CONTEXT DEPENDENT TREE BASED TRANSFORMS FOR PHONETIC SPEECH RECOGNITION

Bernard Doherty   Saeed Vaseghi   Paul McCourt

Queen's University of Belfast, N. Ireland.

email: (b.doherty, s.vaseghi, pm.mccourt) @ee.qub.ac.uk

ABSTRACT

This paper presents a novel method for modeling phonetic context using linear context transforms. Initial investigations have shown the feasibility of synthesising context dependent models from context independent models through weighted interpolation of the peripheral states of a given hidden markov model with its adjacent model. This idea can be further extended, to maximum likelihood estimation of not only single weights, but a matrix of weights or a transform. This paper outlines the application of Maximum Likelihood Linear Regression (MLLR) as a means of modeling context dependency in continuous density Hidden Markov Models (HMM).

1. INTRODUCTION

Context dependent phones preserve inter-phonetic transitional dynamics which often provide important cues for recognition. However some of the potential advantages of context dependent models are mitigated owing to training data scarcity. This paper presents a novel method for modeling phonetic context using linear context transforms. Conventionally the move from context independent models to context dependent triphone models presents two major obstacles: firstly, a lack of available training data and secondly, the problem of unseen distributions. Initial investigations have shown the feasibility of synthesising context dependent models directly from context independent models. This is achieved through the use of weighted interpolation of the peripheral states of a given hidden markov model with its adjacent model. However it is possible that this idea can be further extended to maximum likelihood estimation of not only single weights, but a matrix of weights or a transform.

This paper outlines the application of Linear Context Dependent Transforms within an MLLR framework, as a means of modeling context dependency in continuous density HMMs. MLLR is an adaptation technique which has been previously used for adaptation to speaker and environmental variations [1] [3]. It involves using adaptation data to derive broad-based general transforms. If only a small amount of adaptation data is presented a few transforms are used for all models in the system, and if more data is available the number of transforms used is increased. A regression class tree is employed to order the Gaussians in the system so that the set of transformations to be estimated can be chosen according to the amount and type of adaptation data available.

The utilisation of MLLR presented here deviates from the application of conventional MLLR in that it is extended to encompass the case of individual Gaussian distributions. This is equivalent to a complete re-estimation of the output probability distribution. Consequently, the required amount of training/adaptation data dramatically increases. In this way, model-specific or triphone-dependent transforms can be estimated. The transformation matrices are calculated to maximise the likelihood of the training/adaptation data, using context dependent models, and can be implemented using the Forward-Backward algorithm.

2. LINEAR REGRESSION TRANSFORMS

Initial investigations have shown the feasibility of synthesising context dependent models from context independent models through weighted interpolation of the peripheral states of a given hidden markov model with its adjacent model. In this way the means of a triphone can be generated from a combination of monophone models. This is illustrated in equation 1 where α and β represent linear weights for state 4 of the triphone ih-n+ah.

$$\mu^{CD}_{ih-n+ah(state4)} = \alpha\,\mu^{CI}_{n(state4)} + \beta\,\mu^{CI}_{ah(state2)} \qquad (1)$$

The weights α and β are determined through the partial differentiation of an error function *E*, with respect to α and β, where C_k represents the mean Cepstral Coefficients of the relevant state. The resultant simultaneous equations are then set to zero and solved to find the optimal weights, α and β, which minimise the error function *E*. In short, the weights are derived to minimise the difference between an estimated triphone and a weighted combination of monophones. Only the means of states 2 and 4 of the triphone are synthesised, with the variances and transition matrices as well as the mean of state 3 coming from an estimated triphone. As an example the linear weights for state 4 of triphone ih-n+ah would be calculated using equation 2 as follows,

$$E = \sum_{k=1}^{39}(C_{k(state4(ih-n+ah))} - \alpha C_{k(state4(n))} - \beta C_{k(state2(ah))})^2 \quad (2)$$

2.1 Regression Transform Approach

MLLR [1] [3] is a speaker adaptation technique which uses a set of regression based transforms to tune HMM mean parameters to a new speaker. It produces improvements with small amounts of adaptation information by the sharing of transformations and data. Each of the transformations are applied to a number of HMM mean parameters and estimated from the corresponding data.

A Gaussian distribution, ***S***, is characterised by a mean vector, $\boldsymbol{\mu}_j$, and a covariance matrix, $\boldsymbol{\Sigma}_j$. Given a parameterised speech frame vector, ***o***, the probability density of that vector and distribution ***S*** is given by,

$$b_j(\boldsymbol{o}) = (2\pi)^{-\frac{n}{2}}\left|\boldsymbol{\Sigma}_j\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{o}-\boldsymbol{\mu}_j)^{\boldsymbol{T}}\boldsymbol{\Sigma}_j^{\boldsymbol{-1}}(\boldsymbol{o}-\boldsymbol{\mu}_j)} \quad (3)$$

The adaptation of the extended mean vector is given by:

$$\hat{\boldsymbol{\mu}}_j = \boldsymbol{W}_j\boldsymbol{\xi}_j \quad (4)$$

where,

- $\boldsymbol{W}_j$ is an (n x (n+1)) transformation matrix,
- $\boldsymbol{\xi}_j = [\omega, \mu_1, \mu_2, \cdots, \mu_n]^T$ is the extended mean vector,
- ω determines offset and can be either 1 (offset) or 0 (no offset),
- and $\hat{\boldsymbol{\mu}}_j$ is the adapted mean vector.

A separate transform for each Gaussian distribution, amounts to a complete re-estimation of the output probability distribution. However, the parameters of unseen distributions still present an obstacle. If the same transformation is used for several distributions and estimated using data from all tied distributions, then unseen distributions can be taken into account. The degree of transformation tying is determined by the amount of adaptation data available. The total likelihood of the model set λ generating the observation sequence ***O***, is given by,

$$\mathrm{P}(\boldsymbol{O}|\lambda) = \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{P}(\boldsymbol{O},\boldsymbol{\theta}\,|\lambda) \quad (5)$$

where,

- ***O*** is an observation vector sequence of length *T* samples,
- $\boldsymbol{\theta} = \theta_1, \theta_2 \cdots, \theta_T$ is a state sequence of length *T*,
- λ is the current set of model parameters and,
- $\boldsymbol{\Theta}$ is the set of all possible state sequences of length *T*, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

The standard auxilliary function is adopted,

$$Q(\lambda, \bar{\lambda}) = constant + P(\boldsymbol{O}|\lambda)\sum_{j=1}^{S}\sum_{t=1}^{T}\gamma_j(t)\log b_j(\boldsymbol{o}_t) \quad (6)$$

where,

- $\bar{\lambda}$ is the re-estimated set of model parameters and,
- $\gamma_j(t)$ is the a posteriori probability of occupying state *j* at time *t* given the observation sequence ***O***.

Model parameters which maximise the auxilliary function also increase the value of the objective function unless it is at a maximum. To find the maximum of the auxilliary function we differentiate it (for details refer to [1]) and set the answer equal to zero. This results in the equation,

$$\sum_{t=1}^{T}\gamma_j(t)\boldsymbol{\Sigma}_j^{-1}\boldsymbol{o}_t\xi_j^T = \sum_{t=1}^{T}\gamma_j(t)\boldsymbol{\Sigma}_j^{-1}\overline{\boldsymbol{W}}_j\boldsymbol{\xi}_j\boldsymbol{\xi}_j^T \quad (7)$$

$\overline{\boldsymbol{W}_j}$ is computed row by row from the general equation,

$$\boldsymbol{W}_i^T = \boldsymbol{G}^{(i)^{-1}}\boldsymbol{Z}_i^T \quad (8)$$

where ***Z*** is the LHS of equation 7 and ***G*** is the product of the scaled covariance with the outer product of the two mean vectors on the RHS of equation. Equation 8 can be solved using LU decomposition or Gaussian elimination methods.

MLLR is presented here as a means of deriving context dependent transforms. In this way, context dependent mappings can be created for a 3 state single Gaussian per state HMM,

$$\boldsymbol{W}_j\boldsymbol{\mu}_j^{CI} \rightarrow \boldsymbol{\mu}_j^{CD} \quad (9)$$

where,

- $\boldsymbol{W}_j$ is the state dependent (n x n) transformation matrix,
- $\boldsymbol{\mu}_j^{CI}$ is the state mean vector and
- $\boldsymbol{\mu}_j^{CD}$ is the newly adapted context dependent mean.

2.2 Context -Dependent Transforms

The Linear Regression Transform approach is similar in essence to the least squares approach, however it differs in the treatment of covariance's. Initially this method is applied to a single mixture triphone system. It parallels the least squares approach here in assuming that the shape of the distributions of each acoustic class modelled, are identical. This results in a simplification of equation 7 to the form,

$$\sum_{t=1}^{T}\gamma_j(t)\boldsymbol{o}_t\,\xi_j^T = \sum_{t=1}^{T}\gamma_j(t)\overline{\boldsymbol{W}}_j\,\xi_j\xi_j^T \qquad (10)$$

where,

- $\overline{\boldsymbol{W}}_j$ is an (n x n) transformation matrix and
- $\xi_j = [\mu_1, \mu_2 \cdots, \mu_n]^T$ is the state mean vector.

2.3 Inversion of Transform

In order to compute $\overline{\boldsymbol{W}}_j$ we need to invert $\boldsymbol{G}^{(i)}$. Experiments have demonstrated that this matrix may be ill-conditioned for certain transforms, so even small round off errors that occur during the inversion of a matrix can have a drastic effect on the solution.

$$\boldsymbol{G}^{(i)}\boldsymbol{W}_i^T = \boldsymbol{Z}_i^T \qquad (11)$$

Equation 11 is solved using LU decomposition and back substitution. In this method of solution, round-off errors that accumulate can be magnified to the extent that the matrix is close to singular. Iterative improvement of the solution can be implemented using equation 12,

$$\boldsymbol{G}^{(i)}\delta\boldsymbol{W}_i^T = \boldsymbol{G}^{(i)}(\boldsymbol{W}_i^T + \delta\boldsymbol{W}_i^T) - \boldsymbol{Z}_i^T \qquad (12)$$

In solving for $\delta\boldsymbol{W}$ and subtracting from the original estimate there is a significant improvement in the solution obtained [4].

3. TREE BASED CLUSTERING

One of the major issues concerning tree based clustering is the devising of an optimal data clustering strategy [2] [3]. In real terms this is concerned with how to best cluster components in such a way that they will all have a similar context transformation matrix, thus incorporating the concept of data redundancy. The clustering procedure is based on a measure of similarity between the transforms, rather than defining a set number of clusters to be obtained. A binary tree is constructed and split based on the level of divergence encountered in any given node. If the level of divergence falls below a defined threshold in the given node, then the splitting procedure is stopped. In this way the degree of similarity between matrices can be examined at varying levels of clustering.

3.1 Clustering Algorithm

The clustering algorithm, which is based on a pairwise distance measure, is outlined as follows :

1. In any given node, an average of all transform matrices is calculated.
2. The matrix ***D***, which is furthest away from the average transform is then found. This will form the centre of clustering for the right child of the tree node to be split.
3. Similarly, the procedure is repeated to find the matrix furthest away from the matrix ***D*** found in step 2. This will form the centre of clustering for the left child of the tree node to be split.
4. All the matrices in the current node are then searched and assigned to either the left or right child depending on which has the greater degree of similarity.

3.2 Measure of Divergence

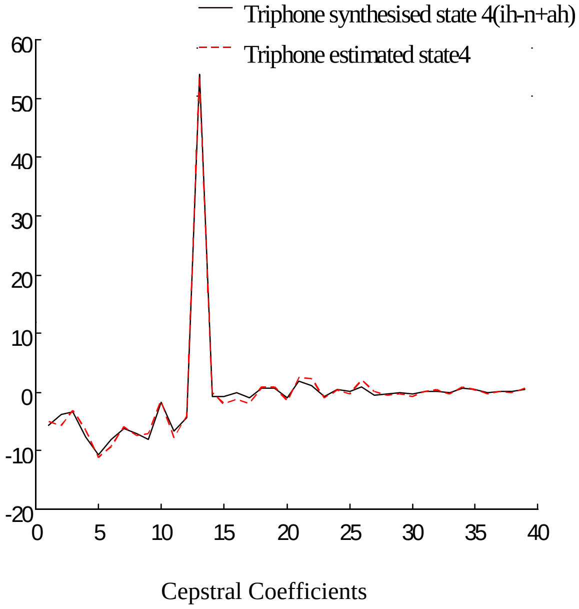
The clustering procedure is based on a distance measure, the assumption being that acoustically similar models will be statistically closer together. The measure used is given as,

$$\|C\|_2 = \frac{\sqrt{\sum_j\sum_k (A_{ij} - B_{ij})^2}}{i \text{ x } j} \qquad (13)$$

where *A* and *B* are the transforms to be compared and *i* and *j* represent the dimensions of a transform. Equation 13 is termed as the Frobenius norm and is used as a measure of the comparability of two matrix transformations or the Mean Squared Error.

4. EVALUATION

Using the least squares approach to synthesising context dependent triphones resulted in an improvement of 2% from a baseline of 57% recognition This was in a single mixture gaussian monophone system using the TIMIT speech database. The graph below details the results of synthesising the mean of state 4 of the triphone ih-n+ah. For this case α was calculated as 0.86084 and β as 0.16995.

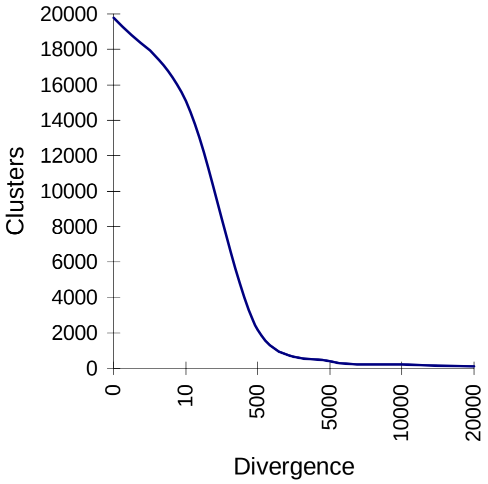


Context dependent transforms were generated using the training set of the TIMIT speech database. 9,573 three state single mixture word internal triphones were used. The speech features consisted of 13 MFCC's supplemented with 1st and 2nd differentials. They were constructed using the HMM Tool Kit (HTK). Transforms were only generated for models which had 3 or more examples. This resulted in 19794 state dependent context transforms.

Divergence	Number of Clusters
20000	123
10000	217
5000	391
500	2157
10	15091

Table 1: The table above gives the clustering of context dependent transforms based on a varying scale of divergence.

It was found that specifying high levels of divergence resulted in an uneven distribution in the number of models per cluster. A single cluster tended to dominate, containing the majority of transforms. As the level of divergence or Mean Squared Error decreased, the number of clusters increased towards the total number of transforms (19794).



The graph details the measure of similarity, or divergence, against the number of clusters obtained. The key point in the graph is that at which the level of divergence is such that 19794 clusters are obtained. This gives the lowest significant difference between context transforms and is a measure of the usefulness of this technique. As illustrated, this is achieved at quite a low level of divergence, signifying quite a diverse range of transforms.

5. DISCUSSION AND CONCLUSIONS

A new application for Maximum Likelihood Linear Regression has been investigated. The concept of cluster based generalisation of context has been examined as a means of transforming a context independent model or group of models to a particular context. Through the employment of this technique robust context transforms can be developed which model the effects of context dependency. The potential similarity of transforms is an indication of the redundancy in the update mechanism for the mean parameters. This may afford a large degree of parameter redundancy in the re-estimation process. Essentially, what the transforms attempt to do is to capture the dynamics of the estimation process in a form not previously examined. This allows the very characteristics of context to be extracted and effectively handles the problem of unknown distributions, as transforms can be re-used for unseen models.

In this respect it has been found that the transforms generated provided a generally poor degree of generalistion of context. However this has to be weighed against the criterion which is used as the benchmark for defining some significant dimension of semblance between transforms. The significance of the current measure is not entirely clear. It is useful, however, as a first approximation. Other criteria are also being examined such as the use of eigenvalues and eigenvectors as a means of characterising a transform. This would hope to yield a more fundamental definition. Further addition to the current clustering procedure is the idea of initial pre-clustering using linguistic knowledge. This would have substantial gains computationally as it would help lessen the already heavy clustering workload. Future research is also being directed towards the possibility of synthesising triphones from pairs of biphones through a process of weighted interpolation. This would be beneficial from the point of view of data availability as there is an abundance of biphones.

6. REFERENCES

1. C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for speaker adaptation of continuous density hidden markov models" , Computer Speech and Language, 9:171-186, 1995.
2. M.J.F. Gales, "The generation and use of Regression Class Trees for MLLR adaptation ", Technical Report CUED/F-INFENG/TR263, Cambridge University, 1996. Available via anonymous ftp from : svr-ftp.eng.cam.ac.uk.
3. C.J. Leggetter, "Improved acoustic modeling for HMM's using Linear Transformations", Ph.D. Thesis, Cambridge University, 1995.
4. "Numerical Recipes in C", pp 49-51, Cambridge University Press, 1988.