Glottal Parameters Estimation on Speech Using the Zeros of the Z-Transform

Nicolas Sturmel[1], *Christophe d'Alessandro* [1], *Boris Doval*[2]

[1] LIMSI-CNRS, B.P. 133, F-91403, ORSAY FRANCE
[2]LAM-IJLRA, UPMC Univ Paris 06, 11 Rue de Lourmel, F-75015, PARIS FRANCE
nicolas.sturmel@limsi.fr, cda@limsi.fr, boris.doval@upmc.fr

Abstract

This paper presents a method for the joint estimation of the open quotient and the asymmetry quotient of the open phase of the glottal flow on speech. An algorithm based on a source/filter decomposition (the Zeros of the Z Transform - ZZT) is presented. This algorithm is first tested on a database of sustained vowels spoken at different voice qualities, then on running speech. Results are evaluated in comparison to the value of the open quotient obtained by analysis of the synchronous ElectroGlottoGraphic (EGG) signal. Results of this test show that open quotient is estimated within the just noticeable difference from the EGG reference in more than 60% of the cases, and that 75% of the estimations give a value within 25% of the reference. The estimation results on asymmetry are also discussed and confirm previous studies.

Index Terms: ZZT, Inverse Filtering, Glottal Flow, Open Quotient

1. Introduction

Voiced speech analysis is partly based on the estimation of glottal parameters [1]. One of them is the ratio between the vocal folds open phase duration T_e and the speech period T_0: the open quotient $O_q = \frac{T_e}{T_0}$. It is valued in both source production, voice quality and speech synthesis for unit selection. In expressive speech analysis, O_q caries a lot of information. Tensed voice is often associated to low values - around 0.3 - and laxer voice is associated to high values - of typically 0.7 to 0.9. On the other hand, the effect of the asymmetry coefficient ($\alpha_m = \frac{T_p}{T_e}$ [2]) has been scarcely studied, due to fact that it is hardly measurable on real speech. Previous study in [3] showed that the speed quotient, which reflects the asymmetry of the glottal flow, increases with vocal effort. The LF model of the glottal flow and its derivative are presented on figure 1, the equivalence between the time based parameters and normalized parameters is given.

The easiest way to measure O_q is the use of electroglottographic signals (EGG). EGG is a direct measure of the vocal fold activity by mean of electrical conductance. If analysis on EGG is carefully done, it is a reliable way of estimating O_q on real speech data ([4]). When EGG data is not available, various methods have been developed to estimate this parameter :

A first approach is inverse filtering using linear prediction, and parametrisation of the estimated glottal flow. Many different approaches have been studied, for instance in [5]. But linear prediction inverse filtering needs a lot of tuning for an accurate estimation of the glottal flow. This measure gives an hybrid parameter correlated with the open quotient O_q, and then also sensitive to other parameters such as the asymmetry coefficient.

A second approach is the detection of singularities in speech signal caused by the opening and the closing of the vocal folds. Multi-scale product of the wavelet transform can estimate these opening and closing instants [6]. The testing protocol gave 76% detection of the opening instant within a 0.25ms window, no data was available for O_q.

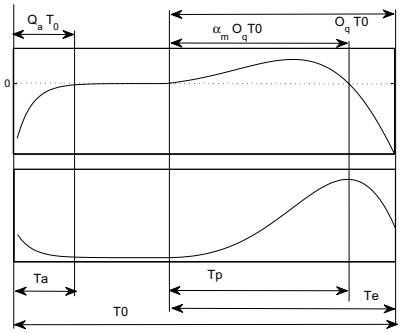

Figure 1: The LF model of the glottal flow (GF) [1]. Top figure: derivative GF; bottom figure: GF. Top arrows: time parameter; bottom arrows: normalized coefficients

Recent study on the causal/anti-causal model of the glottal flow allowed the use of the Zeros of the Z Transform (ZZT [7]) analysis to obtain the glottal flow spectrum. ZZT computes the roots of a windowed speech frame and separates the causal and anti-causal parts based on their absolute value (inside or outside the unit circle). The anti-causal zeros are describing the open phase of the glottal flow. These zeros can be used either for computation of the glottal flow signal. Work done in [8] used both ZZT and inverse filtering combined with non linear prediction of the parameters of the glottal flow model. This work was not based on solid comparison of the parameters estimated on real speech, but rather on a perceptive test.

The present work uses ZZT analysis of speech combined with a general model of the glottal flow for simple and robust estimation of both the open quotient and the asymmetry quotient. Since ZZT needs precise estimation of the glottal closure instant, a wavelet based method (LOMA [9]) is used for their estimation. First the mathematical relation allowing estimation of O_q and α_m from analysis by ZZT will be presented and the associated algorithm will be detailed. This algorithm will then be tested on a running speech database. There is no reference method available for estimation of the asymmetry coefficient but open quotient reference on real speech will be extracted from synchronous EGG recordings. The validation of the estimation will then allow discussion on the quality of the estimation of α_m in regard to voice quality.

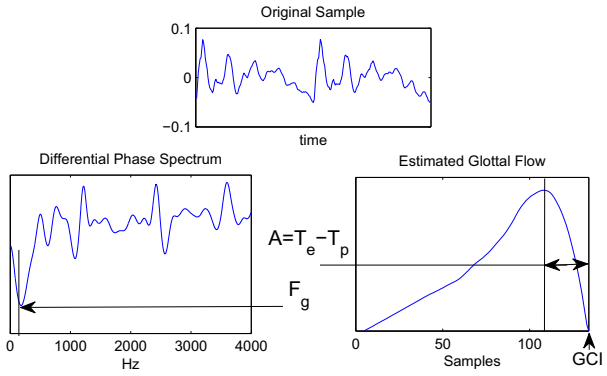

Figure 2: Illustration of the two steps of the algorithms involving the measure of F_g and α_m on a speech frame.

2. Estimation of the parameters

The open phase of the differential glottal flow (DGF) seen on figure 1 follows the expression ($a > 0$) :

$$G_o(t) = sin(2\pi F_g t)e^{at} \quad (1)$$

F_g is the frequency of the glottal formant, the maximum on the spectrum of the differential glottal flow. It directly gives the value of T_p, as T_p is the time between the opening instant - start of the opening phase - and the first zero crossing of the differential glottal flow ; it is equal to half the oscillation period of the glottal formant. Therefore, $T_p = \frac{2}{F_g}$. Glottal formant frequency F_g is estimated on the anticausal part of the ZZT decomposition as seen on figure 2, by looking for a minimum on the derivative phase spectrum [10]. The critical factor for a successful estimation is the radius of the circle used for the computation of the phase spectrum on the complex plane. The closest to 1, to sharper will be the phase spectrum, and therefore the estimation, but the higher the risk will be that incomplete decomposition lead to erroneous detection. This, for instance, may be caused by the the phase ripples of incompletely separated vocal tract formants.

Solely knowing the glottal formant frequency gives the ability to approach O_q as shown in [10] but with low accuracy. Because the frequency of the glottal formant depends also on the value of α_m. In this case, variation of this parameter could lead to mis-estimation of O_q. Information added to F_g for complete estimation of O_q and α_m are extracted from the glottal flow. Two of the most reliable points of the estimated glottal flow with the ZZT are chosen : the maximum of the glottal flow and the glottal closing instant (GCI), already known as it is basically a parameter of the decomposition. Close to the glottal closing instant, the maximum of the glottal flow belongs to the part the most accurately estimated by the method. The time between the GCI and the maximum of the glottal flow gives the knowledge of $A = T_e - T_p$, showed on figure 2.

One can then reach the values of O_q and α_m as follows :

$$O_q = \frac{T_e}{T_0} = \frac{A + \frac{2}{F_g}}{T_0} \quad (2)$$

$$\alpha_m = \frac{T_p}{T_e} = \frac{T_p}{A + \frac{2}{F_g}} \quad (3)$$

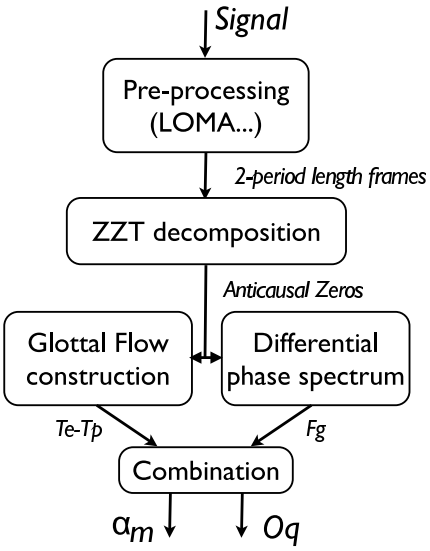

Figure 3: Algorithm used for the joint estimation of α_m and O_q

3. Algorithm

The schematic algorithm of the estimation is given on figure 3 . Each of the steps are presented hereafter :

1. First step : preprocessing is done on the Signal. GCI are estimated using LOMA [9] analysis. Although the placement of the GCI is critical for the decomposition, LOMA was only used to refine the placement from the GCI extracted from EGG.

2. Second step : ZZT source/filter decomposition. The signal is cut in two-period overlapping frames. Following the study done in [11], those frames are weighted using the window of expression (for a N-length window) :

$$w_\beta(n) = \frac{\beta}{2} - \frac{1}{2}cos(2\pi\frac{n}{N}) + \frac{1-\beta}{2}cos(4\pi\frac{n}{N}) \quad (4)$$

With $\beta = 0.7$. The vector obtained is then considered as a polynom and its root are computed. Roots are separated by their relative position to the unit circle on the complex plane. The part used here - the anticausal part - consists of the roots of the polynom located outside the unit circle. This set of zeros allows computation of the so-called anticausal spectrum of the glottal flow. The derivative phase spectrum is also computed on a given circle of radius ρ on the complex plane.

3. Third step : The maximum is found on the glottal flow - right of figure 2. The frequency of the glottal formant is estimated on the differential phase spectrum - left of figure 2. These two values are combined to compute both O_q and α_m following equations 2 and 3.

4. Additional post processing (not shown on figure 3) is applied to the estimated values in order to prevent gross errors. This post processing mainly involve adjustment on the ZZT as explained in previous papers.

Table 1: Results on the varying voice quality database. Two speakers for three vowels and three voice qualities. O_q estimated with the proposed method, $\hat{O}_q$ is the reference from EGG. α_m is given as the mean of the estimations paired with O_q within the JND.

#	Vow.	Quality	O_q	$\hat{O}_q$	JND	25%	α_m
M1	/a/	normal	0.61	0.61	94%	95 %	0.62
M2	/i/	normal	0.50	0.49	66%	76%	0.60
M3	/u/	normal	0.52	0.50	85%	97%	0.60
M4	/a/	tensed	0.46	0.39	43%	86%	0.67
M5	/i/	tensed	0.41	0.39	90%	97%	0.62
M6	/u/	tensed	0.51	0.38	3%	13%	0.73
M7	/a/	lax	0.71	0.75	69%	82%	0.70
M8	/i/	lax	0.79	0.68	53%	74%	0.64
M9	/u/	lax	0.71	0.67	85%	88%	0.53
F1	/a/	normal	0.47	0.44	42%	70%	0.68
F2	/i/	normal	0.39	0.48	20%	31 %	0.90
F3	/u/	normal	0.47	0.49	65%	71%	0.77
F4	/a/	tensed	0.42	0.42	69%	80%	0.67
F5	/i/	tensed	0.34	0.35	50%	74%	0.87
F6	/u/	tensed	0.34	0.29	58%	79%	0.66
F7	/a/	lax	0.71	0.72	88%	93%	0.68
F8	/i/	lax	0.60	0.68	58%	74%	0.91
F9	/u/	lax	0.66	0.68	32%	47%	0.78

4. Experimental results

4.1. Sustained Vowel Database

A database of varying voice quality sustained vowel has been recorded with synchronous EGG signal. It is composed of tensed, modal and lax vowels spoken by two subjects: male - MX samples - and female - FX samples. A total of 9 samples per subject was used for the evaluation, combination of three vowels (/a/, /i/, and /u/) and three voice qualities. Average fundamental frequency was 130Hz for the male speaker and 250Hz for the female speaker. Samples where chosen given the accuracy and stability of $\hat{O}_q$ estimated on the EGG signal. Sample rate is 16kHz and recording was done in a quiet environment using a condenser microphone.

4.2. O_q estimation results

On table 1 are given the results for each sample of this database. Voice quality and mean open quotient from EGG : $\hat{O}_q$ are indicated on columns 2 and 3. Two successful estimation rates are given, one within the just noticeable difference (JND [12]) of 17% on column 6 and one for an error bellow 25% of the actual value of $\hat{O}_q$ at the given time on column 7. 25% is the maximum acceptable error to separate O_q values of 0.3, 0.5 and 0.8 and keep discrimination of voice qualities. Mean values for O_q is also given. This value represents the mean on the whole set of estimation, unlike the mean of α_m presented later.

On some cases, a precise adjustment of the parameter ρ (radius of the DPS computation) was needed. But most of the time, a variation between 0.9 and 1 of this parameter had little to no effect on the estimation success rate.

4.3. α_m estimation results

Because there is no possibility to validate the estimation of α_m, It is considered valid only when the error on O_q is low enough. On the last column of table 1 are presented the estimation results for α_m. The value given is the mean of the values paired with the O_q within the JND error range from the EGG reference.

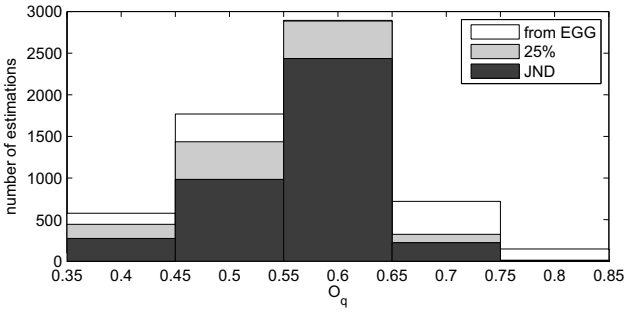

Figure 4: Histogram of the successful estimations on the running speech database. White: EGG, light gray: 25% error, dark gray: JND.

4.4. Test on spoken real speech

In order to support the results presented before, a short database of running speech was also analysed. It included two speakers (male and female) reading french newspaper, totalizing 57 seconds of speech and 6715 speech periods. Because running speech is less stationary than sustained vowels, estimated values of O_q where smoothed with a unity gain, averaging filter on 10 consecutive values to lower the effect of unstable source/filter decompositions. Results give 60% successful estimations for a value of O_q inside the JND and 75% inside a 25% error range. On figure 4 are presented the estimation histogram for five range of values in O_q for the reference from EGG and the two error range.

5. Discussion

Overall results show that the estimation of the open quotient with the proposed method is competitive: on a large part of the database, the method performs well at estimating O_q both in terms of mean on the signal than in terms of success rate. On five of the 18 tested samples, 80% of the estimated points are within the JND from the value measured on EGG. And 14 of the 18 samples are presenting exploitable values (within 25% of the reference more than 70% of the time) for voice classification. On top of that, even in the case of wide dispersion of the estimated points, mean values are very close from the reference mean value.

Subjective effect of a low value of O_q is hardly noticeable by itself, low values are often coupled with a highly resonant vocal tract filter. This is why most of the bad results are obtained for a tensed voice. The more resonant is the vocal tract filter, the harder is the source/filter decomposition. The effect of the speaker is noticeable. The sample files from the female speaker are exhibiting lower precision among the detection. This is mainly due to the precision when computing both the glottal flow (precision of $\frac{F_0}{F_s}$, decreasing when F_0 increases) and the derivative phase spectrum (spectral resolution also decreasing when F_0 increases). At 16kHz, a fundamental frequency of 250Hz implies a precision of $\frac{1}{64}$ of the period, which can be problematic especially for the low values of O_q where the sampling error can cause itself more than 5% of the total error. This could explain the overall lesser results obtained on the female speaker samples.

Although the tests have been performed on sustained vowel at constant voice quality, providing ease of segmentation between different voice qualities, similar analysis on varying

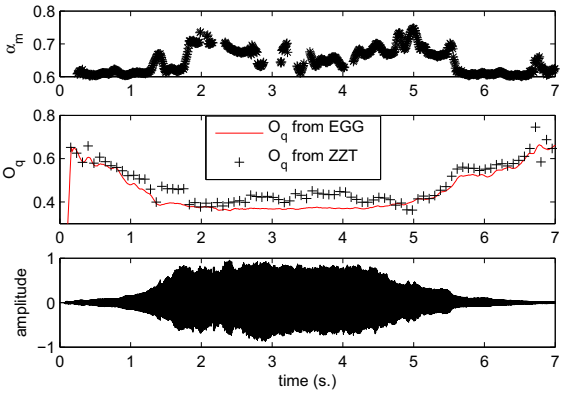

Figure 5: Analysis of a file with variation of the voice quality. Male speaker, vowel /a/, approx. 120Hz F_0.

voice qualities have also been made. Example presented on figure 5 presents the analysis of an /a/ from a male speaker at approximately 120Hz F_0. Top figure shows the reference data $\hat{O}_q$ in solid line and estimated O_q with stars. Bottom figures shows the waveform of the signal. The estimation of O_q follows the EGG reference with some errors. Judging from table 1, it seems that α_m plays little role on the voice quality variation studied here. It may be more related to the vocal effort (perceived loudness) of the speaker than on the stressing factor of the voicing. On figure 5 when the intensity increases, the asymmetry coefficient tends to rise as well and the open quotient decreases. In addition to an increase in RMS power, perception off the loudness is often associated with a decrease of the spectral tilt: increasing α_m increases the high frequency power of the glottal flow spectrum. This could also be a possible explanation of the values obtained for α_m in table 1 for the female speaker on samples F2, F5, and F8. Production of the vowel /i/ may have needed more effort than the other vowel, therefore increasing the value of α_m. Mis-estimation of O_q directly leads to false estimation of α_m as it is can be seen between seconds 1 and 2 of figure 5.

Some voice sample hardly give accurate results. This is caused by an incomplete decomposition and a miss-estimated glottal formant frequency. The algorithm catches on the first formant or one of its ripple in the derivative phase spectrum. One solution is the adjustment of the radius of the DPS ρ between 0.9 and 1, ρ was set to 0.98 except for samples M2, M9, F3-4, F7-8. No value of ρ has been found for samples M4, M6 and F2. For all the other the samples, changing the value of ρ in the 0.92-0.98 range had little to no effect on the success rate of the estimation.

Further validation of the algorithm was done on running speech. Running speech is more problematic in regard to the quality of the source/filter decomposition, because of the variability of the vocal tract filter. Results are comparable to the ones obtained on the sustained vowel database. On figure 4, one can see that the decomposition performs much better for lower values of O_q, below 0.65. Because higher O_q is often associated with additive noise, proper decomposition is harder to achieve. For the 0.55-0.65 range, successful estimation rate is near 100% within 25% error, and above 80% for values of O_q below 0.65. Those results are very promising for further application of the method to expressive speech analysis.

6. Conclusion

A simple method for joint estimation of the open quotient and asymmetry of the glottal flow was presented. Based on the ZZT decomposition, it uses both estimation of the location of the maximum of the glottal flow and the frequency of the glottal formant. This algorithm was tested on sustained vowels covering a wide range of voice qualities with very good results. In more than 70% of the times, O_q is estimated within the just noticeable reference from EGG data. Results were confirmed on a running speech database. Proper estimation of the open quotient with an error inferior to the typical just noticeable difference of 17% is achievable with little adjustment of the algorithm parameters more than 60% of the time on running speech. The proposed method performs better for values of O_q around 0.5-0.6. The values obtained for the asymmetry showed that this parameter has not direct link with the quality of the voice in terms of tenseness but may be linked to the vocal effort similar to what can be found in [3]. These results open the possibility of application of this method to expressive speech analysis. Future work should especially aim at establishing the link between vocal effort and glottal configuration, and studying the role that the asymmetry plays in it.

7. References

[1] G. Fant, J. Liljencrants., and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, pp. 1–13, 1985.

[2] B. Doval, C. d'Alessandro, and N. Henrich, "The spectrum of glottal flow models," *Acta Acustica united with Acustica*, vol. 92, pp. 1026–1046, 2006.

[3] C.Sapienza, E.Stathopoulos, and C.Dromey, "Approximations of open quotient and speed quotient from glottal airflow and egg waveforms : Effects of measurement criteria and sound pressure level," *Journal of Voice*, vol. 12 (1), pp. 31–43, 1998.

[4] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of derivative electroglottographic signals for characterization of nonpatological phonation," *J. Acoust. Soc. Am.*, vol. 115 (3), pp. 1321–1332, 2004.

[5] P. Alku and T. Bäckström, "Normalized amplitude quotient for parametrization of the glottal flow," *J. Acoust. Soc. Am.*, vol. 112 (2), pp. 701–710, 2002.

[6] A. Bouzid and N. Ellouze, "Open quotient measurements based on multiscale product of speech signal wavelet transform," *Research Letters in Signal Processing*, vol. 2007, 2007.

[7] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, "Zeros of z-transform representation with application to source-filter separation in speech," *IEEE signal proc. letter*, vol. 12, pp. 344–347, 2005.

[8] T. Drugman, T. Dubuisson, N. D'Alessandro, A. Moinet, and T. Dutoit, "Voice source parameters estimation by fitting the glottal formant and the inverse filtering open phase," in *EUSIPCO'08, Lausanne, Switzerland*, 2008, p. 4 pages.

[9] N. Sturmel, C. d'Alessandro, and F. Rigaud, "Glottal closure instant detection using lines of maximum amplitudes (loma) of the wavelet transform," in *ICASSP'09, 4 pages*, 2009.

[10] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, "A method for glottal formant frequency estimation," in *International Conference on Spoken Language Processing (ICSLP)*, 2004.

[11] T. Drugman, B. Bozkurt, and T. Dutoit, "Complex cepstrum-based decomposition of speech for glottal source estimation," in *Interspeech09, Brighton, U.K*, 2009, p. 4 pages.

[12] N. Henrich, G. Sundin, D. Ambroise, C. d'Alessandro, B. Doval, and M. Castellengo, "Just noticeable differences of open quotient and asymmetry coefficient in singing voice," *Journal of Voice*, vol. 17, pp. 481–494, 2003.