DATA-DRIVEN PHONETIC REGRESSION CLASS TREE ESTIMATION FOR MLLR ADAPTATION

Reinhold Haeb-Umbach

Philips Research Laboratories Aachen
Weisshausstrasse 2, 52066 Aachen, Germany
Reinhold.Haeb@philips.com

10.21437/ICSLP.2000-669

ABSTRACT

In this paper a method is presented to estimate a broad phonetic class regression tree to be used in MLLR adaptation. The tree is derived from the correlation structure among phone units estimated on the training data. The algorithm is language-independent and showed good results on both an English and a Mandarin Chinese database. In adaptation experiments the tree outperformed a regression tree obtained from clustering according to closeness in acoustic space and achieved results comparable with those of a manually designed broad phonetic class tree.

1. INTRODUCTION

Maximum likelihood linear regression (MLLR) adaptation has proven to be an effective speaker adaptation technique in the presence of limited adaptation data [7]. A set of linear transformations for the mean (and, possibly, variance) parameters of a mixture Gaussian HMM recognizer is estimated such that the likelihood of the adaptation data is maximized. Since in general there is little adaptation data compared to the number of model parameters, it is necessary to cluster model parameters together into regression classes. It is assumed that all components in a given regression class transform in a similar fashion.

A regression class tree consists of a hierarchy of regression classes and a set of base classes as leaves. All base classes below a tree node may share a common transformation matrix. The number of different transformation matrices is chosen according to the amount of adaptation data available. If only few adaptation data are present a single transformation matrix at the tree root is calculated and applied to all base classes. The more adaptation data become available, the further the tree is descended and the more specific transformation matrices are computed.

An important question is which model parameters to group together into regression classes. Ideally, the model parameters should be grouped such that the likelihood of the adaptation data is maximized (for a given number of classes). Since this optimization is typically computationally intractable, one resorts to suboptimal solutions. Two approaches are common practice [4]:

Phonetic knowledge: Here, expert knowledge is used to decide which components are to be transformed together. The components are split according to broad phonetic classes (e.g. nasals, glides) or, at a lower level, into phones.

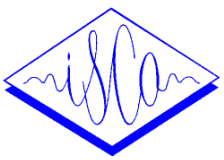
Acoustic space: Components are clustered according to how close they are in acoustic space, irrespective of which phone they belong to. This has the advantage of being a "data-driven" approach with no need for expert knowledge. However, the resulting classes usually cannot be assigned a phonetic identity.

The problem addressed here is as follows: we would like to get rid of the phonetic expertise required in the first approach but still obtain regression classes which represent broad phonetic classes and which deliver adaptation performance comparable to a hand-designed tree. If phonetic expertise is no longer required, the recognizer can be transferred faster to a new language, of which no in-depth phonetic expertise might be available.

We adopt a bottom-up clustering approach starting from base classes which represent either phone models or individual HMM states. We then would like to cluster those classes for which the assumption of having the same transformation matrix is most justified. It is shown that this leads to a correlation criterion as clustering criterion, in contrast to a criterion based on closeness in acoustic space.

The use of correlation among speech units in adaptation is not new, though the use for this particular problem of regression class tree estimation is. The basic idea is to estimate the correlation structure among speech units on the speaker-independent training data and then use this as an additional constraint in the adaptation process (e.g. [1], [2], [3], [6], [9]). Effective adaptation can then be performed with fewer adaptation data. In [6] the MAP estimation is formulated for correlated Gaussian means. Takahashi and Sagayama [9] perform HMM parameter tying based on the correlation of model parameters. It is argued that sets of parameters with high correlation "move together" in acoustic space and should therefore be tied for adaptation. The Adaptation by Correlation algorithm [2] computes linear minimum variance estimates of unobserved mean vectors from observed ones by exploiting the correlation among the speech units. Bocchieri *et al.* [1] used correlation between the bias terms of the MLLR transformations to refine the transformation function, and Doh and Stern [3] used inter-class correlation to improve the estimates of the transformation matrices.

In Section 2 we motivate the use of a correlation criterion



6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

as cluster criterion to obtain regression classes. Section 3.1 describes how the correlation structure is estimated from the training data and presents two sample correlation trees, one obtained on an English and one on a Mandarin database. Then Section 3.2 presents adaptation results to validate the quality of the estimated regression trees, and in Section 4 we summarize the findings and mention another potential application of the presented algorithm.

2. IMPORTANCE OF CORRELATION

Let us assume that for the estimation of the MLLR transformation matrix A_c of class c sufficient observations are present in the adaptation data, such that μ_c, the speaker-adapted mean vector of class c, can be estimated as follows[1] [7]:

$$\mu_c = A_c \xi_c. \tag{1}$$

ξ_c is the extended original (speaker-independent) mean vector: $\xi_c = (1, \xi_{c1}, \ldots, \xi_{cD})^T$, and A_c is the $D \times (D+1)$ transformation matrix of regression class c, D being the vector dimension.

Suppose that for another class c' no observations are available and thus $A_{c'}$ cannot be computed. However, if we know the joint statistics of μ_c and $\mu_{c'}$, we can still estimate $\mu_{c'}$. The best estimate (in the minimum mean square error sense) is

$$\hat{\mu}_{c'} = E\left[\mu_{c'} | \mu_c\right]. \tag{2}$$

In the case of a jointly Gaussian distribution and assuming that the random variables are zero mean, we obtain:

$$\hat{\mu}_{c'} = \Sigma_{\mu_{c'}\mu_c} \Sigma^{-1}_{\mu_c\mu_c} \mu_c \tag{3}$$

where

$$\Sigma_{\mu_{c'}\mu_c} = E\left[\mu_{c'} \mu_c^T\right], \quad \Sigma_{\mu_c\mu_c} = E\left[\mu_c \mu_c^T\right]. \tag{4}$$

Given these covariances, mean vectors with no observations in the adaptation data can be estimated from observed data. This is what is done in the Adaptation by Correlation algorithm [2].

Equation (3) is a linear transformation from μ_c to $\hat{\mu}_{c'}$ which we denote by $B_{\mu_c\mu_{c'}}$:

$$B_{\mu_c\mu_{c'}} := \Sigma_{\mu_{c'}\mu_c} \Sigma^{-1}_{\mu_c\mu_c}. \tag{5}$$

The covariances of $\mu_{c'}$ and μ_c are not known and there are too few adaptation data to estimate them. Therefore they have to be estimated from the speaker-independent training data: $B_{\mu_c\mu_{c'}} \approx B_{\xi_c\xi_{c'}}$. The change of subscripts shall indicate that B is now estimated on the training data.

From the definition of $B_{\xi_c\xi_{c'}}$ it follows that[2]

$$\hat{\xi}_{c'} = B_{\xi_c\xi_{c'}} \xi_c. \tag{6}$$

[1] For simplicity of discussion and notation we assume that each base class has only one Gaussian.

[2] Note the similarity to eq. (1). $B_{\xi_c\xi_{c'}}$ can thus be estimated by using MLLR on the training data.

Now we obtain for eq. (3) by using (1) and (6):

$$\begin{aligned} \hat{\mu}_{c'} &= B_{\xi_c\xi_{c'}} \mu_c \\ &= B_{\xi_c\xi_{c'}} A_c \xi_c \\ &= B_{\xi_c\xi_{c'}} A_c B^{-1}_{\xi_c\xi_{c'}} \hat{\xi}_{c'}. \end{aligned} \tag{7}$$

Thus, even if no adaptation data for class c' are available, a transformation matrix

$$\hat{A}_{c'} := B_{\xi_c\xi_{c'}} A_c B^{-1}_{\xi_c\xi_{c'}} \tag{8}$$

can be estimated by exploiting the correlation structure among the classes. This is done in the "inter-class MLLR algorithm" [3].

In standard MLLR the inter-class correlation is not used, the unknown $A_{c'}$ is replaced by another transformation matrix, for which sufficient observations are available for its estimation. The foregoing derivation shows that the transformation matrix of that class c should be chosen, which has the largest covariance with the class c'.

3. EXPERIMENTAL RESULTS

3.1. Estimation of Regression Trees

We now want to estimate a regression tree by bottom-up clustering of base classes. From the last section we know that a covariance-based criterion is an appropriate clustering criterion. Base classes with 'high' covariance should be clustered first since here the assumption that they transform in the same way is the most justified.

As a clustering criterion we need a scalar criterion. This is obtained by neglecting the correlation across different vector dimensions and by taking the average over all vector dimensions. After normalization, the resulting scalar can be interpreted as correlation coefficient between two classes, see [5] for details.

The base classes can be either the context-independent phone models, the context-independent HMM states or the context-dependent HMM states. These three choices have increasing numbers of classes. For the following experiments we used context-independent phone models as base classes.

Figure 1 shows the regression tree obtained by the above algorithm on the Wallstreet Journal WSJ0+1 142 male speakers training corpus. The clustering started with the 44 phonemes of our lexicon, the symbols can be found in the rectangles at the tree leaves. The values in the ellipses right above two clusters is the value of the correlation coefficient $r_{c,c'}$ of the respective two clusters below.

Although the tree has been obtained fully automatically, it is remarkable that often phonemes of the same broad phonetic class have been clustered together, e.g. the nasal cluster consisting of the phonemes "un", "n", "ng", "um" and "m". Indeed, the tree resembles very much our broad phonetic class tree which was designed by a phonetic expert.

We also used the clustering algorithm on a Mandarin database to cluster the Philips Mandarin phone set called "SAMPA-C". It is based on the European SAMPA standard and consists of a set of 86 *preme/toneme* models [8]. The resulting regression class tree is depicted in Fig. 2. Again, the tree appears to be quite 'reasonable', see e.g. the "S", "tS", "dZ", "dz", "ts" cluster[3].

3.2. Adaptation Experiments

To evaluate the quality of the regression tree, recognition experiments with online supervised MLLR adaptation have been carried out on the WSJ 5k '92 and '93 development and evaluation test sets. Context-dependent gender-dependent models have been trained on the 42 male and 42 female speakers of the WSJ0 training database. The reference word error rate without any adaptation is 9.9% for the male speakers test database and 7.7% for the female speakers, respectively. It can be seen from the results in Table 1 that the regression tree based on correlations performs almost as well in adaptation as the hand-designed tree.

Table 1: Word error rate (WER) on WSJ 5k 92/93 dev/eval test sets for different MLLR regression trees (BPC = broad phonetic class). 20 male (13113 words) and 18 female (11517 words) speakers, bigram language model.

Regression tree	WER [%]		
	all	male subset	female subset
hand-designed BPC tree	7.81	8.68	6.82
correlation tree	7.87	8.76	6.86
euclidian. distance tree	8.08	9.00	7.04

The appropriateness of the correlation as clustering criterion becomes evident when the performance is compared with clustering according to closeness in acoustic space. For this experiment the same clustering as outlined above was carried out, however, the Euclidian distance between mean vectors representing a base class was used as cluster criterion. This resulted in an overall error rate of 8.08%, compared to 7.87% for the correlation-based clustering. Note that the approximate 95% confidence interval for these experiments is $\pm 0.3\%$.

4. SUMMARY AND OUTLOOK

A correlation-based clustering criterion has been shown to be appropriate for the data-driven design of broad phonetic class regression trees to be used in MLLR adaptation. The proposed design method does not require any phonetic knowledge in addition to the definition of the phoneme set and thus simplifies the setup of a recognizer for a new language. This has been demonstrated since the algorithm has been used without any change on both an English and a Mandarin Chinese database. The tree achieved better error rate performance than a tree obtained by clustering according to closeness in acoustic space, and it achieved similar error rate performance as a broad phonetic class tree designed by a phonetic expert.

Note that the regression tree subdivides the phone set into subsets. This partitioning might also be useful in decision-tree clustering of context-dependent phone units. A question could be "is the right context a phoneme of the set of phonemes found below a certain node in the regression tree?" In this way, the question set is derived automatically. However, the development of this approach is left to future research.

[3]Many thanks to L. Liao from Philips Research Taipei for conducting this experiment.

5. REFERENCES

1. E. Bocchieri et al., "Correlation Modeling of MLLR Transform Biases for Rapid HMM Adaptation to New Speakers", in *Proc. ICASSP*, pp 2343-2346, Phoenix, March 1999.
2. S.S. Chen, P. DeSouza, "Speaker Adaptation by Correlation", in *Proc. EUROSPEECH*, pp 2111-2114, Rhodes, Greece, Sep. 1997.
3. S.-J. Doh, R. Stern, "Inter-Class MLLR for Speaker Adaptation", in *Proc. ICASSP*, pp 1755-1758, Istanbul, June 2000.
4. M.J.F. Gales, "The Generation and Use of Regression Class Trees for MLLR Adaptation", Technical Report CUED/F-INFENG/TR263, Cambridge University, 1996.
5. R. Haeb-Umbach, "Automatic Generation of Phonetic Regression Class Trees for MLLR Adaptation", to appear in *IEEE Trans. on Speech and Audio Processing.*
6. M.J. Lasry, R.M. Stern, "A Posteriori Estimation of Correlated Jointly Gaussian Mean Vectors", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. PAMI-6, No. 4, pp 530-535, July 1984.
7. C.J. Leggetter, P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", *Computer Speech and Language* (1995)9, pp 171-185.
8. F. Seide, N. Wang, "Phonetic Modeling in the Philips Chinese Continuous-Speech Recognition System", in *Proc. International Symposium on Chinese Spoken Language Processing*, 1998.
9. S. Takahashi, S. Sagayama, "Tied-Structure HMM Based on Parameter Correlation for Efficient Model Training", in *Proc. ICASSP*, pp 467-470, Atlanta, GA, 1996.

Figure 1: Broad phonetic class tree obtained from clustering of phonemes with correlation as distance criterion. The numbers in the ellipses are the correlation coefficient between the two subtrees below. The letters in the rectangles denote the phonemes.Tree estimated on WSJ0+1 database.

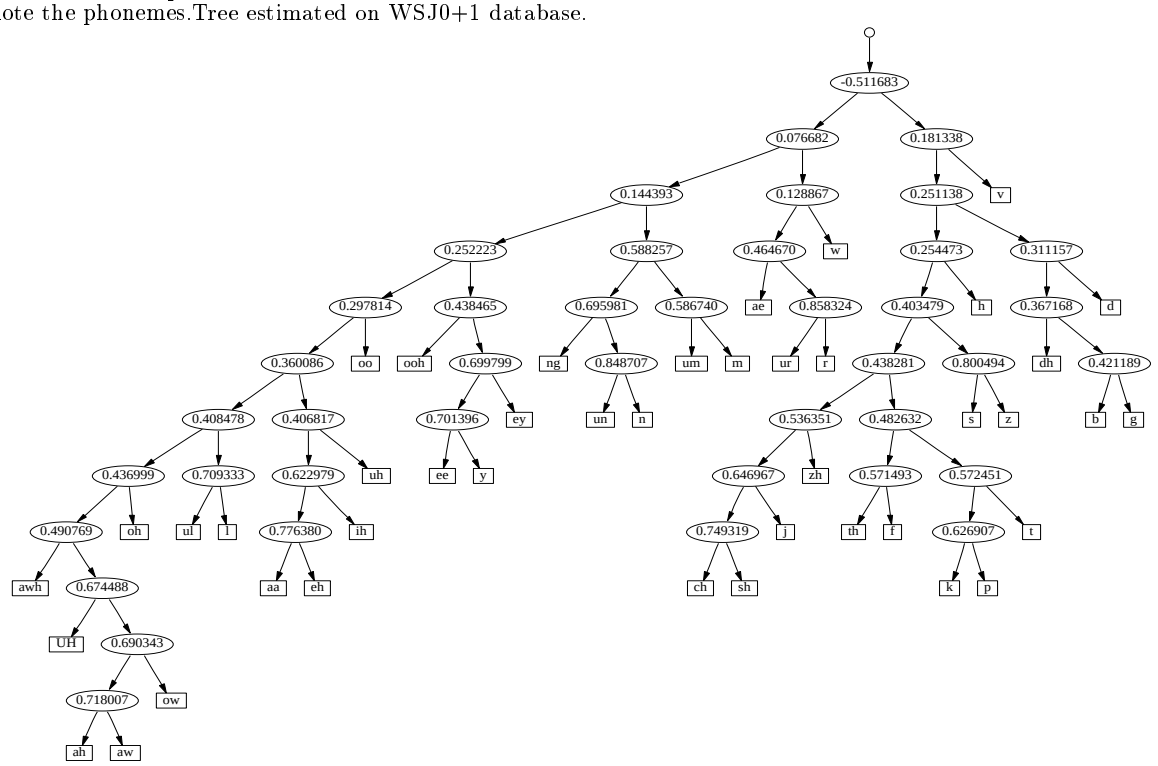


Figure 2: Broad phonetic class tree obtained from clustering the Philips Mandarin phone set on a Mandarin Chinese database.