GlóRí - the *Gl*ottal *R*esearch *I*nstrument

John Dalton, John Kane, Irena Yanushevskaya, Ailbhe Ní Chasaide, Christer Gobl

Phonetics and Speech Laboratory,
School of Linguistic, Speech and Communication Sciences,
Trinity College Dublin

jrdalton@tcd.ie, kanejo@tcd.ie, yanushei@tcd.ie, anichsid@tcd.ie, cegobl@tcd.ie

Abstract

This papers presents GlóRí - the glottal research instrument. GlóRí is a speech analysis interface which offers a flexibility and multiplicity of approaches to voice analysis. The system allows for fully automatic processing, for instance for analysis of large corpora. However, for more fine-grained studies, which may require precise voice source measurements, the systems facilitates manual optimisation of parameter settings. The present paper highlights the main features of the GlóRí system and provides illustrations of the usefulness of this approach.

Index Terms: Glottal source, voice source, phonetic features, voice quality

1. Introduction

The research being carried out by the voice processing group at the Phonetics and Speech Laboratory in Trinity College Dublin is concerned with the development of robust voice source processing methods and analysing the function of the voice source in prosody. As part of this endeavour, we have been developing our speech analysis methods so as to be able to handle the inherently different acoustic characteristics of the speech signal and to be adaptive and flexible according to the phonetic and prosodic context. This paper presents GlóRí - the glottal research instrument. Note that [glo:ɰi:] is the Irish (Gaelic) word for *voices*. GlóRí is the voice analysis system within which our ongoing developments in voice source analysis and data visualisation will be integrated.

The ability to derive precise and robust measurements of the voice source is becoming increasingly important for speech technology applications (e.g., speech synthesis [1, 2], emotion classification [3, 4]), as well for linguistic analysis on the prosody of the voice [5, 6] and also for voice pathology and voice function assessment [7]. Ideally, one would wish to be able to derive robust parameterisation of the voice source completely automatically. However, this process typically involves three non-trivial steps.

The first step, in order to allow glottal pulse-synchronous analysis, is to estimate glottal closure instants (GCIs, [8]). After several decades of research state-of-the-art GCI detection is at a sufficiently high level of performance for the analysis of neutral read speech. Despite this, a recent study [9] demonstrated how different phonation types (and in particular creaky voice) deteriorated GCI detection performance. For disordered voices, where there is indeed a significant excitation within each glottal cycle, the deterioration is likely to be significantly more.

The next step, typically involves glottal inverse filtering, the process of compensating for the effect of vocal tract resonance from the speech signal. The most commonly used vocal tract model is an all-pole model, which can be estimate by linear predictive analysis, or similar methods. For nasals consonants and nasalised vowels there are generally zeros present in the vocal tract spectrum, making the all-pole model less suitable and this has negative effects on subsequent voice source parameterisation [10]. For speech with a low first formant frequency (F1), discrimination of F1 from the glottal formant can be problematic. This is particularly true when combined with a high f_0, which has the additional effect of having more widely spaced harmonics making the vocal tract spectral envelope more difficult to estimate effectively.

Finally, once an estimate of the voice source has been derived, for many purposes one typically then requires a parametric description of the signal. The two main approaches to this are to either take measurements from the voice source estimate directly or fit a mathematical model to the individual glottal pulses. Quotients characterising the timing of important events in the glottal cycle are generally thought to be the most salient parameters, regardless of which of the two approaches are used. One critical event is the instant of glottal opening, a reference point required for most time-quotient parameters. Localisation of this time instant is extremely difficult, not least because the glottal opening does not always display a significant discontinuity (e.g., in lax phonation). For the direct measure approach, the more commonly used parameters are generally amplitude-based [11, 12] or frequency domain correlates of time quotients [13].

From the three steps described above there is clearly wide scope for the introduction of significant errors, particularly in the case of natural, expressive speech or indeed even moderately disordered speech. In many instances, an experienced speech science researcher may be able to make adjustments to the automatic process and improve the overall effectiveness of the analysis. With this firmly in mind, the newly developed analysis system, GlóRí, has been designed to allow both fully automatic analysis while also facilitating manual intervention and optimisation at various stages in the analysis. From the outset there are three main characteristics that are fundamental to our system design.

1. **Adaptive** The system should allow a multiplicity of approaches, e.g., for research on large corpora fully automatic analysis can be deployed, but for more fine-grained analysis the researcher should be able to manually optimise the analysis to ensure maximal precision. Furthermore, the analysis should be adaptive to the phonetic and prosodic context, e.g., allowing glottal inverse filtering with an adaptive vocal tract modelling.
2. **Modular** Ongoing development of various voice source and speech analysis algorithms should be easy to incor-

10.21437/SpeechProsody.2014-177

porate into the system. To this end we created the interface in the Matlab programming environment. As our algorithm development (as well as much of the signal processing development in the speech research community) is done using this environment, it facilitates newly developed algorithms being easily incorporated.

3. **Knowledge** We intend for the system to incorporate various sources of *knowledge* in the analysis. This can be considering the given phonetic class being analysed and adapting the analysis accordingly (e.g., using a vocal tract model which included pole-zeros for nasal regions). It can also include incorporating knowledge from speech production theory, e.g., precluding parameterisation which is outside the physical boundaries of human speech.

1.1. Existing voice source analysis systems

There are a small number of interfaces for voice source analysis available in the literature. The APARAT system is one interface which facilitates automatic glottal inverse filtering and voice source parameterisation using a range of existing parameters from the literature [14]. The system is available under an open-source licence[1] and has encourage using voice source feature extraction in a range of speech-related areas.

Another freely available interface for voice analysis is the Voice Sauce program[2] [15]. Voice Sauce enables a wide range of voice-related speech analysis including f_0 and harmonic extraction, formant tracking and the formant compensation proposed by Hanson [13], harmonic and subharmonic to noise ratio, energy and cepstral peak prominence extraction. Voice sauce also includes a facility for analysis of electroglottographic (EGG) waveforms.

Although there is some overlap with these interfaces, GlóRí is a useful complement and there are several major differences compared with existing systems. First, GlóRí allows manual intervention and optimisation. Second, GlóRí includes some very recently developed voice quality related analysis methods. A third major difference is that we are intending for the system to allow incorporation of speech production knowledge and to involve pre-processing steps which could be used to contrain possible analysis settings and also, where necessary adapt the analysis (e.g., the structure of the vocal tract model) to more closely match the acoustic structure of the given speech segment. Furthermore, GlóRí includes resynthesis and data vizualisation components that facilitate construction of stimuli for perception experiments as well as allow to represent the analysed data for visual inspection in a number of ways.

2. System features

This section serves to illustrate the main system features of the GlóRí system. The system was designed to be user-friendly and to allow manual analysis, if it is deemed necessary, or completely automatic voice source feature extraction.

2.1. Manually-optimised analysis

Voice source analysis, including the possiblity for manually-optimised analysis, can be carried out using the analysis window shown in Figure 1. When a speech sample is loaded into

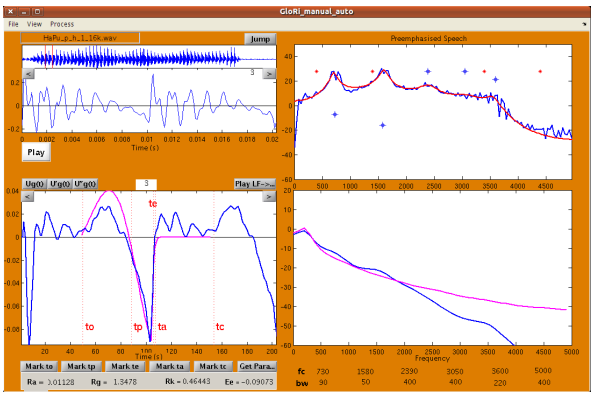


Figure 1: Screenshot of the manual analysis interface of the GlóRí system. Users can choose from a selection of parameterisation approaches, derived from the voice source estimate, using parameters derived from model-fitting and parameters derived from the speech signal.

system, it is resampled to 10 kHz. GCIs are located automatically using our recently developed algorithm (SE-VQ, [9]) and GCIs detected in unvoiced regions are excluded. GCI locations can then be manually edited later if required using the GCI editor. Locations that are judged to be false can be deleted and undetected locations can be added. Although state-of-the-art GCI detection has reached a mature level of performance, this can still degrade when analysing speech involving wide variation in phonation type [9] or the voice is disordered. Allowing a facility for manual intervention here may enable more precise analysis for these types of speech.

For each GCI-centred two pulse length frame, the vocal tract model can be constructed by setting the formants frequencies and bandwidths. The frequency and bandwidth of each formant can be adjusted using the keyboard arrow keys, and a time and frequency domain representation is available to assess the effect of the inverse filter. As each anti-formant nears its optimal location, the oscillations of the corresponding formant will be dampened in the time domain (see bottom left panel of Figure 1), and the formant peak will be largely attenuated in the frequency domain (see bottom right panel). Once the speech signal has been inverse filtered the user can then move to the parameterisation step. The manually optimised system allows parameterisation of the estimated voice source signal by fitting the Liljencrants-Fant (LF) voice source model [16] to the individual glottal pulses. An LF model can be fitted to the inverse-filtered pulse by manually adjusting the time-points of the model (see bottom left panel of 1). Fitting is facilitated with both the time and frequency domains displays in the two adjacent panels, allowing the user to achieve accurate time-point matches while also ensuring close spectral fitting.

2.2. Fully automatic analysis

In contrast to the manually-optimised analysis, a fully automatic analysis approach is included in the GlóRí system (see Figure 2). The analysis relies entirely on the use of automatic algorithms, without any intervention from the user. A folder of speech samples is loaded through the interface, and the desired analysis parameters are selected.

These fall under three categories. The category title “Glottal params” is further subdivided into two different types. Un-

[1] http://sourceforge.net/projects/aparat/
[2] http://www.ee.ucla.edu/~spapl/voicesauce/

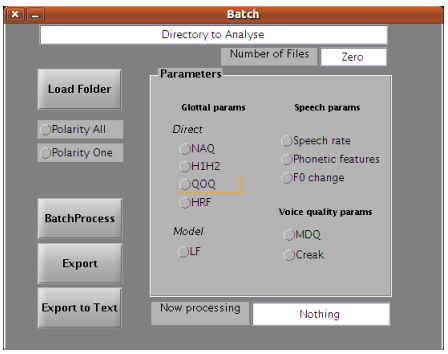


Figure 2: Screenshot of the fully automatic analysis interface of the GlóRí system. Users can choose from a selection of parameterisation approaches, derived from the voice source estimate, using parameters derived from model-fitting and parameters derived from the speech signal.

der "Direct" one can select parameters which are derived using direct measurements of the glottal inverse filtered signal. These include: the normalised amplitude quotient (NAQ; [12]), the difference between the amplitude of the first two harmonics (H1-H2; [13]), the quasi-open quotient (QOQ; [11]), and the harmonic richness factor (HRF; [19]). Under "Model" one can select to have the glottal inverse filtered signal parameterised by fitting LF-model pulses to the individual glottal pulses using our recently developed automatic fitting algorithm [20]. Note that prior to estimation of these parameters, the inputted speech signal is inverse filtered using iterative and adaptive inverse filtering (IAIF, [21]).

Under the category "Voice quality parameters", one can select the maxima dispersion quotient (MDQ; [22]) and Creak [23]. MDQ is a wavelet based algorithm which discriminates breathy and tense voice quality by assessing the dispersion of peaks across a range of frequency bands relative to the GCI. The Creak parameter gives the binary output of a decision tree classifier, using two input features derived from the Linear Prediction (LP) residual signal.

Finally, the "Speech params" category allows selection of parameters related directly to aspects of the speech signal. Phonetic feature extraction selected and this outputs a continous score on the likelihood of the presence of a range of phonetics features {voiced, syllabic, nasal, liquid, fricative, plosive}. This is done using the algorithm recently proposed in [24]. Note that this algorithm provides important information on the underlying manner of articulation is various speech regions which can facilitate analysis strategies which are adaptive to the phonetic context. This algorithm can also be harnessed for deriving a 'speech rate' measurements, in terms of syllables per second.

2.3. Synthesis interface

Once a given speech signal has been analysed, using either manual or automatic methods, one can then load the exported analysis file into a synthesis interface. As shown in Figure 3 a user is provided with parameter contour displays. The user can modify parameter contours by clicking new points on the panel, as has been done for the f_0 parameter (top panel). It is then possible to resynthesise the speech using the modified parametric setting. This is a useful facility for stimuli generation to be used in perception experiments.

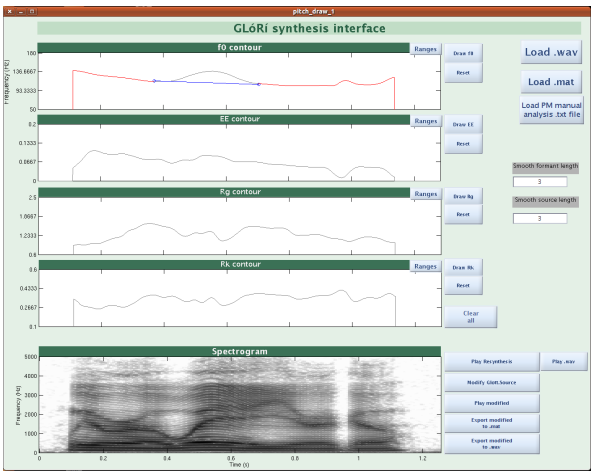


Figure 3: Screenshot of the synthesis interface of the GlóRí system. Users are shown display of parameter settings and can make alterations to these contours and resynthesise the speech.

2.4. Visualisation interface

An interface is also included for easy visualisation of extracted parameter contours. By loading in an analysis file, again either using manual or automatic systems, one can select combinations of parameters to be plotted together with the speech spectrogram. We have also begun to experiment with novel visualisation approaches for showing high dimensional parameter data in a clear single plot, e.g., using spidergrams (illustrated below). These novel developments are incorporated within this interface component of the GlóRí system.

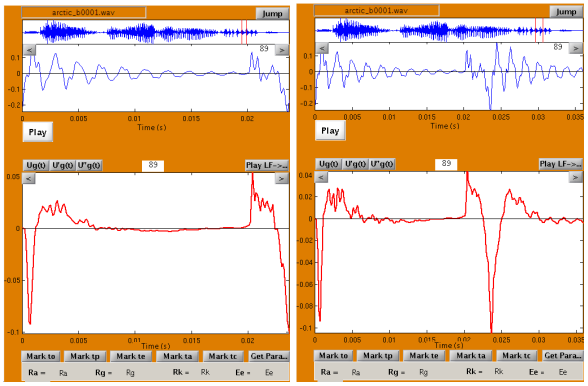
3. Illustrations

This section serves to provide illustrations of how the system features of GlóRí may be beneficial for a range of analysis purposes.

The first illustration highlights the importance of allowing manual intervention to improve the precision of the analysis in certain cases. In the left panel of Figure 4 one can observe the negative impact of a false positive GCI, as often occurs in creaky voice (see [9]), on the overall analysis. The main glottal excitation should be located close to the centre of the analysis panel, and, hence, the false positive observed here will preclude the possiblity of obtaining sensible parameter values if a completely automatic approach was used. However, by exploiting the ability for manual intervention, in this case facilitated by the manual GCI editor, one can easily delete this false positive and proceed to effective voice source modelling as shown in panel (b) of Figure 4.

One crucial intention in the development of the GlóRí system is to facility incorporating knowledge (in its various forms) to help constrain and augment the analysis. In particular, it is desired to facilitate analysis that is sensitive and adaptive to the phonetic environment of the speech signal. As mentioned previous, our phonetic feature extraction algorithm can provide us with initial information on the underlying manner of articulation in a given utterance [24]. Figure 5 shows the output of this feature extraction for a sample utterance.

The information yielded by these phonetic feature extractors may be beneficial for a range of purposes in the analysis



(a) GCI false positive (b) Corrected GCI false positive

Figure 4: Screenshot of analysis of a creaky voice pulse involving a false positive GCI (a) and with the false positive corrected (b).

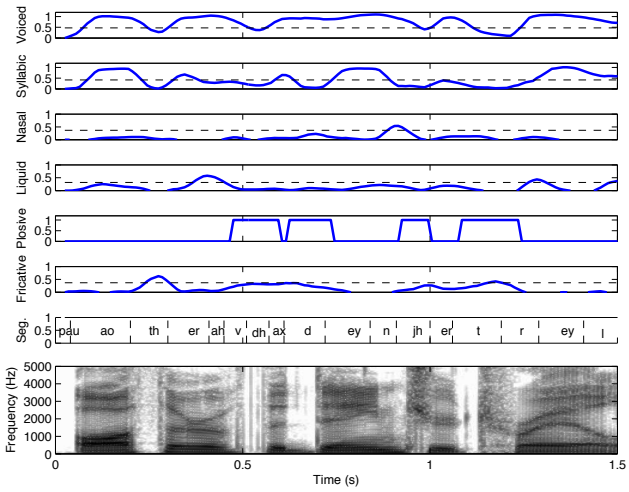


Figure 5: Illustration of phonetic feature extraction for the utterance “Author of The Danger Trail ...”, spoken by an American male.

interface. For voice source feature extraction for the purpose of voice quality classification one may choose to exclude certain phonetic regions known to cause problems for analysis (e.g., areas of frication, or nasality). Another usage could be for adaptive glottal inverse filtering, where the vocal tract model could be adapted to the given region, e.g., in nasalised regions pole-zeros could be incorporated into the vocal tract model. Besides these uses, the phonetic feature extraction may also provide a useful guide for the researcher when carrying out manually-optimised voice source analysis.

Another important component of the GlóRí system is data visualisation. A frequently used approach when analysing voice source parameters is to reduce the data, often to a single shape parameter. However, this approach may at times be premature and may involve losing important information to do with the glottal pulse shape. In order to avoid premature data reduction and to display the voice source parameter data in an accessable form the GlóRí system allows plotting of the data as a “spider-

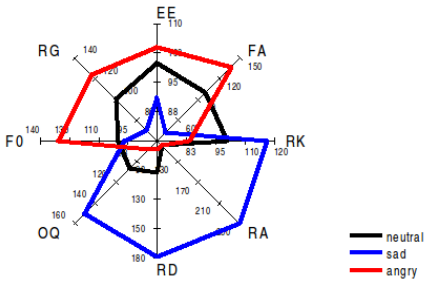


Figure 6: Spidergram plot summarising high dimensional voice source data into a single plot. Note that the axes are the parameter values in percentage relative to neutral.

gram” as shown in Figure 6 [4]. In the spidergram, parameters are arranged in such a way that increased parameter levels above the horizontal axis typically indicate a tenser phonation. Similarly, levels extending below the horizontal axis point to a laxer phonation. The illustration in Figure 6 shows an example spidergram for a sentence spoken with three types of affective colouring: neutral, sad and angry. The blue web for *sad* with it’s increased parameter levels below the horizontal line provides strong evidence of a laxer phonation type, whereas *angry* (with the red web) indicates a tenser phonation.

4. Discussion & conclusion

This paper presented the new voice analysis system, GlóRí. The system is shown to facilitate completely automatic voice source feature extraction, and incorporates a range of state-of-the-art voice source analysis developments as well as existing parameters from the literature. The automatic system may be extremely useful for studies across a range of speech-related disciplines when analysing large corpora, and could in particular be useful for allowing voice source feature extraction for researchers from a non-technical or non-voice related background.

A further benefit of the GlóRí system over existing analysis systems, is that it facilitates manually-optimised analysis. This may be critical for very fine-grained analysis studies which require precise voice source parameter data. Manual intervention here may help reduce the effect of error introduction in the various stages of analysis.

The GlóRí system is intended to be a constantly work-in-progress development. One main direction for ongoing and future research is to bring to bear our knowledge of speech production so we can constrain possible vocal tract model and voice source parameterisation solutions. Our newly developed fully automatic techniques (for instance for deriving information to do with breathy, tense and creaky voice, as well as the underlying phonetic features) can provide prior information that can be used to constrain vocal tract filter and voice source modelling. We intend to make this system publicly available in the near future.

5. Acknowledgements

This research is supported by the Science Foundation Ireland Grant 09/IN.1/I2631 (FASTNET) and the Irish Department of Arts, Heritage and the Gaeltacht (ABAIR project).

6. References

[1] Cabral, J., Renals, S., Richmond, K., Yamagishi, J., (2011) "HMM-based speech synthesiser using the LF-model of the glottal source", Proceedings of ICASSP, Prague, Czech Republic, 4704-4707.

[2] Degottex, G., Lanchantin, A., Roebel, A., Rodet, X., (2012) "Mixed source model and its adapted vocal-tract filter estimate for voice transformation and synthesis", Speech Communication, 55(2), 278-294.

[3] Lugger, M., (2007) "The relevance of voice quality features in speaker independent emotion recognition", Proceedings of ICASSP, Hawaii, USA, 17-20.

[4] Yanushevskaya, I., Gobl, C., Ní Chasaide, A., (2009) "Voice parameter dynamics in portrayed emotions", Proceedings of Maveba, 21-24.

[5] Yanushevskaya, I., Gobl, C., Kane, J., Ní Chasaide, A., (2010) "An exploration of voice source correlates of focus" Proceedings of Interspeech, Makuhari, Japan, 462-465.

[6] Ní Chasaide, A., Yanushevskaya, I., Gobl, C., (2011) "Voice source dynamics in intonation" Proceedings of ICPhS, Hong Kong, 1470-1473.

[7] Gómez-Vilda, P., Fernández-Baillo, R., Rodellar-Biarge, V., Lluis, V. N., Álvarez-Marquina, A., Mazaira-Fernández, L., M., Martínez-Olalla, R., Godino-Llorente, J. I., (2009) "Glottal Source biometrical signature for voice pathology detection", Speech Communication 51(9), 759-781.

[8] Naylor, P., Kounoudes, A., Gudnason, J., Brookes, M., (2007) "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm" IEEE Transactions on Audio Speech and Language processing, 15(1), 34-43.

[9] Kane, J., Gobl, C., (2013) "Evaluation of glottal closure instant detection in a range of voice qualities", Speech Communication 55(2), 295-314.

[10] Gobl, C., Mahshie, J., (2013) "Inverse filtering of nasalized vowels using synthesized speech", Journal of Voice, 27(2), 155-169.

[11] Hacki, T., (1989) "Klassifizierung von glottisdysfunktionen mit hilfe der elektroglottographie" Folia Phoniatrica, 43-48.

[12] Alku, P., Bäckström, T., Vilkman, E., (2002) "Normalized amplitude quotient for parameterization of the glottal flow" Journal of the Acoustical Society of America, 112(2), 701-710.

[13] Hanson, H. M., (1997) "Glottal characteristics of female speakers: Acoustic correlates" Journal of the Acoustical Society of America, 10(1), 466-481.

[14] Airas, M., (2008) "TKK Aparat: An environment for voice inverse filtering and parameterization" Logopedics Phoniatrics Vocology, 33, 49-64.

[15] Shue, Y-L, Keating, P., Vicenik, C., Yu, K., (2011) "Voice sauce: a program for voice analysis" Proceedings of ICPhS.

[16] Fant, G., Liljencrants, J., Lin, Q., (1985) "A four parameter model of glottal flow" KTH, Speech Transmission Laboratory, Quarterly Report, 4, 1-13.

[17] Gobl, C., (1988) "Voice source dynamics in connected speech", KTH, Speech Transmission Laboratory, Quarterly Report, 29, 123-159.

[18] Walker, J., and Murphy, P., (2007) "A review of glottal waveform analysis" *in* Progress in nonlinear speech processing, 1-21.

[19] Childers, D. G., Lee, C. K., (1991) "Voice quality factors: Analysis, synthesis and perception", Journal of the Acoustical Society of America, 90, 2394-2410.

[20] Kane, J., Gobl, C., (2013) "Automating manual user strategies for precise voice source analysis", Speech Communication 55(3), 397-414.

[21] Alku, P., (1992) "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering", Speech Communication, 11(2-3), 109-118.

[22] Kane, J., Gobl, (2013) "Wavelet maxima dispersion for breathy to tense voice discrimination", IEEE Transactions on Audio, Speech and Language Processing, 21(6), 1170-1179.

[23] Kane, J., Drugman, T., Gobl, (2013) "Improved automatic detection of creak", Computer Speech and Language, 27(4), 1028-1047.

[24] Kane, J., Aylett, M., Yanushevskaya, I., Gobl, [Under review] "Phonetic feature extraction for context-sensitive glottal source processing", Speech Communication.