End-point detection of speech under low SNR

A.Revathi
Research Scholar
Email: revathidhanabal@rediffmail.com

Dr.R.Chinnadurai
Asst.Professor
Email: rc@nitt.edu
Dept. of ECE
National Institute of Technology, Tiruchirappalli

Dr.Y.Venkataramani
Principal
Saranathan College of Engineering
Tiruchirappalli
Email: principal@saranathan.ac.in

With the increased deployment of voice-based automatic systems over a variety of voice-based services, accurate discrimination of speech frames from background noise and silence frames for providing satisfactory results is necessary especially in the case of speech recognition and speaker recognition.

Many end-point detection methods have been developed over the past several decades. Short-time energy and zero-crossing-rate-based methods provide satisfactory end detection in silent environments. To achieve effective speech communication even in the presence of background noise, it is better if you develop a robust method that works as well in noisy environments.

There are number of methods that help determine end points[3-9]. The method we propose uses the wavelet transform, a powerful transform that has been increasingly used in the signal processing fields[10-14] to extract the frequency content of the signal by exploiting a special property such as time-frequency localisation of wavelet transforms. In this method wavelet energy in the frequency band of interest tracking the energy changes along the speech duration is used to compute wavelet entropy.

**Discrete wavelet transforms**

The wavelet basis function is a small wave from which other wavelets can be derived by proper translation and dilation. Any function can be written as linear combination of a set of basis functions such as wavelets: equation [1]

\[
Y(n) = \sum_{k=-\infty}^{\infty} x(k) h(2^{m} n-k)
\]

Where Yi(n) are wavelet coefficients and h( ) are the set of basis functions derived from the mother wavelets h( ).These wavelet coefficients are suitably interpolated, filtered and combined to reconstruct the original signal. In the tree structured filter bank analysis of wavelet decomposition: equation [2] and [3]

Where Wj(n) and Sj(n) are detail and approximation wavelet coefficients at jth level of decomposition and m is number of levels of decomposition.

Approximation coefficients in the third level of decomposition (0-689Hz) are chosen based on the criterion that most of the speech is concentrated below 1kHz.

**End point detection**

Table 1 indicates the number of wavelet coefficients in each frequency band considering three levels of decomposition in case of 220 samples/frame and 11,025 Hz sampling rate. Approximation coefficients in the third level of decomposition are used to determine wavelet energy. Figure 1 represents the distribution of wavelet energy in the frequency band 0-689Hz. The modified probability density function is obtained as: equation [4]

Where i denotes the frame number and j denotes the sample in a frame. Wj(i) represents the wavelet energy of jth sample in ith frame. The negative wavelet entropy is derived as: equation [5]

Approximation coefficients in the third level of wavelet decomposition are chosen to compute the wavelet entropy because it could provide more immunisation against noise considering the fact that noise energy is concentrated in the high frequency bands. The negative wavelet entropy

![Figure 1: The distribution of wavelet energy in the frequency band 0-689Hz for clean speech.](image)

<table>
<thead>
<tr>
<th>Level of Decomposition</th>
<th>Frequency range in Hz</th>
<th>Number of wavelet coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2756-5512</td>
<td>110</td>
</tr>
<tr>
<td>2</td>
<td>1378-2756</td>
<td>55</td>
</tr>
<tr>
<td>3</td>
<td>689-1378</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 1: Three levels of wavelet decomposition covering the entire frequency band in case of 11,025 samples/sec and 220 samples/frame
Entropy is more negative for noise and silence frames and less negative for speech frames. This feature is explored to determine end points.

This successfully developed algorithm begins by computing the spectral energy of speech and noise. Then spectral energy of noise is made more than the spectral energy of speech by adjusting the amplitudes of speech and noise suitably. Speech of less average spectral energy is combined with noise of more average spectral energy. Figures 1 and 2 show the clean speech and speech with a noisy background.

Speech with background noise is divided into stationary segments of 20ms. Each segment is wavelet decomposed into three levels of decomposition and approximation coefficients in the third level of decomposition are considered for computing wavelet energy. Figure 3 shows the distribution of wavelet energy in the third level of wavelet decomposition of speech with noise. From the plot it is obvious that wavelet energy is tracking the speech in the noisy speech.

Then modified probability density function is computed with the help of equation [4]. In the equation [4], Wi(j) is wavelet energy of jth sample in ith frame. K is some positive constant added with wavelet energy in order to give significant value for probability density function if the sum of wavelet energy of samples in a frame is supposed to be zero. K value is chosen based on the SNR of speech and noise. Modified negative wavelet entropy is computed using the equation [5].

Figure 4 represents the wavelet entropy plot revealing the fact that how the entropy is more negative for noise frames and less negative for speech frames and how the entropy tracks the boundaries of the speech signal.

**End points**

Using entropy values, energy changes between speech and noise can be easily detected. Computing the modified entropy from the wavelet energy in the third level of decomposition generates the curve as shown in the figure 4. Wavelet entropy plot clearly specifies the noise to speech transition at the beginning and speech to noise transition at the end. The algorithm for determining end points from this curve is as follows.

- **Threshold:** The average of the entropy values is chosen as the threshold.
- **Entropy comparison:** For determining starting point entropy values are compared with the threshold in the following manner. If the entropy value of first two frames is less than the threshold and entropy value of third frame is greater than the threshold, then first frame is chosen as the starting frame. Remaining frames are considered for end frame determination. If the entropy

![Figure 2](image1.png)

**Figure 2:** The distribution of wavelet energy for speech with background noise.

![Figure 3](image2.png)

**Figure 3:** Distribution of wavelet energy in the third level of wavelet decomposition of speech with noise.

![Figure 4](image3.png)

**Figure 4:** Modified wavelet entropy plot for: “should we chase.”

![Figure 5](image4.png)

**Figure 5:** Speech signal and end point markers for the sentence “Should we chase.”

![Figure 6](image5.png)

**Figure 6:** Speech signal and logic series markers for the sentence, “Adjacent elements have very large currents,” with SNR = -3.1318db.

![Figure 7](image6.png)

**Figure 7:** Speech signal and end point markers for the sentence, “This signal is known,” with SNR = -4.9623db.
Experimental results

The speech database used in this experiment contains fifty speech samples obtained from different speakers. Each speaker is required to utter ten sentences in English. Every utterance lasts 3-20s. In our experiment, speech has been combined with variable noise in order to simulate a real noisy environment. After applying the above algorithm to the noisy speech following results are obtained as shown in figures 6.

There are some difficult cases for which the end points cannot be detected accurately in normal noise conditions. These cases are:

1. Weak fricatives at the beginning or end.
2. Weak plosives at the beginning or end.
3. Nasals at the end.
4. Voiced fricatives that become devoiced at the end.

These cases are analysed and following are the results obtained for the sentence that begins with voiced fricative and ends with nasals. Figure 7 shows the performance of the algorithm for the sentence starts with voiced fricative and ends with nasal sound.

The performance of the algorithm is also tested on the completely voiced sentences and the following figure 8 reveals the performance.

This algorithm accurately detects the end-points of a speech with noisy background under low SNR conditions. Entropy model is developed based on the computation of probability density function by taking the approximation coefficients in the third level of wavelet decomposition exploiting the time-frequency localisation of wavelet transform. This model is also used to develop the logical series markers tracking the boundaries of the speech signal.

An evaluation of the system is done by superimposing the manually generated noises with speech signal at different signal-to-noise ratios. The system is also tested on some difficult cases of considering some sentences with voiced fricatives at the beginning and nasals at the end and completely voiced sentences. Experimental results shown reveal the performance of the algorithm even for the speech mixed with noise of high spectral energy.

References