Feature Transformations and Combinations for Improving ASR Performance

Panu Somervuo[1,2], *Barry Chen*[1], *Qifeng Zhu*[1]

1) International Computer Science Institute
1947 Center Street
Berkeley, CA 94704
USA
byc,qifeng@icsi.berkeley.edu

2) Neural Networks Research Centre
Helsinki University of Technology
P.O. Box 5400, FIN-02015 HUT
Finland
panu.somervuo@hut.fi

Abstract

In this work, linear and nonlinear feature transformations have been experimented in ASR front end. Unsupervised transformations were based on principal component analysis and independent component analysis. Discriminative transformations were based on linear discriminant analysis and multilayer perceptron networks. The acoustic models were trained using a subset of HUB5 training data and they were tested using OGI Numbers corpus. Baseline feature vector consisted of PLP cepstrum and energy with first and second order deltas. None of the feature transformations could outperform the baseline when used alone, but improvement in the word error rate was gained when the baseline feature was combined with the feature transformation stream. Two combination methods were experimented: feature vector concatenation and n-best list combination using ROVER. Best results were obtained using the combination of the baseline PLP cepstrum and the feature transform based on multilayer perceptron network. The word error rate in the number recognition task was reduced from 4.1 to 3.1.

1. Introduction

Feature representation is an important part of any pattern recognition system, automatic speech recognition (ASR) being no exception. It is difficult to develop any theoretically optimal feature extraction methods which would minimize the recognition error. In practice, several methods have been experimented and during the long history of ASR, some feature representations have been experimentally proved to be more beneficial than others.

In most systems the speech signal is chunked into overlapping 20-30ms time windows at every 10 ms and the spectral representation is computed from each frame. A common feature vector consists of mel-frequency cepstral coefficients (MFCC). Temporal dynamics is represented by concatenating the first and second order deltas to this feature.

In this work, logarithmic mel-spectra and PLP cepstra have been used as original feature vectors. Linear and nonlinear transformations have then been applied to these features. The resulted features have been used for training mixture-of-Gaussians based hidden Markov models (HMMs).

Besides comparing different feature transformations, the interest was also to combine different feature transformation streams. Two combination methods were experimented. In the first method the feature vectors of two streams were concatenated and the new recognition system was trained by using the new feature vector. In the second method recognizers were trained for each feature stream separately and ROVER was used for combining the outputs of the recognizers.

2. Feature transforms

Logarithmic mel-spectra and PLP cepstra were used as original feature vectors. Feature transformations were computed from single frames and multi-frame windows, see Fig. 1. The basic ideas behind the experimented feature transformations are described in this section, for text book references, see e.g. [1] and [2]. Principal component analysis and independent component analysis are linear, unsupervised methods, whereas linear discriminant analysis and its nonlinear extension, nonlinear discriminant analysis, utilize the class information of the original feature vectors. Linear feature transforms can be implemented by matrix multiplications and nonlinear feature transforms can be implemented using nonlinear multilayer perceptron networks.

2.1. Principal component analysis

Principal component analysis (PCA) is a method to represent the data in the low-dimensional subspace of the data space. The projection into the principal components is called Karhunen-Loeve transform (KLT). Principal components of the data set can be found by computing the covariance matrix of the data set and then finding the eigenvectors corresponding to the largest eigenvalues. KLT decorrelates the feature vectors which enables the modeling of data using Gaussians with diagonal covariance

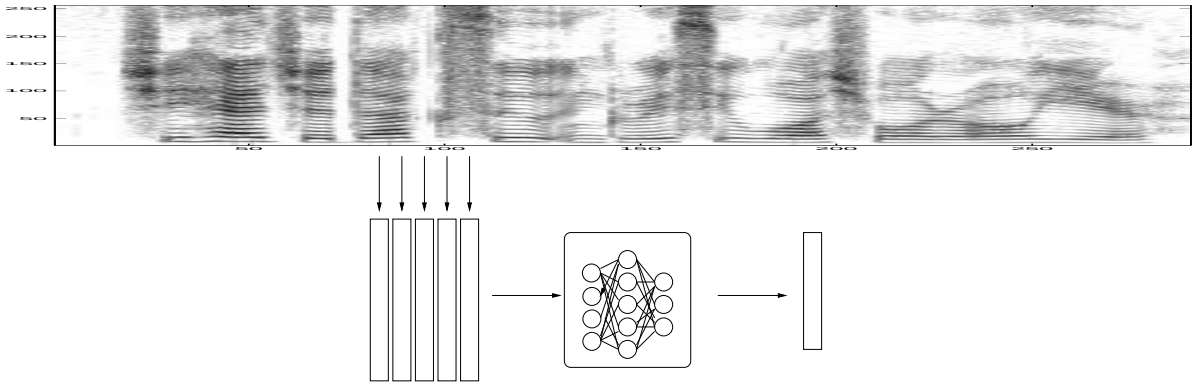

Figure 1: *Feature transform. One or more frames (five in this figure) original feature vectors, e.g. logarithmic mel-spectra are fed to the linear (matrix) or nonlinear (MLP network) feature transform which performs the projection of the original feature vector (or concatenation of them) to the new feature space. The output is used as a feature vector in the mixture-of-Gaussians based HMM system.*

matrices.

2.2. Independent component analysis

The idea behind the use of Independent component analysis (ICA) in the feature extraction is to reduce the redundancy of the original feature vector components. While PCA decorrelates the data, i.e. removes the second-order dependencies of the feature vector components, ICA removes also higher order dependencies. The objective of ICA is to minimize the mutual information between the feature vector components.

ICA tries to find basis vectors onto which the projections of the data are non-gaussian (the sum of two independent components is more gaussian than either of the components alone). Typically in the ASR systems, the features are modeled by mixtures of Gaussians. Because of the nature of the non-gaussianity, this may not be the best choice for ICA based features.

2.3. Linear discriminant analysis

Linear discriminant analysis (LDA) attempts to find such basis vectors that the linear class separability is maximized. Two matrices are computed, the within-class scatter matrix (covariance matrix) S_w and between-class scatter matrix S_b. S_w is a weighted linear sum of classwise covariance matrices and S_b can be defined as $1/n \sum_i n_i(m_i - m)(m_i - m)^T$, where m_i is the mean of the ith class, n_i the sample count, m the global mean, and T denotes the transpose. LDA basis vectors are now the eigenvectors of the matrix $S_w^{-1} S_b$. For c classes, there are at most $c - 1$ linearly independent eigenvectors. Not all of them need be used, but the selection can be based on the eigenvalues, as in PCA. Like PCA, LDA decorrelates the feature components.

The assumption behind the basic LDA is that each class is modeled by a single Gaussian and the covariance matrices of all classes are equal. Depending on the classes and original features, this can be quite far from the true distributions. There are some modifications to the basic LDA which loose these restrictions. E.g. in heteroscedastic LDA, the class covariances are not considered being equal. But the drawback is that the optimization cannot then be done in closed form, and iterative schemes must be used [3].

2.4. Nonlinear discriminant analysis

Modeling limitations of the linear discriminants can be ignored by using nonlinear discriminant functions. Multilayer perceptron (MLP) networks can be used for this purpose. The number of the units in the hidden layer and the nonlinearity of the activation functions determines the complexity and the modeling capacity of the network.

MLP nets are trained for separating target classes, e.g. phonemes. With suitable activation functions in the output layer, the MLP net gives the posterior probabilities of the classes for the given input feature vector. The dimension of the new feature vector can be reduced by computing PCA and taking the projections to the eigenvectors corresponding to the largest eigenvalues.

If soft-max activation function is used for output layer, the outputs tend to be very spiky. If the transformed features are modeled by mixtures of Gaussians, it is beneficial to "gaussianize" them e.g. by taking the logarithm of the output activation vector or simply removing the final nonlinearity [4].

3. Recognition task

For our experimental setup, we have adopted a slightly unconventional training and testing regimen to test the generalizability of our feature transformations. Our training consists of a mixture of conversational telephone speech and read speech from the Macrophone Corpus.

This is a subset of the HUB5 training data consisting of over 60 hours of speech from 6273 speakers. For testing, we use the Numbers corpus collected at OGI which consists of strings of continuous numbers (32 words total) like "five hundred fifty eight" or "six oh four". This corpus was collected from different speakers over the telephone. We report word error rates on the Numbers test set consisting of 1227 utterances (0.6 hours and 4670 words).

We used SRI's recognizer [5] with bigram language model. Acoustic models were gender-independent mixture-of-Gaussians based HMMs. The means and variances of the original feature vectors were normalized speaker-wise before computing the transformations.

3.1. Single-frame transforms

The baseline feature consisted of 12-dimensional PLP cepstrum with energy. Together with the first and second order deltas, the feature vector had 39 components.

Linear feature transformations were first applied to single-frame feature vectors. The original data were 15-dimensional spectral representations (logarithmic critical band energies). PCA, ICA, and LDA was applied and the final feature dimension was reduced to 13. LDA target classes were 48 phones. Together with the deltas, the feature vector contained 39 components like the baseline feature. The results are in Table 1. None of the experimented feature transforms was able to outperform the baseline.

Table 1: *Recognition results for single-frame transforms. Each feature vector contains 39 components including first and second order deltas. Original features for PCA, ICA, and LDA were 15-dimensional logarithmic critical band energies.*

feature	WER
PLPcep	4.1
PCA	5.1
ICA	6.7
LDA	4.9

3.2. Multi-frame transforms

More contextual information can be gained if the feature transform is computed from the multi-frame window. Now the original single-frame feature was 12-dimensional PLP cepstra with energy. Nine consecutive frames was used for computing the transform. The final feature dimension was then reduced from 117 to 48. Results are in Table 2. Here the nonlinear discriminant (MLP) based feature performed best. Comparing PCA, LDA, and ICA results between Tables 1 and 2, it can be seen that it is more beneficial to compute the transform from the multi-frame input.

Table 2: *Recognition results for multi-frame transforms. Each feature vector contains 48 components (47 for LDA feature).*

feature	WER
PCA	4.5
ICA	5.0
LDA	4.5
MLP	4.1

4. Feature combinations

Feature stream combinations were experimented in two levels, first, concatenating the feature vectors from two feature streams, and then, combining the outputs of two feature-specific recognizers with ROVER.

4.1. Feature vector concatenation

The baseline feature vector was appended by an additional feature stream and the resulting feature was decorrelated by KLT (whitening matrix was computed from the training data). Since ICA did not seem to outperform PCA when using mixtures of Gaussians in HMMs, it was not used in this experiment. First, PCA and LDA features were computed from single frames, and then deltas were appended (same features as in Table 1). It was also experimented to concatenate multi-frame PCA to the baseline feature and not use deltas. This gave better performance (denoted as PCA2 in Table 3). MLP features were computed from nine-frame context windows. The differences in the results between linear and MLP features are quite striking. The baseline feature was PLP cepstrum and the linear transforms were computed from 15-dimensional logarithmic critical band energies. PCA and LDA features were thus not just linear transforms of the baseline feature. Nevertheless, for some reason the concatenated linear transform features did not perform well. MLP feature appended to the baseline feature was able to reduce the baseline word error rate.

Table 3: *Recognition results for concatenated feature vectors. PCA and LDA were computed from single-frame inputs and they were used with deltas. PCA2 was computed from multi-frame input and no deltas were used.*

feature	WER
PLPcep + PCA	6.6
PLPcep + LDA	6.4
PLPcep + PCA2	4.7
PLPcep + MLP	3.3

4.2. ROVER

Another feature stream combination was based on ROVER (Recognition Output Voting Error Reduction) [6]. Here the weighted ROVER was used in which the recognition hypotheses (n-best lists) are weighted by their posterior probabilities. Different streams have also a priori weights. Here 0.66 was used for baseline feature and 0.33 for an additional stream. These values were not particularly tuned, but these values gave better results compared to equal 0.5 weights. The results are in Table 4.

These results show improvement over the baseline system. The order of the performance in the results reflect the performance of the additional features when used alone. MLP feature is again the best. But in contrast to the results in Table 3, now all error rates are between 3.0 and 4.0 per cent.

It was also experimented to feed all five feature streams to ROVER, but this did not give further improvement over the system using only two recognizers with PLP cepstrum and MLP feature.

Table 4: *Recognition results when combining the n-best lists of two recognizers with ROVER.*

feature	WER
PLPcep + PCA	3.4
PLPcep + ICA	3.6
PLPcep + LDA	3.2
PLPcep + MLP	3.1

5. Discussion

In earlier work [7] PCA, ICA, LDA, and MLP features were experimented in phone recognition task. Feature concatenation was used as a combination method and then the linear transformations did bring additional information to the combination of the baseline feature and the MLP feature improving the recognition. Also in that work all feature transformations outperformed the baseline recognition system. It is interesting that in the present work none of the feature transformations could outperform the baseline feature when used alone. But there are some differences in the recognition systems used in the present work and in the earlier work. Also the recognition task in the current work was number recognition instead of phone recognition. In [7] the acoustic models were simple monophone HMMs and no speaker normalizations were applied to the original feature vectors. In the present work state-of-the-art recognizer was used (SRI's recognition system), speaker-wise mean and variance normalizations were applied to the original feature vectors and the acoustic models were now triphones instead of monophones.

6. Conclusions

In this work feature transformations were experimented in ASR front end. Unsupervised transformations were based on principal component analysis and independent component analysis. Discriminative transformations were based on linear discriminant analysis and MLP networks. None of the feature transformations could outperform the baseline system when used alone, but improvement in the word error rate was gained when the baseline feature was combined with the feature transformation stream.

When MLP features were combined with the baseline feature, improvement was gained both when the two features were concatenated in order to form a new feature vector, and when two separate recognizers were run separately for each feature stream and the combination was based on ROVER. For linear feature transformations, improvement was gained only when using ROVER.

Based on the results in this work, MLP feature seems to be the best choice for using as an additional feature for the baseline PLP cepstrum. Experimented linear feature transformations did not bring any further improvement to the recognizer using these two streams.

7. Acknowledgements

All authors would like to thank people at ICSI for creating a friendly working atmosphere.

Panu Somervuo would like to thank Academy of Finland for financial support, project no. 44886 "New information processing principles" (Finnish Centre of Excellence Programme 2000-2005).

8. References

[1] Bishop, C., Neural Networks for Pattern Recognition, Oxford, 1995.

[2] Hyvärinen, A., Karhunen, J, and Oja, E. Independent Component Analysis, John Wiley & Sons, 2001.

[3] Saon, G., Padmandabhan, M., Gopinath, R., and Chen, S. "Maximum likelihood discriminant feature spaces", ICASSP 2000, vol. 2, pp. 1129-1132.

[4] Hermansky, H., Ellis, D., and Sharma, S. "Tandem connectionist feature stream extraction for conventional HMM systems", ICASSP 2000, vol. 3, pp. 1635-1638.

[5] Digalakis, V., Monaco, P., and Murveit, H. "Genones, generalized mixture tying in continuous hidden Markov model-based speech recognizers", IEEE Tr SAP, vol 4, no 4, pp. 281-289, 1996.

[6] Fiscus, J. "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)", IEEE ASRU Workshop 1997, pp. 347-352.

[7] Somervuo, P. "Experiments with linear and nonlinear feature transformations in HMM based phone recognition", ICASSP 2003.