ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

ENERGY CONTOUR ENHANCEMENT FOR NOISY SPEECH RECOGNITION

Tai-Hwei Hwang and Sen-Chia Chang

E000/Computer & Communication Labs, Industrial Technology Research Institute,
Chutong, Hsinchu, 310
Email: hthwei@itri.org.tw

ABSTRACT

The environmental noise, known as an additive noise, not only corrupts the spectra of speech signal but also blurs the shape of energy contour. The corruption of energy contour can distort the energy derived feature and degrade the performance of pattern classification of noisy speech. To reduce the distortion of the energy feature, the energy bias in the energy contour has to be removed before the feature extraction. For the purpose, we propose two methods to estimate the noise energy; one is obtained from the speech inactive period, and one is from the noisy speech itself. The methods are evaluated by the connected digit recognition of TIDigits, in which the test speech is corrupted with the white noise, babble, factory noise, and in-car noises. As shown in the experiments, the energy enhancement can provide an additional improvement when it is jointly applied with a spectral subtraction.

1. INTRODUCTION

The robustness of automatic speech recognition (ASR) in the noisy environment is a critical issue to many practical applications. For example, the performance of voice control in a cruising car may become unacceptable because of the high noise level. Many techniques have been proposed to improve the robustness of ASR, such as enhancing the test speech in the spectral domain or equalizing the feature distribution in the cepstrum domain. The spectrum enhancement is proposed to remove the noise spectral bias from the input signal. Therefore, the noise spectrum estimation and the formulation of spectral subtraction are two critical points in the technique [1]. The cepstral equalization is a method to match the cepstral distributions of test speech and of training data. In the category of cepstral equalization, cepstral mean normalization (CMN) could be the most popular one for its simplicity and effectively. Besides of the CMN, a cepstral distribution normalization with the 2nd order statistics of the cepstral feature was proposed as well [2]. Recently, a histogram based equalization method was developed to map the cepstral coefficients of test speech into the same distribution domain of the training data [3].

Besides of the spectrum related features, the frame energy derived features are often adopted as a part of a feature vector as well [4]. In fact, according to the discriminative analysis of speech feature in [5], the energy derived features such as the delta and the delta-delta log energies are the most critical parameters in terms of recognition accuracy. However, the energy features are also distorted by the noise and degrade their ability for pattern classification. It has been shown in [6] that the recognition rate of in-car speech can be improved when the energy contour is extracted from the less corrupted frequency band.

In this paper, we investigate the distortion of the time derivatives of log energy under the noise effect. It can be shown that the magnitude of time-dynamic log energy feature becomes small under the corruption of the additive noise. One method to reduce the distortion is to perform a noise energy removal before the energy feature extraction. For the purpose, we propose two methods to compute the noise energy; one is obtained from the noise energy in the speech inactive period and the other is obtained within the noisy signal. To evaluate the proposed schemes, a connected digit recognition using the TIDigits database was conducted. As shown in the experimental results, the proposed schemes can improve the recognition accuracy of noisy speech significantly, even thought it was jointly applied with a spectral enhancement.

This paper is organized as follows. In section 2, the distortion of the energy feature under the noise effect is introduced. In section 3, two methods for energy contour enhancement are proposed. In section 4, the improvement on the speech recognition is demonstrated by the experimental results. Finally, a conclusion is given in section 5.

2.DISTORTION OF SPEECH ENERGY FEATURE UNDER NOISE EFFECT

The delta parameters of the log converted energy, which are the approximations of the time derivatives, can be defined by [7]

$$\left.\frac{dE_x^l(\tau)}{d\tau}\right|_{\tau=t} \cong \Delta E_x^l[t] = \frac{1}{K}\sum_{i=-L}^{L} i \log E_x[t+i] \,, \tag{1}$$

$$\left.\frac{d^2E_x^l(\tau)}{d\tau^2}\right|_{\tau=t} \cong \Delta^2 E_x^l[t] = \Delta E_x^l[t+1] - \Delta E_x^l[t-1] \,, \tag{2}$$

where $E_x[t]$ is the segmental energy of received signal x at frame t, superscript l means a log conversion, L is the extent range from current frame and constant $K = \sum_{i=-L}^{L} i^2$.

When the speech signal x is corrupted with the additive noise w, based on the assumption of statistical independence, the frame energy of the noisy signal y can be expressed by

$$E_y(\tau) \cong E_x(\tau) + E_w(\tau) \,. \tag{3}$$

If the variation of $E_w(\tau)$ is relative slowly to that of $E_x(\tau)$, we may assume $E_w(\tau)$ to be a time-invariant constant e_w, hence

$$E_y(\tau) \approx E_x(\tau) + e_w \,. \tag{4}$$

time derivative of log converted energy can be expressed by

$$\frac{dE_x^l(\tau)}{d\tau} = \frac{1}{E_x(\tau)} \frac{dE_x(\tau)}{d\tau}. \quad (5)$$

Therefore, the first order time derivative of the noisy speech can be expressed by

$$\frac{dE_y^l(\tau)}{d\tau} \cong \frac{1}{E_x(\tau) + e_w} \frac{dE_x(\tau)}{d\tau}. \quad (6)$$

Since $E_x(\tau) > 0$ and $e_w > 0$, we have

$$\left| \frac{dE_y^l(\tau)}{d\tau} \right| < \left| \frac{dE_x^l(\tau)}{d\tau} \right|. \quad (7)$$

The above shows that the magnitude of the time derivative of log energy will become smaller under the noise effect. This effect produces a mismatch between the test speech and the training data so that the classification accuracy of noisy speech is degraded.

3. ENERGY CONTOUR ENHANCEMENT

To reduce the distortion in the energy feature, an intuitive method is to remove the noise energy from the energy contour before the feature extraction. In this study, we propose two methods to perform the noise energy removal as follows.

3.1. Estimate noise energy from speech inactive period

Assume that the noise energy does not change seriously during the period of the received signal. Thus, the noise energy within noisy speech can be properly estimated by the noise energy obtained from the speech inactive period. To do this, a reliable speech/none speech detector is required. Similar to the spectral subtraction method [1], we apply a non-linear mapping function for the noise energy removal to avoid negative energy by using

$$\hat{E}_x[t] = \begin{cases} E_y[t] - \alpha\, \hat{e}_w, & if\ E_y[t] > \frac{\alpha}{1-\beta} \hat{e}_w \\ \beta\, E_y[t], & otherwise \end{cases} \quad (8)$$

where $\hat{e}_w$ is the estimated noise energy, α and β are predefined factors.

3.2.Estimate noise energy with dynamic range

A dynamic range of speech energy is defined by

$$r_x = \frac{\max_{t=1,T}(E_x[t])}{\min_{t=1,T}(E_x[t])} \quad (9)$$

According to equation (4), the ratio for the noisy speech will be

$$r_y = \frac{\max_{t=1,T}(E_y[t])}{\min_{t=1,T}(E_y[t])} = \frac{\max_{t=1,T}(E_x[t]) + E_w}{\min_{t=1,T}(E_x[t]) + E_w} < r_x. \quad (10)$$

Therefore, if the dynamic range of clean speech, r_x, is given, the noise energy can be computed by

$$\hat{e}_w = \frac{\max_{t=1,T}(E_y[t]) - r_x \min_{t=1,T}(E_y[t])}{1 - r_x}. \quad (11)$$

In practice, it could not be easy to estimate the exact r_x from the received noisy speech. However, a reasonable value of r_x could be set according to the training data of clean speech. In our experiments, the distribution of dynamic range ratio of the training data is investigated and the mostly occurred value $\bar{r}_x$ is taken as the r_x in equation (11) to estimate the noise energy in a test utterance. In fact, as shown in the experimental result, the setting of $\bar{r}_x$ is not so sensitive in terms of the recognition accuracy. Equation (11) can be rewritten by

$$\bar{r}_x = \frac{\max_{t=1,T}(E_y[t]) - \hat{e}_w}{\min_{t=1,T}(E_y[t]) - \hat{e}_w}. \quad (12)$$

Since $\bar{r}_x > r_y > 0$ and $\hat{e}_w > 0$, according to equation (12), we have $\min_{t=1,T}(E_y[t]) > \hat{e}_w$. Therefore, the noise energy removal can be performed by the direct subtraction,

$$\hat{E}_x[t] = E_y[t] - \hat{e}_w. \quad (13)$$

There are two benefits obtained from the method of using dynamic range. At first, it does not need an accurate speech/non speech indicator to estimate the noise energy. The second, the noise energy estimated from within the noisy speech may produce smaller error than that obtained from the outside of the speech period.

4. EXPERIMENTS ON NOISY SPEECH RECOGNITION

4.1.Experimental setting for TI Digits recognition

A connected digit recognition using the TIDigits database was conducted to evaluate the proposed methods. The speech recognizer was a one-stage decoder using continuous density hidden Markov models (HMMs). The speech models were obtained from the training part of the TIDigits database, which was collected from 55 males and 55 females through a desktop microphone in a quiet room. Each digit was expressed with a word model, which was consisted of 12 left-to-right states. The observation function of each speech state was a mixture of 4 Gaussian densities, while the mixture number was 16 for the silence state. The test part of the TIDigits database was equally divided into four subsets. Each subset was combined with one type of noise in signal-to-noise ratio of 20dB, 15dB, 10dB, 5dB, and 0dB. Three types of noise, Gaussian white noise, babble, and factory noise, were selected from the NOISEX 92, and one in-car noise, denoted by Civic, was selected from the NTT-AT database [8]. The in-car noise was collected within a car, which was cruising on a highway in high speed with closed windows. The speech signal was sampled in 8KHz and segmented into frames for feature extraction. The frame length was 30ms. with a shift rate of 50%. The speech features consisted of 26 components; 12 mel-frequency cepstrum coefficients (MFCCs), 12 delta MFCCs, one delta log energy, and one delta-delta log energy.

In the MFCC extraction, a 24-filter bank was applied to band-pass filter the magnitude spectrum of input signal [9]. The cepstrum mean normalization (CMN) was applied to the MFCCs. The log converted energy of speech signal was computed by the scaled c_0,

$$E_y^l[t] \equiv \frac{c_0[t]}{\sqrt{24}} = \frac{1}{24}\sum_{i=1}^{24}\log(S_{y,t}[i]) \qquad (14)$$

where $S_{y,t}[i]$ is the spectral magnitude in the i-th band pass filter at frame t. In this case, $E_y[t]$ will correspond to the geometric average of spectral magnitude of a frame. The delta MFCCs and delta log energy were computed across 5 frames, *i.e.*, L=2 in equation (1).

The experiment without any enhancement applied on the speech features was referred to as the baseline. The experiment with the method using the energy obtained from the background noise is denoted by EBN, in which the subtraction factors α and β were set to 0.95 and 0.05, respectively. The method using the dynamic range to estimate the noise energy was denoted by EDR In our experiments, the dynamic range of embedded speech in each noisy signal was assumed to be a mostly occurred one in the training data. The occurrence probability of dynamic range for each utterance in the training data was depicted to find the one with most occurrence. As shown in Fig. 1, we can see that the mostly occurred one is around 700, and which value was assigned to the experiments of EDR. To demonstrate the effect of the energy enhancement by using EDR and EBN, the log energy contour of noisy speech and its enhanced versions were compared in Fig. 2.

Besides of the methods EBN and EDR, an experiment using noise spectrum subtraction (NSS) to enhance the energy extraction was conducted for comparison. The spectral subtraction was performed in the stage of sub-band filter of the MFCC extraction. The noise spectrum was computed by the spectral average of sub-band filter from the speech inactive period. Equation (8) was applied to each Mel-frequency band for spectral subtraction, in which the setting of α and β were dependent on the sub-band SNR [10].

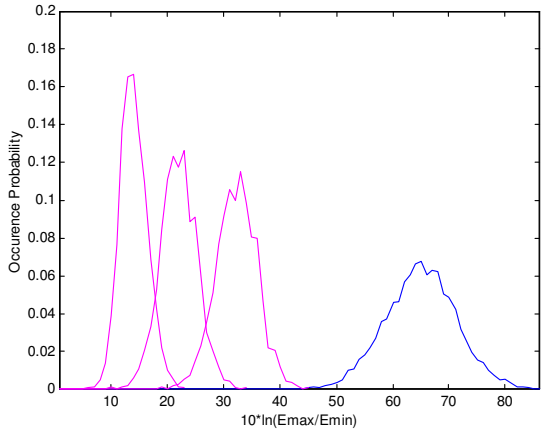

Fig. 1, Occurrence probabilities of dynamic range; the most right one is for the training data and the left three are for noisy speech with babble noise in SNR=0dB, 10dB, and 20dB (from left to right).

Fig. 2, Log energy contours of speech signal corrupted with the babble noise; the solid line is for the original one, the broken line is for the one enhanced by EBN, and the dotted line is for the one enhanced by EDR.

4.2. Experimental results and discussions

The performance resulted from different enhancement schemes were demonstrated in Table 1 with the average of recognition accuracy and relative improving rates. The detailed experimental results for different noisy conditions are given in Table 2 as well. In the first part of the experiment, the spectral features of speech signal were not enhanced by any scheme. In this case, applying EBN and EDR to enhance the energy feature can produce a relative improvement of 13.4% and 10.4%, respectively.

	Acc. (%)	Rel. Impr. (%)
Baseline	67.1	-
EBN	71.5	13.4
EDR	70.5	10.4
E_NSS	76.6	28.7
EBN_NSS	80.0	39.1
EDR_NSS	79.2	36.7
NSS_E	77.7	32.1
NSS_ EBN	79.9	38.9
NSS_ EDR	79.8	38.6

Table 1, Average digit recognition accuracy across 4 noise types in the SNR ranging from 20dB to 0dB, and relative improving rates from the baseline.

	Civic					white					babble					factory				
	20dB	15dB	10dB	5dB	0dB	20dB	15dB	10dB	5dB	0dB	20dB	15dB	10dB	5dB	0dB	20dB	15dB	10dB	5dB	0dB
Baseline	99.1	98.8	97.9	94.7	86.2	88.5	72.2	44.0	19.5	12.7	95.6	88.7	71.1	41.8	20.9	95.4	87.4	67.3	37.4	22.9
EBN	99.2	99.0	98.3	96.3	89.8	90.9	78.3	54.1	26.2	14.7	96.7	92.4	80.1	54.3	28.1	96.5	91.0	75.3	46.1	23.5
EDR	99.1	98.9	98.2	96.2	89.5	91.0	78.2	52.9	25.6	15.0	96.4	91.6	77.3	50.1	24.4	96.3	90.2	73.0	43.5	23.3
E_NSS	99.0	98.9	98.5	97.3	93.5	94.0	86.2	70.0	44.5	18.4	95.1	91.0	80.9	59.9	31.1	96.9	93.6	84.6	64.1	33.7
EBN_NSS	99.2	99.1	98.8	97.9	95.6	95.6	90.2	77.1	53.8	27.7	95.5	92.3	85.1	66.4	38.9	97.5	95.0	88.0	69.5	36.9
EDR_NSS	99.1	99.0	98.7	97.8	95.2	95.3	89.4	75.2	51.1	24.7	95.4	92.1	84.0	64.8	36.7	97.4	94.6	87.1	68.5	37.7
NSS_E	99.1	98.9	98.6	97.5	94.3	94.5	87.1	71.3	46.3	19.3	95.6	92.1	83.7	63.6	33.7	97.2	94.2	85.7	65.7	34.7
NSS_ EBN	99.2	99.1	98.8	98.0	95.5	95.6	90.4	77.8	55.2	29.7	95.3	91.9	84.3	65.0	37.4	97.3	94.7	87.5	68.8	36.7
NSS_ EDR	99.1	99.1	98.8	97.9	95.6	95.5	89.7	76.4	52.5	26.3	95.7	92.6	84.9	66.5	38.3	97.5	95.0	87.7	69.0	38.1

Table 2, Digit recognition accuracy (in %) for each test condition

In the second part, the test utterances were enhanced by the NSS in the feature extraction of MFCC. The experiments with energy feature extracted from the original signal and from the NSS enhanced one, denoted by E_NSS and NSS_E, were compared. As shown in Table 1, NSS_E can produce a further improving rate of 3.4 % with respect to the E_NSS. Applying EBN and EDR methods to enhance the energy feature of E_NSS, denoted by EBN_NSS and EDR_NSS, can produce further improving rates of 10.4 % and 8%, respectively. Again, we apply EBN and EDR methods to enhance the NSS_E, denoted by NSS_EBN and NSS_EDR, and obtain further improving rates of 6.8% and 6.5%, respectively.

According to the results, we can see that either EBN or EDR is able to provide an additional improvement when it is jointly applied with the spectral subtraction. The spectral subtraction method can improve the recognition rate significantly, but the energy feature extracted from the enhanced spectrum can not provide an equal effect as the EBN or EDR does. Since the noise/speech boundary is definitely determined in the experiments, the noise energy estimation in EBN is accurate so that the performance of EBN can outperform EDR. However, in practical applications, the noise energy estimation from speech inactive period may not be so exactly because of the inaccurate boundary decision.

5. CONCLUSION

It is shown that the absolute value of delta log energy and the dynamic range of energy contour will become smaller when the speech signal is corrupted with an additive noise. To reduce the distortion of energy feature, we can perform a noise energy removal before the feature extraction. In this work, we can see that the proposed energy enhancement schemes outperform the energy extraction from the enhanced spectra.

6. ACKNOWLEDGEMENTS

This paper is a partial result of project B33BXT5200 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, Taiwan.

7. REFERENCES

[1]Chen, J, Paliwal, K.K, & Nakamura, "Sub-band based additive noise removal for robust speech recognition", Eurospeech 2001.

[2]Jain, P. and Hermansky, H. "Improved mean and variance normalization for robust speech recognition," Proc. of ICASSP 2001.

[3]Molau, S., Pitz, M., and Ney, H., "Histogram base normalization in the acoustic feature space," Proc. of ASRU 2001.

[4]Wilpon,. J.G., Lee, C.-H., & Rabiner, L.R., "Improvements in connected digit recognition using higher order spectral and energy features," Proc. of ICASSP 91, pp. 349-352, 1991.

[5]Bocchieri,. E.L. & Wilpon, J.G., "Discriminative analysis for feature reduction in automatic speech recognition," Proc. of ICASSP 92, pp. 501-503, 1992.

[6]Hwang, T.-H., "Energy Contour Extraction for In-Car Speech Recognition", Proc. of Eurospeech 2003, pp. 2181-2184, 2003.

[7]Rabiner, L. & Juang, B.-H., Fundamentals of speech recognition, pp. 194-195, published by Prentice Hall, 1993.

[8]NTT-AT, Ambient Noise Database for Telephonometry 1996, http://www.ntt-at.com/index.html, NTT Advanced Technology Corp., 1996.

[9]Davis, S. B. & Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 28, No. 4, pp. 357-366, 1980.

[10]Schless, V. & Class, F.,"SNR-dependent flooring and noise overestimation for joint application of spectral subtraction and model combination," Proc. of ICSLP 98, pp.1495~1498.