A ROBUST METHOD BASED ON LIKELIHOOD ESTIMATION FOR SPEECH SIGNAL DETECTION

Shaoyan CHEN*, Bo XU*, Taiyi HUANG* and Yintao YANG†
† Center of Space Science & Applied Research
* National Laboratory of Pattern Recognition, Institute of Automation
† Chinese Academy of Sciences, Beijing

ABSTRACT
Speech signal detection is found to have a variety of applications in the speech communication. Many methods have been proposed for that purpose. Most of these methods can achieve very high detection accuracy for a reasonable given false alarm probability in clean speech environment. However, these methods become less reliable in the noisy environment. The accurate detection of speech signal is known to be still very difficult in the presence of noise and interference. In this paper, we propose a method to use the likelihood estimated from a noise model to detect the speech signal. We shall address the problems on how to train a noise model, how to use the likelihood to detect the speech signal and how to use an on-line adaptation procedure to adapt the model parameters to a new noisy environment. We will also present experiment results to demonstrate some of the properties and advantages of the method.

1. INTRODUCTION
Speech signal detection (SSD), or sometime called endpoint detection of speech segments has many potential applications in speech communication. For example, in the short-wave-based wireless system, a SSD module is often used for discontinuous transmission to save battery power [1]. Similarly, in a speech recognition system, a SSD module is often used to reduce the computational burden of the recognition process and to improve the recognition accuracy [4].

In the past several decades, various methods have been proposed and investigated for the SSD. Amongst them, the short-term energy method, sometimes with the augmentation of short-term zero-crossing rate, is perhaps the most widely used one [2]. However, this method has many problems. First, if the noise is very strong, this method becomes unlikely reliable. In a short-wave wireless communication system, we also find a fact that the noise becomes much stronger for non-speech segments. That means although the signal-to-noise ratio (SNR) is not very low for a speech segment, the energy of a noise segment approaches or even exceeds that for a speech segment. In this case, the short-term energy method appears much less reliable. Second, the short-term energy method has no attempt to use some speech-specific characteristics.

The fundamental frequency or Pitch is an important parameter for characterizing the excitation source in the speech production model [3]. It conveys the prosodic information about the speech signal and, hence, is useful for applications such as speech signal detection, speech recognition and speech synthesis. Many reported results show that the use of Pitch for speech detection can achieve very high accuracy in clean speech environment [3]. However, Pitch information can not be reliably estimated in the presence of strong noise. The related detection method, therefore, becomes less reliable in such circumstance [4].

In this paper, we will investigate a likelihood based speech signal detection method using the mel-frequency cepstra as features. We show that the new method is more robust to noise than the traditional methods. An on-line adaptation algorithm is proposed to adapt the pre-trained system’s parameters to a new noisy environment, which is shown to be able to further improve the performance.

The overall structure of the method is as shown in Fig. 1. The whole process is divided into three parts, i.e., training, detection and adaptation. The training process is designed to estimate the noise model parameters from the labeled training data. The role of the detection process is to distinguish between noise and speech. This is a well-known classification problem. For an input segment of unknown signal, we use a method to determine it as speech signal segment or noise segment according to the likelihood estimated from signal features and the pre-trained model parameters. Once the segment is classified into noise, an adaptation process will use this noise signal to adapt the model parameters to a new noisy environment.

The layout of the paper is as follows. The following section briefly introduces the representation of MFCC features and the estimation of the model parameters. The automatic detection
process is described in section 3. In section 4, an adaptation method is presented. Section 5 presents some detection results and the comparison with some other methods such as short-term energy based method and Pitch based method. Important conclusions will be given in the last section.

2. TRAINING PROCESS

Speech signal detection is a well-known classification problem. In this paper, we will use a maximum likelihood (ML) classifier to distinguish speech signal from noise. To estimate the ML score for a given frame of speech signal, we will need a noise model. The determination of noise model parameters is made through two steps, i.e., feature representation and parameters’ estimation.

- Feature representation: This includes blocking the labeled noise data into frames, frame windowing, pre-emphasis, Fourier transformation, Mel-filter bank analysis and discrete cosine transform (DCT). The center frequencies of the filters in the filter bank are spaced equally on a linear scale from 100 to 1000Hz and equally on a logarithmic scale above 1000Hz. Each filter’s magnitude frequency response has a triangular shape that is equal to unity at the center frequencies and linearly decreasing to zero at the center frequencies of the two adjacent filters. The spectrum for each frame is passed through the mel-filter bank which consists of 25 triangle filters. The filter bank energies are then converted into 12 Mel-scale cepstral coefficients (MFCC) using a DCT.

- Model parameter estimation: The noise is modeled by a single Gaussian pdf with a mean vector and a diagonal covariance matrix, namely \( g(\mu, \Sigma) \), where

\[
\mu = [\mu_1, \mu_2, ..., \mu_K]
\]

\[
\Sigma = [\Sigma_{ij}], \quad 1 \leq i \leq K, 1 \leq j \leq K
\]

and \( \Sigma_{ij} \) satisfies

\[
\Sigma_{ij} = \begin{cases} 
\sigma^2_i, & \text{if } i = j \\
0, & \text{Otherwise}
\end{cases}
\]

For the ith frame signal, The feature vector is denoted as

\[
X_i = [x_{i1}, x_{i2}, ..., x_{ik}]
\]

where \( K \) is the dimensionality of the feature space. The \( \mu_k \) and \( \sigma^2_{ik} \) are estimated via

\[
\mu_k = \frac{1}{N} \sum_{i=1}^{N} x_{ik}
\]

\[
\sigma^2_{ik} = \frac{1}{N} \sum_{i=1}^{N} (x_{ik} - \mu_k)^2
\]

where \( N \) is the total frame number for the whole training data.

Here only single mixture is used since it is found that in many different noise environments, single mixture is enough to model the noise distribution [9]. However, this method is not limited to single mixture case. If multiple mixtures should be employed, the k-mean method can be used to cluster the training samples to several groups. For each cluster, the above method can be used to estimate the model parameters.

3. DETECTION PROCESS

The detection process is depicted by Fig.2.

[Diagram of detection procedure]

The role of the detection process is to distinguish between noise segment and speech segment. The method we use for identifying the speech signal is based on the comparison of the log-likelihood score with some thresholds. For an unknown frame of input signal, we firstly represent it into a MFCC feature vector. The conversion of signal to the MFCCs is as what described in the above section. We then calculate the log-likelihood of the feature vector using the model parameters estimated from the training process. Much literature has addressed the problem of the estimation of likelihood. For saving the space, we will not describe how to calculate the likelihood in this paper. Details please refer to [6].

The SSD decision is made through two steps. First, an initial decision is made based on a direct comparison of the instantaneous frame likelihood with an initial threshold \( TH_i \). If the likelihood is greater than the threshold, current frame is classified into noise segment. Second, if the likelihood is less than or equal to the threshold, the initial decision is smoothed, taking into account of past neighboring frames.

The threshold used for the initial decision is an empirical value. In our system, we estimate the average log-likelihood for the whole training data and use the average log-likelihood multiplying a factor of 0.8 as the initial threshold. As will be shown later, the initial threshold \( TH_i \) will be updated in the adaptation process.

The initial decision is local. The decision does not take into account the short-term stationary properties of both the noise and the speech signal. A further smoothing and correction algorithm is employed. The algorithm consists of three steps. In the first step, the current frame is classified into speech segment if the likelihood is less than a final threshold which is 70 percent of the initial threshold. In the second step, the speech signal decision is extended to the current frame if the precious 4 frames were determined as speech. In the final step, the current frame is again classified into noise if the conditions in the first and second steps are not met.
4. ADAPTATION

The background noise may change considerably between different conversation environments or even in a same conversation environment. The variation of the noise characteristics may result a serious mismatch between the training condition and the testing environment, which often causes a dramatic degradation in performance of the system. Hence a process of updating the model parameters according to the varying noise characteristics is required. This process is called adaptation. In this paper, we use a method as shown in follows for the adaptation purpose.

Look at the model parameter estimation formulae, where both (5) and (6) can be rewritten as

\[ \mu_k = \frac{1}{N} \sum_{i=1}^{N} x_{i}^k = \frac{1}{N} \left( \sum_{i=1}^{N} x_i^k \right) \]

\[ = \frac{N-1}{N} \mu_k^{i-1} + \frac{1}{N} x_k^i \]  

(8)

and

\[ \sigma_k^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i^k - \mu_k)^2 \]

\[ = \frac{1}{N} \left( \sum_{i=1}^{N} (x_i^k)^2 + (\mu_k^i)^2 \right) - \left( \frac{N-1}{N} \mu_k^{i-1} + \frac{1}{N} \mu_k^i \right)^2 \]

\[ = \frac{N-1}{N} (\sigma_k^i)^2 + \frac{N-1}{N^2} (\mu_k^i - x_k^i)^2 \]  

(9)

From the above expression of model parameter estimation formula, one can see that the model is easy to be updated if given a new feature vector.

Without the loss of generality, we modify the above two equations as follows and use the modified equations for adaptation purpose.

\[ \mu_k^{\text{adp}} = (1-\alpha)\mu_k^{\text{old}} + \alpha \mu_k^{\text{new}} \]  

(10)

\[ (\sigma_k^i)^{\text{adp}} = (1-\alpha)(\sigma_k^i)^{\text{old}} + \alpha (1-\alpha)(\mu_k^{\text{old}} - x_k^{\text{adp}})^2 \]  

(11)

where \( \mu_k^{\text{old}} \) and \( (\sigma_k^i)^{\text{old}} \) denote the old model parameters, \( \mu_k^{\text{new}} \) and \( (\sigma_k^i)^{\text{new}} \) signify the new model parameters after the adaptation. \( x_k^{\text{adp}} \) is the 4th element of the adaptation feature vector, and \( \alpha \) is called an adaptation constant. From (10) and (11), one can see that an on-line adaptation can be performed when an adaptation vector is available without knowing any knowledge of the training data.

The adaptation process is performed in two steps. First, if the current frame of signal is classified into noise, we will use a simple rejection method to ensure the decision is reliable. To achieve this goal, we compare the likelihood with a rejection threshold which is 10 percent greater than the initial threshold. If the likelihood is greater than the rejection threshold and the previous frame is noise, the current frame of signal is treated as noise which can be used to update the noise model parameters. Second, use the algorithm shown in (10) and (11) to modify the model parameters.

Besides the modification of the model parameters, all the thresholds used in detection process should be updated in the adaptation procedure. Similarly with the mean values, the initial threshold \( I_{\text{TH}} \) is adapted via

\[ I_{\text{TH}}^{\text{adp}} = (1-\alpha)I_{\text{TH}}^{\text{old}} + \alpha l(x^{\text{adp}}) \]  

(12)

where \( I_{\text{TH}}^{\text{old}} \) is the initial threshold before adaptation, \( I_{\text{TH}}^{\text{adp}} \) is the threshold estimated from the adaptation feature vector.

Once the initial threshold is modified, all the other thresholds will be updated accordingly.

5. EXPERIMENT AND RESULTS

Several detection experiments were conducted to evaluate the effectiveness of the likelihood based method, and the difference in detection performances as compared to short-term energy based method and Pitch based method.

The database used is collected from a short-wave communication system. It contains ten isolated mandarin digits from zero to nine. The whole database contains about 16 minutes speech spoken half by one male speaker and half by one female speaker. The signals are sampled at 8kHz to 16 bit resolution with two bytes per sample.

This database is designed mainly for the evaluation of speech detection algorithms. The boundaries of speech and noise segments are labeled.

To simulate different noise environments, we directly add some noise samples to the speech. The noise database used is NOISEX database [7]. This database consists of a limited set of noises representative of the military environment. It is recorded from various noise sources such as: jet-planes, helicopters, wheel carriers, tanks, and command rooms.

The noise database is original sampled at 19.9kHz and with a wordlength of 16 bits. We down-sample it to 8kHz with 16 bits per sample.

Three kinds of noise are used in our experiments. They are white noise, pink noise and machine-gun noise. The white noise has equal energy per unit of bandwidth. The pink noise has equal energy per log unit of bandwidth. Whereas the energy of the machine-gun noise are mainly distributed under 2kHz. The details about the characteristics of the three kinds of noises, please refer to [7].

The signal is analyzed every 10ms with a frame width of 25ms (with Hamming window and preemphasis), and each frame is represented in terms of 12 features.

5.1 Detection performance

In the first set of experiment, we are concerned about the detection performance of the proposed method and the comparison of the method with short-term energy method and Pitch-based method in different noise environments. For the Pitch based method, we use an algorithm same as what described in [4].

As mentioned above, the boundaries of noise and speech in the database are given. The detection results are evaluated in such a way that if both labeled and detected results are speech, the determination is treated as one correct detection. If current frame is labeled as speech frame, but it is detected as noise, the detection is treated as error detection. On the contrary, if a labeled noise frame is detected as speech, the result is counted as one false alarm. In the whole detection process, the false alarm probability is controlled being less than 5%.

In the training process, we use 30 frames of the noise in 15dB condition to train the noise model. Detection methods are performed in different noise levels. The detection results are shown in Fig. 3, Fig. 4 and Fig. 5.

Many observations can be made from the results.

• In high SNR conditions, three methods get very similar detection performances.
• In strong noise environments, the likelihood method performs better than the other two methods.
• When background noises are white noise and pink noise, in high SNR conditions, the Pitch-based method is slightly better than the energy-based method (Fig. 3 and Fig. 4). However, in low SNR conditions or in machine-gun noise environment, the energy method is superior to the Pitch-based method.

![Fig. 3 Detection performance in white noise environment](image1)

![Fig. 4 Detection performance in pink noise environment](image2)

![Fig. 5 Detection performance in machine gun noise environment](image3)

5.2 Adaptation performance
The second experiment is performed to evaluate the adaptation algorithm described in the section 4. The adaptation constant is set to 0.03. The Tab. 1 gives the results under condition that the noise model is trained in 15dB condition (white noise) but tested in different noise level.

<table>
<thead>
<tr>
<th>SNR</th>
<th>15dB</th>
<th>10dB</th>
<th>5dB</th>
<th>0dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>No adaptation</td>
<td>6.45</td>
<td>10.27</td>
<td>13.17</td>
<td>16.70</td>
</tr>
<tr>
<td>Adaptation</td>
<td>6.41</td>
<td>7.91</td>
<td>9.89</td>
<td>12.1</td>
</tr>
</tbody>
</table>

Tab. 1 Error rates with and without adaptation in white noise conditions

From Tab.1, one can see that the adaptation greatly improves the detection performance under different noise levels.

Tab.2 provides the results under condition that the noise model is trained in 15dB white noise environment but tested in 15dB pink noise and 15dB Machine-gun noise. From Tab.2, we can see that if the model is trained in white noise condition, but tested in pink noise and machine-gun noise environments, the performance decreases dramatically. However, an adaptation can increase the performance significantly.

<table>
<thead>
<tr>
<th>SNR</th>
<th>15dB white noise</th>
<th>15dB pink noise</th>
<th>5dB machine – gun noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>No adaptation</td>
<td>6.45</td>
<td>12.56</td>
<td>16.10</td>
</tr>
<tr>
<td>Adaptation</td>
<td>6.41</td>
<td>9.11</td>
<td>9.65</td>
</tr>
</tbody>
</table>

Tab. 2 Error rates with and without adaptation in different noise conditions

6. CONCLUSION
A likelihood based speech signal detection method was investigated in this paper. Various experiments showed that this method outperforms the short-term energy based method and Pitch-based method in different noise conditions. An adaptation method was used to adapt the noise model parameters trained in one condition to a new noisy environment. Experiment results showed that the adaptation method improved the system performance significantly.

In this paper, experiments are only performed on small database. Work is in progress to test the detection method and the adaptation on larger database as well as in various noise conditions.

REFERENCE