1) Invariant outputs at various background energy levels, with maximum improvements of speech detection.
2) Accurate location of detected endpoints.
3) Short time delay or look-ahead.

If we use only one algorithm, it’s hard to satisfy the second and the third items simultaneously. If the average SNR level of the current speech signals is above zero, the short-term SNRs at the speech endpoints are usually much lower than those between the endpoints. Hence, we could use different detection schemes for different part of the speech segment.

The proposed algorithm has two steps to separate the speech segments from the background noise. For the first step, we use the double threshold energy detection algorithm [2] to detect the possible endpoints of the speech segments efficiently. However, the detected endpoints are rough. Therefore, for the second step, we use the GMM based MO-LLR algorithm to search around the possible endpoints for the accurate ones.

By doing so, only the signals around the endpoints need the computationally complex algorithm. Therefore, a lot of detecting time could be saved.

B. Empirical Rules Based Energy Detection

The efficient energy detection algorithm is not only to detect the apparent speeches but also to find the approximate positions of the endpoints. But the algorithm is not robust enough when the SNR is low. To enhance its robustness, we integrate it with a group of rules and present it as follows:

Part1 As for the beginning-point (BP) detection, the silence energy and the low/high energy thresholds of the nth observation \( o_n \) are defined as

\[
E_{\text{sil}} = \frac{1}{3} \sum_{j=n-1}^{n+1} E_j \tag{1}
\]

\[
T_{h_{\text{low}}} = \alpha \cdot E_{\text{sil}}, \quad T_{h_{\text{high}}} = \beta \cdot E_{\text{sil}} \tag{2}
\]

where \( E_j \) is the short-term energy of the jth observation, \( \alpha, \beta \) are the user defined threshold factors.

Given a signal segment \{\( o_{n}, o_{n+1}, \ldots, o_{n+N_{B}-1} \}\) with a length of \( N_{B} \) observations, if there are several consecutive observations with a length of \( N_{B} \) in the signal whose energy is higher than \( T_{h_{\text{low}}} \), and if the ratio \( \tilde{N}_{B}/N_{B} \) is higher than an empirical threshold \( \varphi_{B_{\text{low}}} \), the first observation \( o_{B} \) whose energy is higher than \( T_{h_{\text{low}}} \), should be remembered.

And then we detect the given segment starting from \( o_{B} \), if there are another consecutive observations with a length of \( \tilde{N}_{B} \) whose energy is higher than \( T_{h_{\text{high}}} \) and the ratio \( \tilde{N}_{B}/N_{B} \) is higher than another empirical threshold \( \varphi_{B_{\text{high}}} \), one possible beginning-point \( o_{B} \) is detected in the segment \{\( o_{n}, \ldots, o_{n+N_{B}-1} \}\).

Part2 As for the ending-point (EP) detection, suppose that the energy of the current observation \( o_{E} \) is lower than \( T_{h_{\text{low}}} \), we analyze the subsequent signal segment with a length of \( N_{E} \) observations. If there are \( N_{E} \) observations whose energy is higher than \( T_{h_{\text{high}}} \) in the subsequent signal segment, and if the ratio \( \tilde{N}_{E}/N_{E} \) is lower than an empirical threshold \( \varphi_{E_{\text{high}}} \), one possible ending-point is detected as the current observation \( o_{E} \).

C. GMM Based Multiple-Observation Log Likelihood Ratio Algorithm

Although the energy based algorithm is efficient to detect the speech signals roughly, the endpoints detected by it are not sufficiently accurate. Therefore, some computationally complex algorithm is needed to find the endpoints accurately. Here, a new algorithm called the GMM based multiple-observation log likelihood ratio (MO-LLR) algorithm is proposed.

Given the current observation of the VAD \( o_{n} \), a window \{\( o_{n-l}, \ldots, o_{n-1}, o_{n}, o_{n+1}, \ldots, o_{n+m} \}\) is defined over \( o_{n} \). Acoustic features \{\( x_{n-l}, \ldots, x_{n+m} \}\) are extracted from the window. Two \( K \)-mixture GMMS are used to model the speech and noise distributions respectively

\[
P(x_i|H_1) = \sum_{k=1}^{K} \pi_{1,k}N(x_i|\mu_{1,k},\Sigma_{1,k}) \tag{3}
\]

\[
P(x_i|H_0) = \sum_{k=1}^{K} \pi_{0,k}N(x_i|\mu_{0,k},\Sigma_{0,k}) \tag{4}
\]

where \( i = n - l, \ldots, n + m \), \( H_1 (H_0) \) denotes the hypothesis of the speech (noise), and \( \{\pi_k, \mu_k, \Sigma_k\} \) are the parameters of \( k \)th mixture of the GMM.

After above definition, the log likelihood ratio (LLR) \( s_i \) of the observation \( o_i \) can be calculated as

\[
s_i = \log (P(x_i|H_1)) - \log (P(x_i|H_0)) \tag{5}
\]

And the hard decision on the LLR \( s_i \) is obtained by

\[
c_i = \begin{cases} 
1, & \text{if } s_i \geq \epsilon \\
0, & \text{otherwise}
\end{cases} \tag{6}
\]

1) Here, we use the MFCC, its delta and delta-delta features as the feature, which has a total dimension of 39. But the proposed method is not limited to the feature.