An Improved Method for Unsupervised Training of LVCSR Systems

Christian Gollan, Stefan Hahn, Ralf Schlüter, Hermann Ney

Lehrstuhl für Informatik 6 – Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{gollan,hahn,schluter,ney}@cs.rwth-aachen.de

Abstract

In this paper, we introduce an improved method for unsupervised training where the data selection or filtering process is done on state level. We describe in detail the setup of the experiments and introduce the state confidence scores on word and allophone state level for performing the data selection for mixture training on state level. Although we are using a relatively small amount of 180 hours of untranscribed recordings in addition to the available carefully manually transcribed transcriptions of 100 hours, we are able to significantly improve our final speaker adaptive acoustic model. Furthermore, we present promising results by doing system combination using the acoustic models trained on different confidence thresholds. These methods are evaluated on the EPPS corpus starting from the RWTH European English parliamentary speech transcription system. A significant improvement of 7% relative is achieved using less data for unsupervised training than conventional systems require.

Index Terms: Unsupervised Training, Automatic Transcription, Confidence Score

1. Introduction

The setup and the tuning of a statistical automatic speech recognition (ASR) system is time consuming and therefore expensive. It requires large, task specialized databases for estimating the model parameters in the acoustic and in the language model training. Large vocabulary continuous speech recognition (LVCSR) systems can still be improved using thousands of hours of acoustic training data [1]. The effort in setting up specialized training corpora is expensive and not always affordable especially for rarely spoken languages.

The porting of an already existing ASR system to a new domain is a feasible solution. [2] describes rapid cross-domain porting of an American English broadcast news ASR system to an European English Parliamentary Speech Transcription system: the acoustic model of the American English broadcast news system was used in combination with a specialized vocabulary, a specialized pronunciation dictionary and a specialized language model for this particular task. This was a very effective way of creating a transcription system for the European Parliament Plenary Session (EPPS) speeches for the TC-STAR project (further project details in [2]). Although this rapid cross-domain porting gave significant improvements in word error rate (WER), the use of 50 hours of carefully manually transcribed EPPS recordings within the first TC-STAR ASR Evaluation in 2005 lowered the WER by one half, from approximately 33% to about 14% WER.

For the third and last TC-STAR ASR Evaluation in 2007 a total of 280 hours of EPPS recordings were available for acoustic model training, out of which 100 hours were carefully manually transcribed. The manual transcription process took more than one year and is therefore not only a large cost factor but also time consuming.

Since a linear relationship between the WER of an ASR system and the logarithm of the amount of training data exists [3], the manual transcription of speech is only worthwhile for a certain amount of data as the transcription of further training data is too expensive compared to the expected performance gain. The automatic transcription of acoustic data using an ASR system and then using this automatically transcribed data for further training of the ASR system is commonly referred to as *'unsupervised training'*. Automatic transcriptions are not only much cheaper than manual transcriptions but can also be obtained much faster. On the downside, automatically transcribed documents commonly contain more transcription errors than manually transcribed ones. To avoid problems due to this drawback, the unsupervised training can be extended to use confidence scores of the recognized words to remove those which are likely to be subject to errors. Here we investigate filtering the automatically transcribed training data on state level using word posterior confidence scores [4] and allophone state posterior confidence scores.

One of the first publications describing unsupervised training for large vocabulary speech recognition is [5]. Many publications discussed various aspects of unsupervised training. Thousands of hours of data with automatic transcriptions have been used to improve the acoustic model [6]. In [7] it is shown, that systems trained on manually transcribed data outperform systems trained on (the same) data with automatic transcriptions only. There, the word posterior confidence score was used to filter out the most likely errors on word level in an iterative procedure.

Publications have already demonstrated that unsupervised training works in principle for LVCSR. In this paper we present unsupervised training results on the English EPPS task. We show that a well optimized state-of-the-art ASR system can be improved further even with a relative small amount of automatically transcribed speeches in comparison to previously presented unsupervised training experiments.

2. System Description

The acoustic model is trained on 87.5 hours EPPS training data which is a carefully created transcription of 100 hours of recordings leaving out long parts of music, foreign speeches (with respect to English) and other non EPPS specific parts.

We are using the Beep pronunciation dictionary and across-word context dependent triphones modeled by 6 HMM-States where two neighboring states are always tied. The skip, loop and forward transition probabilities are set globally. Silence and non-speech are modelled context independently. The HMM states are top down clustered by CART except the single silence

Table 1: *Performance of the English TC-STAR 2007 Evaluation system with and without unsupervised training (UT) (WER[%]).*

system/method	dev06	eval06	eval07
VTLN+Voicing	15.5	12.7	14.3
+SAT-CMLLR+MPE	13.7	10.4	12.0
+MLLR	12.2	9.9	11.2
+LM-Rescoring	11.6	8.8	10.4
UT transcription system	14.1	11.5	12.9
UT+VTLN+Voicing	13.6	11.7	13.1
+SAT-CMLLR+MPE	12.0	9.8	11.0
+MLLR	11.6	9.2	10.4
+LM-Rescoring	11.0	8.3	9.6
System Combination (TC-STAR 2007)	10.0	7.8	9.0

state. The emission probabilities for the 4,501 generalized states are modelled by Gaussian mixture models (GMMs) with a globally pooled diagonal covariance matrix.

The acoustic front end is based on Mel-Frequency Cepstral Coefficients (MFCC) features with Cepstral mean subtraction and variance normalization for a centered sliding window of 7 seconds. Vocal Tract Length Normalization (VTLN) is applied to the filterbank within the MFCC extraction. For each time frame, a voicing feature is added to the MFCC features. For each frame, the features of a centralized window of 9 consecutive frames are concatenated and projected to a 45 dimensional vector using Linear Discriminant Analysis (LDA).

Maximum Likelihood Speaker Adaptive Training (SAT) based on Constrained Maximum Likelihood Linear Regression (CMLLR) is performed and the acoustic model is refined by discriminative training using the Minimum Phoneme Error (MPE) criterion. For speaker adaptation the automatic generated segments are clustered due to the Bayesian Information Criterion (BIC) if no speaker labels are available.

Lastly, Maximum Likelihood Linear Regression (MLLR) and Language Model (LM) Rescoring are applied.

For the third TC-STAR Evaluation 2007 we have trained 4 ASR systems in parallel and performed system combination using the minimum frame WER (min-fWER) approach for our final recognition hypotheses. A more detailed system description can be found in [8].

We have participated in the restricted and public condition of the TC-STAR Evaluation 2007 and no other participant has achieved better results in these. Table 1 gives an overview of the performance of our TC-STAR Evaluation 2007 system for the English public condition on the development 2006 (dev06), evaluation 2006 (eval06) and evaluation 2007 (eval07) corpus. It also lists the most important methods and presents their results with and without unsupervised training (UT).

The English UT system is our best performing single system which shows the impact in performance of a relatively small amount of additional task-representative data. It can be assumed that the use of manual transcriptions instead of automatic ones would further improve the performance [7]. This leaves room for advanced unsupervised training methods. Therefore, we have compared the well known word posterior confidence scores with state posterior confidence scores filtering the automatic transcriptions on state level.

In the next section we describe the unsupervised training process on word level and state level using state posterior confidence scores. For the TC-STAR Evaluation 2007 we have used Unsupervised Training on word level with a confidence score threshold of 0.7. It should be noted that the results of Table 1 pertain exactly to the corresponding (threshold of 0.7) results of Table 3.

Table 2: *Transcription statistics of the English EPPS data.*

	transcription	
	MT	AT
raw recordings [h]	102.1	182.9
segmented data [h]	87.5	146.6
# segments	66,670	30,557
# running words	704,883	1,240,423

3. Unsupervised Training

3.1. Automatic Transcription System

The automatic transcription setup was optimized on the raw recordings of the English EPPS 2005 development (dev05) corpus. As here, the recordings are almost completely transcribed speeches of interpreters and politicians. The later evaluation corpora cover only politicians and provide manual segmentation for these main sections. The politician speeches held in English cover less than one third of the EPPS speeches and English is still one of the most frequently spoken languages from the 20 official European Parliament languages. As said before, the raw EPPS recordings contain foreign language parts. At almost every speaker change foreign language phrases can be observed as it takes some time before the broadcasting team switches to the correct interpreter channel. The raw dev05 corpus covers all of these raw recording specialties and is therefore most suitable to optimize the automatic transcription system for this task.

The automatic transcriptions were produced by a two pass SAT-CMLLR system derived from our TC-STAR Evaluation 2006 ASR system, but with pruning thresholds optimized to speed up the recognition process. Table 1 sets the used automatic transcription system in relation to acoustic models where different methods were applied. If we take a look at the dev06 corpus, we can observe that the automatic transcription produced with a system which has an error rate of 12.9% was able to improve a state-of-the-art system by 7% relative from 10.4% to 9.6%.

The first unsegmented recognition pass over the raw recordings was done with the VTLN+Voicing model. On basis of this first recognizer output we have done the automatic segmentation and the speaker clustering. For placing segment boundaries we take the length of the non-speech part and the language model probability of a sentence end into account. The segmented first pass transcription was used for the second speaker adaptive recognition pass with the SAT-CMLLR models.

The recognition setup described optimized for raw recordings leads to 12.3% WER on the raw dev05 recordings.

3.2. Transcription Statistics

Table 2 gives an overview of the manual transcriptions (MT) and the automatic transcriptions (AT) that we have produced for the experiments of this work. We present and break down the amount of raw recordings and the amount of segmented data where the most uninteresting acoustic parts are already removed. Note that we have used a heuristic approach to remove the foreign speech segments in the first place. Here, we have used the results of the BIC clustered automatically generated segments. Namely keeping only those segments where one of the two neighboring segments belong to the same speaker cluster. This first heuristic selection method leads then to the AT corpus listed in Table 2. A notable difference between the MT and AT corpora is the average segment length, which is much larger for the AT set due to the automatic segmentation.

3.3. Data Selection / Filtering on State Level

In unsupervised training the performance can be improved by data selection. The goal is to filter out most of the transcription errors and to keep only those parts which are expected to pay off in the acoustic model generalization. Different methods are known for the filtering of error-prone data. If closed captions are available this could be done on basis of an alignment between these and the AT and is known as lightly supervised training. Instead of these methods we have used state confidence scores to select and restrict the automatic transcriptions for the acoustic training.

We have used word graphs to calculate the posterior probability $p([w;\tau,t]|x_1^T)$ for a specific, aligned word hypothesis $[w;\tau,t]$ as presented in [4], where w is the word, τ and t are the word boundaries, which are start and end frame indices of an aligned word. Furthermore, x_1^T is the feature sequence, where T is the last time frame index of a segment. The first best recognition hypothesis for the segment is the sequence of words $\hat{w}_1^M$. $\hat{s}_1^T$ is the corresponding aligned state sequence, where $s_t(w)$ denotes the aligned state of a word w at time frame index t. $\hat{w}(\hat{s}_t)$ denotes the word hypothesis of the first best sequence $\hat{w}_1^M$ which correspond to the state $\hat{s}_t$ at the time frame index t. The decision for taking the pair $[\hat{s}_t, x_t]$ in mixture training into account depends on the state confidence score $\mathcal{C}(t;\hat{s}_1^T)$ and the applied threshold. We rewrite the maximum word posterior confidence score presented in [4] as

$$\mathcal{C}_{mw}(t;\hat{s}_1^T) = \max_{t_{max}\in[\hat{w};\hat{\tau},\hat{t}]=\hat{w}(\hat{s}_t)} \sum_{\substack{[\hat{w};\tau',t']:\\ \tau'\leq t_{max}\leq t'}} p([\hat{w};\tau',t']|x_1^T) \ , \tag{1}$$

and introduce furthermore the allophone state posterior confidence score,

$$\mathcal{C}_{as}(t;\hat{s}_1^T) = \sum_{\substack{[w;\tau',t']:\\ \tau'\leq t\leq t' \,\wedge\, s_t(w)=\hat{s}_t}} p([w;\tau',t']|x_1^T) \tag{2}$$

for data filtering.

We have chosen the most probable correct sequence of state and acoustic feature vector pairs for GMM training by applying a state confidence threshold on word or state level. For the TC-STAR Evaluation 2007 we have done discriminative MPE refinement of the GMM without any data filtering and we are currently performing further MPE trainings with data filtering on state level.

Figure 1 illustrates the allophone state error rate in relation to a certain threshold. The threshold is mapped to the corresponding amount of kept data which makes it possible to compare different threshold dependent data selection methods.

We calculate the allophone state error rate of an automatic transcription word hypothesis state alignment due to the reference state alignment using the reference transcriptions and the acoustic model of the automatic transcription system. For each state of the reference state alignment we count an allophone state error if the corresponding hypothesis state would be used for unsupervised training and if it differs from the reference state.

Further experiment and investigations are necessary to decide on which level we should investigate the state error rate. The units should be smaller than words, such that a correct, aligned pair of state and acoustic feature vector (for GMM training) is not counted as an error. For a wrong word hypothesis at word level, all aligned states would be counted as an error, even though their prefix, stem or postfix alignment states correspond.

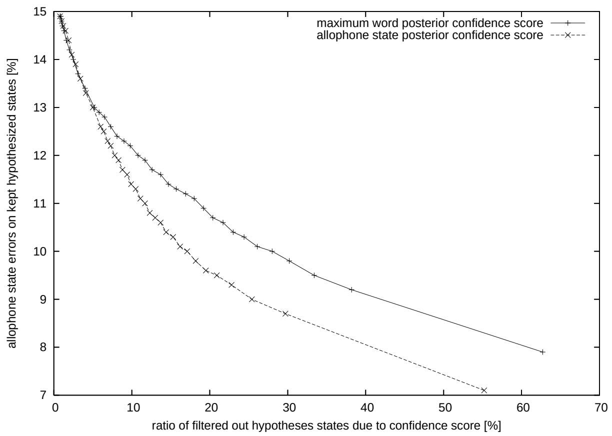

Figure 1: Allophone state error rate on dev05 for the selected states due to their confidence score.

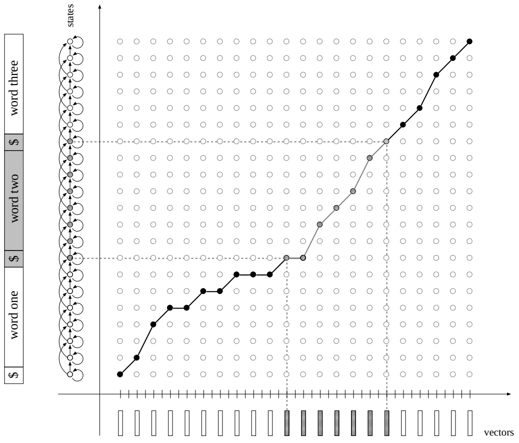

Figure 2: Selected state sequence for unsupervised training.

This is the reason why we think the allophone state posterior confidence score curve drops faster (performs better) than the maximum word posterior confidence score curve. I.e.: in Figure 1 an allophone state error rate of 10% corresponds to 72.0% of reference data kept using the maximum word posterior confidence score, whereas 82.9% of the data can be kept to achieve the same error rate due to allophone state posterior confidence score.

3.4. Practical Aspects

For unsupervised training and data selection on state level a practical aspect has to be considered. Figure 2 illustrates the state alignment of the automatic transcript of three words. The word in the middle was selected for acoustic training and the corresponding sequence of states and acoustic feature vectors are used for the mixture training. We filter the training data on the state alignment of the recognition run to preserve the connection of states and acoustic feature vectors. Thus, we keep the phoneme context (especially for the across-word phonemes) of the best recognition path even though the context could have been filtered out.

Table 3: *Comparison of SAT-CMLLR unsupervised training results on English EPPS Corpora (WER[%]).*

	system ID	data				
transcript.	meth./thr.	[h]	[%]	densities	dev06	eval06
MT	–	87.5	37.4	870 930	14.2	10.9
MT + AT	–	234.1	100.0	1 893 080	13.0	10.6
MT + AT	$\mathcal{C}_{mw}$/0.5	225.6	96.4	1 874 552	12.8	10.4
MT + AT	$\mathcal{C}_{mw}$/0.7	210.2	89.8	1 850 988	12.7	10.3
MT + AT	$\mathcal{C}_{mw}$/0.9	189.9	81.1	1 807 999	12.9	10.2
MT + AT	$\mathcal{C}_{as}$/0.45	226.4	96.7	1 879 389	12.7	10.3
MT + AT	$\mathcal{C}_{as}$/0.78	212.4	90.7	1 845 974	12.7	10.0
MT + AT	$\mathcal{C}_{as}$/0.96	193.5	82.7	1 799 244	12.8	10.2

Table 4: *Comparison of system combination results on English EPPS Corpora (WER[%]).*

ROVER system combination			dev06	eval06
$\mathcal{C}_{as}$/0.45	$\mathcal{C}_{as}$/0.78	$\mathcal{C}_{as}$/0.96	12.1	9.9
$\mathcal{C}_{mw}$/0.5	$\mathcal{C}_{mw}$/0.7	$\mathcal{C}_{mw}$/0.9	12.2	9.8
$\mathcal{C}_{mw}$/0.7	$\mathcal{C}_{as}$/0.78	–	12.2	9.9

4. Experiments

As described in the previous section we have generated the automatic transcription by the SAT-CMLLR model. We have focussed our experiments on this model because of the highly time consuming discriminative training and MLLR adaptation on all the tested data selection setups. On the other hand we are still close to our best recognition arrangement and we are able to preserve the unsupervised training gain for our final recognition system.

The corresponding dev06 and eval06 results of the SAT-CMLLR acoustic models are listed in Table 3. The models are trained on different amounts of training data selected on state level by calculating either the maximum word posterior confidence score $\mathcal{C}_{mw}$ or the allophone state posterior confidence score $\mathcal{C}_{as}$. Experiments with different thresholds show that unsupervised training with both confidence methods achieves the best performance if approximately 10% of the data is filtered out. It can be observed that the $\mathcal{C}_{as}$ trained models never perform worse and sometimes slightly better than the $\mathcal{C}_{mw}$ trained models.

Note that the only difference of the GMM training is the selection method applied or threshold on state level. I.e.: the training state alignment to the corresponding acoustic feature vector sequence were exactly the same for all the listed experiments. To investigate the difference of the acoustic models we have used ROVER for combining their first best system outputs. The system combination improvements in Table 4 are interesting results, as there is only relatively small difference in the data used for mixture training.

5. Conclusion and Outlook

We have successfully applied unsupervised training on the English EPPS task and introduced a more sophisticated method for unsupervised training, namely the data selection or filtering process on state level. In this framework, the automatic selection of training data for GMM training is done on the smallest transcription unit, a pair of a state and an acoustic feature vector. We have adapted the word posterior confidence score definition for this framework and have investigated the state confidence score on word and allophone state level.

Furthermore, we have made an interesting observation: even if the difference in WER between the differently thresholded systems is quite small, we could achieve further gains doing system combination on these.

Significant improvements were achieved with the additional, relatively small amount of 180 hours of untranscribed data, when compared to the 100 hours carefully transcribed recordings. We were able to improve the performance of our final speaker adaptive models by more than 7% relative. The WER was reduced from 10.4% to 9.6% on the English EPPS 2007 evaluation set. For word posterior confidence scores we have shown that the gains could be preserved for the final recognition system.

In the future, we are planning to investigate iterative unsupervised training in combination with state confidence scores on other levels, e.g. on subword, triphone, monophone or mixture id level.

Acknowledgments

This work was partly funded by the European Union under the integrated project TC-STAR (FP6-506738).

6. References

[1] G. Evermann, H. Y. Chan, M. J. F. Gales, B. Jia, D. Mrva, P. Woodland, and K. Yu, "Training LVCSR Systems on Thousands of Hours of Data," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, USA, Mar. 2005, vol. 1, pp. 209 – 212.

[2] C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, and H. Ney, "Cross Domain Automatic Transcription on the TC-STAR EPPS Corpus," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, USA, Mar. 2005, vol. 1, pp. 825 – 828.

[3] R. K. Moore, "A Comparison of the Data Requirements of Automatic Speech Recognition Systems and Human Listeners," in *European Conf. on Speech Communication and Technology*, Geneva, Switzerland, Sept. 2003, pp. 2582 – 2584.

[4] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence Measures for Large Vocabulary Continuous Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, Mar. 2001.

[5] G. Zavaliagkos and T. Colthurst, "Utilizing Untranscribed Training Data to Improve Performance," in *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, USA, Feb. 1998, pp. 301 – 305.

[6] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, "Unsupervised Training on Large Amounts of Broadcast News Data," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006, vol. 3, pp. 1057 – 1059.

[7] F. Wessel and H. Ney, "Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 23 – 31, Jan. 2005.

[8] J. Lööf, Ch. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, Ch. Plahl, D. Rybach, R. Schlüter, and H. Ney, "The RWTH 2007 TC-STAR evaluation system for European English and Spanish," in *Int. Conf. on Spoken Language Processing*, Antwerp, Belgium, Aug. 2007, Submitted.