Pitch Estimation of Noisy Speech Signals using Empirical Mode Decomposition

Md. Khademul Islam Molla[1], *Keikichi Hirose*[1], *Nobuaki Minematsu*[2] *and Md. Kamrul Hasan*[3]

[1]Dept. of Information and Communication Engineering, The University of Tokyo, Tokyo, Japan
[2]Dept. of Frontier Informatics, The University of Tokyo, Tokyo, Japan
[3]Dept. of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
E-mail: [1,2]{molla, hirose, mine}@gavo.t.u-tokyo.ac.jp, [3]khasan@eee.buet.ac.bd

Abstract

This paper presents a pitch estimation method of noisy speech signal using empirical mode decomposition (EMD). The normalized autocorrelation function (NACF) of the noisy speech signal is decomposed into a finite set of band-limited signals termed as intrinsic mode functions (IMFs) using EMD. The periodicity of one IMF is supposed to be equal to the accurate pitch period. A conventional autocorrelation based pitch period detection method is used to select the IMF with pitch period. The accurate pitch period is obtained from the selected IMF. The pitch estimation performance in term of gross pitch error (GPE) of the proposed algorithm is compared with recently proposed methods. The experimental results show that the EMD based algorithm performs better in pitch estimation of noisy speech.

Index Terms: empirical mode decomposition, pitch estimation, normalized autocorrelation.

1. Introduction

The estimation of pitch period of speech signal plays an important role in different speech processing applications including speech enhancement using harmonic model, automatic speech recognition and understanding, analysis and modeling of speech prosody, low-bit-rate speech coding etc. Although many methods of pitch estimation have been proposed, reliable and accurate detection is still a challenging task [1]. The speech waveform is weakly periodic and the instantaneous values of pitch are different even within a frame. The presence of noise further complicates the problem and deteriorates the performance of the pitch estimation algorithms (PEAs) [2].

Various PEAs have been reported in the literature to address this problem. Pitch has been determined in the time-domain [1]-[4], frequency-domain [5]-[6], and also in time-frequency domain [7]-[9]. The discussion of this paper is only confined in the time-domain method. Among the reported methods, the autocorrelation based approaches are popular for their simplicity, low computational complexity, and better robustness to noise. Because of the periodic nature of the voiced speech, its autocorrelation function (ACF) is also periodic with period equal to the pitch value. The accuracy of ACF based PEAs depends on the estimation accuracy of the 'pitch peak' in the ACF. The following introduce additional peak and may obscure the 'pitch peak': the presence of noise, peaks due to the detailed formant structure of the vocal track, quasi-periodic nature of the speech signal, and the choice of analysis frame size and window [10]. A weighted autocorrelation (WAC) method using the average magnitude difference (AMDF) has been proposed in [1]. The main shortcoming of the method is that, it is very sensitive to the double pitch error. Normalized autocorrelation function (NACF) based technique is introduced in [2]. The NACF is also sensitive to the noise. A significant improvement of conventional NACF by signal reshaping is reported in [10]. The EMD based pitch estimation algorithm is implemented in time-frequency domain in [8]-[9]. Several thresholds are used to determine the pitch frequency in [8] without considering noise. The authors have not produced any quantitative evaluation of their proposed algorithm using large speech database. In [9], the derivative of the instantaneous energy computed from the Hilbert spectrum is reported as the basis of pitch determination. There is a big possibility to have the instantaneous energy peak in the places other than glottal pulse position due to the presence of noise with the speech signal.

In this paper a new pitch estimation approach is presented, which is based on the empirical mode decomposition (EMD) developed specially for non-linear and non-stationary time-series analysis [11]. The NACF of the speech signal is decomposed into basis functions called intrinsic mode functions (IMFs) using EMD. The decomposition is started to extract the highest frequency local oscillation and ends up with the lowest frequency global oscillation. One of the basis functions carries the fundamental oscillation with pitch period. Conventional autocorrelation based pitch estimation is used to select that IMF in the EMD domain. Then the actual pitch period is determined from the selected basis function.

Regarding organization of this paper, the basic of EMD is described in Section 2, the proposed pitch estimation algorithm is illustrated in Section 3, experimental results are presented in Section 4 and finally some concluding remarks are described in Section 5.

2. Basics of EMD

Empirical mode decomposition (EMD) represents any temporal signal into a finite set of AM-FM oscillating components which are bases of the decomposition. The key benefit of using EMD is that it is an automatic decomposition and fully data adaptive.

The principle of the EMD technique is to decompose a signal $s(t)$ into a sum of the band-limited functions $C_m(t)$ called intrinsic mode functions (IMFs). Each IMF satisfies two basic conditions: (i) in the whole data set, the number of extrema and the number of zero crossings must be the same or differ at most by one, (ii) at any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero. The first condition is similar to the narrow-band requirement for a stationary Gaussian process and the second condition is a local requirement induced from the global one, and is necessary to ensure that the instantaneous frequency will not have

redundant fluctuations as induced by asymmetric waveforms. There exist many approaches of computing EMD [12]. The following algorithm is employed here to decompose signal $s(t)$ into a set of IMF components.

1. Set $g_1(t) = s(t)$
2. Detect the extrema (both maxima and minima) of $g_1(t)$
3. Generate the upper and lower envelopes $h(t)$ and $l(t)$ respectively by connecting the maxima and minima separately with cubic spline interpolation
4. Determine the local mean as: $\mu_1(t)=[h(t)+l(t)]/2$
5. IMF should have zero local mean; subtract μ_1(t) from the original signal as: $g_1(t)=g_1(t)-\mu_1(t)$
6. Decide whether $g_1(t)$ is an IMF or not by checking the two basic conditions as described above
7. Repeat steps 2 to 6 and end when an IMF $g_1(t)$ is obtained

Once the first IMF is derived, define $C_1(t)=g_1(t)$, which is the smallest temporal scale in $s(t)$. To find the rest of the IMF components, generate the residue $r_1(t)$ of the data by subtracting $C_1(t)$ from the signal $s(t)$ as: $r_1(t)=s(t)-C_1(t)$. The sifting process will be continued until the final residue is a constant, a monotonic function, or a function with only one maxima and one minima from which no more IMF can be derived. The subsequent basis functions and the residues are computed as,

$$r_1(t)-C_2(t)=r_2(t),\cdots,r_{M-1}(t)-C_M(t)=r_M(t) \quad (1)$$

where $r_M(t)$ is the final residue. At the end of the decomposition the signal $s(t)$ is represented as:

$$s(t)=\sum_{m=1}^{M} C_m(t)+r_M(t) \quad (2)$$

where M is the number of IMFs and $r_M(t)$ is the final residue.

Another way to explain how EMD works is that it extracts out the highest frequency oscillation that remains in the signal. Thus locally, each IMF contains lower frequency oscillation than the one extracted just before. Being data adaptive, the basis usually offers a physically meaningful representation of the underlying processes. There is no need of considering the signal as a stack of harmonics and, therefore, EMD is ideal for analyzing non-stationary and nonlinear data. The higher order IMFs contain lower frequency oscillations than that of lower order IMFs. Each IMF is considered as a mono-component contribution such that the derivation of instantaneous amplitude and frequency provides a physical significance. The advantage of this time-space filtering is that the resulting band passed signals preserve the full nonlinearity and non-stationary in physical space. This filtering method is intuitive and direct, its basis is a posteriori and data adaptive, which means, it is based on the data and also derived from data. The completeness of the decomposition is given by the Eq. (2). The original signal can easily be reconstructed by simply adding the bases with the error of the order 10^{-14} (found experimentally) which is negligible in practical sense.

3. Pitch Period Detection

A two-stage pitch period determination algorithm is proposed here. In the first stage, a rough estimation of the pitch period is performed by conventional autocorrelation based method. The EMD based method is applied in the second stage to estimate the accurate pitch period.

3.1. Autocorrelation of noisy speech model

Consider that $x(n)$ and $v(n)$ denote speech and uncorrelated white Gaussian noise with zero mean and variance σ_v^2, respectively. Then the observed signal $y(n)$ is given by: $y(n)=x(n)+v(n)$. The autocorrelation function of $y(n)$ can be expressed as

$$R_{yy}(\tau)=\begin{cases} R_{xx}(\tau)+\sigma_v^2 & for\ \ \tau=0, \\ R_{xx}(\tau) & for\ \ \tau\neq 0, \end{cases} \quad (3)$$

Where $R_{xx}(\tau)$ is the autocorrelation function of the clean-speech signal $x(n)$ estimated as

$$R_{xx}(\tau)=\frac{1}{N}\sum_{n=0}^{N-1-|\tau|} x(n)x(n+|\tau|). \quad (4)$$

Where N is the length of the speech frame under analysis and τ is the lag number. $R_{xx}(\tau)$ essentially exhibits peaks at the periodicity (T) of $x(n)$ (i.e., at $\tau=\rho T$, where ρ is an integer). The basic idea of ACF based method is to use the location of the second largest peak (at $\tau=T$) relative to the largest peak (at $\tau=0$) to obtain the estimate of the pitch period [1]. Such method is noise robust as long as speech and noise are truly uncorrelated, a requirement seldom met in practice. If the noise is white, its effect can be reduced significantly by pre-filtering the observe signal $y(n)$. As the pitch range is known to be 50-500 Hz for most male and female speakers, a significant portion of the high frequency components is filter out. The pre-filtering noisy speech (PFNS) signal, $\varphi(n)$, containing less noise is used in pitch estimation.

The WAC based method [1] is used here to determine the preliminary pitch period. The AMDF weighted ACF, $\psi(\tau)$, is used and is defined as

$$\psi(\tau)=\frac{R_{\varphi\varphi}(\tau)}{\chi(\tau)+\varepsilon} \quad (5)$$

Where ε is a small positive constant, $R_{\varphi\varphi}(\tau)$ is the autocorrelation of $\varphi(n)$, and the AMDF, $\chi(\tau)$, involving $\varphi(n)$ is defined as

$$\chi(\tau)=\frac{1}{N}\sum_{n=0}^{N-1-|\tau|} \left|\varphi(n)-\varphi(n+|\tau|)\right|. \quad (6)$$

It is expected to give deep notches at $\tau=\rho T$, and therefore the true pitch peak in $\psi(\tau)$ is emphasized. Even with the aforementioned pre-processing, the pitch determination algorithm in [1] may give erroneous results under strong noisy condition due to the presence of spurious peaks obscuring the second largest peak and also due to the inherent shortcoming introduced with the AMDF. To reduce the effects of erroneous terms a further pitch determination is performed using EMD which extracts the fundamental oscillating component of the signal.

3.2. EMD based pitch estimation

To weaken the effect of formants and noise, pre-filtered noisy speech (PFNS) signal is used here. Instead of using PFNS signal directly, the NACF [2] of the PFNS signal is

taken as the input to the EMD based decomposition. The pitch peaks become more prominent in NACF than the PFNS signal. The steps of the proposed algorithm are as follows:

a) Pre-filter the noisy speech signal to remove a significant portion beyond the pitch range 50-500 Hz
b) Determine the pitch period of PFNS signal using WAC [1] based method
c) Find the NACF [2] of the pre-filtered noisy speech
d) Apply EMD on NACF. There exists an IMF in EMD domain representing the fundamental oscillation of NACF. The period of that IMF is the pitch period of the analysis speech.
e) The pitch period estimated in step (b) is used to select the IMF with fundamental oscillation. The IMF with periodicity closest to the pitch period obtained in step (b) is the target IMF.
f) The selected IMF is considered as damped sinusoid symmetric about the zero-lag. Its periodicity is the estimated pitch period of the speech segment.

A speech frame and its corresponding PFNS signal are shown in Figure 1(a). The NACF of PFNS signal and its EMD is shown in Figure 1(b).

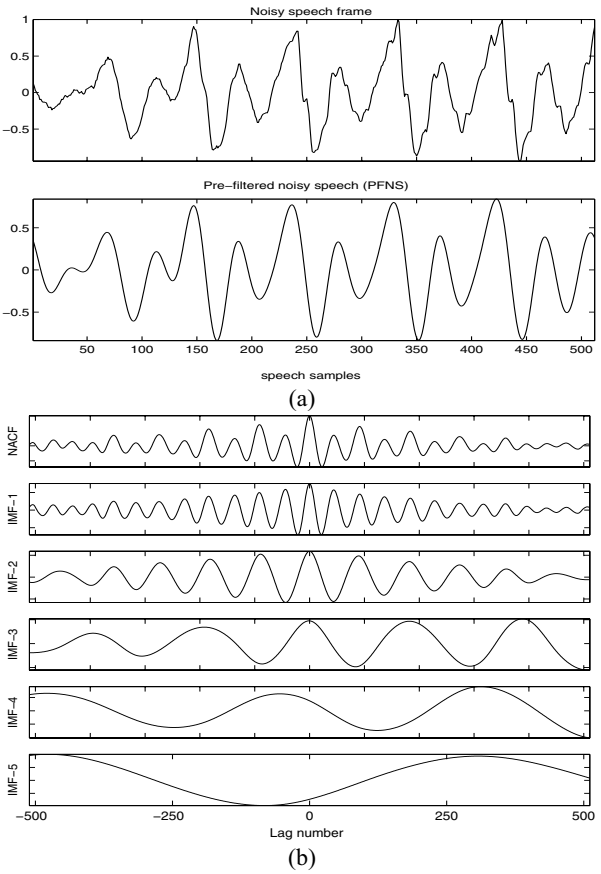

Figure 1: *Pre-filtering and EMD of noisy voiced speech frame; (a) noisy speech frame (upper) and its pre-filtered signal, (b) the NACF of pre-filtered signal (top) and its EMD (different IMFs).*

The pitch peak is emphasized in NACF than the peaks in PFNS signal. Observing the EMD domain, it is clear that the 2nd IMF explicitly corresponds the envelop-like global oscillation of the NACF. It can be considered as a damped sinusoid symmetric with respect to the zero lag position. The peaks of the IMF are matched with the pitch peak in NACF. Since, EMD produces such IMF component as a basis function of NACF, any local change of peaks (close to the pitch peak) due to noise has no significant effect on pitch determination. The difference between two peaks of that IMF will closely match with the pitch period estimated using WAC. The EMD is also treated as dyadic filter-banks [12]. Hence, it is almost not possible to match the preliminary period with the IMF other than the IMF containing the accurate pitch value. Sometimes there exists a significant difference between the accurate pitch period and the period determined by WAC based method as shown in Figure 2. However, the robustness of the EMD based method can accurately estimate the pitch period illustrated in Figure 2.

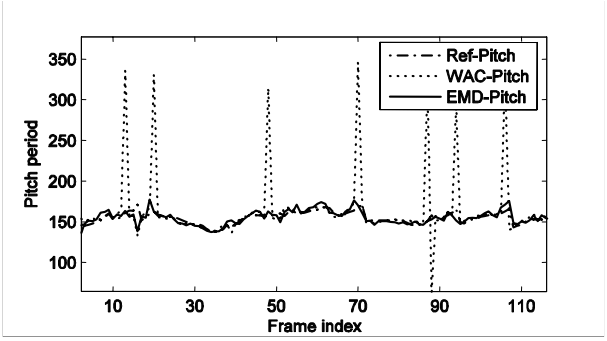

Figure 2: *Comparison of preliminary pitch period estimated by WAC method and EMD based pitch period with the reference pitch. The pitch period estimated by the proposed EMD based method is very much close to the original pitch period*

The proposed EMD based pitch determination does not depends on the prominent peak at glottal pulse position. It recovers the global oscillatory component of NACF representing the pitch period. A small variation of the peak position does not affect estimation of the pitch period.

4. Experimental Results

The Keele pitch extraction reference database obtained from ftp://ftp.cs.keele.ac.uk/pub/pitch/ is used to test the performance of the proposed pitch estimation method. Both male (M2-M3) and female (F2-F3) mature speakers' speech are used here. The speech signal is sampled at 20 kHz with 16-bit resolution. As described in [10], for many frames the original database has reference pitch values where the signal is hardly periodic. Excluding these frames manually after visual inspection a modified database of 'clearly voiced' frames has been developed [10]. We present results using the modified Keele databases. The experiments are conducted by adding white Gaussian noise to the signal. Each 25.6 msec analysis frame is weighted by a 512-point rectangular window $w(n)$ ($w(n)$=1 for $0 \le n \le 511$ and $w(n)$=0, elsewhere). The frame shift is set to be 10 msec as used 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 index GPE is often expressed in percentage denoted as %GPE. The true pitch values are obtained from the original database. The performance comparison of different pitch estimation algorithms (PEAs) is presented

using the modified Keele database which contains 'clearly voiced' reference frames as used in other works [13]. Results obtained using the three PEAs namely, the proposed one using EMD, conventional NACF method [2] and WAC [1] are presented in Tables 1, and in Figure 3. The results of conventional NACF are directly obtained from [10].

Table 1. *Comparison of the proposed EMD based method with conventional NACF and WAC based algorithm in terms of %GPE for male (M2-M3) and female (F2-F3) speakers on the modified Keele pitch database. The white noise is used to corrupt the speech signal.*

SNR(dB)		-15	-5	0	10	20	30
M2	EMD	37.76	12.88	7.05	4.53	4.45	4.23
	NACF	58.83	22.36	14.74	6.96	5.26	4.94
	WAC	61.42	25.20	15.23	7.37	6.17	6.07
M3	EMD	43.07	9.11	3.88	1.62	0.91	0.48
	NACF	67.30	23.37	11.51	3.38	1.69	1.27
	WAC	69.20	24.29	12.07	3.38	1.2	0.98
F2	EMD	61.57	18.09	8.71	1.93	1.32	1.21
	NACF	69.44	21.84	11.91	4.24	2.04	1.70
	WAC	68.56	23.05	11.58	3.97	1.93	1.59
F3	EMD	57.34	16.40	4.87	2.75	1.48	1.34
	NACF	62.16	21.49	12.16	4.73	2.19	1.69
	WAC	67.32	21.78	7.85	4.38	1.64	1.62

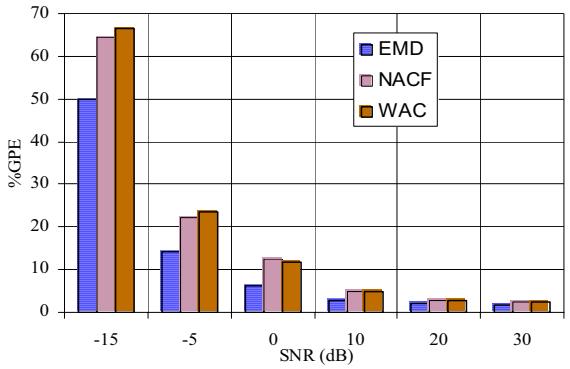

Figure 3: *The average performance in term of %GPE of the EMD based method and a comparison with conventional NACF and WAC.*

As shown in Tables 1, the %GPE obtained by the proposed EMD based technique for both male (M2-M3) and female (F2-F3) speakers is significantly smaller than that of the conventional NACF and WAC methods at all SNRs. The average %GPE for male and female speakers as shown in Figure 3, demonstrates the superior performance of the proposed technique over the conventional NACF and WAC for the whole range of SNR from -15 to 30 dB. The average results are computed for 2650 'clearly voiced' male frames and 3227 'clearly voiced' female frames. It is also observed that the proposed EMD based method is capable of performing better at low SNR compared to the other methods as shown in Figure 3.

5. Conclusions

A new pitch estimation approach for noisy speech signal based on EMD is presented in this paper. The NACF of the pre-filtered signal has been decomposed into AM-FM oscillatory components using EMD. The component oscillating with the period equal to the pitch value is selected by using the pitch period obtained by WAC based technique. It is shown that using the proposed technique, better accuracy in terms of %GPE can be obtained as compared to other recent techniques for a wide range of SNR varying from -15 to 30dB. The superiority of the EMD based method is that, it does not require explicit pitch peak at the glottal pulse positions in ACF. It decomposes the ACF in a non-stationary way in which each IMF preserve the non-stationarity of the original signal. The IMF containing the pitch value obtained by EMD is the envelop-like global oscillation of ACF of PFNS signal. It is relatively more significant to determine the pitch period from the damped sinusoid-like IMF than to find the pitch peak in ACF for a noisy speech. That is why; the proposed method is more noise robust. For fair evaluation of PEAs, Keele database was used in all the experiments.

The selection of the IMF representing the pitch period based on the information in EMD domain without preliminary estimation of pitch period and to observe the performance of the proposed method with difference types of noises will be the future extension of this research.

6. References

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

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

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

[4] Abu-Shikhah, N. and Deriche, M., "A novel pitch estimation technique using the Teager energy", Proc. of ISSPA, 1, 135-138, 1999.

[5] Abe, T., Kobayashi, T. and Imai, S., "Robust pitch estimation with harmonics enhancement in noisy environment based on instantaneous frequency, Proc. ICSLP96, 2, 1277-1280, 1996.

[6] Babu, M. M., "Efficient and accurate pitch estimation using FFT", Proc. IEEE Int. Joint Symposium on Intelligence and Systems, pp.354-358, 1998.

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

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

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

[10] Hasan, M. K., Hussain, S., Setu, M. T. H. and Nazrul, M. N. I., "Signal reshaping using dominant harmonic for pitch estimation of noisy speech", Signal Processing, 86(5):1010-1018, 2005.

[11] Huang, N. E. 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

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

[13] Takrikian, J., Dubnov, S. and Dickalov, Y., "Speech enhancement by harmonic modeling via MAP pitch tracking", Proc. IEEE ICASSP, 1, pp.549-552, 2002.