

EIGENFACES AND EIGENVOICES: DIMENSIONALITY REDUCTION FOR SPECIALIZED PATTERN RECOGNITION

R. Kuhn¹, P. Nguyen^{1,2}, J.-C. Junqua¹, and L. Goldwasser³

¹Panasonic Technologies-STL, Santa Barbara, California

²Institut Eurécom, Sophia-Antipolis Cedex, France

³Now at Marine Science Institute, Univ. Cal. Santa Barbara
(kuhn, jcj@research.panasonic.com; nguyennp@eurecom.fr)

Abstract - There are hidden analogies between two dissimilar research areas: face recognition and speech recognition. The standard representations for faces and voices misleadingly suggest that they have a high number of degrees of freedom. However, human faces have two eyes, a nose, and a mouth in predictable locations; such constraints ensure that possible images of faces occupy a tiny portion of the space of possible $2D$ images. Similarly, physical and cultural constraints on acoustic realizations of words uttered by a particular speaker imply that the true number of degrees of freedom for speaker-dependent hidden Markov models (HMMs) is quite small.

Face recognition researchers have recently adopted representations that make explicit the underlying low dimensionality of the task, greatly improving the performance of their systems while reducing computational costs. We argue that speech researchers should use similar techniques to represent variation between speakers, and discuss applications to speaker adaptation, speaker identification and speaker verification.

EIGENFACES FOR FACE RECOGNITION

“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 ... Such examples include turbulent flows, human speech, and the subject of this correspondence, human faces” ([6], p. 103).

Kirby and Sirovich [6] applied principal component analysis (PCA), which derives a low-dimensional coordinate set from a collection of high-dimensional data points [5], to the analysis of face images. Previously, researchers had modeled faces with general-purpose image processing techniques. The new coordinates consist of the eigenvectors of the covariance or correlation matrix of the data points, ordered by the magnitude of their contribution to these data. Thus, the 0th “eigenface” is the vector obtained by averaging over all original faces, and the other eigenfaces from 1 onwards model variation

from this average face. The expansion is truncated at some point, say after eigenface M . Any face image can then be represented as the average face plus a linear combination of the remaining M eigenfaces. PCA guarantees that for the original set of data points, the mean-square error introduced by truncating the expansion after the M -th eigenvector is minimized.

To match a new image of a person's face to one of a set of stored faces, one may find the Euclidean distances between the vector of M coordinates representing the new face and each of the M -dimensional vectors representing the stored faces, and then choose the stored image yielding the smallest distance. In experiments along these lines [14], where the training data (the 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 $M = 100$ eigenfaces. For face images of size 256 by 256 pixels, the dimensionality goes from 65,536 to 100: a compression factor of 655.

The best introduction to the eigenface literature is [14]. An intriguing series of recent papers discusses probability distributions for eigenfaces [10, 11, 12].

EIGENVOICES FOR SPEAKER ADAPTATION

PCA and related techniques are already widely used in speech recognition and allied fields. However, they have been applied to acoustic feature selection (*e.g.* [9]). As far as we can determine, we were the first to apply such techniques at the level of speaker representation (for other recent work, see [4]).

The obvious analogy to face recognition in the world of speech technology is speaker identification: matching the voice of an unknown person to one of a set of known voices. Our work so far has focused on a different problem, speaker adaptation, though we have conducted some preliminary experiments on speaker identification.

What is speaker adaptation?

In a typical medium- or large-vocabulary speech recognition system, words are represented as sequences of phonemes; each phoneme is represented as a set of hidden Markov models (HMMs). HMM-based speech recognition systems may be speaker-independent (SI), speaker-dependent (SD), or adaptive. SI systems are designed to recognize speech from anyone; their HMMs are trained on data from a large number of speakers. SD systems are designed to recognize speech from a particular individual; their HMMs are trained on data from that individual. Error rates for SI systems are roughly 2 to 3 times higher than those for SD systems, when the latter are tested on the speaker they are trained for [8]. Adaptive systems attempt to combine the advantages of SI and SD systems. When a new user first speaks to an adaptive system, the system employs SI HMMs; once speech data from this user has been

obtained, the parameters of the HMMs are updated to reflect user-specific traits.

Why do SD systems work better than SI systems? Phonemes do not occupy absolute positions in acoustic space, but are perceived relative to each other. If one hears someone’s “uw” and “ih”, one can make a good guess about the sound of his “ae”, because of one’s knowledge about the relative positions of these three phonemes in acoustic space. SI systems contain HMMs that are averaged over many individuals, and thus have much flatter probability distributions than HMMs in SD systems. These distributions overlap: one person’s “ow” in “about” may sound like another person’s “oo” in “room”. Training SI systems on more speakers, or changing the training algorithm, cannot solve this problem.

Many applications of speech recognition (*e.g.* flight reservation over the telephone) involve short-term user-system interactions, so there is considerable interest in fast speaker adaptation techniques. Two currently popular adaptation techniques are maximum likelihood linear regression (MLLR) and maximum *a posteriori* estimation (MAP). In MLLR, certain parameters of the SI system’s HMMs undergo an affine transformation W , which is estimated from the new user’s speech [8]. MAP estimation is a form of Bayesian learning, in which *a priori* knowledge about the parameters of the SI HMMs is combined with observations from the new speaker [3]. Neither MLLR nor MAP employs *a priori* information about **type of speaker**. The eigenvoice approach more closely resembles an older technique, speaker clustering [2], in which training speakers are divided into clusters, and HMMs for the new speaker are obtained from the cluster that best models his speech. However, information isn’t shared across clusters: *e.g.*, a Chinese-accented senior citizen might be assigned to a “Chinese accent” cluster or to a “senior citizen” cluster, but not to both. By contrast, the eigenvoice approach would give the speaker both a “Chinese accent” and an “age” coordinate (if PCA happened to produce eigenvoices correlated with these properties).

The eigenvoice approach

We train T SD models, each consisting of a complete set of HMMs, from T different speakers. Each such SD model is turned into a vector with a large dimension D ; the T vectors thus obtained are the “supervectors”. PCA applied to the set of T supervectors yields T eigenvectors, each of dimension D . By analogy with eigenfaces, we call these eigenvectors “eigenvoices”. Since the first few eigenvoices capture most of the variation in the data, we need to keep only the first K of them, where $K < T \ll D$. These K eigenvoices span “K-space”. We approximate the supervector for a new speaker S by a nearby point in K-space. Once the coordinates of this point have been estimated by means of a technique called maximum-likelihood eigendecomposition (MLEDE; [7]), it can be mapped back into a supervector of D HMM parameters to make a new model for S .

We conducted mean adaptation experiments on the Isolet database [1], which contains 5 sets of 30 speakers, each pronouncing the alphabet twice. Five splits of the data were done, each taking four sets (120 speakers) as training data, and the remaining set (30 speakers) as test data; all results below were obtained by averaging over the five splits. 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) cepstral features. 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%.

Unit accuracy results for three conventional mean adaptation techniques are shown in Table 1: MAP with SI prior (“MAP”), global MLLR with SI prior (“MLLR G”), and MAP with the MLLR G model as prior (“MLLR G => MAP”). *alph. sup.* and *alph. uns.* in Table 1 show supervised and unsupervised adaptation using the first repetition of the alphabet for each speaker as adaptation data; *alph. uns.* used SI recognition for its first pass. The other experiments in the table are for supervised adaptation on one letter from the first alphabet repetition as adaptation data. Since we can’t show all 26 experiments single-letter experiments, we show results for *D* (the worst MAP result), the average result over all single 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
<i>alph. sup.</i>	87.4	85.8	87.3
<i>alph. uns.</i>	77.8	81.5	78.5
<i>D</i> (worst)	77.6	3.8	3.8
<i>ave(1-let.)</i>	80.0	3.8	3.8
<i>A</i> (best)	81.2	3.8	3.8

Table 1: NON-EIGENVOICE ADAPTATION

To carry out experiments with eigenvoice techniques, we performed PCA on the $T = 120$ supervectors (using the correlation matrix), and kept eigenvoices $0 \dots K$ (0 is mean vector). For unsupervised adaptation or small amounts of adaptation data, some of these techniques performed much better than conventional techniques. The results in Table 2 are for the same adaptation data as in Table 1. “Eig(5)” and “Eig(10)” are the results for $K = 5$ and $K = 10$ respectively; “Eig(5)=>MAP” shows results when the Eig(5) model is used

as a prior for MAP (and analogously for “Eig(10)=>MAP”). For single-letter adaptation, we show W (letter with worst Eig(5) result), the average results *ave(1-let.)*, and results for V (letter with best Eig(5) result). Note that unsupervised Eig(5) and Eig(10) (*alph. uns.*) are almost as good as supervised (*alph. sup.*). The SI performance is 81.3% word correct; Table 2 shows that Eig(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	Eig(5)	Eig(5)=>MAP	Eig(10)	Eig(10)=>MAP
<i>alph. sup.</i>	86.5	88.8	87.4	89.0
<i>alph. uns.</i>	86.3	80.8	86.3	81.4
W (worst)	82.2	81.8	79.9	79.2
<i>ave(1-let.)</i>	84.4	83.9	82.4	81.8
V (best)	85.7	85.7	83.2	83.1

Table 2: EIGENVOICE ADAPTATION

We tried to interpret eigendimensions 1, 2, and 3 for these experiments. Dimension 1 is closely correlated with sex: 74 of 75 women in the database have negative values in this dimension, and all 75 men have positive values. Negative values in dimension 2 seem to be associated with loud, quick speakers, while negative values in dimension 3 seem to be associated with a short steady-state portion of vowels relative to the onsets and offglides.

FUTURE WORK

In these small-vocabulary experiments, the eigenvoice approach reduced the degrees of freedom for speaker adaptation from $D = 2808$ to $K \leq 20$ and yielded excellent performance for unsupervised adaptation and for small amounts of adaptation data. The reduction in degrees of freedom - hence, the potential for improved performance - will be much greater for large-vocabulary systems. Before testing large-vocabulary applications, however, we need to understand how a user’s position in eigenvoice space fluctuates over time (some eigendimensions seem to fluctuate more than others).

We have also begun to explore applications of the eigenvoice approach to speaker identification and speaker verification [13]. In speaker identification, one must match a voice to one of a closed set of known voices; in the more difficult (and more economically important) speaker verification task, one must determine whether a voice matches a certain stored voice or is an impostor. Early results on speaker identification are good (our score on 30 speakers is 100%); early results on speaker verification are mediocre, but we believe that with slight modifications the algorithm will do much better. Current work in face recognition theory will continue to suggest interesting avenues of exploration.

References

- [1] R. Cole, Y. Muthusamy, and M. Fanty, "The ISOLET Spoken Letter Database", <http://www.cse.ogi.edu/CSLU/corpora/isolet.html>
- [2] S. Furui, "Unsupervised speaker adaptation method based on hierarchical spectral clustering", *ICASSP-89*, V. 1, pp. 286-289, Glasgow, 1989.
- [3] J.-L. Gauvain and C.-H. Lee, "Bayesian learning for HMM with Gaussian mixture state observation densities", *Speech Comm.*, V. 11, pp. 205-213, 1992.
- [4] 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.
- [5] I. T. Jolliffe, "Principal Component Analysis", Springer-Verlag, 1986.
- [6] M. Kirby and L. Sirovich, "Application of the Karhunen-Loève Procedure for the Characterization of Human Faces", *IEEE Trans. Patt. Anal. Mach. Int.*, V. 12, no. 1, pp. 103-108, Jan. 1990.
- [7] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for Speaker Adaptation", *Int. Conf. Spoken Lang. Proc.*, Sydney, Australia, Nov. 30 - Dec. 4, 1998.
- [8] C. Leggetter and P. Woodland. "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs", *Comp. Speech Lang.*, V. 9, pp. 171-185, 1995.
- [9] N. Malayath, H. Hermansky, and A. Kain, "Towards decomposing the sources of variability in speech", *Eurospeech '97*, V. 1, pp. 497-500, Sept. 1997.
- [10] B. Moghaddam and A. Pentland. "A Subspace Method for Maximum Likelihood Detection". *IEEE Int. Conf. Image Processing*, Oct. 1995.
- [11] B. Moghaddam, C. Nastar, and A. Pentland. "A Bayesian Similarity Method for Direct Image Matching". *Int. Conf. Pattern Recognition*, Aug. 1996.
- [12] B. Moghaddam and A. Pentland. "Probabilistic Visual Learning for Object Representation". *IEEE Trans. Patt. Anal. Mach. Int.*, V. 19, no. 7, pp. 696-710, July 1997.
- [13] *Speech Communication* (special issue), V. 17, no. 1-2, Aug. 1995.
- [14] M. Turk and A. Pentland. "Eigenfaces for Recognition". *Jour. Cognitive Neuroscience*, V. 3, no. 1, pp. 71-86, 1991.