Class Definition in Discriminant Feature Analysis

Jacques Duchateau, Kris Demuynck, Dirk Van Compernolle, Patrick Wambacq

Katholieke Universiteit Leuven - ESAT
Kardinaal Mercierlaan 94
B-3001 Heverlee, Belgium
E-mail: Jacques.Duchateau@esat.kuleuven.ac.be

Abstract

The aim of discriminant feature analysis techniques in the signal processing of speech recognition systems is to find a feature vector transformation which maps a high dimensional input vector onto a low dimensional vector while retaining a maximum amount of information in the feature vector to discriminate between predefined classes.

This paper points out the significance of the definition of the classes in the discriminant feature analysis technique. Three choices for the definition of the classes are investigated: the phonemes, the states in context independent acoustic models and the tied states in context dependent acoustic models.

These choices for the classes were applied to (1) standard LDA (linear discriminant analysis) for reference and to (2) MIDA, an improved, mutual information based discriminant analysis technique. Evaluation of the resulting linear feature transforms on a large vocabulary continuous speech recognition task shows, depending on the technique, the best choice for the classes.

1. Introduction

It is known from information theory that creating new information by transforming data is impossible. Transformations can only lead to information loss [1], so at the first sight it seems that they should be avoided.

However, in speech recognition systems, the dimensionality of the input data – typically filterbank outputs and their time derivatives – is too large to allow direct modelling.

Therefore, in the past years, many research groups investigated transformations that reduce the dimensionality of a feature vector with a minimal loss in information, using standard linear discriminant analysis (LDA) [2, 3] and recently also more elaborate techniques [4, 5, 6, 7, 8, 9].

The aim of discriminant feature analysis is to find the transformation that retains in the remaining features a maximal ability to distinguish the predefined classes.

The topic of this paper is the definition of these classes. At the first sight, the phoneme seems to be the right choice for speech recognition: in phonetics the phoneme set is defined as the minimal set of speech sounds that are necessary to distinguish between all pairs of words. Therefore, if the phonemes can be distinguished, all words can be recognised.

However, other choices for the classes can also be justified. In MIDA, the mutual information based discriminant feature analysis that we proposed in [6], a feature transform is derived that keeps the distance between the original and the transformed conditional class probability functions minimal. As the probability functions used in context dependent acoustic modelling correspond to the tied states, those tied states seem to be an obvious choice to define the classes in the discriminant feature analysis as well.

In this paper, three choices for the definition of the classes – the phonemes, the states in context independent acoustic models and the tied states in context dependent acoustic models – are applied in combination with two discriminant analysis techniques, namely standard LDA and MIDA [6].

In section 2, the investigated techniques for discriminant feature analysis are briefly reviewed. Section 3 describes the experimental setup (signal processing, acoustic model development, recognition task). The experimental results are given and discussed in section 4 and finally in section 5 the conclusions from the proposed work are given.

2. Discriminant Feature Analysis Techniques

2.1. Standard LDA

Linear discriminant analysis (LDA [10]) is a common data driven method that searches for a linear transformation $T : x \rightarrow y, \mathbb{R}^q \rightarrow \mathbb{R}^r$ which maps a $q$ dimensional input vector $x$ onto an $r$ ($r < q$) dimensional vector $y = Tx$ while retaining a maximum amount of class discrimination information.

In order to do so LDA needs the individual class density functions $f(x | c_i)$ and the a priori class probabilities $W_i = p(c_i)$ for all $N_c$ classes $c_1, \ldots, c_{N_c}$. In that, LDA as-

---

* Supported by the research fund of the K.U.Leuven
The main idea behind MIDA is the following. In the same paper we therefore proposed MIDA, a transformation into two scatter matrices only. LDA thus ignores the individual means and the individual scatter matrices, 

\[ \Sigma_i = \sum W_i \Sigma_i \]  

with \( \mu_i \) and \( \Sigma_i \) the mean and covariance of the gaussian.

Next this information is condensed into the following two scatter matrices: (1) the within class scatter matrix 

\[ \Sigma_w = \sum W_i \Sigma_i \]  

and (2) the between class scatter matrix 

\[ \Sigma_b = \sum W_i (\mu_i - \mu) (\mu_i - \mu)^T \]  

with \( \mu = \sum W_i \mu_i \), or alternatively the overall scatter matrix 

\[ \Sigma = \Sigma_b + \Sigma_w \]  

LDA then maximises the following ratio of determinants:

\[ T = \arg\max_T \frac{|T^T \Sigma_a T'|}{|T^T \Sigma_w T'|} = \arg\max_T \frac{|T^T \Sigma_b T'|}{|T^T \Sigma_w T'|} \]

As the determinant of a covariance matrix is related to the volume spanned by the corresponding gaussian, we thus have to minimise the volume spanned by the classes while maximising the volume between the classes, thus maximising the average distance between the classes.

It can be shown that a solution to the above maximisation can be found by solving the generalised eigenvalue problem 

\[ \Sigma_a V = \Sigma_a V \Delta \]  

with \( \Delta \) a diagonal matrix containing the eigenvalues of \( \Sigma_a \) and \( \Sigma_b \). The eigenvectors corresponding to the r largest eigenvalues then make up the rows of matrix \( T \).

2.2. MIDA

The main problem with LDA is that it condenses all information into two scatter matrices only. LDA thus ignores the individual means and the individual scatter matrices of the classes. As explained in detail in [6], this is the reason why LDA can result in suboptimal transformations. In the same paper we therefore proposed MIDA, a mutual information based discriminant analysis technique. The main idea behind MIDA is the following.

The conditional class probability functions 

\[ p(c_i | x) \]

in the original q dimensional feature space contain all relevant available information concerning the classification problem. In order to reduce the dimensionality of the input feature vector with minimal loss of information, the transformed conditional class probability functions should resemble the original ones as good as possible. In other words, the distance between the original and the transformed probability functions should be minimal.

Using the Kullback Leibler distance measure [1], following results were obtained:

\[ D(p(C | X) \, || \, p(C | Y)) = \int_{\mathbb{R}^q} \sum_{i=1}^{N_i} f(c_i, x) \log_2 \frac{p(c_i | y)}{p(c_i | y = T x)} \, dx \]

\[ = \int_{\mathbb{R}^q} \sum_{i=1}^{N_i} f(c_i, x) \log_2 (p(c_i | y = T x)) \, dx \]

\[ - \int_{\mathbb{R}^q} \sum_{i=1}^{N_i} f(c_i, y) \log_2 (p(c_i | y)) \, dy \]

\[ = H(C | Y) - H(C | X, Y) \]

Finding a transformation \( T : \mathbb{R}^q \rightarrow \mathbb{R}^r \) that minimises the Kullback Leibler distance is thus equal to finding a transformation that minimises the conditional class entropy \( H(C | Y) \). As \( x \) uniquely determines \( y \) \((y = T x)\) we can also write:

\[ D(p(C | X) \, || \, p(C | Y)) = H(C | Y) - H(C | X, Y) \]

\[ = I(C ; X | Y) \]

The transformation thus also has to minimise the conditional mutual information \( I(C ; X | Y) \), i.e. the original feature vector \( x \) should provide minimal additional information on the classes on top of the reduced feature vector \( y \).

This mutual information based optimisation problem can be worked out – using some approximations to make the calculations computationally tractable – for class density functions which are (mixtures of) gaussian densities. As a closed form solution to the optimisation problem was not found, we resorted to numerical optimisation techniques, splitting the transformation matrix \( T \) in a sequence of elementary Givens rotations. For more details, the reader is referred to [6, 11].

3. Experiments

3.1. Signal Processing

The input feature vectors – one vector for each 30 msec frame in 16 kHz data, with a 10 msec frame shift – for the feature transforms consist of a set of 24 mean normalised MEL warped spectral coefficients with their first and second order time derivatives, a total of 72 features. As the time derivatives are part of the input vector for the feature transformations, they will be incorporated automatically into the reduced output vector according to their relative importance.

The size of the output vectors of the feature transforms was set to three different values: 39, 25 and 16.
The resulting features are decorrelated using the algorithm described in [12] (with the gaussians in the acoustic models as classes).

3.2. Acoustic Model Development

The context dependent acoustic modelling used to evaluate the different feature transforms is based on a phoneme set with 38 three state phonemes and one noise state. The cross word context dependency is defined with a global phonetic decision tree which results in 575 tied states. The same decision tree was used for all experiments. The number of tied states (and the number of gaussians) for the experiments is rather small due to the size of the training database, namely 6 hours of speech.

The different class choices in the discriminant feature analysis are (1) the 39 phonemes, (2) the 115 context independent states, (3) the 575 context dependent tied states used in the acoustic modelling and (4) the 2246 tied states obtained from a larger decision tree. The latter class choice was only used in standard LDA.

The acoustic models with partially tied gaussians are developed in a similar way as described in [13]. First, for each of the tied states a small number of gaussians is initialised, this number being proportional to the occupancy of the state. Next, all these small sets of gaussians are put together in a large set of 10k gaussians. Then this large set is used to construct a fully tied acoustic model. Finally, this model is trained and the number of gaussians per tied state is reduced using the occupancy criterion as described in [14]. The average number of gaussians per state in the final acoustic models is about 100, 85 and 65 for feature vector sizes 39, 25 and 16 respectively.

The reason why acoustic models with partially tied gaussians are used is that they give better recognition results than models without tying of the gaussians. The average relative improvement in word error rate (WER) of models with tied gaussians over models without tied gaussians is about 10%, even when taking into account the small improvement in recognition results that can be obtained for modelling with 20k non tied gaussians instead of 10k non tied gaussians (the models with tied gaussians do not benefit from an increase in the number of gaussians).

3.3. Recognition Task

We evaluated the different feature transforms on a speaker independent continuous speech recognition task in Dutch, with a 40k word vocabulary and a trigram language model. The test set has the following properties: 12 speakers with 25 sentences each, 3596 words and 20 minutes of speech in total, a 3.5% OOV rate and a test set perplexity of 128.9.

With a similar type of acoustic modelling, a WER of 8.2% was found on the well known speaker independent Wall Street Journal (WSJ) recognition task for the November 92 evaluation test set (standard trigram language modelling, 20k word vocabulary, 1.9% OOV rate).

There are two main reasons for the higher absolute word error rates reported in this paper. The first is typical for languages with compound words like German and Dutch: the OOV rate is high and each OOV word results on average in almost 2 errors. As the described experiments focus on the acoustic modelling, no specific effort was done to tackle this problem (for instance with post processing on the recognised sentences). The second reason is the size of the acoustic model training database, which was chosen small: 6 hours of speech compared to 69 hours for WSJ. This size of the database keeps the computation time for a large number of experiments feasible, while still making comparisons between the results possible.

To obtain the reported results for 39 features, our recognition system works in real time on a single 733 MHz Pentium III processor running Linux. Note that smaller feature vectors result in slower recognition as less accurate acoustic models result in a more complex search for the optimal word string.

4. Results

The results of the comparison between the investigated choices for the classes in the discriminant feature analysis are summarised in table 1, both for standard LDA (at the top of the table) and for MIDA (at the bottom).

As explained, the number of classes in the discriminant analysis was set to 39, 115, 575 and 2246 (the latter for standard LDA only), and the resulting linear feature transforms reduce the number of features from 72 to 16, 25 and 39.

<table>
<thead>
<tr>
<th>Number of features</th>
<th>Number of classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>115</td>
</tr>
<tr>
<td>16</td>
<td>23.1%</td>
</tr>
<tr>
<td>25</td>
<td>18.5%</td>
</tr>
<tr>
<td>39</td>
<td>17.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of features</th>
<th>Number of classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>115</td>
</tr>
<tr>
<td>16</td>
<td>18.1%</td>
</tr>
<tr>
<td>25</td>
<td>15.3%</td>
</tr>
<tr>
<td>39</td>
<td>15.1%</td>
</tr>
</tbody>
</table>

Table 1: Evaluating different definitions of the classes in LDA and MIDA (WER given)

From the table it is clear that the worst results are found when the phoneme is chosen as class in the discriminant analysis, especially for standard LDA. Increas-
ing the number of classes improves the result until a number of classes is reached that seems to be sufficient for the discriminant analysis technique. For standard LDA the result doesn’t improve beyond 575 classes, for MIDA 115 classes is sufficient.

When comparing standard LDA with MIDA, it can be seen that MIDA always outperforms standard LDA, as we already concluded in [6]. MIDA even seems to be able to compress the same amount of information in 25 features as standard LDA in 39 features.

However it is not clear from the above experiments why the result improves with more classes: more classes also means more gaussians and thus a better modelling of the acoustic data. Therefore the reason for the improvement with more classes can be twofold: or the discriminant analysis knows better what to discriminate (more classes) or it gets a better idea about the data distribution in the feature space (more gaussians).

We investigated which reason for the improvement is more important through a further experiment with MIDA. An advantage of MIDA is that the classes can be modelled with mixtures of gaussians (as opposed to standard LDA which uses a single gaussian per class). This way, the better modelling of the acoustic data is obtained with a small number of classes.

In practice, we modelled the 39 phoneme classes in the MIDA with 115 gaussians, putting together in the mixture for one phoneme the gaussians that model the corresponding context independent states. If the same result as for 115 classes in the MIDA is found, then the better modelling with more gaussians is important. However if the same result is found as for 39 classes modelled with a single gaussian, then the discriminant analysis benefits from a better knowledge about what to discriminate when the number of classes increases.

Unfortunately, the outcome of this experiment was unclear: a WER of 17.1%, 15.2% and 15.4% was found for a final number of 16, 25 and 39 features respectively. For 16 and 25 features, the result is close to the result for MIDA with 115 classes, but for 39 features, the result is not better than the result for 39 classes (modelled with a single gaussian).

5. Conclusions

This paper investigated the significance of the definition of the classes in discriminant feature analysis techniques for speech recognition.

It was found that increasing the number of classes improves (or at least does not deteriorate) the recognition result. The 39 phonemes as classes is a bad option for both standard LDA and for MIDA. For standard LDA 575 tied states are sufficient, the MIDA technique can do with 115 context independent states as classes.

6. References