動的分散適応に基づく音声強調と音声認識の統合手法の提案

デルクロアマーク[†] 中谷 智広[†] 渡部 晋治[†]

[†] NTTコミュニケーション科学基礎研究所、619-0237「けいはんな学研都市」光台2-4
E-mail: [†]{marc.delcroix,nak,watanabe}@cslab.kecl.ntt.co.jp

あらまし 一般に、雑音や残響の影響により音声認識率は低下する。これに対し、音声強調を前処理として用いると、時間的に変化する音響的な歪みをある程度低減することができるが、必ずしも音声認識性能を改善できるとはかぎらなかった。また、モデル適応技術を用いることで、音声強調処理後の音声と音響モデルのミスマッチをある程度低減することができるが、動的なミスマッチについては扱うことはできなかった。音声強調とモデル適応のより最適な組み合わせ法の開発が重要であると考えられる。本稿では、動的なミスマッチについても適切に低減できるモデル適応法を提案する。分散を静的な分散と動的な分散で構成されるパラメトリックモデルで表現し、適応処理に基づき、モデルパラメータを最適化する。実験により、残響除去を前処理として用いた場合に、認識誤りを80%削減できること、およびクリーン音声に近い5.4ることを示す。クリーン音声の場合と近い性能が得られた。

キーワード ロバスト音声認識、分散補正、モデル適応

Dynamic feature variance adaptation for robust speech recognition with a speech enhancement pre-processor

Marc DELCROIX[†], Tomohiro NAKATANI[†], and Shinji WATANABE[†]

[†] NTT Communication Science Laboratories, NTT Corporation 2-4, Hikaridai, Seika-cho (Keihanna Science City), Soraku-gun, Kyoto 619-0237 Japan
E-mail: [†]{marc.delcroix,nak,watanabe}@cslab.kecl.ntt.co.jp

Abstract It is well known that the performance of automatic speech recognition degrades severely in presence of noise or reverberation. Speech enhancement techniques may reduce such acoustic perturbations, but often do not interconnect well with speech recognizer. To cope with this problem, model adaptation is usually used to reduce the mismatch between the speech enhanced features and the acoustic model used by the recognizer. However, conventional model adaptation techniques assume *static* mismatch and may therefore not cope well with *dynamic* mismatch arising from noise or reverberation. There seems to be a lack of optimal ways to combine model adaptation and speech enhancement. In this paper we propose a novel adaptation scheme that may cope with *dynamic* mismatch. We introduce a parametric model for variance adaptation that includes *static* components, and *dynamic* components derived from a speech enhancement pre-process. The model parameters are optimized using adaptive training. An evaluation of the method with a speech dereverberation for pre-processing 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%).

Key words Robust ASR, Variance compensation, Model adaptation

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. One way to tackle the problem consists of modifying the acoustic model parameters to fit better with the observed speech features. This is usually referred to as model based approaches, [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, but it may not cope well with *dynamic* mismatch arising for example from non-stationary noise or reverberation.

Alternatives to model based approaches are feature based approaches that 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. Moreover, the estimated variance may be far from the Oracle variance 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 *dynamic* feature variance. 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 *static* adaptation is realized by weighting the acoustic model variances as it is done with conventional *static* variance adaptation. The model parameters 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):

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

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:

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

and $\Sigma_{\hat{x}_t}$ 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]:

$$p(x_t|n)=\int_{-\infty}^{+\infty} p(x_t, b_t|n)db_t = \int_{-\infty}^{+\infty} p(x_t|b_t, n)p(b_t|n)db_t$$
$$=\sum_{m=1}^{M} p(m)N(\hat{x}_t; \mu_{n,m}, \Sigma_{n,m} + \Sigma_{\hat{x}_t}), \quad (5)$$

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 adaptation

3.1 Parametric model of variances

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:

$$(\Sigma_{\hat{x}_t}(\alpha))_{i,j} = \delta_{i,j}\alpha_i(u_{t,i} - \hat{x}_{t,i})^2 \triangleq \delta_{i,j}\sigma^2_{\hat{x}_t,i}, \quad (6)$$

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

Moreover, in order to also account for static mismatch, we introduce a weight λ in the variances of the acoustic models [2]. The sate variance can thus be written as:

$$\sigma^2_{n,m,i}(\lambda) = \lambda_i\sigma^2_{n,m,i}. \quad (7)$$

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 $\lambda_i = 1$ 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:

$$(\theta, W) = \arg\max_{\theta, W} p(X|W, \theta)p(W). \quad (8)$$

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:

$$Q(\theta|\theta')=\sum_{S}\sum_{C}\iint_{X+B=\hat{X}} p(X, B, S, C|\Psi, \theta')$$
$$\log(p(X, B, S, C|\Psi, \theta))dXdB$$
$$\propto\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\iint_{X+B=\hat{X}} p(X, B, n, m|\Psi, \theta')$$
$$(\log(p(b_t|\alpha)) + \log(p(x|n, m, \lambda)))\, dXdB,$$
$$\triangleq Q(\alpha|\alpha', \lambda') + Q(\lambda|\alpha', \lambda') \quad (9)$$

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.(9) 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.(9). We observe that the auxiliary function decomposes into two functions $Q(\alpha|\alpha', \lambda')$ and $Q(\lambda|\alpha', \lambda')$. However, there is no closed form solution for the joint estimation of (α, λ). Therefore, we consider the three following cases, α = *const.* (i.e. *static* Variance Adaptation (SVA)), λ = *const.* (*dynamic* Variance Adaptation (DVA)) and a combination of the two (*static* and *dynamic* Variance Adaptation (SDVA or DSVA)).

3.2.1 Static Variance Adaptation (SVA, α = *const.*)

Let us here consider the maximization of $Q(\theta|\theta')$ with respect to λ for a constant α. By considering the model of Eq.(7) and doing similar calculation as in [2], we can show that a close form solution may be obtained as:

$$\lambda_i = \frac{\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_t(n, m)\frac{A(x_{t,i}, n, m, \Psi, \alpha')}{\sigma^2_{n,m,i}}}{\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_t(n, m)}, \quad (10)$$

where $\gamma_t(n, m)$ is the state occupancy probability, which can be obtained using the forward-backward algorithm, and

$$A(x_{t,i}, n, m, \Psi, \alpha')=\mu^2_{n,m,i} - 2\mu_{n,m,i}E\{x_t|x_t, n, m, \Psi, \alpha'\}$$
$$+E\{x_t^2|x_t, n, m, \Psi, \alpha'\},$$
$$(11)$$

$$E\{x_t|\hat{x}_t, n, m, \Psi, \alpha'\}=\frac{\iint_{X+B=\hat{X}} x_{t,i}p(x_t, b_t, n, m|\Psi, \theta')dx_t db_t}{p(\hat{x}_t|n, m, \Psi, \theta')}$$
$$=\frac{\sigma^2_{\hat{x}_t,i}\sigma^2_{n,m,i}}{\sigma^2_{\hat{x}_t,i} + \sigma^2_{n,m,i}}(\frac{\hat{x}_{t,i}}{\sigma^2_{\hat{x}_t,i}} + \frac{\mu_{n,m,i}}{\sigma^2_{n,m,i}}), \quad (12)$$
$$(13)$$

$$E\{x_t^2|\hat{x}_t, n, m, \Psi, \alpha'\} = \frac{\iint_{X+B=\hat{X}} x_{t,i}^2 p(x_t, b_t, n, m|\Psi, \theta') dx_t db_t}{p(\hat{x}_t|n, m, \Psi, \theta')}$$
$$= \frac{\sigma_{\hat{x}_t,i}^2 \sigma_{n,m,i}^2}{\sigma_{\hat{x}_t,i}^2 + \sigma_{n,m,i}^2} + E\{x_t|\hat{x}_t, n, m, \Psi, \alpha'\}^2. \tag{14}$$

Equations (12) and (14) follow from similar derivations as those in [8]. Note that if $\alpha = 0$, the problem is reduced to conventional *static* model variance adaptation as proposed in [2] and sometimes referred to as variance scaling. 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, $\lambda = const$)

When $\lambda = const$, we can find a close form solution to the maximization problem. By inserting Eqs.(6) and (4) in Eq.(9) and maximizing with respect to α_i, we find the following expression:

$$\alpha_i = \frac{\sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{m=1}^{M} \frac{\gamma_t(n,m)}{(u_{t,i} - \hat{x}_{t,i})^2} E\{b_t^2|\hat{x}_t, n, m, \Psi, \alpha'\}}{\sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_t(n,m)}, \tag{15}$$

where $E\{b_t^2|\hat{x}_t, n, m, \Psi, \alpha'\}$ follows from a similar definition as Eq.(12) and Eq.(14):

$$E\{b_t^2|\hat{x}_t, n, m, \Psi, \alpha'\} = \frac{\sigma_{\hat{x}_t,i}^2 \sigma_{n,m,i}^2}{\sigma_{\hat{x}_t,i}^2 + \sigma_{n,m,i}^2} + E\{b_t|\hat{x}_t, n, m, \Psi, \alpha'\}^2. \tag{16}$$

$$E\{b_t|\hat{x}_t, n, m, \Psi, \alpha'\} = \frac{\sigma_{\hat{x}_t,i}^2}{\sigma_{\hat{x}_t,i}^2 + \sigma_{n,m,i}^2} (\hat{x}_{t,i} - \mu_{n,m,i}). \tag{17}$$

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 or DVSA)

It may not be easy to find a close form solution of the EM algorithm when the maximization relatively to α and λ is done at the same time. 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. This procedure may approach the general case.

Here we investigate two cases. With the first case, we start by removing the *static* bias with *static* variance adaptation as described in section 3.2.1 and setting $\alpha = 0$. Then, using the previously adapted acoustic model, we perform *dynamic* variance adaptation as shown in section 3.2.2. This is referred to as Static and Dynamic Variance Adaptation (SDVA). We also consider the opposite case, where first dynamic variance adaptation is performed followed by static variance adaptation (i.e. Dynamic and Static Variance Adaptation - DSVA).

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 a speech dereverberation for pre-processing recently proposed in [5].

4.1 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. $\alpha = 1$, $\lambda = 1$) 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 %

表 1 *Baseline ASR results.*

4.2 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

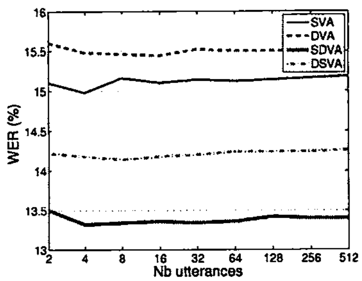
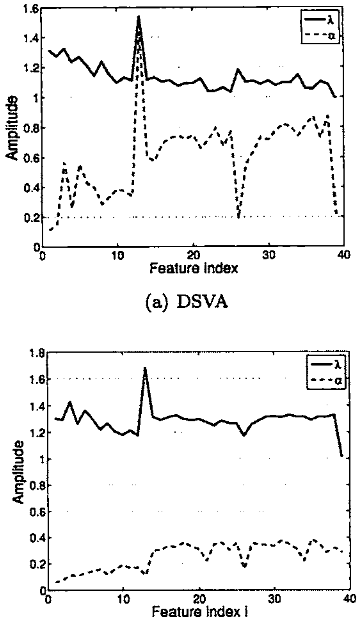


図 1 *WER as a function of the number of adaptation data for SVA (thin solid line), DVA (dash line) and SDVA (thick solid line) and DSVA (dash-dotted)*

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 ($\alpha = 0$), DVA, DVSA 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 achieved using SVA. DVA achieved similarly good results although they were slightly worse than SVA. In contrast, when using DSVA and SDVA the performance improved by an additional 1% and 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.

Figure 2 plots the values of λ and α obtained after adaptation using 16 utterances for SDVA and DSVA. Looking at Fig-(a), when *dynamic* adaptation is performed first, we observe a large peak in α and λ for the 13th components which corresponds to the 0th cepstrum coefficient. It is not surprising that speech enhancement introduces large uncertainty in that coefficient which is related to feature energy. Looking at Fig-(b), we observe the same large peak in λ but the peak has disappeared in α. This suggests that the mismatch in the 13th components cepstrum coefficient is essentially *static*. By using only *dynamic* adaptation, the model may not be well suited for compensating *static* component and consequently, performance are not optimal. By first removing the *static* mismatch with SVA, we may then focus only on optimizing the model for the *dynamic* part and consequently better performance is obtained. This illustrates the need to include both *static* and *dynamic* variance compensation, and suggests that the order in which the optimization is carried influences the results.



(a) DSVA

(b) SDVA

図 2 λ (solid line) and α (dash line) for DSVA and for SDVA.

4.3 Results of variance adaptation combined with MLLR for mean adaptation

Here we investigate the use of feature variance adaptation with mean adaptation using MLLR. Figure 3 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 prove the effectiveness of combining the proposed method with mean adaptation.

Note that with MLLR, SVA + MLLR, DVA + MLLR and DSVA + 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.

4.4 Discussion

Since the proposed adaptation scheme adds a dynamic variance term to the acoustic model variances, the acoustic model variances will become time-varying and therefore

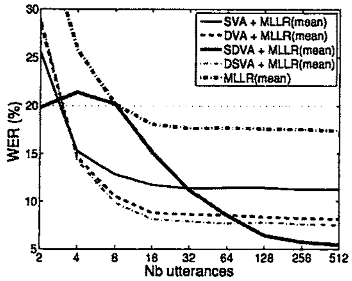
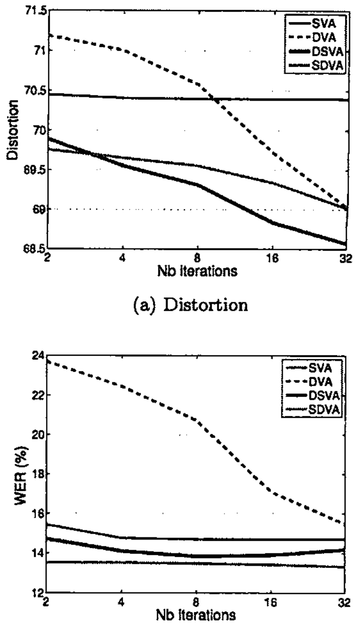


図 3 *WER as a function of the number of adaptation data for MLLR (dotted line), SVA + MLLR (thin solid line), DVA + MLLR (dash-dotted line) and SDVA + MLLR (thick solid line)*

we may expect poor convergence property of the EM algorithm. Here we briefly discuss the convergence of the proposed adaptation scheme. Figure 4 plots the distortion (minus log-likelihood) and the WER as a function of the number of iterations of the EM algorithm for SVA, DVA, DSVA and SDVA. The EM algorithm converges after only 2 iterations for SVA. In contrast, the convergence is much slower with DVA. The poor convergence is due to the difficulty of handling dynamic component, and from the fact that the mismatch may not be well modeled with only a dynamic component, as it was suggested in section 4. 2. Such a poor convergence property would become a problem especially for online applications. Fortunately, we observe that the convergence is improved when both *static* and *dynamic* adaptations are used jointly. In particular, SDVA and DSVA converge in terms of WER after only 2 iterations of the EM algorithm which would be reasonable for online use.

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 investigation on the use of the proposed method with other speech enhancement methods such as spectral subtraction for noise reduction.



(a) Distortion

(b) WER

図 4 Distortion and WER as a function of the number of iterations of the EM algorithm for SVA, DVA, SDVA and DSVA. In this experiment, 4 utterances were used for adaptation.

文 献

[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.