LOG-ENERGY DYNAMIC RANGE NORMALIZATION FOR ROBUST SPEECH RECOGNITION

Author : Weizhong Zhu
Douglas O'Shaughnessy
Professor: 陳嘉平
Reporter: 吳國豪

Outline

- Introduction
- Energy Dynamic Range Normalization
- Experiment

Introduction

- Accuracy of speech recognition degrades rapidly when speech is distorted by noise. ASR must work well in a wide range of unexpected noisy environments.

- We propose **a log-energy dynamic range normalization (ERN)** method to minimize mismatch between training and testing data.

Energy Dynamic Range Normalization

- The log-energy feature sequence of noisy speech with a 10 dB SNR ratio and that of clean speech are shown in Figure 1.

- Comparing with that of clean speech, characteristics of the log-energy feature sequence of noisy speech are
  (1) Elevated minimum value.
  (2) Valleys are buried by additive noise energy, while perks are not affected as much.

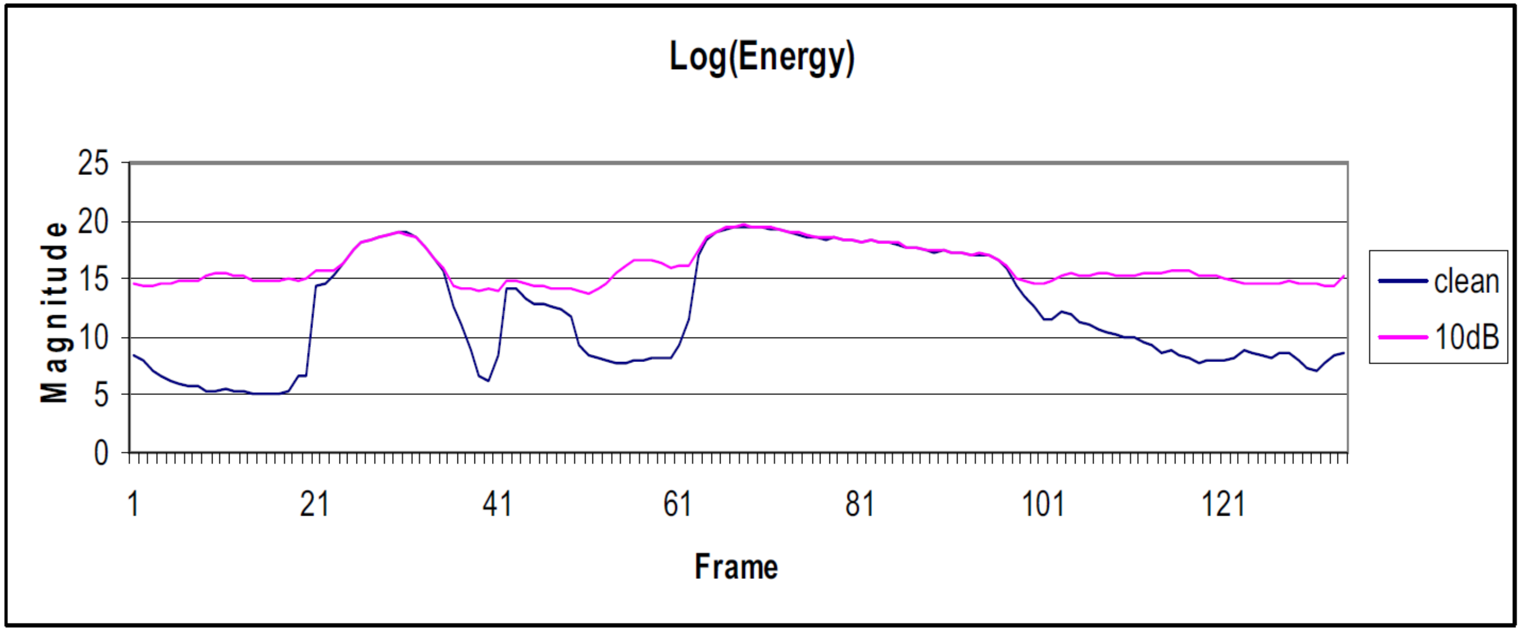

Figure 1: Comparison of log energy feature sequences between clean and noisy speech.

Spanish

	WM	MM	HM
Baseline	86.88%	73.72%	42.23%
Log E replaced	**94.94%**	**89.43%**	**80.99%**
R.I.	61.43%	59.78%	67.09%

Energy Dynamic Range Normalization

- To minimize the mismatch, we suggest an algorithm to scale the log-energy feature sequence of clean speech, in which we lift valleys while we keep peaks unchanged.
- We define **a log-energy dynamic range** of the sequence as follows

$$D.R.(dB) = 10 \times \frac{Max\ (Log(Energy_i)_{i=1...n})}{Min\ (Log(Energy_i)_{i=1...n})} \qquad (1)$$

Energy Dynamic Range Normalization

- As we know, in the presence of noise, $Min\ (Log(Energy_i)_{i=1...n})$ is affected by additive noise, while $Max\ (Log(Energy_i)_{i=1...n})$ is not affected as much. We let

$$Min(Log(Energy_i)_{i=1...n}) = \alpha \times Max(Log(Energy_i)_{i=1...n})$$

- Define **target energy dynamic range** as *X ;* then the above equation becomes

$$X(dB) = \frac{10}{\alpha}$$

Energy Dynamic Range Normalization

- Following are the steps of the proposed log-energy feature dynamic range normalization algorithm:

(1) find $Max = Max(Log(Energy_i)_{i=1...n})$ and

$$Min = Min(Log(Energy_i)_{i=1...n})$$

(2) Calculate target

$$T\_Min = \alpha \times Max(Log(Energy_i)_{i=1...n})$$

Energy Dynamic Range Normalization

(3) If $Min(Log(Energy_i)_{i=1\ldots n}) \; < \; T\_Min$ then (4)

(4) For $i = 1 \ldots n$,

$$Log(Energy_i) = Log(Energy_i) + \frac{T\_Min - Min}{Max - Min} \times (Max - Log(Energy_i)) \qquad (2)$$

Experiment

- Aurora 2

  - (SNRs: -5 dB, 0 dB,5 dB, 10 dB, 15 dB, 20 dB, clean).

  -There are three tests from the Aurora 2 database to evaluate the performance of all considered techniques.

Experiment 1

- In experiment 1, we explore how good the performances are in the sense of relative improvement if we introduce log-energy dynamic range normalization with different target ranges. What is the optimized dynamic range?

- It is shown that as the target log energy dynamic range decreases, performances of Set A and B as well as Overall increase.

Table 1: Relative improvements (%) in different target energy dynamic ranges using linear scaling.

Target energy dynamic range	Set A	Set B	Set C	Overall
30 dB	9.97	10.27	2.57	8.85
25 dB	18.63	19.51	3.69	16.48
20 dB	27.13	30.39	-1.14	23.78
19 dB	27.62	32.57	-5.98	24.12
18 dB	29.13	34.67	-8.96	25.13
17 dB	**29.41**	**36.49**	**-13.23**	**25.32**
16 dB	28.35	37.72	-19.65	24.37
15 dB	24.74	37.02	-14.55	23.53

Experiment 1

- Linear scaling of equation 1 may not be the best solution.
  We modify equation 2 into equation 4.

$$\text{For } \; i = 1 \ldots n,$$

$$Log(Energy_i) = Log(Energy_i) + \frac{T\_Min - Min}{\log(Max) - \log(Min)} \times (\log(Max) - \log(Log(Energy_i))) \qquad (4)$$

Table 2: Relative improvements (%) in different target energy dynamic ranges using non-linear scaling.

Target energy dynamic range	Set A	Set B	Set C	Overall
18 dB	19.29	22.88	-2.29	17.19
17 dB	20.09	24.68	-4.56	17.94
16 dB	22.44	26.88	-3.38	20.03
15 dB	24.24	28.76	-2.50	21.71
14 dB	**34.88**	**41.02**	**-5.55**	**30.83**
13 dB	34.19	37.07	-0.98	29.50
12 dB	32.18	32.99	0.03	27.09

Table 3: Performance comparisons between linear and non-linear normalization methods for average relative improvement (%) at different SNR levels.

Method	20dB	15dB	10dB	5dB	0dB
Linear	8.40	26.42	35.10	26.33	12.29
N.L.	31.75	38.94	40.55	32.59	15.78

Experiment 2

- Here in experiment 2, can the proposed algorithms combine with other techniques get an even better result?

- The results are shown in Table 4, in which CMN refers to cepstral mean normalization, CVN for cepstral variance normalization, ERN(L) and ERN(N) for proposed methods, linear and non-linear respectively .

Table 4: Relative improvement (%) of techniques with respect to a standard MFCC.

Technique	Set A	Set B	Set C	Overall
CMN	12.51	34.05	-3.73	19.30
ERN(L)	29.41	36.49	-13.23	25.32
ERN(N)	34.88	41.02	-5.55	30.83
ERN(L) + CMN	24.46	42.86	7.05	29.67
ERN(N) + CMN	44.81	55.45	25.98	46.33
CVN	44.94	54.43	27.33	46.16
ERN(L) + CVN	44.94	54.43	27.34	46.16
ERN(N) + CVN	53.72	61.27	36.79	54.19