AUTOMATIC MODEL COMPLEXITY CONTROL FOR GENERALIZED VARIABLE PARAMETER HMMS

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

[1]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, xl207@cam.ac.uk, lan.wang@siat.ac.cn

ABSTRACT

An important task for speech recognition systems is to handle the mismatch against a target environment introduced by acoustic factors such as variable ambient noise. To address this issue, it is possible to explicitly approximate the continuous trajectory of optimal, well matched model parameters against the varying noise using, for example, using generalized variable parameter HMMs (GVP-HMM). In order to improve the generalization and computational efficiency of conventional GVP-HMMs, this paper investigates a novel model complexity control method for GVP-HMMs. The optimal polynomial degrees of Gaussian mean, variance and model space linear transform trajectories are automatically determined at local level. Significant error rate reductions of 20% and 28% relative were obtained over the multi-style training baseline systems on Aurora 2 and a medium vocabulary Mandarin Chinese speech recognition task respectively. Consistent performance improvements and model size compression of 57% relative were also obtained over the baseline GVP-HMM systems using a uniformly assigned polynomial degree.
Index Terms: model complexity control, generalized variable parameter HMM, robust speech recognition

1. INTRODUCTION

An important task for automatic speech recognition (ASR) systems is to robustly handle the mismatch against a target environment introduced by time-varying factors such as environment noise. To handle this issue, a range of model based techniques can be used: multi-style training [19] exploits the implicit modelling power of mixture models, and more recently deep neural networks [29], to obtain a good generalization to unseen noise conditions; noise adaptive training [12, 13] structurally models the variability introduced to the observed speech signals by environment noise and other factors; uncertainty decoding [8, 24, 9, 15], propagates the uncertainty that varies with the noise represented by, for example, a conditional distribution of the corrupted speech, into the recognizer. In addition to these approaches, it is also possible to explicitly approximate the continuous trajectories of optimal model parameters against the varying noise condition using a polynomial function [10, 6, 33, 18], for example, as in multiple regression HMMs (MR-HMM) [10] and variable parameter HMMs (VP-HMM) [6, 33]. In order to reduce the interpolation cost incurred at Gaussian component level when mean or variance trajectory modelling are used, an extension to both MR-HMMs and VP-HMMs, generalized variable parameter HMMs (GVP-HMMs), were proposed in [3, 4, 16, 17]. In addition to Gaussian parameters, GVP-HMMs can also provide a more compact trajectory modelling for model space tied linear transformations, and thus provide a flexible form of parameter trajectory modelling.

An important issue associated with MR-HMMs, VP-HMMs and GVP-HMMs is the appropriate polynomial degree to use. In order to reduce the oscillation occurring when higher degree polynomials are used [27], lower degree polynomials of the same order, for example, the second order, were used in previous research [10, 6, 3, 4, 16, 17]. However, there are two issues with this approach. First, the variability introduced by ambient noise manifests itself in a locally varying fashion on different dimensions in the acoustic space. For example, lower order cepstral parameters, which contain richer information of speech than higher order cepstras, are more prone to the distortion introduced by environment noise. A uniformly assigned polynomial degree can cause an under-fitting for the lower order cepstras, while at the same time an over-fitting for the higher order cepstras that are more related to noise in nature and thus more invariant to the distortion. Such lack of modelling flexibility can lead to a poor generalization to unseen noise conditions. Secondly, over-fitting higher degree polynomials increases the interpolation cost during recognition.

To address these issues, this paper investigates a novel model complexity control method for GVP-HMMs. The optimal polynomial degrees of Gaussian mean, variance and model space linear transform trajectories are automatically determined at local level. The rest of the paper is organized as follows. The GVP-HMM framework is reviewed in section 2. An efficient Bayesian model complexity control criterion is presented in section 3. The detailed complexity control algorithm for GVP-HMMs is proposed in section 4. In section 5 various complexity controlled GVP-HMM systems are evaluated on Aurora 2 and a medium vocabulary Mandarin speech recognition task. Section 6 is the conclusion and future research.

2. GENERALIZED VARIABLE PARAMETER HMMS

Generalized variable parameter HMMs (GVP-HMMs) [3, 4, 16, 17] explicitly model the trajectory of optimal acoustic parameters that vary with respect to the underlying noise condition. The type of parameter trajectories are not restricted to those of means and covariances of conventional tied mixture HMMs. Other more compact forms of parameters, such as model or feature space linear transformations [14, 11], may also be considered. In this paper, trajectories of Gaussian mean transforms are modelled. For a D dimensional observation $\boldsymbol{o}_t$ emitted from Gaussian mixture component m, as-

This work is supported by National Natural Science Foundation of China (NSFC 61135003), National Fundamental Research Grant of Science and Technology (973 Project: 2013CB329305) ShenZhen Fundamental Research Program JC01005280621A, JCYJ20130401170306806.

suming P^{th} order polynomials are used, this is given by

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

where $\mathbf{v}_t^{\top}$ is a $(P+1)$ dimensional Vandermonde vector [2], such that $\mathbf{v}_{t,p} = v_t^{p-1}$. v_t is an auxiliary feature, and in this paper, the speech-noise-ratio (SNR) condition [25] at frame t. $\boldsymbol{W}^{(r_m)}(\mathbf{v}_t)$ is the $(D+1)\times D$ mean transform that component m is assigned to at frame t. $\boldsymbol{\mu}^{(m)}(\cdot)$, $\boldsymbol{\Sigma}^{(m)}(\cdot)$ and $\boldsymbol{W}^{(r_m)}(\cdot)$ are the P^{th} order mean, covariance and mean transform trajectory polynomials of component m respectively. Assuming diagonal covariances are used, then the trajectories of the i^{th} dimension of the mean and variance, and the transform element in row i and column j, are

$$\begin{aligned} \mu_i^{(m)}(\mathbf{v}_t) &= \mathbf{v}_t \cdot \mathbf{c}^{(\mu_i^{(m)})} \\ \sigma_{i,i}^{(m)}(\mathbf{v}_t) &= \breve{\sigma}_{i,i}^{(m)} \mathbf{v}_t \cdot \mathbf{c}^{(\sigma_{i,i}^{(m)})} \\ w_{i,j}^{(r_m)}(\mathbf{v}_t) &= \mathbf{v}_t \cdot \mathbf{c}^{(w_{i,j}^{(r_m)})} \end{aligned} \quad (2)$$

where $\mathbf{c}^{(\cdot)}$ is a $(P+1)$ dimensional polynomial coefficient vector such that $\mathbf{c}_p^{(\cdot)} = c_{p-1}^{(\cdot)}$, and $c_{p-1}^{(\cdot)}$ the $(p-1)^{th}$ order polynomial coefficient of the parameter trajectory being considered. $\breve{\sigma}_{i,i}^{(m)}$ is the clean speech based variance estimate. 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 computed as

$$\widetilde{\boldsymbol{\mu}}^{(m)}(\mathbf{v}_t) = \boldsymbol{W}^{(r_m)}(\mathbf{v}_t)\boldsymbol{\zeta}_t^{(m)} \quad (3)$$

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

GVP-HMMs share the same instantaneous adaptation power as standard MR-HMMs and VP-HMMs. For any noise characteristics as indicated by the auxiliary feature, e.g. the SNR level as considered in this work, present or unseen in the training data, GVP-HMMs can instantly produce the matching Gaussian component and mean transform parameters by-design without requiring any multi-pass decoding and adaptation process. GVP-HMMs also provide a more compact and flexible form of parameter trajectory modelling. For example, when only limited amounts of noisy training data is available, to ensure all polynomial coefficients are robustly estimated, only the trajectories associated with the elements of a globally tied mean transform can be considered. When large amounts of noisy training data is used, a more refined modelling resolution can also be obtained by increasing the number of tied transformations, or modelling the trajectories of multiple parameter types simultaneously. The use of locally optimized polynomial degree for different model parameters is expected to further improve their modelling flexibility and generalization.

3. MODEL COMPLEXITY CONTROL

A standard problem in speech recognition, and statistical modelling in general, is how to select a model structure, $\hat{\mathcal{M}}$, that generalizes well to unseen data, from a set of candidate model structures $\{\mathcal{M}\}$. In Bayesian learning, when no prior knowledge over individual model structures is available, the optimal model structure or complexity, is determined by maximizing the *evidence* integral,

$$p(\mathcal{O}|\mathcal{W},\mathcal{M}) = \int p(\mathcal{O}|\lambda,\mathcal{W},\mathcal{M})p(\lambda|\mathcal{M})\mathrm{d}\lambda \quad (4)$$

where λ denotes 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.

For standard HMMs, MR-HMMs, VP-HMMs and GVP-HMMs, it is computationally intractable to directly compute the evidence integral in equation (4). To handle this problem, a variety of approximation schemes can be used: a first order asymptotic expansion based Bayesian Information Criterion (BIC) [28], a second order asymptotic expansion based Laplace's approximation [31, 21, 22, 23], variational approximation [30], and Markov Chain Monte Carlo (MCMC) based sampling schemes [26]. Among these, BIC (or equivalently MDL [1]) is the most widely used technique. It is expressed in terms of a penalized log likelihood evaluated at the maximum likelihood (ML) estimate of model parameters $\hat{\lambda}$. The model selection is based on the following approximation

$$\log p(\mathcal{O}|\mathcal{W},\mathcal{M}) \approx \log p(\mathcal{O}|\hat{\lambda},\mathcal{W},\mathcal{M}) - \rho\cdot\frac{k}{2}\log\mathcal{T} \quad (5)$$

where k denotes the number of free parameters in $\mathcal{M}$ and ρ is a penalization coefficient which may be tuned for the specific task [5, 20]. When $\rho = 1$, BIC was shown to be a first order asymptotic expansion of the evidence integral [28].

One issue with the BIC based complexity control of equation (5) is that the log-likelihood for each model structure is required. For HMMs and their variants such as GVP-HMMs this can be computationally expensive. One method to avoid this is to derive a lower bound that may be assumed to be applicable for multiple different structures. Let $\tilde{\lambda}$ denote the *current* parameterization for $\mathcal{M}$. Using the EM algorithm the following inequality may be derived [7]

$$\begin{aligned} \log p(\mathcal{O}|\lambda,\mathcal{W},\mathcal{M}) &\geq \mathcal{L}_{\mathsf{ml}}^{(\mathcal{M})}(\lambda,\tilde{\lambda}) \\ &= \log p(\mathcal{O}|\tilde{\lambda},\mathcal{W},\mathcal{M}) + \mathcal{Q}_{\mathsf{ml}}^{(\mathcal{M})}(\lambda,\tilde{\lambda}) - \mathcal{Q}_{\mathsf{ml}}^{(\mathcal{M})}(\tilde{\lambda},\tilde{\lambda}) \end{aligned} \quad (6)$$

where the auxiliary function, $\mathcal{Q}_{\mathsf{ml}}^{(\mathcal{M})}(\lambda,\tilde{\lambda})$, is given by

$$\mathcal{Q}_{\mathsf{ml}}^{(\mathcal{M})}(\lambda,\tilde{\lambda}) = \sum_{m,t}\gamma_m(t)\log p(\boldsymbol{o}_t|\theta_t = m,\lambda,\mathcal{M}). \quad (7)$$

$\theta_t = m$ indicates that an acoustic observation $\boldsymbol{o}_t$ was generated by a hidden state m, and the hidden state posterior $\gamma_m(t) = P(\theta_t = m|\mathcal{O},\mathcal{W},\tilde{\lambda},\mathcal{M})$.

Accumulating the above statistics for all possible systems is infeasible. To handle this problem, a range of model structures can use the same set of statistics generated using a single system. This allows the lower bound in (6) to be efficiently computed [21, 22, 23]. For example, when determining the appropriate order of a Gaussian component mean's trajectory polynomial on a particular dimension in equation (2) for an GVP-HMM system, the sufficient statistics $\{\gamma_m(t)\}$ to be used for a range of different polynomial orders to select can be derived from a common baseline HMM system, or a conventional GVP-HMM system that uses a globally assigned polynomial order across all dimensions for every Gaussian mean vector in the system. In the same fashion, sufficient statistics can also be shared when determining the degrees of Gaussian's variance or mean transformation trajectory polynomials in equation (2).

The only term in the lower bound of equation (6) dependent on the model parameters, λ, is the auxiliary function $\mathcal{Q}_{\mathsf{ml}}^{(\mathcal{M})}(\lambda,\tilde{\lambda})$. When multiple model structures use the same set of statistics, the rank ordering derived from the marginalization of $\mathcal{L}_{\mathsf{ml}}^{(\mathcal{M})}(\lambda,\tilde{\lambda})$ is equivalent to that of $\mathcal{Q}_{\mathsf{ml}}^{(\mathcal{M})}(\lambda,\tilde{\lambda})$ [1]. Under these conditions, the op-

[1] When multiple sets of statistics are used, the other terms in the lower bound cannot be ignored and must be computed.

timal model complexity is finally determined by

$$\hat{\mathcal{M}} \quad = \quad \arg\max_{\mathcal{M}} \left\{ \mathcal{Q}_{\mathsf{ml}}^{(\mathcal{M})}(\hat{\lambda}, \tilde{\lambda}) - \rho \cdot \frac{k}{2} \log \mathcal{T} \right\}. \qquad (8)$$

4. MODEL COMPLEXITY CONTROL FOR GVP-HMMS

When using the lower bound based BIC metric of equation (8) for the complexity control of GVP-HMMs, the computation of the ML auxiliary function of equation (7) is required. For the form of GVP-HMMs of equation (1) introduced in section 2, the associated ML auxiliary function is given by [7, 3, 4, 17],

$$\begin{aligned} \mathcal{Q}_{\mathsf{ml}}^{\mathsf{GVP}}(\lambda, \tilde{\lambda}) \quad &= \quad \sum_{m,t} \gamma_m(t) \log p\left(\boldsymbol{o}^{(t)}; \boldsymbol{\mu}^{(m)}(\mathbf{v}_t), \right. \\ &\qquad \left. \boldsymbol{\Sigma}^{(m)}(\mathbf{v}_t), \boldsymbol{W}^{(r_m)}(\mathbf{v}_t)\right) \end{aligned} \qquad (9)$$

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

Combining the above with equations (1) and (2), the corresponding parts of the above auxiliary function associated with the polynomial coefficient vectors of the Gaussian mean, variance scaling and mean transform element trajectories respectively can be re-arranged into convex quadratic forms,

$$\begin{aligned} \mathcal{Q}_{\mathsf{ml}}^{(\mu_i^{(m)})}(\lambda, \tilde{\lambda}) \quad &= \quad -\frac{1}{2}\mathbf{c}^{(\mu_i^{(m)})\top}\mathbf{U}^{(\mu_i^{(m)})}\mathbf{c}^{(\mu_i^{(m)})} \\ &\quad +\mathbf{k}^{(\mu_i^{(m)})}\mathbf{c}^{(\mu_i^{(m)})} + \text{const} \\ \mathcal{Q}_{\mathsf{ml}}^{(\sigma_{i,i}^{(m)})}(\lambda, \tilde{\lambda}) \quad &= \quad -\frac{1}{2}\mathbf{c}^{(\sigma_i^{(m)})\top}\mathbf{U}^{(\sigma_i^{(m)})}\mathbf{c}^{(\sigma_i^{(m)})} \\ &\quad +\mathbf{k}^{(\sigma_i^{(m)})}\mathbf{c}^{(\sigma_i^{(m)})} + \text{const}' \\ \mathcal{Q}_{\mathsf{ml}}^{(w_i^{(r_m)})}(\lambda, \tilde{\lambda}) \quad &= \quad -\frac{1}{2}\mathbf{c}^{(w_i^{(r_m)})\top}\mathbf{U}^{(w_i^{(r_m)})}\mathbf{c}^{(w_i^{(r_m)})} \\ &\quad +\mathbf{k}^{(w_i^{(r_m)})}\mathbf{c}^{(w_i^{(r_m)})} + \text{const}'' \end{aligned} \qquad (10)$$

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

$$\begin{aligned} \hat{\mathbf{c}}^{(\mu_i^{(m)})} \quad &= \quad \mathbf{U}^{(\mu_i^{(m)})-1}\mathbf{k}^{(\mu_i^{(m)})} \\ \hat{\mathbf{c}}^{(\sigma_{i,i}^{(m)})} \quad &= \quad \mathbf{U}^{(\sigma_{i,i}^{(m)})-1}\mathbf{k}^{(\sigma_{i,i}^{(m)})} \\ \hat{\mathbf{c}}^{(w_i^{(r_m)})} \quad &= \quad \mathbf{U}^{(w_i^{(r_m)})-1}\mathbf{k}^{(w_i^{(r_m)})} \end{aligned} \qquad (11)$$

where $\mathbf{c}^{(w_i^{(r_m)})}$ is a $(D+1)\times(P+1)$ dimensional meta polynomial coefficient vector spanning across all elements of row i of transform $\boldsymbol{W}^{(r_m)}$, and the sufficient statistics are

$$\begin{aligned} \mathbf{U}^{(\mu_i^{(m)})} \quad &= \quad \sum_t \gamma_m(t)\sigma_{i,i}^{(m)-1}(\mathbf{v}_t)\mathbf{v}_t^\top\mathbf{v}_t \\ \mathbf{k}^{(\mu_i^{(m)})} \quad &= \quad \sum_t \gamma_m(t)\sigma_{i,i}^{(m)-1}(\mathbf{v}_t)o_i^{(t)}\mathbf{v}_t^\top \\ \mathbf{U}^{(\sigma_{i,i}^{(m)})} \quad &= \quad \sum_t \gamma_m(t)\breve{\sigma}_{i,i}^{(m)}\mathbf{v}_t^\top\mathbf{v}_t \\ \mathbf{k}^{(\sigma_{i,i}^{(m)})} \quad &= \quad \sum_t \gamma_m(t)\left(o_i^{(t)} - \mu_i^{(m)}(\mathbf{v}_t)\right)^2\mathbf{v}_t^\top \end{aligned} \qquad (12)$$

$\mathbf{U}^{(w_i^{(r_m)})}$ is a $[(D+1)\times(P+1)]\times[(D+1)\times(P+1)]$ meta Vandermonde matrix, and $\mathbf{k}^{(w_i^{(r_m)})}$ a $(D+1)\times(P+1)$ dimensional meta regression target vector. The sub-matrices and sub-vectors associated with transform element $w_{i,j}^{(r_m)}$ are

$$\begin{aligned} \mathbf{U}^{(w_{i,j}^{(r_m)})} \quad &= \quad \left[\sum_{m\in r_m,t} \gamma_m(t)\sigma_{i,i}^{(m)-1}(\mathbf{v}_t)\zeta_{t,j}^{(m)}\zeta_{t,1}^{(m)}\mathbf{v}_t^\top\mathbf{v}_t\ , \right. \\ &\quad \ldots, \\ &\quad \sum_{m\in r_m,t} \gamma_m(t)\sigma_{i,i}^{(m)-1}(\mathbf{v}_t)\zeta_{t,j}^{(m)}\zeta_{t,i}^{(m)}\mathbf{v}_t^\top\mathbf{v}_t, \\ &\quad \ldots, \\ &\quad \left. \sum_{m\in r_m,t} \gamma_m(t)\sigma_{i,i}^{(m)-1}(\mathbf{v}_t)\zeta_{t,j}^{(m)}\zeta_{t,D+1}^{(m)}\mathbf{v}_t^\top\mathbf{v}_t\right] \\ \mathbf{k}^{(w_{i,j}^{(r_m)})} \quad &= \quad \sum_{m\in r_m,t} \gamma_m(t)\sigma_{i,i}^{(m)-1}(\mathbf{v}_t)\zeta_{t,j}^{(m)}o_i^{(t)}\mathbf{v}_t^\top \end{aligned} \qquad (13)$$

where the $(D+1)$ dimensional extended mean vector trajectory is given by $\boldsymbol{\zeta}_t^{(m)} = [\boldsymbol{\mu}^{(m)}(\mathbf{v}_t), 1]^\top$, as previously defined in section 2.

When determining the optimal order for a particular polynomial associated with the ith dimension of the mth Gaussian component in the system, $\mu_i^{(m)}(\cdot)$, for example, the above statistics in equations (12) and (13) are accumulated for the highest order P_{max} being considered. The corresponding statistics for any other order $0 \leq P^{(\mu_i^{(m)})} < P_{\mathsf{max}}$ can be derived by taking the associated sub-matrices or subvectors from the full matrix statistics accumulated for P_{max}. Using these statistics and the ML solutions in equation (11), the ML auxiliary function associated with $\mu_i^{(m)}(\cdot)$ in equation (10), can be efficiently evaluated at the optimum for each candidate polynomial degree. The number of free parameters (polynomial coefficients) in the BIC metric of equation (8) is $k = P^{(\mu_i^{(m)})} + 1$. The number of frame samples for the current Gaussian is computed as the component level occupancy counts $\mathcal{T}^{(m)} = \sum_{t,m}\gamma_m(t)$. An overview of this algorithm is shown in figure 1.

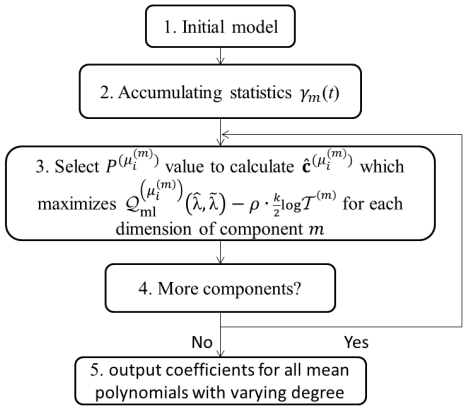

Fig. 1. Complexity control of GVP-HMM mean polynomials using BIC locally for each dimension of all Gaussian components.

The same approach can also be used to determine the optimal order of Gaussian variance and mean transform polynomials, by evaluating the respective auxiliary functions to compute the BIC metric in equation (8).

System	Poly. Types			#PolyCoef (Aurora2)		#PolyCoef (In-Car)	
	mean	var	tran	base	BIC($\rho = 1/2/3$)	base	BIC($\rho = 1/2/3$)
mean	$\surd$	$\times$	$\times$	120K	67.4K/58.1K/53.6K	3.66M	1.97M/1.77M/1.67M
mv	$\surd$	$\surd$	$\times$	240K	119K/105.6K/99.1K	7.32M	3.79M/3.47M/3.3M
tran2	$\times$	$\times$	$\surd$	9.4K	8.68K/8.44K/8.16K	10.8K	7.22K/7.18K/7.18K
tran8	$\times$	$\times$	$\surd$	37.4K	32.56K/32.16K/31.56K	-	-
tran256	$\times$	$\times$	$\surd$	-	-	1.39M	0.98M/0.97M/0.97M
mvt2	$\surd$	$\surd$	$\surd$	249K	126.2K/112K/106K	7.32M	3.79M/3.47M/3.3M

Table 1. Description of various GVP-HMMs: parameter polynomial types and the number of polynomial coefficients.

5. EXPERIMENTAL RESULTS

In this section, complexity controlled GVP-HMM systems are evaluated on two tasks: Aurora2 and a medium vocabulary Mandarin Chinese In-car navigation command recognition task.

5.1. Description of GVP-HMM Variant Systems

As discussed in section 2, in order to adjust the trade-off between modelling resolution, robustness in estimation and computational efficiency, a wide rage of GVP-HMM configurations may be considered to suit different purposes. The description of these GVP-HMM variant systems' configurations and the number of polynomial coefficients used for the standard Aurora2 task and a Mandarin Chinese In-car navigation command recognition system are shown in table 1. Two standard VP-HMM configurations, which allow trajectory modelling of Gaussian component means, and optionally variances, are shown in the first two lines of the table, as "mean" and "mv" respectively. In the 2nd section (line 3 to 5) of table 1, three GVP-HMM systems modelling the polynomial trajectories of 2, 8 or 256 mean transforms are shown as "tran2", "tran8" and "tran256". Finally, the most complex GVP-HMM system that uses trajectory modelling for Gaussian means and variances, plus 2 model space transforms are shown as "mvt2" in the bottom section of the table. The number of polynomial coefficients for various GVP-HMM systems are shown in the last 4 columns of table 1 for the Aurora 2 and In-Car tasks respectively. All the baseline GVP-HMMs using 2nd degree polynomials for all parameter trajectories are shown as "base" in the 5th and 7th columns. The number of polynomial coefficients of BIC based complexity controlled GVP-HMM systems with varying settings, $\rho = 1, 2, 3$, are given in the 6th and 8th columns of table 1. For all polynomials the range of candidate degree to consider is $[0, 5]$.

5.2. Experiments on Aurora 2

The Aurora2 speaker independent digit sequence recognition database contains 4 noisy conditions: subway, babble, car and exhibition. A total of 420 utterances from four different SNR conditions (-5dB, 5dB, 15dB, 25dB) were used to train both the baseline multi-style HMMs and various GVP-HMM systems. A total of 1000 utterances selected from the car noise environment at 0dB, 5dB, 10dB, 15dB and 20dB SNR were used for word error rate (WER) evaluation. 39 dimensional MFCC plus log energy features including their 1st and 2nd order differentials were used. Performance of the baseline multi-style HMM and GVP-HMM systems, as described in table 1 are shown in table 2. Modelling the trajectories of all three parameter types, including Gaussian means, variances and two shared mean transforms, gave the best performance for the baseline GVP-HMMs, as shown in the last line of table 2. Using this "mvt2" system, average WER reductions of 0.79%-0.81% absolute (9% relative) across all SNR conditions were obtained over the "mcond" multi-style baseline, and the mean only VP-HMM/GVP-HMM system shown as "mean" in table 2.

System	0dB	5dB	10dB	15dB	20dB	Ave
clean.base	75.33	41.42	15.63	6.14	3.14	28.34
mcond.base	22.88	9.42	4.29	3.58	2.78	8.95
mean	25.63	9.52	4.32	3.10	2.26	8.97
mv	23.16	9.12	4.30	3.09	2.28	8.39
tran2	23.64	8.77	4.05	2.89	2.32	8.33
tran8	21.73	8.17	4.14	3.37	2.17	7.92
mvt2	22.34	8.96	4.18	3.04	2.29	8.16

Table 2. WER performance of baseline HMM and GVP-HMM systems using a uniform parameter polynomial degree on Aurora 2

ComCtrl	System	0dB	5dB	10dB	15dB	20dB	Ave
BIC ($\rho = 1$)	mean	23.60	9.01	4.29	2.92	2.38	8.44
	mv	19.03	8.38	4.08	2.95	2.35	7.36
	tran2	22.60	8.62	4.17	2.86	2.38	8.13
	tran8	20.95	8.14	4.17	3.37	2.11	7.75
	mvt2	18.85	8.23	4.02	3.22	2.32	7.33
BIC ($\rho = 2$)	mean	23.60	8.92	4.23	3.01	2.35	8.42
	mv	18.71	8.11	4.14	2.98	2.62	7.31
	tran2	22.45	8.56	4.08	2.89	2.38	8.07
	tran8	20.71	7.99	4.02	3.40	2.11	7.65
	mvt2	18.38	7.81	4.11	3.13	2.71	7.23
BIC ($\rho = 3$)	mean	23.12	9.19	3.99	3.01	2.32	8.33
	mv	18.62	7.84	3.93	2.95	2.62	**7.19**
	tran2	22.80	8.59	4.05	2.92	4.41	8.15
	tran8	20.62	7.99	3.99	3.37	2.11	7.62
	mvt2	18.68	7.87	4.08	3.13	2.71	7.30

Table 3. WER performance of BIC optimized GVP-HMM systems with a locally varying polynomial degree on Aurora 2

The performance of a comparable set of GVP-HMMs with a locally varying polynomial degree are shown in table 3. These were derived using the BIC based complexity control method described in sections 3 and 4. For both the standard BIC penalty setting $\rho = 1$ and more aggressive configurations $\rho = 2$ or 3, complexity controlled GVP-HMMs were found to consistently outperform their comparable GVP-HMM baselines in table 2. For example, the complexity controlled "mv" system on average outperformed the baseline GVP-HMM system using a uniform 2nd degree for all polynomials in table 2 by 1.03%-1.20% absolute (12%-14% relative) in

error rate. The best performance was obtained using the BIC complexity controlled "mv" system with $\rho = 3$, as is shown in the 2nd line in the bottom section of table 3. This GVP-HMM "mv" system outperformed the the multi-style baseline "mcond.base" system shown in the 2nd line of table 2 by 1.76% absolute (20% relative). The setting of the BIC penalty ρ was also found to have only a small impact on WER performance in table 3.

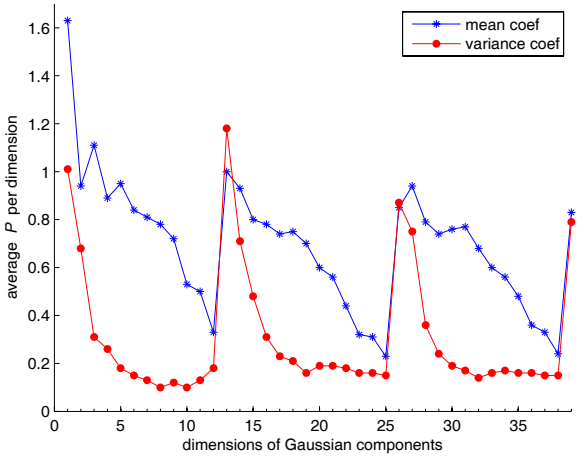

Fig. 2. Avg. polynomial degree P over feature dimensions in complexity controlled GVP-HMM "mv" system using BIC ($\rho = 1.0$)

A consistent reduction in model complexity was also obtained using the BIC complexity controlled GVP-HMM systems over the comparable GVP-HMM baselines. This is shown in the 5th and 6th columns of table 1. For example, using the BIC complexity controlled "mv" system with $\rho = 3$ (shown in the last line, 6th column in table 1), the number of polynomial coefficients was reduced by 57% relative from 249K in the baseline GVP-HMM system (shown in the last line, 5th column in table 1) down to 106K. As discussed in section 1, a locally varying polynomial degree is preferred as the variability introduced by noise manifests itself on a dimension by dimension basis in the acoustic space. This is shown in figure 2 for the BIC complexity controlled GVP-HMM "mv" system ($\rho = 1.0$) across different dimensions in the 39 dimensional feature space constructed by augmenting 1st to 12th order MFCC parameters plus log energy augmented with their 1st and 2nd order differentials. For both the static and differential features, a general trend can be found that lower order cepstras of up to the 3rd order and the log energy, which contain more information of speech, tend to use more complex polynomial trajectories than higher order cepstras.

5.3. Experiments on Mandarin In-Car Task

The medium vocabulary Mandarin In-Car navigation command recognition system was developed using 25 hours of clean training data. A multi-style training data set was constructed by artificially corrupting the clean speech data with added car engine noise. Noise corrupted speech data generated under six sentence level SNR conditions: 0dB, 4dB, 8dB, 12dB, 16dB and 20dB, were used in training, while a corrupted 5 hour test set consists of five sentence level SNR conditions: 2dB, 6dB, 10dB, 14dB, and 18dB, was used for character error rate (CER) evaluation. The baseline HMM acoustic models were ML trained using HTK [32] on 42-dimensional HLDA projected PLP features augmented with smoothed pitch parameters. Decision tree clustered cross-word tonal triphones HMMs were used. A total of 2.4k tied states with 12 components per state were used. A 5k word list and a tri-gram language model was used in decoding.

System	2dB	6dB	10dB	14dB	18dB	Ave
mcond.base	44.15	27.56	20.08	17.76	17.51	25.41
mean	39.95	27.31	21.62	17.87	16.84	24.72
mv	34.22	23.66	20.24	18.47	17.98	22.91
tran2	34.62	20.72	17.12	16.09	14.51	20.61
tran256	32.59	19.85	16.95	16.28	16.47	20.43
mvt2	31.05	21.87	17.88	17.31	16.62	20.95

Table 4. CER performance of baseline HMM and GVP-HMM systems using a uniform parameter polynomial degree on In-Car task

ComCtrl	System	2dB	6dB	10dB	14dB	18dB	Ave
BIC ($\rho = 1$)	mean	32.66	22.76	16.24	13.35	13.28	19.66
	mv	26.73	19.40	15.82	14.69	15.68	18.46
	tran2	33.43	20.32	16.61	15.57	15.90	20.37
	tran256	31.13	19.45	15.68	14.87	14.32	19.09
	mvt2	26.73	19.40	15.65	14.89	15.81	18.50
BIC ($\rho = 2$)	mean	32.65	22.64	16.19	13.37	13.14	19.60
	mv	26.88	19.08	15.45	14.37	15.66	**18.29**
	tran2	31.45	20.32	16.61	15.57	15.90	19.97
	tran256	30.96	19.55	15.70	14.87	14.40	19.10
	mvt2	26.51	19.25	15.51	14.76	15.68	18.34
BIC ($\rho = 3$)	mean	32.71	22.74	15.89	13.40	13.24	19.60
	mv	27.09	19.25	15.31	14.47	15.74	18.37
	tran2	31.45	20.32	16.61	15.57	15.90	19.97
	tran256	31.01	19.55	15.66	14.86	14.37	19.19
	mvt2	26.46	19.21	15.50	14.79	15.51	**18.29**

Table 5. CER performance of BIC optimized GVP-HMM systems with a locally varying polynomial degree on Mandarin In-Car Task

A set of experiments similar to those for Aurora 2 presented in table 3 were conducted on the In-Car data. Performance of the baseline multi-style and GVP-HMM systems, are shown in table 4. Consistent with the trend found in table 3, every BIC complexity controlled GVP-HMM system in table 5 outperformed its comparable GVP-HMM baseline in table 4. For example, the complexity controlled model space transform based GVP-HMM system, "tran256", using a matrix row level varying polynomial degree (4th, 9th and 14th lines in table 5) gave an average CER reduction of 5.7% absolute (23% relative) over the baseline "tran256" GVP-HMM system (5th line in table 4), and a 30% relative reduction in model complexity, as is shown in the 5th line, 7th and 8th columns in table 1. The two more complex BIC GVP-HMM systems, "mv" ($\rho = 2$) and "mvt2" ($\rho = 3$), both outperformed the multi-style trained baseline "mcond.base" system in the 1st line of table 4 by 7.12% absolute (28% relative). They gave the lowest average error rate among all GVP-HMM systems in table 5, and a 52%-55% relative reduction in the number of polynomial coefficients against their respective baselines, as is shown the 2nd and bottom line, 7th and 8th columns in table 1.

6. CONCLUSION

An efficient BIC based model complexity control technique was investigated for GVP-HMMs in this paper. The optimal polynomial degrees of Gaussian mean, variance and mean transform trajectories were automatically determined at local level. The proposed technique was shown to improve both the generalization and computational efficiency of GVP-HMM based acoustic models. Significant error rate reductions of 20%-28% relative obtained on Aurora 2 and a medium vocabulary Mandarin speech recognition task suggest the proposed method may be useful for speech recognition. Future research will focus on discriminative training and modelling multiple sources of acoustic variability.

7. REFERENCES

[1] A. R. Barron, J. J. Rissanen & B. Yu (1998). The Minimum Description Length Principle in Coding and Model ing, *IEEE Transactions on Information Theory*, pp. 2743–2760, vol. 44, no. 6, October 1998.

[2] A. Bjorck & V. Pereyra (1970). "Solution of Vandermonde Systems of Equations", Mathematics of Computation (American Mathematical Society) 24(112): pp. 893-903.

[3] N. Cheng, X. Liu & L. Wang (2011). "Generalized Variable Parameter HMMs for Noise Robust Speech Recognition", in *Proc. ISCA Interspeech2011*, pp. 482-484, Florence, Italy.

[4] N. Cheng, X. Liu & L. Wang (2011). "A flexible framework for HMM based noise robust speech recognition using generalized parametric space polynomial regression" *Science China, Information Sciences*, 54(2), pp. 2481-2491, 2011.

[5] W. Chou & W. Reichl (1999). Decision Tree State Tying Based on Penalized Bayesian Information Criterion, in *Proc. IEEE ICASSP1999*, Vol. 1, Phoenix.

[6] X. Cui & Y. Gong (2007). "A study of variable-parameter Gaussian mixture hidden Markov modeling for noisy speech recognition", *IEEE Transactions on Audio, Speech and Language Processing*, 15(4):1366-1376, 2007.

[7] A. P. Dempster, N. M. Laird & D. B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, 39(1):1-39,1977.

[8] L. Deng, J. Droppo & A. Acero (2005). "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion." *IEEE Transactions on Speech and Audio Processing*, 13.3 (2005): pp.412-421.

[9] J. Droppo, A. Acero & L. Deng (2002). "Uncertainty decoding with SPLICE for noise robust speech recognition", in *Proc. IEEE ICASSP2002* pp. 57-60, Orlando.

[10] K. Fujinaga, M. Nakai, H. Shimodaira & S. Sagayama (2001). "Multiple-Regression Hidden Markov Model", in *Proc. IEEE ICASSP2001*, Vol 1, pp. 513-516, Salt Lake City.

[11] M. J. F. Gales (1998). "Maximum likelihood linear transformations for HMM-based speech recognition", *Computer Speech and Language*, 12(2):75-98, 1998.

[12] O. Kalinli, M. Seltzer, J. Droppo & Alex Acero (2010). Noise Adaptive Training for Robust Automatic Speech Recognition, *IEEE Trans. on Audio, Speech and Language Processing*, 01/2010; 18:1889-1901.

[13] D. K. Kim & M. J. F. Gales (2011). Noisy constrained maximum-likelihood linear regression for noise-robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2), pp. 315-325.

[14] C.J. Leggetter & P.C. Woodland (1995). "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs", *Computer Speech and Language*, 9:171-186,1995.

[15] H. Liao & M. J. F. Gales (2008). "Issues with uncertainty decoding for noise robust speech recognition", *Speech Communication*, 50:265-277, 2008.

[16] Y. Li, X. Liu & L. Wang (2012). "Structured Modeling Based on Generalized Variable Parameter HMMs and Speaker Adaptation," in *Proc. IEEE ISCSLP2012*, pp. 136-140, Hong Kong, China.

[17] Y. Li, X. Liu & L. Wang (2013). "Feature Space Generalized Variable Parameter HMMs for Noise Robust Speech Recognition", to appear in *Proc. ISCA Interspeech2013*, Lyon, France.

[18] S. Lin, B. Chen & Y-M Yeh (2009). Exploring the Use of Speech Features and Their Corresponding Distribution Characteristics for Robust Speech Recognition, *IEEE Transactions on Audio Speech and Language Processing*, 17(1), pp.84-94, Jan. 2009.

[19] R. Lippmann, E. Martin & D. Paul (1987). "Multi-style training for robust isolated-word speech recognition", in *Proc. IEEE ICASSP1987*, pp. 705-708, Dallas, Texas.

[20] X. Liu, M. J. F. Gales & P. C. Woodland (2003). "Automatic complexity control for HLDA systems", in *Proc. IEEE ICASSP2003*, Vol. 1, pp. 132-135, Hong Kong, China.

[21] X. Liu & M. J. F. Gales (2003). "Automatic Model Complexity Control Using Marginalized Discriminative Growth Functions", in *Proc. IEEE ASRU2003*, pp. 37-42, St. Thomas, U.S. Virgin Islands.

[22] X. Liu & M. J. F. Gales (2004). "Model Complexity Control and Compression Using Discriminative Growth Functions", in *Proc. IEEE ICASSP2004*, Vol. 1, pp. 797-800, Montreal.

[23] X. Liu & M. J. F. Gales (2007). "Automatic Model Complexity Control Using Marginalized Discriminative Growth Functions," *IEEE Transactions on Audio, Speech and Language Processing*, vol.15, no.4, pp.1414-1424, May 2007.

[24] T. T. Kristjansson & B. J. Frey (2002). "Accounting for uncertainty in observations: A new paradigm for robust speech recognition", in *Proc. IEEE ICASSP2002*, pp. 61-64, Orlando.

[25] R. Martin (1993). "An efficient algorithm to estimate the instantaneous SNR speech signals", in *Proc. Eurospeech1993*, pp. 1093-1096, Berlin.

[26] R. M. Neal (1993) Probabilistic Inference using Markov Chain Monte Carlo Methods, Technical Report, CGT-TR-93-1, Department of Computer Science, University of Toronto, 1993.

[27] C. Runge (1901). "Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten", *Zeitschrift für Mathematik und Physik*, 46:224-243.

[28] G. Schwartz (1978). Estimating the Dimension of a Model, *The Annals of Statistics*, pp. 461-464, Vol. 6, No. 2, February 1978.

[29] M. Seltzer, D. Yu & Y. Wang (2013). An Investigation Of Deep Neural Networks For Noise Robust Speech Recognition, in *Proc. IEEE ICASSP2013*, Vancouver.

[30] S. Watanabe et al. (2004) Variational Bayesian Estimation and Clustering for Speech Recognition, *IEEE Transactions on Speech and Audio Processing*, pp. 365-381, Vol. 12, 2004.

[31] A. Azevedo-Filho & R. Shachter (1994), "Laplace's Method Approximations for Probabilistic Inference in Belief Networks with Continuous Variables", in R. Mantaras & D. Poole, *Uncertainty in Artificial Intelligence*, San Francisco, CA: Morgan Kauffman.

[32] S. Young et al., "The HTK Book Version 3.4.1 ", 2009.

[33] D. Yu, L. Deng, Y. Gong & A. Acero (2009), "A Novel Framework and Training Algorithm for Variable-Parameter Hidden Markov Models", *IEEE Transactions on Audio, Speech and Language Processing*, Vol 17(7), pp. 1348-1360, 2009.