A LOG-ENERGY SCALING NORMALIZATION SCHEME FOR ROBUST SPEECH RECOGNITION

Hung-Shin Lee, Hung-Bin Chen, Berlin Chen
National Taiwan Normal University, Taipei, Taiwan
{g95470153, james, berlin}@csie.ntnu.edu.tw

ABSTRACT

The log-energy parameter, as an auxiliary but influential feature, has been commonly used to augment Mel-frequency cepstral coefficients (MFCCs) to improve the recognition accuracy in automatic speech recognition (ASR). In this paper, a new and effective scaling approach named log-energy scaling normalization (LESN), which utilizes special nonlinear scaling functions on noisy speech data for log-energy normalization, is investigated. The scaling function is contrived conceptually from the relationship between clean and noisy data. In the experiments carried out with the Aurora-2 database, LESN is shown average word error rate (WER) improvements of 37.53%, 43.93% and 9.66% for Test Sets A, B, and C, respectively, when compared with the baseline processing. The results also show that LESN outperforms other similar approaches, such as log-energy dynamic range normalization (ERN). Furthermore, as well as being integrated with cepstral mean and variance normalization (CMVN), LESN is further applied to the LVCSR task and a considerable gain is achieved.

Index Terms—Speech recognition, Robustness, Acoustic noise, Speech enhancement, Speech processing

1. INTRODUCTION

The performance of automatic speech recognition (ASR) often deteriorates in the presence of noise. Therefore, the robustness of ASR in various noisy environments is a cardinal issue to many practical applications. A large number of techniques have hitherto been proposed to improve speech recognition robustness. In general, they can fall into three main categories [1]: 1) speech enhancement, which is to reduce the noisy effect on the observed speech signals, e.g. spectral subtraction (SS) [2], 2) model-based noise compensation, which is to adjust the recognition model parameters for better matching the testing (noisy) conditions, e.g. maximum a posterior (MAP) [3], and 3) robust speech feature extraction, which is to find robust representations of speech signals less influenced by additive or channel noises, e.g., cepstral mean and variance normalization (CMVN) [4].

In the first category mentioned above, there are many efforts devoted to restoring the original frame energy of speech signals under the influence of noise in the standard MFCC feature extraction scheme. As has been long and widely known, the frame energy of speech signals contains salient information. Despite that, the energy-alone processing or normalization is considered a lightweight task without the need of a complicated computation procedure. In this regard, the energy-related parameters, i.e., energy itself and its dynamic information from the adjacent frames, are particularly to be the focus of our proposed approach. Conventionally, the energy parameters are replaced with the log-energy ones owing to not only the similarity to the logarithm processing performed by the human auditory system [5], but also the stability to the abrupt changes in the level of energy [6].

To date, a few approaches related to the enhancement of log-energy features have been proposed. One of the noticeable approaches is log-energy dynamic range normalization (ERN) [7]. The main idea of ERN is based on a phenomenon that lower-energy frames are more vulnerable to additive noise than higher-energy ones, and actually, each log-energy frame of the noisy speech is higher than its corresponding frame of the clean speech. Thus, the work of ERN is to normalize or elevate the log-energy of the clean speech by using linear or nonlinear interpolations, so that the original (clean) log-energy features can approximate the target (noisy) ones to reduce the environmental mismatch.

In this paper, inspired by ERN, we propose an alternative approach which we call log-energy scaling normalization (LESN). LESN is based on some kind of nonlinear data-fitting techniques conducted on the noisy speech instead of the interpolation-based approach, which is utilized by ERN and conducted on the clean speech. We use two methods to exemplify LESN and the corresponding results reveal that LESN can lead to better performance on the test sets with either stationary or non-stationary additive noises.

The remainder of the paper is organized as follows. ERN [7] is reviewed in Section 2. Then, two alternative methods for LESN are introduced in Section 3. The experimental setup and results are presented in Section 4. Finally, a summary of the conclusions is given in Section 5.

2. LOG-ENERGY DYNAMIC RANGE NORMALIZATION

The dynamic range of a log-energy feature sequence is defined as follows [7]:

\[ D.R.(dB) = 10 \times \frac{\text{Max}(\text{Energy}_i)_{i=1..n}}{\text{Min}(\text{Energy}_i)_{i=1..n}} \]  

(1)

where \( \text{Max}(\text{Energy}_i)_{i=1..n} \) is the maximum value of the log-energy feature sequence, \( \text{Min}(\text{Energy}_i)_{i=1..n} \) is the minimum value, and \( n \) is the number of total frames in one utterance. Suppose \( \text{Max}(\text{Energy}_i)_{i=1..n} \) is the same for both original and target dynamic ranges. The target minimum value \( TMin \) can be calculated based on Eq. (1) for a given target dynamic range. Then, the log-energy normalization algorithm is expressed as follows:

(1) For each utterance of the clean speech, find...
Max = Max(Log(Energy\_i)\_{i=1..n}),
Min = Min(Log(Energy\_i)\_{i=1..n}).

(2) Calculate TMin.
(3) If Min < TMin, then there are two ways to update the log-
energy feature Log(Energy\_i) for each frame i:
For linear scaling:
Log(Energy\_i) = Log(Energy\_i) +
\frac{TMin - Min}{Max - Min} \times (Max - Log(Energy\_i)).
For nonlinear scaling:
Log(Energy\_i) = Log(Energy\_i) +
\frac{TMin - Min}{log Max - log Min} \times (log Max - log(Log(Energy\_i))).

Both of the above two scaling methods belong to interpolation
methods (for calculating a new point between two existing data
points), and it is easy to verify that the above nonlinear scaling
adopts the form y = x + a log x + b as an interpolant.

3. LOG-ENERGY SCALING NORMALIZATION

3.1. The basic idea of LESN

The log-energy features of the clean speech and the corresponding
noisy speech with signal-to-noise ratios (SNRs) of 5 dB and 15 dB,
respectively, are shown in Figure 1. The relationship between the
clean speech and the noisy speech are expressed in quantic (fifth-
order) regression, where the former is taken as the independent
variable. From Figure 1, we can find some clues:
1. The quantic-regression shapes of log-energy features of SNRs
of 5 dB and 15 dB are much similar.
2. Lower log-energy frames are more affected by noise than
higher ones, i.e., the lower the log-energy frame of the clean
speech, the larger the difference between the clean and the
Corresponding noisy speech.
As can be seen in the preceding section, ERN uses linear or
nonlinear interpolation to elevate the log-energy features from the
original (clean) speech to its presumed target (noisy) speech.
Inversely, in LESN, we take the clean speech as the original
and the noisy speech as the target. Our effort is to directly return
the given noisy speech to the target speech by means of some proper
scaling functions. In a sense, the scaling operation is to “lower” the
regression curves in order to fit the straight line in Figure 1. In next
subsections, we will introduce two methods to show how LESN
works. One is the quantile-based method, namely LESN1, and the
other is the weight-based one, namely LESN2.

3.2. Example: LESN1

First, according to the characteristics of the noisy speech, which
are mentioned in the above subsection, a transformation function
can be designed as follows:
\[ Log(Energy\_i) = Log(Energy\_i) + \log N \times \frac{Log(Energy\_i) - Min}{Max - Min}. \quad (2) \]
Following are the steps of LESN1 algorithm:
(1) For each utterance of the clean speech, find

![Figure 1: Plot of log-energy in multi-condition versus clean-
condition with quantic regression.](image)

\[ Max = Max(Log(Energy\_i)\_{i=1..n}), \]
\[ Min = Min(Log(Energy\_i)\_{i=1..n}). \]

(2) Calculate L = \frac{Max - Min}{N} \] (3)
where N is the number of given quantiles and L represents the
width of each quantile.
(3) Define a scaling function as
\[ W(x) = \frac{\log x}{\log N} \]

(4) Calculate
\[ Index_i = \frac{\log(Energy\_i) - Min}{L} \quad i = 1..n, \]
where Index\_i is the index quantile to which the i-th frame
belongs.
(5) Update Log(Energy\_i) for each frame i by the following
updating function:
\[ Log(Energy\_i) = Log(Energy\_i) \times W(Index_i). \]

Therefore, Eq. (2) can be taken as a function of the log-energy
features of the noisy speech and can approximate the target speech
for robust speech recognition.

3.3. Example: LESN2

For LESN2, the weight-based method, we redefine the scaling
function, and the updating function is expressed as follows:
\[ Log(Energy\_i) = Log(Energy\_i) \times \left( \frac{Log(Energy\_i) - Min}{Max - Min} \right)^\beta, \quad (4) \]
where the exponential term $\beta$ is the weight of the scaling function.
In the implementation of LESN2, $\beta$ can be efficiently estimated
by using the data-fitting scheme. Data fitting is a mathematical
optimization method which, when given a series of data points
\((u_i, v_i)\) with $i = 1, \ldots, n$, attempts to find a function $G(u_i)$ whose
output $\hat{v}_i$ closely approximates $v_i$. That is, it minimizes the sum
of the squares error (or the squares of the ordinate differences)
between the points $(u_i, \hat{v}_i)$ and their corresponding points $(u_i, v_i)$
in the data. Given each log-energy feature Log(Energy\_i) of the
noisy speech and its corresponding feature $v_i$ of the clean speech

in the training corpus, we can estimate $\beta$ in Eq. (3) by minimizing the squares error $E(\beta)$ expressed in the following equation:

$$E(\beta) = \sum_{i=1}^{n} |v_i - \left( \log(\text{Energy}_i) \times \left( \frac{(\log(\text{Energy}_i) - \text{Min})}{\text{Max} - \text{Min}} \right)^{\beta} \right)|^2,$$

$$\beta^* = \arg\min_{\beta} E(\beta), \quad (5)$$

where $n$ is the total number of log-energy features and the arg min returns the argument that minimizes. The optimization problem in Eq. (4) can be solved by the gradient descent method for each of the training set with different SNR levels.

### 4. EXPERIMENTAL RESULTS

#### 4.1. Experimental setup

The speech recognition experiments are conducted under various noise conditions using the Aurora-2 database and task [8]. The Aurora-2 database is a subset of the TI-DIGITS, which contains a set of connected digit utterances spoken in English, while the task consists of the recognition of the connected digit utterances interfered with various noise sources at different SNRs, in which Sets A and B are artificially contaminated with eight different types of real world noises (e.g. the subway noise, street noise, etc.) in a wide range of SNRs (-5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB and Clean) and the channel distortion is additionally included in Set C. For the baseline system, the training and recognition tests use the HTK recognition toolkit, which follow the setup originally defined for the ETSI Aurora evaluations.

More specifically, each digit is modeled as a left-to-right continuous density HMM with 16 states and three diagonal Gaussian mixtures per state. Two additional silence models are defined. One had three states with six Gaussian mixtures per state for modeling the silence at the beginning and at the end of each utterance. The other one had one state with 6 Gaussian mixtures for modeling the interword short pause. In the front-end speech analysis, a 39-dimensional feature vector is extracted at each time frame, including 12 Mel-frequency cepstral coefficients (MFCCs), the logarithm of the energy, which is processed by the proposed LESN algorithm described in Section 3 or some other approaches, and the corresponding delta and acceleration coefficients. The frame length is 25 ms and the corresponding frame shift is 10 ms [8].

#### 4.2. Experimental results

In this subsection, we compare the recognition performance of three log-energy-processing approaches including LESN-based methods and ERN. In the experiment with ERN, the linear and nonlinear version of scaling equations were adopted, and when compared with other methods, 17 dB and 12 dB were used respectively as the optimum target energy dynamic ranges that can achieve the best overall recognition accuracy for ERN in our experiments.

In the experiment of quantile-based LESN1, we compare the recognition performance when different numbers of quantiles are applied. Table 1 shows that the number of quantiles $N$ in Eq. (3) can be set to 100 empirically to make better word error rate (WER) though it seems that $N$ does not matter much while it is large enough in LESN1. Figure 2(b) shows the LESN1-based log-energy contours of the clean utterance and its noise-corrupted counterpart with SNR of 15 dB. From this figure, it can be shown that LESN1 preserves the higher log-energy portions, and for the lower log-energy portions, it lowers the effects of additive noise to some extent.

In the experiment of weight-based LESN2, Table 2 shows the $\beta^*$ values according to different SNR levels in the training data. Furthermore, Table 3 shows that, when the $\beta^*$ value derived under the training condition of 20 dB SNR is applied to Eq. (4) for speech recognition, the better WER result can be obtained among all training conditions.

Then, we compare our presented log-energy feature normalization approaches with the ERN-based approaches. The WER results for the baseline MFCC system, as well as ERN (linear), ERN (nonlinear), LESN1 and LESN2, for the entire test data are shown in Table 4, respectively. The results also show that LESN1 and LESN2 attain relative improvements of 35.27% and 31.55%, respectively, when compared with the MFCC baseline. As compared with ERN (linear) and ERN (nonlinear), it can be found that LESN1 provides better performance on all of the test sets, especially on Set C. However, for LESN2, the performance is worse than the nonlinear version of ERN. One of the possible reasons is that the data-fitting method in Eq. (5) may, to some extent, raise the degree of the training-test mismatch, notwithstanding the best parameter $\beta^*$ can be determined with the training data. With the same reason, it can also be explained why the performance of LESN1 is better.

Moreover, LESN can be easily integrated with the well-known cepstral processing approaches since they are performed on different features. Here, we combine LESN1 and LESN2 with the approach of cepstral mean and variance normalization (CMVN), and the associated WER results are shown in Table 5. From the
Table 1. The WER results (%) of LESN1, with respect to different quantile numbers N.

<table>
<thead>
<tr>
<th>N</th>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>25.90</td>
<td>23.29</td>
<td>36.92</td>
<td>27.06</td>
</tr>
<tr>
<td>100</td>
<td><strong>25.65</strong></td>
<td><strong>23.28</strong></td>
<td><strong>36.16</strong></td>
<td><strong>26.80</strong></td>
</tr>
<tr>
<td>500</td>
<td>26.51</td>
<td>24.68</td>
<td>36.23</td>
<td>27.72</td>
</tr>
<tr>
<td>1000</td>
<td>27.07</td>
<td>25.32</td>
<td>36.42</td>
<td>28.24</td>
</tr>
</tbody>
</table>

Table 2. The exponential weight $\beta^*$ in Eq. (5), with respect to different SNR levels.

<table>
<thead>
<tr>
<th>Multi</th>
<th>20 db</th>
<th>15 db</th>
<th>10 db</th>
<th>05 db</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta^*$</td>
<td>0.31</td>
<td>0.23</td>
<td>0.28</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Table 3. Comparison between the WER results (%) in LESN2, with respect to various $\beta^*$ estimated under different training condition.

<table>
<thead>
<tr>
<th>Training Condition</th>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR 05 db with $\beta^* = 0.41$</td>
<td>33.28</td>
<td>28.97</td>
<td>45.16</td>
<td>33.93</td>
</tr>
<tr>
<td>SNR 10 db with $\beta^* = 0.33$</td>
<td>29.98</td>
<td>26.25</td>
<td>42.33</td>
<td>30.96</td>
</tr>
<tr>
<td>SNR 15 db with $\beta^* = 0.28$</td>
<td>28.10</td>
<td>24.82</td>
<td>40.74</td>
<td>29.31</td>
</tr>
<tr>
<td>SNR 20db with $\beta^* = 0.23$</td>
<td><strong>26.87</strong></td>
<td><strong>23.99</strong></td>
<td><strong>39.37</strong></td>
<td><strong>28.22</strong></td>
</tr>
<tr>
<td>Multi-Condition with $\beta^* = 0.31$</td>
<td>29.13</td>
<td>25.64</td>
<td>41.63</td>
<td>30.24</td>
</tr>
</tbody>
</table>

Table 4. Comparison between the WER results (%) of the baseline (MFCC) and various approaches.

<table>
<thead>
<tr>
<th>Method</th>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>41.06</td>
<td>41.52</td>
<td>40.03</td>
<td>41.04</td>
</tr>
<tr>
<td>ERN (linear)</td>
<td>26.92</td>
<td>24.17</td>
<td>40.15</td>
<td>28.47</td>
</tr>
<tr>
<td>ERN (nonlinear)</td>
<td>25.82</td>
<td>23.55</td>
<td>37.48</td>
<td>27.25</td>
</tr>
<tr>
<td>LESN1</td>
<td><strong>25.65</strong></td>
<td><strong>23.28</strong></td>
<td><strong>36.16</strong></td>
<td><strong>26.80</strong></td>
</tr>
</tbody>
</table>

Table 5. Comparison between the WER results (%) of CMVN only and the combinations of LESN and CMVN.

<table>
<thead>
<tr>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMVN</td>
<td>22.73</td>
<td>19.60</td>
<td>27.17</td>
</tr>
<tr>
<td>LERN1+CMVN</td>
<td>19.59</td>
<td>17.02</td>
<td>23.37</td>
</tr>
<tr>
<td>LERN2+CMVN</td>
<td>20.81</td>
<td>17.90</td>
<td>25.01</td>
</tr>
</tbody>
</table>

Table 6. Comparison between the SER, CER and WER results (%) of the baseline (MFCC) and LESN1 on the LVCSR task.

<table>
<thead>
<tr>
<th>Word Graph</th>
<th>Baseline</th>
<th>LESN1 ($N = 500$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SER</td>
<td>22.97</td>
<td>22.10</td>
</tr>
<tr>
<td>CER</td>
<td>30.70</td>
<td>29.72</td>
</tr>
<tr>
<td>WER</td>
<td>39.62</td>
<td>38.80</td>
</tr>
</tbody>
</table>

Table 7. The optimal WER results (%) that the energy based approaches can achieve.

<table>
<thead>
<tr>
<th>Optimum</th>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15.95</td>
<td>14.20</td>
<td>23.42</td>
<td>16.74</td>
</tr>
</tbody>
</table>

4.3. Further evaluations on the LVCSR task

Furthermore, we try to apply LESN1 to the large vocabulary continuous speech recognition (LVCSR) task [9]. The speech corpus consists of about 200 hours of MATBN Mandarin broadcast news, among which 25 hours of gender-balanced speech data is for training and another 1.5 hours of speech data is reserved for testing. Compared with the baseline MFCC system, the results in Table 6 indicate that LESN1 ($N = 500$) achieves considerable improvements on the syllable error rate (SER), character error rate (CER) and WER.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have introduced the scheme of log-energy scaling normalization (LESN) to speech recognition and given two related examples to explain how LESN performs. One of the main benefits of LESN is that it is easy to understand and takes lightweight computation load. Experimental results have also demonstrated its ability to reduce the WER under various noisy conditions.

Despite that, we also attempt to replace the energy-related features of the noisy speech with those of the clean speech to find out the upper bound of the average WER that the log-energy normalization based approaches are expected to achieve. Table 7 shows that there is still much room (about 10% of the absolute WER) for improving the current energy-based robustness approaches.

6. REFERENCES