Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

# Detection of time varying pitch in tonal languages: an approach based on ensemble empirical mode decomposition*

Hong HONG†[1], Xiao-hua ZHU[1], Wei-min SU[1], Run-tong GENG[1], Xin-long WANG[2]

([1]*School of Electronic Engineering and Optoelectronic Techniques, Nanjing University of Science and Technology, Nanjing 210094, China*)

([2]*State Key Laboratory of Modern Acoustics, Institute of Acoustics, Nanjing University, Nanjing 210093, China*)

†E-mail: hongnju@gmail.com

Received Apr. 13, 2011; Revision accepted Aug. 9, 2011; Crosschecked Dec. 29, 2011

**Abstract:** A method based on ensemble empirical mode decomposition (EEMD) is proposed for accurately detecting the time varying pitch of speech in tonal languages. Unlike frame-, event-, or subspace-based pitch detectors, the time varying information of pitch within the short duration, which is of crucial importance in speech processing of tonal languages, can be accurately extracted. The Chinese Linguistic Data Consortium (CLDC) database for Mandarin Chinese was employed as standard speech data for the evaluation of the effectiveness of the method. It is shown that the proposed method provides more accurate and reliable results, particularly in estimating the tones of non-monotonically varying pitches like the third one in Mandarin Chinese. Also, it is shown that the new method has strong resistance to noise disturbance.

**Key words:** Ensemble empirical mode decomposition, Time varying pitch, Tonal language, Noise restraint
**doi:**10.1631/jzus.C1100092 **Document code:** A **CLC number:** TN912.3

## 1 Introduction

Detection or estimation of the time variation of pitch is of crucial importance in speech processing of such tonal languages as Chinese (Wang and Seneff, 1998; Chang *et al.*, 2000), a language spoken by over 1.3 billion of people in the world. Among the characters of Mandarin Chinese are the four speech tones, which are the manifestation of temporally varying patterns or contours of pitch and often bewilder foreigners who try to learn this language. Other than official Chinese, there are roughly seven dialects in Chinese, each still spoken by vast populations in different parts of China. These dialects are characterized by, among others, their intricate and yet unique tonal patterns or contours of time varying pitch. In addition, China, as one of the most ancient civilizations, has a plentiful poetic literature inherited particularly from Tang and Song dynasties. The beauty of the classical poetry lies not only on the elegantly arranged words and the artistic conceptions that are conveyed, but also on their cadent rhythmic expressions in the format of voicing patterns of pitch.

Existing methods for pitch detection can roughly be divided into three classes: frame-, event-, and subspace-based pitch detectors. Most frame-based pitch detectors (Noll, 1967; Talkin, 1995; Li *et al.*, 2000; de Cheveigne and Kawahara, 2002; Chan and So, 2004) are based on the assumption of invariability of pitch over a speech segment of short duration, consisting of several pitch

* Project supported by the National Natural Science Foundation of China (No. 10574070) and the State Key Laboratory Foundation of China (No. 9140C240207060C24)

periods. Thus, the time varying information of pitch within the short duration cannot be accurately extracted. Event-based pitch detectors (Ananthapadmanabha and Yegnanarayana, 1975; Cheng and O'Shaughnessy, 1989; Kadambe and Boudreaux-Bartels, 1992; Boersma, 2002) are based on pitch marking or epoch detection. These methods detect the event of a glottal cycle, for example, the instants of glottal closure, from which the duration of each pitch period can be derived. However, event-based pitch detectors are sensitive to shapes of the speech waveform, which may cause the detectors to fail in cases where instances of glottal closures are not very obvious (Resch *et al.*, 2007). Recently, subspace-based pitch detectors (Christensen *et al.*, 2007; 2008; Zhang *et al.*, 2010) have been shown to have good estimation performance with a high accuracy in low signal-to-noise ratio (SNR) conditions, and these detectors also provide flexibility for robust estimation on inharmonic signals and multi-pitch signals. However, the main disadvantages of subspace-based pitch detectors are the high computational complexity of the subspace decomposition process, which would make these detectors not very effective for fast processing in many real-time applications.

All these factors necessitate a reliable and accurate speech processing algorithm that can extract the time varying pitch information for applications such as speech synthesis and recognition of tonal languages. The Hilbert-Huang transform (HHT) (Huang *et al.*, 1998) seems competent for completing the task. HHT is built essentially on empirical mode decomposition (EMD). With EMD, any signal, whether linear or nonlinear, can be disintegrated into a sequence of intrinsic mode functions (IMFs) that characterize the underlying dynamics and evolution of the system generating the signal. These IMFs act as base functions, like those in Fourier analysis, but are totally data or signal specific. These features further make HHT not only capable of coping with non-stationary signals, but also highly adaptive in general signal processing (Huang and Wu, 2007). Although still in its developing state, HHT has already demonstrated its remarkable power in a number of applications (Liang *et al.*, 2000; Huang *et al.*, 2001; Schlurmann *et al.*, 2001; Jánosi and Müller, 2005; Goska and Krawiecki, 2006; Qi *et al.*, 2007; Pai and Palazotto, 2008; Bekara and Baan, 2009).

Recently, EMD was applied to pitch detection (Huang and Pan, 2006). At first view, it seems promising, since EMD is believed to decompose a speech signal into IMFs, including the one that contains pitch information. However, EMD suffers from the mode mixing effect (Huang *et al.*, 1999; Hong *et al.*, 2009; Xu *et al.*, 2009) and instability against noise disturbance (Lin *et al.*, 2009). The mode mixing effect usually causes the pitch information to be scrapped into pieces in an indefinite number of IMFs in an anomalistic way. Consequently, it becomes difficult to perfectly merge the scattered pitch information into a single piece. In this work, we propose a novel pitch detection algorithm based on EEMD (Wu and Huang, 2009) to avoid these problems and capture the time varying pitch information.

## 2 Formulation and algorithm

### 2.1 Summary of ensemble empirical mode decomposition

EEMD actually is a remarkable improvement of the original EMD for disintegrating any signal $s(t)$, even non-stationary and nonlinear signals, into a sequence of IMFs:

$$s(t)=\sum_{n=1}^{N}\mathrm{IMF}_n(t). \tag{1}$$

$s(t)$ is decomposed by first constructing an ensemble of signal samples $s_m(t)$ by adding to $s(t)$ $M$ independent copies of finite amplitude white noise $n_m(t)$, i.e.,

$$s_m(t)=s(t)+n_m(t),\quad m=1,2,\cdots,M, \tag{2}$$

decomposing every sample $s_m(t)$ into IMFs,

$$s_m(t)=\sum_{n=1}^{N}\mathrm{IMF}_n^{(m)}(t),\quad m=1,2,\cdots,M, \tag{3}$$

in exactly the same way as in the original EMD (Huang *et al.*, 1998), and then calculating the ensemble means

$$\mathrm{IMF}_n=\frac{1}{M}\sum_{m=1}^{M}\mathrm{IMF}_n^{(m)},\quad n=1,2,3,\cdots \tag{4}$$

as the final IMFs. A large number ($M$) of samples and a 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, an unfavorable effect in the original EMD, is effectively avoided, and the 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 (Wu and Huang, 2009). For speech signals sampled at 16 000 Hz, the frequency bands of the decomposed IMFs are approximately 4000–8000 Hz, 2000–4000 Hz, 1000–2000 Hz, $\cdots$, in sequence. Pitch either falls most possibly in one of these bands, say, the band of the 6th or 7th IMF, or transits among several bands belonging to the neighboring modes, usually ranging from $\mathrm{IMF}_5$ to $\mathrm{IMF}_8$. By the decomposition, we therefore are provided with an opportunity to accurately extract time varying pitch from few individual IMFs that are remarkably simpler than the original $s(t)$.

### 2.2 Time varying pitch detection algorithm based on EEMD

Our steps to detect the pitch of a speech signal $s(t)$ can be illustrated in Fig. 1.

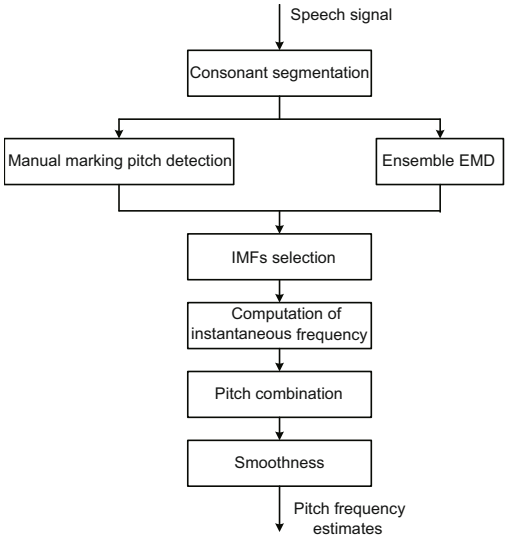


**Fig. 1 Block diagram of the proposed pitch detection algorithm**

First, $s(t)$ is pre-processed by consonant segmentation, so as to remove those segments that are identified as being silent or unvoiced. In general, vowels and consonants appear alternately in Mandarin Chinese. Energy alteration occurs when the sounds change from vowels to consonants or from consonants to vowels. Here we adopt the short-term energy algorithm to realize the consonant segmentation (Deller *et al.*, 1993). The preprocessed $s(t)$ then is decomposed with the EEMD technique, yielding a set of IMFs from $\mathrm{IMF}_1$ (the highest frequency component) to $\mathrm{IMF}_N$ (the lowest frequency component), from which the four IMFs ($\mathrm{IMF}_5$–$\mathrm{IMF}_8$) that may contain the pitch information are picked out and their instantaneous frequencies $\tilde{f}_i(t)$ $(i=5,6,7,8)$ are calculated via Hilbert transform (Huang *et al.*, 1998). Based on the reference contour $\hat{f}_{\mathrm{r}}(t)$ (which will be described later), the pitch contour $F(t)$ can be attained from the four candidates, $\tilde{f}_i(t)$ $(i=5,6,7,8)$, by

$$F(t)=\{\tilde{f}_k(t)|\ 5\le k\le 8, k\in\mathbb{N},\ |\tilde{f}_k(t)-\hat{f}_{\mathrm{r}}(t)| \le |\tilde{f}_i(t)-\hat{f}_{\mathrm{r}}(t)|,\ i=5,6,7,8\}. \quad (5)$$

That is, the $\tilde{f}_i(t)$ that is the nearest to $\hat{f}_{\mathrm{r}}(t)$ for $i=5,6,7,8$ is taken as $F$. Finally, a smoothing operation is applied to $F(t)$ so as to eliminate the possible fluctuation induced by noise and computational errors, yielding a smoothly varying instantaneous frequency $F_0(t)$ taken as the contour of pitch. For simplicity, the moving windowed average is adopted here:

$$F_0(t)=\frac{1}{T}\int_{-T/2}^{+T/2}F(\tau)W(\tau-t)\mathrm{d}\tau, \quad (6)$$

where $W(t)$ is a rectangular window function of width $T$ and height 1, with $T$ being set to be multiple times the dominant oscillatory period of the selected IMF.

Two issues about the procedure should be further elucidated as follows:

1. To ensure success, one needs to determine which of the decomposed IMFs actually contains pitch information. Perhaps the most simple and stable solution, as adopted here, is using manual marking in parallel with EEMD processing to extract the so-called primary pitch frequency $\hat{f}_p$ $(p=1,2,3,...)$. To ensure the resolution, $\hat{f}_p$ is interpolated by means of cubic spline, thereby yielding a smooth frequency curve, which is used as the reference $\hat{f}_{\mathrm{r}}(t)$ to locate the pitch-bearing IMFs.

2. The number of ensemble members and the amplitude of the added white noise are two crucial parameters that need to be set when the EEMD method is used. In this study, we set the ensemble number for each case to 100 in algorithm implementation, and the added white noise amplitude is set to

be 0.02 times the standard deviation of the speech signal (Lei *et al.*, 2009).

An illustration of the operation is shown in Fig. 2. Here we take as an example the decomposition of Chinese monosyllable /á/ (sampled at 16 kHz). Fig. 2a displays the time domain signal of speech /á/, and Fig. 2b shows the four IMFs ($IMF_5$–$IMF_8$) decomposed using EEMD. Fig. 2b shows that only $IMF_7$ includes the pitch information. Assuming $F(t)=\tilde{f}_7(t)$, the time varying pitch contour $F_0(t)$ is captured successfully (Fig. 2c).

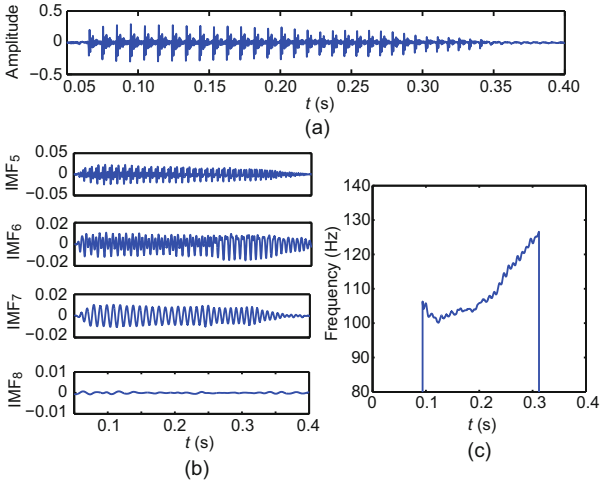


**Fig. 2 The time domain signal of speech /á/ (a), the four IMFs ($IMF_5$–$IMF_8$) decomposed using EEMD (b), and the time varying pitch detected using the proposed algorithm (c)**

## 3 Performance and comparisons

Now we attempt to apply the proposed method to practical speech signals to further demonstrate its performance and effectiveness.

The speech data used are from CLDC, a standard Mandarin Chinese speech database, whose samples, all recorded at a sampling rate of 16 000 Hz with 16-bit amplitude resolution, were acquired from 30 speakers (15 males, 15 females) in a typical quiet environment.

### 3.1 Chinese monosyllable

As we know, the monosyllable is the essential unit in Mandarin Chinese. We applied the proposed method to monosyllable processing to examine whether it could extract the time varying pitch information. Fig. 3 presents the pitch contours detected using the new scheme for two Chinese monosyllables taken from the CLDC database: (a) /á/ in Chinese, meaning 'surprise' in English and (b) /yǎo/ in Chinese, meaning 'bite' in English. Also plotted in the figure for comparison are those evaluated using the autocorrelation function (ACF) method. For /á/ (Fig. 3a), essentially the same results are obtained using the different methods, except that the new scheme greatly improves the frequency-resolving power. For the third tone /yǎo/, however, the ACF method is totally unable to capture the pitch frequency on the trough of the tone contour $f(t)$, causing a gap of vanishing pitch frequency, or provides the misleading information that would cause the confusion of the monosyllable /yǎo/ as a disyllable in tone recognition. The emergence of the gap is most likely due to the non-monotonic variation of the third tone. Nevertheless, the new method gives an encouraging result that correctly delineates the temporal variation of this tone (Fig. 3b).

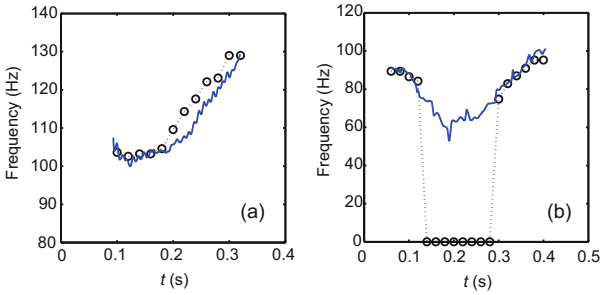


**Fig. 3 Time varying pitch of two Chinese monosyllables detected using the proposed method (solid line) and the autocorrelation function (ACF) method (dotted-line connected circles). (a) For Chinese monosyllable /*á*/; (b) For Chinese monosyllable /*yǎo*/**

To further demonstrate the effectiveness of the proposed method, intensive tests were performed for 3052 third-tone monosyllables drawn from the CLDC database, and comparisons were made with existing methods including the ACF method, the cepstrum method (CEP), the YIN pitch estimation method, a state-of-the-art frame-based pitch detector (de Cheveigne and Kawahara, 2002), the get-f0 method of ESPS (Talkin, 1995), a benchmark routine for pitch detection, and the pitch marking algorithm of Praat, a state-of-the-art event-based pitch detector (Boersma, 2002). When evaluating the candidate methods, the case in which more than 5% evaluated values differ by more than 20% from manual marking pitch estimates is counted as a 'gross

error' (de Cheveigne and Kawahara, 2002). Table 1 summarizes the gross error numbers and rates for each method. Since the proposed method can extract time varying pitch information, it performs better than other methods. In addition, the occasional erroneous detections with our method are largely ascribed to an incautious combination of pitch contours when several IMFs contain the pitch information.

**Table 1 Gross error numbers and rates over 3052 third-tone monosyllables**

| Method | Gross error number | Gross error rate |
|---|---|---|
| ACF | 1221 | 40% |
| CEP | 946 | 31% |
| YIN | 824 | 27% |
| Praat | 611 | 20% |
| get-f0 | 494 | 16% |
| Ours | 184 | 6% |

### 3.2 Chinese disyllable and continuous speech

From Section 3.1, we find that the new algorithm has advantages in the pitch detection of monosyllables, particularly in estimating the tones of non-monotonically varying pitch like the third one. As an extension, here we consider Chinese disyllable and continuous speech signal.

Fig. 4 presents the pitch contours detected using the new scheme: (a) Chinese disyllable '/mō é/' in Chinese, meaning 'touch goose' in English and (b) continuous speech signal '/tā qù wú xī shì/' in Chinese, meaning 'He went to Wuxi city' in English. For the Chinese disyllable, the new algorithm greatly improves the frequency-resolving power (Fig. 4a). The new scheme also gives an encouraging result for the continuous speech signal (Fig. 4b).

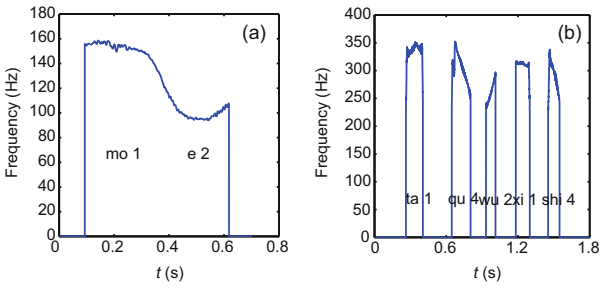


**Fig. 4 Time varying pitch of Chinese disyllable and continuous speech signal detected using the proposed method. (a) For Chinese disyllable /*mō é*/; (b) For continuous speech signal /*tā qù wú xī shì*/**

To further show the excellent performance of our method in rhythm extraction, we present in Fig. 5 the estimated tonal patterns of a Chinese ancient poetry, entitled 'Gu Lang Yue Xing' ('The Classic Bright-Moon Melody' in English), which was written by one of the most famous poets, Li Bai, who lived from 701 to 762 A.D. in the Tang Dynasty. The English translation of the poetry is: Too childish to understand the moon bright, I called it a jade-disk in white. It looks like a mirror in the paradise, flying high in the blue sky.

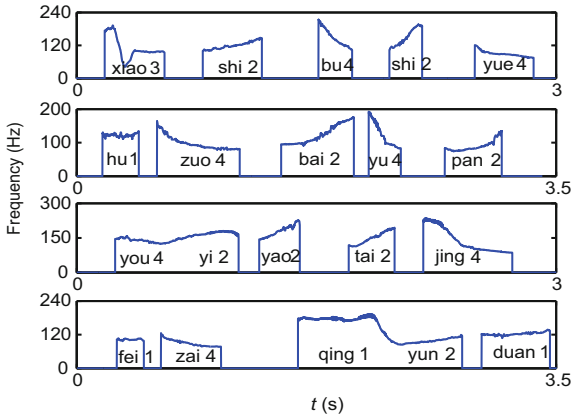


**Fig. 5 Time varying pitch of a Chinese ancient poetry from Li Bai. From the top to the bottom are the four rows of the poetry in sequence**

### 3.3 Noise robustness

An added benefit with the present algorithm is the enhanced tolerance with respect to noise disturbance. This is largely ascribed to the remarkably improved stability of EEMD. Note here that the noise in a noisy signal is irrelevant to the artificially added white noise in forming the ensemble of the signal. Unlike the original EMD, the ensemble decomposition is able to cancel out the noise effect to a large degree. The reason is that the noise in a signal will blend with the additive noise in each ensemble copy, and the noise effect in different ensemble copies counteracts each other by the ensemble average.

By contrast, both frame- and event-based pitch detectors are sensitive to noise. For instance, both ACF and CEP fail to extract the pitch information from the Chinese monosyllable /á/ that is contaminated with Gaussian noise, with SNR ≈ 3 dB, but the present method successfully detects the time varying contour (Fig. 6a). The noise-resistance ability, together with the adaptability of the EEMD

method, makes the present algorithm widely applicable in processing real speech signals with noise contamination. Fig. 6b shows the encouraging performance of the new scheme in extracting the time varying pitch for the Chinese continuous speech signal in Fig. 4b, polluted by an additive Gaussian white noise with SNR ≈ 7 dB.

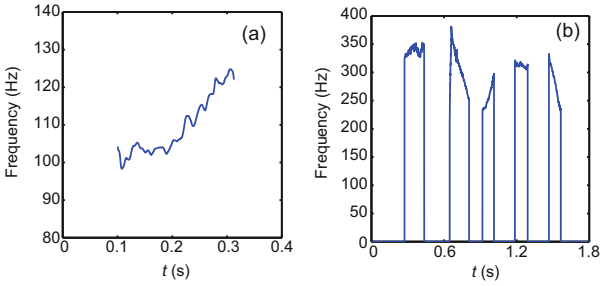


**Fig. 6 Time varying pitch detected by the proposed method from two noise-contaminated speech signals. (a) For noisy Chinese monosyllable /*á*/, with SNR ≈ 3 dB; (b) For noisy Chinese continuous speech signal /*tā qù wú xī shì*/, with SNR ≈ 7 dB**

## 4 Conclusions

In this paper, we have developed a novel pitch detector that is able to accurately detect and estimate the time variation of pitch. Extensive tests on Mandarin Chinese speech signals show that the proposed method outperforms the existing methods, this being largely ascribed to the powerful time-frequency analysis tool, ensemble empirical mode decomposition. Also, its robustness against noise interference ensures its applicability in the speech processing for tonal languages.

## References

Ananthapadmanabha, T., Yegnanarayana, B., 1975. Epoch extraction of voiced speech. *IEEE Trans. Acoust. Speech Signal Process.*, **23**(6):562-570. [doi:10.1109/TASSP.1975.1162745]

Bekara, M., Baan, M.V.D., 2009. Random and coherent noise attenuation by empirical mode decomposition. *Geophysics*, **74**(5):89-98. [doi:10.1190/1.3157244]

Boersma, P., 2002. Praat, a system for doing phonetics by computer. *Glot Int.*, **5**:341-345.

Chan, K.W., So, H.C., 2004. Accurate frequency estimation for real harmonic sinusoids. *IEEE Signal Process. Lett.*, **11**(7):609-612. [doi:10.1109/LSP.2004.830115]

Chang, E., Zhou, J., Di, S., Huang, C., Lee, K., 2000. Large Vocabulary Mandarin Speech Recognition with Different Approaches in Modeling Tones. Proc. Int. Conf. on Spoken Language Processing, p.983-986.

Cheng, Y.M., O'Shaughnessy, D., 1989. Automatic and reliable estimation of glottal closure instant and period. *IEEE Trans. Acoust. Speech Signal Process.*, **37**(12):1805-1815. [doi:10.1109/29.45529]

Christensen, M.G., Jakobsson, A., Jensen, S.H., 2007. Joint high-resolution fundamental frequency and order estimation. *IEEE Trans. Audio Speech Lang. Process.*, **15**(5):1635-1644. [doi:10.1109/TASL.2007.899267]

Christensen, M.G., Stoica, P., Jakobsson, A., Jensen, S.H., 2008. Multi-pitch estimation. *Signal Process.*, **88**(4):972-983. [doi:10.1016/j.sigpro.2007.10.014]

de Cheveigne, A., Kawahara, H., 2002. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, **111**(4):1917-1930. [doi:10.1121/1.1458024]

Deller, J., Proakis, J., Hanson, J., 1993. Discrete-Time Processing of Speech Signals. Prentice Hall, Englewood Cliffs, NJ, USA.

Goska, A., Krawiecki, A., 2006. Analysis of phase synchronization of coupled chaotic oscillators with empirical mode decomposition. *Phys. Rev. E*, **74**(4):046217. [doi:10.1103/PhysRevE.74.046217]

Hong, H., Wang, X.L., Tao, Z.Y., 2009. Local integral mean-based sifting for empirical mode decomposition. *IEEE Signal Process. Lett.*, **16**(10):841-844. [doi:10.1109/LSP.2009.2025925]

Huang, H., Pan, J., 2006. Speech pitch determination based on Hilbert-Huang transform. *Signal Process.*, **86**(4):792-803. [doi:10.1016/j.sigpro.2005.06.011]

Huang, N.E., Wu, Z., 2007. An adaptive data analysis method for nonlinear and nonstationary time series: the empirical mode decomposition and Hilbert spectral analysis. *Wavel. Anal. Appl.*, **1**(4):363-376. [doi:10.1007/978-3-7643-7778-6_25]

Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C., Liu, H.H., 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear non-stationary time series analysis. *Proc. R. Soc. Lond. A*, **454**:903-995. [doi:10.1098/rspa.1998.0193]

Huang, N.E., Shen, Z., Long, S., 1999. A new view of nonlinear water waves: the Hilbert spectrum. *Ann. Rev. Fluid Mech.*, **31**(1):417-459. [doi:10.1146/annurev.fluid.31.1.417]

Huang, N.E., Chern, C.C., Huang, K., Salvino, L.W., Long, S.R., Fan, K.L., 2001. A new spectral representation of earthquake data: Hilbert spectral analysis of Station TCU129, Chi-Chi, Taiwan, 21 September 1999. *Bull. Seismol. Soc. Am.*, **91**(5):1310-1338. [doi:10.1785/0120000735]

Jánosi, I.M., Müller, R., 2005. Empirical mode decomposition and correlation properties of long daily ozone records. *Phys. Rev. E*, **71**(5):056126. [doi:10.1103/PhysRevE.71.056126]

Kadambe, S., Boudreaux-Bartels, G.F., 1992. Application of the wavelet transform for pitch detection of speech signals. *IEEE Trans. Inf. Theory*, **38**(2):917-924. [doi:10.1109/18.119752]

Lei, Y.G., He, Z.J., Zi, Y.Y., 2009. Application of the EEMD method to rotor fault diagnosis of rotating machinery. *Mech. Syst. Signal Process.*, **23**(4):1327-1338. [doi:10.1016/j.ymssp.2008.11.005]

Li, H.B., Stoica, P., Li, J., 2000. Computationally efficient parameter estimation for harmonic sinusoidal signals. *Signal Process.*, **80**(9):1937-1944. [doi:10.1016/S0165-1684(00)00103-1]

Liang, H., Lin, Z., McCallum, R.W., 2000. Artifact reduction in electrogastrograms based on the empirical mode decomposition. *Med. Biol. Eng. Comput.*, **38**(1):35-41. [doi:10.1007/BF02344686]

Lin, S.L., Tung, P.C., Huang, N.E., 2009. Data analysis using a combination of independent component analysis and empirical mode decomposition. *Phys. Rev. E*, **79**(6):066705. [doi:10.1103/PhysRevE.79.066705]

Noll, A.M., 1967. Cepstrum pitch determination. *J. Acoust. Soc. Am.*, **41**(2):293-309. [doi:10.1121/1.1910339]

Pai, P.F., Palazotto, A.N., 2008. Detection and identification of nonlinearities by amplitude and frequency modulation analysis. *Mech. Syst. Signal Process.*, **22**(5):1107-1132. [doi:10.1016/j.ymssp.2007.11.006]

Qi, K., He, Z.J., Zi, Y.Y., 2007. Cosine window-based boundary processing method for EMD and its application in rubbing fault diagnosis application in rubbing fault diagnosis. *Mech. Syst. Signal Process.*, **21**(7):2750-2760. [doi:10.1016/j.ymssp.2007.04.007]

Resch, B., Nilsson, M., Ekman, A., Kleijn, W.B., 2007. Estimation of the instantaneous pitch of speech. *IEEE Trans. Audio Speech Lang. Process.*, **15**(3):813-822. [doi:10.1109/TASL.2006.885242]

Schlurmann, T., Dose, T., Schimmels, S., 2001. Characteristic Modes of the 'Adreanov Tsunami' Based on the Hilbert-Huang Transformation. Proc. 4th Int. Symp. on Ocean Wave Measurement and Analysis, **2**:1525-1534. [doi:10.1061/40604(273)154]

Talkin, D., 1995. A robust algorithm for pitch tracking (RAPT). *Speech Cod. Synth.*, **14**:495-518.

Wang, C., Seneff, S., 1998. A Study of Tones and Tempo in Continuous Mandarin Digit Strings and Their Application in Telephone Quality Speech Recognition. Proc. Int. Conf. on Spoken Language Processing, p.635-638.

Wu, Z., Huang, N.E., 2009. Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv. Adapt. Data Anal.*, **1**(1):1-41. [doi:10.1142/S1793536909000047]

Xu, G.L., Wang, X.T., Xu, X.G., 2009. Time-varying frequency-shifting signal-assisted empirical mode decomposition method for AM–FM signals. *Mech. Syst. Signal Process.*, **23**(8):2458-2469. [doi:10.1016/j.ymssp.2009.06.006]

Zhang, J.X., Christensen, M.G., Jensen, S.H., Moonen, M., 2010. A robust and computationally efficient subspace-based fundamental frequency estimator. *IEEE Trans. Audio Speech Lang. Process.*, **18**(3):487-497. [doi:10.1109/TASL.2010.2040786]