Glottal Closure and Opening Detection for Flexible Parametric Voice Coding

Pamornpol Jinachitra

Center for Computer Research in Music and Acoustics
Stanford University
pj97@ccrma.stanford.edu

Abstract

The knowledge of glottal closure and opening instants (GCI/GOI) is useful for many speech analysis applications. A Pitch-synchronous waveform encoding of voice is one such application. In this paper, a dynamic programming is employed to solve for the global close/open phase segmentation based on the polynomial parametric waveform of the derivative glottal waveform and its quasi-periodicity. Not only does the algorithm identify GCIs, but also the elusive GOIs, and as a by-product, the parameters of the glottal excitation waveform. The results show its effectiveness compared with a classical method. Its application to parametric voice encoding which allows for simple time-pitch scaling as well as voicing quality conversion is demonstrated.

Index Terms: speech coding, glottal closure instant detection, glottal opening instant, voice transformation.

1. Introduction

Accurate detection of glottal closure instants (GCI) allows for many useful pitch-synchronous operations. Closed-phase linear prediction has been shown to give more accurate vocal tract filter due to little source-tract interaction in those periods. Speaker identification, pathological voice detection, pitch tracking and low bit-rate coding can also benefit from the knowledge of GCIs.

Numerous techniques for GCI detection have been proposed in the past. Most are based on detecting discontinuity or peaks from some measurements of speech. The peaks of LPC residual energy were used in [1] while in [2], abrupt change in Kalman filtering innovation error indicates such changes. In [3], an energy-weighted group delay (EWGD) was proposed. This method provides a very effective and efficient GCI detection although its false alarm rate can go up significantly for noisy signal. While a large window used in the averaging of EWGD method improves false alarm rate (FAR), it may compromise the miss-detection rate (MDR) as well as the accuracy, which refers to how close the detection is to the real value (see a quantitative analysis in [4]). Recently, a dynamic programming approach with a variety of cost functions related to pitch deviation, quasi-periodicity, combined with some other heuristic cost terms [5] was proposed, giving considerable improvements over earlier methods.

GCI detection is a classical problem, receiving a great deal of attention. The detection of GOIs, on the other hand, has received relatively less consideration. One reason is because the opening instant is much harder to identify or even defined. Another reason is due to its less crucial effect on the perceptual quality of a voice. While a closure generates significant excitation, with an abrupt change in waveform, an opening instant tends to be gradual. A number of works have been presented on how various tools respond to the opening instants [6][7] but none has really made any evaluation on a real speech corpus.

Recently, an interest of speech or voice coding that is flexible for modification has become more widespread. Instead of just coding for compression and reconstruction faithfulness, a more structured approach has been developed. For example, the HVXC speech coder which encodes the glottal excitation by sines+noise model, allows for time-scaling [8]. When voice is parametrically encoded in such a way that its perceptual properties can be modified, the whole host of applications, such as emotion modification and intelligibility enhancement, can be easily implemented. While spectral-based model is easy to obtain and has been used effectively, a more physical model offers more intuitive control of the parameters to effect different articulations and voicing quality. Often, this requires pitch-synchronous analysis and the identification of the related parameters such as the vocal tract filter and the glottal excitation waveform.

In this paper, an algorithm to identify both GCIs and GOIs is proposed and its performance is investigated. The dynamic programming with cost functions based mainly on fitting a parametric model of the derivative glottal waveform is first described. The evaluation criterion and results are then presented. At the end, its application in encoding a sustained voice parametrically that allows for many types of modification is demonstrated.

2. Dynamic Programming for GCI/GOI Detection

A dynamic programming (DP) calculates the optimal path through a lattice of candidate points where the decision at any particular point only depends on the objective function of that point and the previous ones. For our problem, the cost function to be minimized is based on a combination of polynomial waveform fitting and the quasi-periodicity nature of the derivative glottal waveform, expected from inverse filtering the speech signal. The robustness against non-ideal LPC residual shape is achieved through the flexibility in the cost function as well as other means to constrain the problem toward the right solution. The composite cost function is presented by

$$C(i,j) = C_P(i,j) + \lambda \cdot C_Q(i,j) \tag{1}$$

where C_P and C_Q are waveform error cost function and the cross-correlation measurements respectively while λ is the relative weighting.

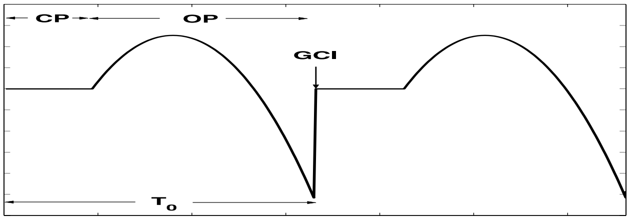

Figure 1: *Two periods of Rosenberg's derivative glottal waveform model showing the period (T_0), glottal closure instant (GCI), closed-phase (CP) and open-phase (OP)*

2.1. Waveform Error Cost

The derivative glottal waveform can be presented by Rosenberg's model [9]

$$g(n) = \begin{cases} 2a_g n/f_s - 2b_g(n/f_s)^2, & 0 \le n \le T_0 \cdot OQ \cdot f_s \\ 0, & T_0 \cdot OQ \cdot f_s \le n \le T_0 \cdot f_s \end{cases} \tag{2}$$

$$a_g = \frac{27 \cdot AV}{4 \cdot (OQ^2 \cdot T_0)} \tag{3}$$

$$b_g = \frac{27 \cdot AV}{4 \cdot (OQ^3 \cdot T_0^2)} \tag{4}$$

where T_0 is the fundamental period, f_s is the sampling frequency, AV is the amplitude parameter, and OQ is the open-quotient of the glottal source. An example of the waveform is shown in Figure 1.

Let s indicate the phase of the glottal waveform where $s = 0$ is the close-phase (CP) and $s = 1$ is the open-phase (OP). The cost function of a segment between time sample t_1 and t_2 for both CP and OP is the squared $L2$-norm

$$C_{P,s}(t_1, t_2) = ||\mathbf{x}_{t_1:t_2} - \hat{\mathbf{x}}||_2^2 \tag{5}$$

where for $s = 0$, $\hat{\mathbf{x}}$ is the mean of $\mathbf{x}_{t_1:t_2}$. Even though the model expects this to be zero, using the mean gives extra robustness to non-ideal waveform. For $s = 1$, $\hat{\mathbf{x}}$ is generated from equation (2) using a_g and b_g estimated from least-squares regression. If a_g or b_g estimates are less than zeros, it is not the right shape and the cost is set to a large number. To increase the robustness, local search is also performed for open-phase fitting, while for close-phase, an offset is allowed to avoid spikes which commonly occur in LPC residual.

2.2. Cross-correlation Cost

To tap into the quasi-periodicity expected in the LPC residual waveform, the cross-correlation cost between two segments, similarly used in [5], is

$$C_{Q,s}(\mathbf{x}_1, \mathbf{x}_2, \gamma) = -\max(CrossCorr_\gamma(\mathbf{x}_1, \mathbf{x}_2)) \tag{6}$$

where γ is the maximum lag used (set to correspond to 1.5 ms in the experiments).

Since both cost terms, C_P and C_Q, are the sum of time-sample products, they are comparable in magnitude so λ is set to one in all experiments. At each candidate point j, for each mode s, the cost functions with respect to a set of preceding marker candidates $i \in \mathcal{I}_s$, are calculated by

$$J_s(i,j) = \min\nolimits_{k \in \mathcal{K}_{s'}}\{J_{s'}(k,i) + C_{s'}(i,j)\} \tag{7}$$

where s' is the negation of s, which means that the constraint of alternate open and close phase is enforced. The set $\mathcal{I}_s$ and $\mathcal{K}_{s'}$ contains allowable candidates for each mode: $\mathcal{I}_s = \{i | j - \Delta^s_{max} < i < j - \Delta^s_{min}\}$ and $\mathcal{K}_s = \{k | i - \Delta^s_{max} < k < i - \Delta^s_{min}\}$. Δ^s_{max} is set to correspond to maximum CP or OP duration. Here, maximum pitch period (20 ms) is used for OP and a fraction smaller for CP. Δ^s_{min} is, however, set to zero for $s = 0$ to allow for non-existent CP which commonly occurs. Together with candidate selection later described, robust identification for such case is possible. For $s = 1$, Δ^s_{min} is set to a fraction of minimum pitch period (2 ms). This helps reduce computation.

Because three points can only cover one period (CP and OP), the calculation of $C_Q(i,j)$ in equation (1), used in equation (7), actually depends on two further points back. In order to keep the global optimality, we have to look back two points and select the point that gives minimum total cost at point k in (7). It is the same back-tracing used in general DP with point k being the end point.

The first period may be assumed to be CP. However, in our experiments, either case is allowed for truncation robustness. Two tables are therefore populated and, at the end of the segment, the end cost that is smaller is chosen and back-tracking is executed to obtain alternate GCIs and GOIs.

2.3. Candidate Selection

A careful selection of candidate points can help reduce the amount of computation time. It also affects the performance, both in terms of detection rate trade-off and the accuracy. One possible way is to choose positive zero crossings of the inverse-filtered signal. To make sure opening instant candidates are included, an extra point has to be added for every positive zero-crossing. Alternatively, one can first apply EWGD method which has been shown to detect impulses at the GCIs as well as some GOIs. By using a narrow averaging window, it is likely that GOIs will be detected. Besides, using a narrow window generates an over-complete set of candidates with higher accuracy for the correct points [4]. The key is to generate a redundant set of candidate points, which also include those accurate ones, for the DP to choose from.

An experiment was carried out to evaluate initial candidate sets against ground truth GCIs, to be described in section 3.1. It is found that using zero crossings as candidates can give more accurate results but with a large number of candidates. On the other hand, EWGD with a very narrow averaging window gives much less amount of data but not very accurate. We combine the two methods by first performing narrow-windowing EWGD and then look for the adjacent zero-crossings on its left (A) (potentially "true" opening instant) and right (B) (potentially more accurate closing instant). From inspections, sometimes CP never exists, especially in female voice. We therefore assign an extra point (B+1) to the candidate set so that the DP can have more choices and at the same time, allows for no CP events which can now be modeled as having CP period of one sample. Table 1 shows the accuracy measured against ground truth data and the number of candidates per voice segment. Accuracy is defined as the percentage of those samples, excluding false alarms and misses, that fall within 25 ms of the references. The RMS measures the standard deviation of all

Table 1: *Initial candidates characteristics.*

	Zero X	EWGD (.1 ms)	Hybrid
Accuracy(%)	80.2	46.4	78.8
RMS (ms)	0.29	0.34	0.30
# per segment	310	102	156

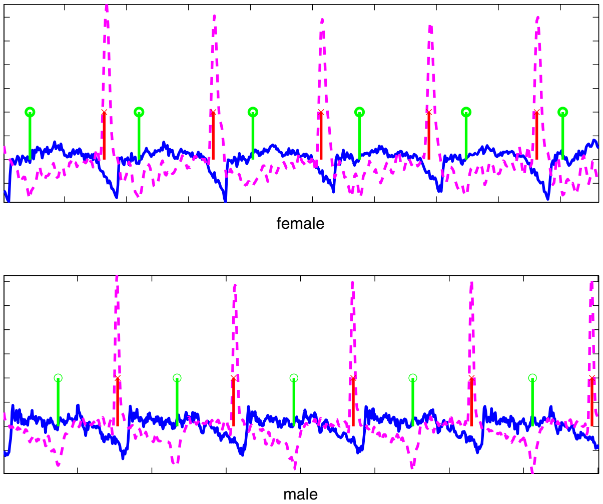

Figure 2: *Examples of the DEGG waveform (pink-dash) and the inverse-filtered derivative glottal waveform (blue-solid) for a female (top) and male (bottom). Reference GCIs and GOIs derived from peak picking DEGG are shown in (x) and (o) respectively*

errors from matched samples.

3. Experiments

3.1. Evaluation Test Set

The test set used here is the Keele's database [10], popularly used for pitch estimation evaluation. It consists of 20-kHz sampling-rate recordings of a roughly 30-second phonetically-balanced passage read by five females and five males. The Electroglottograph (EGG) signal is also simultaneously recorded for each speaker. The pitch masks at a frame resolution have been derived from the EGG and our evaluation periods ignore one frame margin on both sides of the voiced periods. Our reference GCIs are generated by finding peaks of the derivative EGG (DEGG) [11] which are very clear. The ground truth for GOIs are, on the other hand, much harder to identify. Our references are again based on minimal peaks of the DEGG which correspond to the inflexion points in the EGG (see Figure 2. Note that time delay is not compensated). Although easier to identify, such minimal peaks do not correspond to the "opening" instants for our modeling purposes, although it may be useful in other applications. The results presented in table 3 will therefore be only an approximation after forced-alignment (also a DP) of these different types of opening definition, assuming a constant offset between the two as evident in Figure 2.

A forced-alignment is also performed on the GCI results to compensate for a fixed delay difference between speech and EGG ground-truth. The cost of miss-detection and false alarm is taken to be the same during the alignment. False alarm rate (FAR) and miss detection rate (MDR) are defined as the outputs and the references, respectively, left unmatched. Accuracy and RMS are defined as described earlier in section 2.3.

3.2. Results and Discussion

Table 2 shows the results for EWGD method (best result for each gender) and our algorithm. The DP waveform fitting (DPWF) achieves comparable FAR and MDR to the EWGD. The accuracy, however, is clearly more superior. Figure 3 illustrates a typical example of successful performance. Note accurate identification even when CP is non-existent. The algorithm is resilient against deviations from the ideal model. On closer inspection, it is clear that most errors occur during transitioning periods where the LPC residual does not follow Rosenberg's model and when many spurious peaks occur. Results for GOI identification are shown in table 3. FAR and MDR are comparable to those of GCI as should be expected. The accuracy, however, is much lower for the same stringent accuracy criterion used for GCIs, illustrating the difficulty. From inspections, most are nevertheless reliable enough for waveform coding purposes, although parameter smoothing might be required for a good sound.

It is possible that other models are used to fit the waveform. A systematic error of identifying an opening as another closing is quite common and post processing to correct such obvious errors should improve the performance rather easily. Other cost terms like those used in [5] could also help further although hard to control. The algorithm also has a weak point of being polarity dependent. However, this should be easy to spot using simple thresholding.

Table 2: *GCI identification results comparison.*

Performance		EWGD	DPWF
FAR (%)	Females	2.6	2.3
	Males	6.0	6.1
MDR (%)	Females	4.6	4.3
	Males	2.6	2.0
Accuracy (%)	Females	52.8	80.9
	Males	64.1	69.3
RMS (ms)	Females	2.0	0.5
	Males	1.0	0.9

Table 3: *GOI identification results.*

Performance		DPWF
FAR (%)	Females	2.6
	Males	6.8
MDR (%)	Females	4.2
	Males	3.5
Accuracy (%)	Females	27.0
	Males	13.3
RMS (ms)	Females	1.3
	Males	2.0

3.3. Voice Coding and Modification

The direct by-products of the proposed GCI/GOI detection are the parameters of the waveform during the open-phase. Together with

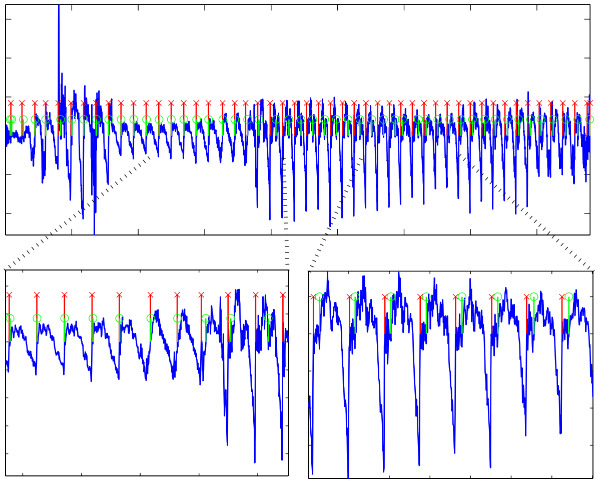

Figure 3: *A female voice with GCIs (x) GOIs (o) identified*

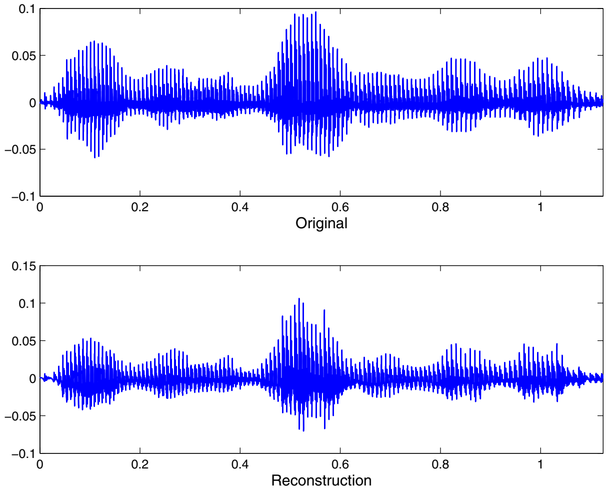

Figure 4: *Original utterance (top) and its parametric reconstruction (bottom)*

the GCIs and GOIs information, an approximated glottal excitation waveform can be generated and passed through the vocal tract filter for a speech output that will sound very much like the original. This can then be used to encode the voice parametrically as a time varying set of vocal tract filter, source's amplitude, pitch and open-quotient coefficients. Figure 4 shows the original male utterance of "Where were you while you were away" and the reconstruction using parameters and temporal marks estimated from the algorithm. Its pitch can be modified easily by changing the closure marks. Its duration can be changed by appropriate addition of extra cycles or omission of some cycles. Rosenberg's model also allows the change in voice quality, from pressed to normal and breathy, by changing the open-quotient, and perhaps, even adding some noise pitch-synchronously. For results with no artifacts, all parameters should be smoothed before using. All audio demonstrations can be found at *http://ccrma.stanford.edu/˜pj97/icslp06_demo.html*.

In addition to the simple polynomial waveform, the residual resulting from the subtraction of the parametric estimate can also be coded using codebook or other forms of compression. How to encode them for optimal perceptual effect during original playback and during modification is the subject of an ongoing research. It is quite clear that pitch-synchronous method will again be the most likely choice especially in breathy voice modeling. It should also be mentioned that these residual signals are likely to be more random while a_g and b_g are slowly varying, and hence, easy to compress.

4. Conclusions

A dynamic programming algorithm which simultaneously identifies the GCIs, GOIs and the glottal waveform parameter has been presented. The evaluation results show identification rates of the GCIs to be comparable to a classical method using group delay. The accuracy of the identified GCIs, however, show improvements. Experiments also show reasonable estimates of the GOIs. The algorithm enables parametric coding of the voice excitation source which is amenable to various types of expressiveness modification.

5. References

[1] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least squares inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1979.

[2] McKenna, "Automatic glottal closed-phase location and analysis by Kalman filtering," in *SSW4*, 2001.

[3] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Transactions on Speech Audio Processing*, pp. 325–333, 1995.

[4] M. Brookes, P. A. Naylor, and J. Gudnason, "A quantitative assessment of group delay methods for identifying glottal closures in voiced speech," *IEEE Transactions on Audio, Speech and Language Processing*, 2006.

[5] A. Kounoudes, P. A. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," in *ICASSP*, 2002.

[6] M. Brookes and H. P. Loke, "Modeling energy flow in the vocal tract with applications to glottal closure and opening detection," in *ICASSP*, 1999, pp. 213–216.

[7] W. Wokurek, "Time-frequency analysis of the glottal opening," in *ICASSP*, 1997.

[8] M. Nishiguchi, A. Inoue, Y. Maeda, and J. Matsumoto, "Parametric speech coding - HVXC at 2.0-4.0 kbps," in *IEEE Speech Coding Workshop*, 1999, pp. 84–86.

[9] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. of Am.*, vol. 49, no. 2, pp. 583–590, 1971.

[10] F. Plante, W. A. Ainsworth, and G. Meyer, "A pitch extraction reference database," in *Proc. Eurospeech Madrid*, 1995, pp. 837–840.

[11] D. G. Childers, D.M.Hicks, G. P. Moore, and Y. A. Alsaka, "A model for vocal fold vibratory motion, contact area, and the electroglottogram," *J. Acoust. Soc. Am.*, vol. 80, no. 5, pp. 1309–20, 1986.