NOVEL APPROACH FOR FINDING PITCH MARKERS IN SPEECH SIGNAL USING ENSEMBLE EMPIRICAL MODE DECOMPOSITION

Sheenam Mehta[1] and R.S.Chauhan[2]

[1]M.Tech Scholar , [2]Astt. Proff.
Department of Electronics and Communication, J.M.I.T, Radaur, Haryana, (India)

ABSTRACT

A novel approach has been described in this paper to find pitch markers (vocal tract excitation) using ensemble empirical mode decomposition (EEMD). EEMD is the method used for time-frequency analysis for any speech signal. Using EEMD, signal decomposed into intermediate function called IMF(Intrinsic mode function). This IMF is used to extract the pitch excitation in speech signal. This paper uses IMF 4 for experimentation. To find out accurate pitch marker zero crossing points determined in IMF and after that to separate voiced, unvoiced and silence segment simple energy based threshold is applied. This proposed algorithm is giving very promising and convincing results.

Keywords: EEMD, IMF, Pitch markers

I INTRODUCTION

Speech is the output of a time-varying vocal tract system excited by a time-varying excitation. However, for analysis purpose, speech is assumed to be quasi-stationary when it is treated in blocks of 10-20 msec. Features are extracted from these blocks for further processing using signal processing techniques. Pitch marking (PM), is used to locate every vibration of the vocal chords. That is, the beginning and end of each pitch cycle is to be located by timing markers. PM does not involve classifying speech into voiced or unvoiced regions but rather may use such pre-existing knowledge for locating pitch cycle markers. Broadly there are two approaches for the analysis of speech that is, pitch-synchronous and pitch-asynchronous. In pitch-synchronous analysis, pitch markers are detected from the speech signal and are used as anchor points for further processing. Alternatively, in pitch-asynchronous analysis no such pitch markers are used for processing. Generally it has been observed that pitch-synchronous analysis gives better performance compared to pitch-asynchronous analysis [1-4].The present study focuses on developing a new method for detecting pitch markers in a computationally efficient manner.

1.1 Significance of Epochs in Speech Analysis

Voiced speech analysis consists of determining the frequency response of the vocal-tract system and the glottal pulses representing the excitation source. Although the source of excitation for voiced speech is a sequence of glottal pulses, the significant excitation of the vocal-tract system is within a glottal pulse. The significant excitation can be considered to occur at the instant of glottal closure, called the epoch. Many speech analysis

situations depend on the accurate estimation of the epoch locations within a glottal pulse. For example, knowledge of the epoch locations is useful for accurate estimation of the fundamental frequency (f_0). Other potential applications of the markings of pitch period markers include analysis of jitter, prosody in speech [5], text-to-speech synthesis [6,7], analysis of voice quality and pitch synchronous speech analysis [8].

1.2 Review of the Existing Methods

Normally, pitch markers are associated to the glottal closure instants (GCIs) of the glottal cycles. Most pitch marker extraction methods rely on the error signal derived from the speech waveform after removing the predictable portion (second-order correlations). The error signal is usually derived by performing linear prediction (LP) analysis of the speech signal [9]. The first contribution to the detection of epochs was due to Sobakin [10]. A slightly modified version was proposed by Strube [11]. In Strube's work, some predictor methods based on LP analysis for the determination of the pitch markers were reviewed. Most of pitch marker determination methods are based on autocorrelation function Autocorrelation method [17], Cepstral method [18], AMDF [19], etc. But, all of these techniques face a few or all of these problems- windowing effect, low time resolution, low frequency resolution, etc. Later on Group delay based method [12,13] and zero frequency resonator based method developed [23,24]. Except zero frequency resonator based method all are short term processing. Only zero frequency resonator based algorithm can use on long duration signal.

This paper work is an attempt to get rid of a few or all of these shortcomings. We can use Empirical Mode Decomposition (EMD) [20] to find the instantaneous pitch. The idea is that one of the Intrinsic Mode Frequencies (IMFs) contains the pitch information. To make sure that there is a unique IMF containing the pitch information, we need to get rid of "Mode-mixing" [22]. This problem can be solve by Ensemble Empirical Mode Decomposition (EEMD) [21].

New proposed method for finding pitch markers using EEMD can be apply on long duration signal (upto 1 sec.) and determine the pitch markers in very good manner as good as other method used for pitch markers.

II BASIS FOR PROPOSED PITCH MARKER METHOD

EMD algorithm has been recently proposed by Huang [14] for adaptively decomposing nonlinear and non stationary signals into a sum of well-behaved AM - FM components, called Intrinsic Mode Functions. This new technique has received the attention of the scientific community, both in its understanding and application. EMD based algorithms suffer the well-known "mode mixing" problem and they use a set of post-processing rules with the intention of alleviate it. The mode mixing is perhaps the major drawback of the original EMD. This effect is defined as a single IMF either consisting of signals of widely disparate scales (energies), or a signal of a similar scale residing in different IMF components [15]. Wu and Huang [15] proposed a modification to the EMD algorithm. This new method, called Ensemble Empirical Mode Decomposition (EEMD), largely alleviates the mode mixing effect.

2.1 Ensemble Empirical Mode Decomposition

Ensemble Empirical Mode Decomposition (EEMD) approach consists of sifting [25] an ensemble of white noise-added signal and treats the mean as the final true result. Finite, not infinitesimal, amplitude white noise is necessary to force the ensemble to exhaust all possible solutions in the sifting process, thus making the different

scale signals to collate in the proper intrinsic mode functions (IMF) dictated by the dyadic filter banks. As the EMD is a time space analysis method, the white noise is averaged out with sufficient number of trials; the only persistent part survives the averaging process is the signal, which is then treated as the true and more physical meaningful answer. The effect of the added white noise is to provide a uniform reference frame in the time-frequency space; therefore, the added noise collates the portion of the signal of comparable scale in one IMF. With this ensemble mean, one can separate scales naturally without any *a priori* subjective criterion selection as in the intermittence test for the original EMD algorithm. This new approach utilizes the full advantage of the statistical characteristics of white noise to perturb the signal in its true solution neighborhood, and to cancel itself out after serving its purpose; therefore, it represents a substantial improvement over the original EMD.

2.2 EMD Algorithm

The standard EMD algorithm was derived using following steps [15]:

(1) Identify all the extreme (maxima and minima) peaks of the signal (DC component of signal was removed before preprocessing), s(t).

(2) Generate the upper and lower envelope by the cubic spline interpolation of the extreme peaks developed in step (1).

(3) Calculate the mean function of the upper and lower envelope, *m(t).*

(4) Calculate the difference signal, *d(t)=s(t)-m(t).*

(5) If d(t) becomes a zero-mean process, then the iteration is stopped and *d(t)* is considered as the first IMF, named $c_1(t)$; otherwise, go to step (1) and replace s(t) with *d(t).*

(6) Calculate the residue signal, *r(t)=s(t)-* $c_1(t)$

(7) Repeat the procedure from steps (1) to (6) to obtain the second IMF, named $c_2(t)$. To obtain $c_n(t)$ continue the steps (1)–(6) after n iterations. The process is stopped when the final residual signal, *r(t)*, is obtained as a monotonic function.

At the end of the procedure, a residue $r(t)$ and a collection of n IMF were derived and named from $c_1(t)$ to $c_n(t)$. Hence, the original signal can be represented as:

$$s(t) = \sum_{i=1}^{n} c_i(t) + r(t) \qquad (1)$$

where $r(t)$ is often regarded as $c_{n+1}(t)$.

The low IMF scales were mainly the high-frequency components of signal, while the high IMF scales were the low-frequency components of signal. Thus, an EMD-based low-pass filter was developed using the partial reconstruction of the selected IMF scale, which is given as:

$$REMD_k = \sum_{i=k}^{n+1} c_i(t) \qquad (2)$$

When k =1, the $REMD_1$ was equivalent to the original noise-contaminated ECG.

2.3 EEMD Algorithm

The EEMD algorithm is as follows [16]:

(1) Add a white-noise series, *n(t)*, to the targeted signal, *x(t)*, in the following description, $x_1(t)=x(t)+n(t)$. The added noise power from 5 to 25 dB was used to investigate the EEMD performance.

(2) Decompose the data $x_1(t)$ using the EMD algorithm, as described above.

(3) Repeat Steps (1) and (2) until the pre-set trial numbers, each time with different added white-noise series of the same power. The new IMF combination $c_{ij}(t)$ is achieved, where *i* is the iteration number and *j* is the IMF scale.

(4) Estimate the mean (ensemble) of the final IMF of the decompositions as the desired output.

$$EEMD_{c_j(t)} = \frac{\sum_{i=1}^{nt} c_{ij}(t)}{nt}$$

where *nt* denotes the trial numbers. Similar to EMD, an EEMD-based partial reconstruction of ensemble IMF can be defined as:

$$REEMD_k = \sum_{j=k}^{n+1} EEMD_{c_j(t)}$$

This method to determine IMF using EEMD is applied on a small segment of speech signal. The resultant IMFs are shown in Figure 1.

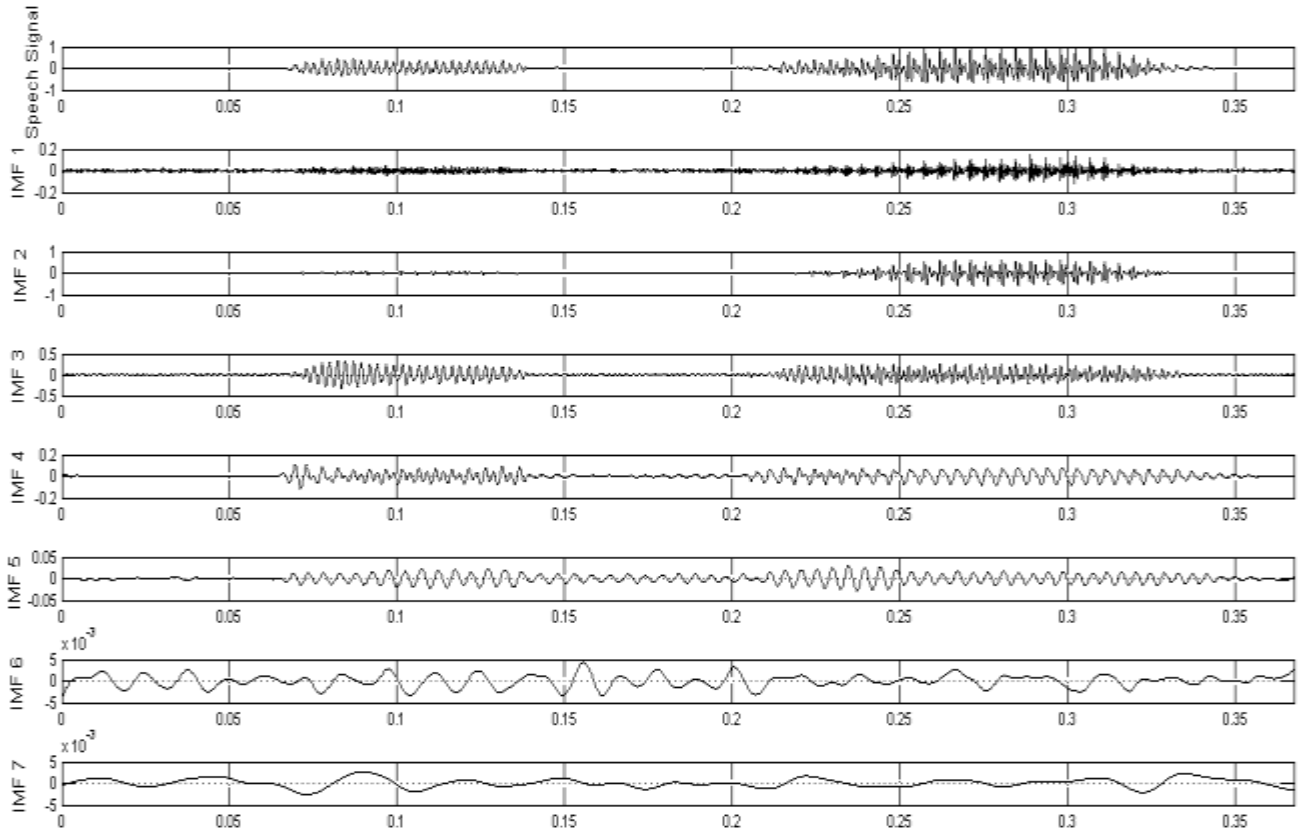

Figure 1: Speech signal and its corresponding IMFs

Now, the idea is that one of the Intrinsic Mode Frequencies contains the pitch information. The IMF having the highest energy is proposed as the IMF containing the pitch information. The amplitude of the Filtered IMF4 is high in the voiced region and is close to zero in the non-voiced part as shown in Figure 2. It can be observed in Figure 3, the plots of the speech signal, its IMF3, IMF4 that IMF 4 contains the pitch information, has the highest fraction of energy, lowest fluctuation and irregularity in the instantaneous frequency. These fractions also represent the confidence of the IMF chosen. The fraction should be as large as possible for the IMF that will be chosen and as low as possible for others. The fourth IMF is almost the full signal, which can produce a sound that is clear and with almost the original audio quality. All other components are also regular and have comparable and uniform scales and amplitudes for each respective IMF component, but the sounds produced by them are not intelligible, they mostly consist of either high frequency hissing or low frequency moaning. The results once again clearly demonstrate that the EEMD has the capability of catching the essence of data that manifests the underlying physics.

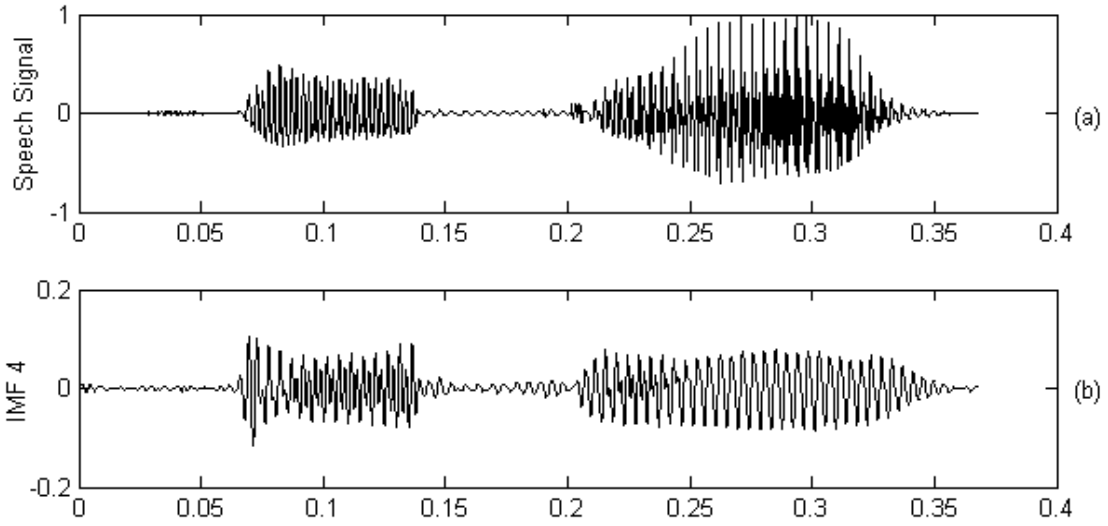

Figure 2: Sample speech signal and its corresponding IMF 4

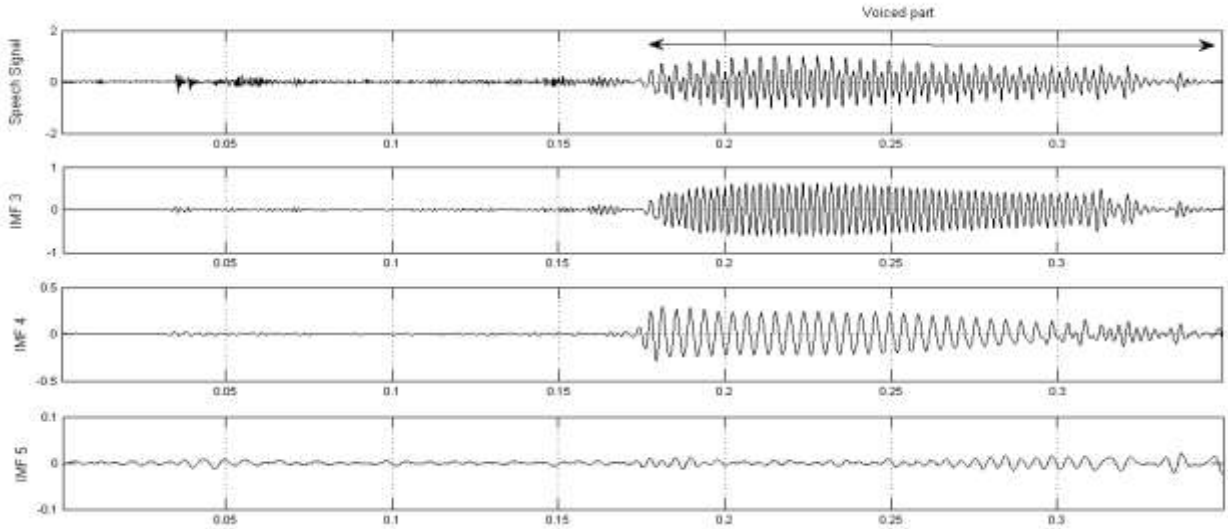

Figure 3: Comparison between IMF 3, IMF 4 and IMF 5

2.4 Finding Pitch Markers Using EEMD

Ensemble Empirical Mode Decomposition is a noise assisted data analysis to take care of mode-mixing. A white Gaussian noise is added to the input speech signal to avoid mode mixing. The same experiment is repeated N (>>1) times using N different sequences of noise. The corresponding IMFs from these N experiments are added. Because, the noise is random, it becomes negligible compared to the signal. Hence, we get only the signal component, ideally. We can thus avoid mode-mixing in Empirical Mode Decomposition. To determine the pitch markers in a speech signal using EEMD, the algorithm can be described as:

Step 1: Initially low pass filter is applied to the sample speech signal with the purpose of eliminating spurious frequency components .This filter is centered in the frequency 0-4khz.

Step 2: EEMD method has been used to decompose the filtered signal into a finite and often small number of frequency modes called Intrinsic Mode Functions (IMF). It defines the true IMF components as the mean of certain ensemble of trials, each one obtained by adding white noise of finite variance to the original signal.

Step 3: Select the IMF having the highest energy , proposed as the IMF containing the pitch information. It can be observed that IMF 4 contains the pitch information, has the highest fraction of energy, lowest fluctuation and irregularity in the instantaneous frequency.

Step 4: Find out zero-crossings in the selected IMF. The zero-crossings accompanied by positive to negative transition are detected as the candidates for pitch markers. For convenience, the positive going zero crossings has been used in this study.

Step 5: Some of the detected zero-crossings may also correspond to excitations like glottal openings in voiced speech and burst and frication in unvoiced speech and these are unwanted. To determine the desired zero crossings for finding the locations of the pitch markers, search back process is applied to the detected zero crossings.

Step 6: Threshold is then applied to the signal to locate the desired pitch markers and to eliminate the unwanted zero crossings from the silent and unvoiced part.

The proposed algorithm has been shown in the form of a block diagram in the Figure 4 according the steps described above.

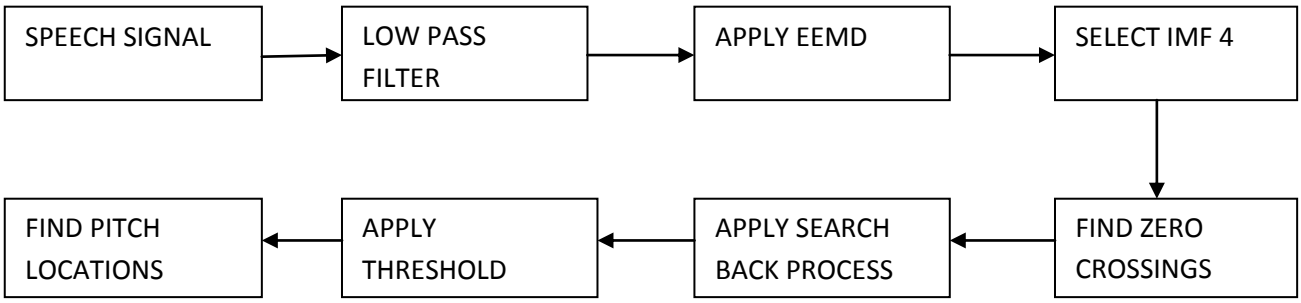

Figure 4: Block Diagram for proposed algorithms

III RESULT AND DISCUSSION

3.1 Experimental Setting

According to the principle of the EEMD, the added white noise would populate the whole time-frequency space uniformly with the constituting components of different scales separated by the filter bank. When signal is

added to this uniformly distributed white background, the bits of signal of different scales are automatically projected onto proper scales of reference established by the white noise in the background. Of course, each individual trial may produce very noisy results, for each of the noise-added decompositions consists of the signal and the added white noise. Since the noise in each trial is different in separate trials, it is canceled out in the ensemble mean of enough trails. The ensemble mean is treated as the true answer, for, in the end, the only persistent part is the signal as more and more trials are added in the ensemble. In this study, the noise standard deviation used is 1.5 and ensemble size is 1000 i.e. no. of trials. These both parameters can vary upto their right combination. The noise standard deviation can vary from 0.2 to 2.5 or so on as per the no. of trials gives the appropriate results.

3.2 Implementation of proposed algorithm

EEMD method has been used to decompose the filtered signal into a finite and often small number of frequency modes called Intrinsic Mode Functions (IMF). It defines the true IMF components as the mean of certain ensemble of trials, each one obtained by adding white noise of finite variance to the original signal.IMF having the highest energy , proposed as the IMF containing the pitch information. It can be observed that IMF 4 contains the pitch information, has the highest fraction of energy, lowest fluctuation and irregularity in the instantaneous frequency. The zero-crossings accompanied by positive to negative transition are detected as the candidates for pitch markers. For convenience, the positive going zero crossings has been used in this study.

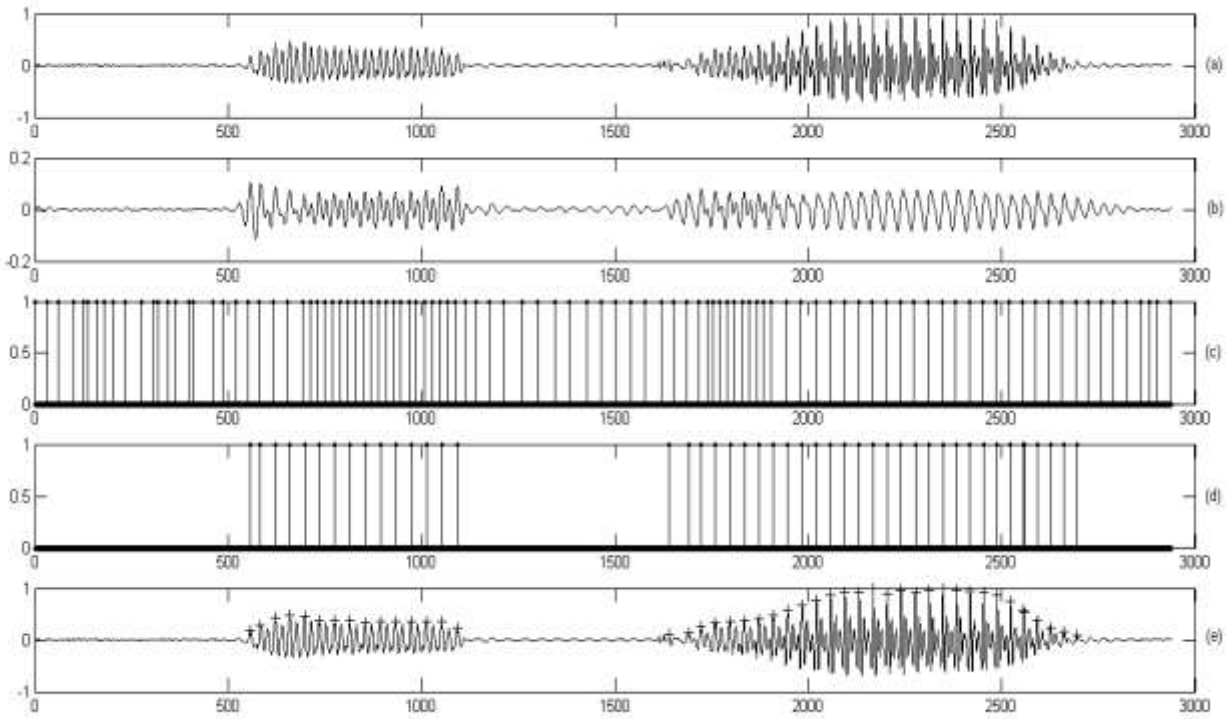

Figure 5: Results from proposed algorithm for detection of pitch markers (a) A segment of speech signal, (b) corresponding IMF 4 of speech signal, (c) zero crossing points in IMF signal, (d) zero crossing points after applying threshold, and (e) pitch marker points corresponding speech segment.

Some of the detected zero-crossings may also correspond to excitations like glottal openings in voiced speech and burst and frication in unvoiced speech and these are unwanted. To determine the desired zero crossings for finding the locations of the pitch markers, search back process is applied to the detected zero crossings. Threshold is then applied to the signal to locate the desired pitch markers and to eliminate the unwanted zero crossings from the silent and unvoiced part. The result obtained by the proposed algorithm has been shown in the Figure 5.

IV CONCLUSION

This paper proposed a novel and effective approach for determining pitch markers in speech signal which operates using the Ensemble Empirical Mode Decomposition (EEMD) technique. The real data with a comparable scale can find a natural location to reside. The EEMD utilizes all the statistical characteristic of the noise: It helps to perturb the signal and enable the EMD algorithm to visit all possible solutions in the finite (not infinitesimal) neighborhood of the true final answer; it also takes advantage of the zero mean of the noise to cancel out this noise background once it has served its function of providing the uniformly distributed frame of scales, a feat only possible in the time domain data analysis. In a way, this new approach is essentially a controlled repeated experiment to produce an ensemble mean for a non-stationary data as the final answer. Since the role of the added noise in the EEMD is to facilitate the separation of different scales of the inputted data without a real contribution to the IMFs of the data, the EEMD is a truly noise-assisted data analysis (NADA) method that is effective in extracting signals from the data. The truth defined by EEMD is given by the number in the ensemble approaching infinity. But the number of the trials in the ensemble, ***N***, has to be large. It is concluded that the EEMD indeed represents a major improvement over the original EMD. As the level of added noise is not of critical importance, as long as it is of finite amplitude to enable a fair ensemble of all the possibilities, the EEMD can be used without any subjective intervention; thus, it provides a truly adaptive data analysis method. By eliminating the problem of mode mixing, it also produces a set of IMFs that bears the full physical meaning and a time-frequency distribution without transitional gaps.It is concluded that the EMD, with the ensemble approach, may be a more mature tool for nonlinear and non-stationary time series (and other one dimensional data) analysis.

REFERENCES

[1] A. K. Krishnamurthy and D. G. Childers, "Two-channel speech analysis," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-34, pp. 730-743, Aug. 1986.

[2] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-27, pp. 350-355, Aug 1979.

[3] B. Yegnanarayana and N. J. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," IEEE Trans. Speech Audio Processing, vol. 6, pp. 313-327, July 1998.

[4]. S.Harbeck , A. Kiebling, R. Kompe , H. Niemann and E. Nöth , “Robust pitch period detection using dynamic programming with an ANN cost function”, Proc. EUROSPEECH, Madrid, vol. 2, pp. 1337-1340, September 1995.

[5]. V.Colotte and Y Laprie, “Higher precision pitch marking for TD-PSOLA”, Proceedings of XI European Signal ProcessingConference (EUSIPCO), Toulouse, 2002.

[6]. Laprie, Yves and Colotte, Vincent, "Automatic pitch marking for speech transformations via TD-PSOLA", European Signal Processing Conference (EUSIPCO), Rhodes, 1998.

[7]. E. Moulines and F, Charpentier., "Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis Using Diphones", Speech Communication, 9: 453-467, 1990.

[8] J. E. Markel and A. H. Gray, Linear Prediction of Speech. New York: Springer-Verlag, 1982

[9] A. N. Sobakin, "Digital computer determination of formant parameters of the vocal tract from a speech signal," Soviet Phys.-Acoust., vol. 18, pp. 84–90, 1972.

[10] K. Rao, S. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using hilbert envelope and group delay function," IEEE Signal Process. Letters, vol. 14, no. 10, pp. 762 –765, 2007.

[11] S. Prasanna and A. Subramanian, "Finding pitch markers using first order gaussian differentiator," in Third Int. Conf. on Intelligent Sensing and Inf. Process., 2005, pp. 140–145.

[12] N.E. Huang, Z. Shen, S.R. Long, Wu, M. C., Shih, E. H., Zheng, Q., Tung, C. C., Liu, H. H.: The empirical mode decomposition method and the Hilbert spectrum for non-stationary time series analysis, Proc. Royal Society London 454A, 1998, p. 903–995.

[13] Z.Wu, N.E. Huang, (2004). "A study of the characteristics of white noise using the empirical mode decomposition method", Proceedings of the Royal Society A, 460, 1597–1611.

[14] Z. Wu and N.E. Huang. "Ensemble Empirical Mode decomposition: a noise-assisted data analysis method". Advances in Adaptive Data Analysis, vol. 1, pp. 1-41, 2009.

[15] Lawrence R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. Assp-25, no. 1, February 1977.

[16] A.M. Noll, Cepstrum pitch determination, J. Acoust. Soc. Amer. 41 (2) (1967) 293–309.

[17] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J.Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Processing*,vol. ASSP-22, pp. 353-362, Oct. 1974.

[18] 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.

[19] G. Schlotthauer, M. E. Torres, and H. L. Rufiner, "A new algorithm for instantaneous F0 speech extraction based on ensemble empirical mode decomposition," in Proc. *European Signal Processing Conference*, Glasgow, Scotland, August 2009.

[20] G. Schlotthauer, M. E. Torres, and H. L. Rufiner, "Voice fundamental frequency extraction algorithm based on ensemble empirical mode decomposition and entropies," in Proc. *11th Int. Congr. of the IFMBE*, Munich, 2009, pp. 984–987.

[21] L. R. Rabiner, M. J. Cheng, A. H. Rosenberg and C. A. McGonegal. "A comparative performance study of several pitch detection algorithms". IEEE Trans. Acoust., Speech, Signal Processing, 24(5): 399-417, 1976.

[22] B. Yegnanarayana and K.Sri Rama Murty, "Event-based instantaneous fundamental frequency estimation from speech signals", IEEE Trans. Audio, Speech and Language Processing, Vol.17, No.4, May 2009.

[23] J.D. Markel, "The SIFT algorithm for fundamental frequency estimation", IEEE Trans. Audio Electroacoust. AU-20 (1972) 367– 377.