Evaluation of Automatic Glottal Source Analysis

John Kane and Christer Gobl

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

Abstract. This paper documents a comprehensive evaluation carried out on automatic glottal inverse filtering and glottal source parameterisation methods. The experiments consist of analysis of a wide variety of synthetic vowels and assessment of the ability of derived parameters to differentiate breathy to tense voice. One striking finding is that glottal model-based parameters compared favourably to parameters measured directly from the glottal source signal, in terms of separation of breathy to tense voice. Also, certain combinations of inverse filtering and parameterisation methods were more robust than others.

1 Introduction

The production of voiced speech can be considered as: the sound source created by the vibration of the vocal folds (glottal source) inputted through the resonance structure of the vocal tract and radiated at the lips. Most acoustic descriptions typically used in speech processing involve characterisation of mainly the vocal tract contribution to the speech signal. However, there is increasing evidence that development of independent feature sets for both the vocal tract and the glottal source components can yield a more comprehensive description of the speech signal. Recent developments in speech synthesis [1], voice quality modification [2], voice pathology detection [3] and analysis of emotion in speech [4] have served to highlight the potential of features related to the glottal source.

However, approaches for analysing the estimated glottal source are at times believed to lack robustness in certain cases. For instance, higher pitch voices are known to be problematic for inverse filtering [5] and particularly when combined with a low first formant frequency. There can be strong source-filter interaction effects [6] which seriously affect the linear model of speech exploited in inverse filtering. Furthermore, precise glottal source analysis is often said to require the use of high-quality equipment to record in anechoic or studio settings [5]. Despite these claims, some studies have found that glottal source parameters derived from speech recorded in less than ideal recording conditions contribute positively to certain analyses [7].

It follows that the purpose of this paper is to investigate the performance of both inverse filtering and parameterisation steps typically used in glottal source analysis. The evaluation of glottal source analysis methods is known to be problematic as it is not possible to obtain 'true' reference values. To deal with this,

T. Drugman and T. Dutoit (Eds.): NOLISP 2013, LNAI 7911, pp. 1–8, 2013.

the current study presents two different evaluation procedures in order to provide a more thorough impression of the performance of the various methods. Some similar work was recently carried out in [8] and the current study builds on this by incorporating model-fitting based parameterisation methods.

2 State-of-the-Art

A description of the state-of-the-art in terms of automatic glottal inverse filtering and glottal source parameterisation methods was previously given in [9] and [5]. For the evaluation in the present study the following glottal inverse filtering methods are evaluated: a closed-phase inverse filtering method (CPIF, [10]), iterative and adaptive inverse filtering (IAIF, [11]) and mixed-phase decomposition based on the complex-cepstrum (CCEPS, [12]). Note that for these methods glottal closure instants (GCIs) are detected using the SE-VQ algorithm [13]. For the CPIF method, the glottal closed phase is determined by detecting glottal opening instants (GOIs) using the algorithm described in [14].

The glottal source parameterisation methods are divided into two groups: direct measures and model fitting. The direct measures used in the current study are: the normalised amplitude quotient (NAQ, [15]), the quasi-open quotient (QOQ, [16]) and the difference between the first two harmonics of the narrow-band glottal source spectrum (H1-H2, [21]). These three parameters are chosen as they were shown to be particularly effective at discriminating breathy to tense voice in a previous study [17]. Two algorithms are included which involve fitting the Liljencrants-Fant (LF) glottal source model [22] to the glottal source signal. A standard time domain method is used (Strik-LF, [18]) and an algorithm based on dynamic programming (DyProg-LF) described in [26]. One further algorithm is used in the evaluation which provides an estimate of the *Rd* parameter of the LF model by minimising a phase-based error criterion (Degott-LF, [20]).

3 Experimental Setup

As any single evaluation of glottal source analysis has its shortcomings, the approach here is to evaluate both on synthetic and natural speech data.

3.1 Synthetic Testing

A frequently used evaluation procedure (see e.g., [18,8]) is to do analysis of synthetic vowel segments where there are known reference values. This has the advantage of allowing straightforward quantitative evaluation where systematic modifications to both vocal tract and glottal source model settings can be investigated. The disadvantage, however, is that the stimuli will be a simplified version of real speech and will not contain some of the known difficulties for glottal source analysis (e.g., the presence of aspiration noise, source-filter interaction effects, etc.). In this paper, analysis is carried out on a large range of

synthetic vowel segments with wide variation of glottal source and vocal tract filter model settings. This is done in a similar fashion to that in [8]. The LF glottal source model is used to generate the synthetic source signal and is varied using f_0 and three parameters which can be used to characterise its shape: Ra, Rk and Rg. With each setting 10 LF pulses are concatenated to create the source signal. An all-pole vocal tract model is used to modulate the source signal. Eight vowel settings are used based on the analysis of spoken vowels (i.e. one vocal tract model used to characterise each of the eight vowels). Note that the first formant frequency (F1) is derived from the vocal tract model, and we consider error rates as a function of F1. In total 198,720 synthetic signals (each containing 10 concatenated synthetic glottal pulses) are generated for analysis. A small proportion of these variations result in improper LF model configurations (i.e. when $Rk > 2Rg - 1$ or when $Ra > 1 - \frac{1+Rk}{2Rg}$), and these signals are not analysed.

In order to evaluate the performance of automatic inverse filtering the following three parameters are considered: NAQ, QOQ and H1-H2. These parameters are calculated from the synthetic source signal, as reference values. Then for each synthetic vowel the three inverse filtering methods: CPIF, IAIF and CCEPS, are used to estimate the source signal, which is subsequently parameterised. Relative error scores are then computed for each parameter and then are analysed as a function of f_0 values and first formant frequency (F1, derived from the all-pole settings).

3.2 Voice Quality Differentiation

One useful application of glottal source analysis is to automatically differentiate voice quality. Furthermore, as NAQ, QOQ, and H1-H2 have been shown to be suitable for separating breathy to tense voice (see for example: [17,15,16,21]) it is reasonable to assume that quality of inverse filtering can be somewhat evaluated on the basis of how well the extracted glottal source parameter differentiates the voice quality. Such an approach has been used in previous studies [23,8] and can allow quantitative evaluation on natural speech.

The speech data consist of all-voiced spoken sentences from two separate databases. The use of all-voiced sentences allowed evaluation independent of the effects of using automatic voicing decision algorithms. Furthermore, as voicing transitions often display characteristics associated with laxer phonation this would affect the results. The speech data from 6 speakers (3 male and 3 female) were selected from the speech database first described in [13]. Participants were asked to read a set of phonetically balanced TIMIT sentences in six different phonation types (though only the 5 all-voiced breathy, modal and tense samples are used here, i.e. 6-speakers $\times$ 5-sentences $\times$ 3-phonation types). Following a final perceptual evaluation 10 of the intended tense utterances were not perceived as such, and hence were discarded from the analysis. A further 10 all-voiced sentences produced by 3 male speakers in breathy, modal and tense voice, were recorded and added to the sentence dataset (giving a total of 60 breathy, 60 modal and 50 tense utterances). The three male speakers are all experienced in voice-related research and individual utterances were re-recorded in several

iterations until the sentences were deemed to properly represent the stated voice quality mode for the entire utterance. All audio was captured in a semi-anechoic studio using high quality recording equipment: a B & K 4191 free-field microphone and a B & K 7749 pre-amplifier. The signals were digitised at 44.1 kHz and were subsequently downsampled to 16 kHz.

For each included speech segment inverse filtering is carried out using CPIF, IAIF and CCEPS, and parameterised using NAQ, QOQ and H1-H2. Furthermore, *Rd* and OQ parameters are derived from the model fitting by the **Strik-LF** and **DyProg-LF** methods, following IAIF inverse filtering. *Rd* is also derived using **Degott-LF**, which does not require prior inverse filtering. In order to have a balanced dataset it is desirable to have a fixed number of datapoints per sentence. To address this, parameter contours are derived using each of the methods. These contours are then resampled to 10 samples which can capture variations in the parameter contour but still maintaining a constant number of datapoints. An explained variance metric is then derived as the squared Pearson's R coefficient by treating median parameter values as the dependent variable and voice quality label as the independent variable. A similar evaluation procedure was carried out in [17].

4 Results

4.1 Synthetic Testing

The results from the synthetic testing are shown in Figure 1, with mean relative error plotted as a function of f_0 setting and F1 (derived from the vocal tract models). The NAQ parameter is shown to be rather insensitive to variations in f_0 (Figure 1a). Below around 240 Hz the IAIF method produces the lowest relative error; however from after this point the three inverse filtering methods yield similar results. Although these results corroborate previous findings in [8] for the performance of NAQ on synthetic data other studies on natural speech have found that NAQ becomes less effective with wide f_0 variation [19].

For F1, NAQ is shown to be insensitive to its variation. Again IAIF provides the lowest relative error scores, although there is a sudden increase for the vowel setting with an F1 of 344 Hz. This can be explained by the fact that this is a /u/ vowel setting with a very low second formant. IAIF may at times treat this as a single formant resulting in incomplete formant cancellation thus affecting NAQ. For QOQ, the closed-phase inverse filtering method (CPIF) provided the lowest relative error scores. This is particularly true for higher f_0 values, with both IAIF and CCEPS showing significant increases in relative error from around 200 Hz. There is a clear effect of certain vowel settings on IAIF and CCEPS, but they are clearly not as a result of F1. CPIF is not affected by the different vowel settings. In the case of H1-H2, however, CPIF gave clearly the highest relative error values. It is apparent from the analysis that even though the extracted time domain waveform, using CPIF, is suitable for deriving time domain parameters, it is considerably less so for the frequency domain one. The CPIF method is unable to reliably extract the relative amplitude of the first few harmonics.

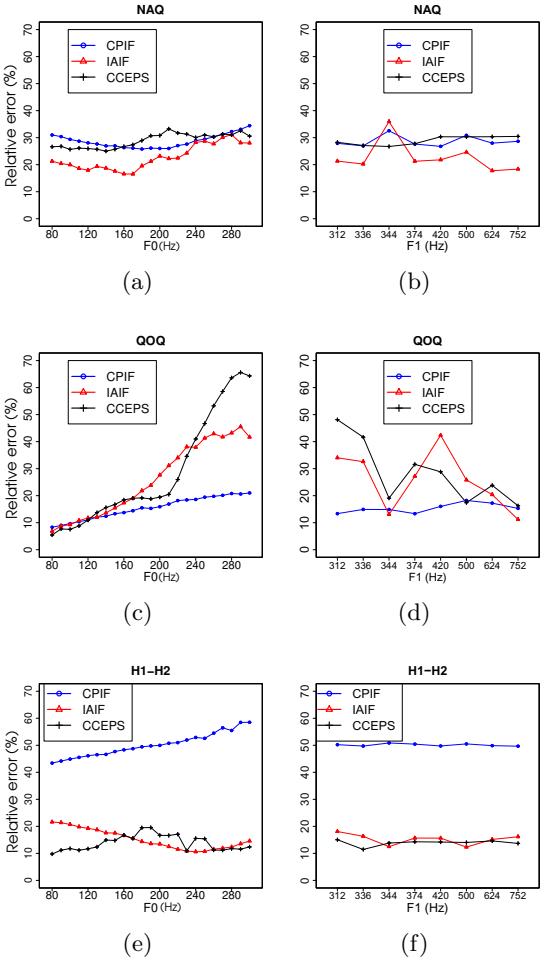

Fig. 1. Mean relative error score for NAQ (top row), QOQ (middle row) and H1-H2 (bottom row) as a function of f_0 (left column) and F1 (right column), for the three inverse filtering methods: CPIF (blue), IAIF (red) and CCEPS (black)

4.2 Voice Quality Differentiation

The results from the voice quality differentiation experiments are summarised Table 1. As expected, overall differentiation of voice quality is reduced when analysing the continuous speech considered here compared to vowel data analysed in [17] (note that our analysis of the same vowel data, not presented here, closely corroborates the trends seen in [17]). This is likely due to the difficulty in inverse filtering some parts of continuous speech (e.g., certain voiced consonants). However, similar trends to those in [17] are maintained with NAQ derived following IAIF giving the best performance for the direct measure parameters

($R^2 = 0.28$). Once more CCEPS is the most suitable decomposition method for applying H1-H2 ($R^2 = 0.26$). For QOQ, a serious degradation in performance is observed for all decomposition methods. CPIF is observed to be the least effective inverse filtering method for voice quality classification. Note that it displays considerably better performance on steady vowels (not presented here). In the synthetic data experiments the glottal closed phase is known *a priori*, whereas for the natural speech data used in these experiments the glottal closed phase has to be estimated with automatic algorithms which will inevitably display a certain degree of error.

Table 1. Explained variance (Pearson R^2) for each parameter and inverse filtering type combination. The glottal source parameter is treated as the dependent variable and voice quality label as the independent variable.

	NAQ	**H1-H2**	**QOQ**	*Strik-LF*		*DyProg-LF*		*Degott-LF*
				Rd	**OQ**	**Rd**	**OQ**	**Rd**
IAIF	0.28	0.22	0.20	0.21	0.24	0.39	0.34	0.28
CPIF	0.06	0.06	0.09					
CCEPS	0.10	0.26	0.05					

The performance for the model fitting methods is considerably better than has previously been reported [17]. Here the DyProg-LF method gave the best performing *Rd* values ($R^2 = 0.39$). This is also the case for OQ ($R^2 = 0.34$) and in fact both *Rd* and OQ derived from DyProg-LF provided considerably better voice quality differentiation than all the direct measure parameters. Another interesting observation is that the traditional OQ method, derived using model fitting methods, consistently outperformed QOQ.

5 Discussion

Perhaps the most striking finding in this study is the strong performance of LF model based parameters at differentiating breathy to tense voice. Whereas the standard time domain LF model fitting algorithm (Strik-LF, [18]) gave comparable performance to that in [17], more recent algorithms for deriving LF model parameters (DyProg-LF, [26] and Degott-LF, [20]) compared favourably with direct measure parameters. This is particularly the case for continuous speech, where direct measure parameters suffered a serious degradation in performance. Specifically for DyProg-LF, both the *Rd* and OQ parameters still provided strong differentiation of the voice quality in continuous speech. The reason for the apparent robustness of the DyProg-LF method to continuous speech can be explained by the suitability of dynamic programming for maintaining sensible parameter contours even in "*difficult*" speech regions.

Although differentiation of voice quality does not directly measure the accuracy of derived parameter values, strong performance does suggest that the particular method is characterising salient glottal features.

Evidence from the testing on synthetic speech signals indicates that certain glottal inverse filtering methods are more suited to certain parameters. For instance, closed-phase inverse filtering (CPIF) is shown to be particularly suitable for deriving NAQ and QOQ, both time domain parameters. These parameters derived following CPIF are also rather insensitive to changes in f_0 and vocal tract filter setting. However, for the frequency domain parameter, H1-H2, the CPIF output is clearly less suitable. This finding may corroborate those in [8] where CPIF is shown to produce higher levels of spectral distortion than the other inverse filtering methods. However, the findings for IAIF conflict with those in [8], as in the present results IAIF had a similar performance to the other methods in terms of relative error on NAQ and QOQ, whereas in [8] it was considerably worse. In fact IAIF displayed relatively stable performance across the experiments and is shown to be particularly useful in combination with NAQ for breathy-tense discrimination and accuracy on synthetic speech signals.

6 Conclusion

This study presents a general assessment of automatic glottal inverse filtering and glottal source parameterisation methods. To overcome the known difficulty of quantitative evaluation of glottal source analysis methods two different experiments are conducted which, in combination, provide a more comprehensive impression of the performance of the methods. Testing on synthetic signals revealed that different glottal inverse filtering methods are more suited to certain parameter estimation methods. The experiments on voice quality differentiation show that more recent LF model fitting methods are more suited to the continuous speech data than direct measures.

Acknowledgments. 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).

References

1. Degottex, G., Lanchantin, P., Roebel, A., Rodet, X.: Mixed source model and its adapted vocal-tract filter estimate for voice transformation and synthesis. Speech Communication 55(2), 278–294 (2013)
2. Degottex, G., Roebel, A., Rodet, X.: Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter. In: Proceedings of ICASSP, pp. 5128–5131 (2011)
3. Drugman, T., Dubuisson, T., Dutoit, T.: On the mutual information between source and filter contributions for voice pathology detection. In: Proceedings of Interspeech, pp. 1463–1466 (2009)
4. Lugger, M., Yang, B.: The relevance of voice quality features in speaker independent emotion recognition. In: Proceedings of ICASSP, pp. 17–20 (2007)
5. Walker, J., Murphy, P.: A review of glottal waveform analysis. In: Stylianou, Y., Faundez-Zanuy, M., Esposito, A. (eds.) COST 277. LNCS, vol. 4391, pp. 1–21. Springer, Heidelberg (2007)

6. Lin, Q.: Speech production theory and articulatory speech synthesis, Ph. D. Thesis (1990)
7. Székely, É., Kane, J., Scherer, S., Gobl, C., Carson-Berndsen, J.: Detecting a targeted voice style in an audiobook using voice quality features. In: Proceedings of ICASSP, pp. 4593–4596 (2012)
8. Drugman, T., Bozkurt, B., Dutoit, T.: A comparative study of glottal source estimation techniques. Computer Speech and Language 26, 20–34 (2011)
9. Alku, P.: Glottal inverse filtering analysis of human voice production - A review of estimation and parameterization methods of the glottal excitation and their applications. Sadhana 36(5), 623–650 (2011)
10. Yegnanarayana, B., Veldhius, R.: Extraction of vocal-tract system characteristics from speech signals. IEEE Transactions on Audio Speech and Language Processing 6(4), 313–327 (1998)
11. Alku, P., Bäckström, T., Vilkman, E.: Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Communication 11(2-3), 109–118 (1992)
12. Drugman, T., Bozkurt, B., Dutoit, T.: Complex cepstrum-based decomposition of speech for glottal source estimation. In: Proceedings of Interspeech, pp. 116–119 (2009)
13. Kane, J., Gobl, C.: Evaluation of glottal closure instant detection in a range of voice qualities. Speech Communication 55(2), 295–314 (2013)
14. Drugman, T., Thomas, M., Gudnason, J., Naylor, P., Dutoit, T.: Detection of Glottal Closure Instants From Speech Signals: A Quantitative Review. IEEE Transactions on Audio Speech and Language processing 20(3), 994–1006 (2012)
15. Alku, P., Bäckström, T., Vilkman, E.: Normalized amplitude quotient for parameterization of the glottal flow. Journal of the Acoustical Society of America 112(2), 701–710 (2002)
16. Hacki, T.: Klassifizierung von Glottisdysfunktionen mit Hilfe der Elektroglottographie. Folia Phoniatrica 41, 43–48 (1989)
17. Airas, M., Alku, P.: Comparison of multiple voice source parameters in different phonation types. In: Proceedings of Interspeech, pp. 1410–1413 (2007)
18. Strik, H.: Automatic parameterization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses. Journal of the Acoustical Society of America, 2659–2669 (1998)
19. Gobl, C., Ní Chasaide, A.: Amplitude-based source parameters for measuring voice quality. In: Proceedings of the ISCA Tutorial and Research Workshop VOQUAL 2003 on Voice Quality: Functions, Analysis and Synthesis, Geneva, pp. 151–156 (2003)
20. Degottex, G., Roebel, A., Rodet, X.: Phase minimization for glottal model estimation. IEEE Transactions on Audio Speech and Language processing 19(5), 1080–1090 (2011)
21. Hanson, H.M.: Glottal Characteristics of female speakers: Acoustic Correlates. Journal of the Acoustical Society of America 10(1), 466–481 (1997)
22. Fant, G., Liljencrants, J., Lin, Q.: A four-parameter model of glottal flow. In: STL-QPSR, Speech, Music, and Hearing, KTH, Stockholm, vol. 26(4), pp. 1–13 (1985)
23. Kane, J., Kane, M., Gobl, C.: A spectral LF model based approach to voice source parameterisation. In: Proceedings of Interspeech, pp. 2606–2609 (2010)
24. Laver, J.: The Phonetic Description of Voice Quality. Cambridge University Press (1980)
25. Gobl, C.: Modelling aspiration noise during phonation using the LF voice source model. In: Proceedings of Interspeech, pp. 965–968 (2006)
26. Kane, J., Gobl, C.: Automating manual user strategies for precise voice source analysis. Speech Communication 55(3), 397–414 (2013)