ON-LINE SPEAKER ADAPTATION ON TELEPHONY SPEECH DATA WITH ADAPTIVELY TRAINED ACOUSTIC MODELS

Diego Giuliani, Roberto Gretter and Fabio Brugnara

Human Language Technology Research Unit
FBK-irst - Fondazione Bruno Kessler
Via Sommarive 18, 38100 Povo (TN), Italy
http://hlt.fbk.eu - giuliani@fbk.eu

ABSTRACT

This paper addresses speaker adaptive acoustic modeling, based on feature space maximum likelihood linear regression, in the context of on-line telephony applications. An adaptive acoustic modeling method, that we previously proved effective in off-line applications, is used to train acoustic models to be used in text-dependent and text-independent on-line adaptation. Experiments on telephony speech data indicate that feature space maximum *a posteriori* linear regression (fMAPLR) greatly helps to cope with sparse adaptation data when performing instantaneous and incremental adaptation with both baseline models and speaker adaptively trained models. The use of speaker adaptively trained models in conjunction with fMAPLR leads to the best recognition results in both instantaneous and incremental adaptation. The proposed text-independent adaptation approach, exploiting speaker adaptively trained models, is also proven effective.

Index Terms— speaker adaptation, on-line adaptation, telephony application, speaker adaptive training, automatic speech recognition

1. INTRODUCTION

To tackle acoustic mismatch between training and testing acoustic conditions, state-of-the-art speech recognition systems embed acoustic adaptation. In this work we are interested in experimenting with fast acoustic adaptation methods in the context of an on-line telephony application. There are a variety of adaptation methods which are commonly adopted by modern recognition systems based on hidden Markov models (HMMs); however, feature space maximum likelihood linear regression (fMLLR) [1] is usually preferred in telephony applications because a small amount of adaptation data is available from a single speaker and there is no need to update acoustic model parameters [2, 3]. fMLLR employs, in fact, a single transformation matrix and a bias vector to linearly transform the input acoustic features. The affine transformation is estimated by maximizing the likelihood of the transformed acoustic observations w. r. t. the speech recognition models, i.e. continuous density HMMs, assuming a word level transcription of the acoustic data [1]. Effectiveness of fMLLR in reducing the acoustic mismatch between the speech recognition models and the input acoustic data was proven on a number of different domains [1, 2, 4]. However, effectiveness of fMLLR adaptation is usually reduced when there are very sparse adaptation data such as in on-line telephony applications [2, 3]. To tackle the problem of unreliable transformation parameter estimation in on-line incremental adaptation, smoothing of fMLLR sufficient statistics [2] and fMAPLR were proposed [3].

On the other hand, fMLLR offers an efficient and simple way for implementing speaker adaptive acoustic modeling allowing transformation of the acoustic data of each training and testing speaker instead of transforming acoustic model parameters [1]. In [5, 6] we proposed a variant of fMLLR-based speaker adaptive training in which transformation parameter estimation is carried out, both during training and testing, with respect to a set of "target" models which are different than the "recognition" models used for performing decoding of the test data. Leveraging on the possibility that the structure of the target and recognition models can be determined independently, in [7] we proposed to estimate transformation parameters, both during training and testing, with the aim of maximizing the likelihood of the transformed features with respect to a Gaussian mixture model modeling the whole training data. With this *text-independent* variant, at recognition stage there is no need of having a transcription of the adaptation data making this approach very appealing when it is important to achieve computational efficiency and low latency adaptation [7, 8].

In this work, fMLLR-based adaptation is exploited for speaker adaptive acoustic modeling as well as for on-line adaptation. Effectiveness of the speaker adaptive acoustic modeling method, proposed in the past for off-line applications [5, 6], is assessed in an on-line task with telephony speech data by addressing issues such as adaptation latency and system complexity. To cope with sparse adaptation data, fMAPLR is employed in both instantaneous and incremental adaptation of baseline models and speaker adaptively trained models. Text-dependent and text-independent adaptation approaches are also investigated and compared.[1]

The remainder of this paper is organized as follows. Section 2 introduces the method used for speaker adaptive training of acoustic models. Section 3 addresses important issues related to the use of fMAPLR. Section 4 introduces the telephony speech corpus while Section 5 describes the speech recognition systems. Section 6 describes unsupervised batch adaptation experiments while Section 7 reports on on-line adaptation experiments. Finally, conclusions are reported in Section 8.

This work was partly supported by Siemens under the grant BMU/536380.

[1] In this paper the term *text-dependent* adaptation denotes an adaptation mode which needs transcription of adaptation data, in contrast the term *text-independent* adaptation denotes an adaptation mode which does not need transcription of adaptation data.

2. SPEAKER ADAPTIVE ACOUSTIC MODELING

In this section we summarize the adaptive acoustic modeling procedure that we introduced in [5, 6, 7] and goes under the name of constrained, or feature space, MLLR-based speaker normalization (CMLSN). It is a variant of fMLLR-based speaker adaptive training introduced in [1]. The aim of adaptive acoustic modeling is to reduce the influence of phonetically irrelevant acoustic variability on the acoustic models.

The CMLSN 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 acoustic observations of each speaker (or acoustic condition) are transformed by applying a transformation matrix and a bias vector estimated, through fMLLR [1], in order to maximize the likelihood of the acoustic observations w. r. t. the target models. Training data are normalized once w. r. t. the target models by applying the estimated transformation. The transformed, or normalized, acoustic 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 w. r. t. other adaptive training methods is that the target models and the recognition models are independent, i. e. they may have a different model structure. We found convenient to adopt as target models triphone HMMs with just one Gaussian per state [5]. We also introduced the use of a very simple target model that is a Gaussian mixture model (GMM). As in this latter case word transcriptions of test utterances are not required for estimating the transformation parameters, acoustic data normalization can be applied at recognition stage without any preliminary decoding pass.

The proposed training method was demonstrated effective in the context of off-line applications such as broadcast news and parliamentary speeches transcription [7, 4]. Relations of the CMLSN method with other popular methods for speaker adaptive acoustic modeling are discussed in [6, 7]. In this work, we investigate the effectiveness of the method in the context of on-line adaptation with telephony speech data.

We trained, on original acoustic data, several sets of target models: a set of triphone HMMs, with a single Gaussian density per state, and GMMs with different number of Gaussian densities, i. e. 128, 256, 512 and 1024. A full transformation matrix was always adopted in experiments reported in the following sections. During training and in batch adaptation experiments three fMLLR iterations were performed for transformation parameters estimation while a single iteration was performed in on-line adaptation experiments.

3. FMAPLR

In fMLLR-based adaptation, sparse adaptation data may result in unreliable transformation parameters estimation leading to recognition performance even worse than with the unadapted system. To achieve robustness to small amount of adaptation data, in [3] a smoothed version of fMLLR statistics was derived in the maximum *a posteriori* (MAP) framework [9]. MAP estimation provides a way to incorporate prior knowledge into the transformation parameter estimation. In this framework, a prior distribution is assumed for the transformation parameters that is the extended transformation matrix including the bias vector. As prior distribution it was proposed to use an elliptically symmetric matrix variate distribution [10, 3]. Hyperparameters of this prior distribution, that is location and scale parameters, can be estimated from transformation matrices estimated w. r. t. speaker-independent models [3]. In this work, data of each training speaker were used to estimate an fMLLR transformation w. r. t. target models that is speaker-independent triphone HMMs, for the text-dependent approach, and a GMM for the text-independent approach. The obtained transformation matrices were used to estimate location vectors and full scale factor matrices to be used in the prior distribution. In principle, a disjoint development set should be used, instead of the training set, to learn the prior distribution of the feature transformation matrix. However, we found that using the training set is an acceptable compromise.

4. TELEPHONY SPEECH DATA

For training and testing we exploited a telephony German speech corpus collected by Siemens for internal use. The corpus contains recordings of phone calls, acquired with a sample rate of 8kHz, in which the speaker was asked to utter several prepared sentences including phonetically rich sentences, dates, sequence of digits, proper names, etc. Each phone call in the corpus was placed by a different speaker. The corpus was conventionally partitioned into training and testing data without speakers overlapping. The training set contains data from 1999 phone calls for a total of about 100 hours. The test set consists of 500 phone calls for a total duration of about 11 hours. There are no phonetically rich sentences in the test set and each speaker uttered on average 14 sentences. On average each sentence lasts 5.8 sec and contains 5.7 words. Silence segments represent almost half of the total duration of an utterance in this test set. Long initial and final silence segments were maintained during recognition experiments.

5. SPEECH RECOGNITION SYSTEMS

Each speech frame was parametrized into a 39-dimensional feature vector formed by 13 mel frequency cepstral coefficients (MFCCs) plus their first and second order time derivatives. Cepstral mean subtraction was performed on static features on an utterance-by-utterance basis. Heteroscedastic linear discriminant analysis was performed on 117-dimensional supervectors resulting composing three adjacent 39-dimensional observation vectors [11]. The resulting transformation matrix was applied to map back 117-dimensional supervectors into 39-dimensional observation vectors.

Acoustic models were cross-word, tied-state, left-to-right triphone HMM. Each model had a six-state Bakis topology and a Gaussian mixture associated to each state with up to 16 Gaussian densities with diagonal covariance matrix. In addition to conventionally trained baseline models, having about 30000 Gaussian densities, several other sets of HMMs were trained by exploiting the CMLSN method described above. All sets of recognition models had similar number of parameters.

Recognition experiments were conducted with a word-loop finite state network with 284 words in parallel and uniform probabilities assigned to the arcs. The recognition vocabulary was made by all words in test set.

6. BATCH UNSUPERVISED ADAPTATION

As a reference, we carried out batch adaptation experiments by assuming data of each test speaker available in one block. A two-pass decoding scheme was first adopted. The first recognition pass was carried out with conventionally trained acoustic models on unnormalized test data, and the recognition hypotheses generated were exploited as supervisions for transformation parameters estimation before performing the final decoding pass on transformed data. While the same supervision was exploited, fMLLR-based estimation of

One-pass Baseline	Baseline + Adaptation	CMLSN
10.3	8.7	7.9

Table 1. Recognition results (% WER) of unsupervised batch adaptation experiments with baseline acoustic models (*Baseline + Adaptation*) and speaker adaptively trained acoustic models (*CMLSN*). As a reference, performance achieved with the one-pass baseline system and unnormalized test data is also reported (*One-pass Baseline*).

CMLSN GMM128	CMLSN GMM256	CMLSN GMM512	CMLSN GMM1024
9.4	9.1	9.1	9.1

Table 2. Recognition results (% WER) of text-independent batch adaptation experiments achieved performing a single decoding pass with speaker adaptively trained acoustic models. Target models, for fMLLR-based speaker normalization, are GMMs with different number components.

transformation parameters, to be applied to acoustic observations before the final decoding pass, was carried out w. r. t. two different sets of models: baseline models for the baseline system and triphone HMMs, with a single Gaussian density per state, for recognition models speaker adaptively trained through the CMLSN method. Recognition results, in terms of word error rate (WER), are reported in Table 1. From the table we can see that a 10.3% WER is achieved with the one-pass baseline, while performing a second decoding pass with baseline models on normalized data leads to 8.7% WER. By performing the second decoding pass on normalized data with speaker adaptively trained models leads to 7.9% WER. This confirms that the CMLSN training procedure is effective.

Table 2 reports text-independent batch adaptation results achieved using speaker adaptively trained recognition models and GMMs target models. Four sets of speaker adaptively trained recognition models were trained, through the CMLSN method, by using as target models GMMs with different number of components: 128, 256, 512 and 1024.

Comparing results of Table 1 and Table 2 we can note that the text-dependent approach performs tangibly better than the text-independent approach. However, we point out that in this second case no preliminary decoding pass is necessary, as for each training and testing speaker the transformation parameters are estimated w. r. t. a GMM. From results reported we can also see that recognition performance does not change varying over 256 the number of Gaussian components in the GMM. Given these results, on-line adaptation experiments were carried out by using the set of speaker adaptively trained recognition models trained on data normalized w. r. t. the GMM with 256 components.

7. ON-LINE ADAPTATION EXPERIMENTS

During a phone call, speech data are progressively made available for unsupervised adaptation. We carried out experiments with two different modalities for collecting fMLLR sufficient statistics and applying the estimated transformation leading to *instantaneous* and *incremental* adaptation [12].

Instantaneous adaptation operates utterance per utterance. It attempts to improve recognition performance on the same data, i. e. the current input utterance, that are used to estimate the transformation parameters. It is better suited when there is a very short interaction, consisting of very few utterances, between the user and the system.

	Baseline + Adaptation	CMLSN	CMLSN GMM256
fMLLR	10.0	9.6	13.0
fMAPLR	9.5	8.5	10.0

Table 3. Recognition results (% WER) of instantaneous adaptation experiments with (*fMAPLR*) and without (*fMLLR*) using a prior distribution on transformation parameters.

Incremental adaptation assumes that test data are progressively available and that incoming utterances come from the same speaker and are uttered in a roughly stationary environment so that fMLLR sufficient statistics can be continuously accumulated over time and a transformation matrix is computed after each utterance and immediately applied to the next incoming utterance before decoding. In incremental adaptation no acoustic normalization is performed on the first input utterance before decoding.

In instantaneous adaptation, removing the need for word level transcription of the input utterance may help in achieving computational efficiency as the input utterance is decoded only once. In incremental adaptation, there is no need of decoding twice an input utterance anyway, as only past data and corresponding recognition hypotheses are exploited for collecting sufficient statistics.

7.1. Instantaneous adaptation

Results of instantaneous adaptation experiments are reported in Table 3. We can see that, for both the cases of baseline models and speaker adaptively trained models, fMAPLR is effective to cope with small amount of adaptation data. Noticeable is the 8.5% WER obtained by the recognition models trained through the CMLSN method using as target models triphone HMMs with a single Gaussian density per state. It is significantly better than the 9.5% WER achieved with the two-pass baseline system. Furthermore, considering that instantaneous adaptation works utterance-by-utterance instead of speaker-by-speaker, this result compares well with the 7.9% WER achieved in the corresponding unsupervised batch adaptation experiment (see Table 1). We point out that in all these cases, supervisions for transformation parameter estimation were generated by the one-pass baseline system (leading to 10.3% WER).

Instantaneous adaptation with the text-independent approach leads to 10.0% WER when transformation parameters are estimated w. r. t. a GMM target model having 256 components and fMAPLR is adopted. In this case the use of fMAPLR is essential, without it we get 13.0% WER. The 10.0% WER achieved is significantly worse than the 9.1% WER achieved in the corresponding unsupervised batch adaptation experiment reported in Table 2, however, a certain improvement is ensured w. r. t. the 10.3% WER achieved with the one-pass baseline.

7.2. Incremental adaptation

Results of incremental adaptation experiments are reported in Table 4. Results again confirm that fMAPLR is effective to cope with sparse adaptation data. With fMAPLR, text-dependent incremental adaptation exploiting speaker adaptively trained models leads to 8.7% WER, which is better than the 9.0% achieved with the baseline system. However, it is worse than the 8.5% WER achieved in instantaneous adaptation (see Table 3). Text-independent incremental adaptation leads to 9.8% WER which is better than the 10.0% WER achieved with text-independent instantaneous adaptation (see Table 3).

	Baseline + Adaptation	CMLSN	CMLSN GMM256
fMLLR	9.5	10.2	10.7
fMAPLR	9.0	8.7	9.8

Table 4. Recognition results (% WER) of incremental adaptation experiments with baseline models and speaker adaptively trained models.

	Baseline + Adaptation		CMLSN	
	fMLLR	fMAPLR	fMLLR	fMAPLR
No Adapt.	10.3	10.3	-	-
1 Utt. Incr.	21.2	10.2	37.8	11.0
2 Utt. Incr.	11.6	9.6	18.8	9.9
3 Utt. Incr.	11.0	9.3	14.5	9.6
4 Utt. Incr.	10.7	9.3	12.4	9.4
5 Utt. Incr.	9.8	9.2	11.5	9.1
Incr.	9.5	9.0	10.2	8.7

Table 5. Recognition results (% WER) of incremental adaptation experiments exploiting different amounts of adaptation data.

We point out that in incremental adaptation, when speaker adaptively trained models are used, a problem is encountered in decoding the first utterance of a speaker. In fact, decoding unnormalized data with speaker adaptively trained models results in systematic suboptimal recognition results as acoustic models are trained on normalized data. To cope with this problem we decided to decode the first utterance with baseline models.

To assess the effect of the amount of data on the level of adaptation achieved [2], we carried out a series of experiments whose results are reported in Table 5. In the table, *No Adapt.* denotes results achieved with the one-pass baseline system without adaptation. *Incr.* denotes results achieved with incremental adaptation where adaptation data are continuously collected and after each utterance a new fMLLR transformation matrix is estimated and then applied to the next input utterance before decoding. The intermediate rows report results obtained by proceeding as for the *Incr.* case but stopping reestimation of the transformation afters a certain number of utterances. For example, *4 Utt. Incr.* means that after the fourth utterance, transformation parameters are fixed and the same transformation matrix is applied to all the remaining utterances of the test speaker before decoding.

Results in Table 5 confirm effectiveness of fMAPLR: some performance improvement, over the one-pass baseline, can be observed exploiting for adaptation just two utterances. We point out that the average number of words per utterance for the first five utterances was 4.4, 3.6, 1.8, 3.9, 3.4, respectively. Furthermore, we can observe that the baseline system adapts faster than the system using speaker adaptively trained models. This may be due to the fact that, in case of speaker adaptively trained models, over-fitting effects may result emphasized as transformation parameters are estimated w. r. t. target models which are different than recognition models.

8. CONCLUSIONS

The use of speaker adaptively trained models has been investigated in the context of on-line adaptation with telephony speech data. An adaptive acoustic modeling method, we previously proved effective in off-line applications, was adopted for training the acoustic models used in text-dependent and text-independent on-line adaptation experiments.

Results have shown that fMAPLR greatly helps to achieve robustness to small amount of adaptation data when performing both instantaneous and incremental adaptation with baseline models and speaker adaptively trained models. The use of speaker adaptively trained models in conjunction with fMAPLR leads to the best results in both instantaneous and incremental adaptation. In incremental adaptation, some performance improvements can be observed exploiting for adaptation just two utterances. However, baseline models adapt faster in incremental adaptation.

The proposed text-independent adaptation approach, exploiting speaker adaptively trained models, is also proven effective and may represent an interesting solution for instantaneous adaptation when it is preferable to avoid multiple decodings of the input utterance.

9. REFERENCES

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

[2] Y. Li, H. Erdogan, Y. Gao, and E. Marcheret, "Incremental online feature space MLLR adaptation for telephony speech recognition," in *Proc. of ICSLP*, Denver,Colorado, Sep. 2002, pp. 1417–1420.

[3] X. Lei, J. Hamaker, and X. He, "Robust Feature Space Adaptation for Telephony Speech Recognition," in *Proc. of INTERSPEECH*, Pittsburgh, PA, Sept. 2006, pp. 773–766.

[4] D. Giuliani and F. Brugnara, "Experiments on Cross-System Acoustic Model Adapatation," in *ASRU Workshop 2007*, Kyoto, Japan, Dec. 2007.

[5] D. Giuliani, M. Gerosa, and F. Brugnara, "Speaker Normalization through Constrained MLLR Based Transforms," in *Proc. of INTERSPEECH*, Jeju Island, Korea, Oct. 2004, pp. 2893–2897.

[6] D. Giuliani, M. Gerosa, and F. Brugnara, "Improved automatic speech recognition through speaker normalization," *Computer Speech and Language*, vol. 20, pp. 107–123, 2006.

[7] G. Stemmer, F. Brugnara, and D. Giuliani, "Adaptive Training Using Simple Target Models," in *Proc. of ICASSP*, Philadelphia, PA, March 2005, pp. I–997–1000.

[8] P. Kenny, V. Gupta, G. Boulianne, P. Ouellet, and P. Dumouchel, "Feature Normalization Using Smoothed Mixture Transformations," in *Proc. of INTERSPEECH*, Pittsburgh, PA, Sept. 2006, pp. 17–21.

[9] J.-L. Gauvain and C.-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[10] Wu Chow, "Maximum a Posterior Linear Regression with Elliptically Symmetric Matrix Variate Priors," in *Proc. of EUROSPEECH*, Budapest, Hungary, Sept. 1999, pp. 1–4.

[11] G. Stemmer and F. Brugnara, "Integration of Heteroscedastic Linear Discriminant Analysis (HLDA) into Adaptive Training," in *Proc. of ICASSP*, Toulouse, France, May 2006, pp. I–1185–1188.

[12] G. Zavaliagkos, R. Schwartz, and J. Makhoul, "Batch, Incremental and Instantaneous Adaptation Techniques for Speech Recognition," in *Proc. of ICASSP*, Detroit, May 1995, pp. I–676–679.