STEREO-BASED STOCHASTIC MAPPING WITH CONTEXT USING PROBABILISTIC PCA FOR NOISE ROBUST AUTOMATIC SPEECH RECOGNITION

Xiaodong Cui[1]*, Mohamed Afify*[2] *and Bowen Zhou*[1]

IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598, USA[1]
Orange Lab, Smart Village, Cairo, Egypt[2]

ABSTRACT

In this paper we investigate stereo-based stochastic mapping (SSM) with context for the noise robustness of automatic speech recognition, especially under unseen conditions. Probabilistic PCA (PPCA) is used in the SSM framework to reduce the high dimensionality of the noisy speech features with context and derive an eigen representation in the noisy feature space for the prediction of clean features. To reduce the computational cost in training, an approximation by single-pass re-training is considered for the estimation of joint GMM. We also show that the SSM estimate under the minimum mean square error (MMSE) in a space where low dimensional representation of clean speech and uncorrelated additive noise can be assumed is related to the subspace speech enhancement. Experiments on large vocabulary continuous speech recognition tasks observe gains from the proposed approach under the conditions with seen, unseen and real noise.

Index Terms— stereo-based stochastic mapping, probabilistic PCA, noise robustness, LVCSR, subspace speech enhancement

1. INTRODUCTION

Stereo-based stochastic mapping (SSM) [1] is one of the noise compensation approaches for automatic speech recognition (ASR). It belongs to a family of stereo-based algorithms relying on the stereo data to learn the statistical relationship between the observed channel and target channel. Such algorithms find wide applications in various areas such as noise robust ASR [1][2][3], audio bandwidth extension [4] and voice conversion [5].

SSM models the joint distribution of clean and noisy speech features by a Gaussian mixture model (GMM) based on which the clean features are predicted from the observed noisy features under certain criterion, for instance, minimum mean square error (MMSE) or maximum a posteriori (MAP). In this work, we aim to achieve decent ASR performance under seen and unseen noisy conditions by noise compensation taking into account the acoustic context in the SSM framework. To that end, the data from the noisy channel are collected from a variety of conditions covering diverse types of noise and signal-to-noise ratios (SNRs). The compensation of noisy features from unseen conditions can be approximated by a condition corresponding to the closest point, in the sense of MMSE, in the space spanned by the seen conditions. To capture the statistical structure of the space of the seen conditions, full covariance is used in the estimation of the joint GMM instead of diagonal covariance used in most of the previous stereo-based approaches [1][2][4].

Including acoustic context in the SSM will dramatically increase the dimensionality of the joint feature space, especially when the context is wide, which may raise an issue for reliable parameter estimation. Probabilistic PCA (PPCA) [6] is employed to reduce the dimensionality of the noisy features with context and, in the meantime, still keep the most significant structural information of the space by the eigen representation for the prediction of clean features. Carried out in a probabilistic framework, mixture of PPCA can be seamlessly embedded into the estimation of the joint GMM. Approximation by single-pass re-training is also investigated to reduce the computational cost and turn-around training time.

As an interesting note, we will show the connection between the MMSE estimate of SSM and subspace speech enhancement [7][8]. In a space where low dimensional representation of clean speech and uncorrelated additive noise can be assumed the MMSE estimate of SSM without context is analogous to the subspace speech enhancement under colored noise. With context being considered, the MMSE based SSM can be seen as an extension of the latter.

The remainder of the paper is organized as follows. Section 2 gives the MMSE estimate of SSM with context. Section 3 describes the eigen representation of the noisy features with context by PPCA and approximation by single-pass re-training for training speedup. Section 4 shows the relationship between the MMSE estimate of SSM and subspace speech enhancement. Experimental results on large vocabulary continuous speech recognition (LVCSR) tasks under seen, unseen and real noisy conditions are presented in Section 5.

2. SSM WITH CONTEXT

Let $z_t = (x_t, \bar{y}_t)$ be the joint vector with x_t being the clean feature at time t and $\bar{y}_t$ being the noisy feature at time t with $2c$ ($c > 0$) frames of context

$$\bar{y}_t = [y_{t-c} \ \cdots \ y_{t-1} \ \ y_t \ \ y_{t+1} \ \cdots \ y_{t+c}]. \tag{1}$$

The distribution of z_t is assumed to be GMM as shown in Eq.2

$$p(z_t) = \sum_{k=1}^{K} w_k \mathcal{N}(z_t; \mu_{z,k}, \Sigma_{zz,k}) \tag{2}$$

where K is the number of Gaussian components, w_k, $\mu_{z,k}$, and $\Sigma_{zz,k}$ are the mixture weight, mean, and covariance of each component, respectively. $p(z_t)$ can be estimated by the EM algorithm [9]. The mean and covariance have a partition with respect to the clean feature x and noisy feature with context $\bar{y}$

$$\mu_{z,k} = \begin{bmatrix} \mu_{x,k} \\ \mu_{\bar{y},k} \end{bmatrix}, \qquad \Sigma_{zz,k} = \begin{bmatrix} \Sigma_{xx,k} & \Sigma_{x\bar{y},k} \\ \Sigma_{\bar{y}x,k} & \Sigma_{\bar{y}\bar{y},k} \end{bmatrix}. \tag{3}$$

Given the observed noisy speech feature with context, $\bar{y}_t$, the MMSE estimate of the clean speech x_t is [1]

$$\hat{x}_t = E[x_t|\bar{y}_t]. \tag{4}$$

From the joint GMM in Eq.2,

$$\hat{x}_t = \sum_k p(k|\bar{y}_t) \left[\mu_{x,k} + \Sigma_{x\bar{y},k}\Sigma^{-1}_{\bar{y}\bar{y},k}(\bar{y}_t - \mu_{\bar{y},k})\right] \tag{5}$$

The posterior probability, $p(k|\bar{y}_t)$, is computed against the marginal noisy distribution $p(\bar{y}_t)$ of the joint distribution $p(z_t)$.

3. EIGEN REPRESENTATION BY PPCA

The noisy channel of the stereo data in the training encompasses a wide range of conditions with diverse types of noise and SNRs. As illustrated in Fig.1, the same acoustic sound under various noisy conditions may possess distinct structures in the feature space. Therefore, even though diagonal covariance Gaussians are a reasonable assumption sometimes for the clean condition, full covariance Gaussians will give a more accurate description of the data structure in the multi-condition case. Furthermore, with the acoustic context being taken into consideration, full covariance Gaussians can also model the correlation of adjacent frames. This is the motivation behind using a full covariance joint GMM in Eq.2 in this work.

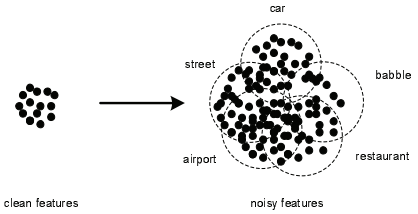

Fig. 1. Illustration of the effect of clean speech corrupted under multiple noisy conditions.

SSM with context may lead to a high dimensionality in $\bar{y}$, especially when the context is wide. With full covariance assumed, a reliable estimation of the joint GMM can be an issue. To deal with it, PPCA is used for the eigen representation of $\bar{y}$ space to reduce the parameter size to be estimated.

Let ξ be a latent variable. Assume the distribution of $\bar{y}$ given ξ in the kth Gaussian component of GMM is

$$p(\bar{y}|\xi_k) \sim \mathcal{N}(\bar{y}; W_k\xi_k + \mu_{\bar{y},k}, \sigma_k^2 I) \tag{6}$$

and the latent variable ξ_k itself obeys a Gaussian distribution with zero mean and unit covariance

$$p(\xi_k) \sim \mathcal{N}(\xi_k; 0, I) \tag{7}$$

From Eqs. 6 and 7, one has

$$p_k(\bar{y}) \sim \mathcal{N}(\bar{y}; \mu_{\bar{y},k}, W_k W_k^{\mathsf{T}} + \sigma_k^2 I) \tag{8}$$

Given the PPCA assumptions on $\bar{y}$ in Eqs. 6-8, the parameters of the joint GMM in Eq.2 has the following update equations based on the EM algorithm:

$$\mu_{x,k} = \frac{\sum_{t=1}^{T}\gamma_k(t)x_t}{\sum_{t=1}^{T}\gamma_k(t)}, \quad \mu_{\bar{y},k} = \frac{\sum_{t=1}^{T}\gamma_k(t)\bar{y}_t}{\sum_{t=1}^{T}\gamma_k(t)} \tag{9}$$

$$\Sigma_{xx,k} = \frac{\sum_{t=1}^{T}\gamma_k(t)(x_t-\mu_{x,k})(x_t-\mu_{x,k})^{\mathsf{T}}}{\sum_{t=1}^{T}\gamma_k(t)} \tag{10}$$

where $\gamma_k(t) = p(k|z_t)$ is the posterior probability of component k given the joint vector z_t.

The autocovariance of $\bar{y}$ under PPCA is

$$\tilde{\Sigma}_{\bar{y}\bar{y},k} = W_k W_k^{\mathsf{T}} + \sigma_k^2 I \tag{11}$$

According to [6], W_k and σ_k^2 can be computed as

$$W_k = U_{k,q}(\Lambda_{k,q} - \sigma_k^2 I)^{1/2}, \quad \sigma_k^2 = \frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_{k,j} \tag{12}$$

where $\Lambda_{k,q} = \text{diag}\{\lambda_{k,1}, \cdots, \lambda_{k,q}\}$ contains the leading eigenvalues and $U_{k,q}$ contains the eigenvectors corresponding to the leading eigenvalues of the autocovariance matrix

$$\Sigma_{\bar{y}\bar{y},k} = \frac{\sum_{t=1}^{T}\gamma_k(t)(\bar{y}_t-\mu_{\bar{y},k})(\bar{y}_t-\mu_{\bar{y},k})^{\mathsf{T}}}{\sum_{t=1}^{T}\gamma_k(t)} \tag{13}$$

In Eq.12, d is the dimension of $\bar{y}$ and q is the number of principal components to construct the eigen subspace.

The cross-covariance of x and $\bar{y}$ with the latent variable ξ_k is computed as

$$\begin{aligned}\tilde{\Sigma}_{x\bar{y},k} &= \frac{\sum_{t=1}^{T}\gamma_k(t)\int_{\xi_k}(x_t-\mu_{x,k})[W_k\xi_k p(\xi_k|\bar{y}_t)]^{\mathsf{T}}d\xi_k}{\sum_{i=1}^{N}\gamma_k(i)} \\ &= \frac{\sum_{t=1}^{T}\gamma_k(t)(x_t-\mu_{x,k})(W_k E[\xi_k|\bar{y}_t])^{\mathsf{T}}}{\sum_{t=1}^{T}\gamma_k(t)}\end{aligned} \tag{14}$$

Since

$$p(\xi_t|\bar{y}_t) \sim \mathcal{N}(\xi_t; M_k^{-1}W_k^{\mathsf{T}}(\bar{y}_t-\mu_{\bar{y},k}), \sigma_k^2 M_k^{-1}) \tag{15}$$

hence

$$E[\xi_k|\bar{y}_t] = M_k^{-1}W_k^{\mathsf{T}}(\bar{y}_t-\mu_{\bar{y},k}) \tag{16}$$

where

$$M_k = W_k^{\mathsf{T}}W_k + \sigma_k^2 I \tag{17}$$

Therefore,

$$\begin{aligned}\tilde{\Sigma}_{x\bar{y},k} &= \frac{\sum_{t=1}^{T}\gamma_k(t)(x_t-\mu_{x,k})\left[W_k M_k^{-1}W_k^{\mathsf{T}}(\bar{y}_t-\mu_{\bar{y},k})\right]^{\mathsf{T}}}{\sum_{t=1}^{T}\gamma_k(t)} \\ &= \frac{\sum_{t=1}^{T}\gamma_k(t)(x_t-\mu_{x,k})(\bar{y}_t-\mu_{\bar{y},k})^{\mathsf{T}}W_k M_k^{-1}W_k^{\mathsf{T}}}{\sum_{t=1}^{T}\gamma_k(t)} \\ &= \Sigma_{x\bar{y},k}W_k M_k^{-1}W_k^{\mathsf{T}}\end{aligned} \tag{18}$$

where

$$\Sigma_{x\bar{y},k} = \frac{\sum_{t=1}^{T}\gamma_k(t)(x_t-\mu_{x,k})(\bar{y}_t-\mu_{\bar{y},k})^{\mathsf{T}}}{\sum_{t=1}^{T}\gamma_k(t)} \tag{19}$$

With $\mu_{x,k}$, $\mu_{\bar{y},k}$, $\Sigma_{xx,k}$, $\tilde{\Sigma}_{yy,k}$ and $\tilde{\Sigma}_{x\bar{y},k}$ in place, the kth Gaussian component in the joint GMM in Eq.2 can be written as

$$\begin{bmatrix}x\\ \bar{y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}x\\ \bar{y}\end{bmatrix}; \begin{bmatrix}\mu_{x,k}\\ \mu_{\bar{y},k}\end{bmatrix}, \begin{bmatrix}\Sigma_{xx,k} & \tilde{\Sigma}_{x\bar{y},k}\\ \tilde{\Sigma}_{x\bar{y},k}^{\mathsf{T}} & \tilde{\Sigma}_{\bar{y}\bar{y},k}\end{bmatrix}\right) \tag{20}$$

It is trivial to see that when $\sigma_k^2 \to 0$,

$$\tilde{\Sigma}_{x\bar{y},k} \to \Sigma_{x\bar{y},k} \tag{21}$$

The above iterative parameter update for the GMM with full covariance can be time-consuming when the context is wide and the number of Gaussian components K is large. To speed up the training process, single-pass re-training [10] is used to reduce the computational cost. A full covariance GMM model $p(y_t)$ with the same number of Gaussian components as Eq.2 is built separately on the noisy features y_t without context. The posterior probabilities $\gamma_k(t)$ in Eqs. 9, 10, 13 and 19 are approximated by computing y_t against $p(y_t)$ instead of z_t against $p(z_t)$

$$\gamma_k(t) \approx p(k|y_t) \tag{22}$$

Since the computation of $\gamma_k(t)$ is conducted in a lower dimensional space and the parameters are updated in only one iteration, the turnaround training time is significantly reduced.

4. SSM AND SUBSPACE SPEECH ENHANCEMENT

In this section, we draw a connection between the MMSE estimate of SSM and subspace speech enhancement [7][8]. Although speech enhancement techniques (including spectral subtraction as a special case) may not always help for the noise robustness of ASR, we would like to point out the relationship between the two as a note from the speech enhancement perspective.

Suppose in a space $\mathcal{R}^d$ the clean speech can be assumed to occupy a low dimensional space

$$x = \Psi s \tag{23}$$

where s is an m-dimensional vector ($m < d$) and Ψ is a matrix with linearly independent component vectors. The noisy speech y can be expressed as

$$y = x + n \tag{24}$$

where the additive noise n is assumed to be uncorrelated with x. For instance, linear spectral or linear power spectral domain is a reasonable space satisfying these assumptions. In such a space, when the context in Eq.1 is zero, the SSM estimate in each Gaussian component (dropping index k for notation simplicity) in Eq.5 is

$$\hat{x} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y) \tag{25}$$

Under the assumption of uncorrelation between x and n, one has

$$\Sigma_{xy} = \Sigma_{xx} + \Sigma_{xn} \approx \Sigma_{xx}, \quad \Sigma_{yy} \approx \Sigma_{xx} + \Sigma_{nn} \tag{26}$$

First, assume n is white noise, based on the eigen-decomposition of Σ_{xx}, it is easy to show that

$$\begin{aligned} \Sigma_{xx} &= U\Lambda_x U^{\mathsf{T}}, \quad \Sigma_{nn} = U(\sigma_w^2 I)U^{\mathsf{T}} \\ \Sigma_{yy} &= U\Lambda_y U^{\mathsf{T}} = U(\Lambda_x + \sigma_w^2 I)U^{\mathsf{T}} \end{aligned} \tag{27}$$

where U is the eigenvector matrix and the diagonal matrix $\Lambda_x = \text{diag}(\lambda_1^x, \cdots, \lambda_m^x, 0, \cdots, 0)$ contains the eigenvalues of Σ_{xx}. Under the assumption of Eq.23, Λ_x has only m non-zero eigenvalues. Since n is white, it shares the same eigenvectors as Σ_{xx} and σ_w^2 is its energy. Therefore, Σ_{yy} in Eq.27 can be written as

$$[U_m, U_{d-m}] \left(\begin{bmatrix} \Lambda_{x,m} & 0 \\ 0 & 0 \end{bmatrix} + \sigma_w^2 \begin{bmatrix} I_m & 0 \\ 0 & I_{d-m} \end{bmatrix} \right) \begin{bmatrix} U_m^{\mathsf{T}} \\ U_{d-m}^{\mathsf{T}} \end{bmatrix} \tag{28}$$

which shows that the noisy speech y space can be decomposed into an m-dimensional signal-plus-noise subspace spanned by U_m and its complimentary noise-only subspace spanned by U_{d-m}. Accordingly, Eq.25 can be written as

$$\hat{x} = \mu_x + U_m G_m U_m^{\mathsf{T}}(y - \mu_y) \tag{29}$$

where G_m is an m-dimensional diagonal matrix with gain g_i on the diagonal

$$g_i = \frac{\lambda_i^x}{\lambda_i^x + \sigma_w^2} \tag{30}$$

which indicates that the denoising is carried out only in the m-dimensional signal-plus-noise subspace. This is analogous to the conventional subspace speech enhancement proposed in [7].

In most cases, noise n is not white so that Σ_{xx} and Σ_{nn} will not have the same eigenvectors. A generalized subspace speech enhancement is discussed in [8] to deal with colored noise n based on the simultaneous diagonalization. Given the two positive-definite matrices Σ_{xx} and Σ_{nn}, there exists a matrix V such that

$$V^{\mathsf{T}}\Sigma_{xx}V = \Lambda_x, \quad V^{\mathsf{T}}\Sigma_{nn}V = I \tag{31}$$

In fact, V is composed of eigenvectors of matrix $\Sigma_{nn}^{-1}\Sigma_{xx}$. The colored noise n is whitened into white noise in the space spanned by V. Given Eq.31, the covariance matrices Σ_{xx} and Σ_{yy} can be written as

$$\Sigma_{xy} \approx \Sigma_{xx} = V^{-\mathsf{T}}\Lambda_x V^{-1} \tag{32}$$

$$\Sigma_{yy} \approx \Sigma_{xx} + \Sigma_{nn} = V^{-\mathsf{T}}(\Lambda_x + I)V^{-1} \tag{33}$$

Accordingly, Eq.25 can be rewritten as

$$\begin{aligned} \hat{x} &\approx \mu_x + \Sigma_{xx}\Sigma_{yy}^{-1}(y - \mu_y) \\ &= \mu_x + V^{-\mathsf{T}}\Lambda_x(\Lambda_x + I)^{-1}V^{\mathsf{T}}(y - \mu_y) \\ &= \mu_x + V_m^{-\mathsf{T}}G_m V_m^{\mathsf{T}}(y - \mu_y) \end{aligned} \tag{34}$$

where m is the rank of Σ_{xx} and the gain on the diagonal of the diagonal matrix G_m is

$$g_i = \frac{\lambda_i^x}{\lambda_i^x + 1} \tag{35}$$

In this case, the colored noise is whitened and normalized to unit energy. The denoising is again carried out in the m-dimensional signal-plus-noise subspace spanned by V_m.

Based on the above discussion, in a particular space where assumptions can be made on the low dimensional representation of the clean speech x and uncorrelated additive noise n, the MMSE estimate of SSM in each Gaussian component is analogous to the subspace speech enhancement. However, the SSM framework has the following distinctions. The MMSE estimate of SSM gives rise to a mixture of linear signal estimator weighted by the posteriors. In addition, the stereo data in the SSM provides an additional channel of clean speech signals from which Σ_{xx} can be precisely estimated. This is the advantage of using stereo data comparing to the conventional subspace speech enhancement in which Σ_{xx} is computed by removing Σ_{nn} from Σ_{yy} and Σ_{nn} is estimated from the noise samples of the speech-absent frames. Furthermore, in the SSM framework, context information from adjacent frames can be trivially taken into account as shown in Section 2, which is not straightforward for the conventional subspace speech enhancement. Therefore, the MMSE estimate of SSM with context in such a space can be considered an extension of the conventional subspace speech enhancement from that perspective.

5. EXPERIMENTAL RESULTS

Experiments on English LVCSR tasks were conducted on the proposed approach. The acoustic model trained on clean speech signals has 5K quinphone states and 100K Gaussians. The trigram language model with 330K n-grams is built on a vocabulary of 45K words and 56K pronunciations. The feature space is constructed by splicing 9 frames of 24-dim PLP features and projecting down to a 40-dim linear discriminant analysis (LDA) space with a global semi-tied covariance (STC) transformation. The acoustic model is trained with both feature and model space discriminative training (FMMI and BMMI) [11]. The SSM is performed in the final 40-dim FMMI space where the noisy features are compensated and decoded using the clean acoustic model.

For the training of SSM, the clean channel consists of 60 hours of continuous speech signals. The noisy channel is artificially generated by corrupting the clean speech with various types of noise from the NOISEX-92 dataset including *M109*, *Buccaneer*, *Leopard*, *wheel carrier*, *destroyer operation room*, *HF radio*, *babble*, *factory*, *car* and *white noise*. Each utterance in the training set is randomly corrupted by one type of the noise from the above set at a random

SNR from 25dB to 10dB. The test sets are composed of three scenarios: seen conditions (Set A), unseen conditions (Set B) and real conditions (Set C). In Set A, *Buccaneer*, *destroyer operation room*, *babble*, *factory* and *car noise* are added to the DARPA Transtac Nov08 offline test set (10 speakers, 0.7 hours) at an SNR randomly selected from the range of 20dB to 5dB. In Set B, *Lynx*, *machine gun*, *STITEL*, *F-16* and *pink noise* are added to the DARPA Transtac Oct09 offline test set (11 speakers, 0.6 hours) at an SNR randomly selected from the range of 20dB to 5dB. The Set C (7 speakers, 1.9 hours) consists of the speech data recorded under the humvee-tank noise with SNRs estimated at 5-8dB.

As a baseline, Table 1 shows the word error rates (WERs) of the three noisy test sets without SSM compensation. In particular, for the two artificially generated noisy test sets A and B, WERs of the original clean speech are also presented.

dataset	Set A	Set B	Set C
clean condition	13.8	21.2	–
noisy condition	37.3	45.7	38.2

Table 1. WERs(%) of the baseline for clean and noisy conditions without SSM compensation.

dataset	Set A	Set B	Set C
K=256, c=0, q=40	**27.8**	**37.6**	**30.0**
K=256, c=1, q=120	27.2	37.3	27.5
K=256, c=1, q=100	27.4	37.0	26.8
K=256, c=1, q=90	27.2	**36.7**	**26.6**
K=256, c=1, q=80	**26.9**	37.0	26.9

Table 2. WERs(%) of SSM compensation using 256 Gaussian components without ($c=0$) and with ($c=1$) context. Results of different numbers of principal components in PPCA are shown when context is 1.

dataset	Set A		Set B		Set C	
components	1024	2048	1024	2048	1024	2048
c=0, q=40	**29.1**	**29.4**	**38.0**	**38.0**	**28.9**	**28.9**
c=1, q=120	28.7	28.8	37.8	37.5	27.9	27.6
c=1, q=100	28.1	27.8	37.1	36.9	**27.3**	27.3
c=1, q=80	**28.0**	**27.8**	**36.9**	**36.9**	27.4	**27.1**
c=1, q=60	28.5	28.1	37.4	36.9	27.4	27.5
c=2, q=200	28.8	29.0	37.7	37.8	27.3	27.4
c=2, q=160	27.9	27.8	37.3	36.9	26.9	26.6
c=2, q=140	27.8	27.8	**37.2**	**36.9**	**26.9**	26.9
c=2, q=120	**27.6**	**27.7**	37.3	37.0	26.9	**26.5**
c=2, q=100	28.1	27.9	37.4	37.0	27.3	26.8

Table 3. WERs (%) of SSM compensation with various Gaussian components, contexts and principal components of PPCA. The estimation of the joint GMM of SSM is approximated by single-pass re-training.

Table 2 shows the WERs with SSM compensation. There are 256 Gaussian components (K) in the joint GMM of SSM. When $c=0$, no acoustic context information is used in the SSM and, accordingly, no PPCA is applied in this case. The number of principal components (PCs) q is set to 40 which is equal to the dimensionality of the noisy features. When the context is set to 1 ($c=1$), the dimensionality of the noisy features with context is 120. Table 2 demonstrates the performance of PPCA with various numbers (q) of PCs. The best performance is achieved in the three test sets when q is somewhere around 90 to 80. From the table, SSM with context obtains superior performance over SSM without context and it helps for all three test scenarios.

Table 3 shows the WERs with SSM compensation under various Gaussian components (K), contexts (c) and PCs (q). The training of the joint GMM of SSM in this case is approximated by single-pass re-training. From the table, it can be observed that single-pass re-training sacrifices certain degree of performance for the training speedup. With wider context, better performance can be yielded. The results indicate that, with 1024 or 2048 Gaussian components, best WERs under PPCA are achieved when q is around 100 to 80 for context $c=1$ and when q is around 140 to 120 for context $c=2$. Again, SSM with context helps for all three test scenarios.

In summary, SSM with context using full covariance is investigated in this paper. PPCA is employed to reduce the dimensionality in the GMM framework for reliable parameter estimation. The estimation of the joint GMM is approximated by single-pass training for a significant reduction of training time. In current experiments, however, the noise compensation time grows when the context gets wider. Strategies for compensation speedup will be studied in the future work.

6. REFERENCES

[1] M. Afify, X. Cui, and Y. Gao, "Stereo-based stochastic mapping for robust speech recognition," *Proc. of ICASSP*, pp. 377–380, 2007.

[2] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the AURORA 2 database," *Proc. of Eurospeech*, pp. 217–220, 2001.

[3] V. Stouten, H. V. hamme, and P. Wambacq, "Joint removal of additive and convolutional noise with model-based feature enhancement," *Proc. of ICASSP*, pp. 949–952, 2004.

[4] M. L. Seltzer, A. Acero, and J. Droppo, "Robust bandwidth extension of noise-corrupted narrowband speech," *Proc. of Interspeech*, pp. 1509–1512, 2005.

[5] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[6] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society, Series B*, vol. 61, no. 3, pp. 611–622, 1999.

[7] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–228, 1992.

[8] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, 2003.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[10] S. Young et. al., "The HTK Book," *University of Cambridge*, 2005.

[11] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," *Proc. of ICASSP*, pp. 4057–4060, 2008.