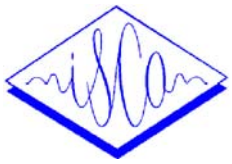

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Evaluation of an Inverse Filtering Technique Using Physical Modeling of Voice Production

Paavo Alku[1], Brad Story[2], Matti Airas[1]

1: Helsinki University of Technology, Espoo, Finland
2: University of Arizona, Tucson, AZ, USA
paavo.alku@hut.fi

Abstract

Glottal flows and sound pressure waveforms of four different fundamental frequencies were generated using a computational model of vocal fold vibration and acoustic wave propagation in order to evaluate the performance of an inverse filtering method. Four time-based parameters of the glottal flow were used in order to assess the accuracy of the inverse filtering technique. The results show that for most of the cases analyzed the relative error was less than 5 % when the time-based parameters extracted from the estimated glottal flows were compared to those obtained from the original flow waveforms produced by the physical model of the vocal fold vibration.

1. Introduction

Inverse filtering (IF) is a method to estimate the source of voiced speech, the glottal volume velocity waveform. The idea behind IF is to first form a model for the vocal tract transfer function. The effect of the vocal tract is then canceled from the produced speech waveform by filtering this through the inverse of the model. The waveform, from which the canceling of the vocal tract contribution is done, can be either the oral flow recorded in the mouth with a flow mask (e.g. [1]) or the pressure waveform captured by a microphone in free field outside the mouth (e.g. [2]). Older methods typically used manually adjustable analog circuits in implementation of the inverse model of the vocal tract. Currently used methods are based on digital modeling of the vocal tract, which is usually implemented either completely automatically or semi-automatically [3-7].

Evaluation of the performance of an inverse filtering method is problematic. When inverse filtering is used in estimation of the glottal flow of natural speech, it is actually *never* possible to assess in detail, how closely the obtained waveform corresponds to the true glottal flow generated by the vibrating vocal folds. There are, however, certain methods that have been used in order to estimate the performance of inverse filtering. It is possible, for example, to compare the estimated glottal flow given by an IF method to other information signals extracted from the fluctuating vocal folds by exploiting such techniques as, for example, electroglottography [8], high-speed filming [9], video-kymography [10] or high-speed digital videoscopy [11]. These methods undoubtedly provide valuable information from the vibration of the vocal folds. They are, however, problematic in evaluation of inverse filtering, because of the unknown mapping between the glottal volume velocity and the corresponding information signal provided by these voice production analysis techniques (e.g. the time-varying impedance signal in the case of electroglottography and the glottal area function in the case of high-speed imaging techniques). On the other hand, it is also possible to assess the accuracy of inverse filtering by using synthetic speech that has been created using a known, artificial waveform of the glottal excitation. This kind of evaluation, however, is not truly objective, because speech synthesis and inverse filtering analysis are typically based on similar models of the human voice production apparatus (e.g. the source filter model [12]).

In the current study, we use a different strategy in order to analyze, how accurately an inverse filtering method can estimate the glottal flow. The idea is to use *physical modeling* of the vocal folds and the vocal tract in order to synthesize vowels with a known, realistic glottal flow waveform. By using the pressure signals given by the physical models as an input to an inverse filtering method, it is possible to analyze, how closely the obtained estimate of the voice source matches the original glottal flow.

The paper first describes in section 2 the straightforward inverse filtering method used. The physical modeling applied in synthesizing the test vowels is explained then in section 3. The results obtained are described in section 4 and the paper is finished with short conclusions in section 5.

2. Inverse filtering

The inverse filtering method used is based on our previous experiments in developing automatic methods to estimate the glottal flow from the speech pressure waveforms with the Iterative Adaptive Inverse Filtering (IAIF) method [5]. The current method, the flow diagram of which is shown in Fig. 1, is a slightly modified version from our previous ones. Parametric spectral models that are used in various blocks of the flow diagram are computed with the Discrete All-pole Modeling (DAP) method [13] instead of the conventional linear predictive analysis. This makes it possible to obtain estimates of the formant frequencies that are less biased by the harmonic structure of the speech spectrum. The description of the various blocks shown in the flow diagram is given below.

Firstly (block no. 1), the speech signal is high-pass filtered in order to remove any distorting low frequency fluctuations captured by the microphone during the recordings. The high-pass filter must be a linear phase FIR-filter, and its cut-off frequency should be adjusted to be smaller than the fundamental frequency (F0) of the speech sound analyzed. In the current study, the cut-off frequency was set to 60 Hz. Secondly, (block no. 2) a first order all-pole filter is computed from the high-pass filtered speech signal in order to get a preliminary estimate for the combined effects of the glottal flow and the lip-radiation effect on the speech spectrum. This stage yields a first-order all-zero filter, the transfer function of

which is denoted by $H_{g1}(z)$ in Fig. 1. Next (block no. 3) the estimated effects of the glottal flow and lip-radiation are canceled from speech by filtering it with $H_{g1}(z)$. The output is analyzed using a *p*th-order DAP analysis (block no. 4) to obtain a model, denoted by $H_{vt1}(z)$, for the vocal tract filtering. Next (block no. 5) the effect of the vocal tract is canceled from speech by inverse filtering it through the inverse of the obtained *p*th-order model. A first estimate for the glottal flow is obtained (block no. 6) by canceling the effect of the lip-radiation by integrating the output of block no. 5. The IAIF-method next computes (block no. 7) a new estimate, denoted by $H_{g2}(z)$, for the contribution of the glottal flow on the speech spectrum by computing DAP-analysis of order *g* to the obtained first estimate of the glottal excitation. By first canceling the effect of the estimated glottal contribution (block no. 8) and the lip-radiation effect (block no. 9) a new model for the vocal tract filtering is obtained by a *r*th-order DAP-analysis (block no. 10). The final result is obtained by canceling the effect of the new vocal tract model (block no. 11) and the lip-radiation effect (block no. 12).

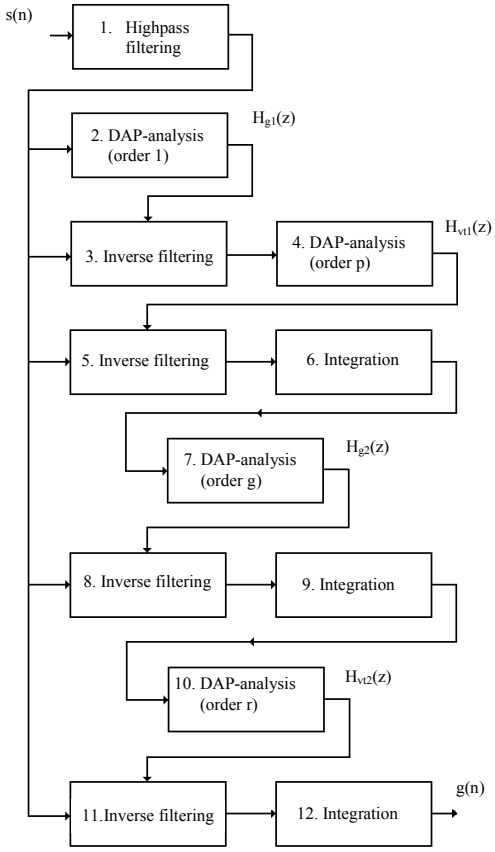

Figure 1: Block diagram of the IAIF method

The IAIF method described in above has limitations. It is based on straightforward linear modeling of speech production without taking into account, for example, the interaction between the glottal source and the vocal tract. Moreover, the digital model of the vocal tract is a pure all-pole filter, which is not accurate for nasals. Despite these inherent limitations, the proposed technique provides a promising method to estimate the glottal flow especially given the fact that the method can be implemented (if desired) in a completely automatic manner with a reasonable computational cost.

3. Physical modeling

The glottal flow and sound pressure waveforms used for testing the inverse filtering technique were generated with a computational model of vocal fold vibration and acoustic wave propagation. Specifically, the body-cover structure of the vocal folds was simulated with three masses coupled to each other through stiffness and damping elements [14]. A schematic diagram of a single vocal fold can be seen on the left side of Fig. 2. For all of the simulations used in this study, bilateral symmetry was assumed such that identical vibrations occur within the right and left vocal folds. The mechanical constants (i.e. mass, stiffness, and damping) were set to produce a typical male fundamental frequency (see [15] for details on setting these constants based on physiologic parameters). The vocal fold model was coupled to the pressures in the trachea and the vocal tract through aerodynamic and acoustic considerations [16], thus allowing for self-sustained oscillation. Acoustic wave propagation in both the trachea and vocal tract was simulated in time-synchrony with the vocal fold model. This was carried out with a wave-reflection approach (digital waveguide) [e.g. 17] that included energy losses due yielding walls, viscosity, and radiation at the lips [18]. The resulting glottal flow is determined by the interaction of the glottal area with the time-varying pressures present just below and just above the glottis.

A conceptualization of the model is given in Fig. 2, where the vocal fold model is shown located between the trachea and the vocal tract. An example glottal flow waveform is indicated at the right side of the figure. Note that the ripples in the waveform are due to the formant oscillations in the vocal tract. The sound pressure waveform radiated at the lips is also shown and can be considered analogous to a microphone signal recorded for a speaker.

In summary, the model is a simplified but physically motivated representation of a speaker. It generates both the signal on which inverse filtering is typically performed (microphone signal) and the signal that it seeks to determine (glottal flow), thus providing an idealized test case for an inverse filtering algorithm.

4. Results

Four /a/ vowels produced by the physical modeling were inverse filtered with the IAIF method by using the following parameters (see Fig. 1): p = r = 8, g = 4. The sampling frequency was converted from its original value (44.1 kHz) used in the generation of sounds to 8 kHz. The length of the analysis window was 100 ms and the number of iterations in the DAP computations was 10. The lip radiation effect (blocks no 6, 9 and 12 in Fig 1) was canceled by a first order all-pole filter with its pole at z = 0.997.

The comparison between the original flows given by physical modeling and those estimated from the pressure signals with inverse filtering was done by parameterizing the corresponding flow signals with time-based quotients. The

following parameters were used [19]: (1) Open Quotient (OQ), which is the length of the open phase of the glottal flow divided by the length of the fundamental period, (2) Speed Quotient (SQ), which is the ratio between the lengths of the glottal opening and closing phases, and (3) Closing Quotient (ClQ), which is the length of the glottal closing phase divided by the fundamental period. In addition to these conventional parameters, we also used the (4) Normalized Amplitude Quotient (NAQ), which measures time-domain characteristics of the glottal closing phase from two amplitude-domain quantities (see [20] for details).

The four parameters were extracted from individual glottal cycles of the flow pulseforms over 10 consecutive fundamental periods and the final parameter value was computed as a mean of the obtained data. Extraction of the critical time-instants required in the computation of the parameters was done automatically as follows. Firstly, the time instant of the negative peak of the flow derivative was defined by searching for the minimum value of the derivative in each glottal cycle. Glottal closure was then defined as the time-instant after the negative peak, when the derivative returned to the zero level. The instant of the maximal flow was next determined by searching for the maximum value of the flow waveform in a window that spanned one fundamental period and ended at the instant of glottal closure. Glottal opening was then determined by searching in backwards the waveform of the flow derivative by starting from the instant of the maximal flow: glottal opening was defined as the instant, when the derivative changed its sign from positive to negative (occasional sign changes were ignored).

An example of the waveforms obtained is shown in Fig. 3, which depicts the original glottal flow computed by physical modeling (upper graph) and its counterpart estimated by inverse filtering (lower graph). The numerical results describing the similarity between the two flow signals are given in Table 1 for each of the four parameters. The data given in the table is expressed as a ratio of the corresponding parameter value obtained from inverse filtering to that extracted from the original flows given by physical modeling. The results show that the maximum error in the parameter values extracted from estimated flows was 8 % when compared to the original flows. As expected, the deviation of the parameter values was larger for voices with higher fundamental frequency. Since the waveform of the glottal flow in the excitation signals given by physical modeling was sharp both at the glottal opening and closure (see Fig. 3, upper graph), these instants could be extracted reliably also from the flow signals given by inverse filtering. Hence, the value of OQ showed little error. However, there was some fluctuation in the flow waveforms near the instant of the maximal flow, which caused some error in the extraction of the instant of maximal flow. Hence, the values of SQ and ClQ showed somewhat larger deviations between the two flow waveforms.

The error caused by inverse filtering was assessed by comparing it to the ratios of the respective parameters extracted from real speech in different phonation types. The speech data used in the comparison consisted of three phonation types (breathy, normal, pressed) acquired in previous studies [20, 21]. The data of these previous experiments show that the ratio of the parameter (computed as a mean of two parameter ratios: breathy vs. normal and normal vs. pressed) equaled 1.17, 0.76, 1.45 and 1.80 for OQ, SQ, ClQ and NAQ, respectively. These data suggest that the error introduced by inverse filtering is clearly smaller than the difference in the parameter value in the three phonation types. Moreover, it can be noticed in Table 3 that inverse filtering caused deviations of NAQ that were approximately of the same order as those of ClQ, even though differences of NAQ values between phonation types were clearly larger than corresponding differences in ClQ [20].

Table 1: Ratio of the glottal flow parameters (see section 4) computed from the inverse filtered flows (denoted by subscript "if") to parameters extracted from the flows given by physical modeling (denoted by subscript "pm").

$F0(Hz)$	$\frac{OQ_{if}}{OQ_{pm}}$	$\frac{SQ_{if}}{SQ_{pm}}$	$\frac{ClQ_{if}}{ClQ_{pm}}$	$\frac{NAQ_{if}}{NAQ_{pm}}$
100	1.01	1.06	0.95	1.00
115	0.97	1.03	0.95	1.01
130	1.00	1.05	0.94	1.08
145	1.01	0.94	1.08	1.08

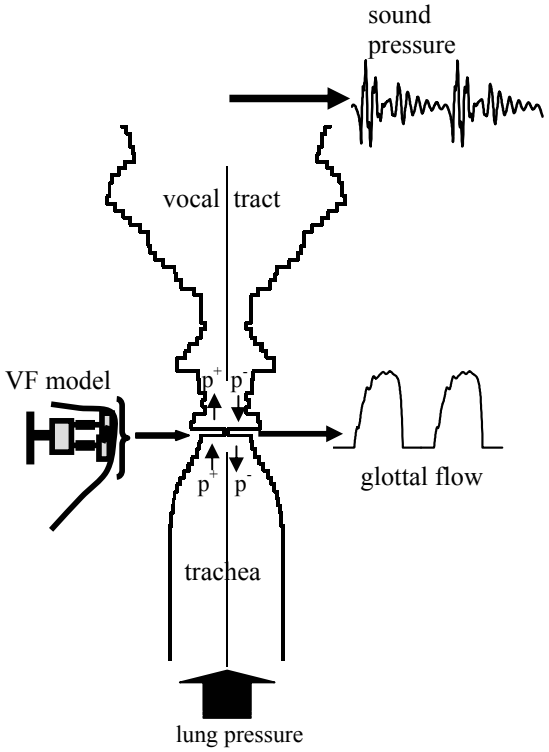

Figure 2: Conceptualization of the physical model

5. Conclusions

Evaluation of inverse filtering methods is problematic because direct measurements of the glottal flow are difficult, if not impossible. In addition, using synthetic speech as test material does not make a fully objective evaluation possible, because voice synthesis and inverse filtering are typically based on the same voice production models.

The present study aimed to avoid these fundamental limitations by using vowels produced with physical modeling in evaluation of inverse filtering. The glottal flows estimated by an inverse filtering technique were compared to the original flows computed by physical modeling by parameterizing the time-based features of the waveforms using four different quotients. The results were encouraging in showing that the differences in the parameters extracted from the original and estimated flows were small.

The experiments of the present study were based on the vowel /a/ alone using only four different values of F0. In order to better understand the limitations of inverse filtering, the characteristics of the test material should be expanded. In particular, the range of F0 values used in the evaluation should be expanded to cover the pitch range of female speech. Moreover, the variation of the vowels should be larger to assess, how sensitive the inverse filtering method analyzed is for the position of the lowest formant. Therefore, our next goal is to expand the test material provided by the physical modeling approach to include different vowels (such as /i/ with a low first formant) and higher F0 values.

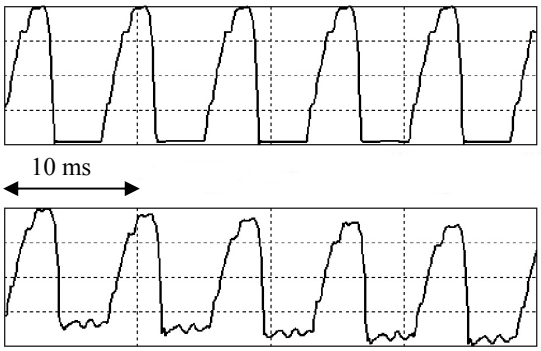

Figure 3: Examples of flow waveforms (F0 = 130 Hz). Upper graph: the original glottal flow computed by physical modeling, lower graph: the flow waveform estimated from the speech pressure waveform with inverse filtering (y-axis arbitrary).

6. Acknowledgements

This study was supported by the Academy of Finland (project no. 200859) and the National Institutes of Health (grant no. R01 DC04789).

7. References

[1] Rothenberg, M., "A new inverse-filtering technique for deriving the glottal airflow waveform during voicing", *J. Acoust. Soc. Amer., 53(6):1632-1645, 1973*.

[2] Wong, D.Y., Markel, J.D., and Gray, A.H., Jr., "Least squares glottal inverse filtering from the acoustic speech waveform", *IEEE Trans. on Acoust., Speech, and Signal Proc., 27(4):350-355, 1979.*

[3] Veeneman, D.E. and BeMent, S., "Automatic glottal inverse filtering from speech and electroglottographic signals", *IEEE Trans. on Acoust., Speech, and Signal Proc.*, *33(2):369-377, 1985.*

[4] Milenkovic, P., "Glottal inverse filtering by joint estimation of an AR system with a linear input model", *IEEE Trans. on Acoust., Speech, and Signal Proc., 34(1):28-42, 1986.*

[5] Alku, P., "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering", *Speech Comm.*, *11:109-118, 1992.*

[6] Krishnamurthy, A.K., "Glottal source estimation using a sum-of-exponentials model", *IEEE Trans. on Signal Proc.*, *40(3):682-686, 1992.*

[7] Gobl, C., Monahan, P., Fitzpatrick, L., and NiChasaide, A., "Automatic source-filter decomposition: A knowledge-based approach", *Proc. of the Speech Maps Workshop*, Vol. 2, paper 9, 1994.

[8] Baken, R.J., "Electroglottography", *J. Voice*, *6(2):98-110, 1992.*

[9] Baer, T., Löfqvist, A., and McGarr, N.S., "Laryngeal vibrations: A comparison between high-speed filming and glottographic techniques", *J. Acoust. Soc. Amer., 73(4):1304-1308, 1983*.

[10] Svec, J.G. and Schutte, H.K., "Videokymography: High-speed line scanning of vocal fold vibration", *J. Voice*, *10(2):201-205, 1996.*

[11] Granqvist, S. and Lindestad, P-Å., "A method of applying Fourier analysis to high-speed laryngoscopy", *J. Acoust. Soc. Amer., 110(6):3193-3197, 2001*.

[12] Fant, G., *The Acoustics Theory of Speech Production*, Mouton, the Hague, 1960.

[13] El-Jaroudi, A. and Makhoul, J., "Discrete all-pole modeling", *IEEE Trans. Signal Proc*., *39:411-423, 1991*.

[14] Story, B.H. and Titze, I.R., "Voice simulation with a body-cover model of the vocal folds", *J. Acoust. Soc. Amer.*, *97(2):1249-1260, 1995.*

[15] Titze, I.R. and Story, B.H., "Rules for controlling low-dimensional vocal fold models with muscle activities", *J. Acoust. Soc. Amer.*, *112(3):1064-1076, 2002.*

[16] Titze, I.R., "Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model", *J. Acoust. Soc. Amer.*, *111(1):367-376, 2002.*

[17] Liljencrants, J., *Speech Synthesis with a Reflection-type Line Analog*, DS Dissertation, Dept. of Speech Comm. and Music Acoust., Royal Inst. of Tech., Stockholm, Sweden, 1985.

[18] Story, B. H., *Physiologically-based Speech Simulation Using an Enhanced Wave-reflection Model of the Vocal Tract*, Ph. D. Dissertation, University of Iowa, 1995.

[19] Holmberg, E., Hillman, R., and Perkell, J., "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice", *J. Acoust. Soc. Amer., 84(2):511-529, 1988.*

[20] Alku, P., Bäckström, T., and Vilkman, E., "Normalized amplitude quotient for parameterization of the glottal flow", *J. Acoust. Soc. Amer., 112(2):701-710, 2002*.

[21] Alku, P., Vilkman, E., "A comparison of glottal quantification parameters in breathy, normal and pressed phonation of female and male speakers", *Folia Phoniatr Logop., 48:240-254, 1996.*