Performance Comparison of Multiple-Model Speech Recognizer with Multi-Style Training Method Under Noisy Environments

Yoon Jang-Hyuk and Chung Young-Joo*
*Department of Electronics Engineering, Keimyung University

(Received May 7, 2010; revised June 1, 2010; accepted June 7, 2010)

Abstract

Multiple-model speech recognizer has been shown to be quite successful in noisy speech recognition. However, its performance has usually been tested using the general speech front-ends which do not incorporate any noise adaptive algorithms. For the accurate evaluation of the effectiveness of the multiple-model frame in noisy speech recognition, we used the state-of-the-art front-ends and compared its performance with the well-known multi-style training method. In addition, we improved the multiple-model speech recognizer by employing N-best reference HMMs for interpolation and using multiple SNR levels for training each of the reference HMM.

Keywords: *HMM, Multi-model Speech Recognizer, Noise Robustness*

I. Introduction

Various research efforts have been made for the noise-robust speech recognition like speech feature extraction, speech enhancement and model parameter compensation [1-3]. These approaches are used independently or combined with each other to improve the performance of the speech recognizer under noisy environments.

As a different approach to those conventional methods, the multiple-model based speech recognizer has been proposed recently and shown quite successful results [4]. In the method, multiple acoustic models corresponding to various noise types and SNR levels are obtained during the training and the trained acoustic models are used altogether in the testing. This approach is contrary to the conventional methods where a single acoustic model corresponding to clean speech is used.

The real situation where the speech recognizer operates include various noisy environments and the distributed speech recognition (DSR) is thought to be one of the most representative noisy conditions. European Telecommunications Standards Institute (ETSI) has developed two standards for the DSR front-ends. The first standard is called FE. It is a basic version and specifies a feature extraction scheme based on the widely used mel frequency cepstral coefficients (MFCC) [5]. As the FE standard did not show successful results in noisy environments, the ETSI has proposed the second standard called AFE which includes some noise adaptive algorithms [6].

In the previous research [4], the multiple-model based speech recognizer has shown superior performance compared with the popular the Multi-

Corresponding author: Chung Young-Joo (yjjung@kmu.ac.kr)
Department of Electronics Engineering, Keimyung University
Daegu, 704-701, Korea

style TRaining (MTR) approach. However, the evaluation was done using the FE front-end instead of the more noise-robust front end, AFE. In this paper, we will evaluate the effectiveness of the multiple-model framework using the AFE front-end and compare its performance with the MTR method. We also propose methods to improve the performance of the multiple-model based speech recognizer. In the previous work, only one acoustic model which is most similar to the input noisy speech is selected for recognition but there are always some errors in this process due to the inaccurate SNR estimation and even the most similar acoustic model will not exactly match to the input noisy speech due to the noise signal variability. To overcome this problem with the multiple-model based recognizer, we propose to select N most similar acoustic models and use them all together in recognition. Also, the SNR range for each acoustic model is extended to generate more robust acoustic models during training.

II. Multiple-model based speech recognizer

2.1. Improved Multiple-Model Based Speech Recognizer.

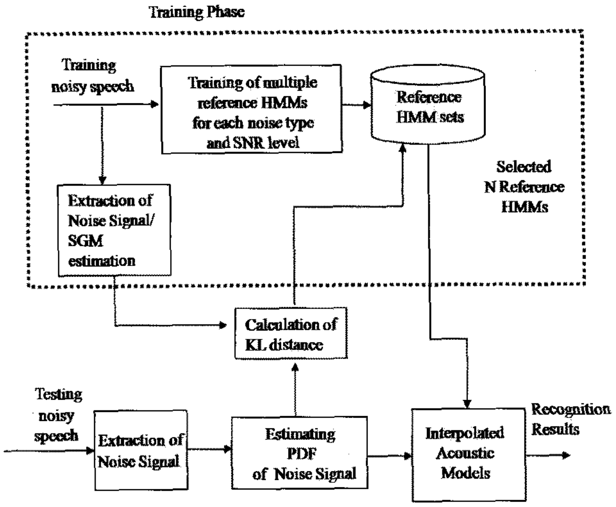
In the multiple-model based speech recognizer, multiple reference HMMs are trained using noisy speech corresponding to various noise types and SNR levels and one reference HMM which is most similar to the testing noisy speech is chosen as the acoustic model for recognition. This approach is advantageous over the conventional method using a single reference HMM because it can improve robustness against various noise characteristics.

In this paper, we modified the structure of the multiple-model based speech recognizer and its architecture is shown in Fig. 1. First, the noise signal extracted from the testing noisy speech is used to measure the similarity of the testing noisy speech to the reference HMMs and the most similar N reference HMMs are selected and they are interpolated for improved recognition performance. The interpolation can compensate for the errors in the selection process and the robustness of the recognizer is generally improved by using multiple acoustic models. When the probability density functions (PDFs) of the N most similar reference HMMs are given by $f_i(O), i=1,\cdots,N$, the interpolated PDF $f(O)$ is defined as follows.

$$f(O)=\sum_{i=1}^{N}\alpha_i f_i(O) \tag{1}$$

where O is the observation vector and $\alpha_i, i=1,\cdots,N$ are the interpolation weights.

In this paper, $\alpha_i=\frac{1}{N}, i=1,\cdots,N,$ are used to equally weight all the PDFs of the N reference HMMs. We experimented with assigning a distinct weight to each reference HMM but no significant performance improvement was observed. Single mode Gaussian models (SGMs) are estimated for each noise type and SNR level during the training. The estimated SGMs are used in selecting the N most similar reference HMMs. The SGM for the D-dimensional noise vector $\boldsymbol{n}$ with mean vector μ and covariance matrix Σ is given as follows.



Training Phase
Training noisy speech
Training of multiple reference HMMs for each noise type and SNR level
Reference HMM sets
Selected N Reference HMMs
Extraction of Noise Signal/ SGM estimation
Calculation of KL distance
Testing noisy speech
Extraction of Noise Signal
Estimating PDF of Noise Signal
Interpolated Acoustic Models
Recognition Results

Fig. 3. The architecture of the modified multiple-model speech recognizer.

$$p(n)=\frac{1}{(2\pi)^{D/2}|\Sigma|^{D/2}}exp\{-\frac{1}{2}(n-\mu)'(\Sigma)^{-1}(n-\mu)\} \quad (2)$$

Given the noise vectors, we can estimate the mean vector μ and covariance matrix Σ by the expectation-maximization (EM) algorithm.

In recognition, the Kullback-Leibler (KL) distances between the Gaussian PDF of the testing noise signal and the SGMs are calculated and those N SGMs with the smallest KL distances are determined and their corresponding N reference HMMs are chosen as the acoustic models for recognition in the multiple-model based speech recognizer.

The KL distance (KLD) between two Gaussian PDFs $N_1(\mu_1,\Sigma_1), N_2(\mu_2,\Sigma_2)$ is defined as follows [8].

$$KLD(N_1,N_2)=\frac{1}{2}\sum_{i=1}^{D}\left[\log\left(\frac{(\Sigma_1)_{ii}}{(\Sigma_2)_{ii}}\right)+\frac{((\mu_2)_i-(\mu_1)_i)^2}{(\Sigma_1)_{ii}}+\left(\frac{(\Sigma_2)_{ii}}{(\Sigma_1)_{ii}}-1\right)\right] \quad (3)$$

where $\Sigma_{1,ii}$ and $\Sigma_{2,ii}$ are the i-th diagonal components of the covariance matrices and $\mu_{1,i}$ and $\mu_{2,i}$ are the i-th components of the mean vectors.

As a second approach for the performance improvement of the multiple-model based speech recognizer, we used multiple SNR levels for training each of the reference HMM. Although a single SNR level is usually assigned to each reference HMM for more discriminative acoustic models, we improved robustness against the selection errors and noise variability by employing multiple SNR levels in the training.

2.2. Standards for the DSR front-ends

ETSI proposed two standard front-ends for the DSR speech recognition. The first standard ES 201 108 which was published in 2000 consists of two separate parts, feature extraction and encoding [5]. The widely used MFCC is generated in the feature extraction part while channel encoding for transmission is done in the encoding part. In this paper, we implemented only the feature extraction part as our concern is on the noise robustness of the front-ends. We call the first standard as FE and its block diagram is shown in Fig. 2.

The feature extraction part includes the compensation of the constant level offset, the pre-emphasis of high frequency components, the calculation of the spectrum magnitude, the bank of mel-scale filters, the logarithmic transform and finally the calculation of the discrete cosine transform. For every frame, a 14 dimensional feature vector consisting of 13 cepstral coefficients and a log energy is generated.

The FE front-end is known to perform inadequately under noisy conditions. Thus, a noise robust version of the front-end was proposed in 2002 [6]. This version called Advanced Front-End (AFE) is known to provide a 53 (%) reduction in error rates on the connected digits recognition task compared to the FE standard [7].

Fig. 3 shows a block diagram of the AFE front-end. Wiener filter based noise reduction, voice activity detection (VAD), waveform processing improving the overall SNR and blind equalization for compensating the convolutional distortion are added in order to improve the recognition rates.

The multiple-model based speech recognizer has shown improved results compared with the previous noise-robust methods like the MTR when they use the FE. However, for the accurate comparison, it is

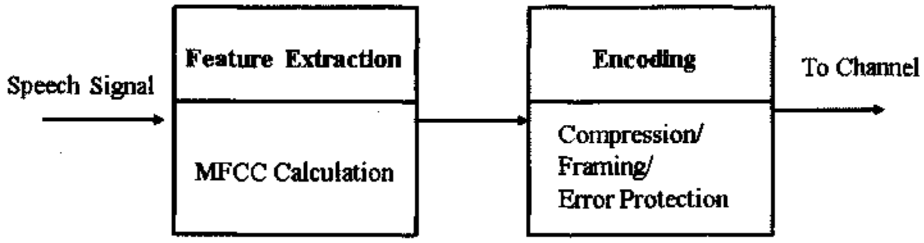


Fig. 2. Block diagram of the AFE

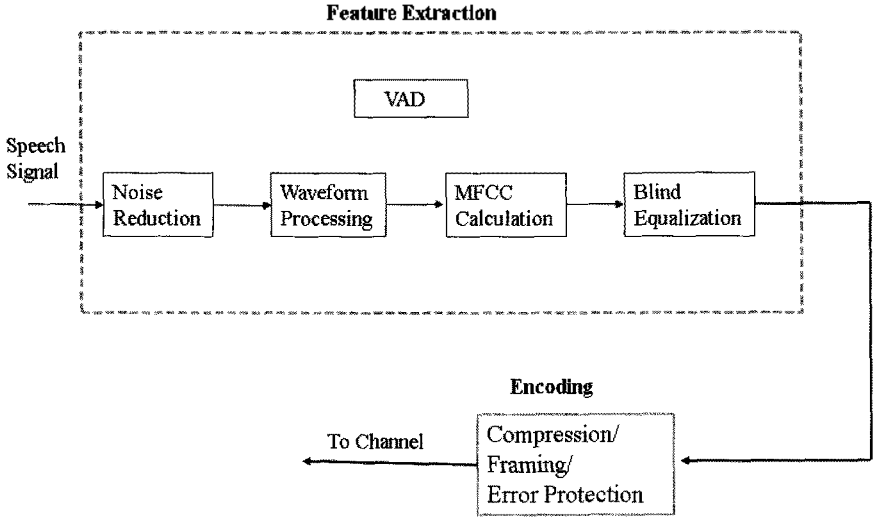


Fig. 3. Block diagram of the AFE.

necessary to compare the recognition rates when they use the AFE as the basic front-end because the AFE generally performs better than the FE in noisy conditions. Thus, in this paper, we evaluated the performance of the multiple-model speech recognizer using the AFE and proposed some methods to improve the recognition rates of the multiple-model based speech recognizer.

III. Experiments and Results

3.1. Databases and Recognition system

We used the Aurora 2 database for the experiments. There are two kinds of training approaches for the Aurora 2 database. The first one called CLEAN uses only clean speech not contaminated with any kinds of noises to train the HMM models. The second training method called MTR uses both clean and noisy speech signals which are contaminated by various kinds (subway, car, exhibition, babble) of noises at several SNR levels. The recognition experiments were conducted for Set A (including 4 known types of additive noise: subway, car, exhibition, babble), set B (including 4 unknown types of additive noise: restaurant, street, airport, train) and set C (including one known and one unknown type of noises with convolutional noises).

The AFE was used for the feature extraction. From the original feature vectors, 13-th order feature vectors which consist of 12-th order MFCCs without 0-th cepstral component and the log energy are generated as the basic feature vectors and their delta and acceleration coefficients are added to construct a 39-dimensional feature vector for each frame.

The HMM for each digit consists of 16 states with 3 Gaussian mixtures for each state. Silence is also modeled by a 3 state HMM with 6 Gaussian mixtures in each state [9]. The approximate Baum-Welch algorithm was used to obtain the acoustic models.

3.2. Experimental Results

To compare the performance of the FE and AFE in noisy speech recognition, we show the word error rates (WERs) when the acoustic models are trained by CLEAN and MTR method.

As we can see in Table 1, the average word error rate (WER) with the FE was 38.78 (%) in CLEAN training mode while the WER with the AFE was 14.37 (%), which means that the AFE reduces the WER by 63(%) in CLEAN training mode. For the case of MTR training, we can also see that the AFE reduces the WER by about 35(%) compared with the FE. From these results, we can conclude that the AFE performs much better both in the CLEAN and

MTR training mode on the Aurora 2 database. This also means that the previous research which demonstrated the superiority of the multiple-model based recognizer using the FE should be re-evaluated using the AFE.

In Table 2, we show the WERs of the multiple-model based recognizer using the AFE as the number of interpolated PDFs in (1) varies. The conventional multiple-model based recognizer corresponds to the case of N=1. Comparing with the results in Table 1, we can see that the conventional multiple-model based speech recognizer performs worse than the MTR method. This is because the use of various types of noise signals in the MTR method improves significantly its robustness against unknown types of noise signals. This resulted in the large difference in recognition rates for set B and consequently the inferior average recognition rate of the multiple-model based speech recognizer[4]. As we increase the number of interpolated PDFs, some performance improvement is observed. We could obtain the best performance when N=4 with the WER of 10.71 (%) reducing the WER of the conventional method by about 3 (%).The decrease in the WER mainly comes from Set C where a 10 (%) error rate reduction is achieved. The improvement may have come from reducing the negative effect of errors in finding the most similar reference HMM using the KL distance. Also, the variability of the noise signal in the testing noisy speech may have been more efficiently compensated by using multiple acoustic models in recognition.

Table 1. Performance comparison between the AFE and FE (WER (%)).

Training method	Front-end	FE	AFE
CLEAN	set A	37.43	13.67
	set B	42.94	14.58
	set C	33.08	15.36
	평균	38.75	14.37
MTR	set A	12.55	8.51
	set B	13.71	8.94
	set C	17.03	9.83
	Average	13.91	8.95

Table 2. The performance of the multiple-model based recognizer using the AFE (WER (%)).

Number of interpolated HMMs (N)	set A	set B	set C	Ave.
1	9.28	13.24	9.95	11.00
2	9.16	13.21	9.49	10.85
4	9.17	13.15	8.92	10.71
6	9.18	13.32	8.8	10.76

In addition to the interpolation approach, we also tried to improve the performance of the multiple-model based speech recognizer by using multiple SNR levels for training each of the reference HMM. In Table 3, we show the two cases of merging SNR levels called SNRMERG, SNRMER2.

In the conventional method, the reference HMM was constructed for each SNR level (0, 5, 10, 15, 20, 25, 30 dB) independently while the SNRMERG method merged 0 and 5, 10 and 15, 20 and 25 to construct the reference HMMs reducing the number of reference HMMs for each noise type from 7 to 4. While SNRMERG2 method is similar to SNRMERG, it allows overlap in SNR levels among different reference HMMs.

In Table 4, we compared the performance of the proposed SNRMERG and SNRMERG2 method.

As we can see in Table 4, the overall recognition rates of the SNRMERG and SNRMERG2 are better than the conventional method. In Table2, the conventional method had the WER of 11.0 (%) when N=1 while the SNRMERG and SNMERG2 had the WERs of 10.80 (%) and 10.54 (%) respectively. Also, the recognition rate of the SNRMERG improves further by increasing the number of interpolated

Table 3. The SNR levels for each noise type and the resulting number of reference HMMs for each noise type.

	Conventional Method	SNRMERG	SNRMERG2
SNR Levels (dB)	{0},{5}, {10},{15}, {20},{30}	{0,5}, {10,15}, {20,25},{30}	{0,5},{5,10}, {10,15}, {15,20}, {20,25},{25,30}, {30}
Number of reference HMMs	7	4	7

HMMs. However, we could not see further improvement with increasing N for the SNRMERG2. This is because we can have similar effect as interpolating PDFs by using multiple SNR levels in training the reference HMMs. So, the results in Table 4 do not show remarkable increase in recognition rates as in Table 2 by increasing N. Although the difference in lowest WERs between the SNRMERG and SNRMERG2 is small, the SNRMERG2 has a merit that it does not need the interpolation to obtain the lowest WER.

We compared the improved multiple-model based speech recognizer with the MTR method which is a very popular approach in noisy speech recognition and the comparison results are shown in Table 5.

In Table 5, SNRMERG (N=4) and SNRMERG2 (N=1) showed lower WERs than the conventional multiple-model based speech recognizer but they were worse than the MTR. This is contrary to the previous research result where the multiple-model based recognizer outperformed the MTR when the FE was used as the basic front-end [4]. The noise reduction algorithm in the AFE may have diminished the relative merit of noise robustness of the multiple-model based speech recognizer.

To increase the recognition rates of the proposed multiple-model based recognizer, we interpolated the PDF of the SNRMER2 (N=1) with that of the MTR to take the advantage of the MTR. Although the average recognition rate of the interpolated acoustic model still falls slightly short of that of the MTR, it shows better recognition rates for Set A and C. The quite inferior results for Set B contributed to the overall performance degradation. As the Set B consists of noisy speech with unknown noise types, the recognition rates of the interpolated acoustic model may be further increased and outperform the MTR by applying model parameter compensation approaches for the multiple-model based speech recognizer, which is the topic of our further study.

IV. Conclusions

As compared to the conventional method where one single reference HMM is chosen as the acoustic model for recognition, we improved the performance of the multiple-model based speech recognizer by selecting N most similar reference HMMs based on the KL distance between the SGM of the training noise signal and the PDF of the noise in the testing noisy speech. We could also increase the recognition rates of the multiple-model based recognizer by using multiple SNR levels for training each of the reference HMM. To further improve the performance of the multiple-model based recognizer, the PDFs of the reference HMMs are interpolated with that of the MTR. The interpolated acoustic model performed better than MTR for the Set A and Set C in the Aurora 2 database. We think that the performance of the multiple-model based recognizer could be further improved by applying model parameter compensation approaches.

Table 4. Performance comparison of the SNRMERG and SNRMER2 method (WER (%)).

	N	set A	set B	set C	Average (95% confidence interval)
SNRMERG	1	9.01	13.12	9.75	10.80 (±0.150)
	2	8.60	13.04	9.02	10.46 (±0.148)
	4	8.94	13.01	8.49	10.48 (±0.148)
	6	9.17	13.07	8.49	10.59 (±0.149)
SNRMERG2	1	8.80	12.72	9.66	10.54 (±0.149)
	2	8.63	13.02	9.38	10.54 (±0.149)
	4	8.70	13.17	9.10	10.57 (±0.149)
	6	8.93	13.28	8.66	10.62 (±0.149)

Table 5. Performance comparison of the multiple-model based speech recognizer with the MTR method (WER (%)).

	set A	set B	set C	Average (95% confidence interval)
Conventional Method	9.28	13.24	9.95	11.00 (±0.151)
SNRMERG (N=2)	8.94	13.01	8.49	10.48 (±0.148)
SNRMERG2 (N=1)	8.80	12.72	9.66	10.54 (±0.149)
MTR	8.51	8.94	9.83	8.95 (±0.138)
SNRMERG2 (N=1)+MTR	8.21	10.66	8.46	9.24 (±0.140)

Acknowledgement

This work was supported by the Korea Research Foundation (KRF) grant funded by the Korea government (MEST) (No. 2009-0075480)

References

1. M. J. F. Gales, "Model based techniques for noise-robust speech recognition" , Ph.D. Dissertation, University of Cambridge, 1995.
2. P. J. Moreno, "Speech recognition in noisy environments" , Ph.D. Dissertation, Carnegie Mellon University, 1996.
3. S. F. Ball, "Suppression of acoustic noise in speech using spectral subtraction" , *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, pp.113-120, 1979.
4. H. Xu, Z.-H. Tan, P. Dalsgaard and B. Lindberg, "Robust Speech Recognition on Noise and SNR Classification - a Multiple-Model Framework" , in *Proc. Interspeech*, 2005.
5. ETSI draft standard doc. *Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm, ETSI Standard ES 202 108.*, 2000.
6. ETSI draft standard doc. *Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced Front-end feature extraction algorithm; Compression algorithm, ETSI Standard ES 202 050*, 2002.
7. D. Macho, L. Mauuary, B. Noe, Y. Cheng, D. Eahey, D. Jouvel, H. Kelleher, D. Pearce, F. Saadoun, "Evaluation of a noise-robust DSR front-end on Aurora databases" , in *Proc. ICSLP*, pp.17-20, 2002.
8. B. H. Juang and L. R. Rabiner, "A Probabilistic Distance Measure for Hidden Markov Models" , *AT&T Technology Journal*, pp. 391-408, 1984.
9. D. Pearce and H. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under conditions" , in *Proc. ICSLP*, pp.29-32, 2000.

[Profile]

- Yoon Jang-Hyuk

He is a student at the Department of Electronics, Keimyung University. His major research interests include speechl recognition and its applications.

- Chung Yong-Joo

He received the B.S. degree in Electronics Engineering from Seoul National University in 1988 and he also received M.S. and Ph.D. degrees in Electrical Engineering from Korea Advanced Institute of Science and Technology in 1990 and 1995, respectively. Since 1999, he has been with Keimyung University as a professor at the Dept. of Electronics Engineering.