CONNECTED DIGIT RECOGNITION USING STATISTICAL TEMPLATE MATCHING

L. Welling, H. Ney, A. Eiden, C. Forbrig

Lehrstuhl für Informatik VI, RWTH Aachen – University of Technology
D-52056 Aachen, Germany

ABSTRACT

In this paper we describe the optimization of 'conventional' template matching techniques for connected digit recognition (TI/NIST connected digit corpus). In particular we carried out a series of experiments in which we studied various aspects of signal processing, acoustic modeling, mixture densities and linear transforms of the acoustic vector. After all optimization steps, our best string error rate on the TI/NIST connected digit corpus was 1.71% for single densities and 0.74% for mixture densities.

1. INTRODUCTION

Over the last five years much progress has been made in connected digit recognition [3, 7, 8, 9]. This paper describes how the systematic optimization of various components of a 'conventional' recognition system leads to high performance comparable with other systems that use much more complicated techniques. Experimental results on the adult corpus of the TI/NIST connected digit corpus are given. The optimization steps presented in this paper are:

1. Several methods for improved signal processing were tested.
2. Different types of density models and the use of derivatives were studied.
3. Linear transforms were employed for improved selection of the acoustic parameters used for recognition.
4. The acoustic resolution was increased by mixture densities.

The organization of the paper is as follows. Section 2 gives an overview of our baseline recognition system. Section 3 is on the results for various experiments with the signal processing part of our recognizer. Aspects of acoustic modeling are studied in Section 4. Finally, Section 5 presents the improvements due to mixture densities. Conclusions are given in Section 6.

2. BASELINE SYSTEM

2.1. Signal Analysis

This section contains a description of our baseline system. First we a perform a preemphasis of the sampled speech signal. The preemphasized samples $d(n)$ are obtained from the original samples $s(n)$ by

$$d(n) \;=\; s(n) - s(n-1).$$

Every 10 ms, a Hamming window is applied to preemphasized 15 ms speech segments. We compute the short term spectrum by a 8192–point fast Fourier transform together with zero padding. For further processing we use only the frequency range from 0 to 5 kHz although the signal is sampled with 20 kHz. We employed two alternative methods to obtain $M = 16$ cepstrum coefficients c_m:

- Method A:
  In this method we take the logarithm of the spectral magnitudes. Then the mel scale

  $$\mathrm{Mel}(f) \;=\; 2719 \tanh\left(\frac{f}{3000Hz}\right)$$

  is applied. The resulting $M = 16$ cepstrum coefficients c_m are calculated from $N = 1024$ mel–warped log magnitudes f_n:

  $$c_m \;=\; \sum_{n=0}^{N-1} f_n \cos\left(\frac{\pi m n}{N}\right) \quad , \quad 0 \le m < M.$$

- Method B:
  This method is based on 20 mel scale triangular filters [6]. We use a mel scale defined by

  $$\mathrm{Mel}(f) \;=\; 2595 \log_{10}\left(1 + \frac{f}{700Hz}\right).$$

  A filter bank in which each filter has a triangle bandpass frequency response with bandwidth and spacing determined by a constant mel frequency interval is applied to the mel spectrum, as can be seen in Fig. 1. For each filter the output is the logarithm of the sum of the weighted spectral magnitudes. Due to overlapping filters, filter bank outputs of adjacent filters are correlated. The covariance matrix of a vector consisting of the filter bank outputs has approximately Toeplitz form. Thus the filter bank outputs are decorrelated by a discrete cosine transform [2]. $M = 16$ cepstrum coefficients c_m are computed from $N = 20$ filter bank outputs f_n by

  $$c_m \;=\; \sum_{n=1}^{N} f_n \cos\left(\frac{\pi m (n-0.5)}{N}\right) \quad , \quad 0 \le m < M.$$

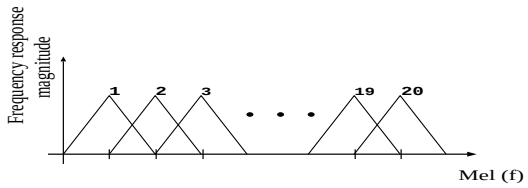

Figure 1: Filter bank with triangle bandpass frequency responses.

In methods A and B the subsequent steps are identical. A cepstral mean normalisation is carried out for every utterance in order to account for different transfer functions. In addition, the zeroth coefficient is shifted so that the maximum value within every utterance is zero (energy normalisation). Every 10 ms, a vector $y(t)$ consisting of $M = 16$ cepstrum coefficients at time t is computed. Each vector $y(t)$ is augmented by first–order and second–order derivatives. The resulting acoustic vector $x(t)$ is used for recognition:

$$x(t) := \begin{bmatrix} y(t) \\ y'(t) \\ y''(t) \end{bmatrix} = \begin{bmatrix} y(t) \\ y(t) - y(t - \Delta t) \\ y(t + \Delta t) - 2y(t) + y(t - \Delta t) \end{bmatrix}$$

In this paper we used $\Delta t = 3 \cdot 10$ ms.

Table 1 compares word error rates (WER), string error rates (SER) and the number of substitutions (sub), deletions (del) and insertions (ins) of our baseline system for analysis methods A and B. We found that the error rate for analysis method B is significantly smaller than for method A. This paper gives recognition results for both methods because analysis method B was not available for our initial experiments.

Table 1: Error rates for analysis methods A and B.

	sub/del/ins	WER [%]	SER [%]
method A	132/70/26	0.80	2.24
method B	108/49/22	0.63	1.77

2.2. Acoustic Modeling

Our baseline recognition system is based on hidden Markov models with continuous observation densities. It is characterized by:

- single Laplacian densities with state dependent deviation vectors,
- gender–dependent word models for 11 English digits including '*oh*' and gender–dependent silence models,
- 357 states plus 1 state for silence per gender,
- maximum likelihood training in the Viterbi approximation [4].

3. SIGNAL PROCESSING STEPS

We conducted a series of experiments in which we investigated the effect of signal processing steps on the error rate. All experiments in this section were carried out with signal analysis method A.

Table 2 shows recognition results for different window lengths. We found that the 20 ms window performs best.

Table 2: Word error rate (WER) and string error rate (SER) for different window lengths.

	sub/del/ins	WER [%]	SER [%]
10 ms	133/76/36	0.86	2.36
15 ms, baseline	132/70/26	0.80	2.24
20 ms	115/63/36	0.75	2.13
25 ms	122/74/64	0.91	2.60

Table 3 summarizes the effects of the preemphasis and the mel scale. The preemphasis results in a relative improvement of 9% for the string error rate. The mel scale yields a reduction in the string error rate by 18%.

Table 3: Effect of analysis steps on the error rate.

	sub/del/ins	WER [%]	SER [%]
baseline system	132/70/26	0.80	2.24
no preemphasis	133/73/40	0.86	2.45
no mel scale	151/90/27	0.94	2.74

By a cepstral mean normalisation the string error rate can be reduced by 14%, as shown in Table 4. This table also shows that the energy normalisation produces only a slight improvement in recognition performance.

Table 4: Impact of normalization steps on the error rate.

	sub/del/ins	WER [%]	SER [%]
baseline system	132/70/26	0.80	2.24
no mean norm.	167/74/38	0.98	2.61
no energy norm.	129/71/36	0.83	2.31

4. ACOUSTIC MODELING

4.1. Density Modeling

For state dependent deviation vectors we observed a relative improvement of string error rate by 13% as opposed to a deviation vector pooled over all states. Replacing Gaussian densities by Laplacians leads to 17% less string errors. These results are summarized in Table 5 for signal analysis method A.

Table 5: Effect of density modeling on the error rate

	sub/del/ins	WER [%]	SER [%]
baseline system	132/70/26	0.80	2.24
pooled dev.	133/90/30	0.89	2.56
Gaussian dens.	156/93/39	1.01	2.69

A pooled deviation vector and two densities per mixture lead to the same number of parameters as a density specific deviation vector and single densities. As shown in Table 6, the density specific deviation vector performs better. In these experiments we used method B for signal analysis.

4.2. Effect of Derivatives

The use of derivatives in the acoustic vector decreases the error rate. Table 7 (method A) shows the enormous improvement in recognition performance due to the first–

Table 6: Error rates for single densities with a deviation vector per density and mixture densities with a pooled deviation vector.

deviation	dens.	sub/del/ins	WER [%]	SER [%]
per dens.	1	108/49/22	0.63	1.77
pooled	2	62/104/26	0.67	2.09

order derivatives. The second–order derivatives do not have an evident effect on the error rate.

Table 7: Error rates for different acoustic vectors.

vector	sub/del/ins	WER [%]	SER [%]
baseline system	132/70/26	0.80	2.24
first–order deriv.	132/74/27	0.82	2.28
no derivatives	317/153/75	1.91	5.62

4.3. Whitening Transform

The error rate can be decreased by adding information about adjacent vectors to the acoustic vector $y(t)$. An alternative to the explicit incorporation of derivatives in the acoustic vector is given by the following approach:

- For time t, we form a so called spliced vector $x_s(t)$ by adjoining acoustic vectors $y(t)$ with no derivatives from frames $t - \Delta t, t, t + \Delta t$ according to:

$$x_s(t) := \begin{bmatrix} y(t) \\ y(t - \Delta t) \\ y(t + \Delta t) \end{bmatrix}$$

  We used $\Delta t = 3 \cdot 10$ ms.
- We perform a whitening transform of the spliced vector.
- We use Gaussian densities instead of Laplacians.

A Toeplitz matrix is only an approximation of the covariance matrix of the filter bank outputs. Thus some correlations among the components of the acoustic vector remain after a cepstral decorrelation. These correlations can, on the average, be removed by a whitening transform (pp. 24 in [5]) based on a pooled covariance matrix of the spliced vector. The pooled covariance matrix was calculated as follows:

1. We performed a time alignment without a whitening transform.
2. The time alignment path was then used for the sequence of spliced vectors.
3. We computed the covariance matrix using one mean vector for each state.

For experimental evaluation the spliced vector was transformed by a matrix containing the eigenvectors of the pooled covariance matrix in training and recognition. Table 8 shows the results for the acoustic vector with first–order and second–order derivatives and for the whitening transform. We used signal analysis method B. The whitening transform leads to an improvement from 1.77% to 1.71% string error rate. This was our best result for the case of single densities.

Table 8: Error rates for whitening transform and derivatives.

	sub/del/ins	WER [%]	SER [%]
derivatives	108/49/22	0.63	1.77
whitening	97/62/13	0.60	1.71

4.4. Linear Discriminant Analysis

Linear discriminant analysis (pp. 118 in [1] and pp. 445 in [5]) has already been successfully utilized for speech recognition [3, 8]. In our experiments with linear discriminant analysis (LDA) and mixture densities, 3 successive 48–component vectors $x(t-1), x(t)$ and $x(t+1)$ which included derivatives were adjoined to form a large input vector. A 48×144 gender–dependent transformation matrix was used to reduce the dimension of the acoustic vector to 48 components. The LDA classes were defined as states.

For single densities we also studied the effect of a 11–frame window of vectors without derivatives. Again the resulting acoustic vector consisted of 48 components. Such a long window performed best with Gaussian densities [10]. In Table 9 (method B) the results for Gaussian and Laplacian densities are summarized. For comparison, Table 9 also shows the error rates of our baseline system with no LDA. A 11–frame window combined with a LDA and Gaussian densities produces a string error rate of 2.03% whereas our baseline system with no LDA produces 1.77%.

Table 9: Effect of linear discriminant analysis and 11–frame window on the error rate for Laplacian and Gaussian densities.

	sub/del/ins	WER [%]	SER [%]
11–frame window:			
Laplacians	103/67/54	0.82	2.24
Gaussians	96/70/38	0.71	2.03
no LDA:			
Laplacians	108/49/22	0.63	1.77

5. MIXTURE DENSITIES

To improve the acoustic resolution, we tested mixture densities. The results are summarized in Table 10 where only method B is used for signal analysis.

5.1. Cepstral Decorrelation

As described in Section 2, we performed a cepstral decorrelation of the filter bank outputs. In order to show the improvements due to a cepstral decorrelation we also tested a filter bank analysis with an acoustic vector consisting of

- 20 filter outputs plus frame energy,
- the corresponding 21 first–order and 21 second–order derivatives.

The results for the filter bank analysis are presented in Table 10a. Table 10b shows the results for a cepstral

Table 10: Error rate reductions due to an increasing number of densities per mixture (densities/mixture) for a) filter bank analysis, b) cepstral decorrelation, c) whitening transform of the spliced vector, d) linear discriminant analysis.

densities/mixture		sub/del/ins	WER [%]	SER [%]
a)	1	177/73/44	1.03	2.98
	2	180/66/25	0.95	2.70
	4	160/54/16	0.80	2.32
	8	135/49/20	0.71	2.07
	16	121/31/17	0.59	1.76
	32	101/33/24	0.55	1.63
b)	1	108/49/22	0.63	1.77
	2	102/40/21	0.57	1.57
	4	82/38/21	0.49	1.36
	8	71/32/25	0.45	1.24
	16	66/28/22	0.41	1.16
	32	67/25/20	0.39	1.10
	64	64/16/19	0.35	0.98
c)	1	97/62/13	0.60	1.71
	2	89/52/14	0.54	1.54
	4	64/41/15	0.42	1.18
	8	46/30/15	0.32	0.97
	16	52/20/29	0.28	0.85
d)	1	96/72/37	0.72	2.08
	2	74/55/31	0.56	1.68
	4	44/43/34	0.42	1.30
	8	27/32/27	0.30	0.94
	16	21/28/25	0.26	0.83
	32	21/23/24	0.24	0.74

decorrelation. The components of the acoustic vector were

- 16 cepstrum coefficients (including energy),
- the corresponding 16 first–order and 16 second–order derivatives.

We found that a cepstral decorrelation reduced the string error rate by more than a third in all tests.

5.2. Whitening Transform

A further improvement was achieved by performing a whitening transform of the spliced vector as described in Section 3. Using 16 densities per mixture we obtained 0.85% string error rate, as can be seen in Table 10c.

5.3. Linear Discriminant Analysis

The error rates for the experiments with a linear discriminant analysis are given in Table 10d. We used 3 successing vectors with derivatives as described in Section 4.4. By performing a linear discriminant analysis we achieved our best string error rate of 0.74% on the TI/NIST connected digit corpus.

6. CONCLUSIONS

This paper has presented several possibilities of optimization of important components of a speech recognition system.

By systematically performing these optimization steps we obtained the following results on the TI/NIST connected digit corpus:

- For single densities we achieved a string error rate of 1.71% without discriminative training.
- Using up to 64 densities per state and no discriminative training, the string error rate goes down to 0.98%.
- The string error rate was reduced from 0.98% to 0.85% by a whitening transform.
- A further reduction in string error rate down to 0.74% was achieved by a linear discriminant analysis.

REFERENCES

1. R. O. Duda, P. E. Hart, "Pattern Classification and Scene Analysis," J. Wiley & Sons, New York, 1973.
2. S.B. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *IEEE Trans. on Acoustic, Speech, and Signal Processing*, ASSP–28, No. 4, August 1980.
3. G.R. Doddington, "Phonetically sensitive discriminants for improved speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, ICASSP–89*, pp. 556-559, Glasgow, May 1989.
4. H. Ney, "Acoustic modeling of phoneme units for continuous speech recognition," in *Proc. Fifth Europ. Signal Processing Conf.*, pp. 65-72, Barcelona, September 1990.
5. K. Fukunaga: "Introduction to Statistical Pattern Recognition," Academic Press, San Diego, CA, 1990.
6. S.J. Young, "HTK: Hidden Markov Model Toolkit V1.4," User Manual, Cambridge University Engineering Department, February 1993.
7. J.L. Gauvain, C.H. Lee, "Improved acoustic modeling with bayesian learning," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, ICASSP–92*, Vol. I, pp. 481-484, San Francisco, CA, March 1992.
8. R. Haeb-Umbach, D. Geller, H. Ney, "Improvements in connected digit recognition using linear discriminant analysis and mixture densities," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, ICASSP–93*, Vol. II, pp. 239-242, Minneapolis, MN, March 1993.
9. Y. Normandin, R. Cardin, R. De Mori, "High-performance connected digit recognition using maximum mutual information estimation," in *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 2, April 1994.
10. K. Beulen, L. Welling, H. Ney, "Experiments with linear feature extraction in speech recognition," in *Proc. Europ. Conf. on Speech Communication and Technology*, Madrid, September 1995.