ACOUSTIC FACTORISATION

M.J.F. Gales

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, UK
mjfg@eng.cam.ac.uk

ABSTRACT

This paper describes a new technique for training a speech recognition system on inhomogenous training data. The proposed technique, *acoustic factorisation*, attempts to explicitly model all the factors that affect the acoustic signal. By explicitly modelling all the factors the trained model set may be used in a more flexible fashion than in standard adaptive training schemes. Since an individual model is trained for each factor, it is possible to factor-in only those factors that are appropriate to a particular target domain, for example the distribution over all training speakers. The target domain specific factors are simply estimated from limited target specific data, for example the target acoustic environment. The theory of this new approach for a particular speaker and environment transforms is described. Initial experiments on a large vocabulary speech recognition task are presented.

1. INTRODUCTION

It is well known that the perceived acoustic signal is influenced by many different factors. The signal varies depending on the words being uttered (the desired variation), the speaker and the acoustic environment, to name a few. When most speech recognition systems are built there is an inherent assumption that the features extracted from the signal are independent of the speaker and acoustic environment. However this assumption is poor. For example the performance of speech recognition systems degrade rapidly as the acoustic environment changes [1]. The standard approach adopted is to ignore this dependence on unwanted factors and to simply train a system on all the data, irrespective of the acoustic environment or speaker associated with the data. Recently the concept of adaptive training [2, 3, 4] has been introduced. Here a transform is associated with each speaker/acoustic environment combination. Usually a linear transform, such as maximum likelihood linear regression (MLLR) [5], is used. A canonical model set is trained given these training speaker/acoustic environment transforms. Hopefully, this canonical model set reflects variations in the underlying signal after the effects of of the unwanted acoustic factors, speaker and acoustic environment, have been taken into account. During recognition a new transform is estimated for a particular target speaker and acoustic environment.

The aim of this paper is to extend the concept of adaptive training so that each of the unwanted factors that affects the acoustic signal is modelled separately. This form of individual source modelling will be referred to as *acoustic factorisation* and the generation of models for each of the individual sources as *factoring-out*. Consider the case where there are two unwanted factors, speaker and acoustic environment variations, in the training data. A separate transform is generated for each speaker, $\mathbf{W}^{(s)}$, and acoustic environment, $\boldsymbol{\lambda}^{(n)}$. A canonical model is then built given these transforms. This form of factorisation has advantages over standard adaptive training. For example if multiple training, or test, speakers are known to be be talking in the same acoustic environment it is possible to explicitly constrain the transform to reflect this. In addition it is possible to train prior distributions for each of the individual factors. This form of explicit modelling allows additional flexibility in the way that the models can be used.

Speaker independent system generation: using a limited number of speakers in the target noise environment it is possible to obtain an estimate of the acoustic environment transform, $\boldsymbol{\lambda}^{(n)}$. Given this transform the prior transform speaker distribution, $p(\mathbf{W})$, which represents the distribution, hopefully, of all speakers may then be *factored-in* to generate a speaker independent system for the target acoustic environment domain.

Multi-environment systems: if a system for a particular speaker in multiple acoustic environments is required then the acoustic environment transform may be factored-in given the estimate of the speaker transform, $\mathbf{W}^{(s)}$.

Posterior adaptation: using the data from a specific speaker in a target acoustic environment, a posterior distribution over the speaker and noise transform parameters may be obtained. These may then be factored-in. By using posterior distributions, rather than ML or MAP estimates, very rapid adaptation can be achieved. Furthermore by factoring the sources fewer transform parameters should be required.

This paper presents the basic theory of acoustic factorisation. First it considers the general form of acoustic factorisation and likelihood expressions that result from applying this factorisation. The paper then describes a particular form of acoustic factorisation that uses MLLR as the speaker transform and cluster adaptive training (CAT) [3] as the noise transform. An approximate scheme for factoring-in the desired factor transform distributions is described. Finally initial experiments on a large vocabulary speech recognition task are described.

2. ACOUSTIC FACTORISATION

For the purpose of this discussion of acoustic factorisation there are assumed to be two unwanted factors affecting the signal; an additive noise with transform $\boldsymbol{\lambda}$ and speaker variations with transform $\mathbf{W}$. Figure 1 shows the dynamic Bayesian network (DBN) that reflects the dependencies on these factors. Also, it is assumed that the noise conditions and speaker are constant over an utter-

This work was partially funded by the European Commission under the Language project Le-5 Coretex. The author also thanks IBM for an SUR equipment award and Phil Wooland for helpful discussions.

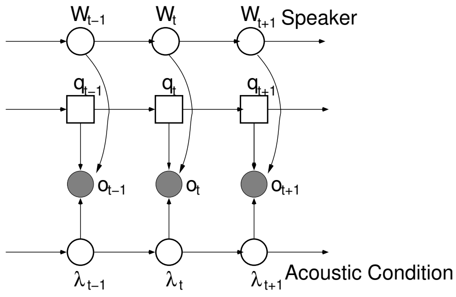

Fig. 1. DBN for acoustic factorisation

ance. Thus $\mathbf{W}_t = \mathbf{W}_{t-1}$ and $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}_{t-1}$. Given these assumptions and using an HMM as the underlying canonical speech model, the likelihood of the utterance $\mathbf{o}_1, \ldots, \mathbf{o}_T$ is given by

$$p(\mathbf{o}_1, \ldots \mathbf{o}_T) = \int_{\mathcal{R}} \sum_{\{\mathbf{Q}_T\}} \left(\prod_{t=1}^{T} p(\mathbf{o}_t | q_t, \mathbf{W}, \boldsymbol{\lambda}) \right) P(\mathbf{Q}) p(\boldsymbol{\lambda}) p(\mathbf{W}) d\mathbf{W} d\boldsymbol{\lambda} \quad (1)$$

where $\{\mathbf{Q}_T\}$ is the set of all valid state sequences through the model of length T and q_t is the state at time t along the particular path $\mathbf{Q}$. From the training data we need to extract two distinct sets of model parameters.

Canonical model parameters: this models the acoustic data given the unwanted acoustic factor transforms. In terms of equation 1 this yields the observation likelihood, $p(\mathbf{o}|q, \mathbf{W}, \boldsymbol{\lambda})$ and the state sequence probability (duration model), $P(\mathbf{Q})$.

Transform prior distributions: these represent the variation over the training speaker transforms, $p(\mathbf{W})$, and acoustic environment $p(\boldsymbol{\lambda})$.

The exact form of the training algorithm depends on the nature of the transforms to represent the speakers and the acoustic environment. In this work the acoustic environment is represented in the form of cluster adaptive training (CAT) [3]. Thus each acoustic environment, n, is represented by a point $\boldsymbol{\lambda}^{(n)}$ in the "noise eigenspace" spanned by the clusters. The speaker is represented as a linear transform of the mean using MLLR [5]. Given these representations the next sections describe how the canonical model and priors are trained and used in recognition.

2.1. Factoring-Out

In common with standard adaptive training schemes an iterative approach is used to estimate the model parameters. Given some initial model set a set of speaker/environment transforms are estimated. Then given the set of transforms a new canonical model set is estimated and the process repeated. From the forms of acoustic environment and speaker transforms, and using an HMM canonical model with an M-component Gaussian mixture model to model each HMM state, the likelihood is given by

$$p(\mathbf{o}|q, \mathbf{W}, \boldsymbol{\lambda}) = \sum_{m=1}^{M} c^{(qm)} \mathcal{N}(\mathbf{o}; \mathbf{A}\mathbf{M}^{(qm)}\boldsymbol{\lambda} + \mathbf{b}, \boldsymbol{\Sigma}^{(qm)}) \quad (2)$$

where $\mathbf{M}^{(qm)} = \left[\begin{array}{ccc} \boldsymbol{\mu}_1^{(qm)} & \ldots & \boldsymbol{\mu}_C^{(qm)} \end{array} \right]$ and $\mathbf{W} = \left[\begin{array}{cc} \mathbf{A} & \mathbf{b} \end{array} \right]$. The training of the model parameters $\mathbf{M}^{(qm)}$, $\boldsymbol{\Sigma}^{(qm)}$ and $c^{(qm)}$ is a restricted version of the extended SAT (ESAT) training scheme described in [4][1]. The relationship to ESAT training is illustrated by writing

$$\mathbf{A}\mathbf{M}^{(qm)}\boldsymbol{\lambda} + \mathbf{b} = \left[\begin{array}{cccc} \lambda_1 \mathbf{A} & \ldots & \lambda_C \mathbf{A} & \mathbf{b} \end{array} \right] \left[\begin{array}{c} \boldsymbol{\mu}_1^{(qm)} \\ \vdots \\ \boldsymbol{\mu}_C^{(qm)} \\ 1 \end{array} \right] \quad (3)$$

The estimation of the factored transforms require simple changes to standard linear transformation schemes. The position in the noise eigenspace $\boldsymbol{\lambda}$ is a modification to the standard CAT scheme using transformed cluster means, $\check{\boldsymbol{\mu}}_c^{(qm)} = \mathbf{A}\boldsymbol{\mu}_c^{(qm)}$, and "bias corrected" observations, $\check{\mathbf{o}}_t = \mathbf{o}_t - \mathbf{b}$. The estimation of the speaker transform, $\mathbf{W}$, is a simple modification to the standard MLLR estimation scheme. The estimation of the MLLR transform is based on the transformed mean, $\hat{\boldsymbol{\mu}} = \mathbf{M}^{(qm)}\boldsymbol{\lambda}$. As the estimation of the position in the noise eigenspace is dependent on the speaker transform, and vice-versa, an iterative scheme looping between estimating the point in eigenspace and speaker transforms is used. Each step is guaranteed to result in a non-decreasing training data likelihood. For further details of this factored transform estimation scheme see [6].

In common with many "factored" transforms there are an infinite number of equal likelihood transforms. From equation 3 by scaling the linear transform to $a\mathbf{A}$ and attenuating the values of the point in eigenspace $\boldsymbol{\lambda}/a$ the likelihood is the same for all values of a. This has no affect on the canonical model trained, however it is important when factoring-in the transforms. Although in this work it is assumed that there is sufficient training speaker/environment data so the prior distribution does not affect the estimates of the transforms, it does determine the scaling value. The scaling $\hat{a}$ is selected such that

$$\hat{a} = \arg\max_{a} \{p(a\mathbf{A})\} \quad (4)$$

where $p(\mathbf{A})$ is the appropriate sections of the speaker transform prior given in equation 5.

2.2. Transform Prior Distribution

An important question is the form of the transform priors. Gaussian prior distributions have been proposed for both CAT [3] and MLLR [7]. However, using a single multi-variate Gaussian distribution to model the distribution over the speaker transforms is limited. For example consider the basic gender split, this should result in a 2-component Gaussian mixture model being required for the speaker transform prior. Hence, the form of prior distribution for MLLR considered in this work is

$$p(\mathbf{W}) = \sum_{p=1}^{P} c^{(p)} \prod_{d=1}^{D} \mathcal{N}(\mathbf{w}_d; \boldsymbol{\mu}_d^{(p)}, \boldsymbol{\Sigma}_d^{(p)}) \quad (5)$$

where P is the number of components in the transform prior distribution and D is the dimensionality of the observation vector.

[1] For the work presented here is it assumed that there is sufficient training data for each training speaker/environment that the variances on the posterior distributions are very small. Thus in training it is assumed that for speaker s in environment n $p(\mathbf{W}) \approx \delta(\mathbf{W} - \mathbf{W}^{(s)})$ and $p(\boldsymbol{\lambda}) \approx \delta(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(n)})$. This allows equation 1 to be directly used in training

There is an assumption of independence between the rows of the transforms. This is consistent with the standard diagonal covariance matrix assumption used in the HMM canonical model.

When using multiple components for the prior transform distribution there are a couple of ways of using the prior distribution. First each of the components of the prior distribution may be separately factored-in. Using the approximate factoring-in approach described in the next section this results in the number of components per state in the resultant system being PM. Alternatively the number of components in the resultant may be restricted to be the same as the original model M. For more details of schemes to do this see [6]. For the work in this paper each prior/canonical model component pairing is kept distinct. Thus there is an additional computational cost when the prior distribution uses more than a single component.

2.3. Factoring-In

Equation 1 gives the expression for the likelihood of an observation sequence. Unfortunately there is no simple way of evaluating this expression during recognition. Note that this problem was ignored during the training process by assuming that there was sufficient training data so that the posterior distributions for the speaker, s, and noise, n, transforms may be approximated as $p(\mathbf{W}) \approx \delta(\mathbf{W} - \mathbf{W}^{(s)})$ and $p(\boldsymbol{\lambda}) \approx \delta(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(n)})$. This is not possible if a distribution over the transform parameters is required to be factored-in.

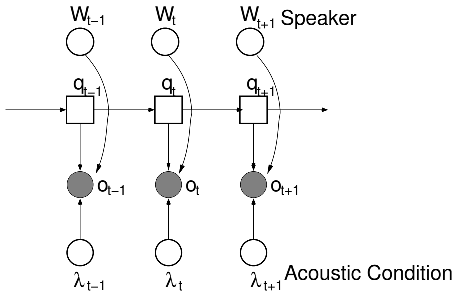

Fig. 2. Approximate DBN for factoring-in

To solve this problem an approximate factoring-in approach is used. The DBN for this approximate factoring-in is shown in figure 2. The speaker and acoustic environment are allowed to change every frame. This in many ways is similar to the standard HMM training scheme which places no constraints on the speaker or acoustic environment. Using this approximation it is possible to obtain likelihoods of producing an observation sequence.

$$p(\mathbf{o}_1, \ldots \mathbf{o}_T) \approx \sum_{\{\mathbf{Q}_T\}} \left(\prod_{t=1}^{T} \tilde{p}(\mathbf{o}_t|q_t) \right) P(\mathbf{Q}) \qquad (6)$$

For the case of factoring in the speaker transform we need to estimate (assuming a P-component prior)

$$\tilde{p}(\mathbf{o}|q) = \sum_{m=1}^{M} \sum_{p=1}^{P} \tilde{c}^{(qmp)} \mathcal{N}(\mathbf{o}; \tilde{\boldsymbol{\mu}}^{(qmp)}, \tilde{\boldsymbol{\Sigma}}^{(qmp)}) \qquad (7)$$

where

$$\tilde{\boldsymbol{\mu}}^{(qmp)} = \int_{\mathcal{R}} \mathbf{o} p(\mathbf{o}|q, m, \mathbf{W}, \boldsymbol{\lambda}) p(\mathbf{W}|p) p(\boldsymbol{\lambda}) d\mathbf{W} d\boldsymbol{\lambda} \qquad (8)$$

and similarly for the covariance matrices. Here the position in the noise eigenspace of the target domain is assume to be accurately estimated, $p(\boldsymbol{\lambda}) \approx \delta(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(n)})$. It can be shown that

$$\tilde{\mu}_d^{(qmp)} = \boldsymbol{\mu}_d^{(p)T} \hat{\boldsymbol{\xi}}^{(qm)} \qquad (9)$$

$$\tilde{\sigma}_d^{(qmp)2} = \hat{\boldsymbol{\xi}}^{(qm)T} \boldsymbol{\Sigma}_d^{(p)} \hat{\boldsymbol{\xi}}^{(qm)} + \sigma_d^{(qm)2} \qquad (10)$$

where $\hat{\boldsymbol{\xi}}^{(qm)} = \left[\begin{array}{cc} \boldsymbol{\lambda}^{(n)T} \mathbf{M}^{(qm)T} & 1 \end{array} \right]^T$ is the standard extended mean vector. The component prior is $\tilde{c}^{(qmp)} = c^{(qm)} c^{(p)}$.

3. RESULTS

The baseline system used for the recognition task was a gender-independent cross-word-triphone mixture-Gaussian tied-state HMM system. This was the same as the "HMM-1" model set used in the HTK 1994 ARPA evaluation system [8]. The number of components per state was 12 for the speech state and 24 for the "silence" states. For more details of the baseline system see [8]. The task considered in this section is how to generate a speaker-independent system in a target acoustic environment using limited target specific data.

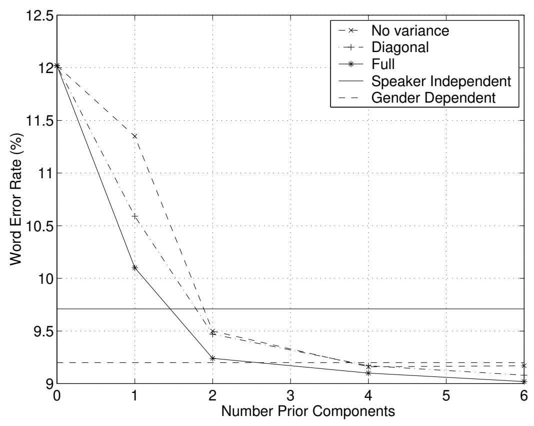

Fig. 3. "Clean" recognition against prior complexity

A SAT [2] version of of the baseline clean (only a single factor, the speaker, is assumed to affect the acoustic signal) WSJ system was built. The 284 training speaker transforms were used to estimate various complexity prior distributions. Figure 3 shows the variation of average performance on the 1994 development and evaluation data for three covariance matrix structures and number of Gaussian components. The three covariance forms were *No variance* where $\boldsymbol{\Sigma}_d^{(n)} = \mathbf{0}$, *Diagonal* where the off-diagonal terms of $\boldsymbol{\Sigma}_d^{(n)}$ were set to zero, and *Full* where $\boldsymbol{\Sigma}_d^{(n)}$ was a full covariance matrix. Zero prior components indicate that an identity matrix was used as the initial transform. Using the identity matrix degrades the recognition performance by around 25%[2]. Using the full covariance structure with at least 2-components gave good performance, yielding approximately the same performance as a gender dependent system. This indicates that the approximate factoring-in approach described yields reasonable performance.

[2]This result disagress with results produced by BBN. Similar degradation in performance was observed on experiments on the Switchboard database

For the second set of experiments noise was artificially added on to the clean data[3]. The "operations room" and the "car" noises from the NOISEX-92 database were used. Two conditions were then selected for training, the original training database, the clean training data, and operations room noise added at 0.05, the noisy training data. Using this data a series of systems were constructed. First a *clean* system was built using the original training data. Using both the clean and noisy training data a multi-environment system, *Mult*, was constructed. A baseline noise eigenspace, *Proj*, was built using CAT [3] with the clean and noisy training data. The weights used to build this eigenspace were fixed, since an artificial database was being used and the eigenspace was only meant to represent noise variability. The weights were fixed at $\begin{bmatrix} 1 & 0 \end{bmatrix}^T$ for the clean data and $\begin{bmatrix} 0 & 1 \end{bmatrix}^T$ for the noisy data. Finally an acoustic factorised system was built. Again the position in the noise eigenspace was fixed at $\begin{bmatrix} 1 & 0 \end{bmatrix}^T$ for the clean data and $\begin{bmatrix} 0 & a \end{bmatrix}^T$ for the noisy data (a was selected using equation 4). Using these fixed values for $\boldsymbol{\lambda}$ models were constructed by iteratively estimating the speaker transform and then the model parameters. Two forms of speaker transform priors were then used, a 2-component prior, *Fact(2)*, and a 4-component prior, *Fact(4)*. As an additional baseline a single-pass-retrained system [1], *SPR*, was built starting with the clean system for all the test conditions. These models were used to generate the lattices that were rescored in all the experiments performed.

System	Operations Room			Car
	Clean	0.025	0.05	0.025
Clean	9.34	13.77	23.89	20.99
SPR	—	11.79	16.76	12.11
Mult	9.34	12.17	17.45	17.76
Mult+MLLR	9.42	12.51	18.41	12.76
Proj	9.84	12.05	17.31	17.24
Fact(2)	9.18	11.84	17.11	16.45
Fact(4)	9.24	11.86	17.18	15.46

Table 1. Baseline system and "adapted" using 2 speakers error rates (%) on the 1994 H1 Development Data

Table 1 shows the baseline performance of the standard systems. As expected the performance of the *Clean* system degrades as the noise level increases. The *SPR* system's performance degrades more slowly. The *Mult* system shows similar performance to the *SPR* systems on the Operations Room noise, but is significantly worse on the Car noise. This is not surprising since the training data was the Operations Room noise data scaled at 0.05.

Table 1 also shows the performance of target domain adapted systems. The target domain data consisted of the supervised adaptation data of the first two test speakers, one male (*4q0*) and one female (*4q1*). The first system, *Mult+MLLR*, used a block-diagonal MLLR transform of the *Mult* system to the target domain data. This adaptation degraded the performance on the Clean and Operations Room noise tasks. However on the Car noise data the performance was reduced from 17,76% to 12.76% error rate. This was not surprising since where the *Mult* system was performing well the adaptation simply tuned to the 2 target domain speakers and degraded the other speakers performance. Where there was a large mismatch, the Car noise, the adaptation both tuned the model set to the noise and target speakers. The gains from tuning to the noise condition offset the degradation from tuning to specific speakers. This "adapted" performance is still about 5% relative worse than the *SPR* performance. The performance of the simple noise eigenspace *Proj*, is also given. Other than for the clean environment this shows slight improvement over the *Mult* system. The performance of acoustic factorisation is also shown. For both the 2-component and 4-component transform prior distributions slight gains were obtained over the *Mult* and *Proj* systems. Again the performance was far worse in the non matched condition, where the noise eigenspace was not matched to the target domain, *Car*.

[3]Since the WSJ0 and WSJ1 databases were recorded at different levels, the WSJ0 databases was attenuated so as to be at approximately the same level as the WSJ1 data (and the test data).In this paper scaling of the noise sources have been given to allow the database to be exactly reconstructed.

4. CONCLUSIONS

This paper has described the basic concepts behind and possible uses for acoustic factorisation. Here each of the unwanted factors affecting the acoustic signal is separately modelled. This allows appropriate factors for a particular target domain to factored-in as required. The factors whose variations are not required for the target domain are estimated for the target domain and fixed. A particular implementation of acoustic factorisation, using CAT and MLLR, is described in detail. Re-estimation formulae for both the factored transform and canonical model estimation are given. Simple initial experiments indicate that this is a useful research direction. Future work will examine real inhomogenous databases and alternative speaker and acoustic environment transforms.

5. REFERENCES

[1] M J F Gales, *Model-Based Techniques for Noise Robust Speech Recognition*, Ph.D. thesis, Cambridge University, 1995.

[2] T Anastasakos, J McDonough, R Schwartz, and J Makhoul, "A compact model for speaker-adaptive training," in *Proceedings ICSLP*, 1996, pp. 1137–1140.

[3] M J F Gales, "Cluster adaptive training for speech recognition," in *Proceedings ICSLP*, 1998, pp. 1783–1786.

[4] M J F Gales, "Multiple-cluster adaptive training schemes for speech recognition," in *Proceedings ICASSP*, 2001, pp. 233–236.

[5] C J Leggetter and P C Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, vol. 9, pp. 171–186, 1995.

[6] M J F Gales, "Acoustic factorisation: Theory and initial evaluation," Tech. Rep. CUED/F-INFENG/TR419, Cambridge University, 2001, Available from: svr-www.eng.cam.ac.uk/~mjfg.

[7] W Chou, "Maximum a-posterior linear regression with elliptical symmetic matrix variate priors," in *Proceedings ICASSP*, 1999, pp. 1–4.

[8] P C Woodland, J J Odell, V Valtchev, and S J Young, "The development of the 1994 HTK large vocabulary speech recognition system," in *Proceedings ARPA Workshop on Spoken Language Systems Technology*, 1995, pp. 104–109.