Sub-band based Log-energy and Its Dynamic Range Stretching for Robust In-car Speech Recognition

Weifeng Li[1], *Hervé Bourlard* [2]

[1]Department of Electronic Engineering / Graduate School at Shenzhen, Tsinghua University, China
[2]Idiap Research Institute, Martigny, Switzerland

Li.Weifeng@sz.tsinghua.edu.cn

Abstract

Log energy and its delta parameters, typically derived from full-band spectrum, are commonly used in automatic speech recognition (ASR) systems. In this paper, we address the problem of estimating log energy in the presence of background noise (usually resulting in a reduction in dynamic ranges of spectral energies). We theoretically show that the background noise affects the trajectories of the "conventional" log energy and its delta parameters, resulting in very poor estimation of the actual log energy and its delta parameters, which no longer describe the speech signal. We thus propose to estimate log energy from the sub-band spectrum, followed by a dynamic range stretching. Based on speech recognition experiments conducted on CENSREC-2 in-car database, the proposed log energy (and its corresponding delta parameters) is shown to perform very well, resulting in an average relative improvement of 27.2% compared with the baseline front-ends. Moreover, it is also shown that further improvement can be achieved by incorporating those new MFCCs obtained through non-linear spectral contrast stretching.

1. Introduction

Standard mel-frequency cepstral coefficients (MFCCs) [1] are extracted from log scaled mel-filterbank (MFB) outputs. However, in the presence of background noise, the dynamic ranges of spectral energies are generally reduced. Figure 1 (second row) shows the first-channel log MFB trajectory (or contour) of speech captured by a close-talking (headset) microphone and a distant microphone (attached to the ceiling above the driver's seat [2]) in a car-driving environment. Compared to close-talking speech, the floor level of the log MFB trajectory of distant speech is elevated and the valleys are buried by noise energy. While the spectral changes of close-talking speech over time are rather apparent, they turn to be obscure for distant speech due to the noise effect. Besides MFCCs, the short-time log energy and its temporal derivatives are often adopted as standard features as well. According to the discriminative analysis of the features used for ASR [3], the frame log energy and its temporal derivatives are the most critical parameters in terms of recognition accuracy. It has been shown that the ASR performance in clean condition improves when short-time log energy and its temporal derivatives are used [4]. However, in low SNR conditions, the trajectory of the short-time log energy, which is derived from full-band spectrum, can be so distorted that it fails in describing the speech signal dynamics, as demonstrated in Fig. 1 (lower part). Therefore, in the presence of background noise, the conventional MFCCs and log energy usually introduce undesirable mismatches between relatively clean speech

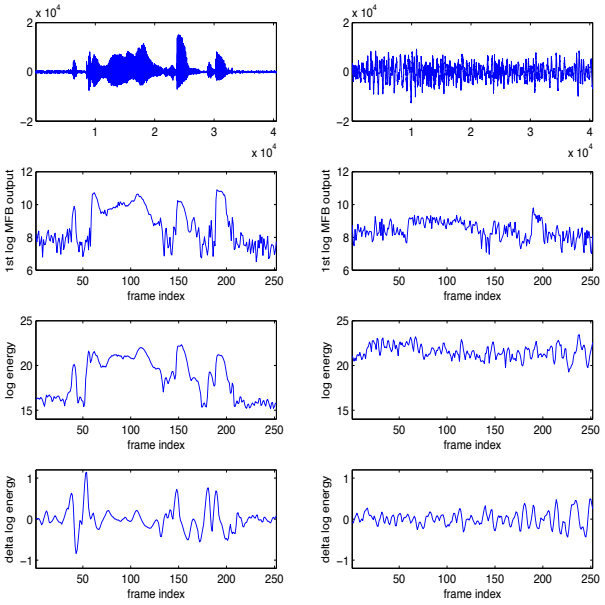

Figure 1: Effect of car noise on log mel-filter bank (MFB) and log energy trajectories. The left subfigures (up to down): waveform, the first log MFB output, log energy, and the delta log energy of close-talking speech; The right subfigures (up to down): waveform, the first log MFB output, log energy, and the delta log energy of distant speech; The speech is "12439" in Japanese.

(used for training) and noisy speech (used for testing), resulting in serious ASR performance degradation.

From the viewpoint of speech perception, it is generally acknowledged that to make the softest speech sounds audible and the loudest still comfortable, a certain degree of speech dynamic range is necessary [5]. Under adverse conditions, background noise generally leads to a reduction of dynamic ranges. Even for normal hearing listeners, serious reductions in dynamic ranges lead to unreliable segmentation, making the task of parsing the speech signal more difficult [6]. It has been suggested that under adverse conditions the auditory system make some adaptations serving to emphasize newly arriving components of the signal and enhance the regions of the signal undergoing spectro-temporal changes [7]. Motivated by the adaptation capabilities of the auditory system, we proposed in [8] a new MFCC front-

ends based on the spectral contrast stretching of the log mel-filterbank (MFB) outputs.

In the present paper, we address the problem of the conventional log energy when the background noise presents. More specifically, we theoretically analyze how the noise affects the trajectories of the conventional log energy and its delta parameters, which makes them no longer describe the variations of the speech, or even hurt the speech recognition performance in low SNR conditions. We then propose to estimate the log energy from sub-band spectrum for better representing the variations of the speech in order to enhance its discriminative power in the speech recognition. To further boost the dynamic variations of speech over time, a dynamic range stretching is followed. Our experiments, conducted on realistic in-car data under different training and test conditions, demonstrate that with the subsequent post-processing, the proposed log energy and its temporal derivatives are capable of significantly reducing the mismatches between the training and test conditions.

2. Log energy estimation mismatch between clean and noisy conditions

We assume that the noisy signal is given by

$$x(i) = s(i) + n(i), \tag{1}$$

where $s(i)$ is the clean speech signal which is assumed to be independent of the additive noise $n(i)$. By using short-time Discrete Fourier Transform (DFT), in the time-frequency domain we have their power spectrogram:

$$|X(k,l)|^2 = |S(k,l)|^2 + |N(k,l)|^2, \tag{2}$$

where we make the assumption of statistical independence between the clean speech and noise, and k and l are the frequency bin and frame indexes. The conventional log energy of the noisy speech is derived from the full-band spectrum (all the K frequency bins):

$$\log E_X(l) = \log\left(E_S(l) + E_N(l)\right), \tag{3}$$

where $E_X(l) = \sum_{k=1}^{K} |X(k,l)|^2$, $E_S(l) = \sum_{k=1}^{K} |S(k,l)|^2$, and $E_N(l) = \sum_{k=1}^{K} |N(k,l)|^2$. Compared to clean speech $\log E_S(l)$, the log energy values of noisy speech are elevated, as demonstrated in Fig. 2-1. More seriously, as analyzed next, in the low SNR conditions the derived log energies fail to describe the variations of speech, making them even hurt speech recognition performance.

The variations of speech can be measured by the dynamic changes of log energy C, which are computed by the difference between the log energies of noisy speech at frame l and its subsequent one (e.g. at frame $l+p, p>0$):

$$\begin{aligned} C &= \log E_X(l+p) - \log E_X(l) \\ &= \log[E_S(l+p) + E_N(l+p)] \\ &\quad - \log[E_S(l) + E_N(l)] \\ &= \log \frac{E_S(l+p) + E_N(l+p)}{E_S(l) + E_N(l)} \qquad (4) \\ &\approx \log\left(1 + \frac{E_S(l+p) - E_S(l)}{E_S(l) + E_N(l)}\right), \qquad (5) \end{aligned}$$

where the approximation is based on the (reasonable) assumption that the energy of the noise does not vary too much across time. From Eq. (5), we can see that in the presence of noise

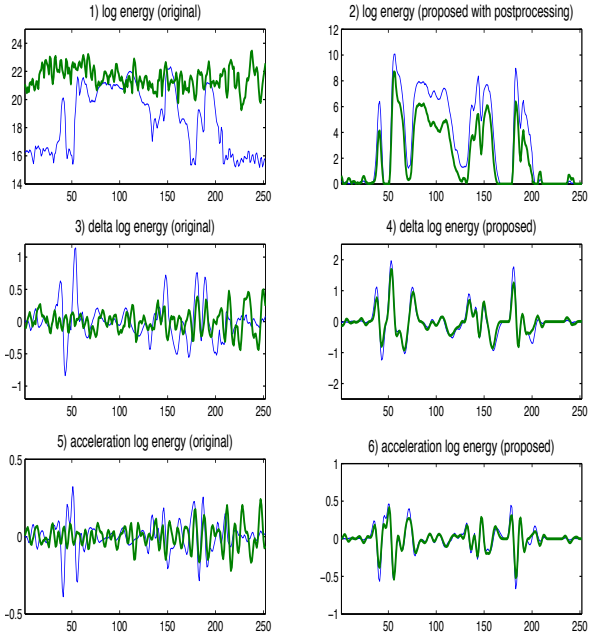

Figure 2: Log energy, delta log energy, acceleration log energy trajectories of the close-talking speech and the distant speech (The speech is the same as Fig. 1). Inside each sub-figure, thin line is for the close-talking speech and bold line is for the distant speech.

the dynamic changes of log energy decreases, and as the noise energy increases the dynamic change of log energy becomes smaller. If the noise is dominant (i.e., $E_N \gg E_S$), Eq. (4) reduces to $\log(E_N(l+p)/E_N(l))$. In this case, the speech signals are buried and the dynamic changes over time reveals the dynamic changes of the noise rather than of the speech. Figures 2-1 and 2-3 illustrate this phenomenon, especially for the first and last 50 frames.

When $E_S(l) = 0$ (i.e., non-speech segments) and $E_S(l+p) > 0$ (i.e., speech segments), from Eq. (5) the dynamic changes of log energy from non-speech segments to speech segments reduce to

$$\log\left(1 + \frac{E_S(l+p)}{E_N(l)}\right) \approx \log\left(1 + SNR(l+p)\right), \tag{6}$$

where $SNR(l+p) = E_S(l+p)/E_N(l+p)$ indicates the signal-to-noise ratio (SNR) at frame $l+p$ (Here we assume $E_N(l) \approx E_N(l+p)$ as already discussed earlier.). Equation (6) reveals that in the speech transition regions the presence of noise is reducing the dynamic changes as a function of the signal-to-noise ratio (SNR)[1]. This phenomenon can be reflected by the trajectories of the delta and acceleration log energy features (i.e., $\triangle$ and $\triangle\triangle$ features), as shown in in Fig. 2-3 and Fig. 2-5. In summary, if the conventional full-band spectrum based static log energy and its dynamic features are fed into an ASR system, they will produce a mismatch between relatively clean speech and noisy speech, which will inevitably degrade the ASR performance.

[1] In the case of the transition from speech to non-speech segments Eq. (4) reduces to $-\log\left(1 + SNR(l)\right)$.

3. The proposed log energy and its Dynamic range enhancement

3.1. The proposed log-energy

To alleviate the problems described above and make log-energy better suited to reflect the variations of the speech over time, the proposed log-energy is calculated from the log Mel-filter-bank (MFB) outputs with the following two considerations: (1) Log MFB outputs are sub-band based, and can capture the dynamic variations of speech signals over time inside a particular sub-band; (2) The log MFB outputs with larger change ranges across the time can better reflect the dynamic variations of speech signals than those with smaller ones, where speech signals are probably more seriously contaminated by the noise. Therefore, we propose to calculate the log-energy from *the log MFB outputs with the largest dynamic ranges*. The dynamic range of their log MFB values for the m-th filter-bank is defined by

$$R(m) = X_{\max}^{(L)}(m) - X_{N}^{(L)}(m) \quad m = 1, 2, \cdots, M \tag{7}$$

where $X_{\max}^{(L)}(m)$ and $X_{N}^{(L)}(m)$ are the maximum values of the m-th log MFB outputs along the frames of the utterance and the estimated noise log MFB value, respectively. Then we sort $R(m)$ over all the M filter-banks and identify the J-th largest value R_J. Finally the proposed log energy is obtained by

$$E(l) = \frac{1}{J} \sum_{m=1}^{M} \Big(U(R(m) - R_J) \cdot X^{L}(m, l) \Big), \tag{8}$$

where $U(\cdot)$ is the unit step function and $X^L(m, l)$ is the m-th log MFB values at frame l. While the conventional log-energy is derived from full-band spectrum, the proposed log-energy is obtained from J sub-bands which can better reflect the speech variations over time[2].

3.2. Dynamic range stretching (DRS)

In order to further boost the dynamic variations of speech signals over time, a dynamic range stretching (DRS) is performed by utilizing the non-linear spectral contrast stretching technique in [8]. More specifically, the dynamic range stretching (DRS) of log energy is implemented by

$$\hat{E}(l) = \begin{cases} \frac{E(l) - E_n}{E_{\max} - E_n} \cdot E(l), & \text{if } E(l) - E_n \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

where $E_{\max}$ and E_n denote the maximum value of the log energies along the frames in an utterance and the estimated log-energy of the noise, respectively.

Through the further dynamic range enhancement, the dynamic range is stretched to $[0, E_{\max}]$ non-linearly (emphasizing the speech variations with larger $E(l)$ values more than the ones with smaller values), and the level of the spectral variations of the speech is enhanced accordingly. Figures 2-2, 2-4, and 2-6 show resulting static, delta, and acceleration log-energies, respectively. Comparing with the original features in 2-1, 2-3 and 2-5, it is clear that the mismatches between the close-talking speech and the distant speech have been reduced significantly.

[2]The zero-order MFCC can be viewed as an average of all the M log MFB outputs, while the proposed log-energy is the mean value of the J log MFB outputs only with prominent dynamic changes.

Table 1: Training and test configurations for each of the four evaluation conditions. HF: hands-free microphone; CT: close-talking microphone.

conditions	Cond.1		Cond.2		Cond.3		Cond.4	
train/test	train	test	train	test	train	test	train	test
microphone	HF	HF	HF	HF	CT	HF	CT	HF
idling	◯	◯	◯		◯		◯	
low speed	◯	◯		◯	◯	◯		◯
high speed	◯	◯		◯	◯	◯		◯

4. Speech Recognition Experiments

4.1. Experimental setup

Evaluation of the proposed algorithm has been performed on the CENSREC-2 in-car speech database [2]. This database comprises a task for continuous digit recognition in real car driving environments. In-car speech data is collected in a specially equipped vehicle under 11 environmental conditions. The speech recorded by a distant (hands-free) microphone (attached to the ceiling above the driver's seat) is used for evaluation. There are four evaluation environments (conditions), as shown in Table 1, and speech recognition performance depends on whether the recording environments and the microphones used between training and testing data are matched or not.

In the baseline system, a 24-channel MFB analysis is applied, and the logarithmic outputs of the filterbanks are computed. The estimated log MFB outputs are transformed into 12 mel-frequency cepstral coefficients, and then the delta and the acceleration coefficients are calculated. Finally, an acoustic vector size of 39 parameters is used for HMM modeling [2]. In our experiments, $J = 10$ is used and the first 15 frames are utilized for estimating the noise.

4.2. Speech recognition results

The experiments can be divided into three sub-parts, as shown in Table 2. The upper part consists of the experiments using the original MFCCs with different log energies:

- Eorg (39 dimensions): the original MFCC features and log-energy, and their delta, and delta-delta features (MFCCs+E+△+△△);
- NoE (36 dimensions): the original MFCC features without log-energy parameter ((MFCCs+△+△△);
- proposeE (39 dimensions): the original MFCC features and the proposed log-energy;
- proposeE+DRS (39 dimensions): the original MFCC features and the proposed log-energy with a subsequent dynamic range stretching (DRS);

The middle part of the table denotes the recognition performance using the convention techniques for robust speech recognition. More specifically, "LSA" uses the MFCCs extracted from the speech enhanced using the minimum mean square error (MMSE) on log-spectral amplitude [9]; "MVN" and "HEQ"denote the cepstral post-processing methods based on Mean and Variance Normalization (MVN) [10] and Histogram Equalization [11], respectively. On the other hand, as shown in Figure 1, the original log MFB outputs of the noisy speech become vague for speech segments. In [8] we have proposed an new MFCC front-ends based on the nonlinear spectral stretching of log MFB outputs. The lower part of this table corresponds to the speech recognition experiments by incorporating

the new MFCC features obtained using the method in [8] with our proposed log energy.

Table 2 summarizes the recognition results obtained form the different methods. From this table, we could draw the following observations:

- The recognition accuracies of "MFCC+Eorg" depend on the evaluation environments. When the recording environments and the microphones used between training and testing data are not matched, the recognition accuracy can degrade into 43.85% (Condition 4).
- When the original log energy and its $\triangle$ and $\triangle\triangle$ are not used, the performance increases for the last three unmatched conditions. This illustrates again that when the training and testing conditions are not matched the conventional log-energy and its $\triangle$ and $\triangle\triangle$ become harmful, and should be discarded.
- The use of the proposed log energy is helpful for the speech recognition, which reveals its discriminative ability. The subsequent dynamic range stretching (DRS) further improves the recognition accuracies, achieving an average relative improvement of 27.2% compared with the original log energy.
- The speech enhancement method "LSA" is effective for all the four evaluation conditions for its noise reduction effects. However, the algorithm introduces much computation cost. Using the normalization methods in cepstral domain is helpful for improving the in-car speech recognition performance, and "MVN" performs better than HEQ. It is noticeable that with only log energy parameter changed our method can be able to achieve comparable average recognition performance to these normalization methods, in which all the MFCCs are normalized. This clearly demonstrate the effectiveness of the proposed log energy parameter and its dynamic range stretching (DRS).
- From the lower part of the table it is found that the new MFCCs derived by the nonlinear spectral stretching of log MFB outputs in [8] is effective for reducing the mismatch between the MFCCs of close-talking speech and distant speech. With the new MFCCs the original log energy is still harmful for speech recognition, while the proposed log energy parameter and its dynamic range stretching (DRS) can improve the recognition performance, resulting in an average relative improvement of 53.6%.

Table 2: Recognition accuracies (in percentages) for different methods.

methods	Cond.1	Cond.2	Cond.3	Cond.4	Ave.
Eorg	81.23	66.85	57.94	43.85	62.46
NoE	80.79	68.04	60.47	44.12	63.35
proposedE	83.06	68.12	62.68	50.22	66.04
proposedE+DRS	83.97	76.65	71.37	58.67	**72.66**
LSA	81.71	80.08	67.77	60.24	72.45
MVN	83.95	80.87	70.54	64.11	**74.86**
HEQ	83.52	80.54	66.97	58.96	72.50
Eorg	84.16	83.80	76.85	72.51	79.33
NoE	83.32	83.74	80.83	77.43	81.33
proposedE+DRS	85.21	84.40	81.93	78.75	**82.57**

5. Conclusions

The log energy and its delta parameters are critical features to good performance of ASR systems. In the presence of background noise, however, those parameters may introduce serious distortions, reducing their discriminative potential, or even seriously hurting performance, especially for low SNR conditions. In this paper, we have theoretically analyzed the impact of background noise on the trajectories of the conventional log-energy and its delta parameters. Based on this, we have proposed a robust log-energy parameter estimation algorithm, which significantly reduces the mismatch between clean speech and noisy speech. The effectiveness of the proposed log-energy and its corresponding delta parameters has been demonstrated on the CENSREC-2 continuous digit recognition task in real in-car environments, resulting in a relative improvement of more than 50%.

6. References

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357 – 366, 1980.

[2] S. Nakamura, M. Fujimoto, and K. Takeda, "Censrec2: Corpus and evaluation environments for in car continuous digit speech recognition," in *Interspeech'06*, 2006, pp. 2330–2333.

[3] E. L. Bocchieri and J. G. Wilpon, "Discriminative analysis for feature reduction in automatic speech recognition," in *ICASSP'92*, 1992, pp. 501–504.

[4] S. Ikbal, H. Hermansky, and H. Bourlard, "Nonlinear spectral transformations for robust speech recognition," in *the IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, 2003, pp. 393–398.

[5] D. B. Pisoni and R. E. Remez, *The Handbook of Speech Perception*, Blackwell Publishing, 2004.

[6] S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. Fay, *Speech Processing in the Auditory System*, Springer-Verlag, New York, 2004.

[7] A. Q. Summerfield, A. Sidwell, and T. Nelson, "Auditory enhancement of changes in spectral amplitude," *Journal of the Acoustical Society of America*, vol. 81, no. 3, pp. 700–708, 1987.

[8] W. Li and H. Bourlard, "Non-linear spectral contrast stretching for in-car speech recognition," in *Interspeech'07*, 2007.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 443–445, 1985.

[10] P. Jain and H. Hermansky, "Improved mean and variance normalization for robust speech recognition," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[11] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust speech recognition," in *Proceedings Eurospeech*, 2001, pp. 1135–1138.