Instantaneous Speaker Adaptation through Selection and Combination of fMLLR Transformation Matrices

Diego Giuliani and Fabio Brugnara

Human Language Technology Research Unit
FBK- Fondazione Bruno Kessler, Trento, Italy
giuliani@fbk.eu, brugnara@fbk.eu

10.21437/Interspeech.2011-658

Abstract

This paper addresses instantaneous speaker adaptation, based on feature-space maximum likelihood linear regression (fMLLR), in the context of an automatic transcription task. We investigate the use of fMLLR-based adaptation when the need of a preliminary decoding pass for a speech segment is removed, as sufficient statistics for adaptation parameter estimation are gathered with respect to a Gaussian mixture model. To cope with limited adaptation data, in addition of using feature-space maximum *a posteriori* linear regression (fMAPLR), an investigation is conducted where the transformation matrix to be applied to the speech segment is estimated through selection and combination of pre-computed fMLLR transformation matrices. For speaker adaptively trained acoustic models results of recognition experiments show that the proposed approach is moderately better than fMLLR but not as good as fMAPLR.
Index Terms: instantaneous speaker adaptation, fMAPLR, fMLLR transformation matrices selection and combination, automatic speech recognition

1. Introduction

In instantaneous adaptation, adaptation occurs on the same speech segment to be decoded and fMLLR offers an efficient and simple way for reducing the acoustic mismatch between training and testing conditions [1]. fMLLR employs, in fact, a single transformation matrix and a bias vector to linearly transform the input acoustic features before decoding. 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 hidden Markov models (HMMs), assuming a word level transcription of the acoustic data. Effectiveness of fMLLR in reducing the acoustic mismatch between the speech recognition models and the input acoustic data has been proved on a number of different application domains [1, 2, 3]. It is well known that effectiveness of fMLLR adaptation is usually reduced when there are very sparse adaptation data [2, 4, 5, 6]. To tackle the problem of unreliable transformation parameter estimation several approaches have been proposed such as smoothing of fMLLR sufficient statistics [2, 6], fMAPLR [4] and eigenspace-based fMLLR [7, 8].

In this work, we investigate fMLLR-based adaptation when the need of a preliminary decoding pass for transformation parameter estimation is removed as sufficient statistics for adaptation parameter estimation are gathered with respect to a Gaussian mixture model (GMM). In [9] we proposed the use of a GMM as target acoustic model for transformation parameter estimation for fMLLR-based speaker adaptive acoustic modeling. Effectiveness of this approach in off-line transcription tasks [9, 3] as well as for on-line incremental adaptation in a telephony application [5] was proved. In this adaptation scenario, we propose an approach based on selection and combination of pre-computed fMLLR matrices for rapid and robust estimate of the transformation matrix to be applied on the incoming speech segment before decoding.

Fast selection of pre-computed matrices was proposed in the context of linear transformation approach to vocal tract length normalization (VTLN). In this approach, a linear transformation of the acoustic features is used to approximate the complex, non-linear, warping of the frequency axis. In linear VTLN [10], for each of the possible warping factors, training data are warped and an fMLLR transformation matrix is estimated to reduce the mismatch between the warped data and baseline models trained on unwarped data. Instead, in [11] a set of VTLN warp-matrices are analytically computed, as opposed to being estimated from training data as in the previous case, by using a method to obtain VTLN-warped features through linear transformation of unwarped mel frequency cepstral coefficients (MFCCs).

In [10, 11], given the speaker data, the best suited pre-computed transformation matrix is selected as the one that maximizes the same auxiliary function used for fMLLR transformation parameters estimation. The pre-computed transformation matrices are ranked by comparing the values of the fMLLR auxiliary function. We use this selection criterion considering different sets of pre-computed transformation matrices e.g. a set of speaker-specific fMLLR transformation matrices estimated using training data. Furthermore, the resulting N best transformation matrices are combined using the technique proposed for eigenspace-based fMLLR [8]. In this technique, under the assumption of diagonal covariance matrices for Gaussian density emission distributions, each row of the fMLLR transformation matrix to be determined is expressed as a bias vector plus a linear combination of eigenvectors with combination coefficients estimated by maximizing the auxiliary function. In our case, for each row, instead of eigenvectors, the corresponding rows in the N selected transformation matrices are linearly combined.

Results of instantaneous adaptation experiments with the proposed approach are presented for a large vocabulary speech transcription task and compared with those obtained with fMLLR and fMAPLR.

The rest of this paper is organized as follows. Section 2 describes the techniques used for obtaining robust estimation of fMLLR transformation matrices in the instantaneous adaptation scenario investigated. The experimental recognition set-up is described in Section 3 while experiment results are reported and discussed in Section 4. Conclusions are reported in Section 5 which concludes the paper.

This work was partly supported by Pervoice under the grant *Speech Analytics*.

Copyright © 2011 ISCA

2. Speaker Adaptation Techniques

In this work we cope with an application scenario in which each speech segment is individually processed. Furthermore, while the segment is available for multiple processing, multiple recognition of the segment is avoided. In all experiment reported, parameters for acoustic feature normalization are always estimated in order to maximize the likelihood of the normalized adaptation data w. r. t. to a GMM modeling the acoustic training data.

In this section, several techniques for obtaining a robust estimate of the fMLLR matrix to be applied to acoustic observations before recognition are described.

2.1. fMAPLR

To achieve robustness to small amount of adaptation data, in [4] a smoothed version of fMLLR statistics was derived in the maximum *a posteriori* (MAP) framework. 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 [12, 4]. Parameters of this prior distribution, that is location and scale parameters, can be estimated from transformation matrices estimated w. r. t. speaker-independent models [4]. In this work, given the data of each training speaker, an fMLLR transformation is estimated w. r. t. a GMM trained on all training data. The obtained transformation matrices are 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 parameters of the prior distribution of the feature transformation matrix. However, we found that using the training set is an acceptable compromise.

2.2. fMLLR transformation matrices combination

Eigenspace-based fMLLR projects the fMLLR transformation onto its dominant principal components [8]. Under the assumption of diagonal covariance matrices for Gaussian density emission distributions, the transformation matrix to be determined can be dealt with row by row. Given the pool of fMLLR transformation matrices, principal component analysis is performed row by row on the collection of row vector examples and the N eigenvectors corresponding to the N most dominant eigenvalues are selected. Given the adaptation data, each row of the transformation matrix to be determined is expressed as a bias vector plus a linear combination of the selected eigenvectors. The combination coefficients for the selected eigenvectors are estimated in an iterative scheme row by row by maximizing the auxiliary function by means of the steepest descent algorithm. In [8] the method was validated in the context of an incremental adaptation task.

In this work, instead of relying on principal component analysis, each row of the transformation matrix to be determined is expressed as a linear combination of row vectors from a set of pre-computed fMLLR transformations. This set of pre-computed fMLLR transformations is determined for each speech segment as described in Section 2.3 below. Combination coefficients are still estimated as proposed in [8] for eigenspace-based fMLLR. For comparison purposes, results of adaptation experiments in which each row of the transformation matrix to be determined is expressed as linear combination having uniform coefficients are also reported. In this latter case, the estimated transformation matrix corresponds to the average of the pre-computed fMLLR transformation matrices to be combined.

2.3. fMLLR transformation matrices selection

Given an incoming speech segment and a set of pre-computed fMLLR transformation matrices, the same auxiliary function used for fMLLR transformation parameters estimation can be used for selecting the best suited transformation to be applied to the incoming segment before decoding [10, 11]. In this method, instead of using the fMLLR auxiliary function for transformation parameter estimation, the fMLLR auxiliary function is used for ranking the pre-computed fMLLR matrices: each pre-computed transformation matrix is plugged into the auxiliary function and the values of the auxiliary function obtained are then compared. In this work, instead of limiting ourselves to consider the best suited transformation as proposed in the original method [10, 11], the N best pre-computed transformation matrices are selected and combined as described in Section 2.2 to estimate the desired transformation matrix. Furthermore, the selection procedure was validated by considering three sets of fMLLR transformations matrices: 1) speaker-specific fMLLR transformation matrices estimated over training speakers; 2) prototype transformation matrices obtained through K-means clustering of the vectorial representation of speaker-specific transformation matrices; 3) transformation matrices estimated for a discrete number of warping factors.

In the former case, given the audio data of a speaker in the training set an fMLLR transformation matrix is estimated w. r. t. a GMM trained on all training data. In our case the data of each training speaker are identified in automatic way. To this end, each audio file in the training set is automatically segmented and speech segments found are clustered based on acoustic similarity [13]. A transformation is then estimated for each cluster of speech segments resulting in about 11,000 transformations, that do not correspond to truly different speakers as speech of the same speaker may occur in different audio files.

A second set of transformations is obtained through K-means clustering of the vectorial representation of speaker-specific transformation matrices using an Euclidean distance measure between supervectors. The K centroids of the obtained partition are taken as prototype transformation matrices. Results of experiments with K set to 256 are reported in this paper.

The third set of transformations is determined by first partitioning the training set by grouping data of speakers sharing the same warping factor. To this end, for each cluster of speech segments an optimal frequency warping factor is estimated as in conventional VTLN procedure [14]. To assign a frequency warping factor to a cluster of speech segments, a grid search over a possible set of frequency warping factors is carried out which maximizes the likelihood of the cluster warped data with respect to a set of speaker independent triphone HMMs trained on the original training data. The warping process is implemented by changing the spacing and the width of the filters in the mel filter-bank while maintaining the speech spectrum unchanged [14]. 18 warping factors equally spaced in the range 0.82 to 1.16 are considered. Original data of speakers sharing a given warping factor are then pooled and a fMLLR transformation matrix is estimated w. r. t. a GMM trained on all original training data. When applied the estimated transformation matrices should compensate for the effect induced by vocal tract length differences.

3. Experimental Set-up

Acoustic models (AMs) were trained on about 223 hours of speech data obtained merging the Italian Broadcast News corpus, about 130 hours of audio recordings (mostly acquired from TV and radio broadcast news programs) and a corpus consisting

of audio recordings of Italian political speeches acquired in the Italian Parliament, about 93 hours of recordings. Two sets of HMMs were trained, in both cases AMs were state-tied, cross-word, speaker-independent triphone HMMs. Output probability distributions were modeled by mixtures of Gaussian probability density functions having diagonal covariance matrices. A phonetic decision tree is used for tying states and for defining the context-dependent allophones.

For baseline models acoustic feature were 13 MFCCs plus their first and second order time derivatives. Cepstral mean subtraction is performed on static features on a segment-by-segment basis. Baseline models were trained with a conventional training maximum likelihood procedure. A GMM with 1024 Gaussian components was also trained and used later for estimating adaptation parameters.

Speaker adaptively trained models were trained as described below. Each speech frame is parametrized into a 52-dimensional observation vector composed of 13 MFCCs plus their first, second and third order time derivatives. Cepstral mean subtraction is performed on static features on a segment-by-segment basis. A projection of acoustic feature space, based on heteroscedastic linear discriminant analysis (HLDA), is embedded in the feature extraction process as follows [15]. A GMM with 1024 Gaussian components is first trained on the original 52-dimensional observation vectors. Acoustic observations in each, automatically determined, cluster of speech segments, are then normalized by applying an affine transformation estimated w. r. t. the GMM through fMLLR [1]. After normalization of training data, an HLDA transformation is estimated w. r. t. a set of state-tied, cross-word, gender-independent triphone HMMs with a single Gaussian per state, trained on the normalized 52-dimensional observation vectors. The HLDA transformation is then applied to project the set of 52 normalized features into a 39-dimensional feature space. Triphone HMMs are trained on these normalized, HLDA projected, acoustic features.

A fourgram language model (LM) with a recognition vocabulary of a million of words is employed for decoding. The LM was trained using an Italian text corpus of 606M words mainly formed of texts from the journalistic domain.

The test set consists of 3h:33m (about 33,000 uttered words) of broadcast news shows from different TV channels. On average each automatically determined speech segment lasts 15s. Recognition results are obtained with a single decoding pass on original acoustic observation or on normalized acoustic observations. For both baseline models and speaker adaptively trained acoustic models, parameters for acoustic feature normalization are estimated for each speech segment in order to reduce the acoustic mismatch w. r. t. a GMM with 1024 components.

4. Experimental Results

Table 1 reports reference recognition results, in word error rates (WERs), for baseline and speaker adaptively trained models, 19.7% and 18.0% WER respectively.

It can be noted that fMLLR adaptation is effective in improving performance of baseline models, from 19.7% to 18.7% WER. Performing more than one fMLLR iteration does not result in significant performance improvement. With speaker adaptively trained acoustic models performing more than one fMLLR iteration worsens recognition performance. It is possible that, in this latter case, performing more iterations induces over-fitting phenomena during transformation parameter estimation. Speaker adaptively trained acoustic models outperform baseline models and this is more evident when fMAPLR is adopted. The use of fMAPLR ensures better performance than using fMLLR especially when speaker adaptively trained acoustic models are used. By performing three iterations with fMAPLR an 18.2% WER is achieved with baseline models and a 16.9% WER with speaker adaptively trained acoustic models.

system	no adapt.	fMLLR		fMAPLR	
		1 iter	3 iter	1 iter	3 iter
Baseline	19.7	18.7	18.6	18.7	18.2
Adaptive	-	18.0	18.2	17.2	16.9

Table 1: *WERs achieved with the baseline and speaker adaptively trained acoustic models by performing fMLLR and fMAPLR acoustic feature normalization.*

system	N best from 11,000 speaker-specific matrices					
	1	3	5	10	20	40
Baseline	19.5	19.2	19.0	19.1	19.1	19.1
Adaptive	18.4	17.9	17.8	17.8	17.9	17.9

Table 2: *WERs achieved with baseline and speaker adaptively trained acoustic models by performing acoustic feature normalization based on uniform-coefficients combination of N pre-computed, speaker-specific, fMLLR transformations.*

Table 2 reports results achieved by performing on each speech segment acoustic feature normalization by applying a transformation matrix computed through uniform combination of pre-computed, speaker-specific, fMLLR matrices selected as described in Section 2.3. With the baseline models, application of the best suited pre-computed transformation matrix leads to some marginal improvement over the unadapted system, from 19.7% to 19.5% WER. Combination of the 5 best pre-computed fMLLR transformations improves performance further to 19.0% WER. With the speaker adaptively trained models a 17.8% WER is achieved by combining the 5 best pre-computed fMLLR transformations. It can be noted that no further benefit is achieved by combining a larger set of pre-computed matrices.

system	N best from 11,000 speaker-specific matrices					
	1	3	5	10	20	40
Baseline	19.5	18.8	18.7	18.7	18.7	18.6
Adaptive	18.4	17.6	17.6	17.3	17.4	17.3

Table 3: *WERs achieved with baseline and speaker adaptively trained acoustic models by performing acoustic feature normalization based on estimated-coefficients combination of N pre-computed, speaker-specific, fMLLR transformations.*

Tables 3 reports results obtained by combining, with estimated coefficients, pre-computed, speaker-specific, fMLLR transformations selected as described in Section 2.3. By comparing these results with corresponding results in Tables 2 we can see that estimation of combination coefficients systematically results in slightly better performance than using uniform coefficients in the combination. With 10 pre-computed fMLLR transformations entering in the combination an 18.7% WER is achieved with baseline models while 17.3% WER with speaker adaptively trained models. For baseline models this result is aligned with the result achieved with fMLLR, while for speaker adaptively trained acoustic models this result is better than that achieved with fMLLR (see Table 1). Furthermore, results achieved are aligned with results achieved with one iteration of

fMAPLR but worse than results achieved with 3 iterations of fMAPLR.

system	N best from 256 prototype matrices					
	1	3	5	10	20	40
Baseline	19.3	19.0	19.0	18.8	18.9	19.0
Adaptive	18.0	17.8	17.8	17.7	17.7	17.8

Table 4: *WERs achieved with baseline and speaker adaptively trained acoustic models by performing acoustic feature normalization based on estimated-coefficients combination of N precomputed prototypes of fMLLR transformations.*

system	N best from 18 warping factor specific matrices				
	1	3	5	10	18
Baseline	19.6	18.8	18.9	18.8	18.9
Adaptive	18.4	17.8	17.8	17.9	18.3

Table 5: *WERs achieved with baseline and speaker adaptively trained acoustic models by performing acoustic feature normalization based on estimated-coefficients combination of N precomputed, warping factor specific, fMLLR transformations.*

Table 4 reports recognition results obtained by combining, with estimated coefficients, pre-computed prototype matrices obtained through K-means clustering of speaker-specific fMLLR transformation matrices. With 10 prototype matrices entering in the combination an 18.8% WER is achieved with baseline models while a 17.7% WER is achieved with speaker adaptively trained acoustic models. These results are slightly worse than those reported in Table 3 for speaker-specific fMLLR transformations. The trend of performance with the number of transformations entering in the combination is quite smooth and the advantage in enlarging the number of transformation matrices entering in the combination appears to be reduced. The result achieved with speaker adaptively trained acoustic models, 17.7% WER, is encouraging as it shows that the combination approach is still slightly better than using fMLLR (i.e. 18.0% WER). This result shows how to cope with computational issues in matrix selection when very large numbers of speaker-specific transformations are available. The optimal number of prototype fMLLR matrices to be used is an open issue that merit further investigations but it is not discussed in this paper.

Table 5 reports results obtained by combining, with estimated coefficients, warping factor specific fMLLR transformations. With 3 fMLLR transformations entering in the combination recognition results are aligned with those achieved with the combination of prototypes of fMLLR transformations reported in Table 4. As expected, in this case, recognition performance does not improve by further enlarging the number of fMLLR transformations entering in the combination.

5. Conclusions

This paper has reported on an initial investigation on instantaneous speaker adaptation based on-the-fly selection and combination of pre-computed fMLLR transformation matrices.

For speaker adaptively trained models, results of recognition experiments show that with pre-computed speaker-specific transformations the proposed approach is slightly better than fMLLR but not as good as fMAPLR. Results confirm, in fact, that fMAPLR is very effective to achieve reliable estimation of transformation parameters.

Future work will focus on the validation of the proposed approach in several application scenarios that, for example, encompass instantaneous and incremental adaptation with extremely limited data such as for example very few words. Future work will be also devoted to experimentation with different sets of pre-computed transformation matrices.

6. 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] D. Giuliani and F. Brugnara, "Experiments on Cross-System Acoustic Model Adapatation," in *ASRU Workshop 2007*, Kyoto, Japan, Dec. 2007, pp. 117–122.

[4] 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.

[5] D. Giuliani, R. Gretter, and F. Brugnara, "On-line speaker adaptation on telephony speech data with adaptively trained acoustic models," in *Proc. of ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 4385–4388.

[6] C. Breslin, H. Xu, K.K. Chin, M. Gales, and K. Knill, "Prior Information for Rapid Speaker Adaptation," in *Proc. of INTERSPEECH*, Chiba, Japan, Sept. 2010, pp. 1644–1647.

[7] K. Visweswariah, V. Goel, and R. Gopinath, "Structuring linear transforms for adaptation using training time information," in *Proc. of ICASSP*, Orlando, Florida, May 2002.

[8] X. Cui, J. Xue, and B. Zhou, "Improving Online Incremental Speaker Adaptation with Eigen Feature Space MLLR," in *ASRU Workshop*, Merano, Italy, Dec. 2009, pp. 136 – 140.

[9] 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.

[10] D.Y. Kim, S. Umesh, M.J.F. Gales, T. Hain, and P.C. Woodland, "Using VTLN for Broadcast News Transcription," in *Proc. of INTERSPEECH*, Jeju Island, Korea, Oct. 2004.

[11] Akhil P.T., Rath S.P., Umesh S., and Sanand D.R., "A computationally efficient approach to warp factor estimation in VTLN using EM algorithm and sufficient statistics," in *Proc. of INTERSPEECH*, Brisbane, Australia, Sept. 2008, pp. 1713–1716.

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

[13] Mauro Cettolo, "Segmentation, classification and clustering of an italian broadcast news corpus," in *Proc. of Content-Based Multimedia Inf. Access Conf. (RIAO)*, Paris, France, 2000, pp. 372–381.

[14] L. Lee and R.C. Rose, "Speaker Normalization Using Efficient Frequency Warping Procedure," in *Proc. of ICASSP*, Atlanta, GA, May 1996, pp. I–353–356.

[15] 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.