DYNAMIC VARIANCE ADAPTATION USING DIFFERENCED MAXIMUM MUTUAL INFORMATION

Marc Delcroix, Atsunori Ogawa, Tomohiro Nakatani, Atsushi Nakamura

NTT Communication Science Laboratories, NTT corporation,
2-4, Hikaridai, Seika-cho (Keihanna Science City), Soraku-gun, Kyoto 619-0237 Japan
{marc.delcroix,ogawa.atsunori,nakatani.tomohiro,nakamura.atsushi}@lab.ntt.co.jp

ABSTRACT

A conventional approach for noise robust automatic speech recognition consists of using a speech enhancement before recognition. However, speech enhancement cannot completely remove noise, thus a mismatch between the enhanced speech and the acoustic model inevitably remains. Uncertainty decoding approaches have been used to mitigate such a mismatch by accounting for the feature uncertainty during decoding. We have proposed dynamic variance adaptation to estimate the feature uncertainty given adaptation data by maximization of likelihood or discriminative criterion such as MMI. For unsupervised adaptation, the transcriptions are obtained from a first recognition pass and thus contain errors. Such errors are fatal when using a discriminative criterion. In this paper, we investigate the recently proposed differenced MMI discriminative criterion for unsupervised dynamic variance adaptation, because it inherently includes a mechanism to mitigate the influence of errors in the transcriptions.

Index Terms— Robust speech recognition, dynamic variance adaptation, unsupervised adaptation, discriminative training, dMMI

1. INTRODUCTION

Robustness of automatic speech recognition (ASR) systems to noise or reverberation is known as one of the major remaining challenges for the ASR community. To tackle this issue, speech enhancement methods are often used to reduce noise prior to recognition. However, most speech enhancement methods fail to remove completely noise or introduce distortions and therefore a mismatch exists between the enhanced speech and the acoustic model used for recognition. The mismatch is time-varying, i.e. *dynamic*, because noise is usually non-stationary and because most speech enhancement methods involve frame-by-frame processing. Such a *dynamic* mismatch cannot be completely compensated with conventional *static* model adaptation approaches such as maximum likelihood linear regression (MLLR) [1].

Recently, uncertainty decoding approaches have been proposed to mitigate the influence of the dynamic mismatch emanating from the speech enhancement pre-processor [2, 3, 4]. These approaches consist of adding a dynamic feature variance term, representing feature uncertainty, to the variance of the Gaussians of the acoustic model. The dynamic feature variance takes large values for uncertain features, therefore reducing their influence on recognition results. In [5], we proposed dynamic variance adaptation (DVA) to estimate the dynamic feature variance using adaptation data. DVA relies on a parametric model of the dynamic feature variance, in which parameters are optimized using maximum likelihood (ML) given adaptation data [5].

We have recently investigated the use of a discriminative criterion for DVA [6]. The motivation was twofold. First, using a discriminative criterion we expect to obtain a better estimate of the dynamic feature variance that would lead to higher recognition performance. Second, since most current recognition systems use discriminatively trained acoustic models, using also a discriminative criterion for DVA may better preserve the discrimination capability of the acoustic model. In [6], we demonstrated performance gains when employing the maximum mutual information (MMI) criterion for supervised adaptation. However, in case of unsupervised adaptation, the transcriptions are obtained from a first recognition pass and therefore inevitably contain errors. Such errors in the transcriptions are particularly harmful when using a discriminative criterion, and may prevent any performance improvement compared to ML [7].

Recently, the differenced MMI (dMMI) [8] criterion was proposed to generalize conventional criteria such as minimum phone error (MPE) [9] or boosted MMI (BMMI) [10]. In a similar way as BMMI, dMMI includes margin terms to boost the contribution of high error recognition candidates during adaptation. In addition, dMMI defines references in a soft manner, i.e. as a summation of low error recognition candidates [11]. This soft definition of the references inherently provides a mechanism to mitigate the influence of errors in the transcriptions. Therefore, the dMMI criterion appears more suitable for unsupervised adaptation than other discriminative criteria. In this paper, we investigate the use of dMMI criterion for DVA. We confirm experimentally with a noisy speech command recognition task that dMMI based DVA can achieve higher and more stable performance than MMI for unsupervised adaptation.

2. DYNAMIC VARIANCE ADAPTATION

Let us first briefly recall the principles of uncertainty decoding and DVA. In this paper, we consider acoustic models represented by hidden Markov models (HMMs) with HMM state output probability density modeled by Gaussian mixture models (GMMs). Uncertainty decoding approaches consider features as distributions instead of point estimates. Therefore, the probability density of the enhanced feature vector at time frame t, $\mathbf{y}_t$, given the HMM state n can be obtained as [4],

$$p(\mathbf{y}_t|n) \quad = \quad \sum_{m=1}^{M} \omega_{n,m} N(\mathbf{y}_t; \boldsymbol{\mu}_{n,m}, \boldsymbol{\Sigma}_{n,m} + \boldsymbol{\Sigma}_t), \qquad (1)$$

where m is the Gaussian mixture component index, M is the number of Gaussian mixtures, $\omega_{n,m}$ is the mixture weight, and $\boldsymbol{\mu}_{n,m}$ and $\boldsymbol{\Sigma}_{n,m}$ are a mean vector and a covariance matrix, respectively. In the following, we consider diagonal covariance matrices. $\boldsymbol{\Sigma}_t$ is a dynamic feature covariance matrix that we assume to be diagonal, with diagonal elements denoted as the dynamic feature variance $\sigma^2_{t,i}$, where i is the feature dimension index and I is the dimension of the feature vector.

Several approaches have been proposed to estimate the dynamic feature variance [2, 3, 4]. We have recently proposed to introduce a parametric model for the feature variance and optimize the model parameters using adaptation data. In [5], we introduced a model for the dynamic feature variance that is based on the assumption that the more features are affected by acoustic distortions, the more features will be uncertain. Moreover, the amount of acoustic distortion is assumed proportional to the acoustic distortion reduction provided by the speech enhancement pre-processor and obtained as the difference between enhanced and observed features. Therefore, we model the dynamic feature variance, $\sigma^2_{t,i}$, as,

$$\sigma^2_{t,i} = \alpha^2_i \left(\mathbf{u}_{t,i} - \mathbf{y}_{t,i}\right)^2, \qquad (2)$$

where α_i is a *pre-processor uncertainty weight* representing the uncertainty of the speech enhancement process, and $\mathbf{u}_{t,i}$ is the observed speech feature. The pre-processor uncertainty weights, α_i, are the parameters that should be optimized. Note that as shown by the index, α_i can take different values for each feature dimension i.

The model of the dynamic feature variance of Eq. (2) depends only on the input and output of the speech enhancement and is therefore general. However, in some cases the model is too weak at representing well the actual dynamic feature variance. The dynamic feature variance changes according to the level of speech sound, which varies with the HMM states. The model representational power could be improved by introducing HMM state dependency. We achieve this by using a cluster-based representation, i.e. the Gaussians of the acoustic model are grouped into clusters and a different pre-processor uncertainty weight is associated with each cluster. The model of the dynamic feature variance is thus expressed as,

$$\sigma^2_{t,c,i} = \alpha^2_{c,i} \left(\mathbf{u}_{t,i} - \mathbf{y}_{t,i}\right)^2, \qquad (3)$$

where c is the index of the HMM state cluster C_c. Accordingly, in the following, we call the method cluster DVA.The Gaussian clusters can be obtained using conventional clustering approaches to create a binary regression tree [1]. Here the clustering is performed according to the Mahalanobis distance since it considers the variance of the Gaussians and may therefore be more suitable for DVA.

The dynamic feature variance model parameters are estimated using adaptation data. Let us define $\boldsymbol{\theta} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_c, \ldots, \boldsymbol{\alpha}_C)$, where $\boldsymbol{\alpha}_c = (\alpha_{c,1}, \ldots, \alpha_{c,i}, \ldots, \alpha_{c,I})$ is a vector containing the uncertainty weights associated with the c^{th} cluster and C is the number of clusters. The model parameters can be obtained by maximizing an objective function as,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{F}_{\boldsymbol{\theta}}(Y, S_r), \qquad (4)$$

where $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_T]$ is a sequence of enhanced speech feature vectors, S_r is the word sequence corresponding to the feature sequence Y and $\mathcal{F}_{\boldsymbol{\theta}}$ is an objective function. In case of unsupervised adaptation, S_r is obtained from a first recognition pass. After optimizing Eq. (4), recognition is performed using Eq. (1) where the dynamic feature variance is obtained with Eq. (3) using the estimated pre-processor uncertainty weights, $\boldsymbol{\theta}$.

Several criteria can be used for the optimization. We have investigated optimization using ML [5] and MMI [6] objective functions. In the following section, we discuss the use of dMMI.

3. DMMI-BASED DVA

The dMMI objective function generalizes MPE and BMMI objective functions. Let us first recall the BMMI objective function which is given by [10],

$$\mathcal{F}^{BMMI}_{\boldsymbol{\theta},\upsilon}(Y, S_r) = \frac{1}{\psi} \log \frac{P(S_r)^{\psi\eta} p_{\boldsymbol{\theta}}(Y|S_r)^{\psi}}{\sum_j P(S_j)^{\psi\eta} p_{\boldsymbol{\theta}}(Y|S_j)^{\psi} e^{\psi\upsilon\mathcal{E}_{j,r}}}, \quad (5)$$

where S_j is a recognition candidate for Y. $\mathcal{E}_{j,r}$ represents the error between the recognition candidate S_j and the reference S_r. $p_{\boldsymbol{\theta}}(Y|S_r)$ corresponds to the acoustic model. The parameter η is the language model scaling and ψ is the acoustic scaling. Note that to simplify the expressions in Eq. (5) we omitted the summation over the adaptation utterances. The numerator of the BMMI objective function corresponds to the contribution to the reference transcription, and the denominator accounts for the contribution of the competing recognition candidates. BMMI includes a margin term with parameter υ in the denominator. The error term, $\mathcal{E}_{j,r}$, can be defined as the phone error, word error or phone frame error. In the following, we use the phone frame error as defined in [12]. Note that by setting υ to a positive value, the contribution of the recognition candidates with high errors is emphasized, and therefore candidates with low errors have to work harder to compensate high error candidates. Consequently, better discrimination is expected. Accordingly, υ can be interpreted as a boosting factor [10].

The dMMI objective function can be derived from the difference of two BMMI objective functions with different margin parameters as [8],

$$\mathcal{F}^{dMMI}_{\boldsymbol{\theta},\upsilon_1,\upsilon_2}(Y, S_r) = \frac{1}{\psi(\upsilon_2 - \upsilon_1)} \log \frac{\sum_j P(S_j)^{\psi\eta} p_{\boldsymbol{\theta}}(Y|S_j)^{\psi} e^{\psi\upsilon_1\mathcal{E}_{j,r}}}{\sum_j P(S_j)^{\psi\eta} p_{\boldsymbol{\theta}}(Y|S_j)^{\psi} e^{\psi\upsilon_2\mathcal{E}_{j,r}}}, \qquad (6)$$

where υ_1 and υ_2 represents the two margin parameters.

One major difference between dMMI and BMMI is that the dMMI objective function contains margin terms in both its numerator and denominator. By setting υ_2 to a positive value, we emphasize recognition candidates with a high number of errors in the denominator, which is equivalent to conventional BMMI as shown in Eq. (5). By setting υ_1 to a negative value, the contribution of recognition candidates with low error are emphasized in the numerator. Therefore, the numerator is equivalent to references defined in a soft manner, i.e., by also accounting for the recognition candidates with few errors compared to the references [11]. Consequently dMMI possesses inherently a mechanism to compensate transcription errors that may have a similar effect to other approaches used to mitigate the transcription errors using for example confidence scores [7]. This lose definition of the references appears particularly interesting in case of unsupervised adaptation where transcriptions inevitably contain errors.

Eq. (4) can be solved using a gradient ascent method [6]. We use the RPROP algorithm [13] to solve the optimization problem of Eq. (4). Note that RPROP is often used for discriminative training as it is simple to implement and has been shown to provide performance competitive with other optimization approaches [14, 15]. We set the initial value for $\boldsymbol{\theta}$ to the value obtained by likelihood maximization. Indeed, having appropriate initial value when using discriminative criterion is essential. In this paper, we did not include any smoothing during discriminative adaptation.

For cluster DVA, in a similar way to MLLR [1], we decide to adapt the parameters at a node of the regression tree if the node occupancy count is lower than an occupancy count threshold. Otherwise, adaptation is carried out at the leafs of the node. When using the dMMI criterion, the denominator term that includes high error recognition candidates would also contribute to the node occupancy counts which then take unrealistic values (e.g. negative values) . To solve this issue, we calculate the node occupancy counts using the ML criterion.

4. EXPERIMENTS

We tested the proposed method on the recent PASCAL 'CHiME' speech separation and recognition challenge task [16].

4.1. Settings

The CHiME task consists of 6-word commands spoken by 34 English speakers. The commands are corrupted by background noise that was collected in a real living room. The noise is highly non-stationary and includes noise sources such as TV, children's voices or music. The recognition target consists of two keywords consisting of a letter followed by a digit, which are included in the command. The training data consist of 17,000 utterances and 6 hours of background noise data. The training utterances are corrupted by reverberation but do not include noise. The test data consist of a development set and an evaluation set that both include 600 reverberant utterances at 6 different SNRs ranging from -6 to 9 dB. Note that the training data set and the test data sets all consist of reverberant speech for the same room (reverberation time of 300 msec.) but with different speaker positions and room configurations (doors open/close ...) and therefore with different reverberant characteristics. A detailed description of the CHiME task can be found in [16].

We used the DOLPHIN enhancement algorithm to extract the target speech from the noisy signals [17]. DOLPHIN is a recently proposed algorithm that performs speech-noise separation using spatial and spectral models for speech and noise. DOLPHIN was one of the components of the system we developed for the CHiME challenge. That system achieved the best performance among the challenge participants [18]. DOLPHIN represents therefore a state of the art speech enhancement method for the CHiME challenge task. We used here the most recent and powerful version of DOLPHIN that is based on a speech model using MFCCs [17].

We employed the speech recognizer platform SOLON [19], which was developed at NTT Communication Science Laboratories, to train the acoustic models, and to perform adaptation and decoding. The acoustic models consisted of conventional left-to-right HMMs with a total of 254 states (including silent states) each modeled by a GMM consisting of 7 Gaussian components. We trained speaker dependent acoustic models according to the CHiME regulation using the dMMI criterion with the clean training data.

We performed speaker dependent, SNR independent unsupervised adaptation. We used all the test data (from all SNR levels) from a given speaker and a given test set to generate transcriptions that were used for adaptation. The amount of adaptation data per speaker was about 3.5 minutes. The same data were used for recognition.

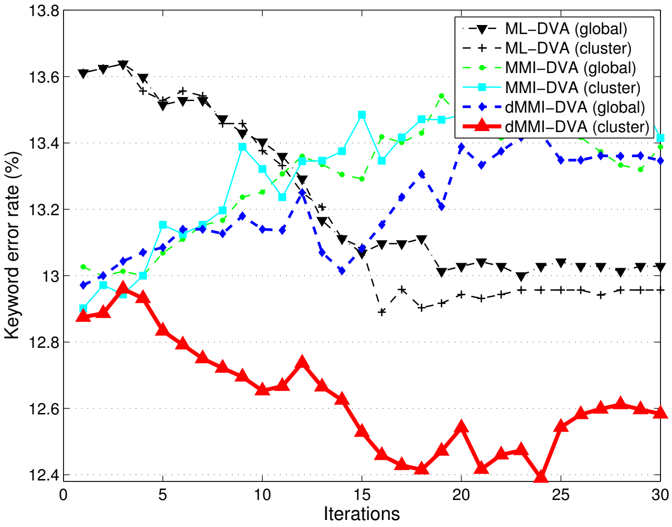

Fig. 1. *Average keyword error rate as a function of the number of iterations for the development set.*

4.2. Results

Figure 1 plots the keyword error rate as a function of the number of iterations for global[1] and cluster DVA, using ML, MMI and dMMI criteria for the development set. The results are expressed in terms of keyword error rate averaged over the 34 speakers and 6 SNR conditions. The baseline system achieved a 3.25 % keyword error rate for recognition of clean speech for the development set. This demonstrates the severeness of the noise conditions since the noisy speech keyword error rate was up to 30.39 %. For the dMMI criterion, the values of the margin parameters were set to $\upsilon_1 = -4$ and $\upsilon_2 = 0.1$. These settings were tuned on the development set for cluster DVA. Consequently, different parameters may provide better performance for global DVA. The occupancy count threshold used for cluster DVA was set to 1000 for all experiments (this corresponds to between around 10 and 30 clusters depending on the speaker). The initial value for the pre-processor weights $\alpha_{i,r}$ was set to 1 for the ML-based adaptation. For MMI and dMMI, the initial value was set to the weights obtained from ML-DVA after convergence.

Figure 1 clearly shows that MMI based DVA fails to bring performance improvement compared to ML. This may be partly caused by the errors in the transcriptions. Note that the error rate of the transcriptions was 14.93 % (as shown in Table 1). In contrast, using the dMMI criterion for cluster DVA provides consistent performance improvement compared to ML or MMI criteria. Note that although we obtained best performance with the above dMMI parameters, other values around $\upsilon_1 = -4$ can provide similar performance, which indicates that dMMI-DVA is not over sensitive to the choice of margin parameters. We also observe that performance degrades when using global DVA. Note that by setting $\upsilon_1 = -1$,

[1] Global DVA consists of the originally proposed DVA were the pre-processor uncertainty weights α_i are common to all the Gaussians of the acoustic model [5].

Table 1. *Average keyword error rate for the development and evaluation sets of the CHiME challenge task.*

	Dev.	Eval.
Noisy	30.39 %	31.00 %
Enhanced (no adaptation)	14.93 %	12.75 %
ML-DVA (global)	13.03 %	11.49 %
ML-DVA (cluster 1000)	12.96 %	11.41 %
MMI-DVA (global)	13.39 %	11.88 %
MMI-DVA (cluster 1000)	13.41 %	12.24 %
dMMI-DVA (global)	13.35 %	11.67 %
dMMI-DVA (cluster 1000)	**12.58 %**	**11.12 %**

we could obtain some improvement with global DVA. This suggest that the values of the margins may depend on the occupancy count threshold used.

Table 1 summarizes the results obtained for the development and evaluation sets. It confirms that improvement can be obtained with the dMMI criterion for cluster DVA for both the development and evaluation sets. These results are close to the state of the art results for this task [18]. However, this paper includes only a part of the system we proposed for the CHiME challenge. By including the proposed method with the other parts of our CHiME system we can expect further improvement.

5. CONCLUSION

We proposed using the dMMI discriminative criterion for DVA when dealing with unsupervised adaptation. dMMI possesses an inherent mechanism to compensate the effect of errors in the transcriptions, which enables unsupervised discriminative adaptation. The effect of dMMI for DVA was confirmed in a noisy speech command recognition task.

Future work will include combining dMMI based DVA with discriminative adaptation of the mean parameters of the Gaussians of the acoustic model, as well as testing the proposed method with large vocabulary tasks.

6. REFERENCES

[1] Leggetter C. J. and Woodland P. C., “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Computer Speech & Language, vol. 9, no. 2, pp. 171-185, 1995.

[2] Droppo, J., Acero, A. and Deng, L., “Uncertainty decoding with SPLICE for noise robust speech recognition,” Proc. ICASSP’02, vol. 1, pp. 57-60, 2002.

[3] Kolossa, D., Haeb-Umbach, R. (eds.), “Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications,” Springer Verlag, 380 pages, 2011.

[4] 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.

[5] Delcroix, M., Nakatani, T. and Watanabe, S., “Static and dynamic variance compensation for recognition of reverberant speech with dereverberation preprocessing,” IEEE Trans. ASLP, vol. 17, no. 2, pp. 324-334, 2009.

[6] Delcroix, M., Watanabe, S., Nakatani, T. and Nakamura, A., “Discriminative approach to dynamic variance adaptation for noisy speech recognition,” In Proc. HSCMA’11, pp. 7-12, 2011.

[7] Wang, L. and Woodland, P.C., “MPE-based discriminative linear transform for speaker adaptation,” In Computer Speech and Language, vol. 22, pp. 256-272, 2008.

[8] McDermott, E., Watanabe, S. and Nakamura, A., “Discriminative training based on an integrated view of MPE and MMI in margin and error space,” In Proc. ICASSP’10, pp. 4894 - 4897, 2010.

[9] Povey, D. and Woodland, P., “Minimum phone error and I-smoothing for improved discriminative training,” In Proc. ICASSP’02, vol. 1, pp. 105-108 , 2002.

[10] Povey, D. , Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G. and Visweswariah, K., “Boosted MMI for model and feature-space discriminative training,” In Proc. ICASSP’08, pp. 4057-4060, 2008.

[11] Nakamura, A., McDermott, E., Watanabe, S. and Katagiri, S., “A unified view for discriminative objective functions based on negative exponential of difference measure between strings,” In Proc. ICASSP’09, pp. 1633-1636, 2009.

[12] Zheng, J. and Stolcke, A., “Improved discriminative training using phone lattices,” In Proc. Interspeech’05, pp. 2125-2128, 2005.

[13] Riedmiller, M. and Braun, H., “A direct adaptive method for faster backpropagation learning: The RPROP algorithm,” In Proc. ICNN’93, pp. 586-591, 1993.

[14] Kanevsky, D., Heigold, G., Wright, S. and Ney, H., “Overview of large scale optimization for discriminative training in speech recognition,” In Proc. ICASSP’12, pp. 5233-5236, 2012.

[15] Igel, C. and Husken, M., “Empirical evaluation of the improved RPROP learning algorithms,” In Neuralcomputing, vol. 50, pp. 105-123, 2003.

[16] Christensen, H., Barker, J., Ma, N., and Green, P., “The CHiME corpus: a resource and a challenge for computational hearing in multisource environments,” In Proc. Interspeech’10, pp. 1918-1921, 2010.

[17] Nakatani, T., Yoshioka, T., Araki, S., Delcroix, M. and Fujimoto, M., “LogMax observation model with MFCC-based spectral prior for reduction of highly nonstationary ambient noise,” In Proc. ICASSP’12, pp. 4029-4032, 2012.

[18] “PASCAL CHiME challenge results,” `http://spandh.dcs.shef.ac.uk/projects/chime/PCC/results.html`, Cited 13 June 2012.

[19] Hori, T., Hori, C., Minami, Y. and Nakamura, A., “ Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition,” IEEE Trans. ASLP, vol. 15, no. 4, pp. 1352-1365, 2007.