IMPROVE SPEECH RECOGNITION PERFORMANCE IN REVERBERANT ENVIRONMENT BASED ON ESTIMATION OF ENERGY FEATURE

Dinh Cuong Nguyen
Information Technology Department - Nha Trang University
cuongnd@ntu.edu.vn

Abstract *- The purpose of this paper is to improve speech recognition performance in reverberant environment with distant talking,which based on directly speech processing or selected Log-energy feature. Speech recognition performance can be improved by changing the value of Log-energy feature or directly speech energy. Experiments used CENSREC-4 corpus to evaluate distant-talking speech under various reverberation environments by single microphone. And likelihood score string is proposed to optimal Energy feature of speech for each utterance to improve recognition performance. The results have showed improvements of speech recognition performance in reverberant environments with our proposals.*

Keywords *- Speech recognition, Log-energy feature, reverberant environment, CENSREC-4.*

I. INTRODUCTION

Speech recognition system has applied much in our life. Most of today's applications still require a microphone located near talker. However, almost all of these applications would benefit from distant-talking capturing, where talkers are enable to speak at some distance from microphone without the encumbrance of hand held or body equipment [1]. The major problem in distant-talking speech recognition is the corruption of speech signals by both interfering sounds and the reverberation caused by the large speaker -to-micro-phone distance. The range of successful technique have been developed since beginnings of speech recognition research to combat the additive and convolution noise caused by interfering sounds, microphone mismatch. In recent years, many research about speech energy estimation gave the improvements of speech recognition performance in reverberant environment. In [2, 3] Log-energy feature is proposed to improve speech recognition performance by adding energy value in MFCCs feature in noise environment.

In this paper, our research considered on directly speech processing and Log-energy feature in MFCC domain through changing value of speech energy in reverberant environment to improve speech recognition performance with CENSREC-4 corpus.

II. REDUCTION OF SPEECH ENERGY

In MFCCs feature domain, the energy is computed as the signal energy, that is, for speech sampling, $\{x(n)$, $n = 1, N\}$.

$$\log E = \log \sum_{n=1}^{N} x^2 \tag{1}$$

We can increase speech recognition performance in noise environment by adding more energy into Log-energy feature (C13 feature). In our research, the value of this feature is find by new method to increase recognition performance in reverberant environment. We used such way to re-estimate Log-energy feature as follows:

$$\log_{new}(E) = \log(E) + \Delta E \tag{2}$$

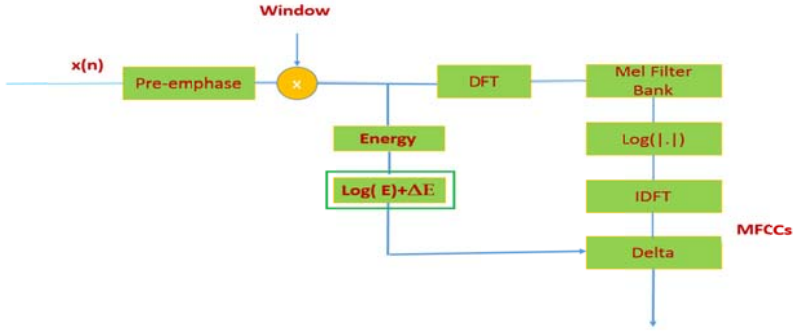

Fig. 1. Process flow of re-estimated for Log-energy feature

With $\log(E)$ is value of C_{13} feature domain, $\log(E)$ is re-estimated for the value of feature. From experimental results, we proposed ΔE in range [-MaxEnergy..MaxEnergy]. We have the proposals to find ΔE following

Proposal 1: get ΔE direct on range [-MaxEnergy .. MaxEnergy] to have maximum recognition performance for the system.

Proposal 2: Using for each utterance

Step 1: get ΔE using likelihood score, using clean utterance is to create optimal likelihood score string by recognition on each speech frame.

Step 2: Log energy (C_{13} feature) of reverberant speech is changed with ΔE. This ΔE value is find by minimum the error score between the likelihood score string of clean utterances and that of reverberant utterances. Fig. 2 showed our proposal to get the ΔE optimal value. We have the math model for ΔE finding. Using the likelihood score function with Decoding in HMM mode.

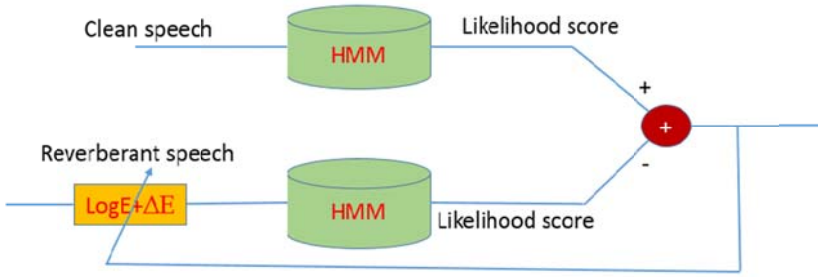

Fig. 2. Process flow to optimal value of *Energy*

Likelihood score calculation [4]: Let each spoken word be represented by a sequence of speech vectors or observations O, defined as $O = o_1, o_2, ..o$ where O_t is the speech vector observed at time t. The isolated word recognition problem can then be regarded as that of computing $\arg\max\{P(\omega_i \mid O)\}$ where i^{th} is the i^{th} vocabulary word. This probability is not computable directly but using Bayes's Rule gives

$$P(\omega_i \mid O) = \frac{P(O \mid \omega_i)P(\omega_i)}{P(O)} \tag{3}$$

Thus, for a given set of prior probabilities $P(O \mid \omega_i)$ the most probable spoken word depends only on the likelihood $P(O \mid \omega_i)$. A Markov model is a finite state machine which changes state once every time unit and each time t that a state j is entered, a speech vector o_t is generated from the probability density $b_j(o_t)$. Furthermore, the transition from state i to state j is also probabilistic and is governed by the discrete probability a_{ij}. The joint probability that O is generated by the model M moving through the state sequence X is calculated simply as the product of the transition probabilities and the output probabilities.

$$P(O, X \mid M) = a_{12}b_2(o_1)a_{22}b_2(o_2)a_{23}b_2(o_3) \tag{4}$$

Given that X is unknown, the required likelihood is computed by summing over all possible state sequences $X = x(1), x(2), x(3)...x(T)$, is the score which can be approximated by considering the most likely state sequence, that is

$$P(O \mid M) = \max\left\{a_{x(0)x(1)}\prod_{t=1}^{T} b_{x(t)}(O_t)a_{x(t)x(t+1)}\right\} \tag{5}$$

Given a set of models M_i corresponding to words ω_i, $P(O \mid \omega_i) = P(O \mid M_i) = score(.)$ function our problem is re-written such as (6)

$$L(\Delta E) = \arg\max_{\Delta E} score(\log E < frame i, -\infty >, \Delta E) \tag{6}$$

$$L(\Delta E) = \arg\max_{\Delta E} score_{\log E_i \mid \log E_{i-1}}(\log E_i \mid \log E_{i-1}, \Delta E) \tag{7}$$

assume that $\log E_{frame\ i}$ and $\log E_{frame\ i-1}$ are independent

$$L(\Delta E) = \arg\max_{\Delta E} score(\log E_i, \Delta E) \tag{8}$$

Here, observation vector at time t to be split into a number of S independent data streams of o_{st} we have the formula for computing $b_j(o_t)$ is then

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} c_{jsm} \eta\left(O_{st}; \mu_{jsm}, \sum_{jsm}\right) \right]^{\gamma_s} \tag{9}$$

where M_s is the number of mixture components in stream s, c_{jsm} is the weight of the *m th* component and $\eta(.;\mu,\Sigma)$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ, that is

$$\eta(o;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(o-\mu)'\Sigma^{-1}(o-\mu)} \tag{10}$$

where n is the demensionality of o, the exponent s is a stream weight. In practicality, if having the silent in speech frame each, we can not calculate the value for C_{13} energy by calculating on log energy function. In this case, we proposed this value equal to zero. Fig. 3. is Log energy feature function

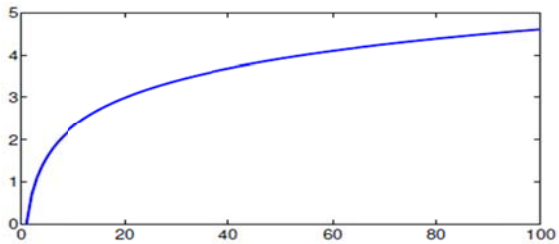

Fig. 3. Log Energy function

III. EXPERIMENTAL RESULTS

In this research, we used CENSREC-4 [5, 6] corpus as a framework for evaluating speech recognition performance in reverberant environment. CENSREC-4 defines two training types, clean and multi-condition. In Test A, clean training, the acoustic model is trained with non-reverberant clean speech data which consists of 8440 utterances spoken by 110 subject speakers (55 female and 55 males). Test B is used to estimate speech recognition performance for open test with data convolved with the remaining four room impulse responses (lounge, Japanese style room, meeting room and Japanese style bath). Each testing environment consists of 4004 utterance spoken 104 subject speakers (52 females and 52 males). The task is connected digit recognition in Japanese. Fig. 4 gives the impulse responses in many reverberant environments.

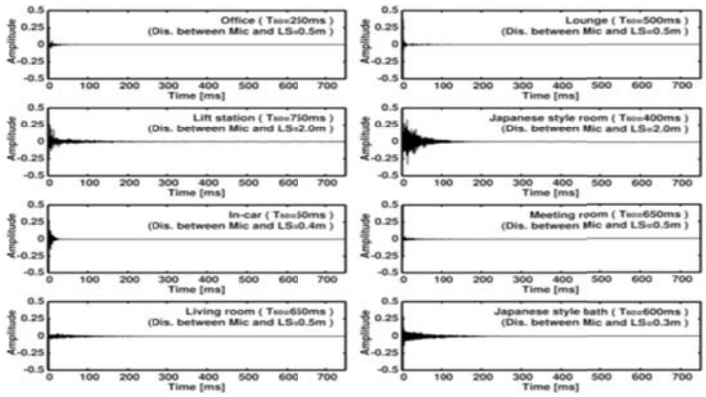

Fig. 4. Impulse response in eight environments of CENSREC-4[6]

In the basic data set, clean speech data was recorded with a close-talking microphone and impulse responses were recorded in eight kinds of rooms. Then reverberant speech data was simulated by convolutions of the data. The reverberant time in each room is showed in Table 1.

Table 1. Reverberant time (T60) of the rooms in CENSREC-4.

Office	Elevator Hall	In-Car	Living-room
0.25 sec	0.75 sec	0.05 sec	0.65 sec
Lounge	Japanese	Meeting	Japanese Style Path
0.05 sec	0.04 sec	0.65 sec	0.60 sec

The feature parameters of the baseline system used 39-dimension feature vectors that consist of 12-dimension MFCCs, 12-dimension delta, 12 delta-delta MFCCs, and one dimension each for log-energy, delta log-energy, and delta-delta log-energy. The analysis conditions used pre-emphasis (1-0.97z−1), hamming windows, 25-ms frame lengths and 10-ms frame shifts. The fast Fourier transformation (FFT) length was 512 and number of Mel-Filter Banks was 24. The sampling frequency of the data was 16kHz.

Experiment 1, Log-energy feature of each frame is re-estimated with our proposal 1. The value of ΔE for each room is selected as -2.5 (Lounge), -2.5 (Japan room), -1.5 (Conference) and -1.5 (Bath room). The experimental results are showed in Table. 2 for open Test B.

Table 2. Speech recognition performance with open Test B for Word (%)

	Lounge	Japan room	Confere -nce	Bath room	Average
baseline	74.06	89.51	89.78	78.06	82.85
Proposal 1	88.15	99.44	90.46	79.57	89.41

Speech recognition performance was increased 9.93 % (in Japan Room) and 14.09% (in Lounge). This results is higher that of baseline.

As we know that in a reverberant environment, speech signals have late reverberation, whose energy is plotted as a long term exponential decay. This tends to cause the Logarithmic Energy feature to keep the constant value for a long time. When speech energy was changed, the reverberation of speech is also changed in this condition. So that we can improve the speech recognition performance by changing of directly speech energy or Log Energy feature.

Experiment 2, using proposal 2 for improving recognition performance with each utterance, we have four of experimental results in Fig 5-8. Likelihood score is used to show recognition performance of speech for each frame. In this case, the recognition results score for reverberant speech is adapted to clean speech.

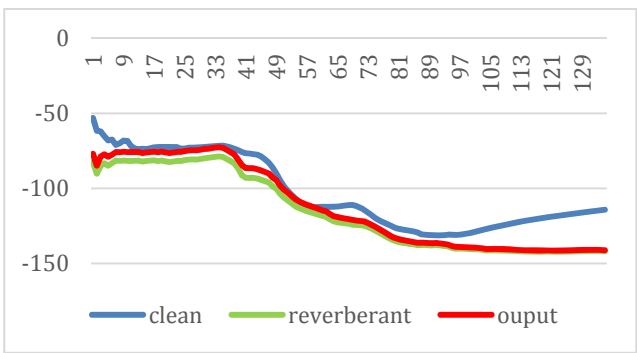

Fig. 5. An Optimal likelihood score speech in Lounge with proposal 2, ΔE=-5.8

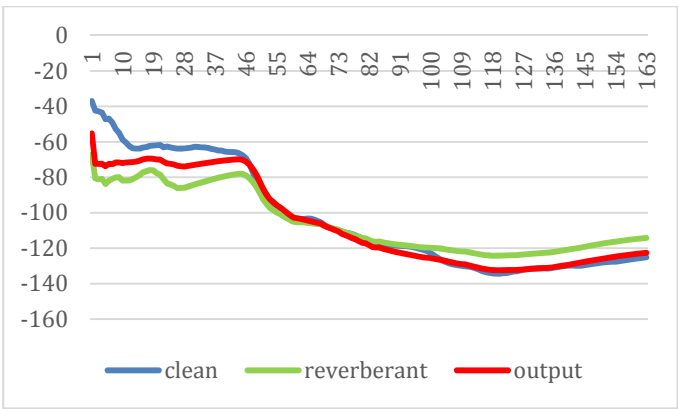

Fig. 6. An Optimal likelihood score speech in Japan room with proposal 2, ΔE=-14.1

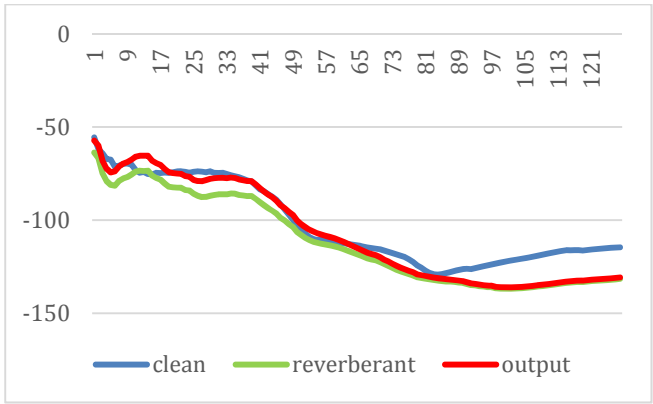

Fig. 7. An Optimal likelihood score speech in Conference room with proposal 2, ΔE=-6.8

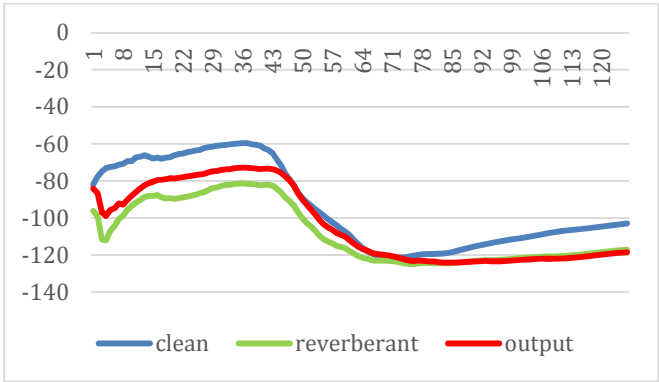

Fig. 8. An Optimal likelihood score speech in Bath room with proposal 2, ΔE=-11.9

IV. CONCLUSION

In this paper, we proposed another way to improve speech recognition performance in reverberant environment for CENSREC-4 corpus. We estimated directly speech energy for improvements of recognition performance by changing the Log energy feature. This experiment applied on MFCCs Log-energy feature is the direction for improvements of speech recognition performance. The experimental results gave that the speech recognition performance is increased 9.93% (in Japan Room) and 14.09% (in Lounge) compared to that of baseline.

Likelihood score is used as an optimal standard for improvements of speech recognition performance to find optimal value of ΔE that need to be changed for each utterance in the system.

V. REFERENCES

[1] T. Yoshioke, A. Sehn, M. Delcroix: "Masking machines understand us in reverberant room", IEEE Journal of Signal Processing, vol.4, no.5, pp.114-125, Oct. 2010.

[2] S. Guanghu, S. Soo-Young and C.Hyun-Yeol:"Improved feature en-hencement using temporal filtering in speech recognition," IEICE Electronics Express, vol.7, no.15, pp.1099-1105, Aug 2010.

[3] W. Z. Zhu and D. O Shaughnessy: "Log-energy dynamic range normalization for robust speech recognition," Proc. ICASSP, pp.245-248, 2005.

[4] HTK Toolkit, http://htk.eng.cam.ac.uk/.

[5] O. Ichikawa , T. Fukuda and M. Nishimura:"Dynamic features in the linear-logarithmic hybrid domain for automatic speech recognition in a reverberant environment," IEEE Journal of Selected Topic in Signal Processing, vol.4, no.5, Oct. 2010.

[6] T. Fukumori, T. Nishiura., M. Nakayama et al:"CENSREC-4: an evaluation framework for distant-talking speech recognition in reverberant environments," The acoustical Society Japan, 2011.

NÂNG CAO HIỆU QUẢ NHẬN DẠNG TIẾNG NÓI TRONG MÔI TRƯỜNG REVERBERANT QUA SỰ ĐÁNH GIÁ CỦA ĐẶC TRƯNG NĂNG LƯỢNG

Nguyễn Đình Cường

Tóm tắt - Bài báo nhằm nâng cao hiệu quả nhận dạng tiếng nói trong môi trường reverberant với điều kiện người nói cách xa microphone, việc nâng cao hiệu quả nhận dạng được dựa trên sự thay đổi trực tiếp năng lượng tiếng nói hoặc đặc trưng Log-Energy. Hiệu quả nhận dạng tiếng nói có thể được nâng cao bởi sự thay đổi giá trị của đặc trưng Log-energy. Thực nghiệm trên bộ dữ liệu CENSREC-4 cho việc đánh giá hiệu quả nhận dạng trong điều kiện người nói cách xa microphone qua nhiều môi trường nhận dạng khác nhau sử dụng đơn microphone. Và chuỗi likelihood score được đề nghị để tối ưu giá trị đặc trưng Log-Energy cho mỗi mẫu câu tiếng nói nhằm nâng cao hiệu quả nhận dạng. Kết quả nhận dạng cho thấy bằng phương pháp này chúng ta có thể nâng cao được hiệu quả nhận dạng tiếng nói trong điều kiện môi trường reverberant.