Factor analysis back ends for MLLR transforms in speaker recognition

Nicolas Scheffer, Yun Lei, Luciana Ferrer

SRI International, Menlo Park, California, USA

Abstract

The purpose of this work is to show how recent developments in cepstral-based systems for speaker recognition can be leveraged for the use of Maximum Likelihood Linear Regression (MLLR) transforms. Speaker recognition systems based on MLLR transforms have shown to be greatly beneficial in combination with standard systems, but most of the advances in speaker modeling techniques have been implemented for cepstral features. We show how these advances, based on Factor Analysis, such as eigenchannel and ivector, can be easily employed to achieve very high accuracy. We show that they outperform the current state-of-the-art MLLR-SVM system that SRI submitted during the NIST SRE 2010 evaluation. The advantages of leveraging the new approaches are manyfold: the ability to process a large amount of data, working in a reduced dimensional space, importing any advances made for cepstral systems to the MLLR features, and the potential for system combination at the ivector level.

Index Terms: speaker verification, MLLR, factor analysis

1. Introduction

The speaker recognition community has seen tremendous changes in the past few years so as to produce robust speaker models and employ advanced modeling techniques. The dominant features used for the speaker verification task are cepstral. After the widely adopted Gaussian mixture model (GMM) approach along with a universal background model (UBM) and the use of support vector machines (SVM), channel compensation techniques such as Nuisance Attribute Projection (NAP) have been drastically reducing the error rates. Recently, these features have been modeled best using the Joint Factor Analysis (JFA) paradigm [1]. One drawback of this paradigm is that it evolves in the GMM mean supervector space of a usually very large dimension. A recent effort, originated by [2] and pursued further during the BOSARIS workshop[1] [3], led to new modeling techniques evolving in a reduced space, called the *ivector space*. The literature reports that ivector-based systems outperform JFA. Since the ivector space is of dimension of several hundred, techniques that could not be practically implemented earlier can now be used.

Maximum Likelihood Linear Regression (MLLR) transforms features have proved to be beneficial for speaker recognition as well as bringing complementary information to standard cepstral systems [4]. These features have not seen their processing evolve over the years. Indeed, the modeling paradigm is based on SVM and NAP for channel compensation. This work is a prelude to showing how MLLR transforms can leverage the advancements that have been done for GMM models and cepstral features, that the back ends for the cepstral and MLLR features can be unified, and that MLLR ivectors can be effectively produced. For the scientist, the advantages of using an ivector approach are numerous: the ability to process a large amount of data, working in a reduced dimensional space, import improvements from cepstral system modeling and the potential for system combination at the ivector level. In this work, we show how the framework of eigenchannel and ivector can be transposed to these features without loss of accuracy.

The first section focuses on the algorithms that work in the highly dimensional space of the MLLR transforms. We report results using inner-product scoring and eigenchannel compensation, as well as the Probabilistic Linear Discriminant Analysis (PLDA) model. The second section shows how we can generate high-performing MLLR ivectors and thus transfer the speaker recognition problem in a reduced dimensional space. We then show the performance of the SRI MLLR-SVM system during NIST SRE 2010 evaluation as a point of reference, and compare it with the presented approaches. Finally, we present system combination experiments at both the score and ivector level.

2. Commonalities

We describe the experimental protocol, the process used to generate the MLLR transforms, and the background data used throughout our work.

2.1. Experimental protocol

All experiments were performed on the NIST SRE 2010 (SRE10) extended data set. We focus on the telephone condition, in both enrollment and test (Condition 5 in the evaluation plan[2]). The amount of speech in the utterances is approximately 2.5 minutes. We report results on both genders, and hyperparameters are chosen so that the performance generalize across genders. For this condition, the number of trials is 7169 true target trials (3704 female, 3465 male) and 408,950 impostor trials (233,077 female, 175,873 male). All results are reported using the new as well as the standard minimum decision cost function (DCF) for the evaluation along with the equal error rate (EER). Our main focus will however be the performance at the standard minimum DCF.

2.2. MLLR transforms

The MLLR system uses the speaker adaptation transforms from the speech recognition system as features for speaker verification. A total of 16 affine 39x40 transforms were used to map the Gaussian mean vectors from speaker-independent to speaker-

This work was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U. S. Government.

[1] http://speech.fit.vutbr.cz/en/workshops/bosaris-2010

[2] http://www.itl.nist.gov/iad/mig/tests/sre/2010/

Table 1: *Inner-product of MLLR transforms using different feature normalization. Results on NIST SRE 2010, Condition 5. New DCF / Standard DCF / EER.*

Feature norm.	Females	Males
raw features	.989 / .972 / 36.1	.997 / .987 / 38.8
mean normalized	.930 / .514 / 12.18	.970 / .428 / 9.44
z-normalized	.878 / .400 / 9.67	.784 / .305 / 7.91
z-norm + s-norm	.670 / .324 / 8.94	.629 / .267 / 7.79

dependent speech models; eight transforms were estimated relative to the male and female recognition models, respectively and independently of the speaker true gender. More details for the speech recognition system used can be found in [5]. The resulting vector is of dimension 24,960.

2.3. Background data

The background data used for channel compensation, subspace estimation, and score normalization is monolingual (English only) given the nature of SRE10. This data was gathered from SRE04, SRE06 and Switchboard (Phase 2 and 3 and Cellular parts 1 and 2), which amount in 12,193 utterances of 1045 female speakers and 9696 utterances of 796 male speakers. Score normalization is applied when indicated using S-norm. S-normalization is computed as the average of the T-normalized scores for the enrollment utterance and the Z-normalized scores for the test utterance. The S-norm cohort uses all the speakers mentioned above.

3. Experiments in high dimensional space

The experiments here focus on algorithms evolving in the highly dimensional space of the input MLLR features.

3.1. Feature normalization and inner-product performance

Since the objective of this work is to revisit the modeling of MLLR transforms for speaker recognition, our first step is to assess their discriminative power using an inner product as a scoring function. Table 1 shows the performance of the system with different feature normalizations as well as the effect of score normalization.

The best performance is obtained using mean-variance normalization (z-normalization) of the features, trained on the background data. For the rest of this work, z-normalized features and s-normalized scores will be used, unless indicated otherwise.

3.2. Eigenchannel compensation

The benefits of channel (or session) compensation for speaker recognition has been demonstrated multiple times in the literature. In the cepstral domain, while NAP was used in the past, the standard method is now eigenchannel [6, 7]. The eigenchannel approach is equivalent to a *Probabilistic NAP*, as both NAP and eigenchannel aim at estimating a subspace that spans the intraspeaker variability. Indeed, the problem is formulated in a probabilistic way with the assumption of an underlying GMM distribution, the UBM. Since there is not any UBM involved for MLLR transforms, adapting this technique consists of considering a single Gaussian, which then reverts to a standard Probabilistic PCA algorithm as in [8]. The standard eigenchannel algorithm for cepstral features can then be used, which is

$$\mathbf{o} = \mathbf{o_s} + \mathbf{U}\mathbf{x} \tag{1}$$

Table 2: *Influence of the eigenchannel subspace rank for MLLR transforms. NIST SRE 2010 Condition 5. New DCF / Standard DCF / EER. The removal of 30 dimensions is enough to get high accuracy.*

Subspace rank	Females	Males
none	.670 / .324 / 8.94	.629 / .267 / 7.79
10	.493 / .192 / 4.32	.399 / .145 / 3.80
30	.462 / .171 / 4.08	.388 / .133 / 3.69
50	.441 / .176 / 3.95	.395 / .137 / 3.95
70	.431 / .175 / 4.08	.400 / .141 / 3.98
90	.433 / .178 / 4.27	.404 / .142 / 3.89
110	.445 / .180 / 4.29	.406 / .144 / 3.98

where o represents the MLLR transform for an utterance, o_s is the speaker model, $\mathbf{U}$ is a low rank tall matrix that spans the subspace of intra-speaker variability, and x is the latent variable of the model, with a prior Gaussian distribution $\mathcal{N}(0, \mathbf{I})$. One key difference with NAP, apart from the Expectation Maximization (EM) algorithm used for estimating the subspace, is that the projection of the input vector is scaled by the covariance of the posterior distribution of the latent variable. This estimate $\mathbf{P}^{-1}$, is constant for all utterances which makes training of the subspace fast. The point estimate x, representing the contribution of the subspace, is given by

$$\mathbf{x} = \mathbf{P}^{-1}\mathbf{U}'\mathbf{o}, \mathbf{P} = \mathbf{I} + \mathbf{U}'\mathbf{U} \tag{2}$$

and is constant for all input utterances.

For the experimental setup, we used a maximum likelihood training of 30 iterations, with minimum divergence steps every 4 iterations. The verification score is produced by removing the subspace contribution from the MLLR transform for the training and test utterances, followed by an inner product. Table 2 reports the performance of the eigenchannel MLLR system with different dimensions of subspace rank.

As predicted, the benefit of eigenchannel is clear, and only 30 dimensions have to be removed to get the maximum benefit. This is consistent with MLLR-SVM systems where a NAP dimension of 50 was usually reported as optimal.

3.3. PLDA

Once the eigenchannel approach is adapted, one can imagine the natural extension to the full JFA model for the MLLR features. However, when dealing with a single Gaussian model, there exists a special case of JFA called Probabilistic Linear Discriminant Analysis (PLDA) model. This model was earlier used in the context of ivector systems to produce log-likelihood ratio scores [9, 10]. In the case of MLLR transforms, we then assume that they are distributed according to

$$\mathbf{o} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x} + \epsilon \tag{3}$$

where $\mathbf{V}$ and $\mathbf{U}$ are low rank matrices describing the speaker and channel subspaces, trained using the EM alogrithm. Using the PLDA model, one can directly evaluate the log-likelihood ratio for the speaker recognition hypothesis test. This can be evaluated analytically, and scoring can be performed very efficiently.

Table 3 shows the results when applying PLDA to the MLLR transforms for different speaker and channel subspace ranks. These results show that the PLDA model can result in impressive improvements similar to the eigenchannel approach. Using 100 dimensions for the speaker space and 30 for the channel space seems to be optimal, which is consistent with the

Table 3: *PLDA approach applied to MLLR transforms for different ranks. SRE10 Cond 5. New / Standard DCF / EER.*

	Rank V / U	Females	Males
none		.670 / .324 / 8.94	.629 / .267 / 7.79
EC	/ 30	.462 / .171 / 4.08	.388 / .133 / 3.69
PLDA	50 / 30	.654 / .208 / 4.83	.523 / .173 / 3.89
PLDA	50 / 50	.806 / .221 / 4.13	.634 / .185 / 3.86
PLDA	100 / 30	.566 / .179 / 4.29	.466 / .150 / 3.60
PLDA	100 / 50	.679 / .184 / 4.13	.484 / .157 / 3.95

eigenchannel approach. However, this system is slighly less performant than the eigenchannel approach. The authors believe that more training data for the PLDA model might help to bridge the gap between the two approaches.

4. Experiments in low dimensional space

Being able to work in a smaller dimensional space has many advantages. The speaker recognition community has lately been successful in working in a space of dimension of several hundred, called the ivector space [2]. The ivector paradigm is integrated in the feature extraction process, since a single ivector is produced in the same way for each utterance. This makes further processing much easier, faster, and memory efficient. This approach aims at describing the total variability in the data, and use a factor analysis approach to produce the vectors. The amount of data available to the scientists in now so large that this algorithm can effectively reduce the dimension of the input features – the GMM supervector space of dimension several thousand – to a few hundred while preserving most of the speaker information.

The ivector paradigm uses a variant of the JFA framework, which assumes that speaker and channel subspaces are not decoupled. A single subspace is used to cover the total variability and employed to produce the vectors. The model is $o = m + \mathbf{T}i$ where o is the MLLR feature, $\mathbf{T}$ spans a linear subspace in the original space representing the total variability, and the latent variables i are called ivectors.

The subspace is estimated using the EM algorithm the same way as in section 3.2. The speaker verification score is produced by computing the cosine distance between the two ivectors.

In Figure 1, we show the performance of this system (*cosine* line). The value of the rank does not seem to influence greatly the performance. However, since the ivector extraction does not factor out the intraspeaker variability, the speaker and channel information is convoluted in the vector. We then need to remove the nuisance information to find the optimal values for this parameter.
We adopted a Linear Discriminant Analysis (LDA) back end followed by within-class covariance normalization (WCCN), as in [2]. Both LDA and WCCN are estimated on the whole background data. The selection of number of dimension to preserve as well as the influence of the ivector dimension on the performance of the system is shown in Figure 1.

From these results, it seems that using a rank of 500 and retaining 100 dimensions of LDA is suitable for both genders and all errors metrics. For these values, the MLLR ivector system matches the performance of the eigenchannel approach as seen in Table 4. This supports the fact that high-performing MLLR ivectors can be effectively generated as for cepstral features.

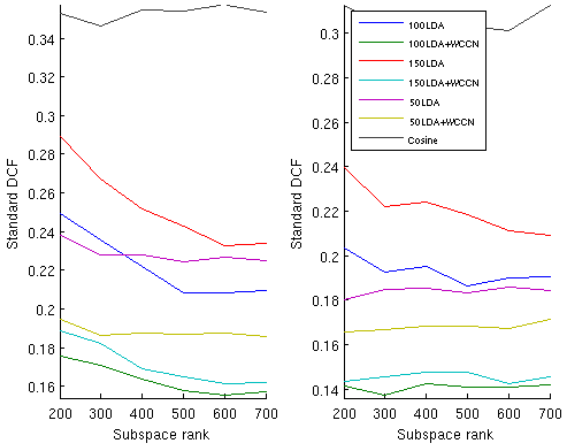

Figure 1: *Using LDA and WCCN backends for the MLLR ivector system for different subspace ranks and LDA dimensions. SRE10 Condition 5. Females/Males. Standard DCF / EER*

5. Comparison with MLLR-SVM

As a reference point to the performance presented in this paper, we show the results of the SRI MLLR-SVM system submitted to the NIST SRE 2010 evaluation, whose performance can be considered state of the art.

The SRI MLLR system uses the NAP-compensated MLLR transform followed by an SVM back end. The configuration of this system is as follows. The within-speaker variance was estimated on SRE04 telephone data, SRE05 microphone data, SRE08 and SRE10 sample data, and SRE08 speakers designated for training. The impostor (background) data for SVM training was from SRE06 telephone and microphone sessions, as well as from SRE08 data designated for training. For more details, the reader can refer to the SRI NIST SRE 2010 system paper [5]. Performance of this system is summarized in Table 4 along with the MLLR eigenchannel and the MLLR ivector system. The approaches employed in our work are able to match the performance of the MLLR-SVM system. Moreover, the relative improvement with respect to the SRI MLLR-SVM system is of around 15% for both genders at the standard DCF.

6. System combination

MLLR features are attractive, as they bring complementarity to the information provided by a cepstral-based system. For instance, in SRI NIST SRE 2010 submission, the MLLR system brought up to 30% improvement to the baseline JFA system.

One of the attractiveness of the ivector paradigm lies in the possibility of system combination at an earlier stage, namely in the ivector space. The following presents a comparison between the score- and ivector-level combination approaches.

Table 4: *SRI MLLR SVM system for the NIST SRE 2010 evaluation. Comparison with the eigenchannel and ivector approaches. NIST SRE 2010 Condition 5. New DCF / Standard DCF / EER*

System	Females	Males
SRI MLLR-SVM	.475 / .193 / 4.63	.462 / .162 / 4.36
MLLR eigenchannel	.462 / .171 / 4.08	.388 / .133 / 3.69
MLLR ivectors	.485 / .158 / 3.32	.432 / .141 / 3.29

Table 5: *System combination with cepstral- and a MLLR- based ivector systems. Score-level combination as well as model-level combination in the ivector space*

System	Females	Males
cepstral	.541 / .181 / 4.10	.435 / .154 / 3.14
MLLR	.485 / .158 / 3.32	.432 / .141 / 3.29
score-level	.394 / .125 / 2.80	.297 / .108 / 2.39
iv-comb. 1	.475 / .163 / 3.78	.364 / .138 / 2.97
iv-comb. 2	.428 / .125 / 2.78	.315 / .111 / 2.40
iv-comb. 3	.415 / .122 / 2.67	.285 / .106 / 2.43

6.1. Cepstral system

The cepstral system used in this experiment is an ivector-based system whose configuration is 19 cepstral coefficients and energy, along with their first- and second- order derivatives. The system utilizes a 1024-component UBM model along with a total variability subspace of rank 400. The back end classifier is composed of LDA where the principal 200 dimensions were selected, followed by WCCN normalization. The final score is produced by taking the cosine distance of the compensated ivectors. Performance of this system is shown in Table 5.

6.2. Score-level combination

Score-level combination is performed using a logistic regression model as in [11]. The parameters for this model are estimated on the same test set, which will result in optimistic performance. Table 5 shows the results of such score-level combination. The improvement by combining the two ivector based systems is similar to that in the SRI NIST submission using a cepstral JFA and an MLLR-SVM system. The relative gain is in the range of 30% for both genders at the standard DCF.

6.3. ivector-level combination

Score-level combination can be tedious. It requires a development set, carefully chosen to match the target evaluation set. Also, while a few systems might be easy to combine, problems can occur when trying to combine a multitude of systems at once. Feature- and/or model- level combinations are attractive for these reasons. The ivector paradigm offers a natural way to combine systems at the ivector level.

We performed three different experiments to illustrate the ivector combination, with results presented as in Table 5. The first one, *iv-comb 1*, consists of concatenating the two i-vectors, after dimensionality reduction through LDA and WCCN, namely, of dimension 100 and 200 for the MLLR and cepstral, respectively, and taking the cosine distance between the two. The second, *iv-comb 2*, replicates this process but adds a WCCN step. The third, *iv-comb 3* uses a similar process but insert a LDA step before WCCN where only 250 dimensions were selected. For both final stages of LDA and WCCN, the covariance matrices were estimated on the same background data.

From these results, we see that system combination in the ivector space can effectively be performed. The best-performing results are produced when using a final stage of LDA. Moreover, the ivector level combination is on par with the score-level combination, even though it was trained on the same data set resulting in an optimistic performance estimation.

7. Conclusion

The focus of this work was to find an alternative method for modeling MLLR transforms for speaker recognition. We leveraged the recent progress in the modeling of cepstral features in order to apply them to MLLR features. We show how one can easily apply the eigenchannel, the PLDA and ivector approach to these features to get very high accuracy. We also show our reference performance from NIST SRE 2010 and report a 15% relative improvement over this state-of-the-art system. Finally, we investigated the possibility of ivector-level system combination and showed that it can match a score-level combination that uses standard logistic regression. The relative improvement with a cepstral system is of around 30%.

This work aimed at showing that the current techniques used for cepstral features and GMM models can be easily transferred to MLLR features. The advantages of producing MLLR ivectors are numerous. Any improvement in the research in the ivector space can be leveraged for MLLR features. Moreover, a unified back end will simplify the speaker recognition pipeline as well as open many possibilities for feature-level system combination. Future enhancements of this work will consist of assessing these approaches on diverse audio and mismatch conditions as well as exploring the use of more sophisticated back ends for system combination such as Gaussian and heavy-tail PLDA.

8. References

[1] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, " A study of inter-speaker variability in speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, July 2008.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, , no. 99, pp. 1–1, 2010.

[3] L. Burget, O. Plchot, S. Cumani, Glembek O., P. Matejka, and N. Brummer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *Proceedings of ICASSP 2011, Prague, CZ, May 2011*, 2010.

[4] A. Stolcke, S. Kajarekar, L. Ferrer, and E. Shriberg, "Speaker recognition with session variability normalization based on mllr adaptation transforms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1987–1998, 2007.

[5] N. Scheffer, L. Ferrer, M. Gracianera, S. Kajarekar, E. Shriberg, and A. Stoclke, "The SRI NIST 2010 Speaker Recognition evaluation system," *ICASSP 2011, Prague, CZ.*, 2011.

[6] R. Vogt, B. Baker, and S. Sridharan, "Modelling session variability in text-independent speaker verification," in *Proceedings of Eurospeech*. ISCA, 2005.

[7] D. Matrouf, N. Scheffer, B. Fauve, and J.F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," *Interspeech 2007, Brighton, UK*, 2007.

[8] M.E. Tipping and C.M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[9] N. Brümmer and E. de Villiers, "The speaker partitioning problem," in *Proceedings of the Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic*, 2010.

[10] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic*, 2010.

[11] L. Ferrer, M. Graciarena, A. Zymnis, and E. Shriberg, "System combination using auxiliary information for speaker verification," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4853–4856.