Noisy Speech Recognition Based on Noise-Adapted HMMs Using Speech Feature Compensation

Yong-joo Chung*

ABSTRACT

The vector Taylor series (VTS) based method usually employs clean speech Hidden Markov Models (HMMs) when compensating speech feature vectors or adapting the parameters of trained HMMs. It is well-known that noisy speech HMMs trained by the Multi-condition TRaining (MTR) and the Multi-Model-based Speech Recognition framework (MMSR) method perform better than the clean speech HMM in noisy speech recognition. In this paper, we propose a method to use the noise-adapted HMMs in the VTS-based speech feature compensation method. We derived a novel mathematical relation between the train and the test noisy speech feature vector in the log-spectrum domain and the VTS is used to estimate the statistics of the test noisy speech. An iterative EM algorithm is used to estimate train noisy speech from the test noisy speech along with noise parameters. The proposed method was applied to the noise-adapted HMMs trained by the MTR and MMSR and could reduce the relative word error rate significantly in the noisy speech recognition experiments on the Aurora 2 database.

Keywords : noisy speech recognition, MTR, Expectation-Maximization, VTS

I. Introduction

Despite many technical advances, accurate speech recognition in noisy environments still remains a difficult problem. The techniques cannot fully overcome the performance degradation caused by channel and additive noise. Broadly categorized, there are two different approaches used to improve the performance in noisy speech recognition. In one of the approaches, noise-robust feature extraction, speech enhancement, feature and model parameter compensation approaches are used independently or in combination with each other to improve the performance of speech recognition under noisy environments[1],[2],[3]. In particular, compensation based on the Vector Taylor Series (VTS) has been known to perform quite well in noisy conditions [4],[5].

In another approach, noisy speech was directly used to produce noise-adapted hidden Markov Models (HMMs) during training. The Multi-condition TRaining (MTR) [6]and Multi-Model-based Speech Recognition framework (MMSR) [7],[8] are representatives of this approach.

* Department of Electronics Engineering, Keimyung University, Daegu, S. Korea.
투고 일자 : 2014. 3. 8. 수정완료일자: 2014. 4. 28.
게재확정일자 : 2014. 5. 2.

The environment-dependent HMMs make it possible to cope with test noisy speech without any compensation algorithm. In the MTR method, noisy speech signals under various noise conditions are collected and used for training the HMM. The MMSR was recently proposed to improve the sharpness of probability density functions in acoustic models of the MTR, and successful results using the MMSR were demonstrated [7],[8],[9]. In contrast to the MTR, where a single HMM set is constructed, multiple HMM sets corresponding to various noise types and signal-to-noise ratio (SNR) values are produced during training, and a single HMM set which is closest to test noisy speech among multiple HMM sets is selected for recognition.

Although the noise-adapted HMM performs rather well by itself, its performance would be improved further by applying compensation. In a previous study, a novel mathematical relation between test and training noisy speech was derived in the log-spectrum domain [9]. After approximating the relation using the VTS, the performance of the noise-adapted HMM could be improved by compensating the feature vectors of the test noisy speech. The Minimum Mean Square Error (MMSE) estimation of training noisy speech (not clean speech) conditioned on the test noisy speech was used for recognition instead of the test noisy speech, which could

reduce the mismatch between the test noisy speech and the acoustic models of the noise-adapted HMM. However, in the previous study, the channel noise was not considered in the compensation, which probably had a negative effect on improving the performance on Set C of the Aurora 2 database. In this study, the previous algorithm was modified to compensate the test noisy speech considering both the channel and additive noise. The detailed mathematical formulation is derived, and the MTR as well as the MMSR are used for producing the noise-adapted HMM.

This paper is organized as follows. A review on the MTR and the MMSR is presented in Section 2, and compensation of the test noisy speech based on the noise-adapted HMM is described in Section 3. The experimental procedure and results are presented and discussed in Section 4. Finally, conclusions are given in Section 5.

Ⅱ. A Review on Noise Adapted HMMs

In this study, both the MTR and MMSR are used to produce the noise-adapted HMM. Although the MMSR is known to have advantages over the MTR method [7],[8], it is rather controversial regarding which method is better in performance for noisy speech recognition. Both will be used to find the more appropriate method in the proposed feature-compensation method.

In the MTR, a collection of clean and noisy speech signals with various noise types (Subway, Babble, Car, Exhibition) and SNR values (0, 5, 10, 15, 20 dB) is used to construct a single set of noise-adapted HMM. In the MMSR, multiple HMM sets are constructed, and each of them corresponds to a different noise type (Subway, Babble, Car, Exhibition) and SNR value (0-30 dB in 2-dB intervals). A single HMM set which is closest to the test noisy speech is selected for recognition based on the estimated SNR value and noise type of the test speech. Since the MTR method combines a number of noise conditions to train a single HMM set, it tends to reduce the phonetic sharpness of the acoustic models in their probability density functions of the HMM. The MMSR method can overcome the weakness of the MTR by choosing a specific single HMM set which is most appropriate to the test noisy speech. However, the errors in selecting the closest HMM set will incur misrecognition, causing performance degradation in the MMSR.

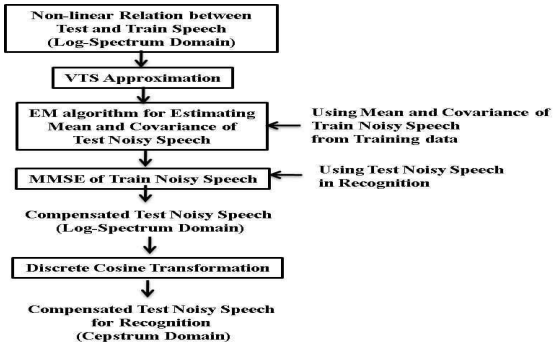

Fig .1. Block diagram of the proposed feature compensation method.

Ⅲ. Feature Compensation

For the test noisy speech feature compensation based on noise-adapted HMM, the relation between training and test noisy speech is first derived in log-spectrum domain. Since the relation is non-linear, it is approximated using the VTS to obtain the mean vectors and covariance matrices of the test noisy speech given the statistics of training noisy speech obtained during the training. The statistics of the test noisy speech are used to estimate MMSE of the training noisy speech, which is used as a feature vector for recognition after Discrete Cosine Transformation (DCT). The block diagram of the whole process is shown in Fig. 1. A more detailed explanation of this process is given in the next subsections.

A. Relation between Test and Train Noisy Speech

Log-spectrum vector of the clean speech and y of the noisy speech are usually assumed to be related as follows:

$$y = x + h + \log(i + \exp(n - x - h)) \tag{1}$$

where n and h are the log-spectrum vector of additive and convolution noise, respectively, and i is a unity vector. Based on (1), the log-spectrum vector of the test noisy speech y and the training noisy speech y can be expressed as follows, assuming that there is no channel noise in the training noisy speech for the convenience of analysis:

$$y_{Tr} = x + g_0(x, n_{Tr}) \tag{2}$$

$$y = x + h + g(x, n, h) \tag{3}$$

$$g_0(x, n_{Tr}) = \log(i + \exp(n_{Tr} - x)) \tag{4}$$

$$g(x, n, h) = \log(i + \exp(n - x - h)) \tag{5}$$

where and n represent the additive noise contained in the test and training noisy speech, respectively. n_{Tr} should be determined during training, and n is estimated using test noisy speech in recognition.

By combining (2) and (3), the test noisy speech can be expressed in terms of the training noisy speech as follows:

$$y = y_{Tr} + h + g(x, n, h) - g(x, n_{Tr}) \tag{6}$$

$$y = y_{Tr} + h + \quad (n, h, y_{Tr}, n_{Tr})$$

$$y = y_{Tr} + h + \log(i + \exp(n - h - y_{Tr}) - \exp(n_{Tr} - y_{Tr}))$$

B. Statistics of Test Noisy Speech

From (6), the mean and covariance of the test noisy speech can be estimated. Equation (6) is expanded using the first-order VTS around the initial value n_0, h_0 of n, h and the mean of the training noisy speech $\mu_y = E\{y_{Tr}\}$ to obtain the following equation.

$$\begin{aligned} y = y_{Tr} + h + G(n_0, h_0, \mu_{y_{tr}}, n_{Tr}) + \\ \nabla_{y_{Tr}} G(n_0, h_0, \mu_{y_{tr}}, n_{Tr})(y_{Tr} - \mu_{y_{Tr}}) + \\ \nabla_n G(n_0, h_0, \mu_{y_{tr}}, n_{Tr})(n - n_0) + \\ \nabla_h G(n_0, h_0, \mu_{y_{tr}}, n_{Tr})(h - h_0) \end{aligned} \tag{7}$$

Using (7), the mean μ_y and covariance Σ_y of the test noisy speech can be expressed from the mean $\mu_{y_{Tr}}$ and covariance $\Sigma_{y_{Tr}}$ of the training noisy speech as follows:

$$\begin{aligned} \mu_y = \mu_{y_{Tr}} + h + G(n_0, h_0, \mu_{y_{Tr}}, n_{Tr}) + \\ \nabla_n G(n_0, h_0, \mu_{y_{Tr}}, n_{Tr})(n - n_0) + \\ \nabla_h G(n_0, h_0, \mu_{y_{Tr}}, n_{Tr})(h - h_0) \end{aligned} \tag{8}$$

$$\begin{aligned} \Sigma_y = \ I + \nabla_{y_{Tr}} G(n_0, h_0, \mu_{y_{Tr}}, n_{Tr})\big)\Sigma_{y_{Tr}} \\ \bullet \ \big(I + \nabla_{y_{Tr}} G(n_0, h_0, \mu_{y_{Tr}}, n_{Tr})\big)^T + \\ \nabla_n G(n_0, h_0, \mu_{y_{Tr}}, n_{Tr}) \Sigma_n G(n_0, h_0, \mu_{y_{Tr}}, n_{Tr})^T \end{aligned} \tag{9}$$

C. Estimation of Noise Parameter

Assuming also that the log-spectrum vector y of the test noisy speech is a mixture of Gaussian distributions, the distribution of y as a function of unknown noise vector n , h can be defined using (8) and (9),

$$p(y|n, h) = \sum_{m=1}^{M} p_m N(\mu_{y,m}, m, \Sigma_{y,m}) \tag{10}$$

where $N(\mu_{y,m}, m, \Sigma_{y,m})$ is the m-th Gaussian distribution with a mean vector $\mu_{y,m}$ and covariance matrix $\Sigma_{y,m}$. p_m is the mixture weight of the m-th component. Note that the mean vector $\mu_{y,m}$ and covariance matrix $\Sigma_{y,m}$ are themselves fully parameterized by the noise vectors n and h, which are treated just as parameters, not random variables; only the noisy speech vectors were treated as random variables.

Given a sequence of log-spectrum vectors for the test noisy speech, the log-likelihood for the sequence is defined as follows using (10):

$$L(Y|n, h) = \sum_{t=1}^{T} \log p(y_t|n, h) \tag{11}$$

An iterative Expectation Maximization (EM) algorithm is used to re-estimate the noise vector maximizing (11). In the EM algorithm, an auxiliary function $Q(\phi|\phi)$ is written as follows:

$$\begin{aligned} Q(\phi|\bar{\phi}) = E\{L(Y|\bar{\phi})|Y, \phi\} = \\ \sum_{t=1}^{T}\sum_{m=1}^{T} p(m|y_t, n, h) \log p(y_t, m|\bar{n}, \bar{h}) \end{aligned} \tag{12}$$

The symbol ϕ represents the noise vector n,h, which is already known and $\bar{\phi}$ is the unknown noise vector $\bar{n},\bar{h}$, which should be estimated. To re-estimate n,h in (12), the derivative of the auxiliary function with respect to $\bar{n},\bar{h}$ must be taken and set equal to 0.

The noise vector $\bar{n},\bar{h}$ derived is substituted into n,h in (8) and (9) to adapt the mean and covariance of the test noisy speech. The likelihood function from (11) and the auxiliary function from (12) are consequently updated. This process is iterated until the log-likelihood function from (11) converges. After the convergence, an MMSE estimation of the training noisy speech is performed and used for recognition.

D. MMSE of Training Noisy Speech

The MMSE of training speech y_{Tr} given the test speech y is expressed as follows:

$$\hat{y}_{Tr,\,MMSE} = E(y_{Tr}|y) = \int y_{Tr} p(y_{Tr}|y) dy_{Tr} \tag{13}$$

From (6),

$$y_{Tr} = y - h - G(n, h, y_{Tr}, n_{Tr}) \tag{14}$$

Substituting (14) into (13) and approximating $G(n, h, y_{Tr}, n_{Tr})$ by the VTS of order zero around $\mu_{y_{Tr},m}$, the following relationship is obtained:

C1	C2	⋯	C12	E	△C1	△C2	⋯	△C12	△E	△ C1	△²C2	⋯	△²C12	△²E

Fig .2. The arrangement of the feature vector used in the experiments.
(: i-th cepstrum coefficient, E: log-energy, $\triangle Ci$,$\triangle E$: delta coefficient, $\triangle^2 Ci$,$\triangle^2 E$: acceleration coefficient)

$$y_{r,MMSE} \cong y - h - \sum_{m=1}^{M} p(m|y)\, G(n, h, \mu_{y,m}, n_{Tr}) \quad (15)$$

The DCT of the log-spectrum vector $\hat{y}_{Tr,MMSE}$ is taken to find a 13-th order cepstrum vector. The 0-th component in the cepstrum vector is replaced with log-energy. The delta and acceleration(delta-delta) coefficients of the cepstrum vector are also calculated to obtain a 39-dimentional feature vector which is used for the speech recognition experiments described in the next section.

IV. Experimental Results

To verify the effectiveness of the proposed feature compensation method, experiments were conducted on the Aurora 2 database. There are two sets of training data, each corresponding to clean training (CLEAN) and multi-condition training (MTR). Each consists of 8,440 sentences of approximately 3~5 s in duration. The MTR set consists of both clean and noisy speech signal that is artificially contaminated by various kinds (subway, car, exhibition, babble) of noise with SNR ranges from 0 to 20 dB in 5-dB intervals.

Recognition experiments were conducted on 3 test sets (sets A, B, C) that are corrupted by a range of noise types with a SNR range of 0, 5, 10, 15, 20 dB. For each noise type and SNR value, there are 1,001 sentences for recognition. Set A and set B are corrupted by an additive noise distortion alone, and set C is corrupted by a combination of convolution noise and additive noise.

For the feature vector, a noise-robust version of Mel-Frequency Cepstral Coefficients (MFCCs) called AFE (Advanced Front-End) was used. AFE is known to significantly reduce the word error rates in noisy speech recognition [10]. The 12-th order MFCCs with the 0-th order cepstral coefficient set aside are appended with the log-energy to form a 13-th order basic feature vector along with their delta and acceleration coefficients to construct a 39-th order feature vector for each frame. The feature vector for each frame is arranged as in Fig. 2.The acoustic models were trained using both the Complex Back End (CBE) and Simple Back End (SBE) scripts, which are each separately defined for the Aurora 2 database. For the SBE model, the HMM for each digit consists of 16 states with 3 Gaussian mixtures in each state. In addition, a three-state silence model with 6 Gaussian mixtures per state and a one-state pause model tied with the center state of the silence model are used. For the CBE, the number of mixtures in each state is increased to 20 and 36 for the digit and silence models, respectively. The hidden Markov model toolkit (HTK) was employed to train and test the HMM used in this study [11].

Table 1 shows the word error rates (WERs) of the proposed feature compensation methods (MTR-MMSE/MMSR-MMSE) in comparison with the conventional methods for the Aurora 2 database. The MTR-MMSE and the MMSR-MMSE differ in the type of noise-adapted HMM used for recognition. The average WER (Ave.) in the last column is calculated by summing the weighted WERs for Set A, Set B and Set C. The weighting factors are 0.4 for Set A and Set B and 0.2 for Set C since the number of test utterances for Set A and Set B is twice as many as that in Set C.

As expected, both the MTR and MMSR method improve the performance of the baseline system, which was trained using clean speech data. The baseline system scores 12.97% WER on average, whereas the MTR and MMSR achieve WERs of 8.22 % and 8.17%, respectively. Although the MMSR performs slightly better than the MTR, their difference is not so significant. The VTS method based on the clean speech HMM improves the performance of the baseline system but is quite inferior to the MTR and MMSR. This demonstrates the superiority of the multi-style training approaches.

By using the proposed feature compensation, the performance of the MTR and the MMSR methods could be improved further. As shown in Table 1, the MTR-MMSE and the MMSR-MMSE achieve 7.81% and 7.80% average WERs, providing 4.98% and 4.52% relative word error rate improvement over the MTR and the MMSR, respectively. The relative word error rate is computed by dividing the WER difference of the two methods with the WER of the reference method.

The MTR-MMSE and the MMSR-MMSE were also applied to the noise-adapted HMM trained with the CBE script to verify whether the proposed method could work as in the SBE script. We could observe similar performance trend as in the SBE script. The result is shown in Table 2 in comparison with other conventional approaches.

Table 1. WER (%) of MTR-MMSE/MMSR-MMSE using SBE models compared to conventional methods for Aurora 2 database.

Method	Set A	Set B	Set C	Ave.
Baseline	12.25	12.90	14.56	12.97
VTS	12.01	12.37	13.87	12.52
MTR	7.70	8.23	9.26	8.22
MMSR	6.78	9.56	8.17	8.17
MTR-MMSE	7.54	7.75	8.45	7.81
MMSR-MMSE	6.71	8.98	7.60	7.80

Compared with the results in Table 1, consistent performance improvement can be observed with the CBE script, and it is most prominent in the MTR. The increased number of mixtures in each state of the HMM may have greatly contributed to sharpening the acoustic modeling in the MTR. Although the MMSR had comparable performance with the MTR in the SBE script, the MTR significantly outperforms the MMSR in the CBE script.

Table 2. WERs (%) of MTR-MMSE/MMSR-MMSE using CBE models compared to conventional methods for Aurora 2 database.

Method	Set A	Set B	Set C	Ave.
Baseline	11.58	12.10	13.68	12.20
VTS	11.42	11.49	12.83	11.73
MTR	6.04	6.82	7.22	6.59
MMSR	6.17	9.0	7.97	7.66
MTR-MMSE	5.9	6.33	6.37	6.16
MMSR-MMSE	5.86	8.17	7.51	7.11

V. Conclusions

In this study, we proposed a VTS-based feature compensation method using noise-adapted HMMs. The approach is distinguished from the conventional VTS-based methods where the clean speech HMM is used instead of the noisy speech HMM. The MTR and MMSR were used to train the noise-adapted HMM, and the speech recognition performance could be significantly improved by employing the proposed feature compensation. The proposed algorithm was applied to HMMs trained with the CBE script as well as the SBE script to test the robustness of the method and we could find improved performance in both of them. The best result (6.16% average WER) was obtained when the feature compensation was applied to the MTR in the CBE script, resulting in 6.5% relative improvement in WER over the conventional MTR method.

References

[1] S.F Ball, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.* Vol. 27, No. 2, pp. 113-120, 1979.

[2] M.J.F Gales, *Model based techniques for noise-robust speech recognition*, Ph.D. Dissertation, University of Cambridge, 1996.

[3] W Kim, J.H.L Hansen, "Feature compensation in the cepstral domain employing model combination," *Speech Communication,* Vol. 51, No. 2, pp. 83-96, 2009.

[4] P.J Moreno, *Speech Recognition in noisy environments*, Ph.D. Dissertation, Carnegie Mellon University, 1996.

[5] D.Y Kim, C.K Un, N.S Kim, "Speech recognition in noisy environments using first-order vector Taylor series." *Speech Communication,* Vol. 24, No. 1, pp. 39-49, 1998.

[6] H.G Hirsch, D Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions." in *Proceedings of the International Conference on Spoken Language Processing,* Bejing, China, pp. 18-20, 2000.

[7] H Xu, Z.H Tan, P Dalsgaard, B Lindberg, "Robust speech recognition on noise and SNR classification - a multiple-model framework." in *Proceedings of INTERSPEECH,* Lisboa, Portugal, pp. 977-980, 2005.

[8] H Xu, X.H Tan, P Dalsgaard, B Lindberg, "Noise condition dependent training based on noise classification and SNR estimation." *IEEE Trans. Audio, Speech, Language Process.* Vol. 15, No. 8, pp. 2431-2443, 2007.

[9] Y. Chung and J.H.L. Hansen, "Compensation of SNR and noise type mismatch using an environmental sniffing based speech recognition solution," *EURASIP Journal on Audio, Speech, and Music Processing,* 2013:12, (2013), pp. 1-14, 2013.

[10] *ETSI draft standard doc., Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm.* ETSI Standard ES 202 050, 2002

[11] S Young, *HTK: Hidden Markov Model Toolkit V3.4.1.* Cambridge Univ. Eng. Dept. Speech Group, 1993.

Yong-joo Chung (Member) 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 30 papers in international peer reviewed journals. His research interests are in the areas of speech recognition, multimedia signal processing and pattern recognition.