KNOWLEDGE-BASED QUADRATIC DISCRIMINANT ANALYSIS FOR PHONETIC CLASSIFICATION

Heyun Huang[1], *Yang Liu*[2], *Louis ten Bosch*[1], *Bert Cranen*[1], *Lou Boves*[1]

[1]Department of Linguistics, Radboud University Nijmegen,
Erasmuslaan 1, 6525 HT, Nijmegen, the Netherlands
[2]Department of Statistics, Yale University,
24 Hillhouse Avenue, New Haven, CT 06511, USA
{h.huang,l.tenbosch,b.cranen,l.boves}@let.ru.nl, yang.liu@yale.edu

ABSTRACT

Modeling the second-order statistics of articulatory trajectories is likely to improve the performance in classifying phone segments compared to using only linear combinations of MFCCs. Nevertheless, the extremely high dimensionality of the feature space spanned by a combination of monomials of degree-1 and degree-2 makes it difficult to effectively exploit the discriminative information in the full covariance matrix. This paper proposes a novel algorithm, dubbed Knowledge-based Quadratic Discriminant Analysis (KnQDA), for reducing the number of dimensions of the space spanned by degree-1 and degree-2 monomials by using phonetic knowledge for selecting the set of degree-2 monomials that are most likely to improve classification. KnQDA seeks a trade-off between overfitting and undertraining, which further improves the learnability. Binary classifications on all pairs of phones in TIMIT show the effectiveness of the proposed method, especially on those phone pairs that overlap strongly in the linear feature space.

***Index Terms*—** Dimensionality Reduction, Knowledge-Based Quadratic Discriminant Analysis, Phone Classification, TIMIT

1. INTRODUCTION

Stacking consecutive frames of speech parameters is perhaps the most straightforward way of capturing information about context-induced articulatory dynamics [1–4]. However, stacking frames may not be the most effective representation for the purpose of automatic speech processing, because the information is represented implicitly. Thus, some further processing is required to make the dynamics explicit [3]. Linear combinations of nearby frames [1,5], such as the velocity and acceleration (Δ, $\Delta\Delta$) of acoustic parameters (MFCC, PLP, etc.) do capture the local and short-term part of the dynamics. However, these parameters cannot capture the gestural dynamics at the level of (demi-)syllables implicit in stacks of consecutive frames. The long-term dynamics of the articulatory gestures results in non-linear correlations between frames at a distance corresponding to (demi-)syllables. Explicit modeling of second-order statistics in [6,7] and implicit modeling by means of a polynomial kernel in [8] have shown to be effective in improving the performance of phonetic classification and automatic speech recognition (ASR). Thus, it is useful to extend the original feature space with degree-2 monomials.

Stacking frames of 15 – 25 frames of MFCC (or PLP) parameters results in a highly redundant feature space (≥ 200 dimensions). Adding degree-2 monomials will make it even more difficult to find the relevant information. Moreover, a large proportion of the monomials are likely to harm a subsequent classifier, rather than help it [9]. Therefore, it is necessary to select a subset of the monomials. As an important side-effect, reducing the number of monomials reduces the model complexity [10] and may help to prevent overfitting in the full space spanned by the monomials.

In order to model the speech trajectories in MFCC stacks, this paper proposes a novel dimensionality reduction algorithm called Knowledge-based Quadratic Discriminant Analysis (KnQDA), which aims to extract the discriminative information from the high-dimensional degree-2 monomial feature space under the guidance of specific linguistic knowledge. KnQDA first utilizes the covariance estimators and linguistic knowledge to learn the interpretable and discriminative monomials from the huge space with a trade-off between overfitting and undertraining, and then generates a low-dimensional projection by optimizing the discriminant objective function.

The rest of this paper is organized as follows. Section 2 proposes our approach. Section 3 describes the data, the feature extraction procedure, and the experimental designs. Experimental results on the TIMIT [11] phonetic classification task are reported in Section 4, followed by Section 5 with general discussion and conclusion.

2. KNOWLEDGE-BASED QUADRATIC DISCRIMINANT ANALYSIS

2.1. Feature Space Spanned by Monomials

Phones are represented by a block of M consecutive N-dimensional MFCC vectors, stacked to form $d(= M \times N)$-dimensional feature vectors $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$. For the purpose of classifying two classes of phones, we model their first and second order statistics with two multivariate normal distributions: $N_1(\mu_1, \Sigma_1)$ and $N_2(\mu_2, \Sigma_2)$. The element in the ith row and jth column of Σ_k is denoted by $\sigma^k_{ij}, k = 1, 2$. Given a new token $\mathbf{x}$, the theoretical optimal

The research of Heyun Huang has received funding from European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 213850. Dr. Louis ten Bosch has received funding from the OPTIFOX project.

classifier can be derived from the logarithm of the likelihood ratio:

$$\mathbf{l}(\mathbf{x}) = \log(\frac{p(\mathbf{x}|\mu_1, \Sigma_1)}{p(\mathbf{x}|\mu_2, \Sigma_2)}) = \mathbf{x}^T\mathbf{A}\mathbf{x} - 2\mathbf{b}^T\mathbf{x} + c, \tag{1}$$

where $\mathbf{A} = \Sigma_1^{-1} - \Sigma_2^{-1}$, $\mathbf{b} = \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2$, and $c = -(\log|\Sigma_1| - \log|\Sigma_2|) + (\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_2^T\Sigma_2^{-1}\mu_2)$.

After discarding the non-discriminative term c, expansion of this optimum classifier results in:

$$\mathbf{l}(\mathbf{x}) = \sum_{i=1}^{d}\sum_{j=1}^{d} a_{ij}x_ix_j + \sum_{i=1}^{d} b_ix_i, \tag{2}$$

in which a_{ij} and b_i stand for the elements of matrix $\mathbf{A}$ and vector $\mathbf{b}$. Eq. (2) means that, provided that the means and the covariances of the distributions of the phone classes are known, a weighted linear combination of monomials of degree-1 and degree-2 in $\mathbf{x}$ provides the optimal separation of the two classes. However, the optimal weights $(\mathbf{A}, \mathbf{b})$, derived from the distributions of the two classes, are difficult to estimate, because estimating the means and the full covariance matrices will require an amount of training data that is seldom available, the more so if the observation vectors consist of some 20 frames of each about 13 parameters. Therefore, ways for estimating optimal weights from a realistic amount of training data must be developed, which is the focus of the remainders of this paper: estimating the optimal weights $(\mathbf{A}, \mathbf{b})$ to construct the quadratic surface that separates the classes.

2.2. Learning Parameters of the Quadratic Hyperplane

From Eq. 2 it can be seen that the number of parameters needed to specify the optimal (quadratic) surface is $\frac{1}{2}d(d+1) + d$ (the covariance matrix is symmetric). In the case of stacks of some 20 MFCC frames this number is comparable to or even larger than the numbers of training tokens of some phonetic classes in the TIMIT corpus [11]. Using that high-dimensional parametric space to fit the training data probably minimizes the training error, but inevitably loses generalization capacity. Therefore, to achieve a balance between the complexity of the classifier and training error minimization [10], we must constrain the number of parameters in $(\mathbf{A}, \mathbf{b})$. This reduction was also addressed for modeling the covariance matrix in [12] and the second-order statistical features in [6].

As mentioned in Section 1, there probably exist a huge number of harmful monomials in the feature space spanned by stacking MFCCs. However, linguistic knowledge can be used to indicate the subset of the monomials whose corresponding x_i and x_j are inherently correlated. In the next subsection, we show how to select those relevant monomials to reach the balance between the complexity of the classifier and minimization of the training error, which is crucial to the subsequent discriminant analysis.

2.2.1. Knowledge-based removal of irrelevant Covariance terms

Linguistic knowledge predicts that neighboring frames will be strongly correlated, while more distant frames are probably uncorrelated, which means that the elements on the ith row and jth column of Σ_1^{-1} and Σ_2^{-1} are also approximately identical [13] and thus x_ix_j does not make any contribution to the optimal classifier (Eq. (1)). Therefore, a_{ij} in Eq. (2) becomes approximately zero and x_ix_j should not be involved in the discriminant analysis.

To realize this idea, the vector $\mathbf{x}$ should be reformatted as a matrix $\mathbf{X}$ with the size as $M \times N$. The time indices of elements x_ix_j in $\mathbf{x}$ are indicated by $x_{m_in_i}x_{m_jn_j}$. We use $\mathbf{G_{ij}^{(1)}}$ to indicate whether x_ix_j should be excluded:

$$\mathbf{G_{ij}^{(1)}} = \begin{cases} 0 & \text{if } |m_i - m_j| > \eta_t, \\ 1 & \text{otherwise.} \end{cases} \tag{3}$$

This monomial selection strategy implies that if x_i and x_j come from two distant frames, the corresponding monomial will be discarded.

2.2.2. Data-based removal of irrelevant Covariance terms

While most irrelevant monomials can be excluded from the complete set by applying the function $\mathbf{G^{(1)}}$, the dimension of the reduced feature space might still be too high to handle in standard classifiers. Therefore, it is important to be able to control the selection of the monomials with the most discriminative power [10].

Consider the monomial x_ix_j: the expected value of its mean is $\mathbf{E}(x_ix_j) = \hat{\sigma}_{ij} + \mathbf{E}(x_i)\mathbf{E}(x_j)$. Therefore, we can use the difference of corresponding $\hat{\sigma_{ij}}$s in two classes to predict the difference of their mean vectors $\mathbf{E}^{(1)}(x_ix_j) - \mathbf{E}^{(2)}(x_ix_j)$, which reflects the discriminative ability of the monomial x_ix_j. Concretely speaking, whether a monomial should be kept depends on the difference between the σs of corresponding elements in the two classes:

$$\mathbf{G_{ij}^{(2)}} = \begin{cases} 1 & \text{if } |\hat{\sigma}_{ij}^{(1)} - \hat{\sigma}_{ij}^{(2)}| > \alpha * \max_{ij}|\hat{\sigma}_{ij}^{(1)} - \hat{\sigma}_{ij}^{(2)}|, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

If an element of the matrix $\mathbf{G_{ij}^{(2)}}$ is one, the corresponding monomial a_{ij} is kept; if it is zero, the corresponding a_{ij} is discarded. In Eq. (4) α varies from 0 to 1, which allows selecting the monomials according to the contribution of the difference between corresponding correlation coefficients to the discrimination between the classes. When α increases, the number of selected monomials will be reduced, resulting in a reduction of the model complexity.

2.3. The Algorithm for Selecting Monomials

Here, we sketch the algorithm for selecting the relevant monomials according to Eq. (3) and Eq. (4). Given a training set with n observations $\mathbf{x}_i$, $(i = 1, 2, \ldots, n)$, each observation is assigned a label $c_{x_i} = c \in \{1, 2\}$. The number of observations in class c is denoted by n_c. The proposed algorithm, dubbed Knowledge-based Quadratic Discriminant Analysis (KnQDA), is described as follows:

- Map each vector $\mathbf{x}$ in the training set via the second-order polynomial kernel:

$$\phi(\mathbf{x}) = \{x_1, \cdots, x_d, x_1^2, x_1x_2, \cdots, x_{d-1}x_d, x_d^2\}. \tag{5}$$

- Given the parameters α, η_t and the estimated parameters of Gaussian distributions of two classes, formulate the matrix $\mathbf{G}$ as the dot product of $\mathbf{G^{(1)}}$ and $\mathbf{G^{(2)}}$ to indicate the selected degree-2 monomials, while keeping all degree-1 monomials.
- Select the subset of $\phi(\mathbf{x})$ according to $\mathbf{G}$ to generate a vector with reduced dimensionality which we will denote by $\mathbf{z}$.This results in a vector containing all degree-1 monomials, augmented with the degree-2 monomials that are most likely to improve the classification.

2.4. Discriminant Analysis

We use the augmented vectors $\mathbf{z}$ as input to the conventional Fisher Discriminant Analysis (FDA) for a two-class problem [5]. This requires finding the linear transformation that maximizes the ratio of the trace of between class variance to the trace of within class variance:

$$\arg\max_{\mathbf{w}} \frac{tr(\mathbf{w}^T\mathbf{S}^b\mathbf{w})}{tr(\mathbf{w}^T\mathbf{S}^w\mathbf{w})}. \tag{6}$$

In Eq. (6), $\mathbf{S}^{(w)} = \sum_{j=1}^{2}\sum_{c_{z_i=j}}(\mathbf{z}_i - \mu_j)(\mathbf{z}_i - \mu_j)^T$ is the *within-class scatter matrix* and $\mathbf{S}^{(b)} = \sum_{j=1}^{2}\frac{n_c}{n}(\mu_j - \mu)(\mu_j - \mu)^T$ is the *between-class scatter matrix*, where μ denotes the overall mean and μ_j $(j = 1, 2)$ denote the mean vectors of two classes. With the projection vector $\mathbf{w}$, the classification of a test vector $\mathbf{z}$, comprising the monomials selected by $\mathbf{G}$, can be performed by judging the sign of $\mathbf{w}^T\mathbf{z}$.

2.5. Comparison with Other Second-Order Methods

The proposed algorithm is related to other discriminant analysis methods that can be also regarded as estimators of $(\mathbf{A}, \mathbf{b})$.

- **Fisher Discriminant Analysis (FDA) [5]**: Conventional FDA does not involve correlation terms; it sets all $a_{ij} = 0$, and then optimizes $\mathbf{b}$. This implies that $\Sigma_1 = \Sigma_2 = S^{(w)}$.
- **Quadratic Discriminant Analysis (QDA) [14]**: QDA computes $(\mathbf{A}, \mathbf{b})$ by Eq. (1) with the maximum likelihood estimators: $(\hat{\mu}_1, \hat{\Sigma}_1)$ and $(\hat{\mu}_2, \hat{\Sigma}_2)$.
- **Regularized Discriminant Analysis (RDA) [14]**: RDA smoothes the covariance estimators of FDA and QDA by:

  $$\hat{\Sigma}_j^{RDA} = (1-\lambda)S^{(w)} + \lambda\hat{\Sigma}_j, j = 1, 2. \tag{7}$$

  RDA adopts the same mean estimators as QDA.
- **Kernel Discriminant Analysis (KDA) [15]**: KDA with the second-order polynomial kernel implicitly estimates $(\mathbf{A}, \mathbf{b})$ without any constraint.

3. EXPERIMENTAL SETUP

3.1. The Data: TIMIT

For our experiments we used the TIMIT database. We used the standard NIST training sets (excluding the 'sa' utterance), the core test set [11], and the development set [16] for training, evaluating the performance, and tuning the parameters, respectively. The 64 phone labels were merged into 48 classes according to [17] and the glottal stops are excluded from the corpus.

3.2. Classification Task

Binary classifications between all pairs of phones are performed with a k-Nearest-Neighbor (kNN) classifier and the feature vectors formulated by stacking MFCCs after dimensionality reduction by means of the approaches described above. Actually, the multi-class classification task can be effectively decomposed into several binary classification tasks [6, 8] and multi-class performance depends crucially on the performance of binary classifications. Furthermore, the parameters involved in binary classifications are likely to be more interpretable and informative to ASR systems than those involved with the whole multi-class classification task, which is another reason of using binary classification task for performance evaluation in this paper.

Table 1: Performance Comparison Among Different Discriminant Analysis Methods on the Most Difficult Tasks (≤ 0.90) and the Less Difficult Tasks (≤ 0.95).

Subset	FDA	KDA	QDA	RDA	KnQDA
C: ≤ 0.90	86.14	85.66	82.24	86.62	**88.02**
D: ≤ 0.95	90.97	91.10	87.66	91.42	**92.83**
D - C	92.26	92.60	89.14	92.71	**94.12**

To demonstrate the effectiveness of the proposed method, we compare its classification accuracy with the results obtained with the four aforementioned discriminant analysis algorithms: FDA [5], KDA [15], QDA [14] and RDA [14]. The classification accuracy will be reported on the core test set, with the parameters optimized on the development set. For example, the parameters η_t, α and k_{KNN} in the kNN classifier are jointly tuned for KnQDA, while only one parameter k_{KNN} is tuned for FDA and QDA, and two parameters k_{KNN}, λ for RDA.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1. Performance of Crucial Binary Classifications

Excluding the silences, there are in total 946 binary pairs of phones. A large number of phone pairs are easy to separate with the original features. For example, confusions between a vowel and a plosive are very rare. Therefore, we focus on classifying confusable pairs. We define the highly confusable subsets as those pairs for which the classification accuracy is lower than some threshold. Subset C is chosen with the thresholds at 0.90 (representing "most difficult task"), which includes, among others, the vowel pair /ix/ and /ih/ and all nasals. Subset D (the "less difficult task") contains the pairs for which the classification accuracy is lower than 0.95. It includes, for example, the consonant pair /b/ and /d/. The average classification accuracy (defined as the ratio of the number of correctly classified tokens and the total number of tested tokens in all cases of a subset) obtained for these subsets is given in Table 1. Set C and D have 65 and 156 pairs of phones, respectively. The last row in the Table shows the accuracy for the pairs with classification accuracy between 0.90 and 0.95, denoted by "D-C".

From Table 1 it can be seen that RDA and KnQDA outperform FDA in all cases, which suggests that there is some discriminative information in the covariances. However, the performance of the other two methods that use statistics of degree-2 monomials, KDA and QDA, falls below the classification accuracy obtained with FDA. Most probably, this is due to the overfitting of training data with a relatively small training sample size. This confirms the prediction in section 2.2 that the complexity of the classifier, represented by the number of parameters to be estimated, is too high to achieve generalization capacity.

It is worth mentioning that the superiority of KnQDA over RDA can be explained by the more direct way to consider the structural risk minimization: when λ in Eq. (7) goes to 0, the elements of matrix $\mathbf{A}$ in Eq. (1) will approach zero and the model complexity will be reduced. However, the training error is crucially dependent on the discriminability of the features, and it cannot be guaranteed

Table 2: The Number of Phonetic Pairs with That of Non-Vowel Pairs in Set C and D-C Won by Four Competitive Methods

Subset	FDA	KDA	RDA	KnQDA
C	12(1)	7(3)	16(1)	**37(11)**
D-C	8(3)	35(18)	20(8)	**41(15)**

Table 3: Performance Comparison Among Different Discriminant Analysis Methods on all Phone Pairs from One Broad Phonetic Class

Broad Class	FDA	KDA	QDA	RDA	KnQDA
Plosives	93.22	93.32	90.64	94.26	**95.44**
Fricatives	96.11	95.66	92.98	96.34	**96.69**
Nasals	83.00	82.81	84.08	85.63	**88.76**
Semi-Vowels	92.36	92.05	89.93	93.21	**94.84**
Vowels	92.98	93.02	88.85	93.02	**93.15**
Diphthongs	96.45	95.66	94.89	**96.52**	95.50

that RDA finds the best features.

Table 2 shows for how many phone pairs each of four competitive methods (excluding QDA) perform best in the sets C and D-C. The proposed KnQDA performs best for most of the pairs in both sets. The numbers in parentheses indicate the non-vowel pairs. In the most difficult set C, KnQDA performs best in the majority of the non-vowel pairs, which might imply that the feature trajectories in the consonants can be better captured using the monomials $x_i x_j$.

4.2. Classification within a Broad Phonetic Class

In this section, we compare the performance of the five classifiers on phone pairs from the same broad phonetic class [18]. The experimental results are given in Table 3. It can be seen that KnQDA outperforms the competing methods (sometimes substantially) for all broad phonetic classes, except for the diphthongs.

For the non-vowel sounds it would seem that $\mathbf{G}^{(1)}$ effectively creates a sequence of short-span trajectories that are concatenated in a 23-frame block. This helps in capturing fast dynamics, as in plosives, but less so for slow dynamics, as in diphthongs. This would also explain the fact that the advantage of KnQDA for classifying vowel pairs (where the role of dynamic trajectories is minimal) is quite small. The finding that KnQDA is outperformed by FDA for the diphthongs may be due to the small number of training tokens. Especially /aw/ and /oy/ have very small sample-to-dimension ratios. Therefore, adopting the degree-2 monomials inevitably overfits the training data.

5. CONCLUSIONS

In this paper we have proposed and tested an effective way to identify the degree-2 monomials that hold most promise for improving TIMIT phone classification. The complete set of degree-2 monomials added to the original features results in an extremely redundant feature space. To alleviate the overfitting problem in a quadratic classifier, the proposed KnQDA method uses a combination of data- and knowledge-driven techniques for identifying the set of degree-1 and degree-2 features that provide the optimal balance between classifier complexity and training error.

The superior performance of KnQDA and RDA in experiments on the TIMIT corpus in which we performed binary classification of all possible phone pairs (except silence segments) have shown that second-order statistics features indeed improve classification performance. The inferior performance of the approaches using all degree-2 monomials, such as QDA, confirms that the feature selection is indeed necessary to avoid overfitting. By design, RDA is always better than FDA. It appears that RDA, although less effective in finding the most discriminative features than KnQDA when there is a sufficient amount of training data, can outperform KnQDA if the amount of training data becomes very small, as is the case with the diphthongs in TIMIT.

6. REFERENCES

[1] H. Huang, Y. Liu, J. Gemmeke, L. ten Bosch, B. Cranen, and L. Boves, “Globality-locality consistent discriminant analysis for phone classification,” in *Proc. of Interspeech*, 2011.

[2] J. Gemmeke, T. Virtanen, and A. Hurmalainen, “Exemplar-based sparse representations for noise robust automatic speech recognition,” *IEEE Transactions on Audio, Speech and Language processing*, vol. 99, 2011.

[3] Y. Han, J. de Veth, and L. Boves, “Trajectory clustering for solving the trajectory folding problem in automatic speech recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15:4, pp. 1425 – 1434, 2007.

[4] S. Serena, M. Magimai.-Doss, J. Pinto, and H. Bourlard, “Posterior features for template-based asr,” in *Proc. of ICASSP*, 2011.

[5] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Ann. Eugen.*, vol. 7, pp. 179–188, 1936.

[6] R. Rifkin, K. Schutte, M. Saad, J. Bouvrie, and J. Glass, “Noise robust phonetic classification with linear regularized least squares and second-order features,” in *Proc. of ICASSP*, 2007.

[7] M. Tahir, R. Schlueter, and H. Ney, “Log-linear optimization of second-order polynomial features with subsequent dimension reduction for speech recognition,” in *Proc. of Interspeech*, 2011.

[8] P. Clarkson and P. Moreno, “On the use of support vector machines for phonetic classification,” in *Proc. of ICASSP*, 1999, pp. 585–588.

[9] V. Brailovsky, O. Barzilay, and R. Shahave, “On global, local, mixed and neighborhood kernels for support vector machines,” *Pattern Recognition Letters*, pp. 1183–1190, 1999.

[10] V. Vapnik, “An overview of statistical learning theory,” *IEEE Trans. on Neural Network*, pp. 988 – 999, 1999.

[11] L. Lamel, R. Kassel, and S. Seneff, “Speech database development: Design and analysis of the acoustic-phonetic corpus,” in *Proc. of the DARPA Speech Recognition Workshop*, 1986.

[12] S. Axelrod, R. Gopinath, P. Olsen, and K. Visweswariah, “Dimensional reduction, covariance modeling, and computational complexity in asr systems,” in *Proc. of ICASSP*, 2003.

[13] A. Dempster, “Covariance selection,” *Biometrics*, vol. 28, pp. 157–175, 1972.

[14] J. Friedman, “Regularized discriminant analysis,” *Journal of the American Statistical Association*, pp. 165 – 175, 1989.

[15] A. Kocsor, “On kernel discriminant analyses applied to phoneme classification,” in *Proc. of ISNN*, 2005.

[16] A. K. Halberstadt, “Heterogeneous acoustic measurements and multiple classifiers for speech recognition,” *Ph.D. Thesis, MIT*, 1998.

[17] K. F. Lee and H. W. Hon, “Speaker-independent phone recognition using hmms,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[18] T. Reynolds and C. Antoniou, “Experiments in speech recognition using a modular mlp architecture for acoustic modelling,” *Information Sciences*, pp. 39 – 54, 2003.