speech and Signal Processing, May 3–5, 1982, Paris, France. p. 759–62.
15. Veeneman D, BeMent S. Automatic glottal inverse filtering from speech and electroglottographic signals. IEEE Trans Acoust Speech Signal Process. 1985;33:369–77.
16. Pinto N, Childers D, Lalwani A. Formant speech synthesis: Improving production quality. IEEE Trans Acoust Speech Signal Process. 1989;37:1870–87.
17. Childers D, Lee C. Vocal quality factors: Analysis, synthesis, and perception. J Acoust Soc Am. 1991;90:2394–410.
18. Plumpe M, Quatieri T, Reynolds D. Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Trans Speech Audio Process. 1999;7: 569–86.
19. Gudnason J, Brookes M. Voice source cepstrum coefficients for speaker identification. In: IEEE International Conference on Acoustics, Speech and Signal Processing, March 30–April 4, 2008, Las Vegas, Nevada, USA. p. 4821–4.
20. Arroabarren I, Carlosena A. Effect of the glottal source and the vocal tract on the partials amplitude of vibrato in male voices. J Acoust Soc Am. 2006;119:2483–97.
21. Akande O, Murphy P. Estimation of the vocal tract transfer function with application to glottal wave analysis. Speech Commun. 2005;46:15–36.
22. Ananthapadmanabha T, Yegnanarayana B. Epoch extraction from linear prediction residual for identification of closed glottis interval. IEEE Trans Acoust Speech Signal Process. 1979;27:309–19.
23. Smits R, Yegnanarayana B. Determination of instants of significant excitation in speech using group delay function. IEEE Trans Speech Audio Process. 1995;3:325–33.
24. Naylor P, Kounoudes A, Gudnason J, Brookes M. Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. IEEE Trans Speech Audio Process. 2007;15: 34–43.
25. Krishnamurthy A, Childers D. Two-channel speech analysis. IEEE Trans Acoust Speech Signal Process. 1986;34:730–43.
26. Larar J, Alsaka Y, Childers D. Variability in closed phase analysis of speech. In: IEEE International Conference on Acoustics, Speech and Signal Processing, March 26–29, 1985, Tampa, Florida, USA. p. 1089–92.
27. Yegnanarayana B, Veldhuis N. Extraction of vocal-tract system characteristics from speech signals. IEEE Trans Speech Audio Process. 1998;6:313–27.
28. Riegelsberger E, Krishnamurthy A. Glottal source estimation: Methods of applying the LF-model to inverse filtering. In: IEEE International Conference on Acoustics, Speech and Signal Processing, April 27–30, 1993, Minneapolis, Minnesota, USA. p. 542–5.
29. Rabiner L, Schafer R. Digital processing of speech signals. Englewood Cliffs: Prentice-Hall; 1978.
30. Bazaraa M, Sherali H, Shetty C. Nonlinear programming: Theory and algorithms. New York: Wiley; 1993.
31. Alku P, Magi C, Yrittiaho S, Bäckström T, Story B. Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. J Acoust Soc Am. In press.
32. Gobl C. A preliminary study of acoustic voice quality correlates. STL-QPSR. 1989;4:9–21.
33. Peterson G, Barney H. Control methods used in a study of the vowels. J Acoust Soc Am. 1952;24:175–84.

HELSINKI UNIVERSITY OF TECHNOLOGY
DEPARTMENT OF SIGNAL PROCESSING AND ACOUSTICS
REPORT SERIES

ISBN 978-951-22-9963-8
ISSN 1797-4267