IMPROVEMENTS IN CONNECTED DIGIT RECOGNTION USING LINEAR DISCRIMINANT ANALYSIS AND MIXTURE DENSITIES

R. Haeb–Umbach, D. Geller, H. Ney

Philips GmbH Research Laboratories Aachen, P.O. Box 1980, 5100 Aachen, Germany
haeb@pfa.philips.de

ABSTRACT

Four methods were applied to reduce the error rate of a continuous-density hidden Markov model based speech recognizer on the TI/NIST Connected-Digits Recognition Task.
1. *Energy thresholding* sets a lower limit on the energy in each frequency channel to suppress spurious distortion accumulation caused by random noise. This led to an improvement in error rate by 15%.
2. *Spectrum normalization* was used to compensate for across-speaker variations resulting in an additional improvement by 20%.
3. The *acoustic resolution* was increased up to 32 component densities per mixture. Each doubling of the number of component densities yielded a reduction in error rate by roughly 20%.
4. *Linear discriminant analysis* was used for improved feature selection. A single class-independent transformation matrix was applied to a large input vector consisting of several adjacent frames resulting in an improvement by 20% for high acoustic resolution.
The final string error rate was 0.84%.

1. INTRODUCTION

The key to achieving high recognition accuracy is detailed acoustic modelling. For speaker-independent recognition the use of gender-dependent models, the inclusion of first- and second-order spectral and energy features into the acoustic feature vector, and a large number of elementary distributions contributing to an emission probability of a hidden Markov model state have been shown to significantly reduce the error rate, for both large and small vocabulary systems. On the TI/NIST Connected-Digits Recognition Task ("TI-Digits") [10], which serves as a kind of standard benchmark for small vocabulary speaker-independent recognition systems, the error rate has been reduced by more than a factor of 5 since the first published results, mainly due to the above means.

With our speech recognizer which employs whole-word hidden Markov models with continuous emission probabilities, the degree of acoustic resolution is easily controlled by the number of Laplacian component densities per mixture [12]. The cost for high recognition accuracy is, however, increased computational complexity and memory consumption. In this paper we present two methods which improved performance on the TI-Digits with only little increase in complexity. First, a lower limit is set on the energy in each frequency channel, in both training and recognition. Then low energy portions which are dominated by random noise do not contribute to the distortion accumulation for the log-likelihood computation. This method is essentially equivalent to Klatt's thresholding method [9] applied however to high-quality data rather than noisy data. For a properly chosen threshold the string error rate could be reduced by 15%.

Second, we introduced a long-term spectrum normalization step to account for across-speaker variations. Low-frequency changes of the log-spectral intensities are suppressed since they are likely to be caused by speaker and acoustic channel variations rather than by the information-bearing speech signal itself [11]. This technique has previously been used successfully to compensate for acoustic channel variations [13]. On the TI-Digits this measure reduced the string error rate by 20%.

Systems achieving very high recognition accuracy often include some discriminative training aspects: Cardin et al. use hidden Markov models trained with maximum mutual information estimation [2], Gauvain and Lee incorporate corrective training [6], and Doddington employs state-specific transformation matrices based on linear discriminant analysis (LDA) [4]. Our approach is different since we use a single class-independent LDA transformation matrix which proves to be very effective, in particular in the case of high acoustic resolution. On the TI-Digits task the error rate was reduced by one fifth to 0.84% string error rate.

0-7803-0946-4/93 $3.00

2. BASELINE RECOGNITION SYSTEM

We carried out our experiments on the TI/NIST Connected-Digits Recognition Task. After filtering and sampling the speech signal at 8 kHz and subsequent preemphasis a 256-point FFT is computed every 12 ms from a Hamming-windowed 32 ms portion of the speech signal. 15 power-spectral density coefficients are determined from a convolution of the power density spectrum with triangular kernels with a nonlinear frequency spacing. The logarithm is computed and the average of the 15 coefficients is subtracted from each channel and included as an additional component in a resulting 16-component feature vector $y(n)$. This feature vector is optionally augmented with slope and curvature information according to the following scheme:

$$x(n) := \begin{bmatrix} y(n) \\ y'(n) \\ y''(n) \end{bmatrix} = \begin{bmatrix} y(n) \\ y(n) - y(n-2) \\ y(n+2) - 2y(n) + y(n-2) \end{bmatrix} \quad (1)$$

Slope and curvature are computed from each static feature resulting in a 32-component (only slope) or 48-component (slope and curvature) feature vector.

The connected-word recognition algorithm is based on whole-word hidden Markov models, the emission probabilities of which are modeled by continuous Laplacian densities with a single 'standard deviation' vector pooled over all states [12]. The transition probabilities $a(s \mid s\prime)$ from state $s\prime$ to state s are not trained but instead are given fixed a-priori values that are nonzero only for loop, skip and forward transitions. For the 11-word vocabulary of the TI-Digits we use 23 models, i.e a separate model for each word and gender, and a single background noise model.

We employ the Viterbi approximation in both training and recognition, i.e. the probability of a word is replaced by the probability of its most likely state sequence. Details of the acoustic modeling can be found in [1].

3. THRESHOLDING

The key idea of thresholding or noise masking as proposed by Klatt [9] is to choose first, for each filter bank output, the masking noise level as the maximum of the noise level in the training and in the testing signal. If the energy is below this threshold it is replaced by the threshold. This helps to prevent spurious distortion accumulation because those channels that are determined to have been corrupted by noise will have the same spectral value in both the training and testing tokens.

This method was originally proposed to cope with mismatches between training and testing noise levels. S. Dobler observed that this simple method is also very effective on the almost noisefree TI-Digits where there is no mismatch between training and recognition environment [3]. A threshold was chosen which lies between the average speech and average non-speech level, using the same value for each of the 15 frequency bands. Thus the logarithm, see Section 2, is replaced by a nonlinearity consisting of a constant for small power spectral density values and the original logarithm for values larger than a threshold. The measure prevents low-energy portions of the input signal from influencing the result of the log-likelihood computation.

4. SPECTRUM NORMALIZATION

The goal of spectrum normalization [11] is to remove the influence of an unknown, slowly time-varying channel transfer function on the computed features. The log-spectral density of the speech signal at the output of a linear, time-varying channel can be written as

$$g(f,t) = s(f,t) + h(f,t), \quad (2)$$

where $s(f,t)$ and $h(f,t)$ are the logarithms of the power-spectral density of the input speech signal and of the channel transfer function, respectively. As indicated, these signals are in general nonstationary, i.e. their spectral density is not only a function of frequency f but also of time t. In general, $h(f,t)$ changes much more slowly with respect to time than the speech signal $s(f,t)$. Therefore a filter with high-pass characteristics may be used to suppress $h(f,t)$. One realization is to compute the sample mean of a whole utterance (sentence) for each frequency channel $f_k, k = 1, ..., 15$, and then subtract this estimated mean vector from each feature vector of the utterance [11]. Note that this operation introduces a processing delay of the length of the utterance.

We successfully applied the above method to the TI-Digits. The suppression of slow changes of the log-spectral intensities improved performance, which confirms that they contain little information for recognition and are mainly caused by speaker and acoustic channel variations.

5. LINEAR DISCRIMINANT ANALYSIS

Linear discriminant analysis (LDA) is a well-known technique in statistical pattern classification for improving the discrimination between classes in a high-dimensional vector space [5]. The basic idea is to

find a linear transformation such that a suitable criterion of class separability is maximized. The transformation is obtained as the eigenvector decomposition of the product of two scatter or covariance matrices, the total-scatter matrix and the inverse of the average within-class scatter matrix. Recently, this technique has been successfully applied to speech recognition, for both small [4], [8] and large vocabulary tasks [1], [7].

Unlike the system reported in [4] we employ a single class-independent transformation matrix [7]. The advantage is greatly reduced computing (and storage) requirements. The matrix multiplication is done as a preprocessing step in the front-end rather than in the comparison process between the test frame and the model states, which would entail many matrix multiplications per test frame. A class in the LDA sense is associated with a hidden Markov model state, and each class contributes to the scatter matrices according to its occurrence in the training corpus, with one exception, though: To avoid a dominant influence of the class corresponding to the hidden Markov model of the background noise, only a fraction of the frames classified as background noise is used for the scatter matrix computation. Details of the training procedure can be found in [7].

LDA is used to extract the information relevant to discrimination from a large input window. m adjacent input vectors $x(n)$, see eq. (1), are adjoined to form a large input vector $z_m(n)$, e.g. in the case of $m = 3$:

$$z_3(n) := \begin{bmatrix} x(n-1) \\ x(n) \\ x(n+1) \end{bmatrix} \qquad (3)$$

i.e. $z_3(n)$ has $3 * 48 = 144$ components. As can be seen from this equation, first- and second-order time differences are explicitly included in the input vector $z_m(n)$. This performed consistently better than using a large window of only static feature vectors $y(n)$. After the LDA-based transformation the dimension of the vector is reduced to 30 - 40 components. This number is rather uncritical since the optimum of the output dimension with respect to error rate is flat.

6. EXPERIMENTAL RESULTS

All results presented are obtained on the adult speakers' portion of the TI/NIST connected digits recognition task [10]. In contrast to the original database we used a sampling rate of 8 kHz and an overall number of 17305 digit strings (training and recognition).

First we verified the effectiveness of the thresholding technique described in Section 3. We observed a relative improvement in error rate by approximately 15%, both for single emission probability densities, see *Table 1*, and for mixtures with up to 32 component densities. For all following results we thus included thresholding.

Table 1: Effect of thresholding on error rate; HMM with single density emission probability, no spectrum normalization. DEL/INS: deletion/insertion rate; WER: word error rate, SER: string error rate. All rates in [%].

Thresholding	DEL/INS	WER	SER
no	0.60/0.17	1.70	4.62
yes	0.57/0.09	1.43	4.00

Table 2 shows the influence of long-term spectrum normalization as described in Section 4: the error rate can be reduced by another 20%. This table shows further that an increase of the number of component densities per mixture and thus the acoustic resolution can significantly reduce the error rate: a doubling of the number of densities results in approximately 20% relative improvement of the error rate. However, the result for single densities is already quite impressive: 1.1% word error rate for speaker-independent recognition! For all following results spectrum normalization was included.

Table 2: Effect of spectrum normalization and acoustic resolution on error rate; L: number of component densities per mixture.

Without spectrum normalization			
L	DEL/INS	WER	SER
1	0.57/0.09	1.43	4.00
2	0.44/0.07	1.06	3.01
4	0.35/0074	0.76	2.22
8	0.23/0.05	0.56	1.68
16	0.20/0.05	0.50	1.52
32	0.16/0.05	0.45	1.40
With spectrum normalization			
L	DEL/INS	WER	SER
1	0.47/0.05	1.11	3.22
2	0.36/0.03	0.82	2.44
4	0.29/0.03	0.59	1.81
8	0.21/0.02	0.46	1.43
16	0.16/0.02	0.36	1.18
32	0.14/0.05	0.34	1.07

In the above experiments, a 32-component feature vector was used which did not contain second-order time differences since they only marginally improved

performance. When LDA was applied, up to 3 successive 48-component input vectors were adjoined to form a large acoustic vector. After the LDA-based transformation only the 32 components corresponding to the largest eigenvalues were retained. The results in *Table 1* lead to the following conclusions: First, LDA is able to take advantage of a large input window and to extract information relevant to recognition. Second, detailed acoustic models seem to be a precondition for LDA. The relative improvement in error rate is 20% in the case of 32 component densities and only 10% for 1 component density per mixture.

Table 1: Error rate in the presence of LDA-based transformation; D: Vector dimension prior to LDA; L: number of component densities per mixture.

D	L	DEL/INS	WER	SER
1*48	1	0.52/0.10	1.17	3.38
2*48	1	0.40/0.08	1.03	2.98
3*48	1	0.42/0.09	1.07	3.11
1*48	2	0.39/0.08	0.87	2.60
2*48	2	0.31/0.06	0.76	2.29
3*48	2	0.35/0.07	0.80	2.39
1*48	4	0.36/0.05	0.64	1.99
2*48	4	0.23/0.03	0.50	1.54
3*48	4	0.25/0.04	0.51	1.55
1*48	8	0.25/0.05	0.49	1.48
2*48	8	0.18/0.04	0.41	1.19
3*48	8	0.15/0.04	0.33	1.06
1*48	16	0.20/0.05	0.38	1.15
2*48	16	0.13/0.05	0.32	0.97
3*48	16	0.13/0.04	0.31	0.97
1*48	32	0.18/0.05	0.36	1.11
2*48	32	0.12/0.05	0.29	0.88
3*48	32	0.12/0.04	0.28	0.84

1. REFERENCES

[1] X. Aubert, R. Haeb-Umbach, and H. Ney. Continuous mixture densities and linear discriminant analysis for improved context-dependent acoustic models. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Minneapolis, MN, Apr 1993.

[2] R. Cardin, Y. Normandin, and R. De Mori. High performance connected digit recognition using codebook exponents. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages I505–I508, San Francisco, CA, Mar. 1992.

[3] S. Dobler. Personal communication, July 1992.

[4] G.R. Doddington. Phonetically sensitive discriminants for improved speech recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 556–559, Glasgow, UK, May 1989.

[5] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* Wiley, New York, 1973.

[6] J.L. Gauvain and C.H. Lee. Improved acoustic modeling with Bayesian learning. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages I481–I484, San Francisco, CA, Mar. 1992.

[7] R. Haeb-Umbach and H. Ney. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages I13–I16, San Francisco, CA, Mar. 1992.

[8] M.J. Hunt and C. Lefebvre. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 262–265, Glasgow, UK, May 1989.

[9] D. H. Klatt. A digital filter bank for spectral matching. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 573–576, Philadelphia, PA, 1976.

[10] R.G. Leonhard. A database for speaker-independent digit recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 42.11.1–42.11.4, San Diego, CA, Mar. 1984.

[11] H. Ney. Automatic speaker recognition using time alignment of spectrograms. *Speech Communication*, 1(2):135–149, 1982.

[12] H. Ney. Acoustic modelling of phoneme units for continuous speech recognition. In *Proc. 5th European Signal Processing Conf.*, pages 65–72, Barcelona, Spain, Sep. 1990.

[13] A. Noll, H.H. Hamer, H. Piotrowski, H.W. Ruehl, S. Dobler, and S. Weith. Real-time connected-word recognition in a noisy environment. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 679–682, Glasgow, UK, May 1989.