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)

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

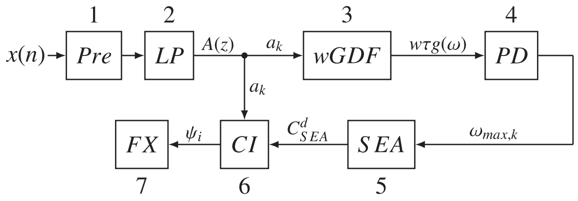

Fig. 1. Overview of the proposed formant extraction method, $Pre$ = short term segmentation and preemphasis of discrete input $x(n)$, $PD$ = peak detection, and $FX$ = output of formants.