CONSTRUCTION OF MODEL-SPACE CONSTRAINTS

Patrick Nguyen[1,2], *Luca Rigazio*[1], *Christian Wellekens*[2] *and Jean-Claude Junqua*[1]

[1] Panasonic Speech Technology Laboratory
Santa Barbara, U.S.A
{nguyen, rigazio, jcj}@research.panasonic.com

[2] Institut Eurécom
Sophia-Antipolis, France
welleken@eurecom.fr

ABSTRACT

HMM systems exhibit a large amount of redundancy. To this end, a technique called Eigenvoices was found to be very effective for speaker adaptation. The correlation between HMM parameters is exploited via a linear constraint called eigenspace. This constraint is obtained through a PCA analysis of the training speakers.

In this paper, we show how PCA can be linked to the maximum-likelihood criterion. Then, we extend the method to LDA transformations and piecewise linear constraints. On the Wall Street Journal (WSJ) dictation task, we obtain 1.7% WER improvement (15% relative) when using self-adaptation.

1. OPTIMAL ESTIMATION OF THE EIGENSPACE

In this section, we show that the expected log-likelihood of the data is related to a sum of squared euclidean distances in the model space. This justifies using the SVD to compute the eigenspace.

First, we will show that the log-likelihood of rows of MLLR matrices defines a quadratic form. Then, we define proper normalization to reduce the ML problem to a standard least-squares problem, that can be solved by SVD.

1.1. Gaussianity of MLLR rows

Speaker dependent models are needed to build the eigenspace. However, for large vocabulary applications, building these models is difficult because of data sparsity and memory requirements. In practice, most systems use MLLR-adapted models [1]. MLLR transforms model means μ_m by a matrix $W = [w_1, ..., w_N]^T$:

$$\hat{\mu}_m = W\xi_m = \begin{bmatrix} w_1^T \\ \vdots \\ w_N^T \end{bmatrix} \begin{bmatrix} 1 \\ \mu_m \end{bmatrix}. \tag{1}$$

The feature space has dimension N. Each row w_k has dimension $N+1$.

We are concerned with the adaptation of mean vectors, with diagonal covariance matrices. The expected log-likelihood after E-step of the Baum-Welch algorithm is

$$Q = -\frac{1}{2}\sum_{t,m} \gamma_m(t)(\mu_m - o_t)^T C_m^{-1}(\mu_m - o_t) + C, \tag{2}$$

where C is a constant independent of the transformation. The index m refers to a Gaussian distribution. Without loss of generality, we only explore the case of a global transformation matrix. By hypothesis C_m^{-1} is a diagonal matrix with elements r_k. The ML estimate [2] for the MLLR row y_k has precision G_k:

$$y_k = G_k^{-1} z_k, \tag{3}$$

$$z_k = \sum_{t,m} \gamma_m(t) r_k o_k^{(t)} \xi_m, \tag{4}$$

$$G_k = \sum_{t,m} \gamma_m(t) r_k \xi_m \xi_m^T. \tag{5}$$

Rearranging the terms of eq(2) as in [3], we obtain:

$$Q = -\frac{1}{2}\sum_{k} (w_k - y_k)^T G_k (w_k - y_k) + C', \tag{6}$$

where C' completes the quadratic form. The sum is over all rows k of the transformation matrix. Eq. (6) states that MLLR rows are Gaussian with mean y_k and precision G_k.

1.2. Eigenvoices with MLLR-adapted models

To be effective in fast speaker adaptation, we choose to reduce the dimensionality of the problem [4]. We define the set of speaker transformation parameters by stacking all rows to form a supervector w:

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_N \end{bmatrix}. \tag{7}$$

The dimension of the supervector is $N(N+1)$. We postulate that speaker supervectors w lie in a low-dimensional space of dimension $E < N(N+1)$. We stack ML estimates of rows y_k to form the supervector y, and we approximate it by:

$$w \approx P^T P y, \tag{8}$$

where P is a projection matrix of dimension $E \times N(N+1)$. The matrix P is called the eigenspace and is estimated as follows. We observe a collection of T training speakers.They form an observation matrix $Y = [y^{(1)} ... y^{(T)}]$. Then we choose P to be the E first eigenvectors of the matrix YY^T. This will minimize the squared error of the approximation:

$$\hat{P} = \arg\max_P \left\{ \varepsilon = \mathrm{tr}(PYY^T P^T) \right\}. \tag{9}$$

Unfortunately, this is not guaranteed to maximize the likelihood. We propose a normalization that ensures optimality of the dimensionality reduction under the maximum likelihood criterion.

1.3. Root modulation

The quadratic form corroborates the fact that the ML row estimates y_k are Gaussian. Maximizing Q can also be seen as minimizing a distortion of observations with covariances G_k^{-1}:

$$\max_w Q = \min_w \sum_k (w_k - y_k)^T G_k (w_k - y_k). \tag{10}$$

The y_k are the ML estimates for the row k. We want to optimize the supervectors $w = [w_1, ... w_N]$ subject to the eigenspace constraint. By assuming the precision matrix to be constant after the E-step, we can modulate the variables by $G_k^{\frac{1}{2}}$:

$$\tilde{w}_k = G_k^{\frac{1}{2}} w_k, \tag{11}$$

$$\tilde{y}_k = G_k^{-\frac{T}{2}} z_k. \tag{12}$$

We may choose $G_k^{\frac{1}{2}}$ to be symmetric. The likelihood becomes:

$$Q = -\frac{1}{2} \sum_k (\tilde{w}_k - \tilde{y}_k)^T (\tilde{w}_k - \tilde{y}_k) + C'. \tag{13}$$

We call this normalization the *root modulation* because the observation z_k is multiplied by the square root of the precision.

1.4. ML dimensionality reduction

Now it becomes clear that the Q function is related to the least-squares problem in the root modulated space. If $\tilde{w}^{(q)}$ corresponds to the modulated estimate of the speaker q, the observation matrix of T speakers is $\tilde{Y} = [\tilde{y}^{(1)} ... \tilde{y}^{(T)}]$. We maximize the likelihood by maximizing the correlation between speakers:

$$\max_P Q = \max_P \left\{ Q = \mathrm{tr}(P\tilde{Y}\tilde{Y}^T P^T) \right\}. \tag{14}$$

Also, we may shift by the bias to get the covariance rather than the correlation. The optimal P is well-known to be a truncation of the eigenvectors decomposition of $\tilde{Y}\tilde{Y}^T$. Thus, the dimensionality reduction step is optimal with respect to the likelihood in the root modulated space. The precision matrix G_k is proportional to the number of frames $\sum_t \gamma_m(t)$. Therefore, our estimate is robust to uneven distribution of data. However, we assume the existence of the inverse of G_k, which does not exist when classes are not seen. Unseen classes are tied with the closest seen class.

This is to be contrasted with the original eigenvoice approach, which reduces the dimensionality based on $y^{(q)}$. The criterion function was distinct from the likelihood:

$$\varepsilon = \sum_q (w-y)^T (w-y) \neq Q = \sum_q (w-y)^T G (w-y), \tag{15}$$

and thus suboptimal under ML. We will call this method the *inverse space transformation*.

1.5. Estimation of the speaker model in the root space

Once we have obtained the optimal eigenspace, we can estimate the constrained MLLR transformation corresponding to a speaker. Let P_k be the matrix P corresponding to row k. The columns of P_k are the eigenvoices $p_j^{(k)}, j = 1..E$.

The models for root and inverse modulations constrain the transformation rows to be:

$$\text{Inverse space: } w_j = P_j \theta, \, \forall j, \tag{16}$$

$$\text{Root space: } w_j = G_j^{-\frac{1}{2}} P_j \theta, \, \forall j, \tag{17}$$

where θ is called the *eigenvalues decomposition*.

For the case of the inverse space transformation, the solution can be obtained by direct differentiation of Q in equation 2. This leads to an inefficient implementation. As noted in [5], one can follow the Markov chain of sufficient statistics

$$\{ \sum_{m,t} \gamma_m(t), \sum_{m,t} \gamma_m(t) o_t \} \rightarrow \{G_k, z_k\} \rightarrow \theta. \tag{18}$$

The inverse space and root space transformation have respectively:

$$w_j = P_j \left(\sum_k P_k^T G_k P_k \right)^{-1} \sum_k P_k^T z_k, \tag{19}$$

$$w_j = G_j^{-\frac{1}{2}} P_j \left(\sum_k P_k^T P_k \right)^{-1} \sum_k P_k^T G_k^{-\frac{T}{2}} z_k. \tag{20}$$

From these equations, the successive projections and mean-square estimation steps become apparent. In the root space, the inverse correlation may be computed offline.

1.6. Reestimation of the eigenspace

As with CAT [1] and MLES [6], we can reestimate the eigenspace in the Baum-Welch algorithm. If we reestimate the eigenspace the solution may not retain orthogonality of the eigenvectors. We embed the eigen decompositions of speaker location into the hidden data of the EM algorithm. The resulting optimal eigenspace is:

$$P_j = \left(\sum_q \theta\theta^T\right)^{-1} \sum_q G_j^{-\frac{T}{2}} z_j \theta^T, \tag{21}$$

where for each speaker q, we estimate the eigen decomposition

$$\theta = \left(\sum_k P_k^T P_k\right)^{-1} \sum_k P_k^T G_k^{-\frac{T}{2}} z_k. \tag{22}$$

The sufficient statistics are accumulators for P_j and the autocorrelation $\mathrm{E}_q \theta\theta^T$.

2. DISCRIMINATIVE PROJECTION

2.1. Objective functions

The supremacy of PCA schemes has been contested by discriminative projection. Among them, the most popular is LDA, which aims at maximizing the Fisher discriminant J:

$$\max_P J = \max_P \frac{|P^T S_B P|}{|P^T S_W P|} \Rightarrow P = \Lambda_E\left(S_W^{-1} S_B\right), \tag{23}$$

where $\Lambda_E(\cdot)$ are the E first eigenvectors of the matrix. The matrices S_B and S_W are called the between-class and with-class scatter matrices. Another popular discriminant objective function is the trace H:

$$\max_P H = \max_P \operatorname{tr}(S_B - \alpha S_W) \Rightarrow P = \Lambda_E^+(S_B - \alpha S_W), \tag{24}$$

where α is a tuning parameter. We take the most positive eigenvectors. In particular when $\alpha = 1$,

$$H = \mathrm{E}_O \log p(O|W)p(W) - \mathrm{E}_O \mathrm{E}_\chi \log p(O|\chi), \tag{25}$$

where W is the correct word sequence and χ is a competing word sequence. In our framework J is understood to be the log-likelihood ratio and H the cross-entropy.

2.2. Definition of Scatter Matrices

The most decisive choice left to the designer of an LDA system is the proper definition of classes. The choice of classes affects the homoscedastic assumption (S_W is global), the reliability of estimates, and fitness to the HMM classification design. We have many criteria to choose from: speaker adaptation gain, intra speaker acoustic variability, and linguistic variability. One has to distinguish between the speaker *regression* problem and the linguistic *classification* problem. LDA is best suited for classification, but may also be used for regression. We define three scatter matrices:

Inter-speaker variability: $S_B \equiv$ Between speakers
Intra-speaker regression: $S_I \equiv$ Average within speaker
Linguistic classification: $S_X \equiv$ Within speaker, using competing candidates

Our final objective is to perform speaker linear regression to minimize linguistic variability.

2.3. Experiments

To extract S_X, we decoded the training set with a unigram decoder. The decoder ran at about 2 times real-time. We retained only the first best solution. We weighted scatter matrices by unigram probabilities. For regression ($S_W = S_I$), we observed no enhancement. The intra-speaker variances S_I are measured by deviation from the true speaker model, on a sentence per sentence basis. The parameter α was set empirically. Best results were obtained for H and $S_W = S_X$ on Table 1.

3. PIECEWISE LINEAR DECOMPOSITION

Because of its simplicity and the presence of closed-form solutions, the linear assumption has proven very effective in many pattern regression problems. However, the linearity constraint has no legitimacy. In this section, we investigate a simple non-linear model. Our model is rooted on the equation

$$w = P_1\theta_1 + P_2(\theta_1)\theta_2. \tag{26}$$

We have a linear model involving θ_1 and P_1. Then, we set

$$P_2(\theta_1) = \begin{cases} P_2^+ & \text{if } \theta_1^T v > 0, \\ P_2^- & \text{elsewhere.} \end{cases} \tag{27}$$

The vector v is called the discriminant. The residual space is modelled by either P_2^+ or P_2^- according to the discriminant. The method is generalized to multiple discriminants by taking all possibilities of the signs, as shown on figure 1. For each region $\mathcal{R}_k$ we grow a different residual eigenspace. Not all dichotomies have a a populated intersection. For our experiments, we chose canonical $v_k = [0_{k-1}^T, 1, 0_{E_1-k}^T]^T$. For the particular case of v_1, it is equivalent to splitting according to the gender. The dimensionality of θ_1 is E_1. The vector 0_j is

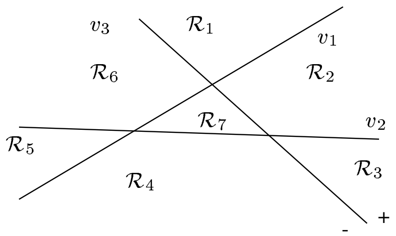

Fig. 1. Discriminants and regions

a zero vector of length j. The regions are the quadrants of the eigenspace.

The MLED location is a linear programming problem. The standard MLED formulae may be used. If the best point falls out of region, then the search resumes on the boundary region. In our experiments, such a case happened very seldom. It is possible to move the region assignment in the EM algorithm. We obtain a soft-weighting comparable to a multi-mixture eigenspace.

We reestimate of the eigenspace the same way we would optimize the linear eigenspace. We can also optimize the discriminative functions. Since the dimensionality of the parameter space is very high in comparison with the number of speakers, speaker points will always be linearly separable. In addition, since the optimal location almost always coincides with its ML region, discriminants are redundant and convergence of the discriminative functions is very quick. The perceptron algorithm [7] can be used to update the discriminant vectors v.

4. EXPERIMENTAL CONDITIONS

For our experiments we chose the Wall Street Journal (WSJ1) Nov92 evaluation test. The training database, called SI-284 consists of 37k sentences produced by 284 speakers. The acoustic frontend uses 39 MFCC coefficients and sentence-based cepstral mean subtraction (CMS). We train a total of 64k Gaussians with diagonal covariances, pooled in 1500 mixtures. The language model (LM) for this task is the standard trigram model provided by MIT. There are about 20k words for decoding.

Our recognizer, called EWAVES [8], is a lexical-tree based, gender-independent, word-internal context-dependent, one-pass trigram Viterbi decoder with bigram LM lookahead. The systems runs at about 3 times real-time, with a search effort of about 9k states (on a Pentium IV at 1.5 GHz).

For all experiments, we used an eigenspace of dimension $E = 40$. There was one full MLLR regression matrix for each of the following classes: silence, vowels, and consonants. For all experiments, we operated in self-adaptation mode: a first pass produces the most likely hypothesis. The second pass exploits adapted models. Five iterations of within-word Viterbi alignments are performed between passes. Table 1 summarizes the results for MLLR only (MLLR), eigenspace-constrained MLLR (MLED-MLLR), eigenvoice estimated on MLLR models with MAP smoothing (MLED-MAP/MLLR). Also, we report the piecewise linear extension applied on MLED-MLLR models in the inverse space, Root space and LDA space results. LDA space provided the best results.

	WER
SI	10.8%
MLLR	10.5%
MLED - MLLR	9.8%
MLED - MAP/MLLR	9.6%
Piecewise-linear	9.6%
Root space	9.5%
LDA space	9.1%

Table 1. Results

5. CONCLUSION

In this paper, we show how to perform the dimensionality reduction under the ML criterion. This is obtained by normalizing the ML speaker estimates by their corresponding precision. Then, we employ a linear discriminant approach to improve classification. Lastly, we relax the linearity constraint by introducing piecewise linear eigenspaces. Results attest the effectiveness of the approaches: the baseline WER is improved by 1.7% (15% relative).

6. REFERENCES

[1] M F J Gales, "Cluster adaptive training of hidden markov models," *IEEE Trans. on SAP*, vol. 8, pp. 417–418, 2000.

[2] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaption of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[3] M. Bacchian, "Using maximum likelihood linear regression for segment clustering and speaker identification," in *Proc. of ICSLP*, Beijing, China, Oct. 2000, vol. 4, pp. 536–539.

[4] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid Speaker Adaptation in Eigenvoice Space," *IEEE Trans. on SAP*, vol. 8, no. 6, pp. 695–707, Nov. 2000.

[5] N. Wang, S. Lee, F. Seide, and L. Lee, "Rapid speaker adaptation using *a priori* knowledge by eigenspace analysis of MLLR parameters," in *Proc. of ICASSP*, 2001, vol. I, pp. 317–320.

[6] P. Nguyen and C. Wellekens, "Maximum likelihood Eigenspace and MLLR for speech recognition in noisy environments," in *Proc. of Eurospeech*, Sep. 1999, vol. 6, pp. 2519–2522.

[7] R. O. Duda and P. B. Hart, *Pattern Classification and Scene Analysis*, Wiley, 1973.

[8] P. Nguyen, L. Rigazio, and J.-C. Junqua, "EWAVES: an efficient decoding algorithm for lexical tree based speech recognition," in *Proc. of ICSLP*, Beijing, China, Oct. 2000, vol. 4, pp. 286–289.