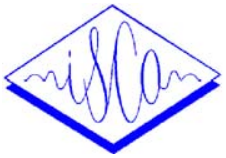

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

AN EFFICIENT F0 DETERMINATION ALGORITHM BASED ON THE IMPLICIT CALCULATION OF THE AUTOCORRELATION OF THE TEMPORAL EXCITATION SIGNAL

Joseph Di Martino, Yves Laprie

LORIA - UMR 7503
B.P. 239 54506 Vandœuvre-lès-Nancy Cedex
FRANCE
Tel: (+33) 3 83 91 21 62 Fax: (+33) 3 83 27 36 84
E-mail: jdm@loria.fr

10.21437/Eurospeech.1999-611

ABSTRACT

In this paper we are presenting a new algorithm for determining the fundamental frequency. The evaluation of pitch is a very difficult problem mainly because of the great variability and irregularity of the speech signals. The algorithm we are presenting is original so far as it relies on the implicit calculation of the autocorrelation of the temporal excitation signal. We have tested our algorithm on the Bagshaw database, created at the Center for Speech Technology Research at Edinburgh, which is primarily dedicated to the evaluation of algorithms estimating the fundamental frequency of speech. The results of our experiments show that our approach is very reliable.

1. INTRODUCTION

The applications of the pitch determination are numerous, for instance the synthesis and recognition of speech, the diagnosis of vocal fold pathologies, the study of prosody in the fields of phonetics and linguistics, the assistance with voice therapies, etc... A great number of algorithms have been put forward [4], but no truly efficient solution has up to now made it possible to move towards a robust solution whatever the voice of the speaker. The purpose of our work is to design a very reliable algorithm by working almost exclusively in the frequency domain as did Markel [5], Martin [7] or Hermes [3] unlike more recent works, that of Medan et al. [8] for instance. The experimental part of our paper shows that results that we obtain are of an excellent quality.

2. OVERVIEW OF OUR ALGORITHM

Our algorithm is based on the model of speech production shown in Fig. 1. The speech wave according to this model is the result of the convolution of the glottal source $E(z)$ and of the impulse response of the vocal tract $H(z)$. The information about the pitch is entirely placed in the glottal wave. It follows that an estimation of $E(z)$ is needed. In the frequency domain $E(z)$ appears like a periodical signal in the case of voiced sounds and like a noise signal in the case of unvoiced sounds. So, a priori, if the excitation signal is isolated a good estimation of the fundamental frequency can be obtained.

Our algorithm works in three steps. First, we determine the amplitude spectrum of the excitation $E(z)$. Then, we determine pitch candidates and finally, we determine the F0 curve by means of a post-processing algorithm.

3. DETERMINATION OF $E(Z)$

Several methods to determine the glottal wave have been put forward. In particular we should mention cepstral smoothing [10], inverse filtering with linear prediction [6] and frequency filtering [11]. In [11] the three methods are discussed and, as indicated in this paper, we consider that the frequency filtering is undoubtedly one of the best existing methods. That is the reason why we accepted this solution to determine $E(z)$.

More precisely, we proceeded in the following manner. The speech signal is windowed in a Hamming window. A Fourier transform is then applied on that signal and the amplitude spectrum logarithm is calculated. The frequency filtering is applied on that spectrum to evaluate the contribution of the vocal tract $H(z)$. We then obtain the excitation spectrum by subtracting the vocal tract contribution to the spectrum.

The frequency filtering is carried out by an FIR filter applied on the amplitude spectrum logarithm. In order to apply the FIR filter to the spectrum it is necessary to have a frequency "signal" which is sufficiently long. We duplicate therefore the spectrum three times. The filter is defined on a large number of points in order to avoid any Gibbs effect. Its frequency characteristics have been designed so as to eliminate, on the spectral level, the frequency components such that $ff> \ _c$ where the low "frequency" f_c corresponds to the frequency wave of a signal whose pitch is 1000 Hz.

From this frequency signal we extract the part which

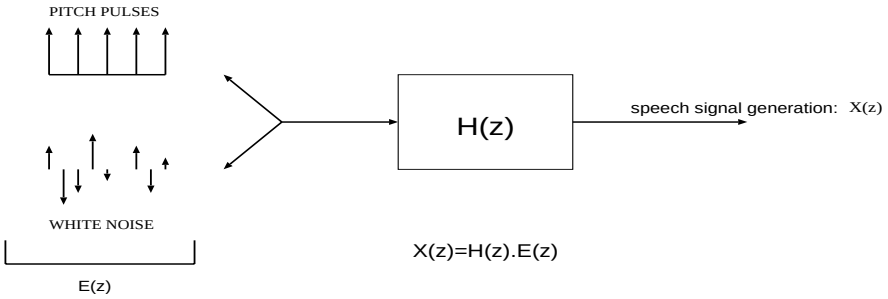

Figure 1: The speech production model

corresponds to the filtered amplitude spectrum. The log-spectrum of the excitation is obtained by subtracting the filtered frequency signal from the initial log-spectrum. An exponential of the log-spectrum gives $E(z)$ (see Fig. 2).

4. FO DETERMINATION

The second step of the algorithm consists in finding F0 candidates and deciding whether speech is voiced or not. We think that the most important point of our algorithm is the calculation of a quasi-autocorrelation of the excitation signal. For that purpose we calculate the modulus of the Fourier transform of $E(z)$. To understand the reason behind this choice, one must keep in mind that $FT^{-1}(|FT|^2)$ (where FT is the Fourier transform and FT^{-1} the inverse Fourier transform) produces the exact circular autocorrelation of the initial signal. A demonstration of that can be found in [1] pp. 93-94 in the particular case where one effects the complex product of two sequences, one sequence being the conjugate of the other. It is clear now that the FT operator followed by the modulus operator applied to the excitation amplitude - not squared - spectrum in fact provides a signal which is similar to the autocorrelation of the temporal excitation signal. We have named that type of signals, the CATE signals (Circular Autocorrelated Temporal Excitation). The interest of our method lies in the fact that the excitation signal can generally not be obtained in a straightforward manner. So, what we are proposing enables us to have easy access to the circular autocorrelation of the temporal excitation signal. Having observed a very large number of circularly autocorrelated excitation signals, it clearly appears that in the case of voiced segments, the CATE signals present a periodic succession of peaks of decreasing amplitude which we have called temporal harmonics in this paper. In the case of unvoiced segments the CATE signals present a series of peaks of irregular amplitudes. Fig. 3 schematically sums up the process we apply to $E(z)$ to extract F0. Fig. 4 (resp. Fig. 5) shows the periodogram (CATE signal) of an excitation autocorrelated signal in the case of a female (resp. male) voice for a voiced speech signal. Fig. 6 shows a periodogram of an excitation autocorrelated signal for an unvoiced signal.

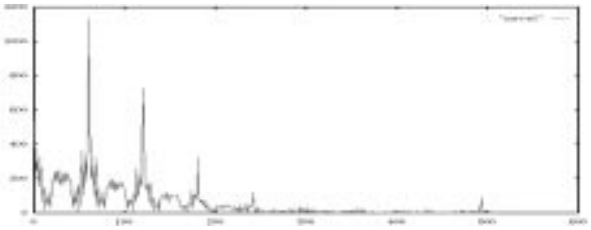

Figure 4: Peridogram of a voiced segment for a female speaker

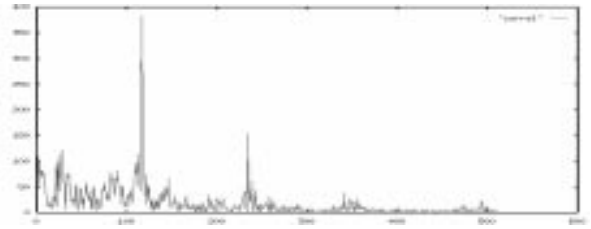

Figure 5: Peridogram of a voiced segment for a male speaker

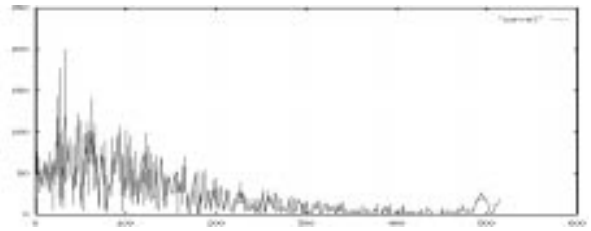

Figure 6: Peridogram of an unvoiced segment

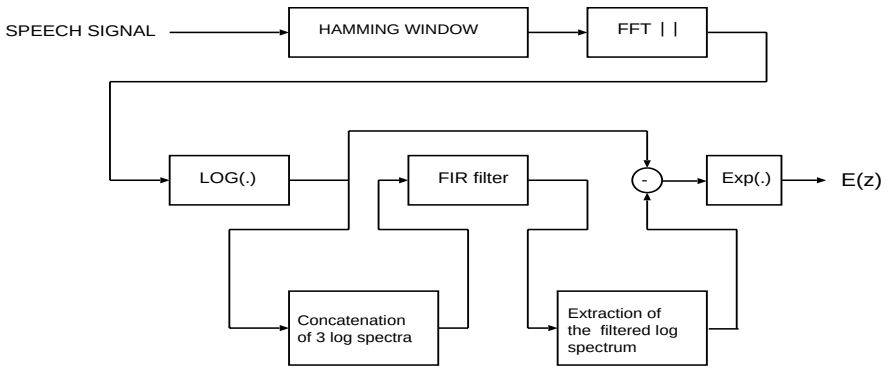

Figure 2: Layout of the dermination of the glottal wave spectrum

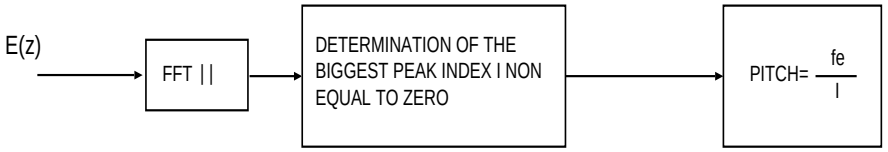

Figure 3: Determination of the F0 candidates from the glottal spectrum

5. POSTPROCESSING ALGORITHM

The aim of this step is the construction of a pitch curve sufficiently smooth and which implicitely includes an optimal voiced/unvoiced decision. Our postprocessing algorithm performs simultaneously both voicing decision and smoothing. It stems from a non-linear smoothing algorithm proposed by Ney [9] which uses dynamic programming to select points from a curve in an optimal way to obtain a sub-curve sufficiently smooth. It operates on an array of values $A = [a(i)](0 \leq i \leq N)$ and searches for a function J which minimizes $D = \sum_{k=2}^{K}(d((j(k), j(k-1)) - B)$, where j represents a selection function which yields a strictly rising sequence of indices $J = [j(k)](0 \leq k \leq K, 0 \leq K \leq N)$ and B represents a bonus which prevents the dynamic programming from giving the empty function for J as a solution.

We have extended the structure of Ney's algorithm in the following way:

- several F0 candidates are considered (3 non-zero F0 values plus a voiceless candidate) in the postprocessing of the CATE algorithm. Using a voiceless candidate allows both smoothing and voicing decision to be performed simultaneously.
- for each candidate a bonus is calculated which represents its reliability by incorporating the periodogram value as well as a voicing criterion. The higher the reliability, the more likely it belongs to the final curve.

Furthermore, the penalization computation has been adapted as follows:

- pitch period doublings as well as pitch period halvings are heavily penalized.
- the second derivative of F0 is taken into account to obtain a sufficiently smooth pitch curve.

6. IMPLEMENTATION

The window length is 51.2 ms which corresponds to 1024 samples for a sampling frequency of 20 kHz. The shift between two consecutive windows is 4 ms in order to achieve an accurate voicing decision.

The normalized frequency of the low pass filter is $1/51.2 = 0.01953$ (51.2 corresponds to 1000 Hz for a spectrum calculated on the 1024 sample window) and the filter order is 400. In fact, one must keep in mind that the Gibbs effect is stronger as the pass band is narrower. This phenomenon introduces an extra oscillation in the vocal tract spectrum, then corrupts the excitation spectrum, and eventually leads to spurious peaks in CATE signals. A solution could be to use a larger pass band but in this case the vocal tract spectrum as well as the excitation spectrum are not correctly estimated because the separation between the two contributions is not sufficient. The 400 point filter allows a correct separation of the two contributions to be achieved and prevents the Gibbs phenomenon we observe otherwise. Our approach has enabled us to obtain a pitch domain ranging from 70 Hz to 1000 Hz.

7. RESULTS

We tested our algorithm on the database of the CSTR (Center for Speech Technology Research). This database was recorded by two speakers, one male, the other female. The average pitch is 216 Hz, the minimum pitch is 60 Hz and the maximum pitch is 400 Hz. The approximative duration of this database is 5 minutes.

We obtained an average frequency error less than 1.5%, which corresponds to a precision in the estimation of the fundamental period less than one sample. The results provided by our algorithm are shown in Tab. 1. They are better than those of Bagshaw et al. [2] especially for gross errors. These results show clearly that our approach

PDA	Unvoiced error (%)	Voiced error (%)	Gross error high (%)	Gross error low (%)	Abs. deviation mean (Hz)	Abs. deviation s. dev (Hz)
eSRPD	4.63	12.07	0.90	0.56	1.40	1.74
CATE	6.13	9.20	0.16	0.21	1.81	2.81
eSRPD	2.73	9.13	0.43	0.23	4.17	5.13
CATE	4.40	6.96	0.29	0.37	4.24	5.81
CATE	5.20	8.01	0.24	0.31	3.29	5.00

Table 1: PDA evaluation for male speaker (top), female speaker (medium) and both (low). eSRPD is the enhanced (by Bagshaw et al.) super resolution pitch determination algorithm proposed originally by Medan et al. [8], CATE is our algorithm.

reduces the gross error rate substantially in the case of the male voice (approximately 75%). In the case of the female voice eSRPD and CATE give equivalent gross error rates. With regards to the voicing decision our algorithm provides better results for the male as well as the female voices. Whatever the case, voicing errors are well distributed for our algorithm which could significate that the behavior of our voicing criterion is better than that of eSRPD. On the other hand, the absolute deviation is slightly greater than that of Bagshaw because of the relatively long window we used (51.2 ms) which should be reduced especially for female voices.

8. CONCLUSION

In this paper we have presented an algorithm of pitch determination whose originality resides in the fact the autocorrelation of the excitation signal is extracted in an indirect manner from the log-spectrum of the speech. F0 candidates given by our algorithm are correct in most cases. The results we presented on the CSTR corpus clearly demonstrate the validity of the approach we adopted.

9. REFERENCES

[1] R. W. Shafer A. V. Oppenheim. *DIGITAL SIGNAL PROCESSING*. Prentice-Hall, 1975.

[2] P. C. Bagshaw, S. M. Hiller, and M. A. Jack. Enhanced pitch tracking and the processing of f0 contours for computer aiding intonation teaching. In *Proceedings of European Conference on Speech Technology*, volume 2, pages 1000–1003, Berlin, September 1993.

[3] D. J. Hermes. Measurement of pitch by subharmonic sommation. *Journal of Acoustical Society of America*, 83(1):257–264, 1988.

[4] W. J. Hess. *Pitch Determination of Speech Signals - Algorithms and Devices*. Springer Berlin, 1983.

[5] J. D. Markel. The sift algorithm for fundamental frequency algorithm. *IEEE Trans. Audio Electroacoustic.*, Au-20(5):367–377, December 1972.

[6] J.D. Markel and A.H. Gray. Automatic formant trajectory estimation. In *Linear Prediction of Speech*, chapter 7. Springer-Verlag, Berlin Heidelberg New York, 1976.

[7] Ph. Martin. Comparison of pitch detection by cepstrum and spectral comb analysis. In *Proc. of Int. Conf. Acoust., Speech, Signal Processing 1982*, pages 180–183, 1982.

[8] Y. Medan, E. Yair, and D. Chazan. Super resolution pitch determination of speech signals. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-39(1):40–48, January 1991.

[9] H. Ney. A dynamic programmation algorithm for nonlinear smoothing. *Signal Processing*, 5(2):163–173, March 1983.

[10] A. V. Oppenheim. A speech analysis synthesis system based on homomorphic filtering. *Journal of Acoustical Society of America*, 45:458–465, 1969.

[11] S. Seneff. System to independently modify excitation and/or spectrum of speech waveform without explicit pitch exctraction. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-30(4):566–578, August 1982.