Evaluation of Glottal Epoch Detection Algorithms on Different Voice Types

João P. Cabral[1], *John Kane*[2], *Christer Gobl*[2], *Julie Carson-Berndsen*[1]

[1]School of Computer Science and Informatics, University College Dublin, Ireland
[2] Centre for Language and Communication Studies, Trinity College, Dublin, Ireland
joao.cabral@ucd.ie, kanejo@tcd.ie, cegobl@tcd.ie, julie.berndsen@ucd.ie

Abstract

According to the source-filter model of speech production, speech can be represented by passing the excitation signal through the vocal tract filter. The epoch or instant of maximum excitation corresponds to the glottal closure instant. Several speech processing applications require robust epoch detection but this can be a difficult task. Although state-of-the-art epoch estimation methods can produce reliable results, they are generally evaluated using speech recorded with a neutral voice quality (modal voice). This paper reviews and evaluates six popular algorithms for the calculation of glottal closure instants on speech spoken with modal voice and seven additional voice qualities. Results show that the performance of each method is affected by the voice type and that some methods perform better than others for each voice quality.

Index Terms: GCI, epoch detection, glottal source

10.21437/Interspeech.2011-523

1. Introduction

During voiced speech the vocal tract is excited by the modulated glottal airflow, produced by the vibration of the vocal folds. The moment of significant excitation or glottal closure instant (GCI) will be referred to as the glottal epoch in the current work. The glottal epoch is thought to have considerable perceptual importance not least because of the influence it has on the perceived pitch of the vocal production. Furthermore, it has been suggested that human listeners rely heavily on the presence of glottal epochs when perceiving speech in degraded conditions [1].

Being able to accurately identify glottal epochs is essential for a range of speech processing applications. For example accurate epoch locations are needed for deriving estimates of the glottal waveform using closed-phase inverse filtering [2] as well as other glottal inverse filtering methods (see e.g. [3]). Epoch locations are also usually required for pitch and speech rate modification using pitch-synchronous overlap-and-add methods, as well as for reducing phase mismatches in unit-selection speech synthesis [4].

A current direction of our research is to improve the fine-grained modelling of the glottal source waveform and to exploit this modelling in parametric speech synthesis [5]. Such approaches provide the potential for more flexible and possibly more expressive speech synthesis. It is, hence, desirable to be able to appropriately model non-modal voice qualities which play an active role in expressive speech. This has implications for epoch detection as non-modal voice qualities may display different glottal closure characteristics than those found in modal speech. For instance, it has been demonstrated that the glottal closure in breathy voice qualities is more of a smooth and sinusoid-like transition than the sharp closure in modal voice [6]. Epoch detection in creaky voice qualities presents an even stronger challenge. Many algorithms involve setting likely F_0 ranges, but as creaky phonation usually produces F_0 values below 70 Hz successful detection of epochs in such cases is unlikely. Furthermore, as certain voice qualities (e.g., creaky voice) can involve multiple excitations within a single glottal cycle the traditional definition of a glottal epoch is challenged.

It may be the case that for successful epoch detection in a range of voice types, certain epoch detection algorithms are robust in certain situations and other algorithms are suited to others. Most evaluation work on epoch detection has been carried out on modal, read speech by different speakers using databases like the ARCTIC database [7]. For the evaluation in the present work we looked to determine the accuracy of popular epoch detection algorithms on a common speech database of neutral voice as well as from sentences produced in different voice qualities. From this we hope to determine which algorithms provide robust epoch detection in modal speech as well as high accuracy in non-modal voice qualities. Furthermore, findings may demonstrate the need for further developments in epoch detection algorithms in the context of particular voice qualities.

2. Epoch Detection Approaches

This section reviews the methods for detection of the glottal closure instants which were evaluated in this work. It should be noted that the code for all the methods used in the present work were original implementations with the exception of the wavelet-based method which, as it was unavailable in its original form, was implemented by authors following the descriptions in the relevant paper [6] (described in Section 2.5).

2.1. Normalised Cross-Correlation on the Residual

A common pitch-tracking technique consists of detecting an optimal sequence of peaks from the quasi-periodic speech signal using dynamic programming applied to the correlation function, e.g. [8]. In this work, we use a popular epoch detection method which computes the normalised cross-correlation function (NCCF) on the linear prediction (LP) residual and uses dynamic programming to select the optimal sequence of NCCF (epoch estimates) across all short-time signals [9] (labelled here as ESPS). This epoch detector is available in the ESPS/*waves+* software package.

2.2. Residual Phase

The Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) [10] is a commonly used method for extracting glottal epochs and its implementation is available in the VOICE-

Copyright © 2011 ISCA

BOX Matlab toolbox. Initially the algorithm finds epoch candidates using an adaptation of the algorithm described in [11], which uses the phase slope of the group delay function. The method then uses a dynamic programming algorithm to minimise a weighted cost function (which is consists of Frobenius norm, pulse similarity and phase slope deviation cost elements).

2.3. Energy Contour of the Speech Signal

The algorithm described in [12] is commonly used particularly in methods for voice conversion/modification (here labelled *find_pmarks*). This method first obtains epoch candidates from the peaks of the filtered energy contour of the speech signal. Dynamic programming is used to find the best path of peaks that produce maximum energy and then to find the most likely epoch location for each frame, which is centered around the location of the epoch candidate.

2.4. Mean-based signal functions

In this study we also evaluate two methods that use a so-called mean-based signal for glottal epoch detection. The first method, which is described in [1], draws on observations that the impulse-like nature of the glottal closure is reflected across all frequencies, including 0-Hz, just as an ideal impulse. The method first derives the difference-speech signal, $x(n)$, by subtracting the previous speech sample from the current sample. $x(n)$ is then twice passed through an ideal resonator at zero frequency to give $y_z(n)$. The zero frequency signal $y(n)$ is then derived as follows

$$y(n) = y_z(n) - \frac{1}{2N+1} \sum_{m=-N}^{N} y_z(n+m) \qquad (1)$$

where $2N + 1$ is the number of samples in the 10 ms interval. Finally, the time instants of the positive zero-crossings are used as the glottal epochs. In this paper this method is labelled ZZF (Zeros of the Zero-Frequency resonator).

The next method, called SEDREAMS (Speech Event Detection using the Residual Excitation And a Mean-based Signal), uses a modified version of (1) to derive the mean-based signal $y(n)$ from the speech signal (originally described in [13]). In this case, the minima and following positive zero-crossings from the $y(n)$ signal are used to define time intervals between which the glottal epoch is expected to lie. In the most recent implementation of the method which was provided by the authors of [13], these intervals are defined as starting from the mean-based signal minima and spanning 0.35 the local pitch period (defined as the distance between consecutive minima). The epoch location is then detected by finding local maxima in the LP-residual (24^{th} order) that fall within the expected intervals.

2.5. Wavelet Transform

Features of the wavelet transform have been used for finding glottal epochs due to their suitability at detecting singularities in signals [6, 14]. In this work, we opted to implement the algorithm described in [6]. This method uses wavelet decomposition of the speech signal at eight octave bands of the spectrum using the mother wavelet, which is given by

$$g(t) = -cos(2\pi F_n t) \cdot exp(-\frac{t^2}{2\tau^2}) \qquad (2)$$

where $F_s = 16$ kHz is the sampling rate, $F_n = \frac{F_s}{2}$ and $\tau = \frac{1}{2F_n}$. Maxima are initially computed in the smallest wavelet scale with significant maxima (the authors suggest the wavelet at 4 kHz). From each of these maxima an optimal line is derived by descending through the scales producing Lines of Maximum Amplitude (LOMA). The epoch LOMA is selected as the LOMA with the highest energy within the expected pitch period. The location of the epoch is then considered to be index of the highest scale maximum. Insertion errors are removed following a post-processing procedure which considers consecutive pitch period and accumulated LOMA amplitude changes. The method was reported to be suited to both sharp and smooth glottal closure (which are usually found in breathy signals or during voice offset).

3. Evaluation

Epochs estimated from the electroglottographic (EGG) signal, which is a measure of the conductivity in the vocal folds, were used as reference epochs to evaluate the epoch detection algorithms. The epochs estimated using each of the methods described in the previous sections were then aligned with the reference epochs for each sentence, in order to calculate error measurements.

3.1. Speech Databases

Two American English voices from the ARCTIC speech database [7] were used for the evaluations, one male and the other female (the *bdl* and *slt* voices respectively). This database contains the same 1312 speech sentences spoken by the two speakers with a neutral voice quality (modal voice) and the respective EGG signals, from which the first 100 sentences were selected for the evaluation. We used this subset of the corpora instead of the whole number of sentences due to time constraints in checking and manually correcting the epochs detected from the EGG signal.

An additional database was used which consisted of ten sentences spoken by a native UK English speaker [5]. Both the speech and EGG signals were also available in this corpus and it included the sentences recorded in seven different voice qualities besides the modal voice: breathy, creaky, falsetto, harsh, lax, tense, and whisper (hence, 80 sentences in total). The sentences recorded with whisper were not used in this experiment because there is no glottal vibration present during the production of whisper. Furthermore, following listening to the sentences by two individuals involved in voice quality research all the sentences labelled 'creaky' were deemed not to display the auditory or acoustic characteristics of that particular voice quality. Instead, they were perceived as tense voice quality. However, they were not excluded from the analysis. In order to make clear that no evaluation of epoch detection of creaky voice qualities is covered in this paper those sentences are labelled as: creaky*. Another difference between this dataset (labelled Male VQ) and the ARCTIC is that the first contains sonorant sentences only (all-voiced), which avoids voicing classification errors.

3.2. Calculation of Reference Epochs

The amplitude of the EGG signal is higher the closer the vocal folds are to each other. The epochs can be estimated using a simple method to detect the maximal peaks in this signal. This peak tracking technique is usually effective and accurate because the EGG signal is highly periodic and has little noise and other aperiodicity effects, which are common in the voiced parts of the speech signal or the glottal source derivative signal estimated from speech. For calculating the epochs from the

EGG signal, we used the *pitchmarks* function of the Edinburgh Speech Tools [15] and a post-processing step to align the candidate epochs to the nearest peaks of the EGG signal. We did not find a significant delay between the EGG and the respective speech signal for the ARCTIC databases (we believe the available signals already have the delay corrected). However, we took into account the delay for the Male VQ dataset (approximately 0.9 ms).

Although the use of the EGG signal for epoch detection can provide accurate results, occasionally incorrect epoch detection or missing epochs may occur. We observed that an irregular shape of the EGG signal, e.g. containing more then one pulse during the glottal period, the ineffective removal of the DC component and noise effects were the most important causes of these errors. False epoch detection was the type of error we found most frequently. This problem was mainly due to the effect of multiple pulses occurring during one glottal period which caused multiple zero crossings in the derivative of the EGG and it was very common for speech spoken with certain voice qualities, such as breathy and harsh voice. The epochs estimated using the EGG signal were manually checked by the authors and the false detected epochs were deleted. Besides removing epochs, no other type of correction was made because the epochs obtained from the EGG were expected to be accurate and we very rarely found that it was necessary to insert epochs.

3.3. Evaluation of Epoch Detection Methods

A recursive algorithm was developed in this work to perform the alignment between the epochs estimated using the EGG signal (*reference epochs* $r(n)$) and the epochs estimated from speech using the different epoch detection techniques (*test epochs* $e(p)$). Figure 1 shows the block diagram of this algorithm.

We found that the epochs estimated using the ZZF method have an approximately constant offset compared to the EGG epochs. We tried to compensate for this delay by shifting the ZZF epochs by a positive factor of 0.7 ms for ARCTIC and 0.3 ms for Male VQ sentences.

In order to avoid the effect of differences between the voicing classification part of the methods (some methods do not include a voiced/unvoiced decision), we removed the test epochs generated by each method which were aligned to unvoiced pitchmarks of the EGG signal (the vector of reference epochs included equally spaced epochs in unvoiced regions). However, these unvoiced pitchmarks were only used for this purpose and they were discarded in the calculation of the evaluation metrics.

We opted to use similar evaluation metrics to those used in previous studies [1, 10, 13] to allow meaningful comparisons with the results obtained in the current work. The identification error, ζ, is measured as the difference between a given reference epoch $r(n)$ and the aligned test epoch $e(p)$. The Identification Accuracy metric (IDA) is defined as the standard deviation of ζ. We also use the mean squared error (MSE) of ζ. Another metric is the Miss Rate (MR), which is the percentage of glottal cycles for which the algorithm does not detect an epoch. In this work, it is calculated as the rate of 'insertions' obtained from the alignment described in Figure 1. Finally, the False Alarm Rate (FAR) is calculated as the percentage of the total number of test epochs not aligned with reference epochs.

For the voice quality sentences we examined whether voice quality as a factor affected the squared identification error, ζ^2, for all the values from the different algorithms combined using a one-way ANOVA. Then to further examine the effects of the voice qualities and the algorithm types we performed a two-way

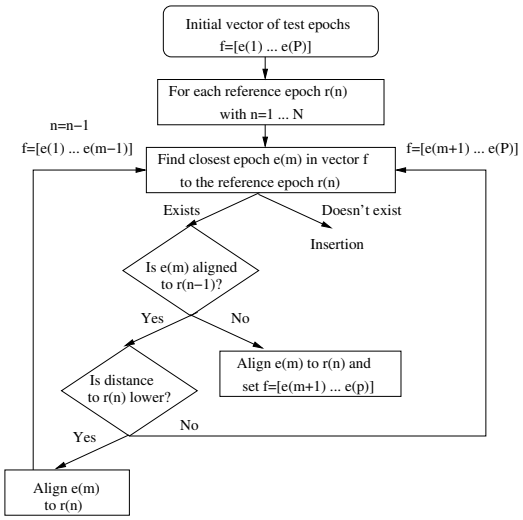


Figure 1: *Block diagram of the algorithm for alignment of the reference epochs (EGG epochs) to the test epochs.*

ANOVA (with voice quality and algorithm type as factors) also on ζ^2. Post-hoc testing was done using Tukey's HSD (Honestly Significant Difference) test.

4. Results

4.1. ARCTIC Dataset

The overall results for the ARCTIC dataset (combined male and female speakers) are presented in the upper part of Table 1. The $find_pmarks$ method performs clearly worse than the rest with considerably higher mean values for all metrics. The ESPS, ZZF, SEDREAMS and *wavelet* methods obtained comparable IDA and MSE values. However, the ESPS and ZZF methods obtained considerably higher MR than the other two, whereas the *wavelet* method obtained higher FAR. Since this was not the original implementation of the wavelet algorithm, the post-processing used in the original implementation may reduce the FAR. Previous studies showed that the SEDREAMS, ZZF and the *wavelet* methods outperformed DYPSA in terms of IDA [1, 6, 13] and this seems apparent again in this study.

4.2. Dataset with Different Voice Qualities

The overall results obtained for the voice quality dataset are presented in the lower part of Table 1. They show that the ESPS and SEDREAMS methods are comparable to each other and better than the other methods (with the exception that ESPS has higher MR than SEDREAMS). The DYPSA method performs relatively better for this dataset than ARCTIC, obtaining similar results to the *wavelet* method in terms of accuracy. However, the FAR and MR are still higher for DYPSA (we observed that the latter was due to a significant high MR for falsetto voice quality). Also, the ZZF method is worse in terms of accuracy compared to the previous methods but obtains good results in terms of MR and FAR. Similar to the ARCTIC voices, the $find_pitchmark$ epoch detector performed worse than the other methods.

One-way ANOVAs revealed a significant effect of voice quality on ζ^2 ($F = 136.26$, $df = 6$, $p < 0.001$). Post-hoc analysis revealed no significant differences ($p > 0.05$) for the pairs: breathy-lax, harsh-creaky*, modal-creaky*, modal-tense and tense-creaky*, but with significant differences for all other

Table 1: *Results of the epoch estimation methods for the two ARCTIC sets (combined) and the Male VQ set in terms of IDA (identification accuracy (ms)), MSE (mean squared error), MR (miss rate %) and FAR (false alarm rate %)*

Dataset	Method	IDA	MSE	MR	FAR
ARCTIC	ESPS	0.63	0.50	4.19	1.08
ARCTIC	DYPSA	0.82	0.68	2.54	4.33
ARCTIC	*find_pmarks*	1.06	1.12	0.94	3.78
ARCTIC	ZZF	0.64	0.45	3.22	1.89
ARCTIC	SEDREAMS	0.65	0.42	1.16	1.86
ARCTIC	*wavelet*	0.69	0.53	1.22	3.88
Male VQ	ESPS	0.45	0.21	1.0	0.7
Male VQ	DYPSA	0.56	0.33	1.5	4.1
Male VQ	*find_pmarks*	1.35	1.86	0.5	4.1
Male VQ	ZZF	0.67	0.63	1.0	0.4
Male VQ	SEDREAMS	0.48	0.27	0.5	0.4
Male VQ	*wavelet*	0.57	0.39	1.3	1.4

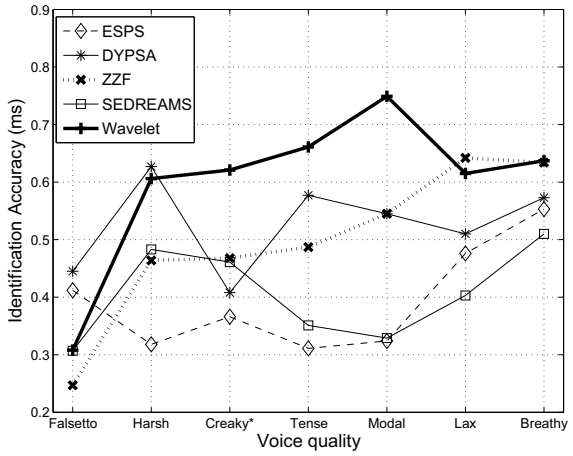


Figure 2: Identification Accuracy (IDA) for the epoch detection methods (excluding $find_pmarks$ which obtained significantly higher values) across the seven voice qualities.

pairwise comparisons. Figure 2 also shows that the epoch detection accuracy (IDA) generally varies in terms of the voice quality. For example, ESPS demonstrates clearly better accuracy for harsh voice compared to the other methods. On the other hand, SEDREAMS seems to be more accurate (lower IDA) than ESPS for lax, breathy and falsetto, although the difference is only significant for falsetto ($p < 0.01$). Figure 2 also suggests that the accuracy of ESPS, SEDREAMS and ZZF tends to be worse for lax and breathy voices compared to modal. This result could be due to the effect of aspiration noise which is characteristic of the breathy and lax voices, especially for the ESPS and SEDREAMS which estimate epochs by detecting amplitude peaks of the LP-residual. Furthermore, as was observed in [6] the smoother glottal closures usually present in lax and breathy voice qualities may also provide difficulties for epoch detection.

5. Conclusion

This study showed that the voice type has an impact on the robustness of epoch detection. Of the algorithms investigated, the ESPS and SEDREAMS methods performed similarly and generally obtained the best epoch identification accuracy across the voice qualities. Moreover, the SEDREAMS method also displayed the lowest false alarm and miss rates. The results obtained in this work can be used to select a suitable method for different types of voice quality in terms of different criteria: identification accuracy, false alarm rate and miss rate.

Future work will involve the inclusion of creaky voice qualities in the evaluation of epoch detection as well as development of methods suited to the characteristics of creaky and breathy voice qualities. The factors which affect the performance of the epoch estimation methods for different voice qualities will also be further investigated. Finally we plan to extend these experiments to include speech data from a larger number of speakers.

6. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (`www.cngl.ie`) at University College Dublin and Trinity College Dublin. The opinions, findings and conclusions, recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

7. References

[1] Murty, K., Yegnanarayana, B., "Epoch extraction from speech signals", IEEE Tr. Aud. Sp. Lan. Proc., 16, 1602–1613, 2008.

[2] Chan, D., Brookes, D., "Variability of excitation parameters derived from robust closed phase glottal inverse filtering", in Proc. of Eurospeech, Paris, 2199–2202, 1989.

[3] Bozkurt, B., Couvreur, L., Dutoit, T., "Chirp group delay analysis of speech signals", Speech Comm., 49 (3), 159–176, 2007.

[4] Stylianou, Y., "Removing linear phase mismatches in concatenative speech synthesis", IEEE Tr. Aud. Sp. Lan. Proc., 9 (3), 232–239, 2001.

[5] Cabral, J. P., "HMM-based speech synthesis using an acoustic glottal source model", Ph.D. Thesis, The Uni. of Edinburgh, 2010.

[6] Sturmel, N., d'Alessandro, C., Rigaud, F., "Glottal closure instant detection using lines of maximum amplitudes (LOMA) of the wavelet transform", in Proc. of ICASSP, 4517–4520, 2009.

[7] Kominek, J. and Black, A., "The CMU Arctic speech databases", in Proc. of 5th SSW", Pittsburgh, USA, 2004.

[8] Secrest, B.G., and Doddington, G.R., "An integrated pitch tracking algorithm for speech systems", in Proc. of ICASSP, 1352–1355, 1983.

[9] Talkin, D., "Voicing epoch determination with dynamic programming", J. Acoust. Soc. Amer., 85, Supplement 1, 1989.

[10] Naylor, P., Kounoudes, A., Gudnason, J., Brookes, M. "Estimation of glottal closure instants in voiced speech using the dypsa algorithm", IEEE Tr. Aud. Sp. Lan. Proc., 15 (1), 34–43, 2007

[11] Smits, R., Yegnanarayana, B., "Determination of instants of significant excitation in speech using group delay function", IEEE Tr. Aud. Sp. Lan. Proc., 3, 325–333, 1995

[12] Goncharoff, V., Gries, P., "An algorithm for accurately marking pitch pulses in speech signals", in Proc. of SIP, USA, 1998.

[13] Drugman, T, Dutoit, T., "Glottal closure and opening instant detection from speech signals", in Proc. of Interspeech, 2009.

[14] Vu Ngoc Tuan, d'Alessandro, C., "Robust glottal closure detection using the wavelet transform", in Proc. of Europ. Conf. Speech Tech., Eurospeech, Budapest, 2805–2808, 1999.

[15] King, S., Black, A.W., Taylor, P., Caley, R., Clark, R., "Available at http://www.cstr.ed.ac.uk/projects/speech_tools", The Uni. of Edinburgh.