EFFICIENT NORMALIZATION BASED UPON GPD

Eric Woudenberg† Alain Biem◇ Erik McDermott◇ Shigeru Katagiri*
†ATR International ◇ATR Human Information Processing Res Labs *ATR Interpreting Telecommunications Res Labs
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02

ABSTRACT

In this paper we propose a simple but powerful method for normalizing various sources of mismatch between training and testing conditions in speech recognizers, based on a recent training methodology called the Generalized Probabilistic Descent method (GPD). In this new framework, a gradient based method is used to adapt parameters of the feature extraction process in order to minimize distortion between new speech data and existing classifier models, while most conventional normalization/adaptation methods attempt to adapt classification parameters. GPD was proposed as a general discriminative training method for pattern recognizers such as neural networks. Up until now this has been used only for classifier design, sometimes in combination with the design of a non adaptive feature extractor. This paper, in contrast, studies the adaptive training benefits of GPD in the framework of normalizing the feature extractor to a new pattern environment. Experiments which use this technique to improve Japanese vowel classification were conducted and demonstrate the ability to reduce error rates by as much as 40%.

1. INTRODUCTION

A speech recognizer usually consists of a front-end feature extractor module and a back-end classifier module. In most cases, the feature extractor is selected based on mathematical models of speech production or hearing mechanisms, i.e., LPC and filter-bank; the classifier is designed by using statistical estimation principles such as the Maximum Likelihood (ML) estimation or the loss (risk) minimization principle.

There are many factors hindering accurate recognition of speech utterances, e.g., speaker variation, speaking style variation, and acoustic channel distortion. In particular, speaker variation has been extensively studied so as to increase speaker-independency of recognizers. Most solutions to this problem are based on the adaptation of classifier parameters. Techniques such as Maximum a Posteriori estimation (MAP) [1] and Vector Field Smoothing (VFS) [2] have been widely used for this purpose. However, since these conventional methods attempt to adjust the classifier parameters, they have often resulted in low post-adaptation recognition scores when the amount of training samples is limited. Note that adaptation of classifier parameters in principle requires class information for all possible classes, and thus usually requires many training samples.

Such adaptation as speaker adaptation and channel-distortion adaptation should be done quickly. A limited number of training samples will provide partial information about classification, and one will need to compensate for this shortage of information using some assumptions which are often heuristic. However, there is no guarantee that these assumptions are mathematically optimal, and therefore an alternative approach to the problem is obviously needed.

Recently, techniques which operate through the adaptation of feature extraction parameters have received attention (for example [3] [5] [4]). In this paper, we propose a simple but quite powerful solution based on the recent training methodology called the Generalized Probabilistic Descent method (GPD). In this new framework, parameters of the feature extraction process are optimized with GPD in order to minimize the mismatch between an adaptation token and the existing classifier models.

2. GPD-BASED FEATURE NORMALIZATION

2.1. Concept - adapting filterbank centers, bandwidths and gains to minimize distortion

The central idea in this paper is the adaptation of centers, bandwidths and gains of a filterbank based feature extractor through gradient-based minimization of the distortion between adaptation speech and existing classifier models. For clarity, we refer to the proposed method as GPD-based Feature Normalization (GPDFN).

Let us assume that a speech recognizer consists of a feature extractor parameterized by Φ, and a classifier parameterized by Λ. Conventionally, the parameter set Λ is adapted, as cited above, so that the recognizer can cope with new environments that have not appeared in the training stage of the recognizer; the parameters Φ of the feature extraction module are fixed. In contrast, GPDFN attempts to directly reduce mismatch (distortion) between incoming speech patterns X and classifier parameters Λ by learning a new state of Φ through probabilistic descent of the distortion measure. This work can be viewed as an extended, recognition-oriented formalization of Adaptive Filtering. Most Adaptive Filtering uses Least Mean Square Method (LMS) for optimization, but GPDFN uses the more general stochastic descent framework of GPD for optimizing an overall distortion defined over the entire computational process including nonlinear time alignment. GPDFN also has a close relation with Discriminative Feature Extraction (DFE) [6] [7], but whereas DFE attempts to learn Φ to minimize classification error during the initial training phase, GPDFN performs the role of normalization, where the learning of Φ attempts to make the pre-designed feature extractor more suited to new environments or data.

2.2. Formalization

The formalization of GPDFN is fundamentally the same as that of the original GPD and DFE. However, since GPDFN is aimed at adaptation and normalization and adjusts only the feature extractor parameter Φ and not the classifier parameters Λ, whereas DFE optimizes Λ and Φ together to optimize overall classifier performance, the definition of the loss function is different.

There can be two implementation approaches to GPDFN: supervised training and unsupervised training. In the supervised training mode, the classes of samples given for normalization/adaptation are known, and therefore training can be performed using the distortion defined over the sample and its corresponding correct class model. In the unsupervised mode, the class of the normalization sample is unknown, entailing that the training must automatically select some reasonable class model.

The supervised training of GPDFN is formalized as follows. Assume that a sample X of class k is given for normalization. Since the class index is now known, we measure the distance of X with respect to the model for class k, i.e., $g_k(X;\Lambda,\Phi)$. Different from GPD and DFE, this distance itself is used as a distortion that must be minimized through training, i.e. $\ell_k(X;\Lambda,\Phi) = g_k(X;\Lambda,\Phi)$.

As shown in the probabilistic descent theorem [8, 9], the following adjustment

$$\Phi[\tau+1] \quad = \quad \Phi[\tau] - \epsilon_\tau \mathbf{U} \frac{\partial \ell(X,\Lambda,\Phi)}{\partial \Phi} \tag{1}$$

will lead to at least a local minimum of $\mathcal{L}(\Phi) = E_X\left[\ell_k(X;\Lambda,\Phi)\right]$; here $\mathbf{U}$ is a positive-definite matrix, ϵ_τ is a small, monotonically-decreasing, positive number, and $\Phi[\tau]$ denotes the status of Φ at training iteration τ. A practical approach is to use the given samples repeatedly; i.e., similar to the motivation of Adaptive Filtering, GPDFN may adapt the feature extractor to at least a locally optimal situation, guided by the available samples.

In the unsupervised mode, the most likely model may be selected in computing the distortion. However, since there is no guarantee that the model selection is correct, the normalization may be less effective than in supervised mode.

2.3. Implementation for filter-bank feature extractor

In our work we use a filterbank model of feature extraction in which parameters such as the channel centers, bandwidths, and gains are normalized.

For computational simplicity, we simulate the filterbank model with DFT techniques. This offers a fast alternative to FIR or IIR-based banks-of-filters. Thus, for a sequence of speech vectors $X = \{\mathbf{x}_1,\ldots,\mathbf{x}_t,\ldots,\mathbf{x}_\mathcal{T}\}$ in which $\mathbf{x}_t = [x_{t,1},\ldots,x_{t,f},\ldots,x_{t,F}]^T$ is the magnitude spectrum of the frame (short time window position); $x_{t,f}$ represents the f-th element of the frame vector. F is the maximum frequency. A Q-channel filter bank model transforms each $\mathbf{x}_t$ into a lower dimensional vector $\mathbf{e}_t = [e_{t,1},\ldots,e_{t,c},\ldots,e_{t,Q}]^T$ such that an output feature $e_{t,c}$ is the windowed log energy of the c-th channel:

$$e_{t,c} = \log_{10}\left(\sum_{f\in B_c} \theta_c(f) x_{t,f}\right), \quad \text{for } c = 1,\ldots,Q, \tag{2}$$

where B_c represents the channel interval and $\theta_c(f)$ the weighting at frequency f provided the c-th filter.

For practical accommodation of gradient-based optimization, we employ a filterbank model consisting of Gaussian-shaped filters defined as:

$$\theta_c(f) = \varphi_c \exp\left(-\beta_c\left(p(\gamma_c) - p(f)\right)^2\right), \tag{3}$$

for $c = 1,\ldots,Q$, where the trainable parameters $\beta_c > 0$ and γ_c determine bandwidth and center frequency, and φ_c is the trainable "gain" parameter in the c-th channel. $p(f)$ maps the linear frequency f onto the perceptual representation. An important feature of this type of filter is related to the fact that it allows a straightforward adjustment of center frequency, bandwidth or gain via GPD while preserving the characteristics of triangular filters.

2.4. Normalization

Normalization is performed as follows: let $\ell(X,\lambda,\Phi)$ be the distortion of the input spectrum X to the model represented by λ, given the feature extractor parameters Φ, and let ϕ be any parameter of the filterbank (e.g. a filter center, gain, or bandwidth). Since ϕ has a physical meaning, we ensure that its value remains positive by using the transformation

$$\phi \quad = \quad \exp(\overline{\phi}) \tag{4}$$

We then perform the following adjustment:

$$\overline{\phi}[\tau+1] \quad = \quad \overline{\phi}[\tau] - \rho_\tau \mathbf{U2} \delta\overline{\phi} \tag{5}$$

where

$$\delta\overline{\phi} \quad = \quad \frac{\partial \ell(X,\lambda,\Phi)}{\partial \overline{\phi}} \tag{6}$$

$$= \quad \sum_{t=1}^{\mathcal{T}} \sum_{c=1}^{Q} \frac{\partial \ell(X,\lambda,\Phi)}{\partial e_{t,c}} \frac{\partial e_{t,c}}{\partial \overline{\phi}} \tag{7}$$

2.5. Gradient Calculation

The key steps for filterbank optimization are the calculation of $\frac{\partial e_{t,c}}{\partial \overline{\phi}}$, for each channel c. The general chain rule is

$$\frac{\partial e_{t,c}}{\partial \overline{\phi}} = \sum_{f=1}^{F} \frac{\partial e_{t,c}}{\partial \theta_c(f)} \frac{\partial \theta_c(f)}{\partial \overline{\phi}}. \tag{8}$$

This signifies that, first, one needs to compute the derivative $\frac{\partial e_{t,c}}{\partial \theta_c(f)}$ for each frequency f, before expanding to $\frac{\partial \theta_c(f)}{\partial \overline{\phi}}$ according to the type of parameter $\overline{\phi}$. According to (2),

$$\frac{\partial e_{t,c}}{\partial \theta_c(f)} \quad = \quad \frac{x_{t,f}}{\log(10)\mathcal{E}_c(\mathbf{x}_t)} \tag{9}$$

where

$$\mathcal{E}_c(\mathbf{x}_t) \quad = \quad \sum_{f\in B_c} \theta_c(f) x_{t,f}$$

$$= \quad \exp_{10}(e_{t,c})$$

is the output energy of frame $\mathbf{x}_t$ in the c-th channel (without the log transformation).

training speaker	testing speaker	before adaptation	after adaptation
male	female	80%	100%
male	child	80%	90%
female	male	90%	100%
female	child	90%	100%
child	male	50%	50%
child	female	60%	70%

Table 1. Recognition accuracies for simple cross speaker adaptation experiment

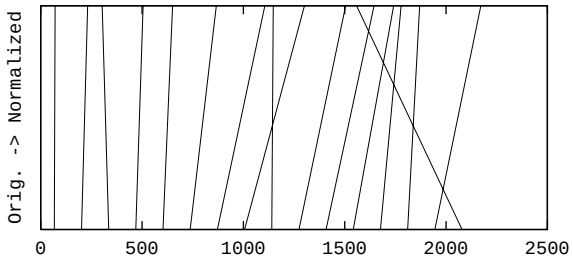

Figure 1. Mel center frequency shift adapting female speech to male speech models

Now, let us expand $\frac{\partial\theta_c(f)}{\partial\overline{\phi}}$ for the case when $\overline{\phi}$ represents the center frequency of a channel $\hat{c}$. Thus $\phi = \tilde{\gamma_{\hat{c}}} = p(\gamma_{\hat{c}})$, where $\tilde{\gamma_{\hat{c}}}$ represents the center frequency of channel $\hat{c}$ in the perceptual domain. For $\tilde{\gamma_{\hat{c}}} = \exp(\overline{\tilde{\gamma_{\hat{c}}}})$ and from (3), it follows that

$$\frac{\partial\theta_c(f)}{\partial\overline{\tilde{\gamma_{\hat{c}}}}} = -2\beta_c\,(\tilde{\gamma_c} - p(f))\,\theta_c(f)\,\tilde{\gamma_c}\,\chi(c,\hat{c}). \quad (10)$$

Derivation of the gradient for gain and bandwidth parameters is also straightforward and will not be elaborated here.

3. EXPERIMENTS

3.1. Cross-gender and cross-age task

As an initial test of GPDFN a simple cross speaker vowel recognition experiment was conducted using a Prototype Based Minimum Error Classifier (PBMEC) [10]. Four repetitions of the 5 Japanese vowels /a/ /i/ /u/ /e/ and /o/ were collected by telephone from a male, female and child speaker. The data was parameterized to feature vectors of 10 Mel-scale log filterbank energies and then used to make a set of simple (1 state, 1 "mixture") speaker specific models for the 5 phonemes plus /pau/ (silence). The experiment tested the recognition accuracy of a testing speaker's speech on a training speaker's models, before and after adaptation. The test used one half of the testing speaker's data for adaptation and the other half for testing. The parameters adapted were the filterbank centers. The results are shown in Tab.1.

Figure 1 illustrates the effect of adaptation, in this case the shifting of Mel frequency centers which occurred during the adaptation of female speech for minimum distortion on the male speech models. The frequency cross overs which occurred reflect the probabilistic nature of the adaptation.

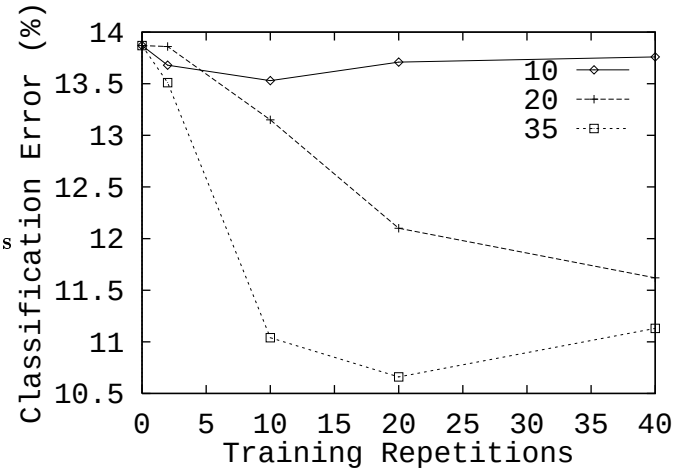

Figure 2. Vowel Classification Error (male)

3.2. Vowel Classification Task

To more thoroughly evaluate the effect of GPDFN on error rates, a larger experiment was conducted on the ATR isolated word database.

Our training data contained a total of 30000 vowel segments from 2 male and 2 female speakers. Speech was parameterized to 24 filterbank log energy coefficients, with a 20ms frame window and 5ms frame shift.

Testing data consisted of 10000 vowel segments from 1 male and 1 female speaker not in the training set.

Experiments involved presenting varying numbers of randomly selected adaptation vowels (10, 20, and 35 tokens), with several different training repetition counts (2, 10, 20 and 40 epochs) and adapting filterbank center frequencies to minimize distortion. The adapted filterbank parameters were then used to reclassifying the 10000 vowels to determine post adaptation accuracy.

Figures 2 and 3 show the effect the number of presentation tokens and presentation epochs have on classification error on testing data. Provided that sufficient training tokens are available to avoid undergeneralization, results show a relative decrease in error rate of up to 42% for the female (17.7% to 10.2% with 40 epochs of 35 tokens) and 23% for the male (13.9% to 10.7% with 20 epochs of 35 tokens). Overtraining from high repetitions of the training tokens does not appear to be a factor until 20 or more epochs are presented.

3.3. Phoneme Recognition Task

In order to evaluate the GPD adaptation technique on a more realistic task, we conducted a set of phoneme recognition experiments on the ATR 5240 Japanese word database. Multispeaker phoneme models were trained from the speech of 5 speakers from the database, totaling 26200 utterances. These models were 3 state 5 mixture HMMs with diagonal covariances matrices. Input speech was sampled at 12kHz and parameterized as 12 Mel-scale log filter bank energy coefficients. In order to maintain consistency in the adaptation framework, the HMM models were not further refined with Minimum Error training.

The adaptation phase consisted of a stochastic gradient based search for the set of filterbank centers, bandwidths and gains that would yield the smallest distortion when adaptation tokens from an unknown speaker were

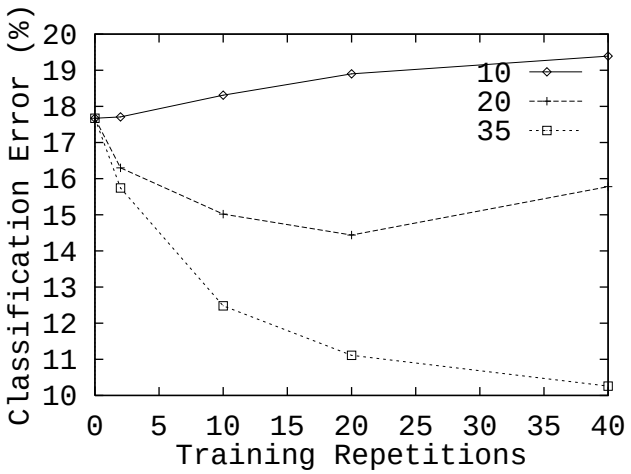

Figure 3. Vowel Classification Error (female)

DP aligned with their correct transcriptions. In this supervised adaptation mode, varying numbers of tokens were presented, and the feature extraction parameters found were then tested for generalization in a phoneme recognition task using 1000 utterances from the unknown speaker. In order to maximize the use of our data, the experiment was done in 6 jackknife runs, each time omitting a different speaker from the training set, whose speech would then be used for adaptation testing.

For comparative purposes, a frequency warping technique was also evaluated. For this, an 18 point grid search over a range of warping values (0.88 to 1.22) was conducted for each set of adaptation tokens from the unknown speaker. The warping value yielding the best phoneme accuracy on the adaptation tokens was then tested for generalization on the multi-speaker models, as with the GPD adapted parameters above. A frequency warping based grid search was also conducted over the entire set of 1000 testing utterances. This enabled us to see the best achievable recognition accuracy over the testing set by the warping-based normalization. Tab.2 shows phoneme recognition accuracies on the 1000 test utterances as a function of the number of adaptation tokens, both for warp and GPD based adaptation, and also shows (in the right most column) the best achievable, grid-search-based warping score over the 1000 test utterance set. The results in the table clearly demonstrate that the GPD-based normalization efficiently improved the given feature extractor only using a limited number of adaptation tokens. For some speakers, such as MAU, the normalization achieved a remarkable improvement, from about 42% to about 55%.

4. CONCLUSION

We have described a new approach to improving classification performance by normalizing sources of distortion in speech recognition systems, via GPD adaptation of front end feature extractor parameters. Experimental results in the vowel and phoneme recognition tasks successfully demonstrated the utility of the proposed method.

The evaluations done in this paper are still preliminary. Further careful investigations of training settings would increase stability and reduce the variation among speakers in the experimental results. Also, the proposed normalization idea is general enough to apply to a more powerful extraction framework than the filter-bank. This point can be an important future research issue.

adapt. speaker	adapt. method	no adapt.	10 tokens	80 tokens	best warp
MAU	warp	41.97	43.64	44.88	45.20
MAU	gpd	41.97	55.36	56.86	
MXM	warp	60.55	60.63	60.26	60.64
MXM	gpd	60.55	62.88	63.23	
MHT	warp	50.70	52.37	54.65	54.69
MHT	gpd	50.70	52.43	57.00	
FMS	warp	57.31	55.75	58.06	58.46
FMS	gpd	57.31	57.65	58.15	
FTK	warp	46.10	46.10	46.10	46.66
FTK	gpd	46.10	48.89	46.21	

Table 2. Phoneme recognition accuracies on 1000 test utterances

The warping based normalization is simple and easy to implement, but its capability is in principle limited in comparison to a more general framework of GPD based adaptation of feature extractor parameters. This is demonstrated in our experimental results.

REFERENCES

[1] Lee, C.-H., and Gauvain, J.-L. (1993). *Speaker Adaptation Based on MAP Estimation of HMM Parameters* Proc. ICASSP93, Vol. 2, pp. 558-561.

[2] Kosaka, T. , Matsunaga, S. , and Sagayama, S. (1995). *Tree-Structured Speaker Clustering for Speaker Adaptation* IEICE Trans. D-II, Vol. J78-D-II, No. 1, pp. 1-9.

[3] Sankar, A., Neumeyer, L., and Weintraub, M. (1995). *An Experimental Study of Acoustic Adaptation Algorithms* Proc. ICASSP96, Vol. 2, pp. 713-716.

[4] Sankar, A., and Lee, C.-H. (1996). *A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition* IEEE Trans. on Speech and Audio Proc. Vol. 4, No. 3, May 1996, pp. 190-202.

[5] Lee, L., and Rose, R., (1995). *Speaker Normalization using Efficient Frequency Warping Procedures* Proc. ICASSP96, Vol. 1, pp. 353-356.

[6] Biem, A., Katagiri, S. and Juang, B.-H. (1993). *Discriminative Feature Extraction for Speech Recognition* IEEE, NNSP III.

[7] Biem, A., McDermott, E., and Katagiri, S. (1995). *A Discriminative Filter Bank Model for Speech Recognition* Proc. Eurospeech95, Vol. 1, pp. 545-548.

[8] Katagiri, S., Lee, C.-H. and Juang, B.-H. (1990). *A Generalized Probabilistic Descent Method.* Proceedings of the of Acoustical Society of Japan, Fall Meeting, pp. 141-142.

[9] Katagiri, S., Lee, C.-H. and Juang, B.-H. (1991). *New Discriminative Training Algorithms Based on the Generalized Descent Method.* Proceedings of the 1991 IEEE Workshop on Neural Networks for Signal Processing, August 1991, pp 1-10.

[10] McDermott, E. and Katagiri, S. (1994). *Prototype Based Discriminative Training for Various Speech Units.* Computer Speech and Language, volume 8, pp. 351-368.