CONFIDENCE SCORES FOR ACOUSTIC MODEL ADAPTATION

*Christian Gollan**

Human Language Technology and
Pattern Recognition,
Computer Science Department 6,
RWTH Aachen University, Germany
gollan@cs.rwth-aachen.de

Michiel Bacchiani

Google Inc.,
76 Ninth Avenue
4th Floor
New York, NY 10011, USA
michiel@google.com

ABSTRACT

This paper focuses on confidence scores for use in acoustic model adaptation. Frame-based confidence estimates are used in linear transform (CMLLR and MLLR) and MAP adaptation. We show that adaptation approaches with a limited number of free parameters such as linear transform-based approaches are robust in the face of frame labeling errors whereas adaptation approaches with a large number of free parameters such as MAP are sensitive to the quality of the supervision and hence benefit most from use of confidences. Different approaches for using confidence information in adaptation are investigated. This analysis shows that a thresholding approach is effective in that it improves the frame labeling accuracy with little detrimental effect on frame recall. Experimental results show an absolute WER reduction of 2.1% over a CMLLR adapted system on a video transcription task.

Index Terms— acoustic model adaptation, confidence scores

1. INTRODUCTION

Acoustic model adaptation is a common component of state-of-the-art large vocabulary continuous speech recognition (LVCSR) systems. Popular methods are linear transform based adaptation algorithms such as maximum likelihood linear regression (MLLR) and constrained MLLR (CMLLR) [1, 2, 3] and state-dependent parameter adaptation such as maximum *a posteriori* (MAP) [4]. These adaptation methods estimate the model parameters so as to maximize the state emission likelihoods for the observed acoustic feature vectors. The state-sequence conditioning can be supervised in the case of supervised adaptation or when the adaptation algorithm is used in training to normalize for speaker variability orthogonal to the speech content. For unsupervised adaptation, at test time, the conditioning state sequence is derived from a prior recognition pass. Although the prior transcript in that case contains errors, adapting on that transcript disregarding that fact generally still results in accuracy improvements. Since the number of free parameters of the adaptation model are limited (e.g. in case of a linear transform, the matrix and offset parameters alone), the estimate has some robustness towards transcription errors. However, when the number of free parameters increase for example by using many regression classes in linear transform-based adaptation or by using MAP adaptation, the errors will limit the potential gain of adaptation.

Another approach to using the erroneous transcriptions in unsupervised adaptation is to include confidence scores and use these to filter or weight the data for adaptation parameter estimation. A lot of prior work has investigated the use of confidence based adaptation [5, 6, 7, 8, 9, 10, 11] and have shown that this approach can improve the recognition performance.

Here we investigate confidence based unsupervised acoustic model adaptation methods and analyze different uses of the confidence estimates and their relationship to the number of free adaptation parameters. In particular, we present the application of state posterior confidences for use in CMLLR, MLLR and MAP adaptation on a LVCSR task. In terms of confidence use, we compare confidence based selection and data weighting on frame-level.

Section 2 gives an overview of the recognition task and a detailed description of the recognition system. Section 3 describes the experimental results. First a performance bound is established through an oracle experiment simulating perfect state confidence scores. We then describe and analyze automatically derived confidences based on state posterior probabilities and uses these probabilities in different adaptation approaches. Performance is analyzed on a development set and the best performing setup is chosen for the final experiment on the evaluation set.

2. TASK AND SYSTEM DESCRIPTION

Internet multimedia content like pod-casts or videos is growing and therefore the demand for efficient search services for these. Automatic annotation can be used in addition to available meta data to build a search index. Automatic transcription system can be used to annotate speech content with text to allow content-based search of the spoken material.

The development of an automatic transcription system for such a task is difficult, even if we focus on a subset of the multimedia content. Narrowing the domain to video data of English speeches, the subset still consists of many different sub-domains like news, talk shows, lectures, etc. This means it is not possible to optimize

Table 1. Statistics for training and test corpora

	data-set		
	train	dev	eval
audio [h]	182.8	4.3	4.0
# running words	1,713,660	41,701	40,085
# segments man.	110,324	2,674	2,441
# segments auto.	–	4,706	4,742
# adapt. clusters	873	304	281
LM perplexity	–	155.6	170.0
OOV [%]	–	1.4	0.9

*Work was perfomed at Google.

Table 2. Proportion of running words per sub-domain ([%])

	CRS	ATASF	UCTV	TTALK	AUTHORS	UPLOAD	leftover
train	17.1	6.5	20.6	2.8	0.2	0.4	52.4
dev	49.3	19.7	19.4	11.2	–	0.4	–
eval	25.9	19.9	15.8	21.5	15.2	1.7	–

the automatic transcription system with a very specific lexicon, language model and acoustic model. Making the domain more specific would dramatically increase the data collection requirements and pose a large effort in system design. This work investigates the performance of a single transcription system and focus on acoustic model adaptation methods.

Table 1 and Table 2 give an overview of the speech corpora used for training, development and evaluation of the LVCSR system. Videos available at Google video search were selected, transcoded and resampled to 16 kHz mono audio data. The audio data was transcribed and divided into training, development and evaluation sets. Each set consists of audio tracks corresponding to whole video shows. Table 2 lists the proportion of the selected video sub-domains, e.g. *Charlie Rose* (CRS), *Tech Talks* (TTALK), *Authors@Google* (AUTHORS).

System Description:

- Speaker and domain independent acoustic model
- 39 dimensional acoustic vectors after applying LDA and semi-tied covariance transform on 9 time consecutive stacked 13 dimensional PLP features
- 3-state left-to-right HMM topology
- 41 phonemes + 4 noise models + silence state
- 8698 decision tree tied cross-word triphone states
- 16 component Gaussian mixture models with diagonal covariances
- Maximum likelihood training using Viterbi approximation
- 81k recognition vocabulary
- 4-gram language model with 9 million n-grams

Table 3 gives an overview of the performance of the baseline system applying commonly used adaptation methods. Prior to adaptation, the audio material was segmented and clustered into speaker clusters of about 30 seconds of speech per cluster using an algorithm very similar to the one described in [12]. We observe a gain of more than 10% relative from the speaker adaptive trained CMLLR (SAT-CMLLR) model over the speaker independent (SI) model. The use of MLLR on top of the CMLLR pass provides a small additional gain. For SAT-CMLLR and MLLR we are using a single transformation matrix and single bias vector.

The refinement of the SAT-CMLLR model with MAP adaptation leads to no performance improvements. Our conjecture is that even though MAP has many more free parameters and hence should have the potential to further improve performance of the adapted system, the transcription errors prevent that gain from materializing.

Table 3. Baseline recognition results (WER [%])

	dev	eval
1st pass: SI	40.6	45.8
2nd pass: SAT-CMLLR	36.3	38.5
3rd pass: MAP	36.3	38.5
3rd pass: MLLR	35.8	37.7

Table 4. Oracle adaptation results on dev. corpus (WER [%])

	2nd pass: SAT-CMLLR	3rd pass: MLLR	3rd pass: MAP
unsupervised baseline	36.3	35.8	36.3
perfect confidence	35.2	33.9	31.5
used frames [%]	60	64	64

3. APPLYING CONFIDENCE SCORES IN ADAPTATION

The most common way for unsupervised adaptation is the use of the automatic transcription of a previous recognition pass without the application of confidence scores. However, many publications have shown that the application of confidence scores for adaptation can improve recognition results. Small improvements for confidence based CMLLR adaptation is reported in [5]. In [8] the authors have investigated lattice-based MLLR applying a confidence threshold and report 2% relative improvement in word error rate (WER) over the 1-best transcription. In [7] 5% relative improvement is reported for MLLR adaptation by performing word confidence selection from the 1-best transcription.

3.1. Oracle experiments

In Table 4 we present an oracle experiment using SAT-CMLLR, MLLR and MAP adaptation. Here the perfect confidence experiments is taken at the frame-level. When aligning the reference and hypothesis state sequences against the observation sequence, we select a subset of the frames for which the state labels match and discard the frames for which the state alignments differ.

The results in Table 4 show performance bounds of using such an adaptation approach with this "perfect state confidence" and should be contrasted to the results in Table 3. Note however that the adaptation performance is not only dependent on the observations it is estimated on but also depends on the number of free parameters. For example, we could increase the MLLR regression classes which would lead to better oracle results. More important, these results show us the negative effect transcription errors have for the estimation of adaptation parameters. Especially the estimation of the many MAP adaptation parameters is prone to errors. Whereas SAT-CMLLR and MLLR are much more error robust due to the fewer parameters and the linear transformation constrains. On the other hand, due to the high number of free parameters MAP outperforms the MLLR method in our oracle experiments.

These results suggest the application of confidence scores for acoustic model adaptation and indicate the higher potential of MAP within our parameter setup for the adaptation methods.

3.2. State Posterior Confidence scores

In automatic speech recognition confidence scores can be developed and optimized for different units like utterances, words, phonemes or states. The optimization at the utterance level is a focus for dialog systems where a confidence based utterance rejection is applied. For unsupervised training or acoustic model adaptation it makes sense to focus on the tied state label since distributions are associated with

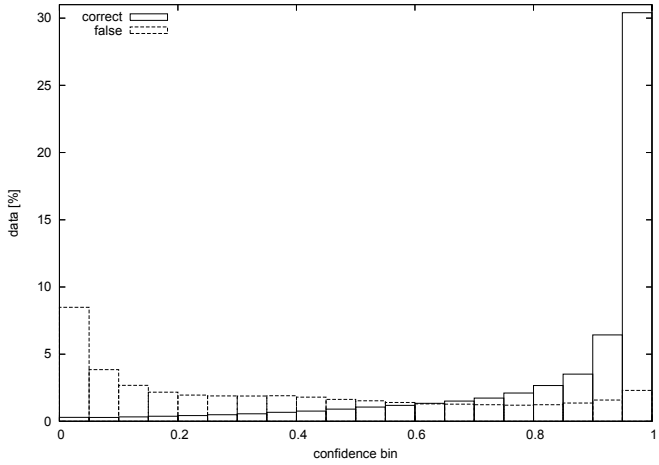

Fig. 1. State data histogram for lattice-weight confidence variant.

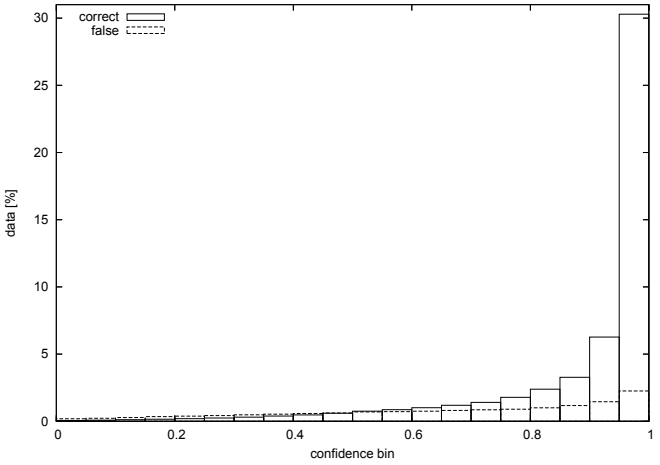

Fig. 2. State data histogram for 1-best-weight confidence variant.

these units. Instead of rejecting an entire utterance or word the system can use state confidence scores to select state-dependent data.

State confidence scores are obtained from computing arc posteriors from the lattice output from the decoder. The arc posterior is the fraction of the probability mass of the paths that contain the arc from the mass that is represented by all paths in the lattice. The posterior probabilities can be computed efficiently using the forward-backward algorithm as, for example, described in [13, 14].

3.3. Application of confidence score

Confidence scores can be used for observation selection through thresholding or observation weighting. We describe three different application variants of confidence scores which are commonly found in the literature for confidence based accumulation of observation statistics.

lattice-weight: Fractionally assign all observations to all states based on their posterior probabilities.

1-best-weight: Fractionally assign all observations to the states seen in the 1-best recognition hypothesis. In other words, intersect the lattice mentioned in the previous approach with the 1-best state sequence and assign observations based on that intersection.

1-best: Find the subset of observations for which the frame confidences exceed a threshold in the 1-best recognition hypothesis. If the confidence threshold is set to 0, this corresponds to the baseline method of unsupervised adaptation without confidences.

3.4. Evaluating state confidence scores

Given that the adaptation algorithms update state-tied distributions, our focus is on frame-level confidence measures and we aim to evaluate the confidence measures using the state as a unit. It is difficult to interpret the influence of word errors for the aligned state sequence or to state which of the deletion, insertion or substitution errors are the most harmful ones. Out-of-vocabulary words are also meant to be harmful for adaptation [7, 11] but even when a word is wrong, the pronunciation or most of the pronunciation can still be correct. An analysis of frame-level assignments will incorporate these considerations and give a more direct view of the effect errors will have on the adaptation sample.

We use the same acoustic model to compute the reference state alignment as we have used to generate the hypothesized state alignment from the automatic transcripts. In the case the phonetic transcriptions or the tied state sequence match we count no errors. However, if one state is mis-labeled it is not counted as a single error but as the sum of the wrongly assigned frames for that state.

In the cases of using fractional frame assignments, errors are counted fractionally as well for the evaluation statistics. The weighted error per frame is equal to its posterior probability in the interval $[0, 1]$.

3.5. Analysis experiments

For each adaptation method we applied the three different confidence variants and multiple thresholds on the development corpus.

To estimate the effectiveness of the different confidence variants for adaptation, we depict the histograms of the lattice-weight confidence variant and the 1-best-weight variant in Figure 1 and Figure 2 respectively. These histograms visualize the weighted sum of correct and falsely labeled frames based on state labels, distributed over 20 confidence bins. It can be observed that the weighted sum of mislabeled states is much higher for the lattice-weight confidence variant than for the 1-best-weight variant. This shows that, due to the quality of the confidence metric, thresholding the confidence scores results in improved accuracy with little loss in recall. As a result, adaptation approaches that have a large number of free parameters and hence will be sensitive to frame state-label errors are expected to benefit from confidence thresholding. Figure 3 shows the recognition performance for CMLLR, MLLR and MAP. For each adaptation approach, 3 curves show the performance when using the lattice-weight, 1-best-weight or 1-best approaches to incorporating confidence scores. The different points on the curves correspond to using different thresholds for selecting the adaptation data sample.

As expected, the limited number of free parameters in the linear transform-based adaptation schemes makes them fairly robust against mis-labeled frames as evident by the little effect on WER by changing the confidence threshold. However, MAP which has a much larger number of free parameters exhibits much better performance when only high confidence frames are used in the adaptation sample. It appears using a confidence threshold of 0.7 and the 1-best-weight approach results in the best performance.

3.6. Evaluation set experiments

In Table 5 we present the results of the confidence based MAP adaptation on top of the SAT-CMLLR model on the evaluation set.

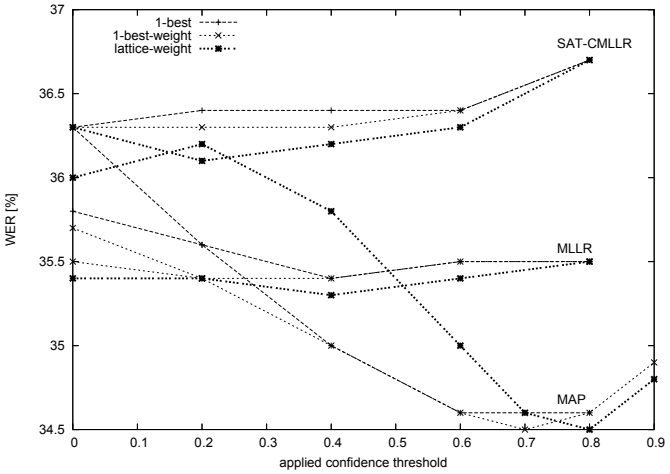

Fig. 3. Recognition performance for confidence based adaptation.

Here we have used the 1-best-weight state confidence variant with a threshold of 0.7 which was found to give good results on the development test set. It shows consistent improvements over use of CMLLR adaptation alone for the overall test set as well as for each sub-domain. Note that no gain was observed when using MAP without confidences showing the importance of using confidences.

4. CONCLUSION

The experimental results in this paper support the conjecture that linear transform-based adaptation approaches, limited in the number of free adaptation parameters, are somewhat robust to frame labeling errors as is the case in unsupervised adaptation on a partially accurate previous pass transcript. The experimental results also show that MAP adaptation implemented in such a condition results in little or no gain since the many free adaptation parameters will reinforce the errors present in the transcripts used for supervision.

The work shows, consistent with previous findings, that the use of confidence scores in adaptation leads to improved performance, in particular for adaptation approaches that have a large number of free parameters. Use of confidences in MAP adaptation shows a 2.1% absolute WER reduction over use of CMLLR alone whereas the use of MAP without confidences shows no gain. In addition, use of confidences in an adaptation approach that uses few free adaptation parameters like CMLLR or MLLR show little additional gain. It is to be expected that using linear transform-based adaptation approaches with multiple regression classes (effectively increasing the number of free adaptation parameters) will see a similar benefit from using confidence. Further analysis of the confidence scores shows that thresholding is an effective strategy for improving the frame labeling accuracy with little detrimental effect on the frame recall and hence has a beneficial effect on adaptation.

Table 5. Recognition results on test corpora (WER [%])

	Overall	CRS	ATASF	UCTV	TTALK	AUTHORS	UPLOAD
dev 2nd pass	36.3	25.4	37.4	38.5	77.4	–	21.2
+ MAP	34.5	24.5	35.2	36.7	72.1	–	20.7
eval 2nd pass	38.5	41.2	28.1	28.1	66.6	21.5	17.5
+ MAP	36.4	39.9	26.0	26.4	63.8	18.6	17.5

The use of adaptation algorithms for the fairly wide domain of English speech video data appears effective as it results in a 9.4% absolute WER reduction over an unadapted system on the evaluation set (45.8% vs. 36.4%).

5. REFERENCES

[1] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[2] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker Adaptation Using Constrained Reestimation of Gaussian Mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 357–366, 1995.

[3] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[4] J. Gauvain and C. Lee, "Maximum a-posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[5] T. Anastasakos and S.V. Balakrishnan, "The Use of Confidence Measures in Unsupervised Adaptation of Speech Recognizers," in *Proc. Int. Conf. on Spoken Language Processing*, Sydney, Australia, 1998.

[6] F. Wallhoff, D. Willett, and G. Rigoll, "Frame Discriminative and Confidence-driven Adaptation for LVCSR," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, 2000.

[7] M. Pitz, F. Wessel, and H. Ney, "Improved MLLR speaker adaptation using confidence measures for conversational speech recognition," in *Proc. Int. Conf. on Spoken Language Processing*, Beijing, China, 2000.

[8] M. Padmanabhan, G. Saon, and G. Zweig, "Lattice-Based Unsupervised MLLR For Speaker Adaptation," in *Proc. ISCA ITRW Automatic Speech Recognition: Challenges for the Millenium*, Paris, France, 2000.

[9] L. Uebel and P. Woodland, "Improvements in linear transforms based speaker adaptation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, 2001.

[10] J. Ogata and Y. Ariki, "Unsupervised acoustic model adaptation based on phoneme error minimization," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Orlando, USA, 2002.

[11] D. Wang and S.S. Narayanan, "A confidence-score based unsupervised MAP adaptation for speech recognition," in *Proc. of Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, USA, 2002.

[12] X. Zhu, C. Barras, S. Meignier, and J. Gauvain, "Combining Speaker Identification and BIC for Speaker Diarization," in *Proc. Int. Conf. on Spoken Language Processing*, Lisbon, Portugal, 2005.

[13] T. Kemp and T. Schaaf, "Estimating Confidence Using Word Lattices," in *Proc. European Conf. on Speech Communication and Technology*, Rhodes, Greece, 1997.

[14] F. Wessel, K. Macherey, and R. Schlüter, "Using Word Probabilities as Confidence Measures," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Seattle, USA, 1998.