IMPROVED UNSUPERVISED SPEECH RECOGNITION SYSTEM USING MLLR SPEAKER ADAPTATION AND CONFIDENCE MEASUREMENT

Mukund Jha[a], Sourabh Sriom[b], Míriam Luján[c], Carlos D. Martínez-Hinarejos[c], Alberto Sanchís[c]

[a]Department of Computer Science, MNNIT, Allahabad, 211004, Allahabad, India
[b]Department of Electronics and Communications, IIT Guwahati, 781039, Guwahati, India
[c]Instituto Tecnológico de Informática, Universidad Politécnica de Valencia,
Camino de Vera, s/n, 46022, Valencia, Spain

ABSTRACT

A robust ASR system needs to perform well in different environment and with different speakers. For this reason speaker adaptation has become an essential part of a state of art ASR system. Here we show how confidence measurement technique can be used to improve the quality of unsupervised speaker adaptation. An initial speaker-independent system is adapted to improve the modelling of a new speaker by modifying HMM parameters using Maximum Likelihood Linear Regression(MLLR) technique. Improvement gained from unsupervised speaker adaptation technique are lowered because of their dependency on the accuracy of recognition in first pass. We use confidence measures to improve the performance by selective adaptation. We present experimental results on the 8 speakers' data from Wall Street Journal.

1. INTRODUCTION

Even after significant amount of improvement in speaker independent speech recognition systems, error rates are still quite higher when compared to speaker dependent (SD) systems, and dependence on large amount of speaker specific data for training make SD unsuitable for many application. Many system make use of speaker adaptation techniques to adapt to new speakers. These techniques can be either supervised, where the correct word transcription of the adaptation data is known or unsupervised, where it is not known. Unsupervised adaptation relies on recognizer to provide a transcript for the spoken utterances in the first pass, which is used to adapt the model during the training. But these transcriptions contain recognition errors and out-of-vocabulary words which degrade the performance of adaptation technique. Confidence measures can be used to classify the words in the recognized transcript as correct or incorrect, which allows the system to use only those words for adaptation which are most probably correct.

The remainder of the paper is organized as follows: first, we give a description of the MLLR technique for speaker adaptation, next we give a brief description of the confidence measurement technique and describe the use of confidence measurement for MLLR adaptation, followed by the details of experiments. We conclude this paper with results and a summary of the work.

2. MAXIMUM LIKELIHOOD LINEAR REGRESSION

Speaker adaptation applied to HMMs mostly involve the techniques that uses the original models as the starting point and add speaker specific information by transforming some of the parameters in the models. The general idea is that the fully trained model should contain some general speech information which will be used for the new system as well. It is also assumed that even the smallest amount of adaptation data would contain some speaker specific information.

The MLLR technique follows the above assumptions for adapting only the mean vectors of continuous density HMMs. However, the adaptation can also be performed for the covariance matrices to improve the results. If a transformation matrix can be estimated specifically for the covariance matrix, it is likely that an improvement in performance can be achieved by transforming the covariances as well. A detailed discussion on this is presented in [1, 2]. In this section we shall discuss the general theory behind the MLLR technique and its evaluation given a set of observation data.

The MLLR adaptation involves transforming the means of the HMM Gaussians. This transformation is preformed by applying a *transformation matrix W*. Therefore, given a gaussian s with mean μ_s, the adaptation consists of re-evaluating the new mean $\hat{\mu_s}$ as below:

$$\hat{\mu_s} = W_s \mu_s \tag{1}$$

where W_s is the adaptation matrix for the gaussian s and an *offset*(or bias) value, ω_s is introduced in the mean vector. This gives us the *extended mean vector*, $\tilde{\mu} = [\omega_s : \mu_s]$. Now the equation (1) can be modified to:

$$\hat{\tilde{\mu}_s} = \tilde{W}_s \tilde{\mu_s} \tag{2}$$

here, if the dimension of μ_s is n, the dimension of $\tilde{W}_s$ would be $n \times (n+1)$.

The transformation can be evaluated for each gaussian in the acoustic models. However, this would require a huge amount of data for the adaptation process. To solve this problem we group the gaussians into what is referred to as *regression classes*, which are the sets of gaussians which share the same transformation matrix. Regression classes are discussed in detail in [3].

Therefore, for a given set of adaptation samples from a particular regression class c, denoted by the sequence of acoustic features vectors $x_1^T = x_1, x_2, ..., x_T$, the adaptation matrix $\hat{\tilde{W}}$ for a Viterbi approximation can be estimated as below:

$$\hat{\tilde{W}}_c = (\sum_{t=1}^{T} x_t \mu'_{s_t})(\sum_{t=1}^{T} \mu_{s_t} \mu'_{s_t})^{-1} \quad (3)$$

where s_t denotes the most likely state (and gaussian) in the Viterbi path at time t and μ' is the transpose of the mean vector.

3. CONFIDENCE MEASUREMENT

The speaker recognition systems that are available to us are not completely free of errors. To develop an efficient speaker independent speech recognition system using MLLR approach, we must have the knowledge about the reliability of the recognized words. Therefore the goal of confidence measurement is to detect words that are likely to have errors in their recognition. In other words confidence measurement would be used for each hypothesized word to classify it as either *correct* or *incorrect*. Such a classification is done using *confidence measures*, which are essentially normalized scores to help the system decide on the reliability of the recognized words. Finally, only those words, which have been tagged as correct would be used for the MLLR adaptation and hence yield better results in the subsequent recognitions.

The Bayes' decision rule is the fundamental rule in all statistical speech recognition systems. The Bayes' rule is based on the posteriori probability $p(w_1^M|x_1^T)$ of a word sequence $w_1^M = w_1, w_2, ..., w_M$ given a sequence of acoustic observations $x_1^T = x_1, x_2, ..., x_T$. That word sequence $[w_1^M]_{opt}$ which maximizes this posteriori probability would also minimize the probability of an error in the recognized sentence:

$$[w_1^M]_{opt} = argmax_{w_1^M} p(w_1^M|x_1^T)$$

$$= argmax_{w_1^M}[\frac{p(x_1^T|w_1^M).p(w_1^M)}{p(x_1^T)}]$$

$$= argmax_{w_1^M}[p(x_1^T|w_1^M).p(w_1^M)]$$

where, $p(w_1^M)$ denotes the language model probability, $p(x_1^T|w_1^M)$ the acoustic model probability and $p(x_1^T)$ is the probability of acoustic observations.

If all these posteriori probabilities are known to us, the posteriori probability $p(w_m|x_1^T)$ for a specific word w_m could be estimated by summing up the posteriori probabilities of all sentences w_1^M containing this word at position m. This posterior word probability can now be used as an efficient measure of confidence.

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. Thus, the decisions during the decoding phase are based on unnormalised scores. These scores can be used for a comparison of competing sequences of words, but can not be used to predict which of the recognized words are correct. The estimation of probability of the acoustic observations thus, is the main problem for the computation of confidence measures.

The usefulness of word graphs in confidence measurement is well known. In [4] the proposed features based on word graphs are the most important predictors. In [5] the confidence measure is estimated on word graphs directly by the posterior probability of a hypothesized word given all the acoustic observations of the utterance. The word posterior probability based on word graphs is used in [6] along with a large set of other authors use a single word graph which is obtained through the recognition process. We have used single word graphs for the evaluation of confidence measures in our experiments and an overview of estimating the posterior probabilities based on a single word graph is discussed below.

3.1. Posterior probabilities on word graphs

A word graph G is a directed, acyclic, weighted graph. The nodes corresponds to discrete points in time. The edges are triplets $[w,\ s,\ e]$, where w is the hypothesized word from node s to node e. The weights are scores associated to the word graph edges. Any path from the initial to the final node forms a hypothesis h.

Given the acoustic observations $\vec{\Theta}_1^T$, the posterior probability for a specific word (edge) $[w,\ s,\ e]$ can be computed by summing up the posterior probabilities of all hypotheses of the word graph containing the edge $[w,\ s,\ e]$:

$$P([w,\ s,\ e] \mid \vec{\Theta}_1^T) = \frac{1}{P(\vec{\Theta}_1^T)} \sum_{\substack{h \in G: \\ \exists [w', s', e']: \\ w'=w, s'=s, e'=e}} P(h,\ \vec{\Theta}_1^T) \quad (4)$$

The probability of the sequence of acoustic observations $P(\vec{\Theta}_1^T)$ can be computed by summing up the posterior probabilities of all word graph hypotheses:

$$P(\vec{\Theta}_1^T) = \sum_h P(h\ , \vec{\Theta}_1^T) \quad (5)$$

These posterior probabilities can be efficiently computed based on the well-known *forward-backward* algorithm [5].

3.2. Scaling of the probabilities

In the evaluation of confidence measures for the recognized words we also include a scaling factor α which plays an important role in the evaluation of the posterior probabilities and their performance as a confidence measure. If the acoustic model probabilities are not scaled appropriately, the sums of the equations mentioned above will be dominated by only a few word graph hypotheses because of very large dynamic range of the acoustic scores (which is the negative logarithm of the unnormalized acoustic probabilities). The differences in acoustic scores arise mainly due to the variance of acoustic features which are presumable underestimated. To avoid the re-evaluation of these variances, it is better to scale the acoustic probabilities in order to have efficient results. Training data for each speaker was tested for different values of α to decide on the optimum α value for each speaker, α giving the minimum *Classification Error Rate (CER)* for a speaker was taken to be optimum. Same value of scaling factor was used during the testing phase for each speaker.

3.3. Threshold for confidence scores

The confidence measurement technique gives us a score (or probability) of a word's reliability. The system now sets a certain threshold τ, a value between 0 and 1 (probability score) for each speaker in the corpus after analyzing the *Classification Error Rate (CER)* for each speaker separately. Threshold for a given speaker and given α is the probability score which gives the minimum *CER*. All words recognized for a given speaker, which have their confidence measures above this threshold are classified as correct while the ones having their measures below this threshold are labeled incorrect. The correct words are now the ones that are used for the MLLR adaptation.

4. EXPERIMENTS

We performed some interesting experiments with confidence measures and MLLR technique. Wall Street Journal was used for all the experiments. Recognition was performed using a trigram language model and 20k lexical model using iATROS, an HMM based continuous speech recognizer. First different parameters of the recognizer were optimized to give low word error rate. Acoustic scaling factor, α, used for confidence measurement, was optimized for each corpus to give better confidence scores and baseline word error rates were determined for each speaker.
Experiments were made with the following kind of trainings.

1. **Full MLLR**: This is the standard MLLR adaptation. We use all the time frames to train the models, thus includes error from the recognizer. This gives us a base on MLLR adaptation on which we try to improve using confidence measure.

2. **MLLR with Confidence Measure**: In this we apply confidence measure on the output of the recognizer before estimation of the adaptation matrix. Only time frame of high confidence words were used for adaptation. Experiment was done with two different types of gaussian means.

   - **Max**: It is the standard method, means from the most probable gaussian in the gaussian mixture of the HMM state were used during adaptation.
   - **Normalized**: Instead of using the means only from the most probable gaussian and ignoring the means from the rest of the mixture, we use a normalized mean from all the gaussians of the mixture. We normalize the means of the gaussian mixture by taking the means from each mixture in the ratio of its emission probability, i.e. in ration of its contribution to the state emission probability.
     Assume a mixture having N gaussians and let μ_{ij} denote the *ith* mean from *jth* mixture, let p_j be the probability of emission of the observed feature frame by *jth* gaussian and p_t be the probability of emission of the state, sum of emissions of all gaussians. Then normalized mean μ_i' is given by the following expression.

     $$\mu_i' = \sum_{j=1}^{N} \frac{p_j}{p_t} \mu_{ij}$$

     Normalized means has given better results for some speakers than using the means from maximum gaussian.

3. **MLLR with ideal CM**: We performed this experiment to find the upper limit which can be achieved through MLLR in unsupervised training. In this only the frames of correctly recognized words were used in calculating the adaptation matrix. We used the correct transcription to compare the recognition output and only correctly recognized words were used for the adaptation. Experiments were performend using means from the most probable gaussian (Max). The results were very close to the supervised training.

4. **Supervised (Ideal Recognizer)**: Lastly, we performed supervised training on the corpus. This was done by performing forced recognition with the actual transcription of the sentences as language model during training. The case idealizes a recognizer and gives an upper limit that can be achieved using MLLR adaptation technique.

Speakers	Baseline	Full	Max	Norm	Improvement	Correct	Supervised
46h	39.94	34.65	33.33	32.82	5.28 %	32.91	32.47
47b	32.52	29.60	29.55	29.85	0.17 %	28.93	28.87
47h	14.26	12.37	12.25	12.52	0.97 %	11.46	11.56
47n	52.85	41.36	40.37	40.81	2.39 %	38.70	34.70
48r	10.63	9.63	9.76	9.86	-1.34 %	8.83	8.69
48v	18.33	18.19	18.33	17.83	1.98 %	17.71	17.56
49n	9.28	8.82	8.93	8.98	-1.22 %	7.40	7.40
4am	31.26	22.92	22.13	21.75	5.10 %	21.92	21.98
Average	26.51	22.19	21.83	21.80	1.66 %	20.98	20.40

Table 1. The numbers indicate the WER and the improvement is the relative decrease in the WER while comparing the Full and the minimum of Max and Norm values. The figures under *Full* indicate the results for MLLR adaptation without the use of confidence measurement, while those under *Max* are the results for MLLR adaptation using CM with most probable means, and the figures under *Norm* represent the results for MLLR adaptation using CM with normalized means.

5. RESULTS

Experiments were performed on data from 8 different speakers of Wall Street Journal corpus, each having about 150 utterances. During the training phase 50 sentences were used and testing was done with the remaining 100 sentences. MLLR training is done using a single regression class for all the frames. We obtained significant improvement in word error rates in case of some speakers after using confidence measures. We gained a relative improvement as high as 5 % in some cases when compared with MLLR technique without using confidence measures.

Also we observed adaptation using normalized means outperformed most probable gaussian's mean (Max), for some speakers. Although on an average, their performances are comparable. A more detailed study is needed for selection criteria between normalized and maximum probable gaussian mean. We also observed MLLR with only correct words (ideal CM) gave word error rate very close to those obtained with supervised adaptation. Table 1 summarizes the results.

6. CONCLUSIONS AND FUTURE WORK

In this work we have shown that use of confidence measures for MLLR adaptation improves the adaptation performance. We have also shown that normalizing the means of gaussian mixture can be an alternative, though more experiments have to be performed to judge the selection criteria for the types of mean. We have shown the performance of supervised adaptation is superior than other unsupervised adaptation, because supervised MLLR is able to reduce the mismatch between the acoustic models and acoustic vectors of incorrectly recognized words in first pass. We also found the performance of unsupervised adaptation with only correct words is close to that of supervised, thus shows a good confidence measure technique can raise the level of unsupervised MLLR adaptation. In future more experiments can be done using different regression classes and affect of different confidence measurement parameters can be tested to see the improvement.

7. REFERENCES

[1] C.J. Legetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, pp. 171–185, 1995.

[2] C.J. Legetter and P.C. Woodland, "Flexible speaker adaptation for large vocabulary speech recognition," *Proceedings EUROSPEECH95*, pp. 1155–1158, 1995.

[3] M. Pitz, F. Wessel, and H. Ney, "Improved mllr speaker adaptation using confidence measures for conversational speech recognition," *Proceedings of ICSLP*, pp. 548–551, 2000.

[4] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," *Proceedings of EUROSPEECH*, pp. 827–830, 1997.

[5] F. Wessel, "Confidence measurement for large vocabulary continuous speech recognition," *IEEE Trans. on Speech and Audio Processing*, pp. 9(3):288–298, 2001.

[6] D. Vergyri, "Use of word level side information to improve speech recognition," *Proceedings of ICASSP*, pp. 1823–1826 vol.3, 2000.