# CONFIDENCE-DRIVEN ESTIMATOR PERTURBATION : BMPC

*Stefan Besling*
Philips GmbH Forschungslaboratorien
Weißhausstr. 2, D-52066 Aachen, Germany
besling@pfa.research.philips.com
*Hans-Günter Meier*
Fachhochschule Düsseldorf
Josef-Gockeln-Straße 9, D-40474 Düsseldorf, Germany
guenter.meier@fh-duesseldorf.de

## ABSTRACT

In most practical applications of speech recognition, like for example in a dictation system (cf. [3]), the acceptance and performance of the system depends strongly on its capability to adapt to the special speaker characteristics. Restricted to the problem of language model adaptation, one has to find an efficient way to combine a typically well-trained a priori estimator for a domain with a regularly updated but undertrained estimator reflecting the actual speaker-specific data so far. To assure a greater impact of reliable speaker-specific information, in this paper we present a new language model estimation technique that makes explicit use of the confidence in estimates obtained on the (typically small) adaptation or training data. Mathematically it attempts to perturb a given reliable a priori distribution in such a way that it fits into the confidence regions given by the training material. Experiments performed on real-life data supplied by US radiologists indicate that the method could improve standard adaptation techniques like linear interpolation.

## 1. INTRODUCTION

Given some basic vocabulary $V$, a language model is given by an estimator $\widehat{P}$ which tries to approximate the unknown probability of some finite word sequence $h_N := (w_1, \ldots, w_N)$, which also indicates the history of words spoken so far. Because of

$$\widehat{P}(w_1, \ldots, w_N) = \prod_{i=1}^{N} \widehat{P}(w_i|h_{i-1})$$

a language model can be interpreted as a collection of discrete conditional distributions $\widehat{P}(\cdot|h)$ over a finite vocabulary. Standard methods to establish $\widehat{P}(w|h)$ for events $(h, w)$ usually count event frequencies on corpora, modify them in a more or less simple way (like subtracting 1 from a count if possible [4]) and use the result as a point estimation for the unknown probability that the event occurs. Especially for small counts, this has the problem that the resulting estimate is of low confidence. In this paper, we present a new estimation technique that makes explicit use of this confidence. Starting with a typically small corpus of adaptation material, of which we know from experience that it is representative of the speaker-specific test data, we compute confidence intervals $[\alpha_h(w), \beta_h(w)]$ for history/word events $(h, w)$, based on the assumption of a binomial distribution. For each history $h$ we now would like to choose our estimator out of the set $\mathcal{C}_h$ of all distributions respecting these confidence intervals by minimizing the Kullback-Leibler distance $D(\cdot||\cdot)$ to a given strictly positive a priori distribution $P(\cdot|h)$. Using Distr$(V)$ to indicate the set of all strictly positive discrete distributions over $V$, this perturbation of $P(\cdot|h)$ is more precisely decribed by

$$\widehat{P}(\cdot|h) := \underset{Q(\cdot|h)\in\mathcal{C}_h}{\operatorname{argmin}} D(P(\cdot|h)||Q(\cdot|h))$$

using the notations

$$D(p||q) := \sum_{w \in V} p(w) \log \left( \frac{p(w)}{q(w)} \right)$$

for $q, p \in \text{Distr}(V)$ and

$$\mathcal{C}_h := \{q \in \text{Distr}(V) | q(w) \in [\alpha_h(w), \beta_h(w)]\}.$$

Note that in case of using a relative frequency distribution for $P(\cdot|h)$, this coincides with a maximum likelihood approach under confidence constraints $\mathcal{C}_h$.

As discussed in more detail in the next Section, it is possible to recognize the *BMPC*[1] estimator $\hat{P}(w|h)$ as a structurally simple cut-off scheme.

Since $P(\cdot|h)$ may be any distribution, particularly one obtained by linear interpolation or other adaptation techniques like fill-up (cf. [1]), our method can be used as a post-processing step to further improve a given model. The tests indicate that it does a good job at this.

## 2. THEORETICAL DISCUSSION

Solving the minimization problem described in Section 1 for a given strictly positive probability distribution $P(\cdot|h)$ and intervals $[\alpha_h(w), \beta_h(w)]$ such that $\sum_w \alpha_h(w) \leq 1 \leq \sum_w \beta_h(w)$ leads to the following structurally simple cut-off scheme of the *BMPC* estimator

$$\hat{P}(w|h) = \begin{cases} \alpha_h(w) & w \in \mathcal{A}_h \\ \beta_h(w) & w \in \mathcal{B}_h \\ \gamma_h P(w|h) & \text{else} \end{cases}$$

where

$$\mathcal{A}_h := \{w \in V \mid \gamma_h P(w|h) < \alpha_h(w)\},$$

$$\mathcal{B}_h := \{w \in V \mid \gamma_h P(w|h) > \beta_h(w)\}.$$

This means that the a priori distribution $P(\cdot|h)$ is scaled by an appropriate positive factor $\gamma_h$ if the result fits into the confidence intervals. In cases where it does not fit, the value is cut off to the nearest interval boundary.

Of course the parameter $\gamma_h$ has to be chosen so as to assure that the estimator sums up to 1 which leads to:

$$\gamma_h \sum_{w \notin \mathcal{A}_h \cup \mathcal{B}_h} p(w|h) = 1 - \sum_{w \in \mathcal{A}_h} \alpha_h(w) - \sum_{w \in \mathcal{B}_h} \beta_h(w)$$

Because $\gamma_h$ also appears in the definition of $\mathcal{A}_h$ and $\mathcal{B}_h$, we are faced with a fixed point equation which can be solved using standard techniques.

For each history $h$, the interval boundaries $\alpha_h(w), \beta_h(w)$ are determined with respect to some level of confidence for each word $w$ under the assumption that the appearance of each event $(h, w)$ is governed by a binomial distribution. To avoid any free parameter the level of confidence is chosen to minimize $\mathcal{C}_h$ while making sure that the relative frequency distribution is within the constraint set. Confidence intervals for the binomial distribution, which are quite well-known (and sometimes called Pearson-Clopper intervals), can be explicitly expressed in terms of the F-distribution (cf. Lehmann [2] pp. 199ff and the appendix) which is available in most mathematics libraries.

## 3. THE RESULTS

For the tests presented in the rest of the Section, we used bigram estimators, which means that for histories ending with the same word, the same conditional distribution was used. In the case of an unseen history (i.e. an unseen predecessor word within the adaptation material) we used a conditional distribution derived from the background data.

In order to evaluate our new estimator, we performed tests on real-life data supplied by US radiologists. These tests were performed using the following strategy :

1. Start with an initial model obtained on background data from one hospital location (not including any of the speaker data).

[1] 'Best Model Perturbation within Confidence'. The suggested pronunciation is 'bempac'.

2. Recognize one utterance using the current model.

3. Obtain the corrected text of the recognized utterance. Add it to the text accumulated so far and train a standard speaker-specific language model on this adaptation material. Build a new model using the one just trained, the initial one, as well as the confidence intervals calculated on the adaptation material.

4. Go to step 2.)

To build the new model mentioned in step 3, we first combined the initial model and the confidence knowledge by using *BMPC*. As a second way of combination we chose standard linear interpolation (*LI*) of the initial model and the speaker-specific model. Furthermore this *LI* model was additionally combined with the speaker-specific confidence intervals, which is referred to as *LI+BMPC*.

As opposed to *BMPC*, which does not know anything about the test material, the free weighting parameter of the linear interpolation was chosen so as to minimize the perplexity of the test set (which of course gives overly optimistic results).

Results are shown in Table 2 where the upper half gives word error rates and the lower half shows perplexities. The 'baseline' column always shows the word error rate obtained with the background material only, i.e. without any adaptation.

It can be seen that the use of *BMPC* leads to a consistent improvement, performing essentially as well as the test-optimized *LI* or the combination of both *LI* and *BMPC* without the need of any additional adjustment of free parameters.

As far as perplexities are concerned, *LI* clearly outperforms *BMPC*. However, by applying *BMPC* in addition to *LI* it is possible to further reduce perplexities by 23% on average, while the word error rate remains roughly the same (2% reduction on average). Once again this indicates that perplexity is not always a reliable measure when comparing the performance of different language models (cf. [1]).

| Speaker | #Utterances | #Words |
|---|---|---|
| m80 | 13 | 3,962 |
| m81 | 7 | 5,743 |
| m90 | 21 | 1,922 |
| f95 | 50 | 6,039 |
| BG | 166 | 115,172 |

Table 1: Sizes of adaptation and background material

| Speaker | Baseline | BMPC | LI | LI+BMPC |
|---|---|---|---|---|
| m80 | 15.5% | 12.3% | 12.2% | 11.8% |
| m81 | 31.3% | 28.0% | 27.7% | 27.4% |
| m90 | 18.4% | 15.8% | 16.2% | 16.0% |
| f95 | 19.4% | 16.3% | 17.3% | 17.1% |
| Aver. Δ | - | -15% | -14% | -15% |
| m80 | 6122 | 181 | 166 | 94 |
| m81 | 4542 | 548 | 254 | 227 |
| m90 | 966 | 155 | 110 | 77 |
| f95 | 838 | 141 | 67 | 60 |
| Aver. Δ | - | -88% | -93% | -94% |

Table 2: Test results (upper half: word error rate; lower half: test set perplexity )

## 4. THE CONCLUSION

From the tests we performed it appears that *BMPC* is a robust way to improve background models and to adapt them to small amounts of speaker-specific data. As compared to linear interpolation, there is no need to optimize free parameters and in fact when used in addition to *LI*, *BMPC* has the ability to compensate for suboptimal parameter settings.

## 5. REFERENCES

[1] BESLING, S.; MEIER, H.-G., *Language Model Speaker Adaptation*, Proceedings EUROSPEECH 1995, Madrid, pp. 1755-1758

[2] LEHMANN, E. L.,

*Testing Statistical Hypotheses*, Chapman&Hall, New York 1994

[3] NEY, H.; STEINBISS, V.; HAEB-UMBACH, R.; TRAN, B.-H.; ESSEN, U., *An Overview of the Philips Research System for Large Vocabulary Continuous Speech Recognition*, Intern. Journal of Pattern Recognition and Artificial Intelligence, Vol. 8 No. 1 (1994), pp. 33-70

[4] NEY, H. ; ESSEN, U. ; KNESER, R., *On Structuring Probabilistic Dependencies in Stochastic Language Modelling*, Computer Speech and Language, Vol. 7 (1993), pp. 101-138

## Appendix

Depending on the level of confidence $1-\epsilon$ the confidence interval $[\alpha_\epsilon(k,N), \beta_\epsilon(k,N)]$ associated with an event occurring exactly $k$ times in a series of $N$ independent samples assures that for all $\wp \in [0,1]$ we have

$$P_\wp(\{k \in \{0,\ldots,N\} \mid \alpha_\epsilon(k,N) \leq \wp \leq \beta_\epsilon(k,N)\}) \geq 1-\epsilon$$

where $\wp$ denotes the probability that $w$ is drawn after history $h$.
Thus $\alpha_h(w)$ and $\beta_h(w)$ depend only on the counts $N(h,w)$, $N(h)$ and a free parameter $\epsilon$ if we set $\alpha_h(w) := \alpha_\epsilon(N(h,w),N(h))$ and $\beta_h(w) := \beta_\epsilon(N(h,w),N(h))$. Its well-known and easy to verify that $\alpha_\epsilon(k,N)$ and $\beta_\epsilon(k,N)$ (sometimes called the *Pearson-Clopper intervals*) can be explicitly expressed in terms of the $F$-distribution as

$$\alpha_\epsilon(k,N) =$$

$$\begin{cases} 0 & k=0 \\ \frac{kF_{2k,2(N-k+1),\frac{\epsilon}{2}}}{(N-k+1)+kF_{2k,2(N-k+1),\frac{\epsilon}{2}}} & k \in \{1,\ldots,N\} \end{cases}$$

and

$$\beta_\epsilon(k,N) =$$

$$\begin{cases} \frac{(k+1)F_{2(k+1),2(N-k),1-\frac{\epsilon}{2}}}{(N-k)+(k+1)F_{2(k+1),2(N-k),1-\frac{\epsilon}{2}}} & k \in \{0,\ldots,N-1\} \\ 1 & k=N \end{cases}$$

Here $F_{n,m,\eta}$ is defined as a quantile

$$\int_0^{F_{n,m,\eta}} f_{n,m}(x)dx = \eta$$

with respect to the probability density function (cf. [2], pp. 199ff)

$$f_{n,m}(x) = \left(\frac{n}{2}\right)^{\frac{n}{2}} \left(\frac{m}{2}\right)^{\frac{m}{2}} \frac{\Gamma(\frac{n}{2}+\frac{m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \frac{x^{\frac{n}{2}-1}}{(\frac{n}{2}x+\frac{m}{2})^{\frac{n+m}{2}}}.$$

For practical use, we choose the free parameter $\epsilon$ as the maximal one such that the relative frequencies $\frac{N(h,w)}{N(h)}$ are still assured to be within the interval $[\alpha_\epsilon(N(h,w),N(h)), \beta_\epsilon(N(h,w),N(h))]$. Since this also makes sure that $\sum_w \alpha_\epsilon(N(h,w),N(h)) \leq 1 \leq \sum_w \beta_\epsilon(N(h,w),N(h))$ for all histories $h$, it can be shown that there is a non-negative solution for $\gamma_h$.