Adaptive Thresholding Approach for Robust Voiced/Unvoiced Classification

Md. Khademul Islam Molla¹,², Keikichi Hirose¹
¹Dept. of Information and Communication Engineering
The University of Tokyo, Tokyo, Japan
Email: {molla, hirose}@gavo.t.u-tokyo.ac.jp
Sujan Kumar Roy², Shamim Ahmad²
²Dept. of Computer Science and Engineering
The University of Rajshahi, Rajshahi, Bangladesh
Email: {sujan, shamim_cst}@ru.ac.bd

Abstract—This paper presents a robust voiced/unvoiced classification method by using linear model of empirical mode decomposition (EMD) controlled by Hurst exponent. EMD decomposes any signals into a finite number of band limited signals called intrinsic mode functions (IMFs). It is assumed that voiced speech signal is composed of trend due to vocal cord vibration and some noise. No trend is present in unvoiced speech signal. A linear model is developed using IMFs of the noise part of the speech signal. Then a specified confidence interval of the linear model is set as the data adaptive energy threshold. If there exists at least one IMF exceeding the threshold and its fundamental period is within the pitch range, the speech is classified as voiced and unvoiced otherwise. The experimental results show that the proposed method performs superior compared to the recently developed voiced/unvoiced classification algorithms with noticeable performance.

I. INTRODUCTION

The efficient 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. The essence of classification is to determine whether the speech production system involves the vibration of the vocal cords. The speech signal originated from the speaker’s vocal cords contains a sequence of periodic correlation. Such signal is also called voiced speech signal and unvoiced with absence of periodically correlated sequences. The voiced-unvoiced (V/Uv) discrimination problem is an important one and has been worked on extensively during the last three decades [1].

Many algorithms have been reported for solving the classification problem [2]-[6]. In [2], Gaussian mixture model with cepstrum based features is proposed for robust V/Uv classification. A higher order statistics (HOS) based method is implemented in [3] for V/Uv detection and pitch estimation simultaneously. The matching pursuit algorithm is used in [4] with Gabor decomposition. The wavelet transform is proposed in pitch and V/Uv detection in [5]. A statistical model applied in autocorrelation domain is also reported in [6]. In most of the existing algorithms, the Fourier transform or wavelet transform are used in signal decomposition. Although speech signal is non-stationary in nature, those transformations assume that it is piecewise stationary. The speech decomposition is performed by fitting some predefined bases without satisfying its non-stationary nature and hence the degradation of the classification performance. The empirical mode decomposition (EMD) based data adaptive technique is used in [7] to implement noise robust V/Uv classification. The mentioned algorithms require intensive training data and threshold for classification. Such requirements are troublesome for the practical applications.

In this paper, the voiced/unvoiced (V/Uv) discrimination is performed by trend detection is the speech signal. If there exists any trend in the signal it is considered there is a periodic correlation and hence classified as voiced speech otherwise unvoiced. The noisy speech is decomposed using EMD into a finite number of subband signals. The single or multiple subbands which contain the energy greater than an adaptive threshold, the subband(s) are treated as responsible for the trend of the signal. The method being data adaptive, it successively decomposes the signal into noise and trends and hence it is more noise robust with noticeable classification performance.

II. VOICED/UNVOICED CLASSIFICATION

Due to the fact that the range of values of any single speech parameter overlaps between different regions, the accuracy of the single-feature V/Uv classification method would be limited. The proposed algorithm does not need any conventional features and threshold value for discrimination. It compares the energies of the decomposed subband signals with an adaptively determined level. Such energy level also calculated from the analyzing speech signal, no training is required. The algorithm is described in the following subsections.

A. Basic of EMD

Speech signal is non-stationary and non-linear signal. The EMD decomposes the speech signal into multiple subbands in which individual subband contain its original non-stationarity. In the EMD process, signals to be analyzed are adaptively
decomposed into a number of IMFs [8]. Every IMF must satisfy two properties [9]: (i) the number of extrema and the number of zero crossings are either equal or differ by one; (ii) the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is constant zero. A special ‘sifting’ process is employed to extract all of IMFs. After completing the sifting process the signal \( s(n) \) is represented as \( s(n) = \sum_{i=1}^{K} c_i(n) + r_s(n) \). The EMD of the noisy speech signal (0dB SNR) is shown in Figure 1. The whole decomposition being only based on elementary subtractions, it obviously allows for a perfect reconstruction of the initial signal \( s(t) \), given the collection of details \{\( c_k(n) \), \( k = 1 \ldots K \)\} and the residual \( r_s(n) \).

### B. Trend Detection Using EMD

The analyzing signal \( s(n) \) consisting of a slowly varying trend superimposed to a fluctuating process \( \tilde{s}(n) \), the trend is captured by IMFs of large indices including the final residue [10]. Detrending \( \tilde{s}(n) \), which corresponds to estimate \( s(n) \), may therefore amount to capturing the partial reconstruction

\[
\tilde{x}_L(n) = \sum_{i=1}^{L} c_i(n),
\]

where \( L \) is the largest IMF index prior contamination by the trend. For Each of the IMF \{\( c_l(n) \), \( l=1, 2, \ldots, L \)\} being close to zero-mean. A rule of thumb for choosing \( L \) is to observe the evolution of the empirical mean of \( \tilde{x}_L(n) \) as a function of a test order \( l \), and to identify for which \( l=L \) it departs significantly from zero. An example of this approach is illustrated in Figure 2, where a noisy voiced speech signal (0dB SNR) is taken into consideration.

The trend detection procedure outlined above is a rough approach that can be improved upon when a more precise model can be advocated for the signal + noise mixture. To this end, a detailed knowledge of IMFs statistics in noise only situations can help identifying the significance of a given mode [11]. This idea, which has been pioneered by Wu and Huang [12], can be followed in two directions, namely detrending (by keeping only those modes which are identified as noise) and denoising (by removing them). The studies in [12] have considered in some detail white Gaussian noise and more generally fractional Gaussian noise (fGn) as a versatile class for broadband noise with no dominant frequency band. The statistical properties of fGn are entirely determined by its second order structure, which depends solely upon single parameter, \( H \) (with \( 0<H<1 \)), its Hurst exponent. The expected IMFs log-variance has been shown to admit a simple linear model controlled by \( H \) of the considered process [11]:

\[
\log_{10} \sigma_{H}[k] = \log_{10} \sigma_{H}[2] + 2(H-1)(k-2)\log_{10} \rho_H,
\]

for \( k>2 \), with \( \rho_H=2 \). As for the variability of this quantity is concerned, a quantitative yet empirical appreciation can be gained from the experiment with different values of \( H \) (\( H=0.2 \), 0.5 and 0.8), the experimental mean, median and various confidence intervals have been observed, together with the model (2). The series of simulations (which has been carried out on 10000 realizations of 2048 data points in each case) are conducted to derive the parameters for the linear model.

The relative confidence intervals can be given by semi-analytical form as a function of the IMF index: the quasi – linear dependences for parameterization of \( \epsilon_H[k] \) corresponding to a chosen confidence interval according to a functional relationship of the form

\[
\log_{10}(\log_{10}(\epsilon_H[\tilde{c}_H]/\epsilon_H[k])) = \alpha_{H}k + \beta_{H}
\]

where \( \epsilon_H[k] \) stands for the \( H \)-dependent variation of some IMF mean energy, considered as a variance estimator. The best linear fit is obtained when choosing for \( \epsilon_H[k] \) the median of the IMFs energy over the realizations, which is in this case very close from the model \( \epsilon_H[k] \).

<table>
<thead>
<tr>
<th>( H )</th>
<th>( \delta_{H} )</th>
<th>( \epsilon_{H}(95%) )</th>
<th>( \beta_{H}(95%) )</th>
<th>( \epsilon_{H}(99%) )</th>
<th>( \beta_{H}(99%) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.487</td>
<td>0.458</td>
<td>-2.435</td>
<td>0.452</td>
<td>-1.951</td>
</tr>
<tr>
<td>0.5</td>
<td>0.719</td>
<td>0.474</td>
<td>-2.449</td>
<td>0.460</td>
<td>-1.919</td>
</tr>
<tr>
<td>0.8</td>
<td>1.025</td>
<td>0.497</td>
<td>-2.331</td>
<td>0.495</td>
<td>-1.833</td>
</tr>
</tbody>
</table>

### Figure 1

Noisy speech and its EMD (individual IMF). The top most represents voiced speech signal of 10dB SNR, and others represent individual IMF of the noisy speech signal.

### Figure 2

Trend detection of noisy speech signal. Left: standardized empirical mean of the fine-to-coarse EMD reconstruction, evidencing \( L=2 \) as the change point. Top right: original signal. Middle right: estimated trend obtained from the partial reconstruction with IMFs 3 to \( 7^* \) the residual. Bottom right: noise (detrended) signal obtained from the partial reconstruction with IMFs 1 to 2.

The relative confidence intervals can be given by semi-analytical form as a function of the IMF index: the quasi – linear dependences for parameterization of \( \epsilon_H[k] \) corresponding to a chosen confidence interval according to a functional relationship of the form

\[
\log_{10}(\log_{10}(\epsilon_H[\tilde{c}_H]/\epsilon_H[k])) = \alpha_{H}k + \beta_{H}
\]

where \( \epsilon_H[k] \) stands for the \( H \)-dependent variation of some IMF mean energy, considered as a variance estimator. The best linear fit is obtained when choosing for \( \epsilon_H[k] \) the median of the IMFs energy over the realizations, which is in this case very close from the model \( \epsilon_H[k] \).
The parameters $c_0$ and $b_0$ that are used as ingredients for modeling the confidence interval (CI) are deduced from simulation results, and their values are reported in the following Table 1 [10]. The considerations above are used here to separate the trend if any in the signal contaminated with $H$ dependent $fGn$ based on the empirical energy $E_{i0}[k]$ of the IMFs $c_i(n)$. In practice, $E_{i0}[1]$ can be estimated as

$$\hat{E}_0[1] = \sum_{i=1}^{N} c_i^2[n]$$

and subsequent values of $E_{i0}[k]$ are given by

$$\hat{E}_m[k] = L_m \rho_H^{-2(i-1)/H}k, k \geq 2,$$

where $L_m$ is given by $\hat{E}_0[1]/\delta_m$. The possible strategy for trend detection is as follows:

- From the first IMF which captures most of the noise, estimate the noise level in the signal by computing $\hat{E}_0[1]$.

- Estimate the “noise only” model from (4) and (5).

- Estimate the corresponding model for a chosen confidence interval using (3) and Table 1.

- Compare the IMF energies with the confidence interval as a threshold.

- The IMF whose energy exceeds the threshold is responsible for the trend in the signal.

The IMFs with energy exceeding the threshold are separated for further processing to accomplish the proposed V/Uv classification. The implementation of trend detection algorithm for the noisy voiced speech signal (0dB SNR) is shown Figure 3.

The parameter, and its fundamental period (in term of sample) lies within the range corresponding to pitch range (50-500Hz), the speech frame is classified as voiced. The proposed algorithm for voiced unvoiced classification is as follows:

- Construct a set $\Gamma$ of IMFs $\hat{c}_i(n)$ collectively representing the signal trend using the trend detection algorithm. Let $\Gamma = \{\hat{c}_i(n) | z=1, 2, ..., Z\}$, for $0 \leq Z < K$.

- If $\Gamma = \phi$ i.e. no IMF’s energy exceeds the threshold, the speech frame is classified as unvoiced.

- For $\Gamma \neq \phi$, if there exists at least one IMF $\hat{c}_i(n) \in \Gamma$ whose fundamental period are within the pitch range, the signal is classified as voiced, otherwise unvoiced. The fundamental period of $\hat{c}_i(n)$ is found as follows [13]:

- Calculate the normalized autocorrelation function (NACF) of $\hat{c}_i(n)$

- The maximum peak is appeared at zero-lag

- The fundamental period is the absolute difference between the index of zero-lag and the index of highest peak in NACF.

In Figure 3, the fundamental periods of IMF 3 to 5 are within the pitch range and hence the speech signal is classified as voiced. The noisy unvoiced speech signal (0dB SNR) and the illustration of trend detection algorithm on it are shown in Figure 4. None of the IMF’s energy exceeds the energy level of 99% confidence interval. No trend is detected in that speech signal and hence labeled as unvoiced speech.

![Figure 3](image1)

![Figure 4](image2)

Figure 3. (a) Noisy speech signal (0dB SNR), (b) The energies of the IMFs 3-5 exceed the threshold of 99% confidence interval and hence they collectively represent the trend of the signal.

**C. V/Uv Discrimination Algorithm**

In this work, the speech is considered as a sum of signal contaminated with the noise. The response of the excitation of vocal cord is treated as the trend signal which is modulated by the noise producing the speech signal. In the proposed method 99% confidence interval is used with $H = 0.8$ is selected experimentally yielding better results. If there exists at least one IMF whose energy exceeds the confidence interval parameter, and its fundamental period (in term of sample) lies within the range corresponding to pitch range (50-500Hz), the speech frame is classified as voiced. The proposed algorithm for voiced unvoiced classification is as follows:

- Construct a set $\Gamma$ of IMFs $\hat{c}_i(n)$ collectively representing the signal trend using the trend detection algorithm. Let $\Gamma = \{\hat{c}_i(n) | z=1, 2, ..., Z\}$, for $0 \leq Z < K$.

- If $\Gamma = \phi$ i.e. no IMF’s energy exceeds the threshold, the speech frame is classified as unvoiced.

- For $\Gamma \neq \phi$, if there exists at least one IMF $\hat{c}_i(n) \in \Gamma$ whose fundamental period are within the pitch range, the signal is classified as voiced, otherwise unvoiced. The fundamental period of $\hat{c}_i(n)$ is found as follows [13]:

- Calculate the normalized autocorrelation function (NACF) of $\hat{c}_i(n)$

- The maximum peak is appeared at zero-lag

- The fundamental period is the absolute difference between the index of zero-lag and the index of highest peak in NACF.

In Figure 3, the fundamental periods of IMF 3 to 5 are within the pitch range and hence the speech signal is classified as voiced. The noisy unvoiced speech signal (0dB SNR) and the illustration of trend detection algorithm on it are shown in Figure 4. None of the IMF’s energy exceeds the energy level of 99% confidence interval. No trend is detected in that speech signal and hence labeled as unvoiced speech.

**III. EXPERIMENTAL RESULTS AND DISCUSSION**

The performance of the proposed method is evaluated 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 20ms shifting. Each frame of the data base is accurately labeled for voiced and unvoiced.

The performance of the proposed method named EMD based linear model [11] (LmEMD) are compared for two existing methods, EMD based noise filtering (NEEMD) and without noise filtering (WNF) for different noise levels [7]. In NEEMD, EMD based time domain noise filtering is performed...
and then V/Uv classification algorithm is applied, whereas in WnF, no noise filtering is performed before V/Uv detection. The white Gaussian noise is added to obtain different levels of segmental SNR (SSNR). Voiced – to – unvoiced (V-V) 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 speech data has taken from TIMIT database and labeled 1 for voiced and 0 for unvoiced frame. Then apply the proposed method to that speech data and similarly labeled the detected voiced frame as 1 and unvoiced frame as 0. Then compare the performance of V-Uv and Uv-V errors. The Figure 5 shows the comparison of original labels versus detected V/Uv results for a 1.6s length speech signal of 10dB SNR. Only one Uv-V error is occurred for 60 frames at the transition of voiced to unvoiced.

Figure 5. (a) Speech signal of 10 dB SNR, (b) Detected frames labels versus original frame labels in terms of ‘1’ as voiced and ‘0’ for unvoiced.

TABLE II. PERFORMANCE COMPARISON OF THE PROPOSED METHOD LmEMD WITH NfEMD AND WnF METHODS.

<table>
<thead>
<tr>
<th>SNR(dB)</th>
<th>Methods</th>
<th>V-Uv(%)</th>
<th>Uv-V(%)</th>
<th>Overall(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LmEMD</td>
<td>0.03</td>
<td>0.40</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>NfEMD</td>
<td>0.65</td>
<td>0.33</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>WnF</td>
<td>0.69</td>
<td>0.38</td>
<td>1.07</td>
</tr>
<tr>
<td>10</td>
<td>LmEMD</td>
<td>0.55</td>
<td>0.40</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>NfEMD</td>
<td>0.74</td>
<td>0.68</td>
<td>1.42</td>
</tr>
<tr>
<td></td>
<td>WnF</td>
<td>1.07</td>
<td>0.85</td>
<td>1.92</td>
</tr>
<tr>
<td>0</td>
<td>LmEMD</td>
<td>0.64</td>
<td>0.40</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>NfEMD</td>
<td>2.92</td>
<td>1.31</td>
<td>4.23</td>
</tr>
<tr>
<td></td>
<td>WnF</td>
<td>4.34</td>
<td>3.29</td>
<td>7.63</td>
</tr>
<tr>
<td>-5</td>
<td>LmEMD</td>
<td>0.83</td>
<td>0.28</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>NfEMD</td>
<td>4.31</td>
<td>2.63</td>
<td>6.94</td>
</tr>
<tr>
<td></td>
<td>WnF</td>
<td>6.47</td>
<td>4.55</td>
<td>11.02</td>
</tr>
<tr>
<td>-10</td>
<td>LmEMD</td>
<td>1.00</td>
<td>0.13</td>
<td>1.13</td>
</tr>
<tr>
<td></td>
<td>NfEMD</td>
<td>5.43</td>
<td>3.98</td>
<td>9.41</td>
</tr>
<tr>
<td></td>
<td>WnF</td>
<td>8.02</td>
<td>7.11</td>
<td>15.13</td>
</tr>
</tbody>
</table>

The performance of the proposed method (LmEMD) for a wide range of SNRs is compared with existing methods as shown in Table 2. It is clearly observed that the proposed algorithm performs excellent even compared to the related existing algorithms as mentioned in [7]. The LmEMD is excessively noise robust because its principle is to separate the trend and noise of the analyzing signal. Another novelty of the algorithm is that no training data or priori threshold is required. It is derived adaptively from each analyzing speech frame and hence it is more suitable for practical application and more robust than existing algorithms.

IV. CONCLUSIONS

A robust voiced unvoiced classification algorithm has been presented in this paper. A data adaptive thresholding approach is proposed here using the linear model of EMD. The threshold is based on the analyzing speech frame and derived by the specified confidence interval of energy distribution of the IMFs of speech frame. No training data or priori information about the analyzing signal is required. The principle of the proposed method is to separate the noise as well as the trend from the analyzing signal and hence it performs even with very low SNR (e.g. 10dB). Furthermore, because it is local in time, this structure can adopt automatically to non-stationary situations with greater flexibility than other. The result of the proposed algorithm proves its superiority and robustness against noise. The future extension of this work is instantaneous V/Uv classification rather the frame based implementation.

REFERENCES