AN ALGORITHM FOR FORMANT ANALYSIS, TRACKING AND MODIFICATION

Géza Németh, Géza Kiss, Tamás Bőhm and József Kiss

Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Hungary

ABSTRACT

Formant tracking has been an intensively studied topic in speech research. In the work reported in this paper formants and their tracks are used for analyzing and modifying the spectral content of speech. The algorithm is based on linear prediction (LP) analysis, on finding the roots of the all-pole filter, and on a constraint-based mapping between poles, formants and formant tracks. Formant frequency and bandwidth information can be obtained from the poles assigned to formants. The main advantage of this method is its accuracy: the results are considerable even with minimal input and it can be improved significantly by providing additional information. With the ability to modify the extracted formant tracks and to re-synthesize speech, it allows a wide range of applications: a visual tool for educational purposes was implemented and the described method was successfully used to conduct an extensive formant analysis for Hungarian vowels.

1. INTRODUCTION

There is a definite need for exact formant tracking together with the ability to modify formant trajectories for research purposes and for specific applications. These 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 can also be applied to voice conversion or voice transformation.

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. Since the problem of re-synthesis has not been extensively studied yet, the authors have not found a solution in the literature.

Formant extraction has been intensively studied in the past decades. There are several approaches in the focus of the scientific community: most of them employs some kind of peak finding on cepstrally smoothed or LP spectra. There are some algorithms based on Hidden Markov Models (HMM or lately HMM2) employing statistical criteria for finding the best fit of formant structures [1,2]. Dynamic programming for trajectory estimation is also applied [3]. Some of these algorithms are highly robust, fast or fit well to a specific application but a general algorithm for accurate formant extraction and tracing trajectories is still an area of research.

In this paper we report a method of formant analysis, tracking and modification that is highly accurate.

2. BASIC CONCEPTS

2.1. LP-based spectrum

A formant is a local maximum in the spectrum of the speech (a resonance point). Linear prediction (LP) based analysis is preferred to DFT in automated formant extraction because it produces a smoothed spectrum approximation where we can set the resolution by the order of prediction. Furthermore, it is nearest to the spectrum at the peaks and can lead to valuable results even in the case of short segments.

With the α_k coefficients of p-order linear prediction, we can give an all-pole model to the vocal tract.

2.2. Formant extraction

Two ways of formant extraction have been most frequently discussed in publications: spectrum-based and pole-based. The former uses amplitude or phase spectrum to find the local maxima corresponding to formants. The method of McCandless defines the log spectrum as the logarithm of the absolute spectrum and looks for the peaks in this function [4]. Christensen, Strong and Palmer worked out a faster way of formant detection. It is applying the same peak-finder algorithm to the negative second derivative of the log spectrum [5]. Yegnanarayana proved that the first derivative of the complex spectrum phase is showing noteworthy similarity to the amplitude spectrum [6]. Employing the second derivative of this function allows more exact formant extraction than the second derivative of the log spectrum. Reddy and Swamy developed a method that is calculating simultaneously in the f- and the

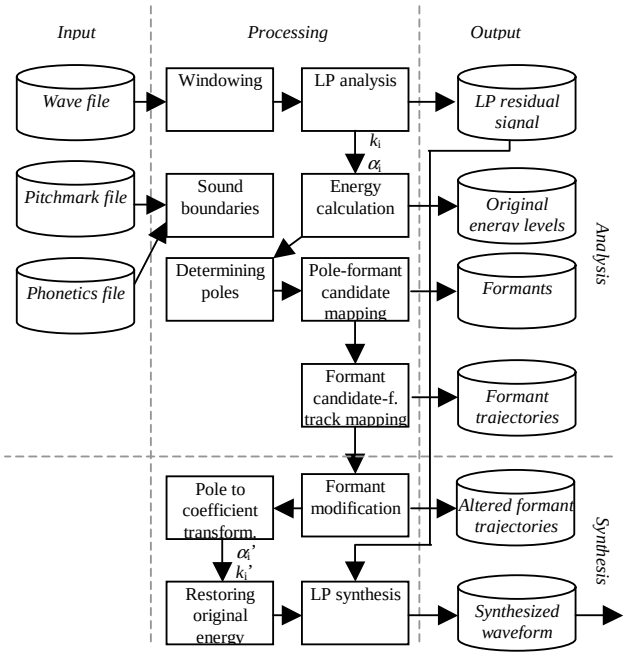

Figure 1. Block diagram of the algorithm

z-domain so it can distinguish between formants near to each other [7].

Although the above-mentioned methods have been implemented and thoroughly studied, none of them fits our needs in terms of accuracy. The algorithms developed so far are usually based on the ones mentioned.

Slifka and Anderson introduced an LP pole-based approach to voice conversion [8]. We decided to use a similar method for formant extraction. The poles of the transfer function are corresponding to the resonance points (the value of the function is increasing without limits when approaching them), i.e. the local maxima of the amplitude spectrum. So the poles of H(z) and the peaks of the amplitude spectrum mutually determine each other. Note that all the poles of the system must be inside the unit circle because the process is stable.

In order to calculate the formants, first we have to find the poles of the transfer function estimation, and then we can get the formant frequencies (F_i) and bandwidths (B_i) from the $p_i = r_i \cdot e^{j\varphi_i}$ form of the complex roots [9]:

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

3. ALGORITHM

We developed our algorithm based on the concepts above. The input data is pitch synchronously segmented speech. Sound boundary information and the phonetic transcript of the utterance can improve the accuracy of the results. We can distinguish two separate stages of signal processing: analysis and synthesis. The former one refers to the tracking of formant trajectories and its output is the formant data throughout the utterance. During synthesis, the modification of the formants and the re-synthesis of speech is implemented (the output is a new waveform). The block diagram of the algorithm is shown in Fig. 1 and the explanation of the specific blocks is given in the following subsections.

3.1. Analysis

3.1.1. LP analysis

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.

The order of LP analysis is 14 at the sampling frequency of 11025 Hz and 26 at 22050 Hz. It is reasonable if we would like to detect six formants: each formant has a corresponding pair of complex conjugated roots and it was observed that the number of real roots is usually two.

We determine the k_m partial correlation (PARCOR) coefficients with the Burg method. Then we convert them to α_k values in order to calculate the transfer function.

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 change, it is advisable 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.

If the pitch-synchronous segmentation of the utterance is not available, the algorithm uses uniform segmentation.

3.1.2. Pole-formant candidate mapping

We used the algorithm of Laguerre [10] to determine the poles of the system. Since multiple roots are almost impossible to occur, the number of poles is extremely unlikely to be less than the order of LP analysis.

The output of the Laguerre algorithm is a real root or a pair of complex conjugated roots. After dividing the polynomials with these, we run the algorithm again until we have all the roots.

It has to be decided which roots we consider formant candidates. This mapping is done by checking several constraints:

- The frequency (calculated from the argument of the pole) of a formant candidate must be above the fundamental frequency.
- The absolute value of the pole must meet a lower limit

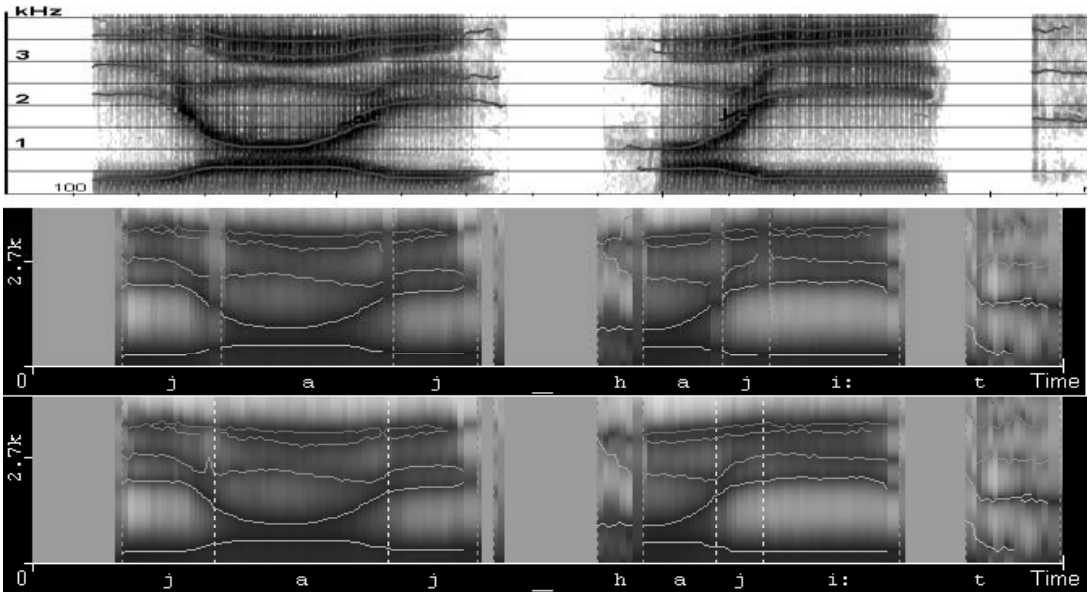

Figure 2. The utterance "jaj hajít" analysed by they Kay 4300B instrument (above) and by our algorithm – with pitch-synchronous segmentation and sound-boundaries (middle) and with phonetic transcript (below) provided. The spectrogram is in the background.

so the formant cannot be too wide-banded.

- The energy of the time frame must exceed a minimum (silence detection).
- Poles with a real part that is zero or almost zero can be excluded in order to eliminate narrowband noise.

The poles that are not considered a formant candidate have to be stored for re-synthesis.

3.1.3. Formant candidate-formant track mapping

The formant candidates have to be arranged into continuous formant tracks, employing continuity constraints. We map a formant candidate to the formant track whose previous assigned formant is nearest in frequency. Formant tracks that have no or minimal collision with each other can be merged. Extremely short trajectories should be excluded.

There is no point in applying continuity constraints before and after plosives, 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. Else, every sound boundary is treated this way. This approach may lead to less accurate information on several sound transitions but it improves the general efficiency of the mapping.

3.2. Synthesis

3.2.1. Modification of the formant trajectories

This stage of processing is about creating the input of re-synthesis from the output of analysis. This mapping can be arbitrary. For example, we might use some kind of interpolation in order to spectrally smooth the output of a concatenative TTS.

The way of mapping is always determined by the actual application. We have to note that the modifications should not be made directly on the formant frequencies, rather on the argument of the complex conjugated pole pair associated with the formant to avoid adding noise.

3.2.2. Re-synthesis

First we have to express the polynomial created from the poles in order to calculate the α-type LPCs. They have to be transformed to PARCOR coefficients before synthesis.

If we did not use overlap-add windowing, discontinuity would appear on segment boundaries. The Hanning function causes small amplitudes to appear on the sides of the window, that helps to restore the continuity of the waveform. The final step is energy restoration.

4. EVALUATION

In order to test and evaluate our method, a demonstration application was developed. All figures in this paper were created with this tool (except where stated otherwise). This application can be used for educational purposes.

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 4300B instrument at the Linguistics Institute of the Hungarian Academy of Sciences. Second, formant frequency values published in [11] were also used. Finally, the formant tracks were compared to the spectrograms produced by the demo application and we measured a mapping error rate as defined in [12].

The corpus used for this latter experiment consisted of 29 utterances of several Hungarian words by a male speaker. It covered all the typical VC and CV transitions. We found a mapping error in any of F_1, F_2 or F_3 at 2 utterances (6.90%) – in one case F_3 was false and in the other, all the three first formants were mapped incorrectly. The published error rate for a formant tracker using nominal formant frequencies is 3.62-3.99%. Our algorithm produced a higher number of errors but it does not use any predefined typical values so it is speaker-, gender- and language-independent. Such a formant tracking method achieved 13.04% in [12].

An example of the output is given in Fig. 2. The utterance was "jaj hajít". According to the figure, formants for voiced sounds were generally well detected, even after the fricative-vowel transition in the second word. This case was highlighted as problematic in [12] for formant analyzers that do not use nominal frequencies.

4.2. Synthesis

The synthesis capabilities of the method are evaluated in the context of prospective applications since the modification method is highly depending on it.

By changing the formant structure, we can modify a recorded vowel to another phoneme. As an example Hungarian "fésü" (fE:SU) was transformed to "fásü" (fA:SU). We calculated the ratios of the formants of the sounds and used them as multiplication factors. Four out of four listeners recognized the formant-modified output as "fásü". This technique was effectively used in initial experiments to extend speech databases of concatenative synthesizers with vowels that were not recorded.

It might be the higher formant frequencies that bring personal characteristics. We conducted several experiments towards voice transformation (where the aim is to modify the personal features so that the original speaker's identity disappears): altering some carefully chosen formants, we could confuse the recognition of speaker identity in several listeners.

We have to note that artifacts can appear when drastically modifying the formants. These may be eliminated by modifying the poles not mapped to formants also [8].

5. SUMMARY

Formant analysis, tracking and modification can be the background of a wide range of applications. In this paper such an algorithm is presented. High accuracy was a requirement and it was accomplished according to the results of the evaluation. The formant track mapping error was compared to an up-to-date algorithm. The results are promising, considering that our method is independent of the speaker and the language. The method can produce considerable results even with minimal input and shows a definite improvement when more information is available.

A demonstration application employing the methods detailed in the paper was developed for evaluation and research purposes.

The algorithm was successfully employed in an extensive formant analysis for Hungarian vowels. The processing of the results is underway.

6. ACKNOWLEDGEMENTS

The authors would like to thank Gábor Olaszy (Linguistics Institute of the Hungarian Academy of Sciences) his valuable advices. Géza Németh acknowledges the support of the National Széchenyi Scholarship Award.

7. REFERENCES

[1] Weber, K., Bengio. S., Bourlard, H. "HMM2 – Extraction of Formant Structures and their Use for Robust ASR", *Proc. of Eurospeech*, Vol. 1, pp. 607-610, Aalborg, 2001.

[2] Acero, A., "Formant Analysis and Synthesis Using Hidden Markov Models," *Proc. Of Eurospeech*, Vol. 1., pp. 411-414., Budapest, 1999.

[3] Xia, K. and Espy-Wilson, C. "A New Strategy of Formant Tracking Based on Dynamic Programming", *Proc. of the ICSLP*, pp. 832-836, Beijing, 2000.

[4] McCandless, S. S., "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra", *IEEE Tr. on A.S.S.P.*, Vol. ASSP-22, No. 2., 1974.

[5] Christensen, R. L., Strong, W. J. and Palmer, E. P., "A Comparison of Three Methods of Extracting Resonance Information from Predictor Coefficient Coded Speech", *IEEE Tr. on A.S.S.P.*, Vol. ASSP-24, No. 1., 1976.

[6] Yegnanarayana, B., "Formant Extraction from Linear Prediction Phase Spectra", *J. Acoust. Soc. Amer.*, Vol. 63., pp. 1638, 1978.

[7] Reddy, N.S., Swamy, M.N.S. "High-Resolution Formant Extraction from Linear Prediction Phase Spectra", *IEEE Tr. on A.S.S.P.*, Vol. ASSP-32, No. 6., 1984.

[8] Slifka, J., and Anderson, T. R., "Speaker modification with LPC pole analysis," *Proc. of ICASSP*, pp. 644-647, 1995.

[9] Lavner, Y., Gath, I., Rosenhouse, J. "The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels", *Speech Communication*, 30:9-36, 2000.

[10] Orchard, H. J. "The Laguerre method for finding the zeros of polinomials", *IEEE Tr. on Circuits and Systems*, Vol. 36, No. 11, p 1377-1381, 1989.

[11] Olaszy, G., *Electronic Speech Synthesis, Acoustics and Formant Synthesis of Hungarian* (original title: Elektronikus beszédelőállítás), Műszaki Könyvkiadó, Budapest, 1989.

[12] Lee, M., van Santen, J., Möbius, B., and Olive, J., "Formant Tracking Using Segmental Phonemic Information," *Proc. Of Eurospeech*, Vol. 6., pp. 2789-2792., Budapest, 1999.