A New Method for Pitch Tracking and Voicing Decision Based on Spectral Multi-scale Analysis

Mohamed Anouar Ben Messaoud anouar.benmessaoud@yahoo.fr
Electrical Engineering Department
National School of Engineers of Tunis
Le Belvédère BP. 37, 1002 Tunis, Tunisia

Aïcha Bouzid bouzidacha@yahoo.fr
Electrical Engineering Department
National School of Engineers of Tunis
Le Belvédère BP. 37, 1002 Tunis, Tunisia

Noureddine Ellouze n.ellouze@enit.rnu.tn
Electrical Engineering Department
National School of Engineers of Tunis
Le Belvédère BP. 37, 1002 Tunis, Tunisia

Abstract

This paper proposes a new method for voicing detection and pitch estimation. This method is based on the spectral analysis of the speech multi-scale product. The multi-scale product (MP) consists of making the product of the speech signal wavelet transform coefficients. The wavelet used is the quadratic spline function. The spectrum of the multi-scale product analysis reveals rays corresponding to the fundamental frequency and its harmonics. We evaluate our approach on the Keele University database. The experimental results show the effectiveness of our method comparatively to the state-of-the-art algorithms.

Keywords: Speech, Wavelet transform, Multi-scale Product, Pitch, Voicing detection.

1. INTRODUCTION

Pre-processing of speech signal is very crucial in the applications where silence or background noise is completely undesirable. Applications like speech and speaker recognition [1] needs efficient feature extraction techniques from speech signal where most of the voiced part contains speech or speaker specific attributes. Silence removal is a well known technique adopted for many years for this and also for dimensionality reduction in speech that facilitates the system to be computationally more efficient. This type of classification of speech into voiced or silence/unvoiced sounds [2] finds other applications mainly in fundamental frequency estimation, formant extraction or syllable marking and so on.

More over, the fundamental frequency is an important parameter in the speech analysis and synthesis. It plays an eminent role in the speech production and perception. In application areas such as speech enhancement, analysis and prosody modeling, low-bit rate coding, and speaker recognition, a reliable pitch estimation is required [3].

A wide variety of sophisticated voicing classification and pitch detection algorithms have been proposed in the speech processing literature [4], [5], [6], [7], [8] and [13].

The voicing decision and pitch estimation from speech signal only are basically done by relying on different types of speech transformation. This transformation can be operated following three domains:
The first approach works in the time domain. The common transformation is the autocorrelation function like the YIN algorithm and the Praat Software application [9], [10] and [11].
The second approach works in the frequency domain. The frequently used transformation is the spectrum [12] and [13].
The third approach combines both time and frequency domains, using the Short Time Fourier Transform (STFT) or the Wavelet Transform (WT) [14].

In this paper, we propose and evaluate a new algorithm for voicing classification and pitch determination operated on a clean speech signal. We are motivated by the work developed in [15] and [16], where the multi-scale product-based approach constitutes an efficient method for glottal closure instant detection. These instants delimit the pitch period.

This paper is organized as follows: Section 2 reminds some properties of the continuous wavelet transform and the multi-scale product method for edge detection. In section 3, we detail our approach for voicing decision and pitch estimation. In section 4, experimental results are presented using the Keele University database. Finally, we conclude this work.

2. MULTI-SCALE ANALYSIS

Wavelet Transform [17], [18] is introduced as an alternative technique for analyzing non stationary signal. It provides a new way for representing the signal into well-behaved expression that yields useful properties. The wavelet is a square integrable function well localised in time and frequency, from which we can extract all basis functions by time shifting and scaling.

Dyadic wavelet transform, is a particular case of continuous wavelet transform when the scale parameter is discretized along the dyadic grid (2^j), $j \in Z$.
The wavelet transform can be used for various applications as edge detection, noise reduction and parameter estimation. When the mother wavelet function is the nth derivative of a smoothing function, it acts as a differential operator. The number of wavelet vanishing moments gives the order of the differentiation. For an appropriately chosen wavelet, the wavelet transform modulus maxima denote the points of sharp variations of the signal [19]. Wavelet transform which is the first derivative of a smoothing function is proved to be convenient for discontinuity detection in a signal.
The wavelet transform is a multi-scale analysis which has been shown to be very well suited for speech processing in many applications as glottal closure instant (GCI) detection, pitch estimation [20], speech enhancement and recognition and so on.

To improve edge detection by wavelet transform, we use a non linear combination of wavelet transform coefficients. The multi-scale product (MP) consists of making the product of wavelet transform coefficients of the function f(n) at some successive dyadic scales as follows [21]

$$p(n) = \prod_{j=j_0}^{j=j_L} w_{2^j} f(n) \tag{1}$$

Where $w_{2^j} f(n)$ is the wavelet transform of the function f at scale 2^j. This is distinctly a non linear function of the input time series f(n).

Singularities produce cross-scale peaks in wavelet transform coefficients, which are reinforced in the product p(n). Although particular smoothing levels may not be optimal, the non linear combination tends to reinforce the peaks while suppressing spurious noisy peaks. The signal peaks align across scale for the first few scales, but not for all scales because increasing the amount of smoothing will spread the response and cause singularities separated in time to interact. Thus choosing scales too large will result in misaligned peaks in p(n). Odd number of terms in p(n) preserves the sign of maxima [22]. Choosing the product of three levels of wavelet decomposition is generally optimal and allows detection of small peaks.

This is intended to enhance multi-scale peaks due to edge, while suppressing noise, by exploiting the multi-scale correlation due to the presence of the desired signal. Bouzid et al. prove that the MP is very efficient for glottal closure and opening instants detection from speech signal only [16].

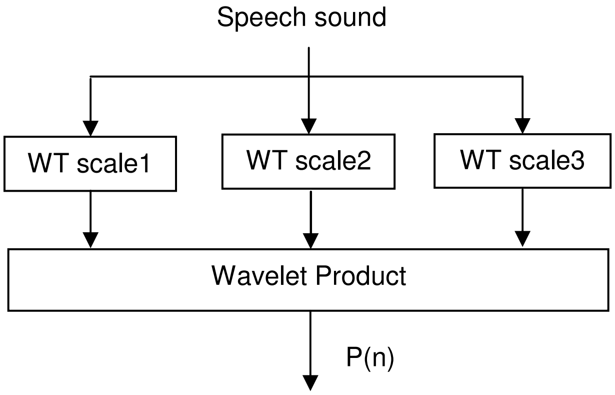

FIGURE 1: The Multi-scale Product scheme.

3. THE PROPOSED METHOD

We propose a new technique to localize voiced sounds with an estimation of the fundamental frequency in the case of a clean speech signal. The method is based on the spectral analysis of the speech multi-scale product (SPM).

Our method can be decomposed in four essential steps. The first step consists of computing the product of wavelet transform coefficients of the speech sound. The wavelet used in this multi-scale product analysis is the quadratic spline function at scales $s_1=2^{-1}$, $s_2=2^0$ and $s_3=2^1$.

The second step consists of calculating the fast Fourier transform (FFT) of the obtained signal over windows with a specific length of 4096 samples. In deed, the product is decomposed into frames of 1024 samples with an overlapping of 512 points at a sampling frequency of 20 kHz.

In fact, the product p[n] is divided into frames of N length by multiplication with a sliding analysis window w[n]:

$$p_w[n,i] = p[n]\,w[n - i\Delta n] \qquad (2)$$

Where i is the window index, and Δn the overlap. The weighting w[n] is assumed to be non zero in the interval [0, N-1]. The frame length value N is chosen in such a way that, on the one hand, the parameters to be measured remain constant and, on the other hand, there are enough samples of p[n] within the frame to guarantee reliable frequency parameter determination.

The third step consists of identifying voiced frames in a speech waveform. And the last step consists of giving an estimation of the pitch frequency for the detected voiced frames. Theses two steps will be detailed in the next subsections.

3.1 Voicing Decision

Figures 2 and 4 show the multi-scale product of the voiced speech signal and the unvoiced one respectively. For the first case, the MP has a periodic structure unlike the second case. The figure 3 shows the SMP corresponding to the voiced speech signal. However the figure 5 illustrates the

SMP corresponding to the unvoiced sound depicted in figure 4. We note clearly the difference between the two cases. So, a voicing detection approach can be derived.

After calculating the FFT of the speech MP in the ith frame, we localize all the peaks stored in the vector Pi. We eliminate ones that don’t belong to the following frequency range [F0min F0max]. If there is no peaks, the frame is declared unvoiced, else we calculate the distance separating two successive peak positions Dij=Pij+1-Pij constituting the Di vector elements. These elements are ranked in the growing order to compose the Ei vector. To make a voicing decision, we look for well defined groups constituted from the Ei vector.

The groups are sorted as follows:

If Ei1-Ei2<10, so Ei1 and Ei2 are in the same group Gi1 and we calculate Ei1-Ei3, else, Ei1 is in Gi1 and Ei2 is in Gi2. Then, we calculate Ei2-Ei3 and so on until reaching the last elements in the Ei vector.

Once the groups are formed, we look for their number Ni. If Ni=1, the ith frame is voiced. If Ni=2 and (card(Gi2)<2/3*card(Gi1)), the ith frame is also declared voiced, else, the frame is unvoiced.

The voicing decision diagram is given in the figure 6.

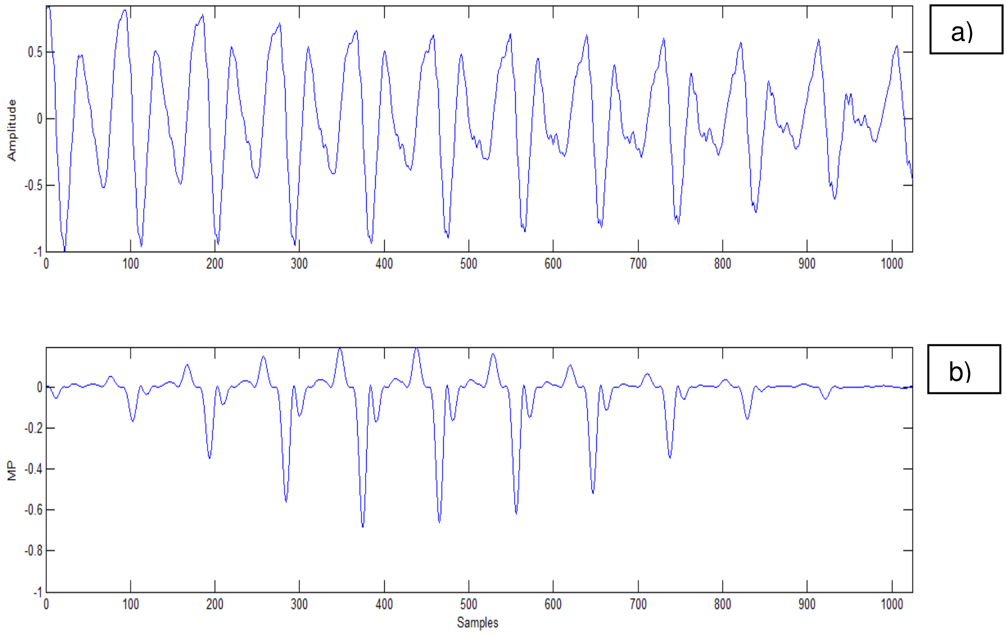

FIGURE 2:. a) Voiced speech of a female speaker. b) its Multi-scale Product.

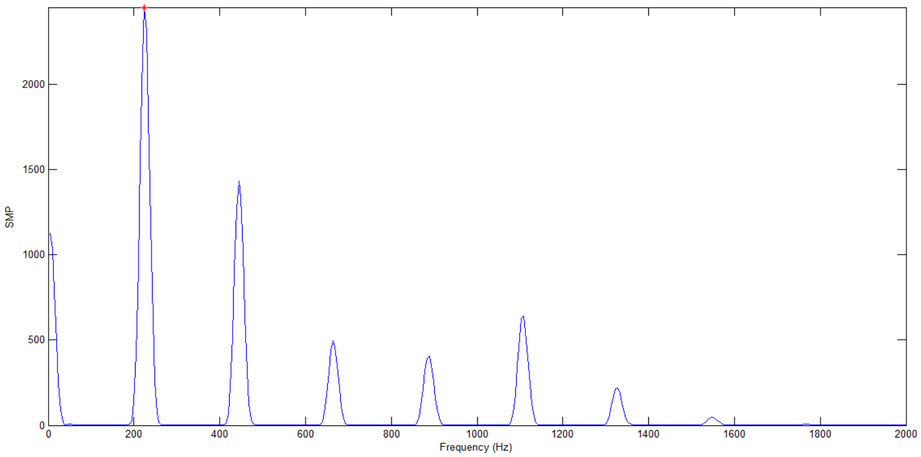

FIGURE 3: Spectral Multi-scale Product Analysis of the voiced speech signal 2(a).

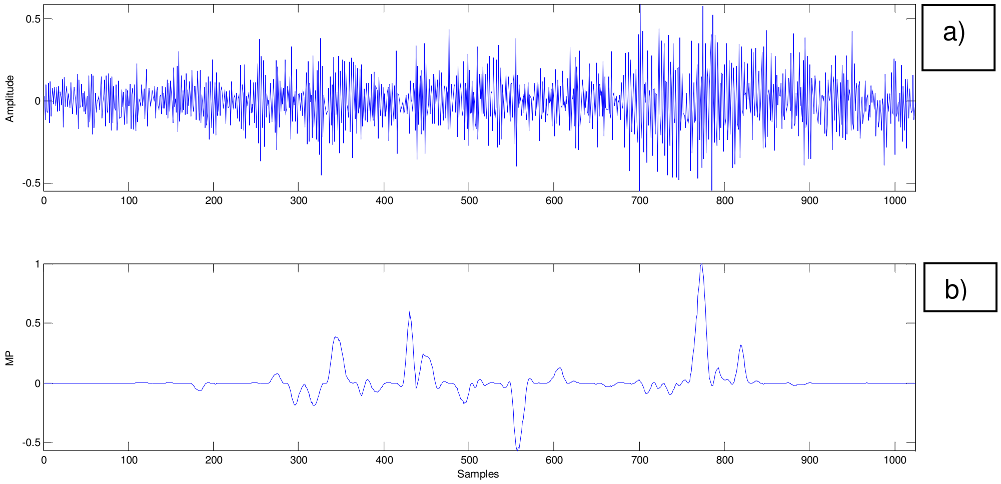

FIGURE 4:. a) Unvoiced speech of a female speaker. b) its Multi-scale Product.

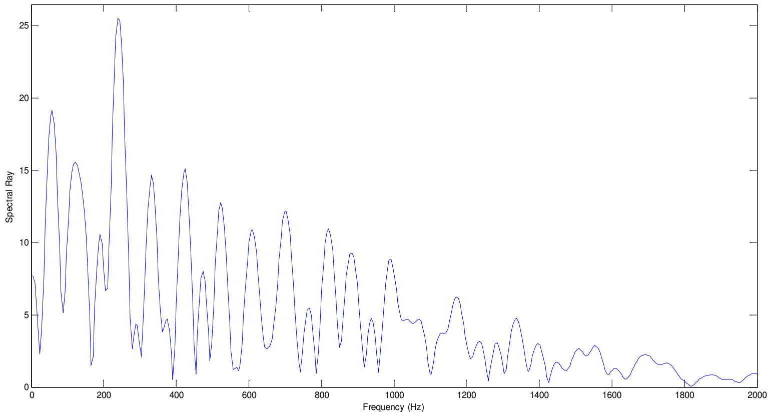

FIGURE 5: Spectral Multi-scale Product Analysis of the unvoiced speech signal 4(a).

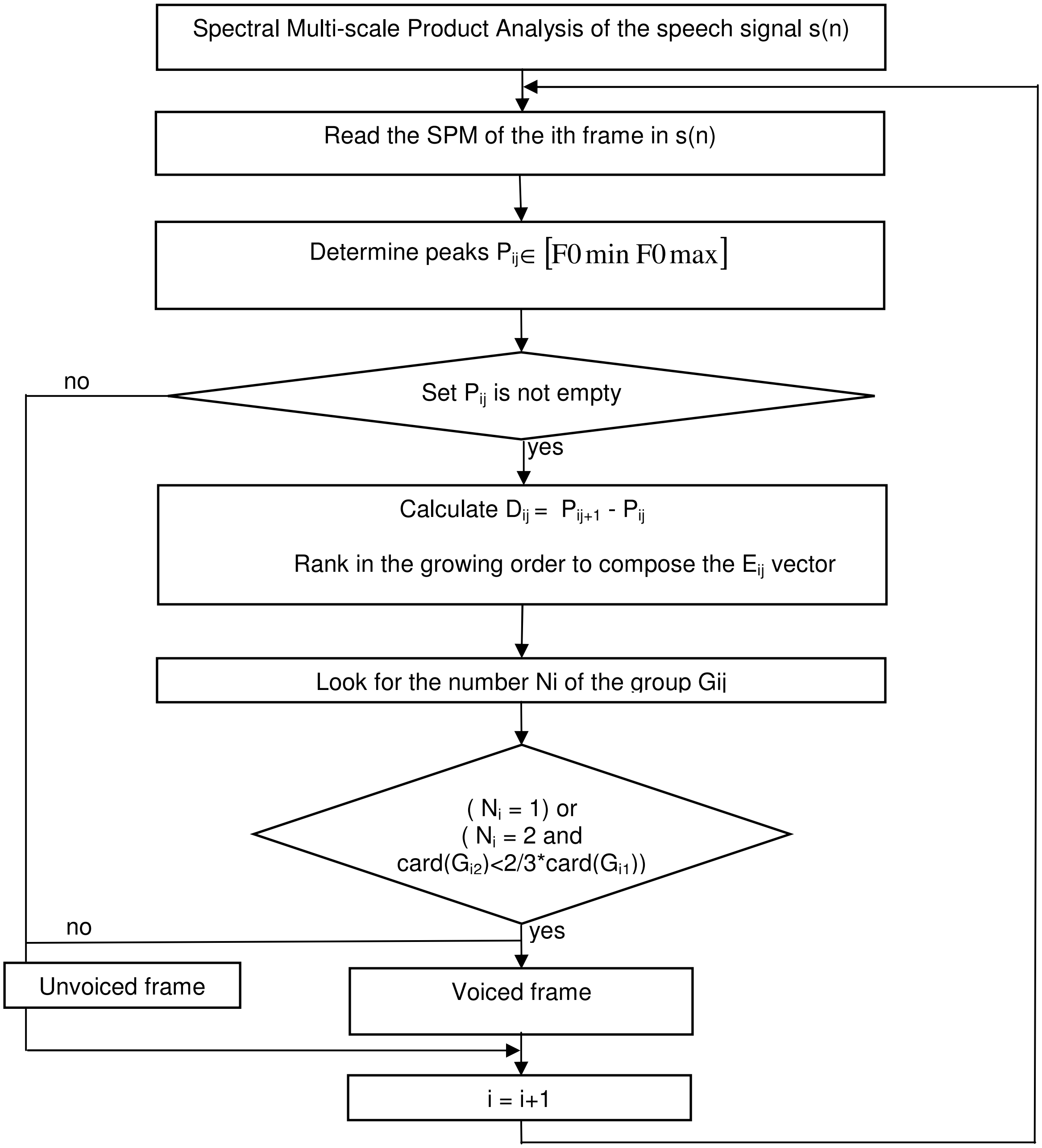
Spectral Multi-scale Product Analysis of the speech signal s(n)
Read the SPM of the ith frame in s(n)
Determine peaks $P_{ij} \in [F0\min\ F0\max]$
Set P_{ij} is not empty
no
yes
Calculate $D_{ij} = P_{ij+1} - P_{ij}$
Rank in the growing order to compose the E_{ij} vector
Look for the number Ni of the group Gij
(N_i = 1) or
(N_i = 2 and
card(G_{i2})<2/3*card(G_{i1}))
no
yes
Unvoiced frame
Voiced frame
i = i+1

FIGURE 6: Algorithm of the proposed voicing classification approach.

3.2 Pitch Estimation

Pitch estimation is ensured on the ith voiced frame detected by the proposed approach. The fundamental frequency is the first element in the group Gi1 described in the previous subsection.

4. Experiments and Results

To evaluate the performance of our algorithm, we use the Keele pitch reference database [19]. This database consists of speech signals of five male and five female English speakers each reading the same phonetically balanced text with varying duration between about 30 and 40 seconds. The Keele database includes reference files containing a voiced–unvoiced segmentation and a pitch estimation of 25.6 ms segments with 10 ms overlapping. The reference files also mark uncertain pitch and voicing decisions. The reference pitch estimation is based on a simultaneously recorded signal of a laryngograph. Uncertain frames are labelled using a negative flag.

For the evaluation of the voicing classification approach, we calculate the error decision probabilities that comprises unvoiced frames detected as voiced,and voiced frames detected as unvoiced as proposed in [9]. Table 1 reports evaluation results for voicing classification of the proposed method in a clean environment. We compare our method to other state-of-the-art algorithms [8], [24], [25] and [26] that are based on the same reference database. As can be seen, our method yields very good results in comparison with well known approaches with the lowest V-UV rate of 2.3%.

Methods	**V-UV (%)**
Proposed Method	*2.3*
RAPT [8]	3.2
NMF [24]	7.7
MLS [25]	7.0
Seg-HMM [26]	8.4

TABLE 1: Performance comparison of some methods for voicing classification.

For pitch estimaton and according to Rabiner [27], the gross pitch error (GPER) denotes the percentage of frames at which the estimation and the reference pitch differ by more than 20%.

Table 2 lists the GPER of our proposed approach compared to others as PRAAT, YIN, and CEPSTRUM for male and female speakers and all the Keele database.
As can be seen, our approach yields good results encouraging us to use it in other hard environments. In fact, the SMP method shows a low GPE rate of 0.75% for all the database.

Methods	**Cep**	**PRAAT**	**YIN**	***Proposed***
	GPE (%)	GPE (%)	GPE (%)	GPE (%)
Female Speakers	4.2	3.3	1.2	*0.4*
Male Speakers	3.7	2.9	3.5	*1.1*
Total	3.95	3.1	2.35	*0.75*

TABLE 2: GPER for pitch estimation using Keele University database.

5. CONSLUSION

In this work, we propose a novel voicing decision and pitch estimation algorithm. This algorithm is based on the spectral analysis of the multi-scale product made by multiplying the wavelet transform coefficients of the speech signal.

The proposed approach can be summarised in four essential steps. First, we make the product of the speech wavelet transform coefficients at three successive dyadic scales (The wavelet is the quadratic spline function with a support of 0.8 ms). Second, we compute the short time Fourier transform of the speech multi-scale product. Thirdly, we select the entire peaks found in the frame spectrum. These peaks are gathered satisfying some criteria. Consequently a decision is made concerning the voicing state and a pitch estimation is given. The experimental results show the efficiency of our approach for clean speech in comparison with the state-of-the-art algorithms.

6. REFERENCES

1. J.P. Campbell. “Speaker Recognition : A Tutorial”. In Proceedings of the IEEE, 85(9): 1437--1462, 1997

2. A. Martin, D. Charlet and L. Mauuary. “Robust Speech/ Non-speech Detection Using LDA Applied to MFCC”. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1: 237--240, 2001

3. D. O. Shaughnessy. “Speech communications: human and machine”. IEEE Press, NY, second edition, (2000)

4. D.G. Childers, M. Hahn and J.N. Larar. “Silence and Voiced/Unvoiced/Mixed Excitation Classification of Speech”. IEEE Trans. On Acoust., Speech , Signal Process, 37(11):1771--1774, 1989

5. L. Liao and M. Gregory. “Algorithms for Speech Classification”. In Proceedings of the 5th ISSPA, Brisbane, 1999

6. W. J. Hess. “Pitch and voicing determination”, Marcel Dekker, Inc., pp. 3-48 (1992)

7. 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 Proceedings of the 3rd European Conference on Speech Communication and Technology, 1993

8. D. Talkin. “A robust algorithm for pitch tracking (RAPT)”. In Speech Coding and Synthesis, W. B. Kleijn and K. K. Paliwal, Eds.,Elsevier Science, pp. 497-518 (1995)

9. L. Rabiner. “On the use of autocorrelation analysis for pitch detection”. IEEE Trans. Acoust., Speech, Signal Processing, 25(1): 24-33, 1977

10. D. A. Krubsack and R. J. Niederjohn. “An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech”. IEEE Trans. Acoust., Speech, Signal Processing, 39(1): 319-329, 1991
11. A. Cheveigné. “YIN, a fundamental frequency estimator for speech and music”. Journal of the Acoustical Society of America, 111(4):1917-1930, 2002

12. P. Boersma. “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound”. In Proceedings of the Institute of Phonetic Sciences, Amsterdam, 1993

13. A. M. Noll. “Cepstrum pitch determination”. J. Acoust. SOC. Amer., 41: 293-309, 1967

14. T. Shimamura and H. Takagi. “Noise-Robust Fundamental Frequency Extraction Method Based on Exponentiated Band-Limited Amplitude Spectrum”. In The 47th IEEE International Midwest Symposium on Circuits and Systems, 2004

15. A. Bouzid and N. Ellouze. “Electroglottographic measures based on GCI and GOI detection using multiscale product”, International journal of computers, communications and control, 3(1): 21-32, 2008

16. A. Bouzid and N. Ellouze. “Open Quotient Measurements Based on Multiscale Product of Speech Signal Wavelet Transform”, Research Letter in Signal Processing, 7: 1687-6911, 2008

17. C. S. Burrus, R. A. Gopinath and H. Guo. “Introduction to Wavelets and Wavelet Transform”, A Primer. Prentice Hall, (1998)

18. S. Mallat. “A Wavelet Tour of Signal Processing”, Academic Press, second edition, (1999)

19. Z. Berman and J. S. Baras. “Properties of the multiscale maxima and zero-crossings representations”, IEEE Trans.on Signal Processing, 42(1):3216-3231, 1993

20. S. Kadambe and G. Faye Boudreaux-Bartels. “Application of the Wavelet Transform for Pitch Detection of Speech Signals”. IEEE Trans. on Info. Theory, 38: 917-924, 1992

21. B. M. Sadler and A. Swami. “Analysis of multi-scale products for step detection and estimation”. IEEE Trans. Inform. Theory, 1043-1051, 1999

22. B. M. Sadler, T. Pham and L. C. Sadler. “Optimal and wavelet-based shock wave detection and estimation”. Journal of the Acoustical Society of America, 104: 955-963, 1998

23. G. Meyer, F. Plante and W. A. Ainsworth. “A pitch extraction reference database”. EUROSPEECH,1995

24. F. Sha and L. K. Saul. “Real-time pitch determination of one or more voices by nonnegative matrix factorization”, L. K. Saul, Y. Weiss, and L. Bottou, Eds., MIT Press, pp. 1233-1240 (2005)

25. F. Sha, J. A. Burgoyne and L. K. Saul. “Multiband statistical learning for F0 estimation in speech”. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Canada, 2004

26. K. Achan, S. Roweis, A. Hertzmann and B. Frey. “A segment-based probabilistic generative model of speech”. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2005

27. L. R. Rabiner, M. J. Cheng, A. H. Rosenberg and C. A. McGonegal. “A comparative performance study of several pitch detection algorithms”. IEEE Trans. Acoust., Speech, Signal Processing, 24(5): 399-417, 1976