Robust Pitch Estimation using Ensemble Empirical Mode Decomposition

Sujan Kumar Roy[1,2], *Md. Khademul Islam Molla*[2] *and Keikichi Hirose*[3]

[1] University of Concordia, ON, Canada, [2]University of Rajshahi, Rajshahi, Bangladesh
[3]Graduate School of Information Science and Technology, The University of Tokyo, Japan

khademul.cse@ru.ac.bd, hirose@gavo.t.u-tokyo.ac.jp

Abstract

This paper presents an efficient pitch estimation algorithm for noisy speech signal using ensemble empirical mode decomposition (EEMD) based time domain filtering. The dominant harmonic of noisy speech is enhanced to make pitch period more prominent. The normalized autocorrelation function (NACF) of the modified signal is then decomposed into time varying subband signals using EEMD. In contrast to the ordinary EMD, it does not introduce any mode mixing during decomposition. The subbands containing pitch component are selected and separated yielding partially reconstructed signal. The pitch period is determined from thus separated signals. The experimental results show that the proposed algorithm performs better compared to other recently reported algorithms in noisy environment.

Index Terms: ensemble empirical mode decomposition, filtering, pitch estimation

1. Introduction

Pitch information is an important prosodic feature which is used in many speech processing applications including speech enhancement, automatic speech recognition, analysis and modeling of speech prosody, low-bit-rate speech coding [1]. Although there are many pitch estimation algorithms (PEAs), the development of an efficient PEA is still demanding.

Some PEAs implemented in time domain contained poor accuracy of pitch estimation [2]-[5]. Autocorrelation Function (ACF) based algorithm is introduced in [2]-[3]. The performance of ACF method is basically depended on pitch peak in the autocorrelation domain which may be quite difficult to perform against noise, quasi-periodic nature of the speech signals [6]-[7]. Normalized autocorrelation function (NACF) based algorithm has been presented in [2] which provides better results than ACF but suffer from robustness. A weighted autocorrelation (WAC) based method has been presented in [3]. It is easy to enhance pitch peak by dividing ACF with the AMDF in a lower portion than succeeding peaks but this leads the algorithm to cause a serious problem called double pitch error. Signal reshaping technique with the improvement of specific harmonic is presented in [6]-[7]. The dominant harmonic (DH) of the noisy speech signals is determined by using a DFT based method and boosted the amplitude of DH in the analyzing signal which is called dominant harmonic modification (DHM). This technique also fails to perform under noisy environment. The data adaptive techniques of pitch estimation are introduced in [8]-[11] using EMD [12, 13]. The ordinary EMD suffers from the 'mode mixing' effect decreases pitch estimation performance. To overcome the mode mixing problem, Wu and Huang [14] proposed a modification to the EMD algorithm what is termed as EEMD.

In this paper, a novel pitch estimation algorithm is proposed using the combination of DHM and EEMD. DHM plays an important role to overcome the double pitch error while EEMD acts as data adaptive time domain filtering to separate the signal containing only the pitch information which minimizes the pitch estimation error.

2. Pitch Estimation Algorithm

The noticeable improvement of the EMD based PEA is possible by overcoming the mode mixing problem. At first, conventional low-pas filtering is performed on the noisy speech signal. If $s(n)$ and $v(n)$ denote the speech and additive noise signals respectively, the observed speech signal $x(n)$ can be represented as: $x(n) = s(n) + v(n)$.

The noise effect can be reduced significantly by pre-filtering the observed signal $x(n)$ in the Fourier domain. As the pitch range of speech signal is well known to be 50-500Hz, a significant portion of the high frequency components is filtered out in frequency domain. The resultant signal is termed as pre-filtered speech (PFS) and represented as $\psi(n)$ which contains less noise. In the second phase of pre-processing, the DHM is applied to the PFS signal which is described in the following subsection.

2.1. Dominant harmonic enhancement

The dominant harmonic (DH) can be estimated as the one which has the largest amplitude in the Fourier domain [6].

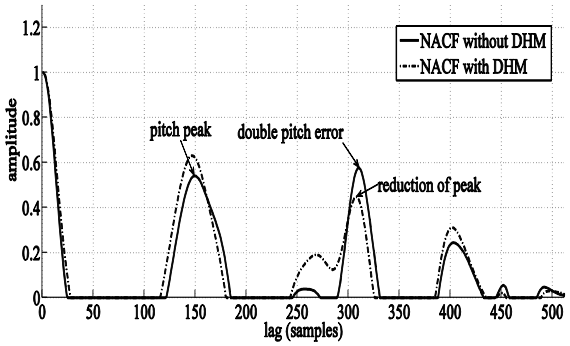

Figure 1: *Pitch peak enhancement using DHM for a voiced frame of a female speaker with 0dB SNR. NACF causes double pitch error but DHM resolves that problem. The true pitch is 151 in samples.*

Let $y_{dh}(n)$ denotes the DH present in the PFS signal $\psi(n)$. The signal after dominant harmonic enhancement denoted by $y(n)$ is then given by:

$$y(n) = \psi(n) + \bar{y}_{dh}(n) \quad (1)$$

where,

10.21437/SpeechProsody.2014-94

$$\bar{y}_{dh}(n) = \rho[y_{dh}(n) - |y_{dh}(n)|] \quad (2)$$

and

$$y_{dh}(n) = \theta_{dh}\cos(\omega_{dh}n + \delta_{dh}) \quad (3)$$

where, θ_{dh}, ω_{dh}, δ_{dh} are amplitude, frequency and phase of the dominant harmonic respectively, ρ is an arbitrary constant, and |.| denotes the absolute value. The parameter ρ controls the mixing ratio to be chosen appropriately. Figure 1 demonstrates the effects of the DHM which shows that the conventional NACF of $\psi(n)$ causes double pitch error while the use of DHM to $\psi(n)$ overcomes such error.

2.2. EEMD in pitch detection

EMD suffers from mode mixing effect which indicates that the oscillations of different time scales coexist in a given IMF, or that oscillations with the same time scale have been assigned to different IMFs [14]. Due to this problem, EMD does not guarantee the accurate frequency scale separation and pitch information can be mixed between two successive IMFs which degrades the pitch detection accuracy. EEMD utilizes the scale separation principle of the EMD. The principle of the EEMD is to add white noise into the signal with many trials. The noise in each trial is different, and the added noise can be canceled out on average, if the number of trials is sufficient. Let $\phi(n)$ is the NACF of $y(n)$. For signal $\phi(n)$, original EMD is defined as [12]:

$$\phi(n) = \sum_{k=1}^{N} c_k(n) \quad (4)$$

where $c_k(n)$ is the k^{th} IMF. EEMD decomposes $\phi(n)$ by first construction an ensemble of signal samples $\phi_m(n)$ by adding to $\phi(n)$, M independent copies of finite amplitude white-noise $\eta_m(n)$, i.e.,

$$\phi_m(n) = \phi(n) + \eta_m(n), (m=1, 2, ..., M), \quad (5)$$

Applying EMD to $\phi_m(n)$ and repeat until the trial number with different added white noise series of the same power at each time, the new IMF combination $c_k^{(m)}$ is obtained, where k is the iteration number and m is the IMF scale.

$$\phi_m(n) = \sum_{k=1}^{N} c_k^{(m)}(n), \qquad (m=1, 2, \ldots, M), \quad (6)$$

where, N is the trial number.

The ensemble means is calculated as:

$$\tilde{c}_k(n) = \frac{1}{M}\sum_{m=1}^{M} c_k^{(m)}(n), \qquad (k=1, 2, 3\ldots), \quad (7)$$

A large number (M) of samples and white noise of finite amplitude are required to force the ensemble to exhaust all possibilities in the sifting process. In this way, possible mode mixing is effectively removed, and components of different time scales embodied in the original signal are well collated in proper IMFs, whose frequency bands essentially approximate those dictated by the dyadic filter banks [14]. The effect of the added white noise should decrease following the well-established statistical rule:

$$\beta_n = \frac{\beta}{\sqrt{M}} \quad (8)$$

or

$$\ln\varepsilon_n + \frac{\varepsilon}{2}\ln M = 0 \quad (8)$$

where, M is the number of ensemble members, β is the amplitude of the added noise and β_n is the final standard deviation of error, which is defined as the difference between the input signal and the corresponding IMF(s) [14].

Mode mixing problem is illustrated in the left column of Figure 2, where a clean speech segment of 100 ms length is decomposed by EMD. The three IMFs with higher energy are shown. The appearance of oscillations of dramatically disparate scales in IMF_3 is clear. Another example can be seen in IMF_4, where two oscillations are marked with circles. These oscillations are very similar to those on IMF 5. On the right side of Figure3, in IMF_4 and IMF_5 there is no mode mixing, where *M=50* and *β=0.1* were used for generating the resulting IMFs decomposed by EEMD.

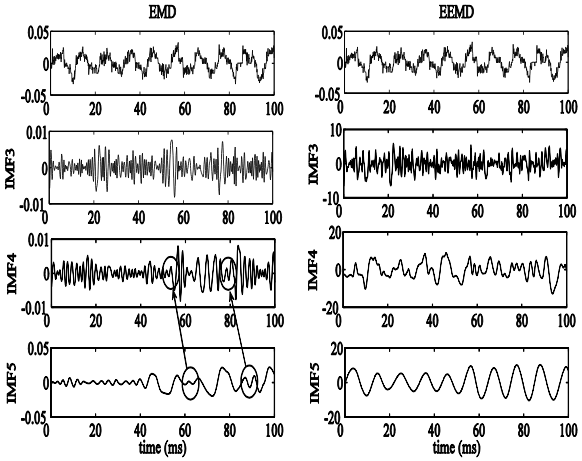

Figure 2: *A clean speech segment analyzed by EMD (left col-umn) and EEMD (right column). The corresponding IMFs 3 to 5 are shown. In IMF4 and IMF5of EMD where "mode mixing" occurs are marked with circles.*

The pitch detection performance using DHM method often reduced especially for low SNR speech signals [6]-[7]. To mitigate this shortcoming effects, EEMD based data adaptive time domain filtering is applied to the NACF of the DH modified signal *y(n)* (equation-2) and a partial reconstruction is made from the EEMD domain to calculate the actual pitch period. The reconstructed signal *λ(n)* is

obtained as $\lambda(n) = \sum \tilde{c}_j(n)$, where $\tilde{c}_j(n)$ are the IMFs and its fundamental periods are within the pitch range 50-500Hz. Let $\xi(n)$ is the NACF of $y(n)$. The NACF $\xi(n)$ and its corresponding partially reconstructed signal $\lambda(n)$ are shown in Figure 3. It is observed that the pitch peak is more prominent in $\lambda(n)$ and overcomes double pitch error.

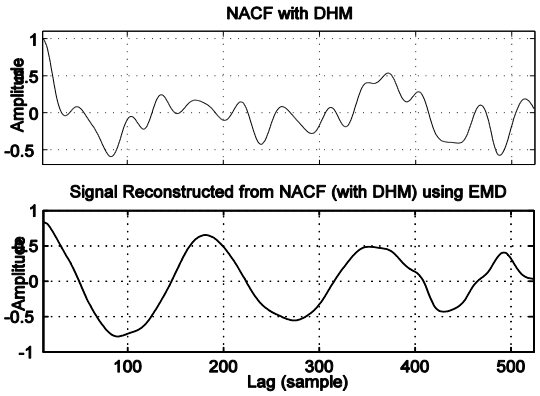

Figure 3: *NACF of the signal after performing the DH enhancement (above), the partially reconstructed signal from NACF with DHM using EMD (bellow). The double pitch error is eliminated (bellow).*

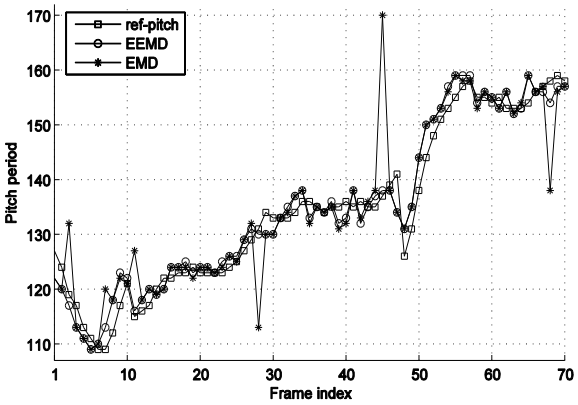

Figure 4: *Performance Comparison of pitch period estimated by EEMD (proposed) and EMD method with the reference pitch for female speech. The proposed EEMD based method performs better than EMD.*

2.3. Proposed pitch detection algorithm

The proposed algorithm for pitch estimation can be summarized as:

- Apply pre-filtering to the noisy speech signal to remove a significant portion beyond the pitch range 50-500Hz.
- Perform signal modification using DHM on PFS and then apply NACF to the modified signal.
- Apply EEMD to the obtained NACF.
- Sum up the IMFs with fundamental period lies within the specified pitch range for partial.
- Take the right half of the reconstructed signal.
- The amplitude at zero-lag is selected as the starting index of the pitch period.
- Find the next highest peak from the right half of the reconstructed signal.
- Calculate the pitch period from the difference between the starting index and the next highest peak index.

The detected pitch response based on the proposed algorithm has shown in Figure 4.

3. Experimental Results and Discussion

The performance of the proposed method is tested using the Keele pitch extraction reference database obtained from ftp://ftp.cs.keele.ac.uk/pub/pitch/ which contains the sampling frequency of 20 kHz with 16-bit resolution. Note that the value of the parameter ρ in Eq. (3) was set to 0.5. Both male (M2~M3) and female (F2~F3) mature speakers' speech are used here. There are 2650 'clearly voiced' male frames and 3227 'clearly voiced' female frames that is a total of 5877 analysis frames are used for experiment. White Gaussian noise and babble noise are used to corrupt the test signal to investigate the robustness. In the experiment, each 25.6 msec analysis frame is weighted by a 512-point rectangular and 10 msec is used as frame shift to generate the reference pitch values given in the database. If the estimated pitch for a frame deviates from the reference by >20%, we recognize the error as a gross pitch error (GPE).

The performance of the proposed EEMD based algorithm is compared with recent four PEAs - EMD based method [11], DHM based method [6], and conventional NACF method [2] and WAC [3] in presence of white noise as shown in Table I. It is observed that the proposed EEMD based algorithm performs better and closer to the original reference pitch (as shown in Figure 4) than all the mentioned PEAs for a wide range of SNRs(-5dB to 30dB). Figure 5 shows the average %GPE for male and female speakers respectively from which we see that the proposed PEA improves pitch estimation accuracy in noisy environment especially with low SNRs.

Table 1. *Performance comparison of EEMD algorithm with recently reported PEAs for male and female data. White noise is used to corrupt the speech signal.*

SNR(dB)		-5	0	10	20	30
M2	EEMD	6.72	5.51	2.02	1.54	1.53
	EMD	13.78	7.69	4.13	2.92	2.67
	DHM	14.26	7.45	2.99	2.26	1.94
	NACF	22.36	14.74	6.96	5.26	4.94
	WAC	25.20	15.23	7.37	6.17	6.07
M3	EEMD	3.12	0.99	0.21	0.14	0.07
	EMD	7.77	2.40	0.35	0.14	0.14
	DHM	12.28	4.73	1.12	0.56	0.35
	NACF	23.37	11.51	3.38	1.69	1.27
	WAC	24.29	12.07	3.38	1.20	0.98
F2	EEMD	5.51	1.98	0.93	0.88	0.55
	EMD	10.98	5.85	1.32	1.05	0.72
	DHM	10.97	6.39	1.93	1.15	0.93
	NACF	21.48	11.91	4.24	2.04	1.70
	WAC	23.05	11.58	3.97	1.93	1.59
F3	EEMD	5.51	2.97	0.98	0.28	0.21
	EMD	11.67	5.37	1.49	0.50	0.49
	DHM	11.38	6.01	2.05	0.70	0.49
	NACF	21.49	12.16	4.73	2.19	1.69
	WAC	21.78	7.85	4.38	1.64	1.62

In the second phase of experiment, we evaluate and compare the pitch detection performances of the proposed PEA with EMD [11, 15] and DHM [6] in noisy environment with bubble noise as illustrated in Table II. The proposed EEMD method provides better results than EMD and DHM based methods for the whole range of SNRs (-5dB to 30dB). The average performance comparison for this experiment has been shown in Figure 6 and it is clearly observed that the proposed PEA exhibits better performance than EMD and DHM based approach.

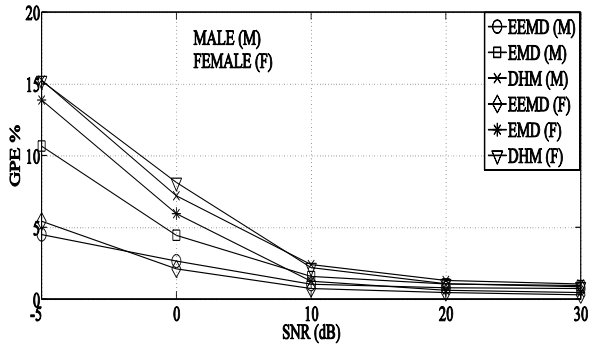

Figure 5: *Average performance results in terms of %GPE for white noise ('M' and 'F' denote male and female, respectively).*

Table 2. *Performance comparison of EEMD algorithm with recently reported PEAs for male and female data. Babble noise is used to corrupt the speech signal*

SNR(dB)		-5	0	10	20	30
M2	EEMD	20.65	10.93	5.26	2.26	1.54
	EMD	24.71	18.15	5.26	2.76	2.75
	DHM	52.83	32.98	9.64	3.81	3.08
M3	EEMD	21.66	14.74	2.04	0.56	0.07
	EMD	29.52	19.49	4.02	0.71	0.28
	DHM	57.98	38.42	8.55	1.06	0.35
F2	EEMD	13.18	5.62	1.48	1.10	0.56
	EMD	58.30	32.98	5.63	1.49	0.93
	DHM	73.08	47.21	9.82	2.31	0.94
F3	EEMD	8.69	5.23	1.06	0.35	0.28
	EMD	63.72	41.02	9.12	2.75	0.49
	DHM	67.05	38.76	6.44	1.98	0.78

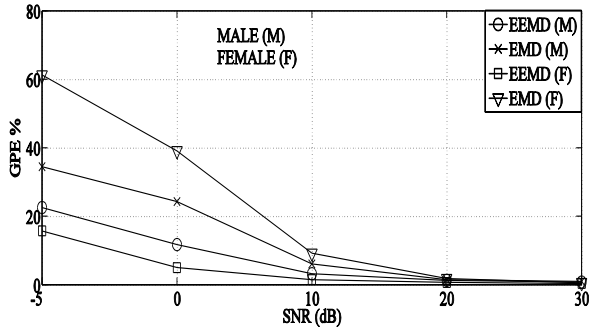

Figure 6: *Average performance results in terms of %GPE for babble noise ('M' and 'F' denote male and female, respectively).*

4. Conclusions

It is observed that the proposed method provides better performance in terms of % GPE than other recently reported PEAs in bubble and white noise environment for a wide range of SNRs(-5dB to 30dB). The EEMD based filtering efficiently extracts the signal components containing pitch information. Then pitch period is determined from partially reconstructed signal of EEMD domain. The pitch peaks become more prominent in the reconstructed signal. The EEMD is being free of mode mixing problem, pitch period does not overlap between any two IMFs and in most cases the detected pitch responses are almost closer to the original reference pitch. This strategy proves the superioty of the proposed algorithm. Also its performance in low SNRs is extremely better than EMD, DHM, NACF and WAC based PEAs. Finally, all sorts of experimental results prove the advantages of the proposed algorithm.

5. References

[1] W. Hess, Pitch Determination of Speech Signals: Algorithms and Devices, Springer, Berlin, 1983.

[2] K. Kasi and S. A. Zahorian, “Yet another algorithm for pitch tracking”, Proc. IEEE ICASSP, pp.361-364, 2002.

[3] T. Shimamura and H. Kobayashi, “Weighted Autocorrelation for Pitch Extraction of Noisy Speech”, IEEE Trans. Speech and Audio Proc., 9(7):727-730, 2001.

[4] M. C. Dogan and J. M. Mendel, “Real-time robust pitch detector”, Proc. of IEEE ICASSP, 1, 129-132, 1992.

[5] N. Abu-Shikhan and M. Deriche, “A novel pitch estimation technique using the Teager energy”, Proc. of ISSPA, 1, 135-138, 1999.

[6] M. K. Hasan et. al., “Signal reshaping using dominant harmonic for pitch estimation of noisy speech”, Signal Processing, 86(5):1010-1018, 2005.

[7] M. K. Hasan, C. Shahnaz and S. A Fattah, “Determination of pitch of noisy speech using dominant harmonic frequency”, Proc. IEEE Int. Symposium on Circuits and Systems, 2, pp.556-559, 2003.

[8] H. Huang and J. Pan, “Speech pitch determination based on Hilbert-Huang transform”, Signal Processing, 86(4):792-803, 2005.

[9] Z. Yang, D. Huang and L. Yang, “A novel pitch period detection algorithm based on Hilbert-Huang transform”, LNCS 3338, pp. 586-593, Sinobiometrics, 2004.

[10] M. K. I. Molla, K. Hirose, N. Minematsu and M. K. Hasan, “Pitch Estimation of Noisy Speech Signals using Empirical Mode Decomposition”, Proc. of EUROSPEECH 2007.

[11] Sujan Kumar Roy, Md. Khademul Islam Molla, Keikichi Hirose and Md. Kamrul Hasan, “Harmonic modification and data adaptive filtering based approach to robust pitch estimation”, International Journal of Speech Technology, Volume 14, Number 4, 339-349, DOI: 10.1007/s10772-011-9112-6.

[12] N. E. Huang et. al., “The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis”, Proc. Roy. Soc. London A, Vol. 454, pp. 903-995, 1998.

[13] P. Flandrin, G. Rilling and P. Goncalves, “Empirical mode decomposition as a filter bank”, IEEE signal processing letters, Vol. 11, No. 2, pp.112-114, 2004.

[14] Z. Wu and N.E. Huang, “Ensemble Empirical Mode Decomposition: a noise-assisted data analysis method,”Ad-vances in Adaptive Data Analysis, vol. 1, pp. 1-41, 2009.

[15] S. K. Roy, M. K. Islam, K. Hirose and M. K. I. Molla, “Dominant harmonic modification with data adaptive filter based algorithm for robust pitch estimation”, Proc. of IEEE Int. Symposium on Circuits and Systems (ISCAS), 2011.