IMPROVING GLOTTAL WAVEFORM ESTIMATION THROUGH RANK-BASED GLOTTAL QUALITY ASSESSMENT

Elliot Moore II, Juan Torres

Department of Electrical & Computer Engineering
Georgia Institute of Technology, Savannah, Georgia
emoore@gtsav.gatech.edu, juan.torres@gatech.edu

Abstract

Information on the glottal waveform is an important part of many speech applications. However, glottal waveform estimation remains one of the more inexact sciences of speech processing. The work presented here describes an enhancement to a recently presented algorithm by a new technique involving Rank-Based Glottal Quality Assessment (RB-GQA). The basic premise is to investigate potential measures of glottal quality and use these measures to mark the general trends for determining which glottal waveform estimations are better than others. The work presented here is the beginning of a new research initiative to identify robust methods of glottal waveform estimation across genders for use in speaker analysis applications of normal voices (i.e., no voice pathology).
Index Terms: glottal quality, voice source, glottal waveform

1. Introduction

Glottal waveform estimation remains one of the more inexact sciences of speech processing. Traditional theory has relied on the modelling of speech as a cascade of linearly separable filters which allow the glottal shaping filter to be estimated via inverse filtering. Although speech production is not truly a linear process [1], this model has served well for many applications and will be assumed for this work. The primary problem in estimating the glottal waveform via inverse filtering is finding an estimate of the vocal tract that is minimally affected by interaction with the glottal source. Closed-phase analysis attempts to take advantage of glottal waveform mechanics to create better vocal tract estimates. The glottal cycle generally consists of three phases: an opening phase (abduction of the vocal folds), a closing phase (adduction of the vocal folds) and a closed-phase. During the closed-phase, it is assumed that the acoustic speech waveform is independent of glottal resonances and can be modelled as a sum of decaying sinusoids (e.g., linear prediction analysis (LPA)). However, the automatic identification of glottal closure instants (GCI's) is a challenging issue. Various studies have studied identifying GCI's [2], [3], [4], [5]. However, the additional problem of identifying GCI's is that in many cases (e.g., females, emotional stress, etc.) they may not exist. External sensors, such as Electroglottographs (EGG) tend to provide insight into the mechanics of glottal motion during speech production and therefore yield fairly accurate estimates of glottal closure. However, it is necessary to collect data from these sensors concurrently with the acoustic data. A recent algorithm [6] presented a structure for creating glottal waveform estimates without the reliance on precise glottal closure information. The algorithm created several potential estimates per speech frame and then selected the "best" estimate based on a simple first order LPA. We present here an improved decision structure for choosing glottal waveform estimates based on merging information from objective measures of glottal quality.

2. Algorithm Structure

Details for the basic structure of the glottal estimation algorithm used in this study can be found in [6]. The algorithm implemented an iterative scheme that searched areas around residual minimum peaks for vocal tract estimates and subsequent glottal waveform estimations. The main components of the algorithm consisted of the following: 1) Average LP coefficients (covariance method) from multiple analysis windows (length = twice the model order) separated by pitch cycles within the frame; 2) Generate and store glottal estimates; 3) Slide all analysis windows by one sample (number of iterations equaled twice the model order); 4) Analyze all stored glottal estimates with a 1^{st} order LPA and choose the "best" one. While much of this algorithm is still intact for this study, a few minor improvements bare mentioning that involve the method of averaging LP estimates and the size of the search region (affecting the number of stored estimates). It was determined that averaging the LP coefficients from each analysis window was inadequate since it did not necessarily imply an averaged spectrum. Additionally, the pitch cycle from one epoch to the next was not exact and could lead to estimates that were at slightly different phases of the cycle being averaged together. The solution was to treat the first analysis window as the primary analysis window and conduct the following steps: 1) Convert LP coefficients to cepstral coefficients; 2) Search the region covered by the next analysis window for the closest estimate via minimum Euclidian distance between cepstral coefficients; 3) Average the cepstral coefficients together and convert to the final set of LP coefficients. This produced smoother and more accurate vocal tract estimates than previously. Also, in [6], the search region for each analysis window was equal to twice the model order. It was determined that this size was insufficient for speakers who did exhibit long closed-phases so the search region was changed to equal three times the model order. The final improvement to the algorithm involved the implementation of a more robust decision structure for choosing the "best" glottal estimate, which will be discussed in the following section.

3. Rank Based Glottal Quality Assessment

The primary feature of this algorithm structure is to automatically choose the "best" glottal waveform estimation from the stored estimates in a frame. This necessitates the ability to objectively com-

Table 1: GQM Descriptions

Name	Description
$hr_{mn(X)}$	Mean ratio of harmonic peaks (0-X Hz, X=1000,3700)
$hr_{mx(X)}$	Ratio of the first harmonic to the maximum harmonic (0-X Hz, X=1000,3700)
$R^2_{(X)}$	Linear regression R^2 statistic over (0-X Hz, X=1000,3700)
GD_{var}	Variance of the group delay function for a glottal cycle
Krt	Kurtosis of the glottal waveform
pp_{cper}	Phase-plane (cycles/period)
pp_{cyc}	Phase-plane(mean sub-cycle length)

pare the stored estimates to one another without the need for visual inspection. While it is difficult to determine exactly what the glottal waveform looks like, it should be possible to evaluate the *expected* characteristics of an ideal glottal waveform. We have investigated the use of Glottal Quality Measures (GQM) (to be discussed in the following section) for assessing glottal waveform estimations. However, decisions are inherently based on the extremum of the measures (e.g., the maximum or minimum value) which are not always accurate. We have found it is more natural to assume that good GQM's should reliably establish *trends* among a set of glottal waveform estimates (i.e., from relatively good to relatively bad) without the "best" or "worst" necessarily being represented by the extreme values. Additionally, no single GQM is designed to measure *all* of the qualities of a glottal estimate and it is likely that the combination of GQM's should produce better results. We propose here a simple new technique for combining multiple GQM's for evaluating the stored glottal estimates in our algorithm structure. We refer to this technique as Rank Based-Glottal Quality Assessment (RB-GQA) and it is implemented in the following steps:

1. For each GQM, rank each stored estimate from '1' to the number of stored estimates available (i.e., for *N* stored estimates, a rank of '1' indicates the "best" of the stored estimates for that GQM and rank of *N* indicates the worst)
2. Compute the average ranking across all GQM's and choose the estimate with the highest average rank

One advantage of RB-GQA is that it allows input from all of the GQM's in making a final decision. Another attractive feature is that the decision is completely self-contained without the need of knowledge of prior or future estimates. For the RB-GQA method to be effective, careful consideration must be given to the GQM's that are used for ranking purposes. We present here an evaluation of ten GQM's including their implementation into the RB-GQA decision structure.

4. Glottal Quality Measures

Glottal quality is a vague concept at best. However, the most notable characteristic of an ideal glottal waveform is that it should exhibit little to no residual formant resonances (e.g., ripple). Additionally, a glottal quality measure (GQM) for the RB-GQA decision structure should adhere to the principle that tracing the trend of it's extremum values reliably tends to better (or worse) estimates. A list of the GQM's evaluated for the RB-GQA decision structure in this study is included in Table 1. Several of the GQM's to be considered in this study have been previously documented and will not be covered in detail here. The motivation for using the group delay (GD) as a GQM was presented in [7]. Work in [7] noted that the phase spectrum over a single cycle of the glottal flow should be essentially constant over a wide frequency range if the vocal tract estimation used to create the glottal estimate was correct. We chose to measure the variance of the group-delay (GD_{var}) for the glottal flow (computed over a single cycle) as it would be expected that better estimates of the glottal waveform should have a variance close to 0. Kurtosis (Krt), which measures the similarity of a distribution to the Gaussian distribution, was proposed as a GQM in [8]. The logic for it's use was based on the understanding that convolution involves summing copies of the input signal at different time delays which should converge to a Gaussian distribution. In [8] it was observed that the sub-gaussian nature of the glottal waveform could me measured using the kurtosis as an indication of the accuracy of the deconvolution operation performed by inverse-filtering (a lower value indicated a higher accuracy). Also presented in [8] were GQM's based on phase-plane analysis. These measures relied on the assumption that the glottal waveform can be modelled as a second order harmonic equation, which implies that its plot in the phase-plane (x(t), dx/dt) should consist of one closed loop per fundamental period. Resonances not completely removed by inverse filtering appeared as sub-cycles within the fundamental loops. The phase-plane plots were quantified by measures reflecting the number of cycles per fundamental period (pp_{cper}) (i.e., fewer cycles reflected better estimates) and the mean sub-cycle length (pp_{cyc}) (i.e., smaller sub-cycles reflected better estimates) as described in [8].

Six new GQM's were proposed for this study based on the harmonics created in a single cycle of the glottal derivative estimate. Ideally, the spectrum of the glottal waveform should exhibit a strictly negative spectral slope due to the lack of resonant structure. If formant residuals are present, this linear trend is disturbed. We proposed GQM's based on the following: the mean ratio of the first harmonic peak to other peaks over a frequency range X ($hr_{mn(X)}$), the ratio of the first harmonic peak to the maximum peak present over a frequency range X ($hr_{mx(X)}$), and the linear regression R^2 statistic over a frequency range X ($R^2_{(X)}$). Ideally, the first harmonic peak *should* tend to be greater than successive peaks to adhere to the negative linear trend expected from an ideal glottal waveform. Deviations from this can create ratios that are greater than one and indicate worse glottal estimates. Additionally, the linear regression R^2 statistic can be used to judge the appropriateness of a linear fit over a given frequency range (better estimates should be indicated by higher values for the R^2 statistic). One of the frequency ranges used for these GQM's was an analysis of the log-spectral peaks from 0-1000 Hz. This frequency range was used to cover the significant area of any residual 1^st formant energy (which is normally the largest culprit in formant ripple). Additionally, the frequency range covering 0-3700 Hz was used to cover the most significant area of human speech production. Figs. 1 and 2 show examples and the resulting measurements of the harmonic ratio and linear regression GQM's for a poor and better glottal estimate, respectively.

5. Experimentation and Results

The purpose of this study was to evaluate the GQM's described above as well as test the usefulness of the RB-GQA decision structure in the glottal estimation algorithm. To accomplish an effective evaluation it was determined that a reasonable requirement for a

Table 2: Percentage of glottal estimations taken from the closed region

	$hr_{mn(1000)}$	$hr_{mx(1000)}$	$R^2_{(1000)}$	$hr_{mn(3700)}$	$hr_{mx(3700)}$	$R^2_{(3700)}$	GD_{var}	Krt	pp_{cper}	pp_{cyc}	**RB-GQA**
S1	0.90	0.75	0.85	0.65	0.75	0.50	0.70	0.30	0.80	1.00	**1.00**
S2	0.00	0.00	0.00	0.00	0.00	0.30	0.65	0.10	0.30	0.10	**0.90**
S3	0.80	0.75	0.50	0.95	0.75	0.05	0.20	0.90	0.45	0.05	**0.75**
S4	0.90	0.90	0.90	0.90	0.90	0.05	0.25	0.00	0.90	0.90	**0.90**
S5	0.65	0.20	0.30	0.45	0.20	0.30	0.35	0.15	0.55	0.60	**0.85**
S6	0.65	0.60	0.65	0.60	0.60	0.50	0.65	0.25	0.85	0.90	**1.00**
S7	0.80	0.85	0.90	0.70	0.85	0.45	0.10	0.45	0.40	1.00	**1.00**
S8	0.95	0.95	0.95	0.95	0.95	0.80	0.80	0.00	0.65	1.00	**1.00**
S9	0.30	0.20	0.65	0.25	0.20	0.60	0.40	0.55	0.65	0.55	**1.00**
AVG	0.66	0.58	0.63	0.61	0.58	0.39	0.46	0.30	0.62	0.68	**0.93**

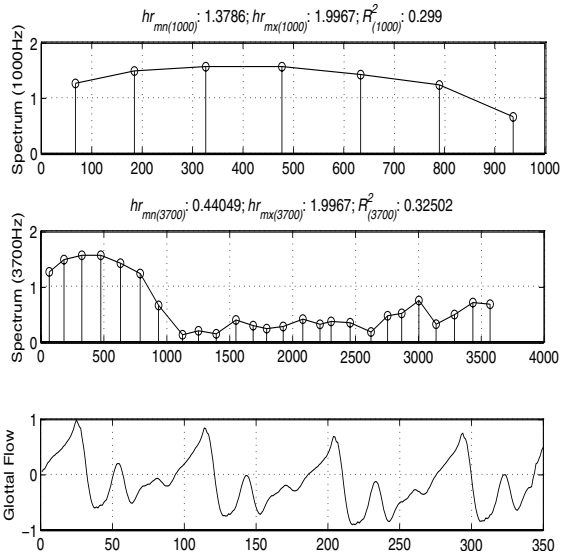

Figure 1: Poor Glottal Estimate

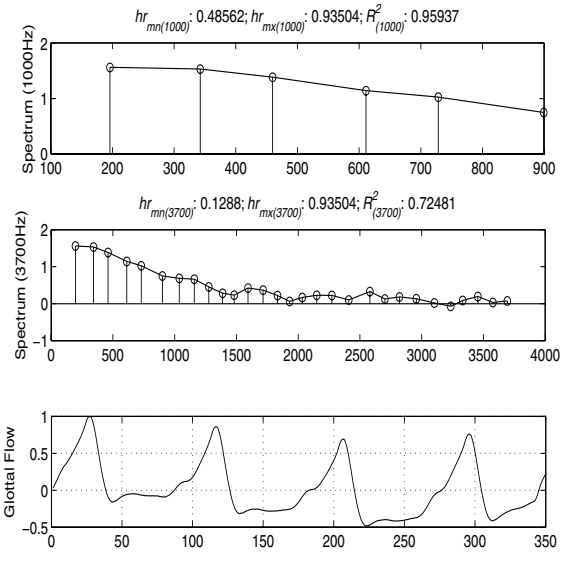

Figure 2: Better Glottal Estimate

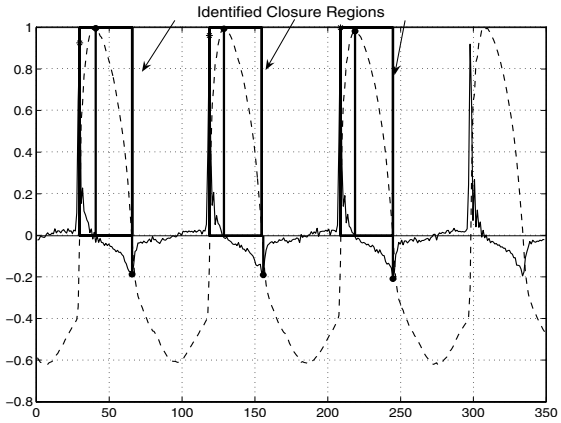

Figure 3: EGG (Dashed) and EGG Derivative (Solid)

GQM was that it should tend to exhibit it's best performance within the closed region of the glottal phase. The closed phase of the glottal waveform is, theoretically, the optimum place for vocal tract estimation and, therefore, glottal waveform estimation. Accurately locating GCI's is non-trivial so the data used for this experiment consisted of 9 males uttering a single vowel for which consecutive EGG recordings had been made. The recordings were made at a 10 kHz sampling frequency and the EGG's signals were shifted appropriately to account for recording delay. The GCI information was extracted from the peaks of the EGG signal. The closure region of the glottal waveform was identified as the area surrounding the GCI peaks in the EGG waveform as shown by the boxes in Fig. 3. While the relationship between the length of the closed/open phase and the EGG waveform is not exact, we defined the glottal closure area around the GCI's by utilizing the EGG derivative to locate the points of maximum rise (to indicate the onset of glottal closure) and maximum descent (to indicate the onset of glottal opening) in the EGG waveform. The glottal algorithm from [6] was modified to use the GCI information from the EGG signal to define a search area for the vocal tract estimate. A sample-by-sample sliding covariance analysis window was utilized to make vocal tract/glottal waveform estimations. Each analysis window was the length of twice the linear prediction model order (model order was manually selected for each speaker between 14 and 18 based on subjective evaluation of performance). The sliding analysis began where the covariance window's rightmost edge touched

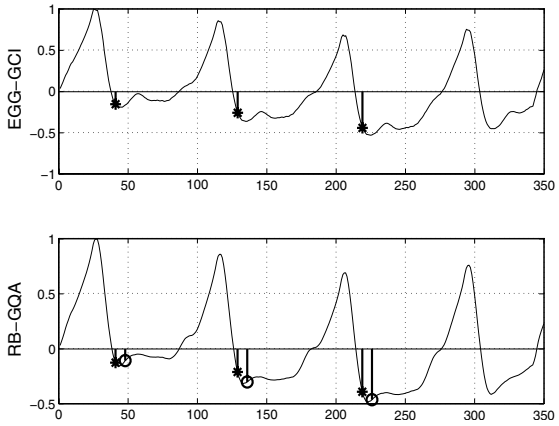

Figure 4: Glottal Wave Estimation (S1)

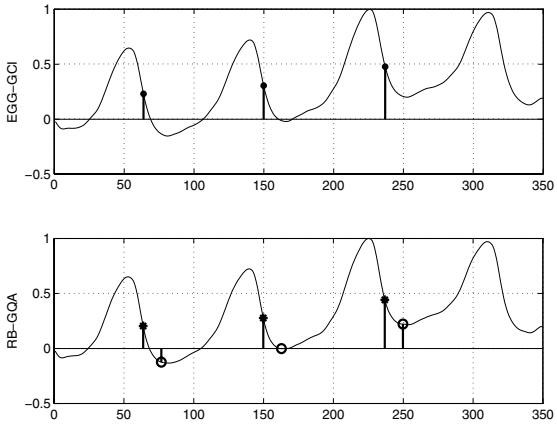

Figure 5: Glottal Wave Estimation (S9)

a GCI (as indicated by the EGG signal) and continued sample-by-sample over a range of three times the model order. In this manner, glottal waveform estimations were created from vocal tract estimates taken before, during, and after the point of closure. The GQM's described previously were then extracted for each of the stored glottal waveform estimates per frame and the "best" glottal waveform was recorded for each GQM. Additionally, all of the GQM's were implemented into the RB-GQA decision structure for choosing the "best" glottal estimate. The percentage of glottal waveform estimates that resulted from choices where the analysis window was actually in the closure region were recorded for each of the GQM's across 20 frames for each of the 9 male speakers (S1-S9). The results are shown in Table 2. On average, the best individual GQM's were related to the mean harmonic ratio and the phase-plane measures where over 60% of the analysis estimations resulted from the closed phase regions across the speakers. However, for S2, none of the GQM's performed well individually as the best estimates indicated by the GQM's rarely (save in the case of GD_{var}) fell in the closed area region. The results from implementing the RB-GQA strategy resulted in marked improvement overall. The best average performance from a single GQM was the phase-plane measure of the mean cycle length (pp_{cyc}) at 68% On average, the RB-GQA decision structure improved the number glottal waveform estimations taken from the closed phase region by 25%. A notable sign of the improvement by the RB-GQA decision structure is for S2 where many of the GQM's had failed individually to choose estimations from within the closed region and the RB-GQA achieved a rate of 90%. Figs. 4 and 5 show two examples of the glottal waveform estimations resulting from using the exact glottal closure instant from the EGG (top graph) and the RB-GQA algorithm. The star indicates the GCI indicated by the EGG signal while the circle indicates the analysis point chosen for the glottal waveform estimation by the RB-GQA. While the RB-GQA decision does not locate the GCI (it was never intended to), it does locate a viable point of analysis within the closed phase region.

6. Conclusion

The work presented here introduces a new structure for combining multiple GQM's in making automated decisions regarding glottal waveform estimations. While the current realization of this algorithm is computationally intensive, this is not a major issue for applications that utilize offline voice analysis. Research is ongoing for narrowing the size of the stored estimates, improving the ranking procedure, and evaluating the use and number of GQM's to reduce the computational burden.

7. References

[1] H. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *NATO ASI: Speech production and speech modeling*, 1989, pp. 1–21.

[2] A. Kounoudes, P. A. Naylor, and M. Brookes, "The dypsa algorithm for estimation of glottal closure instants in voiced speech," in *IEEE Int. Conf. Acous. Spch. Sig. Process.*, 2002, vol. 1, pp. 349–352.

[3] D. M. Brookes and H. P. Loke, "Modeling energy flow in the vocal tract with applications to glottal closure and opening detection," in *IEEE Int. Conf. Acous. Spch. Sig. Process.*, 1999, vol. 1, pp. 213–216.

[4] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 5, pp. 569–585, 1999.

[5] D. Wong, J. Markel, and A. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 4, pp. 350–355, 1979.

[6] E. Moore and M. Clements, "Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information," in *IEEE Int. Conf. Acous. Spch. Sig. Process.*, 2004, vol. 1, pp. 101–104.

[7] P. Alku, M. Airas, T. Backstrom, and H. Pulakka, "Group delay function as a means to assess quality of glottal inverse filtering," in *INTERSPEECH*, 2005, pp. 1053–1056.

[8] T. Backstrom, M. Airas, L. Lehto, and P. Alku, "Objective quality measures for glottal inverse filtering of speech pressure signals," in *IEEE Int. Conf. Acous. Spch. Sig. Process.*, 2005, vol. 1, pp. 897–900.