Noisy Speech Recognition Based on Selection of Multiple Noise Suppression Methods Using Noise GMMs

Norihide Kitaoka, Souta Hamaguchi, Seiichi Nakagawa

Department of Information and Computer Sciences,
Toyohashi University of Technology
Hibarigaoka 1–1, Tempaku-cho, Toyohashi, Aichi,441–8580 Japan
{kitaoka,hamaguchi,nakagawa}@slp.ics.tut.ac.jp

10.21437/Interspeech.2006-643

Abstract

To achieve high recognition performance for a wide variety of noise and for a wide range of signal-to-noise ratio, this paper presents integration methods of four noise reduction algorithms: spectral subtraction with smoothing of time direction, temporal domain SVD-based speech enhancement, GMM-based speech estimation and KLT-based comb-filtering. In this paper, we proposed two types of combination methods of noise suppression algorithms: selection of front-end processor and combination of results from multiple recognition processes. Recognition results on the AURORA-2J task showed the effectiveness of our proposed methods.

Intex Terms: Noisy speech recongition, noise suppression method selection, AURORA-2J

1. Introduction

In recent years, the performance of automatic speech recognition has been improved drastically by applying statistical approaches. However, most speech recognizers still have the serious problem that their recognition performance degrades in noisy environments. It is necessary to realize robust speech recognition under noisy environments for the improvement of recognition accuracy of systems. A variety of noise suppression methods have been proposed as a front-end of speech recognition. The effect of these methods greatly depends on the noise condition.

There are strong and weak points by the kind and SNR of the noise. In general, it is thought that there are no methods which can suppress various noises over SNRs in the wide range effectively. Therefore, it may be effective to select an appropriate method to each noise condition. In this paper, we propose a method to select an appropriate noise suppression method by using GMM for each speech input. We first propose a method to select a noise suppression method suitable for a certain noise condition based on GMM likelihood. The front-end processor first selects a suppression method, applies the method to input speech, and sends the feature to the back-end recognizer.

To this problem, a method for dealing with diversity of noise SNR using Multi-SNR models [7], and a hypothesis combination method which combines hypotheses generated by multiple recognition systems using feature streams obtained from multiple noise suppression methods [8] has been proposed. But they need huge computational cost. In this paper, we also propose a method to suppress the computational cost by using GMM while keeping the advantage of a hypothesis combination method.

We used AURORA-2J [5] for evaluation of our method. The AURORA-2J is a Japanese version of AURORA-2 [1], a common evaluation framework for the noisy connected English digit speech recognition task.

2. Noise suppression algorithms

In this paper, we used four-types of noise suppression methods: spectral subtraction with smoothing of time direction (SS) [2], the temporal domain SVD based speech enhancement (SVD) [3], GMM based speech estimation (GMM) [3] and pitch synchronous KLT (KLT) [4]. The SS estimates a present noise spectrum component from preceding noise spectrum information and subtract it from current observed spectrum. The SVD assumes that the speech components concentrate on the lower-order elements when singular value decomposition is applied to the input speech, so the speech can be separated from the noise. The GMM estimates the distortion of the speech by the noise for each frame in cepstral domain. The KLT projects the input speech to subspace that does not loose the feature of the speech, and eliminates the noise components. We selected these methods because their basic behaivior, for example the signal domain where the methods worked, etc., were different from each other and thus we expected that the effectiveness on the various noises were also different. The above four methods are used individually, or combined sequentially: a single method is applied to the input speech and then the same method or another method is also applied. Sequential uses are denoted as, for example, SS-GMM, etc.

3. Evaluation framework

We used AURORA-2J [5] for evaluation of our method. The AURORA-2J is a Japanese version of AURORA-2 [1], a common evaluation framework for the noisy connected English digit speech recognition task. Two training conditions (clean condition/multi-condition) and three testing sets (sets A/B/C) are defined by the AURORA-2J. Sampling rate is 8 kHz. The training data consists of 8440 utterances. The clean-condition training has acoustic models trained by clean speech only. Because a clean speech is not contaminated with the noise, the noise suppression methods are not applied to clean training data. The multi-condition training has models trained by a corpus consisting of both clean and noisy speech. In the multi-training set, speech data is contaminated with four kinds of noises (subway,babble,car,exhibition) at every SNR in five variations (clean,20dB,15dB,10dB,5dB). The noise suppression methods are applied to the multi-training data as well as the test data. The testing set A includes four different types of noise which were used in the multi-condition training, while the testing set B includes another four different types of noise not used in the multi-condition training. The testing set C then includes noise types from both sets A and B, plus additional convolutional noise. Speech is analyzed using 25 ms frames with a shift of 10ms. Each word-based HMM had 18 states and 20 Gaussian mixtures per state. The feature vectors consist of MFCC features, energy, their delta and their accelation (MFCC_E_D_A) of dimension 39.

Relative performance is defined in the AURORA-2J framework using the accuracy of the target method X_m and the accuracy

Table 1: *Result by selecting the best method (Absolute/relative, %)*

Training	A	B	C	Overall
Clean	85.49 / 72.88	83.82 / 71.13	80.91 / 61.89	83.91 / 70.11
Multicondition	92.86 / 15.77	88.28 / 40.24	90.78 / 34.94	90.61 / 33.28
Average	89.18 / 44.33	86.05 / 55.68	85.84 / 48.42	87.26 / 51.69

Table 2: *Result by GMM-KLT (Absolute/relative, %)*

Training	A	B	C	Overall
Clean	82.31 / 66.93	81.00 / 66.09	78.61 / 57.31	81.05 / 64.79
Multicondition	88.92 / -30.84	84.79 / 22.43	87.84 / 14.19	87.05 / 7.94
Average	85.61 / 18.05	82.89 / 44.26	83.23 / 35.75	84.05 / 36.36

Table 3: *The best method for each condition under multi-training.*

	Subway	Babble	Car	Exhibition
20 dB	GMM	GMM-GMM	SS-GMM	SS
15 dB	SVD	SS-GMM	SS	SS
10 dB	GMM-KLT	SS	SS-GMM	GMM-GMM
5 dB	SVD-KLT	KLT-SVD	SS-GMM	SVD-GMM

of the baseline X_b (that is, without suppression), respectively, as follows:

$$Relative\ performance = \frac{X_m - X_b}{100.0 - X_b} \times 100 \quad [\%] \qquad (1)$$

4. Potential of automatic selection of noise suppression methods

Yamada et al. [6] showed the effectiveness of the selection algorithms from various noise suppression methods and their combinations strongly depend on noise conditions.

Tables 1 show the recognition performance based on the manual selection. A, B, and C express the kind of the test sets. Table 1 shows the average absolute word accuracy and the relative performance of the manual selection in the clean training and the multi-training, and Table 2 shows those of a method (GMM-KLT), whose relative performance was the best. Comparing these performances, selecting the best method for each noise condition obtains the performance improvement than the case to apply the best single method.

5. Noise environment detection based on GMM

The speech data was contaminated with four kinds of noises by five variations of SNRs. Thus there are 20 kinds of noise conditions in the training data. The best suppression method for each noise condition is applied to all the speech under each condition. Table 3 shows the best method for each condition in multi-training data set. The suppression method applied to noisy speech is selected by using GMMs. Figure 1 shows the procedure of the GMM training. In the experiments, we used the first 10 frames of each speech file in the AURORA-2J training data as the noise data. We gathered all the noise data of the noise conditions for which a certain suppression method worked best and trained a GMM corresponding to the suppression methods using the noise data. In the recognition stage, the system compared the GMM likelihoods of the noise preceding to the speech.

6. Automatic selection of noise suppression methods for front-end processing

6.1. Speech recognition based on automatic selection of noise suppression methods

Based on the noise decision, we propose a method of selecting the best noise suppression method in the front-end. After selecting one

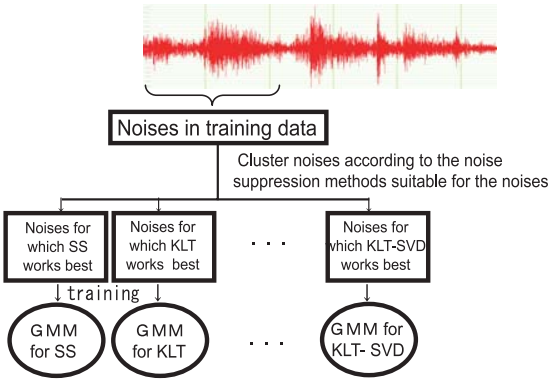

Figure 1: *Training procedure of GMMs for selecting noise suppression methods*

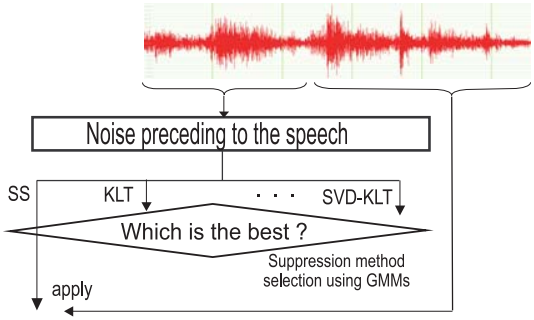

Figure 2: *Recognition procedure using automatic selection of noise suppression methods*

of the suppression methods corresponding to the GMM with the maximum likelihood. The system applies the method to the input speech and then recognize it. We used GMMs with 64 diagonal covariance matrices. The first 10 frames of each speech data were used as the noise. Each noise feature consisted of 12 dimensional MFCC and a log energy.

Figure 2 describes the procedure of the noise suppression using the selection of noise suppression methods. In this figure, SS is selected as the best method as an example.

We expect that the system selects a method for noises similar to the unknown one and the method may be effective for the noise. With this method, the back-end recognizer needs only one HMM set and does not need any special processing. Therefore, this method can be applied to distributed speech recognition.

6.2. Iterative training of acoustic model

In clean training condition, the suppression methods are applied only to the test data. Therefore, there is no modification on the acoustic models even if the front-end applies a different method to each input speech. However, the acoustic models can be retrained by using the training data compensated by various suppression methods in the multi-training condition. Retraining tends to lead the improvement of recognition performance, but the appropriate method for every noise condition may change because of the retraining. So we select the best suppression method for each noise condition and make GMM again (for each noise condition group). Then we can obtain new acoustic models from the training data applied the selected noise suppression method by the new GMMs. We iterate this procedure and stop it when all the correspondences between noise conditions and suppression methods are fixed.

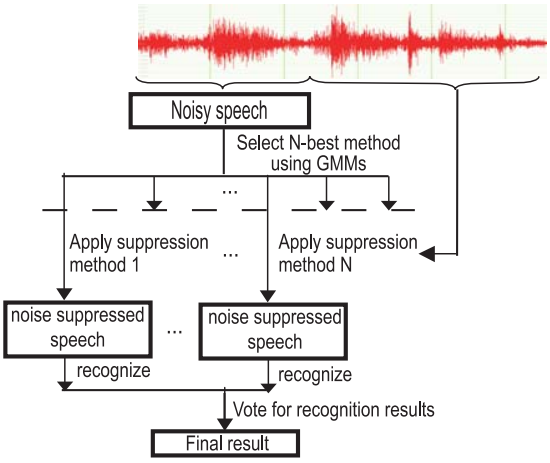

Figure 3: *Voting procedure using GMMs*

7. Integration of recognition results —Integration in Backend—

The integration of the suppression methods in the front-end obtains the accuracy improvement to some degree without increasing computational cost on the back-end processing. On the other hand, the integration of the noise suppression methods in the back-end has been proposed [8]. The integration is done by voting. The recognizer corresponding to each noise suppression method votes for the hypothesis obtained by the recognizer and the hypothesis which gets majority vote is selected as a final result. This method showed the significant improvement of the accuracy. However, a huge computational cost was needed. So, we investigate the method to improve the recognition accuracy with less computational cost using the GMM-based selection of noise suppression method.

To reduce computational cost of voting method, the system first selects some effective suppression methods. The selected methods are performed in parallel and then vote for the results. In this strategy, GMMs are used as the case with the method in the front-end. Figure 3 shows the procedure of the voting algorithm by using GMMs. The training procedure of GMM is similar to Section 5. The noise feature is inputted to each GMM, the likelihood of 21 suppression methods is obtained, and the N-best noise suppression methods are selected. Then, recognition procedures using selected noise suppression methods are performed in parallel. The hypotheses obtained from these procedures are voted, and the hypothesis with maximum vote is adopted as the final result. When the number of votes is the same for plural hypotheses, the hypothesis generated by the method with the highest likelihood of noise-GMM is adopted. Moreover, because there are differences among the effects of the suppression methods, it is natural to assign priorities to the methods according to the noise conditions. Therefore, we use a weighted voting method based on the likelihood (or priority) of GMMs.

8. Experiment

8.1. Front-end processing results

We evaluated the method described in Section 4 on the AURORA-2J. Whole the noise suppression procedure is done in the front-end, so all methods are categorized as *category 0* [1].

Under the clean training condition, we evaluated three noise suppression methods: GMM-KLT, which was the best single (sequential) suppression method among all under clean training condition, the proposed method, and the manual selection of the best suppression method for each noise condition (ideal result). The selection accuracy of noise suppression methods by GMMs was about 54% both under clean and multi-conditions. Figure 4 shows the results in word accuracy and improvement relative to the baseline.

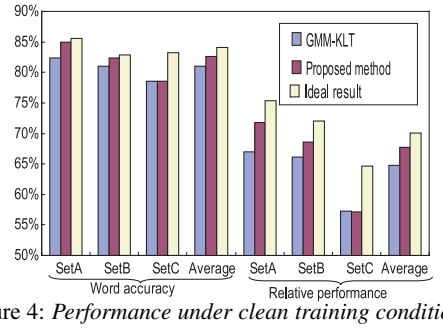

Figure 4: *Performance under clean training condition (%)*

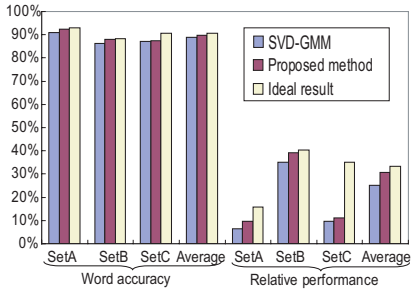

Figure 5: *Recognition performance of multi-condition training (%)*

The proposed method obtained the relative performance improvement of 67.7% as compared to the baseline, which was significantly better than "GMM-KLT" (64.7%). That is to say, we could obtain better performance with the proposed method than all the individual methods included in the selection of the proposed method. The improvement of the recognition accuracy of test set B (speech contaminated with unknown noises) is not so inferior to the improvement of the recognition accuracy of test set A (with known noise). This proved that our proposed method could suppress not only known but also unknown noises robustly.

Under the multi-training condition, we evaluated three noise suppression methods: SVD-GMM, which was the best single (sequential) suppression method among all under multi-training condition, the proposed method, and the manual selection of the best suppression method for each noise condition (ideal result). We used the HMMs trained from the speech applied with "SVD-GMM", which is the best combination method for multi-condition training among the 21 methods. We used the GMM obtained by the training method described in Section 6.2. After the fourth iteration, we obtained the absolute word accuracy improvement of 0.2 %. The word accuracy was 85.93% when the noise is not suppressed (baseline). Figure 5 shows the recognition results for the SVD-GMM, the proposed method, and the ideal method. The proposed method obtained the relative performance improvement of 30.5% as compared to the baseline. Compared with the SVD-GMM, the improvement of relative improvement was 5.4% from "SVD-GMM"(30.5% from 25.1%). So, the proposed method could obtain better relative performance than all the individual method. This method worked well even for unknown noises from the result on test set B.

8.2. Integration in back-end processing

We evaluated the integration method in the back-end. In this method, we modified the back-end processing and thus this method is categorized as *category 5* [1]. The advantage of this method is to be able to use noise suppression method-dependent HMMs, so we evaluated this method under multi-training condition.

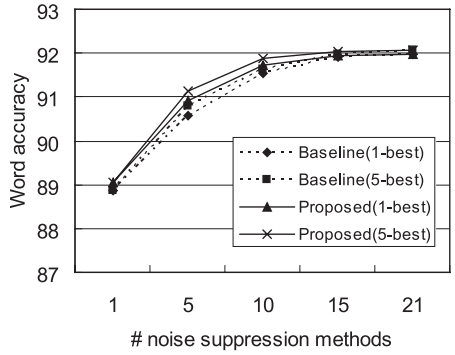

Figure 6: *Recognition accuracy for voting method with N noise suppression methods(%). Comparison between the method with fixed N methods and the method with dynamically selected N method.*

Table 4: *Comparison of proposed methods and baseline.*

	Word accuracy	String accuracy
Fixed 5 methods	90.79 %	81.23 %
Dynamically selected 5 methods	91.15 %	82.41 %
Voting without weight (21 methods)	91.97 %	84.38 %
Weighted voting by GMM (21 methods)	92.20 %	84.60 %

In our method, the noise suppression methods were dynamically selected on the fly. For comparison, we also conducted the voting by fixed N methods. These N methods were selected *a priori* by overall recognition performance on the training data. We conducted the recognition experiment by the voting method with N noise suppression methods with N=1,5,10,15 and 21 (used all suppression methods) on the multi-condition training. We could use multiple hypotheses for voting. So we used the 5-best hypotheses per noise suppression method. Figure 6 shows the results. In Figure 6 'baseline' describes the method with the fixed N noise suppression methods and 'proposed' describes the method with dynamically selected N noise suppression methods. The recognition accuracy of the proposed method was higher than that of baseline. Because all methods were used, the recognition accuracy was the same when using N=21 for the both voting methods. When using N=1,'baseline' was the best single method, and 'proposed' selected a suitable method for every noise condition by using GMM. We found the absolute improvement of 0.34% (2.4% relative) when using N=5.

All the accuracy was slightly improved using 5-best hypotheses for voting and we observed all most same tendency as was in the case of 1-best.

Table 4 shows the results in word accuracy and string accuracy. We tested the improvement of the method with dynamically selected 5 methods from the fixed 5 methods in string accuracy using sign test and proved that there was a significant improvement with the significance level of 1%.

We also evaluated the weighted voting method. We used 1.5 and 0.5 as the weights for the 1/3 of suppression methods with high likelihoods of noise GMMs and for the 1/3 with low likelihoods, respectively. Results are shown in Table 4, and we proved that a significant improvement was achieved with the weighted voting method with the significance level of 1% by a sign test [9]. We obtained the word accuracy improvement of 0.23% (the relative performance improvement of 1.63%) and the string accuracy improvement of 0.22% by the voting with weight.

9. Conclusion

We proposed an automatic selection of noise suppression method using GMM corresponding to each noise suppression method. We also proposed an iterative training of HMMs and GMMs for multi-conditional training. We first proposed to apply the method selection to the front-end processing. We evaluated the proposed method using AURORA-2J Japanese noisy connected digit speech recognition task and obtained better recognition performance than all the individual methods in both clean and multi training. Then, we proposed the integration method in which noise suppression methods were dynamically selected using GMM in back-end. We found the absolute improvement of 0.36% as compared to the method with fixed N noise suppression methods when using N=5 with 5-best hypotheses per suppression methods.

We proved that our method could manage multiple noise suppression methods efficiently to complement each other. Our method, of cause, can adopt other suppression methods to achieve further improvement.

10. Acknowledgment

The authors thank to Prof. Kazuya Takeda of Nagoya University and Dr. Masakiyo Fujimoto of ATR for their provisions of noise suppression softwares. We also thank to Dr. Takeshi Yamada of University of Tsukuba for his help to this study. The presented study was conducted using AURORA-2J database developed by IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group.

11. References

[1] H. G. Hirsh, D. Pearce, "The AURORA experimental frame work for the performance evaluations of speech recognition systems under noisy conditions," ISCA ITRW ASR2000, 2000

[2] N. Kitaoka, S. Nakagawa, "Evaluation of spectral subtraction with smoothing of time direction on the AURORA2 task," Proc. ICSLP2002, pp. 465-468, 2002.

[3] M. Fujimoto, Y. Ariki, "Combination of temporal domain SVD based speech enhancement and GMM based speech estimation for ASR in noise-evaluation on the AURORA2 task," Proc. Eurospeech2003, 2003

[4] M. Ikeda, K. Takeda, F. Itakura, "Speech enhancement by quadratic comb-filtering," Technical Report of IEICE, SP96-45, pp. 23-30, 1996

[5] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto, T. Endo, "AURORA-2J: An evaluation framework for Japanese noisy speech recognition," IEICE Trans. Inf. & Syst., Vol.E88-D,No.3,pp.535-544,2005

[6] T. Yamada, J. Okada, K. Takeda, N. Kitaoka, M. Fujimoto, S. Kuroiwa, K. Yamamoto, T. Nishiura, M. Mizumachi, S. Nakamura, "Integration of noise reduction algorithms for AURORA2 Task," Eurospeech 2003, pp. 1769-1772, 2003

[7] M. Ida, S. Nakamura, "HMM composition-based rapid model adaptation using a priori noise GMM adaptation evaluation on AURORA2 Corpus," Proc. ICSLP2002, pp. 437-440, 2002.

[8] J. Okada, T. Yamada, N. Kitawaki, "Integration of recognition results from multiple noise reduction algorithms," The 2004 spring meeting of the acoustical society of Japan , pp. 157-158, 2004 (in Japanese)"

[9] S. Nakagawa, *Pattern Information Processing*, Maruzen Ltd, 1999 (in Japanese)