Efficient Use of DNN Bottleneck Features in Generalized Variable Parameter HMMs for Noise Robust Speech Recognition

Rongfeng Su[1,3]*, Xurong Xie*[1,3]*, Xunying Liu*[1,2] *& Lan Wang*[1,3]

[1]Key Laboratory of Human-Machine Intelligence-Synergy Systems,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]Cambridge University Engineering Dept, Trumpington St., Cambridge, CB2 1PZ U.K.
[3]The Chinese University of Hong Kong, Hong Kong, China

rf.su@siat.ac.cn, xr.xie@siat.ac.cn, xl207@cam.ac.uk, lan.wang@siat.ac.cn

Abstract

Recently a new approach to incorporate deep neural networks (DNN) bottleneck features into HMM based acoustic models using generalized variable parameter HMMs (GVP-HMMs) was proposed. As Gaussian component level polynomial interpolation is performed for each high dimensional DNN bottleneck feature vector at a frame level, conventional GVP-HMMs are computationally expensive to use in recognition time. To handle this problem, several approaches were exploited in this paper to efficiently use DNN bottleneck features in GVP-HMMs, including model selection techniques to optimally reduce the polynomial degrees; an efficient GMM based bottleneck feature clustering scheme; more compact GVP-HMM trajectory modelling for model space tied linear transformations. These improvements gave a total of 16 time speed up in decoding time over conventional GVP-HMMs using a uniformly assigned polynomial degree. Significant error rate reductions of 15.6% relative were obtained over the baseline tandem HMM system on the secondary microphone channel condition of Aurora 4 task. Consistent improvements were also obtained on other subsets.

Index Terms: generalized variable parameter HMM, deep neural network, bottleneck features, robust speech recognition

1. Introduction

Recently deep neural networks (DNNs) have become increasingly popular for acoustic modelling in automatic speech recognition (ASR) systems [1, 2, 3, 4, 5]. In order to incorporate DNNs, or multi-layer perceptrons (MLPs) in general, into HMM based generative acoustic models, two approaches can be used. In a hybrid architecture HMM state emission probabilities are estimated using DNNs [6]. In tandem systems, an MLP or DNN is used as a feature extractor and trained to produce phoneme posterior probabilities. The resulting probabilistic features [7], or bottleneck features [8] are used to train the back-end GMM-HMMs. As these features capture additional information complementary to standard front-ends, they are often concatenated in tandem systems.

In order to better capture the complex relationship between standard acoustic and bottleneck features, a new approach to incorporate DNN bottleneck features into HMM based acoustic models using generalized variable parameter HMMs (GVP-HMMs) [9, 10, 11, 12, 13, 14] was recently proposed [15]. Motivated by speech production, this technique can fully exploit the hidden variability and contextual factors encoded in the DNN bottleneck features. These features are used as influencing factors to directly introduce controllability to the underlying GVP-HMM based generative acoustic models.

One important practical issue associated with the GVP-HMM based tandem approach is the computational cost incurred during recognition time. As Gaussian component parameter polynomial interpolation is performed dimension wise for each high dimensional DNN bottleneck feature vector at a frame level, conventional GVP-HMMs are computationally very expensive to use in practice. In order to address this issue, several approaches were exploited in this paper to improve the efficiency of using DNN bottleneck features in GVP-HMMs. These include a model selection technique that automatically determined the optimal polynomial degrees in GVP-HMMs, which reduced the number of polynomial interpolation operations performed by over 90%; an efficient GMM based bottleneck feature clustering scheme that gave a further 45% reduction in decoding time; a more compact GVP-HMM trajectory modelling for model space tied linear transforms was also used. The combination of these approaches allow experiments to be conducted efficiently on the Aurora 4 task. A total of 16 time speed up in decoding time over conventional GVP-HMMs using a uniformly assigned polynomial degree was obtained. Significant error rate reductions of 15.6% relative were obtained over the baseline tandem HMM system on the secondary microphone channel condition.

The rest of this paper is organized as follows. GVP-HMMs are reviewed in section 2. Three techniques to improve the efficiency of using DNN bottleneck features in GVP-HMMs are proposed in section 3. A range of tandem GVP-HMM configurations are described in section 4. In section 5 the performance of various tandem GVP-HMM systems efficiently using DNN bottleneck features are evaluated on Aurora 4. Section 6 is the conclusion and future research.

2. Generalized Variable Parameter HMMs

Generalized variable parameter HMMs (GVP-HMMs) [9, 10, 11, 12] explicitly model the trajectory of acoustic parameters, or more compact tied linear transformations, that vary with respect to the underlying influence factors, such as environment noise condition represented by the signal-to-noise ratio (SNR) [9, 10, 11, 12, 13, 14, 16, 17], articulatory features for speech synthesis [18], and more recently DNN bottleneck fea-

This work is supported by National Natural Science Foundation of China (NSFC 61135003, 91420301), Shenzhen Fundamental Research Program JCYJ20130401170306806.

tures [15].

For a D dimensional observation $\boldsymbol{o}_t$ emitted from Gaussian mixture component m, assuming Pth order polynomials modelling a total of N regression variables are used, the form of GVP-HMMs considered in this paper is given by

$$\boldsymbol{o}^{(t)} \sim p\left(\boldsymbol{o}^{(t)}; \boldsymbol{\mu}^{(m)}(\mathbf{v}_t), \boldsymbol{\Sigma}^{(m)}(\mathbf{v}_t)\right). \quad (1)$$

$\mathbf{v}_t^\top$ is a $(P \times N + 1)$ dimensional Vandermonde vector [19],

$$\mathbf{v}_t^\top = \left[1,\ \tilde{\mathbf{f}}_{t,1},\ ..., \tilde{\mathbf{f}}_{t,p},\ ..., \tilde{\mathbf{f}}_{t,P}\right]^\top. \quad (2)$$

Its N dimensional pth order subvector is defined as $\tilde{\mathbf{f}}_{t,p} = [v_{t,1}^p,\ ..., v_{t,j}^p,\ ..., v_{t,N}^p]^\top$, where $v_{t,j}$ is the jth element of an N dimensional factor vector Gaussian parameters are conditioned on at frame t, for example, the DNN bottleneck features,

$$\vec{\mathbf{f}}_t = [v_{t,1},\ ...,\ v_{t,j}, ...,\ v_{t,N}]^\top. \quad (3)$$

$\boldsymbol{\mu}^{(m)}(\cdot)$ and $\boldsymbol{\Sigma}^{(m)}(\cdot)$ are the P^{th} order mean and covariance trajectory polynomials of component m respectively. When diagonal covariances are used, the trajectories of the i^{th} dimension of the mean and variance parameters are computed as

$$\begin{aligned} \mu_i^{(m)}(\mathbf{v}_t) &= \mathbf{v}_t \cdot \mathbf{c}^{(\mu_i^{(m)})} \\ \sigma_{i,i}^{(m)}(\mathbf{v}_t) &= \breve{\sigma}_{i,i}^{(m)} \mathbf{v}_t \cdot \mathbf{c}^{(\sigma_{i,i}^{(m)})} \end{aligned} \quad (4)$$

where $\mathbf{c}^{(\cdot)}$ is a $(P \times N + 1)$ dimensional polynomial coefficient vector and $\breve{\sigma}_{i,i}^{(m)}$ is the conventional HMM variance estimate.

For the form of GVP-HMMs of equation (1) the associated ML auxiliary function is given by [9, 10, 20],

$$\mathcal{Q}_{\mathsf{ml}}^{\mathsf{GVP}}(\theta, \tilde{\theta}) = \sum_{m,t} \gamma_m(t) \log p\left(\boldsymbol{o}^{(t)}; \boldsymbol{\mu}^{(m)}(\mathbf{v}_t), \boldsymbol{\Sigma}^{(m)}(\mathbf{v}_t)\right) \quad (5)$$

where $\gamma_m(t)$ is the posterior probability of frame $\boldsymbol{o}_t$ being emitted from component m at a time instance t.

It can be shown that the corresponding parts of the auxiliary function associated with the Gaussian mean and variance trajectories can be derived as [9, 10, 11, 12, 13, 14, 15],

$$\begin{aligned} \mathcal{Q}_{\mathsf{ml}}^{(\mu_i^{(m)})}(\theta, \tilde{\theta}) &= -\frac{1}{2} \mathbf{c}^{(\mu_i^{(m)})\top} \mathbf{U}^{(\mu_i^{(m)})} \mathbf{c}^{(\mu_i^{(m)})} \\ &\quad + \mathbf{k}^{(\mu_i^{(m)})} \mathbf{c}^{(\mu_i^{(m)})} + \text{const} \\ \mathcal{Q}_{\mathsf{ml}}^{(\sigma_{i,i}^{(m)})}(\theta, \tilde{\theta}) &= -\frac{1}{2} \mathbf{c}^{(\sigma_i^{(m)})\top} \mathbf{U}^{(\sigma_i^{(m)})} \mathbf{c}^{(\sigma_i^{(m)})} \\ &\quad + \mathbf{k}^{(\sigma_i^{(m)})} \mathbf{c}^{(\sigma_i^{(m)})} + \text{const}'. \end{aligned} \quad (6)$$

where the constant terms independent of the coefficient vectors $\mathbf{c}^{(\cdot)}$ can be ignored. Setting the above gradients against the respective polynomial coefficient vectors to zero, the following ML solutions of the coefficient vectors can then be derived

$$\begin{aligned} \hat{\mathbf{c}}^{(\mu_i^{(m)})} &= \mathbf{U}^{(\mu_i^{(m)})-1} \mathbf{k}^{(\mu_i^{(m)})} \\ \hat{\mathbf{c}}^{(\sigma_{i,i}^{(m)})} &= \mathbf{U}^{(\sigma_{i,i}^{(m)})-1} \mathbf{k}^{(\sigma_{i,i}^{(m)})} \end{aligned} \quad (7)$$

and the sufficient statistics are

$$\begin{aligned} \mathbf{U}^{(\mu_i^{(m)})} &= \sum_t \gamma_m(t) \sigma_{i,i}^{(m)-1}(\mathbf{v}_t) \mathbf{v}_t^\top \mathbf{v}_t \\ \mathbf{k}^{(\mu_i^{(m)})} &= \sum_t \gamma_m(t) \sigma_{i,i}^{(m)-1}(\mathbf{v}_t) o_i^{(t)} \mathbf{v}_t^\top \\ \mathbf{U}^{(\sigma_{i,i}^{(m)})} &= \sum_t \gamma_m(t) \breve{\sigma}_{i,i}^{(m)} \mathbf{v}_t^\top \mathbf{v}_t \\ \mathbf{k}^{(\sigma_{i,i}^{(m)})} &= \sum_t \gamma_m(t) \left(o_i^{(t)} - \mu_i^{(m)}(\mathbf{v}_t)\right)^2 \mathbf{v}_t^\top \end{aligned} \quad (8)$$

3. Efficient Use of DNN Bottleneck Features in GVP-HMMs

In this section three approaches improving the efficiency of using DNN bottleneck features in GVP-HMMs are presented.

3.1. Automatic Model Complexity Control for GVP-HMMs

An important issue associated with GVP-HMMs is the appropriate polynomial degree to use. The use of higher degree polynomials significantly increases the number of coefficient parameters in GVP-HMMs. This not only leads to severe over-fitting, but also significantly increases the computation time incurred in interpolation. One solution to this problem is to automatically learn the optimal model structure for GVP-HMMs. The optimal polynomial degrees of parameter trajectories of, for example, Gaussian means and variances, are to be automatically determined at local level [13, 14] by maximizing the Bayesian evidence,

$$p(\mathcal{O}|\mathcal{W}, \mathcal{M}) = \int p(\mathcal{O}|\theta, \mathcal{W}, \mathcal{M}) p(\theta|\mathcal{M}) \mathrm{d}\theta \quad (9)$$

where θ is a parameterization of $\mathcal{M}$, $\mathcal{O} = \{\boldsymbol{o}_1, ..., \boldsymbol{o}_\mathcal{T}\}$ is a training data set of $\mathcal{T}$ frames and $\mathcal{W}$ the reference transcription.

As it is computationally intractable to directly compute the evidence in equation (9) for GVP-HMMs, an efficient approximation using the BIC style first order asymptotic expansion [21] of a lower lower bound [13, 22, 23, 24] of the evidence integral can be used. The optimal model complexity is determined by

$$\hat{\mathcal{M}} = \arg\max_{\mathcal{M}} \left\{ \mathcal{Q}_{\mathsf{ml}}^{(\mathcal{M})}(\hat{\theta}, \tilde{\theta}) - \rho \cdot \frac{k}{2} \log \mathcal{T} \right\}. \quad (10)$$

where the ML auxiliary functions associated with Gaussian mean and variance trajectory parameters in equation (6) are evaluated at the optimal model parameters $\hat{\theta}$ using the statistics given in equations (7) and (8). k is the number of free parameters in $\mathcal{M}$ and ρ is a tunable penalty term [25].

3.2. Clustering of DNN Bottleneck Features

As discussed in section 1, polynomial interpolation is performed for each high dimensional bottleneck feature vector at a frame level in conventional GVP-HMMs. The associated interpolation cost is very expensive in both training and test time. In addition, the large dynamic range of DNN bottleneck features can also introduce numerical issues when accumulating and inverting the sufficient statistics in equation (8). In order to handle these issues, a GMM-based bottleneck feature clustering approach is also considered. The training data bottleneck features are used to estimate the parameters of a GMM first. During both GVP-HMM training and decoding, a given DNN

bottleneck feature vector $\vec{\mathbf{f}}_t$ is quantized as the mean vector of the component Gaussian that assigns the highest posterior probability to $\vec{\mathbf{f}}_t$. As adjacent frames of bottleneck features emitted from the same HMM state can be strongly correlated, a continuous block of DNN features can be quantized into the same vector representation. This allows the interpolated Gaussian parameters in equation (4) to be computed once only for a continuous block of bottleneck features and then efficiently cached.

3.3. Modelling Tied Linear Transformation Trajectories

GVP-HMMs can also use a more compact trajectory modelling for model space tied linear transformations. Considering $(D+1)\times D$ dimensional mean transform $\boldsymbol{W}^{(r_m)}(\cdot)$ in equation (1) assigned to component m, the element wise transform trajectory in row i and column j is computed as

$$w_{i,j}^{(r_m)}(\mathbf{v}_t) \;=\; \mathbf{v}_t\cdot\mathbf{c}^{(w_{i,j}^{(r_m)})} \tag{11}$$

By definition, the mean transform polynomials are modelled on top of the component mean trajectories, thus the final updated mean vector of component m at time instance t is

$$\widetilde{\boldsymbol{\mu}}^{(m)}(\mathbf{v}_t) \;=\; \boldsymbol{W}^{(r_m)}(\mathbf{v}_t)\boldsymbol{\zeta}_t^{(m)} \tag{12}$$

where the $(D+1)$ dimensional extended mean vector trajectory $\boldsymbol{\zeta}_t^{(m)}=[\boldsymbol{\mu}^{(m)}(\mathbf{v}_t),1]^\top$.

The ML auxiliary function associated with the transform polynomial coefficients can also be expressed as

$$\begin{aligned}\mathcal{Q}_{\mathsf{ml}}^{(w_i^{(r_m)})}(\theta,\tilde{\theta}) \;=\;& -\frac{1}{2}\mathbf{c}^{(w_i^{(r_m)})\top}\mathbf{U}^{(w_i^{(r_m)})}\mathbf{c}^{(w_i^{(r_m)})}\\ &+\mathbf{k}^{(w_i^{(r_m)})}\mathbf{c}^{(w_i^{(r_m)})}+\text{const}''\end{aligned} \tag{13}$$

where $\mathbf{c}^{(w_i^{(r_m)})}$ is the $(D+1)\times(P\times N+1)$ dimensional meta polynomial coefficient vector spanning across all elements of row i of transform $\boldsymbol{W}^{(r_m)}$ [9, 10]. Ignoring the constant term and maximizing the above gives a close form solution,

$$\hat{\mathbf{c}}^{(w_i^{(r_m)})} \;=\; \mathbf{U}^{(w_i^{(r_m)})-1}\mathbf{k}^{(w_i^{(r_m)})} \tag{14}$$

where the exact form of the sufficient statistics $\mathbf{U}^{(w_i^{(r_m)})}$ and $\mathbf{k}^{(w_i^{(r_m)})}$ can be found in [9, 10].

4. Tandem GVP-HMM Systems

A range of tandem GVP-HMM configurations are used in this paper to incorporate DNN bottleneck features. Among these, three configurations allow trajectory modelling of Gaussian component means, or variances, and optionally both, are shown in the first section (line 1 to 3) of table 1, as "mean", "var" and "mv" respectively. In the 2nd section (line 4 to 5) of table 1, two more compact systems modelling the polynomial trajectories of 128 or 256 mean transforms are shown as "trans128" and "trans256". In all experiments, only the BIC based complexity controlled tandem GVP-HMM systems were considered. The penalty term in the complexity control metric of equation (10) was fixed as $\rho=2$ in all experiments. For all parameter polynomials the range of candidate degree to consider is [0,5].

In these tandem GVP-HMM systems, the DNN bottleneck features are not only used as the input features to estimate the continuous trajectories of Gaussian component and mean transform parameters in the acoustic feature subspace, but also used as normal features to train the standard static mean and variance parameters in the bottleneck feature subspace. For example, the final mean vector of component m at time instance t using Gaussian mean trajectory modelling is

$$\vec{\boldsymbol{\mu}}_t^{(m)} \;=\; [\boldsymbol{\mu}_{\mathsf{GVP}}^{(m)}(\mathbf{v}_t^{\mathsf{BN}}),\;\boldsymbol{\mu}_{\mathsf{BN}}^{(m)}]^\top. \tag{15}$$

where the $\boldsymbol{\mu}_{\mathsf{GVP}}^{(m)}(\mathbf{v}_t^{\mathsf{BN}})$ is the mean subvector trajectory taking a Vandermonde vector input $\mathbf{v}_t^{\mathsf{BN}}$ constructed using the 26 dimensional continuous or clustered DNN bottleneck features. $\boldsymbol{\mu}_{\mathsf{BN}}^{(m)}$ is the remaining static mean subvector estimated using the bottleneck features.

When mean transform trajectory is also modelled, the final mean vector of component m at time t is computed as

$$\vec{\boldsymbol{\mu}}_t^{(m)} \;=\; [\boldsymbol{W}_{\mathsf{GVP}}^{(r_m)}(\mathbf{v}_t^{\mathsf{BN}})\boldsymbol{\zeta}_{t,\mathsf{BN}}^{(m)},\;\boldsymbol{\mu}_{\mathsf{BN}}^{(m)}]^\top. \tag{16}$$

$\boldsymbol{W}_{\mathsf{GVP}}^{(r_m)}(\mathbf{v}_t^{\mathsf{BN}})$ is the transform trajectory, and $\boldsymbol{\zeta}_{t,\mathsf{BN}}^{(m)}$ is the extended mean subvector trajectory $\boldsymbol{\zeta}_{t,\mathsf{BN}}^{(m)}=[\boldsymbol{\mu}_{\mathsf{GVP}}^{(m)}(\mathbf{v}_t^{\mathsf{BN}}),1]^\top$.

System	mean	var	trans
mean	$\surd$	$\times$	$\times$
var	$\times$	$\surd$	$\times$
mv	$\surd$	$\surd$	$\times$
trans128	$\times$	$\times$	$\surd$
trans256	$\times$	$\times$	$\surd$

Table 1: *Description of tandem GVP-HMM configurations.*

5. Experiments and Results

The performance of the various improved tandem GVP-HMMs were evaluated on the Aurora 4 multi-style training setup. The training set consists of 7138 utterances. One half of the utterances were recorded by the primary Sennheiser microphone and the remaining half were recorded with a secondary microphone. Both halves include clean speech and speech corrupted by one of 6 different noises (street traffic, train station, car, babble, restaurant, airport) under SNR conditions between 10 and 20 dB. A total of 14 subsets was used for Word Error Rate (WER) evaluation. Two of these (330 utterances in each subset) were clean speech data recorded by the primary or secondary microphone. The remaining 12 subsets were obtained by randomly adding the same 6 noise types used in training set at 5-15 dB SNR for each of microphone conditions. A standard bi-gram language model provided for Aurora 4 was used in decoding.

Two baseline HMM systems were multi-style trained [26] using 1206 or 3202 tied states, with 16 Gaussian components per state for for both systems. DNNs with a bottleneck layer are trained on 72 dimensional log mel filter-bank features. The input layer was formed by splicing a context window of 11 frames thus creating a 792 dimensional input vector. Five hidden layers with 2048 nodes each were used between the input layer and the bottleneck layer with 26 nodes. The layer-by-layer RBM based pre-training implemented in the Kaldi toolkit [27] was used in training. After 26 dimensional DNN bottleneck features extracted, they are utterance-level mean normalized and decorrelated using PCA. An extended version of the HTK toolkit [28] was used to train various tandem GVP-HMM systems. As discussed in section 1, directly using higher degree polynomials dramatically increases the number of coefficient parameters in GVP-HMMs. Hence, the model selection technique of section 3.1 was applied to all GVP-HMM systems modelling mean and/or variance parameter trajectories in the experiments.

System	num. of GMMs	clean	noise	channel	both	Ave
tandem	-	5.92	10.05	11.75	22.80	15.34
mean	-	5.36	9.71	10.09	21.78	14.61
	1024	5.31	9.74	10.27	21.62	14.51
	2048	5.31	9.77	10.09	21.53	14.51
	4096	5.27	9.73	10.24	21.65	14.55
var	2048	6.33	10.18	11.34	22.42	15.23
	4096	6.03	10.23	11.06	22.32	15.17
mv	2048	6.31	10.17	11.23	22.41	15.22
	4096	5.49	10.21	10.20	22.10	14.97
trans128	-	5.42	9.64	**9.92**	21.53	**14.45**

Table 2: *Performance of tandem HMM baseline and tandem GVP-HMM systems with 1206 states on Aurora 4.*

The performance of various tandem GVP-HMM systems built using a smaller 1206 tied state configuration are first shown in table 2. First, small but consistent performance improvements were obtained when applying the GMM based DNN bottleneck feature clustering of section 3.2. For example, using a 2048 component GMM in quantization (line 4 in table 2), gave a marginal WER reduction of 0.07% absolute over the comparable GVP-HMM "mean" using non-quantized, continuous DNN Bottleneck features (line 2 in table 2).

Using the more compact mean transform trajectory based GVP-HMM system, "trans128" (last line in table 2), gave the lowest error rate among all systems in table 2. It outperformed the tandem baseline by 0.89% absolute (5.8% relative) in WER. In particular, on the secondary microphone channel condition, significant WER reductions of 1.83% absolute (15.6% relative) were obtained over the baseline tandem HMM system.

System	mean					trans128
#GMMs	-		1024	2048	4096	-
ComCtrl.	×	√	√	√	√	√
#Parm	41.74M	3.74M	2.55M	2.61M	2.63M	2.65M
Time(%)	-	100	54.8	60	65.3	69

Table 3: *Model parameters and computational cost of tandem GVP-HMM "mean" and "trans128" systems in table 2. The cost of complexity controlled tandem GVP-HMM "mean" system with no bottleneck feature clustering taken as reference.*

A detailed comparison of total number of parameters (polynomial coefficients and optionally GMM parameters for feature clustering) and computational cost (decoding time and optionally interpolation cost) is shown in table 3 for the tandem GVP-HMM "mean" and "trans128" systems previously shown in table 2, Using the model selection method described in section 3.1 to remove redundant polynomial parameters, the GVP-HMM "mean" system size is significantly reduced by over 90% (line 3, form column 1 to 2 in table 3). Using a 1024 component GMM based bottleneck feature clustering, the decoding time of the GVP-HMM "mean" system is further reduced by 45.2% relative (last line, from column 2 to 3 in table 3). Compared with the conventional GVP-HMM "mean" system using a uniformly assigned polynomial degree, the combination of these two approaches gave an overall decoding time speed up of over 16 time in total.

In order to further investigate the performance of tandem GVP-HMM systems on the test data associated with unseen noise conditions, a low and high SNR based results breakdown for the "noise" and "both" subsets are shown in table 4. In the both test subsets, tandem GVP-HMM systems with BIC complexity controlled method achieved better WER reduction over the tandem baseline on the unseen data (from 5 dB to 10 dB) compared with the seen data (>10 dB). For example, under unseen "noise" SNR condition from 5 dB to 10 dB, using the GVP-HMM "trans128" system highlighted in bold in table 4, a WER reduction of 0.69% absolute (6.34% relative) over the baseline tandem HMM system. Similar error rate reductions of 6.65% were also obtained on the unseen noise conditions (5dB-10dB) of the "both" set.

System	num. of GMMs	noise (5-10dB)	noise (>10dB)	both (5-10dB)	both (>10dB)
tandem	-	10.89	9.04	27.06	17.77
mean	-	10.37	8.97	25.65	17.19
	1024	10.33	9.04	25.28	17.14
	2048	10.43	9.00	25.29	17.17
	4096	10.33	9.01	25.40	17.20
var	2048	10.84	9.40	26.31	17.82
	4096	10.97	9.34	26.14	17.82
mv	2048	10.89	9.30	26.32	17.79
	4096	10.97	9.31	25.88	17.63
trans128	-	**10.20**	**8.97**	**25.26**	**17.11**

Table 4: *Breakdown Results for tandem HMM baseline and tandem GVP-HMM systems with 1206 states on Aurora 4.*

System	num. of GMMs	clean	noise	channel	both	Ave
tandem	-	5.01	8.85	9.36	20.69	13.69
mean	4096	4.41	8.85	**8.50**	20.70	13.59
	8192	4.46	8.80	8.78	20.75	13.61
var	4096	4.93	8.64	9.28	20.05	13.31
	8192	4.32	8.75	8.59	20.45	13.48
mv	4096	4.45	8.74	8.63	20.04	**13.27**
	8192	4.58	8.89	8.99	20.17	13.42
trans256	-	4.50	8.77	8.61	20.41	13.44

Table 5: *Performance of tandem HMM baseline and tandem GVP-HMM systems with 3202 states on Aurora 4.*

The scalability of tandem GVP-HMM approach was further evaluated on a more competitive tandem HMM baseline with 3202 tied states. A set of experiments comparable to those in table 2 were conducted. The WER performance of this 3202 state tandem baseline system and comparable tandem GVP-HMM systems are shown in table 5. A consistent reduction in WER over the tandem HMM baseline was obtained. Using the best tandem GVP-HMM system highlighted in bold in the 6th line of table 5, modelling both Gaussian mean and variance trajectories, a WER reduction of 0.42% absolute (3.1% relative) over the tandem HMM system was obtained. Consistent with the results shown in table 2, on the secondary microphone channel condition, significant WER reductions of 0.86% absolute (9.2% relative) were obtained over the baseline tandem system.

6. Conclusion

In this paper, several approaches were proposed to improve the efficiency of using DNN bottleneck features in GVP-HMMs. A model selection techniques to optimally reduce the polynomial degrees. An efficient GMM based bottleneck feature clustering method was used to further reduce the interpolation cost. More compact model transform trajectory modelling was also used. A total of 16 time speed up in decoding time over conventional GVP-HMMs was obtained. Significant error rate reductions of 15.6% relative were obtained over the baseline multi-style tandem HMM system on the secondary microphone channel condition of Aurora 4 data. Future research will focus on using more efficient feature space transform based GVP-HMMs to incorporate DNN bottleneck features [12].

7. References

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-Dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Proc. IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[3] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. ISCA INTERSPEECH*, Florence, Italy, 2011, pp. 437–440.

[4] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Proc. ISCA INTERSPEECH*, Florence, Italy, 2011, pp. 237–240.

[5] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. ICASSP*, Vancouver, BC, Canada, 2013, pp. 7398–7402.

[6] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Norwell, MA, USA: Kluwer Academic Publishers, 1993.

[7] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. IEEE ICASSP*, vol. 3, Istanbul, Turkey, 2000, pp. 1635–1638.

[8] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. IEEE ICASSP*, vol. 4, Honolulu, Hawaii, USA, 2007, pp. 757–760.

[9] N. Cheng, X. Liu, and L. Wang, "Generalized variable parameter HMMs for noise robust speech recognition," in *Proc. ISCA INTERSPEECH*, Florence, Italy, 2011, pp. 482–484.

[10] ——, "A flexible framework for HMM based noise robust speech recognition using generalized parametric space polynomial regression," *Science China, Information Sciences*, vol. 54, no. 2, pp. 2481–2491, 2011.

[11] Y. Li, X. Liu, and L. Wang, "Structured modeling based on generalized variable parameter HMMs and speaker adaptation," in *Proc. IEEE ISCSLP*, Hong Kong, China, 2012, pp. 136–140.

[12] ——, "Feature space generalized variable parameter HMMs for noise robust recognition," in *Proc. ISCA INTERSPEECH*, Lyon, France, 2013, pp. 2968–2972.

[13] R. Su, X. Liu, and L. Wang, "Automatic model complexity control for generalized variable parameter HMMs," in *Proc. IEEE ASRU*, Olomouc, Czech Republic, 2013, pp. 150–155.

[14] ——, "Automatic complexity control of generalized variable parameter HMMs for noise robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 102–114, 2015.

[15] X. Xie, R. Su, X. Liu, and L. Wang, "Deep neural network bottleneck features for generalized variable parameter HMMs," in *Proc. ISCA INTERSPEECH*, Singapore, 2014, pp. 2739–2743.

[16] X. Cui and Y. Gong, "A study of variable-parameter gaussian mixture hidden markov modeling for noisy speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1366–1376, 2007.

[17] D. Yu, L. Deng, Y. Gong, and A. Acero, "A novel framework and training algorithm for variable-parameter hidden markov models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1348–1360, 2009.

[18] Z. Ling, K. Richmond, and J. Yamagishi, "Articulatory control of HMM-based parametric speech synthesis using feature space switched multiple regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 207–219, 2013.

[19] A. Bjorck and V. Pereyra, "Solution of vandermonde systems of equations," *Mathematics of Computation (American Mathematical Society)*, vol. 24, no. 112, pp. 893–903, 1970.

[20] 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*, vol. 39, no. 1, pp. 1–39, 1997.

[21] G. Schwartz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[22] X. Liu and M. J. F. Gales, "Automatic model complexity control using marginalized discriminative growth functions," in *Proc. IEEE ASRU*, St. Thomas, U.S. Virgin Islands, 2003, pp. 37–42.

[23] X. Liu, M. J. F. Gales, and P. C. Woodland, "Model complexity control and compression using discriminative growth functions," in *Proc. IEEE ICASSP*, vol. 1, Montreal, Quebec, Canada, 2004, pp. 797–800.

[24] X. Liu and M. J. F. Gales, "Automatic model complexity control using marginalized discriminative growth functions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1414–1424, 2007.

[25] W. Chou and W. Reichl, "Decision tree state tying based on penalized bayesian information criterion," in *Proc. IEEE ICASSP*, vol. 1, Phoenix, Arizona, USA, 1999, pp. 345–348.

[26] R. P. Lippmann, E. A. Martin, and D. B. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. IEEE ICASSP*, vol. 12, Dallas, Texas, USA, 1987, pp. 705–708.

[27] The Kaldi speech recognition toolkit. Available from *http://kaldi.sourceforge.net*.

[28] S. Young *et al.*, *The HTK Book Version 3.4.1*, 2009.