# Augmentation of adaptation data

*Ravichander Vipperla, Steve Renals, Joe Frankel*

The Centre for Speech Technology Research, School of Informatics, University of Edinburgh

r.c.vipperla@sms.ed.ac.uk, s.renals@ed.ac.uk, joe@cstr.ed.ac.uk

## Abstract

Linear regression based speaker adaptation approaches can improve Automatic Speech Recognition (ASR) accuracy significantly for a target speaker. However, when the available adaptation data is limited to a few seconds, the accuracy of the speaker adapted models is often worse compared with speaker independent models. In this paper, we propose an approach to select a set of reference speakers acoustically close to the target speaker whose data can be used to augment the adaptation data. To determine the acoustic similarity of two speakers, we propose a distance metric based on transforming sample points in the acoustic space with the regression matrices of the two speakers. We show the validity of this approach through a speaker identification task. ASR results on SCOTUS and AMI corpora with limited adaptation data of 10 to 15 seconds augmented by data from selected reference speakers show a significant improvement in Word Error Rate over speaker independent and speaker adapted models.

**Index Terms**: ASR, MLLR, MAPLR, Speaker selection

## 1. Introduction

In typical Automatic Speech Recognition (ASR) based interactive voice response and spoken dialogue systems, only a few seconds of speech is generally available from a user to adapt the acoustic models to his/her voice. Linear regression based speaker adaptation techniques such as Maximum Likelihood Linear Regression (MLLR) [1] and Maximum A Posteriori Linear Regression (MAPLR) [2] are widely used in such scenarios. Transformation matrices for regression can be efficiently computed with a reasonable amount of data. However when the transforms are computed using a very small amount of adaptation data, the improvement in recognition accuracy using the adapted models can be low; indeed the accuracy with the adapted models can be lower than that with the speaker independent (SI) models.

To overcome this problem of sparse data, several approaches have been devised to characterise the test speaker and make better use of the data from the existing speakers. Eigenvoices [3] is one such idea where the test speaker is characterised as a linear combination of eigenvectors which are computed from speaker dependent (SD) models of the training set speakers. This approach however has limitations when applied to large vocabulary systems due to the need for generating several SD models and in the computation of speaker coefficients in the high dimensional eigenspace.

Another approach to tackle data sparsity is to augment the adaptation data for the target speaker with speech data from other reference speakers acoustically close to the target speaker. The reference speakers can be a subset of the speakers used to train the SI models as well as other speakers whose data becomes available at a later stage. Such systems where more corpora becomes available for speaker selection can be easily envisaged in practical applications. In telephony based IVR systems, speech data can be collected as the system is used and the collected data can be made available as a pool of reference speakers. In broadcast news, speech content is made available on daily basis from different speakers. Hence it makes sense in such scenarios, to build a speaker independent ASR system and use the data made available consequently, to improve the performance of the system.

Some related work based on this approach of speaker selection include [4], where Gaussian Mixture Models (GMMs) were trained for each reference speaker and the models that maximised the likelihood for the target speaker's adaptation data were chosen as the closest speakers. In [5], custom Hidden Markov Models (HMMs) were built for each reference speaker using MLLR and the speakers whose models maximised the likelihood scores for forced alignment of the adaptation data were chosen as the reference speakers.

Recently, an approach to speaker recognition using MLLR transforms as feature vectors has been proposed [6, 7]. The core idea is to concatenate the coefficients of the adaptation transforms into high dimensional vectors and use these vectors for speaker identification using Support Vector Machine (SVM) classifiers. Inspired by this work, we extend the idea of using transformation matrices as speaker features to identify the reference speakers acoustically closest to the target speaker. However, we do not use SVM classifiers since our task is different from speaker recognition. We use a distance metric based on transformations to compute the distance between speakers. We show, with experimental results on two different corpora, that when the adaptation data available is limited, ASR accuracies can be improved by augmenting the target speakers' adaptation data with data from acoustically close speakers.

In section 2, we describe the distance metric used and some results on a speaker recognition task as a sanity check for the distance metric. The speaker selection approach is described in section 3. ASR experimental setup and the results using augmented adaptation data is described in section 4, followed by discussion and conclusion in sections 5 and 6 respectively.

## 2. Speaker distance measure

In [6], the coefficients of MLLR matrices are concatenated to create high dimensional vectors and these vectors are used as speaker features. Such high dimensional vectors have been shown to have good discrimination properties for classification but the disadvantage of this approach is that it just treats the matrix as a vector and discards the property of the matrices that enable it to transform the means of HMMs to match the target speaker.

In this paper, we propose a distance metric that takes advantage of the transformation defined by the linear regression

matrices. Given transformation matrices from two speakers, sample points in the acoustic space are transformed by the two matrices and the distance between the transformed points is calculated. We cluster the means of all the Gaussian components in the SI model and choose the centroids of the partitions as the sample points. This ensures a good coverage over the acoustic space.

The distance $d_{TS}$ between two speakers whose MLLR transforms are represented by $A_T$ and $A_S$ is given by:

$$d_{TS} = \sum_{k=1}^{K} \frac{|(A_T - A_S)c_k|}{|c_k|} \quad (1)$$

where $c_k$ is the $k^{th}$ mean in $K$ clusters computed from the means of all Gaussian components in the SI Model.

### 2.1. Speaker identification task

In order to understand how well the proposed distance metric helps in identifying the closest transformation matrices, it was applied to a speaker identification task.

The experimental setup consisted of reference MLLR transforms and test MLLR transforms for a set of speakers. For each speaker, the utterances used in computing the reference and test transforms were disjoint. The task is to identify the closest reference transform for each test transform using the distance metric proposed and when the closest reference transform is from the same test speaker, it is treated as correct recognition.

The SCOTUS [1] corpus [8] was used for this task. The corpus was parametrised using Perceptual Linear Prediction (PLP) Cepstral features. A window size of 25ms and frame shift of 10ms were used for feature extraction. Energy along with $1^{st}$ and $2^{nd}$ order derivatives were appended giving a 39-dimensional feature vector. Speaker Independent acoustic models were trained on 90 hours of speech data. The acoustic models were trained using HTK [9] as tied-state context dependent triphone HMMs with 18 Gaussian components per state for all the speech models and 36 components per state for the silence model. In all, the SI acoustic models comprised 59886 Gaussians over 3324 independent states.

A set of 100 speakers was used for the speaker identification task. To compute the reference and test MLLR transforms, about 40 seconds and 12 seconds of speech was used respectively for each speaker. A two class regression tree for speech and silence was used for the MLLR computation and only the speech transforms were used in distance computation.

The sample points in acoustic space to be used in calculating the distance, were selected as the centroids from k-means clustering on all the Gaussian means in the SI model. The task was repeated with various sizes of sample points viz., global mean, 100 clusters, 1000 clusters and using all the Gaussian means in the SI model. A simple euclidean distance between the co-efficients of the matrices was also used for comparison.

The results for this task are shown in Table 1. We observe that a set of 1000 points in the acoustic space is sufficient to achieve acceptable accuracy.

As mentioned earlier, the objective of this task is only to make a sanity check of the distance metric and hence the results have not been compared with other competing methods for speaker recognition.

[1] The Supreme Court Of The United States corpus. http://www.oyez.org

Table 1: *Speaker Identification Task*

| Distance Measure | Accuracy |
|---|---|
| Euclidean Distance | 32% |
| With Global Mean | 36% |
| With 100 Cluster means | 97% |
| With 1000 Cluster Means | 98% |
| With all the Means | 98% |

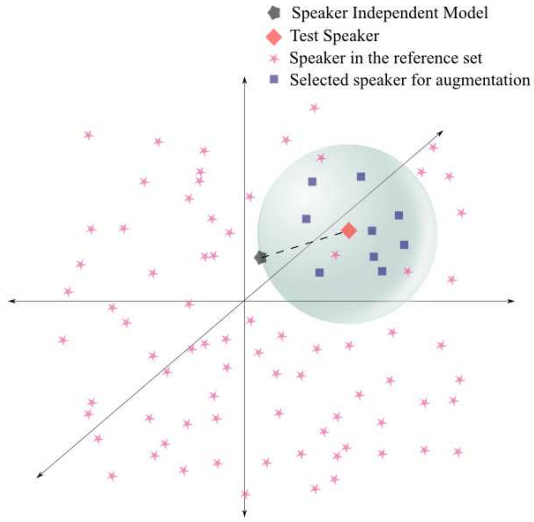


Figure 1: *Speaker Selection.*

## 3. Speaker selection for augmentation

Given a set of $N$ reference speakers, our task is to select a subset of these speakers who are acoustically closest to the target speaker $T$.

Denoting the transformation matrices for the target speaker as $A_T$, the $i^{th}$ reference speaker as $A_{R_i}$ $(i = 1..N)$ and using an identity matrix to represent the SI model $A_I = [I_{mXm} : 0_{mX1}]_{mX(m+1)}$,

1. Compute a linear transform $A_T$ for the test speaker from the available adaptation data.
2. Compute the distances $d_{TR_i}$ for $i = 1..N$ and $d_{TI}$.
3. Choose a subset of speakers satisfying $d_{TR_i} \leq d_{TI}$ to augment the adaptation data.
4. Recompute the linear transform for the test speaker using the augmented data.

To illustrate, the speaker selection process, Figure.1 shows the speakers in 3d space (generated using Multidimensional scaling). The reference speakers selected for augmentation are the ones that lie within the spherical manifold with the target speaker at the center with a radius of $d_{TI}$. In practice, the dimension of the speaker space is large and the selection manifold is a high dimensional ellipsoid.

## 4. Experiments

The experiments were carried out on two different corpora, viz., SCOTUS and AMI[2] [10].

[2] Augmented Multiparty Interaction. http://corpus.amiproject.org/

### 4.1. Corpora

#### *4.1.1. SCOTUS Corpus*

The SI acoustic models used are the same as those described in section 2.1. Back-off bigram language models and the vocabulary were constructed from the transcripts of the Supreme Court of the United States proceedings resulting in 23445 words types.

The reference speaker set consists of speech data from 267 training set speakers and 282 additional speakers not used in the training set. Each reference speaker had about 8 to 20 minutes of available data with an average of 12 minutes per speaker.

The test speaker set comprised 39 speakers disjoint from the training and additional speaker set. Each test speaker had about 60 minutes of data and a small set of about 3 minutes kept aside as the adaptation data.

#### *4.1.2. AMI Corpus*

The waveforms were parametrised into 39 dimensional PLP based features similar to the SCOTUS corpus experimental setup. To build the acoustic models, tied state context dependent triphone HMMs were first trained on 73 hours of meetings data recorded by International Computer Science Institute (ICSI), 13 hours of meeting corpora from the National Institute of Standards and Technology (NIST) and 10 hours of speech corpora from the Interactive Systems Lab (ISL) [11]. These models were then adapted using the Maximum A Posteriori (MAP) approach [12] with 40 hours of speech from the AMI corpus. With 8 and 16 Gaussian components per state for speech and silence respectively, the SI model comprised 3712 independent states with 29720 Gaussians in total. Back-off bigram language models and vocabulary of size 50002 words were built using transcripts of several meeting corpora including Switchboard, Call Home, Fisher, ICSI, NIST, ISL and other web data resources [13].

The reference speaker set comprised of 69 speakers used to MAP adapt the SI models and 78 speakers not used in the training set. Each speaker had about 30 minutes of speech data on average. The test speaker set in this corpus consisted of 42 speakers with 200 utterances as test data per speaker and a small adaptation set separate from the test set.

### 4.2. Procedure

The means of all the Gaussians in the SI acoustic models were clustered into 1000 groups using k-means clustering for each of the two corpora. The centroids of each of these clusters were used as the sample points for computing the acoustic distance between speakers. From each of the reference speakers' data, MLLR and MAPLR mean transforms were computed using a two class regression tree, one for speech and one for non-speech. Three sets of adaptation data were used with different amounts of data for the test speakers viz., 1)10-15 seconds of speech per speaker 2) About 30 seconds per speaker and 3) About 1 minute per speaker. Adaptation transforms were computed from all the adaptation sets using the actual transcripts for supervised case and using the hypothesis from first pass decoding for unsupervised case. For each of the test speakers, acoustically closest speakers were chosen as described in section 3.

### 4.3. Results

The baseline results for the two corpora are shown in Tables 2 and 4. When only 10-15 secs of adaptation data is available, speaker adaptation is not optimal. The word error rates (WER) increase in most cases.

| Speaker Independent | 35.9 | | |
|---|---|---|---|
| ***Adaptation Data*** | ***15s*** | ***30s*** | ***60s*** |
| MLLR Supervised | 36.2 | 35.5 | 35.3 |
| MLLR Unsupervised | 36.5 | 35.8 | 35.5 |
| MAPLR Supervised | 35.6 | 35.2 | 34.9 |
| MAPLR Unsupervised | 36.0 | 35.5 | 35.3 |

Table 2: SCOTUS Corpus: Baseline results (WER %)

| ***Reference Spkrs*** | **Train set** | | | **Train + Add set** | | |
|---|---|---|---|---|---|---|
| ***Adaptation Data*** | ***15s*** | ***30s*** | ***60s*** | ***15s*** | ***30s*** | ***60s*** |
| MLLR Sup | 35.5 | 35.4 | 35.3 | 35.2 | 35.1 | 35.0 |
| MLLR Unsup | 35.5 | 35.4 | 35.4 | 35.2 | 35.2 | 35.1 |
| MAPLR Sup | 35.3 | 35.3 | 35.2 | 35.2 | 35.2 | 35.1 |
| MAPLR Unsup | 35.4 | 35.3 | 35.2 | 35.3 | 35.3 | 35.2 |

Table 3: SCOTUS Corpus: Results with augmented adaptation data (WER %). Notation: Sup - Supervised, Unsup - Unsupervised

| Speaker Independent | 46.3 | | |
|---|---|---|---|
| ***Adaptation Data*** | ***15s*** | ***30s*** | ***60s*** |
| MLLR Supervised | 50.2 | 46.0 | 43.9 |
| MLLR Unsupervised | 51.5 | 47.5 | 45.3 |
| MAPLR Supervised | 48.1 | 45.3 | 43.7 |
| MAPLR Unsupervised | 49.3 | 46.7 | 45.1 |

Table 4: AMI Corpus: Baseline results (WER %)

| ***Reference Spkrs*** | **Train set** | | | **Train + Add Set** | | |
|---|---|---|---|---|---|---|
| ***Adaptation Data*** | ***15s*** | ***30s*** | ***60s*** | ***15s*** | ***30s*** | ***60s*** |
| MLLR Sup | 46.2 | 45.6 | 45.5 | 45.9 | 45.6 | 45.2 |
| MLLR Unsup | 47.4 | 46.2 | 45.8 | 46.8 | 45.9 | 45.4 |
| MAPLR Sup | 46.1 | 45.7 | 45.4 | 45.8 | 45.7 | 45.3 |
| MAPLR Unsup | 47.1 | 46.1 | 45.7 | 46.2 | 45.9 | 45.7 |

Table 5: AMI Corpus: Results with augmented adaptation data (WER%). Notation: Sup - Supervised, Unsup - Unsupervised

Tables 3 and 5 show the results with augmented adaptation data. The tables capture WERs using 1) only the training set speakers as reference speakers and 2) Training set speakers and additional speakers (Train + Add Set). The results show a significant reduction in WER with augmented adaptation data when the adaptation data is limited to 10-15 seconds. The WER reduction is significant at $p < 0.001$ using Matched Pairs Sentence Segment Word Error (MAPSSWE) test. As the adaptation data from the target speaker increases, the benefit from using other speakers' speech reduces.

Augmenting the adaptation data is seen to be particularly advantageous in unsupervised case which is more often the situation in practical systems. It is also observed that accuracies with MAPLR mean adaptation are overall better than MLLR

mean adaptation. With augmented adaptation data, an improvement of 0.8% relative for supervised case and 1.7% relative for unsupervised case on SCOTUS corpus and improvements of 4.2% and 4.5% respectively for AMI corpus are observed.

Using speakers additional to the training set speakers, a further improvement in recognition accuracies can be achieved.

## 5. Discussion

In this paper a simple and efficient method to improve the ASR accuracies with small amounts of adaptation data is described. Other approaches on similar tasks such as eigenvoices have been shown to improve performance in smaller systems, but scaling the eigenvoices approach as described in [3] to our larger system led to Principal component analysis on large matrix (2.5million x 250), which was computationally expensive. Due to the high dimensionality, all the eigenvectors generated had close eigenvalues. Choosing the top 20 of them as basis results in loss of information and an increase in WER of 6.2%. A missing part in our experiments is the comparison of the results with Cluster Adaptive Training [14].

The distance metric proposed in this paper is efficient in memory usage and computational complexity. The storage requirements are one $m\text{x}(m+1)$ matrix per reference speaker and $K$ sample vectors in the acoustic space. The computation of the distance between two speakers only involves a few matrix operations. To speed up the computations $|c_k|$ terms could be precomputed and stored. To save on the time required for computing the regression transforms, sufficient statistics for the reference speakers can be computed offline.

Another feature of this approach is that there is no manual tuning or thresholding involved. If no reference speaker is close enough to the target speaker, only the original adaptation data is used. This approach is expected to work better with the availability of a larger and more varied reference speaker set. Furthermore, if the speech from a target speaker is available in the reference set, it is very likely to be selected first as augmentation data and improve the recognition accuracy significantly.

The MLLR/MAPLR WERs on AMI corpus with 15 seconds adaptation data are significantly higher as compared to the results with SI models. Despite this, the linear transform matrices still capture sufficient information about the speaker to be able to select augmentation data.

In both of our systems, the number of reference speakers were limited to a few hundred. If thousands of reference speakers are available in the selection pool, then computing the distance of the target speaker to all the speakers can be time consuming. A possible solution to this problem is

1. project all the reference speakers($N$) and SI model to a $p$ dimensional space ($p << N$) using MDS.
2. Select $p$ non-coplanar speakers in this $p$ dimensional space as reference points.
3. For a target speaker compute the distance from these $p$ reference speakers and project the target speaker into this reduced dimensionsal space using triangulation method.
4. Select the augmentation speakers satisfying $\tilde{d}_{TR_i} \leq \tilde{d}_{TI}$, where $\tilde{d}$ is the euclidean distance.

## 6. Conclusion

In this paper, we have proposed a simple approach to compute distance between speakers using regression matrices as speaker features. We have shown that speakers acoustically close to the target speaker can be effectively selected from a pool of reference speakers to augment the adaptation data for the target speaker. This approach works well when the adaptation data from the target speaker is very limited and gives significant reduction in WER. It is seen to particularly useful when the adaptation is unsupervised which is often the case in practical deployments of ASR systems. However when sufficient adaptation data is available from the target speaker, augmenting it with speech from other speakers is not beneficial.

## 7. Acknowledgment

This research was funded by the SFC SRDG grant no. HR04016 : MATCH. This work has made use of the resources provided by the Edinburgh Compute and Data Facility which is partially supported by the eDIKT initiative `http://www.edikt.org.uk`.

## 8. References

[1] C. Leggetter and P. Woodland, "Maximum Likelihood Linear Regression for speaker adaptation of continuous density Hidden Markov Models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[2] O. Siohan, T. A. Myrvoll, and C.-H. Lee, "Structural Maximum A Posteriori Linear Regression for fast HMM adaptation," *Computer Speech and Language*, vol. 16, no. 1, pp. 5 – 24, 2002.

[3] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695 – 707, 2000.

[4] S. Yoshizawa, A. Baba, K. Matsunami, Y. Mera, M. Yamada, A. Lee, and K. Shikano, "Unsupervised training of phoneme models using HMM sufficient statistics and a speaker distance function," *Electronics and Communications in Japan, Part 3 (Fundamental Electronic Science)*, vol. 88, no. 9, pp. 33 – 41, 2005.

[5] J. Wu and E. Chang, "Cohorts based custom models for rapid speaker and dialect adaptation," in *Eurospeech*, 2001, pp. 1261–1264.

[6] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR transforms as features in speaker recognition," in *Interspeech*, 2005, pp. 2425–2428.

[7] A. Stolcke, L. Ferrer, and S. Kajarekar, "Improvements in MLLR-transform-based speaker recognition," in *Odyssey*, 2006, pp. 1–6.

[8] R. Vipperla, S. Renals, and J. Frankel, "Longitudinal study of ASR performance on ageing voices," in *Proceedings of Interspeech*, 2008, pp. 2550–2553.

[9] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for Hidden Markov Model Toolkit Version 3.4)*, 2006.

[10] J. Carletta, "Unleashing the killer corpus: Experiences in creating the multi-everything AMI meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[11] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals, "The 2005 AMI system for the transcription of speech in meetings," in *Proceedings of the Rich Transcription 2005 Spring Meeting Recognition Evaluation*, 2005.

[12] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[13] T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, and S. Renals, "Transcription of conference room meetings: an investigation," in *Interspeech*, 2005.

[14] M. Gales, "Cluster adaptive training of hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 417 – 28, 2000.