interspeech 2008 incorporating SST 08

Robust Voiced/Unvoiced Speech Classification using Empirical Mode Decomposition and Periodic Correlation Model

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

[1]Graduate School of Information Science and Technology, [2]Graduate School of Engineering
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[1]{molla, hirose}@gavo.t.u-tokyo.ac.jp, [2]mine@gavo.t.u-tokyo.ac.jp

10.21437/Interspeech.2008-627

Abstract

This paper presents a method of voiced/unvoiced (V/Uv) classification of noisy speech signals. Empirical mode decomposition (EMD), a newly developed tool to analyze nonlinear and non-stationary signals is used to filter the additive noise with the speech signal. The normalized autocorrelation of the filtered speech signal is computed to enhance the periodicity if any. It is considered that the voiced speech signal is periodically correlated and the unvoiced signal is not. A statistical model of determining periodic correlation is used to differentiate voiced and unvoiced speech with low SNR. The experimental results show that the use of EMD improves the classification performance and the overall efficiency is noticeable as compared to other existing algorithms.

Index Terms: empirical mode decomposition, normalized autocorrelation, periodic correlation, voiced/unvoiced speech

1. Introduction

Reliable classification of short time speech signal into voiced and unvoiced is a crucial preprocessing step in many speech processing applications and is essential in most analysis and synthesis system. For example: different strategy could be adopted for voiced and unvoiced parts in speech enhancement using spectral subtraction. The essence of classification is to determine whether the speech production system involves the vibration of the vocal cords [1]. The discrimination problem is an important one and has been worked on extensively during the last three decades [2].

The discrimination can effectively be performed using a single feature or parameter which is closely associated with the voicing and non-voicing activities of speech signal. Many algorithms have been reported for solving the detection problem [3] – [7]. In [3], Gaussian mixture model with cepstrum coefficients features is proposed for robust V/Uv classification. A higher order statistics (HOS) based method is proposed in [4] for V/Uv detection and pitch estimation simultaneously. The matching pursuit algorithm is used in [5] with Gabor decomposition. The wavelet transform is proposed in pitch and V/Uv detection in [6]. A statistical model applied in autocorrelation domain is also reported in [7]. In most of the existing algorithms are not so much noise robust and also the intensive threshold and training data are required for classification. Such requirements are troublesome for the use in application domain.

The proposed method is noise robust and based on the statistical model for periodicity detection in speech signal without any training requirement. To reduce the effect of noise on speech signal, a data adaptive time domain filtering is proposed using newly developed empirical mode decomposition method [8]. Although speech signal is non-stationary in nature, Fourier based frequency domain filtering assumes that it is piecewise stationary. The speech decomposition is performed by fitting some predefined bases without satisfying its non-stationary nature. Whereas, EMD based approach decompose the speech signal as non-stationary time series and hence better performance in noise filtering.

A method for determining whether an observed time series contains a periodically correlated sequence is employed here. It is based on the statistical tests for the coherence between spectral components for the presence of a periodically correlated covariance structure in a time series [9]. The autocorrelation function (ACF) makes the periodicity more prominent if any. The proposed periodic correlation model is applied in the autocorrelation domain rather than original time domain of the speech signal. The periodicity detection method is implemented in spectral domain to classify the speech segment into voiced or unvoiced one based on that it contains periodic correlated sequence or not respectively.

2. Noise filtering using EMD

The principle of EMD technique is to decompose any signal $s(t)$ into a set of band-limited functions $C_n(t)$, which are zero mean oscillating components, simply called the 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 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 [8]. The first condition is similar to the narrow-band requirement for a 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. With this definition, the IMF in each cycle, defined by the zero crossings, involves only one mode of oscillation, no complex riding waves are allowed [8]. IMF is not restricted to a narrow-band signal; it can be both amplitude and frequency modulated, in fact it can be non-stationary.

The idea of finding the IMFs relies on subtracting the highest oscillating components from the data with a step by step process. Although a mathematical model has not been developed yet, different methods for computing EMD have been proposed after its introduction [10, 11]. The very first algorithm is called the sifting process. The sifting process is simple and elegant. It includes the following steps:

1. Identify the extrema (maxima and minima) of $s(t)$
2. Generate the upper and lower envelopes ($u(t)$ and $l(t)$) by connecting the maxima and minima points by cubic spline interpolation

3. Determine the local mean $m_1(t)=[u(t)+l(t)]/2$
4. Since IMF should have zero local mean, subtract out $m_1(t)$ from $s(t)$ to get $h_1(t)$
5. Check whether $h_1(t)$ is an IMF or not
6. If not, use $h_1(t)$ as the new data and repeat steps 1 to 6 until ending up with an IMF

Once the first IMF $h_1(t)$ is derived, it is defined as $C_1(t)=h_1(t)$, which is the smallest temporal scale in $s(t)$. To compute the remaining IMFs, $C_1(t)$ is subtracted from the original data to get the residue signal $r_1(t)$: $r_1(t)=s(t)-C_1(t)$. The residue now contains the information about the components of longer periods. 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 [12]. The subsequent IMFs and the residues are computed as:

$$r_1(t)-C_2(t)=r_2(t),...,r_{B-1}(t)-C_B(t)=r_B(t) \qquad (1)$$

At the end of the decomposition, the data $s(t)$ will be represented as a sum of n IMF signals plus a residue signal,

$$s(t)=\sum_{b=1}^{B} C_b(t)+r_B(t) \qquad (2)$$

A noisy speech signal and some selected IMF components are shown in Figure 1. It can be observed that higher order IMFs contain lower frequency oscillations than that of lower order IMFs. This is reasonable, since sifting process is based on the idea of subtracting the component with the longest period from the data till an IMF is obtained. Therefore the first IMF will have the highest oscillating components; the components with the highest frequencies. Consequently, the higher the order of the IMF, the lower its frequency content will be.

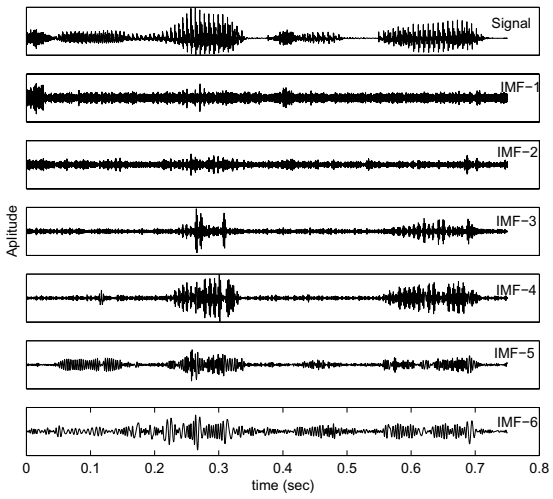

Figure 1: *The illustration of EMD. A noisy speech signal at 10 dB SNR and its first 6 IMFs out of 13.*

2.1. Instantaneous frequency

Instantaneous frequency (IF) represents signal's frequency at any time instance and it is defined as the rate of change of the phase angle at the instant of the "analytic" version of the signal. Every IMF is a real valued signal. The discrete Hilbert transform (HT) denoted by $H_d[.]$ is used to compute the analytic signal for an IMF. Then the analytic version of the b^{th} IMF $C_b(t)$ is defined as:

$$z_b(t)=C_b(t)+jH_d[C_b(t)]=a_b(t)e^{j\theta_b(t)} \qquad (3)$$

where $a_b(t)$ and $\theta_b(t)$ are instantaneous amplitude and phase respectively of the b^{th} IMF. The analytic signal is advantageous in determining the instantaneous quantities such as energy, phase and frequency. The discrete-time IF of b^{th} IMF is then given as the derivative of the phase $\theta_b(t)$ calculated at t i.e.

$$f_b(t)=\frac{d\tilde{\theta}_b(t)}{dt} \qquad (4)$$

where $\tilde{\theta}_b(t)$ represents the unwrapped version of instantaneous phase $\theta_b(t)$. The derivative in Eq. (4), is evaluated at discrete instant of t. It should be noted that such derivative introduces the abrupt fluctuations of IF and hence nonlinear smoothing is required. Here, the moving average smoothing filtering is used to remove such fluctuations. The filtering scheme improves the effectiveness of computing IF using discrete derivative. The concept of IF is physically meaningful only when applied to mono-component signals. In order to apply the concept of IF to arbitrary signals it is necessary to decompose the signal into a series of mono-component contributions. In the recent approaches [12], EMD technique decomposes a time domain signal into a series of mono-component IMFs. Then the IF derived for each component provides the meaningful physical information.

2.2. Noise filtering

Although the IMFs may have frequency overlaps but at any time instant, the instantaneous frequencies represented by each IMF are different. This phenomenon can be well understood in Figure 2 which shows the instantaneous frequencies of the first 6 IMFs. Therefore, EMD is an effective decomposition of non-linear and non-stationary signals in terms of their local frequency characteristics.

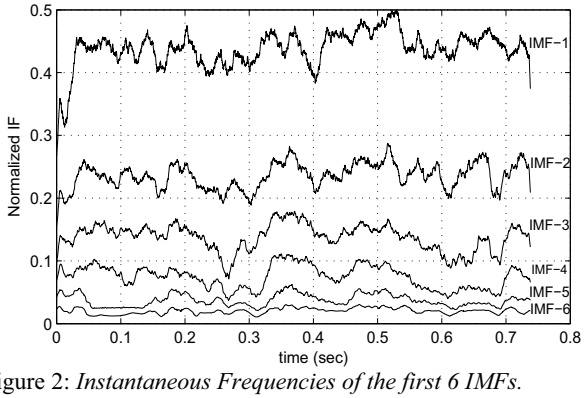

Figure 2: *Instantaneous Frequencies of the first 6 IMFs.*

With these powerful characteristics, recent studies have shown that it is possible to successfully identify and remove a significant amount of the noise components from the IMFs of a noisy speech. Although all IMFs contain energy from both the original speech and the noise, the amount of the energy distribution is different. Since speech signals are mainly concentrated in the low and mid frequency bands, the high frequency noise components dominate the first IMFs. For instance, in case of white noise, most of the noise components are centered on the first two IMFs, while the speech signals dominate between 3rd and 6th IMFs, as can be observed in

Figure 1. Therefore, EMD makes it possible to some extent separate the high frequency noise from the major speech components. The instantaneous frequency vector is normalized between 0 and 0.50 to align with the Nyquist frequency. Then the IMFs with higher frequencies (>1.5kHz) are discarded. Thus most of the high frequency noise will be filtered out. The rest of the IMFs (including residue) is summed up to reconstruct the speech signal with less noise which is termed here as pre-filtered noisy speech (PFNS) what will be processed to detect the periodicity.

3. Detecting periodicity

Periodically correlated (PC) processes are non-stationary but possess many of the properties of stationary processes. Hence, the attempt to apply the model of PC processes is suitable in determining the presence of periodicity in the speech signal. The PC processes exhibit a nature of cyclostationarity which has led to its use in many signal processing applications [13]. In frequency domain the standard tool for detecting the periodicity is the periodogram defined as

$$S_N(\omega_j) = \frac{1}{2\pi N} |F_N(\omega_j)|^2, \tag{5}$$

where $F_N(\omega_j)$ is the discrete Fourier transform and $\omega_j = j/N$, j=1,…,N are the frequencies.

3.1. Statistical model

In many PC processes the periodogram fails to detect the periodicity whereas, the sample coherence can correctly detect that [9]. The sample coherence is defined as

$$|\zeta(p,q,M)|^2 = \frac{\left|\sum_{m=0}^{M-1} F_N(\omega_{p+m})\overline{F_N(\omega_{q+m})}\right|^2}{\sum_{m=0}^{M-1}|F_N(\omega_{p+m})|^2 \sum_{m=0}^{M-1}|F_N(\omega_{q+m})|^2} \tag{6}$$

where 0<p, q≤N, N is the sample length and M is the smoothness coefficient. Sample coherence takes only real values between 0 and 1.

Two tools – coherent and incoherent statistics are proposed in [9] for determining the presence of periodic correlation. The former one is defined as $|\zeta(0,\tau,M)|^2$, i.e. the sample coherence given by Eq. (6), evaluated for $N = M$, whereas the later one is given by

$$\delta(\tau,M) = \frac{1}{L+1}\sum_{p=0}^{L}\left|\zeta(pM, pM+\tau, M)\right|^2 \tag{7}$$

where L = [N-1-τ]/M] and τ = |q-p|. Since both statistics depend on the differences between frequencies, the plots against to τ (or ω_τ) are the most indicative. The statistics are plotted only in the interval (0, N/2), because the values in the interval (N/2, N) are the mirror image of the values in the former one. Peaks at points ω_τ, $\omega_{2\tau}$, $\omega_{3\tau}$ etc. indicate the presence of periodic correlation. The coherent and incoherent statistics are, in general, much better than the periodogram but it is also still fails to detect some PC processes.

To enhance the moderate efficiency of coherent and incoherent statistics, the measure of fitness (MoF) statistics is proposed in based on the bootstrap methodology [14] and is defined by

$$\beta(\tau,M) = \frac{1}{N}\sum_{p=1}^{N}\eta_\alpha(p, p+\tau, M) \tag{8}$$

where

$$\eta_\alpha(p,q,M) = \begin{cases} 1, & |\zeta(p,q,M)|^2 \geq c_\alpha \\ 0, & |\zeta(p,q,M)|^2 < c_\alpha \end{cases} \tag{9}$$

α is the confidence level and c_α is the estimator of the critical value using moving blocks bootstrap procedure. The MoF statistics β(.) takes real values in the interval [0, 1] and due to the symmetry is plotted only in the interval (0, N/2). The peaks at points ω_τ, $\omega_{2\tau}$, $\omega_{3\tau}$ etc. indicate the presence of periodic correlation. What distinguishes the MoF statistics from the former two is the summation scheme in which only the significant of sample coherence are used. It is not the values of the sample coherence that is important but its value relative to values at other frequencies. Thus the MoF statistic detects the periodic correlation even for the processes exhibiting extreme volatility.

3.2. Autocorrelation based implementation

The proposed statistical model is implemented on the normalized autocorrelation (NACF) of the speech signal rather than in time domain. The noisy speech pre-filtered to reduce the noise effect hence obtaining the PFNS signal $\varphi(n)$, 0≤n≤N-1 is used in periodicity detection.

The NACF produces better results in V/Uv detection than the simple autocorrelation function as the peaks are more prominent and the less affected by the rapid variation in the signal amplitude. In this paper, the NACF of the PFNS signal $\varphi(n)$, 0≤n≤N-1 is used to detect the presence of periodicity and is computed as

$$NACF\ (k) = \frac{1}{\sqrt{\lambda_0\lambda_k}}\sum_{n=1}^{N-1}\varphi(n)\varphi(n+k), \tag{10}$$

where

$$\lambda_k = \sum_{n=k}^{k+N-K}\varphi^2(n),\ 0 \leq k \leq K\text{-}1. \tag{11}$$

Even with the aforementioned pre-processing, the periodicity determination with NACF may give erroneous results under strong noisy condition due to the presence of spurious peaks obscuring the actual prominent peaks and also due to the inherent shortcoming introduced with the NACF. The NACF functions of noisy voiced speech signal and its pre-filtered one are shown in Figure 3. It is observed that the PFNS produces more prominent peaks in the NACF domain. The coherent, incoherent and MoF statistics of the NACF of voiced speech (PFNS) are illustrated in Figure 4.

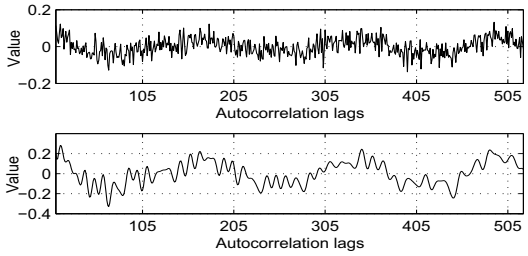

Figure 3: *NACF of noisy speech signal (upper) and of PFNS signal (lower one).*

It is observed that only MoF statistic can really detect the presence of periodicity of the voiced speech signal. The presence of peaks is confirmed using robust outlier detection technique [15]. The presence of peaks with constant duration

proves that the speech signal contains PC process and hence it is a voiced segment and unvoiced otherwise.

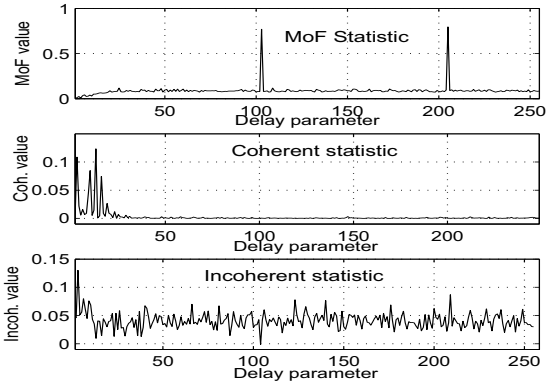

Figure 4: *Different statistics of NACF of voiced speech signal.*

4. Experimental results and discussion

The performance of the proposed method is evaluated by using speech data taken from TIMIT database. The speech material used in this experiment is re-sampled to 20 kHz and segmented into frames of length 30ms with 10ms shifting. 16 bit resolution. Approximately 2010 frames including male and female speech are used. Each frame is accurately labeled for voiced and unvoiced. The error rates are compared for two criterions – with EMD based noise filtering (nfEMD) and without noise filtering (WnF) for different noise levels. The white Gaussian noise is added to obtain different levels of segmental SNR (SSNR). Voiced – to – unvoiced (V-Uv) and unvoiced – to – voiced (Uv-V) error rates denote the accuracy in correctly classifying voiced/unvoiced speech frames. A Uv-V error occurs when an unvoiced frame is classified erroneously as voiced, and a V-Uv error occurs a voiced frame is detected as unvoiced. The overall error rate is obtained by summing up the two error factors. The performance of the proposed method for different SSNRs is shown in Table 1:

Table 1: *Performance of the proposed method with nfEMD and WnF as a function of different SSNR*

SSNR	nfEMD (%)			WnF (%)		
(dB)	V-Uv	Uv-V	Overall	V-Uv	Uv-V	Overall
Clean	0.65	0.33	0.98	0.69	0.38	1.07
10	0.74	0.68	1.42	1.07	0.85	1.92
0	2.92	1.31	4.23	4.34	3.29	7.63
-5	4.31	2.63	6.94	6.47	4.55	11.02
-10	5.43	3.98	9.41	8.02	7.11	15.13

The performance is evaluated under noisy conditions for a wide range of segmental SNRs. The overall classification accuracy with EMD based filtering method is always better than the existing reported algorithms [2]-[5]. With cepstrum-based modified algorithm [2], the overall error is reported as 6.16% for 0dB SSNR and no result is produced for SSNR less than 0dB. Gaussian mixture model (GMM) with cepstral features is proposed in [3] with 8% error for 15dB SNR. In [4], higher order statistics (HOS) based method is employed in V/Uv classification for low SNR (up to 0dB) but no quantitative result is reported. Gabor atomic decomposition method is proposed in [5] with 16% error rate for 5dB SNR speech. Based on the above mentioned performances of the existing algorithms, the proposed method proves its superiority in V/Uv classification of noisy speech signals.

5. Conclusions

An improved and reliable V/Uv classification algorithm is presented in this paper. The voiced speech signal is considered as a time series with periodically correlated process, and, the unvoiced signal does not contain any periodic correlation (PC). A statistical model for detecting the presence of PC is employed herewith. EMD based noise filtering method is proposed to increase the robustness of periodicity detection. The overall classification performance is noticeably improved even for low SNR without any threshold value and training data. The use of the proposed method in robust pitch detection is the future target.

6. References

[1] Atal, B. S. and Rabiner, L. R., "A Pattern Recognition Approach to Voiced – Unvoiced – Silence Classification with Applications to Speech Recognition", IEEE Transaction on Acoustics, Speech and Signal Processing, Vol: 24, No. 3, pp: 201-212, 1976.

[2] Ahmadi, S., and Spanias, A. S., "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Trans. Speech Audio Processing*, vol. 7 No. 3, pp. 333-338, 1999.

[3] Shah, J. K. et. al., "Robust voiced/unvoiced classification using novel features and Gaussian mixture model", in Proc. of ICASSP04, 2004.

[4] Alkulaibi, A., Soraghan, J. J., and Durrani, T. S., "Fast HOS based simultaneous voiced/unvoiced detection and pitch estimation using 3-level binary speech signals", in the proceedings of 8th IEEE Signal Processing Workshop on Statistical Signal and Array Processing, pp. 194-197, 1996.

[5] Lobo, and Loizou, P., "Voiced/unvoiced speech discrimination in noise using Gabor atomic decomposition", in the Proceedings of ICASSP, pp. 820-823, 2003.

[6] Janer, L., Bonet, J. J., and Solano, E. L., "Pitch detection and voiced/unvoiced detection algorithm based on Wavelet transform", in the proceedings of ICSLP, 1996

[7] Giridharan, K., Smolenski, B. Y., and Yantorno, R. E., "Statistical And Model Based Approach To Unvoiced Speech Detection", in the proceedings of ISPACS, 2004.

[8] 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.

[9] Hurd, H. L. and Gerr, N. L., "Graphical methods for determining the presence of periodic correlation", Journal of Time Series Analysis, Vol. 15, No. 127, pp: 337-350, 1991.

[10] Wu, B. Z. and Huang, N. E., "A study of the characteristics of white noise using the empirical mode decomposition method", in the Proc. Roy. Soc. Lond. A (460), pp: 1597-1611, 2004.

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

[12] Rilling, G., Flandrin, P. and Gonqalves, P., "On empirical mode decomposition and its algorithms", in the proceedings of IEEE-EURASIP Workshop on nonlinear signal and image processing (NSIP), 2003.

[13] Gardner, W. A., "Cyclostationarity in Communication and Signal Processing", IEEE Press, New York, 1994.

[14] Broszkiewicz-Suwaj, E, "Methods of determining the periodic correlation based on the bootstrap methodology", Hugo Steinhaus Center Research Report HSC/03/02, 2003.

[15] Yu, R. C. et. al., "Quality control of semi-continuous mobility size-fractionated particle number concentration data", Atmospheric Environment, Vol. 38, pp: 3341-3348, 2004.