Applied Mathematics & Information Sciences
An International Journal

@ 2012 NSP
Natural Sciences Publishing Cor.

Optimal SNR Model Selection in Multiple-Model Based Speech Recognition System

Yongjoo Chung
Department of Electronics Engineering, Keimyung University, Daegu, S. Korea
Corresponding author: Email: yjjung@kmu.ac.kr

Received June 22, 2010; Revised March 21, 2011; Accepted 11 June 2011
Published online: 1 January 2012

Abstract: In the multiple-model based speech recognition system, multiple HMM models corresponding to different types of noise signals and SNR values are trained and the one model which is most close to the input speech is selected for recognition. In the previous research on the multiple-model based speech recognition, it has been thought that the best performance can be obtained by selecting the HMM model which is most similar in SNR values to the input speech. But, from our experimental results, it has been found that better performance can be obtained when there is some mismatch between the SNR values of input speech and the selected HMM model. In this paper, we experimentally determined the optimal HMM models corresponding to the SNR values of the input speech in the multiple-model based speech recognizer. From the recognition experiments on Aurora 2 database, we could see far better recognition results compared with the conventional multiple-model based speech recognizer by using the experimentally determined optimal HMM models.

Keywords: Multi-Model, Noisy Speech Recognition, HMM, SNR

1 Introduction

Various research efforts have been done for the noise-robust speech recognition like speech feature extraction, speech enhancement and model parameter compensation [1][2][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 (MMSR) has been recently proposed and shown quite successful results [4]. In the method, multiple acoustic models corresponding to various noise types and SNR levels which are expected to be present in the testing speech are obtained during the training and the trained acoustic models are used all together in the testing. This is contrary to the conventional methods where a single acoustic model corresponding to clean speech is used.

The MMSR has shown better performance than the conventional noisy speech recognition approaches. It performed better than the model parameter compensation methods like the Parallel Model Combination (PMC) and Jacobian Adaptation (JA) as well as the speech enhancement method like the Spectral Subtraction (SS). And even better, it performs better than the Multi-style TRaining (MTR) which has recently been known to perform very well in noisy speech recognition [5].

The MMSR should classify the type of noise signal in the testing speech and also estimate the SNR of the input speech to determine the reference HMM before the actual speech recognition takes place. As the errors in this process will cause misrecognition, the performance of the MMSR would be improved significantly by minimizing those errors.

According to the previous research results, classification accuracy of the noise signal type based on Gaussian Mixture Model (GMM) is nearly 100(%) [4]. This means that the classification of the noise signal type doesn’t affect adversely the performance of the MMSR. However, the process of estimating the SNR and determining the reference HMM based on the SNR requires some notice. In the previous research on the MMSR, they selected the reference HMM which is closest to the estimated SNR. But, from our preliminary studies, we found that we could further improve the performance by selecting the reference HMM which has a slightly different (higher or lower) SNR value than the estimated SNR.

In this paper, we will do an experimental investigation to determine the reference HMM which gives the best performance given the SNR of the input speech. We expect to improve the performance of the conventional MMSR by employing the experimentally determined SNR mappings rather than using directly the estimated SNR value to select the reference HMM.

2 Multiple-Model Speech Recognizer

In the MMSR, multiple reference HMMs for the assumed various noise environments are constructed during the training and the reference HMM which is most appropriate for the testing noisy speech is selected for recognition. To select the reference HMM, we need to estimate the SNR of the testing noisy speech and classify its noise type. We show the architecture of the MMSR in Figure 2.1. This approach has the advantage of improving the noise-robustness compared with the conventional method in which only a single reference HMM corresponding to the clean speech is considered.

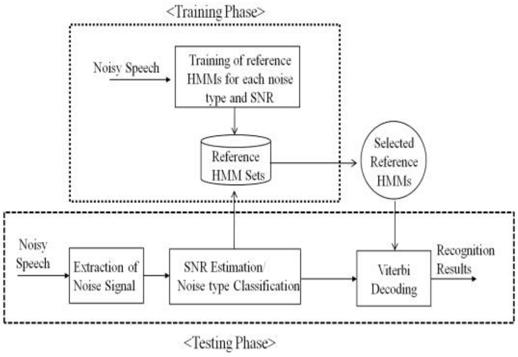

Figure 2.1: The architect of the multiple-model based speech recognizer which is divided into 2 parts: training phase and testing phase.

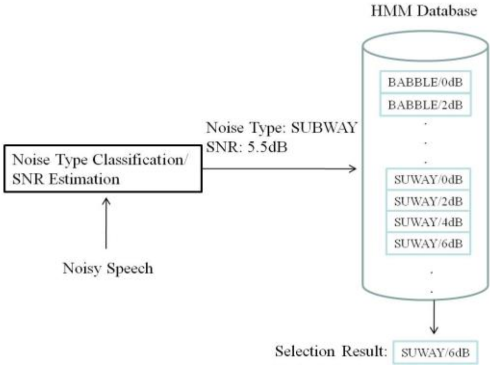

Figure 2.2 : An example of the reference HMM selection process in the multiple-model based speech recognizer

In Figure 2.2, we show an example of the reference HMM selection process in the MMSR. We have constructed the reference HMM in every 2 dB interval for the 4 known types of noise signal (babble, car, subway, exhibition) during the training phase and the trained reference HMMs are stored in the HMM Database. In the example, the type of the noise signal from the testing noisy speech is classified as subway and the SNR is estimated as 5.5 dB. This information is sent to the HMM Database and the reference HMM is determined as SUBWAY/6dB which is most close to the testing noisy speech. In the conventional MMSR, it is thought that choosing the closest reference HMM as described above will result in the best performance. However, in this paper, we experimentally determined the optimal SNR level of the reference HMM given the estimated SNR of the testing speech for further improvement of the performance.

2.1 SNR Estimation and Noise Type Classification

The MMSR utilizes the VAD (Voice Activity Detection) based SNR estimation. In the method, the power of the noise signal $\hat{\sigma}_n^2$ is estimated using samples in non-speech periods obtained by the VAD and the noise signal power is subtracted from the signal power $\hat{\sigma}_x^2$ during the speech periods to estimate the real speech power. The ratio of the real speech power to noise power is determined as the SNR. The expression for the SNR of the noisy speech is as follows.

$$SNR = 10\log\frac{\hat{\sigma}_x^2 - \hat{\sigma}_n^2}{\hat{\sigma}_n^2} \qquad (1)$$

For the noise signal classification, the feature vector of the noise signal n is modeled by the GMM. The GMM represents the weighted linear combination of the Gaussian probability density functions and is expressed as follows.

$$\begin{aligned} p(n) &= \sum_{i=1}^{M} \omega_i p_i \ \ (n) \\ &= \sum_{i=1}^{M} \omega_i \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{D/2}} \cdot \qquad (2) \\ &\qquad \exp\left(-\frac{1}{2}(n-\mu_i)'(\Sigma_i)^{-D}(n-\mu_i)\right) \end{aligned}$$

In (2.2), the weight factor ω_i satisfies $\sum_{i=1}^{M}\omega_i = 1$ and μ_i , Σ_i represent the D-dimensional mean vector and covariance matrix of the Gaussian probability density function $p_i(n)$, respectively. For the noise signal classification, the GMM is trained for each noise type during the training via the Expectation-Maximization (EM) based maximum likelihood estimation.

2.2 Standards for Speech Front Ends

In this paper, we used 2 standards of speech front-ends to more accurately compare the proposed method with the conventional approaches. The European Telecommunications Standards Institute (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 Figure 2.3.

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].

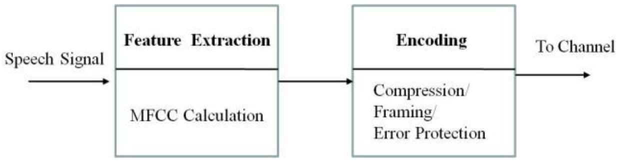

Figure 2.3: The block diagram of the FE front-end which consists of feature extraction and encoding process.

Figure 2.4 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 convolution 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 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 both the FE and AFE and proposed a method to improve the recognition rates of the multiple-model based speech recognizer.

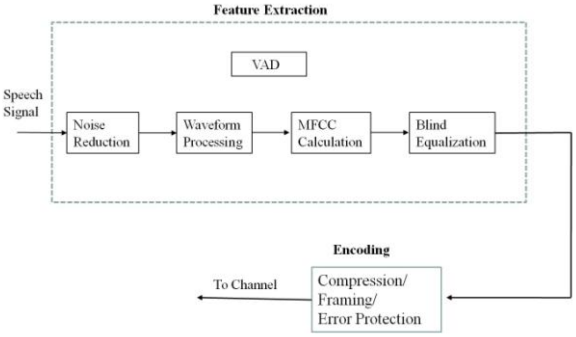

Figure 2.4: The block diagram of the AFE which consists of feature extraction and encoding process.

3 Baseline System and Databases

We used the Aurora 2 database for the experiments. There are two kinds of training approaches for the Aurora 2 database [5]. The first one called CLEAN uses only clean speech not contaminated with any kinds of noises to train the HMMs. The second training method called MTR uses both clean and noisy speech which is contaminated by various kinds (subway, car, exhibition, babble) of noises at several SNR levels. The recognition experiments are 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 convolution noise).

We used 2 widely known speech features for the experiment. The first one called FE in which 12-th order Mel-Frequency Cepstral Coefficients (MFCCs) without 0-th cepstral component and the log energy are used as the basic feature vector and their delta and acceleration coefficients are added to construct a 39-dimensional feature vector for each frame [6]. The noise robust version of the speech feature is called AFE which is known to significantly reduce the word error rate in noisy conditions. We extracted 39-dimentional feature vectors for the AFE as was done in FE.

The HMM for each digit consists of 16 states with 3 Gaussian mixtures in each state. Silence is also modeled by a 3 state HMM with 6 Gaussian mixtures in each state. We used the Hidden Markov Toolkit (HTK) developed at the Cambridge University as the basic speech recognizer. We added the 4 known types of noise signal to the clean speech to generate the noisy speech signal for training the reference HMMs in the MMSR. To construct sufficient number of reference HMMs, the noisy speech signal is generated for every 2 dB interval between 0 and 30 dB so that 16 reference HMMs are constructed for each noise type. The total number of reference HMMs used in the experiment is 4x16=64 and one of them will be selected for recognition depending on the noise type and SNR level of the input noisy speech.

4 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 4.1, the average word error rate (WER) with the FE was 39.94(%) in CLEAN training mode while the WER with the AFE was 14.46(%), 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.

Table 4.1: Performance comparison in word error rates(%) between the AFE and FE for the two training methods: CLEAN and MTR.

FrontEnds / Training Methods		FE	AFE
CLEAN	Set A	38.66	13.81
	Set B	44.25	14.76
	Set C	33.86	15.28
	Average	**39.24**	**14.46**
MTR	Set A	12.23	8.22
	Set B	13.75	8.89
	Set C	16.42	9.43
	Average	**13.68**	**8.73**

In Table 4.2, we show the result of the noise type classification. The highest classification accuracy of 99.9(%) is obtained for the exhibition noise and the lowest accuracy is 99.2(%) in the car noise. For the 4 types of noise signal, the average classification rate is 99.6(%). This means that there will be little performance degradation in the MMSR due to the noise type classification and allows us to focus on the SNR for the performance improvement of the MMSR.

Table 4.2 : The result of the noise type classification accuracy(%) using the GMM for the 4 types of known noise signal.

Reference Noise Type / Testing Noise	car	babble	exhibi-tion	subway
car	99.2	0.8	0.0	0.0
babble	0.4	99.6	0.0	0.0
exhibition	0.1	0.0	99.9	0.0
subway	0.0	0.0	0.3	99.7

In Table 4.3, we compared the word error rate of the MMSR with other approaches using the FE front-end. From the table, we can see that the MMSR outperforms both CLEAN and PMC but it is only slightly better than the MTR method. We think that if we more adequately select the reference HMM in the MMSR, we can see more performance improvement than shown in Table 4.3

Table 4.3: Comparison in word error rates of the MMSR with other approaches using the FE front-end.

	Set A	Set B	Set C	Average
CLEAN	38.66	44.25	33.86	39.94
PMC	20.70	18.82	21.98	20.20
MTR	12.23	13.75	16.42	13.68
MMSR	8.92	16.64	15.09	13.24

We did an experimental investigation on mapping the estimated SNR of the input noisy speech to select the optimal reference HMM model in the MMSR. We generated the testing noisy speech in every 2 dB interval using the 4 types of known noise signal in set A and selected the reference HMM which gives the best performance for each testing noisy speech. In Table 4.4, we show the variation in the word error rate depending on the selection of the reference HMM for the testing noisy speech with additive babble noise. In the table, we can see that the best performance is achieved when there is some mismatch between the SNR of testing noisy speech and the reference HMM. For example, when the SNR of the testing noisy speech is 0dB, the word error rate is 43.65(%) for the reference HMM with 0dB while the word error rate is 32.07(%) for the reference HMM with 6dB. This is contrary to the conventional idea with the MMSR that the best performance will be achieved when the SNRs of the testing speech and the reference HMM are identical. Although prominent in the low SNR regions, the phenomena can be seen in almost all SNR regions in Table 4.4.

Table 4.4: Variation in word error rates depending on the selection of the reference HMM given the SNR of the testing noisy speech (babble noise).

SNR (Testing speech)	SNR (Reference HMM)										
	0	2	4	6	8	10	12	14	16	18	20
0	43.65	36.06	32.59	**32.07**	32.35	-	-	-	-	-	-
2	30.32	23.91	21.19	**19.50**	20.50	21.52	-	-	-	-	-
4	-	19.98	15.24	14.09	**13.18**	**13.18**	15.05	-	-	-	-
6	-	-	10.82	9.46	8.13	**8.01**	8.56	-	-	-	-
8	-	-	-	-	5.83	**5.32**	5.35	6.08	-	-	-
10	-	-	-	-	4.20	4.14	**4.11**	4.17	-	-	-
12	-	-	-	-	-	3.14	3.08	3.05	**2.90**	3.23	-
14	-	-	-	-	-	-	2.36	2.39	2.24	**2.03**	2.30
16	-	-	-	-	-	-	-	1.81	1.69	**1.54**	1.63
	12	14	16	18	20	22	24	26	28	30	
18	-	-	1.75	**1.63**	1.69	1.72	-	-	-	-	
20	-	-	-	**1.36**	**1.36**	1.45	1.42	-	-	-	
22	-	-	-	-	1.48	1.51	1.39	**1.36**	1.45	-	
24	-	-	-	-	-	1.36	1.39	1.36	**1.30**	**1.30**	
26	-	-	-	-	-	1.33	**1.21**	1.24	**1.21**	**1.21**	
28	-	-	-	-	-	1.48	**1.30**	1.33	**1.30**	**1.30**	
30	-	-	-	-	-	1.36	1.27	1.24	1.21	**1.09**	

In addition to the babble noise, we could see similar results in 3 other types of noise signal (subway, car, exhibition) in set A. Based on the experiments, we could determine the optimal SNR of the reference HMM given the testing noisy speech for each noise type and the results are shown in Table 4.5.

SNR (Testing peech)	SNR (Reference HMM showing the best performance)			
	Babble	Subway	Car	Exhibi -tion
0	6	4	2	2
2	6	6	4	4
4	8	6	6	6
6	10	8	8	8
8	10	8	8	10
10	12	12	12	10
12	16	10	14	12
14	18	14	14	16
16	18	16	20	18
18	18	18	18	18
20	20	22	20	20
22	26	22	26	26
24	28	22	28	28
26	28	22	30	28
28	28	22	30	30
30	30	24	30	30

Table 4.5 : The SNR of the reference HMM showing the best word recognition rate as the SNR of the testing speech varies.

As expected, we can see from Table 4.5 that there is some difference between the SNRs of the testing noisy speech and the reference HMM which gives the best word recognition rate. For the testing noisy speech with low SNR, it is advantageous to select the reference HMM with higher SNR than the estimated SNR to improve the word recognition rate.

In Figure 4.1, we show the word error rate of the MMSR when the reference HMM is selected using SNR mappings shown in Table 4.5. For comparison, we also show the word error rate of the conventional MMSR and the MTR method. We used the FE front-end as the speech features. From the results in Figure 4.1, we can see that the conventional MMSR performs better than the MTR method for set A with noisy speech generated from the 4 known types of noise signal and for set C with the convolution noise. However, the MMSR performs worse compared with the MTR method for the set B with unknown types of noise signal. This results from the fact that the reference HMM in the MMSR has been constructed using the noise signal from set A. Overall, the conventional MMSR outperforms slightly the MTR method in average (13.24 (%) vs. 13.61 (%)). However, we could achieve the word error rate of 12.40(%) using the SNR mapping proposed in this paper reducing the word error rate by 6.34 (%) compared with the conventional MMSR. Although the SNR mapping method is inferior to MTR for set B due to the unknown noise signal, it has shown better performance in all cases (set A, set B, set C) than the conventional MMSR. This means that the SNR mappings suggested in Table 4.5 is valid irrespective of the noise type.

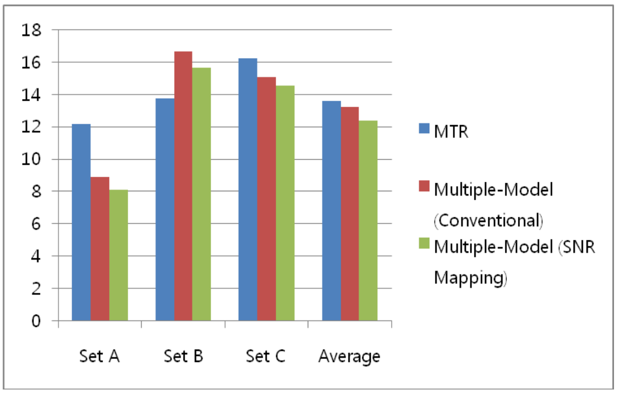

Figure 4.1 Word error rates (%) of the multiple-model based speech recognizer and its comparison with the conventional methods when using the FE front-end

In Figure 4.2, we also show the word error rate when we use the AFE as the front-end. Contrary to the result in Figure 4.1, the performance of the MTR improves significantly and in average, outperforms the conventional MMSR (8.22(%) vs. 9.01(%)). However, we can see that the word error rate of the proposed MMSR also decreases significantly to 8.17(%) which is better than the MTR method. The performance difference between the proposed SNR mapping method and MTR is not as significant as when we used the FE front-end. However, we have more room to improve the performance of the proposed MMSR by adapting the acoustic models to the input noisy speech. JA is a commonly used approach for this purpose and has shown to be quite effective for the MMSR. Improving the noise robustness of the proposed SNR mapping method by the acoustic model adaptation will be our future research topics.

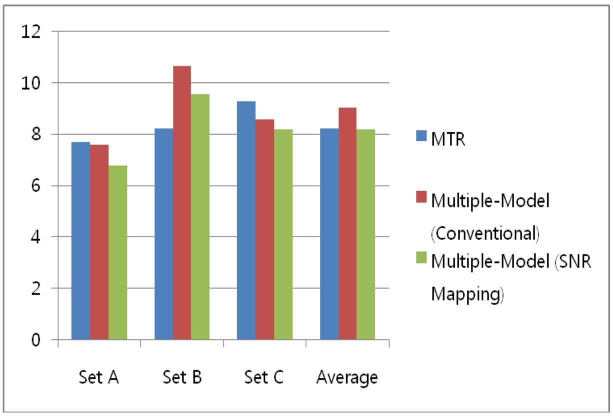

Figure 4.2 : Word error rates (%) of the multiple-model based speech recognizer and its comparison with the conventional methods when using the AFE front-end.

5 Conclusion

In this paper, we proposed to select the reference HMM in the multiple-model based speech recognizer by mapping the estimated SNR of the testing speech. This approach is contrary to the conventional method where the estimated SNR is used directly in the selection process. From the experimental results on the noisy speech recognition, we could achieve far better recognition rates than the conventional multiple-model based speech recognizer. Also, its performance was better than the MTR method, especially with the FE front-end. Although the proposed method is only slightly better than the MTR method when we use the AFE front-end, we expect that its performance will outperform the MTR method by a significant margin when we apply model adaptation methods to the proposed multiple-model based speech recognizer.

Acknowledgements

The present research has been conducted by the Bisa Research Grant of Keimyung University in 2009.

References

[1] Gales, M.J.F. Model Based Techniques for Noise-Robust Speech Recognition, Ph.D. Dissertation, University of Cambridge (1995).

[2] Moreno, P. J. Speech Recognition in Noisy Environments, Ph.D. Dissertation, Carnegie Mellon University (1996).

[3] Ball, S. F.: Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. Acoust., Speech, Signal Process*., Vol.27, (1979), 113-120.

[4] Xu, H., Tan, Z.-H., Dalsgaard, P., Lindberg, B. Robust Speech Recognition on Noise and SNR Classification – a Multiple-Model Framework, *Proc. Interspeech* , (2005), 1123-1126.

[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] F. 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] Macho, D., Mauuary, L., Noe, B., Cheng, Y., Eahey, D., Jouvet, D., Kelleher, H., Pearce, D., Saadoun, F. Evaluation of a noise-robust DSR front-end on Aurora databases, *Proc. ICSLP*, (2002), 17-20.

[8] Juang, B. H. and Rabiner, L. R. A Probabilistic Distance Measure for Hidden Markov Models, *AT&T Technology Journal*, vol. 64, No. 2, (1984), 391-408.

Yongjoo Chung received the PhD degree in electrical engineering from Korea Advanced Science and Technology, He is currently a Professor with the Department of Electronics Engineering at Keimyung University, Daegu, S. Korea. He has published over 20 papers in international peer reviewed journals. His research interests are in the areas of speech recognition, biomedical signal processing and pattern recognition.