Effective Energy Feature Compensation Using Modified Log-energy Dynamic Range Normalization for Robust Speech Recognition

Author : Yoonjae LEE, Hanseok KO

Professor: 陳嘉平

Reporter: 吳國豪

Outline

- Introduction
- Log Energy Dynamic Range Normalization (ERN)
- Proposed Energy Compensation
- Experiments and Discussion

Introduction

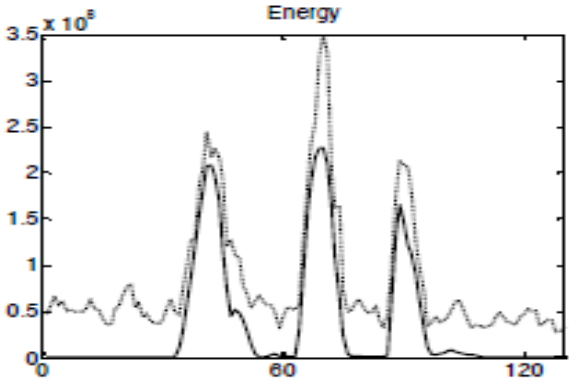
- The mismatch between the training and test conditions is a significant factor in the degradation of speech recognition system performance.

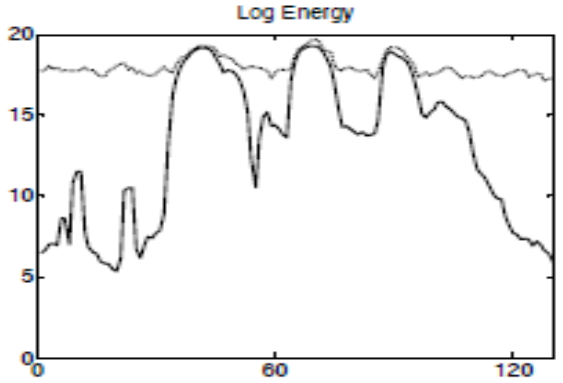
- The energy of a speech signal is widely used as an element of the feature vector for speech recognition.

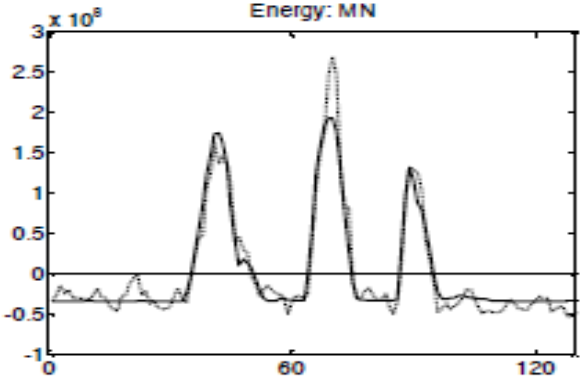
Introduction

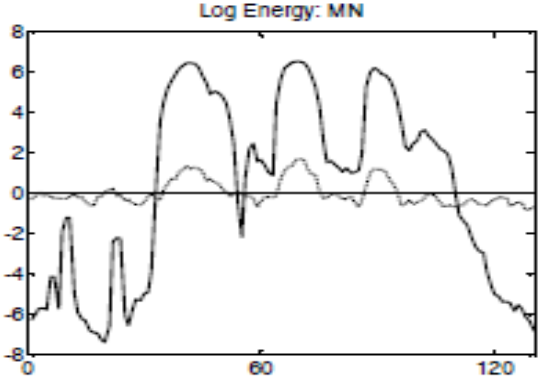
- The problems of the energy feature:
  (1) the energy of a signal can change according to environmental conditions.
  (2) the signal energy in a clean environment is different from that in a noisy environment.

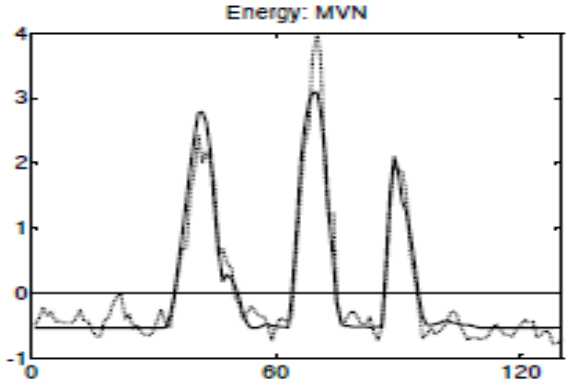
- In order to compensate for these variations, several energy normalization methods have been introduced.

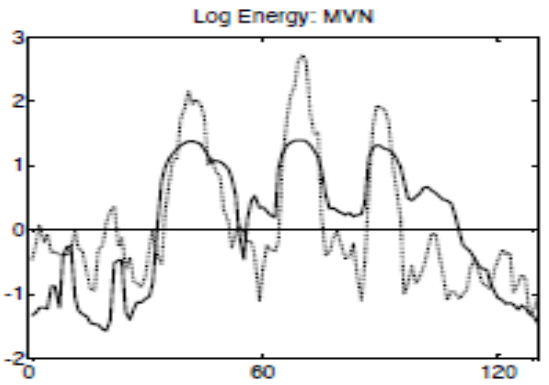

Energy
x 10^6
3.5
3
2.5
2
1.5
1
0.5
0
0
60
120


Log Energy
20
15
10
5
0
0
60
120


Energy: MN
x 10^6
3
2.5
2
1.5
1
0.5
0
-0.5
-1
0
60
120


Log Energy: MN
8
6
4
2
0
-2
-4
-6
-8
0
60
120


Energy: MVN
4
3
2
1
0
-1
0
60
120


Log Energy: MVN
3
2
1
0
-1
-2
0
60
120

ERN

- We define a log-energy dynamic range of the sequence as follows:

$$D.R.(dB) = 10 \times \frac{E_{\max}}{E_{\min}}$$

1. Find the value of $E_{\max}$ *and* $E_{\min}$.

2. Define a fixed value of DR and use it to find $T_{\min}$.

$$T_{\min} = \alpha \times E_{\max}$$

ERN

3. If $E_{\min} < T_{\min}$ *, proceed to step 4.*

4. For all frames,

$$\tilde{E}_i = E_i + \frac{T_{\min} - E_{\min}}{E_{\max} - E_{\min}} \times (E_{\max} - E_i)$$

ERN

1. It is difficult for **the high log energies** including the maximum log energy, to be changed by noise. However, **in low SNR environments**, the maximum and high log energies of clean speech are readily changed by noise, as shown in Fig. 1 (0 dB SNR).

2. If the minimum of test log energy, $E_{\min}$ is larger than $T_{\min}$, no process is applied to the test data.

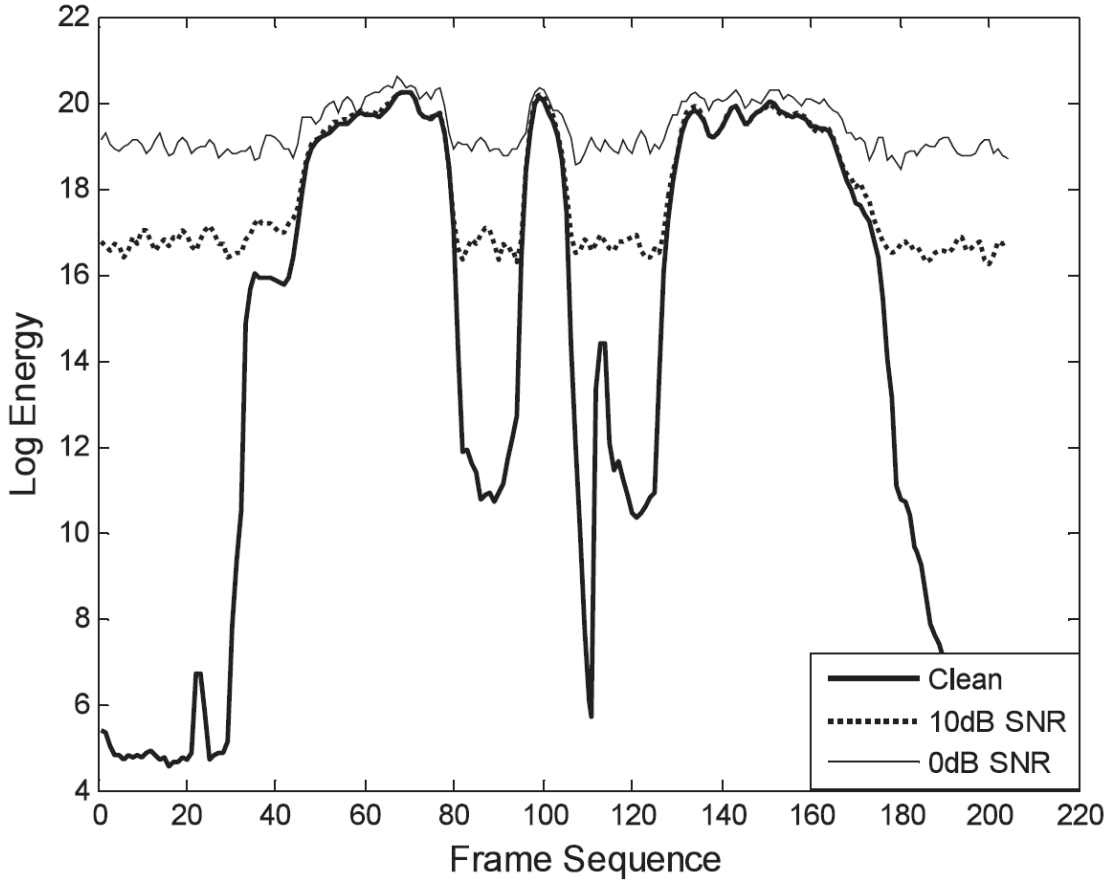


Fig. 1 Comparison of log energy sequences between clean, 10 dB and 0 dB SNR car noisy speech.

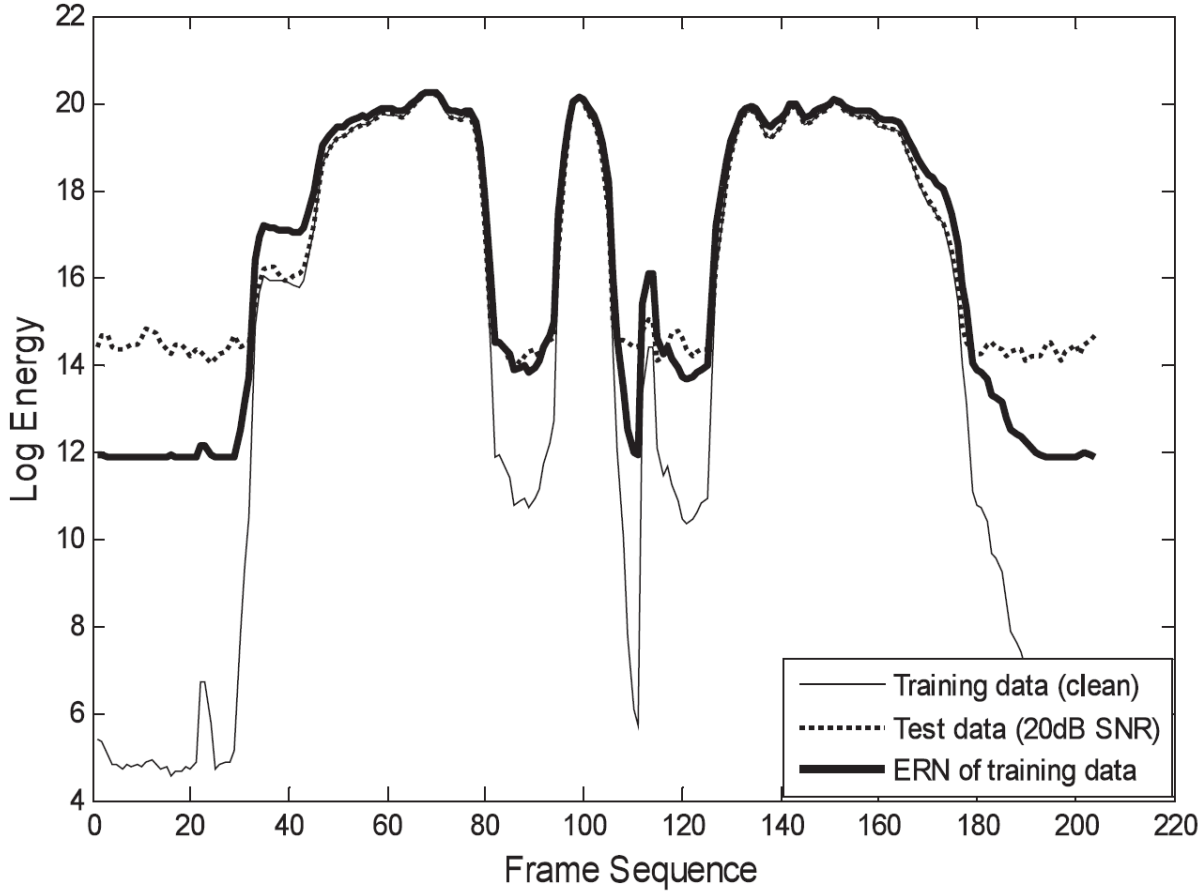


Fig. 3 Comparison of log energy with 20 dB SNR noisy speech in car environment and result of ERN algorithm with DR=17 dB.

Energy Subtraction (ES)

1. The **noise energy** is estimated and subtracted from the entire noisy speech energy.

2. It is assumed that there is no speech signal in the first 10 frames and consequently, the noise energy is estimated as an average of the first 10 frame energies.

3. Smoothing

$$\hat{E}_i = \alpha \hat{E}_{i-1} + (1-\alpha)\widehat{E}_i \quad where\ 0 < \alpha < 1$$

Modified ERN (MERN)

- To determine the low log energy regions, we use voice activity detection (VAD).

- The new transformed log energy:

$$\bar{E}_i = \frac{E_i - K \times E_{\max}}{(1-K)} \quad where\ K = \frac{E_{\min} - T_{\min}}{E_{\max} - T_{\min}}$$

Hybrid Method

1. Classify the non-speech dominant region and speech dominant region.

2. If $E_{\min} < T_{\min}$, proceed with ERN in the non-speech dominant region and proceed with ES in the speech dominant region.

3. If $E_{\min} > T_{\min}$, proceed with MERN in the non-speech dominant region and proceed with ES in the speech dominant region.

Hybrid Method

1. The log energy in the speech dominant region always becomes the energy in a clean environment, like the energy of training data.

2. The log energy in the non-speech dominant region shows that the mismatch between the energy of training data and test data can be reduced.

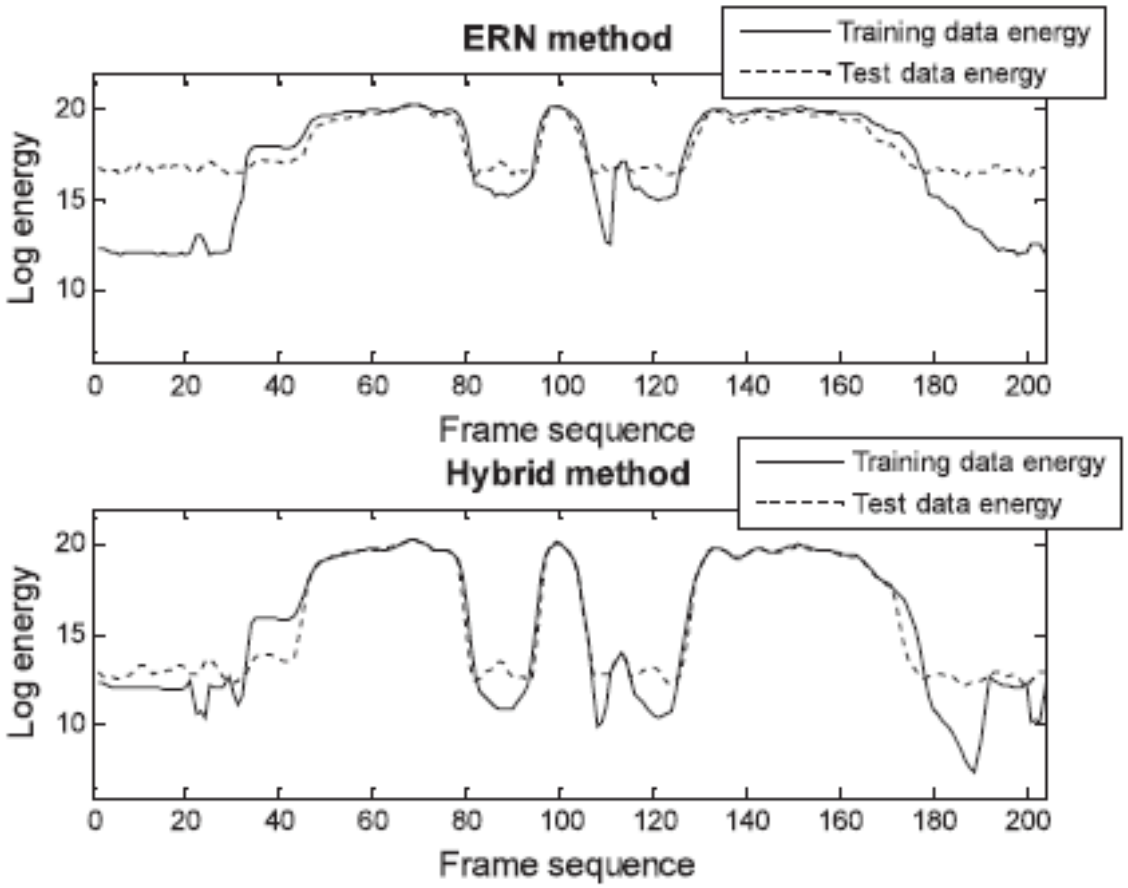


Fig. 4 Comparison of the ERN and proposed hybrid method in the case of a 10 dB car noisy speech energy sequence.

Experiments and Discussion

- Aurora 2.0

- The 13-dimensional Mel Frequency Cepstral Coefficient (MFCC), including the log energy, is used for feature parameter. The delta and the delta-delta features are evaluated. By combining these features, a 39-dimensional feature vector is generated

- Three sets of sentences under several conditions were prepared by contaminating them with noise at SNRs ranging from −5dB to 20 dB and clean condition.

Table 2 Word accuracy of the various algorithms in the case of the car noise condition in Aurora2.0(%).

SNR	Baseline	ES	ERN	MERN	Hybrid
clean	98.96	98.96	98.90	98.90	98.54
20dB	97.41	97.46	96.72	96.96	97.67
15dB	90.04	94.01	94.27	94.78	96.03
10dB	67.01	84.19	85.54	86.76	88.25
5dB	34.09	58.66	59.23	64.99	69.25
0dB	14.46	28.24	27.02	35.31	35.46
-5dB	9.39	11.42	10.71	15.72	14.79
Avg	60.60	72.51	72.56	75.76	77.33

Table 3 Average word accuracy of the various algorithms for all data sets in Aurora2.0(%). (SS: Spectral Subtraction, MVN: cepstral Mean and Variance Normalization)

	SetA	SetB	SetC
Baseline	61.34	55.75	66.14
ES	67.42	66.06	62.57
ERN	72.71	71.90	61.66
MERN	73.82	73.53	62.32
Hybrid	75.64	74.99	65.49
SS	71.40	65.62	73.76
SS + Hybrid	81.65	79.27	74.35
MVN	62.50	62.25	67.70
MVN + Hybrid	80.17	80.97	79.22