1. ABSTRACT

We have devised a new class of fast adaptation techniques for speech recognition, based on prior knowledge of speaker variation. To obtain this prior knowledge, one applies Principal Component Analysis (PCA) [9] or a similar technique to a training set of \( T \) vectors of dimension \( D \) derived from \( T \) speaker-dependent (SD) models. This offline step yields \( T \) basis vectors, which we call “eigenvoices” by analogy with the eigenfaces employed in face recognition [14,18]. We constrain the model for a new speaker \( \tilde{S} \) to be located in \( K \)-space, the space spanned by the first \( K \) eigenvoices. Speaker adaptation then involves estimating the \( K \) eigenvoice coefficients for the new speaker; typically, \( K \) is very small compared to the original dimension \( D \).

We conducted mean adaptation experiments on the Isolet database [2], using PCA to find the eigenvoices. In these experiments, \( D \) (number of Gaussian mean parameters) was 2808, \( T \) was 130, and \( K \) was set to several values between 1 and 20. With a large amount of supervised adaptation data, most eigenvoice techniques performed slightly better than MAP or MLLR; with small amounts of supervised adaptation data or for unsupervised adaptation, some eigenvoice techniques performed much better. For instance, when the supervised adaptation data was four letters pronounced once by the new speaker, the average relative reduction in error rate for an eigenvoice model with \( K = 5 \) was 26% (18.7% error in unit accuracy for SI baseline vs. 13.8% error for eigenvoice); MAP and MLLR showed no improvement. We believe that the eigenvoice approach would yield rapid adaptation for most speech recognition systems, including ones with a medium-sized or large vocabulary.

2. WHAT ARE EIGENVOICES?

“There are many examples of families of patterns for which it is possible to obtain a useful systematic characterization. Often, the initial motivation might be no more than the intuitive notion that the family is low dimensional, that is, in some sense, any given member might be represented by a small number of parameters. Possible candidates for such families of patterns are abundant both in nature and in the literature. Such examples include turbulent flows, human speech, and the subject of this correspondence, human faces” [10].

[10] introduced “eigenfaces” to researchers working on the representation and recognition of human faces. Previously, faces had been modeled with general-purpose image processing techniques. However, the true dimensionality of “face space” is much lower than its apparent dimensionality - outside the œuvre of Pablo Picasso, human faces differ from each other in minor ways. Since the publication of [10], face recognition researchers have applied dimensionality reduction techniques to training images of faces to characterize the space of variation between faces. Often, these researchers use PCA, which generates an orthogonal basis derived from the eigenvectors of the covariance or correlation matrix of the input data [9]. PCA guarantees that for the original data, the mean-square error introduced by truncating the expansion after the \( K \)-th eigenvector is minimized. The dimensionality reduction can be a factor of 50,000 or more [14,18]. However, other dimensionality reduction techniques can be used: e.g., linear discriminant analysis, singular value decomposition, or independent component analysis [3].
3. FINDING EIGENVOICE COEFFICIENTS

3.1. Projection

Let new speaker S be represented by a point P in K-space. We devised two techniques for estimating P from adaptation data. The projection estimator for P is similar to a technique commonly used in the eigenface literature. Let $e(1), \ldots, e(K)$ be the K eigenvectors; then $E = [e(1) \ldots e(K)]$ is a matrix of dimension $(D \times K)$. We now train an SD model on the adaptation data, from which we extract a supervector $V$ of dimension $D \times 1$ and project it into K-space to obtain $P$: $P = E \times E^T \times V$. It is now trivial to generate the adapted HMMs for S from P (if the D parameters in P represent only the Gaussian means, as for the experiments below, the remaining HMM parameters can be obtained from an SI model). The main flaw of the projection method is that for it to work well, all D parameters should be observed at least once in the adaptation data.

3.2. Max. Likelihood Eigen-Decomposition (MLED)

We now derive the maximum-likelihood MLED estimator for P in the case of Gaussian mean adaptation [15,16]. If m is a Gaussian in a mixture Gaussian output distribution for state s in a set of HMMs for a given speaker, let

\[
\begin{align*}
\alpha_k & \quad \text{be the number of features} \\
\omega_k & \quad \text{be feature vector (length n) at time t} \\
C_{m}^{(t)} & \quad \text{be inverse covariance for m in state s} \\
\mu_{m}^{(t)} & \quad \text{be adapted mean for mixture m of s} \\
\gamma_{m}^{(t)}(s) & \quad \text{be the L(m, k|\lambda, \omega_k) (s-m occupation prob.)}
\end{align*}
\]

To maximize the likelihood of observation $O = \omega_1 \ldots \omega_T$ w.r.t. $\lambda$, we iteratively maximize an auxiliary function $Q(\lambda, \lambda)$, where $\lambda$ is current model and $\lambda$ is estimated model [13]. We have

\[
Q(\lambda, \lambda) = -\frac{1}{2}P(O|\lambda) \times \sum_s \sum_m \sum_t \gamma_{m}^{(t)}(s) f(\alpha_k, s, m)
\]

where

\[
f(\alpha_k, s, m) = [n \log(2\pi) + \log[C_{m}^{(t)}] + h(\alpha_k, s, m)]
\]

and

\[
h(\alpha_k, s, m) = (\alpha_k - \mu_{m}^{(t)})^T C_{m}^{(t)}(\alpha_k - \mu_{m}^{(t)})
\]

Consider the eigenvector $e(j)$ with $j = 1 \ldots K$:

\[
e(j) = [e_{1}^{(t)}(j), e_{2}^{(t)}(j), \ldots, e_{m}^{(t)}(j), \ldots]^T
\]

where $e_{m}^{(t)}(j)$ represents the subvector of eigenvector $j$ corresponding to the mean vector of mixture Gaussian $m$ in state $s$. Then we need

\[
\mu = [\mu_{1}^{(t)}, \mu_{2}^{(t)}, \ldots, \mu_{m}^{(t)}]^T = \sum_{j=1}^{K} w(j) e(j)
\]

The $w(j)$ are the K coefficients of the eigenvoice model:

\[
\mu_{m}^{(t)} = \sum_{j=1}^{K} w(j) e_{m}^{(t)}(j)
\]

To maximize $Q(\lambda, \lambda)$, set $\frac{\partial Q}{\partial \mu_{m}^{(t)}} = 0, j = 1 \ldots K$; assuming the eigenvalues are independent, $\frac{\partial Q}{\partial \gamma_{m}^{(t)}} = 0, i \neq j$. We obtain

\[
\begin{align*}
\sum_s \sum_m \sum_t \gamma_{m}^{(t)}(s) \{ & \sum_{i=1}^{K} \omega_i (e_{i}^{(t)}(j)) C_{m}^{(t)}(e_{i}^{(t)}(j)) \} \\
& = \sum_s \sum_m \sum_t \gamma_{m}^{(t)}(s) \{ \sum_{i=1}^{K} \omega_i (e_{m}^{(t)}(k)) C_{m}^{(t)}(e_{m}^{(t)}(j)) \},
\end{align*}
\]

\[
j = 1 \ldots K
\]

Thus, we have K equations to solve for the K unknown $w(j)$ values. The computational cost of this online operation is quite reasonable - for instance, it is much “cheaper” than most implementations of MLLR. To reduce computational cost, one can choose a lower K (at the expense of accuracy). Note also that the adaptation experiments described below involved only one Gaussian per state s (so the K equations we solved for MLED estimation in the experiments were a special case of those just given).

4. EXPERIMENTS

4.1. Protocol and Results

We conducted mean adaptation experiments on the Isolet database [2], which contains 5 sets of 30 speakers, each pronouncing the alphabet twice. After downsampling to 8kHz, five splits of the data were done. Each split took 4 of the sets (120 speakers) as training data, and the remaining set (30 speakers) as test data; results given below are averaged over the five splits. Offline, we trained 120 SD models on the training data, and extracted a supervector from each. Each SD model contained one HMM per letter of the alphabet, with each HMM having six single-Gaussian output states. Each Gaussian involved eighteen “Perceptual linear predictive” (PLP) [7] cepstral features whose trajectories were filtered. Thus, each supervector contained $D = 26 \times 6 \times 18 = 2808$ parameters.

For each of the 30 test speakers, we drew adaptation data from the first repetition of the alphabet, and tested on the entire second repetition. SI models trained on the 120 training speakers yielded 81.3% word percent correct; SD models trained on the first repetition for each new speaker yielded 59.6%. We also tested three conventional mean adaptation techniques, using various subsets of the first alphabet repetition for each speaker as adaptation data. The three techniques (whose unit accuracy results are shown in Table 1) are MAP with SI priors (“MAP”), global MLLR with SI priors (“MLLR G”) and MAP with the MLLR G model as prior (“MLLR G => MAP”). For MAP techniques shown here and below, we set $\tau = 30$ (we verified that results were insensitive to changes in $\tau$).

Using the whole alphabet as adaptation data, we carried out both supervised and unsupervised adaptation experiments (first-pass SI recognition for unsupervised adaptation); the results are denoted as alph. sup. and alph. uns. in Table 1. The other experiments in Table 1 involve supervised adaptation employing
subsets of the alphabet as adaptation data. These include a balanced alphabet subset of size \(1/7\), balanced /= f CD F ... or low rate of change in vowel for-mants, while speakers with /= values show dramatic movement towards the off-glide.

<table>
<thead>
<tr>
<th>Ad. data</th>
<th>MAP</th>
<th>MLLR G</th>
<th>MLLR G =&gt; MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>alph. sup.</td>
<td>87.4</td>
<td>85.8</td>
<td>87.3</td>
</tr>
<tr>
<td>alph. uns.</td>
<td>77.8</td>
<td>81.5</td>
<td>78.5</td>
</tr>
<tr>
<td>bal-17</td>
<td>81.0</td>
<td>81.4</td>
<td>81.9</td>
</tr>
<tr>
<td>AEOW</td>
<td>79.7</td>
<td>14.4</td>
<td>15.4</td>
</tr>
<tr>
<td>ABCU</td>
<td>78.6</td>
<td>17.0</td>
<td>17.5</td>
</tr>
<tr>
<td>D (worst)</td>
<td>77.6</td>
<td>3.8</td>
<td>3.8</td>
</tr>
<tr>
<td>ave(1-let.)</td>
<td>80.0</td>
<td>3.8</td>
<td>3.8</td>
</tr>
<tr>
<td>A (best)</td>
<td>81.2</td>
<td>3.8</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Table 1: NON-EIGENVOICE ADAPTATION

To carry out eigenvoice experiments, we performed PCA on the \(T = 120\) supervectors (using the correlation matrix), and kept eigenvoices 0...K (0 is mean vector). First, we studied the effect of K and of estimation method. For these experiments, shown in Table 2, the whole alphabet was used as supervised adaptation data (alph. sup. data option). "PROJ.K" is eigenvoice model obtained by projection into K-space; "MLED.K" is the maximum-likelihood eigenvoice model in K-space, and "MLED.K => MAP" is MAP using MLED.K as the prior. Comparison with the alph. sup. row of Table 1 shows that MLED.K => MAP outperforms the non-eigenvoice techniques by a small amount.

<table>
<thead>
<tr>
<th>K</th>
<th>PROJ.K</th>
<th>MLED.K</th>
<th>MLED.K =&gt; MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>83.4</td>
<td>84.7</td>
<td>88.3</td>
</tr>
<tr>
<td>5</td>
<td>81.4</td>
<td>86.5</td>
<td>88.8</td>
</tr>
<tr>
<td>10</td>
<td>80.5</td>
<td>87.4</td>
<td>89.0</td>
</tr>
<tr>
<td>20</td>
<td>78.5</td>
<td>87.4</td>
<td>89.1</td>
</tr>
</tbody>
</table>

Table 2: EIGENVOICES: VARYING K (alph. sup.)

Table 3: EIGENVOICES: PARTIAL ALPHABET

<table>
<thead>
<tr>
<th>Ad. data</th>
<th>MLED.5, =&gt; MAP</th>
<th>MLED.10, =&gt; MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>alph. sup.</td>
<td>86.5, 88.8</td>
<td>87.4, 89.0</td>
</tr>
<tr>
<td>alph. uns.</td>
<td>86.3, 80.8</td>
<td>86.3, 81.4</td>
</tr>
<tr>
<td>bal-17</td>
<td>86.5, 86.0</td>
<td>87.0, 86.8</td>
</tr>
<tr>
<td>AEOW</td>
<td>86.2, 85.4</td>
<td>85.8, 85.3</td>
</tr>
<tr>
<td>ABCU</td>
<td>86.3, 85.2</td>
<td>86.4, 85.5</td>
</tr>
<tr>
<td>W (worst)</td>
<td>82.2, 81.8</td>
<td>79.9, 79.2</td>
</tr>
<tr>
<td>ave(1-let.)</td>
<td>84.4, 83.9</td>
<td>82.4, 81.8</td>
</tr>
<tr>
<td>V (best)</td>
<td>85.7, 85.7</td>
<td>83.2, 83.1</td>
</tr>
</tbody>
</table>

4.2. What Do the Eigenvoices Mean?

We tried to interpret the eigendimensions for one of the five splits in these experiments. Figure 1 shows how as more eigenvoices are added, more variation in the training speakers is accounted for. Eigenvoice 1 accounts for 18.4% of the variation; to account for 50% of the variation, we need the eigenvoices up to and including number 14.
5. DISCUSSION

Some other researchers share our belief that fast speaker adaptation can be achieved by quantifying inter-speaker variation. N. Str"om models speaker variation for adaptation in a hybrid AN/NHMM system by adding an extra layer of “speaker space units” [17]. There is one such unit per training speaker; when the system is being trained on speaker $i$, the activity of unit $i$ is set to 1 and all other activities are set to 0. Str"om found moderate improvement for the adapted system over the baseline for four or more words. Examination of the connections in the ANN indicated that male and female speakers form two separate clusters in speaker space ([17], Fig. 2).

After submission of this paper in April 1998, we became aware of some excellent research along similar lines, unpublished at that time. Hu et al [8] focus on vowel classification by Gaussian mixture classifiers, but their approach could be extended to cover all phonemes. PCA is performed on a set of training vectors consisting, for each speaker, of the concatenated mean feature vectors for vowels. Vowel data from the new speaker is projected onto the eigenvectors to estimate the new speaker’s deviation from the training speaker mean vector. Finally, classification is carried out either by subtracting the deviations from the new speaker’s acoustic data (speaker normalization) or by adjusting the Gaussian classifier means to reflect the deviation. This technique can be seen as a special case of the eigenvoice approach for mean adaptation. In this special case, only HMMs for vowels are employed, each HMM has a single state with a single Gaussian output distribution, and the projection technique is used to estimate the eigenvoice coordinates for the new speaker. Hu et al find significant improvements over an SI baseline if their adaptation approach is used, for both supervised and unsupervised adaptation. As it did in our experiments, the first coefficient in their experiments separates men and women (though it accounts for 93.8% of variation vs. only about 18% in our case).

In the small-vocabulary speaker adaptation experiments described in this paper, the eigenvoice approach reduced the degrees of freedom for speaker adaptation from $D = 2898$ to $K <= 20$ and yielded much better performance than other techniques for small amounts of adaptation data. These exciting results provide a strong motivation for testing the approach in medium- and large-vocabulary systems. We also plan to study the robustness of the approach to deterioration in the quantity or quality of the training data: e.g., fewer training speakers or less data per training speaker; mismatch between training and test environments, differences in dialect between training and test speakers. We will also experiment with discriminative training of the original SD models. Other important issues include training of mixture Gaussian SD models (for the resulting eigenvoices to be useful, Gaussian $i$ for phonetic unit $P$ in a given training SD model must mean the same thing as Gaussian $i$ for $P$ for another training speaker - how can this be ensured?) and the performance of eigenvoices found by dimensionality reduction techniques other than PCA. We hope to explore Bayesian versions of the approach: estimate the position $\lambda$ of the new speaker in $K$-space by maximizing $P(O|\lambda) \times P(\lambda)$ (MLED only maximizes the first term). Finally, we have begun to apply the eigenvoice approach to speaker verification and identification, with encouraging early results.

6. REFERENCES