LDA Based Feature Estimation Methods for LVCSR

Janne Pylkkönen

Adaptive Informatics Research Centre
Helsinki University of Technology, Finland
janne.pylkkonen@hut.fi

10.21437/Interspeech.2006-129

Abstract

Features that model temporal aspects of phonemes are important in speech recognition. One method is to use linear discriminant analysis (LDA) to find discriminative features from a spectro-temporal input formed by concatenating consecutive frames of short-time spectrum features. Others use e.g. neural networks to process longer span spectral segments to improve recognition accuracy. Still the most widely used method for including temporal cues is to augment the short-time spectral features with simple time derivatives.

In this paper a new feature estimation method based on pairwise linear discriminants is presented. We compare it and some of its variants to traditional MFCC features and to LDA estimated features in a large vocabulary continuous speech recognition (LVCSR) task. The features obtained with the new estimation method show significant improvements in recognition accuracy over MFCC and LDA features.

Index Terms: speech recognition, feature extraction, linear discriminant analysis, spectro-temporal features

1. Introduction

It is well known that acoustic cues about the identity of a phoneme are spread in time around the actual phone. Yang *et al.* [1] showed that for the best frame accuracy in phoneme recognition the inspection window should span about 200 ms. In features traditionally used for speech recognition the time information is usually taken into account only by augmenting the short-time spectral features like mel-frequency cepstral coefficients (MFCC) with their first and second order time derivatives. This way the features span around 100 ms in time, but the way the time information is used is very restricted.

There have been several attempts to derive more general features for speech recognition which could better use the information in time domain. One of the first approaches was to use linear discriminant analysis (LDA) to combine features in several time frames into one reasonable size feature vector [2, 3]. More recently nonlinear methods like multi-layer perceptrons (MLP) have been used for feature extraction with larger time spans [4]. Surprisingly, it is not that easy to improve the MFCC features for speech recognition. Even the MLP features need to be combined with traditional spectral features for the best performance.

In this paper we experiment with different methods for estimating linear filters for feature extraction. These methods use as input a window of several short-time spectrum vectors, therefore allowing flexible modeling of spectro-temporal patterns. It is shown that already a traditional LDA can improve the performance over the MFCC features in a large vocabulary continuous speech recognition (LVCSR) task. We also present a new feature estimation method based on pairwise linear discriminants and experiment with some of its variants. The features obtained with the new method are shown to outperform both the MFCC and LDA estimated features.

2. Linear feature estimation methods

The goal of this research was to study linear feature extraction methods that can utilize the spectro-temporal patterns of speech signal. These methods can be viewed as different ways to estimate linear filters that are applied to a supervector containing several frames of short-time spectral features such as filter-bank energies. Let $x(t)$ denote a column vector of short-time spectral features at time instance t. The methods presented here operate on a supervector constructed as

$$X(t) = \begin{bmatrix} x(t-\Delta t) \\ \vdots \\ x(t-1) \\ x(t) \\ x(t+1) \\ \vdots \\ x(t+\Delta t) \end{bmatrix}. \quad (1)$$

This way the features can have a time span of $2\Delta t + 1$ frames.

The feature filter estimation methods give as a result a projection matrix A. The final observation features $o(t)$ are then obtained simply by multiplying the supervector with the projection matrix:

$$o(t) = AX(t). \quad (2)$$

The dimension of $o(t)$ is usually much lower than that of $X(t)$. In this work the dimension of $X(t)$ was 315 and the dimension of the final features $o(t)$ was 39.

Note that also the traditional MFCC features can be presented in this formalism. In that case, $x(t)$ is a vector of logarithmic mel-spaced filter bank energies, and the projection matrix A contains the coefficients for the cosine transformation and computation of the time derivatives.

2.1. Linear discriminant analysis

Linear discriminant analysis is a well known method for estimating a linear subspace with good discriminative properties. The idea is to find a projection of the data where the variance between the classes is large compared to the variance within the classes. Under assumptions of Gaussian class distribution and a common within-class covariance matrix this can be stated formally as finding a

projection matrix θ that maximizes the quotient

$$J(\theta) = \frac{\det(\theta \Sigma_b \theta^T)}{\det(\theta \Sigma_w \theta^T)}, \quad (3)$$

where Σ_b is the between-class covariance matrix and Σ_w is the common within-class covariance matrix. Solution to this maximization is to take the first p eigenvectors of the matrix $\Sigma_w^{-1}\Sigma_b$ for a p dimensional projection. For more information about LDA, see [5].

Although the theory behind LDA is well established, there are several choices in how to use it for feature extraction in speech recognition. Being a supervised method, LDA needs class definitions for the estimation process. One choice is to use phonemes as classes [6], which is a popular choice also with MLPs (e.g. in [4]). With LDA, using hidden Markov model (HMM) states as classes have been shown to give improved recognition performance [3]. This is reasonable, as the HMM states can give more compact classes with less variation compared to phonemes, especially when using spectro-temporal input. However, with context-dependent models the number of classes may become rather large. Beulen *et al.* [7] showed that classes consistent with the HMM models are the best choice despite this large number of classes.

Another issue with LDA is the type of the spectral information used as the input for the method. The spectral information can be presented with logarithmic filter bank energies [6], some transformed features like PLP [8] or spectral energies augmented with time derivatives [3]. LDA is invariant to linear transformations, so there should be no need for spectral transformations. Beulen *et al.* [7] noted that with Gaussian mixture densities the time derivatives do not improve LDA over using only the spectral energies. Supported by these findings, we use in our experiments logarithmic mel-spaced filter bank energies as our short-time spectral features.

2.2. Pairwise linear discriminants (PLD)

The assumption about a common within-class covariance matrix in LDA results in unoptimal features if the condition is not met. This is the case with various speech classes which can have very different covariance structures. LDA has been extended to heteroscedastic linear discriminant analysis (HLDA) [9] which removes the equal covariance constraint. However, this requires numerical optimization, which may become restrictive if the dimension of the input vector increases too much. With spectro-temporal input this can occur if a wide input window is required.

Our new method is based on using a two-class LDA to find a one-dimensional projection for each pair of classes that maximize the Mahalanobis distance between those classes. The variance in the distance measure will be the average variance of the two classes along the projection, which results reasonable solutions even with unequal covariance matrices. Now instead of doing a single LDA with all the classes we use LDA to compute linear discriminants between the pairs of classes, and therefore avoid restricting the class covariances to be the same.

When computing pairwise linear discriminants we can not control the resulting number of feature dimensions directly. Furthermore, we want to have features with decorrelated outputs, whereas the projections to the pairwise discriminants may exhibit considerable correlation. To solve these problems we compute principal component analysis (PCA) of the training data projected to the pairwise linear discriminants and take as features the linear combinations of the discriminants as depicted by the p first eigenvectors. We call the resulting features pairwise linear discriminant (PLD) features.

Formally, let W be a $m \times n$ matrix containing the pairwise linear discriminants as its row vectors (n being equal to the dimension of the supervector $X(t)$). The projected covariance matrix C_{PLD} for PCA can then be obtained from the global covariance matrix C of supervectors $X(t)$ as

$$C_{PLD} = WCW^T. \quad (4)$$

Let D_p be a matrix containing the p first eigenvalues of C_{PLD} as its diagonal and let V_p contain the corresponding eigenvectors as its row vectors. The projection matrix A_{PLD} for p dimensional feature vectors is then obtained as

$$A_{PLD} = D_p^{-1/2} V_p W. \quad (5)$$

Due to this formulation, the resulting features are globally decorrelated and have a unit variance. Using PCA to obtain the features corresponds to maximizing the energy of the training data in the discriminative space spanned by the pairwise linear discriminants.

All the computations required for estimating the PLD feature filters are linear matrix operations, except for the computation of eigenvectors with PCA. The computations can therefore be carried out very efficiently. To further lower the computational load we reduced the number of pairwise linear discriminants in the first place by computing them only between the states of the same state position in three-state HMMs used in our acoustic models. Estimating PLD feature filters with context-dependent states as classes would be prohibitive due to the large number of states, so we used monophone HMM state classes instead. This, however, did not prevent the use of context-dependent acoustic models.

2.3. Variants of PLD

One problem with pairwise linear discriminants is the large number of pairs and the high dimensionality of the initial feature transformation. However, it also enables some manipulations which can be used to guide the PCA phase to find the best possible feature filters. One method which was found useful during preliminary testing was to remove those pairs that are easily discriminated. The discrimination effort was measured by the Mahalanobis distance between the classes. In a model used to run the experiments there were originally 693 pairs (corresponding to the state pairs of 22 most frequent Finnish phonemes, including silence), from which 200 pairs with the largest Mahalanobis distance were removed. After estimating the final feature filters it was verified that the pairwise distances were now better preserved in the remaining pairs, while the discrimination in the removed pairs remained good.

Kajarekar *et al.* [6] argued that it was more beneficial to run discriminant analysis in spectral and temporal domains separately than to use joint spectro-temporal LDA as presented in previous sections. Inspired by their report we tried reducing the modeling of correlation among spectral dimensions of different time instances. To achieve this, we first decorrelated the spectral dimension globally, and then when estimating the pairwise linear discriminants, masked the class covariance matrices to only include correlation in the temporal domain. The masking allowed non-zero elements in the covariance matrix only on positions corresponding to temporal correlation. The other positions, which hopefully had small correlations anyway, were replaced with zeros, so in effect the masking introduced smoothing to the covariance matrices.

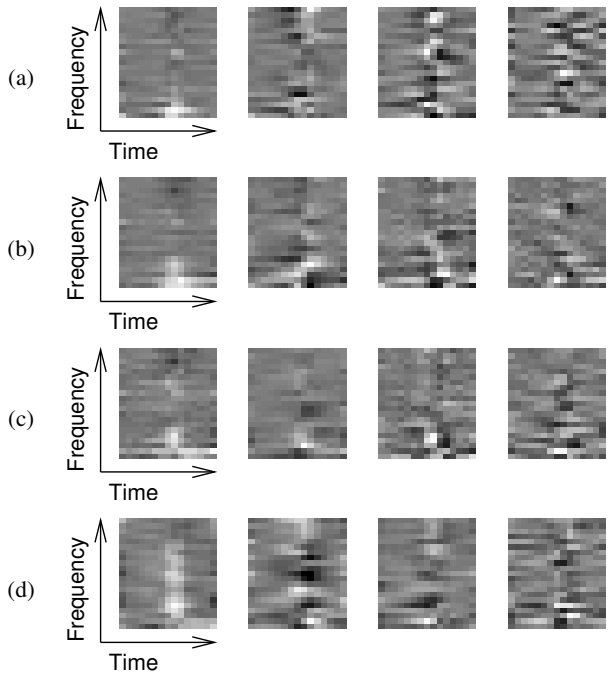

Figure 1: *Four feature filters visualized from different estimation methods. The methods are (a) LDA, (b) PLD, (c) PLD with reduced number of pairs and (d) PLD with masked covariance matrices.*

Figure 1 shows some of the feature filters for four different estimation methods: LDA, PLD, and the two PLD variants previously discussed. From each method four filters are presented: the 1st, 6th, 14th and 27th feature filter out of 39. The 1st feature filter corresponds to the one with the largest eigenvalue in PCA. The time span of the filters is 128 ms. Generally the basic PLD features and the PLD features with reduced number of pairs were rather similar. Features estimated with masked covariance PLD show smoother patterns than these two. LDA features are quite different from the others, and they also show less noise-like patterns than the basic PLD features.

3. Experimental results

3.1. Setup

We tested the presented feature estimation methods in a Finnish large vocabulary continuous speech recognition task. The training data contained both read and spontaneous sentences and word sequences from 207 speakers, with total of 21 hours of speech. The recognition task was the same as the speaker independent task in [10]. Also the speech recognition system was the same as in [10] but with two exceptions: the features had been changed and the n-gram language model had been updated to the one presented in [11]. The acoustic models in the system were based on decision tree tied triphone HMMs with Gaussian mixture densities. Before Gaussian computation the features were transformed with maximum likelihood linear transformation, which has been shown to be an important factor for linear feature extraction methods to work well [9].

The spectrum information used for the feature estimation methods was computed from 16 ms windows, with consecutive frames 8 ms apart. We used 21 logarithmic mel-spaced filter-bank energies as the base features, from which the MFCC and the LDA based features were computed. In preliminary testing the traditional LDA showed the best performance with a window width of 15 frames, resulting in 315 dimensional supervector for the input of transformations, and this setting was used throughout all the experiments. The optimal window width might be larger if more training data was available, now the same 21 hours of data was used for feature estimation as for model training.

The segmentation of the training data used to define the classes for LDA and PLD methods was done using Viterbi segmentation with monophone HMM models and MFCC features. The same segmentation was used for all the methods, except for the triphone state LDA, which used the triphone state segmentation of the LDA model (where classes had been monophone HMM states).

The scaling of language model probabilities with respect to acoustic probabilities was optimized for each method using a held-out development set. Both the development and the actual evaluation set contained only read sentences from disjoint sets of 20 and 31 speakers, respectively. The development set contained 1h and the evaluation set 1.5h of speech.

The speech recognition performance was measured with letter error rate (LER), as it better indicates the recognition performance of a highly inflectional language such as Finnish than the commonly used word error rate (WER). However, also WER is reported for completeness.

3.2. Results

The baseline model for the experiments used 12-dimensional MFCC and energy features augmented with first and second-order time derivatives, resulting in 39 dimensional feature vector. To keep the results of various methods comparable, the dimension of all other feature vectors was also 39.

Two different LDA models were experimented, the difference being only in the class definition. The "LDA monophone" model used monophone HMM states as class labels, whereas "LDA triphone" used triphone HMM states. The acoustic models in both of these and also in all other models experimented here were triphone HMMs, regardless of the LDA class definition. In addition to basic LDA models, a HLDA model with monophone HMM state classes was experimented, using a formulation given in [9].

The pairwise linear discriminant ("PLD") model used the same monophone HMM state class definition as the "LDA monophone" model. We also experimented two variants of PLD: "PLD reduced" had 200 pairs removed, and "PLD masked" was otherwise the same as "PLD reduced" except that the class covariances were masked to reduce the modeling of correlation among spectral dimensions of different time instances. Both of these variants were described in Section 2.3.

The results are summarized in Table 1. It can be seen that the best results, both for the development and the evaluation sets, were obtained using the "PLD reduced" model. The relative improvement compared to the baseline model in the evaluation set was about 18%.

Also the traditional LDA clearly improved the recognition results over the MFCC features in the evaluation set. The class definition (monophone or triphone HMM states) had only minor effect in the results. Resegmenting the triphone state classes with the "LDA triphone" model and retraining the model did not help either. It was a bit surprising that the HLDA model improved

Table 1: *Speech recognition results for different feature estimation methods.*

Model	LER (devel)	WER (devel)	LER (eval)	WER (eval)
MFCC+Δ+ΔΔ	4.80%	17.0%	5.22%	18.4%
LDA monophone	4.86%	17.1%	4.50%	17.2%
LDA triphone	4.94%	17.3%	4.56%	16.8%
HLDA	**4.53%**	16.4%	4.45%	17.0%
PLD	4.83%	16.8%	4.48%	16.7%
PLD reduced	**4.54%**	16.2%	**4.26%**	16.5%
PLD masked	4.74%	17.0%	4.55%	17.4%

the evaluation results compared to the LDA models only slightly. However, in the development set HLDA gave the best results along with the best PLD model, suggesting some variance in the results.

The basic PLD model performed similarly to the LDA models, but the "PLD reduced" achieved a further 5% improvement compared to the "LDA monophone" model. This improvement was statistically significant according to the Wilcoxon signed rank test.

Masking the class covariances degraded the performance from the "PLD reduced" model. This suggests that in an LVCSR task it is useful to analyze spectral and temporal patterns jointly, whereas in [6] a continuous digit recognition task was used to conclude that a separate analysis in spectral and temporal domains was more beneficial.

4. Conclusions

This paper investigated methods for estimating linear filters for extracting features from a window of short-time spectral features in speech recognition. These methods should in principle be able to take into account the spectro-temporal patterns in speech and therefore result in better features than e.g. the commonly used MFCC features with augmented time derivatives. The results in a large vocabulary continuous speech recognition task show that already the well-known linear discriminant analysis is able to estimate improved features compared to the MFCC.

A new method for estimating a discriminative linear subspace was also presented. The basic form of pairwise linear discriminant features showed similar performance as the LDA in the speech recognition experiments. A modified form of the PLD features (with easily discriminated pairs removed) resulted in a significant improvement in recognition accuracy compared to the LDA model. The improvements in the development set were quite different to the improvements observed in the evaluation set, which indicates sensitivity to the speech data. However, the best PLD model outperformed other methods consistently, only the HLDA model gave the same performance in the development set.

The good results of LDA and PLD features show that a modern speech recognition system can benefit from linear feature extraction methods. The PLD results also show that improvements over traditionally used LDA are possible with a better estimation method. A comparison to nonlinear feature extraction methods like MLPs should be carried out. It should be noted, however, that PLD features can be estimated very efficiently without computationally intensive optimization.

The full potential of PLD features was not necessarily seen with the variants experimented in this paper. The role of class definition with PLD features should be investigated more carefully. As removing some of the pairwise linear discriminants before PCA turned out to be useful, also other ideas like pair weighting or more clever pair reduction methods could be considered in the future.

5. Acknowledgments

This work was supported by the Academy of Finland in the project "New adaptive and learning methods in speech recognition" and the Graduate School of Language Technology in Finland.

6. References

[1] Howard Hua Yang, Sarel Van Vuuren, Sangita Sharma, and Hynek Hermansky, "Relevance of time–frequency features for phonetic and speaker-channel classification," *Speech Communication*, vol. 31, no. 1, pp. 35–50, 2000.

[2] Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer, "Speech recognition with continuous-parameter hidden Markov models," in *Proceedings of ICASSP*, 1988, pp. 40 – 43.

[3] Reinhold Haeb-Umbach and Hermann Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proceedings of ICASSP*, 1992, pp. 13–16.

[4] Qifeng Zhu, Barry Chen, Frantisek Grezl, and Nelson Morgan, "Improved MLP structures for data-driven feature extraction for ASR," in *Proceedings of Interspeech*, 2005, pp. 2129–2132.

[5] Keinosuke Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.

[6] Sachin S. Kajarekar, B. Yegnanarayana, and Hynek Hermansky, "A study of two dimensional linear discriminants for ASR," in *Proceedings of ICASSP*, 2001, pp. 137–140.

[7] Klaus Beulen, Lutz Welling, and Hermann Ney, "Experiments with linear feature extraction in speech recognition," in *Proceedings of Eurospeech*, 1995, pp. 1415–1418.

[8] Panu Somervuo, Barry Chen, and Qifeng Zhu, "Feature transformations and combinations for improving ASR performance," in *Proceedings of Eurospeech*, 2003, pp. 477–480.

[9] George Saon, Mukund Padmanabhan, Ramesh Gopinath, and Scrott Chen, "Maximum likelihood discriminant feature spaces," in *Proceedings of ICASSP*, 2000, pp. 1129–1132.

[10] Janne Pylkkönen, "New pruning criteria for efficient decoding," in *Proceedings of Interspeech*, 2005, pp. 581–584.

[11] Vesa Siivola and Bryan L. Pellom, "Growing an n-gram language model," in *Proceedings of Interspeech*, 2005, pp. 1309–1312.