Algorithm for formant tracking, modification and synthesis

TAMÁS BŐHM, GÉZA NÉMETH

BME Department of Telecommunications and Media Informatics
{bohm, nemeth}@tmit.bme.hu

Keywords: formant tracking, formant synthesis, linear prediction, speech character modification

Precise formant tracking has been a challenge for researchers in speech processing for long. In this paper, the authors present a method to track and modify formants in speech signals. It is an efficient tool for analyzing and altering the spectral content of speech, furthermore it provides an opportunity to modify timbre and voice quality. The method is based on the linear prediction model. **(In: 2006/8, pp.11–16.)**

1. Introduction

During the production of voiced speech sounds the vocal folds vibrate in a quasi-periodic manner. The excitation signal produced by this oscillation is modulated by the resonator system of the vocal tract (pharyngeal, nasal and oral cavity): while harmonics near the resonant frequency are boosted, other harmonics are attenuated. These vocal tract resonances are referred to as *formants*.

The resonant frequencies are manifested as local maxima in the vocal tract transfer function. Besides their center frequency, formants can be characterized by their bandwidth and amplitude. The former refers to the width of the boosted frequency region in the transfer function – where the function value does not fall more than 3 dB below the local maximum. The latter is the value of the function at the peak [1].

Although formants and their time course (the *formant tracks* or *formant trajectories*) are clearly visible on the spectrum and on the spectrogram for the human eye, automatic formant measurement and tracking is far from trivial.

There is a definite need for exact formant tracking together with the ability to modify formant trajectories both for research purposes and for specific applications. The latter include the modification of the voice character (such as dialect transformation, speech correction or timbre modification) and smoothing the formant trajectories of waveforms generated by concatenative text-to-speech systems. Such algorithms may also be applied to alter the speaker-specific characteristics of speech so that the listener recognizes another speaker uttering the same sentence.

The prospective applications require a method that can re-synthesize speech after altering the formant structure. This can be achieved only by employing a precise formant extraction algorithm. The problem of re-synthesis for our targeted applications has not been extensively studied, we could not find an appropriate solution in the literature.

Formant extraction has been intensively studied during the past decades. Traditional methods employ peak finding on some kind of non-linearly smoothed spectra [2][3]. One way of doing this is cepstral spectrum smoothing when the maxima due to periodicity (corresponding to glottal excitation) are removed from the cepstrum and it is then Fourier transformed [1]. Rabiner and Schafer's method combines this procedure with "analysis by synthesis" [4]. Another approach is the use of various filter banks [5].

Some techniques borrowed from speech recognition can also be applied to formant tracking. For example, methods based on Hidden Markov Models (HMMs) [6] and Line Spectrum Pairs (LSP) are common. The latter is an implementation of the linear prediction (LP) analysis that is performed in the frequency domain instead of the time domain and the calculated coefficients follow the time course of the high-amplitude regions in the spectrum (that roughly correspond to the formants).

In this paper we report a highly accurate method of formant analysis, tracking and modification. In order to demonstrate the algorithm, we created a graphical computer program that is capable of producing various displays of the formant tracks. Section 2 outlines our approach and Section 3 gives the details of the algorithm. Section 4 is a description of the graphical demo program while Section 5 summarizes our findings.

2. Basic concepts

2.1. Linear prediction-based spectrum

As it was discussed earlier, formants are local maxima in the speech spectrum. The spectrum can be calculated by means of Fast Fourier Transform (FFT) but this gives a function with numerous local maxima and minima. It is not straightforward to reliably find the peaks of such a spectrum. Linear prediction-based spectrum calculation [2] is more common for the purpose of formant extraction because it has several advantages over FFT:

– The "resolution" of the spectrum can be set by the order of prediction (that corresponds to the number of poles in the transfer function).
– Linear prediction gives a good estimate of the spectrum primarily at the peaks (that are of interest here).
– It is capable of producing acceptable spectrum estimates even for short speech segments.

The transfer function estimated from the linear prediction coefficients:

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} \qquad z = r \cdot e^{j2\pi \frac{f}{f_s}}$$

This results in a spectrum estimate much smoother than the FFT while formants are not distorted.

2.2. Formant extraction

Two ways of formant extraction have been most frequently discussed in publications: spectrum-based and pole-based. While the former uses amplitude or phase spectrum to find the local maxima corresponding to formants, the latter calculate with the poles in the z-domain.

One spectrum-based method is the McCandless algorithm that detects peaks in the logarithm of the absolute spectrum [7]. Christensen, Strong and Palmer developed a similar procedure but they apply peak finding on the negative second derivative of the log spectrum [8]. Yegnanarayana proved that the first derivative of the complex spectrum phase is showing noteworthy similarity to the amplitude spectrum [9]. Employing the second derivative of this function instead of the log spectrum allows more accurate formant extraction. The method of Reddy and Swamy is calculating simultaneously in the f- and the z-domain so it can distinguish between formants near to each other [10]. Although the above-mentioned methods have been implemented and thoroughly studied, none of them fits our needs.

Our formant tracker is pole-based, similar to the one developed by Slifka and Anderson for speaker transformation [11]. The H(z) transfer function gives an all-pole (i.e. with no zeros) model for the vocal tract. The poles of this function correspond to the resonances of the system. The poles are the roots of the polynomial in the denominator of the transfer function:

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} = \frac{1}{\prod_{i=1}^{p} (1 - p_i z^{-1})}$$

where α_k are the linear prediction coefficients. Formant frequencies and bandwidths can be obtained from the $p_i = r_i \cdot e^{j\varphi_i}$ form of the poles:

$$F_i = \frac{f_s}{2\pi}\varphi_i \qquad B_i = \frac{f_s}{\pi}\ln\left(\frac{1}{r_i}\right)$$

The results are highly accurate but not all poles correspond to formants (e.g. real poles cannot be formants). Such poles are due to lip radiation or background noise.

3. Algorithm

The input data is pitch synchronously segmented speech with sound boundary labels. The phonetic transcript of the utterance can improve the accuracy of the results. We can distinguish two separate stages of the processing: analysis and synthesis. The former one refers to the tracking of formant trajectories and its output is the formant data throughout the utterance and some side information (such as the linear prediction residuals and gains). During synthesis, the modification of the formants and the re-synthesis of speech is implemented, the output is a new waveform.

3.1. Analysis

3.1.1. LP analysis and calculating the poles

In order to apply LP analysis locally in time that is essential for formant tracking, we need to calculate the LP coefficients (LPCs) for every pitch period separately. In order to reduce spectral distortion, our algorithm is determining LPCs for two adjacent pitch periods (a time frame) instead of one and employs Hanning windowing. The window is shifted from pitch period to pitch period. For unvoiced sounds we create "virtual pitch marks" with a constant time step.

First we obtain the PARCOR coefficients by the method of Burg, then we convert these to linear prediction coefficients in order to calculate the transfer function [2].

LP analysis and synthesis do not guarantee that the energy of the output signal is the same as the energy of the input. To avoid this kind of distortion, it is reasonable to store the energy for each time frame that can be used to restore the original level on the output. If we normalize the energy for the length of the frame, we can also use this value for silence detection. The LP residual signal should also be stored for lossless coding.

As a next step, we need to calculate the poles of the system: the roots of the transfer function denominator. Although the polynomial in the denominator has only real coefficients, the roots can also be complex. This prohibits using the Newton-Ralphson and Brent methods. We chose to employ the Laguerre root finding procedure [12] instead.

We can obtain all the roots by applying the Laguerre algorithm iteratively: one run results in either one real root or a complex conjugated pair of roots. After dividing the polynomial with this/these, we run the root finding again.

We can calculate the formant data (frequency, bandwidth and amplitude) by using both the formulae in Section 2 and the spectrum. Note that in synthesis time we should avoid using these data: the calculation is more accurate if we implement formant frequency changes directly on the corresponding pole.

Poles without a corresponding formant also need to be stored for re-synthesis.

3.1.2. Mapping of poles to formants

Not all the poles have a corresponding formant.

In order to discard these poles from further analysis, several criteria should be examined:

– The formant frequency needs to be higher than the fundamental frequency.
– The absolute value of the pole needs to reach a threshold so that we limit the bandwidth of the formant.
– The energy of the time frame needs to be higher than an energy threshold (silence detection).
– Poles with a real part that is zero or near zero can be excluded (these may be due to low frequency narrowband noise).

Only poles meeting all four criteria are considered formants. Formant data (frequency, bandwidth and amplitude) can be calculated from the selected complex conjugated pole pairs.

The above criteria can only exclude the evidently incorrect results and give a first estimate of the formant-pole mapping. The final mapping is obtained by applying continuity constraints on the formant trajectories.

3.1.3. Formant trajectories

Thus the next step is to arrange the formant candidates into continuous formant tracks by using the initial formant-pole mapping and sound boundary information. We need to map the formants of a time frame to the formants of the next time frame in a way that a continuous formant track should emerge. We map a formant candidate to the formant track whose last assigned formant is nearest in frequency. Formant tracks that have no or minimal collision with each other can be merged. Trajectories that are running parallel near to each other are also merged into one track. Extremely short trajectories are discarded.

There is no point in applying continuity constraints before and after obstruents (stops, fricatives and affricates) because the production of these sounds implies such articulatory movements that can cause a quick change in the formants. If the phonetic transcript of the utterance is available then our method does not try to connect formants through these boundaries. If it is not available, every sound boundary is treated as a break point. This approach may lead to less accurate information on several sound transitions but it improves the general efficiency of the mapping.

3.2. Synthesis

As the first step, the input of the re-synthesis procedure is created from the output of the analysis, so we alter the formant tracks. The way of mapping is always determined by the actual application.

For example, we might use some kind of interpolation in order to spectrally smooth the output of a concatenative TTS system. Note that by modifying the trajectories we move the poles on the z plane.

The second step is re-synthesis. The modified formant frequencies give the new pole locations (by using the same formulae as in the analysis). We construct the polynomial for the denominator of the transfer function from the poles. The linear predictive coefficients are obtained as the coefficients of the polynomial. Using these and also the stored residual signal, linear prediction synthesis can be performed. As a last step we restore the energy of the time frames. The result is a new waveform with modified formant tracts.

4. Results

Evaluation was done for test utterances in Hungarian and separately for the analysis and synthesis stage.

4.1. Analysis

The accuracy of formant analysis and tracking was tested in three ways. First, reference spectrograms for testing was generated by a Kay Elemetrics CSL 4300B digital signal analyzer at the Linguistics Institute of the Hungarian Academy of Sciences. These spectrograms were compared manually with the formant tracks obtained by our algorithm. Second, measured formant frequency values were compared with published data [13].

Finally, we measured a mapping error rate as defined in [14]. For this purpose the formant tracks produced by the demo application (that is to be described later) were compared to spectrograms. The test set consisted of 29 two-word Hungarian recordings by a male speaker. A mapping error in the first three formant tracks was found in only two cases (6.90%) – in one case the third while in the other case all the three formants were mapped incorrectly.

According to [14] an algorithm using nominal formant frequency values achieved a mapping error of 3.62-3.99%. Although this is lower than our error rate, the algorithm presented in this paper does not rely on predefined, typical formant frequencies so it is independent of the physiological attributes, gender and language of the speaker. The paper cited above reports an error rate of 13.04% for a method with similar properties.

An example of the output is given in *Figure 1*. The utterance was “jaj hajít”. According to the figure, formants for voiced sounds were generally well detected, even after the /h/-vowel transition in the second word. This case was highlighted as problematic in [14] for formant analyzers that do not use nominal formant frequencies.

4.2. Synthesis

The synthesis capabilities of the method are evaluated in the context of prospective applications since the modification method is designed to fit these. Evaluation and fine tuning of the system is underway. We report the results of two simple initial experiments here.

By changing the formant structure, we can modify a recorded vowel to another sound. As an example, Hungarian “fésü” (fE:SU) was transformed to “fásü” (fA:SU) by raising the first and lowering the second formant track along the frequency axis. Such a formant modified re-

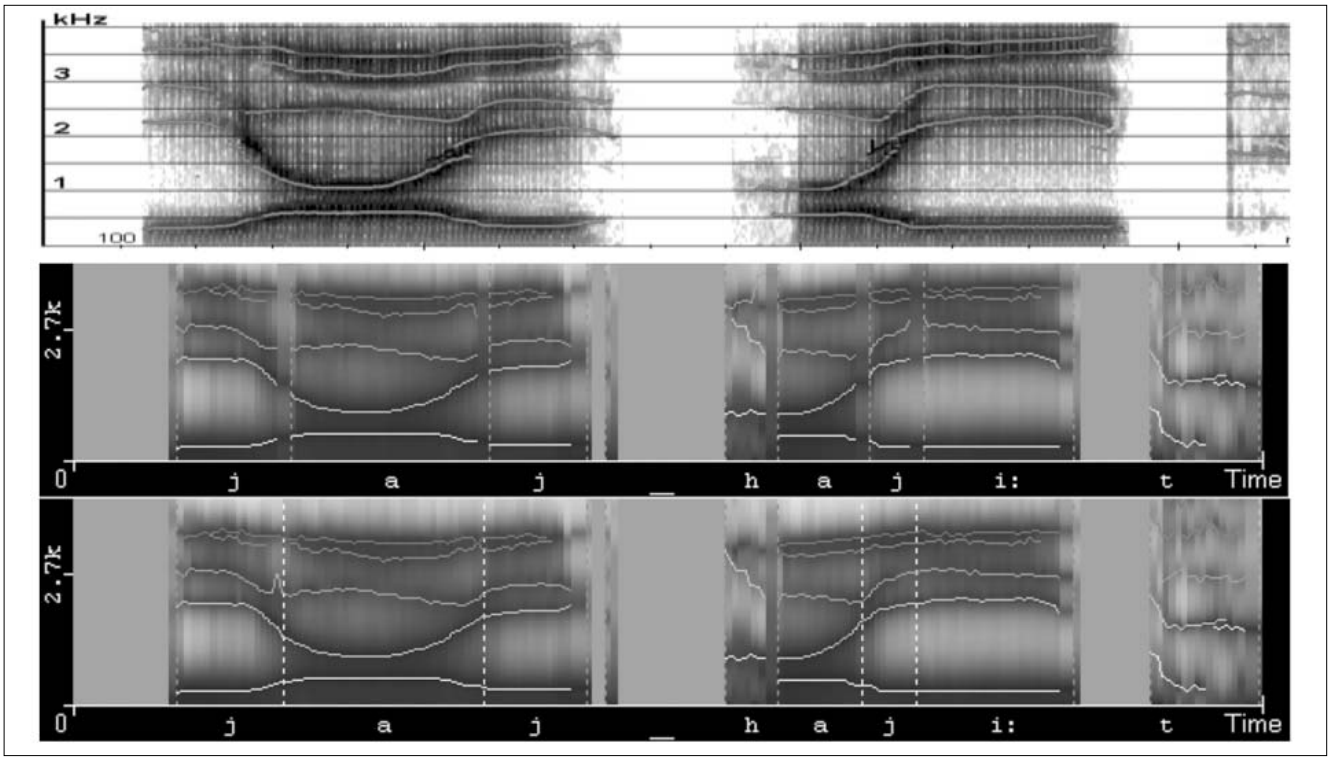
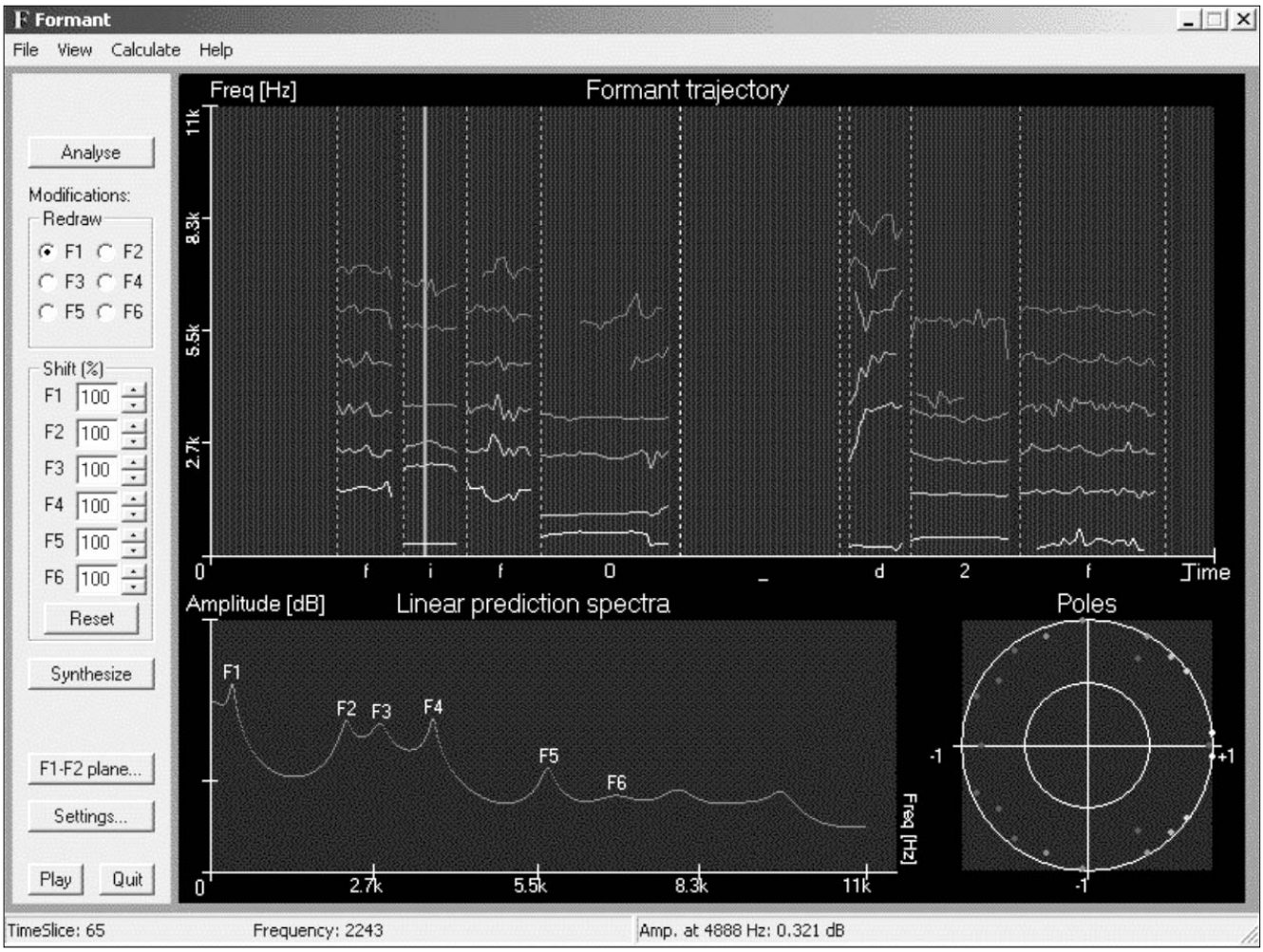


Figure 1. Analysis of a short recording by Kay Elemetrics CSL 4300B (top panel) and by the proposed algorithm – without (middle panel) and with (bottom panel) phonetic transcription

Figure 2. Graphical computer program for formant tracking and modification



cording was listened to by four native Hungarian subjects. All of the listeners were confident that they heard the meaningless word “fásü”. This technique was effectively used in initial experiments to extend speech databases of concatenative synthesizers with vowels that were not recorded (e.g. vowels that do not exist in Hungarian).

It is generally accepted among speech researchers that the higher formant frequencies bear characteristics specific to the speaker. We conducted several initial experiments towards voice transformation where the aim is to modify the personal features so that the original speaker’s identity disappears. By altering some carefully chosen formants, we could confuse the recognition of speaker identity in several listeners.

While manipulating formant tracks we noted that in case of drastic changes audible artifacts may appear in the speech signal. In case of minor (at most about 20%) changes the re-synthesized speech is of good quality and natural-sounding.

4.3. A graphical computer program for formant tracking and modification

A graphical computer program was developed for demonstration and evaluation *(Figure 2)*. This program is publicly available for educational and research purposes [15].

After analyzing the waveform, the program displays all the formant tracks (at most six) in different colors on the time-frequency plane. In order to check the results, the program is also capable of displaying the spectrogram or to show both plots overlaid. After selecting a time frame, the corresponding short-time linear prediction spectrum is drawn (with the formant peaks marked) and the poles of the linear prediction filter appear on the z-plane (in the bottom right corner of the window). The latter is a novel way of visualization that bears the same information as the linear prediction spectrum.

Formant tracks can also be displayed on the F_1-F_2 plane *(Figure 3)*. Instead of the commonly used scatterplots, the program draws the movement of the first two formants in time as a continuous curve.

The figure shows the F_1-F_2 tracks for the vowels in the utterance. The horizontal position of the points of the curve corresponds to the value of the first formant while the vertical position refers to the second formant. The shade of the line represents the time: data from the first pitch period of the vowel is drawn with the darkest color and then it becomes lighter and lighter with each pitch period, the last pitch period being the light-

Figure 3. Formant tracks of vowels on the F_1-F_2 plane

est. In order to improve visibility, circles mark the starting points.

Formant tracks can be altered by redrawing them by hand or by using a factor for each trajectory. The waveform can be re-synthesized with the modified formant tracks and saved to a file.

One can imagine a wide range of applications for this program in the education of phonetics. For example, it can be used as a tool to demonstrate the formant structure of speech sounds or the distinctive features of vowels. It is also capable of visualizing the similarities and differences among different realizations of the same phoneme and to show the effects of co-articulation. It can also be used for phonetics research: for producing the stimulus set of perceptual experiments, for the study of dialects with "analysis by synthesis" and for producing plots of formant tracks.

5. Summary

A general purpose formant tracker and modifier is needed for a wide range of applications. In this paper such an algorithm was described and evaluated. Our method gives acceptable results even with scarce information and by using further input data (phonetic transcript), the results become very accurate. The algorithm was implemented and built into a graphical computer program.

This publicly available program can serve as a tool for education and research – besides in the courses taught by the authors, it is also used at the ELTE Faculty of Humanities. Furthermore, it was successfully applied to extensively examine the formant structure of Hungarian vowels (the data of this study is being analyzed).

References

[1] Gordos, G., Takács, Gy.:
Digital Speech Processing
(Digitális beszédfeldolgozás),
Műszaki Könyvkiadó, Budapest 1983.
pp.52–59; 241–247.

[2] Markel, J. D., Gray, A. H.:
Linear Prediction of Speech,
Springer-Verlag, Berlin 1976., pp.154–158.

[3] Lobanov, B., Levkovskaya, T., Kheidorov, I.:
Speaker and channel – Normalized set of formant
parameters for telephone speech recognition,
Proc. of Eurospeech 1999, Vol. 1., pp.331–334.

[4] Rabiner, L. R., Schafer, R. W.:
Digital Processing of Speech Signals,
Prentice-Hall, Englewood Cliffs, 1978.

[5] Ouni, K., Lachiri, Z., Ellouze, N.:
Formant estimation using Gammachirp filterbank,
Proc. of Eurospeech 2001, Vol. 4., pp.2471–2474.

[6] Weber, K., Bengio. S., Bourlard, H.:
HMM2 – Extraction of formant structures and
their use for robust ASR,
Proc. of Eurospeech 2001, Vol. 1, pp.607–610.

[7] McCandless, S. S.:
An algorithm for automatic formant extraction using
linear prediction spectra, IEEE Trans. on Acoustics,
Speech and Signal Processing, Vol. 22, no.2., 1974.

[8] Christensen, R. L., Strong, W. J., Palmer, E. P.:
A comparison of three methods of
extracting resonance information from predictor
coefficient coded speech, IEEE Trans. on Acoustics,
Speech and Signal Processing, Vol. 24, no.1., 1976.

[9] Yegnanarayana, B.:
Formant extraction from linear prediction phase
spectra, Journal of the Acoustical Society of America,
1978, Vol. 63, pp.1638.

[10] Reddy, N. S., Swamy, M. N. S.:
High-resolution formant extraction from linear
prediction phase spectra, IEEE Trans. on Acoustics,
Speech and Signal Processing, Vol. 32, no.6., 1984.

[11] Slifka, J., Anderson, T. R.:
Speaker modification with LPC pole analysis,
Proc. of ICASSP 1995, pp.644–647.

[12] Orchard, M. T.:
The Laguerre method for
finding the zeros of polynomials,
IEEE Transactions on Circuits and Systems, 1989.
Vol. 36, no.11, pp.1377–1381.

[13] Olaszy, G:
Electronic Speech Production
(Elektronikus beszédelőállítás),
Műszaki Könyvkiadó, Budapest, 1989.

[14] Lee, M., van Santen, J., Möbius, B., Olive, J.:
Formant tracking using segmental
phonemic information, Proc. of Eurospeech 1999,
Vol. 6, pp.2789–2792.

[15] http://fonetika.nytud.hu