MULTIPLE TIME RESOLUTIONS FOR DERIVATIVES OF MEL-FREQUENCY CEPSTRAL COEFFICIENTS

Georg Stemmer, Christian Hacker, Elmar Nöth, Heinrich Niemann

Universität Erlangen–Nürnberg
Lehrstuhl für Mustererkennung (Informatik 5)
Martensstr. 3, 91058 Erlangen, Germany
stemmer@informatik.uni-erlangen.de

ABSTRACT

Most speech recognition systems are based on mel-frequency cepstral coefficients and their first- and second-order derivatives. The derivatives are normally approximated by fitting a linear regression line to a fixed-length segment of consecutive frames. The time resolution and smoothness of the estimated derivative depends on the length of the segment. We present an approach to improve the representation of speech dynamics, which is based on the combination of multiple time resolutions. The resulting feature vector is transformed to reduce its dimension and the correlation between the features. Another possibility, which has also been evaluated, is to use probabilistic PCA (PPCA) for the output distributions of the HMMs. Different configurations of multiple time resolutions are evaluated as well. When compared to the baseline system a significant reduction of the word error rate can been achieved.

1. INTRODUCTION

1.1. Motivation

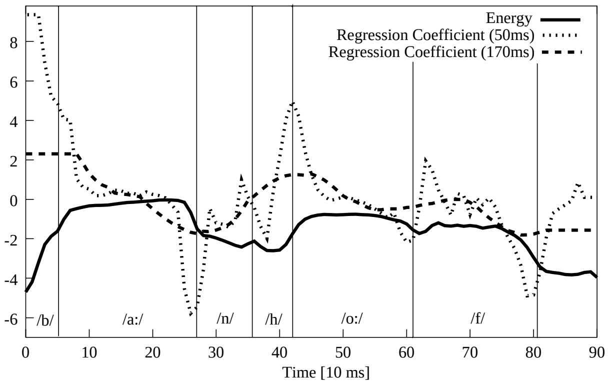

Fig. 1. Energy contour and two different approximations of its first-order time derivative in the German word *Bahnhof*.

Mel-frequency cepstral coefficients (MFCCs) are the standard features for most speech recognition systems. As MFCCs are based on static short-time spectral representations of the speech signal, they are usually extended by dynamic features. Frequently, first- and second-order time derivatives of the static features [1] are employed. The derivative of a time-sampled sequence of the static features can be approximated robustly by the coefficients of that linear regression line which fits best to a segment of several consecutive frames. A simpler approach would be taking the difference of two adjacent frames, a more complex one is the use of a higher-order polynomial instead of the linear regression line [2]. For the computation, all methods employ a time segment of fixed length typically 5, 7 or 9 frames.

Fig. 1 gives an example of the progression of the energy contour in the German word *Bahnhof* (station). The dotted curves depict two approximations of the first-order derivative of the energy contour, the approximations differ in the length of the time segment. It can be seen that both curves are highly correlated. For a larger size of the time segment the curve is smoothed and many details, for instance the fall of the energy contour at the end of the /o:/, are lost. However, the smoother estimation of the derivative eliminates irrelevant variations in energy as within /n/ and /h/. Three relative maxima correspond to the vowels and to the fricative. Both curves provide the acoustic models with different aspects of the speech signal and we can not tell in advance which one should be preferred.

1.2. Approach

Instead of deciding for a certain fixed time segment, the experiments described in this paper attempt to combine the advantages of different resolutions in the dynamic features. For each time frame multiple regression lines corresponding to different sizes of the time segment are computed. Together with the static features all regression coefficients lead to a high-dimensional feature vector. As the components of the feature vector are highly correlated, the obvious thing to do is to apply a transformation for a reduction of the dimension and of the correlation. Another possibility is not to transform the feature vector, but to use an appropriate output distribution for the acoustic models, which is able to cope with the special quality of the feature vector. Section 4 describes an investigation of both approaches.

1.3. Related Work

The idea of the integration of several different features which have been computed on the base of different timescales or time resolutions into one feature vector relates our approach to other works known from the literature. For instance, H. Hermansky and S. Sharma describe an approach to increase the robustness of a recognizer by the incorporation of long-term information [3]. K. Weber [4] describes an approach to achieve a higher robustness against additive noise by combining static features, which have been computed on a much longer timespan (up to two seconds) than the normal MFCCs in a feature vector.

2. SHORT DESCRIPTION OF THE SYSTEM

The system which has been used for the experiments is a speaker independent continuous speech recognizer. It is based on semi-continuous HMMs, the output densities of the HMMs are full-covariance Gaussians. Please refer to [5] for a detailed description of the speech recognizer. Every 10 ms 12 static MFCCs are computed.

Dynamic adaptive cepstral subtraction (DACS) and a temporal filter are applied to the coefficients. The dynamic features of the baseline system are 12 first-order derivatives of the static MFCCs. They are estimated by means of a linear regression covering a time segment of 9 consecutive frames. The experiments will show (Tab. 2) that 9 frames is not optimal for our test set. This may be due to the fact that the parameters of the feature extraction algorithm were optimized on read speech [6].

If the baseline system is only trained on the training data set described in the next section and no other data is used for training or initialization of the acoustic models, it achieves a word error rate of 31.1% on the test data.

3. DATA

Acoustic models are trained on a part of the EVAR data set. It consists of 7438 utterances, which have been recorded by phone with our conversational train timetable information system. A detailed description of this system can be found in [7]. Nearly all utterances are in German language. The total amount of data is ca. 8 hours. 5440 utterances have randomly been selected for training and validation, the rest of 1998 utterances is available for testing.

4. MULTIPLE TIME RESOLUTIONS

4.1. PCA of the Combined Feature Vector

The most natural approach to the integration of multiple time resolutions in the dynamic features is to build one high-dimensional feature vector. As there are 12 static MFCCs, each possible resolution of the regression line results in 12 dynamic features. It is obvious that for a high number of different resolutions the dimension of the feature vector must be reduced with a suitable transformation. We decided to apply principal component analysis (PCA) [8], a popular technique which may be used to compute a low-dimensional principal subspace of a data set. A projection of the feature vectors onto the subspace minimizes the squared reconstruction error and the correlation of the components of the resulting low-dimensional feature vectors is reduced.

We have to decide, whether to put the static MFCCs and the dynamic features together before or after the dimensionality reduction. In our approach the PCA transform is applied only to the dynamic part of the feature vector and the static MFCCs are left unchanged, i.e. correlations between the static and the dynamic part of the feature vector are ignored. The separation has the advantage of protecting the MFCCs from being eliminated from the feature vector by the PCA. The static features themselves are already decorrelated because of the discrete cosine transform (DCT), which is part of the computation of the MFCCs. For comparison purposes we will also give recognition rates when static and dynamic features are transformed jointly with PCA.

The eigenvectors which are needed for the PCA may be derived either from the covariance matrix or from the correlation matrix of the data. For a PCA of MFCCs and their derivatives, the use of the correlation matrix yielded better results on our data. Consequently, the results which are described in the next section have been computed with a correlation-based PCA transformation. Please note that the correlation matrix is equivalent to the covariance matrix, when the standard deviation of the features is one and their mean is zero. Thus, a correlation-based PCA transformation is equivalent to the combination of a re-scaling of the data and a covariance-based PCA.

4.2. Probabilistic PCA

Instead of reducing the dimension of the feature vector, it is also possible to choose a special output distribution for the acoustic models which is able to cope with high-dimensional feature vectors and highly-correlated features. Probabilistic PCA (PPCA) [9] defines a special Gaussian probability density function with a constrained covariance matrix $\mathbf{\Sigma}$ of the form $\mathbf{\Sigma} = \sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^\top$. $\mathbf{I}$ stands for the identity matrix. The columns of the weighting matrix $\mathbf{W}$ span the (low-dimensional) principal subspace of the data. σ^2 represents the noise variance, which is proportional to the variance that is not represented by the principal subspace. The PPCA density is able to model even high dimensional feature vectors in an efficient way. If the output distribution of an HMM is a mixture of PPCA densities, we can expect that the structure of the feature space can be represented in a more flexible way than by combining a single PCA with a mixture of Gaussian densities.

5. EXPERIMENTAL RESULTS

5.1. PCA of the Combined Feature Vector

dim. before	no. of adjacent frames						WER
PCA trafo.	3	5	7	9	11	13	[%]
-				X			31.1
24	X	X					26.8
36	X	X	X				26.2
48	X	X	X	X			26.6
60	X	X	X	X	X		26.7
72	X	X	X	X	X	X	26.9

Table 1. Word error rates (WER) for the PCA transformation of the dynamic features. For each feature vector configuration, the included resolutions are marked by an 'X'. Each resolution adds 12 features to the vector. PCA reduces the dimension to 12. The first row of the table corresponds to the baseline configuration.

The word error rate on the test data for a large number of different combinations of time resolutions has been evaluated and is given in Tab. 1. In order to be able to compare the results with the baseline system the feature vector which is modeled by the HMMs is always 24-dimensional. It consists of 12 static MFCCs, which are the same as in the baseline system. The remaining 12 coefficients are generated by a PCA transformation from a higher dimensional feature vector which contains multiple approximations of the first-order derivative.

The lowest word error rate (WER) of 26.2% has been achieved for the combination of three different approximations of the first-order derivative: three, five and seven adjacent frames are used for the computation of the regression line. It can be seen from Tab. 1 that the improvement relative to the baseline system becomes smaller when many different regression coefficients are combined in a very high dimensional feature vector. This effect may be caused by the PCA, which seems to select the important features not robustly enough for very high dimensional input.

When the PCA transformation is applied to the combination of static and dynamic features in contrast to the dynamic features only, the WER is much higher. For the optimal configuration of three, five and seven adjacent frames, which is mentioned above, the WER is 27.2%.

dim. before	no. of adjacent frames				WER
PCA trafo.	3	5	7	9	[%]
12	X				27.9
12		X			27.5
12			X		28.6
12				X	30.8

Table 2. Word error rates (WER) for the PCA transformation of the dynamic features with single resolutions.

One may argue that the improvement of the new feature vector relative to the baseline system is not caused by the combination of multiple time resolutions, but would be just a consequence from a suboptimal configuration of the baseline features for our test data. The improvement may also be caused by the PCA transformation of the dynamic features. Therefore, we tried to find the best single time resolution for the approximation of the first-order derivative on our test data set. The dynamic features are decorrelated with a PCA transformation, but are not reduced in dimension. As can be seen from Table 2, the optimal single time resolution for the test set is five consecutive frames. The WER for this feature vector is worse than for all experiments with multiple time resolutions. From a comparison of Tab. 2 and Tab. 1 some insight into the multi-resolution approach can be gained. In Tab. 2 five consecutive frames are shown to be the best single time resolution. Adding two additional resolutions of three and seven frames to the feature vector reduces the word error rate by ca. 5% relative. This improvement is still significant at a 95% level. Therefore, the multi-resolution approach has two advances: firstly, it is more likely that the optimal time resolution for the specific test data is contained in the feature vector. Secondly, even the optimal time resolution can be improved by additional smoother and more detailed regression coefficients.

5.2. Probabilistic PCA

For the second approach the output distributions of the HMMs are replaced by PPCA densities. We expect, that the mixture of PPCA densities is able to cope with a high correlation of the input features, so the PPCA densities are used to directly model the 48-dimensional feature vector, which has been shown to be the optimal configuration of the multiple time resolution approach in the previous experiments (36 dynamic features plus 12 static MFCCs). The principal subspace of the PPCA densities is set to a dimension of 24. We improved our results by the application of a prenormalization on the data. Unfortunately, in the standard definition of PPCA densities it is not possible to separate the static and dynamic features in advance of the projection in the principal subspace, as we did successfully in the previous experiments. As a consequence, only a WER of 26.8% is achieved for the optimal configuration of three, five and seven adjacent frames. This is slightly better than a single PCA transformation applied to the full 48-dimensional feature vector (27.2% WER), but it is still worse than if the PCA is only applied to the dynamic part of the feature vector (26.2%). Additional experiments have to be made at this point.

6. CONCLUSION AND OUTLOOK

First- and second-order derivatives of MFCCs are used in many speech recognition systems for the representation of speech dynamics. The derivatives are normally approximated by a linear regression line, which is computed on a fixed segment of consecutive frames. For different sizes of the segment, different aspects of the progression of the cepstral coefficients in time can be accentuated. We presented an approach to combine several different resolutions of the estimated derivative in the feature vector. Since the resulting feature vector is high-dimensional and its components are highly correlated, it is transformed with PCA. Different configurations have been evaluated. When compared to the baseline system, a reduction from 31.1% to 26.2% WER was achieved.

An important bottleneck, which prevented the PPCA densities from outperforming the PCA transformation, is the need for developing a special structure of the covariance matrix, which enables the PPCA density to keep the static and the dynamic features separated. Another transformation which can even be optimized simultaneously with the acoustic models is linear discriminant analysis (LDA) [10, 11]. Further experiments should investigate the application of LDA for the multiple-resolution approach. In addition to this, experiments concerning the application to second-order derivatives are needed as well as evaluations on different speech databases.

7. REFERENCES

[1] S. Furui, “Speaker independent isolated word recognition using dynamic features of speech spectrum,” in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1986, pp. 52–59.

[2] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[3] H. Hermansky and S. Sharma, “Traps - classifiers of temporal patterns,” in *Proc. Int. Conf. on Spoken Language Processing*, Sydney, Australia, 1998.

[4] K. Weber, “Multiple timescale feature combination towards robust speech recognition,” in *KONVENS 2000 / Sprachkommunikation*, Illmenau, Germany, 2000.

[5] F. Gallwitz, *Integrated Stochastic Models for Spontaneous Speech Recognition*, Ph.D. thesis, University of Erlangen-Nuremberg, Erlangen, Germany, to appear.

[6] S. Rieck, *Parametrisierung und Klassifikation gesprochener Sprache (in German)*, Ph.D. thesis, University of Erlangen-Nuremberg, Erlangen, Germany, 1994.

[7] F. Gallwitz, M. Aretoulaki, M. Boros, J. Haas, S. Harbeck, R. Huber, H. Niemann, and E. Nöth, “The Erlangen Spoken Dialogue System EVAR: A State–of–the–Art Information Retrieval System,” in *Proceedings of 1998 International Symposium on Spoken Dialogue (ISSD 98)*, Sydney, Australia, 1998, pp. 19–26.

[8] I. T. Jolliffe, *Principal Component Analysis*, Series in Statistics. Springer-Verlag, 1986.

[9] M. E. Tipping and C. M. Bishop, “Mixtures of probabilistic principal component analyzers,” *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[10] R. Haeb-Umbach and H. Ney, “Linear discriminant analysis for improved large vocabulary speech recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, 1992, vol. 1, pp. 13–16.

[11] D. C. Bateman, D. K. Bye, and M. J. Hunt, “Spectral contrast normalization and other techniques for speech recognition in noise,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, 1992, vol. 1, pp. 241–244.