

PIECEWISE LINEAR CONSTRAINTS FOR MODEL SPACE ADAPTATION

Patrick Nguyen^{1,2}, Luca Rigazio¹, Jean-Claude Junqua¹ and Christian Wellekens²

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

Setting linear constraints on HMM model space appears to be very effective for speaker adaptation. In doing so, we assume that model parameters are jointly Gaussian. While this approach has proven reasonably successful, we question its accuracy in the case of very high dimensionality parameter spaces.

To address this problem, we employ a hierarchical piecewise linear model. Gross speaker variations are modeled with a linear eigenspace, subsuming the joint Gaussian model, and finer residues are modeled using another eigenspace chosen depending on the location of the first values. We perform experiments on Wall Street Journal (WSJ) dictation task, and we observe a cumulative 1.3% WER improvement (11% relative) when using self-adaptation.

1. EIGENVOICES WITH MLLR MODELS

Using the eigenvoices approach in combination with MLLR is not a new idea. In this section, we will briefly introduce the notation and fundamental equations used in the next sections.

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 $W = [w_1, \dots, w_N]^T$:

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

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

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

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 :

$$y_k := G_k^{-1} z_k, \quad (3)$$

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

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

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

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

where C' completes the quadratic form. The sum is over all rows k of the transformation matrix. In eq. (6) we interpret MLLR rows as 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 :

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_N \end{bmatrix}. \quad (7)$$

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:

$$w \approx P^T P y, \quad (8)$$

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 $Y = [y^{(1)} \dots y^{(T)}]$. Then we choose P to be the E first eigenvectors of the matrix $Y Y^T$. This will minimize the squared error of the approximation:

$$\hat{P} = \arg \max_P \{ \varepsilon = \text{tr}(P Y Y^T P^T) \}. \quad (9)$$

Unfortunately, this is not guaranteed to maximize the likelihood. In [5], we propose a normalization that ensures optimality of the dimensionality reduction under the maximum likelihood criterion.

1.2.1. Optimal estimators

Given this model, it is possible to find optimal estimators for the location of a speaker transformation in eigenspace. Let P_j be the matrix of rows of P associated with transformation matrix row j . Given the constraints of the eigenspace, the ML estimate for w_j is:

$$w_j = P_j \left(\sum_k P_k^T G_k P_k \right)^{-1} \sum_k P_k^T z_k. \quad (10)$$

Similarly, the optimal eigenspace may be found by considering the location of training speakers as a hidden variable. The eigen-decomposition is

$$\theta = \left(\sum_k P_k^T G_k P_k \right)^{-1} \sum_k P_k^T z_k. \quad (11)$$

The optimal estimator is given in [1]. We obtain

$$\text{super}(P_k) = \left(\sum_q \theta_q \theta_q^T \otimes G_k^{(q)} \right)^{-1} \sum_q \theta_q \otimes z_k^{(q)}, \quad (12)$$

where $\text{super}(\cdot)$ is the supervector formed by the matrix. Cheaper approaches are discussed in [1, 6].

2. PIECEWISE LINEAR DECOMPOSITION

We shall extend the model to linear piecewise models. Instead of estimating MLLR parameters using a single eigenspace, we approximate them instead using a collection of eigenspaces, each of which are linear within a certain range of eigenvalues.

We describe the new parametric form of model first, and then proceed to detail its implications on maximum-likelihood estimation of location (MLED), and eigenspace (MLES).

2.1. The model

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 relies on the equation

$$w := P_1 \theta_1 + P_2(\theta_1) \theta_2. \quad (13)$$

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

$$P_2(\theta_1) = \begin{cases} P_2^+ & \text{if } \theta_1^T v > 0, \\ P_2^- & \text{elsewhere.} \end{cases} \quad (14)$$

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. Spaces are organized hierarchically. Not all dichotomies have a populated intersection. For our experiments, we

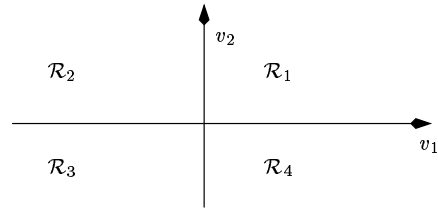


Fig. 1. Discriminants and regions

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

2.2. Estimation of parameters

As with the standard eigenvoices, we are confronted to the estimation of three kinds of parameters:

1. the initial eigenspaces and topology,
2. the eigenspaces in the Baum-Welch retraining,
3. the location of a speaker in the eigenspace.

The first item is the extension of PCA. The second one represents speaker adaptive training. They both have to do with the estimation of hyperparameters. The third one is the actual adaptation process whereby SI models are altered. For the logic of exposition, we answer these questions in reverse order.

2.2.1. Optimal location

The MLED location is a linear programming problem. The standard MLED formula in eq.(10) may be used. If the best point falls out of region, then the search resumes on the boundary region. We optimize the likelihood subject to constraint directly:

$$\theta_1, \theta_2 = \arg \max_{\theta_1, \theta_2} p(O|\theta_1, \theta_2) \quad (15)$$

It is possible to move the region assignment in the EM algorithm. We obtain a soft-weighting comparable to a multi-mixture eigenspace. We only consider the concatenated vector $\theta = [\theta_1^T \theta_2^T]^T$. For all available eigenspaces, we compute $\hat{\theta}^+ = \arg \max_{\theta} p(O|\theta)$ and similarly for $\hat{\theta}^-$. The resulting combination is

$$\theta = \hat{\theta}^+ p(O|\theta^+) + \hat{\theta}^- p(O|\theta^-) \quad (16)$$

The third, and fastest possibility which we have used in our experiments, is to calculate the first part of the eigenlocation θ_1 , find the corresponding, eigenspace, and then θ_2 :

$$\hat{\theta}_1 = \arg \max_{\theta_1} p(O|\theta_1; \theta_2 = 0); \theta_2 = \arg \max_{\theta_2} p(O|\theta_2, \hat{\theta}_1) \quad (17)$$

This may be suboptimal but breaks the complexity into two small MLED problems of eq.(10).

2.2.2. Reestimation of eigenspace parameters

Once eigenvalues and their corresponding associated eigenspaces are discovered, we reestimate the eigenspace the same way we would optimize the linear eigenspace using eq.(12).

2.2.3. Discriminant functions: The Perceptron

We can also optimize the discriminative functions. The perceptron algorithm [7] can be used to update the discriminant vectors v . Suppose we want to find discriminant functions for an arbitrary dichotomy of the set. For instance, in the Wall Street Journal dictation task, the training set comprises data from two databases WSJ0 (or SI84), and WSJ1 (SI200), recorded in two different occasions. To fix ideas, assume that we would like to separate the database component explicitly. It does not appear to be associated to a particular eigenvector. However, we premise that the impact on recognition will be large. To train a sub-eigenspace per database, consider the following problem. Let θ be a speaker base eigenvector. We would like to obtain

$$v^T \theta > 0 \quad \text{if speaker is in WSJ0, and} \quad (18)$$

$$v^T \theta < 0 \quad \text{if speaker is in WSJ1.} \quad (19)$$

If we switch the sign of all θ corresponding to WSJ1 data, we are left with the problem of solving the inequalities with respect to v :

$$v^T \theta > 0, \quad \forall \theta. \quad (20)$$

If the system has a solution, it is called linearly separable. Among all 2^T possible dichotomies, there are only

$$2 \sum_{k=0}^E \binom{T-1}{k}; \quad T > E \quad (21)$$

which are linearly separable. For $E = 20$ and $T = 284$, this amounts to about 17% of all possible dichotomies. The system of inequalities is solved by defining first the optimization criterion

$$J(v) := \sum_{\theta \in \Omega} v^T \theta; \quad \Omega := \text{all misclassified.} \quad (22)$$

By descending the gradient we obtain the notorious *perceptron algorithm*, which at each iteration k computes the set of all misclassified θ as Ω_k and update the discriminant vector v_k using the learning rule:

$$v_{k+1} \leftarrow v_k + \sum_{\theta \in \Omega_k} \theta. \quad (23)$$

If v is a solution, then we will converge in at most K steps,

$$K = \frac{\max_j \|\theta_j\|^2 \|v\|^2}{(\min_j \theta_j^T v)^2} < \infty. \quad (24)$$

There are many extensions to this algorithm, in particular in the case of non-separability. In last resort, we can increase E .

The perceptron approach is very effective when we would like to specify some prior knowledge manually. It is also useful when we need to update discriminant functions when the eigenspaces are reestimated. Positive signs are enforced when the discriminant maps θ to the eigenspace with highest likelihood.

2.2.4. Regression Trees: Alternative to unsupervised clustering

The power of piecewise linear models is introduced by the dependency between eigenvalues and eigenspaces. One possibility, especially popular in mixture modeling [8], initiates the algorithm with unsupervised clustering. This leads to lack of genericity in cases where the amount of data is sparse.

The use of hierarchical binary dichotomies for clustering is a proven approach with well-known efficiency. It is called Classification and Regression Trees (CART). The algorithm uses a finite set of candidate discriminants. It splits each cluster into two sub-clusters, choosing the best

discriminant according to a goodness of fit function. We asserted gaussianity of samples and used entropy as the optimization function. Unfortunately, however, since the number of speakers is rather small ($T = 284$), trees must be rather small. Another limitation arises in the speaker adaptation task since there are only a few characteristics that are known (age, accent, etc).

As discriminant functions, we used the quadrant functions. We discarded the database discriminant since all test data belong to WSJ0.

3. 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 [9], 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 1.7 times real-time each pass, with a search effort of about 9k states (on a Pentium IV at 1.5 GHz).

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). Best results for MLED-MLLR were obtained using $E = 40$. In piecewise linear functions, best results were obtained using $E_1 = 15$ dimensions as primary eigenspace and $E_2 = 15$ for residual eigenspaces. Surprisingly, only minor improvements were obtained by splitting gender (GD-MLED). No improvements were obtained using CART over simple discriminants.

	WER
SI	10.8%
MLED - MLLR	9.8%
GD - MLED	9.7%
Piecewise MLED	9.5%

Table 1. Results

4. CONCLUSION

In this paper, we have introduced a non-linear scheme for model space parameters. The models take the form of piecewise linear functions or mixture models. We assert that tying high energy coefficients of the SVD transformation allows for more robust processing. Training schemes may employ *a priori* knowledge to train dichotomies using variants of the perceptron algorithm. CART techniques were attempted as a clustering mechanism that aims at generality. The use of hard functions allows for a faster decoding. Training of eigenspaces use the EM algorithm and off-the-shelf techniques developed in [1] and [8]. We have experimented on the WSJ large-vocabulary dictation task. We observe improvements over the standard gender-dependent eigenvoices.

5. REFERENCES

- [1] M. J. F. 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 adaptation 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 Eigenspace," *IEEE Trans. on SAP*, vol. 8, no. 6, pp. 695–707, Nov. 2000.
- [5] P. Nguyen, L. Rigazio, C. Wellekens, and J.-C. Junqua, "Construction of model space constraints," in *Proc. of ASRU*, 2001, p. To appear.
- [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] M. E. Tipping and C. M. Bishop, "Mixtures of Probabilistic Principal Component Analysers," Tech. Rep., Neural Computing Research Group, Aston University, July 1998.
- [9] 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.
- [10] 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.