A CONFIDENCE-SCORE BASED UNSUPERVISED MAP ADAPTATION FOR SPEECH RECOGNITION

Dagen Wang, Shrikanth S. Narayanan

Department of Electrical Engineering
University of Southern California, Los Angeles
[dagenwan,shri]@sipi.usc.edu

ABSTRACT

In this paper, a method of confidence-score based MAP (maximum a posteriori) adaptation in speech recognition is proposed and evaluated. Using confidence scores to dynamically decide the weight of the priors is shown to have good performance improvement in unsupervised incremental adaptation. The side effect of vocabulary mismatch in adaptation is also effectively controlled by this way. This paper first gives theoretical analysis and then shows some experimental results. Several extensions are made and also discussed.

1. INTRODUCTION

Speech adaptation in statistical automatic speech recognition (ASR) provides a way for progressively reducing model mismatches between training and testing (usage) data. There are two major model adaptation approaches that are popular. Namely, MAP (maximum a posteriori) adaptation and maximum likelihood linear regression (MLLR) adaptation. MLLR has been shown to yield significant improvement in ASR performance with relatively small amounts of data. But in the long run, with increasing amounts of data, MLLR does not converge to the right model. On the other hand, while MAP adaptation is known to need far more data to demonstrate good performance, it has good convergence properties toward the "true" model in the long run. There are recent research directions that aim at combining these two methods, to benefit from their respective merits.[7]

The motivation for the work in this paper comes from potential applications to embedded speech recognition. For adaptation in embedded scenarios, MAP appears to be preferable since it is difficult to construct and manipulate an MLLR tree with limited memory and processing resources. MAP needs relatively simple operations. [4]

Each of the aforementioned speech adaptation approaches can operate in one of two modes: supervised or unsupervised. Certainly supervised adaptation has better performance because it has correct and ready to use transcriptions. But providing the correct transcription generally requires manual intervention and often is an off-line procedure. In scenarios such as with embedded distributed speech recognition in mobile devices, it is additionally challenging to deal with the non-stationary operational environments. That is why unsupervised adaptation is often used. Based on the Viterbi decoded transcription, the HMM model parameters are updated by weighting between the priors and the input data statistics. In this paper, a confidence score based approach to decide this weight is addressed.

Finally, it should be mentioned that adaptation can be performed in either batch or incremental modes. Again, for real time adaptation considerations in our work, incremental adaptation is used for its small memory footprint.

2. CONFIDENCE MEASURES

In this paper, we propose a simple, but shown to be efficient, method to calculate the confidence score for the automatically recognized speech.

2.1 Basic Idea

Let x denote the event that the speech transcription is in grammar, f denote the feature vector; By Bayes' law:

$$P(x \mid f)=\frac{P(f \mid x)P(x)}{P(f)}$$

Note that the lexical (vocabulary, grammar) constraints dictate whether an utterance is in vocabulary or out of vocabulary (OOV). Statistically speaking, P(x) and P(f) are constant. So measuring P(f | x) can serve as a confidence measure. In speech recognition, Viterbi decoding always provides the result for in vocabulary/grammar, the log probability just reflect the P(f | x). For in vocabulary data, these probabilities are expected to be high. For OOV data, these probabilities are expected to be low because the transcription has a bad decoding path on the word net.

2.2. Confidence Measure Training

For this purpose, in vocabulary data are needed. The training is implemented by feeding the data to the decoder, and recording the output word-based log probabilities. In order to be consistent for different length words, normalization by frame number is needed. With these model based log probabilities, we can estimate the log probability distribution. Here we apply the Gaussian Parzen Window estimation method.

Let x_i be series the output normalized log probabilities. The pdf of the log probabilities can be estimated in terms of these x_i by[9]

$$P(x)=\frac{1}{J}\sum_{i=1}^{J}\Delta(x-x_i)$$

while Δ is the parzen window, J is the total number of sample log probabilities. To make the pdf smooth, we choose a Gaussian window.

0-7803-7576-9/02$17.00 © 2002 IEEE

$$\Delta(x-x_i)=\frac{1}{\sqrt{2\pi}h_i}\exp\{-\frac{1}{2}(\frac{x-x_i}{h_i})^2\}$$

h_i is chosen as the window width:

$$h_i=\frac{1}{\sqrt{J}}$$

Applying the Gaussian Parzen window method, we can get the pdf.

By comparison, we can feed a set of OOV data to the same decoder. A similar OOV log probability pdf can be obtained.

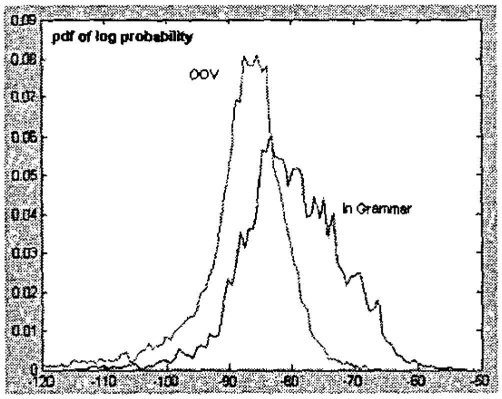

Figure 1: estimated pdf

It is apparent that the in grammar data are expected to have a higher mean of the log probabilities while the mean of OOV data are expected to be low. So by setting a threshold we can statistically distinguish in grammar data and OOV data.

In operation, the pdf is not normalized and not easy to be used. The most straight forward idea to transform the density function to cumulative distribution function (pdf to cdf) . The following illustrated this idea.

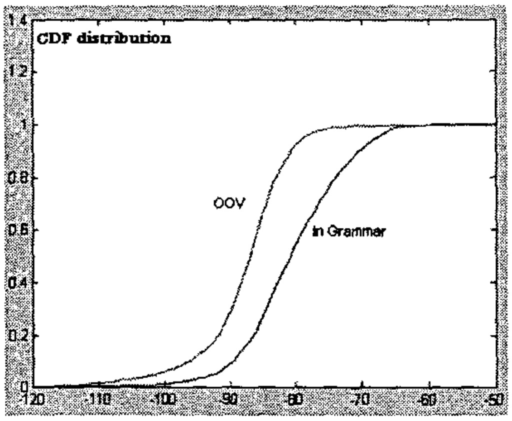

Figure 2: estimated CDF

The cdfs have a nice property that all the scores are between 0 and 1. If the score approaches 1, it conveys that it has high log probability and vice versa. For a given log probability, we can simply lookup the cdf table and get a confidence score which is between 0 and 1.

2.3 Three Types of Confidence Scores

A variety of possibilities have been proposed in the past for the confidence score problem, including model-based, word-based, and utterance-based confidence measure. The method described in Sec 2.2 is word-based confidence score (normalized by frame). This method could be easily extended to model based confidence scoring. For utterance-based confidence scores, we can just sum up the word log probabilities which are normalized by the number of words or even number of frames.

In processing-constrained adaptation, the first two kinds of confidence scoring need a great more computations in adaptation since the weighting coefficients calculations are needed every model/word. So in this context, utterance level for confidence scoring is preferable. For every new utterance, the adaptor applies the same weighting coefficients to each model. It is shown we can get significant recognition improvement even with this very simple approach of score computation.

3. WEIGHTING THE PRIORS

In unsupervised speech recognition, manual labor of labeling speech data is not needed but the performance suffers from incorrect alignment and out-of-vocabulary (OOV) word or utterances. Experiments[6] show that these tend to have a detrimental effect on the recognition accuracy. This in turn implies adaptation tends to degrade, instead of improving, performance. The idea here is to control the effect of OOV by means of confidence measures.

3.1 The learning parameter τ in MAP Adaptation

The key operation in MAP estimation is to choose the appropriate priors and their weights. For speaker adaptation using HMM with Gaussian Mixture states, it is normally assumed that the priors to be adequately represented as a product of Dirichlet and normal-Wishart densities.

Based on the assumption about conjugate priors, the procedure of EM algorithm can be applied to MAP estimation. Considering all the possible output state sequences, the adaptation procedure for mean and precision estimation can be obtained as: [4]

$$m_{ik}=\frac{\tau_{ik}\mu_{ik}+\sum_{t=1}^{T}c_{ikt}x_t}{\tau_{ik}+\sum_{t=1}^{T}c_{ikt}}$$

$$r_{ik}^{-1}=\frac{u_{ik}+\sum_{t=1}^{T}c_{ikt}(x_t-m_{ikt})(x_t-m_{ikt})'+\tau_{ik}(\mu_{ik}-m_{ik})(\mu_{ik}-m_{ik})')}{(\alpha_{ik}-p)+\sum_{t=1}^{T}c_{ikt}}$$

It is apparent that the estimation will be a weighted sum of the prior and the experiment data statistics. The question is how to decide the weights for each of them. For the MAP estimation procedure above, we have two choices:

1). It is more beneficial to use more "data effect" on data likely to contribute positively toward adaptation and use more prior effect otherwise. (Soft Weighting)

2). Use the data for adaptation only if it has a high confidence. In another words, reject the low confidence utterances. (Hard Weighting)

3.2 "Soft" Weighting

This method implies no out right rejection is applied. We apply adaptation to any input utterances. But we choose the weighting coefficients in relation to the confidence score. Given the equation in 3.1, the general principle is: for sentences with higher confidence scores, we choose a relative small τ and vice versa.

There are infinite many such functions, we choose the following functions which reflect different "flat" conditions around 0 or 1 confidence scores.

- **Constant Weighting**

No matter what the input is, the adaptation uses the same weighting coefficients.

$$\tau = K$$

This is the default setting for many of the speech recognizer implementations. Since it has no weighting schemes, all input utterances are deemed at the same confidence level. It has the worst performance comparing with other weighting methods. This acts as our comparison baseline.

- **Linear Weighting**

From the above MAP relations, for higher confidence score, we like to use smaller tau to incorporate more data effects with the final model. The natural implementation is to use a linear function.

$$\tau = 2 * K * (1 - conf\_score)$$

This method is shown to have better performance than constant weighting. The reason we use a factor of 2 here is to balance the overall mean of τ.

- **Exponential Weighting**

Linear weighting is not optimal in general. Since for OOV data, we need to remove its effect more aggressively. For in grammar data, we might wish to maximize its effect. In other words, we need a more flat curve when its confidence score is close to 0 or 1. Towards this goal, we choose exponential function as follows:

$$for\ 0.5 \le \tau \le 1$$
$$\tau = 2 * K * \ln(e - conf\_score)$$
$$for\ 0 \le \tau < 0.5$$
$$\tau = 2 * K * \exp(-conf\_score - e^{-1})$$

In this case, there exist a threshold 0.5. Which makes the mapping function discontinuous and change the slope of the mapping curve. Optimal choice of this threshold can improve the performance, but it is beyond the scope of this paper. Exponential Weighting is shown to have some improvement than the linear weighting in our experiment.

- **Step Weighting**

To be more aggressive, we apply step function which makes the mapping function a straight line in each slope.

$$for\ 0.5 \le \tau \le 1$$
$$\tau = C_{min}$$
$$for\ 0 \le \tau < 0.5$$
$$\tau = 2 * K$$

We use a nonzero constant C_{min} here is to avoid tau being zero. The result is almost the same with exponential Weighting.

Below is the picture of the above 4 methods.

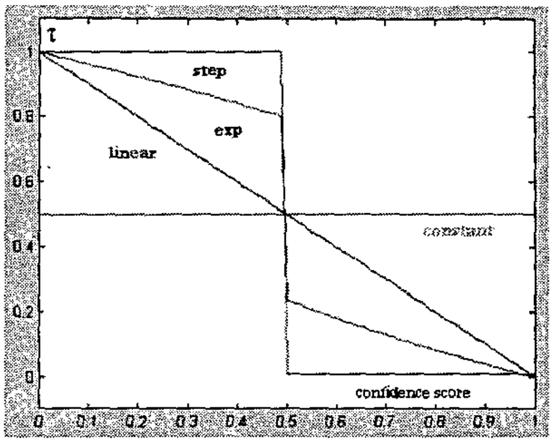

Figure 3: different weighting method

3.3 "Hard" Weighting

For adaptation processes, it is always harmful to use out of vocabulary speech to do the adaptation. Of course, the best strategy is to abandon the OOV data (if we can). For the purpose of implementing this idea, we tried a decision step before the adaptation. That is, first see if the input speech is OOV by comparing the threshold with the confidence score. If deemed OOV, then we abandon the data from adaptation and continue for next utterances.

If OOV decision can be made reliably, the performance of hard weighting in our case is better than soft step weighting. The fact is, the biggest challenge is whether we can robustly identify OOV occurrences. For different environment and people, it might have different solution. For the case of small number utterances, it has risk of degrading the performance by making the wrong decision. But, statistically speaking, it might converge to the same model as soft weighting. Even though, soft weighting might have a more steady performance in general.

4. Experiment and Results

Experiments were done using a hidden Markov model based phoneme recognizer. A standard clean speech database of phonetically balanced utterances (TIMIT) was used to bootstrap speaker independent phoneme models. Noisy speech data from 19 speakers in a military training simulation domain were used to train the domain specific acoustic models. The adaptation experiments were performed using speech from one specific speaker and results evaluated by the same speaker's speech.

For language model, we choose finite state grammar (FSG). It is easy to setup and widely used in embedded speech recognizer. It might have the same effect if we choose other kind of language model. The only difference is applying different language model weight to the decoding path. It is always true that OOV utterances have lower confidence scores.

In order to see the performance, we mixed the in vocabulary data (50%) and OOV data (50%) together. The adaptation procedure can be shown as follows:

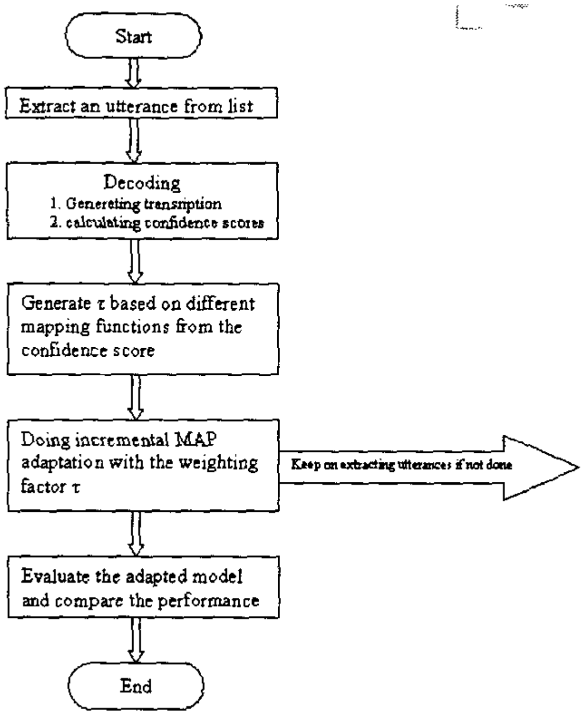
Start
Extract an utterance from list
Decoding
1. Genereting transnption
2. calculating confidence scores
Generate τ based on different mapping functions from the confidence score
Doing incremental MAP adaptation with the weighting factor τ
Keep on extracting utterances if not done
Evaluate the adapted model and compare the performance
End

Figure 4: the flowing chart of experiment

First, incremental MAP adaptation with constant weight priors was done to serve as the baseline. These results are then compared with the proposed confidence score weighted results. Results are shown for different size of adaptation data size.

WER	50 utterances	80 utterances	100 utterances	200 utterances
Const Wt	21.28	18.24	17.57	16.88
Linear Wt	19.93	17.22	15.54	13.85
Exp.Wt	17.91	16.89	15.20	13.51
Step Wt.	18.24	16.55	15.27	13.51
Hard Wt	18.24	15.86	13.85	13.51

The experimental results show that:

1. The confidence score weighted procedure we proposed is a statistical method. On small number of utterances' case, the performance improvement is uncertain due to large variance. In some cases, it is even worse than constant weighting. But, in the long run, the weighted procedure provides steady performance improvement.

2. In the case of sufficient data, all methods converge to the right model except for the constant weighting.

3. While not reaching saturated number of utterances, the difference between different soft weighting methods is: exponential weighting and step weighting provides the best adaptation rates. Linear weighting is not that good as those two but better than constant weighting.

4. Different weighting methods did not differ too much in performance as the amount of adaptation data increases. This is the same as the property as MAP.

5. DISCUSSIONS

- Better confidence score

In this method, the availability of reliable confidence score is a key requirement. In our case, we get performance improvement by using only log probabilities as a single feature to calculate confidence scores. Actually, there are many other useful features such as frequency of appearance in N-best list, etc. [1] With a more robust confidence score measuring, the weighting method proposed in this paper are expected to perform better.

- About combined weighting

In this paper we also discussed the hard weighting scheme. It has rather good performance in our experiment. As mentioned previously, the performance of this method is more related to the confidence score measuring. Actually, the performance of hard weighting can even be improved further by combining with soft weighting. That is, for the utterances which are not rejected, we can apply soft weighting "again". The combined method are expected to have better performance than hard weighting by itself.

- Number of utterances

As the experiment results show, as the number of utterances becomes large, the performance of both the weighting method and the overall performance are better. But in the case of too small amount of data, the performance is very uncertain. (which is not shown on the result table) In that case, the weighting sometimes even performs worse than not weighting. That is because large variance of speech data which can not be controlled by too few data. The true parameter was buried by heavy "noise".

6. EXTENSIONS

An important extension is to do the recursive MAP estimation. The above segmental method does not have memory to remember past data. By applying recursive MAP method, it will be more likely to converge. [5]

Another method in adaptation is MLLR adaptation. By creating a regression node tree, several optimization methods has been provided. [3][6] Confidence score optimization can also be applied to that approach.[8] Comparing with MAP, it needs far less data, but it does not guarantee convergence. Also the memory overhead is more than MAP. It is attractive to use confidence measure to improve MAP estimation performance and accelerate the convergence speed.

Combined MLLR and MAP is also an attractive method. Applying confidence score to this method is also an interesting topic.[7]

7. REFERENCES

[1] Timothy J. Hazen and Issam Bazzi, "A comparison and combination of methods for OOV word detection and word confidence scoring," *Proceedings of the 2001 International Conference on Acoustics, Speech, and Signal Processing,* Salt Lake City, May, 2001

[2] Ananth Sankar and Ashvin Kannan, "Automatic confidence score mapping for adapted speech recognition systems", *ICASSP 2002, 2334*

[3] Chafic Mokbel, "Online Adaptation of HMMs to Real-Life Conditions: A Unified Framework" *IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 4, May 2001*

[4] L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing, vol. 2, pp. 291–298, Apr. 1994.*

[5] Shaojun Wang and Yunxin Zhao, "Online Bayesian Tree-Structured Transformation of HMMs With Optimal Model Selection for Speaker Adaptation", *IEEE Transactions On Speech and Audio Processing, Vol. 9, No. 6, September 2001 663*

[6] J. Chien and J. Junqua, "Unsupervised hierarchical adaptation using reliable selection of cluster-dependent parameters," *Speech Commun., vol. 30, pp. 235–253, 2000.*

[7] Steve Young et al. "The HTK Book (for HTK Version 3.1)"

[8] Ka-Yan Kwan, Tan Lee and Chen Yang, "Unsupervised N-best based model adaptation using model level confidence measures", *ICSLP, 2002*

[9] R.O. Duda, P.E. Hart, and D.G.Stork, "Pattern Classification", *Second Edition(Wiley-Interscience, John Wiley and Sons, Inc., New York, 2001)*

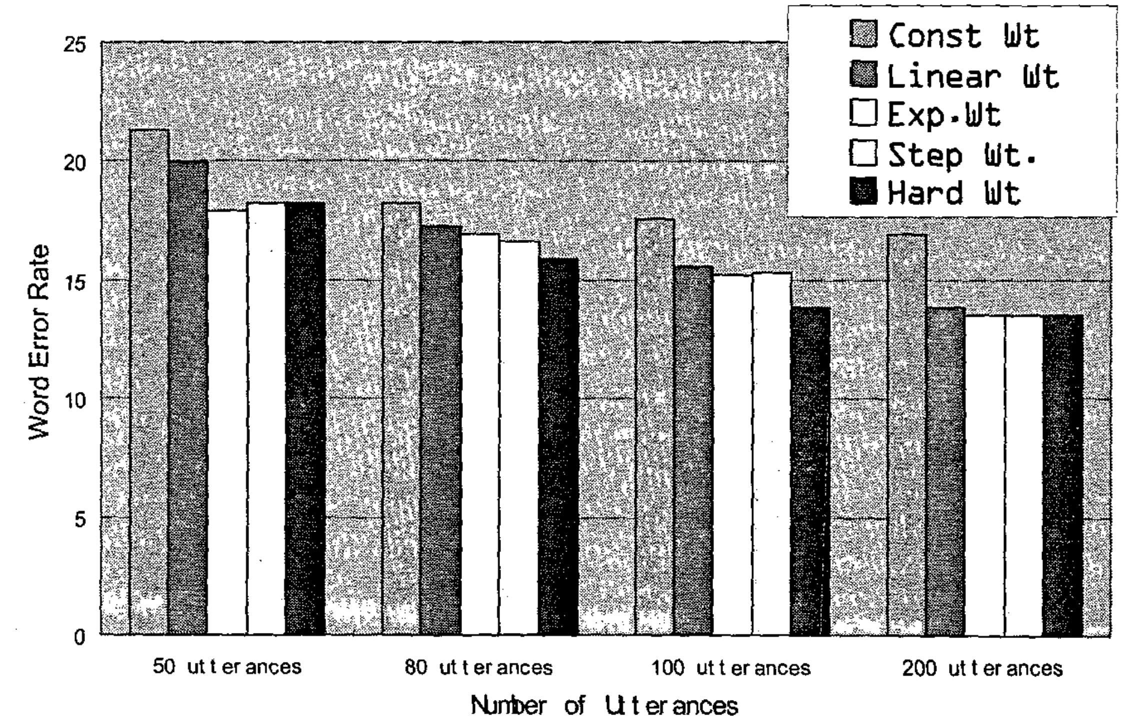

Figure 5: Different Weighting Method Compare