FAST SPEAKER ADAPTATION USING *A PRIORI* KNOWLEDGE

R. Kuhn[1], P. Nguyen[1,2], J.-C. Junqua[1], R. Boman[1], N. Niedzielski[1], S. Fincke[1], K. Field[1], M. Contolini[1]

[1]Panasonic Technologies Inc., Speech Technology Laboratory,
Santa Barbara, California, USA
[2]Institut Eurécom, Sophia-Antipolis Cedex, France
(kuhn, jcj@research.panasonic.com; nguyen@eurecom.fr)

1. ABSTRACT

Recently, we presented a radically new class of fast adaptation techniques for speech recognition, based on prior knowledge of speaker variation. To obtain this prior knowledge, one applies a dimensionality reduction technique to T vectors of dimension D derived from T speaker-dependent (SD) models. This offline step yields T basis vectors, the **eigenvoices**. We constrain the model for new speaker S to be located in the space spanned by the first K eigenvoices. Speaker adaptation involves estimating K eigenvoice coefficients for the new speaker; typically, K is very small compared to original dimension D. Here, we review how to find the eigenvoices, give a maximum-likelihood estimator for the new speaker's eigenvoice coefficients, and summarize mean adaptation experiments carried out on the Isolet database. We present new results which assess the impact on performance of changes in training of the SD models. Finally, we interpret the first few eigenvoices obtained.

2. THE EIGENVOICE APPROACH

2.1. Introduction

In two recent papers [8-9], we showed that dimensionality reduction techniques could be applied to SD models to find a low-dimensional representation for speaker space, the topography of variation between speaker models. This greatly simplifies speaker adaptation: instead of estimating the position of the new speaker in the original high-dimensional space of all possible speaker models, we need only locate this speaker in the low-dimensional space. The inspiration for this idea came from "eigenfaces" in face recognition [12]. Applicable dimensionality reduction techniques include principal component analysis (PCA) [6], independent component analysis (ICA), linear discriminant analysis, and singular value decomposition; such techniques are already widely used in speech recognition, but at the level of acoustic features rather than at the level of complete speaker models.

In the eigenvoice approach, a set of T well-trained SD models is first "vectorized". That is, 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, one applies 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 [3] and MLLR [10]; neither employs *a priori* information about type of speaker. Like speaker clustering [2], our approach employs such prior knowledge. 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.

Some other researchers share our belief that fast speaker adaptation can be achieved by quantifying inter-speaker variation. N. Ström modeled speaker variation for adaptation in a hybrid ANN-HMM system by adding an extra layer of "speaker space units" [15]. Hu *et al.* carried out speaker adaptation in a Gaussian mixture vowel classifier by performing PCA on a set of mean feature vectors for vowels derived from training speakers. They then projected vowel data from the new speaker onto the resulting eigenvectors to obtain adapted estimates for the parameters of the classifier [5].

2.2. Estimating Eigenvoice Coefficients

Let new speaker S be represented by a point P in K-space. In [8], we derived the maximum-likelihood eigen-decomposition (MLED) estimator for P in the case of Gaussian mean adaptation. 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 $O = \mathbf{o}_1 \ldots \mathbf{o}_T$ w.r.t. current model λ, we iteratively maximize an *auxiliary function* $Q(\lambda, \hat{\lambda})$, where $\hat{\lambda}$ is estimated model [10].

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

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

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

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

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

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

For maximal $Q(\lambda, \hat{\lambda})$, solve K equations for the K unknown $w(j)$ values:

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

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

$$j = 1 \ldots K$$

In the Isolet experiments described below, there was only one Gaussian per state s, so the K equations were a special case of those just given. Once they had been solved to yield MLED estimates for the Gaussian means, the other HMM parameters were obtained from a speaker-independent (SI) model.

3. EXPERIMENTS

3.1. Conventional vs. Eigenvoice Techniques

We conducted mean adaptation experiments on the Isolet database [1], 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) cepstral features whose trajectories were bandpass 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, whose unit accuracy results 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"). For MAP techniques shown here and below, we set $\tau = 20$ (we verified that results were insensitive to changes in τ). *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 involve supervised adaptation

employing subsets of the first alphabet repetition 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 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
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 experiments with eigenvoice techniques, we performed PCA on the $T = 120$ supervectors (using the correlation matrix), and kept eigenvoices $0...K$ (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. "MLED.5" and "MLED.10" are the results for the maximum-likelihood estimator with $K = 5$ and $K = 10$ respectively; the "=>MAP" after "MLED.5" shows results when the MLED.5 model is used as a prior for MAP (and analogously for the "=>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 2 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 2: EIGENVOICE ADAPTATION

3.2. Robustness to Changes in SD Training

The eigenvoice approach relies heavily on SD models obtained from training data. How robust is it to reduction in the diversity

or coverage of the training data? How sensitive is it to the method for training the SD models?

We examined these questions in a new series of experiments. The adaptation data consist of the entire first repetition of the alphabet by the new speaker, the estimation method is MLED, the test data consist of the second alphabet repetition, and all results are averaged over five training *vs.* test splits; only the set of SD models from which eigenvoices are obtained is varied.

In Table 3, we lower the number of training speakers of a particular sex. All training SD models were obtained by maximum-likelihood (ML) training on both alphabet repetitions (because of an improvement in a detail of training, these results are not strictly comparable with those in Table 2). The "K" column shows dimension, "Test" shows the test corpus (males M or females F), "Full" shows results for the full training set (60 M SD models plus 60 F SD models), "60M" shows results when only M SD models are used for PCA, and "60F" shows results for only F SD models. Finally, the "60M+4F" column shows results for 60 M models, plus 4 F models which are each copied 15 times before PCA takes place (so that males and females are weighted equally, but the male data are far more diverse); "60F+4M" gives results for the mirror-image experiment (much greater female than male diversity). As expected, performance on test speakers of a given sex deteriorates if the eigenvoices have been trained only or mainly on speakers of the other sex.

In Table 4, we vary the type of training undergone by the SD models, and also the training data corpus. In the "Type" column, "ML" stands for maximum-likelihood training (used in all other experiments), "ad" stands for adaptive training of SD models: first carry out global MLLR adaptation, then MAP adaptation, on speaker-specific data. The "Full" column gives results when both alphabet repetitions are used as training data for 120 training speakers, "2r-60s" gives results for both repetitions for only 60 training speakers (balanced by sex), "1r-120s" gives results for one repetition for all 120 training speakers. "*bal-17*" gives results for training on one repetition of the *bal-17* subset of the alphabet (defined in 3.1 above) for each of the 120 training speakers; "*rand-17*" gives results for one repetition of an alphabet subset of 17 letters (on average) by the 120 speakers, with the letters chosen randomly for each speaker. Note from "1r-120s" vs. "2r-60s" results that it is better to keep all the speakers and discard half of each speaker's data rather than the other way round. From "*bal-17*", note that SD models all trained by ML on the same incomplete letter set yield poor eigenvoices; adaptive training of SD models on the same data yields eigenvoices that perform as well as the "*rand-17*" ones.

K	Test	Full	60M	60F	60M+4F	60F+4M
1	M	85.9	84.7	74.1	85.9	83.6
1	F	84.2	74.4	83.9	81.8	84.5
5	M	87.8	87.6	79.9	86.9	84.2
5	F	86.5	82.3	85.5	82.6	85.2
10	M	89.0	88.6	82.2	89.0	85.2
10	F	87.1	84.3	86.9	84.0	87.0

Table 3: SEX EXPERIMENTS

K	Type	Full	2r-60s	1r-120s	*bal-17*	*rand-17*
1	ML	85.0	82.0	84.7	81.8	84.3
1	ad	84.9	82.0	84.4	84.1	84.2
5	ML	87.1	86.1	86.2	81.0	85.6
5	ad	87.4	86.4	87.1	86.1	85.9
10	ML	88.1	86.3	87.5	81.0	85.9
10	ad	88.0	87.2	87.4	86.1	86.6

Table 4: TRAINING TYPE AND CORPUS EXPERIMENTS

4. WHAT DO THE EIGENVOICES MEAN?

We tried to interpret the eigendimensions for one of the five data splits (with PCA performed on 120 SD models obtained by ML training on both alphabet repetitions). 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.
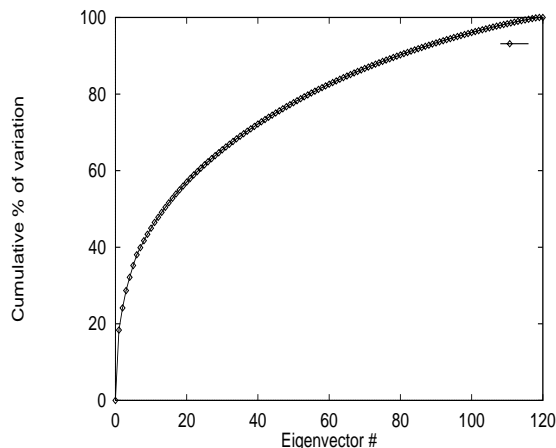

Figure 1: **Cumulative variation by eigenvoice number**

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. Note that both pitch and amplitude may be strongly correlated with other types of information (*e.g.*, locations of harmonics, spectral tilt). 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.

We also analyzed mutual information between the first ten dimensions (for all 150 speakers, both-repetition coordinates). The mutual information $I(X;Y)$ is the amount of information provided about X by Y, or vice versa [11]. It is given by $I(X;Y) =$

$H(X) - H(X|Y)$ where

$$H(X) = -\sum_x [p(X = x) \log_2 p(X = x)],$$

$$H(X|Y) = \sum_y [p(y) \sum_x [p(x|y) \log_2(1/p(x|y))]].$$

Mutual information and correlation are different: two variables may have high mutual information and no correlation. In our analysis, for each dimension the mean was subtracted from all observations, which were then quantized into bins with a width of 0.1 standard deviations. We then calculated the normalized mutual information $N(X;Y) = I(X;Y)/H(X)$. This will always be between 0.00 (Y has no information about X) and 1.00 (Y predicts X perfectly). Each of the ten dimensions has about 0.57 information about the other dimensions - this is high, and suggests there may be nonlinear dependencies between them. It also suggests that ICA might yield even better eigenvoices than the PCA-derived ones we used. Dimension 1 has 1.00 (perfect) information about sex, while the other dimensions have between 0.2 and 0.3 information about sex. Each of the dimensions gives about 0.68 information about the identity of the current speaker. Table 5 shows mutual information for dimensions $1 - 3$, and also the mutual information these dimensions give about sex and speaker ID. Each dimension gives considerable information about speaker ID, indicating the potential of eigenvoice-based speaker identification.

X	Y	N(X; Y)	N(Y; X)
Dim 1	Dim 2	0.56	0.55
Dim 1	Dim 3	0.58	0.56
Dim 1	Sex	0.21	1.00
Dim 1	Speaker ID	1.00	0.66
Dim 2	Dim 3	0.59	0.59
Dim 2	Sex	0.06	0.30
Dim 2	Speaker ID	1.00	0.68
Dim 3	Sex	0.06	0.29
Dim 3	Speaker ID	1.00	0.68

Table 5: NORMALIZED MUTUAL INFORMATION

5. DISCUSSION

In the small-vocabulary 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. For such systems, which typically contain thousands of context-dependent allophones, the issue of training the SD models which will yield the eigenvoices becomes critical. What amount of data is needed per speaker to train each allophone? If only a small amount of data is available for some allophones of some speakers, can it be leveraged in some way? One approach would be to train the SD models adaptively (as in the Table 4 "ad" experiments); we have also devised other approaches. Other important issues include training of mixture Gaussian SD models and the performance of eigenvoices found by dimensionality reduction techniques other than PCA. Eigenvoices might be trained in a way that took into account environment, as well as speaker, variability: for instance, by combining PCA with source normalization training [4]. 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. 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. "Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains". *IEEE Trans. Speech Audio Proc.*, V. 2, pp. 291-298, Apr. 1994.

4. Y. Gong. "Source Normalization Training for HMM Applied to Noisy Telephone Speech Recognition". *Eurospeech-97*, V. 3, pp. 1555-1558, Sept. 1997.

5. 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.

6. I. T. Jolliffe. "Principal Component Analysis". Springer-Verlag, 1986.

7. R. Kuhn. "Eigenvoices for Speaker Adaptation". Internal tech. report, STL, Santa Barbara, CA, July 30, 1997.

8. R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini. "Eigenvoices for Speaker Adaptation", *ICSLP-98*, Sydney, Australia, Nov. 30 - Dec. 4, 1998.

9. R. Kuhn, P. Nguyen, J.-C. Junqua, R. Boman, L. Goldwasser. "Eigenfaces and Eigenvoices: Dimensionality Reduction for Specialized Pattern Recognition", *1998 IEEE Workshop on Multimedia Sig. Proc.*, Redondo Beach, CA, Dec. 7-9, 1998.

10. 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.

11. R. McEliece. "The Theory of Information and Coding", *Encyclopedia of Mathematics and Its Applications*, V. 3, Addison-Wesley Inc., 1977.

12. B. Moghaddam and A. Pentland. "Probabilistic Visual Learning for Object Representation". *IEEE PAMI*, V. 19, no. 7, pp. 696-710, July 1997.

13. P. Nguyen. "ML linear eigen-decomposition". Internal tech. report, STL, Santa Barbara, CA, Jan. 22, 1998.

14. P. Nguyen. "Fast Speaker Adaptation". Industrial Thesis Report, Institut Eurécom, June 17, 1998.

15. N. Ström. "Speaker Adaptation by Modeling the Speaker Variation in a Continuous Speech Recognition System". *ICSLP-96*, V. 2, pp. 989-992, Oct. 1996.