ACOUSTIC FRONT–END OPTIMIZATION FOR LARGE VOCABULARY SPEECH RECOGNITION

L. Welling, N. Haberland and H. Ney

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

ABSTRACT

In this paper we describe experiments with the acoustic front–end of our large vocabulary speech recognition system. In particular, two aspects are studied: 1) linear transforms for feature extraction and 2) the modelling of the emission probabilities. Experiments are reported on a 5000–word task of the ARPA Wall Street Journal database.
For the linear transforms our main results are:

- Filter–bank coefficients yield a word error rate of 9.3%.
- A cepstral decorrelation reduces the error rate from 9.3% to 8.0%.
- By applying a linear discriminant analysis (LDA) a further reduction in the error rate from 8.0% to 7.1% is obtained.
- Recognition results are similar for a LDA applied to filter–bank outputs and to cepstral coefficients.

The experiments with density modelling gave the following results:

- Gaussian and Laplacian densities yield similar error rates.
- One single vector of variances or absolute deviations outperforms density–specific or mixture–specific vectors.

1. INTRODUCTION

This paper is about the optimization of the acoustic front–end of our large vocabulary speech recognition system. In particular, we study two aspects: 1) the use of linear transforms for feature extraction and 2) the modelling of emission probabilities.

Linear transforms for feature extraction are a subject of intense research interest. In particular, linear discriminant analysis (LDA) [8] has been shown to improve recognition performance on small [7, 11] and large vocabulary [1, 2] recognition tasks. A disadvantage of LDA is its poor performance in case of a mismatch in training and testing conditions [6]. Cepstral decorrelation of filter–bank outputs [4] does not suffer much from this drawback since the transformation matrix is data–independent. In this paper we compare the LDA and the cepstral decorrelation in the framework of large vocabulary continuous speech recognition. A comparison for a small vocabulary task and telephone speech was carried out in [6].

The modelling of emission probabilities in our recognition system is based on continuous mixture densities [9]. This paper compares Gaussian and Laplacian component densities and reports on experiments with a pooled, a mixture– and a density–specific vector of absolute deviations.

The paper is organized as follows. Section 2 describes the acoustic front–end of the recognizer. Section 3 is on the acoustic modelling. Experimental results on the ARPA Wall Street Journal database are reported in Section 4. Conclusions are drawn in Section 5.

2. ACOUSTIC FRONT–END

2.1. Critical band filter bank

The speech signal is sampled at 16 kHz. Every 10 ms, a Hamming window is applied to preemphasised 25–ms segments and a 1024–point fast Fourier transform (FFT) is performed. The magnitude spectrum is warped according to the mel scale [12]:

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

The obtained spectral magnitudes are integrated within 20 triangular filters arranged on the mel–frequency scale. The mid–frequency of filter n is $n/2 \cdot 270.48$ and the bandwidth is 270.48 for all filters. The filter output is the logarithm of the sum of the weighted spectral magnitudes. To suppress channel distortions, a mean normalisation is carried out for each sentence. This yields 20 normalized spectral intensities which form the 'raw' acoustic vector. In the following we describe how different acoustic vectors used for recognition are calculated from the 'raw' acoustic vector.

Table 1: Number of coefficients, first– and second–order derivatives and resulting dimension of the acoustic vector for filter bank and cepstrum coefficients.

	coeff.	Δ coeff.	$\Delta\Delta$ coeff.	dim.
filter bank	21	21	1 (energy)	43
cepstrum	16	16	1 (c_0)	33

Table 2: Dimension of the acoustic vector before and after the multiplication with the LDA matrix.

	before LDA	after LDA
filter bank	$3 \cdot 43 = 129$	43
cepstrum	$3 \cdot 33 = 99$	33

Table 3: Effects of linear transforms on the word error rate on a 5000–word task (WSJ0 Nov.'92 development/evaluation set: 10/8 speakers, 410/330 sentences, 6779/5353 spoken words; bigram language model with a perplexity of $PP_{bi} = 107$; deletions (DEL), insertions (INS) and word error rate (WER) are given in percent).

	LDA	#Dens.	Nov.'92 development set			Nov.'92 evaluation set			both sets
		(m+f)	states (m+f)	DEL	INS	WER	states (m+f)	DEL	INS	WER	WER
filter bank	no	64k+57k	5675 + 4300	2.1	1.0	9.9	4911 + 5545	1.4	1.1	8.5	9.3
cepstrum	"	61k+58k	10363 + 12944	1.6	0.7	8.3	17356 + 8100	1.0	0.9	7.6	8.0
filter bank	yes	74k+72k	2860 + 5496	1.8	0.5	7.4	6438 + 6820	1.1	0.8	6.5	7.0
cepstrum	"	75k+86k	3959 + 2870	1.6	0.6	7.4	3375 + 3457	0.9	0.9	6.7	7.1

2.2. Filter–bank Coefficients

For each 'raw' acoustic vector the average of the components is calculated, subtracted from each component and included into the vector as an approximation to the frame energy. A vector of filter–bank coefficients with a dimension of 21 is obtained.

2.3. Cepstrum Coefficients

Due to overlapping filters, the components of the 'raw' acoustic vector are correlated and the covariance matrix has approximately Toeplitz form. Therefore a decorrelation by a discrete cosine transform [4] is performed. $M = 16$ mel–frequency cepstral coefficients (MFCC) c_m are computed from $N = 20$ components of the 'raw' acoustic vector f_n by

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

The coefficients c_m form a vector of cepstrum coefficients.

2.4. Spectral Dynamic Features

Temporal derivatives are calculated by two alternative methods:

- Time differences are calculated as described in [11].
- Linear regression coefficients are calculated over a window covering 5 vectors as described in [10, pp. 194]. This method leads to smoother estimates of the derivatives than the direct difference operation.

Table 1 contains the number of derivatives and the resulting dimension of the acoustic vector for the cepstrum and the filter–bank coefficients.

2.5. Linear Discriminant Analysis

We apply linear discriminant analysis [5, 8] to acoustic vectors containing either cepstrum or filter–bank coefficients. In both cases, 3 successive vectors from times $t-1$, t and $t+1$ which include spectral dynamic features are adjoined to form a large input vector [7]. A gender–independent transformation matrix is employed to reduce the dimension of the acoustic vector. The LDA classes are defined as states. The dimensions of the acoustic vector before and after the multiplication with the LDA matrix are shown in Table 2.

2.6. Front–End Configurations

So far, we have described three linear transforms for feature extraction, namely a cepstral decorrelation, the inclusion of spectral dynamic features and a linear discriminant analysis. Figure 1 shows how the transforms are used in four alternative configurations of the acoustic front–end.

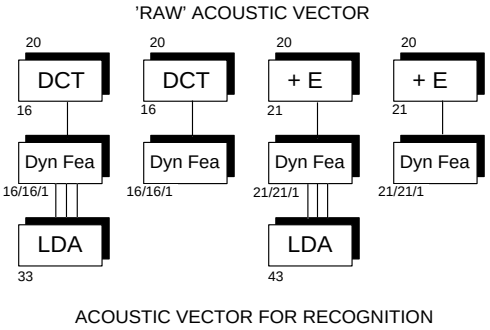

Figure 1: Four different configurations of the acoustic front–end based on cepstral decorrelation (DCT), inclusion of frame energy (+E), inclusion of spectral dynamic features (DynFea) and linear discriminant analysis (LDA).

3. ACOUSTIC MODELLING

The emission probabilities attached to each state are modelled by continuous mixture densities. The parameters of the emission probabilities are trained using the maximum likelihood criterion in combination with the Viterbi approximation, so only the best state sequence is used. For the calculation of emission probabilities the sum over all component densities of a mixture is approximated by the maximum [9]. The transition probabilities are set to a constant value which depends only on the type of the transition. Generalised word–internal triphones are derived by a decision tree method [3].

We use either Gaussian or Laplacian models with a diagonal covariance or deviation matrix that can be either pooled over all states, mixture–specific or density–specific.

4. EXPERIMENTAL RESULTS

All experiments were carried out on the ARPA Wall Street Journal (WSJ) corpus. Training was done on

Table 4: Recognition results for different spectral dynamic features on WSJ0 Nov.'92 test sets (bigram).

Dynamic features	#Dens. (m+f)	Nov.'92 development set			Nov.'92 evaluation set			both sets
		states (m+f)	DEL	INS	WER	states (m+f)	DEL – INS	WER	WER
linear regression	78k+87k	3001 + 4916	1.8	– 0.7	8.0	4916 + 3911	1.2 – 1.0	7.5	7.7
differences	78k+78k	4347 + 5141	1.8	– 0.8	7.8	2914 + 3451	1.1 – 1.4	7.7	7.8

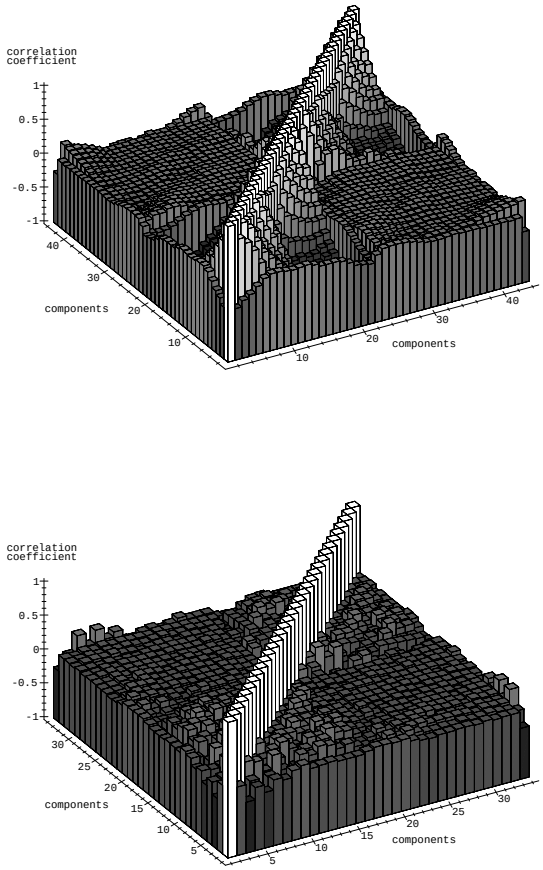

Figure 2: Correlations between the components of the acoustic vector: filter–bank coefficients *(top)* and cepstrum coefficients *(bottom)*.

the WSJ0 84–speaker corpus and testing on the WSJ0 Nov. '92 development and evaluation data. We used gender–dependent models. The size of the vocabulary was 5000 words. The number of tied states including silence was 2001. Recognition was done with a bigram model with a perplexity of 107 on the development and evaluation data.

All tables in this section show the word error rates on the development and evaluation set and the resulting error rates for both sets. In addition, the tables contain the number of component densities and the average number of active states during recognition per gender.

4.1. Linear Transforms

Table 3 summarizes the word error rates obtained with different configurations of the acoustic front–end as depicted in Figure 1. The experiments were done with Laplacian component densities, one single vector of absolute deviations and linear regression coefficients as dynamic features.

Using filter–bank coefficients as described in Section 2.2. for recognition yielded a word error rate of 9.3% on both test sets. A cepstral decorrelation as described in Section 2.3. reduced the error rate by 14% relative from 9.3% to 8.0%.

Figure 2 shows the correlation between the components of the acoustic vector for filter bank and cepstrum coefficients. For the filter–bank vector the figure indicates significant correlations. However, the correlations between the original filter–bank coefficients (components 1 to 21) and the linear regression coefficients (components 22 to 43) are relatively small. Therefore it appears to be reasonable to apply the cepstral decorrelation only to the original filter–bank coefficients and not to spectral dynamic features. Figure 2 shows that the cepstral decorrelation leads to an acoustic vector without strong correlations between its components.

However, applying a LDA as outlined in Section 2.5. to the cepstrum coefficients provided a further reduction of the word error rate from 8.0% to 7.1%. A similar error rate of 7.0% was obtained by applying the LDA to filter–bank coefficients. This result is in accordance with the invariance of the LDA criterion [5, p. 120] under linear transforms.

In summary, we found that the LDA method performed better than a cepstral decorrelation. Filter–bank coefficients yielded the highest error rates. This was also the case for mixture–specific deviation vectors for which the error rates given in Table 3 were increased by approximately 10% relative. In the following we will discuss this result.

The LDA method and the cepstral decorrelation both perform an approximative decorrelation: One step of the LDA method is a whitening transform of the within–class covariance matrix. The cepstral decorrelation results in a diagonalization of the covariance matrix of all acoustic vectors regardless of their class assignments. Since both kinds of decorrelation are beneficial for the subsequent acoustic modelling where diagonal covariance or deviation matrices are used, the error rates are decreased.

The reason why the LDA outperforms the cepstral decor-

Table 5: Effects of density modelling on the word error rates on WSJ0 Nov.'92 test sets (bigram).

Density type	Deviation/ variance	#Dens. (m+f)	Nov.'92 development set			Nov.'92 evaluation set			both sets
			states (m+f)	DEL	INS WER	states (m+f)	DEL	INS WER	WER
Laplacian	pooled	75k+86k	3959 + 2870	1.6	– 0.6 7.4	3375 + 3457	0.9	– 0.9 6.7	7.1
"	per mix.	61k+58k	3001 + 4916	1.8	– 0.7 8.0	4916 + 3911	1.2	– 1.0 7.5	7.7
"	per dens.	59k+59k	3451 + 8124	1.6	– 0.9 8.1	3793 + 4147	0.9	– 0.9 6.9	7.6
Gaussian	pooled	68k+82k	9915 + 6720	1.6	– 0.7 7.3	2744 + 5186	1.0	– 0.9 6.4	6.9

relation seems to be that the LDA method performs an optimal feature reduction using a criterion of class separability. Thus the LDA method can concentrate the relevant information for classification that is contained in a large input vector in a vector of a low dimension [6].

Another experiment was carried out to check the effect of different spectral dynamic features, namely linear regression coefficients and time differences, on the error rate. In this experiment, we applied LDA to cepstrum coefficients and we used a mixture–specific vector of absolute deviations. As Table 4 shows, the word error rates for both methods do not differ significantly.

4.2. Density Modelling

In the experiments described in this section we applied LDA to the cepstrum and we used linear regression coefficients as dynamic features.

Table 5 illustrates the effects of deviation modelling on the recognition performance for Laplacian density models. A single vector of absolute deviations pooled over all states yielded the lowest word error rate of 7.1% on both test sets. Mixture– and density–specific deviation vectors increased the error rate by less than 10% relative.

The performance of Gaussian and Laplacian densities is also compared in Table 5. The comparison was done for the case of one single vector of deviations or variances pooled over all states. As Table 5 indicates, the error rate for both test sets using Gaussian models was 6.9%. An error rate of 7.1% was obtained with Laplacian models. These results show that Gaussian and Laplacian models perform comparable.

5. CONCLUSIONS

We compared a cepstral decorrelation and a linear discriminant analysis in the framework of large vocabulary continuous speech recognition. Both transforms decreased the word error rates significantly. However, the cepstral decorrelation does not provide the recognition performance of the linear discriminant analysis.

In addition, we described experiments with density modelling. In our recognition system which uses diagonal covariance or deviation matrices and Viterbi training, one single vector of variances or deviations performed better than mixture– or density–specific vectors. The error rates obtained with Gaussian and Laplacian densities were comparable.

REFERENCES

1. X. Aubert, C. Dugast, H. Ney, V. Steinbiss, "Large vocabulary continuous speech recognition of wall street journal data," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Vol. II, pp. 129-132, Adelaide, April 1994.
2. K. Beulen, L. Welling, H. Ney, "Experiments with linear feature extraction in speech recognition," in *Proc. Europ. Conf. on Speech Communication and Technology*, Madrid, pp. 1415-1418, September 1995.
3. K. Beulen, E. Branch, H. Ney "State tying for context dependent phoneme models," in *Proc. Europ. Conf. on Speech Communication and Technology*, Rhodos, September 1997.
4. 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.
5. R. O. Duda, P. E. Hart, "Pattern Classification and Scene Analysis," J. Wiley & Sons, New York, 1973.
6. T. Eisele, R. Haeb-Umbach, D. Langmann, "A comparative study of linear feature transformation techniques for automatic speech recognition," in *Proc. Int. Conf. on Spoken Language Processing*, Vol. I, pp. 252-255, Philadelphia, PA, October 1996.
7. 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*, Vol. II, pp. 239-242, Minneapolis, MN, March 1993.
8. M. J. Hunt, C. Lefèbvre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," in *Proc. 1989 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp. 262-265, Glasgow, May 1989.
9. H. Ney, "Acoustic modeling of phoneme units for continuous speech recognition," in *Proc. Fifth Europ. Signal Processing Conf.*, pp. 65-72, Barcelona, September 1990.
10. L. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
11. L. Welling, H. Ney, A. Eiden, C. Forbrig, "Connected digit recognition using statistical template matching," in *Proc. Europ. Conf. on Speech Communication and Technology*, Madrid, pp. 1483-1486, September 1995.
12. S. J. Young, "HTK: Hidden Markov Model Toolkit V1.4," User Manual, Cambridge University Engineering Department, February 1993.