IMPROVED MLLR SPEAKER ADAPTATION USING CONFIDENCE MEASURES FOR CONVERSATIONAL SPEECH RECOGNITION

Michael Pitz, Frank Wessel, Hermann Ney

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen – University of Technology
D-52056 Aachen, Germany
{pitz, wessel, ney}@cs.rwth-aachen.de

ABSTRACT

Automatic recognition of conversational speech tends to have higher word error rates (WER) than read speech. Improvements gained from unsupervised speaker adaptation methods like Maximum Likelihood Linear Regression (MLLR) [1] are reduced because of their sensitivity to recognition errors in the first pass. We show that a more detailed modeling of adaptation classes and the use of confidence measures improve the adaptation performance. We present experimental results on the VERBMOBIL task, a German conversational speech corpus.

1. INTRODUCTION

The adaptation of emission probability distribution parameters of speaker-independent speech recognition systems is a commonly used technique for improving the recognition performance. We distinguish between two main adaptation schemes: supervised adaptation, where the correct word transcription of the adaptation data is known, and unsupervised adaptation, where it is not known. In unsupervised adaptation a preliminary transcription is generated in a first recognition pass, which usually contains recognition errors. Adaptation with this erroneous transcription degrades the performance of the adaptation step compared to supervised adaptation. Confidence measures can be used to automatically label individual words in the preliminary transcription with either *correct* or *false* thus enabling the system and subsequent adaptation to use only those words which are most probably correct. This is especially significant in the context of conversational speech recognition, as the first pass tends to have higher WER than read speech.

Additionally, in conversational speech recognition there is usually less adaptation data available for each individual speaker, which also affects the adaptation performance as a detailed modeling of different classes for adaptation is difficult.

The remainder of the paper is organized as follows: first, we describe a more detailed approach for using different classes for MLLR adaptation. This approach allows us to use more classes for adaptation even with less adaptation data. We present experiments on the German VERBMOBIL task. Then we describe the computation of confidence measures based on word posterior probabilities. Finally, we describe the use of this confidence measure for MLLR adaptation. The paper will be concluded by a summary.

2. MAXIMUM LIKELIHOOD LINEAR REGRESSION

Within the MLLR adaptation framework, an affine transformation is used to adapt the acoustic parameters of the speech recognition system. The adaptation parameters are divided into several classes based on phonetic classes. The more adaptation classes are used, the more detailed is the adaptation but the more adaptation data is required. We will consider the case of unconstrained adaptation of continuous density Hidden Markov Models (CDHMM) with Gaussian emission probability distributions, i.e. only the means μ of the Gaussians are transformed:

$$\hat{\mu}_c = W_c \mu_c + b_c \ ,$$

where μ denotes the n-dimensional mean vector, W is an $n \times n$ matrix, and b is a bias vector. The subscript c denotes the adaptation class and $\hat{\mu}$ the adapted mean vector. To simplify the notation, the transformation can be written as

$$\hat{\tilde{\mu}}_c = \widetilde{W}_c \ \tilde{\mu}_c \ ,$$

with $\tilde{\mu} = [1, \mu]^T$. The transformation matrix $\widetilde{W}$ of dimension $n \times (n+1)$ is estimated using a Maximum Likelihood (ML) approach [1]:

$$\widetilde{W}_{ML} = \underset{\widetilde{W}}{\operatorname{argmax}} P(x_1^T | \widetilde{W})$$

with adaptation data x_1^T given as a sequence of acoustic feature vectors $x_1^T = x_1, \ldots, x_T$. For Gaussian emission probability distributions and Viterbi approximation (i.e. only the state with the highest probability is taken into account at each time frame), the solution becomes

$$\widetilde{W}_{ML} = \left(\sum_{t=1}^{T} x_t \, \mu_{s_t}^T \right) \cdot \left(\sum_{t=1}^{T} \mu_{s_t} \, \mu_{s_t}^T \right)^{-1} \ , \qquad (1)$$

where s_t denotes the most likely state at time t in the Viterbi path.

3. REFINED MODELING OF ADAPTATION CLASSES

In the conventional approach [1], an equal number of matrices and bias vectors is used for adaptation.It has been shown, however, that a more detailed approach can lead to better adaptation results [2]. We have studied the effect of using different numbers of classes for the MLLR matrix W and the bias vector b:

$$\hat{\mu}_c = W_{c'}\mu_c + b_c \ ,$$

where the adaptation class c' is a function of the more detailed classes c

$$c' = c'(c) \ ,$$

and the refined classes c themselves are functions of the HMM states $c = c(s)$ (i.e. several HMM states share the same adaptation class). The estimation formula remain almost the same, only the estimation of the bias vector b (being the first column in the extended matrix $\widetilde{W}$) has to be refined:

$$b_c = \frac{1}{T}\sum_{t=1}^{T}\left(x_t - W_{c'(c)}\mu_{s_t}\right) \qquad (2)$$

We carried out recognition tests on the German VERBMOBIL II corpus [3] (speaker independent, spontaneous speech, 10k word vocabulary, scheduling task). Statistics for this corpus are given in Table 1. We used the RWTH large vocabulary speech recognition system, which has been described in detail in [4]. The recognizer is a continuous Gaussian mixture density speech recognition system with a trigram Viterbi decoder. We use 3-state left-to-right HMM within-word models, a globally pooled variance vector, phonetically tied triphone models, and a linear discriminant analysis (LDA).

Table 1: Training and test corpora

Corpus	VERBMOBIL II	
	Training	Test
Name	CD1-41	DEV-99
Acoustic Data	61.5h	1.6h
Silence Portion	13%	11%
# Speakers	857	16
# Sentences	36,015	1,081
# Running Words	701,512	14,662
Perplexity (Trigram LM)	-	62.0

We tested several modelings of adaptation classes:

a) one global transformation

b) one global matrix/bias vector for phonemes and one matrix/bias vector for silence/noise

c) two matrices/bias vectors for phonemes and one matrix/bias vector for silence/noise

d) three matrices/bias vectors for phonemes and one matrix/bias vector for silence/noise

e) one matrix for all phonemes, one bias vector for each phoneme (i.e. 49 bias vectors) plus one matrix/bias vector for silence/noise

We found that using only few full matrices but more bias vectors performs better than using an equal number of matrices and biases (Table 2). With an equal number of transformations, the optimum was reached with 3 matrices and 3 biases (WER = 23.2%) and could not be further improved by increasing the number of parameters. The use of more classes for the biases gave an additional gain of 2.6% rel. (WER = 22.6%). We also run experiments with diagonal matrices but could not gain any further improvements. Block matrices, which are widely used and are often superior to full matrices, did not improve our results. The reason is the use of LDA as there is no specific order of cepstral coefficients in our acoustic feature vector any more after the LDA transform.

Table 2: Recognition results for different adaptation models

Number of classes for		N. of adpt.	WER [%]
matrices W	bias vectors b	parameters	
-	-	-	24.6
1	1	1122	23.6
2	2	2244	23.4
3	3	3366	23.2
4	4	4488	23.5
2	50	3828	22.6

4. CONFIDENCE MEASURES

The fundamental rule in all statistical speech recognition systems is Bayes' decision rule which is based on the posterior probability $p(w_1^M|x_1^T)$ of a word sequence $w_1^M = w_1, \ldots, w_M$, given a sequence of acoustic observations $x_1^T = x_1, \ldots, x_T$. That word sequence $\{w_1^M\}_{opt}$ which maximizes this posterior probability also minimizes the probability of an error in the recognized sentence:

$$\begin{aligned}\{w_1^M\}_{opt} &= \underset{w_1^M}{\operatorname{argmax}} \ \ p(w_1^M|x_1^T) \\ &= \underset{w_1^M}{\operatorname{argmax}} \ \ \left\{\frac{p(x_1^T|w_1^M)\cdot p(w_1^M)}{p(x_1^T)}\right\} \\ &= \underset{w_1^M}{\operatorname{argmax}} \ \ \{p(x_1^T|w_1^M)\cdot p(w_1^M)\ \} \ ,\end{aligned}$$

where $p(w_1^M)$ denotes the language model probability, $p(x_1^T|w_1^M)$ the acoustic model probability and $p(x_1^T)$ the probability of the acoustic observations. Strictly speaking, the maximization is also over all sentence lengths M.

If these posterior probabilities were known, the posterior probability $p(w_m|x_1^T)$ for a specific word w_m could easily be estimated by summing up the posterior probabilities of all sentences w_1^M containing this word at position m. This posterior word probability can directly be used as a measure of confidence [5, 6, 7].

Unfortunately, the probability of the sequence of acoustic observations $p(x_1^T)$ is normally omitted since it is invariant to the choice of a particular sequence of words. The decisions during the decoding phase are thus based on unnormalized scores. These scores can be used for a comparison of competing sequences of words, but not for an assessment of the probability that a recognized word is correct. This fact, and in other words the estimation of the probability of the acoustic observations, is the main problem for the computation of confidence measures.

For the following considerations it is very useful to introduce explicit boundaries between the words in a word sequence w_1^M. Let τ denote the starting time and t the ending time of word w. With these definitions, $[w;\tau,t]$ is a specific hypothesis for this word. A sequence of M word hypotheses can thus be formulated as $[w;\tau,t]_1^M = [w_1;\tau_1,t_1],\ldots,[w_M;\tau_M,t_M]$, where $\tau_1 = 1$, $t_M = T$ and $t_{n-1} = \tau_n - 1$ for all $n = 2,\ldots,M$. In order to determine these word boundaries, we consider the following modified Bayes' decision rule. $p([w;\tau,t]_1^M|x_1^T)$ denotes the posterior probability for a sequence of word hypotheses, given the acoustic observations and $p(x_1^T|[w;\tau,t]_1^M)$ the acoustic model probability:

$$\begin{aligned}
\left\{[w;\tau,t]_1^M\right\}_{opt} &= \underset{[w;\tau,t]_1^M}{\operatorname{argmax}}\; p([w;\tau,t]_1^M|x_1^T) \\
&= \underset{[w;\tau,t]_1^M}{\operatorname{argmax}}\; \frac{p(x_1^T|[w;\tau,t]_1^M)\cdot p(w_1^M)}{p(x_1^T)} \\
&= \underset{[w;\tau,t]_1^M}{\operatorname{argmax}}\; \frac{\prod_{m=1}^{M}\left[p(x_{\tau_m}^{t_m}|w_m)\cdot p(w_m|w_1^{m-1})\right]}{p(x_1^T)} .
\end{aligned}$$

We assume that the generation of the acoustic observations $x_{\tau_m}^{t_m} = x_{\tau_m},\ldots,x_{t_m}$ depends on word w_m only. With these word boundaries, the posterior probability $p([w;\tau,t]|x_1^T)$ for a specific word hypothesis $[w;\tau,t]$ can be computed by summing up the posterior probabilities of all sentences which contain the hypothesis $[w;\tau,t]$:

$$p([w;\tau,t]|x_1^T) = \\
= \sum_{\substack{[w;\tau,t]_1^M:\\ \exists\, n\in\{1,\ldots,M\}:\\ [w_n;\tau_n,t_n]=[w;\tau,t]}} \frac{\prod_{m=1}^{M}\left[p(x_{\tau_m}^{t_m}|w_m)\cdot p(w_m|w_1^{m-1})\right]}{p(x_1^T)} .$$

The posterior probability for a word hypothesis can be computed on the basis of word graphs. In the style of the forward-backward algorithm we compute the forward probability and the backward probability for a word hypothesis and combine both probabilities into the posterior probability of this hypothesis. In contrast to the forward-backward algorithm on a Hidden-Markov-Model state level, the forward-backward algorithm is now based on a word hypothesis level.

These posterior hypothesis probabilities turned out to perform poorly as a confidence measure. In fact, this observation is not surprising since the fixed starting and ending time of a word hypothesis determine which paths in the word graph are considered during the computation of the forward-backward probabilities. Usually, several hypotheses with slightly different starting and ending times represent the same word and the probability mass of the word is split among them. In order to solve this problem, the posterior probabilities of all those hypotheses which represent the same word have to be summed up. More details are given in [7].

5. USING CONFIDENCE MEASURES FOR MLLR ADAPTATION

If using an unsupervised two-pass adaptation strategy, the accuracy of the first pass is crucial for the adaptation performance. Recognition errors and out-of-vocabulary words degrade the adaptation to the new speaker or environment. Recently, the use of confidence measures for unsupervised speaker adaptation has been investigated [8, 9] with only little success.

In contrast, our experiments show a significant improvement in word error rate when using confidence measures as shown in Table 3. The experiments were carried out on the same data as in section 3, and 2 matrices and 50 biases were used for adaptation. We gain a relative improvement of 5% (WER 21.5%). We attribute this improvement to our refined confidence measures that are basically posterior probabilities derived from word graphs described in section 4. Similar relative improvements were reported in [10], however only on clean, read speech (WSJ0 task) and with a relatively high baseline word error rate for this task.

We used the confidence measures as follows: during the first recognition pass, a word graph was generated and a confidence score was computed for each word in the recognized sentence. When estimating the adaptation matrix W and the bias vector b, only those time frames belonging to words with confidence score higher than a given threshold were taken into account in equation (1) and (2). This means all time frames for a word with low confidence were not considered for adaptation.

To further investigate the performance of our method, we present a "cheating" experiment where we used only the correctly recognized words for adaptation ("ideal" confidence measure). This experiment shows the maximum possible improvement by using confidence measures. As can be seen from Table 3, the result obtained by using computed confidence measures is close to the result of "ideal" confidence measures (21.5% WER compared to 21.0% WER).

Another interesting contrast experiment is to use the correct transcription of the test data in the first pass for a supervised adaptation. Evidently, this is not a feasible experiment for real applications but gives more insight in the performance MLLR: This experiment gives an upper boundary for possible improvements to be gained by MLLR adaptation at all for a given modeling of adaptation classes (WER = 18.3%). When we compare supervised MLLR with the "ideal" confidence measures MLLR, there is still a major difference in recognition performance.

We interpret this result as follows: first, with supervised adaptation there is in the order of 20% more adaptation data available, not just the words that were correctly recognized in the first pass. More important, however, seems to be the ability of supervised MLLR to adapt to also those critical words which would be incorrectly recognized in the first pass of unsupervised MLLR.

Table 3: Improvements obtained by using confidence measures

adaptation method	WER[%]
no adaptation (baseline)	24.6
without conf. meas.	22.6
with conf. meas.	21.5
with "ideal" conf. meas.	21.0
supervised	18.3

6. SUMMARY

In this work we have shown that a more sophisticated modeling of adaptation classes leads to better recognition results even with less adaptation parameters. We have also shown that the use of confidence measures improves MLLR adaptation and that posterior probabilities derived from word graphs perform nearly as good as "ideal" confidence measures (using only correctly recognized words for adaptation). Further we have found that supervised adaptation is still superior to a two-pass strategy with "ideal" confidence measures. The reason is that supervised MLLR is able to reduce the mismatch between the acoustic models and the acoustic vectors of those words which would have been incorrectly transcribed in a first pass.

7. REFERENCES

[1] C.J. Leggetter, P.C. Woodland: "Maximum Likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer, Speech and Language,* vol. 9, pp. 171–185, 1995.

[2] V. Digalakis, S. Berkowitz, E. Bocchieri, C. Boulis, W. Byrne, H Collier, A. Corduneanu, A. Kannan, S. Khudanpur, A. Sankar: "Rapid Speech Recognizer Adaptation to New Speakers" *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Phoenix, AZ, USA, pp. 765–768, May 1999.

[3] T. Bub and J. Schwinn: "VERBMOBIL: The evolution of a complex large speech-to-speech translation system," in *Proceedings International Conference on Spoken Language Processing 1996*, Philadelphia, PA, USA, October 1996, vol. 4, pp. 2371–2374.

[4] H. Ney, L. Welling, S. Ortmanns, K. Beulen, F. Wessel: "The RWTH Large Vocabulary Continuous Speech Recognition System", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, WA, USA, pp. 853–856, May 1998.

[5] F. Wessel, K. Macherey, R. Schlüter: "Using Word Probabilities as Confidence Measures", *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Seattle, WA, USA, pp. 225–228, May 1998.

[6] F. Wessel, R. Schlüter, H. Ney: "Using Posterior Probabilities for Improved Speech Recognition", *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, pp. 1587–1590, June 2000.

[7] F. Wessel, R. Schlüter, K. Macherey, H. Ney: "Confidence Measures for Large Vocabulary Continuous Speech Recognition" *IEEE Trans. on Speech and Audio Processing*, to be published, 2000.

[8] T. Anastasakos, S.V. Balakrishnan: "The Use of Confidence Measures in Unsupervised Adaptation of Speech Recognizers", *Proc. Int. Conf. on Spoken Language Processing*, Sydney, Australia, pp.2303–2306, Dec. 1998.

[9] P. Ngyen, Ph. Gelin, J.-C. Junqua, J.-T. Chien: "N-Best Based Supervised and Unsupervised Adaptation for Native and Non-Native Speakers in Cars", *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* Phoenix, AZ, USA, pp.173–176, May 1999.

[10] F. Wallhoff, D. Willet, G. Rigoll: "Frame-Discriminative and Confidence-Driven Adaptation for LVCSR" *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, pp. 1835–1838, June 2000.