

Improved Speech Recognition Using a Subspace Projection Approach

Philipos C. Loizou[†]

Andreas S. Spanias^{*}

[†]Department of Applied Science
Univ. of Arkansas at Little Rock
Little Rock, AR 72204-1099
(e-mail: loizou@ualr.edu)

^{*}Department of Electrical Engineering,
Arizona State University, Tempe, AZ 85287-7206

Abstract

Two class-separability criteria based on the divergence measure are proposed to improve speech recognition performance. The average and weighted average divergence measures are used as criteria for finding a transformation matrix which maps the original features into a more discriminative subspace. Results are presented for a highly confusable task.

Address correspondence to:

Philipos C. Loizou

Department of Applied Science

Univ. of Arkansas at Little Rock

Little Rock, AR 72204-1099

Phone:(501) 569-8067

Fax:(501) 569-8020

E-mail: loizou@ualr.edu

I. INTRODUCTION

The performance of speech recognition systems degrades considerably when vocabularies consist of confusable words. Various discriminative methods have been proposed in the literature to deal with confusable vocabularies [1-5]. Some of these methods used optimization criteria, other than the maximum-likelihood, to estimate Hidden Markov Model (HMM) parameters such that the separation between the competing classes was maximized [1][2]. Other methods used subspace projection approaches which mapped the feature space into a subspace by maximizing an appropriately chosen class-separability criterion [3]-[5]. One such class-separability criterion that has been used was the ratio of between-to-within class variance, i.e., the F-ratio.

In this paper, we propose the use of weighted average divergence measure as an alternative class-separability criterion. Unlike the F-ratio, the proposed method takes into account the fact that certain pairs of words or phonemes are harder to discriminate than others. In alphabet recognition for example, letter B is often confused with letter D, but is rarely confused with letter W [5]. The weighted average divergence measure is derived from the average divergence measure [6] by properly weighting the confusable pairs, i.e., the pairs which are difficult to discriminate. The weighted average divergence measure is used to find a transformation matrix which maps the original features into a more discriminative subspace. Results are presented for a highly confusable task, namely speaker-independent recognition of E-set letters recorded over the telephone.

The organization of this paper is as follows. Sections II and III describe the average and weighted average measures respectively, followed by a description of the training procedure in section IV. Section V presents our results, and section VI gives our conclusions.

II. AVERAGE DIVERGENCE

Consider two pattern classes, ω_i and ω_j , characterized by the conditional probability density functions $p_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)$ and $p_j(\mathbf{x}) = p(\mathbf{x}|\omega_j)$ respectively. Now consider a linear transformation matrix \mathbf{A} which maps the original observations \mathbf{x} into \mathbf{y} , that is

$$\mathbf{y} = \mathbf{A}^T \mathbf{x} \tag{1}$$

where \mathbf{x} is an n -dimension vector, \mathbf{y} an m -dimension vector, $m \leq n$, and \mathbf{A} is $n \times m$ matrix with m linearly independent columns. Assuming that the conditional density functions $p_i(\mathbf{x})$ are normally distributed according to $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ with means $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$, then it can be shown that the corresponding density functions $p_i(\mathbf{y})$ are also normally

distributed according to $N(\mathbf{A}^T \boldsymbol{\mu}_i, \mathbf{A}^T \boldsymbol{\Sigma}_i \mathbf{A})$. The divergence $J(i, j)$ between the two classes in the y -space is then given by [9]

$$J(i, j) = \frac{1}{2} \text{tr} [\mathbf{C}_i^{-1} \mathbf{C}_j + \mathbf{C}_j^{-1} \mathbf{C}_i] - m + \frac{1}{2} \text{tr} [(\mathbf{C}_i^{-1} + \mathbf{C}_j^{-1}) \mathbf{A}^T \boldsymbol{\Delta}_{ij} \mathbf{A}] \quad (2)$$

where $\text{tr}[\cdot]$ is the trace of a matrix, $\mathbf{C}_i = \mathbf{A}^T \boldsymbol{\Sigma}_i \mathbf{A}$, $\mathbf{C}_j = \mathbf{A}^T \boldsymbol{\Sigma}_j \mathbf{A}$ are $m \times m$ matrices, and $\boldsymbol{\Delta}_{ij} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T$ is an $n \times n$ matrix. The divergence is a measure of dissimilarity between the two pattern classes ω_i and ω_j , and therefore signifies how similar or different the two classes are. In the case that there are more than two classes, an average divergence measure (ADIV) can be defined as [6]

$$\bar{J} = \sum_{i=1}^K \sum_{j=1}^K P_i P_j J(i, j) \quad (3)$$

where K is the number of classes (e.g., words, phonemes, etc.), and P_i is the a priori probability of class i . Assuming a common covariance matrix \mathbf{V} among all classes, the average divergence reduces to

$$\bar{J} = \sum_{i=1}^K \sum_{j=1}^K P_i P_j \text{tr} [(\mathbf{A}^T \mathbf{V} \mathbf{A})^{-1} \mathbf{A}^T \boldsymbol{\Delta}_{ij} \mathbf{A}] \quad (4)$$

The average divergence \bar{J} can be considered to be a statistical measure of how clustered or dispersed the K classes are. Misclassifications often occur because the classes are closely clustered. Naturally, in order to minimize the recognition errors, the classes need to be separated from each other as much as possible. This can be done by finding a transformation matrix \mathbf{A} that will maximize the average divergence \bar{J} . Matrix \mathbf{A} can be found as follows. Let the matrix \mathbf{M} ($n \times n$) be defined as

$$\mathbf{M} = \sum_{i=1}^K \sum_{j=1}^K P_i P_j (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \quad (5)$$

then the average divergence in (4) can be written as

$$\bar{J} = \text{tr} [(\mathbf{A}^T \mathbf{V} \mathbf{A})^{-1} (\mathbf{A}^T \mathbf{M} \mathbf{A})] \quad (6)$$

It can be shown [8] that the matrix \mathbf{A} that maximizes the average divergence \bar{J} can be formed by selecting the m eigenvectors of the matrix $\mathbf{V}^{-1} \mathbf{M}$ corresponding to the m largest eigenvalues. It is noted here that the criterion \bar{J} is not the same as the Fisher's discriminant

ratio because the matrix \mathbf{M} is different from the between-class scatter matrix M_B used in the F-ratio [7]

$$M_B = \sum_{i=1}^K n_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \quad (7)$$

where n_i is the number of vectors in class i , and $\boldsymbol{\mu}$ is the total mean vector.

III. WEIGHTED AVERAGE DIVERGENCE

Each pair of classes in (3) is weighted according to the class a priori probabilities $P_i P_j$. Therefore, classes which are more likely to occur are weighted more. Although this might be the case in some applications, it is not the case in isolated word recognition. Words are chosen with equal probability, i.e., $P_i = 1/K$, and therefore all the pairs in the computation of the $\bar{\mathcal{J}}$ are weighted equally. Certain pairs, however, are more confusable than others. In digit recognition, for example, digits one and nine are often confused with each other, while digits one and six are rarely confused. Hence, in order to account for the fact that certain pair of words are harder to discriminate than others, we propose a weighted average divergence measure $\bar{\mathcal{J}}_w$ (WADIV) as

$$\bar{\mathcal{J}}_w = c \sum_{i=1}^K \sum_{j=1}^K w(i, j) \text{tr} [(\mathbf{A}^T \mathbf{V} \mathbf{A})^{-1} \mathbf{A}^T \boldsymbol{\Delta}_{ij} \mathbf{A}] \quad (8)$$

where c is a constant and $w(i, j)$ is the weight assigned to pair (i, j) . The transformation matrix \mathbf{A} that maximizes $\bar{\mathcal{J}}_w$ can be found as per previous section, however the computation of \mathbf{M} in (5) needs to be modified to incorporate the weights $w(i, j)$. By introducing the weights $w(i, j)$ we hope to enhance the discrimination of confusable pairs and thus reduce the misclassification errors. These weights need to be chosen such that more emphasis is placed on pairs which are hard to discriminate, and less emphasis on pairs which are easy to discriminate. The weights $w(i, j)$ are thus chosen as follows:

$$w(i, j) = \begin{cases} 1 & \text{if pair } (i, j) \in \mathcal{P} \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where \mathcal{P} is the set containing the pairs that need to be emphasized. Two methods are considered in this paper for constructing the set \mathcal{P} . The first method selects the most confusable pairs based on confusion statistics. The second method selects the pairs which have the smallest divergence. The divergence is used here as a means of quantifying how “confusable” two classes are. A small value of divergence indicates that the two classes are confusable. Simulation results with both methods are presented in the following sections.

IV. TRAINING

The ADIV and WADIV measures were applied in speaker-independent recognition of the E-set letters (B,C,D,E,G,P,T,V,Z) recorded over the telephone. This is a challenging task, because of the existing acoustic similarities among the letters. For that reason, the E-set task is often used as a test bed in evaluating discriminative algorithms [2][3].

E-set letters, taken from a telephone-based alphabet database [10], were used to train a phoneme-based HMM recognizer. A total of 100 speakers were used for training and testing, 50 for training and 50 for testing with equal number of female and male speakers. The original features used in training were the log-filterbank energies obtained by applying 22 mel-spaced¹ triangular windows to the magnitude spectrum. The feature vector consisted of 22 log-filterbank energies appended by a normalized energy term. These features were used to derive the transformation matrix \mathbf{A} as follows:

1. Compute the common sample covariance matrix \mathbf{V} and the sample mean vectors $\boldsymbol{\mu}_i$ ($i = 1, \dots, K$) from the training data. In our application, $K = 9$ corresponding to the nine consonants in the E-set.
2. Construct matrix \mathbf{M} as per (5), and compute the eigenvalues and eigenvectors of $\mathbf{V}^{-1}\mathbf{M}$. Use the m eigenvectors corresponding to the m largest eigenvalues of $\mathbf{V}^{-1}\mathbf{M}$ as the m columns of matrix \mathbf{A} .

Matrix \mathbf{A} was then used to project the original observations \mathbf{x} (22×1) to \mathbf{y} , of dimension $m < 22$. The projected observations \mathbf{y} , appended with a normalized energy term, were used to train phoneme-based HMMs.

There is a total of 11 phonemes in the E-set, namely /iy/,/b/,/s/,/d/,/q/,/jh/,/p/, /t/,/v/,/z/ and /pv/². Each phoneme was characterized with a 3-state continuous density HMM with diagonal covariance and one mixture per state. A 2-state HMM was used for /b/ and /q/. Context-independent models were trained by first running five iterations of the segmental k-means procedure, followed by five iterations of the Baum-Welch algorithm. The context-independent models were then used to obtain context-dependent models by running one iteration of the Baum-Welch re-estimation procedure. The context-dependent models were used for recognition. The Viterbi algorithm was employed at recognition with a bigram language model.

¹i.e., linear up to 1000 Hz and logarithmic thereafter.

²The label /pv/ is used for prevoicing, occurring sometimes in letters B, D and G.

V. RESULTS

Since the proposed approach was formulated assuming single-mixture models, we separated the training and test data based on gender. We therefore report results for male and female speakers separately. That is, we trained the HMM models using female (male) speakers and then tested them on female (male) speakers. In the last part of this section we report results obtained using both male and female speakers for training and recognition.

We first established the baseline performance, namely the performance obtained when the average divergence was not used to project the original features. Performance was computed as percent correct, i.e., as: (number of correct utterances/ total number of test utterances) $\times 100$. A total of 450 ($=50 \times 9$) E-set letters taken from the telephone-based alphabet database[10] were used for testing, half produced by female speakers and half produced by male speakers. Baseline performance of 44.88% and 36.88% was obtained with the original features, i.e., with $m = 22$, for female and male speakers respectively. It should be noted that the baseline performance was obtained using 3-state continuous density HMMs with diagonal covariance and one mixture per state. Single mixture models were used in order to make a fair comparison between the baseline approach and the proposed approach, which was formulated assuming single-mixture Gaussian densities. Results obtained with Linear Discriminant Analysis (LDA) -using the F-ratio as a class separability criterion- and ADIV approaches are given in Table 1 for various values of m . As it can be seen both methods yielded a significant improvement in recognition performance. The performance obtained with the ADIV measure was higher than the performance obtained with the LDA approach by as much as 3% for $m = 8$ and by as much as 8% for $m = 4$ (for female speakers). A smaller improvement in performance was observed for male speakers. These results clearly illustrate the effectiveness of the ADIV measure in enhancing discrimination of confusable words.

Further improvement was realized with the WADIV measure. Two methods were investigated for constructing the set \mathcal{P} . In the first method the set \mathcal{P} contained the most confusable pairs found from confusion statistics. In particular, the following 12 pairs were found: B/D, B/E, D/E, D/G, D/T, G/T, P/T, V/Z, Z/C, V/E, V/B, and V/G. These 12 pairs were used in (5) to compute the matrix \mathbf{M} . The rank of matrix \mathbf{M} was found to be 8, hence $\mathbf{V}^{-1}\mathbf{M}$ had only 8 eigenvectors corresponding to 8 non-zero eigenvalues. These 8 eigenvectors were then used to form matrix \mathbf{A} . Results obtained with this method are

given in Table 1, indicated as method 1. As it can be seen, a small improvement is observed when confusable pairs are emphasized in the computation of the average divergence. It is assumed in this method that the set of confusable pairs, i.e., the set \mathcal{P} , is known. In the second method, the confusable set \mathcal{P} is constructed automatically as follows. All possible pairs are ranked according to their divergence, and the L pairs with the smallest divergence are selected to form the set \mathcal{P} . Several experiments were run to determine a good value for L . Good performance was obtained by selecting 15 pairs with the lowest divergence, i.e., with $L = 15$. Results are given in Table 1, indicated as method 2. A small improvement was observed for male speakers. Both weighted average divergence (WDIV) methods (methods 1 and 2) yielded an improvement in performance compared to the ADIV and LDA methods. This clearly indicates the importance of weighting confusable pairs for improved discrimination.

Further overall improvement was achieved by appending to the projected feature vectors the delta coefficients and the delta energy. Results are given in Table 2, showing a big improvement in performance. This is not surprising, given that the delta coefficients are known to be very effective features in tracking dynamic spectral information, which is important in the discrimination of the E-set.

Finally, we investigated the performance of the proposed methods using both male and female speakers for training and recognition. Although the single-mixture Gaussian models were inappropriate for this case, we wanted to see whether the proposed methods would yield any improvements over the traditional LDA method. Comparative results between the four methods are given in Table 3. As shown in Table 3, the performance obtained with the divergence-based methods was higher than the performance obtained with the LDA method. Table 4 shows the results obtained after appending to the feature vector the delta coefficients and the delta energy. Finally, we compared the performance of various conventional feature representations, such as cepstrum coefficients (CEP) and mel-frequency cepstrum coefficients (MFCC), with the projected features obtained with the WADIV measure. Best results were obtained with features derived from the WADIV measure (see Table 5).

VI. CONCLUSIONS

We have presented two different class-separability criteria that could be used to enhance discrimination of confusable words. Of the two criteria, the WADIV measure yielded the

highest speech recognition performance. Furthermore, features derived from the WADIV measure were found to yield better performance than other conventional features, such as cepstrum or mel-frequency cepstrum coefficients. Features derived using the proposed criteria have also been found to perform better than features derived using LDA.

References

- [1] L.R. Bahl, P.F. Brown, P.V. Souza, and R.L. Mercer, "Maximum mutual information estimation of hidden Markov models for speech recognition," *Proc. ICASSP'86*, pp. 49-52, April 1986.
- [2] P.-C Chang and B-H. Juang, "Discriminative training of dynamic programming based speech recognizers," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 2, pp. 135-143, April 1993.
- [3] P. F. Brown, *The Acoustic-Modeling Problem In Automatic Speech Recognition*, Ph.D Thesis, Carnegie Mellon University, 1987.
- [4] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," *Proc. ICASSP'92*, pp. 13-16, 1992.
- [5] P. Loizou and A. Spanias, "High-performance alphabet recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 430-445, November 1996.
- [6] T. L. Grettenberg, "Signal selection in communication and radar systems," *IEEE Trans. on Information Theory*, vol. IT-9, pp. 265-275, October 1963.
- [7] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons: New York, 1973.
- [8] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1972.
- [9] S. Kullback, *Information Theory and Statistics*, John Wiley and Sons, New York, 1958.
- [10] R. Cole, M. Fanty and K. Roginsky, "A telephone speech database of spelled and spoken names," *Proc. of ICSLP-92*, pp. 891-893, October 1992.

Speakers	Dimension (m)	LDA	ADIV	Method 1 WADIV	Method 2 WADIV
Female	22	44.88%	44.88%	44.88%	44.88%
	8	48.44%	51.11%	52.00%	51.55%
	4	40.88%	48.00%	46.22%	37.78%
Male	22	36.88%	36.88%	36.88%	36.88%
	8	52.50%	53.33%	54.22%	54.67%
	4	50.50%	52.89%	50.67%	44.44%

Table 1: Comparative recognition performance of four different methods as a function of projected feature dimension. Dimension 22 corresponds to the case where no transformation is applied to the input feature vectors (baseline results).

Speakers	Dimension (m)	LDA	ADIV	Method 1 WADIV	Method 2 WADIV
Female	22+22+2	59.11%	59.11%	59.11%	59.11%
	8+8+2	65.44%	67.11%	65.77%	68.00%
	4+4+2	57.78%	59.55%	60.00%	56.89%
Male	22+22+2	63.11%	63.11%	63.11%	63.11%
	8+8+2	69.88%	71.11%	74.67%	72.88%
	4+4+2	64.00%	68.44%	64.88%	61.77%

Table 2: Comparative performance of LDA, ADIV and WADIV measures after appending to the projected features the delta coefficients and the delta energy.

Dimension (m)	LDA	ADIV	Method 1 WADIV	Method 2 WADIV
22	40.73%	40.73%	40.73%	40.73%
8	51.21%	53.00%	54.77%	55.61%
4	48.12%	51.68%	51.40%	52.24%

Table 3: Comparative recognition performance of four different methods as a function of projected feature dimension. Training and recognition was performed using data consisting of both male and female speakers.

Feature Dimension (m)	LDA	ADIV	WADIV
22+22+2	51.96%	51.96 %	51.96 %
8+8+2	65.33%	67.69 %	68.53 %
4+4+2	62.12%	65.73 %	65.73 %

Table 4: Comparative performance of LDA, ADIV and WADIV (method 2) measures after appending to the projected features the delta coefficients and the delta energy. Training and recognition was performed using data consisting of both male and female speakers.

Dimension (m)	LDA	CEP	MFCC	WADIV
8	51.21%	50.0 %	53.6 %	55.6 %
4	48.12%	46.9 %	50.2 %	52.2 %

Table 5: Comparative performance of various feature representations on the E-set task.