EIGENVOICES FOR SPEAKER ADAPTATION

R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, M. Contolini

Panasonic Technologies Inc., Speech Technology Laboratory
3888 State Street, Suite 202, Santa Barbara, CA 93105, U.S.A.
Tel. (805) 687-0110; fax: (805) 687-2625; email: kuhn, jcj@research.panasonic.com

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 new speaker 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 120, 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].

[11] proposed that such a technique be applied to SD models to find speaker space, the topography of variation between speaker models. Dimensionality reduction techniques are already widely used in speech recognition, but at the level of acoustic features rather than of complete speaker models. In the eigenvoice approach, a set of T well-trained SD models must first be "vectorized". *I.e.*, for each speaker, one writes out floating-point coefficients representing all HMMs trained on that speaker, creating a vector of some large dimension D. In our Isolet experiments, only Gaussian mean parameters for each HMM state were written out in this way, but covariances, transition probabilities, or mixture weights could be included as well. The T vectors thus obtained are called "supervectors"; the order in which the HMM parameters are stored in the supervectors is arbitrary, but must be the same for all T supervectors. In an offline computation, we apply PCA or a similar technique to the set of supervectors to obtain T eigenvectors, each of dimension D - the "eigenvoices". The first few eigenvoices capture most of the variation in the data, so we need to keep only the first K of them, where $K < T << D$ (we let eigenvoice 0 be the mean vector). These K eigenvoices span "K-space".

Currently, the most commonly-used speaker adaptation techniques are MAP [6] and MLLR [13]; neither employs *a priori* information about **type of speaker**. The EMAP ("extended MAP") or RMP ("regression-based model prediction") approach is an exception: here, phoneme correlations estimated from training data allow observations of any phoneme from the new speaker to update the HMMs for all phonemes [1,4,12]. Like speaker clustering [1,5], our approach employs prior knowledge about speaker types. However, clustering diminishes the amount of training data used to train each HMM, since information is not shared across clusters, while the eigenvoice approach pools training data independently in each dimension.

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 $e(1), ..., e(K)$ be the K eigenvoices; then $E = [e(1)...e(K)]$ 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

n	be the number of features
$\mathbf{o}_t$	be feature vector (length n) at time t
$C_m^{(s)-1}$	be inverse covariance for m in state s
$\hat{\mu}_m^{(s)}$	be adapted mean for mixture m of s
$\gamma_m^{(s)}(t)$	be the $L(m, s\|\lambda, \mathbf{o}_t)$ (s-m occupation prob.)

To maximize the likelihood of observation $O = \mathbf{o}_1 \dots \mathbf{o}_T$ w.r.t. λ, we iteratively maximize an *auxiliary function* $Q(\lambda, \hat{\lambda})$, where λ is current model and $\hat{\lambda}$ is estimated model [13]. We have

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

where

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

and

$$h(\mathbf{o}_t, s, m) = (\mathbf{o}_t - \hat{\mu}_m^{(s)})^T C_m^{(s)-1} (\mathbf{o}_t - \hat{\mu}_m^{(s)})$$

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

$$e(j) = \left[e_1^{(1)}(j), e_2^{(1)}(j), \dots, e_m^{(s)}(j), \dots \right]^T$$

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

$$\hat{\mu} = \left[\hat{\mu}_1^{(1)}, \hat{\mu}_2^{(1)}, \dots, \hat{\mu}_m^{(s)}, \dots \right]^T = \sum_{j=1}^{K} w(j) e(j)$$

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

$$\hat{\mu}_m^{(s)} = \sum_{j=1}^{K} w(j) e_m^{(s)}(j)$$

To maximize $Q(\lambda, \hat{\lambda})$, set $\frac{\partial Q}{\partial w(j)} = 0, j = 1 \dots K$; assuming the eigenvalues are independent, $\frac{\partial w(i)}{\partial w(j)} = 0, i \neq j$. We obtain

$$\sum_s \sum_m \sum_t \gamma_m^{(s)}(t) (e_m^{(s)}(j))^T C_m^{(s)-1} \mathbf{o}_t =$$

$$\sum_s \sum_m \sum_t \gamma_m^{(s)}(t) \{ \sum_{k=1}^{K} w(k) (e_m^{(s)}(k))^T C_m^{(s)-1} e_m^{(s)}(j) \},$$

$$j = 1 \dots K$$

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 Isolet 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 $D = 26 * 6 * 18 = 2808$ 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 entire 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 prior ("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 $\tau = 20$ (we verified that results were insensitive to changes in τ).

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 17, *bal-17* = $\{C\ D\ F\ G\ I\ J\ M\ N\ Q\ R\ S\ U\ V\ W\ X\ Y\ Z\}$, and two subsets of size 4, *AEOW* and *ABCU*, whose membership is given by their names. Finally, since we can't show all 26 experiments using a single letter as adaptation data, we show results for D (the worst MAP result), the average result over single all letters *ave(1-let.)*, and the result for A (the best MAP result). For small amounts of data MLLR G and MLLR G => MAP give pathologically bad results.

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

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 $0 \ldots K$ (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.

K	PROJ.K	MLED.K	MLED.K => MAP
1	83.4	84.7	88.3
5	81.4	86.5	88.8
10	80.5	87.4	89.0
20	78.5	87.4	89.1

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

For unsupervised adaptation or small amounts of adaptation data, some of the eigenvoice techniques performed much better than conventional techniques (Table 3). Here, we tested eigenvoice techniques with $K = 5$ and $K = 10$ and the same adaptation data as in Table 1. Thus, we tried MLED.5, MLED.5 => MAP ("=>MAP" after "MLED.5" in Table 3), MLED.10, and MLED.10 => MAP ("=>MAP" after "MLED.10"). For single-letter adaptation, we show W (letter with worst MLED.5 result), the average results *ave(1-let.)*, and results for V (letter with best MLED.5 result). Note that unsupervised MLED.5 and MLED.10 (*alph. uns.*) are almost as good as supervised (*alph. sup.*). The SI performance is 81.3% word correct; Table 3 shows that MLED.5 can improve significantly on this even when the amount of adaptation data is very small. We know of no other equally rapid adaptation method.

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

Table 3: EIGENVOICES: PARTIAL ALPHABET

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.

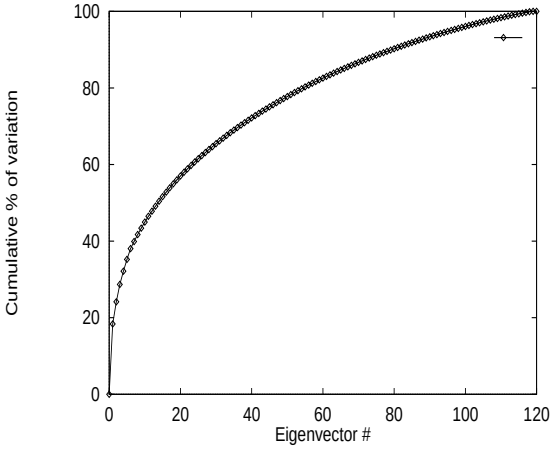

Figure 1: **Cumulative variation by eigenvoice number**

We looked for acoustic correlates of high ($+$) or low ($-$) coordinates, estimated on both alphabet repetitions, for the 150 Isolet speakers in dimensions 1, 2, and 3. Dimension 1 is closely correlated with sex (74 of 75 women in the database have $-$ values in this dimension, all 75 men have $+$ values) and with F0. Dimension 2 correlates strongly with amplitude: $-$ values indicate loudness, $+$ values softness. Both findings are rather surprising: PLP cepstral features should not contain pitch or amplitude information. However, both pitch and amplitude may be strongly correlated with other types of information (*e.g.*, locations of harmonics, spectral tilt) which are likely to survive PLP cepstral parametrization. Finally, $+$ values in dimension 3 correlate with lack of movement or low rate of change in vowel formants, while speakers with $-$ values show dramatic movement towards the off-glide.

5. DISCUSSION

Some other researchers share our belief that fast speaker adaptation can be achieved by quantifying inter-speaker variation. N. Ström models speaker variation for adaptation in a hybrid ANN/HMM 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. Ström 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 $D = 2808$ 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 λ 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

1. S. Ahadi-Sarkani. "Bayesian and Predictive Techniques for Speaker Adaptation". *Ph.D. thesis*, Cambridge University, Jan. 1996.
2. R. Cole, Y. Muthusamy, and M. Fanty. "The ISOLET Spoken Letter Database", $http://www.cse.ogi.edu/CSLU/corpora/isolet.html$
3. P. Comon. "Independent component analysis, a new concept?". *Sig. Proc.*, V. 36, No. 3, pp. 287-314, Apr. 1994.
4. S. Cox. "Predictive speaker adaptation in speech recognition". *Comp. Speech Lang.*, V. 9, pp. 1-17, Jan. 1995.
5. S. Furui. "Unsupervised speaker adaptation method based on hierarchical spectral clustering". *ICASSP-89*, V. 1, pp. 286-289, Glasgow, 1989.
6. J.-L. Gauvain and C.-H. Lee. "Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains". *IEEE Trans. Speech Audio Proc.*, V. 2, pp. 291-298, Apr. 1994.
7. H. Hermansky, B. Hanson, and H. Wakita. "Low-dimensional representation of vowels based on all-pole modeling in the psychophysical domain". *Speech Comm.*, V. 4, pp. 181-187, 1985.
8. Z. Hu, E. Barnard, and P. Vermeulen. "Speaker Normalization using Correlations Among Classes". To be publ. *Proc. Workshop on Speech Rec., Understanding and Processing*, CUHK, Hong Kong, Sept. 1998.
9. I. T. Jolliffe. "Principal Component Analysis". Springer-Verlag, 1986.
10. M. Kirby and L. Sirovich. "Application of the Karhunen-Loève Procedure for the Characterization of Human Faces". *IEEE PAMI*, V. 12, no. 1, pp. 103-108, Jan. 1990.
11. R. Kuhn. "Eigenvoices for Speaker Adaptation". Internal tech. report, STL, Santa Barbara, CA, July 30, 1997.
12. M. Lasry and R. Stern. "*A Posteriori* Estimation of Correlated Jointly Gaussian Mean Vectors". *IEEE PAMI*, V. 6, no. 4, pp. 530-535, July 1984.
13. C. Leggetter and P. Woodland. "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models". *Comp. Speech Lang.*, V. 9, pp. 171-185, 1995.
14. B. Moghaddam and A. Pentland. "Probabilistic Visual Learning for Object Representation". *IEEE PAMI*, V. 19, no. 7, pp. 696-710, July 1997.
15. P. Nguyen. "ML linear eigen-decomposition". Internal tech. report, STL, Santa Barbara, CA, Jan. 22, 1998.
16. P. Nguyen. "Fast Speaker Adaptation". Industrial Thesis Report, Institut Eurécom, June 17, 1998.
17. N. Ström. "Speaker Adaptation by Modeling the Speaker Variation in a Continuous Speech Recognition System". *ICSLP-96*, V. 2, pp. 989-992, Oct. 1996.
18. M. Turk and A. Pentland. "Eigenfaces for Recognition". *Journ. Cognitive Neuroscience*, V. 3, no. 1, pp. 71-86, 1991.