Automatic Complexity Control of Generalized Variable Parameter HMMs for Noise Robust Speech Recognition

Rongfeng Su, Xunying Liu, *Member, IEEE*, and Lan Wang, *Member, IEEE*

***Abstract*—An important part of the acoustic modelling problem for automatic speech recognition (ASR) systems is to handle the mismatch against a target environment created by time-varying external factors such as ambient noise. One possible solution to this problem is to introduce controllability to the underlying acoustic model to allow an instantaneous adaptation to the underlying noise condition. Along this line, the continuous trajectory of optimal, well matched model parameters against the varying noise can be explicitly modelled using, for example, 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 60% relative were also obtained over the baseline GVP-HMM systems using a uniformly assigned polynomial degree.**

***Index Terms*—Complexity control, generalized variable parameter HMMs, robust speech recognition, variable noise.**

I. Introduction

A CRUCIAL task of automatic speech recognition (ASR) systems is to robustly handle the mismatch against a target environment introduced by external factors such as environment noise. When these factors are of time-varying nature, this problem becomes even more challenging. To handle this issue, a range of model based techniques can be used: multi-style training [27] exploits the implicit modelling power of mixture models, or more recently deep neural networks [38], to obtain a good generalization to unseen noise conditions; uncertainty decoding [11], [12], [32], [22], propagates the uncertainty that varies with the noise represented by, for example, a conditional distribution of the corrupted speech, into the recognizer; noise adaptive training [17], [19], [21] techniques remove the variability introduced to the multi-style speech data by the environment noise using, for example, vector Taylor series (VTS) [18] expansion or joint uncertainty decoding techniques [21].

An alternative approach to the above techniques is to directly introduce controllability to the underlying acoustic model. It is hoped that by explicitly learning the underlying effect imposed by evolving acoustic factors, such as noise, on the acoustic realization of speech, an instantaneous adaptation to these factors becomes possible. One class of statistical models along this line includes multiple regression HMMs (MR-HMM) [13], [26] and variable parameter HMMs (VP-HMM) [9], [43], [44], [45]. They explicitly approximate the continuous trajectories of optimal HMM model parameters against time-varying acoustic factors using polynomial functions. Under this parameter trajectory modelling framework, several forms of acoustic factors have been investigated in previous research. These include prosody [13], environment noise condition represented by the signal-to-noise ratio (SNR) [9], [43], [44], [45], as is also considered in this paper, and more recently articulatory features for speech synthesis [26].

Under the MR-HMM or VP-HMM framework, Gaussian component level polynomial modelling of mean and optionally variance trajectories are used. This often results in a dramatic increase in the number of free parameters in the system. In order to robustly estimate the desired polynomial coefficients, a large amount of multi-style speech data is usually required. As Gaussian component level polynomial interpolation is performed during recognition for each noise condition in the test data, conventional MR-HMMs or VP-HMMs are also computationally expensive to use in recognition time. Hence, more compact forms of parameter trajectory modelling techniques are preferred. An extension to both MR-HMMs and VP-HMMs, generalized variable parameter HMMs (GVP-HMMs), were proposed in [6], [7], [23], [24]. In addition to Gaussian parameters, GVP-HMMs can also provide a more compact trajectory modelling for model or feature space tied linear transformations, and thus provide a flexible form of parameter trajectory modelling. For a given noise condition, present or unseen in the training data, GVP-HMMs can instantaneously produce the matching Gaussian component or linear transform parameters by-design without requiring any multi-pass decoding and adaptation process. An important issue associated with MR-HMMs, VP-HMMs and GVP-HMMs in general is the appropriate polynomial degree to use. The underlying polynomial degree being used determines the precise nature

Manuscript received December 04, 2013; revised June 08, 2014; accepted November 07, 2014. Date of publication November 20, 2014; date of current version January 14, 2015. This work was supported in part by the National Natural Science Foundation of China (NSFC61135003) and in part by National Fundamental Research Grant of Science and Technology (973 Project: 2013CB329305) Shenzhen Fundamental Research Programs JC01005280621A and JCYJ20130401170306806. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Murat Saraclar.

R. Su and L. Wang are with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences/The Chinese University of Hong Kong, Shenzhen 518055, China (e-mail: rf.su@siat.ac.cn; lan.wang@siat.ac.cn).

X. Liu is with the Engineering Department, Cambridge University, Cambridge CB2 1PZ, U.K. (e-mail: xl207@cam.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2372901

of the approximated effect of various factors, such as the noise condition, on the actual acoustic realization. When higher degree polynomials are used, an oscillation effect known as the *Runge's phenomenon* and the resulting increase in interpolation error can occur [35]. For this reason, lower degree polynomials uniformly assigned with the same order, for example, the second order, were used in previous research [6], [7], [9], [13], [23], [24].

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, in the presence of car engine noise that are more concentrated in the lower frequency range, lower order cepstral parameters, which contain richer information of speech than higher order cepstra, are more prone to the distortion introduced by noise. A uniformly assigned polynomial degree can cause an under-fitting for the lower order cepstra, while at the same time an over-fitting for the higher order cepstra that are more related to noise in nature and thus more invariant to the distortion. Such lack of modelling flexibility limits the power of GVP-HMMs to more accurately capture the underlying effect of noise on the acoustic realization of speech signals, and the possible improvements that can be obtained from such systems. Secondly, over-fitting higher degree polynomials in practice further increases the interpolation cost during recognition. Hence, a more flexible locally varying polynomial degree configuration is required.

In this paper the above task is converted to a classic automatic model selection problem. A novel and efficient Bayesian model complexity control method for GVP-HMMs is proposed. The optimal polynomial degrees of Gaussian mean, variance and model space linear transform trajectories against environment noise are automatically determined at local level. The rest of the paper is organized as follows. The GVP-HMM framework is reviewed in Section II. An efficient Bayesian model complexity control criterion is presented in Section III. The detailed complexity control algorithm for GVP-HMMs is proposed in Section IV. In Section V 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.

II. Generalized Variable Parameter HMMs

Generalized variable parameter HMMs (GVP-HMMs) [6], [7], [23], [24] can explicitly model the trajectory of optimal acoustic parameters that vary with respect to a scalar variable, such as the underlying noise condition characterized by the signal-to-noise ratio (SNR). The type of parameter trajectories are not restricted to those of means and covariances of conventional decision tree tied Gaussian mixture HMMs. Other more compact forms of model parameters, such as model or feature space linear transformations [14], [20], may also be considered. In this paper, trajectories of Gaussian mean transforms are used. For a D dimensional observation $\boldsymbol{o}_t$ emitted from Gaussian mixture component m, assuming Pth order polynomials are used, this is given by

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

where $\mathbf{v}_t^{\top}$ is a $(P+1)$ dimensional Vandermonde vector [5], such that $\mathbf{v}_{t,p} = v_t^{p-1}$. v_t is an auxiliary feature, and in this paper, the SNR condition [33] measured 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. $\boldsymbol{\mu}^{(m)}(\cdot)$, $\Sigma^{(m)}(\cdot)$ and $\boldsymbol{W}^{(r_m)}(\cdot)$ are the Pth order mean, covariance and mean transform trajectory polynomials of component m respectively. When diagonal covariances are used for computational efficiency, the trajectories of the ith dimension of the mean, variance, and the transform element in row i and column j, are expressed as

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

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

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

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 performance.

III. 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 classic Bayesian complexity control techniques, it is assumed that by increasing the likelihood on some unseen data, the underlying ASR system's error rate on the same data will decrease. When no prior knowledge over individual model structures is available, the optimal model structure or complexity, is determined by maximizing the following Bayesian *evidence* integral,

$$\begin{aligned}
\hat{\mathcal{M}} &= \arg\max_{\{\mathcal{M}\}} \{p(\mathcal{O}|\mathcal{W}, \mathcal{M})\} \\
&= \arg\max_{\{\mathcal{M}\}} \left\{ \int p(\mathcal{O}|\theta, \mathcal{W}, \mathcal{M}) p(\theta|\mathcal{M}) \mathrm{d}\theta \right\}
\end{aligned} \quad (4)$$

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

For conventional HMMs and their extended forms such as 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) [37], a second order asymptotic expansion based Laplace's approximation [3], [29], [30], [31], variational approximation [40], and Markov Chain Monte Carlo (MCMC) based sampling schemes [34]. Among these, BIC (or equivalently MDL [4]) 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 $\hat{\theta}$. The model selection is based on the following approximation

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

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 [8], [28]. When $\rho = 1$, BIC was shown to be a first order asymptotic expansion of the evidence integral [37].

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 during model selection. Let $\tilde{\theta}$ denote the *current* parameterization for $\mathcal{M}$. Using the EM algorithm the following inequality may be derived [10]

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

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

$$\mathcal{Q}_{\mathrm{ml}}^{(\mathcal{M})}(\theta, \tilde{\theta}) = \sum_{m,t} \gamma_m(t) \log p(\boldsymbol{o}_t|\psi_t = m, \theta, \mathcal{M}). \qquad (7)$$

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

For acoustic model training of ASR systems, the majority of the time is spent accumulating these sufficient statistics to estimate the model parameters. Thus, 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 [29], [30], [31][1]. For example, when determining the appropriate order of a Gaussian component mean's trajectory polynomial on a particular dimension in equation (2) for a GVP-HMM system, the sufficient statistics $\{\gamma_m(t)\}$ to be used for a range of candidate polynomial degree settings 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 single Gaussian mean vector in the system. In the same way, sufficient statistics can also be shared when determining the degrees of Gaussian variance or mean transformation trajectory polynomials in equation (2).

It is clear that the only term in the lower bound of equation (6) dependent on the model parameters, θ, is the auxiliary function $\mathcal{Q}_{\mathrm{ml}}^{(\mathcal{M})}(\theta, \tilde{\theta})$. When multiple model structures use the same set of statistics, the rank ordering derived from the marginalization of $\mathcal{L}_{\mathrm{ml}}^{(\mathcal{M})}(\theta, \tilde{\theta})$ is equivalent to that of $\mathcal{Q}_{\mathrm{ml}}^{(\mathcal{M})}(\theta, \tilde{\theta})$[2].

$$\hat{\mathcal{M}} = \arg\max_{\mathcal{M}} \left\{ \int \exp\left\{ \mathcal{Q}_{\mathrm{ml}}^{(\mathcal{M})}(\theta, \tilde{\theta}) \right\} p(\theta|\mathcal{M}) \mathrm{d}\theta \right\} \qquad (8)$$

To further reduce the computational cost, the above integral over the auxiliary function in equation (8) is efficiently computed using a BIC style approximation in this paper. Under these conditions, the optimal model complexity is finally determined by

$$\hat{\mathcal{M}} = \arg\max_{\mathcal{M}} \left\{ \mathcal{Q}_{\mathrm{ml}}^{(\mathcal{M})}(\hat{\theta}, \tilde{\theta}) - \rho \cdot \frac{k}{2} \log \mathcal{T} \right\}. \qquad (9)$$

IV. Model Complexity Control for GVP-HMMs

When using the lower bound based BIC metric of equation (9) 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 II, the associated ML auxiliary function is given by [6], [7], [10], [24],

$$\begin{aligned} \mathcal{Q}_{\mathrm{ml}}^{\mathrm{GVP}}(\theta, \tilde{\theta}) = \sum_{m,t} \gamma_m(t) \log p \Big(&\boldsymbol{o}^{(t)}; \boldsymbol{\mu}^{(m)}(\mathbf{v}_t), \\ &\Sigma^{(m)}(\mathbf{v}_t), \boldsymbol{W}^{(r_m)}(\mathbf{v}_t) \Big) \end{aligned} \qquad (10)$$

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,

$$\begin{aligned} \mathcal{Q}_{\mathrm{ml}}^{(\mu_i^{(m)})}(\theta, \tilde{\theta}) &= -\frac{1}{2} \mathbf{c}^{(\mu_i^{(m)})\top} \mathbf{U}^{(\mu_i^{(m)})} \mathbf{c}^{(\mu_i^{(m)})} \\ &\quad + \mathbf{k}^{(\mu_i^{(m)})} \mathbf{c}^{(\mu_i^{(m)})} + \mathrm{const} \\ \mathcal{Q}_{\mathrm{ml}}^{(\sigma_{i,i}^{(m)})}(\theta, \tilde{\theta}) &= -\frac{1}{2} \mathbf{c}^{(\sigma_i^{(m)})\top} \mathbf{U}^{(\sigma_i^{(m)})} \mathbf{c}^{(\sigma_i^{(m)})} \\ &\quad + \mathbf{k}^{(\sigma_i^{(m)})} \mathbf{c}^{(\sigma_i^{(m)})} + \mathrm{const}' \\ \mathcal{Q}_{\mathrm{ml}}^{(w_i^{(r_m)})}(\theta, \tilde{\theta}) &= -\frac{1}{2} \mathbf{c}^{(w_i^{(r_m)})\top} \mathbf{U}^{(w_i^{(r_m)})} \mathbf{c}^{(w_i^{(r_m)})} \\ &\quad + \mathbf{k}^{(w_i^{(r_m)})} \mathbf{c}^{(w_i^{(r_m)})} + \mathrm{const}'' \end{aligned} \qquad (11)$$

[1]Computing the difference between the left and right hand side of equation (6), expressed as the KL divergence term between the posterior distributions computed using θ and $\tilde{\theta}$ respectively, requires explicitly computing the former posterior distribution over the entire training data set using each possible GVP-HMM structural configuration and its associated parameterization. Given the vast number of candidate GVP-HMM models to consider, this would be computationally infeasible. Hence, an approximation is required.

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

where the constant terms independent of the coefficient vectors $\mathbf{c}^{(\cdot)}$ can be ignored.

After setting the above gradients against the respective polynomial coefficient vectors to zero, the following ML solutions of the coefficient vectors can then be derived,

$$\begin{aligned}\hat{\mathbf{c}}^{(\mu_i^{(m)})} &= \mathbf{U}^{(\mu_i^{(m)})-1}\mathbf{k}^{(\mu_i^{(m)})} \\ \hat{\mathbf{c}}^{(\sigma_{i,i}^{(m)})} &= \mathbf{U}^{(\sigma_{i,i}^{(m)})-1}\mathbf{k}^{(\sigma_{i,i}^{(m)})} \\ \hat{\mathbf{c}}^{(w_i^{(r_m)})} &= \mathbf{U}^{(w_i^{(r_m)})-1}\mathbf{k}^{(w_i^{(r_m)})}\end{aligned} \tag{12}$$

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

$$\begin{aligned}\mathbf{U}^{(\mu_i^{(m)})} &= \sum_t \gamma_m(t)\sigma_{i,i}^{(m)-1}(\mathbf{v}_t)\mathbf{v}_t^\top\mathbf{v}_t \\ \mathbf{k}^{(\mu_i^{(m)})} &= \sum_t \gamma_m(t)\sigma_{i,i}^{(m)-1}(\mathbf{v}_t)o_i^{(t)}\mathbf{v}_t^\top \\ \mathbf{U}^{(\sigma_{i,i}^{(m)})} &= \sum_t \gamma_m(t)\,\breve{\sigma}_{i,i}^{\,(m)}\mathbf{v}_t^\top\mathbf{v}_t \\ \mathbf{k}^{(\sigma_{i,i}^{(m)})} &= \sum_t \gamma_m(t)\Big(o_i^{(t)}-\mu_i^{(m)}(\mathbf{v}_t)\Big)^2\mathbf{v}_t^\top\end{aligned} \tag{13}$$

$\mathbf{U}^{(w_i^{(r_m)})}$ is a $[(D+1)\times(P+1)]\times[(D+1)\times(P+1)]$ meta Vandermonde matrix,

$$\mathbf{U}^{(w_i^{(r_m)})} = \begin{bmatrix}\mathbf{U}^{(w_{i,1}^{(r_m)})} \\ \vdots \\ \mathbf{U}^{(w_{i,j}^{(r_m)})} \\ \vdots \\ \mathbf{U}^{(w_{i,D+1}^{(r_m)})}\end{bmatrix} \tag{14}$$

and $\mathbf{k}^{(w_i^{(r_m)})}$ a $(D+1)\times(P+1)$ dimensional meta regression target vector. This is given by

$$\mathbf{k}^{(w_i^{(r_m)})} = \begin{bmatrix}\mathbf{k}^{(w_{i,1}^{(r_m)})} \\ \vdots \\ \mathbf{k}^{(w_{i,j}^{(r_m)})} \\ \vdots \\ \mathbf{k}^{(w_{i,D+1}^{(r_m)})}\end{bmatrix}. \tag{15}$$

The sub-matrix $\mathbf{U}^{(w_{i,j}^{(r_m)})}$ and sub-vector $\mathbf{k}^{(w_{i,j}^{(r_m)})}$ in the above that are associated with transform element $w_{i,j}^{(r_m)}$ are computed as

$$\begin{aligned}\mathbf{U}^{(w_{i,j}^{(r_m)})} = \Bigg[&\sum_{m\in r_m,t}\gamma_m(t)\sigma_{i,i}^{(m)-1}(\mathbf{v}_t)\zeta_{t,j}^{(m)}\zeta_{t,1}^{(m)}\mathbf{v}_t^\top\mathbf{v}_t\ ,\ldots, \\ &\sum_{m\in r_m,t}\gamma_m(t)\sigma_{i,i}^{(m)-1}(\mathbf{v}_t)\zeta_{t,j}^{(m)}\zeta_{t,i}^{(m)}\mathbf{v}_t^\top\mathbf{v}_t,\ldots, \\ &\sum_{m\in r_m,t}\gamma_m(t)\sigma_{i,i}^{(m)-1}(\mathbf{v}_t)\zeta_{t,j}^{(m)}\zeta_{t,D+1}^{(m)}\mathbf{v}_t^\top\mathbf{v}_t\Bigg] \\ \mathbf{k}^{(w_{i,j}^{(r_m)})} = &\sum_{m\in r_m,t}\gamma_m(t)\sigma_{i,i}^{(m)-1}(\mathbf{v}_t)\zeta_{t,j}^{(m)}o_i^{(t)}\mathbf{v}_t^\top\end{aligned} \tag{16}$$

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 II.

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 (13) and (16) are accumulated for the highest order $P_{\mathtt{max}}$ being considered. The corresponding statistics for any other order $0 \le P^{(\mu_i^{(m)})} < P_{\mathtt{max}}$ can be derived by taking the associated submatrices or subvectors from the full matrix statistics accumulated for $P_{\max}$. Using these statistics and the ML solutions in equation (12), the ML auxiliary function associated with $\mu_i^{(m)}(\cdot)$ in equation (11), 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 (9) 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 below.

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

accumulate sufficient statistics $\{\gamma_m(t)\}$ and those of Eq. (13) using a baseline system with no complexity control;
for each Gaussian component m in the system **do**
  for each dimension of component m's mean vector **do**
    for each polynomial degree $P^{(\mu_i^{(m)})} \in [0, P_{\mathtt{max}}]$**do**
      evaluate the model selection metric in equation (9) for the current $P^{(\mu_i^{(m)})}$ degree mean polynomial $\mu_i^{(m)}(\cdot)$ using the auxiliary function $\mathcal{Q}_{\mathrm{ml}}^{(\mu_i^{(m)})}(\theta,\tilde{\theta})$ of equation (11) and sufficient statistics, $\mathbf{U}^{(\mu_i^{(m)})}$, $\mathbf{k}^{(\mu_i^{(m)})}$ in equation (13) and $\hat{c}^{(\mu_i^{(m)})}$ in equation (12).
    end for
    select $\hat{P}^{(\mu_i^{(m)})}$ that maximizes the above criterion;
    take the ML estimate $\hat{\mathbf{c}}^{(\mu_i^{(m)})}$ associated with $\hat{P}^{(\mu_i^{(m)})}$;
  end for
end for
output coefficients for all mean polynomials $\{\hat{\mathbf{c}}^{(\mu_i^{(m)})}\}$ in the system with a locally varying polynomial degree.

The same approach can also be used to determine the optimal degree of Gaussian variance and mean transform polynomials, by evaluating the respective auxiliary functions with their respective sufficient statistics to compute the BIC metric in equation (9). Unless otherwise stated, in all experiments of this paper, the sufficient statistics $\{\gamma_m(t)\}$ and those of equation (13) required for the above complexity control algorithm are accumulated using a baseline HMM system.

V. Experimental Results

In this section, complexity controlled GVP-HMM systems are evaluated on two tasks: the Aurora 2 speaker independent digit sequence recognition task and a medium vocabulary Mandarin Chinese In-car navigation command recognition task. In all experiments utterance level SNR features are used

TABLE I
DESCRIPTION OF VARIOUS GVP-HMMS ON AURORA 2: PARAMETER POLYNOMIAL TYPES AND THE AVERAGE NUMBER OF POLYNOMIAL COEFFICIENTS ACROSS SUBWAY, BABBLE, CAR AND EXHIBITION

System	Poly. Types			#Ave. PolyCoef (Aurora2)	
	mean	var	tran	base	BIC($\rho = 1/2/3$)
mean	√	×	×	120K	64.75K/56.1K/52.05K
mv	√	√	×	240K	114.5K/101.38K/95.53K
tran2	×	×	√	9.4K	8.51K/8.35K/8.26K
tran8	×	×	√	37.4K	32.75K/32.45K/32.15K
mvt2	√	√	√	249K	121.3K/107.93K/102.23K

for GVP-HMM systems. For all GVP-HMM parameter polynomials the range of candidate degree to consider is from 0 to 5.

A. Experiments on Aurora 2

The multi-style training data provided by the Aurora 2 speaker independent digit sequence recognition database [16], [36] covers 4 different types of noisy types: subway, babble, car and exhibition noise. For each of these 4 training noise types, a total of 420 utterances from four different SNR conditions (-5 dB, 5 dB, 15 dB, 25 dB) were used to train both the baseline multi-style HMM baseline system and a range of GVP-HMM systems with different modelling configurations associated with each particular noise type. 39 dimensional MFCC plus log energy features including their 1st and 2nd order differentials were used in acoustic model training. Word error rate (WER) evaluation was performed on two standard Aurora 2 test sets: test set A which is based on the same four noise types as the multi-style training data, and test set B which is based on four unseen noise types: restaurant, street, airport and station. For the both test sets, a total 1001 utterances selected from each of five different test SNR conditions: 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, were used in recognition performance evaluation.

Description of Aurora 2 GVP-HMM Systems: As discussed in Section II, 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 for different purposes. The description of various GVP-HMM systems and the number of polynomial coefficients used on the Aurora 2 data is shown in Table I.

This table is partitioned into three sections. In the first section, there are two standard VP-HMM systems, which model the mean, and optionally variance, of each Gaussian component, shown as "mean" and "mv" systems respectively. In the second section of Table I, two transform based GVP-HMM systems "tran2" and "tran8" are given. In these two systems the polynomial trajectories of 2 or 8 mean transforms are modelled respectively. In the bottom section of Table I, the most complex GVP-HMM system configuration is presented. This "mvt2" GVP-HMM system models the trajectories of both Gaussian means and variances, and those of 2 mean transforms. In the last 2 columns, the number of polynomial coefficients for various GVP-HMM systems are given. All the baseline GVP-HMMs with no complexity control using 2nd degree polynomials for all parameter trajectories are shown as "base" in the 5th column. The average number of polynomial coefficients of complexity controlled GVP-HMM systems computed over the four training noise types with varying settings, $\rho = 1, 2, 3$, are in the 6th column of Table I.

TABLE II
WER PERFORMANCE OF MULTI-STYLE TRAINED BASELINE HMM SYSTEM, BASELINE GVP-HMM SYSTEMS WITH A UNIFORMLY ASSIGNED PARAMETER POLYNOMIAL DEGREE, AND COMPLEXITY CONTROLLED GVP-HMM SYSTEMS USING A LOCALLY VARYING POLYNOMIAL DEGREE ON SUBWAY NOISE DATA

ComCtrl	System	0dB	5dB	10dB	15dB	20dB	Ave
×	mcond.base	21.25	7.55	3.78	2.36	2.27	7.44
×	mean	19.31	6.79	2.98	1.87	1.57	6.82
	mv	17.16	6.88	3.04	2.18	1.96	6.24
	tran2	20.17	6.57	3.29	2.06	1.66	6.75
	tran8	20.76	6.60	3.38	1.90	1.60	6.85
	mvt2	20.69	7.18	3.25	2.24	2.03	7.48
BIC ($\rho = 1$)	mean	19.22	7.03	2.82	1.84	1.66	6.51
	mv	16.64	6.63	2.98	1.93	2.03	6.04
	tran2	18.82	6.97	3.65	2.43	2.09	6.79
	tran8	17.72	6.72	3.41	2.15	1.96	6.39
	mvt2	17.07	6.72	2.95	2.06	2.03	6.17
BIC ($\rho = 2$)	mean	19.10	6.94	2.89	1.81	1.78	6.50
	mv	16.58	6.32	2.86	2.00	1.96	**5.94**
	tran2	18.70	6.97	3.65	2.43	2.09	6.77
	tran8	17.87	6.72	3.41	2.15	1.96	6.42
	mvt2	17.32	6.60	2.73	2.06	1.93	6.13
BIC ($\rho = 3$)	mean	19.22	6.75	2.92	1.87	1.84	6.50
	mv	16.86	6.51	2.79	2.06	2.12	6.07
	tran2	18.73	6.97	3.65	2.43	2.09	6.77
	tran8	17.62	6.88	3.41	2.09	1.96	6.39
	mvt2	17.44	6.94	2.82	2.09	2.09	6.28

TABLE III
WER PERFORMANCE OF MULTI-STYLE TRAINED BASELINE HMM SYSTEM, BASELINE GVP-HMM SYSTEMS WITH A UNIFORMLY ASSIGNED PARAMETER POLYNOMIAL DEGREE, AND COMPLEXITY CONTROLLED GVP-HMM SYSTEMS USING A LOCALLY VARYING POLYNOMIAL DEGREE ON BABBLE NOISE DATA

ComCtrl	System	0dB	5dB	10dB	15dB	20dB	Ave
×	mcond.base	30.47	12.09	6.53	4.59	4.08	11.55
×	mean	31.08	10.25	4.35	2.78	2.84	10.26
	mv	28.60	9.85	4.63	3.36	3.17	9.92
	tran2	32.95	12.24	5.80	3.45	3.17	11.52
	tran8	32.62	11.12	4.96	2.78	2.84	10.86
	mvt2	35.40	12.00	4.59	3.08	3.02	11.62
BIC ($\rho = 1$)	mean	30.23	9.64	4.32	2.87	3.08	10.03
	mv	30.02	9.28	4.26	3.08	3.23	9.97
	tran2	31.38	11.52	5.62	3.23	3.30	11.01
	tran8	28.87	10.49	4.35	2.90	2.78	9.88
	mvt2	30.77	9.52	4.14	3.26	3.17	10.17
BIC ($\rho = 2$)	mean	29.63	9.55	4.29	2.96	3.20	9.93
	mv	29.29	8.83	4.23	3.36	3.51	9.84
	tran2	31.38	11.58	5.62	3.23	3.30	11.02
	tran8	30.53	10.22	4.20	2.93	2.78	10.13
	mvt2	30.14	9.79	4.14	3.26	3.20	10.11
BIC ($\rho = 3$)	mean	29.08	9.58	4.29	2.99	3.36	9.86
	mv	27.63	9.19	4.44	3.36	3.72	**9.67**
	tran2	31.38	11.58	5.62	3.23	3.30	11.02
	tran8	30.53	10.25	4.17	2.87	2.81	10.13
	mvt2	29.72	9.64	4.23	3.17	3.63	10.08

Performance on Set A of Matched Noise Types: For the four different noise types in test set A, the WER performance of the baseline multi-style HMM systems, a range of standard GVP-HMM systems of no complexity control using a uniform parameter polynomial degree, and a comparable set of complexity controlled GVP-HMM systems using a locally varying polynomial degree as described in Table I, are shown from Table II to V. The "mv" GVP-HMM system in 3rd line of each table, which models both Gaussian mean and variance polynomials consistently outperformed a simpler "mean" system across all four noise types. On the subway data, for example, the associated "mv" system (3rd line in Table II) outperformed

TABLE IV
WER PERFORMANCE OF MULTI-STYLE TRAINED BASELINE HMM SYSTEM, BASELINE GVP-HMM SYSTEMS WITH A UNIFORMLY ASSIGNED PARAMETER POLYNOMIAL DEGREE, AND COMPLEXITY CONTROLLED GVP-HMM SYSTEMS USING A LOCALLY VARYING POLYNOMIAL DEGREE ON CAR NOISE DATA

ComCtrl	System	0dB	5dB	10dB	15dB	20dB	Ave
×	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
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 V
WER PERFORMANCE OF MULTI-STYLE TRAINED BASELINE HMM SYSTEM, BASELINE GVP-HMM SYSTEMS WITH A UNIFORMLY ASSIGNED PARAMETER POLYNOMIAL DEGREE, AND COMPLEXITY CONTROLLED GVP-HMM SYSTEMS USING A LOCALLY VARYING POLYNOMIAL DEGREE ON EXHIBITION NOISE DATA

ComCtrl	System	0dB	5dB	10dB	15dB	20dB	Ave
×	mcond.base	23.46	10.27	4.91	3.09	2.72	8.89
×	mean	21.45	8.15	4.38	2.75	2.10	7.77
	mv	19.69	7.99	4.07	2.56	2.31	7.32
	tran2	22.99	9.87	4.60	2.84	2.47	8.55
	tran8	22.53	9.78	4.44	2.65	2.28	8.34
	mvt2	25.19	9.63	4.07	2.81	2.34	8.81
BIC ($\rho = 1$)	mean	20.77	8.02	4.17	2.62	2.19	7.55
	mv	19.10	7.71	3.89	2.50	2.34	7.11
	tran2	21.60	9.44	4.50	2.81	2.53	8.18
	tran8	21.30	9.13	4.41	2.65	2.34	7.97
	mvt2	19.32	7.84	4.01	2.56	2.34	7.21
BIC ($\rho = 2$)	mean	20.49	8.45	4.26	2.68	2.41	7.66
	mv	19.04	7.90	3.95	2.65	2.31	7.17
	tran2	21.60	9.47	4.50	2.81	2.53	8.18
	tran8	21.20	9.07	4.47	2.62	2.31	7.93
	mvt2	19.38	8.05	3.92	2.53	2.41	7.26
BIC ($\rho = 3$)	mean	20.59	8.55	4.32	2.59	2.44	7.70
	mv	19.14	7.59	3.89	2.56	2.38	7.11
	tran2	21.60	9.47	4.50	2.81	2.53	8.18
	tran8	21.17	9.13	4.35	2.62	2.31	7.92
	mvt2	18.86	7.56	3.86	2.56	2.47	**7.06**

the comparable "mean" system (2nd line in Table II) by 0.58% absolute (8.5% relative).

For both the standard BIC penalty setting $\rho = 1$ and more aggressive configurations $\rho = 2$ or 3, complexity controlled GVP-HMMs were also found to consistently outperform their comparable GVP-HMM baselines using a uniformly assigned 2nd degree from Table II to V across all four noise types. Take the most complex GVP-HMM configuration "mvt2" (previously also shown in the last line of Table I) as an example, a significant WER reduction of 15.44% relative was obtained across all four noise types on average using the BIC complexity control method proposed in Section IV. On the exhibition noise data in Table V, for example, the BIC ($\rho = 3$) complexity controlled "mvt2" GVP-HMM system (highlighted in bold, last line Table V) reduced the WER from 8.81% produced by the comparable baseline (6th line in Table V) down to 7.06% (19.86% relative). This system gave the best performance on the exhibition noise data and outperformed the comparable multi-style HMM baseline "mcond.base" system by 1.83% absolute (20.58% relative). For the other three noise types, the complexity controlled GVP-HMM systems with the lowest error rate are also highlighted in bold in each table. An average WER reduction of 7.5% absolute (19% relative) over the multi-style baseline "mcond.base" HMM systems was obtained. All these results confirms the hypothesis raised in Section I that a more flexible locally optimized polynomial degree improves the underlying power of GVP-HMMs to accurately capture the effect of noise on the acoustic representation of speech. The setting of the BIC penalty ρ was also found to have only a small impact on WER performance in these four tables.

As discussed in Sections III and IV, in order to improve the efficiency during complexity control for GVP-HMMs, a range of GVP-HMM model structures can share the same set of statistics generated using a single system. This allows the lower bound in (6) to be efficiently computed. So far in all experiments of this paper, these sufficient statistics are accumulated using a baseline HMM system. In practice, the choice of the system to generate these shared sufficient statistics was found to have minimum effect on the recognition performance of GVP-HMM systems. A set of contrast experiments were conducted using baseline GVP-HMM systems with no complexity control to generate these sufficient statistics, as are shown in Tables VI, VII, VIII and IX. Compared with those results shown in Tables II, III, IV and V, only a marginal error rate reduction of 0.1%-0.2% was obtained for the best GVP-HMM systems (highlighted in bold) across 4 noise types.

Compared with the baseline GVP-HMM systems using no complexity control, a consistent reduction in model complexity was also obtained by using BIC complexity controlled GVP-HMM systems. A series of contrasts for various GVP-HMM system configurations are shown in the 5th and 6th columns of Table I. As expected, when increasing the setting of ρ, the average size of the resulting complexity controlled GVP-HMM system decreases. For example, using a setting $\rho = 3$, the average number of polynomial coefficients of complexity controlled "mv" system in the 2nd line of Table I was reduced by 60% relative from 240 K in the baseline GVP-HMM system down to 95.53 K.

As discussed in Section I, a locally varying polynomial degree is preferred when the variability introduced by noise manifests itself on a dimension by dimension basis in the acoustic space. This is shown in Fig. 1 to 4 for the mean and variance polynomials of BIC complexity controlled GVP-HMM "mv" systems ($\rho = 1.0$) trained on each of the four noise types across different dimensions in feature space. The standard 39 dimensional Aurora 2 acoustic frontends were derived 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 cepstra of up to the 3rd order and the log energy, which

TABLE VI
WER PERFORMANCE OF MULTI-STYLE TRAINED BASELINE HMM SYSTEM, BASELINE GVP-HMM SYSTEMS WITH A UNIFORMLY ASSIGNED PARAMETER POLYNOMIAL DEGREE, AND COMPLEXITY CONTROLLED GVP-HMM SYSTEMS USING A LOCALLY VARYING POLYNOMIAL DEGREE ON SUBWAY NOISE DATA: BASELINE GVP-HMM SYSTEM WITHOUT COMPLEXITY CONTROL USED TO GENERATE SUFFICIENT STATISTICS

ComCtrl	System	0dB	5dB	10dB	15dB	20dB	Ave
×	mcond.base	21.25	7.55	3.78	2.36	2.27	7.44
×	mean	19.31	6.79	2.98	1.87	1.57	6.82
	mv	17.16	6.88	3.04	2.18	1.96	6.24
	tran2	20.17	6.57	3.29	2.06	1.66	6.75
	tran8	20.76	6.60	3.38	1.90	1.60	6.85
	mvt2	20.69	7.18	3.25	2.24	2.03	7.48
BIC ($\rho = 1$)	mean	18.27	6.79	2.64	1.81	1.63	6.23
	mv	16.03	6.48	2.70	1.90	1.90	5.80
	tran2	18.27	6.91	3.56	2.46	2.09	6.66
	tran8	18.42	6.66	3.53	2.06	1.93	6.52
	mvt2	17.22	6.66	2.89	2.09	1.84	6.14
BIC ($\rho = 2$)	mean	18.48	6.54	2.52	1.87	1.66	6.21
	mv	16.49	6.29	2.58	1.90	1.84	5.82
	tran2	18.27	6.91	3.56	2.46	2.09	6.66
	tran8	18.48	6.66	3.50	1.96	1.93	6.51
	mvt2	17.78	6.48	2.76	1.93	1.81	6.15
BIC ($\rho = 3$)	mean	18.27	6.54	2.64	1.96	1.75	6.23
	mv	15.97	6.05	2.58	1.96	1.96	**5.70**
	tran2	18.24	6.88	3.56	2.46	2.09	6.65
	tran8	18.27	6.63	3.53	1.96	2.00	6.48
	mvt2	17.10	6.39	2.82	1.96	2.00	6.05

TABLE VII
WER PERFORMANCE OF MULTI-STYLE TRAINED BASELINE HMM SYSTEM, BASELINE GVP-HMM SYSTEMS WITH A UNIFORMLY ASSIGNED PARAMETER POLYNOMIAL DEGREE, AND COMPLEXITY CONTROLLED GVP-HMM SYSTEMS USING A LOCALLY VARYING POLYNOMIAL DEGREE ON BABBLE NOISE DATA: BASELINE GVP-HMM SYSTEM WITHOUT COMPLEXITY CONTROL USED TO GENERATE SUFFICIENT STATISTICS

ComCtrl	System	0dB	5dB	10dB	15dB	20dB	Ave
×	mcond.base	30.47	12.09	6.53	4.59	4.08	11.55
×	mean	31.08	10.25	4.35	2.78	2.84	10.26
	mv	28.60	9.85	4.63	3.36	3.17	9.92
	tran2	32.95	12.24	5.80	3.45	3.17	11.52
	tran8	32.62	11.12	4.96	2.78	2.84	10.86
	mvt2	35.40	12.00	4.59	3.08	3.02	11.62
BIC ($\rho = 1$)	mean	30.23	9.34	3.69	2.63	2.81	9.74
	mv	28.84	9.07	3.75	2.78	3.33	**9.55**
	tran2	29.08	11.97	5.86	3.54	3.45	10.78
	tran8	28.26	10.61	4.53	2.75	2.75	9.78
	mvt2	31.95	9.82	4.26	2.81	3.17	10.40
BIC ($\rho = 2$)	mean	29.56	9.34	3.72	2.75	3.08	9.69
	mv	28.60	8.98	4.20	2.96	3.51	9.65
	tran2	29.08	11.97	5.86	3.54	3.45	10.78
	tran8	28.26	10.55	4.66	2.72	2.72	9.78
	mvt2	32.41	10.49	4.59	2.96	3.48	10.79
BIC ($\rho = 3$)	mean	28.93	9.34	3.99	2.93	2.99	9.64
	mv	27.90	9.16	4.14	3.02	3.57	9.56
	tran2	29.08	11.97	5.86	3.54	3.45	10.78
	tran8	28.02	10.49	4.69	2.78	2.78	9.75
	mvt2	32.95	10.91	4.69	2.87	3.51	10.99

TABLE VIII
WER PERFORMANCE OF MULTI-STYLE TRAINED BASELINE HMM SYSTEM, BASELINE GVP-HMM SYSTEMS WITH A UNIFORMLY ASSIGNED PARAMETER POLYNOMIAL DEGREE, AND COMPLEXITY CONTROLLED GVP-HMM SYSTEMS USING A LOCALLY VARYING POLYNOMIAL DEGREE ON CAR NOISE DATA: BASELINE GVP-HMM SYSTEM WITHOUT COMPLEXITY CONTROL USED TO GENERATE SUFFICIENT STATISTICS

ComCtrl	System	0dB	5dB	10dB	15dB	20dB	Ave
×	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
BIC ($\rho = 1$)	mean	23.34	9.25	4.32	2.95	2.41	8.45
	mv	18.53	8.32	4.14	2.98	2.53	7.30
	tran2	21.97	8.14	4.02	3.10	2.87	8.02
	tran8	20.71	8.14	3.96	3.25	2.47	7.71
	mvt2	19.00	7.87	3.99	3.07	2.50	7.29
BIC ($\rho = 2$)	mean	23.16	9.22	4.26	2.98	2.35	8.39
	mv	18.76	7.63	3.87	3.04	2.56	7.17
	tran2	22.00	8.14	4.02	3.10	2.87	8.03
	tran8	20.50	8.11	3.99	3.22	2.47	7.66
	mvt2	18.88	7.96	3.96	3.31	2.53	7.33
BIC ($\rho = 3$)	mean	23.05	8.80	4.23	2.92	2.47	8.29
	mv	18.56	7.72	3.84	2.92	2.62	**7.13**
	tran2	22.03	8.14	4.02	3.10	2.87	8.03
	tran8	20.56	8.14	4.02	3.19	2.44	7.67
	mvt2	18.46	7.66	3.87	3.25	2.62	7.17

TABLE IX
WER PERFORMANCE OF MULTI-STYLE TRAINED BASELINE HMM SYSTEM, BASELINE GVP-HMM SYSTEMS WITH A UNIFORMLY ASSIGNED PARAMETER POLYNOMIAL DEGREE, AND COMPLEXITY CONTROLLED GVP-HMM SYSTEMS USING A LOCALLY VARYING POLYNOMIAL DEGREE ON EXHIBITION NOISE DATA: BASELINE GVP-HMM SYSTEM WITHOUT COMPLEXITY CONTROL USED TO GENERATE SUFFICIENT STATISTICS

ComCtrl	System	0dB	5dB	10dB	15dB	20dB	Ave
×	mcond.base	23.46	10.27	4.91	3.09	2.72	8.89
×	mean	21.45	8.15	4.38	2.75	2.10	7.77
	mv	19.69	7.99	4.07	2.56	2.31	7.32
	tran2	22.99	9.87	4.60	2.84	2.47	8.55
	tran8	22.53	9.78	4.44	2.65	2.28	8.34
	mvt2	25.19	9.63	4.07	2.81	2.34	8.81
BIC ($\rho = 1$)	mean	21.05	8.42	4.32	2.62	2.07	7.70
	mv	19.23	7.59	4.07	2.50	2.65	7.21
	tran2	21.11	9.56	4.54	2.72	2.50	8.09
	tran8	21.36	9.29	4.29	2.59	2.22	7.95
	mvt2	18.61	7.84	4.01	2.16	2.19	6.96
BIC ($\rho = 2$)	mean	20.99	8.55	4.41	2.59	2.22	7.75
	mv	19.17	7.65	4.13	2.59	2.34	7.18
	tran2	21.11	9.56	4.54	2.72	2.50	8.09
	tran8	21.23	9.19	4.26	2.53	2.22	7.89
	mvt2	18.86	7.74	3.73	2.13	2.31	**6.95**
BIC ($\rho = 3$)	mean	20.49	8.67	4.41	2.78	2.31	7.73
	mv	19.04	7.62	4.04	2.59	2.25	7.11
	tran2	21.11	9.50	4.47	2.75	2.50	8.07
	tran8	21.33	9.29	4.23	2.53	2.22	7.92
	mvt2	19.88	7.93	3.86	2.19	2.41	7.25

contain more information of speech, tend to use more complex polynomial trajectories than higher order cepstra. In all four figures, Gaussian variance polynomials use consistently lower degrees than those of component mean vectors. As discussed in Section I, the underlying polynomial degree being used represents the approximated effect of different noise types on the actual acoustic realization. For example, complexity controlled GVP-HMM "mv" system trained on the car noise data of Fig. 3 favored higher order polynomials for both Gaussian means and variances compared with the comparable three other "mv" systems trained on other noise types. The locally varying polynomial degree over feature space dimensions for an example Gaussian component in this BIC ($\rho = 1.0$) optimized GVP-HMM "mv" system trained on the car noise data is also shown in Fig. 5.

As discussed in Section I, the use of uniformly assigned higher degree polynomials for GVP-HMMs in practice increases the interpolation cost during recognition. A detailed comparison

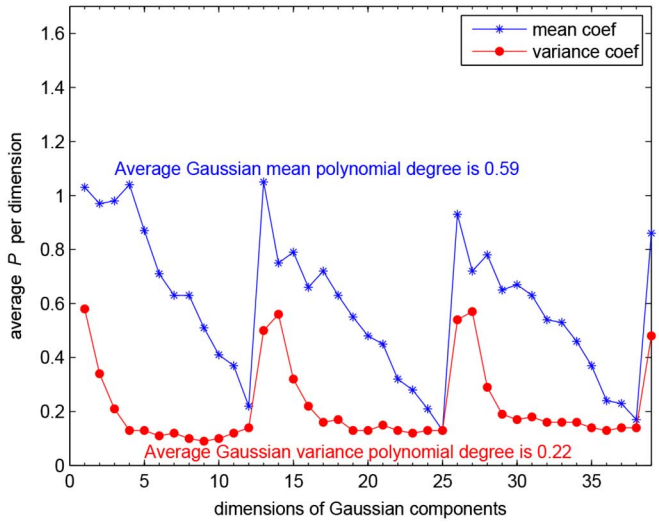

Fig. 1. Avg. polynomial degree P over feature dimensions in BIC ($\rho = 1.0$) optimized GVP-HMM "mv" system trained on subway noise data.

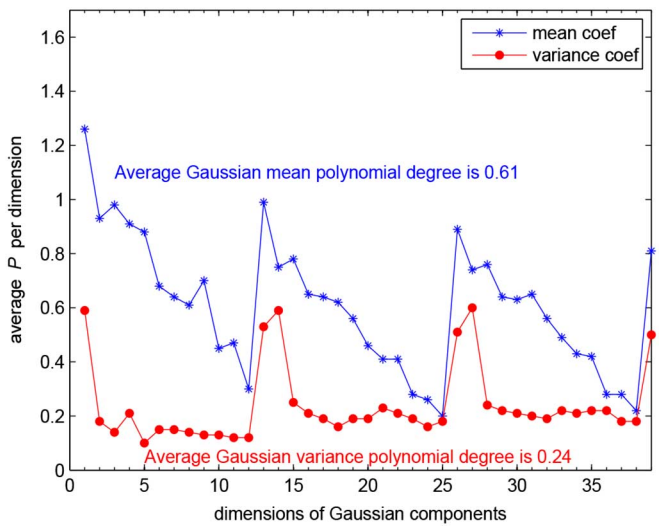

Fig. 2. Avg. polynomial degree P over feature dimensions in BIC ($\rho = 1.0$) optimized GVP-HMM "mv" system trained on babble noise data.

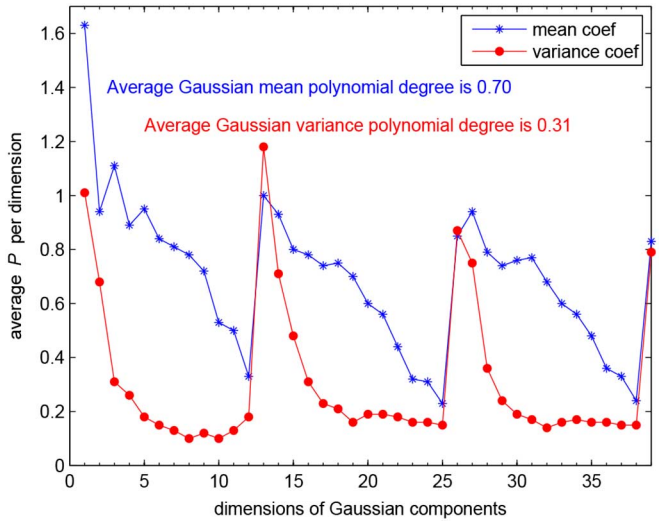

Fig. 3. Avg. polynomial degree P over feature dimensions in BIC ($\rho = 1.0$) optimized GVP-HMM "mv" system trained on car noise data.

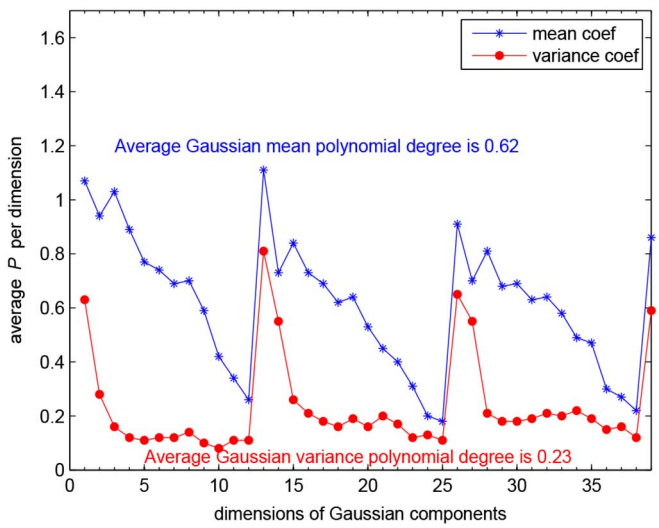

Fig. 4. Avg. polynomial degree P over feature dimensions in BIC ($\rho = 1.0$) optimized GVP-HMM "mv" system trained on exhibition noise data.

of total computational cost (decoding time and optionally interpolation cost for GVP-HMMs) incurred during recognition between the baseline HMM system and various GVP-HMM systems measured on the Aurora 2 test set A is shown in Fig. 6. The standard cost of using the baseline HMM "mcond.base" is taken as the reference (100% shown in the figure), to be contrasted with the relative costs of using various baseline and complexity controlled GVP-HMM systems. As expected, the use of complexity control reduced the decoding time by up to 40% relative for the GVP-HMM "mv" system in the figure. The more compact transform based "trans2" GVP-HMM systems with complexity control increased the total run time only by 6%-7% over the baseline HMM "mcond.base" system.

Performance on Test B of Mismatched Noise Types: In all the experiments presented so far, the performance of complexity controlled GVP-HMM systems were evaluated on test data of noise types well-matched to the training data. In order to further evaluate their generalization performance, complexity controlled GVP-HMM systems were then used on the mismatched noise test set B, which covers four unseen types different from those used in model training.

As discussed in Section I, the variability introduced by background noise on acoustic realization and the resulting spectral characteristics can vary significantly between different types of noisy data. The underlying effect on the production of speech signals is jointly determined by both the particular noise type and the SNR condition. One approach to acquire generalization to noisy data of both unknown noise types and SNR conditions is to construct the multi-style training set by including noisy data associated with multiple noise types and SNR conditions.

An alternative approach used in this paper is to exploit the similarity and commonality between training and test data noise types. An unsupervised maximum likelihood noise type detection procedure is first performed to find a training noise type as the closest approximation to the unknown test data noise

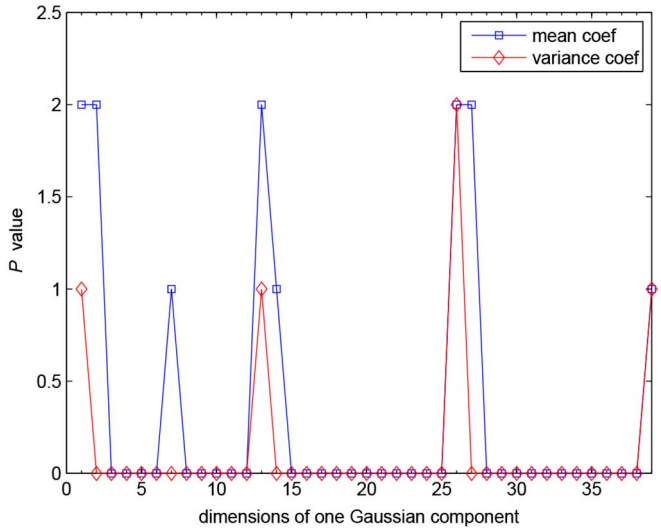

Fig. 5. Locally varying polynomial degree P over feature dimensions for an example Gaussian component in the BIC ($\rho = 1.0$) optimized GVP-HMM "mv" system trained on the car noise data.

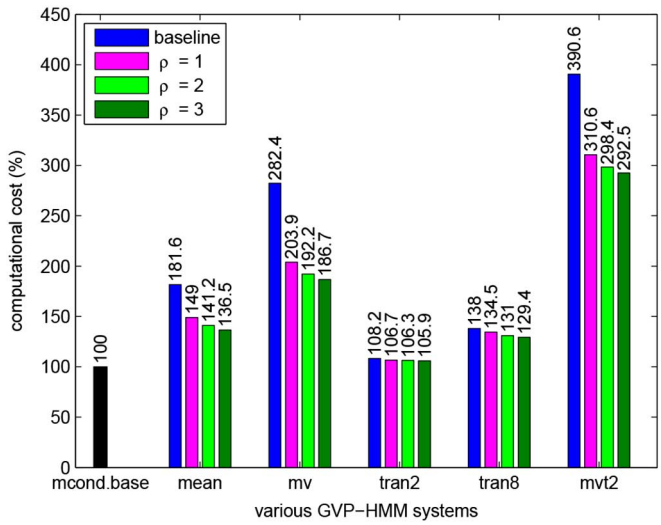

Fig. 6. Comparison of computational cost during recognition between the baseline HMM system and various GVP-HMM systems on Aurora 2 test set A. The standard cost of using the baseline HMM "mcond.base" is taken as the reference (100% shown in the figure), to be contrasted with the relative costs of using various baseline and complexity controlled GVP-HMM systems.

type. The associated baseline HMM system, GVP-HMM systems and complexity controlled GVP-HMM systems trained on such selected noise type are then used in recognition. A baseline HMM system multi-style constructed on all four types of noisy training data is first used to generate the initial recognition outputs for the test data. Four noise dependent HMM baseline systems produced by *maximum a-posteriori* (MAP) [15] adapting this baseline multi-style HMM system (noise type and SNR condition both vary) were then used to force-align the initial recognition outputs. The noise type of the particular noise dependent system giving the highest log-likelihood score was selected as the closest approximation to the unknown test data noise. In practice this approach was found to produce the same noise detection results as using the WER based manually derived ground truth noise labels, as are shown in Table X. Some

TABLE X
NOISE TYPE DETECTION: SELECT THE CLOSEST TRAINING NOISE TYPE FOR EACH UNSEEN NOISE TYPE IN AURORA 2 TEST SET B.

Test Data Noise Type	Noise Detection	
	Manual(WER)	Auto
Restaurant	Babble	Babble
Street	Car	Car
Airport	Babble	Babble
Station	Car	Car

TABLE XI
WER PERFORMANCE OF MULTI-STYLE TRAINED BASELINE HMM SYSTEM, BASELINE GVP-HMM SYSTEMS WITH A UNIFORMLY ASSIGNED PARAMETER POLYNOMIAL DEGREE, AND COMPLEXITY CONTROLLED GVP-HMM SYSTEMS USING A LOCALLY VARYING POLYNOMIAL DEGREE ON RESTAURANT NOISE DATA

ComCtrl	System	0dB	5dB	10dB	15dB	20dB	Ave
×	mcond.base	27.66	10.53	5.10	3.16	3.53	10.00
×	mean	26.50	8.81	3.50	1.75	1.14	8.34
	mv	27.82	10.78	4.21	1.75	1.38	9.19
	tran2	32.08	11.48	5.28	2.52	1.84	10.64
	tran8	30.55	10.53	4.21	1.96	1.63	9.78
	mvt2	34.54	11.70	4.70	1.81	1.29	10.81
BIC ($\rho = 1$)	mean	26.47	8.81	3.59	1.66	1.26	8.36
	mv	25.51	8.60	3.22	1.87	1.54	**8.15**
	tran2	28.09	11.02	4.73	2.43	1.75	9.60
	tran8	27.88	10.41	3.93	2.15	1.60	9.19
	mvt2	25.51	8.63	3.38	1.84	1.44	8.16
BIC ($\rho = 2$)	mean	26.01	9.27	3.84	1.84	1.41	8.47
	mv	25.36	8.84	3.72	1.78	1.54	8.25
	tran2	28.12	11.11	4.73	2.43	1.78	9.63
	tran8	27.82	10.19	3.90	2.12	1.60	9.13
	mvt2	25.39	9.15	3.75	1.90	1.47	8.33
BIC ($\rho = 3$)	mean	26.16	9.40	3.78	1.90	1.44	8.54
	mv	25.30	9.18	3.78	2.00	1.63	8.38
	tran2	28.25	11.08	4.76	2.43	1.78	9.66
	tran8	28.03	10.19	3.90	2.12	1.57	9.16
	mvt2	25.73	9.36	3.84	2.00	1.54	8.49

interesting observations can be made from Table X with respect to the noise selection. For instance, the closest training noise type for "restaurant" and "airport" in test set B is "babble", while for "street" and "station" noise data, "car" noise was selected.

Using the above noise detection method, the WER performance of baseline HMM systems, GVP-HMM systems with no complexity control and complexity controlled GVP-HMM systems are evaluated on test set B. These are shown from Table XI to 14. In common with the previous results presented on test A of matched noise types from Table II to V, a consistent WER reduction was also obtained by using the BIC complexity controlled GVP-HMM systems over their comparable baseline GVP-HMM systems with no complexity control. For example, for the airport noise data in Table XIII, the BIC ($\rho = 1$) complexity controlled "mv" GVP-HMM systems (8th line of Table XIII) outperformed the baseline HMM "mcond.base" system (1st line in Table XIII) by 2.59% absolute (26.5% relative). When a more complex modelling configuration was used for GVP-HMM systems of no complexity control, the over-fitting issue and the resulting poor generalization discussed in Section I can be clearly found on test data of mismatched noise types.

TABLE XII
WER PERFORMANCE OF MULTI-STYLE TRAINED BASELINE HMM SYSTEM, BASELINE GVP-HMM SYSTEMS WITH A UNIFORMLY ASSIGNED PARAMETER POLYNOMIAL DEGREE, AND COMPLEXITY CONTROLLED GVP-HMM SYSTEMS USING A LOCALLY VARYING POLYNOMIAL DEGREE ON STREET NOISE DATA

ComCtrl	System	0dB	5dB	10dB	15dB	20dB	Ave
×	mcond.base	28.26	11.88	5.83	3.54	3.51	10.60
×	mean	27.36	10.37	4.56	2.45	1.87	9.32
	mv	26.81	12.30	5.83	2.48	2.18	9.92
	tran2	29.20	11.67	5.44	3.23	2.36	10.38
	tran8	28.78	11.79	5.08	2.72	2.18	10.11
	mvt2	35.22	14.48	6.08	2.48	2.24	12.10
BIC ($\rho = 1$)	mean	26.15	10.04	4.35	2.42	1.87	**8.97**
	mv	27.18	11.15	4.59	2.18	1.87	9.39
	tran2	27.27	11.25	5.05	3.23	2.33	9.83
	tran8	26.33	11.12	4.66	2.84	2.12	9.41
	mvt2	27.72	11.15	4.50	2.24	1.90	9.50
BIC ($\rho = 2$)	mean	25.88	10.34	4.47	2.45	1.96	9.02
	mv	26.03	11.06	4.20	2.15	1.87	9.06
	tran2	27.27	11.25	5.05	3.23	2.33	9.83
	tran8	26.57	11.00	4.72	2.78	2.12	9.44
	mvt2	27.33	11.00	4.11	2.24	1.96	9.33
BIC ($\rho = 3$)	mean	26.27	10.40	4.59	2.51	1.96	9.15
	mv	27.00	11.00	4.41	2.39	1.90	9.34
	tran2	27.60	11.28	5.11	3.26	2.33	9.92
	tran8	26.57	11.00	4.66	2.81	2.12	9.43
	mvt2	27.66	11.28	4.35	2.51	2.00	9.56

TABLE XIII
WER PERFORMANCE OF MULTI-STYLE TRAINED BASELINE HMM SYSTEM, BASELINE GVP-HMM SYSTEMS WITH A UNIFORMLY ASSIGNED PARAMETER POLYNOMIAL DEGREE, AND COMPLEXITY CONTROLLED GVP-HMM SYSTEMS USING A LOCALLY VARYING POLYNOMIAL DEGREE ON AIRPORT NOISE DATA

ComCtrl	System	0dB	5dB	10dB	15dB	20dB	Ave
×	mcond.base	25.08	10.26	5.70	3.61	4.15	9.76
×	mean	22.73	8.59	3.55	1.97	1.40	7.65
	mv	23.80	9.33	4.09	1.58	1.37	8.03
	tran2	25.41	10.83	5.64	2.65	2.27	9.36
	tran8	24.60	9.84	4.32	2.33	1.70	8.56
	mvt2	29.11	9.99	4.21	1.79	1.49	9.32
BIC ($\rho = 1$)	mean	22.13	8.68	3.64	2.03	1.43	7.58
	mv	20.97	8.08	3.70	1.85	1.25	**7.17**
	tran2	23.50	10.62	4.98	2.68	2.30	8.82
	tran8	22.93	9.54	4.26	2.09	1.70	8.10
	mvt2	21.26	8.11	3.67	1.97	1.22	7.25
BIC ($\rho = 2$)	mean	22.19	8.62	3.76	1.97	1.52	7.61
	mv	21.32	8.74	3.73	2.24	1.25	7.46
	tran2	23.41	10.65	4.98	2.68	2.30	8.80
	tran8	22.93	9.54	4.29	2.06	1.70	8.10
	mvt2	21.53	8.65	3.85	2.21	1.22	7.49
BIC ($\rho = 3$)	mean	21.89	8.83	3.76	1.91	1.55	7.59
	mv	21.92	8.59	3.88	2.27	1.28	7.59
	tran2	23.41	10.68	4.98	2.68	2.27	8.80
	tran8	23.08	9.54	4.44	2.06	1.70	8.16
	mvt2	21.86	8.44	4.06	2.24	1.28	7.58

B. Experiments on Mandarin In-Car Task

A set of experiments similar to those for Aurora 2 presented from Table II to XIV were then conducted on the medium vocabulary Mandarin In-Car navigation command recognition task. The baseline HMM system was developed using 25 hours of clean training data. A multi-style training data set was then constructed by artificially corrupting the clean speech data with added car engine noise. Noise corrupted speech data generated under six sentence level SNR conditions: 0 dB, 4 dB, 8 dB, 12 dB, 16 dB and 20 dB, were used in training,

TABLE XIV
WER PERFORMANCE OF MULTI-STYLE TRAINED BASELINE HMM SYSTEM, BASELINE GVP-HMM SYSTEMS WITH A UNIFORMLY ASSIGNED PARAMETER POLYNOMIAL DEGREE, AND COMPLEXITY CONTROLLED GVP-HMM SYSTEMS USING A LOCALLY VARYING POLYNOMIAL DEGREE ON STATION NOISE DATA

ComCtrl	System	0dB	5dB	10dB	15dB	20dB	Ave
×	mcond.base	27.34	11.23	6.05	3.95	2.62	10.24
×	mean	25.79	9.69	3.98	2.31	1.36	8.63
	mv	26.29	11.63	5.00	2.38	1.17	9.29
	tran2	28.69	10.83	4.78	2.96	1.48	9.75
	tran8	27.80	10.15	4.26	2.78	1.20	9.24
	mvt2	32.03	13.42	5.68	2.44	1.20	10.95
BIC ($\rho = 1$)	mean	24.13	9.29	3.80	2.34	1.27	**8.17**
	mv	25.79	10.89	3.86	2.31	1.20	8.81
	tran2	25.21	10.09	4.41	2.84	1.51	8.81
	tran8	24.41	9.44	3.95	2.72	1.27	8.36
	mvt2	25.73	11.08	4.07	2.19	1.23	8.86
BIC ($\rho = 2$)	mean	24.59	9.50	3.70	2.34	1.27	8.28
	mv	25.08	10.80	3.98	2.28	1.27	8.68
	tran2	25.21	10.09	4.41	2.84	1.51	8.81
	tran8	24.34	9.72	3.95	2.65	1.23	8.38
	mvt2	24.96	11.29	4.17	2.31	1.27	8.80
BIC ($\rho = 3$)	mean	24.41	9.56	4.01	2.62	1.23	8.37
	mv	24.87	11.05	4.13	2.47	1.30	8.76
	tran2	25.15	10.15	4.41	2.87	1.51	8.82
	tran8	24.28	9.66	3.95	2.65	1.27	8.36
	mvt2	24.71	11.39	4.47	2.53	1.30	8.88

TABLE XV
DESCRIPTION OF VARIOUS GVP-HMMS ON THE MANDARIN IN-CAR TASK: PARAMETER POLYNOMIAL TYPES AND THE NUMBER OF POLYNOMIAL COEFFICIENTS

System	Poly. Types			#PolyCoef (In-Car)	
	mean	var	tran	base	BIC($\rho = 1/2/3$)
mean	√	×	×	3.66M	1.97M/1.77M/1.67M
mv	√	√	×	7.32M	3.79M/3.47M/3.3M
tran2	×	×	√	10.8K	7.22K/7.18K/7.18K
tran256	×	×	√	1.39M	0.98M/0.97M/0.97M
mvt2	√	√	√	7.32M	3.79M/3.47M/3.3M

while a corrupted 5 hour test set consists of five sentence level SNR conditions: 2 dB, 6 dB, 10 dB, 14 dB, and 18 dB, was used for character error rate (CER) evaluation. The baseline HMM acoustic models were ML trained using HTK [42] 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. A set of GVP-HMM configurations similar to those described in Table I were used for the In-Car data. These are shown in Table XV. Consistent with the model compression obtained on the Aurora 2 data previously shown in Table I, a significant model size reduction was obtained on the In-Car data using the proposed complexity control scheme. These are shown in 5th and 6th columns of Table XV.

The WER performance of the baseline multi-style and GVP-HMM systems, are shown in Table XVI. Consistent with the trend found on the test A and B of Aurora 2 task in sections 5.1.2 and 5.1.3, every BIC complexity controlled GVP-HMM system in Table XVII outperformed its comparable GVP-HMM baseline in Table XVI. For example, the complexity controlled mean based GVP-HMM system, “mean”, (1th, 6th and 11th

TABLE XVI
CER PERFORMANCE OF MULTI-STYLE TRAINED BASELINE HMM SYSTEM AND BASELINE GVP-HMM SYSTEMS WITH A UNIFORMLY ASSIGNED PARAMETER POLYNOMIAL DEGREE ON MANDARIN IN-CAR TASK

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 XVII
CER PERFORMANCE OF BIC COMPLEXITY CONTROLLED GVP-HMM SYSTEMS USING A LOCALLY VARYING POLYNOMIAL DEGREE ON MANDARIN 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**

lines in Table XVII) gave an average CER reduction of 5.12% absolute (21% relative) over the baseline "mean" GVP-HMM system (2nd line in Table XVI), and a 54% relative reduction in model complexity, as is shown in the 1st line in Table XV. The two highlighted 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 XVI by 7.12% absolute (28% relative). They gave the lowest average error rate among all GVP-HMM systems in Table XVII, and a 52%-55% relative reduction in the number of polynomial coefficients against their respective baselines, as are shown the 2nd and bottom line, 5th and 6th columns in Table XV. The performance comparison of this BIC GVP-HMM "mvt2" system against its comparable baseline GVP-HMM system of no complexity control, and the HMM baseline "mcond.base" system is intuitively shown in Fig. 7. The BIC complexity controlled "mvt2" system in Fig. 7 consistently outperformed its comparable GVP-HMM baseline system by 7%-15% relative, and the baseline HMM "mcond.base" system by 11%-40% relative under various SNR conditions.

VI. 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%

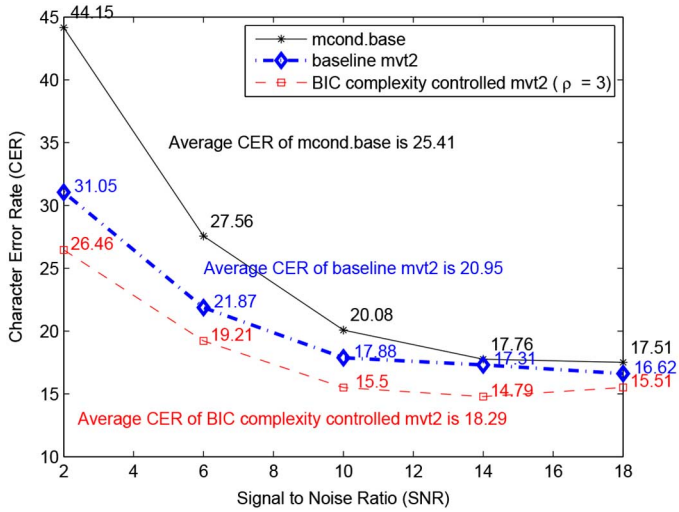

Fig. 7. CER performance of "mvt2" system on Mandarin In-Car Task.

relative obtained on Aurora 2 and a medium vocabulary Mandarin speech recognition task suggest the proposed method may be useful for speech recognition. These results show the applicability of complexity controlled GVP-HMMs to multiple languages and tasks, and their strong generalization performance to noisy speech data of both well-matched and mismatched noise types and SNR conditions. Future research will focus on discriminative training and modelling multiple sources of acoustic variability.

To date automatic model complexity control remains a challenging statistical modelling problem for many practical applications. This is particularly true for speech and language processing systems. As human language is so varied and complex, many aspects of it, for example, the effect imposed by environment noise on the acoustic model parameters as considered in this work, are often investigated using sophisticated computational models. This results in a very large number of possible system configurations to consider. The desired complexity control techniques suitable for these systems therefore should provide a good trade-off between performance and computational feasibility. The EM lower bound based Bayesian complexity control approach proposed for GVP-HMMs in this paper is inspired by a range of precursor techniques derived for conventional HMMs in early research to determine, for example, the optimal phonetic context dependent state tying [41], the number of Gaussian components and feature space dimensions [29], [30], [31]. In order to improve the robustness of the proposed technique, an iterative complexity control approach that enforces a maximum structural mutation constraint at each model selection iteration will also be investigated in future research. By restricting the number of possible systems to share the same set of sufficient statistics, this approach is expected to tighten the underlying lower bound being used.

REFERENCES

[1] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP'96*, Philadelphia, PA, USA, 1996, pp. 1137–1140.
[2] J. A. Arrowood and M. A. Clements, "Using observation uncertainty in HMM decoding," in *Proc. ICSLP*, 2002, pp. 1561–1564.

[3] A. Azevedo-Filho and R. Shachter, , R. Mantaras and D. Poole, Eds., "Laplace's method approximations for probabilistic inference in belief networks with continuous variables," in *Uncertainty in Artificial Intelligence*. San Francisco, CA, USA: Morgan Kauffman, 1994, CiteSeerX: 10.1.1.91.2064.

[4] A. R. Barron, J. J. Rissanen, and B. Yu, "The minimum description length principle in coding and model ing," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2743–2760, Oct. 1998.

[5] 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.

[6] N. Cheng, X. Liu, and L. Wang, "Generalized variable parameter HMMs for noise robust speech recognition," in *Proc. ISCA Interspeech'11*, Florence, Italy, 2011, pp. 482–484.

[7] N. Cheng, X. Liu, and L. Wang, "A flexible framework for HMM based noise robust speech recognition using generalized parametric space polynomial regression," *Sci. China, Inf. Sci.*, vol. 54, no. 2, pp. 2481–2491, 2011.

[8] W. Chou and W. Reichl, "Decision tree state tying based on penalized Bayesian information criterion," in *Proc. IEEE ICASSP'99*, Phoenix, AZ, USA, 1999, vol. 1.

[9] X. Cui and Y. Gong, "A study of variable-parameter gaussian mixture hidden markov modeling for noisy speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1366–1376, May 2007.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. 39, no. 1, pp. 1–39, 1977.

[11] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 412–421, May 2005.

[12] J. Droppo, A. Acero, and L. Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition," in *Proc. IEEE ICASSP2002*, Orlando, 2002, pp. 57–60.

[13] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-regression hidden markov model," in *Proc. IEEE ICASSP'01*, Salt Lake City, UT, USA, 2001, vol. 1, pp. 513–516.

[14] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, no. 2, pp. 75–98, 1998.

[15] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[16] H. G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR'00*, Paris, France, Sep. 2000, pp. 181–188.

[17] O. Kalinli, M. Seltzer, J. Droppo, and A. Acero, "Noise adaptive training for robust automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 1889–1901, Nov. 2010.

[18] D. Y. Kim, C. K. Un, and N. S. Kim, "Speech recognition in noisy environments using first-order vector Taylor series," *Speech Commun.*, vol. 24, no. 1, pp. 39–49, 1998.

[19] D. K. Kim and M. J. F. Gales, "Noisy constrained maximum-likelihood linear regression for noise-robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 315–325, Feb. 2011.

[20] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Comput. Speech Lang.*, vol. 9, pp. 171–186, 1995.

[21] H. Liao and M. J. F. Gales, "Adaptive training with joint uncertainty decoding for robust recognition of noisy data," *Proc. ICASSP*, pp. 389–392, 2007.

[22] H. Liao and M. J. F. Gales, "Issues with uncertainty decoding for noise robust speech recognition," *Speech Commun.*, vol. 50, pp. 265–277, 2008.

[23] Y. Li, X. Liu, and L. Wang, "Structured modeling based on generalized variable parameter HMMs and speaker adaptation," in *Proc. IEEE ISCSLP'12*, Hong Kong, China, 2012, pp. 136–140.

[24] Y. Li, X. Liu, and L. Wang, "Feature space generalized variable parameter HMMs for noise robust speech recognition," in *Proc. ISCA Interspeech'13*, Lyon, France, 2013.

[25] S. Lin, B. Chen, and Y.-M. Yeh, "Exploring the use of speech features and their corresponding distribution characteristics for robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 1, pp. 84–94, Jan. 2009.

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

[27] R. Lippmann, E. Martin, and D. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. IEEE ICASSP'87*, Dallas, TX, USA, 1987, pp. 705–708.

[28] X. Liu, M. J. F. Gales, and P. C. Woodland, "Automatic complexity control for HLDA systems," in *Proc. IEEE ICASSP'03*, Hong Kong, China, 2003, vol. 1, pp. 132–135.

[29] X. Liu, M. J. F. Gales, S. Thomas, and U. S. Virgin Islands, "Automatic model complexity control using marginalized discriminative growth functions," in *Proc. IEEE ASRU'03*, 2003, pp. 37–42.

[30] X. Liu and M. J. F. Gales, "Model complexity control and compression using discriminative growth functions," in *Proc. IEEE ICASSP'04*, Montreal, QC, Canada, 2004, vol. 1, pp. 797–800.

[31] X. Liu and M. J. F. Gales, "Automatic model complexity control using marginalized discriminative growth functions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1414–1424, May 2007.

[32] T. T. Kristjansson and B. J. Frey, "Accounting for uncertainty in observations: A new paradigm for robust speech recognition," in *Proc. IEEE ICASSP'02*, Orlando, FL, USA, 2002, pp. 61–64.

[33] R. Martin, "An efficient algorithm to estimate the instantaneous SNR speech signals," in *Proc. Eurospeech'93*, Berlin, Germany, 1993, pp. 1093–1096.

[34] R. M. Neal, Probabilistic inference using Markov chain Monte Carlo methods Dept. of Comput. Sci., Univ. of Toronto, Tech. Rep., CGT-TR-93-1, 1993.

[35] C. Runge, "Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten," *Zeitschrift für Mathematik und Physik*, vol. 46, pp. 224–243, 1901.

[36] D. Pierce and A. Gunawardana, "Aurora 2.0 speech recognition in noise: Update 2. complex backend definition for aurora 2.0," 2002 [Online]. Available: http://icslp2002.colorado.edu/special_sessions/aurora

[37] G. Schwartz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Feb. 1978.

[38] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. IEEE ICASSP'13*, Vancouver, BC, USA, 2013.

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

[40] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 4, pp. 365–381, Jul. 2004.

[41] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modeling," in *Proc. ARPA Human Lang. Technol. Workshop*, 1994, pp. 307–312, Morgan Kaufman.

[42] S. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, "The HTK Book Version 3.4.1," , 2009.

[43] D. Yu, L. Deng, Y. Gong, and A. Acero, "Discriminative training of variable-parameter HMMs for noise robust speech recognition," in *Proc. ISCA Interspeech'08*, Brisbane, Australia, 2008, pp. 285–288.

[44] D. Yu, L. Deng, Y. Gong, and A. Acero, "Parameter clustering and sharing in variable-parameter HMMs for noise robust speech recognition," in *Proc. ISCA Interspeech'08*, Brisbane, Australia, 2008, pp. 1253–1256.

[45] D. Yu, L. Deng, Y. Gong, and A. Acero, "A novel framework and training algorithm for variable-parameter hidden Markov models," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1348–1360, Sep. 2009.

Rongfeng Su was born in 1983. He received the B.S. degree from East China University of Science and Technology, Shanghai, China, in 2007, and the M.S. degree from the Institute for Information Systems, Technische Universität Braunschweig, Germany, in 2011. Currently, he is a Ph.D. student at the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences.

His main areas of research interest are robust speech recognition and large vocabulary continuous speech recognition.

Xunying Liu (M'06) was born in 1978. He received the Ph.D. degree in speech recognition in 2006 and MPhil degree in computer speech and language processing in 2001 both from University of Cambridge, prior to a bachelor's degree from Shanghai Jiao Tong University in 2000.

He is currently a Senior Research Associate at the Machine Intelligence Laboratory of the Cambridge University Engineering Department. He is the lead researcher on the EPSRC funded Natural Speech Technology and the DARPA funded Broad Operational Language Translation Programs at Cambridge. He was the recipient of best paper award at ISCA Interspeech2010.

His current research interests include large vocabulary continuous speech recognition, language modelling, noise robust speech recognition, speech and language processing. Dr. Liu Xunying is a member of IEEE and ISCA.

Lan Wang (M'05) received her bachelor degree in electrical engineering in 1993 from Beijing Institute of Technology, the her M.S. degree in information science in 1996 from Peking University, China, and the Ph.D. degree in speech signal processing in 2006 from Cambridge University Engineering Department (CUED), UK.

She worked in the Center of Information Science, Peking University as a lecture from 1996 to 2001. From 2005 to 2006, she was a Research Associate in the Machine Intelligence Lab of CUED. In 2007, she joined Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences (CAS), where she is the Professor and Vice Director of the Lab of Ambient Intelligence and Multimodal Systems. Her research interests are large-vocabulary continuous speech recognition, speech visualization and speech-centric human–machine interaction.