PREDICTIVE ADAPTATION AND COMPENSATION FOR ROBUST SPEECH RECOGNITION

Arun C. Surendran and Chin-Hui Lee

Bell Labs, Lucent Technologies, 600 Mountain Ave, Murray Hill NJ 07974.
Phone: (908) 582-6351 Fax: 582-7308, email: {acs,chl}@research.bell-labs.com

ABSTRACT

Earlier work in parametric modeling of distortions for robust speech recognition has focussed on estimating the distortion parameter using maximum likelihood and other techniques as a point in the parameter space, and treating this estimate as if it is the *true* value in a plug-in maximum *a posteriori* (MAP) decoder. This approach is deficient in most real environments where, due to many reasons, the value of the distortion parameter varies significantly. In this paper we introduce an approach which combines the power of parametric transformation and *Bayesian prediction* to solve this problem. Instead of approximating the distortion parameter with a point estimate, we average over its variation, thus taking into consideration the distribution of the parameter as well. This approach provides more robust performance than the conventional maximum-likelihood approach. It also provides the solution that minimizes the overall error given the distribution of the parameter. We present results to demonstrate the robustness and effectiveness of the predictive approach.

1. INTRODUCTION

In many speech recognition applications, the distorting mechanism that degrades performance varies significantly, even during a single utterance (e.g. wireless channel). Also while modeling such distortions, the choice of model is often inaccurate (e.g. modeling a varying channel distortion as a single bias in the cepstral domain). In these situations the traditional compensation schemes, where the distortion is modeled as a nuisance parameter and then estimated as a single point in the parameter space, do not adequately capture the characteristics of the distortion. The usual maximum likelihood estimators (MLEs) used in such techniques are only asymptotically consistent (i.e. the variance of the estimates match the actual variance only as the amount of data used for estimation is unlimited), hence with limited amount of data, such estimates are not robust. These techniques treat the *estimated* parameter as it's *actual* value and use it in a plug-in maximum *a posteriori* (MAP) decoder without considering the modeling errors or the uncertainty in the estimation.

2. BAYESIAN PREDICTION

In this paper, we solve this very important problem using *predictive compensation*, which is a Bayesian solution to the point estimation problem. It is a novel combination of two powerful approaches in the area of robustness - (1) the parametric transformation approach [5, 4] and (2) Bayesian prediction [1, 2]. If Λ_Y, the model of the test data $\mathbf{Y}$ is related to the trained model Λ_X through a functional distortion $\Lambda_Y = F_\theta(\Lambda_X)$, then $P(\mathbf{Y}|\theta, \Lambda_X)$ is the conditional distribution of the data given the parameters of the transformation θ, where $\theta \in \Theta$, the parameter space. The form of this conditional distribution depends upon the characteristics of the the functional transformation. If θ has a prior distribution $P(\theta)$, then instead of estimating θ, we smooth out the conditional distribution averaging over the uncertainty of the parameter to obtain the predictive distribution:

$$P(\mathbf{Y}|\Lambda_X) = \int_{\theta\in\Theta} P(\mathbf{Y}|\theta, \Lambda_X)P(\theta)d\theta. \quad (1)$$

Since $P(\mathbf{Y}|\Lambda_X)$ is the *actual* distribution of the observed values of $\mathbf{Y}$, it can "predict" what values $\mathbf{Y}$ can take and hence is called the predictive distribution [1]. It can be shown that this approach is more robust, and it minimizes the overall error given the prior distribution [2]. A predictive approach was recently proposed for robust classification in [3].

Usually, if the prior has a sharp peak, i.e., if the testing environment is not widely varying, and when reasonable amount of data is available to obtain the ML estimate accurately, there is not much difference in the performance between the predictive and ML approaches since the final predictive distribution can be approximated in terms of the conditional distribution calculated at the ML estimate: $\int_{\theta\in\Theta} P(\mathbf{Y}|\theta, \Lambda_X)P(\theta)d\theta \approx \beta P(\mathbf{Y}|\theta_{ML}, \Lambda_X)$. We shall see from our experiments that the advantage of the predictive distribution is apparent when there is usually a wide difference between the training and testing conditions, and that the performance of the predictive approach is good even in the presence of small amounts of data. This technique can be used for both adaptation (i.e., when limited amount of data from the testing environment is available in advance) or compensation (i.e., when the process is performed using the testing data only).

The steps in our predictive approach are (1) determine the functional form of the prior, (2) estimate the hyper-parameters, (3) compute the predictive density, and (4) use it in a plug-in MAP decoder to compute the word sequence.

3. DETERMINATION OF PRIOR DENSITY

An important element of the predictive approach is the prior probability distribution over Θ, the parameter space. Such a prior should (1) have a functional form such that the integral in Equation 1 can either have a closed form solution or have a reasonably good approximation, and (2) be consistent with the data and the models and adequately quantify the prior information concerning the distortion mechanism.

The shape of the prior (and the subsequent calculation of the predictive distribution) is determined by the choice of the functional transformation $F_\theta(\cdot)$. The form of $F_\theta(\cdot)$ is chosen by examining the data. In our paper we start with a very simple functional

model of the distortion which transforms the means of the model using an additive bias: $\mu_Y = \mu_X + \theta$, where X is the undistorted data and μ represents the mean. If X and θ are independent processes and X is modeled by a hidden Markov Model (HMM) with Gaussian mixture states such that the distribution of the data given the state s_t and the mixture c_t is $P(x_t|s_t = s, c_t = c) = \mathcal{N}(\mu_{sc}, \sigma^2_{sc})$, then the conditional distribution of the test data is $P(y_t|s_t = s, c_t = c, \theta) = \mathcal{N}(\mu_{sc} + \theta, \sigma^2_{sc})$. The parameters of the distribution of θ are called the *hyper-parameters* of the prediction.

One way to tackle the problem of prior selection is to use an empirical Bayes approach, where the priors are computed from the available adaptation data. First we assume that each component of θ is independent of each other, i.e. the distortion affects each cepstral coefficient of the data independently. Then the prior for each component of the D dimensional vector is computed separately, and then the final prior is calculated as the product of the marginal priors: $P(\theta) = \Pi_{i=1}^{D} P(\theta_i)$. Here, the parameter space is $\Theta \subset \mathcal{R}^D$. Through the rest of the paper, we will talk of θ as though it is a single parameter without the loss of generality.

Prior Computed Directly from Stereo Data: To study the nature of the distortion and to obtain preliminary information about the shape of the prior, we used a database where utterances were recorded simultaneously over a close talking microphone ("clean" data) and over a dial-up telephone line ("noisy" data). Such a database is called a "stereo" database. Since such databases are impractical in real applications, this procedure was intended as a tool for understanding the nature of the prior.

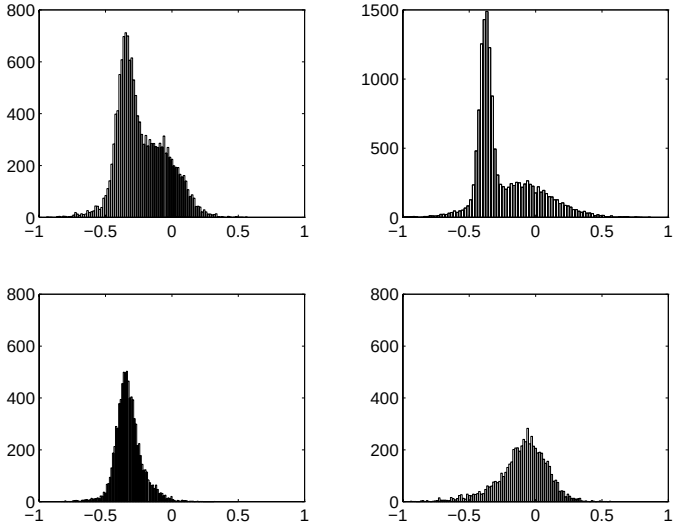

Figure 1: Histograms of bias computed from 20 sentences using (a) stereo data (b) without stereo data (c) for silence segments only (d) for speech segments only.

Consider the functional transformation $Y = X + Z$. Given a stereo database with data $((x_1, y_1), ..., (x_T, y_T))$, we can see that the limit of the estimated mean of Z is

$$\lim_{T\to\infty} \tilde{z} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} (y_t - x_t) = \bar{\mu_Y} - \bar{\mu_X} = \mu_\theta. \qquad (2)$$

Similarly

$$\lim_{T\to\infty} \tilde{\sigma_Z^2} = \sigma_\theta^2. \qquad (3)$$

Hence Z can be a reasonably good estimator of θ and given sufficient data, the histogram of $z_t = (y_t - x_t),\ t = 1, ..., T$ should closely approximate its prior. From the histogram we can hypothesize a parametric functional form of the prior and estimate the hyper-parameters. Figure 1(a) shows this histogram for the first coefficient estimated using 20 sentences spoken over a telephone line. We can determine that the histogram is bimodal, and this bimodal nature can be captured by constructing priors for speech and silence segments separately (Figures 1(c) and (d)). The two priors are asymmetric Laplacians ($\mathcal{L}_{\alpha_1\alpha_2}(\theta) = \frac{\alpha_1\alpha_2}{\alpha_1+\alpha_2}\{\exp(-\alpha_1|\theta-m|)u(-(\theta-m)) + \exp(-\alpha_2|\theta-m|)u(\theta-m)\},\ \alpha_1, \alpha_2 > 0$, where m is the mode of the distribution and $\mu_\theta = m + \frac{1}{\alpha 2} - \frac{1}{\alpha 1}$ is the mean.

If the conditional distribution is a Gaussian $\mathcal{N}(\mu_X + \theta, \sigma_X^2)$ (or a mixture of Gaussians) and the prior is an asymmetric Laplacian given before, then we can show that the predictive distribution is also an asymmetric Laplacian (or a mixture of asymmetric Laplacians) with each component having a mode at μ_X, and the mean at $\mu_X + \mu_\theta$. Even though such flat-tailed priors are considered robust [1], there are many problems with this approach - (1) when the mode of the predictive distribution is not equal to its mean, the MAP decoder runs into problems, and (2) we know that this marginal does not match the distribution of our speech data. To step around these problems, we can choose the shape of the prior such that the generated marginal is consistent with our prior knowledge of the data. We discuss this issue next.

3.1. Using Marginal Distribution to Determine Priors: Minimum Divergence Approximation

One common way to determine the functional form of the prior is from our knowledge of the marginal distribution. Since we know that a hidden Markov model with Gaussian mixture distribution is a good model of speech for both the marginal *and* the conditional distributions, we can chose the conjugate prior of the conditional distribution to model our prior. In our case, that would be a normal distribution. We can determine the hyper-parameters of the normal distribution such that the *Kullback-Leibler information divergence* [2], a standard measure of dissimilarity between distributions, is minimized between the Gaussian and the asymmetric Laplacian. This is also called the maximum-entropy solution [1]. The divergence between two distributions $\mathcal{L}_{\alpha_1,\alpha_2}(\theta)$ and $\mathcal{N}(\mu, \sigma^2)$ is defined as $E\{\log(\mathcal{L}/\mathcal{N})\}$ where $E\{\}$ is the expectation taken over the distribution $\mathcal{L}$. After eliminating terms not containing μ and σ, the divergence is equal to:

$$D(\mathcal{L}||\mathcal{N}) = -E\{\log\sigma\} - E\{\frac{(\theta-\mu)^2}{2\sigma^2}\}. \qquad (4)$$

Setting the derivatives to zero provides

$$\frac{\partial D}{\partial \mu} = 0 \quad \Rightarrow \quad \mu = E\{\theta\} = \mu_\theta \text{ and,} \qquad (5)$$

$$\frac{\partial D}{\partial \sigma} = 0 \quad \Rightarrow \quad \sigma^2 = E\{(\theta-\mu)^2\} = \sigma_\theta^2. \qquad (6)$$

Thus the normal distribution is chosen such that its mean and variance are equal to the mean and variance computed from the asymmetric Laplacian mentioned before.

The choice of the conjugate prior makes the computation in Equation 1 quite simple. (It is considered that the choice of conjugate prior is not necessarily robust. But if the likelihood is concentrated in the center portion of the prior, then the use of the natural conjugate will be reasonably robust [1].) If the conditional distribution is given as $\mathcal{N}(\mu_X + \theta, \sigma_X^2)$ and the prior is $\mathcal{N}(\mu_\theta, \sigma_\theta^2)$, then the marginals are computed to be $\mathcal{N}(\mu_X + \mu_\theta, \sigma_X^2 + \sigma_\theta^2)$. The hyper-parameters μ_θ and σ_θ^2 can be determined from the data as $\mu_\theta = \frac{1}{T}\sum_{t=1}^{T} z_t$ and $\sigma_\theta^2 = \frac{1}{T-1}\sum_{t=1}^{T}(z_t - \mu_\theta)^2$.

3.2. Unsupervised Approach to Hyper-parameter Computation

Computing the prior from the stereo data is not robust and highly impractical. If $X = (x_1, .., x_T)$ and $Y = (y_1, ..., y_T)$ are modeled by HMMs, the ideal alternative would be to use the sequence of means corresponding to each mixture component that generated the data i.e. $(\mu_{X_1}, ..., \mu_{X_T})$ and $(\mu_{Y_1}, ..., \mu_{Y_T})$. Unfortunately such a sequence is not available to us because we neither know the distribution of Y nor have access to the training data X. But if we are given Λ_X then we can estimate the *most likely generator sequence* $\{\mu_{\tilde{X}_t}\}$ $t = 1, ..., T$ by performing a forced alignment using the MAP decoder and the transcriptions estimated from the recognition output. We can now define $Z = Y - \mu_{\tilde{X}}$ and we can use the earlier method to plot the histogram. This method should give a more robust estimate of the prior provided that the mismatch between the models and the data is not too severe. Figure 1(b) shows the histogram computed using the telephone data and models adapted on utterances spoken through a close talking microphone. If separate adaptation data is available, then such a prior can be computed in a supervised environment as opposed to the unsupervised environment described above.

ML-II priors: When the hyper-parameters are chosen such that value of the marginal (i.e., the likelihood of the data) is maximized, this choice of prior is called the maximum-likelihood or ML-II prior [1, 2]. A very similar approach is discussed in a much different context using the same database in [4], and hence will not be discussed here.

4. EXPERIMENTAL RESULTS

Sentences from the 991-word DARPA resource management (RM) task were recorded simultaneously through two channels: (1) A close talking microphone, and (2) a telephone handset over a dial-up line. The sentences were spoken by a non-native speaker. The data consisted of 300 sentences for adaptation and 75 sentences for testing. 1769 context dependent (CD) subword unit models were built, with a maximum of 16 mixtures per state. The RM word pair grammar which gives a perplexity of about 60 was used for the experiments. Starting from a set of gender specific HMMs built on male speakers (Λ_{SIM}) and using the 300 sentences recorded over a close-talking microphone by the non-native speaker, a speaker-specific model (Λ_{MIC}) was generated using MAP adaptation [6]. The 300 sentences recorded over the telephone channel were used as adaptation data. 75 sentences were recorded separately over the telephone line (TEL) for testing. A 38-dimensional feature vector (with 12 LPC-derived cepstral vectors, 12 Δ and 12 $\Delta\Delta$ coefficients, Δ and $\Delta\Delta$ energy) was used for recognition. Only the 12 cepstral coefficients were used for prior determination in this paper.

Recognition results are provided for two mismatch cases: (1) Λ_{SIM} models used with TEL data.(2) Λ_{MIC} tested on TEL data. In each of these cases, either a single prior (or a single ML bias value) is computed for the entire data (this case is referred to as S0) or separate calculations are done for speech and silence segments (this case is indicated as S1).

Tables 1 and 2 show results for predictive compensation i.e., when only the test data and the trained models are available and the transcriptions are unknown. One sentence is used for compensation in all the following tables.

Λ_{SIM}-TEL Baseline: 36.7%		
	Unsupervised	
	Predictive	ML
S0	54.4	45.9
S1	54.4	46.2

Table 1: Performance when SIM models and TEL data were used for compensation

Λ_{MIC}-TEL Baseline: 79.6%		
	Unsupervised	
	Predictive	ML
S0	90.4	87.4
S1	90.4	92.5

Table 2: Performance when MIC models and TEL data were used for compensation.

Table 1 shows the results when the model built on male speakers is tested on the telephone data. The mismatch between the training and testing environments is due to differences in speaker, channel and microphone, and is quite severe. The baseline word accuracy is 36.7%. Using the simple bias transformation the predictive approach gives a 28% reduction in error rate whereas the ML estimation only gives a 15% reduction in error rate. The S1 case gives similar results. These results show how predictive compensation is robust even under severe conditions.

The Λ_{MIC} model gives a baseline word accuracy of 79.6% on the TEL data (Table 2). Since the speaker mismatch has been removed using MAP adaptation, only microphone and channel mismatches remain. The predictive compensation gives a 53% reduction in error rate as compared to a 38% reduction in error rate for the ML estimation in the S0 case. When the ML estimator is allowed different estimates for speech and silence segments (S1), it does better than the predictive approach. This was expected from our analysis before - if the prior has a sharp peak and a clear mode, then the accuracy of the ML estimate is high and it is capable of doing well. Also, in the Λ_{MIC}-TEL case, the segmentation accuracy is already quite reliable thus giving better ML

estimates. The sharp peak in the prior distribution and the presence of a clear mode for this case is apparent from Figure 1(c) and (d). The advantage of the predictive approach is clear from Table 1 where the ML estimates are not reliable. Further, predictive compensation does not require iterative calculations that are common in ML approaches [5, 4] and hence is computationally less intensive and simpler to implement.

As mentioned before, the predictive approach can be used for adaptation also. When transcriptions of the utterances are known, a supervised computation of the prior is possible for *predictive adaptation*. In Tables 3 and 4, one sentence from the 300 sentences of the adaptation data was used for prior computation.

Λ_{SIM}-TEL Baseline: 36.7%		
	Supervised	
	Predictive	ML
S0	58.1	43.1
S1	66.7	65.2

Table 3: Performance when SIM models and TEL data were used for adaptation

Λ_{MIC}-TEL Baseline: 79.6%		
	Supervised	
	Predictive	ML
S0	90.8	87.7
S1	92.2	92.3

Table 4: Performance when MIC models and TEL data were used for adaptation.

Tables 3 and 4 show the results for predictive adaptation. Since the adaptation data in this case is a good representative of the test data, the results are similar to those for compensation. The exception is the performance for the S1-ML case in Table 3. The difference between this result (65.2%) and the corresponding result for unsupervised learning in Table 1 (46.2%) shows how sensitive the ML approach is to the accuracy in initial segmentation. The predictive approach does not exhibit this high degree of dependence, once again demonstrating its robustness.

Figure 2 shows further proof of the dependence of the ML approach on the availability of data and the relative insensitivity of the predictive approach. The improvement in word accuracy with increase in adaptation data (from 1 to 50 sentences taken from the 300 adaptation sentences) for unsupervised prior computation using Λ_{SIM} models and TEL data is plotted in Figure 2. The performance curve of the predictive approach is flatter than the ML technique showing that even when using small amounts of data, the predictive approach performs better and is more consistent.

5. SUMMARY

In this paper we have introduced a new Bayesian robustness paradigm that combines the power of functional transformation with predictive techniques. In this technique, instead of estimating the parameter of the distortion, we integrate over the prior distribution of the parameter, thus compensating for its uncertainty.

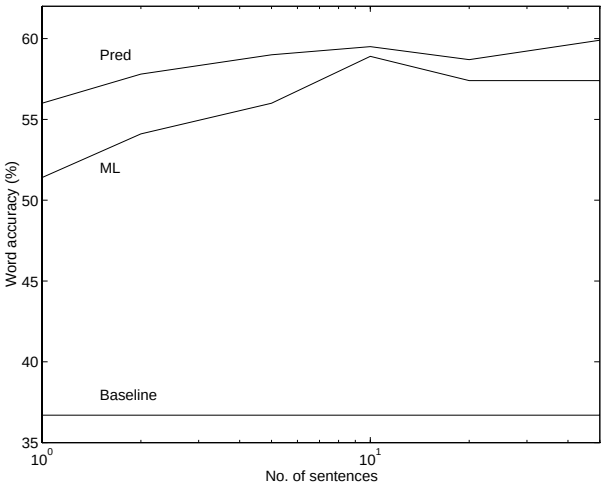

Figure 2: Word accuracy (%) vs. number of sentences for SIM model and TEL data for S1 case

We estimate the prior and the hyper-parameters using empirical Bayes approaches and approximate it using suitable distributions. Predictive compensation is more robust, simpler and faster and needs lesser amount of data than the corresponding maximum-likelihood approach. We have demonstrated the effectiveness of this approach for both adaptation and compensation under different mismatch conditions.

6. ACKNOWLEDGMENTS

The authors thank Dr. Frank Soong for his suggestions and discussions.

7. REFERENCES

1. J. O. Berger, *Statistical Decision Theory and Bayesian Analysis, 2nd Ed.*, Springer-Verlag New York Inc.
2. B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, 1996.
3. Q. Huo, H. Jiang, and C.-H. Lee, "A Bayesian Predictive Classification Approach to Robust Speech Recognition", Proc. ICASSP-97, pp. II-1547-1550.
4. A. Sankar and C.-H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition", IEEE Trans. on Speech and Audio Proc., Vol. 4, No. 3, pp.190-202, May 1996.
5. A. C. Surendran, "Maximum Likelihood Stochastic Matching Approach to Non-linear Equalization for Robust Speech Recognition", Ph.D. Thesis, Rutgers University, May 1996.
6. C.-H. Lee, and J.-L. Gauvain, "Speaker Adaptation based on MAP Estimation of HMM Parameters", Proc. ICASSP-93, pp. II-558-561, 1993.