Automatic estimation of the lip radiation effect in glottal inverse filtering

Manu Airaksinen[1], *Tom Bäckström*[2], *Paavo Alku*[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Finland
[2]International Audio Laboratories Erlangen, Friedrich-Alexander University (FAU), Germany
manu.airaksinen@aalto.fi, tom.backstrom@audiolabs-erlangen.de

Abstract

In the analysis of speech production, glottal inverse filtering has proved to be an effective yet non-invasive method for obtaining information about the voice source. One of the main challenges of the existing methods is blind estimation of the contribution of the lip radiation, which must often be manually determined. To obtain a fully automatic system, we propose an automatic method for determining the lip radiation parameter. Our method is based on a physically-motivated quality criteria for the glottal flow, which can be approximated by minimization of the norm-1. Experiments show that the parameters obtained by the automatic method are mostly within the 95% confidence intervals of the mean values obtained by manual tuning by experts.
Index Terms: glottal inverse filtering, lip radiation

10.21437/Interspeech.2014-92

1. Introduction

Modeling of the voice source lies at the heart of several areas of speech technology; speech codecs employ a source-filter model to enable effective encoding of the sound waveform, and statistical speech synthesizers parameterize the sound waveform into feature streams using models of speech production. Furthermore, study of human voice production is important also in its own right, since information about the speech production system can be used in other science areas such as medicine (e.g. study of occupational voice or pathological speech), phonetics (e.g. prosody) and psychology (e.g. brain imaging of speech perception). Methods for analyzing the voice source are thus, in practice, widely applicable both in several core areas of speech technology as well as in disciplines outside engineering.

Probably the most daunting task of speech analysis is extracting information from the physiological apparatus generating the voice source, the vocal folds. Due to their hidden location in the larynx, behind cartilages, as well as due to their rapid oscillations, the vocal folds lend themselves poorly to direct observations. For example, high-speed video imaging of the vocal folds (e.g. [1]) requires inserting a sensor close to the vibrating vocal folds, which might hinder natural production of speech. Video imaging also requires plenty of light, whereby experimenters have to deal with the practical problems of a heat-source in the vicinity of sensitive tissues.

Non-invasive analysis methods of voice production are therefore much preferred and one of the most widely used of such methods is known as glottal inverse filtering (GIF). It is an indirect method where airflow through the glottis is estimated from an acoustic pressure signal captured by a free-field microphone outside the lips. (In principle, the GIF analysis can also be conducted using the oral flow recorded with a pneumotachograph mask [2]). Even though many different GIF methods have been developed in the past decades (see [3], for a review), they are almost exclusively based on source-filter theory, according to which the production of a voiced sound is modeled as a cascade of three processes: the glottal flow, the vocal tract and the lip radiation effect [4]. GIF estimates the first of these three parts, the glottal flow, by using a two-step procedure. Firstly, the acoustic effect of the vocal tract and lip-radiation is estimated blindly, and, secondly, their contribution is cancelled from the speech signal by inverse filtering. This approach has several benefits, for example, application of inverse filtering requires only a microphone and computer, whereby hardware costs are low. In addition, speech can, in principle, be recorded in any environment, whereby the measurement setup does not significantly interfere with natural communication.

One of the main drawbacks in current GIF algorithms is that they typically involve parameters which require manual tuning. This introduces two problems. Firstly, manual tuning is not possible when GIF is applied in modern data-driven technologies, such as statistical speech synthesis, where large amounts of trainig data need to be processed with GIF (e.g. 1 hour was used in [5]). Secondly, even in applications where relatively small amount of data is to be analyzed, manual tuning of parameters introduces a subjective component, whose effect on the validity of results is difficult to quantify or remove. Only fully automatic GIF methods can provide objective results in the analysis of speech production.

While most previous GIF studies have focused either on the parameterization of the glottal flow or on the computation of the vocal tract, the third part of the source-filter model, the lip radiation effect, has remained less explored. The goal of the current investigation is to propose a new method for automatic computation of the lip radiation effect in an adaptive manner as a part of a GIF-based estimation of the voice source. Our approach is based on formulating a quality criterion for the glottal flow estimate based on known physical properties. More specifically, the lip radiation effect is adaptively computed by searching for a lip radiation parameter yielding a positive time-domain glottal flow pulse with the smallest norm-1.

2. Lip radiation effect modeling

The production of speech according to the source-filter model can be expressed in the digital domain as follows:

$$S(z) = G(z)V(z)R(z), \tag{1}$$

where $S(z)$ corresponds to speech, $G(z)$ is the glottal flow, $V(z)$ is the vocal tract transfer function, and $R(z)$ is the lip radiation effect. Once the vocal tract transfer function has been computed (e.g using methods such as [6] or [7]), the glottal flow is obtained as $G(z) = \frac{S(z)}{V(z)R(z)}$. It is worth noting here that the denominator corresponds to the product of $V(z)$ and $R(z)$.

In the acoustic theory of speech production, the lip radiation effect corresponds to transforming the volume velocity wave-

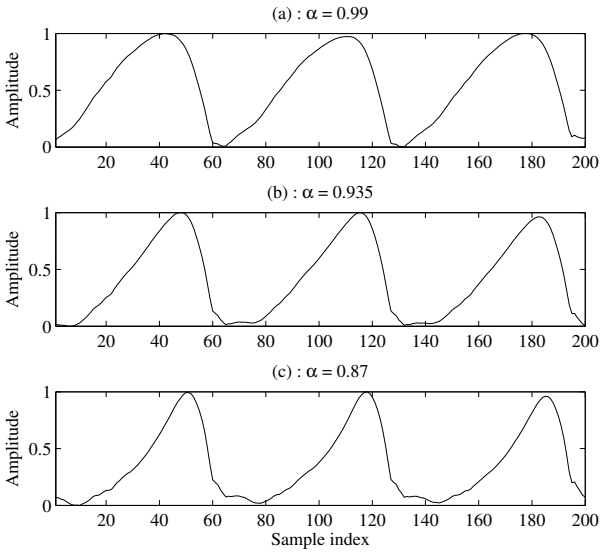

Figure 1: *Glottal flow waveforms estimated by inverse filtering. Parameter α of Eq. 3 is adjusted to be (a) too high, (b) correct, and (c) too low.*

form at lips into an acoustic pressure waveform some distance away from the lips [8] that is caused by the ending boundary conditions of the vocal tract resonator tube [4]. The acoustic model of lip radiation presented in [4] assumes radiation from an infinite plane baffle. This results in radiation impedance (or "radiation load") $Z_L(\Omega)$, converting the volume velocity at the lips $U(\Omega)$ to a pressure $P(\Omega)$ by $P(\Omega) = Z_L(\Omega) \cdot U(\Omega)$, which can be presented as

$$Z_L(\Omega) = \frac{j\Omega L_r R_r}{R_r + j\Omega L_r}, \tag{2}$$

where Ω is the angular frequency, R_r the radiation resistance, and L_r the radiation inductance. It can be seen that $Z_L(\Omega) \approx j\Omega L_r$ for $\Omega L_r \ll R_r$. Hence, the lip radiation effect can be simplified for low frequencies by a first order time-derivative. In digital signal processing, this implies using the following first-order FIR filter as the discrete filter model for the lip radiation effect:

$$R(z) = 1 - \alpha z^{-1}, \tag{3}$$

where $\alpha \lesssim 1$. When glottal flow is computed in GIF algorithms, the lip radiation effect needs to be *canceled* by using the inverse filter of Eq. 3. Since $1/R(z)$ is an IIR filter, using an ideal integrator (i.e. $\alpha = 1.0$) results in a marginally stable filter. Therefore, the root of Eq. 3 is commonly shifted slightly towards the inside of the unit circle in order to guarantee the stability of the $1/R(z)$. Authors of [9] also argue that with a microphone approximately 30 cm away from the lips, the analysis has not totally left the acoustic near field, therefore not fully justifying a 6dB/octave pre-emphasis. Based on these properties, the value for coefficient α is most commonly assumed to be fixed in the range $\alpha \in [0.98, 0.999]$ [9, 10].

The first-order digital model of the lip radiation effect given in Eq. 3 is straightforward and widely used in different GIF methods. Using this kind of a simple filter with a fixed α-value, however, causes distortion that has, to the best of our knowledge, not been discussed before in any GIF study. In defining the vocal tract model $V(z)$ with an all-pole modeling method, such as linear prediction [8] or discrete all-pole modeling [11], the focus is on finding a good spectral model for the formants. However, the lowest frequencies below ca. 200 Hz play a lesser role in defining $V(z)$ for most GIF methods. This, in turn, might result in excessive boosting (or attenuation) of the low frequencies in the spectrum of $V(z)$. If a lip radiation model with a fixed α-coefficient is used, the inconsistent amplitude behaviour of $V(z)$ at low frequencies results in low-frequency distortion in the estimated glottal flow. This distortion, as indicated by an example shown in Fig. 1, affects particularly the closed phase of the glottal flow pulse: using a too large α-coefficient (Fig. 1(a)) resulted in this example in a glottal pulse with a very short, yet clear closed phase. Using a too small α-coefficient (Fig. 1(c)), however, resulted in a pulse which shows a "knee" at the end of the closing phase, suggesting the occurrence of glottal closure, but the airflow still continues to decrease after this instant. Traditionally this kind of ambiguity has warranted manual tuning of α in a manner similar to that which is used for the vocal tract parameters [12]: the lip radiation parameter is adjusted by searching for a setting that yields a flow estimate with the maximally flat horizontal closed phase. Objective quality measures for the tuning of GIF parameters have previously been developed e.g. in [10, 13], and a method for automated voice source analysis based on manual strategies is presented in [14], but their focus has not been in the effect of α. The next section presents the proposed method to automatically adjust α in order to obtain glottal flows, such as the one shown in Fig. 1(b), with plausible closed phase behavior.

3. Proposed method

Our approach in optimizing α begins by assuming the following properties for the ideal glottal flow g (a vector of length N, whose time-variable is omitted from this analysis): 1) It is always positive ($g \geq 0$), but 2) aims to remain as small in amplitude as possible ($g_{\text{opt}} = \min(\|g\|_1)$). Issue 2) is motivated by the human's tendency to minimize the use of air in production of speech. We chose to use norm-1 since it is well-known that minimization of the norm-1 tends to provide sparse results [15], that is, a large part of the signal components are zero. In the current context, this property of the norm-1 aligns nicely with our objective of obtaining a glottal flow estimate with a long closed phase (i.e. zero flow). Concurrently, this approach also minimizes the overall excitation magnitude. This study also assumes that the effective driving excitation signal, $E(z) = G(z)R(z) = S(z)/V_{\text{est}}(z)$, has been computed from speech with GIF. The estimated vocal tract transfer function is denoted by $V_{\text{est}}(z)$.

When the lip radiation model presented in Eq. 3 is coupled with these assumptions, our optimization model becomes:

$$g_{\text{opt}} = \min_{\alpha} \left(\left\| Z^{-1} \left\{ \frac{E(z)}{1 - \alpha z^{-1}} \right\} \right\|_1 \right), \tag{4}$$

where Z^{-1} denotes the inverse Z-transform. From Eq. 4, the following observation can be made: As $\frac{1}{1-\alpha z^{-1}}$ corresponds to a *leaky integrator*, the area of $g(\alpha) = Z^{-1}\{\frac{E(z)}{1-\alpha z^{-1}}\}$ becomes smaller as α decreases. Using a too low value for α distorts the closed phase of the glottal flow as depicted in Fig. 1(c). This phenomenon is further described in Fig. 2 which depicts $\|g\|_1$ as a function of α when effective driving function was estimated from a vowel sound. As a general trend, it can be seen that the norm-1 of g increases as α rises. However, there is a distinct value of α, indicated by a circle in Fig. 2, where the slope of the

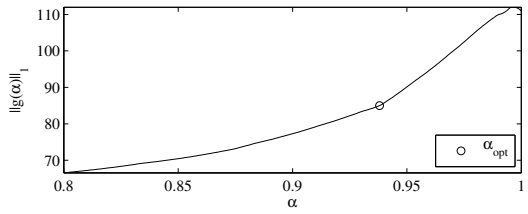

Figure 2: *The norm-1 (area) of g(α) for* $\alpha \in [0.8, 1]$. *The point with the abrupt slope increase is denoted with 'o'.*

curve changes. Interestingly, this value of α coincides with the manually determined "ideal" value of α presented in Fig. 1(b). The sudden increase in the slope of the curve can be attributed to the tilting of the closed phase that is caused by a too high value of α. In that case, the increase in the norm-1 of g as a function of α is caused by two factors: First, the reduction in the leakiness of the integrator, and second, the tilting of the closed phase. The first is a property that we want to maximize, and the second is the property that we wanted to minimize in the original formulation of the problem. Therefore α_{opt} can be detected as the smallest value of α for which the slope $\frac{dA(\alpha)}{d\alpha}$ of $A(\alpha) = \|g(\alpha)\|_1$ exceeds a threshold value of κ that denotes the additive slope increase.

To generalize the proposed method for the value of κ, the area function should be normalized with respect to the frame duration N. As $g(\alpha)$ is normalized between [0, 1], the normalized area function becomes the mean of $g(\alpha)$:

$$A_{\text{norm}} = \frac{1}{N}\|g(\alpha)\|_1 = \text{mean}(g(\alpha)) \tag{5}$$

A simple brute force algorithm for the computation of α_{opt} can be presented as:

```
begin
  for α := α_min to 1 step Δ do
    if (A_norm(α) − A_opt)/Δ < κ
      then A_opt = A_norm(α);
           α_opt = α;
      else return α_opt;
    fi
  od
  return α_opt;
end
```

where $\alpha_{\text{min}} = 0.8$ and $\Delta = 0.001$ for the remainder of this article. Initial experiments on real speech indicated that a fixed value of $\kappa \in [1.1, 1.4]$ provided the best results. Too low κ values resulted in too early detection of the slope boosting, whereas too high values resulted in too late detection and thus too high α values. The next section presents an experiment conducted on real speech where the best value for κ was determined based on the results of a subjective expert test.

4. Experiments

The objective evaluation of GIF methods is problematic because it is impossible to measure the real glottal flow waveform from natural speech. Thus the most justified way to evaluate the performance of the proposed method for real speech is to ask experienced experimenters to conduct GIF analysis by allowing them to manually tune α and to compare these results to the values computed automatically by the proposed method.

A subjective expert test was arranged with five experimenters, all of whom had an average experience of 9.80 (median of 6) years in voice source analysis. These experts were asked to adjust the lip radiation parameter using a range of $\alpha \in [0.8, 1]$ for a test set of glottal flow derivative waveforms that were computed beforehand from real speech frames. The expert-based results were then compared to the α values given by the proposed method with varying values of κ, and the value κ_{opt} that gave the best least-squares fit to the expert data was selected as the reference value.

The speech test set was composed of natural sustained Finnish vowels ([a], [e], [i], [o], [u], [y], [ae], and [oe]), uttered by two male and two female speakers. Speech signals were produced using two phonation types (modal and pressed) which are supposed to show distinct closed phases [16]. In total the test set consisted of 64 speech signals. A sampling rate of 8 kHz was used, and inverse filtering was computed with the QCP method [17] by assigning its parameters as DQ = 0.5, PQ = 0.01, and $N_{\text{ramp}} = 7$.

Finally, the proposed method is demonstrated with two real-speech samples where differentiated electroglottography signal is used as a control reference.

5. Results

The effect of the parameter κ in the proposed method is presented in Fig. 3. The mean squared error of the estimated α values to the expert-based mean value is minimized in the range between $\kappa \in [1.1, 1.4]$. This range was also found to be good in the initial experiments. The best least squares fit to the expert data was achieved with $\kappa_{\text{opt}} = 1.37$.

The automatic estimation results, computed with κ_{opt}, are compared to the expert-based results in Fig. 4. The speech samples in Fig. 4 are sorted in an ascending order with respect to the standard deviation of expert answers. It can be seen that the proposed method is able to produce very similar results compared to the manual tuning by the experts. 50 out of 64 ($\approx$ 78%) samples were automatically estimated to be inside the 95% confidence interval of the mean of the expert answers. The averaged absolute value of the error for all frames was $\frac{1}{N_{\text{all}}}\sum_{\text{all}}|\alpha_{\text{exp}} - \alpha_{\text{auto}}| = 0.0116$, and for the samples outside of the confidence intervals $\frac{1}{N_{\text{out}}}\sum_{\text{out}}|\alpha_{\text{exp}} - \alpha_{\text{auto}}| = 0.0151$.

The statistical values computed from the test results are presented in Table 1. It can be seen from these data that the proposed method has non-biased ($\mu_{E_\alpha} \approx 0$) error values to the expert-based mean with a standard deviation that is similar in scale to the average standard deviation in the expert answers ($\sigma_{\text{Expert,avg}} \approx \sigma_{E_\alpha}$). This suggests that the method oper-

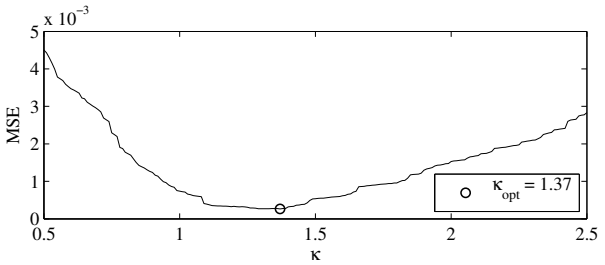

Figure 3: *The mean squared error of the estimated* α *values to the mean value of experts as a function of* κ. *Best fit obtained with* $\kappa_{opt} = 1.37$.

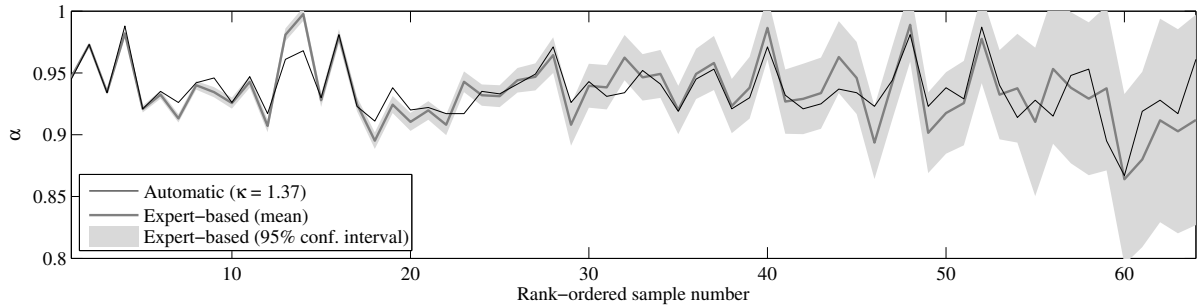

Figure 4: *The automatically estimated α values (black line), the expert-based mean values (gray line), and the expert-based 95% confidence intervals (light gray area). Samples sorted by the standard deviation of expert-based results in ascending order.*

Table 1: *Rows 1-2: Statistics for the expert-based and automatically estimated α-values. The standard deviation for expert-based results was calculated from each test sample, and then averaged over all test samples. Rows 3-5: Statistics for the error between automatic estimation and the expert-based mean values. For the proposed method, $E_{\alpha,\,all}$ denotes the error for all samples, and $E_{\alpha,\,out}$ the error for samples outside the 95% confidence intervals. $E_{\alpha=0.935,\,fixed}$ denotes the error for optimal fixed-valued lip radiation modeling.*

	mean (μ)	std (σ)
Expert-based α (avg)	0.935	0.020
Automatic α	0.937	0.022
$E_{\alpha,\,\mathrm{all}}$	-0.001	0.016
$E_{\alpha,\,\mathrm{out}}$	-0.000	0.018
$E_{\alpha=0.935,\,\mathrm{fixed}}$	0.000	0.026

ates within a reasonable scope of accuracy. Compared to the optimal fixed-variable lip radiation modeling (for the test set $\alpha = 0.935$), the standard deviation of the error for the proposed method is 38% smaller. Finally, the error values to the expert-based means outside the 95% confidence intervals can be seen to be within the same magnitude as for the general case, which suggests that the proposed method provides robust values without extreme outliers.

Two representative examples are depicted in Fig. 5, where estimated glottal flows obtained with the proposed method, Fig. 5($b_{1\text{-}2}$), are shown together with pulse forms, Fig. 5($a_{1\text{-}2}$), computed with a fixed $\alpha = 0.99$. The differentiated electroglottography (EGG) signals are shown in Fig 5($c_{1\text{-}2}$). It can be seen that the pulse forms computed with the proposed lip radiation model show distinct closed phases that correspond well with the glottal closure and opening instants indicated by EGG.

6. Conclusions

This study presents a method for the automatic estimation of the lip radiation coefficient α in glottal inverse filtering. The method is based on computing the norm-1 of a glottal flow waveform as a function of α and by searching for a distinct point where the slope of the curve exceeds a fixed threshold value of κ for the first time. This point corresponds to the value of α that is as close to 1.0 as possible, the ideal value assumed in the acoustical theory of lip radiation, yet producing a flat closed phase for signals inverse filtered from natural speech.

The method was evaluated with a subjective expert-based test where five test subjects manually tuned the ideal α coefficients for glottal flow waveforms estimated from natural speech. The expert results were then compared to the values estimated with the proposed method. The test indicated that κ should be within the range of $\kappa \in [1.1, 1.4]$, and the best least-squares fit was achieved with $\kappa_{\mathrm{opt}} = 1.37$. With the best fit, 78% of the automatically estimated values were within the 95% confidence intervals of the expert-based mean values, and the results did not show any extreme outliers.

The obtained results indicate that the proposed method is suitable for the automatic estimation of the lip radiation parameter in GIF, clearly surpassing the traditional fixed-valued modeling of α. The method can be used with any existing GIF method that operates by estimating the vocal tract transfer function from the speech signal.

7. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287678 and from the Academy of Finland (project no. 256961).

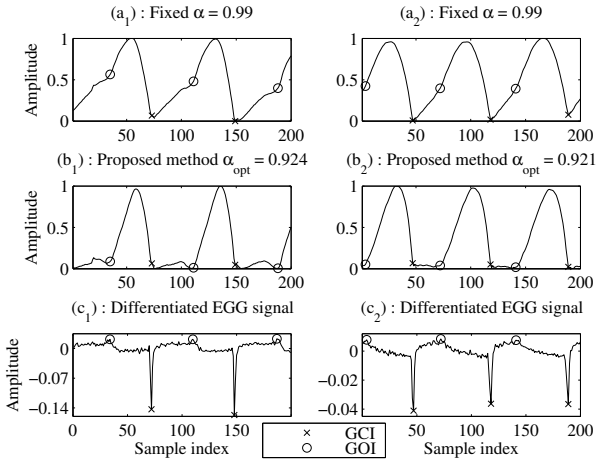

Figure 5: *Examples of glottal flows estimated with (a) fixed α and (b) the proposed method. The differentiated electroglottography (EGG) signals are shown in (c). The glottal closure (GCI) and opening (GOI) instants detected from the EGG signals are marked with 'x's and 'o's, respectively.*

8. References

[1] G. Chen, J. Kreiman, and A. Alwan, "The glottaltopogram: A method of analyzing high-speed images of the vocal folds," *Computer Speech & Language*, vol. 28, no. 5, pp. 1156 – 1169, 2014.

[2] E. Holmberg, R. E. Hillman, and J. S. Perkell, "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice," *Journal of the Acoustical Society of America*, vol. 84, no. 2, pp. 511–529, 1988.

[3] P. Alku, "Glottal inverse filtering analysis of human voice production — a review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, pp. 623–650, 2011.

[4] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, ser. Prentice-Hall signal processing series. Prentice-Hall, 1978.

[5] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 153–165, 2011.

[6] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2–3, pp. 109–118, 1992.

[7] D. E. Veeneman and S. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 369–377, 1985.

[8] J. D. Markel and A. H. Gray Jr., *Linear Prediction of Speech*. Springer-Verlag, Berlin, 1976.

[9] J. J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, 2nd ed. Wiley-IEEE Press, 1999.

[10] T. Bäckström, M. Airas, L. Lehto, and P. Alku, "Objective quality measures for glottal inverse filtering of speech pressure signals," *Acoustics, Speech and Signal Processing (ICASSP), 2005 IEEE International Conference on*, vol. 1, pp. 897–900, 2005.

[11] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *Signal Processing, IEEE Transactions on*, vol. 39, no. 2, pp. 411–423, 1991.

[12] M. Rothenberg, "A new inverse–filtering technique for deriving the glottal air flow waveform during voicing," *Journal of the Acoustical Society of America*, vol. 53, no. 6, pp. 1632–1645, 1973.

[13] E. Moore and J. Torres, "A performance assessment of objective measures for evaluating the quality of glottal waveform estimates," *Speech Communication*, vol. 50, no. 1, pp. 56–66, 2008.

[14] J. Kane and C. Gobl, "Automating manual user strategies for precise voice source analysis," *Speech Communication*, vol. 55, no. 3, pp. 397–414, 2013.

[15] N. Hurley and S. Rickard, "Comparing measures of sparsity," *Information Theory, IEEE Transactions on*, vol. 55, no. 10, pp. 4723–4741, 2009.

[16] P. Alku and E. Vilkman, "A comparison of glottal voice source quantification parameters in breathy, normal, and pressed phonation of female and male speakers," *Folia Phoniatrica et Logopaedica*, vol. 48, no. 5, pp. 240–254, 1996.

[17] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 3, pp. 596–607, 2014.