ADAPTIVE TRAINING USING SIMPLE TARGET MODELS

Georg Stemmer, Fabio Brugnara, Diego Giuliani

ITC-irst, Centro per la Ricerca Scientifica e Tecnologia
38050 Povo (Trento), Italy
stemmer@itc.it

ABSTRACT

Adaptive training aims at reducing the influence of speaker, channel and environment variability on the acoustic models. We describe an acoustic normalization approach to adaptive training. Phonetically irrelevant acoustic variability is reduced at the beginning of the training procedure w. r. t. a set of target models. The set of target models can be a set of HMMs or a Gaussian mixture model (GMM). CMLLR is applied to normalize the acoustic features. The normalized data contains less unwanted variability and is used to generate and train the recognition models. Employing a GMM as a target model leads to a text-independent procedure that can be embedded into the acoustic front-end. On a broadcast news transcription task we obtain relative reductions in WER of 7.8% in the first recognition pass over a conventionally trained system and of 3.4% in the second recognition pass over a SAT-trained system.

1. INTRODUCTION

A major challenge in speech recognition is to achieve good results in tasks where recording conditions, acoustic environment, quality of the transmission channel or speaker frequently change. Such a task is the automatic transcription of broadcast news, i. e. recorded news shows from television and radio networks. For this paper we distinguish between phonetic variability, i. e. differences between the speech sounds that are relevant for their discrimination, and the phonetically irrelevant variability which is caused by the diversity of speakers, transmission channels and acoustic environments. For increased readability the single term speaker variability subsumes here all kinds of speaker, channel and environment variability. The approach introduced here belongs to the group of *adaptive training* schemes [1]. These algorithms have in common that they estimate speaker-specific transformations –either of the models or of the acoustic features– to exclude phonetically irrelevant variability from the training. Ideally, the acoustic models only learn phonetic variability. Well-known representatives of adaptive training algorithms are *Speaker Adaptive Training (SAT)* [2] and *Vocal Tract Length Normalization (VTLN)* [3]. Both procedures have shown to outperform conventional training methods.

1.1. Approach

In this paper we apply an acoustic normalization approach to reduce the influence of irrelevant variability on the acoustic models [4]. The training procedure consists of three stages: Firstly, preliminary acoustic models are trained on the original features. The resulting models are called *target models*. Secondly, the features are transformed w. r. t. the target models and the current speaker or acoustic condition using *Constrained Maximum Likelihood Linear Regression (CMLLR)* [5]. The transformed, or normalized, features are supposed to contain less speaker variability. Thirdly, the *recognition models* are generated and trained on the transformed data. One of the differences of this approach to other adaptive training methods is that the target models and the recognition models are independent, i. e. they may have a different model structure. Therefore we measure the performance that can be achieved for different types of target models, comparing target models with just one Gaussian per state and complex target models. The former is shown to be advantageous; this is consistent with previous work [4]. We also introduce a very simple target model that is just a *Gaussian Mixture Model (GMM)*. As in this case word transcriptions of test utterances are not required for estimating the feature transformation, acoustic data normalization can be applied at recognition stage without any preliminary decoding step. Another contribution of the paper is to compare the proposed approach with a popular variant of SAT introduced by Gales [6].

1.2. Related work

Several adaptive training procedures are well-known from the literature. SAT [2] is one of the most popular approaches. Supervised adaptation is performed for each training speaker in each iteration of the Baum-Welch algorithm. As adaptation requires a sufficiently trained acoustic model, usually 2–3 iterations of SAT are added on top of a conventional training procedure. In recognition two passes are performed: the first generates a preliminary transcription of the utterance which is used to adapt the SAT-trained acoustic models to the test speaker. A second recognition pass using the adapted models generates the final result. Usually, *modified SAT* [7, 6] is employed that allows for an efficient implementation. In broadcast news transcription tasks this variant of SAT leads typically to 3–5% relative reduction in word error rate (WER) over an adapted, conventionally trained baseline system (e. g. [7]). Both modified SAT and the proposed acoustic normalization approach use CMLLR for feature transformation. While in SAT adaptation and training of the models are alternated, in the proposed method two sequential steps are performed for acoustic normalization and training. Thus, speaker variability is reduced not at the end but at the beginning of the training of the recognition models. A detailed discussion of the consequences of this difference can be found in Sec. 2. It is shown in this paper that the proposed approach can be applied in a text-independent manner that avoids the preliminary recognition step required by SAT.

Being an acoustic normalization procedure, the approach

This work was partially financed by the European Commission under the project TC-STAR, contract reference number FB6-506738.

described in this paper is similar to the one introduced in [8] or to VTLN [3]. The main difference of the proposed approach to these methods is that we employ CMLLR for normalizing the data. In [4] it has already been shown that the proposed method compares favorably to VTLN. Note also that VTLN is limited to normalize vocal tract shape differences of speakers while CMLLR can be applied more generally to normalize acoustic variability due to a variety of sources.

The paper is structured as follows: Sec. 2 is dedicated to a detailed description of our approach, and its application to improve the performance of the first and second recognition passes. At the end of this section the differences to other adaptive training approaches are discussed. The data corpus and the baseline system are introduced in Sec. 3. Experiments in Sec. 4 verify the effectiveness of the approach. Finally, we give a conclusion and a short prospect on our future work in Sec. 5.

2. ADAPTIVE TRAINING USING TARGET MODELS

The proposed procedure is motivated by the aim of reducing the influence of speaker variability in the early stages of acoustic model training.

We exploit the fact that a set of continuous density HMMs can be effectively adapted using CMLLR [5, 6]: A transformation $\{\mathbf{M}, \mathbf{d}\}$ is applied to mean $\mu_m \mapsto \mathbf{M}\mu_m + \mathbf{d}$ and covariance $\mathbf{\Sigma}_m \mapsto \mathbf{M}\mathbf{\Sigma}_m\mathbf{M}^\star$ of each Gaussian density $\mathcal{N}(\mathbf{o}_t \mid \mu_m, \mathbf{\Sigma}_m)$, resulting in adapted densities $\mathcal{N}(\mathbf{o}_t \mid \mathbf{M}\mu_m + \mathbf{d}, \mathbf{M}\mathbf{\Sigma}_m\mathbf{M}^\star)$. All densities have the same transformation in common. For each speaker s a single transformation $\{\hat{\mathbf{M}}_s, \hat{\mathbf{d}}_s\}$ is estimated using the EM-algorithm. The objective is to maximize the log-likelihood $\mathcal{L}(\mathbf{O}_s \mid \{\mathbf{M}, \mathbf{d}\}, \mathbf{\Lambda}, \mathbf{W})$ of the acoustic models $\mathbf{\Lambda}$ for the speaker's utterance $\mathbf{O}_s = \mathbf{o}_{s,1}, \ldots, \mathbf{o}_{s,T}$. $\mathbf{W}$ stands for the word-level transcription of the utterance $\mathbf{O}_s$. It can be shown that the model-transformation $\{\hat{\mathbf{M}}_s, \hat{\mathbf{d}}_s\}$ can be implemented using a transformation $\{\mathbf{A}_s := \hat{\mathbf{M}}_s^{-1}, \mathbf{b}_s := -\hat{\mathbf{M}}_s^{-1}\hat{\mathbf{d}}_s\}$ of the feature vectors $\mathbf{o}_{s,t}$. The transformed feature vectors $\mathbf{o}_{s,t}^{\mathbf{\Lambda}}$ are computed according to $\mathbf{o}_{s,t}^{\mathbf{\Lambda}} := \mathbf{A}_s\mathbf{o}_{s,t} + \mathbf{b}_s$. In this work we apply CMLLR as a feature transformation. In the experiments reported here $\mathbf{O}_s$ does not necessarily correspond to a speaker's utterance but to a cluster of acoustically similar speech segments that have been determined in a data-driven manner.

While CMLLR adaptation is usually performed w. r. t. the recognition models, our approach for adaptive training is to use *separate* sets of models. For acoustic normalization so-called *target models* $\mathbf{\Lambda}^{\mathrm{n}}$ are used. For the acoustic representation in the decoding phase *recognition models* $\mathbf{\Lambda}^{\mathrm{r}}$ are employed. All parameters of the two model sets, like initialization, definition of the context-dependent allophones or model structure are completely independent. Therefore we investigate different types of target models. Using a GMM as a target model leads to a text-independent approach; when the target models are triphone HMMs the procedure becomes text-dependent.

2.1. Adaptive training procedure

The proposed training algorithm proceeds as follows:

1. train target model $\mathbf{\Lambda}^{\mathrm{n}}$ on untransformed feature vectors $\mathbf{O}$. $\mathbf{\Lambda}^{\mathrm{n}}$ may be either a GMM or a set of triphone HMMs.
2. for each speaker s, estimate $\{\mathbf{A}_s, \mathbf{b}_s\}$ w. r. t. $\mathbf{\Lambda}^{\mathrm{n}}$ for the feature vectors $\mathbf{O}_s$. Apply $\{\mathbf{A}_s, \mathbf{b}_s\}$ to $\mathbf{O}_s$, yielding transformed feature vectors $\mathbf{O}_s^{\mathrm{n}}$.
3. use the conventional training procedure to initialize and to train the recognition models $\mathbf{\Lambda}^{\mathrm{r}}$ on $\mathbf{O}^{\mathrm{n}}$; including state tying and the definition of the context-dependent allophones.

2.2. Recognition procedure

When a GMM is used as a target model $\mathbf{\Lambda}^{\mathrm{n}}$ CMLLR normalization can be performed without the need for word-level transcriptions. Thus, in this case no word-level transcriptions are required for the acoustic normalization and recognition starts from an untranscribed utterance $\mathbf{O}_s$. For triphone HMMs as target models, a transcription of the utterance has to be available from a previous recognition pass. Normalization and decoding is performed as follows:

1. estimate $\{\mathbf{A}_s, \mathbf{b}_s\}$ w. r. t. $\mathbf{\Lambda}^{\mathrm{n}}$ for the feature vectors $\mathbf{O}_s$. Apply $\{\mathbf{A}_s, \mathbf{b}_s\}$ to $\mathbf{O}_s$, yielding transformed feature vectors $\mathbf{O}_s^{\mathrm{n}}$.
2. decode $\mathbf{O}^{\mathrm{n}}$ using $\mathbf{\Lambda}^{\mathrm{r}}$.

Of course, using CMLLR for acoustic feature normalization does not prevent us from employing *Maximum Likelihood Linear Regression (MLLR)* [9] to adapt the recognition models $\mathbf{\Lambda}^{\mathrm{r}}$ for the second recognition pass. Note that in this work we never observed a performance gain when we used acoustic feature normalization in conjunction with MLLR adaptation of the recognition models $\mathbf{\Lambda}^{\mathrm{r}}$. Thus, in this case both adaptation and decoding are based on *untransformed* feature vectors $\mathbf{O}_s$.

2.3. Relation to SAT

VTLN [3] and SAT [2, 6] are the most popular adaptive training algorithms. As we have already confronted our proposed method to VTLN in [4], we give here a comparison to SAT. The most prominent difference between the two approaches is that in SAT adaptation and training steps are alternated and not performed one after another like in acoustic normalization. As alternation of adaptation and training requires sufficiently trained models, SAT is added on top of a conventional training procedure. Thus, the acoustic models have already taken over speaker variability in the conventional training phase, and subsequent SAT iterations have to alleviate this. The authors of [2] expressed their believe that other initialization methods for SAT should be taken into consideration. In particular, the decision tree for defining the state tying and the context-dependent allophones is determined using the conventionally trained models and cannot be adjusted any more during SAT [10]. For the proposed acoustic normalization procedure the decision tree is determined from models that have already been trained on normalized features. Consequently the state tying and the context-dependent allophones are consistent with the adaptively trained models. Furthermore, in SAT the first recognition pass is always performed using unadapted baseline models which include the speaker variability. For a high mismatch between training and testing the initial transcription will be poor and in consequence the transformation parameters may as well by poorly estimated [1]. As we have described in this section, the proposed method can be applied already in the first recognition pass and is able to reduce the errors in the initial utterance transcription.

Reducing speaker variability in SAT using simple acoustic models has already been investigated by others. Huo and Ma discuss in [10] the presence of inter-speaker variance in the allophone definitions. They propose to train untied HMMs with a single Gaussian per state using SAT. The decision tree for state tying is then constructed from these models. The main difference to our work is that Huo and Ma concentrate only on the decision tree. Our approach goes further as we use the simple target models also to reduce speaker variability learned by the recognition models. McDonough and Byrne introduce in [11] a variant of SAT which is named *Single-Pass Adapted Training (SPAT)*. Simple HMMs with a single Gaussian per state are trained using SAT. The resulting adaptation parameters are then applied to a set of mixture-density baseline models. Finally, several iterations of conventional SAT are added. Thus, this procedure is mainly a different initialization method for SAT using simple acoustic models.

3. DATA SETS AND BASELINE SYSTEM

We used the BN-E data released by the LDC in 1997 and 1998 for training of the acoustic models. The corpora contain a total of about 143 hours of usable speech data. For evaluation we use the 1998 Hub4 evaluation data consisting of two files, each with 1.5 hours of speech *(Eval98)*. Results are reported w. r. t. the focus conditions *(F-conditions)* marked in this test set:

- F0: baseline planned broadcast speech, clean background.
- F1: spontaneous broadcast speech, clean background.
- F2: speech over telephone, clean background.
- F3: speech with background music.
- F4: speech with degraded acoustics (noise, other speech).
- F5: planned speech by non-native speakers, clean background.
- FX: all other conditions that cannot be classified into F0-F5.

Language models were trained on ≈132 million words of broadcast news transcripts distributed by LDC and on the transcripts of the BN-E training data.

We used the ITC-irst speech recognition system for the experiments. The front-end combines 13 Mel-frequency Cepstral Coefficients and their first and second order time derivatives into a 39-dimensional feature vector. In the baseline system *Cluster-based Mean and Variance Normalization (CMVN)* is applied to the static features: input speech data is segmented using a Bayesian Information Criterion. Segments are classified into acoustic conditions. An automatic clustering is performed for all segments that belong to the same class. For each cluster, mean and variance of the features are normalized. No manual segmentation or clustering is used in training and recognition. The acoustic models are state-tied, cross-word, gender-independent, bandwidth-independent triphone HMMs. A phonetic decision tree is used for tying states and defining the context-dependent allophones. The baseline system has 9079 tied states and about 146000 Gaussians; all other systems in this paper have a similar number of parameters.

4. EXPERIMENTAL RESULTS

We report experimental results for one and two recognition passes: For experiments with two passes MLLR adaptation is applied to the recognition models. This requires a preliminary transcription (supervision) which is generated in a first recognition pass. The second pass uses the adapted models to generate the final result.

4.1. First-pass recognition results

As in the first pass no transcription is available, a GMM has to be used as a target model. We compare two different recognizers with the baseline system *(baseline)*. For the first one, *GMM*, the mixture model has been trained using ten iterations of the EM-algorithm. The GMM is trained on the 39-dimensional feature vectors. Only segment-based normalization of the mean of the static features is applied and no variance normalization. Based on preliminary experiments we decided to use 512 mixture components. The second recognizer is denoted by *SAT-GMM*. The only difference to *GMM* is that the mixture model is trained with ten iterations of SAT. This is a variant of the basic procedure described in Sec. 2. For the estimation of the feature transformation both *GMM* and *SAT-GMM* use three iterations of CMLLR. For each cluster of segments one CMLLR transformation is estimated. Thus, the feature transformation is comparable to CMVN that is applied in the baseline system, except that CMLLR transforms the whole 39-dimensional feature vectors and not only the static features. The recognition models are trained using the same procedure as for the baseline system. The results are shown in Tab. 1. Clearly, a GMM is much

F-cond.	all	F0	F1	F2	F3	F4	F5	FX
proportion	100.0	30.7	19.3	3.4	4.3	28.2	0.7	13.5
baseline	20.5	12.9	20.0	30.0	24.0	20.7	20.9	34.3
GMM	19.1	12.2	18.7	27.9	23.9	19.3	18.3	31.4
SAT-GMM	18.9	11.6	17.7	30.4	23.3	19.1	17.0	32.6

Table 1. Recognition results (WER) for the first recognition pass on Eval98.

better in reducing irrelevant speaker variability than CMVN: the relative improvements in WER for *GMM* and *SAT-GMM* are 6.8% and 7.8%, respectively. The use of SAT training for the target model in *SAT-GMM* leads, however, only to a small additional reduction in overall WER.

4.2. Second-pass recognition results

All systems that are compared here use MLLR adaption of the recognition models with two regression classes. MLLR is based on the transcription results of the first recognition pass and is applied both to mean and variances. The baseline system is improved by adding three iterations of SAT, the resulting recognizer is denoted by *baseline+SAT*. For SAT we use in training and recognition the standard procedure as it is described in [6]. The system *complex-target* is built using the acoustic models of *baseline+SAT* as target models. After CMVN the features are transformed w. r. t. these models and then a complete new recognizer is trained. The system *simple-target* uses target models that have been trained with just a single Gaussian per state. For *simple-target* only segment-based mean normalization is applied to the features before CMLLR.

In a first experiment we measure the effectiveness of the acoustic normalization approach for the broadcast news task. Tab. 2 confronts the baseline system *(baseline)*, the SAT-trained baseline system *(baseline+SAT)*, and the proposed method *(complex-target, simple-target, SAT-GMM)* exploiting the word transcriptions that have been generated using the baseline recognizer. SAT leads to an improvement of 5.3% relative in WER over the adapted baseline system. Acoustic normalization using a simple target models is more effective: the corresponding relative reduction for the best system *simple-target* is 8.6%. A complex

F-cond.	all	F0	F1	F2	F3	F4	F5	FX
proportion	100.0	30.7	19.3	3.4	4.3	28.2	0.7	13.5
baseline	18.7	11.7	18.7	24.7	23.0	19.0	19.6	30.8
baseline+SAT	17.7	11.4	17.3	24.3	21.5	17.7	19.6	29.9
complex-target	17.9	11.3	17.9	22.8	22.2	18.1	19.6	29.6
simple-target	17.1	10.9	16.8	21.4	21.1	17.4	18.3	28.4
SAT-GMM	17.4	11.0	17.2	21.6	21.4	17.8	17.9	29.1

Table 2. Recognition results (WER) for the second recognition pass on Eval98 using the baseline supervision.

target model improves over the conventionally trained system, but performs slightly worse than the SAT-trained baseline. The acoustic normalization procedure using the SAT-trained GMM performs surprisingly good (7.0% relative reduction in WER).

Secondly, the different recognition systems are adapted using the improved supervision, i. e. the first-pass recognition result of *SAT-GMM* (18.9% WER). This way we measure the influence of the improved supervision on the recognition performance. The corresponding results are shown in Tab. 3. It can be seen

F-cond.	all	F0	F1	F2	F3	F4	F5	FX
proportion	100.0	30.7	19.3	3.4	4.3	28.2	0.7	13.5
baseline	17.8	11.0	17.9	23.1	21.6	18.1	18.7	29.8
baseline+SAT	17.6	10.8	16.9	24.3	21.8	17.6	18.3	30.8
complex-target	17.3	10.9	16.8	21.2	21.9	17.4	15.7	29.5
simple-target	17.0	10.6	16.6	21.5	20.4	17.5	16.6	29.0
SAT-GMM	17.3	10.7	17.0	22.4	21.4	17.6	16.6	29.2

Table 3. Recognition results (WER) for the second recognition pass on Eval98 using the improved supervision obtained with the *SAT-GMM* system.

that there is no consistent influence of the improved supervision on the recognition result: only for the *baseline* system and the *complex-target* system there are significant relative reductions in WER (4.8% and 3.4%, respectively). For all other systems, improvements are rather small. Thus, the difference between *baseline* and *baseline+SAT* diminishes (see Tab. 3). However, the relative improvement of the *simple-target* system compared to *baseline+SAT* remains the same as for the baseline supervision (3.4%).

Next, we compare the performances achieved by the different types of target models. We found it noticeable that *SAT-GMM* performs so well. Based on a very simple acoustic normalization using a SAT-trained GMM, this recognizer leads to a slightly better performance than *baseline+SAT*. Lowest overall WER is achieved by the system *simple-target* for different supervisions. This result is consistent with previous experiments for other types of speech corpora [4]. A simple target model has the advantage, that it is not able to represent too much speaker variability in its output densities when it is trained on unnormalized data and thus may force a stronger normalization on the data. Thus, it seems reasonable to us to prefer a simple target model over a complex one.

5. CONCLUSION AND FUTURE WORK

In this paper we described an alternative adaptive training procedure based on feature normalization, that allows to reduce unwanted variability at an early stage of the training procedure. Moreover, if a GMM target model is used, the normalization can become a part of the acoustic front-end. We obtained relative reductions in WER of 7.8% in the first recognition pass over a conventionally trained system and of 3.4% in the second recognition pass over a SAT-trained system. In the future we will investigate the use of specialized GMMs for different acoustic conditions, e. g. low-bandwidth speech. Finally, we hope that the GMM-based feature transformation turns out to be useful also in applications where it is difficult or impossible to apply conventional two-pass adaptation methods, e. g. neural network-based/segment model-based recognizers, or real time applications like spoken dialogue systems.

6. REFERENCES

[1] M.J.F. Gales, "Adaptive training for robust ASR," in *IEEE ASRU Workshop*, 2001.

[2] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 1996, pp. 1137–1140.

[3] L. Lee and R.C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1996, pp. 353–356.

[4] D. Giuliani, M. Gerosa, and F. Brugnara, "Speaker normalization through constrained MLLR based transforms," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 2004, vol. 4, pp. 2893–2897.

[5] V. Digalakis, D. Rtischev, 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, 1995.

[6] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[7] S.S. Chen, E. Eide, M.J.F. Gales, R.A. Gopinath, D. Kanvesky, and P. Olsen, "Automatic transcription of broadcast news," *Speech Communication*, vol. 37, pp. 69–87, 2002.

[8] Y. Zhao, "An acoustic-phonetic-based speaker adaptation technique for improving speaker-independent continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 380–394, 1994.

[9] 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.

[10] Q. Huo and B. Ma, "Irrelevant variability normalization in learning HMM state tying from data based on phonetic decision-tree," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1999, pp. 577–600.

[11] J. McDonough and W. Byrne, "Speaker compensation with all-pass transforms," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, 1999, pp. 2737–2740.