EXPLOITATION OF FEATURE VECTOR STRUCTURE FOR SPEAKER ADAPTATION

Eric H.C. Choi, Trym Holter, Julien Epps, Arun Gopalakrishnan

Motorola Australian Research Centre, Motorola Labs
{Eric.Choi, Trym.Holter, Julien.Epps, Arun.Gopal}@motorola.com

ABSTRACT: In this paper we suggest that rather than modelling speaker mismatch as an affine transform of the entire feature vector, it can be modelled by an affine transform of the static coefficients with additional constraints imposed by the temporal relationships of the streams of coefficients. This results in the different streams sharing the same rotation matrix, and thus reduces the complexity and memory requirements for speaker adaptation, as well as minimises the adaptation data requirements. We present the solution for the case where temporal structure constrained transforms (TSCT) are optimised using the maximum likelihood criterion. The experiments presented in the paper show that with the proposed approach, a relative improvement in accuracy of around 26% can be observed for a digit recognition task over the use of block-diagonal transforms, with as little as two adaptation utterances.

1. INTRODUCTION

It is well known that speaker-independent (SI) automatic speech recognition (ASR) systems, despite the steady improvements generated over recent years, still have error rates that are much larger than corresponding speaker-dependent (SD) systems [Zhang et. al. 2002]. However, it takes large amounts of data to properly estimate the model parameters of an ASR system, and it is therefore in most cases not possible to collect sufficient SD data to create such models. More commonly, data from a large pool of different speakers are used to generate SI acoustic models. Even though some speakers experience very good performance with such models, a large variation in accuracy can be expected in the population, depending on how well a particular user's voice characteristics (including both physiological and sociological aspects) are represented in the training data set.

A common solution to the speaker mismatch problem is to employ *speaker adaptation* techniques [Zhang et. al. 2002; Gunawardana & Byrne 2000; Kwong 1999; Leggetter & Woodland 1995]. Such methods modify the parameters of the SI acoustic models, using only small amounts of speaker specific data, to generate the speaker-adapted (SA) models. The goal is to approach the accuracy that can be achieved using SD ASR systems, while at the same time minimising the training load on a user.

One of the most successful approaches to speaker adaptation in the hidden Markov model (HMM) framework is *maximum likelihood linear regression* (MLLR) [Leggetter & Woodland 1995]. The most important feature of an HMM is the probability density functions (pdfs) that specify the state output distributions. These are typically Gaussian mixture models. In MLLR, the mean vectors of the Gaussians are grouped into clusters, and each mean vector ($\boldsymbol{\mu}$) in a cluster is adapted using an affine transformation of the form:

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\Gamma}\boldsymbol{\mu} + \boldsymbol{\beta}. \qquad (1)$$

By carefully selecting the number of clusters to be used, not only MLLR can help to create good SA models when a sufficient amount of adaptation data has been collected, but also it yields improvements when only small amounts of data are available. In this case, the clustering approach will help adapting mean vectors for which there are no data available in the adaptation set. However, with small amounts of data, the number of transformation parameters that can be reliably estimated

will ultimately limit the extent of the performance improvement. In this paper we present a procedure that exploits the temporal relationship that is typically used in ASR feature vectors. The objective of this approach is to reduce the dimensionality of the affine transformations. We will show how this can help reducing the computational complexity, as well as increasing the adaptation rate. The latter is possible because fewer parameters will be required to specify each transformation.

In the next section we will develop the solution for this new transformation under the maximum likelihood (ML) criterion. In section 3, we will then report some experimental results with this novel method for a noisy English digit database, and compare them to a standard MLLR approach. The results are discussed in section 4 while our conclusion can be found in section 5.

2. TEMPORAL STRUCTURE CONSTRAINED TRANSFORMATION (TSCT)

The feature vectors typically used in ASR consist of a stream of static coefficients, augmented by their individual 1st order and 2nd order time derivatives. It is often assumed that there is no cross-correlation among these streams of coefficients, and this is commonly exploited to reduce the complexity in transformation-based adaptation. The implication of this assumption is that the rotation matrix is reduced to a block-diagonal form, and thus a reduced number of adaptation parameters need to be estimated.

Our approach is different, in that rather than modelling speaker mismatch as an affine transform of the entire feature vector, it is modelled by an affine transform *of the static coefficients*. The additional constraints are then imposed by the temporal relationships of the streams of coefficients. This results in the different streams sharing the same rotation matrix, and therefore we can further reduce the number of non-zero elements in a transformation to one-third of that of a block-diagonal transformation. This reduces the complexity and memory requirements for the adaptation, as well as minimises the adaptation data requirements.

2.1 Definition of TSCT

For an *n*-dimensional input feature vector $\underline{\mathbf{X}}$, its output vector ($\underline{\mathbf{Y}}$) after transformation is given by:

$$\underline{\mathbf{Y}} = \mathbf{\Gamma}\underline{\mathbf{X}} + \boldsymbol{\beta} \tag{2}$$

The transformation is thus defined by the *n x n rotation matrix* $\mathbf{\Gamma}$ and the *n* x 1 bias vector $\boldsymbol{\beta}$.

We now assume that the feature vector consists of the static coefficients ($\mathbf{x}$), augmented by their individual 1st order ($\dot{\mathbf{x}}$) and 2nd order ($\ddot{\mathbf{x}}$) time derivatives. We can write this as $\underline{\mathbf{X}} = [\mathbf{x}^T \ \dot{\mathbf{x}}^T \ \ddot{\mathbf{x}}^T]^T$. Assuming that speaker mismatch can be modelled by an affine transform *of the static coefficients*, these coefficients are transformed to

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}\ , \tag{3}$$

where $\mathbf{A}$ is a *n/3 x n/3* rotation matrix and $\mathbf{b}$ is a *n/3* x 1 bias vector. We can now introduce the constraints imposed by the temporal relationships between the streams of coefficients. It follows from equation (3) that:

$$d\mathbf{y} / dt = \mathbf{A}\, d\mathbf{x} / dt = \mathbf{A}\,\dot{\mathbf{x}} \tag{4}$$

$$d^2\mathbf{y} / dt^2 = \mathbf{A}\, d^2\mathbf{x} / dt^2 = \mathbf{A}\,\ddot{\mathbf{x}} \tag{5}$$

The TSCT transformation can thus be written:

$$\underline{\mathbf{Y}} = \begin{bmatrix} \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \dot{\mathbf{x}} \\ \ddot{\mathbf{x}} \end{bmatrix} + \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{A}\mathbf{x} + \mathbf{b} \\ \mathbf{A}\dot{\mathbf{x}} \\ \mathbf{A}\ddot{\mathbf{x}} \end{bmatrix} \tag{6}$$

By contrast, the standard block-diagonal transformation is reached by assuming that speaker mismatch can be modelled by an affine transformation of the entire feature vector and at the same time assuming that the cross-correlation between the streams of the feature vector is zero. For this conventional case the transformation can be written:

$$\underline{\mathbf{Y}} = \begin{bmatrix} \mathbf{y} \\ \dot{\mathbf{y}} \\ \ddot{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mathbf{A_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A_2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A_3} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \dot{\mathbf{x}} \\ \ddot{\mathbf{x}} \end{bmatrix} + \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \mathbf{b}_3 \end{bmatrix} \tag{7}$$

In this case, each $\mathbf{A}_i$ is of dimension *n*/3 x *n*/3, while each of the $\mathbf{b}_i$ is of dimension *n*/3 x 1.

2.2 Estimation of TSCT

The general criterion function to be maximised in MLLR for estimating a transformation is given by [Leggetter & Woodland 1995]:

$$\begin{aligned} J &= -\frac{1}{2}\sum_{t=1}^{\tau}\sum_{k\in\Omega}\left[\gamma_t(k)\left(\underline{\mathbf{o}}_t - \boldsymbol{\Gamma}\underline{\boldsymbol{\mu}}_k - \boldsymbol{\beta}\right)^T \underline{\mathbf{R}}_k \left(\underline{\mathbf{o}}_t - \boldsymbol{\Gamma}\underline{\boldsymbol{\mu}}_k - \boldsymbol{\beta}\right)\right] \\ &= -\frac{1}{2}\sum_{t=1}^{\tau}\sum_{k\in\Omega}\left[\gamma_t(k)\left(\underline{\mathbf{o}}_t - \underline{\mathbf{W}}\underline{\tilde{\boldsymbol{\mu}}}_k\right)^T \underline{\mathbf{R}}_k \left(\underline{\mathbf{o}}_t - \underline{\mathbf{W}}\underline{\tilde{\boldsymbol{\mu}}}_k\right)\right] \end{aligned} \tag{8}$$

where τ is the total number of feature vectors in an adaptation data set, Ω is the set of Gaussians within a regression class, $\underline{\mathbf{o}}_t$ is the feature vector at time *t*, $\gamma_t(k)$ is the posterior probability of $\underline{\mathbf{o}}_t$ occupying the *k*-th Gaussian at time *t*, $\underline{\boldsymbol{\mu}}_k$ is the mean vector of the *k*-th Gaussian and $\underline{\mathbf{R}}_k$ is the corresponding diagonal covariance matrix, $\boldsymbol{\Gamma}$ is the rotation matrix of the transform, $\boldsymbol{\beta}$ is the bias vector, $\underline{\mathbf{W}} = [\boldsymbol{\Gamma}\ \boldsymbol{\beta}]$ and $\underline{\tilde{\boldsymbol{\mu}}}^T = [\underline{\boldsymbol{\mu}}^T \quad 1]$.

In order to simplify the subsequent equations, we add a superscript (i) to a vector or a matrix to identify the corresponding stream of coefficients that it is referred to. For example, $\mathbf{x}^{(1)}$ refers to the static coefficients, $\mathbf{x}^{(2)}$ refers to delta coefficients (1st order time derivatives) and so on. By imposing the temporal structure constraints, equation (8) can be re-written as:

$$J = -\frac{1}{2}\sum_{t=1}^{\tau}\sum_{k\in\Omega}\gamma_t(k)\left[\sum_{i=1}^{3}\left(\mathbf{o}_t^{(i)} - \mathbf{W}\hat{\boldsymbol{\mu}}_k^{(i)}\right)^T \mathbf{R}_k^{(i)}\left(\mathbf{o}_t^{(i)} - \mathbf{W}\hat{\boldsymbol{\mu}}_k^{(i)}\right)\right] \tag{9}$$

where $\hat{\boldsymbol{\mu}}_k^{(1)} = \begin{bmatrix} \boldsymbol{\mu}_k^{(1)} \\ 1 \end{bmatrix}$, $\hat{\boldsymbol{\mu}}_k^{(2)} = \begin{bmatrix} \boldsymbol{\mu}_k^{(2)} \\ 0 \end{bmatrix}$, $\hat{\boldsymbol{\mu}}_k^{(3)} = \begin{bmatrix} \boldsymbol{\mu}_k^{(3)} \\ 0 \end{bmatrix}$, $\mathbf{W} = [\mathbf{A}\ \mathbf{b}]$.

To further simplify the notation, we introduce

$$\bar{\gamma}_k = \sum_{t=1}^{\tau}\gamma_t(k) \text{ and } \bar{\mathbf{o}}_k^{(i)} = \sum_{t=1}^{\tau}\gamma_t(k)\mathbf{o}_t^{(i)}$$

By differentiating the above criterion function in equation (9) with respect to the transform matrix $\mathbf{W}$ and equating the resultant matrix equation to 0, we obtain the following equation system:

$$\sum_{k\in\Omega}\sum_{i=1}^{3}\mathbf{R}_k^{(i)}\bar{\mathbf{o}}_k^{(i)}\hat{\boldsymbol{\mu}}_k^{(i)T} = \sum_{k\in\Omega}\bar{\gamma}_k\sum_{i=1}^{3}\mathbf{R}_k^{(i)}\mathbf{W}\hat{\boldsymbol{\mu}}_k^{(i)}\hat{\boldsymbol{\mu}}_k^{(i)T} \tag{10}$$

To allow more flexibility in this framework, we now introduce a scaling factor for the occupation count corresponding to each stream of a feature vector, i.e., $\gamma_t(k)$ is multiplied by $\lambda^{(i)}$. The equation system can then be written as:

$$\sum_{k\in\Omega}\sum_{i=1}^{3}\lambda^{(i)}\mathbf{R}_k{}^{(i)}\bar{\mathbf{o}}_k{}^{(i)}\hat{\boldsymbol{\mu}}_k^{(i)T} = \sum_{k\in\Omega}\bar{\gamma}_k\sum_{i=1}^{3}\lambda^{(i)}\mathbf{R}_k{}^{(i)}\mathbf{W}\hat{\boldsymbol{\mu}}_k{}^{(i)}\hat{\boldsymbol{\mu}}_k^{(i)T} \tag{11}$$

The above equation system can be solved row-by-row and the solution is given by:

$$\sum_{k\in\Omega}\sum_{i=1}^{3}\lambda^{(i)}r_k{}^{(i)}(j)\bar{\mathbf{o}}_k{}^{(i)}(j)\hat{\boldsymbol{\mu}}_k^{(i)T} = \mathbf{W}(j)[\sum_{k\in\Omega}\bar{\gamma}_k\sum_{i=1}^{3}\lambda^{(i)}r_k{}^{(i)}(j)\hat{\boldsymbol{\mu}}_k{}^{(i)}\hat{\boldsymbol{\mu}}_k^{(i)T}] \tag{12}$$

where $r_k{}^{(i)}(j)$ is the *j*-th diagonal element of $\mathbf{R}_k{}^{(i)}$, $\bar{\mathbf{o}}_k{}^{(i)}(j)$ is the *j*-th element of $\bar{\mathbf{o}}_k{}^{(i)}$ and $\mathbf{W}(j)$ is the *j*-th row of $\mathbf{W}$.

Note that if we choose $\lambda^{(1)}=\lambda^{(2)}=\lambda^{(3)}$, the solution to this equation system provides the maximum likelihood (ML) solution to the TSCT optimisation problem. If we choose to let these scaling factors have different values, we will deviate from this solution. However, this framework increases the flexibility of the approach by allowing the contribution from the different parameter sets to be weighted differently. It would for instance give us the opportunity to calculate the transform based only on the static coefficients by setting $\lambda^{(1)}=1$ and $\lambda^{(2)}=\lambda^{(3)}=0$, thus incorporating the technique suggested in [Choi 1996] as a special case.

3. EXPERIMENTS

3.1 Experimental Setup

In order to evaluate the efficacy of the temporal structure constrained transformations relative to conventional MLLR, an experiment was performed on a digit-dialling task using a proprietary speech recognition search engine. MFCC, delta-MFCC and delta-delta-MFCC features were extracted for each 10 ms frame of input speech using the advanced front-end [Macho 2002]. SI acoustic models were generated using digit portions of the Macrophone speech corpus [Bernstein 1994], which consists of many different speakers reciting strings of between three and eighteen digits (from /oh/, /zero/, /one/, . . . /nine/) in length in clean (non-noisy) conditions. 4184 digit strings comprising 39484 digits spoken by 2003 speakers were used to generate eleven whole word models with 14 emitting states and 6 Gaussian components per state, and one silence model. From these models, a regression tree was formed with 98 regression classes.

Adaptation and testing were performed using a noisy condition ("hands-free" mode speech inside a car with engine running) from a proprietary speech corpus that comprises digit strings of the same format as the Macrophone corpus used to train the SI models. Supervised adaptation (i.e. where adaptation is performed with knowledge of the correct digit sequence) was then performed for each of the 16 speakers, on a subset of the data available for that speaker. This produced a set of transformations for that speaker. This process was repeated using 1, 2, 3, 5, 10 and 20 digit strings as adaptation data, for each speaker. Preliminary experiments testing various scaling factor values revealed that the recognition accuracy of the TSCT was maximised for $\lambda^{(1)}=\lambda^{(2)}=\lambda^{(3)}$, and accordingly settings of $\lambda^{(1)}=\lambda^{(2)}=\lambda^{(3)}=1$ were employed in this experiment. During adaptation, an occupation count threshold of 200 was employed.

The transformations produced by the previous step were then applied to the SI models and the transformed models were tested on 40 digit strings formed from another subset of the internal speech corpus (disjoint from the subset used during adaptation). The average string recognition accuracy over all 40 strings was then calculated for each speaker, and then averaged again across all speakers to produce a series of recognition accuracies corresponding to adaptation over 1, 2, 3, 5, 10

and 20 digit strings. Note that string accuracy is defined here as the percentage of strings for which every digit in the string is correctly recognized.

The entire preceding experimental process was performed once for conventional block-diagonal transformations, and once for block-diagonal TSCT's. In both cases, the same experimental settings and data were used at every stage.

3.2 Results

The recognition accuracy results, as shown in Figure 1, show that for the parameter settings used in this experiment, temporal structure constrained transformations perform better for 1~5 utterances, while conventional transformations perform better for greater amounts of adaptation data. It can be observed that the adaptation rate has been improved significantly by using TSCT's. With only one adaptation utterance, the improvement in string accuracy is more than 3% absolute. It demonstrates that the temporal constraints provide a more efficient means of capturing speaker characteristics with sparse speech data.

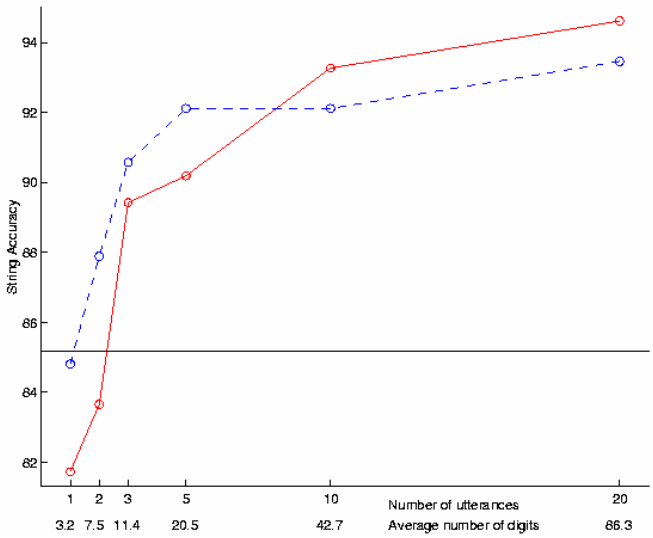

Figure 1. Average string recognition accuracy (%) across all (16) speakers resulting from SI models transformed using conventional (solid) and temporal structure-constrained (dashed) block-diagonal transformations. The result from the SI (untransformed) models is also shown (solid, horizontal).

4. DISCUSSION

Due to the intentional constraints on the transformation structure of the novel approach, the TSCT results in section 3 were obtained using exactly one-third of the memory required for the storage of the conventional block-diagonal transformations. This is a considerable saving that could be critical in an embedded implementation. Not only the use of TSCT requires less storage, but also, more importantly, it needs less amount of speech data to obtain the same improvement in accuracy comparing with conventional approach. The improvement in adaptation rate is crucial in enhancing the initial experience of a new user for using a voice interface.

Although the TSCT approach does not provide better accuracy beyond 5 utterances as shown in Figure 1, it is entirely feasible to make parameter adjustments (e.g. an increase in the number of transforms) as the number of utterances increases in order to provide virtually identical performance to that of the conventional approach beyond 5 utterances. No parameter adjustment of the

conventional approach, however, will provide better performance than that produced by the TSCT approach over 1~5 utterances. As a matter of fact, our further investigation has revealed that for a given recognition accuracy, the TSCT approach always requires the least amount of total transformation parameters (i.e. number of transforms times the number of coefficients in a transform). In other words, fewer transformation parameters are needed to efficiently capture the differences in voice characteristics of different speakers by using TSCT.

The graph in Figure 1 indicates that the accuracy after using one adaptation utterance based on TSCT is close to the corresponding SI performance. Although the accuracy is still a little bit lower than that obtained by just using the SI models, it is not an issue at all since other techniques can be incorporated into the TSCT framework for achieving a better accuracy than the SI performance. An example of such techniques is the use of discounted likelihood in computing the transformations [Gunawardana & Byrne 2000].

The derivation here assumes that the rotation matrix $\mathbf{A}$ is of full-rank. Therefore further simplification of the TSCT is possible if we also impose that $\mathbf{A}$ is a diagonal matrix. In this case, equation (10) becomes a set of independent scalar equations that can then be solved individually without involving any matrix inversions. This particular form of transformation is most suitable for embedded system implementation for its simplicity in computation.

5. CONCLUSION

Rapid adaptation on very limited data is a difficult task for which there are few improvements on conventional approaches. It is during these first few utterances that a user will have their critical first experience of a particular speech recognition system. Thus, it is claimed that the TSCT approach provides a strong advantage. In particular, we have demonstrated the better performances of the TSCT approach on a digit recognition task, in terms of both memory storage and accuracy after adaptation. The experimental results verified that temporal structure of feature vectors can be used to reduce the complexity of a transformation and at the same time to enhance the capability of a transformation in capturing the different voice characteristics.

REFERENCES

Bernstein J., Taussig K. and Godfrey J. (1994), Macrophone: An American English Telephone Speech Corpus for the POLYPHONE Project, in *Proc. IEEE ICASSP'94*, vol. 1, pp. 81-84.

Choi E. (1996), *Spectral Transformation for Speaker Adaptation in HMM Based Speech Recognition*, Ph.D. thesis, University of Sydney.

Gunawardana A. and Byrne W. (2000), Robust Estimation for Rapid Speaker Adaptation Using Discounted Likelihood Techniques, in *Proc IEEE ICASSP'00*, vol. 2, pp. 985-988.

Kwong S. et. al. (1999), Speaker Adaptation Technique for HMM Model, in *Electronics Letters*, vol. 35, no. 21, pp. 1817-1818.

Leggetter C. J. and Woodland P. C. (1995), Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models, in *Computer Speech and Language*, vol 9, no. 2, pp. 171-185.

Macho D. et. al. (2002), Evaluation of a Noise-Robust DSR Front-End on Aurora Databases, in *Proc. Int. Conference on Spoken Language Processing (ICSLP'02)*, vol. 1, pp. 17-20.

Zhang Z, Furui S. and Ohtsuki K. (2002), On-line Incremental Adaptation for Broadcast News Transcription, in *Speech Communication*, issue 37, pp. 271-281.