SPEAKER CLUSTERING USING DIRECT MAXISATION OF THE MLLR-ADAPTED LIKELIHOOD

S.E. Johnson *P.C. Woodland*

Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK.
{sej28, pcw}@eng.cam.ac.uk

ABSTRACT

In this paper speaker clustering schemes are investigated in the context of improving unsupervised adaptation for broadcast news transcription. The various techniques are presented within a framework of top-down split-and-merge clustering. Since these schemes are to be used for MLLR-based adaptation, a natural evaluation metric for clustering is the increase in data likelihood from adaptation. Two types of cluster splitting criteria have been used. The first minimises a covariance-based distance measure and for the second we introduce a two-step E-M type procedure to form clusters which directly maximise the likelihood of the adapted data. It is shown that the direct maximisation technique produces a higher data likelihood and also gives a reduction in word error rate.

1. INTRODUCTION

For speech transcription problems with widely varying speaker, channel and background conditions, such as broadcast news transcription, it is beneficial to include unsupervised test-data adaptation to individual speakers and data types. The standard technique for such adaptation is Maximum Likelihood Linear Regression (MLLR) [4].

To be effective MLLR requires that similar data segments be grouped together so that MLLR transforms can be robustly estimated. This paper presents two strategies for clustering broadcast news data segments (found by an automatic segmentation algorithm) for subsequent MLLR adaptation. The first uses standard covariance techniques and the performance is evaluated using the increase in the MLLR-adapted log-likelihood when applying a single MLLR transform to each cluster. The second method maximises this performance criterion directly using a two step expectation-maximisation (E-M) procedure. The different clustering schemes are implemented in a common top-down split-and-merge clustering framework.

The overall clustering framework is first described and then details of the covariance based and direct maximisation techniques are given. The effectiveness of the schemes is examined using both the increase in adapted data likelihood and change in word error rate.

2. CLUSTERING FRAMEWORK

The clustering schemes work top-down with each node being split into up to four child nodes at each stage. This delays the implementation of a hard local decision and allows greater flexibility than binary-splitting. Each speech frame is represented by a vector of modified PLP-cepstral coefficients along with the first and second derivatives [6].

For each split an initial assignment of segments is made to the child nodes with segment order being retained to exploit any existing temporal correlation. The segments are then moved between child nodes to either decrease a covariance-based distance or maximise the adapted data likelihood. A minimum occupancy count is specified and any segments belonging to nodes which fall below the occupancy limit are re-assigned to the closest other child node. Once there is no change to the segment assignments and all child nodes satisfy the occupancy criterion the parent node is said to have been split successfully. Any parent node which cannot be split becomes a leaf node.

Recombination or merging can be applied at this stage. This involves combining child nodes which are similar. Once clusters have been merged they are considered to represent similar data and the combined child becomes a leaf node so it cannot be split further. When no further child nodes satisfy the recombination criterion the overall splitting/merging of the parent node has been completed.

This entire process of splitting (with or without subsequent merging) is performed until all the unsplit nodes are leaf nodes. The differences between the methods discussed in this paper lie in the choice of splitting and merging criteria.

Since the performance, measured by the increase in log-likelihood of the MLLR adapted data, is dependent on the number of clusters produced, (splitting a node will always result in an increase in likelihood) only schemes which produce a similar number of leaf nodes can be easily compared. Occupancy and recombination parameters can be chosen by specifying an operation point on a Receiver Operating Curve (ROC) of increase in likelihood against the number of clusters. To compare performance with varying numbers of clusters, word error rates can be found with an MLLR-adapted recognition system.

3. COVARIANCE METHODS

Several clustering systems have been built using the mean vector and/or covariance matrix to represent a segment of data [5, 6]. The system allows a full or diagonal covariance or correlation matrix to model each segment using the following choice of distance metrics:

Arithmetic Harmonic Sphericity (AHS) [1]

$$d(X,Y) = \log[tr(\Sigma_y \Sigma_x^{-1}) * tr(\Sigma_x \Sigma_y^{-1})] - 2\log(D)$$

Gaussian Divergence

$$\begin{aligned} d(X,Y) &= 0.5tr(\Sigma_x^{-1}\Sigma_y + \Sigma_y^{-1}\Sigma_x - 2I) \\ &+ 0.5(\mu_x - \mu_y)^T(\Sigma_x^{-1} + \Sigma_y^{-1})(\mu_x - \mu_y) \end{aligned}$$

where D represents the dimensionality of the data and μ_x and Σ_x represent the mean and covariance (or correlation) of the segment X respectively.

Segments which represent only a short period of data (e.g. <0.5s) may form singular or ill-conditioned covariance matrices. Since the inverse covariance is used in the standard distance calculations, the small segments are stored separately whilst the clustering procedure takes place. Once the leaf nodes have been determined, each small segment is assigned to the node with the closest mean in a Euclidean sense.

3.1. Splitting Procedure

Each active node is split into a maximum of four child nodes all satisfying a minimum occupancy criterion. The segments are assigned to four initial child nodes and then reassigned if necessary with the cluster statistics being recalculated, until no segments move or the maximum number of iterations is reached. Algorithmically:

```
foreach (parent) node to be split:
    Initialisation: Assign the segments into 4 child
                    nodes with approximately the same
                    number of segments in each child.
                    Maintain the order of the segments
                    to exploit any temporal correlation.
    Until no movement or max. iterations reached:
        foreach child node:
            Calculate the mean and covariance
            of the node.
        end
        foreach segment:
            Calculate the distance from the segment to
            each child node.
            Assign the segment to "closest" child node.
        end
        while any child node < min. occupancy limit:
            Disperse the segments in that node
            into the other children.
            Reduce the number of children by 1.
        end
    end
end
```

3.2. Merging Procedure

Initially the final number of clusters was determined by specifying the minimum occupancy allowed in each cluster and clustering was done until no further split could be made. However, this does not take into account how "similar" the child nodes are and fails to exploit the gains which could be made by keeping a large cluster whose segments all originate from the same speaker or acoustic condition. To overcome this deficiency, a recombination or merging procedure was introduced which allowed similar clusters to be merged at each stage of the overall splitting.

The intra-node cost of a node is defined as the average distance from the centre of the node to its segments. If the parent intra-node cost is sufficiently larger than the sum of the child intra-node costs then the split is allowed to go ahead, and no recombination takes place. If the absolute gain in splitting is not large enough, however, then child nodes containing single segments are dispersed to the closest other child node to compensate for the fact that their intra-node cost is zero. The inter-node costs between the child nodes are subsequently found and a test on the intra- and inter-node costs is then used to determine whether the nodes overlap sufficiently to justify recombination (see Figure 1).

This process is repeated until all the children no longer satisfy the recombination criteria. If recombination takes place the resulting node becomes a leaf node, whereas unaffected nodes remain active.

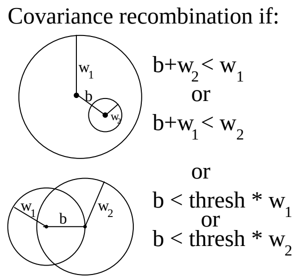

Figure 1: Recombination from Covariance Information

4. DIRECT MAXIMISATION METHOD

The increase in log-likelihood of the MLLR-adapted data is a natural way to evaluate the effectiveness of a clustering scheme which is to be used for MLLR-based adaptation. Clearly "optimal" clusters can be produced by maximising the likelihood directly. Furthermore there is some evidence that the MLLR transform matrix captures speaker-specific information from speech segments of over 5 seconds duration [3].

It should be noted that, unlike the previous schemes the MLLR based methods require an (approximate) transcription of the data to be available. However this is also required for subsequent adaptation and can be obtained from an initial decoding pass with an unadapted speech recogniser.

4.1. Splitting Procedure

The splitting process for direct maximisation of the MLLR-adapted likelihood uses a two-step EM algorithm which maximises the likelihood of the adapted data in the child nodes given the number of child nodes, for each split.

Algorithmically this is represented as:

```
foreach (parent) node to be split:
    Initialisation: Assign the segments in the parent
                    node to the child nodes.
                    This can either be time-ordered,
                    or based on covariance methods.
    Until converged or max. iterations reached:
      Step 1:       Calculate a single MLLR transform.
                    for each child node. This ensures
                    a monotonic increase in likelihood
                    when the transform is applied.
      Step 2:       Calculate the likelihood of the
                    data in each segment using the
                    transform of each node in turn.
                    Assign each segment to the node
                    that gives it the greatest
                    likelihood. This also guarantees a
                    monotonic increase in likelihood.
    end
end
```

4.2. Merging Procedure

The merging procedure for MLLR clustering is:

```
Repeat until no recombination or 1 child left:
    foreach child node i:
        Update the statistics for node i.
        Calculate the MLLR transform for node i, A_i.
        Apply the transform A_i to the basic model set λ
        to give λ_i.
        foreach segment k:
            Calculate the new likelihood of the segment k
            given the transformed model λ_i.
        end
        foreach child node j:
            Sum the weighted likelihoods of the segments
            in j to give the average log likelihood, L(i,j),
            of node j given the transformed model λ_i.
        end
    end
    foreach possible merge (i,j):
        calculate XProb(i,j) = L(i,i)+L(j,j)-L(i,j)-L(j,i)
    end
    find min(XProb) = XProb(i_min, j_min).
    if (min(XProb) < threshold)
        combine i_min and j_min.
    endif
end
```

A metric similar to cross entropy is calculated to determine when merging should be applied. The measure:

$$XProb(i,j) = L(i,i) + L(j,j) - L(i,j) - L(j,i)$$

where L(i,j) is the log likelihood of the data in node j given the models transformed using the data in node i, can also be expressed in terms of the likelihood p:

$$\frac{\prod_{x=y} p(x|y)}{\prod_{x \neq y} p(x|y)} = \frac{p(i|i)p(j|j)}{p(i|j)p(j|i)} = \exp(XProb(i,j))$$

This measure offers a good, robust method of selecting clusters to merge. Data can be combined and a new transform calculated with very little loss of performance if two different transforms produce very similar likelihoods for the data after adaptation. Since often a given number of clusters is required, this method is better than raising a minimum occupancy count to reduce the number of clusters as it takes similarity into account.

5. EXPERIMENTS

Experiments on various sets of broadcast news data have been carried out to evaluate the effect on the adapted data likelihood using a single mean-only MLLR transform per cluster. The changes in the word error rate of adapted recognition systems were then found. Before clustering, the data was automatically segmented and labelled by bandwidth and gender using the method described in [2].

5.1. Comparing Likelihoods

The 52 narrowband-male, 188 wideband-female and 277 wideband-male segments from the 1996/7 Hub4 broadcast news development data were clustered. As a baseline the CMU clustering software distributed by NIST [5] was used. Covariance-based clustering using the Gaussian Divergence and direct MLLR-based clusterering were applied to the segments where the parameters were chosen to produce approximately the same number of clusters as the CMU scheme. The increase in log likelihood of the data from using the clustered MLLR transforms for the different schemes is given in Table 1, with the number of clusters produced in parenthesis.

Clustering Method	Telephone Male	Wideband Female	Wideband Male
CMU clusterer [5]	1.774 (13)	1.596 (45)	1.598 (53)
covariance-based	1.808 (13)	1.746 (42)	1.596 (53)
cov. with recomb.	1.850 (13)	1.761 (44)	1.617 (54)
direct-maximisation	1.872 (14)	1.811 (47)	1.668 (57)
d-m with recomb.	1.890 (14)	1.832 (46)	1.689 (56)

Table 1: Increase in log likelihood on dev97 data - fixed number of clusters

These results clearly show the advantages of adding recombination to the strategies. The superior performance of the MLLR-based methods over the covariance-based scheme is also illustrated.

5.2. Choosing Parameters

In most cases so long as the clusters are of sufficient size to allow computation of the MLLR transforms, the number of clusters produced is not critical.

The relationship between the number of clusters and increase in log likelihood produced by varying the recombination parameter was investigated. For this experiment, the minimum occupancy was set to 25s, the small segment cut-off to 0.5s, the overlap covariance threshold covariance to 1.0 and the distance metrics to the AHS with a full correlation matrix and the Gaussian Divergence with a full covariance matrix. Figure 2 gives the ROC graphs for these conditions for different levels of recombination on the Hub4 development data. A central operating point was chosen for further experiments.

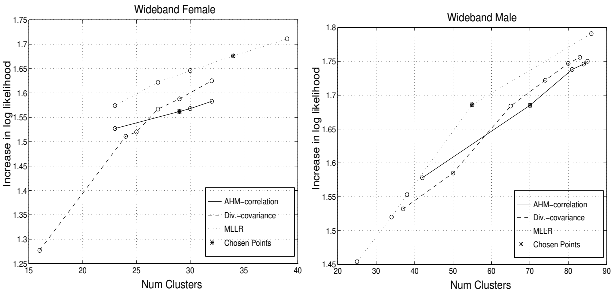

Figure 2: ROCs for Hub4 development data

5.3. Comparing Word Error Rates

The parameters obtained in Sec. 5.2 were used to cluster the 1997 Hub4 evaluation data (34 narrowband-female, 95 narrowband-male, 261 wideband-female, 359 wideband-male segments).

Method	NB Fem	NB Male	WB Fem	WB Male
AHS corr	1.344 (9)	1.586 (31)	2.238 (84)	2.142 (114)
+ recomb	1.269 (5)	1.543 (29)	2.209 (75)	2.109 (102)
MLLR	1.375(10)	1.572 (29)	2.255 (83)	2.144 (117)
+ recomb	1.249 (4)	1.513 (21)	2.145 (52)	2.090 (96)

Table 2: Increase in log likelihood on Hub4 1997 eval data

Method	Corr.	Sub.	Del.	Ins.	WER	Num.Cl.
Unclustered	82.3	13.7	4.0	2.2	19.9	749
AHS corr	83.8	12.8	3.4	2.1	18.4	238
+ recomb	83.8	12.8	3.4	2.2	18.4	211
MLLR-based	83.9	12.7	3.4	2.1	18.3	239
+ recomb	83.9	12.7	3.4	2.1	18.2	173

Table 3: % word error on Hub4 1997 evaluation data

Table 2 gives the absolute increase in log likelihood and number of clusters, whilst Table 3 gives the corresponding word error rates. These were computed with the HTK large vocabulary speech recogniser using gender independent cross-word state-clustered triphone HMMs and a 4-gram broadcast news language model [7]. The error rate for unclustered segments is included as a baseline.

The likelihood results again show the best performance is given by the direct maximisation method. Recombination (with fixed clustering parameters) reduces the number of clusters and the results show a reduction in the number of clusters of 25% can have a beneficial effect on word error rate since the MLLR transforms are more robustly estimated. An additional benefit of having fewer clusters is that further improvements in performance are likely since more MLLR transforms per cluster could be supported.

6. CONCLUSION

This paper has presented two top-down split-and-merge algorithms. One is based on standard covariance methods, whilst the other is based on a new method for directly maximising the MLLR-adapted likelihood. Recombination schemes for both these methods have been presented and shown to increase performance. Use of these new clustering schemes has been shown to improve word error rates on the Hub4 1997 broadcast news evaluation data.

Acknowledgements

This work is in part supported by an EPSRC grant on "Multimedia Document Retrieval" reference GR/L49611.

7. REFERENCES

1. F Bimbot & L Mathan. *Text-Free Speaker Recognition using an Arithmetic Harmonic Sphericity Measure.* Proc. Eurospeech, 1993, Vol. 1. pp. 169-172
2. T Hain, S E Johnson, A Tuerk, P C Woodland & S J Young. *Segment Generation and Clustering in the HTK Broadcast News Transcription System* Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop pp. 133-137
3. S E Johnson *Speaker Tracking* M.Phil Thesis 1997, Cambridge University Engineering Dept. UK
4. C J Leggetter & P C Woodland. *Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs.* Computer Speech and Language, Vol. 9, pp. 171-185
5. M A Siegler, U Jain, B Raj & R M Stern. *Automatic Segmentation, Classification and Clustering of Broadcast News.* Proc. 1997 DARPA Speech Recognition Workshop, pp. 97-99
6. P C Woodland, M J F Gales, D Pye & S J Young *The Development of the 1996 HTK Broadcast News Transcription System* Proc. 1997 DARPA Speech Recognition Workshop, pp. 73-78
7. P C Woodland, T Hain, S E Johnson, T R Niesler, A Tuerk, E W D Whittaker & S J Young. *The 1997 HTK Broadcast News Transcription System* Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop pp. 41-48