BROAD PHONETIC CLASS RECOGNITION IN A HIDDEN MARKOV MODEL FRAMEWORK USING EXTENDED BAUM-WELCH TRANSFORMATIONS

Tara N. Sainath, Dimitri Kanevsky and Bhuvana Ramabhadran

IBM T. J. Watson Research Center
Yorktown, NY 10598, U.S.A.
tsainath@mit.edu, {kanevsky, bhuvana}@us.ibm.com

ABSTRACT

In many pattern recognition tasks, given some input data and a model, a probabilistic likelihood score is often computed to measure how well the model describes the data. Extended Baum-Welch (EBW) transformations are most commonly used as a discriminative technique for estimating parameters of Gaussian mixtures, though recently they have been used to derive a gradient steepness measurement to evaluate the quality of the model to match the distribution of the data. In this paper, we explore applying the EBW gradient steepness metric in the context of Hidden Markov Models (HMMs) for recognition of broad phonetic classes and present a detailed analysis and results on the use of this gradient metric on the TIMIT corpus. We find that our gradient metric is able to outperform the baseline likelihood method, and offers improvements in noisy conditions.

Index Terms— Gradient Methods, Viterbi Decoding, Hidden Markov Models, Speech Recognition

1. INTRODUCTION

The EBW transformations [1] are one of a variety of discriminative training techniques ([2], [3]) that have been explored in the speech recognition community to estimate model parameters of Gaussian mixtures. Given an initial model and input data, [4], [5] derive an explicit formula to measure the gradient steepness required to estimate a new model via the EBW transformations. This gradient steepness measurement is an alternative to likelihood to describe how well the initial model explains the data.

We have observed the advantages of this gradient steepness measurement in a variety of tasks. In [6] we redefined the likelihood ratio test, typically used for unsupervised segmentation, with this measure of gradient steepness. We showed that our EBW unsupervised audio segmentation method offered improvements over the Bayesian Information Criterion and Cumulative Sum methods. In [7], we used this gradient metric to develop an audio classification method which was able to outperform both the likelihood and SVM techniques.

Hidden Markov Models (HMMs) [8] have been the most dominant frame-based acoustic modeling technique for automatic speech recognition tasks to date. When HMMs are used for acoustic modeling, generally the Viterbi algorithm is used during decoding to find most likely sequence of HMM states and corresponding words. This is done by computing likelihood scores for each frame given all HMM states, and then doing a dynamic programming Viterbi search to find the most likely sequence of states. In this work, we look at replacing the likelihood scores computed in Viterbi decoding with the EBW gradient steepness measurement.

Specifically, we focus on recognition of broad phonetic classes (BPCs). Our motivation for looking at BPC recognition is twofold. First, the simple nature of the task allows us to investigate various properties of the EBW transformations in the context of HMMs. Because HMMs are so widely used in speech recognition, success of our gradient steepness measure in an HMM framework will introduce a new decoding metric that can be explored for more complicated medium and large vocabulary speech recognition tasks. Secondly, broad phonetic class recognition is important in a wide variety of contexts. For example, [9] explores using BPCs to speed up lexical access, while [10] investigates expert classifiers specific to each broad phonetic class and performs phonetic classification by combining scores from the different experts. We are interested in exploring BPC recognition to aid in the placement of segments in a segment-based speech recognition system in noisy conditions, one idea which was initially demonstrated in [11].

In this paper, we continue to expand on previous work ([6], [7]) and demonstrate that the EBW gradient steepness measure appears to be a general technique to explain the quality of a model used to represent the data. First, we introduce a novel change to our EBW gradient measurement and explain model fit to the data by looking at a relative, rather than absolute, change in the gradient. We find this novel EBW metric outperforms the standard likelihood method in BPC recognition on the TIMIT corpus. Secondly, we investigate specific properties of EBW transformations. Specifically, we introduce a novel idea to minimize parameter training required to estimate new models via the transformations. Also, we explore the advantages of EBW model re-estimation in noisy environments, demonstrating the improved performance of our gradient steepness metric over likelihood across a variety of signal-to-noise ratios.

In the following sections, we describe the EBW transformations. Our implementation of our EBW gradient metric in an HMM framework is described in Section 3. Section 4 presents the experiments performed, followed by a discussion of these results in Section 5. Finally, Section 6 concludes the paper and discusses future work.

2. EXTENDED BAUM-WELCH TRANSFORMATIONS

2.1. Motivation of using EBW Transformations

Given some input data, there are many different approaches used to calculate how well a model represents this data. One common approach is to calculate the likelihood, that is $p(data|model)$. Another method is to calculate the gradient, as shown in Figure 1. Given an initial model for our data and an objective function, we can estimate a new model for our data by finding the best step along the gradient of the objective function. We can think of the gradient slope as measuring how much we have to adapt an initial model to fit the data. A steep slope indicates the initial model does not fit the data well, while a flat slope indicates the initial model is a good fit for the data. The

EBW transformations provide solutions to estimate this new model, and also provide a measure of the gradient steepness to explain the quality of the initial model to fit the data.

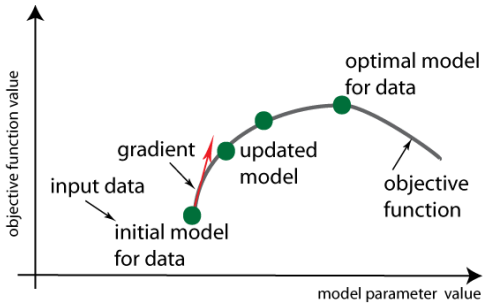

Fig. 1. EBW Model Update Graph

2.2. Derivation of EBW Transformations

The EBW procedure involves continuous transformations that can be described as follows. Assume that frame x_i is drawn from Gaussian mixture model (GMM) λ^k, with each component $j \in k$ parameterized by the following mean and variance parameters $\lambda_j^k = \{\mu_j^k, \sigma_j^k\}$, and weight w_j^k. Thus GMM λ^k includes all the parameters of the individual components, in other words $\lambda^k = \{\lambda_1^k, \dots \lambda_N^k\}$ and weights $w^k = \{w_1^k, \dots w_N^k\}$. Let us define the probability of frame x_i given mixture component j as $p(x_i|\lambda_j^k) = z_{ij}^k = \mathcal{N}(\mu_j^k, (\sigma_j^k)^2)$ and similarly $z_i^k = \sum_{j=1}^N w_j^k z_{ij}^k$. Let $F(z_{ij}^k)$ be some objective function over z_{ij}^k and $c_{ij}^k = z_{ij}^k \frac{\delta}{\delta z_{ij}^k} F(z_{ij}^k)$. Given this function and initial model parameters λ_j^k, the EBW transformations provide formulas to re-estimate parameters $\lambda_j^k(D) = \{\mu_j^k(D), \sigma_j^k(D)\}$ as:

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

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

Here D is a large constant chosen such that the objective function increases with each iteration, that is $F(\hat{z}_{ij}^k) \geq F(z_{ij}^k)$. Using EBW transformations (1) and (2) such that $\lambda_j^k \rightarrow \hat{\lambda}_j^k(D)$ (thus $\lambda^k \rightarrow \hat{\lambda}^k(D)$) and $z_i^k \rightarrow \hat{z}_i^k$, [4], [5] derives linearization formula between $F(\hat{z}_i^k)$ and $F(z_i^k)$ for large D as:

$$F(\hat{z}_i^k) - F(z_i^k) = T_i^k/D + o(1/D) \tag{3}$$

Here T measures the gradient required to adapt initial model λ^k to data x_i. [4], [5] also show that T is always non-negative and only equals zero when $\hat{\lambda}^k$ is a local maximum of $F(z_i^k)$. This guarantees that $F(z_i^k)$ increases per iteration and provides some theoretical justification for using gradient metrics T and $\left(F(\hat{z}_i^k) - F(z_i^k)\right) \times D$ as measures of quality of fitness of models to data.

A large value in T means the gradient to adapt the initial model to the data is steep and $F(\hat{z}_i^k)$ is much larger than $F(z_i^k)$. Thus the data is much better explained by the updated model $\hat{\lambda}^k(D)$ compared to the initial model λ^k. However a small value in T indicates that the gradient is relatively flat and $F(\hat{z}_i^k)$ is close to $F(z_i^k)$. Therefore, the initial model λ^k is a good fit for the data. In the next section, we derive our EBW gradient steepness metric for HMMs using the objective function given in Equation 3.

3. EBW GRADIENT MEASURE IN HMMS

Given a set of acoustic observations $O = \{o_1, o_2 \dots o_T\}$ associated with a speech waveform, the goal of the acoustic model is to find which sequence of sub-word units $\hat{W} = \{w_1, \dots w_k\}$ that most likely produced the given observation sequence. In other words, we want to maximize $\hat{W} = \arg\max_W P(O|W)$. While we represented each sub-word unit as a GMM in [7], here we look at representing each sub-word unit as a state from a HMM, and will subsequently extend our EBW implementation in this context.

3.1. HMMs

Given observation sequence O, HMMs can be used in decoding tasks to find the most optimal state sequence through time $Q = \{q_1, q_2 \dots q_T\}$ that produced the given observations. An HMM is defined over a set of states $S = \{s_1, s_2 \dots s_N\}$ and observations O, and is represented by the following three parameters [8]:

- State Transition Probability Distribution:
  $a_{ij} = P(q_t = s_j | q_{t-1} = s_i)$
- Observation Symbol Probability Distribution:
  $b_i(o_t) = P(o_t | q_t = s_i)$
- Initial State Distribution:
  $\pi_i = P(q_1 = s_i)$

Let us assume that the output distribution for each state s_k is drawn from a mixture of L gaussians where z_{tj}^k is the likelihood of observation o_t given component j from GMM k and w_j^k the *a priori* weight of component j. Then we can define the log-likelihood of o_t from model λ^k as follows:

$$b_k(o_t) = P(o_t | q_t = s_k) = \sum_{j=1}^L w_j^k z_{tj}^k \tag{4}$$

Given the set of states S and corresponding models $\Lambda = \{\lambda^1, \lambda^2 \dots \lambda^N\}$, one possible optimization criterion to find the best state sequence is to choose at each time instance, the state q_t which is individually most likely. In other words,

$$q_t(j) = \arg \max_{1 \leq i \leq N} [P(q_t = s_i | O, \Lambda)] \tag{5}$$

We could replace this likelihood function with the EBW gradient steepness measurement in Equation 3, as was done in [7], and score each frame individually. However, when we model sub-word units via states of an HMMs, Equation 5 does not take into account neighboring state dependencies [8] and therefore is not the best criterion to find the optimal state sequence. Instead, the Viterbi algorithm is generally used to find the most optimal state sequence. To find this sequence, first define $\delta_t(i)$ as the best score along a single path up to time t which ends in state s_i at time t as:

$$\delta_t(i) = \max_{q_1, q_2, \dots, q_{t-1}} P(q_1 q_2 \dots q_t = S_i, O_1 O_2 \dots O_t | \Lambda) \tag{6}$$

By induction, the probability of the best path up to time t which ends in state s_j at time $t+1$ is defined as:

$$\delta_{t+1}(j) = \max_i [\delta_t(i) + log(a_{ij})] + log(b_j(o_{t+1})) \tag{7}$$

Therefore, we see that the best state at each time instance is not simply the individual most likely state as given by Equation 5. Instead, it depends on the scores assigned to previous states as well as transition probabilities a_{ij} between states, capturing the inherent HMM structure. In the next section, we discuss how to find the best state sequence using the EBW gradient metric.

3.2. EBW-F

Instead of scoring each observation frame using standard likelihood, we can score it using the EBW gradient steepness measurement given in Equation 3. Let us define objective function $F(z_t^k)$ to be the log-likelihood of observation o_t given state model λ^k as:

$$F(z_t^k) = \log \sum_{j=1}^{L} w_j^k z_{tj}^k \tag{8}$$

and similarly c_{tj}^k as:

$$c_{tj}^k = z_{tj}^k \frac{\delta}{\delta z_{tj}^k} F(z_{tj}^k) = \frac{z_{tj}^k w_j^k}{\sum_{l=1}^{N} w_l^k z_{tl}^k}. \tag{9}$$

Using Equation 3 and the objective function for $F(z_t^k)$ given by Equation 8, we can compute the state output score at frame o_t as:

$$b_k(o_t) = \left(F(\hat{z}_t^k) - F(z_t^k)\right) \times D \tag{10}$$

Here D is a constant chosen in the EBW model re-estimation formulas, given by Equations 1 and 2. If D is very large then training is very slow (but stable) but if D is too small model re-estimation may not increase the objective function on each iteration.

Using the EBW score assigned to each state from Equation 10, the best path is again found through Viterbi algorithm given in Equation 7. However, the better a model fits the data, the smaller the EBW score, so we define $\delta_t(i)$ as the set of best (smallest) EBW scores along a single path up to time t which ends in state s_i.

$$\delta_t(i) = \min_{q_1, q_2, \ldots, q_{t-1}} EBW(q_1 q_2 \ldots q_t = S_i, O_1 O_2 \ldots O_t | \Lambda) \tag{11}$$

Therefore, by induction, $\delta_{t+1}(j)$ is defined as:

$$\delta_{t+1}(j) = \min_i [\delta_t(i) - log(a_{ij})] + b_j(o_{t+1}) \tag{12}$$

Note to reflect this minimum change, we also compute the negative log-likelihood of a_{ij}. The objective function in Equation 10 is the same as that used in [7], though now applied to HMMs. In the next section, we discuss a novel change to this objective function which is more appropriate for an HMM framework.

3.3. EBW-F Normalization

As shown in Equation 10, we score how well model λ^k fits x_t by looking at the difference in likelihood given the updated model $F(\hat{z}_t^k)$ compared to the likelihood given the initial model $F(z_t^k)$. Using this absolute measure allows us to compare model scores for a given input frame, as was done in [7]. However, we have observed that the magnitude of these scores loses meaning if we compare them across different frames. In other words, a lower absolute EBW score for one frame and one model does not mean a better model than a higher EBW score for another frame and another model. However, having an EBW measure that we can compare across frames is particularly important in HMMs, as scores for a state sequence are computed by summing up scores assigned to individual frames.

Therefore, we compute the EBW score as the *relative* difference in likelihood given the updated model $F(\hat{z}_t^k)$ compared to the initial model likelihood $F(z_t^k)$. To compute this relative EBW score, we normalize Equation 10 by the original likelihood $F(z_t^k)$ as:

$$b_k(o_t) = \frac{\left(F(\hat{z}_t^k) - F(z_t^k)\right) \times D}{F(z_t^k)} \tag{13}$$

Using this relative EBW score provides a metric which can be compared across frames, which is important in the context of HMMs. In the next section, we introduce a novel idea to minimize parameter training required to estimate an optimal D.

3.4. EBW Adaptive D

D controls the rate at which updated models given in Equations 1 and 2 are trained. In [7], we explored using a global D, as well as specific D for each state, both of which required a lot of hand-tuning and training. To minimize the work needed to heuristically tune D, various approaches have been explored to set D [2].

Conceptually, the better our original models, the less we want to train our updated models and the larger we want D. And similarly, the better our original models, the larger the log-likelihood will be. Thus, in this work we investigate adapting the rate of model training at each frame based on the likelihood. Specifically, we look at the following linear relationship between D and log-likelihood:

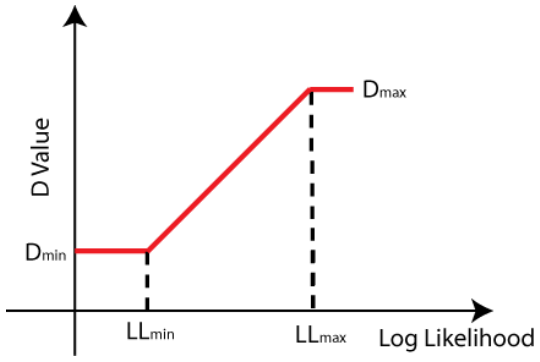

Fig. 2. Linear Transformation of Likelihood used to determine D

Here LL_{max} and LL_{min} are the upper and lower limits of the log likelihood determined from training data, and D_{max} and D_{min} are the upper and lower limits that we allow D to take. Between the likelihood limits, D is set linearly proportional to the likelihood. Intuitively, we can think of the log-likelihood as a confidence measure to determine how quickly we need to estimate the updated model.

4. EXPERIMENTS

We perform BPC recognition experiments using the TIMIT corpus. The 61 TIMIT labels are first mapped into 7 broad phonetic classes (BPC) as shown in Table 4, ignoring the glottal stop 'q'.

Our experiments use 13 dimensional, perceptual linear prediction (PLP) features obtained from a Linear Discriminant Analysis (LDA) projection that are mean and variance normalized on a per utterance basis. In addition, each BPC is modeled as a three-state, left-to-right context-independent HMM with no skip states. The output

Broad Phonetic Class	TIMIT Labels
Vowels/Semivowels	aa ae ah ao aw ax axh axr ay eh er ey ih ix iy ow oy uh uw el l r w y
Nasals/Flaps	em en eng m n ng nx dx
Strong Fricatives	s z sh zh ch jh
Weak Fricatives	v f dh th hh hv
Stops	b d g p t k
Closures	bcl pcl dcl tcl gcl kcl epi pau
Silence	h#

Table 1. Broad Phonetic Classes and corresponding TIMIT Labels

distribution in each state is modeled by a mixture of 32 component diagonal covariance Gaussians. All models were trained on the standard NIST training set (3969 utterances) in clean speech conditions. To analyze phonetic recognition performance in noise, we simulate noisy speech by adding pink noise from the Noisex-92 database [12] at signal-to-noise ratios (SNRs) in the range of 0dB to 30dB in 5dB increments. We train the EBW-F methods to find the adaptive D ranges using the dev set (400 utterances). We report recognition results on both the dev set and the full test set (944 utterances).

5. RESULTS

In this section, we discuss three different experiments performed on the TIMIT corpus. First, we analyze the BPC recognition performance of the EBW-F and EBW-F Norm and likelihood methods. Secondly, we explore the behavior of the EBW-Adaptive D metric and EBW model re-estimation in noisy environments.

5.1. EBW-F Normalization

Table 5.1 shows the phonetic recognition error rates for the likelihood, EBW-F and EBW-F Norm metrics on both test sets, with the best performing method highlighted in bold.

Method	Dev	Test
Likelihood	18.4	19.5
EBW-F	18.7	19.9
EBW-F Norm	**17.7**	**18.9**

Table 2. BPC Phonetic Error Rates on TIMIT

The EBW-F Norm method outperforms the likelihood on both the dev and test sets, but the EBW-F method performs worse than the likelihood. To explain these results, let us first look at the relationship between EBW-F and likelihood scores, evaluated on a per-frame basis, in Figure 3. Note that likelihood score shown is actually the negative log-likelihood, so the better a model explains an observation, the smaller the negative log-likelihood and EBW scores are.

First, we see there is a strong positive correlation between the EBW-F and likelihood scores. However, the variance of EBW scores for a particular likelihood score is quite large. This is mainly because the EBW-F score is an absolute measure and cannot really be compared across frames. Because Viterbi decoding determines the best path based on scores of all individual frames in that path, if the EBW score for one frame is large it dominates and throws off the entire score for the path. This is one reason the EBW-F metric performs worse than likelihood when used in an HMM context. This motivated our reason for looking at the EBW-F score in terms of relative change, and thus introducing the EBW-F Norm metric.

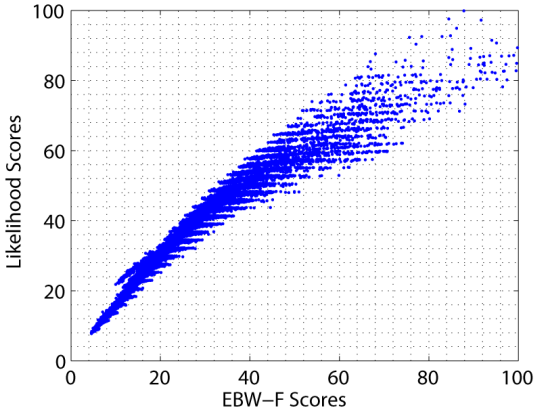

Fig. 3. EBW-F vs. Likelihood scores

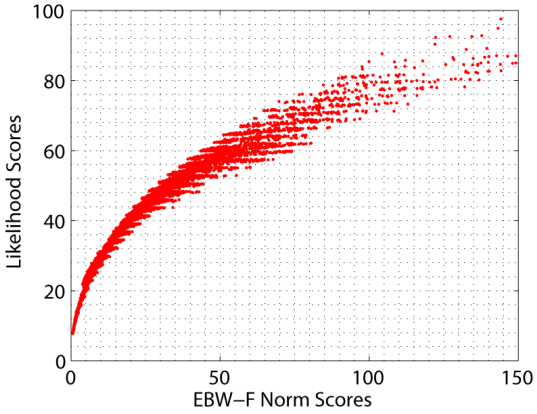

Fig. 4. EBW-F Norm vs. Likelihood scores

To understand why the EBW-F Norm metric outperforms likelihood, let us look at the relationship between the EBW-F Norm and likelihood scores as shown in Figure 4. Again, we see there is positive correlation between scores from the two metrics. However, the variance of EBW-F Norm scores for a given likelihood score is much less compared to the EBW-F metric, showing that using the relative measure allows for a more direct comparison across frames. Also, notice that as the likelihood increases and models become worse, the EBW scores move even faster there is a slight curve to the graph. As shown by Equation 13, EBW-F Norm captures the relative difference between the likelihood of a data given the initial model and the likelihood with a model estimated from the current data sequence being scored, while the likelihood just calculates the former. Thus, when models are poor to explain the data, we see that we must move the initial models quite a bit to explain the current input, and therefore the EBW scores are quite large compared to likelihood.

To better understand the advantages of this curve in Figure 4, we looked at transforming the EBW-F Norm scores to produce a more linear relationship with the likelihood scores. The Box-Cox transformations [13] are a commonly method used to make the relationship between two variables more linear. The transformations are defined

as follows:

$$\tau(EBW;\lambda) = \begin{cases} \frac{(EBW^{\lambda}-1)}{\lambda} \text{ if } \lambda \neq 0 \\ ln(EBW) \text{ if } \lambda = 0 \end{cases}$$

Here λ is the transformation parameter which controls the degree to which we transform the EBW scores. Figure 5 shows the correlation between likelihood and Box-Cox transformed EBW scores for different values of λ.

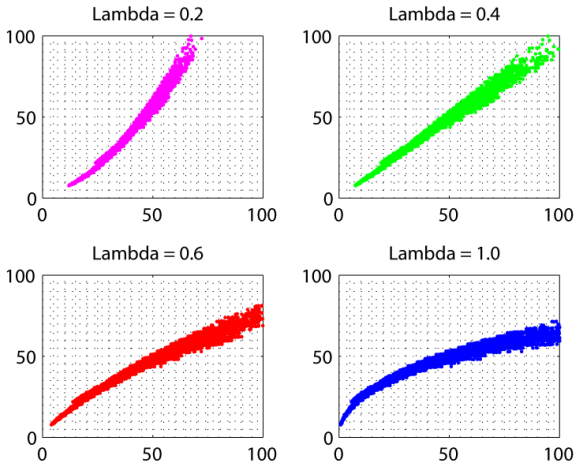

Fig. 5. EBW Box-Cox Transformed (x-axis) vs. Likelihood (y-axis) scores for different values of λ

We see that as we decrease λ, we move the correlation of EBW and likelihood from convex down to convex up. $\lambda = 0.4$ produces the most linear relationship between EBW-F Norm and likelihood, with a much smaller variance compared to Figure 3. Table 5.1 shows the PER for the EBW Box-Cox transformed scores as we vary λ.

λ	PER
0.2	20.5
0.4	19.6
0.6	19.1
1.0	**18.9**

Table 3. BPC Phonetic Error Rates on the TIMIT Test Set using EBW Box-Cox Transformed scores for variable λ

Notice that as we decrease λ and make the relationship of EBW and likelihood more linear, the PER decreases. This shows that the true benefit of EBW over the likelihood occurs when models are poor, and the EBW scores are much higher relative to likelihood, producing the curve in Figure 4. Because we sum up scores from local frames to determine the best path, the large EBW scores for poor models allows us to disregard these paths more confidently.

5.2. EBW Adaptive D

As discussed in Section 3.4, to minimize the work needed to heuristically tune D, we explore using an adaptive D which is linearly proportional to the log-likelihood. Figure 6 shows the performance

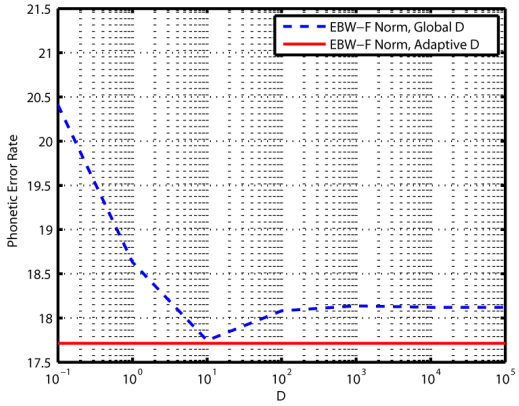

Fig. 6. Phonetic Error Rate vs. D

of the EBW-F Norm Global D and Adaptive D classifiers on the development set as we globally vary D.

First, notice that the performance of the Global D method is quite sensitive to the choice of D, and has huge changes in performance as we vary the rate at which we re-estimate updated models. If we make D smaller and train the updated model quicker, we are able to still get an appropriate estimate for the updated model while allowing the objective function to increase. However, if we take D too small then we train our models too quickly, we do not increase the value of the objective function on each iteration and therefore the performance of the Global D metric decreases.

If we use the likelihood scores as a confidence measure to linearly adapt D, the Adaptive D metric has similar performance to the global D, without having to heuristically tune D.

5.3. Model Re-estimation

In Section 5.1, we showed that the EBW-F Norm metric outperformed the likelihood method due to the model re-estimation inherent in EBW. In this section, given models trained in clean conditions, we analyze the rate of model re-estimation to adapt these models to noisy environments. Recall that D controls the rate at which we train updated models. We would expect that as models become a worse fit for the data, we must make D smaller and re-estimate models faster. Figure 7 shows the PER for the EBW-F Norm metric on the dev set for different SNRs as we vary D. Please note that we are using the Adaptive-D metric discussed in Section 5.2, and here D indicates the average range over which we adapt D.

As the SNR decreases and the clean speech models become poorer estimates of the noisy data, we must increase D and train models quicker for better performance, as indicated by the circles in Figure 7. This shows the importance of the rate of model re-estimation, particularly when models are not a good fit for the data.

Table 5.3 shows the PER rate on the dev and test sets for the EBW and likelihood methods across a variety of SNRs when models are trained in clean conditions. We see that as we increase the SNR, the model re-estimation inherent in EBW allows for significant improvement over the likelihood metric. Thus, we see that the EBW-F Norm metric also provides a simple yet effective noise robust technique over the likelihood measure.

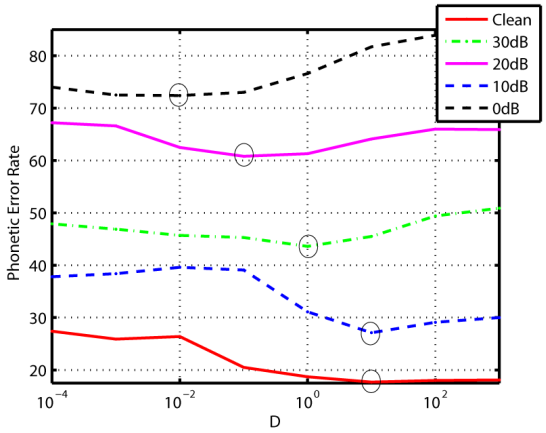

Fig. 7. Phonetic Error Rate vs. D for different SNRs. Circles indicate D at each SNR which gives lowest PER

Set	Method	clean	30dB	20dB	10dB	0dB
	Likelihood	18.4	28.2	45.0	65.2	75.6
dev	EBW-F Norm	**17.7**	**27.1**	**43.6**	**60.8**	**72.4**
	% Err. Red.	3.8	3.9	3.1	7.7	4.2
	Likelihood	19.5	29.7	46.7	66.2	75.9
test	EBW-F Norm	**18.9**	**28.6**	**45.0**	**61.5**	**71.7**
	% Err. Red.	3.1	3.7	3.6	7.1	5.5

Table 4. BPC Phonetic Error Rates on Noisy TIMIT

6. CONCLUSIONS AND FUTURE WORK

In this paper, we expanded on previous work ([6], [7]), showing that the EBW transformations appears to be a general technique to explain the quality of a model used to represent the data. Specifically, we explored doing BPC recognition using a relative EBW measurement, which we found was able outperform the standard likelihood metric. Secondly, we introduced a novel idea to minimize parameter training required to estimate updated EBW models. Finally, we explored the benefits of EBW model re-estimation in noisy environments, demonstrating the improved performance over likelihood across a variety of SNRs.

In the future, we would like to expand this work in a number of directions. Recently, we have applied EBW decoding to a Large Vocabulary Continuous Speech Recognition (LVCSR) task, namely for transcription of English Broadcast News in the distillation portion of the GALE evaluation. Some of the issues related to the choice of D and normalization on a per state basis, are being explored in this context. Our work on the TIMIT corpus provides a good understanding of the behavior of the EBW transformations and serves as a precursor to understand the issues in the LVCSR task better.

In addition, we are also interested in using BPC recognition as a preprocessing step for a variety of tasks. For example, it is well known that phones within a BPC convey similar spectral and temporal characteristics, while phones in different BPCs have quite different characteristics [10]. However, the basic properties of the same BPC across two different languages also differ. We would like to explore if this is a quick and easy way to do language detection.

In addition, we are interested in exploring BPC detection to aid in the placement of segments in a segment-based speech recognition system. In [11], we observed that while the our segment-based speech recognition systems performs well in clean speech, the system has difficulty placing landmarks (representing phonetic transitions) in the presence of noise and often produces poor recognition hypotheses. Transitions between broad phonetic classes represent places of largest acoustic change within an utterance and also represents at minimum where landmarks should be placed. Thus, we would like to explore if BPC recognition can aid in hypothesizing landmarks, particularly in noisy environments.

7. ACKNOWLEDGEMENTS

We would like to thank Brian Kingsbury of IBM for his help with model training and initial recognizer setup. Also, thank you to Emmanuel Yaschin of IBM for his guidance on the Box-Cox transformations. Finally, thanks to David Nahamoo of IBM for his suggestion to look at using the EBW gradient measurement in HMMs.

8. REFERENCES

[1] P.S. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo, "A Generalization of the Baum Algorithm to Rational Objective Functions," in *ICASSP*, 1989.

[2] D. Povey and B. Kingsbury, "Evaluation of Proposed Modifications to MPE for Large Scale Discriminative Training," in *Proc. ICASSP*, April 2007.

[3] F. Sha and L. Saul, "Comparison of Large Margin Training to Other Discriminative Methods for Phonetic Recognition by Hidden Markov Models," in *Proc. ICASSP*, April 2007.

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

[5] D. Kanevsky, "Extended Baum Transformations For General Functions, II," Tech. Rep. RC23645(W0506-120), Human Language Technologies, IBM, 2005.

[6] T. N. Sainath, D. Kanevsky, and G. Iyengar, "Unsupervised Audio Segmentation using EBW Transformations," in *Proc. ICASSP*, April 2007.

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

[8] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," in *IEEE*, February 1989, vol. 77, pp. 257–286.

[9] D. P. Huttenlocher and V. Zue, "A Model of Lexical Access for Partial Phonetic Information," in *Proc. ICASSP*, March 1984.

[10] P. Scanlon, D. Ellis, and R. Reilly, "Using Broad Phonetic Group Experts for Improved Speech Recognition," in *IEEE Transactions in Ausio, Speech and Language Processing*, March 2007, vol. 15, pp. 803–812.

[11] T. N. Sainath and T. J. Hazen, "A Sinusoidal Model Approach to Acoustic Landmark Detection and Segmentation for Robust Segment-Based Speech Recognition," in *Proc. ICASSP*, 2006.

[12] A. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The Noisex-92 Study on the Effect of Additive Noise on Automatic Speech Recognition," Tech. Rep., DRA Speech Research Unit, 1992.

[13] G. Box and D. Cox, "An Analysis of Transformations," *Journal of Royal Statistical Society*, vol. B, no. 26, 1964.