



Performance evaluation of glottal quality measures from the perspective of vocal tract filter consistency

Juan Torres, Elliot Moore

Georgia Institute of Technology, School of Electrical and Computer Engineering
210 Technology Circle, Savannah, GA 31407, USA

juan.torres@gatech.edu, emoore@gtsav.gatech.edu

Abstract

The main difficulty in glottal waveform estimation is the separation of the unknown vocal tract and glottal components of the speech signal. Several glottal quality measures (GQM's) have been proposed to objectively assess the quality of source-tract separation by exploiting known properties of glottal waveforms. In this paper, we present a performance evaluation of 10 GQM's based on the consistency of estimated vocal tract filters (VTF's) on sustained vowel utterances. We compare the results obtained using GQM's to select the optimal estimates to the case where the linear prediction window is aligned exactly with the glottal closure instant (GCI). Although GCI use resulted in the most consistent VTF's, there was a significant benefit from combining several GQM's for selecting optimal estimates. In addition, the GQM-derived estimates were shown to have higher divergence than the GCI estimates across some phoneme-pairs, suggesting higher class-separability.

Index Terms: Source-Tract Separation, Glottal Waveform, Voice Source, Inverse Filtering

1. Introduction

Improved consistency and quality in glottal waveform estimation is directly applicable to several areas of speech processing, including speaker recognition, diagnosis of mood disorders, voice conversion, and vocal affect classification. The main difficulty in glottal waveform estimation stems from the need to separate the unknown vocal tract and glottal components of the speech signal. For many years, the vast majority of speech analysis research has operated on the approximation of a short segment of speech by a cascade of linearly separable time-invariant (LTI) filters, where the effects of the voice source and vocal tract are considered independent and completely separable. It has been widely accepted that the best method for extracting the voice source and vocal tract components under the LTI assumption is to find a point of glottal closure (GCI), which corresponds to the point of maximal excitation of the vocal tract. In addition, the closed glottal phase that follows the GCI corresponds to the region of maximal decoupling between the voice source and vocal tract, where the LTI assumption is maximally valid. Glottal waveform estimation algorithms that exploit the concept of glottal closure include [1, 2, 3, 4]

With exact *a priori* knowledge of the glottal cycle phases, high quality glottal estimates may be obtained by using the covariance method of linear prediction analysis (LPA) with a small analysis window that starts at the GCI sample and extends through the glottal closure region. However, the automatic identification of glottal closure instants is a challenging issue [5, 2], and although newer algorithms have improved accuracy [6, 7],

there is still some uncertainty in determining the exact sample at which the GCI occurs. An additional problem with closed-phase analysis is that in many cases (e.g., females, emotional stress, various voice types, etc.) GCI's may be very short or nonexistent. Another class of glottal waveform estimation algorithms assume a parametric model of the glottal waveform and simultaneously estimate the glottal waveform parameters and the vocal tract filter (VTF) using speech data from all phases of the glottal cycle [8, 9, 10]. While this class of algorithms is better able to account for source-tract interaction, the resulting glottal waveform estimates may in some cases be limited by the assumed parametric model.

A recent algorithm [11] made the assertion that quality glottal waveform estimates could be obtained without the need to find exact GCI locations. The algorithm created several potential estimates by sliding a small (covariance) analysis window across a speech frame and then required a decision structure for selecting the "best" estimate from those available. To perform this decision automatically, it was necessary to objectively quantify the concept of "good" and "poor" estimates of the glottal waveform. Several glottal quality measures (GQM's) have been proposed to objectively perform this assessment [12, 13, 14]. To evaluate the performance of these GQM's, we developed in [14, 15] an experimental setup that produced sets of glottal waveforms estimated from the closed and open glottal regions, respectively. Under the assumption that estimates obtained from the closed phase have higher quality, this data was then used to objectively assess GQM performance by measuring the percentage of pitch cycles for which the GQM's selected a best-estimate that had been obtained from the closed phase. We then presented a rank-based method that allows an arbitrary subset of GQM's to be combined for selecting the best estimate. In [15], an exhaustive search was conducted on the GQM subset space using the same performance metric as above, and a combination of 4 GQM's was found to perform optimally across the entire speech dataset.

Within the sliding covariance analysis framework, it is expected that if the GQM's are truly selecting the best possible glottal estimates, then the application of GQM's should result in an overall improvement in source-tract separation, which would be observable in the form of more accurate vocal tract estimates. Because the true VTF is unknown, directly evaluating the quality of inverse filtering is a difficult task (fig. 1). However, a reasonable assumption that can facilitate evaluation is that the vocal tract remains fairly constant across pitch cycles during sustained vowel utterances. In this paper, we evaluate the performance of source-tract separation via GQM's by measuring the consistency of the estimated VTF's across pitch cycles of individual sustained vowel utterances, in terms of mean cep-

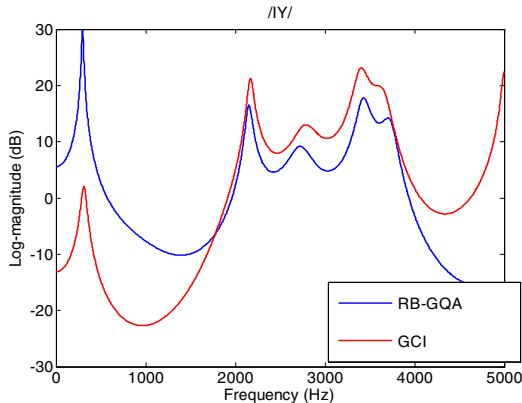


Figure 1: VTF spectra for the same pitch cycle; estimated using a combination of GQM’s (RB-GQA) and the glottal closure instant (GCI), respectively. The selected analysis windows were only 3 samples apart, but the spectra differ significantly.

stral distance. We compare the results obtained using GQM’s to select the optimal LPA analysis window positions to the case where the LPA window is aligned exactly with the GCI’s (as determined from the EGG signal). In addition, to compare the usefulness of the VTF’s extracted by each algorithm, we measure the class separability between the cepstral vectors extracted from different phonemes, with a higher separability indicating more useful VTF estimates for speech recognition.

The paper is organized as follows: Section 2 describes the speech dataset and the data extraction procedure. Section 3 briefly reviews the glottal quality measures (GQM’s) under consideration; The rank-based method that allows GQM’s to be combined for glottal quality assessment is briefly reviewed in Section 4; Experiments and results are described in Section 5; Finally, section 6 presents conclusions and suggests directions for future work.

2. Data Extraction

The data extraction procedure is identical to [15] and will only be described here briefly. The speech data used for evaluation consisted of sustained utterances of the U.S. English vowels /IY/ as in “beet,” /UW/ as in “boot,” and /AA/ as in “Bach” under modal phonation, obtained from [16]. This dataset contains a single utterance of each vowel from each of 25 male speakers and was recorded with a high-quality microphone. The utterances in the dataset also include an accompanying Electrolaryngograph (EGG) signal, which simplifies the identification of glottal closure instants (GCI’s). The acoustic recordings were sampled at 10 kHz and the EGG signals were shifted appropriately to account for recording delay.

To evaluate GQM performance, we wish to obtain a sufficient number of estimates from each pitch cycle. While the relationship between the length of the closed/open phase and the EGG waveform is not exact, we divided the pitch cycle into two halves, with glottal closure defined as the region between the point of maximum rise (corresponding to the GCI) and maximum descent (corresponding to the onset of glottal opening) in the EGG waveform. Likewise, the open glottal phase was defined as the region between the point of maximum descent and maximum rise in the EGG waveform.

For every pitch cycle, we use the covariance method of lin-

ear predictive analysis (LPA) to estimate the vocal tract using a set of analysis windows entirely confined to the closed glottal phase and another set of analysis windows confined entirely to the open glottal phase. After inverse filtering the speech signal using these two sets of all-pole filters and integrating it to remove the effects of lip radiation, we are left with a set of glottal waveform estimates. To ensure that the GQM’s had enough estimates to choose from for each pitch cycle, we only kept pitch cycles containing at least 10 estimates. In addition, to ensure that each utterance contained enough pitch cycles for a statistically significant performance assessment, utterances that did not contain at least 100 analysed pitch cycles were discarded. After these pruning steps, we were left with a total of 16 speakers, with 10 speakers for each phoneme, an average of 206 analysed pitch cycles for each utterance, and an average of 29 estimates per pitch cycle.

The LPA model order for each utterance varied from 12 to 16 and was manually adjusted by visual inspection of the spectrogram of the entire utterance, the LPA spectra, and the estimated glottal waveforms. In all cases, LPA was performed on the pre-emphasized speech signal. After inverse filtering the original speech signal with an LPA-estimated VTF, glottal waveforms were obtained from the glottal derivatives using a lip radiation coefficient of 1.0 (ideal integrator).

3. 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 or no residual formant resonances (e.g., ripple). For this paper, we consider the evaluation of several previously proposed measures that may serve as GQM’s, which are briefly described below:

In [12], Backstrom et al. presented two GQM’s based on phase-plane analysis that rely on the assumption that the glottal waveform can be modeled as a second order harmonic equation. This implies that its plot in the phase-plane ($x(t), \frac{dx}{dt}$) should consist of one closed loop per fundamental period. Resonances not completely removed by inverse filtering should appear as sub-cycles within the fundamental loops. Glottal quality can be assessed by measuring the number of cycles per fundamental period (pp_{per}), with fewer cycles reflecting better estimates, and the mean sub-cycle length (pp_{cyc}), with smaller sub-cycles reflecting better estimates.

Kurtosis (Krt), which can be used to measure the similarity of a distribution to the Gaussian distribution, was another GQM proposed in [12]. The logic for its use is based on the understanding that convolution involves summing copies of the input signal at different time delays, so that the distribution of the convolution output should be closer to Gaussian, with a kurtosis value closer to 3. Thus, they concluded that the kurtosis can be used to measure the accuracy of the deconvolution operation performed by inverse filtering, with a lower value indicating higher accuracy. However, in [15] it was shown that for this dataset there exists a relationship between a higher kurtosis value and higher glottal quality, and that is the definition adopted here.

The motivation for using the group delay (GD) as a GQM was presented by Alku et al. [13], who observed 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 estimate used to create the glottal waveform estimate was correct. We chose to measure the variance of the group delay (GD_{var}) for the glottal flow (computed over a single cycle and using an

FFT size of 4096), as it would be expected that better estimates of the glottal waveform should have a variance closer to zero.

In [14], we presented several GQM’s based on the idea that the spectrum of the glottal waveform should exhibit a strictly negative spectral slope due to the lack of resonant structure, which is disturbed if formant residuals are present. 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 of the log-spectral peaks over a frequency range X ($R^2_{(X)}$). One of the frequency ranges used for these GQM’s was 0–1000 Hz, which covers the significant area of any residual 1st formant energy (normally the largest culprit in formant ripple). Additionally, the frequency range of 0–3700 Hz was used to cover the most significant area of human speech production.

A summary of the GQM’s presented above is shown in Table 1.

Table 1: *GQM Descriptions*

$hr_{mn(X)}$	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 over a single pitch cycle
Krt	Kurtosis of the glottal waveform
pp_{cper}	Phase-plane cycles per period
pp_{cyc}	Phase-plane mean sub-cycle length

4. Rank-Based Glottal Quality Assessment

In the previous section, we discussed the use of glottal quality measures (GQM’s) for assessing the quality of glottal waveform estimates. However, no single GQM is designed to measure *all* of the qualities of a glottal waveform estimate and it is likely that the combination of GQM’s should produce better results. In [14], we introduced a simple new technique for combining multiple GQM’s for evaluating the relative quality of a set of estimates, referred to as rank-based glottal quality assessment (RB-GQA). It is implemented as follows:

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 rank across a subset of GQM’s and sort the estimates by increasing average rank. The estimate with the lowest average rank (i.e., closest to 1) is selected as the highest-quality estimate.

An advantage of RB-GQA is that it is invariant to any monotonic transformation of the GQM values, thus allowing input from any subset of the GQM’s to be effectively combined in the final quality assessment. In [15], it was determined that the combination of GQM’s that maximized the percentage of selected closed-phase estimates was $\{hr_{mx(3700)}, GD_{var}, pp_{cper}, pp_{cyc}\}$. Using this GQM combination on the speech

dataset described in Section 2, RB-GQA selected an average of 94% of the estimates from the closed glottal phase.

5. Experiments and Results

To evaluate the consistency of estimated all-pole vocal tract filters, we quantify their similarity using the quefrency-weighted cepstral distance, which was shown in [17] to yield higher separability between the LPA spectra of different phonemes than the euclidean cepstral distance [18]. The first 10 cepstral coefficients were obtained directly from the VTF’s using the recursive relation in [17], with the VTF’s normalized to have a 0th cepstral coefficient of 1. The consistency of vocal tract estimates is then quantified, for a particular utterance, by the mean of the quefrency-weighted cepstral distances across pitch cycles:

$$v = \frac{1}{N} \sum_{i=1}^N \left[\sum_{j=1}^M \{j(x_{ij} - m_j)\}^2 \right]^{1/2}, \quad (1)$$

where x_{ij} is the j^{th} component of the cepstral vector \mathbf{x}_i obtained from the i^{th} pitch cycle, and m_j is the j^{th} component of the utterance’s mean cepstral vector $\mathbf{m} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$.

Using the sustained vowel dataset described in Section 2, we compare the consistency of the VTF’s obtained via maximization of each individual GQM and with RB-GQA using the GQM subset given in Section 4. For baseline comparison, we also measured the consistency of the estimates obtained by aligning the start of the LPA analysis window with the GCI sample obtained from the EGG signal. Table 2 shows the mean cepstral distances v of the VTF estimates, averaged over the utterances in each phoneme.

Table 2: *Mean cepstral distances (v), averaged over the utterances in each phoneme. The first 10 rows show the results obtained by maximizing glottal quality according to individual GQM’s; The last two rows show, respectively, the results obtained by application of RB-GQA using the GQM subset $\{hr_{mx(3700)}, GD_{var}, pp_{cper}, pp_{cyc}\}$ and by fixing the LPA window at the glottal closure instants (GCI’s).*

	/IY/	/UW/	/AA/	Mean
$hr_{mn(1k)}$	3.071	3.025	2.926	3.01
$hr_{mx(1k)}$	2.768	2.807	3.032	2.87
$R^2_{(1k)}$	3.082	3.001	2.912	3.00
$hr_{mn(3.7k)}$	3.008	2.868	2.976	2.95
$hr_{mx(3.7k)}$	2.708	2.780	3.007	2.83
$R^2_{(3.7k)}$	3.059	2.614	2.691	2.79
GD_{var}	2.669	2.747	2.676	2.70
Krt	2.716	2.852	3.301	2.96
pp_{cper}	3.182	3.215	2.648	3.01
pp_{cyc}	2.763	2.799	2.752	2.77
RB-GQA	2.634	2.632	2.461	2.58
GCI	2.365	2.431	2.043	2.28

The results show that for each phoneme, maximum VTF consistency was obtained by aligning the LPA windows with the glottal closure instants, which is consistent with the notion of GCI’s as the optimal place to perform source-tract separation (when the closed-phase exists and is long enough). While the consistency in all other cases could not match that of the GCI-derived estimates, we do see some improvement from combining several GQM’s via RB-GQA. Statistical significance tests

(paired T-tests) were performed, and it was determined that the RB-GQA results were better to a 95% significance level than the results of each individual GQM.

To further investigate the quality of the VTF estimates extracted using RB-GQA or GCI's, we also measured the class separability between phonemes using the LPA-derived cepstra as features. The divergence [19], defined as

$$d_{ij} = \int_{-\infty}^{\infty} (p(\mathbf{x}|\omega_i) - p(\mathbf{x}|\omega_j)) \ln \frac{p(\mathbf{x}|\omega_i)}{p(\mathbf{x}|\omega_j)} d\mathbf{x}, \quad (2)$$

can be used as a distance measure between the probability distributions $p(\mathbf{x}|\omega_i)$ and $p(\mathbf{x}|\omega_j)$ of classes ω_i and ω_j . Clearly, $d_{ij} = 0$ if $\omega_i = \omega_j$, and it increases as the separation between the distributions of two classes grows. Assuming a multivariate Gaussian distribution for the cepstral features, we computed the divergence between each pair of phonemes using the equation given in [19]. The results in Table 3 show that RB-GQA resulted in a lower divergence between phonemes /IY/ and /UW/, but a higher divergence between the remaining phoneme pairs. This result suggests that in some cases, despite producing less consistent results, performing source-tract separation by choosing the glottal estimates that maximize a combination of GQM's may result in more useful VTF's for speaker-independent speech recognition purposes, relative to the case where separation is performed via GCI-aligned analysis windows.

Table 3: Divergence between the cepstral features of each phoneme pair, for VTF's extracted via (a) GCI's, and (b) RB-GQA.

	(a)		(b)	
	/IY/	/UW/	/IY/	/UW/
/UW/	18.57		11.26	
/AA/	91.44	90.58	105.5	100.5

6. Conclusion and Future Work

In this study, we have presented an alternative way of measuring the suitability of glottal quality measures (GQM's) for source-tract separation of real speech signals. Using sustained vowel utterances, it was found that LPA-based inverse filtering produced maximally consistent vocal tract filter (VTF) estimates when the analysis window is aligned with the instants of glottal closure (GCI's). However, the combination of 4 GQM's significantly reduced VTF inconsistency when compared to individual GQM's. Interestingly, the increased variability in the GQM-derived VTF estimates was in some cases accompanied by higher class-separability between phonemes using data from several speakers. Future work includes the evaluation of GQM performance on female speech and on speech modalities where GCI's and/or closed phases are not well-defined.

7. References

- [1] 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.
- [2] 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.
- [3] O. O. Akande and P. J. Murphy, "Estimation of the vocal tract transfer function with application to glottal wave analysis," *Speech Communication*, vol. 46, no. 1, pp. 15–36, May 2005.
- [4] H. Deng, R. K. Ward, M. P. Beddoes, and M. Hodgson, "A new method for obtaining accurate estimates of vocal-tract filters and glottal waves from vowel sounds," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 2, pp. 445–455, Mar. 2006.
- [5] 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.*, vol. 1, 1999, pp. 213–216.
- [6] A. Kounoudes, P. A. Naylor, and M. Brookes, "The dyspa algorithm for estimation of glottal closure instants in voiced speech," in *IEEE Int. Conf. Acous. Spch. Sig. Process.*, vol. 1, 2002, pp. 349–352.
- [7] M. Brookes, P. A. Naylor, and J. Gudnason, "A quantitative assessment of group delay methods for identifying glottal closures in voiced speech," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 2, pp. 456–466, Mar. 2006.
- [8] M. Frohlich, D. Michaelis, and H. W. Strube, "Simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals," *Journal of the Acoustical Society of America*, vol. 110, no. 1, pp. 479–88, Jul. 2001.
- [9] P. Alku, M. Airas, and B. Story, "Evaluation of an inverse filtering technique using physical modeling of voice production," in *INTERSPEECH*, 2004, pp. 497–500.
- [10] Q. Fu and P. Murphy, "Robust glottal source estimation based on joint source-filter model optimization," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 2, pp. 492–501, Mar. 2006.
- [11] 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.*, vol. 1, 2004, pp. 101–104.
- [12] 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.*, vol. 1, 2005, pp. 897–900.
- [13] 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.
- [14] E. Moore and J. Torres, "Improving glottal waveform estimation through rank-based glottal quality," in *INTERSPEECH*, 2006, pp. 1694–1697.
- [15] —, "A performance assessment of objective measures for evaluating the quality of glottal waveform estimates," submitted to: *Speech Communication*, Mar. 2007.
- [16] D. Childers, *Speech processing and synthesis toolboxes*. John Wiley and Sons, Inc., 2000.
- [17] K. K. Paliwal, "On the performance of the quefrency-weighted cepstral coefficients in vowel recognition," *Speech Communication*, vol. 1, no. 2, pp. 151–154, 1982.
- [18] A. H. Gray and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.
- [19] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. San Diego, CA: Elsevier, 1999.