A GENERALIZED FAMILY OF PARAMETER ESTIMATION TECHNIQUES

Dimitri Kanevsky[1]*, Tara N. Sainath*[2] *and Bhuvana Ramabhadran*[1]

[1]IBM T.J. Watson Research Center, Yorktown, NY 10598, USA
[2]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA
kanevsky@us.ibm.com[1], tsainath@mit.edu[2], bhuvana@us.ibm.com[1]

ABSTRACT

The Extended Baum-Welch (EBW) Transformations is one of a variety of techniques to estimate parameters of Gaussian mixture models. In this paper, we provide a theoretical framework for general parameter estimation and show the relationship between these different techniques. We introduce a general family of model parameter updates that generalizes a Baum-Welch (BW) recursive process to an arbitrary objective function of Gaussian Mixture Models, and show how other common parameter estimation techniques belong to this family of model update rules. Furthermore, we formulate the construction of an even more general family of update rules that has any specified value as a gradient steepness which belongs to the family of EBW gradient steepness, measuring how much an initial model is moved to an estimated updated model.

***Index Terms*—** Pattern recognition, gradient methods.

1. INTRODUCTION

There has been extensive research in estimating parameters for Gaussian Mixture Models, applied to a wide range of natural language processing tasks, such as part-of-speech tagging, word segmentation, optical character recognition, as well as acoustic modeling in speech recognition.

One of the most popular methods for parameter estimation is the Extended Baum-Welch (EBW) Transformations [1]. However, many other parameter estimation/adaptation techniques such as Maximum Likelihood (ML), Maximum A-Posteriori (MAP), Maximum Likelihood Linear Regression (MLLR), Constrained Line Search (CLS), have been used to do parameter estimation. The purpose of this paper is to provide a theoretical framework for general parameter estimation and illustrate the relationship between these different approaches. We introduce a general family of model parameter updates that generalizes a Baum-Welch (BW) recursive process to an arbitrary objective function of Gaussian Mixture Models. We show that popular estimation/adaptation techniques as ML, MLLR, MAP and CLS, belong to this family of model update rules.

In addition, we show that this family of model update rules has a gradient steepness, measuring how much an initial model is moved to an estimated updated model, similar to the gradient steepness derived for the EBW-based model updates in [2]. Since this gradient steepness is non-negative it guarantees that this family of update rules increases the value of the objective function with each iteration. Using linearization techniques, we formulate the construction of an even more general family of update rules that has any real value for the gradient steepness. Specifically, model update rules for Gaussian parameters are derived for the case where the gradient steepness is an approximation of the Kullback-Leibler (KL) distance between updated and initial models.

The rest of the paper is structured as the following. In Section 2 we introduce the generalized representation and establish the analogy between EBW and BW while providing a family of EBW update rules. In Section 3 we show that various other estimation and adaptation techniques belong to the generalized family of update rules. In Section 4 we show how to construct update rules that have a prespecified gradient steepness.

2. AN IN-DEPTH LOOK AT THE EBW TRANSFORMATIONS

2.1. Background in Statistical Optimization

Given an initial model for the data and an objective function, there are many statistical optimization techniques to estimate a new model for the data. In the simplest case, maximizing the objective function directly will lead to a new model estimate. However, in situations where the objective function cannot be maximized directly, an auxiliary function is defined, where maximizing the auxiliary function leads to an increase in the objective function. Standard techniques to re-estimate model parameters by maximizing the auxiliary function include both the Expectation Maximization (EM) and BW algorithms. The disadvantage of these methods is that the auxiliary function is only defined if the objective function is a likelihood function.

To address this issue, another optimization technique involves finding the extremum (that is minimum or maximum) of an associated function. The benefit of the associated function is that it is defined for any rational objective function. The EBW Transformations [1] provide closed form solutions to re-estimate model parameters such that the re-estimated model parameters increase (or decrease) the associated and corresponding objective functions. In the next section, we derive these EBW Transformations in more detail.

2.2. Extended Baum-Welch Transformations

Assume that data $X = (x_1, ...x_M)$, from frames 1 to M, is drawn from a Gaussian λ_j parameterized by the following mean and variance parameters $\lambda_j = \{\mu_j, \sigma_j\}$. Let us define the probability of frame $x_i \in X$ given model λ_j as $p(x_i|\lambda_j) = z_{ij} = \mathcal{N}(\mu_j, (\sigma_j)^2)$. Let $F(z_{ij})$ be some objective function over z_{ij} and $c_{ij} = z_{ij}\frac{\delta}{\delta z_{ij}}F(z_{ij})$.

Given this function and initial model parameters λ_j, the model parameters $\hat{\lambda}_j$ (and $\hat{z}_{ij}$) are re-estimated by finding solutions which increase the associated function Q, given as follows:

$$Q(\lambda_j, \hat{\lambda}_j) = \sum_i z_{ij} \frac{\delta F(\{z_{ij}\})}{\delta z_{ij}} \log \hat{z}_{ij} \quad (1)$$

Optimizing Equation 1 will lead to closed-form update rules, known as the EBW transformations, that are generally not obtainable by optimizing F directly[1] . These update rules to re-estimate model parameters $\lambda_j(D) = \{\mu_j(D), \sigma_j(D)\}$ are given as follows:

$$\hat{\mu}_j = \hat{\mu}_j(D) = \frac{\sum_{i=1}^{M} c_{ij} x_i + D\mu_j}{\sum_{i=1}^{M} c_{ij} + D} \tag{2}$$

$$(\hat{\sigma}_j)^2 = \hat{\sigma}_j(D)^2 = \frac{\sum_{i=1}^{M} c_{ij} x_i^2 + D\left((\mu_j)^2 + (\sigma_j)^2\right)}{\sum_{i=1}^{M} c_{ij} + D} - (\hat{\mu}_j)^2 \tag{3}$$

Here D is a large constant chosen such that the associated function, and corresponding objective function increases with each iteration, that is $F(\hat{z}_{ij}) \geq F(z_{ij})$.

2.3. Linearization of EBW Transformations

Let us take a deeper look at the meaning of these transformations, by linearizing the means and variance parameters. First, let us re-write Equation 2 as follows:

$$\hat{\mu}_j = \frac{\frac{\sum_{i=1}^{M} c_{ij} x_i}{D} + \mu_j}{\frac{\sum_{i=1}^{M} c_{ij}}{D} + 1} \tag{4}$$

Furthermore, we assume the following Taylor series expansion for the denominator where terms with $1/D^2$ are combined together.

$$\frac{1}{\frac{\sum_{i=1}^{M} c_{ij}}{D} + 1} = 1 - \frac{\sum_{i=1}^{M} c_{ij}}{D} + o\left(\frac{1}{D^2}\right) \tag{5}$$

Substituting Equation 5 into 4, we get the following

$$\hat{\mu}_j = \frac{\sum_{i=1}^{M} c_{ij} x_i}{D} + \left(1 - \frac{\sum_{i=1}^{M} c_{ij}}{D}\right)\mu_j + o\left(\frac{1}{D^2}\right) \tag{6}$$

Assuming $\alpha_j = \frac{\sum_{i=1}^{M} c_{ij}}{D}$, Equation 6 can be re-written

$$\hat{\mu}_j = \alpha_j \left(\frac{\sum_{i=1}^{M} c_{ij} x_i}{\sum_{i=1}^{M} c_{ij}}\right) + (1-\alpha_j)\mu_j + o\left(\frac{1}{D^2}\right) \tag{7}$$

Intuitively, we see that the EBW update for $\hat{\mu}_j$ is a weighted combination of the initial mean μ_j and the extremum of the associated function.

Let us derive a similar linearization for the EBW variance given in Equation 3. Assuming the same Taylor series expansion given in Equation 5, we can rewrite Equation 3 as follows:

$$\hat{\sigma}_j^2 = \alpha_j \left(\frac{\sum_{i=1}^{M} c_{ij} x_i^2}{\sum_{i=1}^{M} c_{ij}}\right) + (1-\alpha_j)(\mu_j^2 + \sigma_j^2) - \hat{\mu}_j^2 + o\left(\frac{1}{D^2}\right) \tag{8}$$

Now, rewriting the linearization for the updated mean $\hat{\mu}_j^2$ as

$$\hat{\mu}_j^2 = \mu_j^2 + 2\alpha_j\mu_j \frac{\sum_{i=1}^{M} c_{ij}(x_i - \mu_j)}{\sum_{i=1}^{M} c_{ij}} + o(\frac{1}{D^2}) \tag{9}$$

[1]Note that when the objective F is the log-likelihood function (e.g., standard MLE estimation in HMM, using the Baum-Welch method), then Q coincides with the auxiliary function

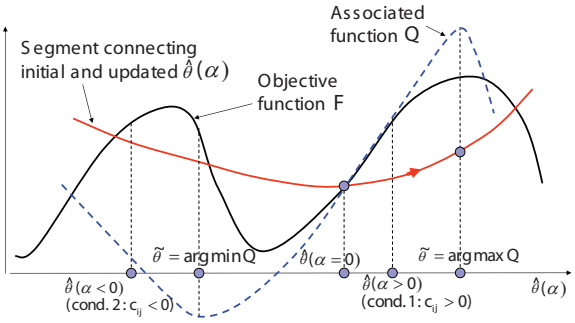

Fig. 1. Illustration of Family of EBW Update Rules

and substituting this into Equation 8 gives the following equation for the updated variance after simplification:

$$\hat{\sigma}^2 = \alpha_j\left(\frac{\sum_{i=1}^{M} c_{ij} x_i^2 - 2\mu_j \sum_{i=1}^{M} c_{ij}(x_i - \mu_j) - \sum_{i=1}^{M} c_{ij}\mu_j^2}{\sum_{i=1}^{M} c_{ij}}\right)$$
$$+(1-\alpha_j)\sigma_j^2 + o(\frac{1}{D^2}) \tag{10}$$

Given Equation 10, we can also rewrite the EBW update for $\hat{\sigma}_j^2$ as a weighted combination of the initial variance σ_j^2 and the extremum of the associated function, as similarly done for $\hat{\mu}_j$.

$$\hat{\sigma}_j^2 = \alpha_j \left(\frac{\sum_{i=1}^{M} c_{ij}(x_i - \mu_j)^2}{\sum_{i=1}^{M} c_{ij}}\right) + (1-\alpha_j)\sigma_j^2 + o\left(\frac{1}{D^2}\right) \tag{11}$$

In the next section, we look at more general parameter re-estimation techniques which can be written as a weighted combination of the initial model and extremum of the associated function.

2.4. Family of EBW Update Rules

In Section 2.3, we showed that the EBW mean and variance re-estimation formulas could be written as a weighted combination of the initial model and extremum of the associated function, plus some higher order taylor series term. More formally, we can describe model updates of this form by the following equation:

$$\hat{\lambda}_j \propto g_j(\alpha_j)\lambda_j^{extremum} + (1 - g_j(\alpha_j))\lambda_j^{init} + f_j(\alpha_j). \tag{12}$$

Here $\lambda_j^{extremum}$ is an extremum of an associated function (Equation 1) and $g_j(\alpha_j)$ is just a function over the weight α_j. We require that $g_j(\alpha_j)$ be differentiable and $g_j(0) = 0$. In Section 2.3 when the mean and variances were linearly derived, we assumed that $g_j(\alpha_j) = \alpha_j$.

A graphical representation of these model updates is shown in Figure 1($\lambda = \theta$). When $\alpha_j > 0$, the updated model $\hat{\lambda}_j$ is a weighted combination of the initial model and *maximum* of the associated function, with the weighting controlled by $g_j(\alpha_j)$. Similarly, when $\alpha_j < 0$, $\hat{\lambda}_j$ is a weighted combination of the initial model and the *minimum* of the associated function. We will refer to model updates given in Equation 12 as belonging to the family of EBW update rules.

3. MODEL ADAPTATION TECHNIQUES IN THE FAMILY OF EBW UPDATE RULES

In this section, we show how other commonly used model adaptation techniques are all members of the family of EBW-based update rules.

3.1. Maximum A-Posteriori Estimation

Maximum A-Posteriori (MAP) adaptation [3] is another popular model adaptation technique to move models from a source to target domain. The MAP mean estimate is in the family of EBW update rules and can be written as Equation 12.

The MAP mean re-estimation formula given by Equation 13, is exactly the same as the EBW formula for mean, and therefore by definition is in the family of EBW update rules.

$$\hat{\mu}_j = \hat{\mu}_j(D) = \frac{\sum_{i=1}^{M} c_{ij} x_i + D\mu_j}{\sum_{i=1}^{M} c_{ij} + D} \tag{13}$$

Similarly, using equations for variance updates in ([3]) it is easy to show that they are proportional to the updates rules for variances in the EBW family (12).

3.2. Constrained Line Search

The Constrained Line Search (CLS) [4], is another popular model adaptation technique used in discriminative training. CLS uses the same mean parameters as EBW but the variance update is based on a summation of the log variance, which translates into a multiplicative form of model updates for variance rather than an additive form. To show that this multiplicative form of EBW-based update rules is also in the family of EBW update rules, first, let us write the model update for CLS as follows:

$$\hat{\lambda}(\alpha') = \tilde{\lambda}^{(\alpha')} \lambda^{(1-\alpha')} \tag{14}$$

Here $\tilde{\lambda}$ is an extremum of an associated function (Equation 1). Next, let us chose $g(\alpha)$ to be

$$g(\alpha) = \frac{\hat{\lambda}(\alpha') - \lambda}{\tilde{\lambda} - \lambda} \tag{15}$$

We can then rewrite Equation 14 as

$$\hat{\lambda}(\alpha') = \tilde{\lambda}^{(\alpha')} \lambda^{(1-\alpha')} = g(\alpha)\tilde{\lambda} + (1 - g(\alpha))\lambda \tag{16}$$

Therefore the updated CLS variance can also be represented as a weighted combination of initial and extremum of associated function models, and is therefore in the family of EBW update rules.

The difference between CLS and EBW as it is used in [5] lies in the fact that in CLS a controlling parameter α is chosen which is inversely proportional to the KL distance between initial and updated models, which prevents significant changes between model updates. However, in [5] α is chosen inversely proportional to state occupancy counts.

3.3. MLLR

MLLR [6] is another common model adaptation technique. Update model $\hat{\mu}_j$ is defined by taking a linear transformation of initial model μ_j, as defined by Equation 17

$$\hat{\mu}_j = A\mu_j + b \tag{17}$$

Here $\eta = \{A, b\}$ is chosen to maximize the following

$$\{A^*, b^*\} = \arg\max_{\eta} p(\mu_j|\eta)$$

Before going on to show that MLLR is in the family of EBW, we must first consider the following Generalized Family of Parameter estimation Techniques given in Equation 12 in matrix form.

$$\Lambda_j \left(\frac{\sum_{i=1}^{M} c_{ij} x_i}{\sum_{i=1}^{M} c_{ij}} \right) + (I - \Lambda_j)\mu_j + o(|\Lambda|_1) \tag{18}$$

Here $x_i,\ \mu_j \in R^n$ are vectors, $I \in R^{n \times n}$ is the identity matrix and $\Lambda \in R^{n \times n}$ is an $n \times n$ matrix. The 1-norm $|\Lambda|_1 = \sum_{ij} |\lambda_{ij}|$ where sumis taken for all entries λ_j.

In order to see how the model update rules given in Equation 17 relate to the family of Matrix EBW Update rules given in Equation 18, we first rewrite Equation 18 such that the μ_j term is independent as follows:

$$\hat{\mu}_j = \mu_j + \Lambda_j \sum_{i=1}^{M} \frac{c_{ij}}{\sum_{i=1}^{M} c_{ij}} (x_i - \mu_j)$$

$$= \mu_j + \Lambda_j(\tilde{\mu}_j - \mu_j) = \mu_j(1 - \Lambda_j) + \Lambda_j \tilde{\mu}_j \tag{19}$$

Intuitively, Equation 19 is another representation for the family of EBW update rules in matrix form. Using Equations 17 and 19 we see that $\hat{\mu}_j$ can be written as:

$$\hat{\mu}_j = A^*\mu_j + b^* \ = \mu_j + \Lambda_j(\tilde{\mu}_j - \mu_j) \tag{20}$$

In general, if $\tilde{\mu}_j \neq \mu_j$, it is easily to see that Equation 20 is solvable for Λ_j (and in fact has infinite number of solutions). A "degenerate" case $\tilde{\mu}_j = \mu_j$ means that μ_j is already an ML estimate of $p(\mu_j|\eta)$. If $\tilde{\mu}_j = \mu_j$ for all j then one can take $A^* = 1$ and $b^* = 0$. Similarly, one deduct a variance representation in MLLR as a matrix form of EBW for variance. Thus, MLLR model update rules are, in general (i.e. exept of degenerated cases), also in the family of EBW.

4. DISTANCE TECHNIQUES IN THE FAMILY OF EBW GRADIENT STEEPNESS METRICS

4.1. EBW Gradient Steepness Metric

The purpose of this section is to demonstrate that our linearization technique allows us to introduce update rules when gradient steepness metrics are given in advance. We specifically apply this to the KL gradient metric. We use notation of the section 2.2 and for simplicity assume that there only one set of parameter models μ, σ, i.e. we drop a subscript j for model parameters. Also let $\Phi, \Psi \in R$ be some real numbers and $z_i = z_i(\mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2)$. First we derive our linearized EBW gradient steepness metric and then go on to show the relation between this gradient metric and the KL distance.

4.2. Linearization of EBW Gradient Steepness

The EBW gradient metric can be derived by the following theorem:

Theorem 1 *Let $F(\{z_i\})$ be a differentiable function of z_i. Let $c_i = \frac{z_i \delta F(\{z_i\})}{\delta z_i}$. Let $T \in R$ be any real number and $\Psi, \Phi \in R$ be such that the following equality holds:*

$$\Psi \sum_i c_i \left[-1/2 + \frac{(y_i - \mu)^2}{2\sigma^2} \right] + \Phi \sum_i c_i (y_i - \mu) = \sigma^2 T \tag{21}$$

Then for the following mean and variance transformations

$$\mu(C) = \mu + \Phi/C, \text{ and } \sigma(C)^2 = \sigma^2 + \Psi/C \quad (22)$$

we get the following gradient steepness:

$$F(\{z_i(\lambda(C))\}) - F(\{z_i(\theta)\}) = T/C + o(1/C^2) \quad (23)$$

where $\lambda(C) = \{\mu(C), \sigma(C)\}$ *Specifically, if* $T > 0$ *then for sufficiently large* C $F(\lambda(C)) > F(\lambda)$.

Here T measures the gradient required to adapt initial model λ to data y_i. The larger the value in T indicates that the gradient to adapt the initial model to the data is steeper and the data is much better explained by the updated model $\hat{\lambda}_j(C)$. These gradient steepness metrics were succesfully used in various speech recognition classification, segmentation and decoding tasks (see [7], [8], [9]). Below is a brief outline of the proof of the above theorem (that uses the same linearization technique as a proof of a theorem in [2])

Proof Substituting (22) into $z_i(\lambda(C))$ and using the first order Taylor series expansion we get the following equalities.

$$\hat{z}_i = z_i(\lambda(C)) \sim z_i + \frac{z_i}{\sigma^2 C}\left\{\Psi\left[-1/2 + \frac{(x_i-\mu)^2}{2\sigma^2}\right] + \Phi(x_i-\mu)\right\} \quad (24)$$

Next we have (assuming $a_i = \frac{\delta F(\{z_i\})}{\delta z_i}$ and $a_i z_i = c_i$)

$$l(\{\hat{z}_i\}) = \sum_i a_i \hat{z}_i = \sum_i a_i z_i +$$

$$\frac{1}{\sigma^2 C}\left\{\Psi \sum_i a_i z_i \left[-1/2 + \frac{(x_i-\mu)^2}{2\sigma^2}\right] + \Phi \sum_i a_i z_i (x_i-\mu)\right\} =$$

$$l(\{z_i\}) + \frac{1}{\sigma^2 C}\left\{\Psi \sum_i c_i \left[-1/2 + \frac{(x_i-\mu)^2}{2\sigma^2}\right] + \Phi \sum_i c_i (x_i-\mu)\right\}$$

This implies (see [2]):

$$F(\{\hat{z}_i\}) - F(\{z_i\}) \sim l(\{\hat{z}_i\}) - l(\{z_i\}) = \frac{1}{\sigma^2 C}\Psi$$

$$\sum_i c_i \left[-1/2 + \frac{(x_i-\mu)^2}{2\sigma^2}\right] + \frac{1}{\sigma^2 C}\Phi \sum_i c_i (x_i - \mu) = T(\Phi, \Psi)/C$$

Specifically for EBW transformations we get:

$$\Phi = \sum_i c_i (x_i - \mu), \text{ and } \Psi = \sum_i c_i [(x_i-\mu)^2 - \sigma^2] \quad (25)$$

and an EBW distance metric

$$T(\Phi, \Psi) = \frac{1}{\sigma^2}\frac{\left\{\sum_i c_i \left[(x_i-\mu)^2 - \sigma^2\right]\right\}^2}{2\sigma^2} + \frac{1}{\sigma^2}\left[\sum_i c_i (x_i - \mu)\right]^2$$

4.3. Update rules with KL gradient steepness

This theorem can be applied to a situation when $T = KL(\lambda(C), \lambda)$ Specifically, one can consider the following approximation to KL:

$$KL(N(\tilde{\mu}, \tilde{\sigma}) || N(\mu, \sigma)) = 1/2\left[\log\frac{\sigma^2}{\tilde{\sigma}^2} - 1 + \frac{\tilde{\sigma}^2}{\sigma^2} + \frac{(\tilde{\mu}-\mu)^2}{\sigma^2}\right] \quad (26)$$

Let

$$\tilde{\mu} = \frac{\sum_i c_i x_i}{\sum_i c_i}, \text{ and } \tilde{\sigma}^2 = \frac{\sum_i c_i (x_i - \mu)^2}{\sum_i c_i} \quad (27)$$

with these notations we can represent (21) as the following:

$$\frac{\sum_i c_i}{2\sigma^2}\Psi\left[-1 + \frac{\tilde{\sigma}^2}{\sigma^2}\right] + \sum_i c_i \Phi \frac{(\tilde{\mu}-\mu)}{\sigma^2} = T \quad (28)$$

We can derive the following solutions of (28) for $T = KL(N(\tilde{\mu}, \tilde{\sigma}) || N(\mu, \sigma))$.

$$\Phi_{KL} = \frac{1}{2\sum_i c_i}(\tilde{\mu} - \mu), \text{ and } \Psi_{KL} = \frac{\sigma^2}{\sum c_i}\left[1 + \frac{\log\frac{\sigma^2}{\tilde{\sigma}^2}}{\frac{\tilde{\sigma}^2}{\sigma^2} - 1}\right] \quad (29)$$

This gives the following update rules with KL gradient steepness:

$$\hat{\mu}_{KL} = \mu + \Phi_{KL}/C, \text{ and } \hat{\sigma}^2_{KL} = \sigma^2 + \Psi_{KL}/C \quad (30)$$

Thus model updates can be constructed given the KL metric.

5. CONCLUSION

In this work, we extended a BW concept to an arbitrary objective function by introducing the concept of associated functions and a family of EBW-based update rules. We showed that some popular estimation and adaptation techniques as MAP, CLS and MLLR belong to an EBW family. We also demonstrated that our linearization technique allows to introduce update rules that have any gradient steepness metrics given in advance. Specifically we applied it to a gradient steepness metrics that approximate KL distances.

6. REFERENCES

[1] P.S. Gopalakrishnan, D. Kanevsky, D. Nahamoo, and A. Nadas, "An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems," *IEEE Trans. Information Theory*, vol. 37, no. 1, January 1991.

[2] D. Kanevsky, "Extended Baum Transformations for General Functions," in *Proc. ICASSP*, 2004.

[3] J.L. Gauvain and C.H. Lee, "Maximum a Posteriori Estimationg for Multivariaet Gaussian Observations of Markov Chains," *IEEE Trans. on Speech and Audio Processing*, 1994.

[4] C. Liu, P. Liu, H. Jiang F. Soong, and R. Wang, "Constrained Line Search Optimization for Discriminative Training in Speech Recognition," in *ICASSP*, 2007.

[5] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*, Ph.D. thesis, Cambridge University, 2003.

[6] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *CSL*, 1995.

[7] T. N. Sainath, V. Zue, and D. Kanevsky, "Audio Classification using EBW Transformations," in *Proc. Interspeech*, 2007.

[8] T. Sainath, D. Kanevsky, and B. Ramabhadran, "Broad Phoentic Recognition in a Hidden Markov Model Framework Using Extended Baum-Welch Transformations," in *Proc. ASRU*, 2007.

[9] T. N. Sainath, D. Kanevsky, and B. Ramabhadran, "Gradient Steepness Metrics Using Extended Baum-Welch Transformations for Universal Pattern Recognition Tasks," in *Proc. ICASSP*, April 2008.