Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 893–902

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Fast and robust formant detection from LP data

Thorsten Smit *, Friedrich Türckheim, Robert Mores [1]

University of Applied Sciences Hamburg, Finkenau 35, 20081 Hamburg, Germany

Received 5 August 2010; received in revised form 13 September 2011; accepted 5 March 2012
Available online 17 March 2012

Abstract

This paper introduces a method for real-time selective root finding from linear prediction (LP) coefficients using a combination of spectral peak picking and complex contour integration (CI). The proposed method locates roots within predefined areas of the complex z-plane, for instance roots which correspond to formants while other roots are ignored. It includes an approach to limit the search area (SEA) as much as possible. For this purpose, peaks of the group delay function (GDF) serve as pointers. A frequency weighted wGDF will be introduced in which a simple modification enables a parametric emphasis of the GDF spikes to separate merged formants. Thus, a nearly zero defected separation of peaks is possible even when these are very closely spaced. The performance and efficiency of the proposed wGDF-CI method is demonstrated by comparative error-analysis evaluated on a subset of the DARPA TIMIT corpus.

Keywords: Speech recognition; Formant tracking; Speech analysis; Male and female voices

1. Introduction

In speech processing the formant structure is the entry point for many kinds of analyses and applications. Formants are resonances of the human vocal tract and correspond to resonances within the spectral shape of speech signals (Markel and Gray, 1976). So far, this correspondence was used for numerous speech processing applications such as speaker recognition (Snell et al., 1983) or forensic analyses (Kuwarabara and Sagisaka, 1995).

Numerous methods have been proposed to automatically determine formants from speech signals. For real time speech applications such proposals often aim at relaxing the trade-off between accuracy and computational cost.

For instance, the inverse filter control (Welling and Ney, 1998) or the iterative energy separation algorithm (Hanson et al., 1994) are rarely used due to their computational complexity (Ueda et al., 2007). Furthermore, it is not possible to extract 3-dB bandwidths of formants by means of an iterative energy separation algorithm.

More traditional formant extraction methods can be roughly divided into spectral peak picking (SPP) (Schafer and Rabiner, 1970) and root finding (RF) approaches (Atal and Hanauer, 1971).

SPP procedures locate formants at peaks of cepstrally smoothed or linearly predicted (LP) spectra. The main problem of these techniques is the unwanted extraction of merged and spurious peaks Kim et al., 2006. This problem cannot be overcome by just increasing the LP order or by widening the cepstral liftering (McCandless, 1974). Moreover, it is not possible to extract a correct formant bandwidth by just simply relying on the shape of the spectral envelope.

The RF method provides solutions for these restrictions by locating roots of the complex LP polynomial which permit simple transformations between complex locations within the z-plane and their spectral equivalents. Such transformation allows for both, avoiding problems coming along with merged or spurious formants, and providing for

* Corresponding author.
E-mail addresses: thorsten.smit@mt.haw-hamburg.de (T. Smit), friedrich.tuerckheim@haw-hamburg.de (F. Türckheim), mores@mt.haw-hamburg.de (R. Mores).
URL: http://www.mt.haw-hamburg.de/akustik/ (R. Mores).
[1] Principal corresponding author.

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.03.002

the wanted bandwidth extraction. Standard root solvers (SRS) are computationally intensive (Dellar et al., 1999) so that several methods have been proposed to approximate true root locations.

Line spectral pair (LSP) roots on the unit circle, contain properties to frame LP roots Itakura, 1975. An empirical method approximates such LSP roots by means of the logarithmic spectral difference function (LSDF). It has been shown that the turning points of the LSDF correspond to roots of the LSP (Kim and Lee, 1999). Furthermore, LSP roots of different orders are interlaced, so tangential boundaries of LP roots can be narrowed iteratively. The main restriction of this procedure is its high computation effort needed to evaluate the SDFs of successive orders. Additionally, it is not readily possible to estimate formant bandwidths by just knowing the tangential boundaries of corresponding LP roots.

Other RF methods make use of findings from investigations on vowel quality perception (Peterson and Barney, 1951; Pfitzinger, 2005). In these works it has been shown that mainly the first three formants $F1$, $F2$ and $F3$ are reliable indicators for most applications in speech processing. Therefore, in most cases it is sufficient to find only the three corresponding roots. That means, that it is possible to significantly reduce the required root-solving computation time. For this reason, selective root finding techniques have been introduced, as proposed by Snell and Milinazzo (1993), Sandler (1991). These methods avoid redundant root determinations by limiting the root SEA in the complex z-plane. It becomes evident that the performance of selective root finding approaches are largely dependent on well predicted SEAs. For achieving such prediction quality, several SEA limitation strategies have been introduced in the past. An exemplary helpful working approach can be found in (Reddy and Swamy, 1984). Here peaks of several interacting spectral functions, the log-magnitude spectrum, the GDF, the second derivative of the log-magnitude spectrum and the second derivative of the GDF are combined to predict and bound respective SEAs.

This paper proposes the wGDF-CI method to greatly simplify mentioned SEA boundary functions. For this purpose, several spectral functions will be combined into an efficient singular routine while maintaining likewise reliable SEA prediction results. The proposed method utilizes peaks of a weighted group delay function (wGDF) to point at the center of respective SEA in the z-plane. Ambiguities of an interacting function such as the one mentioned in (Reddy and Swamy, 1984) are avoided here due to the fact that only one singular pointing function will be used. Once the SEA are defined, a subsequent CI approach facilitates highly accurate center frequency and bandwidth extraction for each formant by narrowing the predicted SEA.

The structure of this paper is as follows: in Section 2, LP analysis, wGDF, and CI implementations will be discussed. Additionally, a simple example will demonstrate the proposed method step by step. Section 3 includes formant extraction error evaluations on the DARPA TIMIT corpus Garofolo et al., 1993, results, and detailed performance discussions for recent processors. Section 4 will briefly summarize the present study. Finally, a discussion of the more common theory of contour integration and its numerical implementation will be given in Appendices A and B.

2. The proposed formant extraction method

This section shows how to combine spectral peak picking and contour integration for reliable formant extraction. In a first step it will be shown how to avoid merged peaks and how to predict the SEA while using spectral peak picking. Secondly the contour integration will be introduced which narrows the SEA iteratively to find accurate root locations of the LP polynomials.

An overview of the proposed formant extraction method is given in Fig. 1. In the following the processing blocks are separately discussed according to their integer label to encourage using this section as orientation or guide while becoming acquainted with the proposed wGDF-CI method.

2.1. Pre-block 1

Discrete speech signals will be segmented into hamming windowed short term frames $x(n)$ of length N.

A subsequent preemphasis flattens the spectral dynamics of the glottal waveform and the lip radiation as proposed by Flanagan et al. (1964) and later by Schafer and Rabiner (1970). This facilitates improved results for formant matching at higher frequencies. Therefore, a first-order FIR filter are used which are given by

$$H_{pre}(z) = 1 - bz^{-1}, \tag{1}$$

where $b = 0.94$, see Wong et al., 1980.

2.2. LP-block 2

LP analysis allows for following the source-filter theory of speech production Fant, 1960. Its autoregressive (AR) model is given by Schafer and Rabiner (1970)

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k \cdot z^{-k}} = \frac{1}{A(z)}, \tag{2}$$

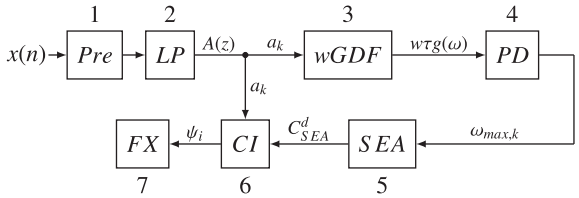

Fig. 1. Overview of the proposed formant extraction method, Pre = short term segmentation and preemphasis of discrete input $x(n)$, PD = peak detection, and FX = output of formants.

where $z = r \cdot \exp(-j\omega/f_s)$ is a complex number with magnitude r and angle ω/f_s. In LP analysis the coefficients a_k of LP order p are computed. Order p depends on the sampling rate f_s and is commonly approximated by

$$p = \text{round}\left\{\frac{f_s}{1000\ \text{Hz}}\right\} + \text{const.} \tag{3}$$

In literature *const* varies between 2 and 4.

With $r = 1$ we follow the unit circle and so Eq. (2) can be simplified to $H(\exp(-j\omega/f_s)) \triangleq H(\omega)$. The power spectrum $|H(\omega)|^2$ contains peaks at formant frequencies (see Fig. 2). Such spectral peaks occur at frequencies where respective poles are crossed while following the unit circle of the complex z-plane. All corresponding peak magnitudes are inversely related to their bandwidth.

2.3. wGDF-block 3

The spectral peak picking will be included in form of a GDF peak picking procedure. First, a brief review of conventional GDF will be given. After that, the GDF of zero padded LP coefficients and, finally, the wGDF will be introduced.

The Fourier transform of the discrete sequence $x(n)$ of length N is given by

$$X(\omega) = |X(\omega)| \exp(j\theta(\omega)), \tag{4}$$

and the GDF is defined as

$$\tau g(\omega) = -\frac{d\theta(\omega)}{d\omega}. \tag{5}$$

Computation of Eq. (5) contains phase wrapping at multiples of $\pm\pi$. These discontinuities can be avoided if the GDF is computed directly from the time signal $x(n)$ in the discrete form Oppenheim and Schafer (1975),

$$\begin{aligned}\tau g(\omega) &= -\text{Imag}\left\{\frac{d(\log X(\omega))}{d\omega}\right\} \\ &= \frac{X_R(\omega)X'_R(\omega) + X_I(\omega)X'_I(\omega)}{|X(\omega)|^2},\end{aligned} \tag{6}$$

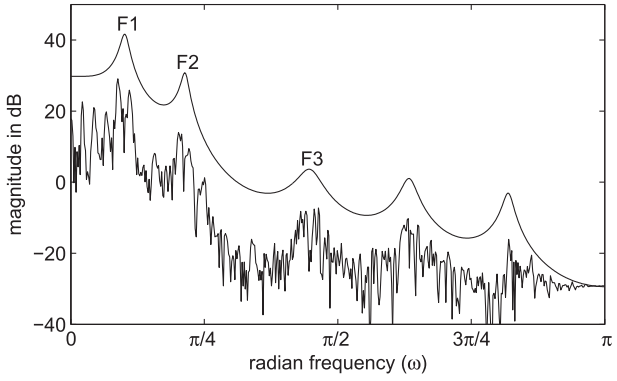

Fig. 2. Bottom: short term spectrum of a male spoken vowel [ɐ], top: LP-derived amplitude spectrum (10 dB offset), the first three formants are labeled.

where $X(\omega) = X_R + jX_I$ is the Fourier transform of $x(n)$ and $X'(\omega) = X'_R + jX'_I$ is its derivate, the Fourier transform of $nx(n)$.

In the same way, the GDF can be computed from LP data. This GDF is appropriate for spectral formant representations since poles close to the unit circle in the z-plane cause peaks within the GDF shape. This is a well known relationship of zero-pole filter design Williams, 1986. Such property can be used when the input signal $x(n)$ will be substituted by $y(n)$ of an arbitrary length M which is given by

$$y(n) = \begin{cases} 1 & \text{for} \quad n = 0, \\ a_n & \text{for} \quad 1 \leqslant n \leqslant p, \\ 0 & \text{for} \quad p < n \leqslant M-1, \end{cases} \tag{7}$$

where a_n are the LP coefficients from Eq. (2) and p is the LP order.

Now, similar to conventional SPP methods, the unweighted GDF of $y(n)$ shows merged peaks of closely spaced resonances. However, the GDF can be used for finding formant regions. This becomes possible through a simple modification of Eq. (6), so that even merged peaks will be separated and be identified as peaks within a weighted GDF shape.

For this purpose and similar to Murthy et al., 1989, the denominator term $|X(\omega)|^2$ of Eq. (6) will be replaced by $|X(\omega)|^{\alpha(l)}$. With the formal notation $X \rightarrow Y$ for the zero padded input signal $y(n)$, Eq. (6) takes the form

$$w\tau g(\omega) = -\frac{Y_R(\omega)Y'_R(\omega) + Y_I(\omega)Y'_I(\omega)}{|Y(\omega)|^{\alpha(l)}}. \tag{8}$$

Due to the fact, that the denominator coefficients of $H(z)$ are used in (8), the group delay function has to be negated.

Lowest formant error rates on the TIMIT corpus were achieved when the arbitrary frequency dependent parameter $\alpha(l)$ follows the linear approach

$$\alpha(l) = \frac{0.4}{M} \cdot l + 0.8, \quad l = 0, \ldots, M/2 - 1. \tag{9}$$

Nonlinear approaches generally show a much higher error value in empirical studies on the TIMIT corpus.

For comparison, extraction results of both the unweighted approach GDF-CI using Eq. (5) and the weighted approach wGDF-CI method using Eq. (8) are shown in Table 3 of Section 3.3. It could be shown, that especially under difficult conditions, for instance nasalized sounds, the error-rate of the wGDF-CI method is often well below the error rate of the unweighted GDF-CI approach.

2.4. PD-block 4

Peak detection in the PD block will deliver predictors for initial SEA. The radian frequencies of all local maxima of length L

$$\omega_{\max,k} = \max_{\omega}\{w\tau g\}, \quad k = 1,\ldots,L \tag{10}$$

indicate potential formant locations and therefore reasonable search areas. Points of local maxima fulfill gradient conditions which can be expressed by

$$\frac{d\tau}{d\omega} = 0 \quad \text{and} \quad \frac{d^2\tau}{d\omega^2} < 0. \tag{11}$$

2.5. SEA-block 5

The transition from spectral indicators $\omega_{\max,k}$ to geometrical SEA within the z-plane is the task of this block (see Fig. 1).

The contour of any single SEA is defined as a particular circular ring, and it is appropriate to divide this specific geometric area into radial components $\omega_{\text{low}}, \omega_{\text{up}}$ and tangential components $r_{\text{inner}}, r_{\text{outer}}$ (see Fig. 3). Such SEA shape facilitates reductions to half of the SEA size for each iteration step to follow.

2.5.1. SEA-radial boundaries

The radial SEA boundaries consist of an inner and an outer radius. First, the stability of LP analysis allows for a direct indication of the constant outer radius r_{outer}. This is because of the fact that all roots of $A(z)$ have to lay within the unit circle, so the outer radius can be defined by

$$r_{\text{outer}} = 1. \tag{12}$$

Furthermore, formants will be formed by roots of $A(z)$ which are close to 1. This fact allows for defining minimum radius ($r_{\text{inner},i}$, see Fig. 3). According to Eq. (23) and Dunn (1961), criteria for the inner radius can be defined by

$$r_{\text{inner},i} = \exp\left(-\frac{B_{\max,i}\cdot\pi}{f_s}\right), \quad i = 1,2,3\ (F1,F2,F3), \tag{13}$$

where $B_{\max,i}$ is the maximum bandwidth of the ith formant.

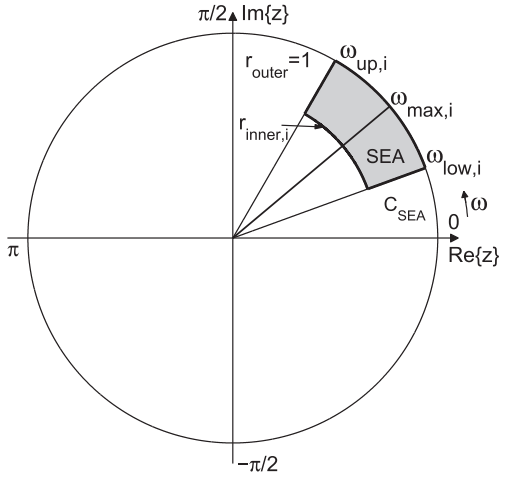

Fig. 3. Shape of the SEA in the z-plane (highlighted in gray), note that boundary dimensions are not scaled here to facilitate ease of reading, for component descriptions see text.

In (Dunn, 1961) Dunn suggested bandwidths of 160, 200, and 300 Hz for $F1$, $F2$ and $F3$, respectively. Kim et al. later proposed a constant bandwidth $B_{\max}$ of 570 Hz for the first three formants, see Kim et al. (2006). In our studies, best results were achieved by defining a fixed maximum bandwidth criterion with $B_{\max} = 750$ Hz.

2.5.2. SEA-tangential boundaries

The tangential components of SEA consist of two parts, a lower and an upper boundary $\omega_{\text{low,k}}$, and $\omega_{\text{up,k}}$. Both boundaries base on $\omega_{\max,k}$. If the normalized spectral resolution is defined by

$$\frac{\Delta\omega}{f_s} = \frac{2\pi}{M}, \tag{14}$$

and L is the number of all detected local peaks $\omega_{\max,k}$, tangential boundaries of all possible SEA can be expressed by

$$\omega_{low,k} = \omega_{\max,k} - q\Delta\omega, \tag{15}$$

$$\omega_{up,k} = \omega_{\max,k} + q\Delta\omega, \tag{16}$$

$$k = 1,\ldots,L,$$

where q is an arbitrary SEA spreading factor which ensures that the corresponding root will be inside the SEA frequency range $2q\Delta\omega$.

2.6. CI-block 6

With the complex contour integration (CI), a particular form of Cauchy's integral formula, roots within an SEA can simply be identified. Especially, the principle of the winding numbers (PWN) will be used here. According to this principle the winding number $n(C_w)$ in Eq. (18) results in a positive integer if the border C_{SEA} of an SEA surrounds a root, see Eq. (17) and Fig. 3:

$$C_w = A(C_{SEA}), \tag{17}$$

$$n(C_w) = \frac{1}{2\pi j}\oint_{C_w}\frac{dw}{w}. \tag{18}$$

For more details and the theoretical background of CI methods see Appendix A. Finally, numerical implementations will be discussed in Appendix B.

2.6.1. SEA sampling

The numerical implementation of the CI block requires to sample the closed curve C_{SEA}, see also Appendix B. Discrete segments which relate to the main boundaries found in Section 2.5 will be defined as follows

$$\left.\begin{aligned} z_{\text{outer},k} &= r_{\text{outer}}\cdot\exp(j\omega_a)\\ z_{\text{inner},k} &= r_{\text{inner},i}\cdot\exp(j\omega_a)\end{aligned}\right\} = \omega_{low,k} \leqslant \omega_a(m_1) \leqslant \omega_{up,k},$$
$$\left.\begin{aligned} z_{low,k} &= v\cdot\exp(j\omega_{low,k})\\ z_{up,k} &= v\cdot\exp(j\omega_{up,k})\end{aligned}\right\} = r_{\text{inner},i} < v(m_2) < r_{\text{outer}}, \tag{19}$$

where the discrete steps are given with $m_1 = 1,2,\ldots,A$ and $m_2 = 1,2,\ldots,B$. Now the sampled C_{SEA} is given as

$C^d_{SEA} = C_{SEA}(m_1, m_2) = \{z_{low,k}, z_{\text{outer},k}, z^{-1}_{up,k}, z^{-1}_{\text{inner},k}\}$ where $z_{up,k}$ and $z_{\text{inner},k}$ have to be reversed.

The parameters A and B were defined after evaluations on the TIMIT database (sampling rate $f_s = 16000$ Hz). Comparison studies between root locations of the wGDF-CI method and root locations of a Newton–SRS approach have shown that the conditions $A = 1$ and $B = 5$ represent the minimum number of samples that would allow zero-defected CI implementations.

The remaining parameters for the tests, $q = 2.5, N = 640$, and $M = 512$, allow for an adequate FFT resolution and non-overlapping SEAs in the z-plane (note, we roughly assume a minimum formant frequency spacing of about 150 Hz according to Stevens (2000)). In conclusion, these settings yield a total of 12 sample points per C^d_{SEA} for each iteration, see Fig. 4.

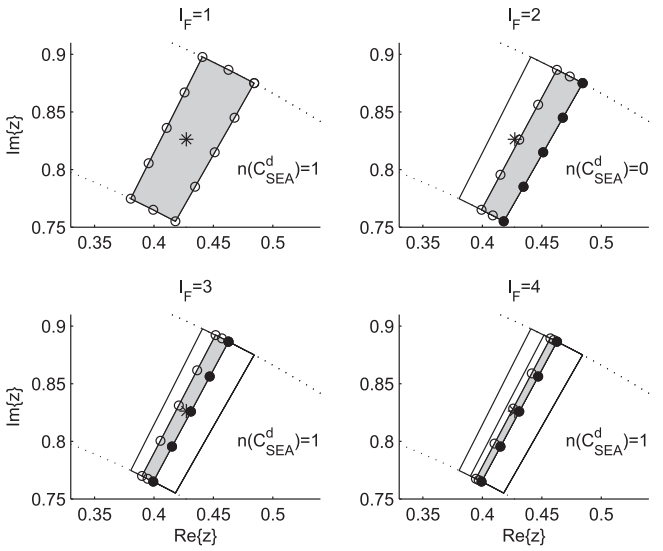

Fig. 4. Exemplary iteration procedure of a single SEA. Each current SEA at iteration step I_F is highlighted in gray and each reused sample is highlighted by a filled marker.

2.6.2. SEA reduction

Iterative SEA reduction helps to find precise root locations within the complex z-plane. Within each iteration the SEA will be halved.

The specific SEA shape facilitates determination of formant frequencies and bandwidths separately for each formant. For this purpose, tangential and radial reduction procedures are separated into two independent computation loops. Computational cost can be scaled to the target aspect of analysis, using a formant-frequency-only or a formant-bandwidth-only implementation of the wGDF-CI approach, e.g. for vowel identification in (International Speech Communication Association, 1995). Additionally, the method allows for scaling the computational cost along the desired frequency or bandwidth resolution by setting the number of iterations for each formant individually.

Fig. 4 shows an exemplary iteration procedure of a single SEA in tangential direction (formant frequency). Obviously at each iteration step with $I > 1$ the sample points of one SEA segment of a given SEA can be reused for the next iteration step, see the filled markers in Fig. 4. Such reuse helps reducing the computational loads, see Sections 2.6.1 and 3.1. This means, only for the first iteration step, 12 sampling point have to be determined. In consequence, the required sample points to be determined are given according to A and B with $S_{\text{tang}} = \{12, 7, 7, \ldots\}$ and $S_{\text{rad}} = \{12, 9, 9, \ldots\}$ for tangential and radial iterations, respectively.

The required number of iterations to achieve arbitrary frequency (ΔF) and bandwidth (ΔB) resolutions can be expressed respectively as

$$I_F = \left\lceil \frac{-\ln\left(\frac{\Delta F}{2q\Delta\omega}\right)}{\ln(2)} \right\rceil \quad \text{and} \quad I_B = \left\lceil \frac{-\ln\left(\frac{\Delta B}{B_{\max}}\right)}{\ln(2)} \right\rceil - 1. \tag{20}$$

Additional remark: bandwidth criteria are checked in the first iteration $I = 1$ which means, in case of a two-stage application, the radial reduction (bandwidth) immediately starts with an SEA halving, see I_B in Eq. (20).

2.7. FX-block 7

The FX processing block transforms complex formant locations into spectral equivalents. For a better understanding, the denominator of Eq. (2) will be written as

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} = \prod_{k=1}^{p} \left(1 - \psi_k z^{-1}\right). \tag{21}$$

where each single root ψ_k can be written as a complex number $\psi_k = r_k \cdot \exp(j\omega_k)$.

With these conditions each formant/root location from the CI block output ψ_i within the complex z-plane can simply be transformed to spectral equivalents by Markel and Gray (1976)

$$F_i = \frac{f_s}{2\pi}\omega_i, \tag{22}$$

$$B_i = -\frac{f_s}{\pi}\ln(r_i), \tag{23}$$

$$\text{with} \quad \omega_i = \tan^{-1}\left(\frac{\text{Imag}\{\psi_i\}}{\text{Real}\{\psi_i\}}\right), \quad \text{and} \quad r_i = |\psi_i|,$$

where again $i = 1, 2, 3$ $(F1, F2, F3)$ and ψ_i are the root locations from the CI block output. Each root location ψ_i has been iteratively determined by successive reductions of the SEA until the user defined frequency and bandwidth accuracy has been reached (see Section 2.6.2).

2.8. An example

An exemplary LP polynomial of order $p = 6$ will be discussed. Its roots are given at the complex points

$\psi_1 = 0.9 \cdot \exp(j0.23\pi), \psi_2 = 0.99 \cdot \exp(j0.25\pi)$ and $\psi_3 = 0.9 \cdot \exp(j0.35\pi)$ and $f_s = 11025$ Hz (see Figs. 5 and 6).

The following steps have to be processed for each SEA (for notation see Appendix B):

1. Calculate $\tau g(\omega)$ of the zero padded LP coefficients $y(n)$, see Eq. (2).
2. Calculate $\omega_{\max} = \max_\omega\{\tau g\}$ and determine number of maxima L.
3. Define $\omega_{low,k}$ and $\omega_{up,k}$ for L peaks at $\omega_{\max}$.
4. Use radial boundaries $r_{\text{inner},i}$ and r_{outer} to complete boundaries for each SEA.
5. Look for roots within current SEA using CI, $i = 1$.
   - *if* root was found $\rightarrow n\left(C_w^d\right) = 1$ for $I = 1$: *while* $It < I_F(i)$
     - Successive division of tangential SEA components into halves for determining formant frequency until a predefined number of iterations has been reached, see Fig. 4.
     - Successive radial SEA halving for determining formant bandwidth until predefined number of iterations has been reached.

     end while
     - Update indexes $k = k + 1$, set formant F_i, set $i = i + 1$ for the next formant, and go to the next SEA.
   - *else if* no root was found $\rightarrow n\left(C_w^d\right) = 0$ for $I = 1$: skip current SEA, update index $k = k + 1$ but $i = i$.
6. If $i = 4$ the first three formants have been found, skip to the next time frame of input $x(n)$.

3. Results

3.1. Computational complexity

The presented wGDF-CI reduces algorithm complexity for defining SEAs in comparison to previous works, see Reddy and Swamy (1984); Sandler (1991). It can be shown that the introduced wGDF facilitates reliable separation of merged peaks, so that interactive functions are no longer necessary.

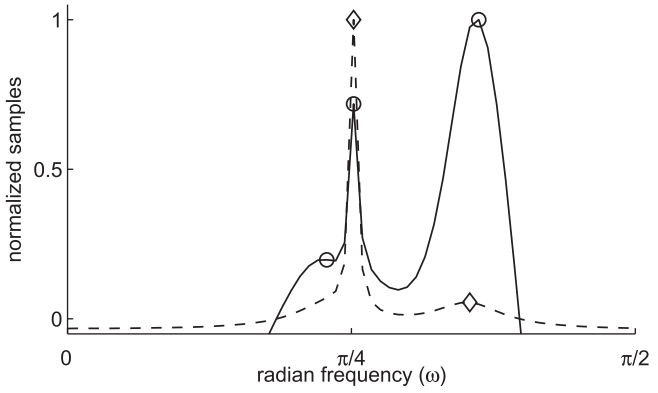

Fig. 5. The wGDF (solid line) and GDF (dotted line) of the exemplary polynomial, note the merged peaks in the unweighted GDF graph.

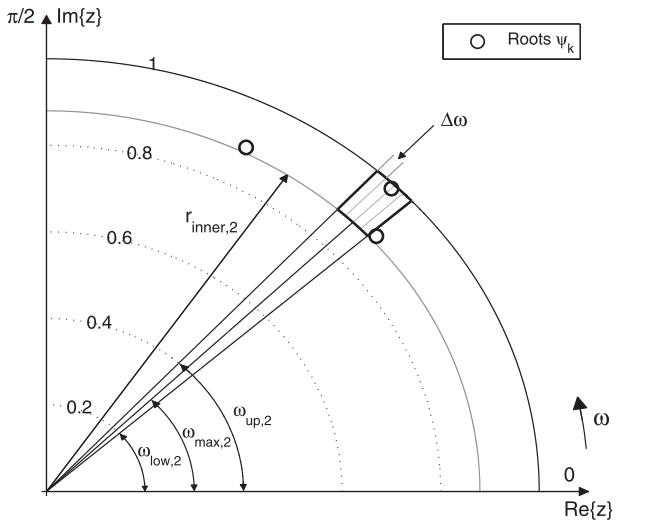

Fig. 6. First quadrant of the complex z-plane with an exemplary SEA which depends on the exemplary polynomial. For SEA boundary determinations see the text.

The computational costs of the wGDF-CI mainly depend on the total number of iterations to meet predefined formant frequency and bandwidth resolutions. The main wGDF-CI parameters which have been used for the presented results on the TIMIT corpus are given in Table 1.

3.1.1. Number of operations

For performance evaluations of the CI block we assume required formant frequency resolutions of $\Delta F \leqslant \{10, 10, 20\}$ Hz for $F1$, $F2$ and $F3$, respectively, and a required bandwidth resolution of $\Delta B \leqslant 12$ Hz for all formants. Under these conditions and with Eq. (20) the numbers of required iterations are given as $I_F = \{4, 4, 3\}$ and $I_B = \{5, 5, 5\}$. This amount of iterations combined with the maximum amount of sampling points that have to be evaluated for individual steps, $S_{\text{tang}} = \{12, 7, 7, 7\}$ and $S_{\text{rad}} = \{12, 9, 9, 9, 9\}$ (see Section 2.6.2), yield in a total of 236 polynomial evaluations (PE) per frame. Such PE rate results in a total of 8024 PE per second which leads to 0.6 million real operations per second (MOPs, multiplications and additions), see Knuth (1998).

Objective complexity of the wGDF block can be expressed with $2O(3Mlog_2M) + O(6M/2)$ real OPs for two times FFT plus numerical GDF. Such complexity leads to a total of 1.0 real MOPs per second so the total cost of both the wGDF and the CI block is about 1.6 real MOPs per second which can be well implemented by mod-

Table 1
Parameter set for all results of this section.

f_s in Hz	q	N (framesize)	M	p	$B_{\max}$ in Hz
16,000	2.5	640 (40 ms)	512	18	750

Table 2
Syllables included in this study of the TIMIT corpus.

	Phonetic transcriptions	Total number of frames
Vowels	iy, ih, eh, ey, ae, aa, aw, ay, ah, ao, oy, ow, uh, uw, uh, uw, ux, er, ax, ix, axr, ax-h	61238
Semivowels	l, r, w, y, hh, hv, el	13826
Nasals	em, m, n, ng, nx, eng, en, or	9981
CVt and VCt	*See the text*	Each 6968

ern processors, e.g. TMS320C5505 from Texas Instruments with 1 GFLOP per second. In contrast to most SRS approaches the wGDF-CI is very robust against degradations of accuracy which provides for high-speed implementations on fixed-point arithmetic based hardware.

For comparison purposes, heuristics on the TIMIT corpus result in 540 PE per frame for a reference Newton's–SRS method after Schleicher (2002). Such frame rate yields in a total of 1.4 real MOPs per second.

3.1.2. Effective computational speed

The computational costs of the wGDF block mainly depends on the FFT complexity. FFT libraries such as the FFTW (IEEE, 1381-1384) and specific FFT hardware accelerators for digital signal processors (e.g. TMS320C5505 from Texas Instruments) provide high-performance FFT implementations. In our studies on a standard Pc (Intel P8400 processor) the wGDF block runs up to two times faster than the CI block when using the FFTW c-library. Under these circumstances the presented method becomes superior to Newton's SRS method.

3.2. Applications

In speech processing, some applications require determination of formant frequencies only, for instance vowel identification. In such cases, the computational costs will be further reduced since the iterations for radial SEA reduction can be set set to $I_B = 0$. In comparison to the two stage iteration with $I_B = \{5, 5, 5\}$ saves about 60% of the computational loads required to run the CI block.

In contrast to most SRS methods which need to determine all roots with high accuracy, the proposed wGDF-CI method allows for determining roots at lower precision because no additional error propagation occurs. This relationship allows for optimal adaption to different system requirements. For example, a formant-frequency-only wGDF-CI approach with $I_F = \{1, 1, \ldots\}$ and $I_B = 0$ provides full access to formant frequencies under defined bandwidth criteria but requires minimal computational costs.

3.3. Formant extraction results

Following methods are compared: wGDF-CI method, the unweighted GDF-CI method and the two SRS based tools WaveSurfer (ESPS/xwaves) (Sjölander and Beskow, 2000) and Praat (Boersma and Weenink, 2005). The performance is benchmarked on a subset of the TIMIT data corpus. Since WaveSurfer includes a viterbi formant trajectory tracker after Talkin, 1987 the raw data of the presented wGDF-CI, the GDF-CI and the Praat method are likewise post processed by means of a three point running median filter.

The following results include error evaluations of vowels, semivowels, nasals, vowel-consonant transitions (VC transitions) and consonant-vowel transitions (CV transitions). Obstruent speech regions of the TIMIT corpus are not discussed in this work, for details see Table 2. VC and CV transitions contain fixed lengths of 4 frames with 2 frames to the left and 2 frames to the right of vowel boundaries.

"Ground truth" values for error calculations were taken from the MSR-UCLA VTR-Formant Database which has been specifically created to investigate formant frequency trackers, see Deng et al. (2006). The MSR-UCLA VTR-Formant Database includes a randomly chosen subset of 538 SX and SI utterances from each speaker of the TIMIT train and test corpus and its reference data provides 10 msec framed, hand labeled formant trajectories. This time resolution of the reference formant data was matched by applying a 10 ms hop size for all methods used in this section.

Table 3 shows the averaged absolute error in Hz for all methods in test. For all passages even for nasal regions the results of the presented method are well comparable with the results of both benchmark formant trackers. The *F*3 extraction results of the wGDF-CI method are generally

Table 3
Formant extraction results. For details see the text.

	Averaged absolute error per frame in Hz					
	wGDF-CI			GDF-CI		
	*F*1	*F*2	*F*3	*F*1	*F*2	*F*3
Vowels	57	86	131	57	87	133
Semivowels	76	118	185	77	129	196
Nasals	122	317	352	125	323	365
CVt	99	137	169	105	133	175
VCt	156	232	310	168	234	331
	Praat			WaveSurfer		
	*F*1	*F*2	*F*3	*F*1	*F*2	*F*3
Vowels	90	116	167	53	84	172
Semivowels	128	201	268	66	106	242
Nasals	180	346	301	93	301	317
CVt	145	179	218	82	127	205
VCt	238	284	358	101	181	294

more precise than the results of both benchmark methods, despite the viterbi processing procedure of WaveSurfer. This means simpler post-processing complexity for the presented method due to the computational simplicity of a three-point running-median filter.

4. Conclusion

In this paper, a method has been presented to locate selective roots of LP data. It combines a modified group delay function approach (wGDF) and the contour integration (CI) of the complex analysis. The problem of merged peaks in speech processing is solved by means of a singular function based on the wGDF. Thus, the complexitiy of determining initial search areas (SEA) in the z-plane has been significantly reduced.

The proposed wGDF-CI method was compared to Praat and WaveSurfer. Under the same conditions, the presented method consistently achieves higher precision than Praat, and almost the same precision as WaveSurfer as evaluated on the TIMIT corpus.

The objective computational complexity of the wGDF-CI method equals the complexity of the standard root solver after Newton's method, however, its effective computational speed is higher due to existing high-performance FFT implementations.

Since the CI approach is robust against degradations of accuracy the present method can be used for implementations on both fixed-and floating-point processors. Several setting options makes the wGDF-CI approach simple to adapt to fixed-and floating-point processors, and can be scaled down to processors with low performance.

Acknowledgment

The authors wish to thank the German Federal Ministry of Education and Research (AiF Project No. 1767X07).

Appendix A. CI: Principle of the winding numbers

This section gives a brief review of how the CI (residue theorem) (Rudin, 1974) can be applied for geometrical root finding as used in Section 2.6.

In complex analysis a specific case of Cauchy's integral formula defines the winding number n of how many times a complex closed path C rotates around the origin. It is given by

$$n(C) = \frac{1}{2\pi j} \oint_C \frac{dz}{z}, \quad z = x + jy. \tag{A.1}$$

This relation can be expressed more generally, as

$$n(C, q) = \frac{1}{2\pi j} \oint_C \frac{1}{z - q} dz, \tag{A.2}$$

where integer n expresses the number of times the closed curve C rotates around the complex point $q = x_q + jy_q$ without containing it, see Figs. A.7 and A.8.

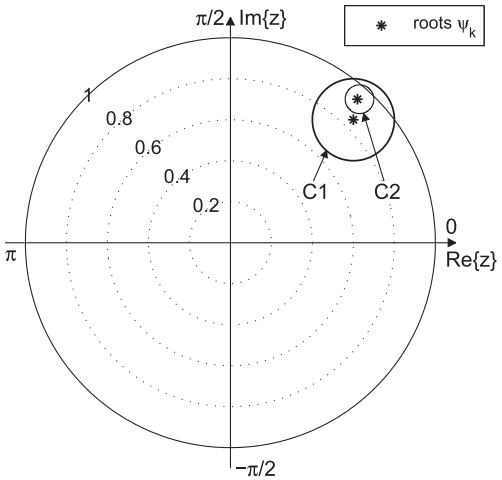

Fig. A.7. Complex z-plane, two exemplary roots ψ_k are enclosed by two simple closed curves $C1$ and $C2$, the polynomial transformation of $C1$, $C2$ from the z-plane into the complex w-plane is shown in Fig. A.8.

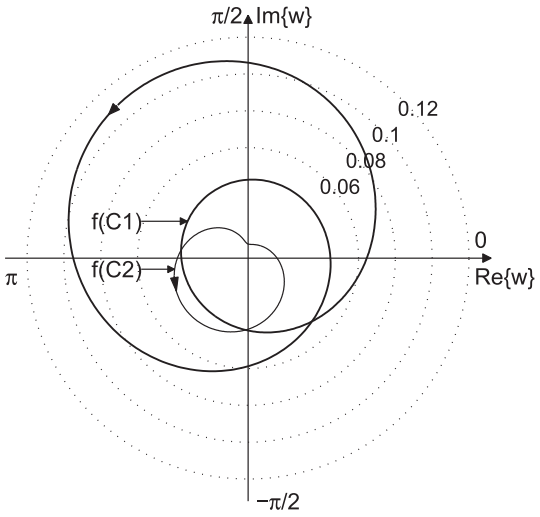

Fig. A.8. Complex w-plane, counting the winding numbers of both transformed curves $C1$ and $C2$ around the origin.

Relation (A.2) can be extended to PWN of the analytical function $f(z)$. With

$$\frac{1}{2\pi j} \oint_C \frac{f'(z)}{f(z)} dz = z_0(C) - z_\infty(C) = n(C), \tag{A.3}$$

the number of zeros z_0 minus the number of poles z_∞ of the function $f(z)$ inside the simple closed curve C is determined (see Figs. A.7 and A.8).

Now we apply the transformation $w = f(z)$ so that formula (A.3) can be written as

$$\frac{1}{2\pi j} \oint_C \frac{f'(z)}{f(z)} dz = \frac{1}{2\pi j} \oint_{f(C)} \frac{dw}{w}. \tag{A.4}$$

In other words, after transforming curve C from the complex z-plane into the likewise complex w-plane, the rotations of $f(C)$ around the origin in the w-plane are

equivalent to the number $n(C)$ in (A.3), see Figs. A.7 and A.8.

Since the integration path C is arbitrary, we come back to the SEA of Section 2.5. With $C = C_{SEA}$, a singular SEA and $f(z) = A(z)$, Eq. (A.3) contains only zeros z_0. In the present term zeros z_0 are commonly known as roots of LP data $A(z)$.

$$C_w = A(C_{SEA}), \tag{A.5}$$

$$n(C_w) = \frac{1}{2\pi j} \oint_{C_w} \frac{dw}{w}, \tag{A.6}$$

In Eq. (A.5) index w indicates the transformation of C_{SEA} into the complex w-plane.

Appendix B. PWM-a numerical implementation

Discrete approaches for CI, especially PWM, commonly utilize point-in-polygon (PiP) algorithms as published in (Alciatore and Miranda, 1995). The most commonly used PiP algorithm bases on the x-axis-crossing approach. Such a procedure is simple to implement and it is characterized by its robustness and computational efficiency.

This section demonstrates numerical determinations of winding numbers of a closed curve as published in (Alciatore and Miranda, 1995). For this purpose the continuous curve C_{SEA} is then discretized, see Section 2.6.1.

With

$$C_w^d = A\left(C_{SEA}^d\right), \tag{B.1}$$

we accomplish the transformation from the complex z-plane into the likewise complex w-plane. After simplifying the notation by

$$x = \mathrm{Real}\left\{C_w^d\right\} \quad \text{and} \quad y = \mathrm{Imag}\left\{C_w^d\right\}, \tag{B.2}$$

the parameters of PiP can be expressed by

$$q = (0,0), \tag{B.3}$$

$$v_b = (x_b, y_b), \quad \text{with} \quad b = 1, \ldots, S, \tag{B.4}$$

$$v_{(\mathrm{end})+1} = v_1, \tag{B.5}$$

$$D = \bigcup_{b=1}^{S} \overline{v_b v_{b+1}}, \tag{B.6}$$

where S denotes the number of SEA sample points S_{tang} or S_{rad}, see Section 2.6.1. With

$$a_b = y_b \cdot y_{b+1}, \tag{B.7}$$

$$r_b = x_b + \frac{y_b(x_{b+1} - x_b)}{(y_b - y_{b+1})}, \tag{B.8}$$

positive x-axis-crossings are easily determined. If both $a_b < 0$ and $r_b > 0$ are true, the positive x-axis has been crossed by D. Crossing directions can be found by verifying $y_b < y_{b+1}$ or $y_b > y_{b+1}$ which depict different update criteria for the winding number n, see Fig. B.9b and c.

An exemplary simple closed curve D is given in Fig. B.9a. PiP steps are calculated as follows:

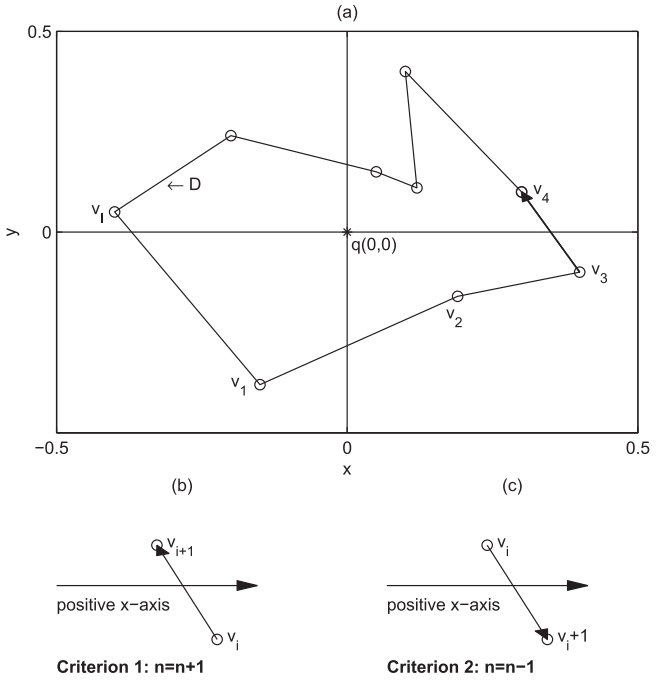

Fig. B.9. (a) An exemplary simple closed curve D in the complex w-plane (b and c) update criterions of the winding number algorithm.

1. $a_3 < 0$ and $r_3 > 0 \rightarrow D(\overline{v_3 v_4})$ crosses the positive x-axis.
2. $\overline{v_3 v_4}$ with $y_3 < y_4$ satisfies criterion 1 in Fig. B.9b.
3. n will be updated to $n = n + 1$.
4. no more positive x-axis crossing detected $\rightarrow n = 1$.
5. D rotates one time around the origin and due to $n > 0 \rightarrow D$ rotates counterclockwise.

References

Alciatore, D.G., Miranda, R., 1995. A winding number and point-in-polygon algorithm, Technical report, Colorado State University.

Atal, B.S., Hanauer, S.L., 1971. Speech analysis and synthesis by linear prediction of the speech wave. J. Acoust. Soc. Am. 50 no 2b, 637–655.

Boersma, P., Weenink, D., 2005. Praat: doing phonetics by computer. Computer program. Retrieved from: http://www.praat.org/.

Dellar, J.R., Proakis, J.G., Hansen, J.H.L., 1999. Discrete-Time Processing of Speech Signals. Wiley–IEEE Press, New York, USA.

Deng, L., Cui, X., Pruvenok, R., Huang, J., Momen, S., Chen, Y., Alwan, A., 2006. A database of vocal tract resonance trajectories for research in speech processing. In: Proc. ICASSP, pp. 60–63.

Dunn, H.K., 1961. Methods of measuring vowel formant bandwidths. J. Acoust. Soc. Am. 33 (12), 1737–1746.

Fant, G., 1960. Acoustic Theory of Speech Production. Mouton & Co., The Hague.

Flanagan, J.L., Meinhart, D.I.S., Cummiskey, P., 1964. Digital equalizer and deequalizer for speech. J. Acoust. Soc. Am. 36 (5), 1030.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., 1993. DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus, CDROM.

Hanson, H., Maragos, P., Potamianos, A., 1994. A system for finding speech formants and modulations via energy separation. IEEE Trans. Speech Audio Process. 2 (3), 436–443.

IEEE, FFTW: An Adaptive Software Architecture for the FFT. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, vol. 3. IEEE. pp. 1381–1384.

International Speech Communication Association (ISCA), Dynamic Vowel Quality: A New Determination Formalism Based on Perceptual Experiments, Eurospeech'95, vol. 1.

Itakura, F., 1975. Line spectrum representation of linear predictive coefficients of speech signals. J. Acoust. Soc. Am. 57, 35.

Kim, H.K., Lee, H.S., 1999. Interlacing properties of line spectrum pair frequencies. IEEE Trans. Acoust. Speech Signal Process. 7, 87–91.

Kim, C., deok Seo, K., Sung, W., 2006. A robust formant extraction algorithm combining spectral peak picking and root polishing. EURASIP J. Appl. Signal Process. 2006, 1–16.

Knuth, D.E., 1998, third ed.. In: The Art of Computer Programming-Seminumerical Algorithms, vol.2 Addison-Wesley.

Kuwarabara, H., Sagisaka, Y., 1995. Acoustic characteristics of speaker individuality: control and conversion. Speech Commun. 16, 165–173.

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

McCandless, S.S., 1974. An algorithm for automatic formant extraction using linear prediction spectra. IEEE Trans. Acoust. Speech Signal Process. 22 (2), 135–141.

Murthy, H., Murthy, K., Yegnanarayana, B., 1989. Formant extraction from phase using weighted group delay functions. IEEE Electron. Lett. 25, 1609–1611.

Oppenheim, A.V., Schafer, R.W., 1975. Digital Signal Processing, Englewood Cliffs, New York, USA.

Peterson, G., Barney, H., 1951. Control methods used in a study of the vowels. J. Acoust. Soc. Am. 24 (2), 1441–1444.

Pfitzinger, H. 2005. Towards functional modelling of relationship between the acoustics and perception of vowels. ZAS papers in Linguistics, vol. 40, pp. 133–144.

Reddy, N.S., Swamy, N., 1984. High-resolution formant extraction from linear-prediction phase spectra. IEEE Trans. Acoust. Speech Signal Process. ASSP-32 6, 1136–1144.

Rudin, W., 1974. Real and Complex Analysis. McGraw-Hill, New York, USA.

Sandler M, (Ed.), 1991. Algorithm for high precision root finding from high order LPC models. IEE Proc. 9(6).

Schafer, R.W., Rabiner, L.R., 1970. System for automatic formant analysis of voiced speech. J. Acoust. Soc. Am. 47, 637–648.

Schleicher, D., 2002. On the number of iterations of newton's method for complex polynomials. Ergodic Theory and Dynamical Systems, vol. 22, No. 3, Cambridge University Press, pp. 935–945.

Sjölander, K., Beskow, J., 2000. Wavesurfer-an open source speech tool. In: Proc. Internat. Conf. on Spoken Language Processing.

Snell, R.C., Milinazzo, F., 1993. Formant location from lpc analysis data. IEEE Trans. Speech Audio Process. 1 (2), 129–134.

Snell, R.C., Dickson, B.C., 1983. Investigation of a speaker verification system using the parameters derived from crc vocoder. Technical report. Centre for Speech Technology Research, Victoria, B.C., Canada.

Stevens, K.N., 2000. Acoustic Phonetics. Mit Press.

Talkin, D., 1987. Speech formant trajectory estimation using dynamic programming with modulated transition costs. J. Acoust. Soc. Am. S1, 55.

Ueda, Y., Hamakawa, T., Sakata, T., Hario, S., Watanabe, A., 2007. A real-time formant tracker based on the inverse filter control method. Acoust. Sci. Technol. 28 (4), 271–274.

Welling, L., Ney, H., 1998. Formant estimation for speech recognition. IEEE Trans. Speech Audio Process. 6 (1), 36–48.

Williams, C.S., 1986. Designing digital filters. Prentice/Hall International, Inc..

Wong, D., Hsiao, C., Markel, J., 1980. Spectral mismatch due to preemphasis in LPC analysis/synthesis. IEEE Trans. Acoust. Speech Signal Process. 80 (2), 263–264.