LOG-ENERGY DYNAMIC RANGE NORMALIZATON FOR ROBUST SPEECH RECOGNITION

Weizhong Zhu and Douglas O'Shaughnessy

INRS-EMT, University of Quebec
800 De la Gauchetiere West, Montreal, Quebec, H5A 1K6, Canada
{zhuw, dougo}@inrs-emt.uquebec.ca

ABSTRACT

Cepstral Mean Normalization (CMN) has proved to be a simple noise robust feature processing technique. In its computation, the log-energy feature or C0 is treated in the same way as other cepstral coefficients. Mean normalization is not an effective way to remove effects of additive noise for the log-energy feature. We propose a log-energy dynamic range normalization (ERN) algorithm which normalizes log-energy sequences of an utterance to a target dynamic range. The AURORA 2.0 Database together with HTK speech recognition toolkits are used to evaluate the proposed algorithm. The proposed algorithm improves the recognition result by 30.83% over the reference front-end algorithm in clean-condition training. It is superior to the result of CMN, which is 19.30%. It is also confirmed that this technique can be combined with CMN to achieve a 46.33% performance gain. The proposed algorithm is fairly simple and it only requires a very small extra computation load.

1. INTRODUCTION

Automatic Speech Recognition (ASR) has been in commercial application for decades but still has severe limitations. Accuracy of speech recognition degrades rapidly when speech is distorted by noise. ASR must work well in a wide range of unexpected noisy environments. Methods to overcome the effects of noise must be applied in order to achieve good recognition accuracy in real speech recognition applications where various types of noises may exist. Robust speech recognition is one of the most challenging areas of speech recognition [1].

Methods of robust speech recognition can be classified into two approaches. If we divide a speech recognition system into front-end processing for speech feature extraction and back-end processing for HMM decoding, then methods to compensate for noise can be implemented at the front-end or the back-end or both. The front-end processing method is to suppress the noise and get more robust parameters, while back-end processing is to compensate for noise and adapt the parameters inside the HMM system.

In this paper, we focus on the first approach. Mel Frequency Cepstral Coefficients (MFCCs) are widely accepted by the speech recognition community right now [2][3]. In order to get more noise robust features, there are numerous efforts (1) to add pre-processing, like noise reduction [4][5] or speech enhancement, (2) to incorporate algorithms in an MFCC calculation framework, like frequency masking [6], SNR-Normalization [7], and (3) to add feature post-processing techniques [8][9].

CMN is known to be a simple noise robust post-feature processing technique. It its calculation, the log-energy feature (or C0) is treated in the same way as other cepstral coefficients. Comparing with cepstral coefficients, the log-energy feature has quite different characteristics. We try to find a more effective way to remove the effects of additive noise for the log-energy feature. We propose a log-energy dynamic range normalization (ERN) method to minimize mismatch between training and testing data. The dynamic range of log-energy feature sequences of an utterance is normalized to a target dynamic range.

The proposed method is evaluated using the AURORA 2.0 Database [10]. The framework of the AURORA 2.0 Database is available to compare different algorithms and systems on a common basis. In the Aurora digital string recognition task, the evaluation focuses on robustness against additive noise and distortion by an unknown transmission channel. Although it is a small vocabulary, it is suitable to evaluate different feature extraction methods. It keeps the recognition engine fixed to a pre-defined HTK reference setup, and it also includes a program to compute the ETSI standard feature vector of the MFCCs and log energy as a reference.

The paper is organized as follows. We introduce the energy dynamic range normalization concept and describe the proposed algorithm as a post-feature processing after a standard MFCC calculation in section 2. In section 3, experimental results and comparisons with cepstral mean and variance normalizations are given. The conclusions are stated in section 4.

0-7803-8874-7/05/$20.00 ©2005 IEEE

2. 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.

The larger difference on valleys leads to a mismatch between the clean and noisy speech. Obviously, mean normalization is not an optimized solution. 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

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

in which $Max(Log(Energy_i)_{i=1..n})$ is the maximum value of the log-energy feature sequence, $Min(Log(Energy_i)_{i=1..n})$ is the minimum value. 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 $Min(Log(Energy_i)_{i=1..n}) = \alpha \times Max(Log(Energy_i)_{i=1..n})$ and define target energy dynamic range as X; then the above equation becomes $X(dB) = \frac{10}{\alpha}$. If we set the target log-energy dynamic range to $X = 20(dB)$, then $\alpha = 0.5$. In this way, we can use α to set the target minimum value based on a given target dynamic range. We scale $Min(Log(Energy_i)_{i=1..n})$ to the target minimum while we keep $Max(Ln(Energy_i)_{i=1..n})$ unchanged.

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

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

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

(4) For i=1..n,

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

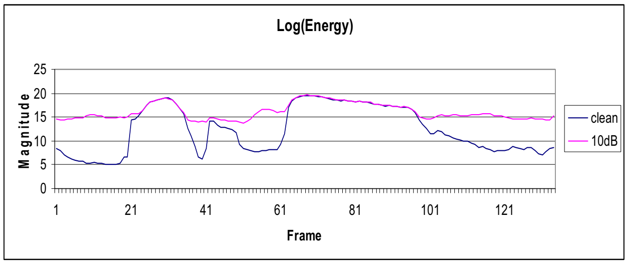

Figure 1: Comparison of log energy feature sequences between clean and noisy speech.

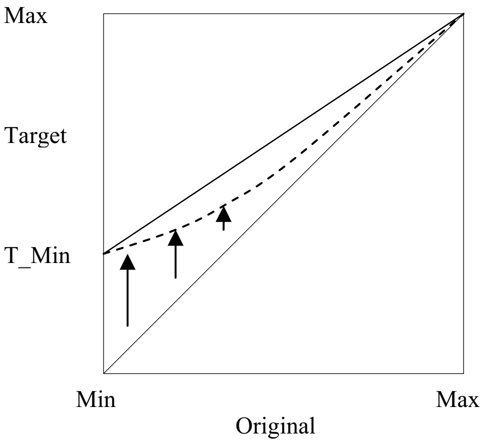

Figure 2: Schematic representation of scaling effect of log-energy dynamic range normalization algorithm.

Figure 2 shows a schematic representation of the scaling effect of the proposed algorithm. If the dynamic range of a log-energy sequence is greater than the target dynamic range, those log-energy features close to the minimum value are scaled to a target minimum; the scaling effect is decreased as its own value goes up and the maximum of the sequence is unchanged.

3. EXPERIMENTAL STUDY

3.1. System setup

The proposed method was evaluated on the Aurora 2 database. All recognition tests were conducted using the HTK recognition toolkit with the setting defined for evaluation. The task is speaker-independent recognition of digit sequences. The distortions are artificially added to the clean TIdigits database [11]. The original 20 kHz data have been downsampled to 8 kHz. Selections of 8 different real-world noises have been added to the speech over a range of signal-to-noise ratios (SNRs: -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, clean – no noise added).

In the baseline system, speech models are eleven whole word HMMs fixed to 16 states with 3 diagonal Gaussian mixtures per state. Two silence models are defined. One has 3 states, modeling pauses before and after utterances, and the other has one state with 6 Gaussian mixtures modeling the interword silence. Delta and acceleration coefficients are computed over five- and three-frame windows. Thus a vector size of 39 parameters is used for HMM modeling. HMM models are trained by a pre-defined clean training script. We confirmed that, using the original front-end program, we can get exactly the reference performance scores. Post-feature processing with modification of log-energy features is added to evaluate our proposed algorithm.

There are three tests from the Aurora 2 database to evaluate the performance of all considered techniques. 8 different real noises are divided into two groups for testing. Data in Test A are added to by noises of Subway, Babble, Car and Exhibition. Data in Test B are added to noises of Restaurant, Street, Airport and Station. In Test C, besides the additive noise, channel distortion is also included.

3.2. Recognition results

The results in this section are defined in terms of relative improvement (*R.I.*). According to the Aurora 2 protocol, it is calculated as

$$R.I.(\%) = \frac{NewScore - Baseline}{100 - Baseline} \times 100\% \qquad (3)$$

where *NewScore*, *Baseline* are recognition accuracies for each test using proposed and reference algorithms, respectively. The mean recognition accuracy for each test set is obtained by taking the average of the recognition accuracies measured in 20, 15, 10, 5 and 0 dB SNR. The overall accuracy is calculated as 0.4*Set A + 0.4*Set B + 0.2*Set C.

3.2.1. 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?

Results of table 1 show relative improvements with the different target log-energy 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. It achieves a 25.32% highest overall relative improvement when the target range is set to be 17 dB. We note that there is a negative effect as the target dynamic range goes down. It seems that it is difficult to use energy dynamic range normalization to deal with channel distortion.

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

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

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

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

The effect of equation 4 is shown as the dotted line in figure 2. Instead of a linear scaling effect, the dotted line shows a non-linear effect. The results of relative improvement in different target dynamic ranges using this non-linear normalization method are shown in Table 2. It achieves a 30.83% highest overall relative improvement when the target range is set to 14 dB.

Performance comparisons between linear and non-linear normalization methods for average relative improvement at different SNR levels are shown in table 3. The non-linear method has a better performance in all different SNR ratios.

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

3.2.2. Experiment 2

Here in experiment 2, we answer the questions: (1) what are the results of techniques like cepstral mean and variance normalizations? (2) Can the proposed algorithms combine with these 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 . If we use CMN, we can only get an overall 19.39% relative improvement, which is lower than the proposed methods. CMN has to process with all 13 parameters, while the proposed ERN method processes log-energy only. Both linear and non-linear algorithms can be combined with CMN to get an even better result. The result of the proposed non-linear algorithm combined with CMN is better than that of CVN and the proposed algorithm can be combined with CVN to achieve a performance gain up to 54.19%.

4. CONCLUSIONS

A log-energy dynamic range normalization technique is introduced to improve ASR performance in noisy conditions. It is evaluated on the Aurora 2.0 digit recognition task. The proposed log-energy dynamic range normalization algorithm can have overall about a 30.83% relative performance improvement when systems were trained on a clean speech training set. It is also confirmed that the proposed algorithm can be combined with the cepstral mean or variance normalization techniques to achieve an even better result.

Like cepstral mean normalization, the proposed method does not require any prior knowledge of noise and level. It is effective to improve the performance of speech recognition for eight different noise conditions at various SNR levels. The proposed algorithm is fairly simple and can be combined with CMN or CVN, and it only needs a very small extra computation load.

Reducing mismatch in log-energy leads to a large recognition improvement. It may result from whole-word HMM models and low complexity of this task. We will extend our experiments on sub-word HMM models and more complex tasks such as large vocabulary continuous speech recognition.

5. ACKNOWLEDGMENTS

This work has been partly supported by the Canadian Government under the NSERC Strategic Partnerships Program at SpeechWorks (now Scansoft) and INRS-EMT (Telecommunications).

6. REFERENCES

[1] D. O'Shaughnessy, *Speech Communications: Human and Machine*, IEEE PRESS, New York, 2000.

[2] L. Rabiner and B-H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, New Jersey, 1993.

[3] J.R. Deller, J.H.L. Hansen and J.G. Proakis, *Discrete-time Processing of Speech Signal*, IEEE PRESS, New York, 2000.

[4] S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoust., Speech Signal Proc.*, Vol.27, pp.113-120, Apr. 1979.

[5] W. Zhu and D. O'Shaughnessy, "Using noise reduction and spectral emphasis techniques to improve ASR performance in noisy conditions," *ASRU 2003,* Nov. 2003, US Virgin Islands.

[6] W. Zhu and D. O'Shaughnessy, "Incorporating Frequency Masking Filtering in a Standard MFCC Feature Extraction Algorithm," *ICSP 2004,* Vol.1, pp. 617-620, Aug. 31-Sept. 4, Beijing.

[7] T. Claes and D. Van Compernolle, "SNR-Normalization for Robust Speech Recognition," *ICASSP'1996*, Vol.1, pp.331-334, Atlanta.

[8] A. Rosenberg, C.-H. Lee and F. Soong, "Cepstral Channel Normalization Techniques for HMM-based speaker verification," *ICSLP'1994*, Vol.4, pp.1835-1838, Yokohama.

[9] C-P. Chen, J. Bilmes and K. Kirchhoff, "Low-Resource Noise-Robust Feature Post-Processing on AURORA 2.0," *ICSLP'2002*, Vol.4, pp.2445-2448, Denver.

[10] H.G. Hirsch and D. Pearce, "The AURORA Experiment Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," ISCA ITRW ASR2000 "Automatic Speech Recognition Challenges for the Next Millennium" Paris, France, Sept. 18-20, 2000.

[11] R.G. Leonard, "A Database for Speaker Independent Digit Recognition," *ICASSP'1984*, Vol.3, pp.42.11.1-4, San Diego.