NORMALIZED DISCRIMINANT ANALYSIS WITH APPLICATION TO A HYBRID SPEAKER-VERIFICATION SYSTEM

Qi Li, S. Parthasarathy, Aaron E. Rosenberg, and Donald W. Tufts†

Speech Research Department
AT&T Bell Laboratories
Murray Hill, NJ 07974

†Dept. of Electrical Engineering
University of Rhode Island
Kingston, RI 02881

ABSTRACT

A modified linear discriminant analysis technique for speaker verification, referred to here as normalized discriminant analysis (NDA), is presented. Using this technique it is possible to design an efficient linear classifier with very limited training data and to generate normalized discriminant scores with comparable magnitudes for different classifiers. The NDA technique is applied to a classifier for speaker verification based on speaker specific information obtained when utterances are processed with speaker independent models. In experiments conducted on a network based telephone database, the NDA technique provides an equal-error rate of 6.13% while the classifier using Fisher linear discriminant analysis has an equal-error rate of 18.18%. Furthermore, when the NDA combined with HMM approach in a hybrid speaker verification system, the rate was reduced from 5.30% (HMM with cohort normalization) to 4.32%.

1. INTRODUCTION

A text dependent, connected digit, speaker verification system often consists of different classifiers for different words for each speaker. Two kinds of problems occur when linear discriminant analysis (LDA) is used to design these classifiers: the amount of training data is usually small, and the discriminant scores obtained from different classifiers are scaled differently so that it is hard to compare and combine them. A *normalized discriminant analysis* technique (NDA) is presented in this paper to address these problems.

NDA is applied to design a hybrid speaker verification (HSV) system. As reported by Setlur *et al* [1], the system combines two types of word models or classifiers (we use the term classifier when discriminant analysis is used). The first type of classifier used is a speaker dependent, continuous density, Gaussian mixture Hidden Markov Model (HMM). This representation has been shown to provide good performance for connected digit password speaker verification [2]. The second type of classifier is based on speaker specific information that can be obtained when password utterances are processed with speaker independent (SI) Gaussian mixture HMM's. The mixture components of a speaker independent Gaussian mixture HMM are found by clustering training data from a wide variety of speakers and recording conditions. For a particular state of a particular word, each such component is representative of some subset of the training data. When a test utterance is processed, the score for each test vector is calculated using a weighted sum of mixture component likelihoods. Because of the way the mixture component parameters are trained, it is reasonable to expect that each test speaker will have different characteristic distributions of likelihood scores across these components. The second type of classifier is based on these characteristic mixture component likelihood distributions obtained over training utterances for each speaker.

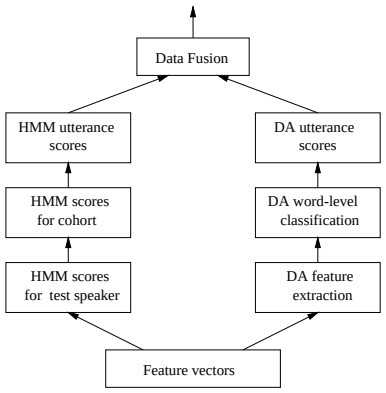

Figure 1: The structure of a hybrid speaker verification (HSV) system.

Although the second type of speaker classifier yields significantly lower performance than the first type, it has been shown in [1] that, when combined, the two

representations yield significantly improved verification performance over either one by itself. The features which distinguish this study from Setlur *et al* [1] are in two areas. First, NDA is used instead of Fisher linear discriminant analysis; second, verification is carried out on a database recorded over the long-distance telephone network under a variety of recording and channel conditions; third, only small amounts of training data are available per speaker.

The HSV system as reported in [1] is shown in Figure 1. The HSV system consists of three modules: a Type 1, HMM classifier, a Type 2, discriminant analysis classifier, and a data fusion layer. We note that an HSV system could include more classifiers as long as each individual classifier can provide independent information.

2. NORMALIZED DISCRIMINANT ANALYSIS

Word-level discrimination between a true speaker and impostors is a two-class classification problem. Using LDA, a weight vector $\mathbf{w}$, is found, such that the projected data from the true speaker and impostors is maximally separated. In brief, for two-class LDA, $\mathbf{w}$ can be solved directly as

$$\mathbf{w} = S_W^{-1}(\mathbf{m}_T - \mathbf{m}_I) \tag{1}$$

where $\mathbf{m}_T$ and $\mathbf{m}_I$ are the sample means of the two classes, true speaker and impostors, and S_W, is usually defined as

$$S_W = \sum_{\mathbf{x}\in\mathbf{X}_T} (\mathbf{x}-\mathbf{m}_T)(\mathbf{x}-\mathbf{m}_T)^t + \sum_{\mathbf{x}\in\mathbf{X}_I} (\mathbf{x}-\mathbf{m}_I)(\mathbf{x}-\mathbf{m}_I)^t, \tag{2}$$

where $\mathbf{X}_T$ and $\mathbf{X}_I$ are the data matrices of a true speaker and impostors. S_W must be non-singular. Each row in the matrices represents one training data vector $\mathbf{x}$. More details on LDA can be found in [3].

However, in practical speaker verification applications , there are usually only a few training vectors for each true speaker (for example, there are only 5 vectors available in our experiments). To compensate for this lack of training data, we redefine the S_W in (2) as

$$\hat{S}_W = R_T + \gamma R_{CT} + R_I + \delta R_{CI}, \tag{3}$$

where R_T and R_I are the sample covariance matrices from the true speaker and impostors and R_{CI} and R_{CT} are compensating covariance matrices from another available group of speakers (not used in the evaluation). R_{CI} is the sample covariance matrix of additional speakers, pooling their data. Actually, R_I and R_{CI} can be combined except we may want to weight the associated data sets differently. R_{CT} is defined as

$$R_{CT} = \frac{1}{T_s}\sum_{i=1}^{T_s} R_i, \tag{4}$$

where R_i is the sample covariance matrix of Speaker i in the other group, and T_s is the total number of speakers in the group. γ and δ are weight factors determined experimentally.

An LDA score p of a data vector $\mathbf{x}$ is obtained by projecting $\mathbf{x}$ onto a weight vector $\mathbf{w}$, $p = \mathbf{w}^t\mathbf{x}$. To make the scores comparable across different words and different speakers, we use the following normalization.

$$\hat{p} = \alpha\mathbf{w}^t\mathbf{x} + \beta \tag{5}$$

where $\alpha = \frac{2}{d}$, $\beta = -1 - \frac{2\mu_I}{d}$, and $d = \mu_T - \mu_I$. The μ_T and μ_I are the means of projected data from true speaker and impostors. $\hat{p}$ is the NDA score.

3. APPLYING NDA IN THE HYBRID SPEAKER-VERIFICATION SYSTEM

NDA can be used to design Type 2 classifiers for speaker verification. The classifiers can be used separately or as a module in the HSV system [1].

3.1. Training of the NDA System

3.1.1. Feature Extraction

As described earlier, the Type 2 classifier features are determined from speaker-independent (SI) HMM's. Each training or test utterance is first segmented into words and states. As shown in Figure 2, we use the averaged outputs of the Gaussian components on the HMM states as one fixed-length feature vector for the NDA training. The elements of the feature vector are defined as follows.

$$x_{jm} = \frac{1}{T_j}\sum_{t=1}^{T_j} \log(\mathcal{N}(\mathbf{o}_t, \mu_{jm}, R_{jm})),$$
$$j = 1, ..., J;\ m = 1, ..., M_j. \tag{6}$$

where $\mathbf{o}_t$ is the cepstral feature vector at time frame t, μ_{jm} and R_{jm} are the mean and covariance of the mth mixture component for state j, M_j is the total number of mixture components for state j, $\mathcal{N}(.)$ is a Gaussian function, and T_j is the total number of frames segmented into state j.

Thus, a sequence of cepstral feature vectors associated with one segmented word is mapped onto a fixed-length feature vector. The length of the feature vector

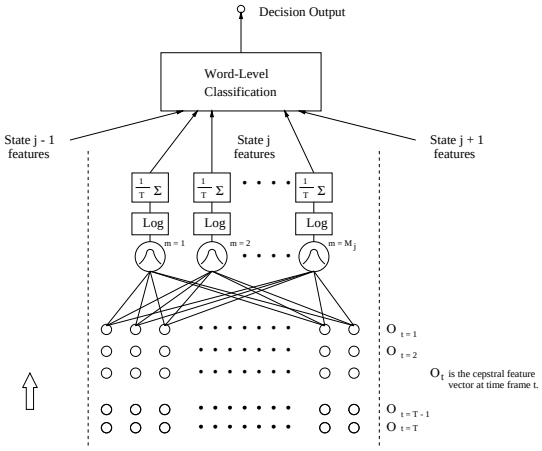

Figure 2: The NDA feature extraction.

is equal to the total number of Gaussian components of the word HMM ($J \times M$).

Feature extraction is almost identical to the technique in [1] except that the HMM mixture weights are omitted since they are absorbed in the NDA calculation.

3.1.2. Word and Utterance Level Verifications

The structure of the word and utterance verification for one speaker is shown in Figure 3. There is an NDA classifier for each word.

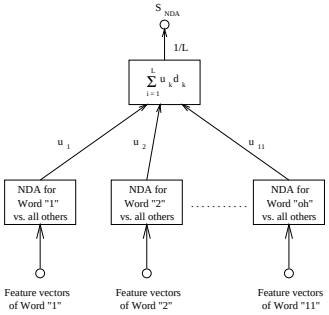

Figure 3: The Type 2 classifier (NDA system) for one speaker.

An utterance score $S_{NDA}(\mathbf{O})$ is a weighted sum of NDA scores of all words in the utterance.

$$S_{NDA}(\mathbf{O}) = \frac{1}{L}\sum_{i=1}^{L} u_{k_i}\hat{p}_{k_i}, \quad k_i \in \{1,..,11\}. \qquad (7)$$

where L is the length of the utterance $\mathbf{O}$ (the total number of words in the utterance), $\hat{p}_{k_i}$ is the NDA score for the ith word. Equation (7) specifies a linear node with associated weight vectors u_{k_i} that can be determined by optimal data fusion [4] to equalize the performance across words if sufficient training data is available. Otherwise, we can just use $u_{k_i} = 1$.

3.2. Training of the HMM System

For the Type 1 classifier, the HMM scores are calculated from speaker dependent (SD) HMM models. Cohort normalization is applied by selecting 5 scores from a group of speakers not in the evaluation group. Cepstral mean normalization is applied in both the HMM and NDA classifiers.

3.3. The Training of the Data Fusion Layer

The final decision on a given utterance $\mathbf{O}$ is made based on a score S which is calculated by combining the NDA score $S_{NDA}(\mathbf{O})$ and HMM score $S_{HMM}(\mathbf{O})$.

$$S = v_1 S_{NDA}(\mathbf{O}) + v_2 S_{HMM}(\mathbf{O}) \qquad (8)$$

where v_1 and v_2 are weight values trained by LDA in the same way that $\mathbf{w}$ in (1) and (2) is determined where $\mathbf{X}_T$, $\mathbf{m}_T$ and $\mathbf{X}_I$, $\mathbf{m}_I$ are replaced by the HMM and NDA scores and associated means. The scores are obtained from a group of speakers not used in the evaluation. This is a speaker independent output node of the HSV system.

4. SPEAKER VERIFICATION EXPERIMENTS

4.1. Experimental Database

The database consists of approximately 6000 connected digit utterances recorded over dialed-up telephone lines. The vocabulary includes 11 words. These are the digits "0" through "9" plus "oh". The database is partitioned into 4 subsets as shown in Table 1. There are 43 speakers in the Roster A, and 42 in Roster B. For each speaker, there are 11 5-digit utterances designated for training recorded in a single session from a single channel in A_s and B_s. These utterances are designed to have each digit appear 5 times in different contexts. Each speaker has a group of test utterances in A_m and B_m. These utterances are recorded over a series of sessions with a variety of handsets and channels. The test utterances in A_m and B_m are either fixed 9-digit utterances or randomly selected 4-digit utterances.

An SI HMM-based digit recognizer [2] is used to segment each utterance into words (digits), and to generate raw feature vectors. In the digit recognizer, 10th order autocorrelation vectors are analyzed over a 45 ms window shifted every 15 ms through the utterance. Each set of autocorrelation coefficients is converted to

Table 1: Segmentation of the Database

	Roster A	Roster B
Training utterances	A_s	B_s
Test utterances	A_m	B_m

a set of 12 cepstral coefficients from linear predictive coding (LPC) coefficients. These cepstral coefficients are further augmented by a set of 12 delta cepstral coefficients calculated over a 5-frame window of cepstral coefficients. Each "raw" data vector has 24 elements consisting of the 12 cepstral coefficients and the 12 delta cepstral coefficients [2].

4.2. NDA System Results

Experiments were conducted first to test the NDA classifier. The SI HMM's used to obtain NDA features as in (6) were trained from a distinct database of connected digit utterances. These HMM's have 6 states for words "0" through "9" and 5 states for word "o". Each state has 16 Gaussian components. So, for a 6 state HMM, the NDA features have $6 \times 16 = 96$ elements. For each true speaker in Roster A, R_T in (3) was calculated using utterances from A_s; R_I was obtained from B_s, R_{CT} from both B_s and B_m, and R_{CI} from B_m. The γ and δ parameters are not very sensitive for these data sets. To calculate (7), we use $u_{k_i} = 1$ due to a lack of training data.

The results in terms of averaged individual equal-error rates are listed in Table 2. An equal-error rate of 6.13% was obtained with NDA using both score normalization (5) and pooled covariance matrices (3). With only score normalization (5) the equal-error rate is 10.12%. Without score normalization and compensating covariance matrices (as in [1]), the equal-error rate was 18.18%.

Table 2: Results on Discriminant Analysis

Algorithms	Scores	Cov. Matrices	Eq-Er %
NDA	Normalized	Pooled	6.13
NDA	Normalized	Unpooled	10.12
LDA (as in [1])	Unnormalized	Unpooled	18.18

1,514 true speaker utterances
23,730 impostor utterances

4.3. Hybrid Speaker-Verification System Results

For the Type 1 classifier, SD HMM's were trained using the utterances in A_s. Five cohort models were constructed from utterances in Roster B. The utterances in A_m were used for testing. The S_{HMM} scores were obtained from the experiments in [2]. The NDA scores were obtained from the current experiments. To obtain the common weight values v_1 and v_2 in (8) for all speakers, both Type 1 and Type 2 classifiers were trained using the data set B_s. Then v_1 and v_2 are formed by LDA using the output scores from the data set B_m. The major results are listed in Table 3.

Table 3: Major Results

	Equal-error rates (%)	
Systems	Mean	Median
HSV with NDA	4.32	3.14
HMM-cohort	5.30	4.35
HMM	9.41	7.42
NDA	8.68	8.15

1,514 true speaker utterances
11,620 impostor utterances

With respect to storage requirements, the HMM classifier needs 51.56 Kb space per speaker for model parameters. The NDA classifier needs 4.12 Kb storage space per speaker, so the HSV system needs only slightly more storage than the HMM system.

5. REFERENCES

[1] A. R. Setlur, R. A. Sukkar, and M. B. Gandhi, "Speaker verification using mixture likelihood profiles extracted from speaker independent hidden Markov models," *Submitted to ICASSP*, 1996.

[2] A. E. Rosenberg, J. DeLong, C. H. Lee, B. H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification," in *Proceedings of the Int'l Conf. on Spoken Language Processing*, (Banff, Alberta, Canada), pp. 599–602, Oct. 1992.

[3] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. NY: John & Wiley, 1973.

[4] Z. Chair and P. K. Varshney, "Optimal data fusion in multiple sensor detection systems," *IEEE Trans. on Aerospace and Electronic Systems*, vol. AES-22, pp. 98–101, Jan. 1986.