PULSE-BY-PULSE PITCH ANALYSIS THROUGH ZERO PHASE LOW-PASS FILTERING

Gunnar Hult

R&D Division
Swedish Telecom

ABSTRACT - A recently proposed pitch detection algorithm is based on iterative low-pass filtering of the speech signal. We discuss some modifications to the proposed algorithm, such as appropriate halting criteria for the iterative filtering, and also how the pitch can be determined from the final, sinusoidal-like filter output. Finally, we compare these pitch estimates to those of two well-known pitch analysis methods, cepstral filtering and time domain parallel processing.

INTRODUCTION

Accurate determination of the pitch frequency of a speech signal is often required in speech based applications such as systems for mobile telephony. In some cases, e.g. when analyzing microprosodic features in systems for automatic speech recognition, it is of interest to obtain pitch estimates which are instantaneous and time-synchronized with the original speech signal and which do not contain any averaging effects.

A RECENTLY PROPOSED PITCH DETECTION ALGORITHM

The recently presented pitch detection algorithm of Dologlou & Carayannis (1989) uses iterative low-pass filtering of the speech signal followed by pulse-to-pulse measurements on the iteratively filtered signal. The filter is a zero phase, non-causal, second order recursive filter

$$H(z) = \frac{1}{4}z + \frac{1}{2} + \frac{1}{4}z^{-1} \qquad (1)$$

The zero phase requirement implies that no phase shift is introduced into the filtered signal and that the timing information is retained relative to the speech signal.

The halting criterion for the iterative filtering is that a frequency measure f_1 derived from an autocorrelation analysis of the signal is sufficiently close to a frequency measure f_2 derived from a second order LPC analysis of the signal.

For the case of a single sinusoid of frequency f_1, amplitude A and sampled at a rate f_S, the proposed algorithm (Dologlou & Carayannis, 1989) relates the frequency f_1 to the first lag autocorrelation coefficient r(1) through

$$f_1 = \frac{f_S}{2\pi} \arccos\left[\frac{2r(1)}{A^2}\right] \qquad (2)$$

The frequency measure f_2 in Dologlou & Carayannis (1989) is determined from a second order LPC analysis. With a second order predictor polynomial

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} \qquad (3)$$

the coefficients a_1 and a_2 are found from the autocorrelation formulation of linear prediction (Markel & Gray, 1976) and we obtain a frequency estimate

$$f_2 = \frac{f_S}{2\pi} \arccos\left[\frac{a_1}{-2\sqrt{a_2}}\right] \qquad (4)$$

which is the second frequency estimate of Dologlou & Carayannis (1989).

MODIFICATIONS TO THE PROPOSED ALGORITHM

In the frequency estimate (2), it is not very clear how the constant A is to be interpreted for anything except a single sinusoid. The most reasonable generalization of (2) to make it applicable to arbitrary signals is

$$f_1 = \frac{f_s}{2\pi} \arccos\left[\frac{r(1)}{r(0)}\right] \tag{5}$$

where r(0) and r(1) are the first two autocorrelation coefficients. For a single sinusoid, (2) and (5) are equivalent.

With an input consisting of a single sinusoid of amplitude A, frequency ϕ, and phase θ,

$$y(n) = A \sin\left[\frac{2\pi\phi n}{f_s} + \theta\right] \quad n=0,1, \ldots \tag{6}$$

it is straightforward to show that the two frequency measures (4) and (5) produce identical and correct estimates

$$f_1 = f_2 = \phi \tag{7}$$

This is in fact the basis for the proposed algorithm of Dologlou & Carayannis (1989): Filter the signal repeatedly until it is sufficiently sinusoidal. The iterative filtering is stopped when f_1 and f_2 are sufficiently close.

For signals which are not purely sinusoidal, e.g. speech, such a halting criterion will run into problems, however. Consider an input signal consisting of two sinusoids:

$$y(n) = A \sin\left[\frac{2\pi\phi_1 n}{f_s} + \theta_1\right] + B \sin\left[\frac{2\pi\phi_2 n}{f_s} + \theta_2\right] \quad n=0,1,\ldots \tag{8}$$

where the amplitudes are A and B, the harmonically unrelated frequencies are ϕ_1 and ϕ_2 and the phases are θ_1 and θ_2.

Denote the amplitude ratio B/A by a real number ß,

$$ß = \frac{B}{A} \tag{9}$$

It can be shown (Hult, 1990) that the algorithm will halt prematurely for certain values of ß, ϕ_1, ϕ_2, and f_s. For instance, with

$$\phi_1 = 600 \text{ Hz} \tag{10}$$

$$\phi_2 = 1250 \text{ Hz} \tag{11}$$

$$f_s = 4000 \text{ Hz} \tag{12}$$

and an initial amplitude ratio $ß > ß_0$, where

$$ß_0 = 1.2393... \tag{13}$$

the iterative filtering will gradually decrease ß until $ß \approx ß_0$ and will then halt.

With ϕ_1 corresponding to the desired fundamental frequency F_0 and ϕ_2 corresponding to the undesired first formant frequency F_1, such values may appear (Fant, 1969) for high-pitched female and children's voices during certain vowels, such as the open, unrounded front vowel /a/. Speech signals with such initial amplitude ratios could occur in band-limited telephony systems where the first harmonics of the fundamental frequency are often severely damped.

In the remainder of this paper, we use a fixed number of filter iterations on the order of 400. This also has the advantage of avoiding a large number of computationally expensive autocorrelation and LPC analyses.

After N iterations with the low-pass filter in (1), the signal component at frequency ϕ will be damped by a factor

$$\cos^{2N}\left[\frac{\pi\phi}{f_s}\right] \qquad (14)$$

relative to its original value. In order to avoid numerical underflow, we may have to introduce a DC gain factor >1 for large values of N. In this paper, the signal is normalized to full 16-bit dynamic range after each iteration.

CALCULATING THE PITCH FROM THE ITERATIVELY FILTERED SIGNAL

No indication is given in Dologlou & Carayannis (1989) on how the pitch frequency should be determined once the sinusoidal-like filter output is obtained. Due to the filter transfer function (14), the filtered signal will have a frequency-modulated character, as seen in Figure 1. below. This frequency modulation must not influence the subsequent pitch estimates.

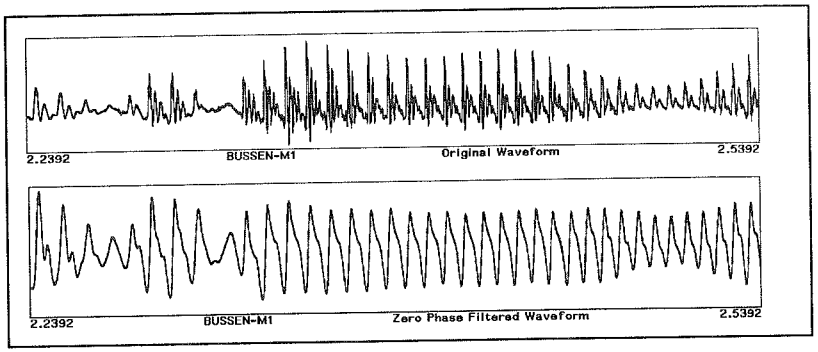

Figure 1. From the top: Part of a speech signal (/dzɑ:g/ sequence)
The iteratively filtered signal after 400 iterations

We have adopted a time-domain technique originally proposed in Gold & Rabiner (1969). An impulse is placed at each valley of the filtered signal. To avoid any influence from the frequency-modulated nature of the filtered signal, the impulse has a fixed height. This impulse train is then processed by a time-varying non-linear system consisting of a dead-zone followed by an exponential decay. Any sufficiently large impulse will set the output level to the height of the impulse and then remain at that level during a certain time, known as the blanking interval. During the blanking interval no pulse can be detected, but during the following exponential decay any pulse that exceeds the decaying exponential will restart the process. A pitch period estimate is obtained as the time between two adjacent process restarts.

In this paper we use a blanking time of 4 ms and a 2 ms time constant for the exponential decay. These values have been selected to fit male voices. A more flexible approach may be to use the strategy of Gold & Rabiner (1969), where the blanking time and decay time constant are changed adaptively based on previously obtained pitch estimates.

The original algorithm of Gold & Rabiner (1969) uses additional impulse trains determined by peaks in the signal and also by combinations of adjacent peaks and valleys. The final pitch estimate is then obtained as the most consistent estimate among all the non-linearly processed impulse trains. The rationale for this is that the pre-processing proposed in Gold & Rabiner (1969), low-pass filtering under 900 Hz, can lead to signals which retain strong higher harmonics. The significantly more filtered signals in Dologlou & Carayannis (1989) should eliminate the need for these additional processors.

The locations of the valleys in the filtered signal have been shown (Dologlou & Carayannis, 1989) to correspond very closely to the location of the maximum derivative of the corresponding laryngograph signal. They are consequently good indicators of the moment of closure of the vocal chords.

Figure 2. below shows an example with part of a speech signal, the iteratively filtered signal, the impulse train and the output of the non-linear impulse train processor.

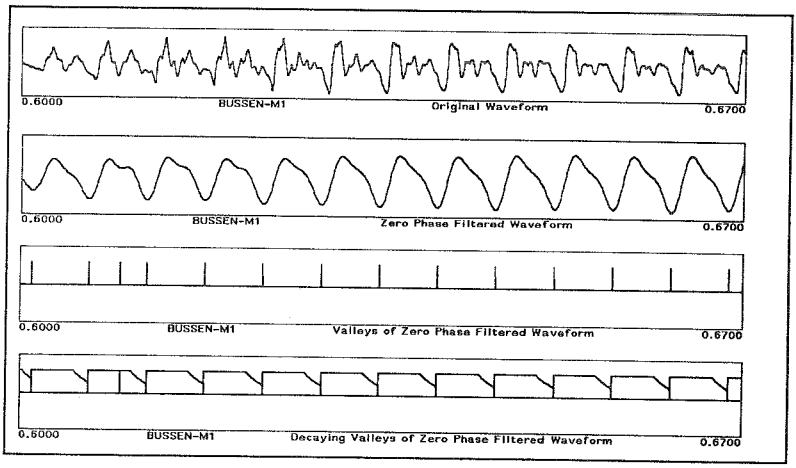

Figure 2. From the top: Part of a speech signal (unrounded front vowel /ɛ/)
The iteratively filtered speech signal
The impulse train
The output of the non-linear processor

PITCH CALCULATION ON SYNTHETIC SPEECH

To be able to judge the pitch estimates from the new algorithm we need a signal where the underlying, "true" pitch values are known. Such a signal may be obtained from a speech synthesizer with user-controllable pitch frequency. For this experiment we used the Klatt-Talk formant synthesizer (Klatt, 1980) where the pitch frequency is specified as a numerical value (in Hz) every 5 ms. The rather simple excitation function of the Klatt-Talk synthesizer, an impulse train filtered by a second order recursive filter, was replaced by a more realistic model (Anantha-padmanabha, 1984).

As comparisons to the new algorithm we used two established pitch detection methods, the time-domain parallel processing of Gold & Rabiner (1969) and a cepstral filtering method (Noll,1967). When all the voiced segments of two synthetic utterances are analyzed, the new algorithm produces more reliable pitch estimates, as seen from Table 1. below.

Method	Average error sentence 1	Average error sentence 2
Zero phase filtering	2.99 %	3.29 %
Gold-Rabiner	7.69 %	8.64 %
Cepstrum filtering	4.86 %	6.49 %

Table 1. The average pitch magnitude error for three pitch detection algorithms running on all voiced segments of two synthetic speech utterances.

It can be argued that the comparison in Table 1. is biased since the zero phase filtering pitch algorithm does not attempt to make a voiced-unvoiced decision but always generates a pitch estimate. Any voiced frame erroneously classified as unvoiced by the other two methods will strongly influence the error figures of Table 1. We therefore made a second comparison between the three methods, where we excluded any frame containing a voiced-to-unvoiced error for any of the other two methods. The results from this second comparison are shown in Table 2. below.

Method	Average error sentence 1	Average error sentence 2
Zero phase filtering	1.01 %	1.13 %
Gold-Rabiner	3.82 %	4.07 %
Cepstrum filtering	3.07 %	2.98 %

Table 2. The average pitch magnitude error for three pitch detection algorithms running on voiced segments of two synthetic speech utterances. All frames where the second and third method make a voiced-to-unvoiced error are excluded.

DETECTION OF MICROPROSODIC FEATURES

Many aspects of pitch behavior are important for speech recognition. For example, in a plosive-vowel combination, the presence or not of voicing during the occlusion of the plosive and the vowel-initial pitch frequency transients may both help to distinguish the voiced plosives /b,d,ɖ,g/ from the unvoiced plosives /p,pʰ,t,tʰ,tʲ,k,kʰ/. Figure 3. below shows the vowel-initial perturbation of the pitch that is often caused by a preceding voiceless stop consonant (Silverman, 1984).

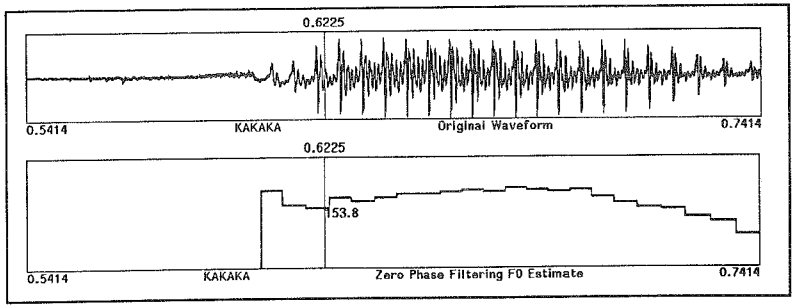

Figure 3. From the top: Part of a speech signal (second syllable in /kʰakʰaˈkʰɑː/ The pitch estimate obtained from zero phase filtering where the vertical pitch range is 100 Hz to 200 Hz.

We have seen ample evidence that the new pitch algorithm is better than both the Gold-Rabiner algorithm and the cepstral algorithm at performing many such distinctions. These phenomena are typically not found by methods that smooth the pitch estimates over a window (e.g. the

cepstrum method) or by methods that have an inherent delay of several pitch pulses (e.g. the Gold-Rabiner algorithm).

The analysis of certain types of voice pathologies also benefits from the new algorithm. Figure 4. below shows analysis results from a voice suffering from diplophonia, a pathological condition (Максимов, 1987) where the excitation exhibits paired pulses.

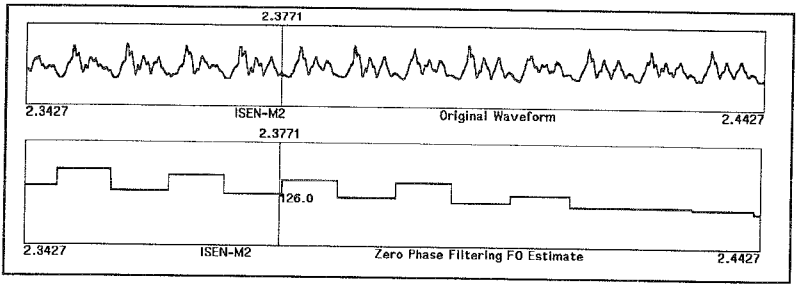

Figure 4. From the top: Part of a diplophonic speech signal (long vowel /æ:/)
The pitch estimate obtained from zero phase filtering
where the vertical pitch range is 100 Hz to 150 Hz.

CONCLUSIONS

We have described an algorithm which in several ways seems to give improved estimates of the pitch frequency of speech signals. Among the remaining problems with the algorithm is finding an appropriate, signal-dependent halting criterion for the iterative filtering, making the blanking time adaptive on previous pitch estimates and extending the algorithm to voiced-unvoiced decisions.

REFERENCES

Ananthapadmanabha, T.V. (1984) *Acoustic analysis of voice source dynamics,* Speech Transmission Laboratory, Quarterly Progress and Status Report 2-3, Stockholm.

Dologlou, I. & G. Carayannis (1989) *Pitch detection based on zero-phase filtering,* Speech Communication 8, pp. 309-318.

Fant, G. (1969) *Formant frequencies of Swedish vowels,* Speech Transmission Laboratory, Quarterly Progress and Status Report 4/1969, Stockholm.

Gold, B. & L.R. Rabiner (1969) *Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain,* J. Acoust. Soc. Am., Vol. 46, No. 2, Pt. 2, pp. 442-448.

Hult, G. (1990) *Some remarks on the halting criterion for iterative low-pass filtering in a recently proposed pitch detection algorithm,* submitted to Speech Communication.

Klatt, D. (1980) *Software for a cascade/parallel formant synthesizer,* J. Acoust. Soc. Am., Vol. 67, pp. 971-995.

Максимов, И. М. (1987) *Фониатрия,* Издательство Медицина, Москва.

Markel, J. & A.H. Gray, Jr. (1976) *Linear Prediction of Speech,* Springer-Verlag, Berlin.

Noll, A.M. (1967) *Cepstrum Pitch Determination,* J. Acoust. Soc. Am., Vol. 41, pp. 293-309.

Silverman, K. (1984) *F0 Perturbations as a Function of Voicing of Prevocalic and Postvocalic Stops and Fricatives, and of Syllable Stress,* Proceedings of the Institute of Acoustics, Autumn Conference, Windermere.