The Use of Confidence Measures in Unsupervised Adaptation of Speech Recognizers

Tasos Anastasakos, Sreeram V. Balakrishnan

Motorola, Lexicus Division
3145 Porter Drive, Palo Alto CA 94304
E-mail: `tasos@lexicus.mot.com`

ABSTRACT

Confidence estimation of the output hypothesis of a speech recognizer offers a way to assess the probability that the recognized words are correct. This work investigates the application of confidence scores for selection of speech segments in unsupervised speaker adaptation. Our approach is motivated by initial experiments that show that the use of mis-labeled data has a significant cost in the performance of particular adaptation schemes. We focus on a rapid self-adaptation scenario that uses only a few seconds of adaptation data. The adaptation algorithm is based on an extension to the MLLR transformation method that can be applied to the observation vectors. We present experimental results of this work on the ARPA WSJ large vocabulary dictation task.

1. INTRODUCTION

Recently, confidence estimation of the output string of a speech recognition system has become an important topic of research [4, 5, 8, 9]. As speech recognition systems find their way into real world applications, confidence provides a way to assess the imperfect recognition results, and detect out-of-vocabulary (OOV) words or generate repair dialogs in a natural language system. In this paper, we investigate the use of confidence annotation of the recognizer output in an unsupervised adaptation scheme.

In many applications it is not feasible to obtain adaptation data of the new condition or speaker prior to the use of the system. In such cases, the adaptation data consist of utterances of the speaker that are spoken during the transaction. This *on-line* adaptation process makes use of the data as they sequentially become available, adjusting the system parameters dynamically to the speaker. Typically these methods operate in unsupervised mode, that is, the correct transcription of the speech data is not known. Instead, the most likely hypothesis that is generated by the recognizer is used to align the speech waveforms for the adaptation process.

We examine a rapid adaptation scheme, *self-adaptation,* that uses only a few seconds of data. In this approach, we use the speech data and recognition result of a single sentence to adapt the system and then recognize the same sentence again. We present experimental results that show that self-adaptation provides a significant reduction in the word error rate of 10%. Previous adaptation results [6] using several minutes of adaptation data have shown little performance differences between supervised and unsupervised adaptation schemes. The effectiveness of this particular adaptation process is greatly affected by the mis-labeled data due to the limited amount of adaptation data. This observation motivates our use of confidence metrics to guide the adaptation process by selecting or emphasizing speech segments with high confidence.

The rest of the paper is organized as follows: in section 2 we briefly outline the adaptation transform that we used throughout our experiments, in section 3 we define the confidence measures and in section 4 we describe the experimental setup and the current results of our work.

2. CONSTRAINED MODEL-SPACE ADAPTATION

Adaptation methods have become important components of large vocabulary speech recognition systems as they compensate for mismatches between training and testing conditions, which are caused due to different speaker characteristics and channel or environment conditions. Model-based approaches, such as *Maximum A Posteriori* (MAP) estimation [3] and linear regression adaptation [1, 7] have shown significant improvements in recognition accuracy by adjusting the parameters of a speaker independent (SI) system based on speech material (adaptation data) which is representative of the new condition. In this work we have applied an extension of the Maximum Likelihood Linear Regression (MLLR) approach termed constrained model-space adaptation. It is a maximum likelihood matrix transformation that is applied to the means and the variances of the Gaussian densities, and it is constrained in the sense that the transformation applied to the variances must correspond to that of the mean vectors. One advantage of the constrained transformation compared to the original MLLR is that it can be applied to the observed feature vectors, thus avoiding the computationally expensive update of the model parameters. This transform has been originally proposed in [2] where the interested reader will find a detailed presentation of the topic.

The constrained model-based transform has the general form

$$\hat{\boldsymbol{\mu}} = \boldsymbol{A}_M \boldsymbol{\mu} - \boldsymbol{b}_M \quad (1)$$

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{A}_M \boldsymbol{\Sigma} \boldsymbol{A}_M^T \quad (2)$$

In [1], the problem was solved for the diagonal transformation case. Our application follows [2] where a solution for the full matrix case is provided assuming that the original models have diagonal covariance. It is easily shown that the transformation of the Gaussian parameters corresponds to an equivalent transformation of the feature vector $\boldsymbol{o}_t$:

$$\hat{\boldsymbol{o}}_t = \boldsymbol{A}\boldsymbol{o}_t + \boldsymbol{b} = \boldsymbol{A}_M^{-1}\boldsymbol{o}_t + \boldsymbol{A}_M^{-1}\boldsymbol{b}_M \tag{3}$$

It is now evident that the constrained model-space transform may be implemented as a transformation of the observed feature and the likelihood of an observation $\boldsymbol{o}_t$ for a particular Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is computed as:

$$\mathcal{L}(\boldsymbol{o}_t; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}, \boldsymbol{b}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{o}_t + \boldsymbol{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \log(|\boldsymbol{A}|) \tag{4}$$

thus avoiding the computationally expensive update of the model parameters.

3. CONFIDENCE METRICS

Most speech recognizers assign scores to word and sentence hypotheses that are not absolute measures of probability but rather relative measures used to rank order hypotheses. Therefore the recognizer scores are not comparable across different sentences and not useful as confidence measures. Instead, we compute two confidence metrics, one derived from word lattice densities and one based exclusively on acoustic scores.

3.1. Lattice Density

During recognition, hypotheses whose likelihood scores fall below certain thresholds are considered unlikely and pruned from the search space. In time segments where the likelihood for a particular word is much higher than the likelihood of other competing hypotheses, many of the competing other words are pruned. Conversely, if a large number of words has similar likelihood, the number of propagating hypotheses will be relatively high. It has been shown in the literature [4, 5, 8, 9] that the uncertainty of the recognizer expressed by the number of hypotheses in a time segment correlates with the word error rate for that time segment. For each word in the word lattice we compute, CM_{lat}, the number of competing hypotheses that end at the same time, normalized in the range $[0, 1]$.

3.2. Acoustic Confidence Measure

The CM_{lat} confidence metric depends implicitly on the language model and dictionary constraints that control the breadth of the recognition search space. Our second confidence metric CM_{ac} is solely based on acoustic scores and can be applied either on the word or the phone level. It is defined as the posterior probability that a particular phone or word w is uttered during a time segment, given the sequence of acoustics observations $\boldsymbol{O}$ for that segment

$$CM_{ac} = P(w|\boldsymbol{O}) = \frac{P(\boldsymbol{O}|w)\,P(w)}{\sum_{q \in Q} P(\boldsymbol{O}|q)\,P(q)} \tag{5}$$

where Q is the set of all possible phone sequences in the time segment.

For computational expediency, we estimate a set of monophone Gaussian mixture densities, so that each monophone is modeled by a 3-state HMM with all states tied to the same Gaussian mixture density. Using Viterbi alignment, we find the phone-frame correspondence of the recognizer output and compute the numerator using the monophone Gaussian densities. By tying all states of the monophone models, the denominator is found by a fairly inexpensive dynamic programming application while enforcing a minimum three frame time constraint. In practice we further simplified the denominator by computing the maximum score instead of the sum. In this work we assume that all prior probabilities are uniform. This formulation can be extended by incorporating statistics for the phone sequences in the form of bigram statistics at the phone level.

4. EXPERIMENTS AND DISCUSSION

The baseline recognition system is a speaker independent, continuous density, tied-state, cross-word triphone HMM system developed at Motorola, Lexicus. The speech was parameterized into a 39 dimensional feature vector that includes 12 MFCCs, the normalized log energy and the first and second differences of these parameters. The acoustic training data consists of 7,200 sentences from the SI-84 WSJ0 corpus. The resulting system has approximately 30,000 Gaussians. The recognition experiments were conducted on the 20,000 word open vocabulary and the 5,000 word closed vocabulary sets from the November 1992 DARPA evaluation that consist of 333 and 330 sentences respectively. A time-synchronous single pass decoder using the standard bigram languange models supplied the data was used in the experiments. In order to perform Viterbi alignment of the reference word transcription, the 20K word pronunciation dictionary is augmented to include any words of the reference transcription which would otherwise be OOV.

We first calculated the confidence score estimates CM_{lat} and CM_{ac}. The word and phone sequences hypothesized during recognition are aligned to the reference word and phone sequence in order to label each decoded hypothesis as either correct or incorrect. The labeling takes into account time information and marks as incorrect segments with less than 80% overlap.

Fig. 1 shows the histograms of the lattice density based confidence scores CM_{lat} for the correctly and the incorrectly recognized hypotheses for the 20K test set. The second plot in Fig. 1 shows the cumulative probability functions of the confidence scores for the correct and incorrect hypotheses. Fig. 2 shows corresponding plot for the phone-level acoustic confidence. These graphs show that CM_{lat} is a indicator of word confidence, a result consistent with [4, 9]. The acoustic based confidence, even though it provides less separation between correct and incorrect hypotheses, was found useful in selecting phone segments that are correctly labeled in an otherwise mis-recognized word which simply matches part of the pronunciation of the reference word.

We conducted a number of experiments to evaluate the effect of supervision in the adaptation process in the incremental and the self-adaptation schemes. During incremental adaptation, the adaptation parameters are updated after the recognition of a test utterance using

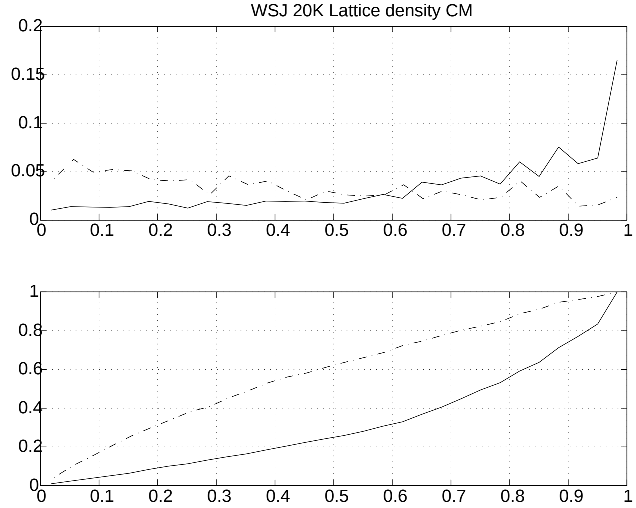

Figure 1: Distribution of lattice density based confidence scores for correctly labeled (solid line) and incorrectly labeled (dashed line) hypotheses. The second subplot shows the cumulative probability functions of the two distributions

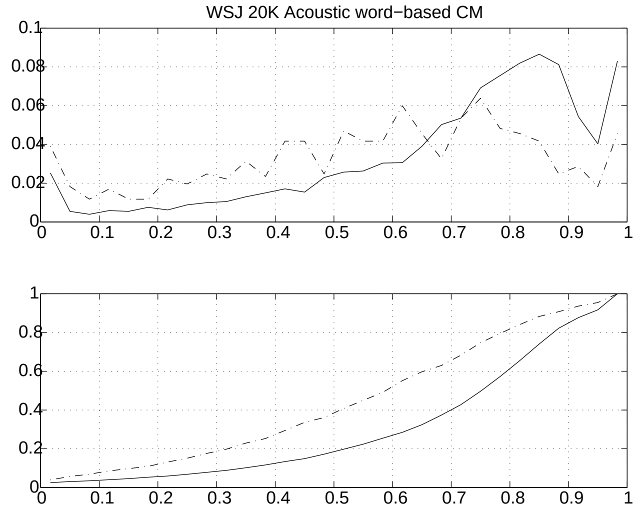

Figure 2: Distribution of phone-level acoustic confidence scores for correctly labeled (solid line) and incorrectly labeled (dashed line) hypotheses. The Second subplot shows the cumulative probability functions of the two distributions

the accumulated statistics of all incoming test utterances up to that point. The updated parameters are used to recognize the following utterance. These statistics are gathered by aligning the observation sequence to the recognized hypothesis in the unsupervised mode or to the reference word transcription in the supervised model. Table 1 summarizes the results of these experiments. These results show that supervised adaptation has little or no effect in the adaptation process, possible due to the relatively low error rate of the recognized hypotheses. In the self-adaptation scheme, each utterance is initially recognized using adaptation parameters that are initialized to identity. Then the adaptation parameters estimated using the statistics from the alignment of the recognized hypothesis. The adaptation transform is applied to the observation stream and the utterance is recognized again. To evaluate the effect of supervision, we performed a wizard experiment whereby we labeled the recognized hypothesis with correct and incorrect tags based on the reference word transcription. We then used only the time segments that were labeled as correct to estimate the adaptation parameters. The results of these experiments are shown in Table 2. We observe that adaptation on a few seconds of speech is enough to provide a significant reduction in word accuracy, comparable to the incremental adaptation scenario that uses considerably more data (at least for the 20K test). Furthermore, the wizard experiment shows that misrecognized segments greatly affect the performance of the adaptation due to the limited amount of data, which is an encouraging result for the application of confidence metrics to guide the adaptation process.

Condition	05K	20K
Baseline	7.7%	12.5%
Unsupervised Incr. Adaptation	6.0%	10.8%
Supervised Incr. Adaptation	6.0%	10.7%

Table 1: Effect of supervision on the adaptation process for incremental adaptation on the ARPA November 1992 05K and 20K test sets. Adaptation is performed using the constrained model-space transform method.

Condition	05K	20K
Baseline	7.7%	12.5%
Unsupervised Self Adaptation	6.9%	11.0%
Correct-only Self Adaptation	6.5%	10.3%
Confidence + Self Adaptation	6.7%	10.8%

Table 2: Effect of supervision on the adaptation process for self-adaptation on the ARPA November 1992 05K and 20K test sets. Adaptation is performed using the constrained model-space transform method. Correct-only self adaptation indicates the wizard experiment, where mis-recognized segments are discarded.

We applied a heuristic selection of speech segments for adaptation based on the word-level lattice density confidence scores and the phone-based acoustic confidence scores. We rejected all segments that correspond to words with CM_{lat} below a threshold θ_{lat}, but retained any phone subsegments that had CM_{ac} higher than an acoustic confidence threshold $\theta_{ac,1}$. Similarly, for the speech segments with CM_{lat} greater than θ_{lat} that are accepted, we discarded phone subsegments that had CM_{ac} lower than a second threshold $\theta_{ac,2}$ such that $\theta_{ac,2} < \theta_{ac,1}$. These confidence thresholds are experimentally determined based on the distribution of the confidence scores. Unfortunately, our results so far (Table 2) have shown small incremental improvements over the unsupervised scenario.

5. CONCLUSIONS

We reported on the application of the constrained model-space transform, a new formulation of the Maximum Likelihood Linear Regression transform for speaker adaptation. An attractive property of this approach is that it can be applied as a transformation on the observation space, thus incurring little computational cost for on-line adaptation schemes. We then examined the effect of supervision on two on-line adaptation schemes, incremental and self-adaptation. Our experimental results showed that the use of the reference word transcription provides very little additional benefit in the context of on-line incremental adaptation.

Self-adaptation, the process of adapting on a single utterance and then recognizing this utterance again, is better suited for very short interactions with a speech recognition system, where the system needs to adapt rapidly based on a few seconds of speech. In this case the use of mis-labeled speech segments greatly affect the performance of adaptation. We proposed the use of two confidence measures to discard speech segments with low confidence that probably correspond to mis-recognitions. We used a word-based lattice density metric and a phone-based acoustic confidence metric for our experiments. The results have only shown marginal improvement over the unsupervised self-adaptation case. This could be attributed to the accuracy of the particular confidence metrics and the heuristics that we employed for the selection of the high confidence speech segments. Our current work addresses these issues, to improve the accuracy of the confidence scores and the selection algorithm, as well as investigate alternative confidence metrics and decision strategies.

6. REFERENCES

1. V. Digalakis, D.Rtichev, and L.G. Neumeyer, “Speaker adaptation using constrained estimation of Gaussian mixtures”, *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 357–366, September 1995.

2. M.J.F. Gales, “Maximum likelihood linear transformations for HMM-based spech recognition”, Tech. Rep. CUED/F-INFENG/TR 291, Cambridge University Engineering Department, May 1997.

3. J.-L. Gauvain and C.-H. Lee, “Maximum a posteriori estimate for multivariate Gaussian mixture observations of Markov chains”, *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, April 1994.

4. L. Gillick, Y. Ito, and J. Young, “A probabilistic approach to confidence estimation and evaluation”, in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997, pp. 879–882.

5. T. Kemp and T. Schaaf, “Estimating confidence using word lattices”, in *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, 1997, pp. 827–830.

6. C.J. Leggetter and P.C. Woodland, “Speaker adaptation of continuous density HMMs using multivariate linear regression”, in *Proceedings of International Conference in Spoken Language Processing*, 1994.

7. 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, pp. 171–185, 1995.

8. M. Siu, H. Gish, and F. Richardson, “Improved estimation, evaluation and applications for confidence measures for speech recognition”, in *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, 1997, pp. 831–834.

9. G. Williams and S. Renals, “Confidence measures for hybrid HMM/ANN speech recognition”, in *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, 1997, pp. 1955–1958.