Rapid Speaker Adaptation in Eigenvoice Space

Roland Kuhn, Jean-Claude Junqua, *Member, IEEE*, Patrick Nguyen, and Nancy Niedzielski

***Abstract*—This paper describes a new model-based speaker adaptation algorithm called the eigenvoice approach. The approach constrains the adapted model to be a linear combination of a small number of basis vectors obtained offline from a set of reference speakers, and thus greatly reduces the number of free parameters to be estimated from adaptation data. These "eigenvoice" basis vectors are orthogonal to each other and guaranteed to represent the most important components of variation between the reference speakers. Experimental results for a small-vocabulary task (letter recognition) given in the paper show that the approach yields major improvements in performance for tiny amounts of adaptation data. For instance, we obtained 16% relative improvement in error rate with one letter of supervised adaptation data, and 26% relative improvement with four letters of supervised adaptation data. After a comparison of the eigenvoice approach with other speaker adaptation algorithms, the paper concludes with a discussion of future work.**

***Index Terms*—Eigenvoice approach, principal component analysis, speaker adaptation, speaker clustering.**

I. Introduction

When a speaker-dependent (SD) system trained on speech from a given speaker S is tested on other speech data from S, the error rate may be as low as a half to a third that of a similar speaker-independent (SI) speech recognition system tested on the same data [18], [29]. The goal of research on speaker adaptation is to achieve performance on each new speaker approaching that of an SD system for that speaker, while avoiding the need for unacceptably large amounts of adaptation data for each new speaker. The meaning of "unacceptably large" depends on the application. Requiring the purchaser of a dictation system to train the system for 30 to 40 min may be acceptable, since he or she is planning to use the system for years to come. On the other hand, in many commercially attractive applications, such as ordering items over the telephone, one can only count on a few seconds of unsupervised speech.

This paper addresses the latter case, describing a new model-based rapid speaker adaptation algorithm called the *eigenvoice approach*. Model-based algorithms differ from other adaptation algorithms (such as speaker normalization [39]) in that they adapt to a new speaker by modifying the parameters of the system's speaker model. Standard model-based algorithms such as maximum *a posteriori* (MAP) adaptation [14], [15], [42] and maximum likelihood linear regression (MLLR) adaptation [12], [31], [32] require significant amounts of adaptation data from the new speaker in order to perform better than a similar SI system. However, in the last two years, model-based algorithms that achieve rapid speaker adaptation have been devised [13], [17], [18], [25]–[27]. These "speaker space" algorithms constrain the adapted model to be a linear combination of a small number of basis vectors obtained offline from a set of reference speakers, and thus greatly reduce the number of free parameters to be estimated from adaptation data. These algorithms are related to an older approach, speaker clustering [11], [24]. In their application of *a priori* constraints derived from reference speakers, they also resemble extended MAP (EMAP) [28], [38], which employs precomputed correlations between acoustic units to estimate unseen distributions, though the details are quite different. All these model-based speaker adaptation algorithms, and some others, are discussed in the paper. Unlike other speaker space algorithms, the eigenvoice approach finds basis vectors that are orthogonal to each other and guaranteed to represent the most important components of variation between the reference speakers. Furthermore, the approach allows the number of basis vectors employed during recognition (i.e., the number of degrees of freedom for the adapted model) to vary dynamically.

We give experimental results for the eigenvoice approach on a small-vocabulary task (letter recognition), showing that eigenvoice-based Gaussian mean adaptation can produce major improvements in performance for tiny amounts of adaptation data. For instance, we obtained 16% relative improvement in error rate with one letter of supervised adaptation data, and 26% relative improvement with four letters of supervised adaptation data. We look at what the eigenvoices tell us about inter-speaker variation. Finally, we outline future work, discussing how the approach could be extended to estimate other HMM parameters besides the Gaussian means, and how it could be modified for application in large-vocabulary systems.

II. Eigenvoice Approach

A. Eigenfaces

> "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. . ." ([23, pg. 103]).

Our work on eigenvoices was inspired by current research on face recognition. As the quotation above suggests, there are

Manuscript received May 17, 1999; revised April 14, 2000. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rafid A. Sukkar.

The authors are with the Panasonic Speech Technology Laboratory, Panasonic Technologies, Inc., Santa Barbara, CA 93105 USA (e-mail: kuhn@stl.research.panasonic.com).

Publisher Item Identifier S 1063-6676(00)09262-2.

hidden similarities between the study of faces and the study of voices; techniques applied in one area may be helpful in the other. Face recognition is the problem of trying to match a given two-dimensional (2-D) face image to a set of face images in a database. Initially, researchers applied general-purpose image processing techniques to this problem. However, building on the work of Kirby and Sirovich [23], they soon realized that the dimensionality of "face space"—the space of variation between photographs of human faces with the same orientation and scale lit in the same way—is much smaller than the dimensionality of a single face considered as an arbitrary 2-D image. As a useful approximation, one may consider an individual face image to be a linear combination of a small number of face components or "eigenfaces" derived from a set of reference face images. One converts each reference image into a vector of floating point numbers representing light intensity in each pixel, calculates the covariance or correlation matrix between these reference vectors, and then applies principal component analysis (PCA) [22] to find the eigenvectors of the matrix: the eigenfaces.

An important advantage of PCA is that the eigenvectors are ordered by the magnitude of their contribution to the variation between the reference images. In the face recognition literature, the vector obtained by averaging over all reference images is called "eigenface 0," and the other eigenfaces from "eigenface 1" onwards model variation from this average face. The expansion is truncated at some point, say after eigenface K. Faces are represented as eigenface 0 plus a linear combination of the remaining K eigenfaces; PCA guarantees that for the original set of data points, the mean-square error introduced by truncating the expansion after the Kth eigenvector is minimized. Incidentally, the eigenfaces are not themselves usually plausible faces, only directions of variation between faces.

To find the best match for an image of a person's face in a set of stored facial images, one may calculate the Euclidean distances between the vector of K coordinates representing the new face and each of the K-dimensional vectors representing the stored faces, and then choose the stored image yielding the smallest distance [40]. In these experiments, for which the training data (the collection of images used to calculate the eigenfaces) and the test data consisted of faces with the same orientation and scale lit in the same way, excellent results were obtained with about $K = 100$ eigenfaces. Recently, more sophisticated methods [6], some employing probability distributions over eigenface space [34], have been studied.

B. What are Eigenvoices?

The obvious parallel in speech research to face recognition is not speech recognition but speaker recognition (usually called "speaker identification and verification"). However, if speaker adaptation is viewed in a rather unfamiliar way, the applicability of PCA to this problem becomes apparent. Instead of thinking of speaker adaptation as a way of improving an initial SI model, we can think of it as finding a good SD model for the new speaker in the space of possible SD models.

The eigenvoice approach is summarized in Fig. 1. We begin with a reference set of R well-trained SD models, plus an SI model. From each of the SD models, we extract a "supervector" containing the parameters to be adapted. In the experiments described below, we extracted from the SD models the means of the HMM output Gaussians [25]–[27]. Parameters may be put into the supervectors in an arbitrary order, as long as the number D of extracted parameters and the order are the same for all speakers. That is, index i of the supervector from speaker 1 must represent the same parameter as does index i for any of the other $R-1$ speakers. Applied to the R supervectors, each of dimension D, a dimensionality reduction technique (DRT) yields R eigenvectors, each of dimension D. Eigenvector 0 is the mean supervector, and the rest are the principal components (ordered by the magnitude of their contribution to variation in the reference supervectors). We then throw away the higher-order eigenvectors, keeping only those numbered from 0 up to K (where $K < R \ll D$); we call those that remain the "eigenvoices."

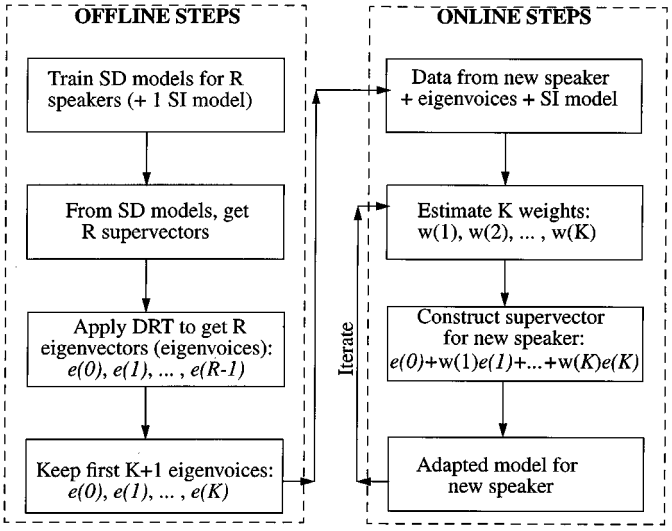

Fig. 1. Block diagram for eigenvoice speaker adaptation

For the experiments described here, the DRT employed was PCA, but we could have used instead independent component analysis (ICA), linear discriminant analysis (LDA), or even a DRT specifically designed to facilitate speaker adaptation (see Section VII-A). As summarized in Jolliffe's book on PCA, the standard reference on the topic [22], there are two variants of PCA: one in which the principal components are derived from the covariance matrix of the variables, and one in which they are derived from the correlation matrix. The latter is recommended in cases where the variables have different units of measurement or are of different types, to prevent variables with large absolute values from dominating the analysis [22, pp. 16–17]. Since acoustic features in HMM represent different physical quantities, the correlation variant of PCA was employed in all experiments described below.

As shown in Fig. 1, the computationally intensive SD training and PCA steps are carried out offline, before recognition begins. Adaptation to a new speaker is computationally cheap. As described in the next subsection, we assume that the supervector for the new speaker is a linear combination of the eigenvoices, of form $e(0)+w(1)*e1+\cdots+w(K)*e(K)$; the space spanned by the vector defined by this expression is called "K-space." Since K is much smaller than D, this is a very powerful constraint, and

only a small amount of adaptation data is required to estimate the $w(j)$ (with the help of the SI model—see below). Having constructed the supervector for the new speaker, we then copy the values in the supervector into the appropriate parameters of the adapted model for the new speaker. Parameters not represented in the supervector must be obtained from elsewhere—for instance, in our mean adaptation experiments, the variances and transition probabilities were obtained from the SI model. Finally, the $w(j)$ can be re-estimated (with the help of the adapted model).

The eigenvoice approach greatly reduces the number of parameters to be estimated for the new speaker. In the letter recognition experiments described below, the use of eigenvoices enabled us to transform a problem requiring calculation of 2808 parameters (estimation of the Gaussian means for the adapted model) into a problem requiring calculation of five or ten free parameters. The price paid for this simplification of the problem is the assumption that all new speakers the system will encounter are located in K-space, or so near K-space that it makes no practical difference. To make sure this assumption is realistic, the set of reference speakers should be as diverse as possible.

C. Estimating the Eigenvoice Coefficients

Let each new speaker S be represented by a point P in K-space

$$P = e(0) + w(1) * e1 + \cdots + w(K) * e(K). \tag{1}$$

The problem is to estimate the weights $w(j)$ from adaptation data. After some experimentation with a method based on projection we decided to use a maximum-likelihood estimator called maximum likelihood eigen-decomposition (MLED), which can be applied in the case of Gaussian mean adaptation in a CDHMM system [25]. For each observation in the adaptation data from the new speaker, one needs to estimate the occupation probability for each output Gaussian represented in the supervector for the new speaker. In the first iteration of MLED, these occupation probabilities are estimated from an SI model; in subsequent iterations, they are estimated from the adapted model. The covariance matrix for each distribution also comes from the SI model (unlike occupation probabilities, the covariance matrices are not re-estimated in subsequent iterations).

If m is a Gaussian in a mixture Gaussian output distribution for state s in a set of HMM's for a given speaker, first "normalize" each observation for m by subtracting the corresponding component of mean reference vector $e(0)$, *i.e.*, the part of $e(0)$ pertaining to m in s. Let

n	number of features;
$C_m^{(s)-1}$	inverse covariance for m in state s;
$\mathbf{o}_t$	normalized observation vector (length n) at time t;
$\mu_m^{(s)}$	adapted mean for mixture m of s;
$\gamma_m^{(s)}(t)$	$P(m, s\|\lambda, \mathbf{o}_t)$ (s-m occupation prob.).

To maximize the likelihood of observation $O = \mathbf{o}_1 \cdots \mathbf{o}_T$ with respect to λ, one iteratively maximizes an *auxiliary function* $Q(\lambda, \hat{\lambda})$, where λ is current model and $\hat{\lambda}$ is estimated model (as in [31]). Thus

$$Q(\lambda, \hat{\lambda}) = -\tfrac{1}{2} P(O|\lambda) \times \sum_s \sum_m \sum_t \gamma_m^{(s)}(t) f(\mathbf{o}_t, s, m) \tag{2}$$

where

$$f(\mathbf{o}_t, s, m) = [-n \log(2\pi) - \log |C_m^{(s)}| + h(\mathbf{o}_t, s, m)] \tag{3}$$

and

$$h(\mathbf{o}_t, s, m) = \left(\mu_m^{(s)} - \mathbf{o}_t\right)^T C_m^{(s)-1} \left(\mu_m^{(s)} - \mathbf{o}_t\right). \tag{4}$$

Consider the eigenvoice vectors $e(j)$ with $j = 1 \cdots K$

$$e(j) = \begin{bmatrix} e_1^{(1)}(j) \\ e_2^{(1)}(j) \\ \vdots \\ e_m^{(s)}(j) \\ \vdots \end{bmatrix} \tag{5}$$

where $e_m^{(s)}(j)$ represents the subvector of eigenvoice j corresponding to the mean vector of mixture Gaussian m in state s. The Gaussian mean estimates are

$$\hat{\mu} = \begin{bmatrix} \hat{\mu}_1^{(1)} \\ \hat{\mu}_2^{(1)} \\ \vdots \\ \hat{\mu}_m^{(s)} \\ \vdots \end{bmatrix} = \sum_{j=1}^{K} w(j) e(j) \tag{6}$$

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

$$\mu_m^{(s)} = \sum_{j=1}^{K} w(j) e_m^{(s)}(j). \tag{7}$$

To maximize $Q(\lambda, \hat{\lambda})$, set $(\partial Q/\partial w(j)) = 0$, $j = 1 \cdots K$; assuming the eigenvalues are independent, $(\partial w(i)/\partial w(j)) = 0$, $i \neq j$. One obtains for $j = 1 \cdots K$

$$\begin{aligned} &\sum_s \sum_m \sum_t \gamma_m^{(s)}(t) \left(e_m^{(s)}(j)\right)^T C_m^{(s)-1} \mathbf{o}_t \\ &\quad = \sum_s \sum_m \sum_t \gamma_m^{(s)}(t) \\ &\qquad \times \left\{ \sum_{k=1}^{K} w(k) \left(e_m^{(s)}(k)\right)^T C_m^{(s)-1} e_m^{(s)}(j) \right\}. \end{aligned} \tag{8}$$

Thus, there are K equations to solve for the K unknown weights ($w(j)$ values). The new model thus obtained yields new values for the occupation probabilities $\gamma_m^{(s)}(t)$; this estimation process can be iterated until the weights $w(j)$ converge.

III. Related Work on Speaker Adaptation

A. Introduction

The eigenvoice approach is quite different from two frequently-used model-based algorithms for speaker adaptation,

MAP estimation and MLLR. MAP estimation applied to continuous-density HMMs is described in [14], [15]. Unlike the eigenvoice approach, standard MAP only updates the parameters of Gaussians that have observations $\mathbf{o}_t$ assigned to them, and thus converges slowly in a system with many output Gaussians. For instance, when MAP adaptation was applied to a recognizer for the 991-word RM task, 94.4% of the Gaussian means were completely unaffected after one adaptation sentence; after 40 adaptation sentences, 35.7% of the Gaussian means still had not changed from the original *a priori* values ([2, Table I]).

Like the eigenvoice approach, MLLR updates the parameters of all Gaussians, whether observed or unobserved [10], [31], [32]. However, standard MLLR adaptation is much less constrained by prior knowledge, other than that contained in the regression class definitions and in the initial speaker-independent model on which the transformations are performed. By contrast, the eigenvoice approach puts a heavy emphasis on prior knowledge about systematic patterns of variation between the reference speakers. The number of degrees of freedom during global MLLR adaptation is roughly equivalent to the number of parameters in the global transformation matrix, while the number of degrees of freedom during eigenvoice adaptation is equivalent to the number of eigenvoices. Thus, MLLR adaptation has far more degrees of freedom (in the experiments reported below, global MLLR had about 500 degrees of freedom, and the eigenvoice approach 1, 5, or 10 degrees).

The techniques most closely resembling the eigenvoice approach are those which update parameters for both observed and unobserved Gaussians during adaptation, while applying strong constraints derived from prior knowledge. These include the EMAP and RMP variants of MAP, hard and soft speaker clustering, and reference speaker weighting. A technique developed by Hu *et al.*for vowel classification is also relevant.

B. EMAP and RMP

A variant of MAP called "extended maximum *a posteriori* estimation" (EMAP) attempts to achieve faster convergence by looking at correlations between the acoustic units employed by a recognizer (e.g., phonemes or words). Use of these correlations by EMAP enables an observation of a speech unit U from the new speaker to modify not only the model for that unit, but also models for unseen units U' that are correlated with U [28], [37], [38]. Cox extended this technique to utilize both pairwise and multiple correlations in an alphabet recognition task [9]. With three utterances of 13 "enrollment" letters as adaptation data for each speaker, he obtained a drop in the error rate on 13 test letters from 17.0% for the SI model to 3.0% for the adapted models. However, the experimental protocol assumed that one knows in advance the source (enrollment) and target (test) distributions; this would not be true for most practical applications. More recently, Ahadi–Sarkani applied multiple-correlation EMAP to continuous-density HMMs, calling this method "regression-based model prediction" (RMP) [2], [3]. The selection of a particular set of source (seen) parameters to predict a particular target (unseen) parameter is made during adaptation, since it depends partly on the amount of adaptation data for each potential source element. Although RMP adaptation is computationally intensive, it is faster than MAP: in one experiment, 96% of the mean parameters had been updated after a single adaptation sentence, compared with just 5.6% updated parameters for MAP. RMP always outperforms MAP, and converges to it for a large number of adaptation sentences [2], [3].

C. Hard and Soft Speaker Clustering

Researchers who believe, as we do, in the importance of different types of speakers have traditionally resorted to "hard" speaker clustering (since [11]). Instead of creating a single SI model, one creates several clusters of reference speakers according to some measure of acoustic similarity and trains a complete model set for each cluster; when data from the new speaker become available, one chooses one of the cluster models. Variants of this approach are found in [21], [24], [30], [33]; a hybrid technique that combines speaker clustering with MLLR is presented in [36].

Like the eigenvoice approach and reference speaker weighting (below), Gales's "soft clustering" constrains the model for the new speaker to be a linear combination, given by an ML estimator, of models obtained from reference speakers [13]. The reference models are obtained by an iterative scheme called "cluster adaptive training" (CAT) which is related to speaker-adaptive training (SAT) [4], [5]. Given initial cluster definitions, one can estimate interpolation weights for each reference speaker; given the weights for the reference speaker, one can re-estimate the cluster means. This intricate training scheme comes in two flavors: a model-based variant in which clusters are represented by a set of Gaussian component means, and a transform-based variant in which they are represented by MLLR transforms from a canonical model (as in SAT). For a fixed number of clusters, the latter approach has considerably fewer degrees of freedom. Gales quotes a figure of about 5% relative improvement in error rate when a CAT soft-clustered model is used as the prior for MLLR, as compared to using an SI prior. This was for a system with 2755 states and 12 Gaussian components per state, with multiple MLLR transforms estimated on 50 adaptation sentences [13].

The eigenvoice approach and CAT soft clustering are both ways of pooling data from the reference speakers and reducing the dimensionality of speaker space. It would be interesting to learn more about the relationship between the initial cluster definitions provided to CAT and the final ones: perhaps the final cluster means depend strongly on the original cluster definitions, or alternatively, perhaps CAT "learns" dimensions of variation that are independent of the original definitions (maybe CAT even learns dimensions of variation similar to those provided by PCA).

D. Reference Speaker Weighting

Reference speaker weighting (RSW) is the technique that most closely resembles the eigenvoice approach. To implement this technique, Hazen and Glass train a model $m(r)$ for each of R reference speakers [17], [18]. They assume that the adapted model m for the new speaker must be a linear combination of the R reference models

$$m = w(1)m(1) + \cdots + w(R)m(R).$$

Finally, they derive a maximum-likelihood estimator for the set of $w(r)$ in much the same way as was shown above for the eigenvoices.

Since Hazen and Glass were working on the medium-vocabulary DARPA RM task, it was difficult to obtain sufficient training data for the R reference speaker models. The same problem will arise if an attempt is made to extend the eigenvoice approach to medium-vocabulary or large-vocabulary systems; thus, Hazen and Glass's solution is of great interest (see the discussion in Section VII-B). The reference models $m(r)$ in their experiments represent the *centroid* for each speaker. To implement RSW, these researchers first define P phone classes in such a way that there are enough data to estimate the mean feature vector (centroid) for each of these classes for each reference speaker. Mixture component mean $\mu_m^{(s)}(p, r)$ for state s of an HMM for class p and speaker r is modeled as the sum of the centroid $c(p, r)$ plus an offset ${\nu_m}^{(s)}$

$$\mu_m^{(s)}(p, r) = c(p, r) + \nu_m^{(s)}.$$

The offsets ν are estimated over all reference speakers (*i.e.*, are assumed to be speaker-independent); each model vector $m(r)$ is obtained by concatenating the P centroids for speaker r

$$m(r) = [c(1, r), c(2, r), \cdots, c(P, r)].$$

To obtain the complete HMMs for the new speaker S at runtime, the vector $m(S)$ is obtained by means of the ML estimator. The remaining parameters—the νs, the covariances, and the transition probabilities—can then be obtained from the speaker-independent model.

There is a single difference between basic RSW and the eigenvoice approach—the application of a dimensionality reduction technique (PCA) to the reference models. However, introduction of this step may have several advantages.

- by keeping only the most important dimensions of variation, it throws away noise in the reference speaker models and reduces dependence on the identity of the R reference speakers;
- by generating an orthogonal basis for K-space, it makes estimation of the linear coefficients for the new speaker more robust;
- as the number R of reference speakers grows, the RSW approach becomes more expensive in terms of memory and computation during adaptation—an eigenvoice-based system can increase R to get more accurate eigenvoices, while holding their number K fixed (increasing offline but not online computation);
- study of the first few eigenvoices may yield insights into the most important sources of variation between speakers (see Section VI below).

E. Vowel Classification and PCA

After submission of our original eigenvoice paper [25], we became aware of work by Hu *et al*. applying PCA to a set of vectors representing SD models [19]. The goal of these researchers was to improve a Gaussian-mixture vowel classifier. PCA based on the eigenvectors of the covariance matrix was performed on a set of vectors consisting, for each reference speaker, of the concatenated mean feature vectors for vowels. Vowel data from the new speaker was projected onto the eigenvectors to estimate the new speaker's deviation from the reference speaker mean vector. Finally, classification was 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 (adaptation).

In experiments on TIMIT involving ten different vowels, a speaker-independent single mixture Gaussian classifier employing six formant-related acoustic features attained 62.8% correct vowel classification. With supervised adaptation data consisting of three other vowels pronounced by the same speaker, projection into the space spanned by the first two principal components yielded 71.6% correct vowel classification; unsupervised adaptation on the same adaptation data yielded 71.8% correct. Both results were better than those obtained by gender-dependent modeling (70.9% correct). Interestingly, these researchers found that the distribution of projections onto the first principal component separated all male speakers from all female speakers, suggesting that this component primarily conveys information about pitch and vocal tract length.

The main difference between the work done by Hu *et al*.and our work is that we applied PCA to estimate parameters of HMM's, rather than parameters of a vowel classifier. A minor difference is that we employed the variant of PCA that uses the correlation matrix rather than the covariance matrix, on the grounds that the acoustic parameters in HMM's represent a variety of physical quantities (see [22, pp. 16–17]).

IV. Eigenvoice Experiments

A. Database and Experimental Approach

Mean adaptation experiments were carried out on the Isolet database [8], which contains five sets of 30 speakers, each pronouncing the alphabet twice. From now on, each sequence of 26 letters pronounced by a given speaker will be "called a production" of the alphabet. After downsampling to 8 kHz, five splits of the data were done. Each split took four of the sets (120 speakers) as the reference data, and the remaining set (30 speakers) as test data. All results given in this paper are averaged over the five splits. Offline, maximum-likelihood (ML) training was carried out to obtain 120 SD models, and a supervector was extracted 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) cepstral features whose trajectories were bandpass filtered. Thus, each supervector contained $D = 26 \times 6 \times 18 = 2808$ parameters, and for each split, PCA was applied to a set of 120 of these supervectors.

For each of the 30 test speakers in a given split of the data, adaptation data were drawn from the first production of the alphabet, and tested on the entire second production (26 letters). To implement the eigenvoice technique, PCA was performed on the $R = 120$ supervectors (using the correlation matrix), yielding 120 eigenvoice vectors. For eigenvoice adaptation, only the first few eigenvoices $0, \cdots, K$, where 0 is the mean of the

reference speaker supervectors, are retained; two iterations of MLED (see Section II-C) are used to estimate the weights on supervectors $1, \cdots, K$. Note that although MLED is general enough to handle estimation for mixture-Gaussian HMMs, in these experiments it was applied to single-Gaussian HMMs.

B. Results

In preliminary experiments, 150 SD models, each ML-trained on the entire first production for a given speaker and tested on that speaker's second alphabet production, yielded an average recognition result of 59.6%, indicating the inadequacy of this way of employing the information in the first production. SI models ML-trained on the 120 reference speakers (i.e., the speakers not in the same Isolet set as the test speaker) yielded 81.3% word percent correct (on average over the five splits). This figure was the baseline against which adaptation methods were evaluated.

Fig. 2 shows the performance of three variants of the eigenvoice method, three other adaptation methods, and the SI baseline on this letter recognition task as a function of the amount of supervised adaptation data provided. Here, "MLED.5" denotes the eigenvoice method with $K = 5$, "MLED.10" denotes the eigenvoice method with $K = 10$, and "MLED.10=>MAP" denotes MAP with MLED.10 model as prior. "MAP" is MAP with SI prior, "MLLR" is MLLR with SI prior applied globally (*i.e.*, a single W transformation matrix is estimated and applied to all Gaussians), and "MLLR=>MAP" is MAP taking the MLLR-trained model as prior. The results given for one letter of adaptation data were obtained by averaging over results for each of the 26 letters of the alphabet in the second production (see next section for more details). Similarly, the points in the graph for 4, 10, and 17 letters were each obtained by choosing different letter subsets from the first alphabet production as adaptation data, and testing each adapted model on the entire second production for that speaker. For MAP, τ was always set to 20.

Qualitatively, the overall pattern can be summarized as follows. As expected, MAP is conservative, giving performance that is close to that of the SI system initially and gradually improving as more letters are seen. For less than about six letters of adaptation data, MAP performs worse than SI. This seems surprising, until one considers the mechanics of ML recognition. If a speaker X has provided the system with a single letter of training data (say "E"), MAP yields an adapted model in which the HMM for "E" reflects the speech idiosyncrasies of X, but the HMM's for all the other letters are derived purely from reference speakers (*i.e.*, not from X). When the recognizer encounters another letter from X—say, "B"—its "B" model, not trained on speech from X, must compete with an "E" model partly trained on speech from X. Clearly, the computed likelihood of the "E" model (wrong letter, but right speaker) having generated the "B" data will sometimes be higher than the computed likelihood for the "B" model (right letter, wrong set of speakers).

The performance of MLLR is very low for a small number of letters on which to adapt (results not shown on the graph: MLLR yields 64.9% for ten letters of adaptation data, 15.7% recognition for four letters, and 3.8% for one letter). MLLR=>MAP performance seems to be similar to, but slightly better than, that of MLLR (results not shown on the graph: 65.0% recognition for ten letters, 16.5% recognition for four letters, and 3.8% for one letter). However, the difference between MLLR and MLLR=>MAP performance may be due to chance.

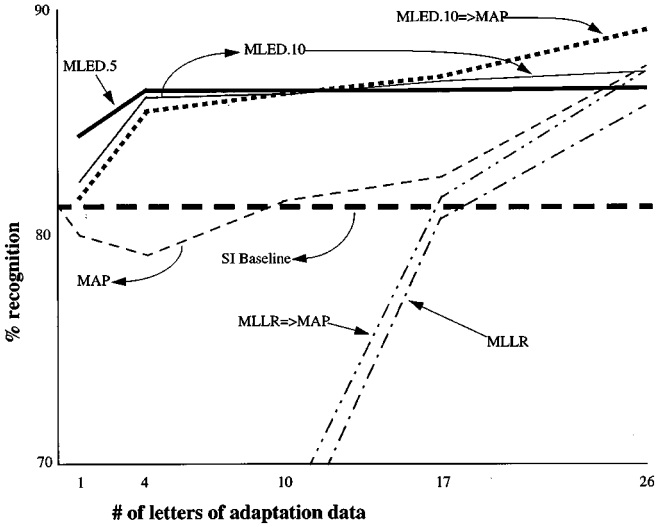

Fig. 2. Performance of Six Adaptation Methods versus SI Baseline (Supervised Adaptation)

One reader of an earlier version of this article commented that it was surprising that MAP performed better than MLLR in our experiments, as MLLR typically does better for small amounts of adaptation data. The definition of "small" is crucial here: many studies comparing MAP and MLLR do not examine performance for less than 10 s or so of adaptation data. The performance of MLLR for the tiny amounts of adaptation data shown here is consistent with results reported in [32] (see [32, p. 182, Fig. 2], which shows poor MLLR performance for one adaptation utterance).

The two pure eigenvoice variants perform better than any other method for up to 17 letters of adaptation data, but seem to reach a plateau quite quickly. Although MLED.5 performs best for very small amounts of data, it seems to reach its plateau sooner than does MLED.10; the former reaches its plateau around four letters, the latter around 17 letters. MLED.5 yields 84.4% for one letter of adaptation data (16% improvement in relative error rate) and 86.3% for four letters (26% improvement). The MLED.10=>MAP hybrid method seems to be inferior to the pure eigenvoice variants for less than ten letters, but to outperform them (and all other methods) for more than ten letters of adaptation data. To avoid cluttering the graph, we have omitted the MLED.5=>MAP results, which are very similar to the MLED.10=>MAP ones (slightly better for small amounts of adaptation data, slightly worse when the full alphabet is given—but for the full alphabet, still better than all methods except MLED.10=>MAP). It is not clear whether differences within the eigenvoice family of adaptation algorithms (MLED.5, MLED.5=>MAP, MLED.10, MLED.10=>MAP) were due to chance (we did not carry out statistical tests).

For unsupervised adaptation, experiments in which the adaptation data consisted of the entire first alphabet production

were carried out (no unsupervised adaptation experiments with subsets of the alphabet were performed). In these experiments, the SI recognizer assigns letters in the first production to the models to be adapted (with an 81.3% SI recognition rate, slightly under a fifth of adaptation letters will be assigned to the wrong letter models). Both MLED.5 and MLED.10 yielded 86.3% recognition, outperforming MAP with 77.8%, MLLR=>MAP with 78.5%, and MLLR with 81.5%. On the other hand, MLED.5=>MAP and MLED.10=>MAP performed rather badly (80.8% and 81.4%, respectively).

C. Quantitative Analysis

The most important result given in this paper is that MLED can yield significantly improved performance over the SI baseline on as little as one letter of supervised adaptation data. To validate this claim, details of these single-letter experiments are given.

Consider a particular letter, e.g., "A." For each speaker in Isolet, the sound file for the speaker's first utterance of this letter was given to the MLED algorithm as adaptation data. MLED then computed the speaker's coordinates in the eigenspace obtained from the current 120 reference speakers (i.e, all Isolet speakers not in the same Isolet set as the current speaker); the resulting adapted model for the current speaker was then tested on all 26 utterances in that speaker's second production of the alphabet. Thus, for a particular letter and a given speaker, 26 isolated-word recognition results were obtained. Each of the 150 speakers ended up being tested (this was arranged by carrying out experiments for each of the five possible splits of the Isolet sets into four training sets and one test set). Thus, $150 \times 26 = 3900$ recognition results were obtained for a particular letter provided as adaptation data.

Table I shows the MLED.5 recognition rate as a function of the single letter provided as supervised adaptation data, with results ordered from best (rank 1) to worst (rank 17). Each of the 26 entries in the table is averaged over 3900 isolated-word recognition results. Note that MLED.5 performed better than the SI baseline (81.3%) even for the worst letter, (W); average single-letter performance is 84.4%. The first column of the table lists the 13 letters yielding the best single-letter performance—these are exactly the E-set and A-set letters. When given a letter from the E-set as adaptation data, MLED produced models that did a good job of recognizing all nine letters in this large set, yielding a high recognition rate (and similarly for the four letters in the "A" set).

To test the significance of these results, McNemar's test [16] was applied to the null hypothesis: that the difference between MLED.5 and SI results is due to chance. Ideally, this hypothesis would be tested on data pooled over all 26 single-letter adaptation data subsets. However, it is difficult to devise a way of doing this that does not violate the assumptions underlying McNemar's test. Accordingly, the test was applied by treating each of the 26 letters separately (a way of evaluating significance that favors the null hypothesis). For each letter, the probability of co-occurrence of the MLED.5 and SI results for that letter was computed, under the assumption that the true performance of both methods is the same. Because of the large numbers involved, the normal approximation of the Binomial probability was employed ([16, p. 533]). At the 1% level of significance, the null hypothesis can be rejected for 20 letters; at the 5% level, the null hypothesis can be rejected for 23 letters.

TABLE I
MLED.5 RECOGNITION RATE (1 LETTER OF ADAPTATION DATA)

RANK	LETTER	% CORRECT	RANK	LETTER	% CORRECT
1	'V'	85.7	10	'L'	84.0
2	'D'	85.6	10	'X'	84.0
2	'T'	85.6	10	'Y'	84.0
3	'G'	85.5	11	'F'	83.9
3	'J'	85.5	11	'I'	83.9
4	'C'	85.3	11	'S'	83.9
4	'E'	85.3	12	'N'	83.8
5	'A'	85.2	13	'U'	83.7
5	'B'	85.2	14	'Q'	83.5
6	'P'	85.0	15	'R'	83.2
7	'H'	84.7	16	'M'	82.8
8	'K'	84.6	16	'O'	82.8
9	'Z'	84.4	17	'W'	82.2

McNemar's test was also employed to evaluate the performance of MLED.5 against SI for four and ten adaptation letters. For each of the two four-letter subsets on which adaptation experiments were carried out, the null hypothesis of no difference between the methods could be rejected at the 1% level. For each of the three 10-letter subsets on which adaptation experiments were carried out, the null hypothesis could also be rejected at the 1% level.

Finally, the performance of MLED.5 in the unsupervised adaptation experiments using the whole alphabet as adaptation data was assessed separately against each of the noneigenvoice methods. In each case, the null hypothesis that MLED.5 was no better than the other method could be rejected at the 1% level (even for MLLR, the competing method with the highest score).

D. Discussion

The main difference between the pure eigenvoice approach on the one hand, and MAP and MLLR on the other, is that the former will typically have a much smaller number of degrees of freedom during the adaptation phase. For instance, in the experiments described above MLED.5 had five degrees of freedom during adaptation, while global MLLR had approximately 522 degrees of freedom (18 features times 19 parameters per feature). The amount of online computation required by the eigenvoice technique is small.

The price paid for the powerful constraints on the adapted model in the eigenvoice approach is large offline data, memory, and computational requirements (because of the need for SD data, SD model training, and PCA applied to the SD models); furthermore, during adaptation somewhat larger amounts of memory storage are required than for other methods. For instance, to obtain the MLED.5-estimated model from the first alphabet production of a given speaker in the above

experiments, it was necessary to train 120 SD models and then perform PCA on them (offline). Since each SD model has 2808 Gaussian mean parameters, with the other parameters coming from the SI model, more than 336 960 parameters were calculated during SD training. Fortunately, the existence of extremely efficient algorithms for PCA ([22, pp. 235–246]) means that this step is less computationally intensive than SD training. During online adaptation, the system must access five eigenvoice vectors, each containing 2808 Gaussian mean parameters—a total of 14 040 parameters. For some applications, these requirements may be too high a price to pay for rapid adaptation.

The results above show that the eigenvoice approach can yield very rapid adaptation: 16% relative improvement in error rate with one letter of supervised adaptation data, 26% relative improvement in error rate with four letters of supervised adaptation data. However, these results should not be interpreted as meaning that the eigenvoice approach is globally "better" than MAP or MLLR. Rather, the approach has both the advantages and disadvantages of heavy reliance on prior information from the population of reference speakers. The advantages are most apparent when the amount of adaptation data is very small; in this situation, strong constraints on the form of the adapted model are helpful. However, as more information about the characteristics of the current speaker arrives, the larger number of degrees of freedom for MAP and MLLR turn into an advantage: these techniques continue to adapt themselves to the new speaker, while the pure eigenvoice model cannot escape from a low-dimensional hyperplane. One would also expect MAP and MLLR to perform better (even for small amounts of data) in cases where the new speaker's voice characteristics are dissimilar to those of the reference speakers. The development of hybrid techniques combining rapid and long-term adaptation will be an important area for future research (see [35] for some exciting recent results).

V. Robustness to Changes in Training

Since the eigenvoice approach relies heavily on the reference speaker models, it is important to know how sensitive it is to changes in these models, such as reductions in the amount of training data. Table II sheds light on this question. The results were obtained by using the first alphabet production for each test speaker as adaptation data, then testing on the second alphabet production by the same speaker (as before, the results were averaged over the five splits of the 150 Isolet speakers into reference and test sets). The column "2 prod, 120 spkrs" gives results when both alphabet productions were used to train each of 120 reference models (the values for MLED.5 and MLED.10 are marginally higher than those shown in Fig. 2, because of an improvement in a detail of ML training). The columns "2 prod, 60 spkrs" and "2 prod, 30 spkrs" show results for 60 and 30 reference models respectively, where each reference model is trained on both productions of the alphabet. Finally, "1 prod, 120 spkrs" shows results when only one alphabet production is used to train each of 120 reference models.

We also tried lowering the number of reference speakers of a particular sex. Unsurprisingly, performance on test speakers of a given sex deteriorated when all or most reference speakers are of the opposite sex; performance on the sex that is predominant among the reference speakers also deteriorated, but not as much.

TABLE II
MLED—Changing Training Data Quantity

Type	2 prod, 120 spkrs	2 prod, 60 spkrs	2 prod, 30 spkrs	1 prod, 120 spkrs
MLED.1	85.0	82.0	81.5	84.7
MLED.5	87.1	86.1	85.4	86.2
MLED.10	88.1	86.3	85.6	87.5

Finally, we experimented with adaptive training of the reference speaker models (global MLLR followed by MAP) where only a subset of one production of the alphabet is supplied for each reference speaker. In Table III, the third column ("Full") gives results when reference speaker models were trained on both production of the alphabet; the fourth column ("balanced 17-letters") gives results when reference speaker models were trained on 17 letters from the first production ("C," "D," "F," "G," "I," "J," "M," "N," "Q," "R," "S," "U," "V," "W," "X," "Y," and "Z"); the fifth column ("random") gives results when each reference speaker model was trained on a random set of letters drawn from the first production (average number of letters in set: 17). Note that adaptive training had no significant advantage over ML training when both alphabet productions were given ("Full"). On the other hand, when 9 letters were missing completely from the training data ("balanced 17-letters") adaptive training performed almost as well as for "Full" training data, while ML performance deteriorated considerably. Choosing random letter subsets to train each reference speaker ("random") brought ML performance almost level with that of adaptive training, presumably because every letter had been seen at least once in the training data.

The Isolet database that served as the testbed for the experiments in this paper is rather homogeneous in terms of age and regional origin (for instance, there are only two non-Americans among the 150 speakers, and very few people from the southern US). Given a more heterogeneous database, it would be fascinating to explore the impact on eigenvoice-based speaker adaptation of including and excluding speakers from the reference set on the basis of various demographic criteria.

VI. Interpretations of the Eigenvoices

Researchers often try to find physical interpretations for the first few principal components (PCs) generated by PCA. As Jolliffe points out ([22, p. 50]): "The results of a PCA are much more satisfying if intuitively reasonable interpretations can be given to some or all of the m retained PCs." However, Jolliffe also warns [22, p. 51]: "It must be emphasized that although in many examples the PC's can be readily interpreted, this is by no means universally true."

We tried to find interpretations for the first few eigendimensions obtained from one of the data splits. First, PCA was performed on the speaker-dependent models obtained from Isolet sets 1–4 (120 models, each trained on both productions of the alphabet by a given speaker); the first ten principal

TABLE III
ADAPTIVE TRAINING EXPERIMENTS

Type	Training	Full	balanced 17 letters	random (17-let. average)
MLED.1	ML	85.0	81.8	84.3
MLED.1	adaptive	84.9	84.1	84.2
MLED.5	ML	87.1	81.0	85.6
MLED.5	adaptive	87.4	86.1	85.9
MLED.10	ML	88.1	81.0	85.9
MLED.10	adaptive	88.0	86.1	86.6

components were retained as eigenvoices. Subsequently, the coordinates of all 150 speakers in the resulting space were estimated (via MLED.10 applied to both alphabet productions for each speaker). Thus, each of the 150 speakers was associated with a 10-dimensional vector of weights: $[w(1), \cdots, w(10)]$ (in order from most to least important component of variation).

Dimension 1 is closely correlated with sex: 74 of 75 women in the database have negative weights in this dimension, and all 75 men have positive weights. Dimension 1 also correlates with pitch, even within the same sex.

One of the authors (Niedzielski), who is an acoustic phonetician, studied a large number of spectrograms from the 150 speakers in order to find acoustic correlates for dimensions 2 and 3. Once a relationship had been hypothesized from visual inspection of the spectrogram, it was confirmed by quantitative measurement. Dimension 2 correlates strongly with amplitude: a negative weight indicates loudness, a positive weight softness. Fig. 3 shows this relationship graphically for the most extreme positive and most extreme negative male (M) and female (F) in dimension 2 from each of Isolet's five speaker sets. The X-axis is the log of the peak rms amplitude for letter "A" (averaged over both productions), and the Y-axis is dimension 2 weight. We chose the letter "A" for the figure arbitrarily, since the pattern is clear for all letters.

Dimension 3 may be correlated with second-formant movement. Many names for letters in the alphabet contain diphthongs, which are produced by beginning with the tongue in a relatively low position, and then "gliding" to a more fronted position associated with high $F2$ (as in "I," "E," "A") or a more backed position associated with low $F2$ (as in "U," "O"). For each of hundreds of spectrograms, $F2$ was measured at the beginning and end of a vowel; high absolute magnitude for the slope was associated with a negative dimension 3, while slopes near zero were associated with a positive dimension 3. Thus, speakers with negative dimension 3 weights seem to produce more dramatic tongue glides than those with positive dimension 3 weights. Fig. 4 was obtained from the most extreme positive and negative males and females in dimension 3 for each of Isolet's sets. This time the Y-axis comes from both productions of "U" (we chose "U" for the figure because the phonemes making up the diphthong, "y" and "uw," have tongue positions that are particularly far apart). For this letter, $F2$ values for negative speakers tend to decrease more dramatically than for positive speakers (for a letter like "A," $F2$ values for negative speakers tend to *increase* more dramatically than for positive speakers).

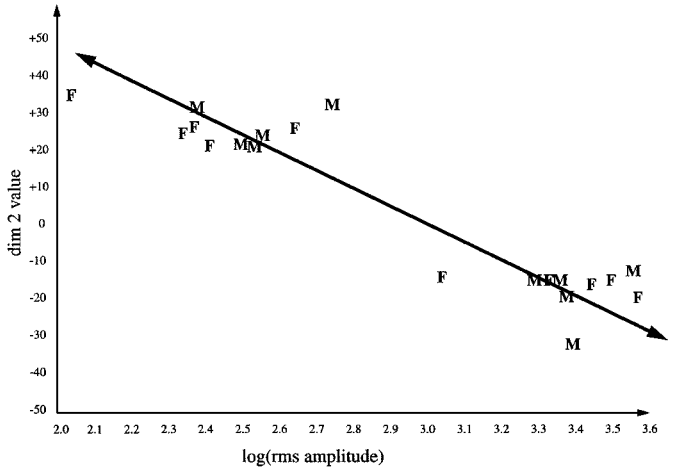

Fig. 3. Dimension 2 versus log(rms amplitude) for "A," extreme M and F in each speaker set

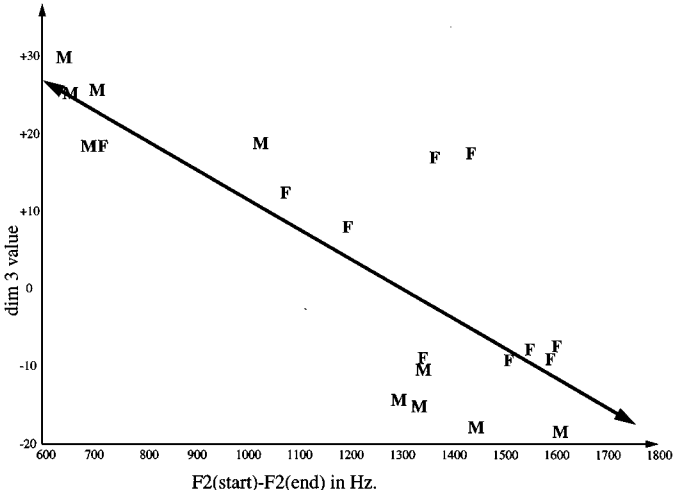

Fig. 4. Dimension 3 versus F2(start)–F2(end) for "U," extreme M and F in each speaker set

Finally, though quantitative measurements were not made, dimension 4 may be correlated with changes in pitch (a negative weight in this dimension seems to imply rising $F0$, a positive weight seems to imply falling $F0$).

These conclusions are tentative ones, given to provide an intuitive appreciation of the kind of phenomena modeled by the eigenvoices, and to suggest hypotheses for further exploration. It is possible that acoustic attributes we did not consider (or combinations and transformations of the attributes we did consider) would correlate even more strongly with the most important eigenvoices.

The most interesting correlation observed was that between dimension 1 (the most important principal component) and the speaker's sex and pitch. Recall that this relationship between the first principal component and sex was also found by Hu *et al.* (see 3.5 above), even though these researchers employed a different database, different acoustic features, and a single-Gaussian classifier instead of HMM's. It seems likely to us that this result will generalize to other databases that contain roughly equal numbers of male and female speakers.

Some of the other relationships we observed may apply only to Isolet, not to the English language in general. Isolet's vocabulary is highly unrepresentative, e.g., the spoken English alphabet contains more than its fair share of diphthongs, which may help explain the correlation between dimension 3 and $F2$ glide. Eigenvoices obtained from a different database with a larger vocabulary and more typical phoneme distribution might well assign very minor importance to $F2$ glide. However, although some of the detailed observations above may not apply elsewhere, the use of PCA as a tool for exploring the major sources of variation between speaker-dependent models is applicable to other databases.

VII. Future Work

A. Extensions of the Eigenvoice Approach

This section discusses possible extensions of the eigenvoice approach that could be applied within the context of either small-vocabulary or large-vocabulary recognition. The next section discusses the specific extensions required to apply the eigenvoice approach to large-vocabulary recognition.

Fig. 2 gives a very clear picture of the disadvantages and advantages of the pure eigenvoice approach. Because it obliges the supervector for the new speaker to be located in K-space, the adapted model is highly constrained. This is a disadvantage for larger amounts of adaptation data (performance quickly reaches a plateau), but an advantage for very small amounts. To retain the advantages of the approach for small amounts of adaptation data, while improving performance for larger amounts, one could employ a pure eigenvoice model initially and then apply MAP or MLLR, with the eigenvoice model as prior, once a threshold for the amount of adaptation data has been exceeded. This would allow the adapted model to escape from K-space, and thus deal with characteristics of the new speaker's voice not seen among the reference speakers, once enough observations have been made. Note the excellent performance of MAP with MLED.10 prior on Isolet for more than ten letters of adaptation data ("MLED.10=>MAP" line in Fig. 2). Another possibility would be to allow K to rise as the amount of adaptation data increases. However, since almost every new speaker probably manifests some speech characteristics not found in the R reference speakers, good long-term performance requires that the adapted model eventually be permitted to leave K-space, so MAP, MLLR or another technique with far more than R degrees of freedom should always be applied once sufficient adaptation data are available. Recent experiments in which the eigenvoice approach is combined with MLLR suggest this may be a highly effective strategy [35].

Besides these hybrid models, we are investigating several other extensions of the eigenvoice approach for small-vocabulary tasks:

- discriminative training of the reference SD models;
- application of the technique to environment adaptation;
- application of the technique to speaker verification and identification (here, LDA rather than PCA would probably be the best way of finding the basis);
- modification of the initial basis vectors by gradient descent so they more closely reflect the directions of maximum likelihood variation between models on the training data (the vectors found by PCA represent directions of maximum variation between the features of the models, which is not the same thing)—this idea is explored in [35];
- eigenvoice adaptation of state transition probabilities and Gaussian standard deviations.

With regard to the last point, the underlying concept is that there might be correlations between different types of parameters in HMMs: for instance, knowledge about the Gaussian means in an SD model might enable one to make more accurate predictions about the state transition probabilities and the values for the standard deviations. To exploit such correlations, one would apply PCA (or some other dimensionality reduction technique) to a set of "extended" supervectors which include information about transition probabilities and standard deviations derived from SD models. The correlation rather than the covariance version of PCA would be used, for the reasons given in Section II-B. It may seem strange to treat Gaussian means, their standard deviations, and HMM state transition probabilities in a uniform way. However, PCA is often used to find correlations between variables of very different types ([22, pp. 51–63]).

The transition probabilities are, by definition, compositional values that sum to unity. Jolliffe discusses this type of variable in detail in his book ([22, pp. 209–211]) and suggests application of the Aitchison transformation prior to PCA. Assuming these probabilities turn out to be correlated with other HMM parameters, this seems the best way of incorporating them in the approach (the reverse transformation would applied to the transition parameters after estimation of the new speaker's extended supervector).

If the relationship between the mean value of a given Gaussian parameter and its standard deviation is roughly linear (across SD models), the linearity approximation underlying PCA implies that one could get good estimates for standard deviations by entering them directly into the supervector. In cases where linearity does not hold, one could try a simple transformation prior to PCA ([22, p. 51]); again, this would imply application of the reverse transform later to obtain the new speaker's standard deviations.

Given a low-dimensional eigenspace derived from extended supervectors, a method for estimating the weights on the eigenvectors is needed. As a heuristic, it would be reasonable to use the MLED formula given earlier to estimate the weights (*i.e.*, to optimize for the Gaussian means only). Use of MLED could conceivably lead to unrealistic parameter estimates (e.g., nonpositive estimates for standard deviations). This possibility is inherent in the use of PCA and thus difficult to eliminate. Devising a form of constrained optimization for transition probabilities and standard deviations in this framework is an open problem. Of course, the really important issue is whether the Gaussian means do in fact contain useful information about the other HMM parameters; this question can only be answered by further experimental work. The next section discusses a simpler approach to estimation of variability in Gaussian models.

B. Training Eigenvoice Models for Large-Vocabulary Systems

"We do not currently have much faith in those approaches to "speaker independent" ASR which are based

> on merely collecting many examples of how different speakers say the same words, without systematising the relationship between them. It seems to us necessary for large vocabulary ASR systems to estimate characteristics of the speaker's voice and allow for them in the recognition process.... We hope that much of the variability among voices can be captured using a few orthogonal dimensions in a 'speaker space' " [7].

The work on the eigenvoice approach described in this paper can be seen as an exploration of the issues raised by the quotation above in the context of small-vocabulary recognition (surprisingly, the quotation dates to 1983). The most interesting remaining challenge is the extension of the eigenvoice approach to large-vocabulary applications. Clearly, this requires a way of dealing with allophones. In principle, one could train R context-dependent, speaker-dependent models (initializing with an SI system) and then proceed exactly as described above for the small-vocabulary alphabet task: extract a supervector from each SD model, perform PCA, and so on. With the databases available today, however, there will be insufficient data per reference speaker to yield good reference SD models (many allophones will be unrepresented in the training data from a given speaker). In addition, the computational and storage requirements of this naive extension of the small-vocabulary methodology would be onerous.

Consider the structure of the eigenvoice models for the small-vocabulary task. In these speech models, there was a clear separation between modeling of inter-speaker variability, represented by position in K-space, and modeling of intra-speaker variability, represented by the Gaussians in each speaker-adapted model. We believe that this separation was crucial to the success of the eigenvoice approach. Thus, the goal of our large-vocabulary work will be to pool data from different speakers to train allophone models in a way that still allows inter-speaker variability to be modeled separately from intra-speaker variability. Similar considerations underlie speaker-adaptive training [4], [5].

However, rather than base the structure of the large-vocabulary system on that of an initial SI model, we plan to remove speaker-dependent characteristics from training data at the very beginning of training, prior to growing state trees for allophones [41] and to creating mixture Gaussians. We hope that this methodology will lead to more efficient pooling of training data from different speakers (compared to systems derived from SI models) since the output distributions will from the beginning model mainly intra-speaker phenomena, rather than irrelevant speaker-dependent phenomena.

This training methodology is applicable with any of the "speaker space" techniques. Huo and Ma recently showed how it yields better allophone models in the SAT paradigm [20]. In the eigenvoice approach, one could attempt to obtain better allophone models by building on the work of Acero and Huang [1] and Hazen and Glass [17], [18] (see Section III-D). These researchers built models in which phoneme means are assumed to be speaker-dependent, but variability around the mean is assumed to be speaker-independent. Thus, covariance parameters can be estimated by subtracting each speaker's "centroid" (mean vector) for each phoneme from his or her training data for that phoneme, then pooling the resulting normalized training data from different speakers.

To support rapid adaptation of context-dependent models, the Acero and Huang paradigm requires a method for rapidly estimating phoneme centroids for new speakers. Given an eigenspace obtained from a population of speaker-dependent, context-independent models, an "eigencentroid" could be subtracted from each reference speaker's data before pooling the training data to train the allophone models. The result would be what we call the "eigencentroid plus delta tree" model: a context-dependent phoneme for a given speaker is modeled as the sum of a speaker-dependent phoneme centroid with an offset given by a decision tree (the delta tree) that models context-dependency. We hope that MLED will support rapid, accurate estimation of the phoneme centroids for a new speaker, yielding (in conjunction with the delta trees) a reasonably good context-dependent, adapted model from a small amount of adaptation data. Experiments with the "eigencentroid plus delta tree" model will be carried out in the near future.

VIII. Summary

We present the idea of using "eigenvoices," a set of orthogonal basis vectors derived from the parameters of speaker-dependent models trained on reference speakers, for rapid speaker adaptation, and give experimental results on the Isolet database. We applied principal component analysis to find the eigenvoices, and employed a maximum likelihood estimator called MLED to find the coordinates of each new speaker in the eigenvoice space. Compared to the performance of a speaker-independent baseline system, the eigenvoice approach yielded remarkably good performance on very small amounts of supervised adaptation data, e.g., 16% relative error rate reduction for a single letter of adaptation data, and 26% relative error rate reduction for four letters of adaptation data. With these amounts of adaptation data, it significantly outperformed the MAP and MLLR adaptation techniques. However, as the amount of adaptation data increased, the performance of eigenvoice adaptation seemed to reach a plateau. We suggest that this behavior arises from the small number of degrees of freedom allowed to the adapted model in our framework, and argue in favor of a hybrid approach in which eigenvoices are used to estimate priors for techniques with more degrees of freedom. Finally, we indicate how the technique could be applied to train context-dependent models capable of rapid speaker adaptation for large-vocabulary tasks.

References

[1] A. Acero and X. Huang, "Speaker and gender normalization for continuous-density hidden Markov models," in *Int. Conf. Acoustics, Speech, Signal Processing '96*, vol. 1, Atlanta, GA, 1996, pp. 342–345.

[2] S. Ahadi-Sarkani, "Bayesian and predictive techniques for speaker adaptation," Ph.D. diss., Cambridge Univ., Cambridge, U.K., Jan. 1996, submitted for publication.

[3] S. Ahadi and P. Woodland, "Combined Bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 11, pp. 187–206, 1997.

[4] T. Anastakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Int. Conf. Speech Language Processing '96*, vol. 2, Philadelphia, PA, 1996, pp. 1137–1140.

[5] T. Anastakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: A maximum likelihood approach to speaker normalization," in *Int. Conf. Acoustics, Speech, Signal Processing '97*, vol. 2, Munich, Germany, Apr. 1997, pp. 1043–1046.

[6] J. Atick, P. Griffin, and A. N. Redlich, "Statistical approach to shape from shading: Reconstruction of 3D face surfaces from single 2D images," *Neural Comput.*, 1996.

[7] J. Bridle and M. Ralls, "An approach to speech recognition using synthesis-by-rule," Joint Speech Research Unit, Cheltenham, U.K., Res. Rep. 1018, Dec. 1983.

[8] R. Cole, Y. Muthusamy, and M. Fanty. The ISOLET spoken letter database. [Online] http://www.cse.ogi.edu/CSLU/corpora/isolet.html

[9] S. Cox, "Predictive speaker adaptation in speech recognition," *Comput. Speech Lang.*, vol. 9, pp. 1–17, Jan. 1995.

[10] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained estimation of mixture Gaussians," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 357–365, Sept. 1995.

[11] S. Furui, "Unsupervised speaker adaptation method based on hierarchical spectral clustering," in *Int. Conf. Acoustics, Speech, Signal Processing '89*, vol. 1, Glasgow, U.K., 1989, pp. 286–289.

[12] M. Gales and P. Woodland, "Mean and variance adaptation within the MLLR framework," *Comput. Speech Lang.*, vol. 10, pp. 250–264, Oct. 1996.

[13] M. Gales, "Cluster adaptive training for speech recognition," in *Int. Conf. Speech Language Processing '98*, vol. 5, Sydney, Australia, Nov. 30–Dec. 4, 1998, pp. 1783–1786.

[14] J.-L. Gauvain and C.-H. Lee, "Bayesian learning for hidden Markov model with Gaussian mixture state observation densities," *Speech Commun.*, vol. 11, pp. 205–213, 1992.

[15] ——, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291–298, Apr. 1994.

[16] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Int. Conf. Acoustics, Speech, Signal Processing '89*, pp. 532–535.

[17] T. Hazen and J. Glass, "A comparison of novel techniques for instantaneous speaker adaptation," in *Eurospeech '97*, vol. 4, Rhodes, Greece, Sept. 1997, pp. 2047–2050.

[18] T. Hazen, "The use of speaker correlation information for automatic speech recognition," Ph.D. diss., Mass. Inst. Technol., Cambridge, Jan. 1998.

[19] Z. Hu, E. Barnard, and P. Vermeulen, "Speaker normalization using correlations among classes," in *Proc. Workshop Speech Recognition, Understanding, Processing*, Hong Kong, Sept. 1998.

[20] Q. Huo and B. Ma, "Irrelevant variability normalization in learning HMM state tying from data based on phonetic decision tree," in *Int. Conf. Acoustics, Speech, Signal Processing'99*, vol. 2, Phoenix, AZ, 1999, pp. 577–580.

[21] A. Imamura, "Speaker-adaptive HMM-based speech recognition with a stochastic speaker classifier," in *Int. Conf. Acoustics, Speech, Signal Processing'91*, vol. 2, Toronto, ON, Canada, May 1991, pp. 841–844.

[22] I. T. Jolliffe, *Principal Component Analysis*. Berlin, Germany: Springer-Verlag, 1986.

[23] M. Kirby and L. Sirovich, "Application of the Karhunen–Loève procedure for the characterization of human faces," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 103–108, Jan. 1990.

[24] T. Kosaka, S. Matsunaga, and S. Sagayama, "Tree-structured speaker clustering for speaker-independent continuous speech recognition," in *Int. Conf. Speech Language Processing'94*, Yokohama, Japan, 1994, pp. 1375–1378.

[25] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," in *Int. Conf. Speech Language Processing '98*, vol. 5, Sydney, Australia, Nov. 30–Dec. 4, 1998, pp. 1771–1774.

[26] R. Kuhn, P. Nguyen, J.-C. Junqua, and L. Goldwasser, "Eigenfaces and Eigenvoices: Dimensionality reduction for specialized pattern recognition," in *Proc. 2nd IEEE Workshop Multimedia Signal Processing*, Redondo Beach, CA, Dec. 7–9, 1998, pp. 71–76.

[27] R. Kuhn, P. Nguyen, J.-C. Junqua, R. Boman, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Fast speaker adaptation using *a priori* knowledge," in *IInt. Conf. Acoustics, Speech, Signal Processing '99*, vol. 2, Phoenix, AZ, March 1999, pp. 749–752.

[28] M. Lasry and R. Stern, "*A posteriori* estimation of correlated jointly Gaussian mean vectors," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 530–535, July 1984.

[29] C.-H. Lee, C.-H. Lin, and B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. Signal Processing*, vol. 39, pp. 806–814, 1991.

[30] K.-F. Lee, *Automatic Speech Recognition: The Development of the SPHINX System*. Norwell, MA: Kluwer, 1989.

[31] C. Leggetter and P. Woodland, "Speaker adaptation of continuous density HMM's using linear regression," in *Int. Conf. Speech Language Processing '94*, vol. 2, Yokohama, Japan, 1994, pp. 451–454.

[32] ——, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, pp. 171–185, 1995.

[33] L. Mathan and L. Miclet, "Speaker hierarchical clustering for improving speaker-independent HMM word recognition," in *Int. Conf. Acoustics, Speech, Signal Processing '90*, vol. 1, Albuquerque, NM, Apr. 1990, pp. 149–152.

[34] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 696–710, July 1997.

[35] P. Nguyen, C. Wellekens, and J.-C. Junqua, "Maximum likelihood Eigenspace and MLLR for speech recognition in noisy environments," in *Proc. Eurospeech '99*, vol. 6, Budapest, Hungary, Sept. 1999, pp. 2519–2522.

[36] M. Padmanabhan and D. Nahamoo, "Speaker clustering and transformation for speaker adaptation in speech recognition systems," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 1, Jan. 1998.

[37] W. Rozzi and R. Stern, "Speaker adaptation in continuous speech recognition via estimation of correlated mean vectors," in *Int. Conf. Acoustics, Speech, Signal Processing '91*, vol. 2, Toronto, ON, Canada, May 1991, pp. 865–868.

[38] R. Stern and M. Lasry, "Dynamic speaker adaptation for feature-based isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 751–763, June 1987.

[39] M. Tonomura, T. Kosaka, and S. Matsunaga, "Speaker adaptation based on transfer vector field smoothing using maximum *a posteriori* probability estimation," *Comput. Speech Language*, vol. 10, pp. 117–132, Apr. 1996.

[40] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[41] S. Young, J. Odell, and P. Woodland, "Tree-based state-tying for acoustic modeling," in *ARPA Workshop Human Language Tech.*, Mar. 1994, pp. 286–291.

[42] G. Zavagliakos, R. Schwartz, and J. McDonough, "Maximum *a posteriori* adaptation for large scale HMM recognizers," in *Int. Conf. Acoustics, Speech, Signal Processing '96*, vol. 1, Atlanta, GA, 1996, pp. 725–728.

Roland Kuhn received the B.Sc. (Hon.) degree in mathematics and biology from the University of Toronto, Toronto, ON, Canada, in 1981 and was Episkopon Scribe CXVII, Trinity College, Toronto. He received the M.Sc. degree in theoretical biology from the University of Chicago, Chicago, IL, in 1984 and the Ph.D. degree in computer science from McGill University, Montreal, PQ, in 1993.

From 1992 to 1996, he was with the Centre de Recherche Informatique de Montréal first as a Research Scientist (Chercheur) (1992–1995) and then as a Lead Researcher (Chercheur principal) (1995–1996). In 1996, he became a Senior Researcher with the Speech Technology Laboratory (STL), Panasonic Technologies Inc., Santa Barbara, CA. Since 1998, he has been a Lead Engineer at STL. His current research interests are rapid speaker adaptation, acoustic modeling, and speaker identification/verification. He has worked on language modeling (cache language model), automatic learning of rules for natural language understanding (semantic classification trees), topic spotting, automatic learning of spelling-to-sound rules, and dialogue systems. Since coming to STL, he has been listed as co-inventor on 21 patent applications, two of which have been granted to date.

Jean-Claude Junqua (M'90) received the Eng. degree in electronics and automation in 1980 from ENSEM, France, and the M.S. and Ph.D. degrees, in 1981 and 1989, respectively, and the Habilitation á diriger des recherche degree in computer science in 1993 from the University of Nancy I, France.

From 1981 to 1986, he was Visiting Researcher with Panasonic's Speech Technology Laboratory, Santa Barbara, CA. In 1989, he joined the Speech Technology Laboratory. From 1992 to 1993, he was Visiting Researcher with Matsushita, Osaka, Japan. He is currently Vice President and Director of Panasonic Speech Technology Laboratory, where he has focused his research on improving robustness of automatic speech recognizers for small to medium size vocabularies. His current interests cover all aspects of automatic speech recognition, e.g., the study of noisy Lombard and channel distorted speech recognition, context-dependent phone modeling, adaptive speech recognition and the design of dialogue systems. He is the author of more than 100 papers and patents in the above areas and the books *Robustness in Automatic Speech Recognition* (Norwell, MA: Kluwer, 1996) and *Robust Speech Recognition in Embedded Systems and PC Applications* (Norwell, MA: Kluwer, 2000). He served as a chairman at several international conferences and participated in various international scientific committees. In 1992 and 1997 he co-organized two ESCA workshops: "Speech Processing in Adverse Conditions" and "Robust Speech Recognition for Unknown Communication Channels." He is currently on the Editorial Board of the *Speech Communication Journal*.

Dr. Junqua was a tutorial speaker for several ESCA/IEEE workshops and for the 1999 International Conference on Acoustics, Speech, and Signal Processing. He was an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING.

Patrick Nguyen received the Dipl.Eng. degree in 1998 from the École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, and Institut Eurécom, France, winning the Hitachi Award for his studies at Institut Eurécom. He is currently pursuing the Ph.D. degree at EPFL.

He was an Intern at Panasonic's Speech Technology Laboratory (STL), Santa Barbara, CA, from January 1998 to September 1998. From September 1998 to September 1999, he was hosted by Eurécom. He returned to STL in January 2000. His research focuses on fast speaker adaptation.

Nancy Niedzielski received the Ph.D. degree in linguistics from the University of California, Santa Barbara, in 1997.

She came to linguistics via an unorthodox route which included stand-up comedy. She was a Speech Scientist with Panasonic Technologies, Inc., Santa Barbara, CA, working mainly on concatenative speech synthesis. She is an Assistant Professor of linguistics at William Marsh Rice University, Houston, TX. Her interests include sociolinguistics, dialectology, speech perception, and acoustic phonetics. She has publications in the areas of sociolinguistics, language and gender, speech synthesis, and acoustic phonetics, and a recent book entitled *Folk Linguistics* (Berlin, Germany: Mouton de Gruyter, 1999). She is co-recipient of a patent for a device to aid in the production of esophageal speech, and is named as co-inventor on two other active patent applications in speech technology.