ROBUST PITCH DETECTION ALGORITHM FOR NOISY SPEECH

[1]WIDED JLASSI, [2]AICHA BOUZID

Electrical Engineering Department National School of Engineers of Tunis Le Belvédère BP. 37, 1002 Tunis, Tunisia
Email: j_wided@yahoo.com, bouzidacha@yahoo.fr

Abstract—The fundamental frequency estimation from noisy speech is a challenging problem for speech and signal processing. In this paper,we propose a pitch estimation algorithm based on the analysis of the voiced speech cepstrum by the multiscale product (MP) using continuous wavelets (CAMP) for noisy environment.The CAMP method consists in framing the voiced signal, calculating the real cepstrum of each frame and then computing the MP of the obtained cepstrum. Our CAMP approach is evaluatedonthe COOK database.Experiments show that our approach substantially outperforms state-of-the-art

Index Terms—speech, real cepstrum, wavelet transform, Multi-scale Product, pitch, noise

I. INTRODUCTION

The fundamental frequency is considered as one of the most important prosodic parameters. Thus, a reliable pitch estimation is a challenging problem in speech processing. It plays an eminent role in the speech perception and production. It is criticalformany speech application such ascoding [2], analysis or speech recognition [3].

Speech signals are seldom available in pure form for speech processing applications, and are often corrupted by acoustic interference like background noise, simultaneous speech from another speaker, distortion etc.Therefore, the problem of estimating the fundamental frequency from noisy speech has attracted a variety of researchesover the last few decades but is still a challenging problem.Great number of algorithms have been proposed for estimating the fundamental frequency of speech signals [4][5][6]. Most of them provide good results on good quality speech signals, but their performance degrades when dealing noisy signals.

This paper is concerned with fundamental frequency estimation when speech is severely corrupted by noise or interfering speakers. A number of studies have investigated this problem. Gu[7] propose a speech separation via frequency bin nonlinear adaptive filtering.Tolonen [8] propose a model for multipitch and periodicity analysis of complex audio signals. The model essentially divides the signal into two channels, computes a "generalized" autocorrelation of the low-channel signal and of the envelope of the high-channel signal, and sums the autocorrelation functions. The summary autocorrelation function (SACF) is further processed to obtain an enhanced SACF (ESACF).

In this paper, we propose and evaluate a new algorithm for pitch estimation operated in corrupted speech. Our algorithm is based on the MP of the voiced speech cepstrum using a wavelet having two vanishing moment.Our approach is compared toother states of arts.

The paper is organized as follows. The next section describes the proposed algorithm. Results and comparisons are given in Section 2, followed by a conclusion section.

II. THE PROPOSED METHOD

In this section, we present our approach for the fundamental frequency estimation.

A. Stage 1

The first stage consists in framing the voiced signal and multiplying each frame by a Hamming window. Then, we proceed to calculate the real cepstrum for each frame [9].

$$s(n) = e(n) * \theta(n) \quad (1)$$

Where s (n) is the speech signal, e (n) is the excitation signal and θ (n) is the contribution of the vocal tract.

The cepstral analysis consists in ensuring the deconvolution[10]. The real cepstrum is defined as:

$$c[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log |S(w)| e^{jnw}\, dw \quad (2)$$

Where:

$$S(w) = \sum_{n=-\infty}^{+\infty} s(n) e^{-jwn} \quad (3)$$

B. Stage 2

The second stage of the algorithm is to calculate the WT [11][12] [13] [14] and [15] of the cepstrum at three scales, operate the product of WT coefficient [16] and [17], and then detect the maxima, whose position gives the pitch estimation.

$$p(n) = \prod_{i=1}^{3} W_{s_i} c[n] \quad (4)$$

Where $W_{s_i}c[n]$ is the WT of the real cepstrum at the scales s_i.
This approach is applied using the second derivative of a Gaussian as a wavelet having two vanishing moments.

C. Stage 3

The third stage is to operatea post-processing phase [18].
The focus of the post-processing is to reduce the F_0 estimation error under noisy conditions in order to have more reliable pitch detection algorithm.
This smoothing phase consists in comparing each value of the F_0 by its previous and next one and if there is a difference bigger than 30 Hz, we consider that is a false measure and is replaced by the average of the last two ones.
The hamming window has a length of 2048 samples, whereas the frame shift is fixed to 200 samples. The three scales (s1, s2 and s3) are (0.5, 1 and 2).

Figures 1 and 2 shows the robustness of our method in the elimination of noise even when it is a second speaker

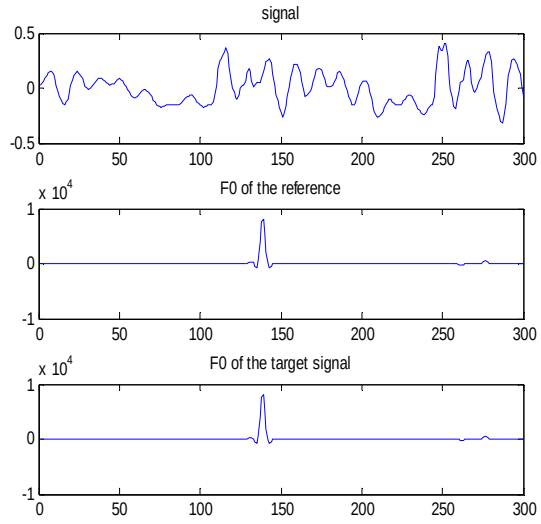


Figure 1the speech signal corrupted by an impulse noise, the reference signal and the CAMP

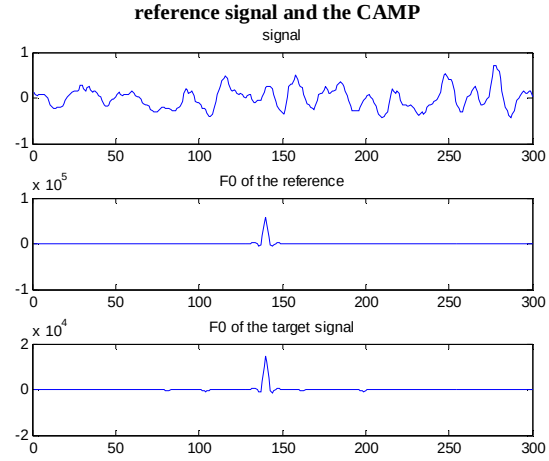


Figure 2the speech signal with competitive noise, the reference signal and the CAMP for speech corrupted by an impulse noise

III. EXPERIMENTS AND RESULTS

To evaluate the performance of our algorithm, we use the Cooke speech database[14] which contains 100 noisy utterances constructed by mixing 10 voiced speech utterances with 10 different types of interference signals. This corpus is commonly used for evaluating PDA performance. This database is a collection of composite sounds obtained by mixing ten male voiced speech signals with ten other signals representing a variety of sounds called interferences that can be classified into three categories:

- Interferences without pitch (N1: White noise and N2: Impulse noise),
- Interferences having a pitch quality (N0: Pure frequency of 1 Khz, N3: Cocktail party noise, N4: Rock music, N5: Siren and N6: Ringtone).
- Speech interferences (N7: Speech signal uttered by a woman 1, N8: Speech signal uttered by a man 2 and N9: Speech signal uttered by a woman 2).

For pitch estimation and according to Rabiner [4], the gross pitch error (GPER) denotes the percentage of frames at which the estimation and the reference pitch differ by more than 20%.
Tables 1, 2 and 3 summarizes the performance of the CAMP for the threecategories

SNR	N1: White noise	N2: Impulse noise
	GPE (%)	GPE (%)
V0	5,73	0
V1	4,23	0
V2	0	0
V3	3,52	0
V4	0	0
V5	1,62	0
V6	3,96	0
V7	4,05	0
V8	3,03	0
V9	1,44	0
Average	2,758	0

Table1The performances of the CAMP method in the presence of Interferences without pitch

For the first category, in the presence of impulse noise, all speakers have no gross pitch error. On the other hand, in an environment contaminated by a white noise our method is less efficient.

SNR	N0: Pure frequency	N3: Cocktail party noise	N4: Rock music	N5: Siren	N6: Ringtone
	GPE (%)	GPE (%)	GPE (%)	GPE (%)	GPE (%)
V0	0,8	8,19	5,64	4,83	0
V1	0	19	0	0	0
V2	0	0	0	0	0
V3	0	0	0	0	0
V4	0	2,47	0	0	0
V5	0	6,5	0	0	0
V6	0	3,22	0	0	0
V7	0	1,35	1,35	0	0
V8	0	0	0	0	0
V9	0	0	0	0	0
Average	0,08	4,073	0,699	0,483	0

Table 2The performances of the CAMP method in the presence of Interferences having a pitch quality

For the second category, most speakers have no gross pitch error.

RSB	N7: voix femme 1	N8: voix homme	N9: voix femme 2
	GPE	GPE	GPE
V0	6,25	0	8,88
V1	0	0	25
V2	0	0	16,6667
V3	0	0	3,125
V4	0	0	0
V5	0	0	2,08
V6	0	0	0
V7	3,57	0	0
V8	0	0	0
V9	0	0	0
Average	0,982	0	5,57517

Table 3The performances of the CAMP method in the presence of Speech interferences

For the third category, as shown in the table the proposed method remains efficient even in the presence of a second speaker.

COMPARISON

Our algorithm is compared to Gu [6] and TOLONEN [7] algorithms for the three categories.
The comparison of our method with other state of the art is tabulated in Table 1

	CAMP	GU	TOLONEN
Category 1	1.39	0.36	2.38
Category 2	1.18	2.10	4.53
Category 3	2.29	7.7	4.28

Table 4Comparison of our approach with existing methods using the Cooke basis for all categories

For the first category, in the presence of Interferences without pitch, the CAMP and Gu methods shows a perfect noise robustness while the Tolonen method is less efficient
For the second category, in the presence of Interferences having a pitch quality, CAMP method surpasses Gu and Tolonen methods. It shows the efficiency of our algorithm.
For the third category, in the presence of competitive noise, CAMP method has the lowest gross pitch error rate compared to Tolonen and Gu methods

CONCLUSION

Perfect Pitch estimation is a tricky problem in speech analysis particularly in noisy environment.In this paper, we propose a noise robust pitch detection algorithm based on the analysis of the voiced speech cepstrum by the multiscale product (MP) using continuous wavelets (CAMP).
The algorithm has been developed with the aim of estimating the fundamental frequency from speech corrupted by any kind of interference.
The proposed method showed better performance compared to other method in speech corrupted with different colored noise with the lower values of %GPEs. These results show that the proposed method is satisfactory to extract the fundamental frequency of voiced speech signals accurately in noisy conditions
The proposed method will be evaluated on a larger dataset in the future.

REFERENCES

[1] Spanias, A. S. 'Speech coding: a tutorial review', Proc. IEEE 1994, 82, pp 1541-1582.
[2] De Bot, K. 'Visual feedback of intonation I: Effectiveness and induced practice behavior', Language and Speech 1983, vol 26, pp 331-350.
[3] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal. 'A comparative performance study of several pitch detection algorithms' in Acoustics, Speech and Signal Processing ICASSP. IEEE, 1976, vol. 24, pp. 399–418.
[4] W. Hess. 'Pitch determination of speech signals' algorithms and devices', Springer-Verlag Berlin and Heidelberg, 1983.
[5] B. Kotnik, H. Hage, Z. Kacic, "Evaluation of Pitch Detection Algorithms in Adverse Conditions", *Proc. 3rd International Conference on Speech Prosody*, pp. 149-152, 2006
[6] GU, Y. H. et VAN BOKHOVEN, W. Co-channel speech separation using frequency bin non-linear adaptive filtering. In: Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on. IEEE, 1991. p. 949-952.
[7] T. Tolonen, et M. Karjalainen, A computationally efficient multipitch analysis model. IEEE Trans. Speech and Audio Process., Vol. 8, No. 6, pp. 708-716, 2000
[8] A.M.NOLL.'Cepstrum pitch determination', Acoustic Society of America, February 1967 vol.41,pp.293-309
[9] Rashmi Makhljanl, Urmila Shrawankar Hrawankar,V. M. Thakare, Rashmi Makhljanl, Urmila Shrawankar , Dr. V. M. Thakare. 'Speech enhancement using pitch detection approach for noisy environement', International Journal of Engineering Science and Technology (IJEST), 2011, Vol. 3 No. 2
[10] F.J Ancin, B.L. Burrows and R.A. Carrasco. 'A Novel DyWTVT approach for continuous speech estimation', In

Proceedings EUSIPCO, 1994, volume 3, pp 7P.13 1677-1680.

[11] F.J Ancin, B.L. Burrows and R.A. Carrasco. 'A Novel DyWTVT approach for continuous speech estimation', In Proceedings EUSIPCO, 1994, volume 3, pp 7P.13 1677-1680.

[12] Léonard Janer. 'Modulated gaussian wavelet transform based speech analyser (MGWTSA) pitch detection algorithm (PDA) ', In Proceedings EUROSPEECH,1995, volume 1, pp 401-404.

[13] S.Kadambe and G.F. Bourdreaux-Bartels 'A comparison of wavelet Function for pitch Detection of speech Signals' In Proceedings ICASSP, 1991

[14] Mikel L. Larreategui, F.J. Ancin, and Rolando A. Carrasco. 'An improved epoch detection alghoriyhm based on sinusoidal modelling of speech', In Proceedings EUROSPEECH, 1995 volume 1, pp 409-412

[15] Sadler et al., 'On multiscale wavelet analysis for step estimation ', in Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'98), Seattle, WA, Apr. 1998, vol. 3, pp. 1517–1520

[16] Sadler and Swami. 'Analysis of multiscale products for step detection and estimation', IEEE Transactions on Information Theory, Volume 45 Issue 3, April 1999, pp 1043-1051

[17] JLASSI, Wided, BOUZID, Aicha, et ELLOUZE, Noureddine. A new method for pitch smoothing. In : Advanced Technologies for Signal and Image Processing (ATSIP), 2016 2nd International Conference on. IEEE, 2016. p. 657-661.

★★★