Restructuring Exponential Family Mixture Models

Pierre L. Dognin, John R. Hershey, Vaibhava Goel, Peder A. Olsen

IBM T.J. Watson Research Center

{pdognin, jrhershe, vgoel, pederao}@us.ibm.com

Abstract

Variational KL (varKL) divergence minimization was previously applied to restructuring *acoustic models* (AMs) using Gaussian mixture models by reducing their size while preserving their accuracy. In this paper, we derive a related varKL for *exponential family mixture models* (EMMs) and test its accuracy using the weighted local maximum likelihood agglomerative clustering technique. Minimizing varKL between a reference and a restructured AM led previously to the *variational expectation maximization* (varEM) algorithm; which we extend to EMMs. We present results on a clustering task using AMs trained on 50 hrs of *Broadcast News* (BN). EMMs are trained on fMMI-PLP features combined with frame level phone posterior probabilities given by the recently introduced sparse representation phone identification process. As we reduce model size, we test the word error rate using the standard BN test set and compare with baseline models of the same size, trained directly from data.

Index Terms: KL divergence, variational approximation, variational expectation-maximization, exponential family distributions, acoustic model clustering.

1. Introduction

A problem commonly encountered in probabilistic modeling is to approximate a model using another model with a different structure. Model restructuring techniques can change the number of components (or parameters), share parameters, or simply modify some other constraints, so a model can better match the needs of an application.

When restructuring models, it is necessary to preserve similarity between *reference* and *restructured* model. Minimizing the *Kullback-Leibler* (KL) divergence [1] between these two models is equivalent to maximizing the likelihood of the restructured model under data drawn from the reference model. Unfortunately, this is intractable for general mixtures of continuous random variables, without resorting to expensive Monte Carlo approximation techniques. However, it is possible to derive a variational approximation to the KL divergence [2] as well as a *variational expectation-maximization* (varEM) algorithm [3] that will update the parameters of a model to better match a reference model. These model restructuring methods were previously applied to reducing the size of *Gaussian mixture models* (GMMs) used in speech recognition. This paper applies restructuring to the broader class of *exponential family mixture models* (EMMs). A greedy clustering algorithm based on these variational methods is used. It provides clustered models of any size, as presented in [4]. For other approaches, based on minimizing the mean–squared error between the two density functions, see [5], or based on compression using dimension–wise tied Gaussians optimized using symmetric KL divergences, see [6]. For a discussion of the properties of EMM representation for GMMs, see [7].

Results, expressed as *word error rates* (WERs), are presented on models built from *perceptual linear prediction* (PLP) features, transformed using *feature space maximum mutual information* (fMMI), as well as models combining fMMI-PLP features and a set of phone-based posterior probabilities known as *sparse representation phone identification features* (SPIF) [8]. This paper expands previous work in two distinct ways. First, it validates restructuring techniques for models built on discriminative features (fMMI-PLP features). Second, it extends restructuring techniques to *acoustic models* (AMs) based on exponential family distributions.

2. Exponential Families

An exponential family is a class of distributions with the form

$$f(\boldsymbol{x}|\boldsymbol{\lambda}) \stackrel{\text{def}}{=} \frac{\mathrm{e}^{\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{x})}}{Z(\boldsymbol{\lambda})} \tag{1}$$

$$= \mathrm{e}^{\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{x})-\log Z(\boldsymbol{\lambda})}, \tag{2}$$

where $\boldsymbol{x}$ is a base observation in some domain $\mathcal{A}$. The *features* are generated by the function $\boldsymbol{\psi}(\boldsymbol{x}) : \mathcal{A} \rightarrow \mathbb{R}^D$, which characterizes the family of distributions, and $\boldsymbol{\lambda} \in \mathbb{R}^D$ is the parameter selecting a specific distribution within that family. $Z(\boldsymbol{\lambda})$ is the normalizing constant defined as

$$Z(\boldsymbol{\lambda}) = \int \mathrm{e}^{\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}. \tag{3}$$

$Z(\boldsymbol{\lambda})$ has the interesting and useful property that

$$\frac{\partial \log Z(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \int f(\boldsymbol{x})\boldsymbol{\psi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \mathrm{E}_f[\boldsymbol{\psi}(\boldsymbol{x})]\,. \tag{4}$$

For the purpose of restructuring, we refer to $f(\boldsymbol{x}|\boldsymbol{\lambda})$ as the reference model and $g(\boldsymbol{x}|\boldsymbol{\theta})$ as a restructured model within the same family. There are many different exponential families. In this paper we focus on multivariate normal and exponential distributions.

Multivariate Normal Distribution: A multivariate normal (or Gaussian) distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, is defined as

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2}\mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}. \tag{5}$$

When $\boldsymbol{\Sigma}$ is a full covariance matrix, $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ can be written as an exponential family with $\boldsymbol{\psi}(\boldsymbol{x})$ and $\boldsymbol{\lambda}$ given by

$$\boldsymbol{\psi}^{\mathrm{F}}(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{x} \\ -\frac{1}{2}\operatorname{vec}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \end{bmatrix},\ \boldsymbol{\lambda}^{\mathrm{F}} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ \operatorname{vec}\boldsymbol{\Sigma}^{-1} \end{bmatrix}, \tag{6}$$

Copyright © 2010 ISCA

where vec A rearranges the elements of A into a column vector. The normalizer has an analytical form given by

$$\log Z\left(\boldsymbol{\lambda}^{\mathrm{F}}\right) = \frac{1}{2}\left[\log|2\pi\boldsymbol{\Sigma}| + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]. \tag{7}$$

For normal distributions with diagonal covariance, (6) becomes

$$\boldsymbol{\psi}^{\mathrm{D}}(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{x} \\ -\frac{1}{2}\operatorname{diag}\left(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right) \end{bmatrix}, \boldsymbol{\lambda}^{\mathrm{D}} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ \operatorname{diag}\left(\boldsymbol{\Sigma}^{-1}\right) \end{bmatrix}. \tag{8}$$

Using (4), the expected value is

$$\mathrm{E}_{\mathcal{N}}\left[\boldsymbol{\psi}^{\mathrm{D}}(\boldsymbol{x})\right] = \begin{bmatrix} \boldsymbol{\mu} \\ -\frac{1}{2}\operatorname{diag}\left(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\right) \end{bmatrix}. \tag{9}$$

Exponential Distribution: The classic exponential distribution, with non-negative scalar $x \in \mathbb{R}^{+}$, and parameter $\lambda \in \mathbb{R}^{+}$, $\mathcal{E}(x|\alpha) = \alpha\exp(-\alpha x)$, is an exponential family with

$$\psi^{\mathrm{E}}(x) = x, \quad \lambda^{\mathrm{E}} = -\alpha, \quad \log Z\left(\lambda^{\mathrm{E}}\right) = -\log\alpha. \tag{10}$$

Using (4), the expected value is $\mathrm{E}_{\mathcal{E}}\left[\psi^{\mathrm{E}}(x)\right] = -1/\alpha$. We can generalize to multidimensional $\boldsymbol{x}$ by having λ^{E} be a vector, in which case all operations become element-wise.

Combined Distributions: Exponential families can be combined together to form new families by concatenating their parameters and features. We define a *combination exponential family* $f(\boldsymbol{z}|\boldsymbol{\lambda}^{\mathrm{C}})$ using $\boldsymbol{z} = (\boldsymbol{x}, y)$, where $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a diagonal covariance Gaussian, and $y \sim \mathcal{E}(y|\alpha)$, with

$$\boldsymbol{\psi}^{\mathrm{C}}(\boldsymbol{z}) = \begin{bmatrix} \boldsymbol{\psi}^{\mathrm{D}}(\boldsymbol{x}) \\ \psi^{\mathrm{E}}(y) \end{bmatrix}, \quad \lambda^{\mathrm{C}} = \begin{bmatrix} \boldsymbol{\lambda}^{\mathrm{D}} \\ \lambda^{\mathrm{E}} \end{bmatrix}. \tag{11}$$

The normalizer is given by

$$\log Z\left(\boldsymbol{\lambda}^{\mathrm{C}}\right) = \log Z\left(\boldsymbol{\lambda}^{\mathrm{D}}\right) + \log Z\left(\lambda^{\mathrm{E}}\right), \tag{12}$$

and the expected value is

$$\mathrm{E}_{f}\left[\boldsymbol{\psi}^{\mathrm{C}}(\boldsymbol{z})\right] = \begin{bmatrix} \mathrm{E}_{\mathcal{N}}\left[\boldsymbol{\psi}^{\mathrm{D}}(\boldsymbol{x})\right] \\ \mathrm{E}_{\mathcal{E}}\left[\psi^{\mathrm{E}}(y)\right] \end{bmatrix}. \tag{13}$$

We use this combination exponential family to model combined fMMI-PLP and SPIF features in the rest of the paper.

3. Exponential Family Mixture Models

In probabilistic modeling, we often resort to use mixture models to approximate complex distributions. An exponential family mixture model $f(\boldsymbol{x})$ is a mixture of exponential family distributions defined as

$$f(\boldsymbol{x}) = \sum_{a}\pi_a f_a(\boldsymbol{x}|\boldsymbol{\lambda}_a) = \sum_{a}\pi_a\frac{\mathrm{e}^{\boldsymbol{\lambda}_a^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{x})}}{Z(\boldsymbol{\lambda}_a)}, \tag{14}$$

where a indexes components of f, π_a is the prior probability, and $f_a(\boldsymbol{x}|\boldsymbol{\lambda}_a)$ is an exponential family *probability density function* (pdf). $\boldsymbol{\psi}(\boldsymbol{x})$ is identical for each component a, so all components are in the same family.

4. KL Divergence

The KL divergence [1] is a commonly used measure of dissimilarity between two pdfs $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$,

$$D_{\mathrm{KL}}(f\|g) \stackrel{\mathrm{def}}{=} \int f(\boldsymbol{x})\log\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} \tag{15}$$

$$= L(f\|f) - L(f\|g), \tag{16}$$

where $L(f\|g)$ is defined as the expected log likelihood of g under f. For f and g members of the exponential family distributions defined in (2), $L(f\|g)$ becomes

$$L(f\|g) = \int f(\boldsymbol{x})\log g(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{17}$$

$$= \int f(\boldsymbol{x})\left[\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{x}) - \log Z(\boldsymbol{\theta})\right]\mathrm{d}\boldsymbol{x} \tag{18}$$

$$= \boldsymbol{\theta}^{\mathrm{T}}\mathrm{E}_{f}[\boldsymbol{\psi}(\boldsymbol{x})] - \log Z(\boldsymbol{\theta}). \tag{19}$$

Similarly, $L(f\|f)$ is given by

$$L(f\|f) = \boldsymbol{\lambda}^{\mathrm{T}}\mathrm{E}_{f}[\boldsymbol{\psi}(\boldsymbol{x})] - \log Z(\boldsymbol{\lambda}), \tag{20}$$

and $D_{\mathrm{KL}}(f\|g)$ can be expressed as

$$D_{\mathrm{KL}}(f\|g) = (\boldsymbol{\lambda} - \boldsymbol{\theta})^{\mathrm{T}}\mathrm{E}_{f}[\boldsymbol{\psi}(\boldsymbol{x})] + \log\frac{Z(\boldsymbol{\theta})}{Z(\boldsymbol{\lambda})}. \tag{21}$$

For exponential family distributions f and g, $D_{\mathrm{KL}}(f\|g)$ has an analytic solution *only* if $\mathrm{E}_{f}[\boldsymbol{\psi}(\boldsymbol{x})]$, $\log Z(\boldsymbol{\theta})$, and $\log Z(\boldsymbol{\lambda})$ do. When no closed–form expressions exist, sampling techniques are usually used [9].

4.1. Generalized KL Divergence

The generalized KL divergence in the Bregman divergence family was proposed in [10] to extend the KL divergence to weighted densities $\alpha f(\boldsymbol{x})$ and $\beta g(\boldsymbol{x})$. The generalized KL divergence is given as

$$D_{\mathrm{KL}}(\alpha f\|\beta g) = \int\alpha f(\boldsymbol{x})\log\frac{\alpha f(\boldsymbol{x})}{\beta g(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} + \int\beta g(\boldsymbol{x}) - \alpha f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \tag{22}$$

$$= \alpha D_{\mathrm{KL}}(f\|g) + \alpha\log\frac{\alpha}{\beta} + \beta - \alpha. \tag{23}$$

The corresponding generalized expected log likelihood is

$$L(\alpha f\|\beta g) = \alpha L(f\|g) + \alpha\log\beta - \beta. \tag{24}$$

5. Variational KL Divergence

For f and g mixture models with weighted individual components $\pi_a f_a(\boldsymbol{x})$ and $\omega_b g_b(\boldsymbol{x})$, computing $D_{\mathrm{KL}}(f\|g)$ becomes intractable. Indeed, the expression for $L(f\|g)$ becomes

$$L\big(f\|g\big) = \sum_{a}\pi_a\int f_a(\boldsymbol{x})\log\sum_{b}\omega_b g_b(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \tag{25}$$

where the integral $\int f_a\log\sum_b\omega_b g_b$ has no closed–form solution. As a consequence, $D_{\mathrm{KL}}(f\|g)$ is not known in general for mixture models like GMMs and EMMs.

One solution presented in [3] provides a variational approximation to $D_{\mathrm{KL}}(f\|g)$. This is done by first providing variational approximations to $L(f\|f)$ and $L(f\|g)$ and then using (16). In order to define a variational approximation to (25),

variational parameters $\phi_{b|a}$ are introduced as a measure of the affinity between the Gaussian component f_a of f and component g_b of g. The variational parameters must satisfy the constraints

$$\phi_{b|a} \geq 0 \quad \text{and} \quad \sum_b \phi_{b|a} = 1. \tag{26}$$

Using Jensen's inequality, a lower bound is obtained for (25),

$$L\big(f\|g\big) \geq \sum_a \pi_a \sum_b \phi_{b|a} \left(\log \frac{\omega_b}{\phi_{b|a}} + L(f_a\|g_b) \right) \tag{27}$$

$$\stackrel{\text{def}}{=} \mathbb{L}_\phi\big(f\|g\big). \tag{28}$$

The lower bound on $L\big(f\|g\big)$, given by the variational approximation $\mathbb{L}_\phi\big(f\|g\big)$, can be maximized w.r.t. ϕ and the best bound is given by

$$\widehat{\phi}_{b|a} = \frac{\omega_b \, \mathrm{e}^{L(f_a\|g_b)}}{\sum_{b'} \omega_{b'} \, \mathrm{e}^{L(f_a\|g_{b'})}}. \tag{29}$$

By substituting $\widehat{\phi}_{b|a}$ from (29) into (27), the following expression for $\mathbb{L}_{\widehat{\phi}}\big(f\|g\big)$ is obtained:

$$\mathbb{L}_{\widehat{\phi}}\big(f\|g\big) = \sum_a \pi_a \log \left(\sum_b \omega_b \, \mathrm{e}^{L(f_a\|g_b)} \right). \tag{30}$$

$\mathbb{L}_{\widehat{\phi}}\big(f\|g\big)$ is the *best* variational approximation of the expected log likelihood $L\big(f\|g\big)$ and is referred to as *variational likelihood*. Similarly, the variational likelihood $\mathbb{L}_{\widehat{\varphi}}\big(f\|f\big)$, which maximizes a lower bound on $L\big(f\|f\big)$, is

$$\mathbb{L}_{\widehat{\varphi}}\big(f\|f\big) = \sum_a \pi_a \log \left(\sum_{a'} \pi_{a'} \, \mathrm{e}^{L(f_a\|f_{a'})} \right). \tag{31}$$

The variational KL divergence $\mathbb{D}_{\mathrm{KL}}(f\|g)$ is obtained directly from (30) and (31) since $\mathbb{D}_{\mathrm{KL}}(f\|g) = \mathbb{L}_{\widehat{\varphi}}\big(f\|f\big) - \mathbb{L}_{\widehat{\phi}}\big(f\|g\big)$,

$$\mathbb{D}_{\mathrm{KL}}(f\|g) = \sum_a \pi_a \log \left(\frac{\sum_{a'} \pi_{a'} \, \mathrm{e}^{-D_{\mathrm{KL}}(f_a\|f_{a'})}}{\sum_b \omega_b \, \mathrm{e}^{-D_{\mathrm{KL}}(f_a\|g_b)}} \right), \tag{32}$$

where $\mathbb{D}_{\mathrm{KL}}(f\|g)$ is based on the KL divergences between all individual components of f and g. The variational likelihood and KL divergence generalize to weighted mixture models $\alpha f(\boldsymbol{x})$ and $\beta g(\boldsymbol{x})$ in exactly the same way as the likelihood and KL divergence given in (24) and (23).

6. Variational Expectation-Maximization

In model restructuring, the variational KL divergence $\mathbb{D}_{\mathrm{KL}}(f\|g)$ can be minimized by updating the parameters of the restructured model g (with a given model structure) to match the reference model f. Since the variational KL divergence $\mathbb{D}_{\mathrm{KL}}(f\|g)$ gives an approximation to $D_{\mathrm{KL}}(f\|g)$, we can minimize $\mathbb{D}_{\mathrm{KL}}(f\|g)$ w.r.t. the parameters of g. Each g_b component of g has parameters $\{\omega_b, \boldsymbol{\theta}_b\}$. It is sufficient to maximize $\mathbb{L}_\phi(f\|g)$, as $\mathbb{L}_\psi(f\|f)$ is constant in g. Although (30) is not easily maximized w.r.t. the parameters of g, $\mathbb{L}_\phi(f\|g)$ in (27) can be maximized leading to an *expectation-maximization* (EM) algorithm. This leads to a *variational* expectation-maximization (varEM) algorithm where we first maximize $\mathbb{L}_\phi(f\|g)$ w.r.t ϕ. With ϕ fixed, we then maximize $\mathbb{L}_\phi(f\|g)$ w.r.t the parameters of g. Previously, we found that the best lower bound on $L(f\|g)$ is $\mathbb{L}_{\widehat{\phi}}\big(f\|g\big)$ given by $\widehat{\phi}_{b|a}$ in (29). This is the *expectation* (E) step. For a fixed $\phi_{b|a} = \widehat{\phi}_{b|a}$, it is now possible to find the parameters $\{\omega_b, \boldsymbol{\theta}_b\}$ of g that maximize $\mathbb{L}_\phi\big(f\|g\big)$. This leads to the following equation

$$\mathrm{E}_{g_b}[\boldsymbol{\psi}(\boldsymbol{x})] = \frac{\sum_a \pi_a \phi_{b|a} \mathrm{E}_{f_a}[\boldsymbol{\psi}(\boldsymbol{x})]}{\sum_{a'} \pi_{a'} \phi_{b|a'}}. \tag{33}$$

where the expected value, for our combination exponential family, is given by (13). The *maximization* (M) step is then:

$$\omega_b^\star = \sum_a \pi_a \phi_{b|a}, \tag{34}$$

$$\boldsymbol{\mu}_b^\star = \frac{\sum_a \pi_a \phi_{b|a} \boldsymbol{\mu}_a}{\sum_{a'} \pi_{a'} \phi_{b|a'}}, \tag{35}$$

$$\boldsymbol{\Sigma}_b^\star = \frac{\sum_a \pi_a \phi_{b|a} \left[\boldsymbol{\Sigma}_a + (\boldsymbol{\mu}_a - \boldsymbol{\mu}_b^\star)(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b^\star)^{\mathrm{T}} \right]}{\sum_{a'} \pi_{a'} \phi_{b|a'}}, \tag{36}$$

$$\frac{1}{\alpha_b^\star} = \frac{\sum_a \pi_a \phi_{b|a} \alpha_a^{-1}}{\sum_{a'} \pi_{a'} \phi_{b|a'}}. \tag{37}$$

The algorithm alternates between the E–step and M–step, increasing the variational likelihood in each step.

7. Weighted Local Maximum Likelihood

To determine the model structure for g, we perform an agglomerative clustering using *weighted local maximum likelihood* (wLML) proposed in [3]. This is a measure of the loss in expected log likelihood due to the merge of components. It has been successfully used in model clustering in [3, 4] for GMMs and it is extended to EMMs in this section.

Let us consider merging two components $\pi_i f_i$ and $\pi_j f_j$ of the EMM f defined in (14). We define $g = \mathrm{merge}(\pi_i f_i, \pi_j f_j) = \exp(\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\psi}(\boldsymbol{x}) - \log Z(\boldsymbol{\theta}))$, with weight ω. The wLML for components f_i and f_j is defined as

$$\mathrm{wLML}_{i,j} = (\pi_i + \pi_j) \mathbb{D}_{\mathrm{KL}}(f_i + f_j \| g). \tag{38}$$

We find the parameters $\{\omega, \boldsymbol{\theta}\}$ associated to g that maximize the generalized expected log likelihood $L(\pi_i f_i + \pi_j f_j \| \omega g)$. Clearly, (23) is minimized w.r.t β when $\beta = \alpha$, which gives $\omega = \pi_i + \pi_j$. To find $\boldsymbol{\theta}$, we use

$$\begin{aligned} L(\pi_i f_i + \pi_j f_j \| \omega g) = \boldsymbol{\theta}^{\mathrm{T}} \left(\pi_i \mathrm{E}_{f_i}[\boldsymbol{\psi}(\boldsymbol{x})] + \pi_j \mathrm{E}_{f_j}[\boldsymbol{\psi}(\boldsymbol{x})] \right) \\ + (\pi_i + \pi_j) \left[\log Z(\boldsymbol{\theta}) + \log \omega \right]. \end{aligned}$$

Setting $\partial L(\pi_i f_i + \pi_j f_j \| \omega g) / \partial \boldsymbol{\theta}$ to zero, and using the fact that $\partial \log Z(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} = \mathrm{E}_g[\boldsymbol{\psi}(\boldsymbol{x})]$ yields

$$\mathrm{E}_g[\boldsymbol{\psi}(\boldsymbol{x})] = \overline{\pi}_i \mathrm{E}_{f_i}[\boldsymbol{\psi}(\boldsymbol{x})] + \overline{\pi}_j \mathrm{E}_{f_j}[\boldsymbol{\psi}(\boldsymbol{x})], \tag{39}$$

where $\overline{\pi}_i = \pi_i / (\pi_i + \pi_j)$ and $\overline{\pi}_j = \pi_j / (\pi_i + \pi_j)$. When f_i is in our combination exponential family, from (13) we have

$$\mathrm{E}_{f_i}[\boldsymbol{\psi}(\boldsymbol{x})] = \begin{bmatrix} \boldsymbol{\mu}_i \\ -\frac{1}{2} \operatorname{diag}\left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathrm{T}} \right) \\ -1/\alpha_i \end{bmatrix}, \tag{40}$$

and similarly for f_j and g. Substituting into (39) gives $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha\}$ for g:

$$\boldsymbol{\mu} = \overline{\pi}_i \boldsymbol{\mu}_i + \overline{\pi}_j \boldsymbol{\mu}_j, \tag{41}$$

$$\boldsymbol{\Sigma} = \overline{\pi}_i (\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathrm{T}}) + \overline{\pi}_j (\boldsymbol{\Sigma}_j + \boldsymbol{\mu}_j \boldsymbol{\mu}_j^{\mathrm{T}}) - \boldsymbol{\mu} \boldsymbol{\mu}^{\mathrm{T}}, \tag{42}$$

$$\alpha^{-1} = \overline{\pi}_i \alpha_i^{-1} + \overline{\pi}_j \alpha_j^{-1}. \tag{43}$$

For diagonal covariance, only $\operatorname{diag}(\boldsymbol{\Sigma})$ is of interest.

WER (%) vs. Model Size (K)							
Models	10K	20K	30K	40K	60K	80K	100K
Baseline	23.0	21.5	21.3	21.2	20.9	20.5	20.4
KL	23.9	22.5	21.5	20.9	20.7	20.5	–
wLML	23.5	21.9	21.3	21.0	20.8	20.6	–
+logP	20.4	20.0	19.7	19.3	19.3	19.0	19.2
KL	20.7	20.2	19.6	19.4	19.1	19.1	–
wLML	20.6	19.9	19.6	19.5	19.3	19.2	–
wLML*	20.6	19.9	19.6	19.4	19.3	19.2	–

Table 1: WERs for models trained from data based on GMM (Baseline) and EMM with SPIF feature (+logP). Reference models (100K) are clustered down using KL, wLML, and wLML with model-based assignment (wLML*).

8. Experiments

The variational methods discussed in this paper are applied to restructuring EMMs by reducing their size using a greedy clustering algorithm in an approach similar to [4]. We present results on a *Broadcast News* (BN) LVCSR task. The training set comprises 50 hours of randomly selected shows from the '96 and '97 English BN speech corpora (LDC97S44, LDC98S71). The EARS Dev-04f set (dev04f) testing set is a collection of 3 hours of audio from 6 shows collected in November '03. Acoustic features are based on 13-dimensional PLP features with speaker-based mean, variance, and vocal tract length normalization. Nine such PLP frames are concatenated and projected to a 40-dimensional space using LDA. Speaker adaptive training is performed on these LDA features with one *feature-space maximum likelihood linear regression* (fMLLR) transform per speaker. A fMMI transform is estimated and baseline GMM models are built in this final fMMI-PLP feature space.

EMM models are built from SPIF and fMMI-PLP features using the combination exponential family defined in (11). SPIFs are phone-based posterior probabilities [8]. We simply use the logarithm of these SPIF posteriors as features (logP features), and model their distribution with (10). AMs are based on 44 phones, each one modeled as three-state, left-to-right hidden Markov models with no skip states. Context dependency trees provide 2206 *context dependent* (CD) states while states that model silence are context independent. Each CD state is modeled using EMMs/GMMs for a total of 100K components in our reference models. Recognition uses a decoder with a statically compiled and minimized word network, allowing for multiple pronunciations and contexts spanning more than one word. The language model is a 54M 4-gram, interpolated back-off model trained on 335M words. The lexicon contains 84K word tokens (1.08 pronunciation variants average per word). When possible, pronunciations are based on PRONLEX (LDC97L20).

Baseline models were built using fMMI-PLP features for GMMs, and fMMI-PLP + logP features for EMMs. GMMs and EMMs models were built from data with a range of sizes shown in Table 1. WERs for all models on the dev04f test set are presented in Table 1. WER for our reference GMM (100K) is 20.4% and 19.2% for EMMs, a significant improvement. WERs for GMM and EMM reference models clustered with KL and wLML show that, in both cases, wLML performs better than KL for smaller size models. wLML performance is very close to baseline performance for GMMs down to 30K with some small differences below that. At 10K, wLML gives 23.5% for 23.0% for baseline model, a 2.1% relative difference. For EMMs, wLML is very close to baseline models across all sizes. At 10K, wLML gives 20.6% for 20.4% for baseline model (a 1% relative difference). KL is performing very well across all sizes, a notch behind wLML for models below 20K.

One difference between baseline and clustered models comes from the number of components assigned to each CD state. This assignment may not be optimal for clustered models, while it is partly optimized during training for baseline models. By using the same assignment as for baseline models built from data, wLML* in Table 1, we obtain WERs that are almost identical to wLML. Hence, assignment difference does not account for the difference between wLML and baseline at 10K. Overall, wLML performs well for our combined feature EMMs.

9. Conclusions

In this paper, we introduced the variational KL divergence for EMMs, and derived a related varEM algorithm. From varKL, we derived wLML for EMMs. These variational techniques were used in the context of model restructuring given the task of clustering down reference models so to closely match performance of models built from data. This paper not only extends previous work by defining varKL, varEM and wLML for the broad class of EMMs, but also presents results for restructuring techniques applied to models built on discriminative features (fMMI-PLP features). Future work includes restructuring discriminatively trained models using boosted maximum mutual information (bMMI) criterion.

10. Acknowledgements

The authors would like to thank Tara Sainath and Bhuvana Ramabhadran for providing us with SPIF features.

11. References

[1] S. Kullback, *Information Theory and Statistics*. Dover Publications, Mileona, New York, 1997.

[2] P. L. Dognin, J. R. Hershey, V. Goel, and P. A. Olsen, "Refactoring acoustic models using variational density approximation," in *ICASSP*, April 2009, pp. 4473–4476.

[3] ——, "Refactoring acoustic models using variational expectation-maximization," in *Interspeech*, September 2009, pp. 212–215.

[4] ——, "Restructuring acoustic models for client and server-based automatic speech recognition," in *SQ2010*, Mar 2010. [Online]. Available: www.spokenquery.org

[5] K. Zhang and J. T. Kwok, "Simplifying mixture models through function approximation," in *NIPS 19*. MIT Press, 2007, pp. 1577–1584.

[6] X.-B. Li, F. K. Soong, T. A. Myrvoll, and R.-H. Wang, "Optimal clustering and non-uniform allocation of Gaussian kernels in scalar dimension for HMM compression," in *ICASSP*, March 2005, pp. 669–672.

[7] P. A. Olsen and K. Visweswariah, "Fast clustering of Gaussians and the virtue of representing Gaussians in exponential model format," in *ICSLP*, October 2004.

[8] T. Sainath, D. Nahamoo, R. Ramabhadran, and D. Kanevsky, "Sparse representation phone identification features for speech recognition," Speech and Language Algorithms Group, IBM, Tech. Rep., 2010.

[9] V. Goel and P. A. Olsen, "Acoustic Modeling Using Exponential Families," in *Proc. Interspeech*, 2009.

[10] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Annals of Statistics*, vol. 19, no. 4, pp. 2032–2066, 1991.