On the Use of Wavelets and Cepstrum Excitation for Pitch Determination in Real-Time

Fadoua Bahja, El Hassan Ibn Elhaj, Joseph Di Martino

On The Use of Wavelets and Cepstrum Excitation for Pitch Determination in Real-Time

Fadoua BAHJA*
[§] Laboratoire LRIT Unité associée au CNRST, URAC 29, Faculté des sciences
Rabat, Morocco

El Hassan Ibn Elhaj
[§] INPT
Madinat Al Irfane
Rabat, Morocco

Joseph Di Martino
[§] INRIA / LORIA
Université de Lorraine (UHP)
Vandoeuvre-lès-Nancy, France

***Abstract*—In the current paper, we propose a new pitch tracking technique based on a wavelet transform in the temporal domain. Our algorithm is designed to determine the pitch frequency of the speech signal using a simple voicing decision algorithm. The pitch period is extracted from the cepstrum excitation signal processed by a wavelet transform; then the pitch contour is refined by thresholding and correction algorithms without any post-processing. The results obtained show that the proposed algorithm provides very good pitch contours compared to those furnished by the Bagshaw database.**

***Keywords-component*—Wavelet, cepstrum excitation, real-time, pitch tracking.**

I. INTRODUCTION

Pitch detection plays an essential part on speech processing and has a large field of applications in speech related domains. For this reason, many techniques for tracking the pitch have been proposed.

Some important problems are encountered in pitch detection: the first one is the variation of the fundamental frequency F0 in time (therefore pitch tracking in real time is still difficult but desired), the second one is the appearance of sub-harmonics that false the detection; and the third one is the difficulty to realize the voiced/unvoiced decision on the pitch contours.

This paper proposes a new pitch tracking algorithm: the WCEPD algorithm (Wavelet and Cepstrum Excitation for Pitch Determination), which is based on the cepstrum excitation signal and the use of the Discrete Wavelet Transform (DWT) in temporal domain in order to try to solve these three problems.

The scope of wavelets [1, 2, 3] is large enough and varied: speech processing, image compression, signal and image denoising are examples of problems treated. In this paper we focus on its use in pitch tracking in real time for speech signals, using the DWT which is an efficient decomposition technique [4], in temporal domain.

To evaluate the performance of the proposed algorithm, we tested and compared the provided results using the Bagshaw database [5].

This article is organized as follows: section 2 describes the principle of the WCEPD algorithm; section 3 performs the various decisions concerning voicing and section 4 gives experimental results using the Bagshaw database; and finally, we conclude this paper in Section 5.

II. WAVELET AND CEPSTRUM EXCITATION FOR PITCH DETERMINATION

A. Concept

The main purpose of the WCEPD algorithm described in this paper is the determination of the pitch from the log-cepstrum excitation signal [6] with the use of DWT.

The main idea of our approach consists in extracting the pitch period from the cepstrum excitation signal in order to suppress the possible disruptive vocal tract effect.

The proposed pitch detection WCEPD has two main steps:

- In a first step, we start by multiplying the speech signal with a Hamming window of 1024 samples followed by a Fast Fourier Transform (FFT) and the modulus of this output signal. By calculating the log we get the Fourier log spectrum. By applying an Inverse Fast Fourier Transform (IFFT) we get the real log cepstrum related to the analyzed frame. From the real log-cepstrum, we eliminate the Oc = 27 first coefficients (which represent the vocal tract) and we obtain the log cepstrum excitation signal as shown in figure 1.

- In a second step, to detect the pitch period, a wavelet decomposition by DWT is carried out to level 3 on the log cepstrum excitation obtained before.
  Then, we search the local maximum peak index in the log cepstrum excitation signal from the 3 levels of the approximation coefficients, in order to extract the global peak index which represents the pitch (see figure 2).

To obtain the period pitch, a search is performed in order to find the highest value in the log cepstrum excitation signal, which is decomposed into 3 levels, which gives the peak related to the pitch frequency.

[§] This study has been realized in the framework of the INRIA Euro-Mediterranean 3+3 M09/02 Oesovox project with help of the European COADVISE- IRSES (FP7) program.

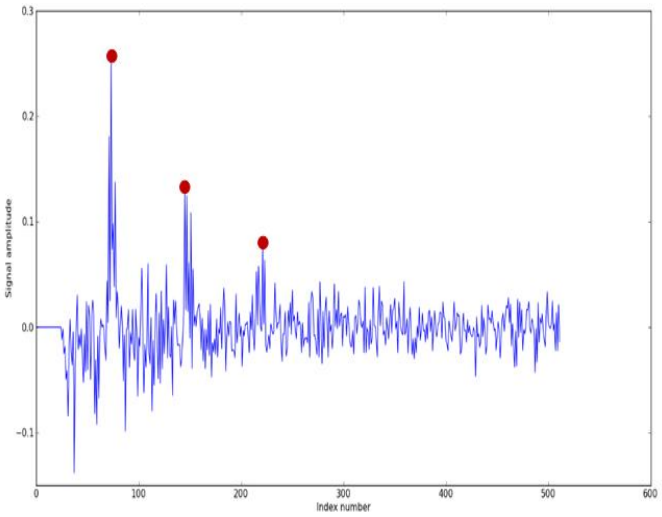

Figure 1. log cepstrum of an excitation signal for a female voice. The red dotted points are related to F0 and its harmonics.

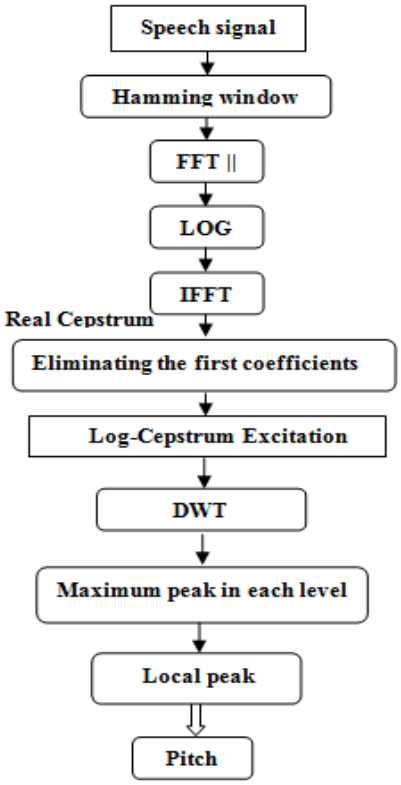

Figure 2. Flowshart of the proposed WCEPD algorithm.

According to the following formula 1, the maximum peak index is related to the pitch by:

$$Pitch = \frac{Sampling\ \ frequency}{Maximum\ \ peak\ index}\ Hz \qquad (1)$$

B. *The role of DWT*

The role of DWT is to attenuate the high frequencies and to divide the signal in sub-bands. It is designed by generalizing the filter bank concept [4]. The first step produces, in one level, from the log cepstrum excitation signal, two sets of coefficients: the approximation coefficients, and the detail coefficients. These vectors are obtained by convolving the log-cepstrum excitation with the low-pass filter for approximation, and with the high-pass filter for detail, followed by a downsampling (see figure 3). In our method, we use only the approximation coefficients in the DWT step. The same steps are repeated for the 3 levels.

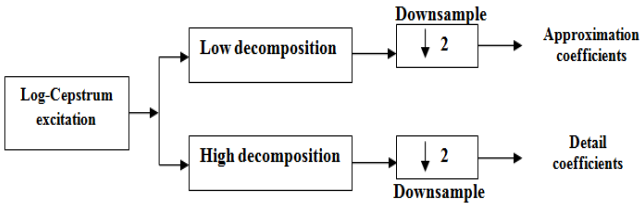

Figure 3. One-dimensional DWT for log-cepstrum excitation.

For the filter used, we choose the Haar filter for its simplicity and efficiency.

III. THE VOICING DECISION

In this section, we present a smart and easy technique for voicing decision which respects real time and uses only the preceding frames [8].

A. *Thresholding*

The voicing decision follows two main criteria explained in the following:

- When the region is voiced, the indexes of the maximum peaks of the cepstrum signals vary slowly. Formula 2 gives the first thresholding concerning the quantity S.

$$S = \sum_{k=0}^{L-1} \left| Index[j-k] - Index[j-k-1] \right| \qquad (2)$$

  Where:

  - L is the number of calculated frames (for our experiment L = 8);
  - *Index* is the maximum peak index of the resulting signal;
  - j is the index of the temporal frame analyzed.

  When S, has a small value, the analyzed frame has a high probability of being voiced. On the contrary, when the value of S is great, the analyzed frame has a significant probability of being unvoiced.

- When the log of the energy of the speech signal (windowed by hamming window) is below a threshold (experimentally 76 dB) the analyzed frame concerns either a silence or an unvoiced region (Formula 3).

$$Ener(x) = 10\log_{10}\left(\sum_{i=0}^{N-1} x_i^2\right) \qquad (3)$$

B. *Corrections*

To correct the pitch contour from parasitic peaks and valleys (see figure 4), we use the following technique:

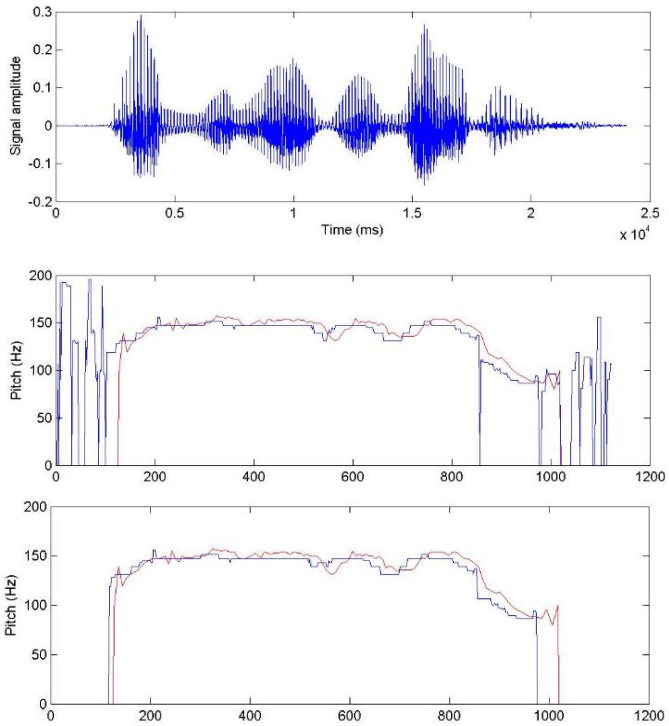

Figure 4. (above) Input signal for a male voice; (middle) the pitch estimated before correction by the WCEPD algorithm (blue curve) and the reference contour extracted in the database (red curve); (below) the pitch estimated after correction.

- Starting with isolated peaks: the pitch peak is eliminated if its duration is bellow 18ms.
- Concerning valleys: we rebuilt the pitch contour linearly if the duration of the valley is below 18ms.

Figure 5 and figure 6 represent the pitch estimated after correction by the WCEPD algorithm using the criteria explained in the sequel, respectively for a male and a female voice.

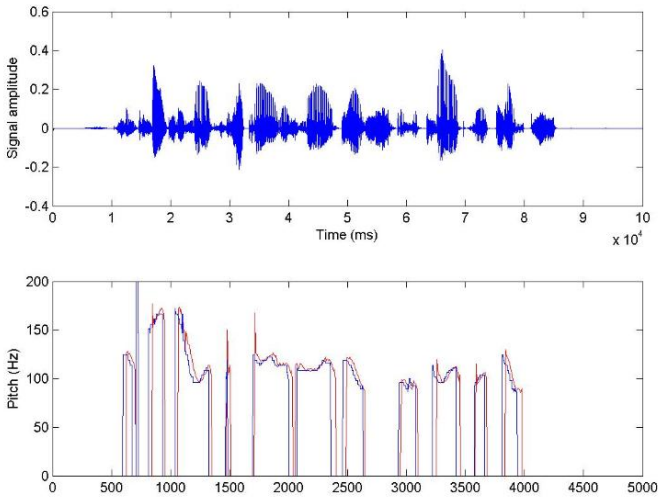

Figure 5. The pitch estimated after correction by the WCEPD algorithm (blue curve) and the reference contour extracted in the database (red curve) for a male voice.

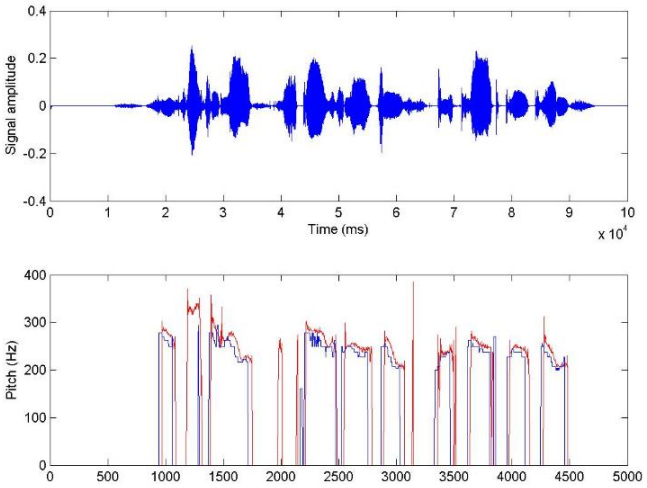

Figure 6. Example of signals obtained for a female voice.

IV. EXPERIMENTAL RESULTS

To evaluate the WCEPD algorithm, we used the corpora of BAGSHAW. Recordings are made simultaneously by a microphone and a laryngograph in an acoustically isolated room. The signal is used to calculate the reference contour of the pitch. The corpora contain 50 sentences; each one spoken in English by a male and a female speaker.

The sampling frequency is 20 kHz. The length of the analysis window is 1024 samples (51.2 ms).

We choose a small shift between two consecutive windows, where the shift is 40 samples (2 ms) in order to obtain a precise voiced/unvoiced decision.

The real time domain is estimated in term of the latency which represents the time between signal onset and pitch determination. For our experiments, the latency is 18ms.

The results obtained by the proposed algorithm (Table 1 and Table 2) are very promising.

Experiments were carried out to compare our algorithm with 9 PDAs, listed below, using the same database.

- Cepstrum pitch determination (CPD) [9]
- Feature-based pitch tracking (FBPT) [10]
- Harmonic product spectrum (HPS) pitch determination [9,10]
- Super resolution pitch determination(SRPD)[13]
- Enhanced SRPD (eSRPD)[5]
- Circular autocorrelation of the temporal excitation (CATE) [14]
- Adaptative least squares (ALS) estimation [15]
- Enhanced CATE (eCATE) [7],
- Modified eCATE (Ecate+) [8].

The “gross low error” concerning the male voice is the lowest among all the pitch determination algorithms (PDAs) listed in Table 1 and Table 2, more of which do not operate in real-time.

TABLE I. PDA RESULTS FOR THE MALE CORPUS

		PDA	Unvoiced error (%)	Voiced error (%)	Gross error		Abs-deviation	
					High (%)	Low (%)	Mean (Hz)	S. dev (Hz)
Male voice	Non real-time	CPD	18.11	19.89	4.09	0.64	2.94	3.60
		FBPT	3.73	13.90	1.27	0.64	1.86	2.89
		HPS	14.11	7.07	5.34	28.15	3.25	3.21
		SRPD	4.05	15.78	0.62	2.01	1.78	2.46
		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
	Real-time	ALS	4.20	11.00	0.05	0.20	---	3.24
		eCATE	5.40	11.25	0.05	0.20	1.63	2.28
		eCATE+	5.38	11.40	0.05	0.11	1.63	2.29
		WCEPD	7.35	12.19	0.41	0.06	3.15	2.84

TABLE II. PDA RESULTS FOR THE FEMALE CORPUS

		PDA	Unvoiced error (%)	Voiced error (%)	Gross error		Abs-deviation	
					High (%)	Low (%)	Mean (Hz)	S. dev (Hz)
Female voice	Non real-time	CPD	31.53	22.22	0.61	3.97	6.39	7.61
		FBPT	3.61	12.16	0.60	3.55	5.40	7.03
		HPS	19.1	21.06	0.46	1.61	4.59	5.31
		SRPD	2.35	12.16	0.39	5.56	4.14	5.51
		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
	Real-time	ALS	4.92	5.58	0.33	0.04	---	6.91
		eCATE	4.33	8.80	0.39	0.45	4.27	5.52
		eCATE+	4.92	7.99	0.41	0.41	4.31	5.60
		WCEPD	12.58	11.76	0.54	0.22	10.86	7.29

Our algorithm yields the pitch period accurately and furthermore concerning the classification errors related to the Bagshaw database, WCEPD allowed us to obtain the lowest "gross error low" rate, with 0.06 % for the male corpus, without any post-processing.

V. CONCLUSIONS

We have presented in this article a new wavelet algorithm WCEPD using a cepstrum analysis for F0 detection.

In order to improve its performance, a voicing decision method easy to implement, fast and robust has been implemented.

The experimental results obtained show that our method is more, or as efficient as, most other classical pitch determination algorithms.

Our algorithm yields the pitch period accurately and furthermore concerning the classification errors related to the Bagshaw database, WCEPD allowed us to obtain the lowest "gross error low" rate (for the male corpus) without any post-processing.

The principal aim of our approach is to determine the elected pitch under different multiresolution levels in the cepstrum excitation signal when the F0 and its harmonics appear clearly. Also, our contribution consists in correcting the pitch period.

Furthermore, the WCEPD algorithm respects real time with a very low latency.

REFERENCES

[1] H. Weiping, W. Xiuxin and P. Gomez, "Robust Pitch Extraction in Pathological Voice Based on Wavelet and Cepstrum", in *Proc.* EUSIPCO 2004, pp. 297–300, Vienna, Austria, September 2004.

[2] M. I. Abdalla and H. S. Ali, "Wavelet-Based Mel-Frequency Cepstral Coefficients For Speaker Identification Using Hidden Markov Models," in Journal of Telecommunications, Volume 1, Issue 2, pp16-21, March 2010.

[3] K. L. Neville and Z. M. Hussain, "Effects of wavelet compression of speech on its Mel-Cepstral coefficients," in International Conference on Communication, Computer and Power (ICCCP'09) MUSCAT, FEBRUARY 15-18, 2009.

[4] S. Mallat, "A wavelet tour of signal processing", second edition. Academic Press, 1999.

[5] P.C. Bagshaw, S. M. Hiller and M.A. Jack, "Enhanced Pitch Tracking and the Processing of F0 Contours for Computer aided Intonation Teaching", in *Proc.* ECST 1993, volume 2, pp. 1000-100, Berlin, September 1993.

[6] A. V. Oppenheim and R. W. Schafer, "Homomorphic Analysis of Speech", in IEEE Trans, Audio Electroacoust. AU-16, N. 2, pp. 221-226, 1968.

[7] F. Bahja, J. Di Martino and E. Ibn Elhaj, "Real-Time Pitch Tracking using the eCATE Algorithm" in ISIVC 2010, Rabat, Morocco, 30 September /2 October 2010.

[8] F. Bahja, J. Di Martino, E. Ibn Elhaj and D. Aboutajdine, "An improvement of the eCATE algorithm for F0 detection", in ISCIT 2010, Tokyo, Japan, 2010.

[9] A. M. Noll, "Cepstrum Pitch Determination", in Journal of the Acoustic Society of America, 41 (2), pp. 293-309, 1967.

[10] M. S. Philips, "A Feature-based Time Domain Pitch Tracker", in Journal of the Acoustical Society of America, 77:S9-S10, 1985.

[11] A. M. Noll, "Pitch Determination of Human Speech by the Harmonic Product Spectrum, the Harmonic Sum Spectrum, and a Maximum Likelihood Estimate", in *Proc.* Symposium on Computer Processing in Communication, pp. 779–798, April 1969.

[12] M. R. Schroeder, "Period Histogram and Product Spectrum: New Methods for Fundamental Frequency Measurement", in Journal of the Acoustical Society of America, 43(4):829–834, 1968.

[13] Y. Medan, E. Yair and D. Chazan, "Super Resolution Pitch Determination of Speech Signals", in IEEE Transactions on Signal Processing, ASSP-39(1), pp. 40-48, January 1991.

[14] J. Di Martino and Y. Laprie, "An Efficient F0 Determination Algorithm based on the Implicit Calculation of the Autocorrelation of the Temporal Excitation Signal", in 6th EUROSPEECH 1999, Budapest, Hungary, 1999.

[15] K. Saul, D. Lee, C. Isbell and Y. LeCun, "Real Time Voice Processing with Audiovisual Feedback: Toward Autonomous Agents with Perfect Pitch", in NIPS 2002.