MODELING UNCERTAINTY OF DATA OBSERVATION

Andreas Wendemuth

Philips Research Laboratories
Man-Machine-Interfaces
Weisshausstrasse 2, 52066 Aachen, Germany
email: Andreas.Wendemuth@philips.com

presently: Otto-von-Guericke Universität
Institute for Electronics, Signal
Processing and Communications Technology
Universitätsplatz 2, 39106 Magdeburg, Germany

ABSTRACT

An approach is presented both theoretically and experimentally which overcomes a number of existing conceptual and performance problems in density estimation. The theoretical approach shows methods for incorporating or estimating uncertainties into speech recognition. In the MMI and ML case, precise formulae are given for estimation of densities for uncertainty variances small compared to the curvature of the posteriors.
For implementation, the theoretical formulae are presented in such a way that the additional computation effort goes linearly with the number of densities.
Experiments on car digits show relative improvements in word error rate of at most 4.8% relative. Uncertainty modelling is shown to help remedy effects of the sparse data problem in density estimation.

1. INTRODUCTION

Density estimation in the methodology normally used in speech recognition has some known drawbacks: [5]

- Density estimation does not follow a suitable, consistent statistical approach.
- The underlying Maximum Likelihood (ML) theory has dangers of false (too narrow) variances which at the moment are only covered by variance pooling and other ad-hoc methods.
- ML theory is ill-posed in the presence of any finite data.
- Other risk functions like Maximum Mutual Information (MMI) or Structural Risk minimization outperform ML.

Uncertainty modelling as proposed here attacks several of these problems without paradigm shifts in theory, and without heavy redesign effort. It is a compromise between non-parametrical and parametrical statistics [1].

- The problem of finite data is approached by explicit modelling of uncertainty of the observed data. This modelling can be done analytically and only results in modifications of the estimation formulae, not in data diffusion or other data-related, expensive methods.
- This modelling can take into account, if available, the **known** uncertainty in given data. This knowledge can be available externally (database quality, etc.) or internally (confidence measures).
- Having modelled uncertainty, variance narrowing cannot occur any more. The dangers of abandoning variance pooling are therefore overcome; unpooled or less pooled variances can be estimated.

This paper is organized as follows: The theory will be developed in section 2 for the case of estimating single densities in the MMI case where linear computation effort is particularly explained. Using striking similarities, in section 3 mixture density estimation under ML is considered. An extension to reestimation in Hidden Markov Models is discussed in section 4. Uncertainty variances will be treated as a *model parameter* in section 5. In speech recognition, viterbi paths and maximum approximations are treated in section 6. Experiments will be presented for the ML case in section 7, for the example of car digit recognition.

2. SINGLE DENSITY ESTIMATION UNDER THE MMI CRITERION

Explicit uncertainty modelling assumes that the data x_n is not a pointed observation (delta-function), but instead the mean of a continuous (!) distribution of observations y modeled as Gaussians with diagonal covariance matrices in an M-dimensional space:

$$p(y|x_n) = \frac{1}{\prod_{i=1}^{M}\sqrt{2\pi v_i^n}} \exp -\frac{1}{2}\sum_{i=1}^{M}\frac{(y-x_i^n)^2}{v_i^n} \qquad (1)$$

This transforms a function $f_{\mathrm{crisp}}(\{x_n\})$ into

$$f_{\mathrm{noisy}}(\{x_n\}) = \int_{-\infty}^{+\infty} dy \left(\sum_n p(y|x_n)\right) f_{\mathrm{crisp}}(y) \qquad (2)$$

Let k,c, l be classes, x data, and λ the parameters of the given model. Under MMI, we optimize for crisp data the following criterion:

$$\mathrm{MMI}_{\mathrm{crisp}} = \sum_{n=1}^{N} \ln p(k_n|x_n) \tag{3}$$

Using Bayes formula for posterior probabilities

$$p(k|x) = \frac{p(k)p(x|k)}{\sum_c p(c)p(x|c)} \tag{4}$$

The optimum is obtained at setting zero

$$\frac{\partial \mathrm{MMI}}{\partial \lambda} = \sum_{n=1}^{N} \left(\frac{\partial \ln p(x|k)}{\partial \lambda} - \sum_c p(c|x_n) \frac{\partial \ln p(x_n|c)}{\partial \lambda} \right) \tag{5}$$

Using a single Gaussian diagonal density model

$$p(x|k) = \frac{1}{\prod_{i=1}^{N} \sqrt{2\pi\sigma_i^k}} \exp -\frac{1}{2} \sum_{i=1}^{M} \frac{(x_i - \mu_i^k)^2}{\sigma_i^k} \tag{6}$$

and with the conventional [3] setting $p(l|x_n) = p_{\mathrm{old}}(l|x_n) - D/N$ one obtains for zero derivatives with respect to the parameters $\lambda = \mu_i^l$, σ_i^l the solution:
($\delta_{c,l} = 1$ if $c = l$, 0 else)

$$\mu_i^l = \frac{D(\mu_i^l)_{\mathrm{old}} + \sum_{n=1}^{N} [\delta_{k_n,l} - p_{\mathrm{old}}(l|x_n)] x_i^n}{D + \sum_{n=1}^{N} [\delta_{k_n,l} - p_{\mathrm{old}}(l|x_n)]} \tag{7}$$

$$\sigma_i^l = \frac{D(\sigma_i^l)_{\mathrm{old}} + \sum_{n=1}^{N} [\delta_{k_n,l} - p_{\mathrm{old}}(l|x_n)](x_i^n - \mu_i^l)^2}{D + \sum_{n=1}^{N} [\delta_{k_n,l} - p_{\mathrm{old}}(l|x_n)]}$$

With D large, changes are slight, and the method works iteratively. The inclusion of uncertainty after eq. 2 now changes the MMI formula, using vector notation:

$$\mathrm{MMI}_{\mathrm{uncertain}} = \int_y \sum_{n=1}^{N} p(y|x_n) \times \tag{8}$$
$$\times (\ln p(k_n) + \ln p(y|k_n) - \ln \sum_c p(c)p(y|c))$$

Since integration is linear, it goes that summation is replaced everywhere with summation + integration:

$$\mu_i^l = \frac{D(\mu_i^l)_{\mathrm{old}} + \int_y \sum_{n=1}^{N} p(y|x_n)[\delta_{k_n,l} - p_{\mathrm{old}}(l|y)] \quad y_i}{D + \int_y \sum_{n=1}^{N} p(y|x_n)[\delta_{k_n,l} - p_{\mathrm{old}}(l|y)]}$$

$$\sigma_i^l = \frac{D(\sigma_i^l)_{\mathrm{old}} + \int_y \sum_{n=1}^{N} p(y|x_n)[\delta_{k_n,l} - p_{\mathrm{old}}(l|y)](y_i - \mu_i^l)^2}{D + \int_y \sum_{n=1}^{N} p(y|x_n)[\delta_{k_n,l} - p_{\mathrm{old}}(l|y)]}$$

The integrations can be treated analytically if:

- the uncertainty variance is small compared to the curvature of the posterior at x_n
- the curvature of the posterior is smooth almost everywhere, in particular in regions where classification is (almost) definite, i.e. within the range of the variance of a density.

Then it is appropriate to expand the posterior $p(l|y)$ about the observation x_n. For the Gaussian model $p(y|c)$, eq. 6 this yields in second order:

$$\frac{P(l|y)}{P(l|x_n)} \simeq 1 + \sum_{i=1}^{N} G_i^l (y_i - x_i^n) + \frac{1}{2} \sum_{i,j} H_{ij}^l (y_i - x_i^n)(y_j - x_j^n)$$

with the reduced Gradient

$$G_i^l = \sum_c (P(c|x_n) - \delta_{l,c}) \frac{x_i^n - \mu_i^c}{\sigma_i^c} \tag{9}$$

and the reduced Hessian

$$H_{ij}^l = \left(\sum_c (P(c|x_n) - \delta_{l,c}) \frac{x_i^n - \mu_i^c}{\sigma_i^c} \right) \times \tag{10}$$
$$\times \left(\sum_k (P(k|x_n) - \delta_{l,k}) \frac{x_j^n - \mu_j^k}{\sigma_j^k} \right)$$
$$+ \sum_c \frac{P(c|x_n) - \delta_{l,c}}{\sigma_j^c} \delta_{ij}$$
$$+ \sum_c P(c|x_n) \frac{x_i^n - \mu_i^c}{\sigma_i^c} \sum_k (P(k|x_n) - \delta_{c,k}) \frac{x_j^n - \mu_j^k}{\sigma_j^k}$$

In particular,

$$G_i^l = -\frac{x_i^n - \mu_i^l}{\sigma_i^l} + \sum_c P(c|x_n) \frac{x_i^n - \mu_i^c}{\sigma_i^c} \tag{11}$$

$$H_{ij}^l = G_i^l G_j^l - \frac{\delta_{i,j}}{\sigma_i^l} + \sum_c \frac{P(c|x_n)}{\sigma_i^c} [\delta_{i,j} + (x_i^n - \mu_i^c) G_j^c]$$

where the additional computation effeort is *linear* with the number of densities, since the latter sums are only dependent on the dimensions i, j, but independent of l. Hence they need only be evaluated *once* per dimension, and can then be used for all G_i^l, H_{ij}^l. Performing the integrations for the estimation formulae now allows integration of the uncertainty y.

$$\mu_i^l = \rho D(\mu_i^l)_{\mathrm{old}} \quad + \rho \sum_{n=1}^{N} P(l|x_n) A_i^l(x_n)$$
$$+ \rho \sum_{n=1}^{N} [\delta_{k_n,l} - p_{\mathrm{old}}(l|x_n)] x_i^n \tag{12}$$

$$\sigma_i^l = \rho D(\sigma_i^l)_{\mathrm{old}} + \rho \sum_{n=1}^{N} P(l|x_n) B_i^l(x_n)$$
$$+ \rho \sum_{n=1}^{N} [\delta_{k_n,l} - p_{\mathrm{old}}(l|x_n)] [(x_i^n - \mu_i^l)^2] \tag{13}$$

$$\text{with} \quad \rho = \left(D + \sum_{n=1}^{N} [\delta_{k_n,l} - p_{\mathrm{old}}(l|x_n)] \right)^{-1} \tag{14}$$

where the dependency w.r.t. the uncertainty variance components v_i^n goes with the correction factors:

$$A_i^l(x_n) = G_i^l v_i^n + \frac{1}{2} x_i^n \sum_{j=1}^{N} H_{jj}^l v_j^n \quad (15)$$

$$B_i^l(x_n) = v_i^n \left(1 + 2G_i^l (x_i^n - \mu_i^l)\right) + \tfrac{1}{2}(x_i^n - \mu_i^l)^2 \sum_{j=1}^{N} H_{jj}^l v_j^n \quad (16)$$

Both mean and variance are shifted due to data uncertainty.

3. MIXTURE DENSITY ESTIMATION UNDER THE ML CRITERION

We consider a scenario where each observation x_n is associated with exactly one class c. For each class c (index omitted in the following), we aim at estimating the parameters λ of a mixture density $\sum_{i=1}^{N} c_k p(x_n|k,\lambda)$ from the associated observations. Under ML, we then have to maximize in the crisp case the Kullback-Leibler-statistics over component densities k (see e.g. [4])

$$Q(\lambda, \lambda^*) = \sum_{n=1}^{N} \sum_{k} p(k|x_n, \lambda) \ln \left(c_k^* p(x_n, k|\lambda_k^*)\right) \quad (17)$$

with respect to new (starred) model parameters, where the component weights $\sum_k c_k^* = 1$. The posterior is given

$$p(k|x,\lambda) = \frac{c_k p(x|k,\lambda)}{\sum_l c_l p(x|l,\lambda)} \quad (18)$$

The component densities $p(x|k,\lambda)$ are modelled as Gaussians, cf. eq. 6. With inclusion of uncertainty, integration goes linearly as above. Setting to zero the derivatives of $Q + \Lambda(\sum_k c_k^* - 1)$ (Lagrange parameter Λ) with respect to the starred parameters results in (k = component index, i = dimension index)

$$c_k^* = \frac{1}{N} \int_y \sum_n p(k|y,\lambda) p(y|x_n) \quad (19)$$

$$\mu_i^{k*} = \frac{\int_y \sum_n p(k|y,\lambda) y_i p(y|x_n)}{N c_k^*} \quad (20)$$

$$\sigma_i^{k*} = \frac{\int_y \sum_n p(k|y,\lambda)(y_i - \mu_i^k)^2 p(y|x_n)}{N c_k^*} \quad (21)$$

There is now a striking similarity to the estimation formulae in MMI. In fact the *only* difference is the absense of the terms with $\delta_{k_n,l}$ and D.

Expanding the posteriors $p(k|y,\lambda)$ in eq. 18 to second order in $(y - x_n)$ changes the terms as above in eqns. 12 - 14, with $P(k|x_n)$ replaced by $P(k|x_n,\lambda)$. The results are, again for diagonal uncertainty covariance matrices, eq. 1:

$$c_k{}^* = \frac{1}{N} \sum_n P(k|x_n,\lambda)\left[1 + C^k(x_n)\right] \quad (22)$$

$$\mu_i^{k*} = \frac{\sum_{n=1}^{N} P(k|x_n,\lambda)\left[x_i^n + A_i^k(x_n)\right]}{\sum_n P(k|x_n,\lambda)} \quad (23)$$

$$\sigma_i^{k*} = \frac{\sum_{n=1}^{N} P(k|x_n,\lambda)\left[(x_i^n - \mu_i^k)^2 + B_i^k(x_n)\right]}{\sum_n P(k|x_n,\lambda)} \quad (24)$$

with the correction factors A_i^k and B_i^k given in eqns. 15, 16 above (whereas here k = component), and

$$C^k(x_n) = \frac{1}{2} \sum_{j=1}^{N} H_{jj}^k v_j^n \quad (25)$$

So all variables, weights, means and variances, are shifted due to uncertainty of the data.

4. HIDDEN MARKOV MODELS

In HMM, not only do we have the parameters of the output probability density model, but as well the initial probabilities Π_i and transition probabilities $a_{i,j}$ [4]. Furthermore, in contrast to data observations, the probabilities of observing a state depend on the *full* set of data observations. One can show [6] that the Π_i, $a_{i,j}$ and mixture weights c^k do *not change* under the following conditions:

- The independency $\prod_{t=1}^{T} p(y_t|x_t)$
- linearization of the posteriors

If one expands the posteriors to second order, or if one computes the means and variances of the densities's emission probabilities even with linearization of the posteriors, second order terms emerge which do not vanish. These terms depend on the current path taken out of the N^T many possible paths, hence the usual computation of posteriors with forward- and backward probabilities does not work anymore. Including uncertainty modelling into the forward- and backward probability scheme will be presented in [6].

5. UNCERTAINTY VARIANCES AS MODEL PARAMETERS

So far we have treated the uncertainty variances as *properties* of the data, which we either known in advance, e.g. from corpus information, or which we obtain in an unsupervised mode, e.g. from SNR investigations on the data or from confidence measures in a two-pass approach.

We may however also treat the uncertainty variance as a speech recognition *parameter* which we set in a supervised mode, e.g. by optimizing the WER in an evaluation set. Alternatively, one can use a *global* uncertainty variance vector v as an additional model parameter which we estimate according to our optimization criterion.

We will demonstrate this optimization here for the scenario of section 3. The Kullback–Leibler statistics $Q(\lambda, \lambda^*)$ eq. 17, integrated for uncertain data, has to be optimized with respect to new (starred) model parameters under the condition $\sum_k c_k^* = 1$. Note that now the vector v^* belongs to the starred model parameters λ^*, rather than being a property of the data. Setting to zero the derivatives of $Q + \Lambda(\sum_k c_k^* - 1)$ with respect to the starred parameters results in (k = component index, i = dimension index)

$$c_k^* = \frac{1}{N} \int_y \sum_n p(k|y,\lambda) P(y|x_n, v^*) \tag{26}$$

$$\mu_i^{k*} = \frac{\int_y \sum_n p(k|y,\lambda)\, y_i\, P(y|x_n, v^*)}{N c_k^*}$$

$$\sigma_i^{k*} = \frac{\int_y \sum_n p(k|y,\lambda)(y_i - \mu_i^{k*})^2 P(y|x_n, v^*)}{N c_k^*}$$

$$v_i^* = \frac{1}{Q(\lambda,\lambda^*)} \sum_n \int dy P(y|x_n, v^*)(y_i - x_i^n)^2 \times$$

$$\times \sum_{\text{components } k} p(k|y,\lambda) \ln \left(c_k^* p(y,k|\mu^{k*}, \sigma^{k*}) \right) \tag{27}$$

The last equation of this set is an implicit equation in $\{v_k^*\}$. Solution of this set of equations can to very good approximation be done by taking the *old* v instead of v^* in the first three equations 26, which can then be explicitly solved as demonstrated above. $Q(\lambda, \lambda^*)$ can as well be expressed with these parameters, giving for components k

$$Q(\lambda,\lambda^*) = N \sum_k c_k^* \left(-\frac{M}{2} + \ln \frac{c_k^*}{\prod_{j=1}^M \sqrt{2\pi\, \sigma_j^{k*}}} \right) \tag{28}$$

Using Q and the resulting starred parameters c^*, μ^*, σ^*, we iterate the forth equation 27 with respect to v^*.

6. VITERBI PATHS AND MAXIMUM APPROXIMATION

Under the Viterbi approximation in speech recognition, the association of observations with states is fixed by the path. It is therefore necessary only to estimate the parameters of the densities associated with the state q given all observations x^n corresponding to this state according to the Viterbi path. This set of observations then acts as uncorrelated observations for the estimation of a mixture, see section 3.
Rather than estimating parameters of the densities from posteriors P(component $k|$ observation x), the *maximum approximation* [4] can be used in the estimation formulae with uncertainty modelling. Then eqns. 11 for the Gradient and the Hessian both become 0. The reason for this is that $P(c|x) = \delta_{l,c}$ since the observation x will only contribute to density l if it is closest to μ_l. So in the case of maximum approximation, what remains for the updates are

$$A_i^l(x_n) = 0 \quad , \quad B_i^l(x_n) = v_i^n \quad , \quad C^l(x_n) = 0 \tag{29}$$

which amounts to a simple addition of the uncertainty variance to the former variance estimate. Hence we have given a thorough theoretical justification for variance biasing.

7. EXPERIMENTS WITH CSDC DIGITS

In CSDC digits [2], the training set are 19477 words, the test set are 7731 words (incl. garbage). Experiments were done for density specific variances, Viterbi paths, ML estimation and sum of likelihoods for density estimation. Uncertainty modellings in training and recognition match. We compare the baseline (no uncertainty modelling) word error rates with a global uncertainty variance which is moderate compared to the baseline density variances, both for models resolved with a range of densities. Standard error bar size is about 1% absolute. While the baseline performance is such that WER initially reduces with higher number of densities, and later increases again due to sparse data for density estimation, uncertainty modelling reaches a lower WER minimum than the baseline at higher number of densities, and is not subject to severe WER increases due to sparse data. This underlines the proposed effect of prohibiting variance narrowing. However, with a global uncertainty variance chosen, the powers of uncertainty modelling are shown here only at a first stage.

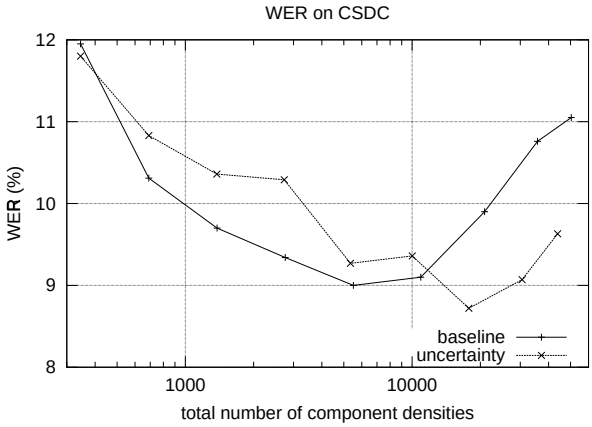

Fig. 1. WER on CSDC

8. ACKNOWLEDGEMENTS

It is a pleasure to acknowledge numerous discussions with Christoph Neukirchen, Peter Beyerlein, Hans Dolfing and Georg Rose.

9. REFERENCES

[1] H. Hayes, "Statistical Digital Signal Processing and Modeling", *John Wiley*, 1996.

[2] D. Langmann et al., "CSDC- The MoTiV Car-Speech Data Collection", *First International Conference on Language Resources and Evaluation*, Granada, Spain, May 1998.

[3] Y. Normandin, "Hidden Markov Models, Maximum Mutual Information Estimation, and the Speech Recognition Problem" *PhD thesis*, Dep. of Electr. Eng., Mc Gill University, Montreal, 1991.

[4] L. R. Rabiner and B.-H. Huang. "Fundamentals of Speech Recognition". *Prentice Hall*, Englewood Cliffs, New Jersey, 1993.

[5] V. Vapnik, "The nature of statistical learning theory", *Springer*, 1995.

[6] A. Wendemuth and C. Neukirchen, "Uncertainty modelling in Hidden Markov Models", in preparation.