Robust Voice Activity Detection Based on Adaptive Sub-band Energy Sequence Analysis and Harmonic Detection

Yanmeng Guo, Qiang Fu, Yonghong Yan

ThinkIT Speech Lab, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080, P.R.China
{yguo,qfu,yyan}@hccl.ioa.ac.cn

Abstract
Voice activity detection (VAD) in real-world noise is a very challenging task. In this paper, a two-step methodology is proposed to solve the problem. First, segments with non-stationary components, including speech and dynamic noise, are located using sub-band energy sequence analysis (SESA). Secondly, voice is detected within the selected segments employing the proposed method concerning its harmonic structure. Therefore, speech segments can be accurately detected by this rule-based framework. This algorithm is evaluated in several databases in terms of speech/non-speech discrimination and in terms of word accuracy rate when it is used as the front-end of automatic speech recognition (ASR) system. It provides a more reliable performance over the commonly used standard methods.

Index Terms: voice activity detection, harmonic structure, noise robustness, automatic speech recognition

1. Introduction
Robust voice activity detection plays an important role in the real-world application of speech processing, especially ASR systems, whose recognition rate and computation complexity are highly relied on the accuracy of voice activity detection.

Numerous approaches [1–9] have been proposed to locate speech in noisy environments. Some of them are effective in certain kinds of noise even with low signal-to-noise-ratio (SNR). However, in real conditions, noise often comes from multiple time-varying sources, and some of them may change abruptly. In these cases, the methods based on statistical models, noise estimation or noise reduction, which may work in stationary or slow-varying non-stationary noise, become insufficient.

Let us assume the input digital signal $y(i)$ as the summation of speech signal $s(i)$ and uncorrelated additive noise $n(i)$, where $i$ is the time index:

$$y(i) = s(i) + n(i)$$

In methodology, some approaches take $y(i)$ as $n(i)$ added with $s(i)$, and isolate $s(i)$ by tracking $n(i)$ [1,2,4]. These tracking algorithms, which tend to estimate $n(i)$ by the minimum statistic [1] or low-variance component of $y(i)$ [4], are often bothered by underestimate time delay and computation complexity. Some methods consider $y(i)$ as $s(i)$ corrupted by $n(i)$, and try to find the speech properties from $y(i)$ [5] [8]. The performance of such systems may be badly affected when the speech properties become unextractable due to the corruption of noise. There are also some approaches based on classifier [3] or statistics [4] by utilizing the statistical information of both $s(i)$ and $n(i)$, but these algorithms are highly relied on the consistency between the training and test data.

The most crucial obstacle for practical voice activity detection is the non-stationarity of noise. In this paper, we build the solution by decomposing it into two sub-problems. First, the signal is interpreted as stationary noise added by non-stationary components, which can be located by analyzing energy sequences in sub-bands, as described in section 3. Second, the non-stationary components are considered as the noise-corrupted-speech, from which speech can be detected by its harmonic properties, as detailed in section 4. The system overview is introduced in section 2, and the hangover scheme is described in section 5. In section 6, the performance evaluation is given. Finally, the conclusions are drawn in section 7.

2. System overview
With the assumption that $y(i)$ and $n(i)$ are uncorrelated, we can conclude from (1):

$$E_y(r) = E_s(r) + E_n(r)$$

where $E_y(r)$, $E_s(r)$ and $E_n(r)$ denote the short-time energy of $y(i)$, $s(i)$ and $n(i)$ in frame $r$ respectively, $\{r = 1, 2, \cdots \}$. And speech can be detected by locating the segments where $E_y(r) > E_n(r)$.

Here we classify the real-world noise into 5 categories: stable noise, slow-varying noise, impulse noise, fluctuant noise and step noise, in which step noise denotes the noise whose energy may step up or down abruptly. Among them, stable noise, slow-varying noise and placid segments of step noise can be viewed as stationary in a short duration of time and the sum of their energy is denoted as $\{E_p(r)\}$, while noises of other kinds are considered non-stationary, and we denote the sum of their energy as $\{E_q(r)\}$. Thus $E_y(r)$ in (2) can be further modeled as

$$E_y(r) = E_s(r) + E_p(r) + E_q(r)$$

In the frames where $E_s(r) = E_q(r) = 0$, $y(i)$ appears to be stationary, from which a noise model describing the characters of $E_y(r)$ can be updated iteratively. This model will be used to distinguish the non-stationary segments (either $E_q(r) > 0$ or $E_p(r) > 0$, or both) from the stationary ones. After this, a harmonic structure detection will be implemented in the non-stationary segments to decide whether they contain speech or not. Furthermore, a hangover scheme is used to finally determine the boundaries of speech. The flowchart of the proposed system is illustrated in Figure 1.
3. Sub-band energy sequence analysis

We assume that \( \{E_p(r)\} \), the energy sequence of stationary noise, follows the ergodic Gaussian distribution with a probability density function (pdf) of

\[
f(E_p) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{(E_p - \mu)^2}{2\sigma^2}\right)
\]

where \( \mu \) and \( \sigma \) represents the mean and the variance, respectively. Define \( \lambda = \sigma / \mu \) to normalize the dynamic range, and (4) can be rewritten as

\[
f(E_p) = \frac{1}{\sqrt{2\pi} \lambda \mu} \exp\left(-\frac{(E_p - \mu - 1)^2}{2\lambda^2}\right)
\]  (5)

Take a 100 ~ 300ms window for analysis, as shown in Figure 2. If the current noise model based on (5) could reasonably describe the energy distribution of the analysis window, the current segment is identified as stationary.

As shown in Figure 2, the Gaussian coefficients \( \mu_j \) and \( \lambda_j \) is extracted from the first \( z \) frames of the \( j \)th analysis window, and the noise model is updated as \( \mu = \mu_j, \lambda = \lambda_j \) if one of the following criteria is met.

1. \( \mu_j < \mu \cdot c \) and \( \lambda_j < \lambda \cdot c \), where \( 1 < c < 1.5 \);
2. \( \mu_j \cdot (1 + \lambda_j) < \mu \cdot (1 + \lambda) \);
3. \( \lambda_j < \Lambda \), where \( \Lambda \) is a threshold and 0.01 < \( \Lambda < 0.1 \).

Criteria 1 and 2 correspond to the slow-varying and more stationary model, and criteria 3 is to adjust the model after \( E_p(r) \) steps up and stays on a stationary condition.

The current model (possibly updated) is used to examine the remaining signal in the window by setting a threshold as \( \theta = \mu + \mu \cdot \lambda / \alpha \), where \( \alpha \) is the sensitivity coefficient and 0 < \( \alpha < 1 \). If during several consecutive frames in the analysis window the mean energy is higher than \( \theta \), the whole window is identified as non-stationary.

In some adverse conditions of low SNR, this method will be less effective. However, the spectral energy distributions of speech and noise are often different, so the SNR in different frequency bands are not equal. The performance of the algorithm will be more robust if applied in several specific sub-bands. In the proposed system with a 8kHz sampling rate, the input signal is decomposed into 4 bands: 0 ~ 500Hz, 500 ~ 1000Hz, 1000 ~ 2000Hz and 2000 ~ 4000Hz. Non-stationary component is detected if the signals in two or more sub-bands are not stationary.

4. Robust harmonic structure detection

The harmonic structure is a peculiar character of voice, which appears as obvious evenly-distributed energy peaks in spectrum. The spectral distance between adjacent peaks equals to the fundamental frequency, which is in range of 60Hz to 450Hz. The harmonic structure is clear even with strong background noise, so it is very useful in discriminating speech from noise. However, in real conditions, the most clear harmonic structure in each segment may appear in different frequency region. In this section, an adaptive method of searching clear harmonic structure is described. Two properties of harmonic structure are utilized, which make this method robust against distortions and time-varying noise.

The first property is the uniformity of harmonic structure, which means that the harmonics are a series of energy peaks, and the distance between adjacent ones is equal to the fundamental frequency \( F_0 \). Under most circumstances, even when some peaks are submerged in the corruption of noise, at least 3 ~ 4 consecutive harmonics still keep clear. To check this property in frame \( r \), the spectral energy of the \( k \)th frequency bin is calculated using short-time Fourier transform (STFT), and represented as \( P_k(k) \). All the local peaks are picked from \( P_k(k) \), and peaks too low or too narrow are removed to eliminate the irrelevant peaks that may mislead the detection. The remaining set of peaks are denoted as \( \{Q_k(k)\} \), and they are possibly the clear harmonics. We take various \( F_0 \) to see if \( \{Q_k(k)\} \) contains its multiples, in which \( F_0 \) is increased in a step length of \( \delta F = 1.5Hz \) within the range of 60 to 450Hz. If \( \{Q_k(k)\} \) contains peaks in the position of (or quite near to) \( nF_0 \sim (n + 3)F_0 \), those peaks are detected as potential harmonics for \( F_0 \) with the pattern of \( n \). The \( n = 0 \) pattern means the 1, 2, 3 multiples of \( F_0 \) are all matched.

Secondly, the continuity of \( F_0 \) and patterns are checked to eliminate the spurious harmonics. If \( F_0 \) fluctuates within a limited extent and its harmonics are all matched with pattern \( n \) for a specific number of consecutive frames, these frames are identified as voice. Figure 3 shows an example. All of frame \( r \) to
frames that are correctly detected as pause or speech: which are defined as the fraction of all actual pause or speech non-speech hit-rate (HR0) and the speech hit-rate (HR1) [2], detection algorithms are evaluated in terms of the probability of noises of AURORA at SNRs of 20, 10 and 0 dB.

V AD, then the silence before and after the utterance are both cut in TIMIT test set, the speech pauses are cut by an energy-based AURORA noise database [11]. For each of the 1680 utterances from standard V AD algorithms in G.729 [6], ETSI AFE [7], ETSI AMR [9] are used for comparison.

6. Experiments and analysis

Several experiments are carried out for the evaluation of the proposed VAD algorithm. In the first experiment, the speech/non-speech discrimination is evaluated. In the following two, the influence of VAD on a speech recognition system is assessed. The proposed VAD algorithm also outperforms the other VAD algorithms, as shown in Table 2.

Point of the speech segment.

After finding the start point, the noise model will be updated more conservatively in order to protect speech components. That is, update is implemented only when criterion 2 in Section 3 is met. When in a certain analysis window, all of the sub-bands appear to be stationary, or no harmonic structure is detected, the first frame of this window is identified as the start frame of speech.

6.1. Performance on speech/noise discrimination

An audio corpus is synthesized from TIMIT test set [10] and the AURORA noise database [11]. For each of the 1680 utterances in TIMIT test set, the speech pauses are cut by an energy-based VAD, then the silence before and after the utterance are both prolonged to 2 seconds. At last, the utterances are added with 8 noises of AURORA at SNRs of 20, 10 and 0 dB.

As shown in table 1, the performance of voice activity detection algorithms are evaluated in terms of the probability of non-speech hit-rate (HR0) and the speech hit-rate (HR1) [2], which are defined as the fraction of all actual pause or speech frames that are correctly detected as pause or speech:

$$HR0 = \frac{N_{0,0}}{N_{0,ref}}$$
$$HR1 = \frac{N_{1,1}}{N_{1,ref}}$$

where \(N_{0,0}\) and \(N_{1,1}\) are the number of real non-speech and speech frames in the whole database, respectively, while \(N_{0,ref}\) and \(N_{1,ref}\) are the number of non-speech and speech frames correctly classified.

It can be seen that the proposed VAD provides the best balance between HR1 and HR0, especially in noise of street and train station. G.729 VAD suffers poor HR1 with increasing noise level, though the HR0 keeps stable. The AFE VAD has high HR1 in all conditions, but it happens at the cost of low HR0. Both AMR1 and AMR2 yield high HR1 in all conditions, but the HR0 of them keeps low in some noise conditions. The proposed VAD yields high HR1 and HR0 in all the noise conditions, while HR1 and HR0 decrease slightly at poor SNRs.

### Table 2: Word accuracy rate for database 1

<table>
<thead>
<tr>
<th>Channel</th>
<th>G.729</th>
<th>AFE</th>
<th>AMR1</th>
<th>AMR2</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDMA</td>
<td>81.8</td>
<td>72.0</td>
<td>74.0</td>
<td>76.0</td>
<td><strong>85.6</strong></td>
</tr>
<tr>
<td>GSM</td>
<td>89.2</td>
<td>88.4</td>
<td>89.4</td>
<td>91.0</td>
<td><strong>91.0</strong></td>
</tr>
<tr>
<td>PAS</td>
<td>88.2</td>
<td>87.6</td>
<td>86.8</td>
<td>88.4</td>
<td><strong>89.8</strong></td>
</tr>
<tr>
<td>Phone</td>
<td>88.2</td>
<td>83.2</td>
<td>83.0</td>
<td>87.0</td>
<td><strong>88.2</strong></td>
</tr>
</tbody>
</table>

### Table 3: Word accuracy rate for database 2

<table>
<thead>
<tr>
<th>Environment</th>
<th>G.729</th>
<th>AFE</th>
<th>AMR1</th>
<th>AMR2</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Campus</td>
<td>77.0</td>
<td>75.0</td>
<td>77.0</td>
<td>76.0</td>
<td><strong>77.0</strong></td>
</tr>
<tr>
<td>Gymnasium</td>
<td>46.0</td>
<td>67.0</td>
<td>58.0</td>
<td>73.0</td>
<td><strong>75.0</strong></td>
</tr>
<tr>
<td>Office</td>
<td>82.0</td>
<td>84.0</td>
<td>86.0</td>
<td>85.0</td>
<td><strong>87.0</strong></td>
</tr>
<tr>
<td>Quiet room</td>
<td>85.0</td>
<td>85.0</td>
<td>87.0</td>
<td>86.0</td>
<td><strong>87.0</strong></td>
</tr>
<tr>
<td>Restaurant</td>
<td>69.0</td>
<td>72.0</td>
<td>71.0</td>
<td>75.0</td>
<td><strong>75.0</strong></td>
</tr>
</tbody>
</table>

6.2. Speech recognition performance

The proposed VAD algorithm is evaluated on its influence as a front-end of an ASR system, and the recognizer is based on hidden Markov model toolkit software package (HTK) [12]. The monophone set for the recognizer is trained as SCHMM with the following parameters: 3 states per phone, simple left-to-right models, 16-Gaussian mixture per state and using diagonal covariance. The 39-dim feature vector consists of 12 cepstral coefficients, the logarithmic frame energy plus the corresponding delta and acceleration coefficients. The model is trained as a cross-word model. The training corpus includes 1 standard Mandarin set plus 4 main Chinese dialects sets.

The experiments are performed on 2 databases. Database 1 is 2000 streams of Chinese phrases spoken by 5 male speakers and 5 female speakers. For each speaker, 50 utterances are recorded through wired phone, CDMA, GSM and Personal Access System (PAS) separately. Database 2 includes 600 audio streams recorded by 5 male and 5 female speakers through GSM in 6 different environments: campus, gymnasium, office, quiet room, street and restaurant, and 100 streams per environment. Table 2 and Table 3 give the results of the word accuracy rate for database 1 and 2.

As can be seen from Table 2, the proposed VAD algorithm is not sensitive to the communication channel. The speech communication channels distort the signal in different ways, but both the sub-band energy sequence analysis and harmonic structure detection are robust to the spectral distortion, and that keeps the performance reliable for all the communication systems.

In the environment with dynamic noise, the proposed algorithm also outperforms the other VAD algorithms, as shown in Table 3. And in the quiet environment, its performance is comparable to that of the reference algorithms. The sub-band energy sequence analysis is effective to detect the non-stationary segments, while the harmonic structure is reliable to avoid the non-stationary noise, so the algorithm performance keeps robust in most environments, and its advantage is outstanding in the dynamic noise.
### Table 1: Performance of the VAD algorithms in HR1 and HR0(%)

<table>
<thead>
<tr>
<th>Noise</th>
<th>SNR(dB)</th>
<th>HR1 (%)</th>
<th>HR0 (%)</th>
<th>Proposed</th>
<th>G.729</th>
<th>AMR1</th>
<th>AMR2</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20</td>
<td>95.4</td>
<td>100</td>
<td>100</td>
<td>99.4</td>
<td>53.0</td>
<td>33.7</td>
<td>52.1</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>94.3</td>
<td>100</td>
<td>100</td>
<td>99.4</td>
<td>55.1</td>
<td>26.0</td>
<td>41.1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>93.4</td>
<td>100</td>
<td>100</td>
<td>99.4</td>
<td>53.0</td>
<td>33.7</td>
<td>52.1</td>
</tr>
<tr>
<td>Airpot</td>
<td>20</td>
<td>84.3</td>
<td>100</td>
<td>99.7</td>
<td>99.9</td>
<td>99.3</td>
<td>53.0</td>
<td>34.4</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>81.0</td>
<td>100</td>
<td>99.8</td>
<td>99.8</td>
<td>99.3</td>
<td>55.0</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>77.1</td>
<td>100</td>
<td>98.5</td>
<td>99.7</td>
<td>99.3</td>
<td>56.4</td>
<td>29.4</td>
</tr>
<tr>
<td>Babble</td>
<td>20</td>
<td>93.4</td>
<td>100</td>
<td>99.9</td>
<td>100</td>
<td>99.4</td>
<td>55.1</td>
<td>26.0</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>88.4</td>
<td>100</td>
<td>99.8</td>
<td>99.8</td>
<td>99.3</td>
<td>55.0</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>85.9</td>
<td>100</td>
<td>98.5</td>
<td>99.7</td>
<td>99.3</td>
<td>56.4</td>
<td>29.4</td>
</tr>
<tr>
<td>Car</td>
<td>20</td>
<td>93.6</td>
<td>100</td>
<td>99.9</td>
<td>100</td>
<td>99.4</td>
<td>55.1</td>
<td>26.0</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>87.8</td>
<td>100</td>
<td>99.8</td>
<td>99.8</td>
<td>99.3</td>
<td>55.0</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>85.9</td>
<td>100</td>
<td>98.5</td>
<td>99.7</td>
<td>99.3</td>
<td>56.4</td>
<td>29.4</td>
</tr>
<tr>
<td>Exhibition hall</td>
<td>20</td>
<td>93.6</td>
<td>100</td>
<td>99.9</td>
<td>100</td>
<td>99.4</td>
<td>55.1</td>
<td>26.0</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>88.4</td>
<td>100</td>
<td>99.8</td>
<td>99.8</td>
<td>99.3</td>
<td>55.0</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>85.9</td>
<td>100</td>
<td>98.5</td>
<td>99.7</td>
<td>99.3</td>
<td>56.4</td>
<td>29.4</td>
</tr>
<tr>
<td>Restaurant</td>
<td>20</td>
<td>94.3</td>
<td>100</td>
<td>99.9</td>
<td>100</td>
<td>99.4</td>
<td>55.1</td>
<td>26.0</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>89.6</td>
<td>100</td>
<td>99.8</td>
<td>99.8</td>
<td>99.3</td>
<td>55.0</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>87.8</td>
<td>100</td>
<td>98.5</td>
<td>99.7</td>
<td>99.3</td>
<td>56.4</td>
<td>29.4</td>
</tr>
<tr>
<td>Street</td>
<td>20</td>
<td>93.6</td>
<td>100</td>
<td>99.9</td>
<td>100</td>
<td>99.4</td>
<td>55.1</td>
<td>26.0</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>88.4</td>
<td>100</td>
<td>99.8</td>
<td>99.8</td>
<td>99.3</td>
<td>55.0</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>85.9</td>
<td>100</td>
<td>98.5</td>
<td>99.7</td>
<td>99.3</td>
<td>56.4</td>
<td>29.4</td>
</tr>
<tr>
<td>Subway</td>
<td>20</td>
<td>94.3</td>
<td>100</td>
<td>99.9</td>
<td>100</td>
<td>99.4</td>
<td>55.1</td>
<td>26.0</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>89.6</td>
<td>100</td>
<td>99.8</td>
<td>99.8</td>
<td>99.3</td>
<td>55.0</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>87.8</td>
<td>100</td>
<td>98.5</td>
<td>99.7</td>
<td>99.3</td>
<td>56.4</td>
<td>29.4</td>
</tr>
<tr>
<td>Train station</td>
<td>20</td>
<td>94.3</td>
<td>100</td>
<td>99.9</td>
<td>100</td>
<td>99.4</td>
<td>55.1</td>
<td>26.0</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>89.6</td>
<td>100</td>
<td>99.8</td>
<td>99.8</td>
<td>99.3</td>
<td>55.0</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>87.8</td>
<td>100</td>
<td>98.5</td>
<td>99.7</td>
<td>99.3</td>
<td>56.4</td>
<td>29.4</td>
</tr>
</tbody>
</table>

### 7. Conclusion

In this paper, a robust VAD algorithm is proposed to detect speech in the noisy, highly variable environments. The problem is decomposed into two parts, and solved by sub-band energy sequence analysis and harmonic structure detection, respectively. Both of them are robust to the environment noise and channel distortion. Extensive evaluations show that this algorithm is reliable in most noisy environments, especially with dynamic noise.

### 8. References


