Energy Feature Normalization for Robust Speech Recognition in Noisy Environments*

Yoonjae Lee · Hanseok Ko****

ABSTRACT

In this paper, we propose two effective energy feature normalization methods for robust speech recognition in noisy environments. In the first method, we estimate the noise energy and remove it from the noisy speech energy. In the second method, we propose a modified algorithm for the Log-energy Dynamic Range Normalization (ERN) method. In the ERN method, the log energy of the training data in a clean environment is transformed into the log energy in noisy environments. If the minimum log energy of the test data is outside of a pre-defined range, the log energy of the test data is also transformed. Since the ERN method has several weaknesses, we propose a modified transform scheme designed to reduce the residual mismatch that it produces. In the evaluation conducted on the Aurora2.0 database, we obtained a significant performance improvement.

Keywords: Log-energy Dynamic Range Normalization (ERN), Energy-Subtraction, Inverse Transform ERN, Speech Recognition

1. Introduction

The mismatch between the training and test conditions is a significant factor that degrades the performance of the speech recognition system. Thus, finding techniques that make both conditions equal is one of the most essential and important issues in ASR (Automatic Speech Recognition) system.

The energy of a speech signal is helpful in discriminating between different sounds,

---

* This work was supported by grant No. A17-11-02 from Korea Institute of Industral Technology Evaluation & Planning Foundation.

** Dept. of Electronics and Computer Engineering, Korea University

such as voiced sounds or unvoiced sounds. Hence, it is widely used as an element of the feature vector for speech recognition. However, the energy of a signal can change according to the environmental conditions. For example, the energy of an utterance is dependent on the loudness of the voice or speakers. In addition, the signal energy in a clean environment is different from that in a noisy environment. To compensate for these variations, several energy normalization methods have been introduced.

In conventional energy normalization schemes, the maximum log energy is subtracted from all of the frames. As a result, the energy is always less than zero, regardless of the environments. The Hidden Markov Model Toolkit (HTK) supports this method[1]. However, it is not robust in noisy environments. The mean and variance normalization method and the log energy dynamic range normalization (ERN) method were recently introduced[2][3]. In the ERN method, the log energy in a clean environment is transformed into the log energy in a noisy environment using a predefined dynamic range (D.R). The acoustic model is obtained using the feature vector with the transformed log energy, and it is assumed that the maximum log energy of the clean speech is not affected by the additive noise. However, this assumption does not hold when the signal to noise ratio (SNR) is low. In addition, the acoustic model with energy parameters transformed by a fixed D.R is not effective over a wide range of SNRs, although it shows the highest average relative improvement using the given fixed D.R.

In this paper, we propose two methods. In the first method, we assume that the relationship between the energy of clean speech and the energy of noise is additive. Therefore, we are able to subtract the noise energy component from the noisy energy. In the second method, we modify the ERN algorithm in order to reduce the residual mismatch using the inverse transform of the ERN and the first method.

This paper is organized as follows. First, we review the original ERN algorithm and describe its weakness in Section 2. We then describe the proposed scheme in Section 3. In Section 4, we present and discuss the experimental results. Finally, in Section 5, we make our concluding remarks and discuss future works.

2. Log Energy Dynamic Range Normalization (ERN)

The log energy sequence of clean speech is changed by additive noise, as shown in Figure 1.

In the ERN method, it is assumed that the maximum value of the log energy of the clean speech is not affected by the additive noise. The mismatch between the clean and noisy speech is larger at low log energy. As a result of this mismatch, the energy sequence is transformed, as shown in Figure 2. To reduce this mismatch, more weight is given to low log energies[3].

First, the dynamic range is defined as follows.

$$D.R(dB) = 10 * \frac{Max(\log - energy_i)}{Min(\log - energy_i)}, i = 1,..,n \tag{1}$$

$Max(\log - energy_i)$ and $Min(\log - energy_i)$ are the maximum and minimum values of the log energy sequence, and D.R is the ratio of the maximum to the minimum log energy. Next, the Target Minimum (T_Min) is defined. The minimum log energy of clean speech is transformed so as to increase its value to T_Min. T_Min is obtained by the substitution of $Min(\log - energy_i)$ into equation 1 to obtain equation 2.

$$D.R(dB) = 10 * \frac{Max(\log - energy_i)}{T\_Min} \tag{2}$$

If we set the value of D.R, we can find the value of T_Min. the following steps are used in the ERN algorithm.

1. Find the value of $Max = Max(\log - energy_i)$ and $Min = Min(\log - energy_i)$
2. Define the fixed value of D.R and use it to find T_Min.
3. If $Min < T\_Min$, proceed to step 4 to increase the energy.
4. For all frames,

$$new\_\log\_energy_i = \log\_energy_i + \frac{T\_Min - Min}{Max - Min} * (Max - \log\_energy_i) \tag{3}$$

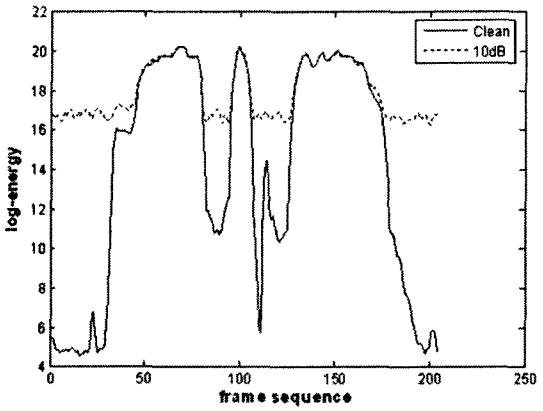

Figure 1. Comparison of log energy sequences between clean and 10 dB car noisy speech.

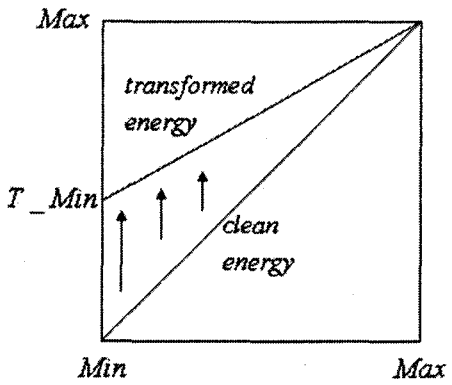

Figure 2. Schematic representation of the energy transform.

However, the ERN algorithm has the following drawbacks.

1. Figure 3 shows the speech energy transformed using the ERN algorithm with an optimal D.R of 17 dB. The minimum energy of the noisy speech is higher than that of the target minimum even though it is in a high SNR environment (20 dB). Therefore, the ERN algorithm is not applied to the test data. In addition, mismatch still exists at low log energy and can be made worse when the SNR is lower.

2. It is difficult for the maximum log energy to be changed by the noise because of the characteristics of the log function. However, in low SNR environments, the maximum log energy of clean speech is readily changed by the noise, as shown in Figure 4.

3. Since the energy of training data is transformed with fixed optimal D.R, it is not

robust for all SNR environments.

Because of these drawbacks, it is difficult to reduce the mismatch effectively for all SNR environments. To solve the above problems, we propose the energy subtraction method and the inverse transform ERN combined with energy subtraction.

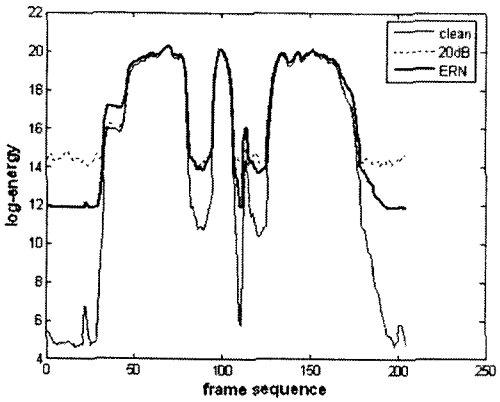

Figure 3. Comparison of log energy with 20 dB noisy speech in car environment and result of ERN algorithm with D.R=17 dB.

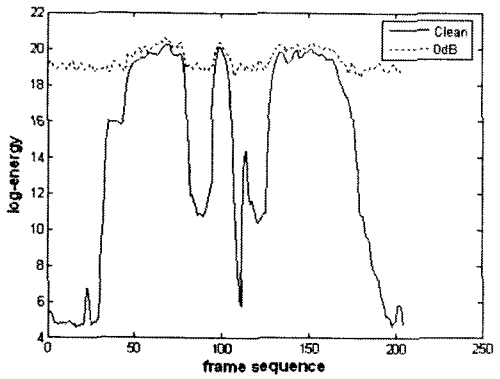

Figure 4. Comparison of log energy with clean speech and 0 dB noisy speech in car environment.

3. Proposed Methods

3.1 Energy Subtraction method (ES)

We assume that the relationship between the clean speech energy and noise energy is

additive before the log function is applied. Therefore, we estimate the noise energy and subtract it from the entire noisy speech energy. We assume that there is no speech signal in the first 10 frames and consequently, we estimate the noise energy as the average of the first 10 frame energies.

We set the threshold energy to 150, in order to prevent the energy from becoming very small or negative. Since the amplitude of the signal is not quite zero in the silence region, the threshold is experimentally determined log energy value of 5.

3.2 Inverse transform ERN combined with ES

The acoustic model having the transformed energy parameter with fixed D.R is not reliable in variable SNR environments. Especially, the mismatch of the speech energy is a serious problem, because the speech energy is important for robust speech recognition. In practice, we can see that the performance of the ERN algorithm is slightly degraded in high SNR environments and that the ERN algorithm with a fixed D.R does not obtain the highest performance at all SNRs. In addition, the assumption that the maximum log energy of the clean speech is not affected by the noise does not hold at low SNRs.

In an attempt to overcome these difficulties, we refrain from transforming the high log energy of the training data in the clean environment, and instead compensate for the high log energy of the test data. In this paper, we set the D.R to a fixed value of 17 dB. This D.R is an optimal value in the case of the ERN algorithm. Then we apply the ERN algorithm to those values of the training data which are less than the threshold Th, as in equation 4.

$$Th = \alpha * Min + (1-\alpha) * Max \tag{4}$$

We let $\alpha = 0.5$ (Half-ERN). Min and Max mean the minimum and maximum log energies of the all frames, respectively. If the minimum log energy of the test data is less than the target minimum, we also apply the Half-ERN method.

If the minimum log energy of the test data is larger than the target minimum, we apply the ES or inverse transform ERN (IT-ERN) to the test data. We also set the threshold to the log energy of the test data using equation 4 with $\alpha = 0.5$.

In the case of energies below the threshold, we apply the IT-ERN. In contrast to the ERN, the IT-ERN is the technique to fall off the low log energy in order to reduce the

mismatch observed in Figure 3. Therefore, we decrease the energy, so that the minimum energy reaches to the target minimum.

To accomplish this, in equation 3, we interchange Min and T_Min. Then, we calculate the inverse function. As a result, we can obtain the new transformed log energy as follows.

$$estimated\_e_i = \frac{noisy\_e_i - K * Max}{(1-K)} \tag{5}$$

$$where\ K = \frac{Min - T\_Min}{Max - T\_Min}$$

$noisy\_e_i$ refers to the i^{th} frame log energy of the test data.

For energy values above the threshold, we apply the ES method. Since we estimate the noise energy from the test data, we can compensate for the relatively high energy of the noisy speech regardless of the SNR. If the estimated noise energy is larger than the noisy energy, we do not perform the ES. In this way, we can guarantee that the energy is always positive. Finally, the variation of the log energy is large for small changes at low energy. So, as a post-processing step, we utilize a 3-point moving average procedure to smooth this large variation. The entire proposed algorithm is depicted in Figure 5. Figure 6 shows the results of the ERN and the IT-ERN combined with ES. We can see that the mismatch of the proposed method is less than that of the ERN.

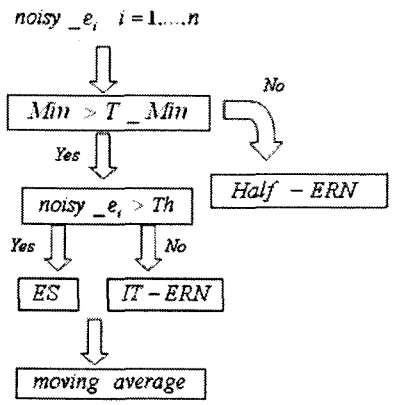

Figure 5. Block diagram of the proposed method

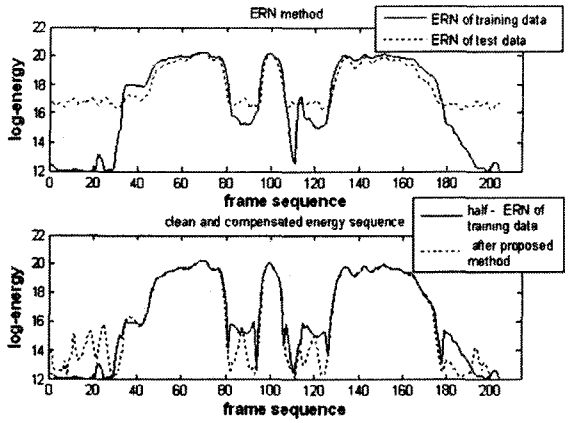

Figure 6. Comparison of the ERN and proposed algorithm in the case of a 10 dB car noisy speech energy sequence.

4. Experiments and Results

In this paper, we followed the Aurora2.0 evaluation, the procedure for performance verification, along with identical conditions suggested in the Aurora2.0 procedure[4][5].

The Aurora2.0 database contains English connected digits recorded in clean environments. Three sets of sentences under several conditions (e.g. Set A: subway, car noise, Set B: restaurant, street and train station noise, Set C: subway and street noise) were prepared by contaminating them with noise at SNRs ranging from -5 dB to 20 dB and clean condition. A total of 1001 sentences are included in each noise condition.

Table 1. Word accuracy of the various algorithms in the case of the car noise condition in Aurora2.0(%)

SNR	Baseline	ERN	ES	IT-ERN With ES
clean	98.96	98.90	98.96	98.54
20 dB	97.41	96.72	97.46	97.55
15 dB	90.04	94.27	94.01	95.17
10 dB	67.01	85.54	84.19	86.64

5 dB	34.09	59.23	58.66	65.73
0 dB	14.46	27.02	28.24	34.30
-5 dB	9.39	10.71	11.42	14.97
Avg	60.60	72.56	72.51	75.88

Table 2. Average word accuracy of the various algorithms for all data sets in Aurora2.0(%)

	Baseline	ERN	ES	IT-ERN With ES
Set A	61.34	72.71	67.42	76.08
Set B	55.75	71.90	66.06	76.80
Set C	66.14	61.66	62.57	64.97

In Table 1, ERN is the original algorithm with a fixed D.R of 17 dB. In the case of the ERN method, the performance in the 20 dB noisy environment is slightly degraded because the variation of the transformed speech energy is larger than the real variation induced by the noise. This lack of robustness to various SNR environments is a problem of the ERN algorithm. The performance of the ES algorithm is similar to that of the ERN algorithm. However, we can see that the performance of ES falls off, as compared with the ERN algorithm in various environments in Table 2. Since the variance of the car noise is relatively small, we can estimate the noise energy well using the energy of the first 10 frames. However, the variances of the other noises are not as small as that of car noise, so it is not sufficient to use the energy of the first 10 frames to estimate their noise energies.

In the case of the IT-ERN combined with ES algorithm, we obtained a substantial improvement over the original ERN and ES methods. In this algorithm, we do not transform the relatively high log energy values of the training data. Instead, we subtract the estimated noise energy (not the log energy) from the relatively high energy values of the test data. Since we compensate for the energy of the noisy test data by taking the noisy environment into considerations, it is very robust to variable environments. In addition, we reduced the mismatch at low log energy to a greater extent than in original ERN method using the inverse transformation. That is the reason for the improvement.

However, we obtained degraded results in the case of Set C database with additive

noise and channel distortion. We believe that the channel distortion makes it difficult to estimate the energy of the noise. Although all of the algorithms cause the degraded results in the case of Set C database, the proposed method is more robust than the original ERN algorithm.

In this paper, we obtained average word accuracies of up to 76.08%, 76.80% and 64.97% for the Set A, Set B and Set C database, respectively, using only energy normalization.

5. Conclusions

In this paper, we proposed the log energy feature normalization algorithm for robust speech recognition. The two proposed algorithms consist of the energy subtraction and inverse transform ERN combined with energy subtraction. The energy subtraction is effective in a stationary noisy environment. Also, we obtained an improvement over the original ERN using the inverse transform ERN combined with energy subtraction. In the high log energy region, we subtract the estimated noise energy from the noisy speech energy and we fall off the log energy to the target minimum in the low log energy region. By means of the proposed algorithm, it is possible to make the log energy feature more robust to various environments. In a future work, we intend to focus on the problem of robust energy normalization for channel distortion.

References

[1] Young, S. et al., 2002. *The HTK book (for HTK version 3.2*, Cambridge University Engineering Department.

[2] Ahidi, S. M., Sheikhzadeh, H., Brennan, R. L. & Freeman, G. H. 2004. "An Energy Normalization Scheme for Improved Robustness in Speech Recognition", *ICSLP2004*, 3, 1649-1652.

[3] Zhu, W. & O'Shaughnessy, D. 2005. "Log-energy Dynamic Range Normalization for Robust Speech Recognition", *ICASSP2005*, 1, 245-248.

[4] Hirsch, H. G. & Pearce, D. 2000. "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions",

ISCA ITRW ASR2000.

[5] ETSI standard document, Speech Processing, Transmission and Quality aspects (STQ), 2000; Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms, ETSI ES 201 108 v1.1.3 (2000-04).

received: January 30, 2006
accepted: March 15, 2006

▲ Yoonjae Lee
Dept. of Electronics Engineering, Korea University
5Ka-1, Anam-dong, Sungbuk-ku, Seoul, 136-701, Korea
Tel: +82-2-922-8997 Fax: +82-2-3291-2450
H/P: 016-564-9619
E-mail: yjlee@ispl.korea.ac.kr

▲ Hanseok Ko
Dept. of Electronics Engineering, Korea University
5Ka-1, Anam-dong, Sungbuk-ku, Seoul, 136-701, Korea
Tel: +82-2-3290-3239 Fax: +82-2-3291-2450
H/P: 011-9001-3239
E-mail: hsko@korea.ac.kr