COMBINED STATIC AND DYNAMIC VARIANCE ADAPTATION FOR EFFICIENT INTERCONNECTION OF SPEECH ENHANCEMENT PRE-PROCESSOR WITH SPEECH RECOGNIZER

Marc Delcroix, Tomohiro Nakatani, Shinji Watanabe

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho (Keihanna Science City), Soraku-gun, Kyoto 619-0237 Japan

ABSTRACT

It is well known that automatic speech recognition performs poorly in presence of noise or reverberation. Much research has been undertaken on model adaptation and speech enhancement to increase the robustness of speech recognizers. Model adaptation is effective to remove *static* mismatch between speech features and acoustic model parameters, but may not cope well with *dynamic* mismatch. Speech enhancement approaches can reduce *dynamic* perturbations, but often do not interconnect well with speech recognizer. There seems to be a lack of optimal way to combine these two approaches. In this paper we propose introducing the *dynamic* capabilities of speech enhancement into a *static* adaptation scheme. We focus on variance adaptation, and propose a novel parametric variance model that includes *static* and *dynamic* components. The *dynamic* component is derived from a speech enhancement pre-process, and the parameters of the model are optimized using an adaptive training scheme. An evaluation of the method with a speech dereverberation for preprocessing revealed that a 80 % relative error rate reduction was possible compared with the recognition of dereverberated speech, and the final error rate was 5.4 % which is close to that of clean speech (1.2%).

Index Terms— Robust ASR, Variance compensation, Model adaptation, Dereverberation

1. INTRODUCTION

It is well known that the performance of Automatic Speech Recognition (ASR) is severely degraded when attempts are made to recognize speech in the presence of noise and/or reverberation. The problem arises from a mismatch between the clean speech data used for training the ASR system and the noisy observed data used for testing. Solving this problem remains one of the main challenges of ASR research. Conventionally there are two approaches for reducing the mismatch between the training and test conditions, namely model based approaches and feature based approaches.

Model based approaches consist of modifying the acoustic model parameters to fit better with the observed speech features [1] [2] [3]. For example, adaptive training, such as Maximum Likelihood Linear Regression (MLLR) [1], estimates a new acoustic model using the clean speech model and observed speech features. The model adaptation relies on likelihood maximization, which assures a reduction in the mismatch. Adaptive training is effective in removing *static* mismatch caused for example by speaker variations. However, it may not cope well with any mismatch arising for example from non-stationary noise or reverberation.

Feature based approaches consist of estimating clean speech features using the observed speech. For example, speech enhancement methods can be used as a pre-process to ASR [4] [5]. Many speech enhancement algorithms can efficiently reduce non-stationary noise. However, remaining noise or the excessive removal of noise may introduce distortions that prevent high recognition performance.

Recently, there have been several proposals suggesting the use of information on feature reliability to improve the ASR performance of speech enhancement pre-process [6] [7]. The idea consists of focusing during decoding on reliable feature components. As an example, *dynamic* variance compensation proposes increasing the model variance for unreliable feature components by adding the variance of enhanced feature. In [6], substantial ASR improvement has been reported when accurate feature variance could be obtained as in an Oracle experiment. However, with estimated feature variance, the performance was much poorer than that obtained with Oracle. There have been several proposals as regards estimating the variance of enhanced feature [6] [7], but the methods are usually dependent on the speech enhancement pre-process and therefore lack generality. For example, in [6] feature variance is derived from a speech enhancement method based on a Gaussian mixture model of clean speech. The generality of the feature variance calculation could be increased by approximating it with, for example, the estimated observed noise (given by the distance between enhanced features and observed noisy features). However, the estimated variance may be far from the Oracle variance (i.e. the distance between clean and enhanced speech features) and therefore, high levels of performance may not be obtained.

In this paper, we aim at interconnecting a speech enhancement pre-processor with a speech recognizer by simultaneously realizing good performance and generality. To this end, we propose introducing a *dynamic* variance compensation scheme into a *static* adaptive training framework. We design a novel parametric model for the feature variance that includes *static* and *dynamic* components. The *dynamic* component can be derived from the speech enhancement pre-processor output as the estimated observed noise. This calculation can be performed for any pre-processor, thus assuring the generality of the proposed method. The parameters of the feature variance model are optimized using an adaptive training approach and therefore may approach better Oracle feature variance. Moreover, the proposed variance adaptation method could be combined with conventional mean adaptation techniques such as MLLR to further reduce the mismatch.

The organization of the paper is as follows. In Section 2, we introduce some notations and review the principles of *dynamic* variance compensation. In section 3, we introduce the parametric model of feature variance and show how the parameters can be estimated using an adaptive training scheme. In section 4, we show simulation results we obtained when using the proposed method in combination

with a speech dereverberation pre-processor. Finally, we conclude the paper and discuss some future research directions.

2. DYNAMIC VARIANCE COMPENSATION

2.1. Notations

Recognition is usually achieved by finding a word sequence, W, that maximizes a likelihood function as:

$$W = \arg\max_{W} p(X|W)p(W), \tag{1}$$

where $X = [x_1, ..., x_T]$ is a sequence of speech features and $p(W)$ is a language model. Speech is modeled using a Hidden Markov Model (HMM) with state density modeled by a Gaussian Mixture (GM):

$$p(x_t|n) = \sum_{m=1}^{M} p(m)p(x_t|m) = \sum_{m=1}^{M} p(m)N(x_t; \mu_{n,m}, \Sigma_{n,m}), \tag{2}$$

where n is the state index, m is the Gaussian mixture component index, M is the number of Gaussian mixtures, and $\mu_{n,m}$ and $\Sigma_{n,m}$ are a mean vector and a covariance matrix respectively. In the following, we consider diagonal covariance matrices and denote the diagonal elements of $\Sigma_{n,m}$ by $\sigma^2_{n,m,i}$, where i is the feature dimension index. The parameters of the acoustic model are trained with clean speech data.

In practice, speech features used for recognition $\hat{x}_t$ may differ from clean speech features used for training, x_t, because of noise, reverberation or distortions induced by speech enhancement pre-processing. In this paper, we focus on the latter case. Let us model the mismatch, b_t, between clean speech feature x_t and enhanced speech feature $\hat{x}_t$ as:

$$\hat{x}_t = x_t + b_t, \tag{3}$$

where b_t is modeled by a Gaussian as:

$$p(b_t) = N(b_t; 0, \mathbf{\Sigma}_{\hat{x}_t}), \tag{4}$$

and $\mathbf{\Sigma}_{\hat{x}_t}$ represents the feature variance, or uncertainty, which may be time-varying.

2.2. Principles

Recently, a new ASR decoding rule has been proposed to account for the mismatch between the acoustic model and the speech feature [6]. The likelihood of a speech feature given a state n, can be obtained by marginalizing the joint probability over mismatch b_t as [6]:

$$\begin{aligned} p(x_t|n) &= \int_{-\infty}^{+\infty} p(x_t, b_t|n)db_t = \int_{-\infty}^{+\infty} p(x_t|b_t, n)p(b_t|n)db_t \\ &= \sum_{m=1}^{M} p(m)N(\hat{x}_t; \mu_{n,m}, \mathbf{\Sigma}_{n,m} + \mathbf{\Sigma}_{\hat{x}_t}), \end{aligned} \tag{5}$$

where they assumed the mismatch to be state independent, i.e. $p(b_t|n) \approx p(b_t)$. It is shown in [6] that *dynamic* variance compensation is very effective, especially when Oracle feature variance is used. In practice, such an accurate feature variance estimation may not be available, and therefore the performance of *dynamic* variance compensation is not optimal. Here, in an effort to improve the performance of variance compensation, we propose a novel parametric model for the feature variance, and a procedure for estimating the model parameters using adaptive training.

3. PROPOSED METHOD FOR VARIANCE CALCULATION

3.1. Parametric model of feature variance

In theory, feature variance should be computed as the squared difference between clean and pre-processed speech features. However, this calculation may not be possible because clean speech features are unknown. Here we assume that the feature variance is proportional to the estimated observed noise, i.e. the squared difference between observed noisy and pre-processed speech features. Intuitively, this means that speech enhancement introduces more distortions when a great amount of noise is removed. One way to model feature variance is thus:

$$\begin{aligned} (\mathbf{\Sigma}_{\hat{x}_t}(\alpha, \lambda))_{i,j} &= \delta_{i,j} \left(\alpha_i (u_{t,i} - \hat{x}_{t,i})^2 + \lambda_i \sigma^2_{n,m,i} \right) \\ &\triangleq \delta_{i,j} \sigma^2_{\hat{x}_t,i}, \end{aligned} \tag{6}$$

where $\delta_{i,j}$ is the Kronecker symbol, u_t is the observed noisy speech feature and α_i and λ_i are model parameters.

This model contains a *dynamic* variance part, $\alpha_i(u_{t,i} - \hat{x}_{t,i})^2$, and a *static* bias, $\lambda_i \sigma^2_{n,m,i}$. We make the bias state dependent and proportional to the model variance $\sigma^2_{n,m,i}$ [2]. The parameters α_i and λ_i can be optimized by using adaptive training. Note that if $\alpha_i = 0$ the model is equivalent to that of conventional *static* variance compensation [2] and if α_i is constant and $\lambda_i = 0$ it is equivalent to the model of conventional *dynamic* variance compensation [6]. The proposed model enables us to combine both *static* and *dynamic* variance compensation within an adaptive training framework.

It is important to note that the proposed method can be further combined with mean adaptation techniques such as MLLR [1], in order to further reduce the gap between model and speech features.

3.2. Adaptation of variance model parameters

The model variance parameters, $\theta = (\alpha, \lambda)$, can be obtained by maximizing the likelihood as:

$$(\theta, W) = \arg\max_{\theta, W} p(X|W, \theta)p(W). \tag{7}$$

For simplicity, we consider supervised adaptation, where the word sequence W is known. The maximum likelihood estimation problem can be solved using the Expectation Maximization (EM) algorithm. We define an auxiliary function $Q(\theta|\theta')$ as:

$$\begin{aligned} Q(\theta|\theta') &= \sum_{S} \sum_{C} \iint_{X+B=\hat{X}} p(X, B, S, C|\Psi, \theta') \\ &\quad \log(p(X, B, S, C|\Psi, \theta))dXdB \\ &\propto \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{m=1}^{M} \iint_{X+B=\hat{X}} p(X, B, n, m|\Psi, \theta') \\ &\quad \log(p(b_t|\theta))dXdB, \end{aligned} \tag{8}$$

where B is a mismatch feature sequence, S is a set of all possible state sequences, C is a set of all mixture components, N is the number of states, and Ψ represents the acoustic model parameters. The auxiliary function of Eq.(8) is similar to that used for stochastic matching [2]. The difference arises from the model of the mismatch given by Eq.(6) that includes a *dynamic* part. θ should be obtained by maximizing Eq.(8). However, there is no closed form solution for the joint estimation of (α, λ). Therefore, we consider the three following cases, $\alpha = 0$ (i.e. *static* Variance Adaptation (SVA)), $\lambda = 0$ (*dynamic* Variance Adaptation (DVA)) and a combination of the two (*static* and *dynamic* Variance Adaptation (SDVA)).

3.2.1. Static Variance Adaptation (SVA, $\alpha = 0$)

If $\alpha = 0$, the problem is reduced to conventional *static* model variance adaptation as proposed in [2] and sometimes referred to as variance scaling. The scaling factor λ_i can be calculated using the EM algorithm as:

$$\lambda_i = \frac{\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_t(n,m)\frac{(\hat{x}_{t,i}-\mu_{n,m,i})^2}{\sigma^2_{n,m,i}}}{\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_t(n,m)} - 1, \quad (9)$$

where $\gamma_t(n,m)$ is the state occupancy probability, which can be obtained using the forward-backward algorithm. From Eq. (9) we can interpret λ_i as the average of the ratio between the enhanced feature variance and the model variance.

3.2.2. Dynamic Variance Adaptation (DVA, $\lambda = 0$)

When $\lambda = 0$, we can find a close form solution to the maximization problem.[1] By inserting Eqs.(6) and (4) in Eq.(8) and maximizing with respect to α_i, we find the following expression:

$$\alpha_i = \frac{\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_t(n,m)\frac{E\{b^2_{t,i}|\hat{x}_t,n,m,\Psi,\alpha'\}}{(u_{t,i}-\hat{x}_{t,i})^2}}{\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_t(n,m)}, \quad (10)$$

where

$$E\{b_{t,i}|\hat{x}_t,n,m,\Psi,\alpha'\} = \frac{\sigma^2_{\hat{x}_t,i}}{\sigma^2_{\hat{x}_t,i}+\sigma^2_{n,m,i}}(\hat{x}_{t,i}-\mu_{n,m,i}), \quad (11)$$

$$E\{b^2_{t,i}|\hat{x}_t,n,m,\Psi,\alpha'\} = \frac{\sigma^2_{\hat{x}_t,i}\sigma^2_{n,m,i}}{\sigma^2_{\hat{x}_t,i}+\sigma^2_{n,m,i}} + E\{b_{t,i}|\hat{x}_t,n,m,\Psi,\alpha'\}^2. \quad (12)$$

Equations (11) and (12) follow from similar derivations as those in [8]. Note that α_i can be interpreted as the average of the ratio between the mismatch variance, i.e. $(\hat{x}_{t,i}-x_{t,i})^2$, and estimated noise variance $(u_{t,i}-\hat{x}_{t,i})^2$.

3.2.3. Static and Dynamic Variance Adaptation (SDVA)

It may not be easy to find a close form solution of the EM algorithm when the feature variance is modeled as in Eq.(6). However, we saw that solutions could be found if we considered the maximization relative to α and λ separately. As these two maximization problems involve the same likelihood function, the likelihood would also increase if we perform maximization relatively to each parameter in turn. Here we propose first removing the *static* bias by performing *static* variance adaptation as described in section 3.2.1. Then, using the previously adapted acoustic model, we perform *dynamic* variance adaptation as shown in section 3.2.2. This procedure may approach the general case when the variance is modeled as in Eq.(6).

4. EXPERIMENTS

Reverberation is a good example of dynamic mismatch that is challenging for conventional *static* model adaptation techniques. Therefore, here we test the proposed method with speech dereverberation for pre-processing.

[1] For simplicity we considered $\lambda = 0$, although similar results could be obtained for $\lambda = const$.

4.1. Dereverberation method

For pre-processing, we use the blind speech dereverberation recently proposed in [5]. The method first estimates late reverberation by using multi-step forward linear prediction. Dereverberation is then achieved by subtracting the estimated late reverberation from the observed reverberant signal, using conventional spectral subtraction. The method has been shown to remove a large amount of reverberation when using multiple microphones, but the performance deteriorates when using only one microphone. Here we investigate the use of variance compensation to improve the performance in the most challenging single microphone case.

4.2. Experimental settings

To test the proposed method, we used the SOLON recognizer [9] modified to account for the decoding rule of Eq.(5). The recognition task consisted of continuous digit utterances. The acoustic model consisted of speaker independent word based HMMs with 16 states and 3 Gaussians per state. The HMMs were trained using clean speech drawn from the TI-Digit database. The sampling rate was 8 kHz. The acoustic features consisted of 39 coefficients: 12 MFCCs, 0th cepstrum coefficient, delta and acceleration. Cepstral mean normalization (CMN) was applied to the features. We generated reverberant speech by convolving clean speech with a room impulse response. The impulse response was measured in a room with a reverberation time of around 0.5 sec., and a distance between the speaker and the microphones of 1.5 m. The clean speech utterances were obtained from the TI-Digit clean test set. The test set consists of 561 utterances spoken by 104 male and female speakers. The average duration of the utterances is around 6 sec.

We measure the ASR performance using the Word Error Rate (WER). Table 1 gives the baseline recognition results for clean speech, reverberant speech and dereverberated speech. We observed severe degradation induced by reverberation. Only a small error reduction was achieved when using single channel dereverberation. We also show the result obtained using variance compensation with variance given by the estimated observed noise (without adaptation, i.e. $\alpha = 1$, $\lambda = 0$) and with ideal variance (Oracle). Variance compensation reduces the error especially with Oracle variance, in which case the WER is very close to that of clean speech. Our objective is to approach Oracle performance.

Clean	1.2 %
Reverberant	32.7 %
Dereverberated	31 %
Variance Compensation (without adaptation)	15.9 %
Oracle	3.3 %

Table 1. *Baseline ASR results.*

4.3. Results of variance adaptation

We use speaker independent adaptation data to adapt the model to the speech enhanced data without performing speaker adaptation. The adaptation data consists of 520 utterances spoken by the same female and male speakers as the test set. To test the influence of the number of adaptation data, we used subsets of adaptation data containing from 2 to 512 utterances extracted randomly from the 520 adaptation utterances. Figure 1 plots the WER as a function of the number of adaptation utterances for SVA, DVA and SDVA. The results are averaged over 5 randomly generated adaptation data sets.

We observe that in all cases, convergence is almost achieved after 2 utterances. A great reduction in the WER from 31% to 15.2% is

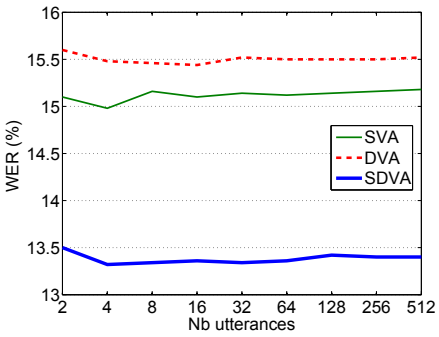

Fig. 1. *WER as a function of the number of adaptation data for SVA (thin solid line), DVA (dash-dotted line) and SDVA (thick solid line)*

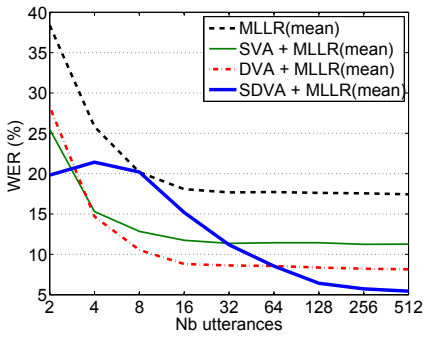

Fig. 2. *WER as a function of the number of adaptation data for MLLR (dotted line), SVA + MLLR (thin solid line), DVA + MLLR (dash-dotted line) and SDVA + MLLR (thick solid line)*

achieved using SVA. DVA achieved similarly good results although they were slightly worse than SVA. In contrast, when using SDVA the performance improved by an additional 2%. These results show that even though there remains a gap compared with the clean speech case or Oracle results shown in Table 1, the proposed method could significantly improve the ASR performance by reducing the error by 56% compared with the recognition of dereverberated speech. This experiment proves the effectiveness of combining static and dynamic variance adaptation.

4.4. Results of variance adaptation combined with MLLR for mean adaptation

Here we investigate the use of feature variance adaptation with mean adaptation using global MLLR with full transformation matrix. Figure 2 plots WER as a function of the number of utterances when using only MLLR (mean), SVA + MLLR (mean), DVA + MLLR (mean), and SDVA + MLLR (mean). Note that SVA + MLLR (mean) is somewhat similar to conventional mean and variance MLLR [1]. With only MLLR, WER converges to around 17%. By combining SVA with MLLR WER is reduced to up to 11%. Using DVA + MLLR can reduce WER further to 8%. Finally, SDVA + MLLR converges to a WER close to 5% which corresponds to more than 80 % relative error rate reduction compared with the recognition of dereverberated speech. This WER is pretty close to that of clean speech. This experiment proves the effectiveness of combining the proposed method with mean adaptation.

Note that with MLLR, SVA + MLLR and DVA + MLLR, more than 16 utterances may be needed to converge. When SDVA is used, better performance is achieved at the cost of more adaptation data (here more than 128 utterances). When using SDVA + MLLR, we obtain poorer results when too few utterances are used. The problem may arise from instabilities that occur when performing the EM algorithm in turns. One way to solve this problem may be to perform the variance update and mean update in the same step of the EM algorithm. Future work will include an investigation of this matter.

5. CONCLUSION

We investigated the use of variance adaptation to improve the ASR performance of speech pre-processed with a speech enhancement method. We proposed a novel method for calculating the feature variance, which involves a parametric model whose parameters are estimated using adaptive training. By combining *static* and *dynamic* adaptation, we designed a general and high performance way of interconnecting a speech enhancement pre-processor and a speech recognizer. We tested the method with a blind dereverberation algorithm for pre-processing. We showed that variance adaptation was very effective in reducing the WER, especially when we combined both *static* and *dynamic* adaptation. We also demonstrated that the proposed method could be combined with conventional mean adaptation methods such as MLLR. In which case the ASR performance was comparable to that of clean speech. Future work will include investigations on a cluster based adaptation method designed to improve estimation of the feature variance model parameters as well as testing the proposed method with other speech enhancement methods and on larger vocabulary tasks.

6. REFERENCES

[1] Gales, M.J.F. and Woodland, P.C., "Mean and variance adaptation within the MLLR framework," Computer Speech & Language, vol. 10, pp. 249-264, 1996.

[2] Sankar, A. and Lee, C. H., "A maximum-likelihood approach to stochastic matching for robust speech recognition," IEEE Trans. SAP, vol. 4, no. 3, pp. 190-202, 1996.

[3] Gales, M. J. F. and Young, S. J., "Robust continuous speech recognition using parallel model combination," IEEE Trans. SAP, vol. 4, no. 5, pp. 352-359, 1996.

[4] Deng, L., Acero, A., Plumpe, M. and Huang, X., "Large-vocabulary speech recognition under adverse acoustic environments," Proc. ICSLP'00, vol. 3, pp. 806-809, 2000.

[5] Kinoshita, K., Delcroix, M., Nakatani T. and Miyoshi, M., "A linear prediction-based microphone array for speech dereverberation in a realistic sound field," Proc. AES'07, 2007.

[6] Deng, L., Droppo, J. and Acero, A., "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," IEEE Trans. SAP, vol. 13, no. 3, pp. 412-421, 2005.

[7] Kolossa, D., Sawada, H., Astudillo, R. F., Orglmeister, R. and Makino, S., " Recognition of convolutive speech mixtures by missing feature techniques for ICA," Proc. ACSSC'06, pp. 1397-1401, 2006.

[8] Rose, R. C., Hofstetter, E. M. and Reynolds, D. A., "Integrated models of signal and background with application to speaker identification in noise," IEEE Trans. SAP, vol. 2, no. 2, pp. 245-257, 1994.

[9] T. Hori, "NTT speech recognizer with OutLook on the next generation: SOLON," Proc. NTT Workshop on Communication Scene Analysis, SP-6, 2004.