USING CONTEXTUAL INFORMATION IN JOINT FACTOR EIGENSPACE MLLR FOR SPEECH RECOGNITION IN DIVERSE SCENARIOS

Oscar Saz, Thomas Hain

Speech and Hearing Research Group, The University of Sheffield, Sheffield, UK

ABSTRACT

This paper presents a new approach for rapid adaptation in the presence of highly diverse scenarios that takes advantage of information describing the input signals. We introduce a new method for joint factorisation of the background and the speaker in an eigenspace MLLR framework: Joint Factor Eigenspace MLLR (JFEMLLR). We further propose to use contextual information describing the speaker and background, such as tags or more complex metadata, to provide an immediate estimation of the best MLLR transformation for the utterance. This provides instant adaptation, since it does not require any transcription from a previous decoding stage. Evaluation in a highly diverse Automatic Speech Recognition (ASR) task, a modified version of WSJCAM0, yields an improvement of 26.9% over the baseline, which is an extra 1.2% reduction over two–pass MLLR adaptation.

Index Terms— Speech recognition, adaptation, eigenspace MLLR, joint factorisation, metadata

1. INTRODUCTION

One of the main challenges in Automatic Speech Recognition (ASR) is to achieve robustness in highly diverse scenarios. Speaker adaptation techniques, such as Maximum Likelihood Linear Regression (MLLR) [1], are used to deal with the variability introduced by multiple speakers. Equally, noise compensation techniques, such as SPLICE [2, 3], are used in degraded acoustic conditions. However, the presence of diverse scenarios, with unbalanced or unreliable data, also degrades the performance of adaptation and normalisation techniques.

Acoustic model factorisation techniques in ASR [4, 5, 6, 7], which separate the many factors in an audio signal, are gaining relevance in dealing with diverse acoustic conditions, as they did before in speaker identification tasks [8, 9]. Early techniques for providing factorisation in ASR were Cluster Adaptive Training (CAT) [10] and eigenspace MLLR [11, 12]. The latter was proposed to represent the speaker variability as a set of eigenvoices extracted via Principal Component Analysis (PCA) [13] from a set of training speakers. For a new speaker, an MLLR transformation is estimated as a linear combination of these eigenvoices. This technique was also shown to deal with the variablity in the background [14, 15].

Recently, there is a growing trend for augmenting audio–visual data with contextual information and metadata describing its content. This metadata already exists for some ASR tasks where the diversity of scenarios is of more concern, like meetings or media [16] and can provide useful information regarding speakers and acoustic conditions. So far, the only metadata commonly used are speaker labels in order to perform speaker adaptation. But the information available can be much richer than that and it is, mostly, unexploited.

Our goal is to make all this existing metadata useful for ASR tasks. Since there are many factors that can be described by this metatada, the use of acoustic factorisation will be required. For that, we propose a new joint factor approach in eigenspace MLLR and a novel methodology to apply adaptation based only in context information, such as tags. This does not require any prior processing of the signal, since only tag information is used, resulting in instant adaptation.

This paper is organised as follows: Section 2 reviews eigenspace MLLR adaptation. Section 3 presents our proposal for joint factorisation. In Section 4 we derive how to use contextual information to estimate the eigenspace coefficients. Then, in Section 5 we present our experimental setup with WSJCAM0; followed by Sections 6 and 7 where the results and conclusions to this work are presented.

2. EIGENSPACE MLLR

MLLR is an adaptation technique that learns a transformation (W), consisting of transformation matrix A and bias vector b, on the means of the Gaussian distributions (μ) of a speaker independent Hidden Markov Model (HMM), as Equation 1.

$$\hat{\mu} = A\mu + b = \begin{bmatrix} b & A \end{bmatrix} \begin{bmatrix} 1 \\ \mu \end{bmatrix} = W\mu \tag{1}$$

Eigenspace MLLR was proposed when little adaptation data exists [11, 12]. In it, eigenvoices (W_m) are trained performing PCA on a set of MLLR transformations from a large number of training speakers (a hundred or more) [11]. For a new speaker, an MLLR transformation ($\hat{W}$) is estimated as a linear combination of the $M+1$ most relevant eigenvoices, as

This work was supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

shown in Equation 2. The first eigenvoice (W_0) is the mean of the space, and its coefficient (α_0) is forced to be 1 [14].

$$\hat{W} = \sum_{m=0}^{M} \alpha_m \cdot W_m = W_0 + \sum_{m=1}^{M} \alpha_m \cdot W_m \qquad (2)$$

The coefficients α_m are obtained via Maximum Likelihood Eigen Decomposition (MLED) [11, 14], in a way similar to [10], maximising the likelihood of the data given the model, although discriminative approaches exist [17]. Using a small number of eigenvoices ($M \approx 50$) [11], just a few seconds of speech are required. Other works in eigenspace MLLR [14] have proposed using also background–specific MLLR transformations when calculating the eigenvoices. In this case, background variability is also modelled, leading to improvements in noisy background conditions.

3. JOINT FACTOR EIGENSPACE MLLR (JFEMLLR)

Eigenspace MLLR with speaker and background transforms [14] defines a full space where some eigenspace MLLR transformations represent background characteristics while other eigenspace MLLR transformations represent speaker characteristics. Follow–up work [15] proposed separating the background and speaker influence by learning two different sets of eigenspaces, with a small improvement in performance.

We propose to perform joint factorisation of the background and speaker spaces in a similar fashion to Joint Factor Analysis (JFA) [8]. The formulation of this proposal is shown in Equation 3, where the final MLLR transformation estimated for an input utterance is the combination of three elements: The mean of the space (W_0); N background eigenspace MLLR transformations (W_n^{bgd}); and P speaker eigenspace MLLR transformations (W_p^{spk})

$$\hat{W} = W_0 + \sum_{n=1}^{N} \beta_n \cdot W_n^{bgd} + \sum_{p=1}^{P} \gamma_p \cdot W_p^{spk} \qquad (3)$$

To achieve joint factorisation, the background and speaker eigenspaces are trained in separated stages. Initially, PCA is used on the training data to create the full variability eigenspace (W_m) and a background eigenspace (W_n^{bgd}). These eigenspaces are applied to the training data to define the MLLR transformation for each utterance in both the full variability and background spaces. Substracting the background–space transformation from the full variability transformation obtains the speaker residual for each utterance. Finally, performing PCA on these residuals will provide the set of speaker–specific eigenbases (W_p^{spk}).

For each test utterance, the combination coefficients for each set of bases (β_n for background and γ_p for speaker) are estimated separately using the MLED algorithm and the final transformation is calculated as in Equation 3.

4. USING CONTEXTUAL INFORMATION IN EIGENSPACE MLLR

The main proposal of this work is the use of contextual information when estimating the vector of combination coefficients $\phi = [\alpha_0, ..., \alpha_M]^T$ in eigenspace MLLR and, by extension, in JFEMLLR (β_n and γ_p instead of α_m). At this point, contextual information will be defined as a set of discrete tags used to describe properties of a speech signal. A tag cloud of T tags ($Tags = \{Tag^1, Tag^2, ..., Tag^T\}$) can be set for any input signal, describing different characteristics of the speaker (gender, age, ...) or the background (channel, noise, ...).

By calculating the coefficients in the eigenspace MLLR framework for all the training utterances using the MLED algorithm, it is possible to estimate a probability distribution of these coefficients for any given tag Tag^t that appears in the training data. Using a Gaussian Mixture Model (GMM) of G Gaussians, the likelihood of the coefficients ϕ given tag Tag^t is calculated as in Equation 4.

$$P(\phi|Tag^t) = \sum_{g=1}^{G} c_g^{Tag^t} \cdot N(\phi; \mu_g^{Tag^t}, \Sigma_g^{Tag^t}) \qquad (4)$$

For a signal with tags $Tags$ the coefficients that maximise the posterior probability of those tags are estimated in Equation 5.

$$\hat{\phi} = \arg\max_{\phi} P(Tags|\phi) \qquad (5)$$

Assuming the tags to be independent and the a priori distribution of tags and coefficients to be uniform, it is possible to maximise the log–likelihood following Equations 6 and 7.

$$Q(\phi) = \sum_{t=1}^{T} \sum_{g=1}^{G} log\Big(c_g^{Tag^t} \cdot N(\phi; \mu_g^{Tag^t}, \Sigma_g^{Tag^t})\Big) \qquad (6)$$

$$Q(\phi) = \frac{1}{2} \sum_{t=1}^{T} \sum_{g=1}^{G} \Big[(\phi - \mu_g^{Tag^t})^T (\Sigma_g^{Tag^t})^{-1} (\phi - \mu_g^{Tag^t})\Big] \qquad (7)$$

$Q(\phi)$ is maximised by equalling the derivative in Equation 8 to zero, leading to Equation 9 where ϕ is calculated.

$$\frac{\partial Q(\phi)}{\partial \phi} = \sum_{t=1}^{T} \sum_{g=1}^{G} \Big[(\Sigma_g^{Tag^t})^{-1} \phi - (\Sigma_g^{Tag^t})^{-1} \mu_g^{Tag^t}\Big] \qquad (8)$$

$$\sum_{t=1}^{T} \sum_{g=1}^{G} \Big[(\Sigma_g^{Tag^t})^{-1} \mu_g^{Tag^t}\Big] = \Big[\sum_{t=1}^{T} \sum_{g=1}^{G} (\Sigma_g^{Tag^t})^{-1}\Big] \cdot \phi \qquad (9)$$

In Equation 9, if there is only one tag modelled with a single Gaussian, the coefficients will result to be the mean of such Gaussian, since this is the only information available to estimate them. If more tags are used, the resulting coefficients will be an interpolation among the means of their models.

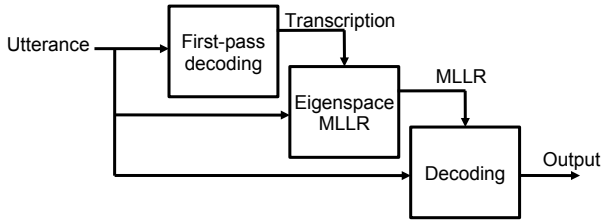

Fig. 1. *Eigenspace MLLR framework*

The eigenspace MLLR framework in Figure 1 is replaced by Figure 2. The first decoding to obtain a transcription for the MLLR estimation is not required when tags are used. Now, the MLLR transform can be calculated instantly.

5. EXPERIMENTAL SETUP

The proposed algorithms were evaluated on a modified version of the WSJCAM0 corpus. WSJCAM0 is a re–recording of the WSJ corpus by British speakers [18]. 7,387 utterances from 86 speakers were used for model training and 4 sets totalling 1,315 utterances from 18 speakers were used for evaluation. Two test sets are 5,000–word closed vocabulary tasks with a bigram language model, the other two sets are 20,000–word open vocabulary tasks with a trigram language model.

Modified train and test sets were generated introducing highly diverse background conditions. These sets will be further referred to as *Diverse* sets, opposed to the original *Clean* sets. The diverse backgrounds are defined by three variables:

- Channel: Close–talk microphone (50%) or table–top microphone (50%).
- Background: Clean background (33%) or music background (33%), divided equally in orchestral and popular contemporary, or noisy background (33%), divided equally in traffic, outdoors, cocktail party and applause.
- Signal–to–Noise Ratio (SNR): Uniform distribution from 5 to 15dB if music or noise are present.

Metadata is available in the form of tags. Each signal has 3 tags describing the speaker (gender, age and accent) as defined in the WSJCAM0 corpus and another 3 describing the background (channel, noise type and SNR) as explained previously. The tags were augmented with combinations of them (for instance, the combination of gender and age).

The ASR system used was a Hidden Markov Model Toolkit (HTK) [19] setup. Crossword triphone models were trained using Maximum Likelihood (ML), with 16 Gaussian mixtures per state. 39-dimension feature vectors were used with 13 Perceptual Linear Predictive (PLP) features [20] and their first and second derivatives. Cepstral Mean Normalization (CMN) was applied to the static features.

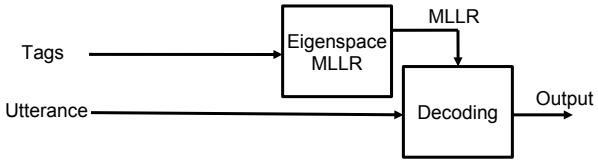

Fig. 2. *Eigenspace MLLR with contextual information*

5.1. Baseline results

The baseline Word Error Rates (WER) for speaker independent ML models are shown on Table 1. The results on *Clean* data show an average WER of 9.5%. Using the *Clean* models on the *Diverse* sets, the WER rose significantly to 33.1%. Training models on the *Diverse* training data the results on the *Diverse* test sets saw their WER reduced to 20.3%.

Table 1. *Baseline WER(%) on the modified WSJCAM0 data.*

Train	Test	5K set	20K set	Total
Clean	Clean	5.8	13.0	9.5
Clean	Diverse	27.0	39.1	33.1
Diverse	Diverse	14.9	25.5	20.3

6. RESULTS

The first adaptation methods evaluated in the decoding of the *Diverse* test sets with models trained on *Clean* data were MLLR and eigenspace MLLR. All transformations used were full matrices with 5 regression classes. The results in Table 2 show that speaker adaptation yielded a 25.7% relative improvement over the baseline; since none of the test speakers appeared in the training set, speaker adaptation was performed unsupervised after a first decoding. Adaptation to the background, by training transformations for each combination of channel, background and SNR in a supervised fashion from the *Diverse* training data, provided 26.9% improvement. Finally, an eigenspace of 30 bases was created from speaker and background transformations trained from the *Diverse* training data. The relative improvement with eigenspace MLLR was 31.1%, significantly higher than using MLLR adaptation. Eigenspace MLLR was able to provide solid improvement due to its ability to factorise the space.

Table 2. *WER(%) with MLLR and eigenMLLR adaptation.*

Adaptation	5K set	20K set	Total
Unsupervised speaker MLLR	19.3	29.8	24.6
Supervised background MLLR	18.5	29.9	24.2
Eigenspace MLLR	17.0	28.5	22.8

The JFEMLLR approach was evaluated afterwards. The

results in Table 3 show how it achieved equal performance to the full variability eigenspace MLLR (30.5% improvement). Within JFEMLLR, it was possible to perform adaptation using each of the subspaces separately. The factorised background eigenspace MLLR yielded 24.1% WER and the factorised speaker eigenspace MLLR gave 25.0% WER. In this case, 15 bases were used for each subspace, so the jointly factorised space was modelled with 30 bases.

Table 3. *WER(%) with JFEMLLR adaptation.*

Eigenspace	5K set	20K set	Total
Background	18.1	29.9	24.1
Speaker	18.8	31.1	25.0
Background & Speaker	17.3	28.6	23.0

Finally, the use of the tags assigned to each utterance was introduced in the estimation of eigenspace MLLR and JFEMLLR transformations. The distribution of the combination coefficients for each tag was modelled as a single Gaussian. Table 4 shows the results for these experiments, where all the cases showed solid improvements over the baseline. Modelling speaker tags alone provided less improvement than modelling background tags alone (21.5% to 27.5%, respectively). In this case, the background tags were more informative and more useful for adaptation. Combining both sources of tags degraded the performance in eigenspace MLLR with respect to using only background tags, while it provided a further 1.2% improvement in JFEMLLR. This indicated that having both spaces jointly factorised was better when it came to model the information provided by contextual information.

Table 4. *WER(%) with tags in eigenMLLR and JFEMLLR.*

Space	Tags used	5K set	20K set	Total
EigenMLLR	Background	18.6	30.5	24.6
	Speaker	20.0	32.3	26.3
	Backg. + Spk.	18.9	30.8	24.9
JFEMLLR	Background	18.6	30.4	24.6
	Speaker	19.8	32.0	26.0
	Backg. + Spk.	18.2	30.1	24.2

6.1. Influence of the amount of training data

One of the main advantages of this proposed method is that it does not require any prior processing or first pass of decoding of the input utterance. In this way, it can be used to provide instant adaptation in a single pass of decoding. However, it requires training data from the diverse scenarios to extract the eigenbases and train the a priori models of the tags. In most cases, this type of diverse data is only available in small amounts, which could become a drawback. An evaluation on the robustness of this approach to the amount of *Diverse* training data available was performed. For that, an approximate one fifth of the original *Diverse* training data (1,515 utterances) was randomly extracted. A comparison was made among training ML models, supervised background MLLR transformations and JFEMLLR with tag models, using this reduced training set. The results are presented in Table 5. ML models and supervised MLLR transformations suffered a significant degradation with limited data, while JFEMLLR using tag models gave the same performance as with the full training set. This indicated that large amounts of training data are not strictly required to provide adaptation in this framework.

Table 5. *WER(%) with limited* Diverse *training data.*

Condition	5K set	20K set	Total
Diverse ML models	19.7	32.8	26.3
Supervised background MLLR	19.3	31.3	25.4
JFEMLLR with tag models	18.1	30.2	24.2

7. CONCLUSIONS

With the results seen we can conclude that using contextual information in a joint factorisation scheme is able to provide instant adaptation to diverse backgrounds and speakers. This adaptation only requires the processing of the utterance's tags to provide an MLLR transformation that can be used directly in decoding. The joint factorisation scheme proposed (JFEMLLR) has been shown to be useful to deal with background and speaker metadata separately.

Although our approach requires training data in similar conditions to those existing in the evaluation sets, it was shown to be robust to data sparsity compared to other approaches that also require training data like ML training or supervised MLLR. Furthermore, not all the possible conditions are required to exist in the training set, since the model can even be used with very broad categories, such as speaker gender or a rough description of the type of background noise.

In future work, we aim to investigate how to deal not only with discrete tags but with naturally occurring metadata. This metadata could be, for instance, agenda briefs in meeting recordings or synopsis in media programmes. This requires processing of unstructured data into a set of categories that can describe either speaker or background in a way that can be handled by the proposed models.

As metadata and other sources of information keep growing and becoming more usual in audio–visual data sets, techniques like the one proposed can help improve the performance of ASR systems. This includes transcription tools for tasks where there is a high diversity of speakers and scenarios, but where metadata is usually available, and that could benefit from it.

8. REFERENCES

[1] M.J.F. Gales and P.C. Woodland, "Mean and Variance Adaptation within the MLLR Framework," *Computer, Speech and Language*, vol. 10(4), pp. 249–264, October 1996.

[2] J. Droppo, L. Deng and A. Acero, "Evaluation of the SPLICE Algorihtm on the AURORA2 Database," in *Proceedings of 7th European Conference on Speech Communication and Technology*, pp. 217–220, Aalborg, Denmark.

[3] L. Buera, E. Lleida, A. Miguel, A. Ortega and O. Saz, "Cepstral Vector Normalization based on Stereo Data for Robust Speech Recognition," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15(3), pp. 1098–1113, 2007.

[4] M.J.F. Gales, "Acoustic factorisation," in *Proceedings of the 2001 Automatic Speech Recognition and Understanding Workshop*. Madonna di Campiglio, Italy.

[5] M.L. Seltzer and A. Acero, "Separating Speaker and Environmental Variability Using Factored Transforms," in *Proceedings of 12th Annual Conference of the International Speech Communication Association*, pp. 1097–1100, Florence, Italy.

[6] Y. Wang and M. J. F. Gales, "Speaker and noise factorisation for robust speech recognition," in *IEEE transactions on audio, speech and language processing*, vol. 20(7), pp. 2149–2158, 2012.

[7] O. Saz and T. Hain, "Asynchronous factorisation of speaker and background with feature transforms in Speech Recognition," in *Proceedings of 14th Annual Conference of the International Speech Communication Association*, pp. 1238–1242, Lyon, France.

[8] S.C. Yin, R.C. Rose and P. Kenny, "A Joint Factor Analysis Approach to Progressive Model Adaptation in Text-Independent Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15(7), pp. 1999–2010, 2007.

[9] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-end factors analysis for speaker verification", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19(4), pp. 788–798, 2011.

[10] M.J.F. Gales, "Transformation smoothing for speaker and environmental adaptation," in *Proceedings of the 5th European Conference on Speech Communication and Technology*. Rhodes, Greece, 1998.

[11] R. Kuhn, J.C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K, Field and M. Contolini, "Eigenvoices for speaker adaptation," in *Proceedings of the 5th International Conference on Spoken Language Processing*. Sydney, Australia, 1998.

[12] K. Chen, W. Liau, H. Wang and L. Lee, "Fast speaker adaptation using eigenspace-based Maximum Likelihood Linear Regression," in *Proceedings of the 6th International Conference on Spoken Language Processing*. Beijing, China, 2000, pp. 742–745.

[13] I. T. Jolliffe, *"Principal Component Analysis,"* Springer-Verlag, 1986.

[14] Y. Liao, H. Fang and C. Hsu, "Eigen-MLLR environment/speaker compensation for robust speech recognition," in *9th Annual Conference of the International Speech Communication Association*. Brisbane, Australia, 2008, pp. 1249–1252.

[15] Y. Liao, H. Fang and C. Yang, "Reference eigenenvironment and speaker weighting for robust speech recognition," in *Proceedings of 6th International Symposium on Chinese Spoken Language Processing*. Kunming, China, 2008, pp. 1–4.

[16] P. Lanchantin, P.J. Bell, M.J.F. Gales, T. Hain, X. Liu, Y. Long, J. Quinnell, S. Renals, O. Saz, M.S. Seigel, P., Swietojanski and P.C. Woodland, "Automatic Transcription of Multi-genre Media Archives," in *Proceedings of the First Workshop on Speech, Language and Audio in Multimedia*. Marseille, France, 2013.

[17] Y. Miao, F. Metze and A. Waibel, "Learning discriminative basis coefficients for eigenspace MLLR unsupervised adaptation," in *Proceedings of the 2013 International Conference on Acoustics, Speech and Signal Processing*. Vancouver, Canada, 2013, pp. 7927–7931.

[18] T. Robinson, J. Fransen, D. Pye, J. Foote and S. Renals, "WSJCAM0: A British English Speech Corpus for Large Vocabulary Contiuous Speech Recognition," in *Proceedings of the 1995 International Conference on Acoustics, Speech and Signal Processing*. Detroit MI, USA, pp. 81–84.

[19] S.J. Young, G. Evermann, M.J.F. Gales, T. Hain, D. Kershaw, G. Moore, J.J. Odell, D.G. Ollason, D. Povey, V. Valtchev and P.C. Woodland, *"The HTK Book version 3.4,"* Cambridge University Engineering Department, Cambridge, UK, 2006.

[20] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustic Society of America*, vol. 87(4), pp. 1738–1752, 1990.