Contemporary Engineering Sciences, Vol. 7, 2014, no. 22, 1121 - 1129
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2014.49140

Noise Estimation Employing Variational Model Composition for Speech Enhancement in Time-Varying Noise Conditions

Sunmee Kang

Department of Electronic Engineering
Seokyeong University, Seoul, Korea

Wooil Kim*

School of Computer Science & Engineering
Incheon National University, Incheon, Korea
*Corresponding Author

Abstract

This paper proposes an effective noise estimation method for speech enhancement to improve speech recognition in time-varying noise conditions. The proposed noise estimation scheme employs the Variation Model Composition (VMC) method. The VMC method generates multiple noise models by selectively applying perturbation factors to the mean parameters of a basis noise model. The resulting collection of the noise models is expected to effectively reflect the unseen noise signal included in the speech segments. The obtained noise models are used to generate multiple environmental models employing the parallel model combination method. The noise estimate is obtained by using the posterior probability of the multiple environmental models. The proposed noise estimation method is applied to the Spectral Subtraction. The proposed speech enhancement scheme is evaluated within the Aurora 2.0 evaluation framework over speech babble and background music noise conditions. Experimental results demonstrate that the proposed method is effective at increasing speech recognition accuracy in time-varying background noise conditions.

Keywords: Variational Model Composition, Noise Estimation, Spectral Subtraction, Speech Recognition, Time-Varying Noise

1 Introduction

Acoustic mismatch between training and operating conditions of an actual speech recognition system is one of the primary factors severely degrading recognition performance. To minimize this mismatch, extensive research has been conducted in recent decades including many types of speech/feature enhancement methods such as Spectral Subtraction, Cepstral Mean Normalization, and a variety of feature compensation schemes [1]-[3]. However, the conventional methods continue to suffer from ineffectiveness in time-varying background noise conditions, where the noise characteristics need to be effectively estimated as time elapses.

In this study, a novel noise estimation method for speech enhancement is proposed to address time-varying background noise for improved speech recognition. Here our previous study of the Variational Model Composition (VMC) method [4][5] is employed for noise estimation. The motivation of the VMC is that each order of the cepstral coefficients represents the frequency degree of the changing components in log-spectrum envelope [6]. In the VMC method, variational noise models are generated by selectively applying perturbation factors to a basis model in the cepstral domain in order to obtain various types of spectral patterns. The variational model composition method showed the effectiveness by being employed to generate multiple environmental models for our feature compensation method [5]. In this study, the posterior probability of each environmental model is used for estimating time-varying background noise. The proposed method will be evaluated on two types of time-varying background noise conditions including speech babble and background music within the Aurora 2.0 evaluation framework [7].

2 Variational Model Composition

In the VMC (Variational Model Composition) method [4][5], it is assumed that (i) a basis noise model can be obtained from periods of silence within the speech stream, and (ii) the target time-varying noise included in the speech duration would reflect variations of the estimated basis model. The variational models are generated by selectively applying weights on each component of the mean vector of the basis model in the cepstral domain.

First, a basis noise model is obtained from non-speech segments within the input speech, which generally exists at the beginning and end parts of an utterance. The model is estimated as a Gaussian pdf (μ, Σ) in the cepstral domain. In

general the variance Σ is estimated as a form of diagonal matrix, resulting in a vector σ^2. Next, the *V* largest components $\{v_1, v_2, \dots, v_V\}$ in the variance vector σ^2 are selected. They are named Variational Components, which are considered highly variable components in a size-ordered rank.

Finally, a variation of the mean vector is generated by selectively applying the perturbation factor f_p on the determined variational components of the cepstral coefficients v_1 to v_V as follows,

$$\tilde{\mu}_i = \begin{cases} \mu_i(1+f_p), & if\ i \in \{v_1, v_2, \dots, v_V\} \\ \mu_i, & otherwise, \end{cases} \quad (1)$$

where $f_p = 0, -\alpha$ or $+\alpha$ and the α is a small positive value which we determine heuristically. The obtained model collection $\{\tilde{\lambda}_e = (\tilde{\mu}_e, \Sigma)\}$ consists of a total 3^V number of generated variational models as a result of combinations of the 3-type gains (i.e., $0, -\alpha$ or $+\alpha$) of the *V* variational components.

3 Noise Estimation Employing Variational Model Composition Method

In this section, a novel noise estimation method is proposed, which employs the Variation Model Composition presented in Sec. 2. In our previous study, Parallel Combined Gaussian Mixture Model (PCGMM) based feature compensation method was proposed, showing robust speech recognition performance in various types of background noise conditions [8]. A series of experiments in that study confirmed that the noise corrupted general speech model (i.e., Gaussian mixture model) employed by the PCGMM method effectively represents the input noise corrupted speech. Based on this motivation, we integrate the PCGMM-based model estimation method for obtaining the speech model into our noise estimation method in this study.

3.1 Speech Model Estimation

The distribution of the clean speech feature **x** in the cepstral domain is represented with a Gaussian Mixture Model consisting of *K* components as follows:

$$\mathrm{p(x)} = \sum_{k=1}^{K} \omega_k \mathrm{N}\left(x; \mu_{x,k}, \Sigma_{x,k}\right). \quad (2)$$

In this study we have multiple noise models obtained by the VMC method presented in Sec. 2. Therefore multiple noise-corrupted speech models are generated through a model combination procedure using the clean speech model

and each noise model $(\tilde{\mu}_e, \Sigma)$ of the variational model collections.

$$\left(\mu_{y,e,k}, \Sigma_{y,e,k}\right) = \mathcal{F}\left[\left(\mu_{x,k}, \Sigma_{x,k}\right), (\tilde{\mu}_e, \Sigma)\right], \tag{3}$$

where $\mathcal{F}[\cdot]$ denotes a function representing the model combination. For the model combination of the PCGMM method, we employ "log-normal approximation" method, where it is assumed that the addition of two log-normal distributions also results in a log-normal formulation [8][9]. Finally, the resulting GMMs of the noise-corrupted speech are represented in the cepstral domain as follows,

$$\mathrm{p}(\mathrm{y}|G_e) = \sum_{k=1}^{K} \omega_k \mathrm{N}\left(x; \mu_{y,e,k}, \Sigma_{y,e,k}\right). \tag{4}$$

3.2 Noise Estimation

The multiple noise models obtained by the Variational Model Composition method are used to generate the multiple environmental models $\{G_e\}$. They are estimated through the model combination procedure using the clean speech GMM and the obtained variational noise models as described in Sec. 3.1. With *V* number of variational components, $3^V (= E)$ environmental models are generated.

The utilization of multiple environmental models is considered to be effective for compensating input features adaptively under time-varying noisy conditions [8]. In the multiple model method, a sequential posterior probability of each possible environment is estimated over the incoming noisy speech. Given the input noisy speech feature vectors $Y_t = [y_{t-d+1}, y_{t-d+2}, \dots, y_t]^T$ over a *d* interval, the sequential posterior probability of a specific environment GMM G_e among all models can be written as,

$$\mathrm{p}(G_e|Y_t) = \frac{P(G_e)p(Y_{t-1}|G_e)p(y_t|G_e)}{\sum_{e'=1}^{E} P(G_{e'})p(Y_{t-1}|G_{e'})p(y_t|G_{e'})}, \tag{5}$$

where $\mathrm{p}(Y_{t-1}|G_e) = \prod_{T=t-d+1}^{t-1} p(y_T|G_e)$ and $\mathrm{P}(G_e)$ is a prior probability of each environment G_e represented as a GMM.

Based on Eq. (5), the noise signal in the cepstral domain at frame t is estimated by the weighted combination of the mean parameters of the variational noise models obtained from a set of *E* multiple environments using the posterior probability as follows,

$$\tilde{n}_t = \sum_{e=1}^{E} p(G_e|Y_t)\, \tilde{\mu}_e \quad . \tag{6}$$

We believe that the noise estimate $\tilde{n}_t$ obtained by Eq. (6) would represents the

change of the unseen noise signal during the speech segments in the time-varying background noise.

3.3 Spectral Subtraction with Noise Estimate

The obtained noise estimate $\tilde{n}_t$ is the cepstral domain, therefore it needs to be converted to the linear spectral domain for applying to the Spectral Subtraction. It can be first converted to the log-spectral domain using an inverse DCT (Discrete Cosine Transform) as follows,

$$\tilde{n}_t^{\{log\}} = C^{-1}\tilde{n}_t \quad . \tag{7}$$

The noise signal for the Spectral Subtraction is obtained as follows,

$$\widetilde{N}_{t,k} = \exp\left(\tilde{n}_{t,i}^{\{log\}}\right), if\ k \in i \quad \text{th Mel filter-bank.} \tag{8}$$

4 Experimental Results

Our evaluations of the proposed method were performed within the Aurora 2.0 evaluation framework as developed by the European Language Resources Association (ELRA) [7][10]. The task is connected English-language digits consisting of eleven words. The acoustic models of the speech recognizer were trained using a database that contains 8,440 utterances of clean speech from the Aurora 2.0 database. In order to evaluate performance under time-varying background noise conditions, speech babble condition was selected from the Aurora 2.0 test database, and a new test data set was generated by combining clean speech samples with background music which consists of prelude parts of ten Korean popular songs with varying degrees of beat and tempo. Each test set consists of 1,001 samples at five different SNRs: 0, 5, 10, 15, and 20 dB.

In our experiment two types of background noise estimation methods were employed for the conventional Spectral Subtraction. In the first method, the noise signal is estimated from the beginning and end non-speech parts of every input speech signal as the same way where the basis noise model is estimated for the proposed noise estimation employing the Variational Model Composition in this study. We used the first 12 frames and the last 12 frames, which are the identical numbers to ones used by the proposed method. The second noise estimation method for the conventional Spectral Subtraction employs minimum statistics based estimation [11], where the previous 25 frames was used for estimation of minimum statistics. SS and MSS indicate the Spectral Subtraction employing noise estimation using non-speech segments (SS) and minimum statistics (MSS) respectively in this paper.

Table 1 shows speech recognition performance (i.e., Word Error Rate, WER) of the baseline system (i.e., no processing), the conventional algorithms (SS and MSS) and the proposed method (VSS) over each SNR condition of speech babble noise. The performance combined with CMN is also shown. In the table, the Spectral Subtraction employing noise estimation during non-speech segments (SS) is slightly better compared to the proposed speech enhancement method (VSS). Here the proposed method (VSS+CMN) shows the best performance compared to the conventional methods (SS and MSS) over all SNR conditions except 0 dB SNR. The SS significantly outperforms the MSS, however, the MSS is considerably better compared to the SS when combined with CMN. We would believe that the SS is effective to increase SNR by suppressing the noise, however, the reconstructed speech is distorted due to the residual noise, resulting in low recognition performance for combination with CMN. It is worth to note that the proposed speech enhancement method is effective for the both cases (i.e., without and with CMN) in the speech babble noise condition.

Table 2 shows speech recognition performance without CMN and with CMN respectively over each different SNR of background music condition. Here the SS is consistently best for the both cases, while the MSS is not effective even compared to the baseline (i.e., no processing). The proposed method shows the comparable performance to the SS at low SNR conditions such as 10, 5 and 0 dB SNR when combined with CMN.

It is interesting to compare the performance of SS and MSS for speech babble and music noise when combined with CMN in Table 1 and 2. The MSS is highly effective for the speech babble noise, however it shows lowest recognition performance (i.e., highest WER) for the background music noise. On the contrary, the SS shows the best performance for the music noise, however it presents lowest performance for the speech babble. Such results show that the conventional noise estimation method for the Spectral Subtraction is not consistently effective for different time-varying background noise conditions. The proposed noise estimation employing the VMC with the Spectral Subtraction shows the consistent effectiveness for both noise conditions.

Table 3 shows speech recognition performance without CMN and with CMN respectively over each different SNR in average for speech babble and music noise conditions. Here the proposed method (VSS+CMN) shows relative improvements 9.02% and 4.86% in WER for SS+CMN and MSS+CMN respectively. These results demonstrate that the proposed noise estimation is effective at improving speech enhancement for speech recognition system in different types of time-varying noise conditions such as speech babble and background music noise..

Table 1. Recognition performance over speech babble noise condition (WER, %)

	0dB	5dB	10dB	15dB	20dB	Avg.
No Processing	86.64	68.75	45/65	24.09	9.76	46.98
SS	79.99	54.78	26.45	10.13	3.48	34.97
MSS	81.92	62.88	42.14	24.91	12.18	44.81
VSS	**77.27**	**52.42**	**27.33**	**13.30**	**5.38**	**35.14**
SS + CMN	69.41	38.27	14.69	4.96	2.33	25.93
MSS + CMN	56.29	28.08	11.85	5.14	2.78	20.83
VSS + CMN	**63.69**	**26.24**	**9.31**	**4.02**	**2.33**	**21.12**

Table 2. Recognition performance over background music noise condition (WER, %)

	0dB	5dB	10dB	15dB	20dB	Avg.
No Processing	72.69	50.60	28.42	13.55	5.86	34.22
SS	66.65	41.01	20.77	9.53	3.64	28.32
MSS	72.72	50.76	32.34	17.96	9.10	36.58
VSS	**69.08**	**45.66**	**27.55**	**15.64**	**7.96**	**33.18**
SS + CMN	59.73	32.68	13.73	6.20	2.90	23.05
MSS + CMN	58.44	36.07	19.16	10.92	5.65	26.05
VSS + CMN	**56.68**	**31.47**	**15.80**	**9.19**	**4.23**	**23.47**

Table 3. Recognition performance averaged over speech babble and background music noise conditions (WER, %)

	0dB	5dB	10dB	15dB	20dB	Avg.
No Processing	79.67	59.68	37.04	18.82	7.81	40.60
SS	73.32	47.90	23.61	9.83	3.56	31.64
MSS	77.32	56.82	37.24	21.44	10.64	40.69
VSS	**73.18**	**49.04**	**27.44**	**14.47**	**6.67**	**34.16**
SS + CMN	64.57	35.48	14.21	5.58	2.62	24.49
MSS + CMN	57.37	32.08	15.51	8.03	4.22	23.44
VSS + CMN	**60.19**	**28.86**	**12.56**	**6.61**	**3.28**	**22.30**

5 Conclusions

This study has proposed an effective noise estimation method for robust speech recognition in time-varying noise conditions. The proposed noise estimation scheme employed the Variation Model Composition (VMC) method, where multiple noise models are generated by selectively applying perturbation factors to the mean parameters of a basis noise model. The resulting collection of the noise models is expected to accurately reflect the unseen noise signal included in the speech segments. The obtained noise models were used to generate multiple

environmental models employing the model combination method. The noise estimate was obtained by using the posterior probability of the multiple environmental models. The proposed noise estimation method was employed for the Spectral Subtraction (SS). The proposed speech enhancement scheme was evaluated within the Aurora 2.0 evaluation framework over speech babble and background music conditions. Experimental results demonstrated that the proposed method is effective at increasing speech recognition performance in time-varying background noise conditions.

Acknowledgements. This Research was supported by Seokyeong University in 2013, and partially by a grant (12-TI-C01) from Advanced Water Management Research Program funded by the Ministry of Land, Infrastructure and Transport of Korean government.

References

[1] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. on Acoustics, Speech and signal Proc.*, vol.27, pp.113-120, 1979.

[2] Y. Ephraim and D. Malah, "Speech Enhancement Using Minimum Mean Square Error Short Time Spectral Amplitude Estimator," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol.32, no.6, pp.1109-1121, 1984.

[3] J.H.L. Hansen and M. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition," *IEEE Trans. on Signal Proc.*, vol.39, no.4, pp.795-805, 1991.

[4] W. Kim and J.H.L. Hansen, "Variational Model Composition for Robust Speech Recognition with Time-Varying Background Noise," *Interspeech-2009*, pp. 2399-2402, 2009.

[5] W. Kim and J.H.L. Hansen, "Variational Noise Model Composition Through Model Perturbation for Robust Speech Recognition with Time-Varying Background Noise," *Speech Communication*, vol.53, no.4, pp.451-464, April 2011.

[6] J.R.Jr. Deller, J.H.L. Hansen, and J.G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, 2000.

[7] H.G. Hirsch, and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," *ISCA ITRW ASR2000*, 2000.

[8] W. Kim and J.H.L. Hansen, "Feature Compensation in the Cepstral Domain Employing Model Combination," *Speech Comm.*, 51(2), pp.83-96, 2009.

[9] M.J.F. Gales and S.J. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination," *IEEE Trans. on Speech and Audio Proc.*, vol.4, no.5, pp.352-359, 1996.
[10] ETSI standard document, *ETSI ES 201 108 v1.1.2 (2000-04)*, 2000.
[11] R. Martin, "Spectral Subtraction Based on Minimum Statistics," *EUSIPCO-94*, pp.1182-1185, 1994

Received: August 8, 2014