An Efficient Method for Fundamental Frequency Determination of Noisy Speech

Mohamed Anouar Ben Messaoud[1,2], Aïcha Bouzid[1], and Noureddine Ellouze[1]

[1] Université de Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis, LR11ES17
Laboratoire du Signal, Images et Technologies de l'Information, 1002, Tunis Tunisie
[2] Université de Tunis El Manar, Faculté des Sciences de Tunis, 1002, Tunis Tunisie
{anouar.benmessaoud,bouzidacha}@yahoo.fr, n.ellouze@enit.rnu.tn

Abstract. In this paper, we present a fundamental frequency determination method dependent on the autocorrelation compression of the multi-scale product of speech signal. It is based on the multiplication of compressed copies of the original autocorrelation operated on the multi-scale product. The multi-scale product is based on realising the product of the speech wavelet transform coefficients at three successive dyadic scales. We use the quadratic spline wavelet function. We compress the autocorrelation of the multi-scale product a number of times by integer factors (downsampling). Hence, when the obtained functions are multiplied, we obtain a peak with a clear maximum corresponding to the fundamental frequency. We have evaluated our method on the Keele database. Experimental results show the effectiveness of our method presenting a good performance surpassing other algorithms. Besides, the proposed approach is robust in noisy environment.

Keywords: Speech, pitch estimation, multi-scale product, autocorrelation, compression analysis.

1 Introduction

The fundamental frequency extraction is one of the most crucial tasks in speech processing. Pitch is used for speech in many applications including determination of emotional characteristics of speech, speaker recognition systems, and aids to the handicapped. Because of its importance, many solutions to this problem have been proposed [1]. All of the proposed schemes have their limitations due to the wide range of applications, and operating environments. Thus various methods for pitch determination have been developed and a comprehensive review of these methods can be found in [2], [3]. However, due to the non-stationarity and quasi-periodicity of the speech signal, the development of more robust pitch determination algorithms still remains an open problem.

Most of the subsequent wavelet-based Pitch Detection Algorithms (PDAs) are originally inspired by the work presented by Kadambe and al [4].

There are two important issues which need to be improved in the PDAs. First, we show the efficacy of a PDA at the beginning of a vowel. Second, we obtain a robust PDA in a noisy environment.

T. Drugman and T. Dutoit (Eds.): NOLISP 2013, LNAI 7911, pp. 33–41, 2013.

We present an approach for estimation and detection of the pitch, extracted from speech signals, in this paper. Our proposed algorithm operates an autocorrelation compression on the voiced speech multi-scale product analysis. This analysis produces one peak corresponding to the fundamental frequency F_0.

The evaluation of the PDAs is an indispensable stage. Eventually, evaluating a pitch detection algorithm means simultaneously evaluating the Gross Pitch Error and the Root Mean Square Error.

The paper is presented as follows. After the introduction, we present our approach based on the multi-scale product analysis to provide the derived speech signal and the Autocorrelation Compression operated on the Multi-scale Product (ACMP) approach for the fundamental frequency estimation. Section 3 describes the pitch period estimation algorithm in clean and noisy voiced speech. In section 4, we give evaluation results and compare them with results of approaches for clean speech. Estimation results are also described for speech mixed with environmental noises at various SNR levels.

2 Proposed Approach

We propose an approach to estimate the fundamental frequency F_0 based on the Autocorrelation Compression (AC) of the voiced sound Multi-scale Product (MP). It can be decomposed into three essential stages, as shown in figure 1. The first stage consists of computing the product of the voiced speech wavelet transform coefficients (WTC) at successive scales. In accordance with the fast change of the instantaneous pitch, the wavelet used in this analysis is the quadratic spline function at scales $s_1=2^{-1}$, $s_2=2^0$ and $s_3=2^1$. It is a smooth function with property of derivative. The second stage consists of calculating the Autocorrelation Function (ACF) of the obtained signal. Indeed, the product is decomposed into frames of 512 samples with an overlapping of 50% points at a sampling frequency of 20 kHz. These two stages were described in our work reported by Ben Messaoud and al in [5]. The last stage consists of generating the functions obtained by the Autocorrelation compression and then multiplying them to provide a signal with a reinforced peak allowing an efficient estimation of the fundamental frequency value.

The product $p(\mathrm{n})$ of wavelet transforms coefficients of the function $f(\mathrm{n})$ at some successive dyadic scales is given as follows:

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

Where $W_{2^j} f(\mathrm{n})$ is the wavelet transform of the function f at scale 2^j.

For the second stage, the product $p(\mathrm{n})$ is split into frames of N length by multiplication with a hanning window $w[n]$:

$$p_w[n,i]=p[n]w[n-i\Delta n]. \tag{2}$$

Where i is the window index, and Δn the overlap.

Then, we calculate the short-term autocorrelation function of each weighted block $p_{wi}[n]$ as follows:

$$R_i(k)=\sum_{j=0}^{N-1} p_{wi}(j)\,p_{wi}(j+k)$$
$$ACF_i(k)=\frac{R_i(k)}{R_i(0)} \quad . \tag{3}$$

In the third stage, the Autocorrelation of the MP is compressed by integer factors (c = 1, 2, 3) and the obtained functions are multiplied. So the fundamental frequency F_0 became stronger.

The compression of each autocorrelation of the multi-scale product is described as follows:

$$ACMP_i(k)=\prod_{c=1}^{C-1}\left|\left(R_i(c*k)\right)\right| . \tag{4}$$

Where C is the number of harmonics to be considered.

The first peak in the original Autocorrelation Multi-scale Product (AMP) coincides with the second peak in the AMP compressed by a factor of two, which coincides with the third peak in the AMP compressed by a factor of three. Finally, we multiply to obtain one peak corresponding to pitch.

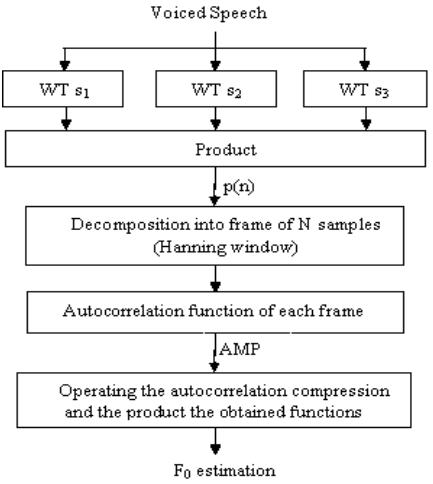

Fig. 1. Block diagram of the proposed approach for pitch estimation

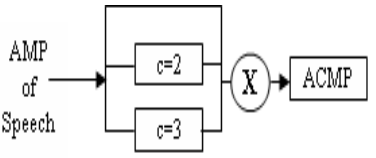

Fig. 2. Product of the compression of the speech AMP

The motivation for using the compression of the AMP is that for clean and noisy speech signals, multiplying the delay scale by integer factors should cause the peaks to coincide at F_0. Indeed the AMP of a voiced speech frame is zero between the peaks,

the product of compression functions cancels out all the peaks falling between two harmonics of the F_0. Thus, in general, finding the largest peak reflecting the product of the shifted AMP would mean finding the F_0. The product of the functions issued from the compression of the AMP is presented in figure 2.

3 Pitch Estimation Algorithm

3.1 Pitch Estimation in Clean Voiced Speech

Figure 3 shows a clean voiced speech signal followed by its MP. The MP has a periodic structure and reveals extrema according to the glottal closure and opening instants.

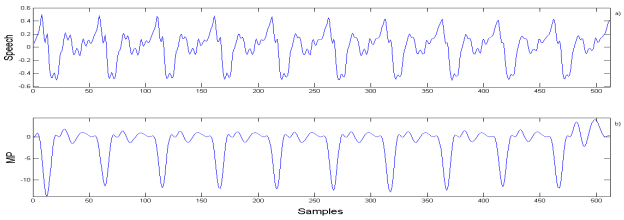

Fig. 3. a) Voiced clean speech. b) Its multi-scale product.

Figure 4.a) illustrates the multi-scale product autocorrelation function of a clean voiced speech signal. The calculated function is obviously periodic and has the same period as the MP. The obtained ACMP shows one peak occurring at the pitch period. The signals of the figures 4.b) and 4.c) represent respectively the AMP compressed with a factor c = 2 and c =3 of the voiced clean speech signal of the figure 3.a). The figure 4.d) corresponds to the multiplication of the functions issued from the compression of the AMP and shows one clear peak at the fundamental frequency.

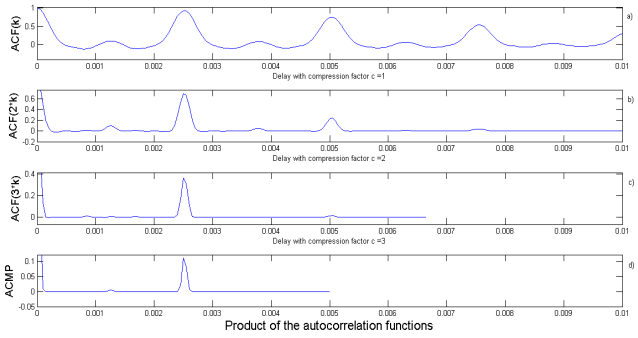

Fig. 4. ACMP of a voiced clean speech. a) Autocorrelation compression of MP with c=1. b) Autocorrelation compression of MP with c=2. c) Autocorrelation compression of MP with c=3. d) Autocorrelation functions multiplication.

Figure 5 treats the beginning of a voiced speech followed by its MP. Figure 6 shows the efficacy of the ACMP method for pitch estimation particularly at the beginning of a vowel. Signals represented in the figure 6.a), 6.b), 6.c) and 6.d) illustrate the compression of the autocorrelation multi-scale product of the speech depicted in 5.a).

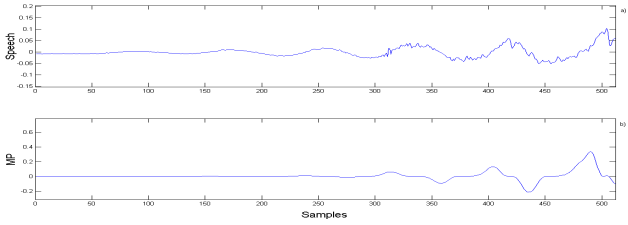

Fig. 5. a) The beginning of a vowel. b) Its multi-scale product.

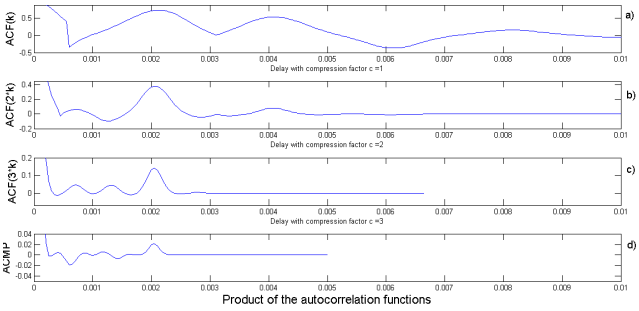

Fig. 6. ACMP of a vowel beginning. a) Autocorrelation compression of MP with c=1. b) Autocorrelation compression of MP with c=2. c) Autocorrelation compression of MP with c=3. d) Autocorrelation functions multiplication.

Figure 6 illustrates the efficacy of our approach for the fundamental frequency determination during a vowel onset. While the experimental results show that the other state of the art methods in literature give an F0 equals to zero at the beginning of vowel at this voiced region.

3.2 Pitch Estimation in a Noisy Environment

In this subsection, we try to show the robustness of our approach in the presence of the noise with high SNR levels.

Figure 7 depicts a noisy voiced speech signal with an SNR of -5 dB followed by its MP.

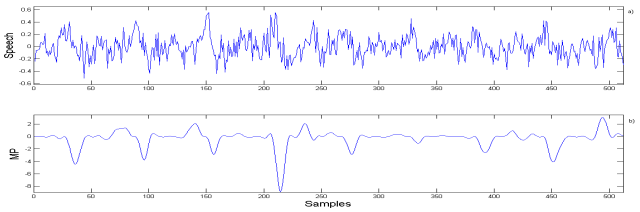

Fig. 7. a) Voiced speech signal corrupted by -5dB white noise. b) Its multi-scale product.

Figure 8 illustrates the ACMP approach. The MP in figure 7.b) lessens the noise effects leading to an autocorrelation function with clear maxima. The signal illustrated in figure 8.d) shows the autocorrelation compression of the MP with a peak giving the pitch estimation.

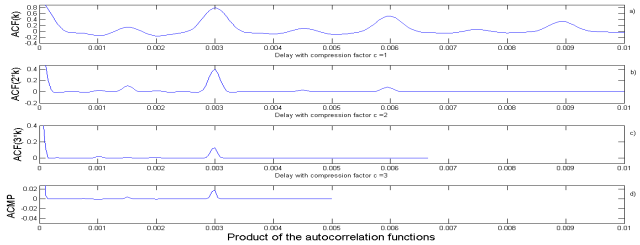

Fig. 8. ACMP of a voiced noisy speech. a) Autocorrelation compression of MP with c=1. b) Autocorrelation compression of MP with c=2. c) Autocorrelation compression of MP with c=3. d) Autocorrelation functions multiplication.

4 Results

4.1 Evaluation Databases

To evaluate the performance of our algorithm, we use the Keele pitch reference database [6]. The Keele database contains ten speakers sampling frequency of 20 kHz. It includes reference files containing a pitch estimation of 25.6 ms segments with 10 ms overlapping. The reference pitch estimation is based on a simultaneously recorded signal of a laryngograph.

We use common performance measures for comparing PDAs: The Gross Pitch Error (GPE) and the Root Mean Square Error (RMSE) [7]. The gross Pitch Error (GPE) is a standard error measure for the pitch tracking. It is defined as the percentage of estimated F_0 deviates from the referenced F_0 by more than 20% of voiced speech. The RMSE is defined as square root of the average squared estimation error with estimation errors which are smaller than the GPE threshold of 20 Hz. It should be noted that the pitch range of speech is 50 – 800 Hz.

4.2 Evaluation in a Clean Environment

For comparison, the four PDAs are based on the same reference database. The speech signal must be segmented into frames of 25.6 ms segments with 10 ms overlapping and is weighted by a Hanning window. The PDA's are only tested in the voiced frame.

Table 1 presents the evaluation results of the proposed approach (ACMP) for fundamental frequency determination in a clean environment and compared to the existed methods [8], [9], and [11].

Table 1. Pitch estimation Performance in a clean environment

Method	GPE (%)	RMSE (Hz)
ACMP	0.64	1.43
SWIPE'[8]	0.62	3.05
SMP [9]	0.75	2.41
NMF-PI [11]	0.93	2.84

The ACMP shows a reduced GPE rate of 0.64 % and the lowest RMSE of 1.43 Hz. It's obviously more accurate than the other methods.

4.3 Evaluation in a Noisy Environment

To test the robustness of our algorithm, we add various background noises (white, babble, and vehicle) at three SNR levels to the Keele database speech signals. For this, we use the noisex-92 database [10].

Table 2 presents the GPE of the ACMP, SMP, and NMF-PI methods in a noisy environment.

Table 2. Pitch estimation Performance of GPE in a noisy environment

		GPE (%)		
Type of noise	SNR level	ACMP	SMP [9]	NMF-PI [11]
White	5 dB	0.84	1.00	1.08
	0 dB	1.02	1.20	1.14
	-5 dB	1.09	1.40	1.32
Babble	5 dB	1.03	2.61	1.51
	0 dB	1.46	4.56	2.93
	-5 dB	1.67	7.62	5.10
Vehicle	5 dB	3.67	6.41	3.94
	0 dB	4.92	7.04	5.22
	-5 dB	5.80	8.98	8.74

As depicted in table 2, when the SNR level decreases, the ACMP algorithm remains robust even at -5dB.

Table 3 presents the RMSE of the ACMP, SMP and NMF-PI methods in a noisy environment.

Table 3. Pitch estimation Performance of RMSE in a noisy environment

		RMSE (Hz)		
Type of noise	SNR level	ACMP	SMP [9]	NMF-PI [11]
White	5 dB	2.45	3.23	4.63
	0 dB	2.86	3.73	4.84
	-5 dB	3.57	4.67	4.95
Babble	5 dB	3.67	4.28	3.81
	0 dB	4.59	4.93	4.92
	-5 dB	5.21	6.38	6.53
Vehicle	5 dB	2.08	5.67	4.53
	0 dB	3.36	7.89	4.60
	-5 dB	5.09	11.57	6.28

As depicted in table 3, the ACMP method presents the lowest RMSE values showing its convenience for pitch estimation in hard situations.

5 Conclusion

In this paper, we presented a pitch estimation method that relies on the compression of the autocorrelation applied on the speech multi-scale product. The proposed approach can be recapitulated in three essential stages. First, we have constituted the product of the voiced speech WTC at three successive dyadic scales (The wavelet is the quadratic spline function with a support of 0.8 ms). The voiced speech MP has a periodic and clean structure that matches well with the speech signal singularities and lessens the noise effects. Second, we have calculated the autocorrelation function of each weighted frame. Third, we have operated the compression of the obtained autocorrelation with various scales and their product.

The experimental results show the robustness of our approach for noisy speech, and its efficacy for clean speech in comparison with state-of-the-art algorithms. Future work concerns the extension of the proposed approach to estimate F_0 in monophonic music.

References

1. Hess, W.J.: Pitch Determination of Speech Signals, pp. 373–383. Springer (1983)
2. Shahnaz, C., Wang, W.P., Ahmad, M.O.: A spectral Matching Method for Pitch Estimation from Noise-corrupted Speech. In: IEEE International Midwest Symposium on Circuits and Systems, pp. 1413–1416. IEEE Press, Taiwan (2009)
3. Chu, C., Alwan, A.: A SAFE: A Statistical Approach to F0 Estimation Under Clean and Noisy Conditions. IEEE Trans. Audio, Speech and Language Process. 20, 933–944 (2012)
4. Kadambe, S., Boudreaux-Bartels, G.F.: Application of the Wavelet Transform for Pitch Detection of Speech Signals. IEEE Trans. Information Theory 38, 917–924 (1992)

5. Ben Messaoud, M.A., Bouzid, A., Ellouze, N.: Autocorrelation of the Speech Multi-scale Product for Voicing Decision and Pitch Estimation. Springer Cognitive Computation 2, 151–159 (2010)
6. Meyer, G., Plante, F., Ainsworth, W.A.: A Pitch Extraction Reference Database. In: 4th European Conference on Speech Communication and Technology EUROSPEECH 1995, Madrid, pp. 837–840 (1995)
7. Rabiner, L., Cheng, M., Rosenberg, A., McGonegal, C.: A comparative performance study of several pitch detection algorithms. IEEE Trans. on Acoustic, Speech, and Signal. Process. 24, 399–418 (1976)
8. Camacho, A.: SWIPE: a Sawtooth Waveform Inspired Pitch Estimator for Speech and Music, Ph.D. dissertation, Dept. Elect. Eng., Florida Univ., USA (2007)
9. Ben Messaoud, M.A., Bouzid, A., Ellouze, N.: Using Multi-scale Product Spectrum for Single and Multi-pitch Estimation. IET Signal Process. Journal 5, 344–355 (2011)
10. Varga, A.: Assessment for Automatic Speech Recognition: II. Noisex-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems. Speech Communication 12, 247–251 (1993)
11. Joho, D., Bennewitz, M., Behnke, S.: Pitch Estimation Using Models of Voiced Speech on Three Levels. In: 4th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, pp. 1077–1080. IEEE Press, Honolulu (2007)