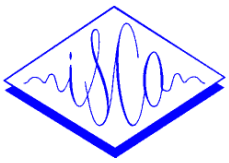

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

# INVERSE FILTERING OF THE GLOTTAL WAVEFORM USING THE ITAKURA-SAITO DISTORTION MEASURE

*Paavo Alku*

Helsinki University of Technology, Acoustics Laboratory
Otakaari 5 A, SF-02150 Espoo, FINLAND

## ABSTRACT

Estimation of the glottal pulseform with inverse filtering is studied in this paper. Automatic computation of the glottal source is usually performed by applying LPC-analysis in order to estimate the vocal tract transfer function. However, the performance of linear prediction decreases as the fundamental frequency of speech increases. Therefore, a new algorithm for computing the vocal tract model is presented. The new method applies the discrete version of the Itakura-Saito distortion measure when the poles of the vocal tract filter are determined. The preliminary results when synthetic and natural vowels were analysed are discussed shortly.

## 1. INTRODUCTION

Accurate estimation of the source of voiced speech, the glottal waveform, is an important task for several applications in speech research. In the computation of the glottal source inverse filtering has become a widely used technique. The earliest methods were based on the manual adjustment of the vocal tract antiresonances [8]. However, this kind of analysis is often troublesome and time consuming to be arranged. The results are also dependent on subjective criteria applied by the user. Therefore, automatic inverse filtering has become a popular approach in the estimation of the glottal flow.

Automatic computation of the glottal waveform is usually performed by estimating the vocal tract transfer function using linear predictive coding (LPC-analysis) [7]. LPC-analysis is a well-known technique that can be implemented using fast algorithms. The performance of linear prediction in the estimation of formants is satisfactory for many applications as far as male voices are concerned.

The performance of LPC-analysis is decreased when voices of high fundamental frequency are analysed [6]. This feature is demonstrated by a simple example depicted in Fig. 1. In Fig. 1(a) a stable all-pole filter, H(z), of order eight is excited by an impulse train, where the number of samples between individual impulses is 80. In Fig. 1(b) the same all-pole filter is excited by an impulse train where the distance between consecutive pulses is only 20 samples. LPC-analysis was computed from the filter outputs in both of the cases using the autocorrelation method together with Hamming-windowing. The order of the prediction was eight and the length of the analysis block was 256 samples. Spectra of the LPC-filters are shown together with the spectrum of filter H(z) in Fig. 2. It can be observed that LPC-analysis is able to estimate filter H(z) fairly accurately in case 1(a). However, in case 1(b), where the fundamental frequency of the excitation is higher, the difference between the spectra of filter H(z) and the corresponding LPC-filter is significant.

The reason why the performance of LPC-analysis deteriorates as the fundamental frequency increases can be explained in various domains [3,6]. A time-domain difference between the cases of Fig. 1(a) and (b) is that in the latter case the response of filter H(z) to an individual impulse of the excitation will not be attenuated enough before the next impulse arrives. Hence, aliasing occurs in the autocorrelation of the output of filter H(z). Thus, linear prediction that takes advantage of the autocorrelation of signal $y_2(n)$ is unable to exactly identify all-pole filter H(z) even though the order of LPC-analysis equals the number of poles of H(z). In the frequency domain this implies that the poles of the LPC-filter in case 1(b) will move towards the nearest harmonic peak.

In the estimation of the glottal source LPC-analysis is usually applied as a tool to model the vocal tract transfer function. After the vocal tract has been modelled the glottal source can be computed by cancelling the effect of formants by inverse filtering. If linear prediction fails to model the vocal tract accurately enough the effect of formants will not be removed from speech completely by inverse filtering. Hence, especially in high pitched voices the resulting glottal waveform will be distorted by a formant ripple.

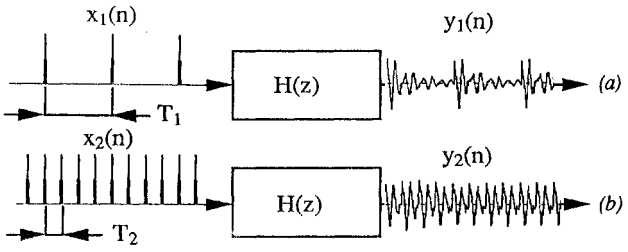


*Fig. 1 Exciting all-pole filter H(z) with an impulse train*
*(a): fundamental period $T_1$ = 80 samples*
*(b): fundamental period $T_2$ = 20 samples*

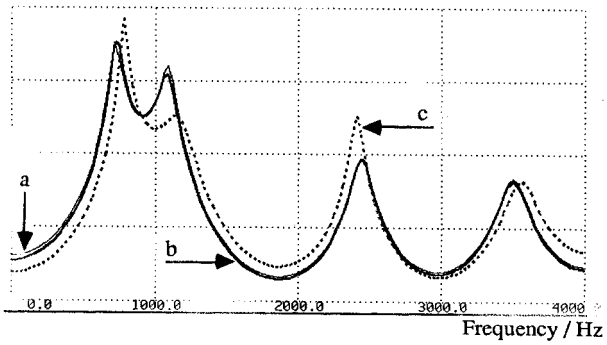


*Fig. 2 Spectra of all-pole filters of Fig. 1:*
*H(z) (curve a), LPC-filter computed from $y_1(n)$ (curve b),*
*LPC-filter computed from $y_2(n)$ (curve c).*

In this paper a new automatic inverse filtering technique is presented to estimate the glottal pulseform. The main purpose is to study whether drawbacks of LPC-analysis in modelling of the formants could be reduced by changing the error criterion that is used in the determination of the vocal tract transfer function. Instead of using the error criterion that is applied in conventional LPC-analysis, i.e. minimising the square of the error of the prediction, we have determined the vocal tract using the discrete version of the Itakura-Saito distortion measure [7]. The study takes advantage of a previously developed algorithm, the IAIF-method [1]. An important background for this paper is the research reported in [3].

## 2. METHOD

In the following the structure of the new glottal wave analysis algorithm is presented in two parts. In section 2.1 the Itakura-Saito distortion measure is described using the simplified scheme of Fig. 1. The application of the proposed error criterion in the inverse filtering of the glottal pulseform is described in section 2.2.

### 2.1 The Itakura-Saito Distortion Measure in the Estimation of an All-Pole Filter Excited by an Impulse Train

Let us further study the simple example of Fig. 1(b) where a stable all-pole filter H(z) is excited by an impulse train $x_2(n)$. The poles of filter H(z) are depicted by circles in the z-domain in Fig. 3. The poles of the LPC-filter, $H_{LPC,2}(z)$, computed from $y_2(n)$ are shown in Fig. 3 by crosses. Since the excitation waveform was of high fundamental frequency, linear prediction failed to yield accurate estimates for the poles of H(z).

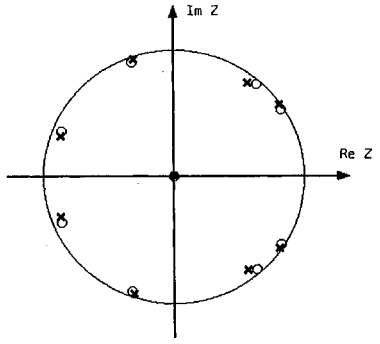


*Fig. 3 Poles of H(z) (circles) and $H_{LPC,2}(z)$ (crosses).*

Conventional LPC-analysis is based on the error criterion according to which an optimal predictor is the one that yields the smallest prediction error (residual) energy. However, if the excitation waveform is an impulse train as the case is in Fig. 1, another error criterion could also be used: the optimal predictor is the one that yields a residual that is as close to an impulse train as possible. This criterion can be quantified using a spectral flatness measure, the Itakura-Saito (I-S) distortion measure [7]. For a given power spectrum (P(k), $0 \le k \le N-1$) the I-S measure can be defined as the geometric mean of the power spectrum samples divided by their arithmetic mean:

$$E_{IS} = \frac{\left[\prod_{k=0}^{N-1} P(k)\right]^{1/N}}{1/N \sum_{k=0}^{N-1} P(k)} \qquad (1)$$

The value of $E_{IS}$ is between zero and one (a perfectly flat spectrum).

To take advantage of the I-S distortion measure in order to improve the estimation of filter H(z) from signal $y_2(n)$ a new algorithm was developed. The method is based on the idea that the pole locations of the conventional LPC-filter are slightly changed in order to generate a residual that maximises the I-S measure. The proposed method is sub-optimal: one complex conjugate pair of poles of the LPC-filter is processed at a time. A small group of pole candidates is generated for each of the LPC-poles. The original LPC-pole is then replaced by the candidate that yields the largest $E_{IS}$-value.

The algorithm comprises the following main stages. Let us denote $H_{IS}(z)$ as the all-pole filter which is to be determined.

1: Conventional LPC-analysis of order $p$ is first computed pitch asynchronously from signal $y_2(n)$. The obtained all-pole filter is denoted $H_{LPC,2}(z)$. Initially, let us make a substitution $H_{IS}(z) = H_{LPC,2}(z)$.
2: Poles of $H_{LPC,2}(z)$, marked as crosses in Fig. 3, are solved and they are sorted according to their angle ($z_{LPC,1}$, $z_{LPC,2}$, etc.).
3: A group of $p$th order all-pole filter candidates ($H_i(z)$, $0 \le i \le R-1$) are created by slightly changing the angle and radius of $z_{LPC,1}$ and $z^*_{LPC,1}$ but keeping the rest of the poles as in $H_{IS}(z)$.
4: Signal $y_2(n)$ is filtered through the inverses of each of the new all-pole filter candidates $H_i(z)$. The corresponding residual signals are denoted $e_i(n)$ ($0 \le i \le R-1$).
5: Power spectrum $|E_i(k)|^2$ is computed from one fundamental period of each of $e_i(n)$ using the FFT.
6: The I-S distortion measure (Eq. 1) is computed from the power spectrum. The filter candidate $H_i(z)$ that gives the largest $E_{IS}$ is denoted $H_{max}(z)$.
7: Substitute $H_{IS}(z) = H_{max}(z)$.
8: Go through steps from 3 to 7 for all those complex conjugate pairs of poles of $H_{LPC}(z)$ that are needed to be adjusted.

Fig. 4 shows the result that was obtained when the example of Fig. 1(b) was processed with the algorithm described above. The poles of H(z) and $H_{IS}(z)$ are denoted by circles and crosses, respectively. By comparing Fig. 3 and 4 it can be observed that changing the error criterion gives an all-pole filter, $H_{IS}(z)$, that more accurately fits the poles of filter H(z) than the one given by conventional linear prediction, $H_{LPC,2}(z)$.

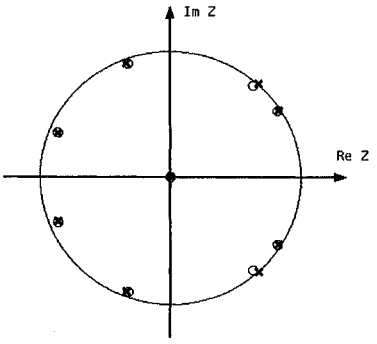


*Fig. 4 Poles of H(z) (circles) and $H_{IS}(z)$ (crosses).*

### 2.2 The Itakura-Saito Distortion Measure in the Estimation of the Glottal Excitation

A method that, in comparison to conventional LPC-analysis, improves estimation of an all-pole process was described in the previous section. An important feature in the scheme of section 2.1 is that the all-pole filter which is to be estimated is excited by an impulse train. Accurate estimation of an all-pole process is important since the vocal tract transfer function is in the case of vowels most

often modelled by an all-pole filter. However, a direct application of the method of section 2.1 in the estimation of the glottal excitation is not possible since the real excitation of the vocal tract is not an impulse train.

In order to combine the method described in the previous section with the estimation of the glottal waveform let us start from the separated speech production model developed by Fant [4]. As shown in Fig. 5(a) this model assumes that speech is produced as a combination of three processes: the glottal excitation, the vocal tract, and the lip radiation effect. Let us first divide the glottal excitation into two parts: an impulse train u(n) and a "glottal wave generator" G(z) (Fig. 5(b)). The lip radiation effect can be modelled as a differentiator. Hence, its effect can be cancelled by integrating the speech signal. This has been shown in Fig. 5(c), where $s_i(n)$ is the integral of s(n). Before we can use the method developed in section 2.1 for the computation of the vocal tract we should somehow be able to estimate filter G(z). This can be done using a previously developed inverse filtering method, IAIF [1]. The output of the IAIF-method, an estimate for the glottal source, is analysed by low order linear prediction. The inverse of the resulting LPC-filter is then used to cancel the estimated effect of G(z) from $s_i(n)$. The result of filtering $s_i(n)$ through the inverse of a low order LPC-filter is denoted $s_{ii}(n)$ in Fig. 5(d). Hence, the separated speech production model (Fig. 5(a)) can be changed into a form (Fig. 5(d)) that is similar to the scheme (Fig. 1) of the previous section.

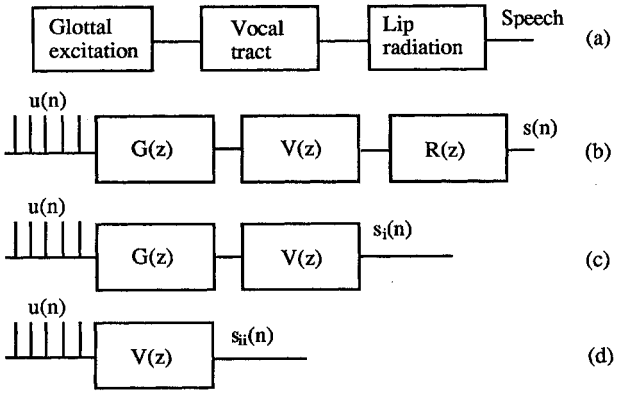


*Fig. 5*
*(a): Separated speech production model*
*(b): The model excited by an impulse train (u(n) = impulse train, G(z) = glottal wave generator, V(z) = vocal tract, R(z) = lip radiation effect, s(n) = speech)*
*(c): Cancelling the lip radiation effect ($s_i(n)$ = integrated version of s(n))*
*(d): Cancelling the estimated effect of the glottal wave generator ($s_{ii}(n) = s_i(n)$ filtered through the inverse of the low order all-pole filter that estimates G(z))*

The final algorithm in order to estimate the glottal excitation can now be presented by combining the algorithm of section 2.1 with the speech production model of Fig. 5(d):

1: Compute a preliminary estimate, $g_p(n)$, for the glottal waveform with the IAIF-method [1].
2: Compute low order (2 or 4, typically) LPC-analysis from signal $g_p(n)$.
3: Cancel the estimated effect of the glottal source by filtering the original speech signal through the inverse of the obtained low order LPC-filter.
4: Cancel the lip radiation effect by integrating the output of stage no. 3.
5: Process the output of stage no. 4 by the algorithm described in section 2.1. The result is an all-pole filter, $H_{IS}(z)$, which forms the final model for the vocal tract.
6: Cancel the effect of the vocal tract by filtering the original speech signal through the inverse of $H_{IS}(z)$.
7: The final result is obtained by cancelling the lip radiation effect by integrating the output of stage no. 6.

## 3. RESULTS

### 3.1 General

The new algorithm was tested using both synthetic and natural speech. Since the method uses an all-pole filter in the modelling of the vocal tract only vowels were analysed. Natural speech material was recorded in an anechoic chamber using the Brüel&Kjær 4134 condenser microphone together with a video cassette recorder (Sony PCM-F1 and Sony SL-F1E). The algorithm was implemented on a Symbolics Lisp-machine.

### 3.2 Synthetic Speech

Synthetic "a"-vowels were created using a procedure described in [5]. As a synthetic source we used a waveform developed in [2]. The vocal tract was modelled using an all-pole filter of order eight. Synthetic speech signals were created by changing the fundamental frequency from 100 Hz to 267 Hz (i.e. the length of the glottal period was varied from 80 to 30 samples). Glottal pulseforms were generated in order to simulate (very) pressed phonation type.

The analysis was computed using eighth order all-pole filtering in the modelling of the vocal tract. The block length of the analysis was 512 samples. (The IAIF-analysis was performed using the following parameters (see Fig. 1 in [1]): p=r=10). The waveform given by the IAIF-analysis was analysed using second order linear prediction in order to get filters to eliminate G(z) (Fig. 5(d)).

Fig. 6 shows typical results that were obtained when synthetic speech material was analysed. Curves 6(a) and (b) correspond to the fundamental frequency of 100 Hz. Curve 6(a) shows the shape of the original glottal excitation. The result given by the new algorithm is shown by curve 6(b). It can be observed that the waveform of Fig. 6(b) is very similar to the original glottal pulseform.

Fig. 6(c) and (d) correspond to the fundamental frequency of 267 Hz. By comparing the original glottal source (curve 7(c)) to the result given by the new algorithm (curve 7(d)) it can be observed that also in the case of high fundamental frequency the result is quite satisfactory. A small formant ripple can be observed during the closed phase of the glottal cycle.

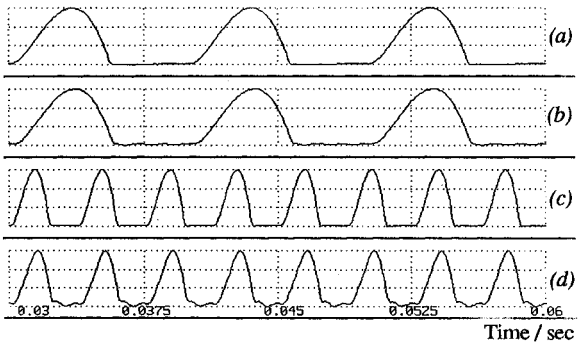


*Fig. 6*
*(a): Original synthetic glottal waveform, fundamental frequency 100 Hz*
*(b): Estimated glottal waveform, fundamental frequency 100 Hz*
*(c): Original synthetic glottal waveform, fundamental frequency 267 Hz*
*(d): Estimated glottal waveform, fundamental frequency 267 Hz*

### Natural Speech

The new method was used in the analysis of natural speech by studying utterances that were produced by one female and one male speaker. Both of the subjects were of healthy voices. The speech material comprised all eight vowels of the Finnish language. The speakers were asked to produce the vowels using their normal phonation type.

The analysis was computed using 10th order all-pole filtering in the modelling of the vocal tract. The block length of the analysis was 512 samples. (The IAIF-analysis was performed using the following parameters (see Fig. 1 in [1]): p=12, r=10.) The waveform given by the IAIF-analysis was analysed using second and fourth order linear prediction in order to get filters to eliminate G(z) (Fig. 5(d)).

Fig. 7 shows typical results that were obtained for the male speaker. Curves 7(a) and (b) correspond to the analysis of vowels "a" and "i", respectively. Both of the curves of Fig. 7 correspond well with an *a priori* knowledge of a typical glottal pulseform for a male speaker. Analysis of the glottal waveform from vowel "i" is in general difficult due to the low first formant. The results showed that the new method yields a fairly reliable glottal wave estimate also for "difficult" vowels like "i" and "u".

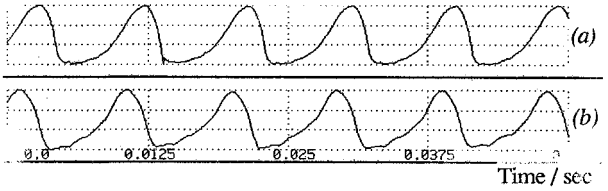


Fig. 7
*(a): Estimate for the glottal excitation, male speaker, vowel "a"*
*(b): Estimate for the glottal excitation, male speaker, vowel "i"*

Fig. 8 shows some of the results that were obtained when female speech was analysed. Curves 8(a) and (b) correspond to the analysis of vowels "u" and "ae", respectively. Glottal pulseforms that were obtained from the female vowels were in general quite smooth. Some of the pulseforms (curve 8(a), for example) were almost sinusoidal, which is characteristic for the glottal excitation of female speakers. However, part of the results (curve 8(b), for example) comprised a clear closed phase.

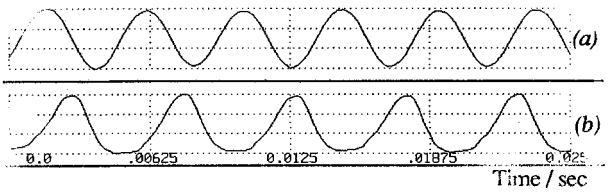


Fig. 8
*(a): Estimate for the glottal excitation, female speaker, vowel "u"*
*(b): Estimate for the glottal excitation, female speaker, vowel "ae"*

## 4. CONCLUSIONS

The application of the Itakura-Saito distortion measure in the automatic analysis of the glottal excitation was reported in this preliminary study. The research was motivated by the fact that the most frequently used automatic inverse filtering tool, LPC-analysis, has certain inherent drawbacks. The target of the study was to find out if changing the mean square error criterion that is applied in conventional LPC-analysis to the IS-distortion measure in the modelling of the vocal tract transfer function could improve the estimation of the glottal excitation.

The preliminary results showed that the new technique was able to give more reliable results compared to the previously developed method, IAIF, that is based on conventional LPC-analysis. The improvements were most significant when vowels of high F0 and/or low F1 were analysed. This results mainly from the fact that the estimation of the vocal tract transfer function, especially the first formant, can be performed more accurately when the analysis is based on the IS-distortion measure. The new method has some features that motivate further studies: the algorithm is automatic, non-invasive, and it uses only one input signal, the acoustical speech pressure wave. However, in comparison to conventional techniques based on LPC-analysis the computational load of the new algorithm is larger.

## REFERENCES:


[1] Alku, P., Vilkman, E., and Laine U.K. (1991). "Analysis of glottal waveform in different phonation types using the new IAIF-method" Proc. XIIth Int. Congress of Phonetic Sciences, Vol. 4, pp. 362-365.

[2] Ananthapadmanabha, T.V. (1984). "Acoustic analysis of voice source dynamics", STL-QPSR, No. 2-3, pp. 1-24, Stockholm: Royal Institute of Technology.

[3] El-Jaroudi, A., Makhoul, J. (1991). "Discrete all-pole modelling," IEEE Trans. Signal Proc., Vol. 39, pp. 411-423.

[4] Fant, G. (1960). *Acoustic theory of speech production* (Mouton, The Hague).

[5] Gold, B., Rabiner, L.R. (1968). "Analysis of digital and analog formant synthesizers", IEEE Trans. Audio and Electroacoustics, Vol. 16, pp. 81-94.

[6] Makhoul, J. (1975). "Linear prediction: A tutorial review", Proc. IEEE, Vol. 63, No. 4, pp. 561-580.

[7] Markel, J.D., and Gray, A.H., Jr. (1976). *Linear prediction of speech* (Springer-Verlag, New York).

[8] Miller, R.L. (1959). "Nature of the vocal cord wave", J. Acoust. Soc. Am., Vol. 31, No. 6, pp. 667-677.