A Method For Glottal Formant Frequency Estimation

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 a method for estimation of glottal formant frequency (Fg) from speech signals. Our method is based on zeros of z-transform decomposition of speech spectra into two spectra : glottal flow dominated spectrum and vocal tract dominated spectrum. Peak picking is performed on the amplitude spectrum of the glottal flow dominated part. The algorithm is tested on synthetic speech. It is shown to be effective especially when glottal formant and first formant of vocal tract are not too close. In addition, tests on a real speech example are also presented where open quotient estimates from EGG signals are used as reference and correlated with the glottal formant frequency estimates.

1. Introduction

Glottal flow parameter estimation is one of the key issues in speech analysis and considerable effort has been made for long years to design efficient and robust methodologies. However, due to difficulty of the source-tract separation problem, efficiency of even the most recent state of the art methods are limited. Therefore, there is still much room for studying new methodologies and approaches in this area.

This study presents a spectral parameter estimation method for one of the most important features of the glottal flow: the frequency of glottal formant. In [1], it has been discussed that the amplitude spectrum of the glottal flow can be considered to have two main components: the so-called glottal formant and the spectral tilt. The glottal formant is named as such though it does not correspond to an acoustic resonance. However, the spectrum of the first phase of the glottal flow (i.e. glottal flow without a return phase) has an asymptotic behavior which is very close to that of a second order linear filter. By studying such behavior of glottal flow models, it has been shown [1] that, the glottal flow can be modeled as a causal/anticausal linear filter where the first phase of the glottal flow constitutes the anticausal part (forming the glottal formant) and spectral tilt constitutes the causal part. This model is called the Causal-Anticausal Linear Model (CALM). For this glottal flow model, the frequency of glottal formant is exactly inversely proportional to the open quotient and the quality factor is controlled by the asymmetry coefficient, which is defined as the instant of maximum of the glottal flow relative to the period and open quotient.

In [2], we have presented a method for estimation of all formant frequencies for speech signals including the glottal formant frequency (Fg). The algorithm presented is based on processing peaks of differential phase spectra calculated at circles other than the unit circle. It has been shown that the technique is capable of tracking causal and anticausal resonances of mixed-phase speech signals in tests with synthetic speech. However, the tests performed on real speech showed that the method was not robust and the dependencies were unclear. Our further efforts on the same path made it possible to understand the theoretical background of the problem and improve our methodology. Our tests showed that improved differential phase peak picking is an efficient method for vocal tract formant tracking but is comparatively less robust for glottal formant tracking. In this paper we will present an alternative method based on processing of glottal flow dominated amplitude spectra obtained by ZZT-decomposition, a spectral decomposition method for source-tract separation [3]. This paper presents new advances on this technology.

In Section 2, we summarize our method to obtain glottal flow dominated spectrum from speech signals. We show on a synthetic speech example that the obtained glottal flow dominated amplitude spectrum is very close to the actual glottal flow amplitude spectrum and have the same spectral peak corresponding to the glottal formant. In Section 3, we present the new version of our algorithm for glottal formant frequency (Fg) estimation. Section 4 presents some tests on synthetic and real speech.

2. Extraction of glottal flow dominated spectrum from speech

2.1. Motivation for studying zeros of z-transform

In our previous method for Fg estimation[2], the main motivation for processing differential phase spectra computed on circles other than the unit circle was to get rid of spiky noises created by zeros of z-transform (roots of the z-transform polynomial) which mask formant peaks on group delay functions. However, our observations on real speech signals showed that we cannot get rid of all zero effects by just calculating z-transform at some distance from the unit circle. The zeros can appear nearly everywhere on the z-plane though most of them are closely located to the unit circle and even a single zero close to the analysis circle introduces wrong estimates in differential phase spectrum peak picking based estimation. For this reason we have performed a systematic study of zeros (roots) of z-transform for windowed speech signals. A summary of the study is presented in [3].

The study on zeros of z-transform (ZZT) [3] 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. 1, ZZT representation for a GCI synchronously windowed synthetic speech frame is presented. The synthetic speech frame is synthesized by filtering an LF model [4] glottal flow derivative pulse with an all-pole filter response (with resonances at 600Hz, 1200Hz, 2200Hz and 3200Hz).

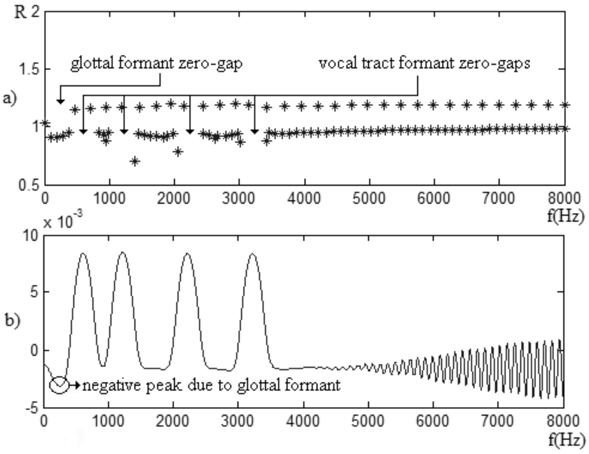

Figure 1: Spectral representations of GCI synchronously windowed synthetic speech frame, a) ZZT representation, b)group delay function.

The zeros are aligned on two sides of the unit circle with gaps corresponding to the formants as labeled on the Fig. 1. The zero gap located outside the unit circle (labeled as the glottal formant zero-gap), creates an anti-causal resonance-like peak in the speech spectrum. A weak spectral peak exists (though it cannot be observed most of the time due to relative strength of F1 peak) in the amplitude spectrum and a negative peak is observed in the group delay function (Fig. 1b), like the effect of an anti-causal pole at the frequency of the gap). This observation is in accordance with the theoretical model of CALM [1]. The main aim of this study is to estimate the frequency of this peak, which is named as glottal formant peak.

2.2. Zero-decomposition for estimation of glottal flow dominated spectrum

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 [3]. 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. Two spectra can be calculated from the two sets of zeros: glottal flow dominated spectrum and vocal tract dominated spectrum. Calculation of DFT for each group of zeros is straightforward using the Eq. 1 (N is the number of zeros, G is the gain factor and Zm are the zeros).

$$X(e^{j\varphi}) = Ge^{(j\varphi)(-N+1)} \prod_{m=1}^{N-1} (e^{j\varphi} - Z_m) \qquad (1)$$

In Fig. 2, we present the ZZT-decomposition result for the signal in Fig. 1. For comparison, the estimated glottal flow dominated spectra is plotted together with the actual spectrum of the glottal flow excitation and the estimated vocal tract dominated spectra is plotted together with the actual vocal tract filter response used for synthesis. The decomposition results are high quality though not complete. Small variations due to vocal tract formants are observable on the glottal flow dominated signal amplitude spectrum. But still, glottal formant frequency detection can be performed effectively on this amplitude spectrum.

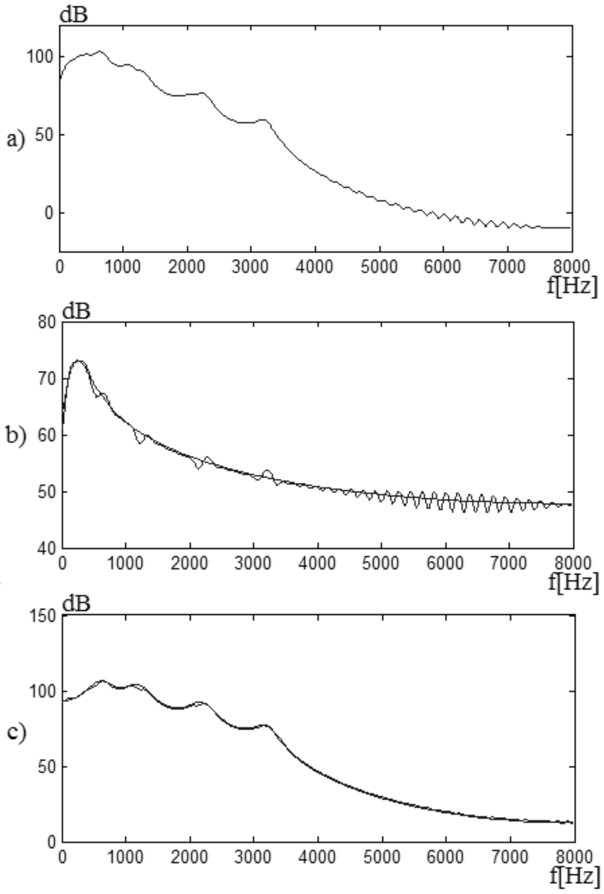

Figure 2 : ZZT-decomposition results. a) amplitude spectrum of windowed speech, b) amplitude spectrum of glottal flow dominated spectrum superimposed on the amplitude spectrum of windowed actual glottal flow signal, c) amplitude spectrum of vocal tract dominated spectrum superimposed on the amplitude spectrum of actual vocal tract response.

The most important detail for ZZT-decomposition algorithm is a single zero on the real axis due to the anti-causal portion of the signal, which in some cases falls inside the unit circle. Actually, including or excluding this single zero in the glottal flow zero set has a great impact on the frequency location of the maximum peak of the glottal flow dominated spectrum. At the moment no governing rule has been found for the location of this zero (inside or outside?) and an engineering approach is used for this problem; if no zero has been found on the real axis in the range R=[1 1.1], then the closest zero on the real axis to point (R=1, φ=0) is removed from the set of zeros inside the unit circle set and put in the set of zeros outside the unit circle.

In addition, in most of the examples with real speech data, we observed that the alignment of zeros on two sides of the unit circle is up-to 4000Hz for 16000Hz sampled signals. For this reason ZZT-decomposition is effective mainly in the 0-4000Hz range. However, this limitation is not a problem for estimation of glottal formant frequency.

3. Glottal formant frequency (Fg) estimation

3.1. The algorithm

Tracking the maximum valued peak location of the amplitude spectrum calculated from zeros outside the unit circle, we can easily get an estimate for Fg. In the next section, we test this simple parameter estimation algorithm on synthetic speech. Due to lack of reference methods and data, systematic tests on real speech cannot be performed, but we present an output for a real speech example and show correlation of our Fg estimate to the open quotient estimate obtained by processing EGG signals recorded in parallel to speech as described in [5].

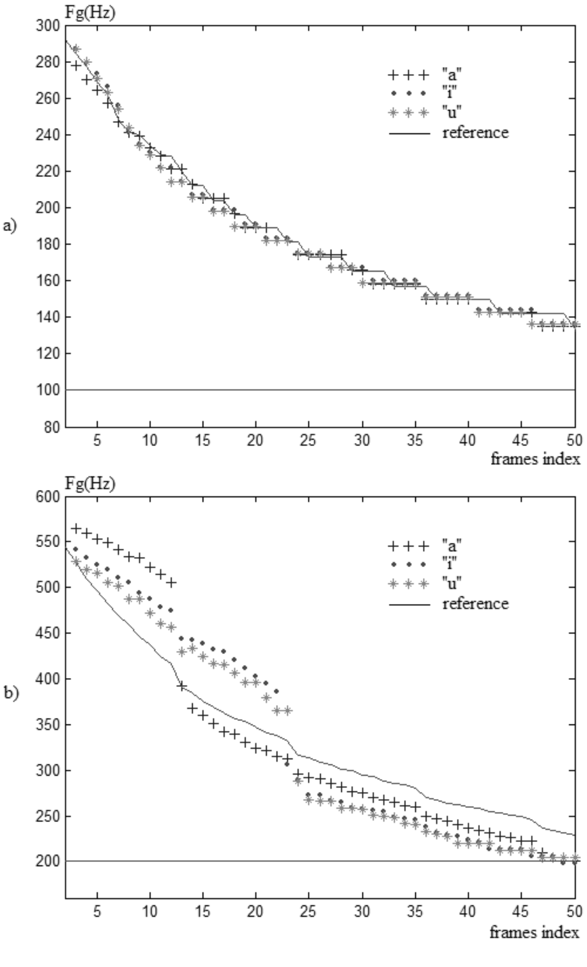

Figure 3: Test results, a) f0=100Hz, b) f0=200Hz

3.2. Testing Fg estimation

3.2.1. Tests with synthetic speech

As discussed in [1], Fg is inversely proportional to the open quotient. Our testing procedure is built on this relation. As test signals, two periodic excitation (glottal flow derivative) signals are synthesized with the LF model [4] at constant pitch frequencies: 100Hz and 200Hz. All the parameters of the glottal flow have been kept constant except the open quotient, which is varied linearly in the range (0.3 0.98). No spectral tilt component is included for simplicity (Ta, the return phase decaying exponential time coefficient is set to zero) and the asymmetry coefficient is set to 0.7. These two excitation signals are then passed through three second order resonant filters for the vowel formants; /a/ (F1=600Hz, F2=1200Hz, F3=2200Hz, F4=3200Hz), /i/ (F1=300Hz, F2=1800Hz, F3=2200Hz, F4=3200Hz) and /u/ (F1=300Hz, F2=800Hz, F3=2200Hz, F4=3200Hz), thereby obtaining 6 synthetic speech signals. Then Fg is estimated using the proposed algorithm on all the signals (in addition, peak picking is performed on amplitude spectrum of the pure excitation signals to obtain a reference estimate). Since Fg is independent from formant variations, we expect to obtain the same Fg estimates for all synthetic vowels created and the glottal flow signal itself. The results are presented in Fig 3.

As expected, the Fg estimate plots have the basic form $y=1/x$ since open quotient is linearly varied and Fg is inversely proportional to open quotient. The robustness of the estimation depends on the relative location of glottal formant to the first formant of the vocal tract (F1). For the frames where Fg is lower than 300Hz, the estimates for the vowels compared to the reference estimate are very close. For higher Fg values, the maximum peak location of the amplitude spectrum corresponding to zeros outside the unit circle is more affected by the ripple due to incomplete separation of F1, and peak picking is sensitive to this effect. The ZZT-decomposition result for the worst estimation in the test is shown in Fig 4.

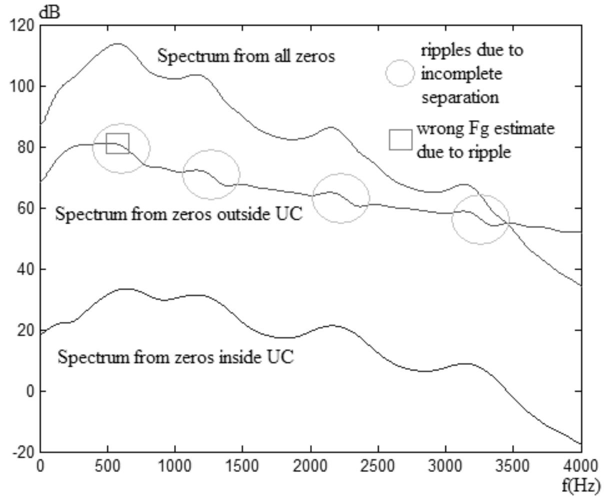

Figure 4: Amplitude spectrums of ZZT-decomposition for the worst glottal formant frequency estimation with peak picking (12th frame from the f0=200Hz test where the vowel is “a”, Fig. 3b)

The ripples due to incomplete separation of vocal tract from glottal flow are marked with circles on the glottal flow dominated spectrum. The ripple due to F1 causes the maximum valued point of the spectrum to appear at a higher frequency (marked with a rectangle) introducing an error in the Fg estimate. To improve robustness of our algorithm, we plan to apply a curve fit method instead of peak picking in our future studies.

3.2.2. Test on real speech

The Fg estimation method has also been tested on a real speech signal for which we could obtain a reliable open quotient estimate. A sustained vowel "a" with flat pitch and decreasing open quotient has been uttered by one of the authors and EGG signals were recorded in parallel. Using the method described in [5], open quotient (*Oq*) estimate was obtained. As defined in [1] with the Eq. 2, Fg location is also dependent on the asymmetry coefficient (αm) and the pitch period (*T0*).

$$F_g = \frac{f_g(\alpha_m)}{O_q T_0} \qquad (2)$$

However, as mentioned in [1], the effect of the asymmetry coefficient variation to Fg variation is rather minor. For this reason, for our speech example with almost constant pitch, we expect the Fg estimate to be highly correlated with the inverse of the open quotient estimate. Below, in Fig. 5, we present the Fg estimate plotted with f0 and inverse of the open quotient estimate scaled with a constant.

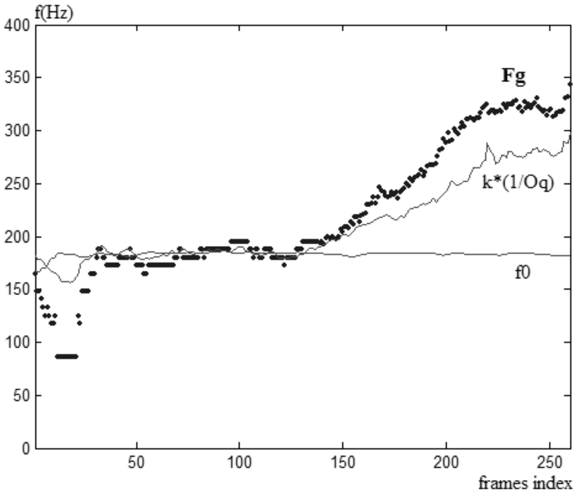

Figure 5: Comparing glottal formant frequency estimate with inverse of open quotient estimate and f0 estimate (k=115)

The glottal formant frequency estimate and inverse of open quotient estimate plotted in Fig. 5 have high correlation which indicates that our algorithm is not only effective on synthetic signals but can tract glottal flow variations of real speech signals.

4. Discussion and Conclusions

We have presented a method for estimating glottal formant frequency. It is mainly composed of picking the maximum valued peak on the amplitude spectrum of glottal flow, obtained by ZZT-decomposition [3] from GCI synchronously windowed speech signals.

The ZZT-decomposition is of high quality though not complete. Small amplitude ripples are observed on the glottal-flow-dominated spectrum due to vocal tract formants. Equally, the vocal tract dominated spectrum is also affected by the glottal flow but these effects are hardly observed.

The proposed algorithm is tested on synthetic speech and the results show that the ripples on the glottal flow spectrum due to incomplete separation introduces errors when Fg and F1 values are close. It is also the sensitivity of peak picking to small ripples which causes the wrong estimates. For our future studies we target improving the quality of the Fg estimation method by replacing peak picking with a more robust method.

The proposed algorithm is very easy to implement but computationally heavy due to the need of finding roots of high degree polynomials. A second disadvantage is the need for glottal closing instant marking, which in fact is also needed in many other glottal flow parameter estimation methods. The reader should refer to [6] for a reliable GCI marking algorithm.

5. 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. The authors would like to thank Nathalie Henrich for providing the recorded speech file and the corresponding open quotient estimate from EGG for the glottal formant frequency estimation test.

6. References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Henrich, N., Doval, B., d'Alessandro, C. and M. Castellengo, "Open quotient measurements on EGG, speech and singing signals", *Proc. 4th Int. Workshop on Advances in Quantitative Laryngoscopy, Voice and Speech Research,* Jena, Apr. 2000.

[6] 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.