Improved Differential Phase Spectrum Processing For Formant Tracking

Baris Bozkurt, Boris Doval, Christophe D'Alessandro, Thierry Dutoit

Faculté Polytechnique De Mons, TCTS Lab,
Initialis Scientific Park, B-7000 Mons, Belgium, {bozkurt,dutoit}@tcts.fpms.ac.be,
LIMSI,
CNRS, Po Box 133 – F91403 Orsay, France
{boris.doval,cda}@limsi.fr

Abstract

This study presents an improved version of our previously introduced formant tracking algorithm. The algorithm is based on processing the negative derivative of the argument of the chirp-z transform (termed as the differential phase spectrum) of a given speech signal. No modeling is included in the procedure but only peak picking on differential phase spectrum. We discuss the effects of roots of z-transform to differential phase spectrum and the need to ensure that all zeros are at some distance from the circle where chirp-z transform is computed. For that, we include an additional zero-decomposition step in our previously presented algorithm to improve its robustness. The final version of the algorithm is tested for analysis of synthetic speech and real speech signals and compared to two other formant tracking systems.

1. Introduction

Automatic tracking of acoustic resonance frequencies of vocal tract, the formant frequencies, has been an important part of speech analysis for many years. Formant estimation methodologies are useful in many areas of speech processing especially in parametric speech synthesis.

In this study, we present a modified version of our previously introduced formant tracking methodology [1]. The algorithm we propose is completely composed of spectral processing and no modeling of speech is included. It is based on picking peaks of differential phase spectrum of speech signals.

The closest methodology to ours is described in [2]. The authors present an algorithm based on picking peaks of group delay spectrum obtained from amplitude spectrum of speech with the minimum phase assumption. In fact, such an algorithm can rather be considered to be a form of cepstrally smoothed amplitude spectrum processing since no actual phase information is processed. Our algorithm mainly differs from such an approach by processing directly the phase information of the z-transform. In addition, we do not perform any smoothing operation but smooth spectrum is automatically obtained by setting the analysis circle radius appropriately. One additional feature of our method is the removal of glottal flow effects from the speech signal, which reduces the risk to label the glottal formant [3] as a vocal tract formant.

As explained in [3], the spectrum of the glottal flow excitation signal contains a resonance-like peak, which can be called the glottal formant (although it does not correspond to an acoustic resonance). The authors indicate that the frequency region for this peak is rather low and does not exceed a few times the fundamental frequency. However, despite its low frequency, this peak may be labeled to be the first formant by automatic formant trackers if its relative amplitude to the first formant peak is not very low. In various studies like [4], researchers have observed such errors in formant frequencies estimates and included filters to exclude low frequency peaks from formant candidates. Therefore, removal of the glottal flow contribution from speech may have some value to get rid of such problems without manually fixed filters. Such filters may also remove actual formants if the threshold frequency is high or may include the glottal formant in vocal tract formant candidates if the threshold frequency is low. Our improved algorithm includes a source-tract separation step which reduces the risk of such errors.

In Section 2, we discuss the effect of location of roots (zeros) of z-transform polynomial on the differential phase spectra. We show that a study of actual zero locations on the z-plane leads to better comprehension of the problems of the differential phase spectrum processing and provides new solutions. In Section 3, we present the new version of our algorithm for formant frequency estimation and in Section 4, we present the tests performed with synthetic speech and real speech.

2. Effects of zeros of z-transform on differential phase spectra

The main motivation for processing differential phase spectra computed on circles other than the unit circle is to get rid of spiky peaks created by zeros of z-transform (ZZT) which mask formant peaks on group delay functions [1]. In Fig. 1, a typical group delay function and ZZT of the signal are provided. Polar coordinates is preferred for the ZZT plots since comparative visual study with amplitude and phase spectra is easier. The zeros close to the unit circle are superimposed to show the spiky effects created by them on the group delay function. Indeed, the group delay function seems to be composed of only a DC function with spikes due the zeros close to the unit circle. However, we know from filter theory that group delay function should include well resolved peaks at the frequencies of resonances corresponding to the formants.

In our previous study, we have shown that a way to get rid of domination of spikes due to ZZT close to the unit circle is to compute the group delay function away from the unit circle. We named the result as "the differential phase spectrum" since group delay is defined for a computation on the unit circle. The differential phase spectrum is simply the

negative derivative of the argument of the chirp-z transform spectrum. Given the chirp-z transform $X(re^{j\varphi})$, the differential phase spectrum $DP(\varphi)$ is defined by the Eq. 3.

$$X(re^{j\varphi}) = a(\varphi) + jb(\varphi) \quad (1)$$

$$\vartheta(\varphi) = \arctan\left(\frac{b(\varphi)}{a(\varphi)}\right) \quad (2)$$

$$DP(\varphi) = -\frac{d(\vartheta(\varphi))}{d\varphi} \quad (3)$$

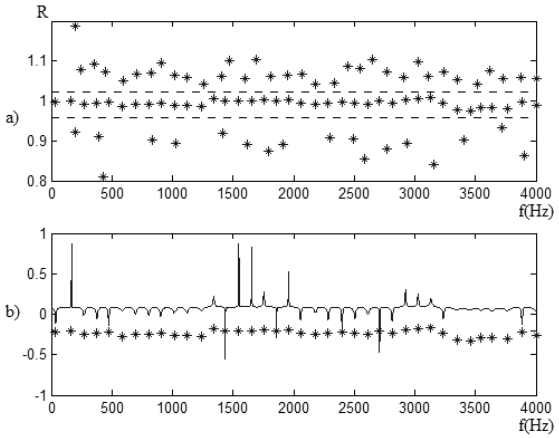

Figure 1: Effect of zeros on group delay function, a) Zeros of Z-Transform (ZZT) plotted in polar coordinates (region of zeros close to the unit circle indicated by dashed lines), b) group delay function with ZZT close to unit circle superimposed.

We have shown that, if the radius of the analysis circle is appropriately set, peaks due to formants of speech signals can be observed on differential phase spectra even with better resolution than amplitude spectrum. However, our further experiments showed that such an approach is sensitive to even a single zero close to the analysis circle and we need to find a means to guarantee certain distance between zeros and the analysis circle. The zeros can appear nearly everywhere on the z-plane though most of them are closely located to the unit circle. In Fig. 2, we demonstrate this type of sensitivity, which causes errors in formant tracking with peak picking.

The peak at 1000Hz in Fig. 2b is due to the single zero close to R=1.1(Fig. 2c). Though it does not correspond to a resonance of the vocal tract, a peak picking algorithm would label this peak as a formant. This peak even hides the F2 peak which should be observed around 1200Hz. For the rest of the frequency axis the differential phase spectrum includes vocal tract formant peaks (around 1900Hz and 3400Hz) as well as the glottal formant peak (lower than 250Hz).

For this reason we have performed a systematic study of ZZT for windowed speech signals (since windowing was observed to be one of the most important factors affecting ZZT of signals. In addition, this part of our study is inspired by Hideki Kawahara's presentation for [5], on group delay functions and windowing.). A summary of our study on ZZT is presented in [6].

The study on zeros of z-transform (ZZT) [6] lead to the following interesting result : for a glottal closing instant (GCI) synchronously windowed speech frame, all the zeros outside the unit circle are mainly due to the first phase of the glottal flow (i.e. glottal flow without a return phase) and the zeros inside the unit circle are mainly due to the vocal tract filter (which also includes the spectral tilt component). In Fig. 3, ZZT representation for a GCI synchronously windowed synthetic speech frame is presented. The synthetic speech frame is synthesized by filtering an LF pulse [7] with an all-pole filter response (with resonances at 600Hz, 1200Hz, 2200Hz and 3200Hz).

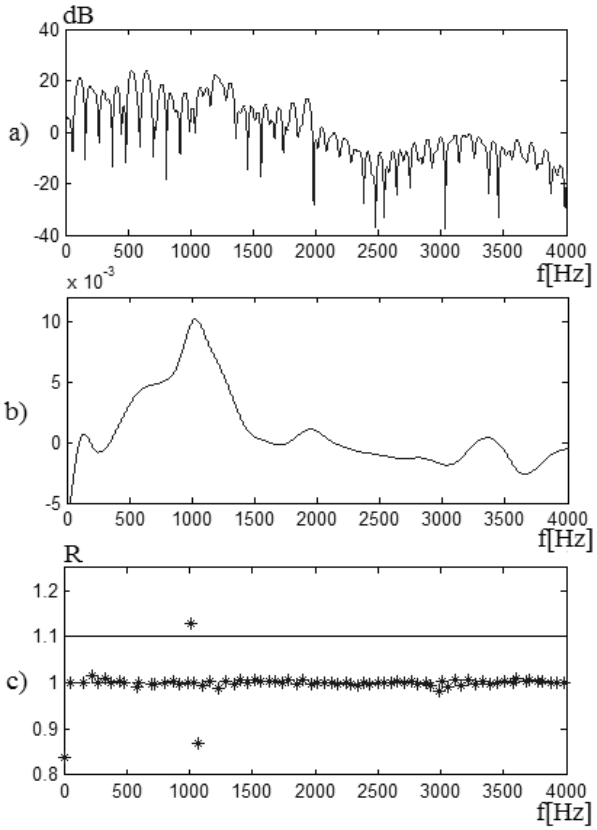

Figure 2: ZZT effect on differential phase spectrum, a) amplitude spectrum of a real speech frame, b) differential phase spectrum computed on R=1.1, c) ZZT representation on z-plane (analysis circle indicated by a line at R=1.1)

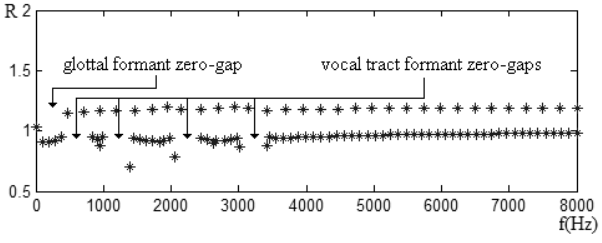

Figure 3: ZZT representation of GCI synchronously windowed synthetic speech frame.

The existence of separate grouping of zeros on two sides of the unit circle made it possible to design a decomposition algorithm which we named ZZT-decomposition [6]. ZZT-decomposition is simply based on splitting the set of zeros in the ZZT representation into two subsets according to their

position with respect to the unit circle: outside zeros are classified as glottal flow zeros and inside zeros are classified as vocal tract zeros.

As discussed in the previous paragraphs, the inefficiency of the differential phase spectra processing is due to the sparseness of zeros on the z-plane. Using the ZZT-decomposition as a priori step, we can guarantee non-existence of zeros on one side of the unit circle. For example, differential phase spectrum calculation can be performed on a circle outside the unit circle after ZZT-decomposition and removal of zeros outside the unit circle. Therefore, we have modified our differential phase spectra processing algorithm for formant tracking and added the initial step of ZZT-decomposition. In addition, the discrete chirp-z transform is calculated directly from the set of ZZT obtained by ZZT-decomposition by Eq. 4. (*N* is the number of zeros, *G* is the gain factor and *Zm* are the zeros).

$$X(re^{j\varphi}) = Gr^{(-N+1)}e^{(j\varphi)(-N+1)}\prod_{m=1}^{N-1}(re^{j\varphi} - Z_m) \quad (4)$$

Once the z-transform is calculated on a circle outside the unit circle with radius *R=1+deltaR*, the closest zero is at least *deltaR* away from the analysis circle since all zeros are inside the unit circle. As *deltaR* increases, the differential phase spectrum gets smoother. Defining an optimum value is a compromise between smoothness and high resolution and *deltaR=0.05* is amprically found to be an appropriate value.

3. Modified Differential-Phase Peak Tracking (DPPT) algorithm for formant tracking

In Fig. 4, we present the new DPPT algorithm. It is simply composed of sequentially performing ZZT-decomposition, differential phase calculation outside the unit circle from ZZT inside the unit circle and peak picking.

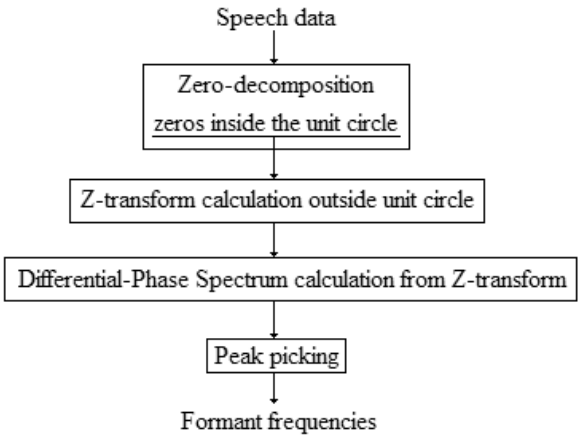

Figure 4: The DPPT algorithm

4. Comparative testing of DPPT

Finally, we have tested our algorithm for formant tracking on both synthetic speech signals and real speech signals. DPPT is compared to two publicly available formant trackers : that of Praat [8] and WinSnoori [9]. The three formant trackers were run on a synthetic speech chunk and four real speech examples (2 male and 2 female). All outputs together with the wave files are presented at : <http://www.tcts.fpms.ac.be/demos/zzt/formantTrackTesting/testingdppt.htm>

For the tests with synthetic speech, a single synthetic speech chunk with pitch frequency and formant frequency variations, which somehow uniformly samples the f0-F1-F2-F3-F4 parameter space is designed. The parameter space plots are available on <http://www.tcts.fpms.ac.be/demos/zzt/formantTrackTesting/prmspc/parameterspaces.htm >

Speech is synthesized by all-pole filtering of a periodic excitation signal. The excitation signals is created by using the LF model [7] with fixed open quotient (0.65) and asymmetry coefficient (0.7) and f0 is varied from 200Hz to 100 Hz by a sinusoidal function.

The formant tracks obtained for this synthetic speech signal by DPPT is presented in Fig. 5 together with the actual formant tracks used in synthesis. In addition, average percentage error rates and formant missing rate for the outputs of the three systems are provided in Table 1. Average percentage error rates are calculated only for the formants, which are not missed (i.e. a formant missed does not contribute to the error rate by 100% but is simply not included in calculations). Plots for outputs from two other formant trackers are available on the web page referred above.

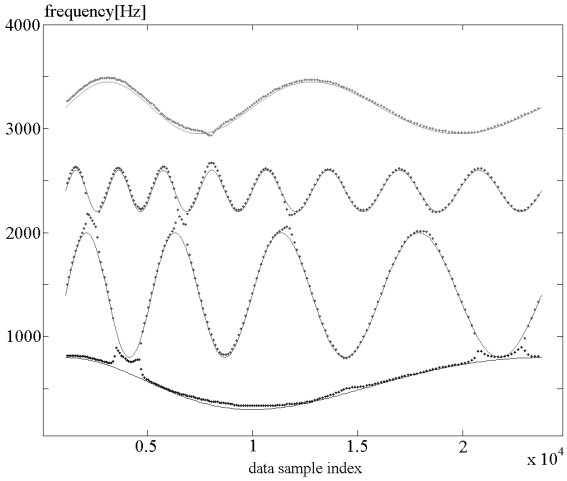

Figure 5: DDPT formant tracks (dots) and formant synthesis parameters (solid lines)

Table 1: Formant tracking error rates for DPPT, Winsnoori (WSN) and Praat

	Average percentage error				Formant miss rate			
	F1	F2	F3	F4	F1	F2	F3	F4
DPPT	6.8	1.8	1.0	0.8	0	17.1	3.5	0
WSN	2.8	1.9	0.6	-	0	0	0	100
Praat	3.8	3.8	4.7	13.8	0	0	0	24.4

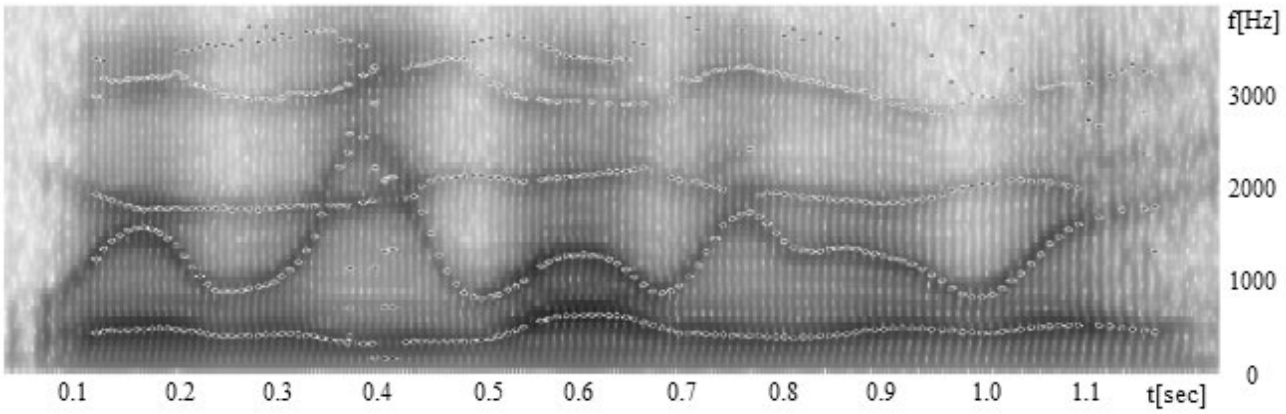

Figure 6: Differential phase spectrum peaks indicated on spectrogram for "where were you while we were away?" from The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, NIST Speech Disc CD1-1.1, DR1/msjs1/sx9.wav

For demonstrating the results of tests with real speech, we provide only plots but no error rates due to unavailability of reference data for formant frequencies of real speech chunks. Here, we provide the DPPT peaks picked for one of the real speech examples (Fig. 6), for which the formant tracks on spectrogram are obvious. (All the outputs from the three systems for the four real speech examples are presented on the web page referred above.)

5. Discussion and Conclusion

It is interesting to observe that the robustness of methods for tracking formants on real speech and synthetic speech are quite different. In tests with synthetic speech, Winsnoori results are best for the first three formants. But the fourth formant track is not available so we cannot compare the results for F4. DPPT's quality is very close to Winsnoori's except in frames where formants get very close to each other (can be easily observed on Fig. 6) and it provides the F4 track with high precision. For frames where formant frequencies are very close, DPPT tracks a single peak instead of two peaks which causes a high formant miss rate for F2. This seems to be the most important drawback of the DPPT algorithm. Praat's robustness on analysis of synthetic speech is lowest except for the F1 track.

However, the robustness of the three methods for analyzing real speech are quite different than that for synthetic speech. For three of the four real speech examples, DPPT is either the best or among the best two. But it gives the worst results among the three techniques for the fourth example. Praat's quality seems to be more constant than the other two methods when all four examples are considered. Winsnoori is effective mainly for tracking F1. It has moderate quality in tracking F2 and fails to provide reliable F3 estimates for most the frames (and cannot provide F4 track).

The results show that DPPT is an effective method for formant tracking. Its main advantage is in tracking high order formants and the main reason for this is the spectral tilt-free property of the differential phase spectra. The main drawbacks are: need for GCI marking and high computational load due to need for calculation of roots of high degree polynomials. However we think that fast algorithms can be designed for DPPT and we are currently working on this. In addition, a post-processing for the peaks picked would be useful, so we also target designing a post-processing algorithm as the final step for DPPT.

6. Acknowledgements

This research was realized during Baris Bozkurt's visit to Limsi-CNRS/Paris. Baris Bozkurt was funded by Region Wallonne, Belgium, grant FIRST EUROPE #215095.

7. References

[1] Bozkurt, B. and Dutoit, T., "Mixed-Phase Speech Modeling and Formant Estimation, Using Differential Phase Spectrums,", *Proc. ISCA ITRW VOQUAL 2003,* Geneva, Switzerland, Aug. 2003, pp. 21–24.

[2] Murthy, H.A. and Yegnanarayana, B., "Formant extraction from group delay function", *Speech Communication*, vol.10, no.3, pp. 209-221, August 1991.

[3] Doval, B., d'Alessandro, C. and N. Henrich, "The voice source as a causal/anticausal linear filter,", *Proc. ISCA ITRW VOQUAL 2003,* Geneva, Switzerland, Aug. 2003, pp. 15–19.

[4] Schafer, R.W. and Rabiner, L.R., "System for automatic formant analysis of voiced speech", *Journal of Acoustical Society America*, vol47, no:2, pp:634-648,1970.

[5] Kawahara, H., Atake, Y., and Zolfaghari, P., "Accurate vocal event detection method based on a fixed-point to weighted average group delay", *Proc. ICSLP*, pp. 664–667, Beijing, 2000.

[6] Bozkurt, B., Doval, B., d'Alessandro, C. and T. Dutoit, "Zeros of Z-Transform (ZZT) decomposition of speech for source-tract separation", Submitted to ICSLP 2004, Jeju Island, Korea.

[7] Fant, G., "The LF-model revisited. Transformation and frequency domain analysis", *Speech Trans. Lab.Q.Rep., Royal Inst. of Tech. Stockholm*, vol.2-3, pp 121-156,1995.

[8] <http://www.praat.org)>

[9] <http://www.loria.fr/~laprie/WinSnoori/>