SPEAKER IDENTIFICATION AND VERIFICATION USING EIGENVOICES

O. Thyes, R. Kuhn, P. Nguyen, and J.-C. Junqua

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, nguyen, jcj@research.panasonic.com

1. ABSTRACT

Gaussian Mixture Models (GMMs) have been successfully applied to the tasks of speaker ID and verification when a large amount of enrolment data is available to characterize client speakers ([1],[10],[11]). However, there are many applications where it is unreasonable to expect clients to spend this much time training the system. Thus, we have been exploring the performance of various methods when only a sparse amount of enrolment data is available. Under such conditions, the performance of GMMs deteriorates drastically. A possible solution is the "eigenvoice" approach, in which client and test speaker models are confined to a low-dimensional linear subspace obtained previously from a different set of training data. One advantage of the approach is that it does away with the need for impostor models for speaker verification.

After giving a detailed description of the eigenvoice approach, the paper compares the performance of conventional GMMs on speaker ID and verification with that of GMMs obtained by means of the eigenvoice approach. Experimental results are presented to show that conventional GMMs perform better if there are abundant enrolment data, while eigenvoice GMMs perform better if enrolment data are sparse. The paper also gives experimental results for the case where the eigenspace is trained on one database (TIMIT), but client enrolment and testing involve another (YOHO). For this case, we show that performance improves if an environment adaptation technique is applied to the eigenspace. Finally, we discuss priorities for future work.

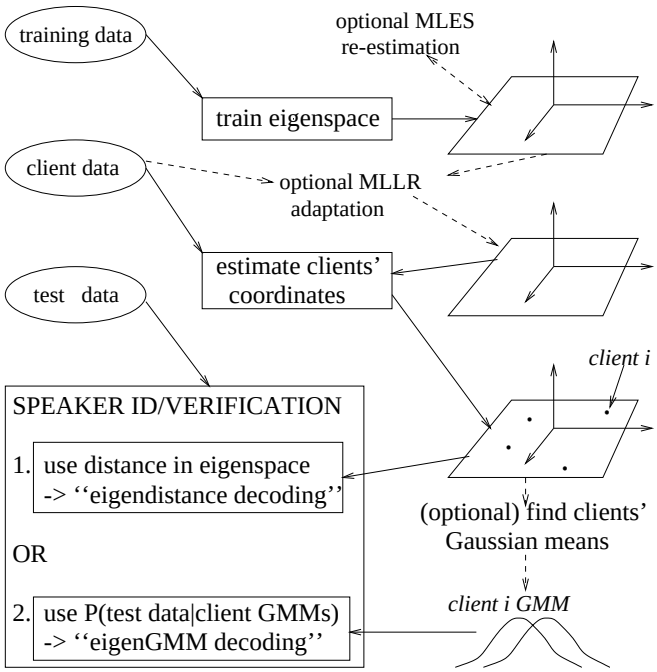

Figure 1: The eigenvoice approach

2. THE EIGENVOICE APPROACH

2.1. Introduction

The exact amount of enrolment data required by state-of-the-art speaker ID and verification systems varies according to the nature of the task. For instance, to distinguish between about 100 client speakers for a high-security application, 60 seconds or more of enrolment speech might be required for each client. However, for some tasks (especially low-security ones) clients might prefer to enrol with as little as 5 sec. of speech. Unfortunately, as experimental results given in this paper show, conventional GMMs do not perform well if enrolment data are sparse.

To solve this problem, we have drawn on earlier work on "eigenvoice" speaker adaptation, in which we employed prior knowledge about speaker space to constrain the adapted model for the new speaker ([4-6],[9]). In traditional speaker ID and verification, the system's knowledge about speech comes entirely from the client speakers. In the eigenvoice approach to the problem, we add an extra step that comes **before** enrolment of client speakers. In this extra step, speech is elicited from a diverse set of training speakers (typically disjoint from the client speakers) and then analyzed to obtain a low-dimensional speaker space called the "eigenspace". Subsequently, when clients are enrolled, the model for each client is represented as a point in the eigenspace. Thus, the approach constrains the client and test speakers to be located in a linear subspace derived from training data.

Forcing client models to be located in eigenspace is a powerful constraint that greatly reduces the number of degrees of freedom. A client GMM with 32 mixture Gaussians and 26 acoustic features per Gaussian has, ignoring mixture weights, 832 degrees of freedom. If we now impose on the model the additional constraint that it must be located in a 20-dimensional eigenspace obtained from training data, the number of degrees of freedom has shrunk by a factor of more than 40. Whether or not it's a good idea to impose this constraint depends on the amount of enrolment data. For small amounts of enrolment data, the eigenspace constraint makes it possible to estimate a reasonable model for

each client quickly (in the example, only 20 parameters would have to be estimated). For large amounts of enrolment data, it's better not to impose the constraint, since it implies that unusual aspects of a client's voice (i.e., phenomena not seen in the training data) will not be represented. Thus, our technique is designed for tasks where clients need to be enrolled quickly, with a minimal amount of enrolment data collected per client. It is also well-suited for tasks where memory must be minimized, since each additional client model only requires a small number of stored parameters.

2.2. Applying the Approach

Our approach is summarized in Figure 1. First, we obtain a set of models for training speakers (in the experiments described here, these models were conventional GMMs). Training data are collected only once, in an offline step; ideally, they will be provided by a large and diverse set of speakers, with large amounts of speech collected from each speaker. Next, we apply a technique such as PCA (Principal Component Analysis) or LDA (Linear Discriminant Analysis) to the means of the training speaker GMMs to obtain a low-dimensional eigenspace made up of "eigenvoice" basis vectors. Optionally, we may apply a re-estimation technique called MLES (Maximum Likelihood EigenSpace) to obtain a better eigenspace. PCA, LDA, and MLES are described in the next subsection.

Our goal in the client enrolment step is to minimize the annoyance to the clients by minimizing the amount of speech collected per client. Since the acoustic environment for the speaker ID/verification task may differ from the environment in which training speakers were recorded, data from the clients (or from other speakers recorded under the task conditions) may optionally be used to adapt the eigenspace to the task environment adaptation via a method such as MLLR [7].

To estimate each client's coordinates in the eigenspace from a few seconds of data, a technique called MLED (Maximum Likelihood EigenDecomposition) is used [4-6]. Each point in the eigenspace represents a possible speaker model (a GMM in these experiments) - thus, once can also build a GMM for a client, given his or her position in the eigenspace. Since each eigenspace point only carries information about Gaussian means, the variances must be obtained from somewhere else (typically, from a speaker-independent model).

Finally, there is the speaker ID/verification step, in which the system must assign data from a test speaker to one of the clients, or decide that he/she is an impostor. There are two ways of doing this:

1. Project the test speaker into the eigenspace using MLED, then find the distance between the test speaker point and the client point(s) in the eigenspace - we call this "eigendistance decoding";
2. Use speaker models (*e.g.*, GMMs) generated from client points in eigenspace to calculate the likelihood of the test data - we call this "eigenGMM decoding".

For speaker ID, the test speaker is assigned to the closest client in eigenspace (eigendistance decoding) or the client whose model derived from eigenspace yields the highest likelihood on test data (eigenGMM decoding). For speaker verification, eigendistance thresholds or eigenGMM likelihood thresholds are applied to decide if the test speaker is a client or an impostor. In the case of eigendistance speaker verification, there is no need for an impostor model to normalize for utterance likelihood dependencies. The reason for this is that the eigenspace itself implicitly normalizes for utterance likelihood: two utterances with very different likelihoods (as calculated by a GMM or HMM) may map to the same point in the eigenspace.

2.3. Eigenspace Training Techniques

As described in our papers on speaker adaptation, PCA discovers the directions that account for the largest variability among training speakers [4-6]. In the experiments reported here, each training speaker's Gaussian means were concatenated to form a "supervector" of dimension D. PCA was applied to the set of T supervectors obtained from the T training speakers, yielding $T-1$ eigenvoice vectors ordered by the magnitude of their contribution to the between-speaker scatter matrix. This matrix is:

$$S_B = \sum_{s=1}^{T} N_s(\mu_s - \mu)(\mu_s - \mu)^T \quad (1)$$

where N_s is the number of training utterances of speaker s, μ_s the mean of all N_s samples, and μ is the overall mean. Typically, we discard the higher-order eigenvoices (which mainly contain noise) to obtain an eigenspace of dimension less than $T-1$.

In pure PCA, the means of the Gaussians in each training speaker's GMM are treated as vectors and we aim to find the maximally varying directions. However, the GMMs are actually probabilistic models. To better model the speaker space, we can apply Maximum Likelihood EigenSpace (MLES) estimation [9] which reestimates the initial PCA eigenspace so as to maximize the likelihood of the training data, given the speaker's identity: i.e., $P(O_S|\lambda_S)$ is maximized (where O_S and λ_S represent an observation and the GMM of a given speaker respectively).

Linear Discriminant Analysis is particularly relevant to speaker ID and verification, since it tries to increase discrimination between classes (in our case, a class consists of all speech from a given speaker). For other recent work applying LDA to this task (though in a completely different way) see [11]. LDA was much less relevant to our earlier work on speaker adaptation for speech recognition systems, since no-one cares whether an adapted recognizer distinguishes between speakers if it performs well for the current speaker.

Fisher's Linear Discriminant (FLD) tries to "shape" the scatter in a set of data samples to make classification easier [2]. Consider an orthogonal transformation W mapping each D-dimensional supervector x_k into eigenspace:

$$y_k = W^T x_k \quad (2)$$

(where y_k is the transformed vector of dimension T). The transformation matrix W is selected so as to maximize the ratio between the between-class scatter S_B and the within-class scatter

$$S_W = \sum_{s=1}^{T} \sum_{x_k \in X_s} (x_k - \mu_s)(x_k - \mu_s)^T \quad (3)$$

where μ_s is the mean of speaker s. The optimal transformation matrix W_{lda} will then be chosen so as to maximize the ratio of the determinant of $\tilde{S}_B = W_{lda} S_B W_{lda}^T$ of the projected samples to the determinant of $\tilde{S}_W = W_{lda} S_W W_{lda}^T$ of the projected samples:

$$\begin{aligned} W_{lda} &= \arg\max_W \frac{\left|W^T S_B W\right|}{\left|W^T S_W W\right|} \\ &= [e(1)e(2)\ldots e(K)] \end{aligned} \quad (4)$$

where $\{e(i)|i = 1, \ldots, K\}$ are the generalized eigenvectors of S_B and S_W corresponding to the K largest eigenvalues $\{\lambda_i|i = 1, \ldots, K\}$:

$$\begin{aligned} & S_B e(i) = \lambda_i S_W e(i), \quad i = 1, \ldots, K \\ \Leftrightarrow \quad & S_W^{-1} S_B e(i) = \lambda_i e(i). \end{aligned} \tag{5}$$

The rank of S_W is at most $N - T$, where N is the total number of utterances in the training database and T the number of speakers. Thus, for each GMM used to build the eigenspace W_{lda}, we require more than D sample utterances (D is the dimension of the supervectors). Given the nature of human speech, this is unlikely to be a problem. For an interesting discussion of LDA applied to face recognition (where obtaining a sufficient number of face images **is** a problem), see [2].

3. EXPERIMENTS

Two databases were used in these experiments: the YOHO Speaker Verification database of "combination lock" phrases and the TIMIT database of acoustically varied continuous speech [8]. However, only YOHO was used for client enrolment and testing (as opposed to eigenspace training). To obtain eigenspaces, speaker-dependent GMMs were initialized on a simple "SILENCE speech SILENCE" segmentation obtained by means of a silence model and a speaker-independent model. The sampling rate was 8 kHz (TIMIT data were downsampled to 8 kHz). There were 26 MFCC acoustic features (13 static, 13 dynamic), to which cepstral filtering was applied.

3.1. Results for abundant enrolment data

In an initial set of experiments on YOHO, we tried several speaker ID approaches on 82 speakers with 360 seconds of enrolment data per client. When 5 seconds of test speech not used for enrolment was presented for each of the 82 clients, the conventional GMM approach with 32 Gaussians yielded 98.8% correct identification. For the eigenvoice approaches, the eigenspace was obtained from 72 of the 82 client speakers (implying that the maximum possible dimensionality of the eigenspace is 71).

Although this overlap between training speakers and enrolment speakers favours the eigenvoice approaches, none of them performed as well as the conventional GMM approach. The best eigenvoice result, 98.0%, occurred in the case where LDA was used for eigenspace training, the dimensionality of the eigenspace was set to 70, and eigenGMM decoding was used. Among all the eigenvoice approaches, training the eigenspace with LDA (rather than PCA, PCA followed by MLES, or LDA followed by MLES), setting the dimensionality high, and carrying out eigenGMM decoding (rather than eigendistance decoding), always contributed to better performance. We concluded that for abundant enrolment data, no eigenvoice approach can outperform the conventional approach, since projecting the client data into a linear subspace causes reliably estimated information about the client to be lost. Thus, we did not perform a comparison of speaker verification techniques for this condition.

3.2. Results for sparse enrolment data

Figure 2 shows speaker ID results for 5 sec. of test speaker data and sparse enrolment data: 10 sec. enrolment for each of 10 clients. Here, the eigenvoice methods all used 64-Gaussian models and eigenGMM decoding. However, the baseline of 77.8% correct ID shown was obtained from the best conventionally obtained GMM, which had 8 Gaussians, rather than from the 64-Gaussian conventional GMM (whose performance was around 30%). Each horizontal line on the figure represents a fixed number of dimensions for the eigenspace: e.g., the line "20 eigenvoices" shows results for 20 dimensions. For the experiments in this figure, the 72 training speakers for the eigenspace were disjoint from the 10 clients. The best result was 95.0% correct for 70 dimensions and an LDA-trained eigenspace.

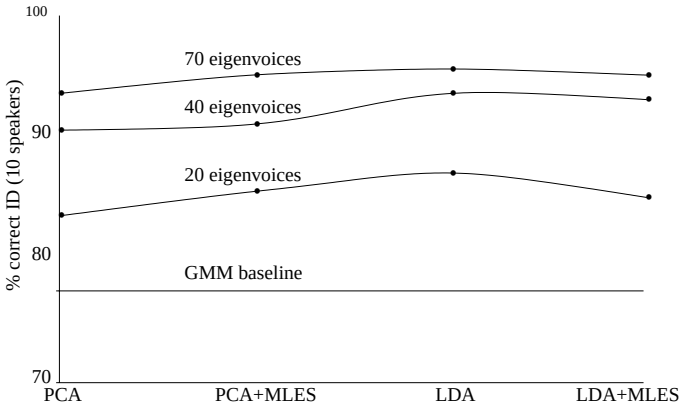

Figure 2: Speaker ID: 10 sec. enrolment data, 5 sec. test data

Clearly, eigenspace dimensionality has a powerful impact on performance. The method of training the eigenspace is also important. Note that LDA always performs better than any other method, beating PCA, PCA initialization with MLES re-estimation, and LDA with MLES re-estimation. Not shown here are experimental results where ID is carried out via eigendistance decoding. We tried three eigendistance metrics - angle, Euclidean distance, and a metric which weighted each eigenspace dimensions by its eigenvalue - but found that there was little difference between them, and that eigenGMM decoding typically outperformed eigendistance decoding by a small amount (about 5% relative error).

Experimental results for speaker verification (using a speaker-independent impostor model for eigenGMM decoding) are shown in table 1 for a 40-dimensional eigenspace on 64-GMMs obtained from 72 speakers (disjoint from the 10 client speakers). For speaker verification, eigendistance decoding outperforms eigenGMM decoding, and both outperform conventional GMM decoding. The best conventional GMM result for 5 sec. of enrolment data is for a 4-Gaussian model, and the best conventional GMM result for 10 sec. is for an 8-Gaussian model.

3.3. Eigenspace adaptation

In practical applications, the eigenvoice will have to handle mismatch between the training environment, on one hand, and the enrolment and testing environments on the other. Thus, we trained an eigenspace for 64-GMMs on the 630 TIMIT speakers, each supplying 10 sentences, and carried out enrolment and testing on YOHO.

Table 2 compares the results from a YOHO-trained 64-GMM eigenspace for 10 sec. of enrolment and 5 sec. of test data (these results were shown in Figure 2) with those obtained for the same 10 YOHO speakers on the TIMIT eigenspace, and on

5 seconds enrolment		
Best GMM baseline (4G)		21.5%
Decoding	PCA	LDA
Euclidian Distance	9.6%	7.0%
GMM Decoding	11.0%	9.9%
10 seconds enrolment		
Best GMM baseline (8G)		14.4%
Decoding	PCA	LDA
Euclidian Distance	7.1%	6.4%
GMM Decoding	10.0%	9.0%

Table 1: Speaker verification (Equal Error Rate): 64-GMM, 40 eigenvoices, YOHO training, enrolment, and testing

Eigenvoice dimension	20	40	70
YOHO eigenspace			
PCA without MLES	84.3%	89.0%	93.0%
PCA with MLES	86.8%	89.3%	92.8%
LDA	87.8%	94.3%	95.0%
TIMIT eigenspace			
PCA without MLES	76.5%	86.0%	91.5%
PCA with MLES	79.0%	85.5%	92.0%
LDA	77.3%	83.5%	82.8%
MLLR-adapted TIMIT eigenspace			
PCA without MLES	78.5%	88.5%	92.3%
PCA with MLES	79.3%	88.8%	92.5%
LDA	79.3%	86.8%	84.0%

Table 2: Speaker ID: 64-GMM, YOHO vs. TIMIT vs. MLLR-adapted TIMIT for eigenspace training

an eigenspace obtained by applying global MLLR environment adaptation to the TIMIT eigenspace (and also to the TIMIT silence model). The adaptation was performed on the enrolment data from the 10 clients; we observed no significant difference when much larger amounts of adaptation data were used.

The results show that although the eigenspace trained on YOHO via LDA performs best on YOHO enrolment and test data, the eigenspace trained on TIMIT via LDA performs worse than TIMIT eigenspaces obtained by the other two methods - whether or not MLLR is applied subsequently. Compared to YOHO, TIMIT is unsuitable for LDA in two ways: it contains only 10 sentences per speaker (YOHO has 96), and it contains far more allophonic variability, making it easier to confound this type of variability with speaker-dependent variability (YOHO has only "combination lock" phrases). Thus, our choice of TIMIT for training the initial eigenspace may have been a mistake.

4. CONCLUSIONS

The eigenvoice approach forces models for the client and test speakers to be confined to a low-dimensional subspace obtained from training data. For sparse amounts of enrolment data ($5-10$ sec.) this approach consistently outperforms conventional GMM training. For larger amounts of enrolment data, the loss of degrees of freedom caused by restriction to eigenspace leads to inferior performance. For speaker verification, an advantage of the approach is that, in its "eigendistance decoding" variant, it dispenses with the need for impostor models.

Of the eigenspace training methods tested, LDA appears to be the most promising. However, all the eigenvoice methods may run into difficulty when trained on acoustically diverse databases with small amounts of data per speaker. For instance, speaker-dependent variability in TIMIT is less important than phoneme identity, channel effects, and phonetic context [3]; this makes it likely that eigenspaces trained on TIMIT and similar databases will confound speaker-dependent information with these other types of information. Clearly, the top priority for future work is the development of more robust eigenspace training techniques.

5. REFERENCES

1. M.E. Forsyth. "Hidden Markov Models For Automatic Speaker Verification". *PhD Thesis*, University of Edinburgh, 1995.
2. João P. Hespanha, Peter N. Belhumeur and David J. Kriegman. "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", *IEEE Trans. PAMI*, no. 17, pp. 711-730, 1997.
3. S. Kajarekar, N. Malayath and H. Hermansky. "Analysis of Speaker and Channel Variability in Speech". *ASRU Workshop*, Keystone, Colorado, Dec. 1999.
4. R. Kuhn, P. Nguyen, J.-C. Junqua, *et al*. "Eigenvoices for Speaker Adaptation". *ICSLP-98*, V. 5, pp. 1771-1774, Sydney, Australia, Nov. 30 - Dec. 4, 1998.
5. R. Kuhn, P. Nguyen, J.-C. Junqua, *et al*. "Fast Speaker Adaptation using *A Priori* knowledge". *ICASSP-99*, V. 2, pp. 749-752, Phoenix, Arizona, March 15-19, 1999.
6. R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski. "Rapid Speaker Adaptation in Eigenvoice Space". *IEEE Trans. Speech Audio Proc.* (to appear around Nov. 2000).
7. C.J. Legetter and P.C. Woodland. "Maximum likelihood linear regression for speaker adaptation of continuous density Hidden Markov Models". *Computer Speech and Language*, V. 9, pp. 171-185, 1995.
8. Linguistic Data Consortium. "YOHO Speaker Verification" and "TIMIT Acoustic-Phonetic Continuous Speech Corpus". http://morph.ldc.upenn.edu/Catalog/
9. P. Nguyen, C. Wellekens and J.-C. Junqua. "Maximum Likelihood Eigenspace and MLLR for Speech Recognition in Noisy Environments". *Eurospeech-99*, V. 6, pp. 2519-2522, Budapest, Hungary, 1999.
10. D.A. Reynolds. "Speaker Identification and Verification using Gaussian Mixture Speaker Models". *Speech Communication*, V. 17, pp. 177-192, 1995.
11. R.A. Sukkar, M.B. Gandhi and A.R. Setlur. "Speaker Verification Using Mixture Decomposition Discrimination". *IEEE Trans. Speech Audio Proc.*, V. 8, no. 3, pp. 292-299, May 2000.