

Canonical State Models for Automatic Speech Recognition

M.J.F. Gales and K. Yu

Cambridge University Engineering Department, Trumpington Street, Cambridge, UK

mjfg@eng.cam.ac.uk, ky219@eng.cam.ac.uk

Abstract

Current speech recognition systems are often based on HMMs with state-clustered Gaussian Mixture Models (GMMs) to represent the context dependent output distributions. Though highly successful, the standard form of model does not exploit any relationships between the states, they each have separate model parameters. This paper describes a general class of model where the context-dependent state parameters are a transformed version of one, or more, canonical states. A number of published models sit within this framework, including, semi-continuous HMMs, subspace GMMs and the HMM error model. A set of preliminary experiments illustrating some of this model's properties using CMLLR transformations from the canonical state to the context dependent state are described.

Index Terms: acoustic modelling, adaptive training, Gaussian mixture models.

1. Introduction

Hidden Markov models (HMMs) with state-specific Gaussian Mixture model (GMM) are the most popular acoustic model in speech recognition. An important issue for training these models is to ensure that there is sufficient data to robustly estimate the context-dependent phone models (usually triphones). The standard approach for this is to use decision tree tying to determine the set of context-dependent states [1]. Given this decision tree, GMMs are then trained for each state. This has two, related, limitations. First, there must be sufficient data to robustly estimate the GMM parameters. Thus diagonal covariance matrices are often used. The second issue is that the number of context dependent states, the depth of the decision tree, is limited due to the need to robustly estimate the GMM parameters. Though good performance has been achieved with these models, any relationship between the context dependent states is not exploited. This motivates the use of a different form of model that attempts to take advantage of this relationship, the *canonical state model*.

The canonical state model (CSM) uses a transformation of some context-independent, or global, canonical state to represent a particular context dependent state in the system. It is possible to specify transformations of all the state parameters: component priors, component means and covariance matrices. Examples of canonical state models in this general framework include semi-continuous HMMs [2], the HMM error model (HEM) [3] and the subspace GMM (SGMM) [4]. It is important to emphasise that the final model that results from these transformations will still sit within the class of standard HMMs. The difference to standard systems is that a *soft* tying scheme [3], rather than the standard tying approach, is being used.

One of the advantages of this form of model is that it is very flexible. Currently published systems clearly demonstrate this. For the HMM error model [3], a powerful transforma-

tion (a mixture of Constrained Maximum Likelihood Linear Regression (CMLLR) transforms) is used to map a simple canonical state model to a context-dependent state. Conversely in the subspace GMM a highly complicated canonical model with a large number of components and full covariance matrices is mapped using a simple transformation to the context-dependent state [4]. This additional flexibility allows the model structure to be tuned to the particular task and available training data. As transformations of the canonical state are being used to generate the context dependent models, the parameters are estimated in a similar fashion to speaker adaptive training approaches [5, 6, 7]. However now rather than the transforms modifying a *canonical speaker* into each of the target speakers, a *canonical state* is transformed into each of the context dependent states.

An interesting aspect of these models, that differs from standard adaptation, is that it may be important to adapt the component priors as well as the means and covariance matrices. The simplest approach to doing this is to use an ML estimate of the priors. This requires there to be sufficient data associated with a context to allow robust estimates to be made [2]. Alternatively a subspace representation can be used [4]. The importance of the component transformations depends on the configuration, in particular the number of components in the canonical state. For example for the HEM no component prior transformation was used, whereas for the subspace GMM it was found to be very important [4].

The best configuration of the canonical state model will also depend on the nature of the task being examined. For the HMM error model a standard recognition task was examined [3]. For this case relatively complex transforms, a mixture of diagonal CMLLR transforms, was used to map from the canonical state to the context dependent state. Conversely, the subspace GMM was examined for multilingual recognition [8] where the data associated with a particular language may be limited. Thus the transformation of the canonical state to the context dependent state is simple, an interpolation of parameters. The form and number of canonical states will also depend on the configuration. If there is, for example, a natural mapping of context-dependent states to context-independent states (i.e. a consistent phone set is used for all training and test), then multiple context-independent canonical states can be estimated [3]. However in the case of multilingual systems a single global canonical model is required [8].

There are links between canonical state models and compact representations of acoustic models. For these compact representations a set of canonical distributions or vectors are used to represent individual components in the context-dependent system. Examples of this form of model are subspace distribution models [9] and structured precision and mean models [10]. However these forms do not maintain the concept of some underlying state, rather they are used for compact component representations, so are not considered further in this paper.

The next section discusses the general canonical state model framework along with some examples of how it can be used, including the semi-continuous HMM, subspace GMM and the use of MLLR [11]-based systems. A more in-depth discussion of CMLLR-based systems, which includes the HMM error model, is then given. Various forms of this CMLLR-based model are then evaluated and conclusions drawn.

2. Canonical State Models

Canonical state models comprise two sets of model parameters: one or more sets of canonical states, $\mathbf{s}_g = \{\dots, \{c_g^{(m)}, \mu_g^{(m)}, \Sigma_g^{(m)}\}, \dots\}$, these canonical states are typically at the global level or at the context-independent phone level; and a set of transform parameters, $\mathcal{T} = \{\dots, \{w_s^{(n)}, \theta_s^{(n)}\}, \dots\}$, from the canonical state, \mathbf{s}_g , to a context dependent state, \mathbf{s} , where $\theta_s^{(n)}$ are the set of transform parameters for component n and $w_s^{(n)}$ the associated prior.

Both sets of model parameters need to be estimated: the context-dependent transforms; and the canonical state model. The parameter estimation follows the same general concepts as speaker adaptive training [5, 6, 7]. However now rather than the transforms modifying a canonical speaker into each of the target speakers, a canonical state is transformed into each of the context dependent states. The general training process is split into two stages, first update the transform parameters given the current canonical model parameters. Second the canonical model parameters are updated given the current context-dependent transformations.

The next section describes the general expression for the canonical state model. This is followed by some example forms of canonical state model including the Subspace GMM.

2.1. General Model

In this section only a single canonical state is considered. The extension to a set of canonical states, where there is a unique mapping from context-dependent state to canonical state, is trivial. The likelihood for context-dependent state \mathbf{s} using canonical state \mathbf{s}_g is given by¹

$$p(\mathbf{x}|\mathbf{s}) = \sum_{n=1}^N w_s^{(n)} \left(\sum_{m \in \mathbf{s}_g} c_s^{(mn)} \mathcal{N}(\mathbf{x}; \mu_s^{(mn)}, \Sigma_s^{(mn)}) \right) \quad (1)$$

where N is the number of transform components

$$c_s^{(mn)} = \mathcal{F}_c(\mathbf{s}_g, m; \theta_s^{(n)}) \quad (2)$$

$$\mu_s^{(mn)} = \mathcal{F}_\mu(\mathbf{s}_g, m; \theta_s^{(n)}) \quad (3)$$

$$\Sigma_s^{(mn)} = \mathcal{F}_\Sigma(\mathbf{s}_g, m; \theta_s^{(n)}) \quad (4)$$

and the canonical state is a standard Gaussian mixture model whose parameters have the form

$$p(\mathbf{x}|\mathbf{s}_g) = \sum_{m \in \mathbf{s}_g} c_g^{(m)} \mathcal{N}(\mathbf{x}; \mu_g^{(m)}, \Sigma_g^{(m)}) \quad (5)$$

This form of mixture of linear transforms was originally used for speaker adaptation [12] and later for the HMM Error Model [3], an example of a canonical state model. Various

¹This form of model is also related to factorial-style models, see [3] for more details and the general concept of transformation streams. The canonical model in (5) is modified if interpolation transforms such as CAT are used.

forms of transform are possible for example, MLLR [11], CMLLR [6] and cluster adaptive training (CAT) [7]. For this general model there is also the transformation of the component priors, for example in [2, 4].

2.2. Semi-Continuous HMM

The simplest form of canonical state model is the semi-continuous HMM [2]. Here only a single transform component is used. The context-dependent state distribution is given by,

$$p(\mathbf{x}|\mathbf{s}) = \sum_{m \in \mathbf{s}_g} c_s^{(m)} \mathcal{N}(\mathbf{x}; \mu_g^{(m)}, \Sigma_g^{(m)}) \quad (6)$$

where the context-dependent component priors are estimated using

$$\mathcal{F}_c(\mathbf{s}_g, m; \theta_s) = c_s^{(m)} = \frac{\sum_t \gamma_{st}^{(m)}}{\sum_{\tilde{m} \in \mathbf{s}_g} \sum_t \gamma_{st}^{(\tilde{m})}} \quad (7)$$

and $\gamma_{st}^{(m)}$ is the posterior probability of canonical component m , and context-dependent state \mathbf{s} generating the observation at time t . Writing the transformations of the means and covariance matrices in the full canonical model form yields

$$\mathcal{F}_\mu(\mathbf{s}_g, m; \theta_s) = \mu_g^{(m)} \quad (8)$$

$$\mathcal{F}_\Sigma(\mathbf{s}_g, m; \theta_s) = \Sigma_g^{(m)} \quad (9)$$

This form of prior transformation will be referred to as an *ML-transformation*. It yields the maximum likelihood estimate of the model parameters for the context-dependent state parameters. If applied to the component priors, means and covariance matrices this will give exactly the same system as standard GMM training (assuming that all states have the same number of components).

2.3. Subspace GMM

For this description of the Subspace GMM (SGMM) the speaker adaptation aspects of the model are ignored. This speaker adaptation aspect may be viewed as being related to CAT acting at the canonical state level. The form of SGMM is parameterised by a single vector of weights for each context dependent state, \mathbf{s} and transform component, n , $\lambda_s^{(n)}$

$$\mathcal{F}_c(\mathbf{s}_g, m; \theta_s^{(n)}) = \frac{\exp(\mathbf{v}_g^{(m)\top} \lambda_s^{(n)})}{\sum_{\tilde{m} \in \mathbf{s}_g} \exp(\mathbf{v}_g^{(\tilde{m})\top} \lambda_s^{(n)})} \quad (10)$$

$$\mathcal{F}_\mu(\mathbf{s}_g, m; \theta_s^{(n)}) = [\mu_g^{(m1)} \dots \mu_g^{(mP)}] \lambda_s^{(n)} \quad (11)$$

$$\mathcal{F}_\Sigma(\mathbf{s}_g, m; \theta_s^{(n)}) = \Sigma_g^{(m)} \quad (12)$$

where $\mathbf{v}_g^{(m)}$ is the P -dimensional subspace prior vector for component m . The transforms associated with the means and covariance matrices have the same form as CAT².

One of the elegant aspects of the SGMM is that efficient likelihood calculation can be performed even when full covariance matrices are used [4]. For example in the large vocabulary configuration in [13], the canonical model had 750 components each with a full covariance matrix, and a total of 200K context-dependent transformations. The number of context-dependent transformations per context-dependent state was made a function of the state occupancy.

²The original form of CAT [7] used diagonal covariance matrices, whereas the SGMM form published in [4] was derived for full covariance matrices. Additionally for simplicity of notation the ‘‘bias cluster’’ for both the means and components is not considered.

2.4. MLLR-Based Systems

An alternative to using a CAT form of transformation is to consider an MLLR transform [11, 5], or mixture of transforms, from the canonical state to the context dependent state. Thus in the general CSM notation, the context dependent state means and covariance matrices of component m with transformation component n are given by

$$\mathcal{F}_\mu(\mathbf{s}_g, m; \boldsymbol{\theta}_s^{(n)}) = \mathbf{A}_s^{(n)} \boldsymbol{\mu}_g^{(m)} + \mathbf{b}_s^{(n)} \quad (13)$$

$$\mathcal{F}_\Sigma(\mathbf{s}_g, m; \boldsymbol{\theta}_s^{(n)}) = \boldsymbol{\Sigma}_g^{(m)} \quad (14)$$

An issue to consider when using MLLR transforms is that if a simple canonical model is used (in the sense of small numbers of components) then situations can occur when the canonical state model yields no benefit over the standard GMM-based scheme. For example with a single full MLLR transformation is under-specified if only a 16 component canonical state model is used. Thus training the transform will yield *exactly* the same system as standard training for the mean, it will be an ML-transformation. However if large canonical state models, or mixtures of simple (for example diagonal) transforms, are used then there may be benefit in this form of model.

3. CMLLR-Based Systems

This paper describes an extension to the version of the HEM described in [3]. The transformation from the canonical state to the context dependent state is a mixture of CMLLR transforms. The original form of the HEM can be expressed in the same form as the general canonical state model form as

$$\mathcal{F}_c(\mathbf{s}_g, m; \boldsymbol{\theta}_s^{(n)}) = c_g^{(m)} \quad (15)$$

$$\mathcal{F}_\mu(\mathbf{s}_g, m; \boldsymbol{\theta}_s^{(n)}) = \mathbf{A}_s^{(n-1)} (\boldsymbol{\mu}_g^{(m)} - \mathbf{b}_s^{(n)}) \quad (16)$$

$$\mathcal{F}_\Sigma(\mathbf{s}_g, m; \boldsymbol{\theta}_s^{(n)}) = \mathbf{A}_s^{(n-1)} \boldsymbol{\Sigma}_g^{(m)} \mathbf{A}_s^{(n-1)\top} \quad (17)$$

Expressing this in the more usual form for the CMLLR transform, the likelihood for context dependent state \mathbf{s} component m and transformation component n can be expressed as

$$p(\mathbf{x}|\mathbf{s}, n, m) = |\mathbf{A}_s^{(n)}| \mathcal{N}(\mathbf{A}_s^{(n)} \mathbf{x} + \mathbf{b}_s^{(n)}; \boldsymbol{\mu}_g^{(m)}, \boldsymbol{\Sigma}_g^{(m)}) \quad (18)$$

The HEM likelihood can then be written as

$$p(\mathbf{x}|\mathbf{s}) = \sum_{n=1}^N w_s^{(n)} \sum_{m \in \mathbf{s}_g} c_g^{(m)} p(\mathbf{x}|\mathbf{s}, n, m) \quad (19)$$

One of the differences between the canonical state model examined in this work and the HEM is that the component priors are also made context dependent. In this work a simple form of transformation similar to the semi-continuous system is used. The likelihood calculation has the form³

$$p(\mathbf{x}|\mathbf{s}) = \sum_{n=1}^N w_s^{(n)} \sum_{m \in \mathbf{s}_g} c_s^{(m)} p(\mathbf{x}|\mathbf{s}, n, m) \quad (20)$$

where $p(\mathbf{x}|\mathbf{s}, n, m)$ is given by (18). In this case of the training of the component prior is

$$\mathcal{F}_c(\mathbf{s}_g, m; \boldsymbol{\theta}_s^{(n)}) = c_s^{(m)} = \frac{\sum_{n=1}^N \sum_t \gamma_{st}^{(mn)}}{\sum_{n=1}^N \sum_{\tilde{m} \in \mathbf{s}_g} \sum_t \gamma_{st}^{(\tilde{m}n)}} \quad (21)$$

³An alternative would be to make the priors transform component and canonical state component. In this case $c_s^{(mn)}$ would be used. Note this would naturally subsume the initial $w_s^{(n)}$ if estimated in an ML-transform fashion.

where $\gamma_{st}^{(nm)}$ is an extension of the posterior $\gamma_{st}^{(m)}$ to include the transform component n . Note for this form of component transformation there is no dependence on the component prior transformation on the transformation component.

The estimation of the canonical state means can be expressed as

$$\boldsymbol{\mu}_g^{(m)} = \frac{\sum_s \sum_{n=1}^N \sum_t \gamma_{st}^{(mn)} (\mathbf{A}_s^{(n)} \mathbf{x}_t + \mathbf{b}_s^{(n)})}{\sum_s \sum_{n=1}^N \sum_t \gamma_{st}^{(mn)}} \quad (22)$$

A similar expression can be derived for the canonical state variances. The statistics to update the context-transformations can be expressed as, for each dimension i ,

$$\mathbf{G}_{si}^{(n)} = \sum_{m \in \mathbf{s}_g} \frac{1}{\sigma_{gi}^{(m)2}} \sum_t \gamma_{st}^{(mn)} \zeta_t \zeta_t^\top \quad (23)$$

$$\mathbf{k}_{si}^{(n)} = \sum_{m \in \mathbf{s}_g} \frac{\mu_{gi}^{(m)}}{\sigma_{gi}^{(m)2}} \sum_t \gamma_{st}^{(mn)} \zeta_t^\top \quad (24)$$

where $\zeta_t^\top = [1 \quad \mathbf{x}_t^\top]$, as well as the transform occupancy count. The estimation of the canonical state transform parameters, $\mathbf{A}_s^{(n)}$ and $\mathbf{b}_s^{(n)}$, can be estimated using the standard CMLLR update formulae [6]. For more details of the derivation of these parameters see [3].

The likelihood calculation for these CMLLR-based schemes does not have the same efficient form as the SGMM. For each context-dependent state there are effectively $N \times M$ (M is the number of canonical state components) Gaussian to calculate⁴. However using the ML-transformation for the component priors allows pruning of the component priors. This was not examined in this work.

4. Preliminary Experiments

The performance of the CMLLR-based canonical state model was evaluated on a Broadcast News transcription task. In contrast to state-of-the-art systems, only basic ML acoustic models were built. There was no decorrelating transformation associated with front-end, and no discriminative training. Furthermore only unadapted experiments were performed.

The acoustic model training configuration was the same as the *bntr-375h* acoustic models described in [14]. These models were trained on 375 hours of acoustic data - the Hub4 training data (144 hours) and TDT2 data (231 hours). A PLP front-end with delta and delta-delta parameters, yielding a 39 dimensional feature vector, was used. Approximately 7000 distinct state-clustered triphone states were used with 16 Gaussian components per-state. Only wide-band gender-independent acoustic models were built in this work. The same narrow-band results, based on the baseline GMM system, were used for all systems. The language model used in these experiments was the *RT03* language model described in [14]. This was trained on approximately a billion words of data, and a 59K vocabulary was used. The *eval03* test data, comprising 6 shows and 3 hours of data, was used to evaluate the performance of the systems⁵ See [14] for more details of the acoustic and language model training sources and additional information.

⁴The calculation can be made more efficient, though not to the same extent as the SGMM.

⁵The baseline number for this system, quoted in [14] is 17.2%. For this work the HDecode decoder was used. The parameters were not extensively tuned, which resulted in a baseline performance of 17.7%.

For the CSM, the number of components in the canonical state was increased using the standard *mix-up* approach in HTK. To increase the number of transform components, the same approach as that described in [3] was used. A separate canonical state was used for each context independent state.

Table 1: *Viterbi decoding results.*

System	#comp	#tran	struct	WER
GMM	1	—		26.2
	16			17.7
HEM	1	16	diag	17.7
	3			17.4
CSM	16	1	diag	20.8
		8		17.9
		12		17.4
		15		17.1
	16	1	blk	18.2
	32			17.7
	48			17.4
	48	2	blk	16.9
	16	1	full	16.5

Table 1 shows the performance of the baseline, standard, GMM system and a range of configurations for the CSM. The initial contrasts used configurations similar to the HEM [3], these are labelled HEM. Given the form of model used, 1 component 16 diagonal CMLLR transforms is the same as the baseline 16 component system, with the same number of free parameters. As expected from [3], using multiple components in the canonical state improves performance. For the more general CSMs, marked CSM, a number of configurations were evaluated. As well as WERs it is interesting to compare the number of parameters in the systems. Using a mixture of 15 diagonal transforms for each context dependent state with 16 components in the canonical state has about the same number of model parameters as the standard 16 component GMM and the 2 block-diagonal transform component with 48 component-GMM.

Table 2: *Confusion network decoding and combination results.*

System	#comp	#tran	struct	WER
GMM	16	—		17.6
CSM ₁	32	1	diag	17.6
CSM ₂	16	15	diag	16.9
CSM ₃	48	2	blk	16.8
GMM \oplus CSM ₁	CNC			16.5
CSM ₂ \oplus CSM ₃	CNC			16.0

One of the interesting aspects about this form of CSM is that it can achieve similar performance to the standard system but constructed in a very different fashion with a similar number of model parameters (but not necessarily decoding cost). These models are thus candidates for system combination. Table 2 shows the combination of the baseline GMM system and a similarly performing CSM system (CSM₁), as well as two similarly performing CSMs. For all systems small gains were obtained from confusion network (CN) decoding over the baseline Viterbi decoding results in table 1. Gains of 1.1% and 0.8% absolute were obtained from CN Combination (CNC). Though these experiments used only basic, unadapted and ML-trained, they illustrate another potential benefit of this style of model.

5. Conclusions

This paper has described a general class of model, the canonical state model. The basic concept behind the model is that the parameters of a canonical state are transformed to become the parameters of a context-dependent state. Thus the model bears some resemblance speaker adaptive training. However instead of adapting a model to a speaker it is adapted to context dependent state. The same range of transforms as speaker adaptation can be applied, including CMLLR, MLLR, and CAT. An interesting modification to the standard adaptation framework, is that it can be advantageous to also adapt the component priors. Existing models within this framework include semi-continuous HMMs, the HMM error model and the subspace GMM.

The form of model examined in this paper is based on context-dependent CMLLR transforms. Various configurations were evaluated on a broadcast news English system. A number of systems yield comparable performance to the standard GMM-based system and each other. They are found to be complementary to one another. This paper has only described preliminary experiments on a particular set-up. There are a large number of possible systems and configurations that could be used. Additionally as the models are usually trained on supervised data, discriminative training of all model parameters, including the transformations, can be applied.

6. References

- [1] S. Young, J. Odell, and P. Woodland, "Tree-Based State Tying for High Accuracy Acoustic Modelling," in *Proc Human Language Technology Workshop*, Morgan Kaufman Publishers Inc, 1994.
- [2] M. Hwang and X. Huang, "Shared-Distribution Hidden Markov Models for Speech Recognition," *IEEE Trans Speech and Audio Processing*, 1993.
- [3] M. Gales, "Transformation Streams and the HMM Error Model," *Computer Speech and Language*, 2002.
- [4] D. Povey *et al*, "Subspace Gaussian Mixture Models for Speech Recognition," in *ICASSP'10*, Dallas, 2010.
- [5] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A Compact Model for Speaker Adaptive Training," in *ICSLP'96*, Philadelphia, 1996.
- [6] M. Gales, "Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.
- [7] —, "Cluster Adaptive Training of Hidden Markov Models," *IEEE Trans Speech and Audio Processing*, 2000.
- [8] L. Burget *et al*, "Multilingual Acoustic Modeling for Speech Recognition Based on Subspace Gaussian Mixture Models," in *ICASSP'10*, Dallas, 2010.
- [9] B. Mak and E. Bocchieri, "Training of context independent subspace distribution clustering hidden Markov models," in *Proceedings ICSLP*, 1998, pp. 2959–2962.
- [10] S. Axelrod, R. Gopinath, and P. Olsen, "Modeling with a Subspace Constraint on Inverse Covariance Matrices," in *ICSLP 2002*, Denver, CO, 2002.
- [11] C. Leggetter and P. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.
- [12] V. Diakouloukas and V. Digalakis, "Maximum Likelihood Stochastic Transformation Adaptation of Hidden Markov Models," *IEEE Trans. On Speech and Audio Processing*, 1999.
- [13] G. Saon *et al*, "The IBM 2008 GALE Arabic Speech Transcription System," in *ICASSP'10*, Dallas, 2010.
- [14] M. Gales, D. Kim, P. Woodland, D. Mrva, R. Sinha, and S. Tranter, "Progress in the CU-HTK broadcast news transcription system," *IEEE Trans Speech and Audio Processing*, September 2006.