IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

MULTIPLE TIMESCALE FEATURE COMBINATION TOWARDS ROBUST SPEECH RECOGNITION

Katrin Weber

IDIAP-RR 00-29

September 2000

TO APPEAR IN

KONVENS 2000 / Sprachkommunikation

5. Konferenz zur Verarbeitung natürlicher Sprache

6. ITG-Fachtagung “Sprachkommunikation”

9.-12. Oktober 2000, Technische Universität Ilmenau, Germany

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 - 27 - 721 77 11
fax +41 - 27 - 721 77 12
email secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP-RR 00-29

MULTIPLE TIMESCALE FEATURE COMBINATION TOWARDS ROBUST SPEECH RECOGNITION

Katrin Weber

September 2000

TO APPEAR IN

KONVENS 2000 / Sprachkommunikation
5. Konferenz zur Verarbeitung natürlicher Sprache
6. ITG-Fachtagung "Sprachkommunikation"
9.-12. Oktober 2000, Technische Universität Ilmenau, Germany

Abstract: While a lot of progress has been made during the last years in the field of Automatic Speech recognition (ASR), one of the main remaining problems is that of robustness. Typically, state-of-the-art ASR systems work very efficiently in well-defined environments, e.g. for clean speech or known noise conditions. However, their performance degrades drastically under different conditions. Many approaches have been developed to circumvent this problem, ranging from noise cancellation to system adaptation techniques. This paper investigates the influence of using additional information from relatively long timescales to noise robustness. The multiple timescale feature combination approach is introduced. Experiments show that, while maintaining recognition performance for clean speech, robustness could be improved in noisy conditions.

1 INTRODUCTION

In state-of-the-art Automatic Speech Recognition systems, feature extraction techniques analyze the speech waveform and produce an acoustic vector, a representation of the speech signal suitable for further processing by Hidden Markov Models (HMM). Typically, this analysis is performed on rather short windows of the speech signal. Some contextual information is provided by adding derivatives to the original feature vector. However, there is no information covering a longer timespan, e.g., spanning the length of one syllable.

Recently it has been shown that this kind of long-term information could improve robustness of ASR systems. For examples, refer to [6] about TempoRAl Patterns (TRAPs), using additional information of up to one second), and [12] where information regarding syllables is considered. The present paper investigates another way of exploiting information on longer timespans: the multiple timescale feature combination approach. This approach is based on the combination of features from different information streams, which is related to a particular multi-band approach described in [11]. It can also be seen as one kind of temporal filtering [9]. Features calculated using windows covering different time spans of the original signal are combined to form a single feature vector, which is then processed in Hidden Markov Models the usual way.

2 MULTIPLE TIMESCALE FEATURE COMBINATION

Usually, a feature vector contains only information obtained from an analysis window of 32ms length and first and second order derivatives, and a frame's context is modeled entirely by the structure and parameters of the HMM. To introduce some additional contextual information, we added new features, obtained either by analysis over a longer timespan or by averaging over a number of subsequent feature vectors, as a second, time-synchronous, stream. For example, we took the average over 9 vectors, looking at four adjacent frames on either side of a feature vector, thus covering a time span of 112ms in total for our second timescale (see Figure 1). This corresponds roughly to the amount of contextual information frequently employed in hybrid HMM/ANN systems [1], as well as to approximately half the length of a syllable. Later experiments used even longer timespans of up to 2s for the additional information stream.

The feature combination method described above leads to comparatively big feature vectors with possibly correlated components. Therefore, we used Linear Discriminant Analysis (LDA) with the aim to reduce the vector dimension as well as the correlation and at the same time extract the most relevant information for classification [4]. LDA has been applied successfully in ASR before, e.g. in [5] and in the IMELDA system [7].

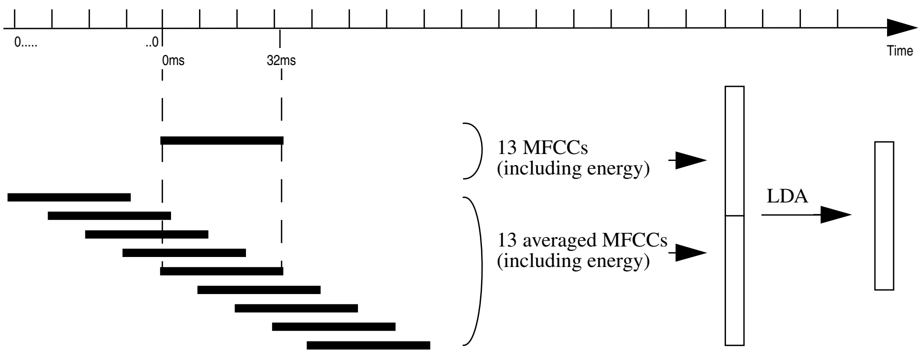

Figure 1 Combination of features from two timescales: Calculation of a feature vector from a speech file using an averaging technique for the second timescale. (The zero padding at the beginning and end of a speech file is introduced for practical reasons for the calculation of the longer timescale features.)

3 EXPERIMENTS AND RESULTS

Experiments on the feature combination approach were run using the HTK recognizer [15] on the Numbers95 [2] as well as on the Aurora [13] databases. Feature extraction, however, was in both cases performed by a software module developed at our institute. Mel frequency cepstral coefficients (MFCC) with their first and second order derivatives were used throughout. Analysis windows had a length of 32ms, and features were extracted every 10ms. Spectral subtraction and cepstral mean subtraction (or blind equalization, as in the case of the Aurora baseline system) were applied. For the tests involving LDA, another program employing HTK modules was developed.

3.1 Tests on the Numbers95 database

First tests were carried out on Numbers95, a small-vocabulary telephone speech database containing naturally spoken numbers. Final HMMs were 80 triphone models, each comprising 3 states with a mixture of 10 Gaussians in every state. Diagonal covariance matrices were used. Hyper-parameters were optimized for the baseline system and kept constant throughout the experiments. It has to be stated that keeping the same number of Gaussian mixtures for different feature vector sizes results in considerable differences in the number of parameters of the overall system, and is unlikely to be an optimal choice.

While training was done exclusively on clean data of the Numbers95 training set, testing was also executed with different types of noise (factory, lynx and car noise) from the Noisex database [14] added to the signal at different signal-to-noise ratios (SNR). The second time-scale was the average over 9 frames, as shown in Figure 1. In some experiments, LDA was used, and the 26, 21, 17, and 13 most distinctive components were extracted. Table 1 shows a comparison of the multiple timescale feature combination approach with the baseline system for the case of factory noise (word error rate (WER) on the development test set).

System (vector size)	Baseline (13)	Two Timescales (26)	Two Timescales + LDA (13)
Clean	5.6	**5.7**	7.4
SNR=18dB	7.2	**7.3**	9.5
SNR=12dB	13.0	**11.2**	15.6
SNR=6dB	26.7	**22.0**	30.7
SNR=0dB	54.4	**48.0**	59.7

Table 1 WER: Numbers95 (full development test set of about 1200 utterances) with additive Noisex FACTORY noise. In the heading, the size of the feature vector before adding first and second order derivatives is given in brackets.

As shown in this table, the introduction of a second timescale (column “Two Timescales”) improved recognition results in the case of noise with low signal-to-noise ratios, otherwise the word error rate remained about constant. The column “Two Timescales + LDA” shows results on two timescale features in combination with LDA, reducing the vector dimension to be the same as the baseline’s. Increasing the number of LDA features did not improve results. In all these experiments, LDA did not show the expected effects, as the word accuracy after this transformation decreased. This might partly be due to a non-optimal segmentation of the database for the calculation of the transformation matrix.

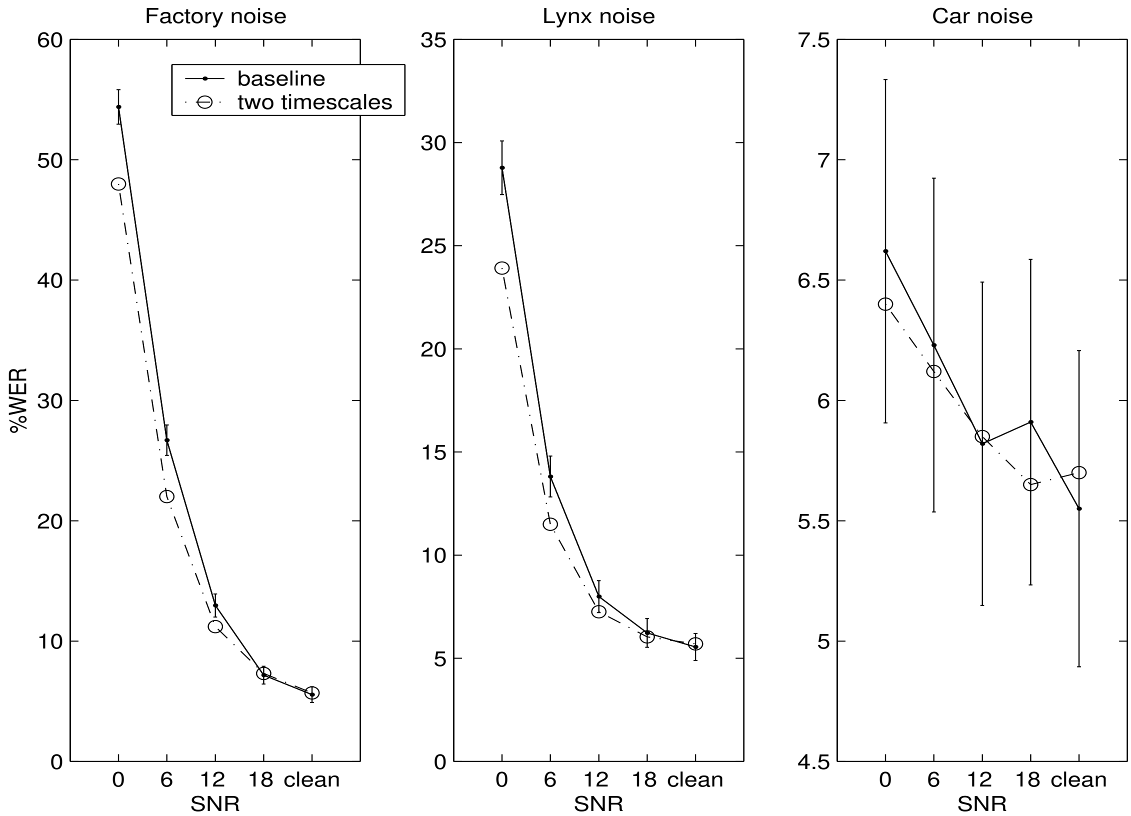

Figure 2 Comparison of the "Two timescales" system with the baseline: WER on Numbers95 with different additive noises from Noisex database. On the baseline results, error bars (corresponding to the 95% confidence interval) are shown.

Tests on the other noises lead to the same conclusions. Figure 2 shows results on the "Two timescales" system compared to the baseline for factory, lynx, and car noise. In the case of good performance of the baseline system (i.e., for high SNR, or for the case of car noise, which does not degrade recognition results very much even for SNR=0), the differences in WER between the "Two timescales" and the baseline systems are not significative, as indicated in the figure by the error bars (showing the 95% confidence interval). However, for more difficult noises and low SNR, we obtain significant improvements.

In the experiments described above, the second time-scale data were calculated on a window centered around the original one. Tests have been run using windows which started long before and finished at the current window for the second timespan, thus only looking at the 'history'. This was repeated with windows starting at the current window, so only considering the 'future' for the second timespan. None of those systems performed better than the one described above.

It should be noted that, with an optimized hyper-para-meter set (e.g. the number of Gaussians), which is adapted to the actual features used in the different experiments (and not to those of the baseline system), it may be possible to achieve even better recognition results.

3.2 Tests on the Aurora database

On the Aurora database (which, in contrast to Numbers95, was designed for training in multiple conditions, i.e. also on noisy data), the effects of introducing a second timescale (without LDA, corresponding to "Two Timescales" in Table 1) were confirmed (see Table 2). Results are compared to our single-timescale Aurora baseline system, yielding an average WER (for 0..20dB) of 13.4%. It should be noted that, apart from changing the feature extraction part (and thereby also the feature vector and model dimension), the same parameters as defined in the Aurora specification were used throughout the tests, which might not be optimal for some cases.

WER	Noise1: Exhibition Hall	Noise2: Babble Noise	Noise3: Train	Noise4: Car moving	Average of Noises 1..4
Clean	1.5	1.7	1.5	1.5	1.5
20 dB	2.1	3.0	2.0	1.7	2.2
15 dB	3.7	6.4	2.8	1.6	3.6
10 dB	8.0	15.4	5.0	2.3	7.7
5 dB	17.4	34.1	11.6	4.7	17.0
0 dB	36.0	58.6	27.5	10.5	33.2
-5 dB	66.8	78.6	53.4	29.0	57.0
Average 0..20dB					12.7

Table 2 Word error rate on the Aurora database: Two timescales, the second timescale being the average over 9 subsequent frames of the first timescale and consisting of 13 coefficients.

WER	Noise1: Exhibition Hall	Noise2: Babble Noise	Noise3: Train	Noise4: Car moving	Average of Noises 1..4
Clean	1.5	1.7	1.5	1.1	1.5
20 dB	2.2	2.6	1.5	1.0	1.8
15 dB	3.7	5.5	2.5	1.3	3.3
10 dB	7.5	13.8	4.7	2.3	7.0
5 dB	14.5	31.7	11.5	4.3	15.5
0 dB	36.7	60.0	28.0	10.5	33.8
-5 dB	70.9	80.1	59.7	29.4	60.0
Average 0..20dB					12.3

Table 3 Word error rate on the Aurora database: Two timescales, the second timescale being the average over about 2s of the energy coefficients of the first timescale.

Given the large amount of test data, the 95% confidence interval is quite small [13.2..13.6], so that we can consider the word error rate of 12.7% achieved by our multiple timescale system a significant improvement. More tests have been carried out, e.g., calculating conventional MFCCs for the second timescale features, but on a window of 128ms (yielding an average WER of 12.5%); taking the average over 17 frames (WER= 13.1%); and taking the average over 201 frames which corresponds to roughly 2 seconds (WER=37.7%). The last experiment was repeated, but for the second timescale only one coefficient (the energy) was calculated and appended to the original feature vector. This way, a WER of 12.3% was obtained (see Table 3), which is the best result obtained on all our multiple timescale systems. The same setting, but only regarding a window of one second, yielded a WER of 13.7%.

Figure 3 shows the performance of the multiple time-scale systems from Tables 2 and 3 in comparison to the baseline. The relative WER ratio

$$RR = \frac{WER(baseline) - WER(2timescales)}{WER(baseline)} \bullet 100$$

is given, positive values meaning a decrease in WER (and thus better performance) for the multiple timescale system. It can be seen that in most cases both multiple timescale systems perform better than the baseline, and that the system with just one additional component for the second timescale, calculated over 2s (light bars in the figure), performs better.

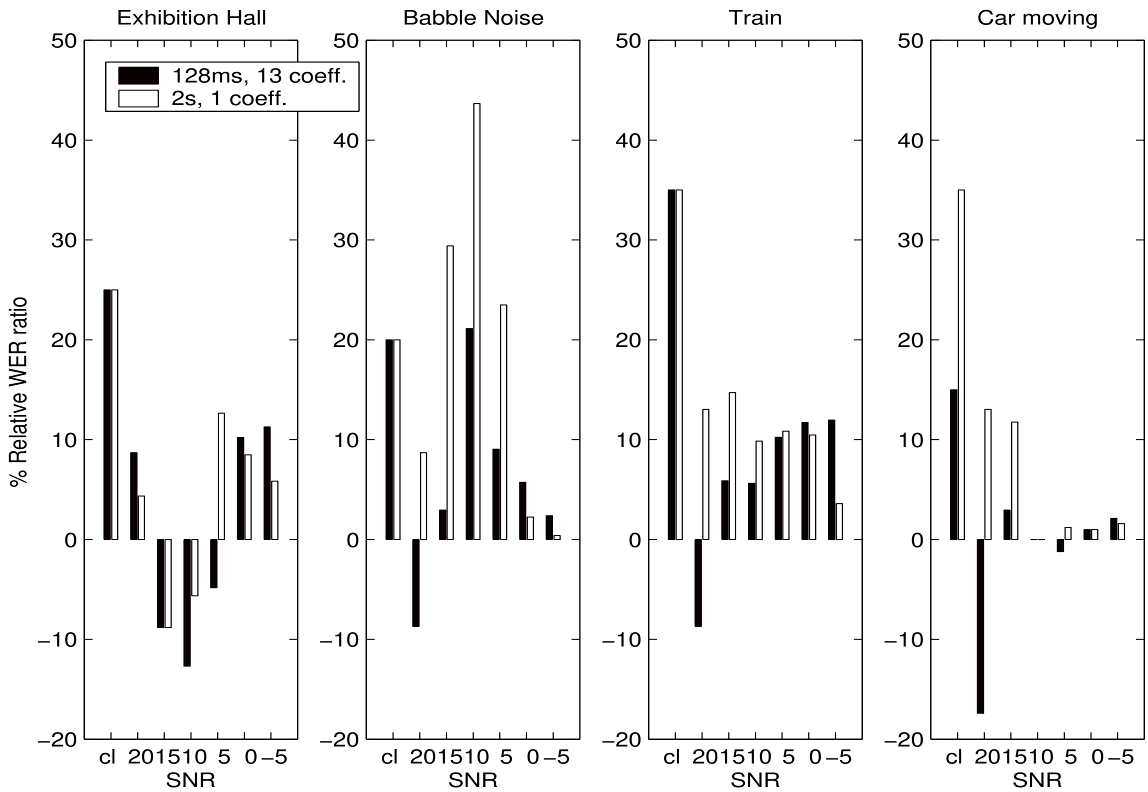

Figure 3 Aurora database: Percentage of relative WER ratio of the multiple timescale systems from Tables 2 and 3 as compared to the baseline. Positive values mean a decrease in WER, i.e., a better recognition performance.

In summary, the best tested multiple timescale system uses 14 coefficients plus their first and second order derivatives. 13 MFCCs (including energy) were calculated on 32ms of speech and one long-term energy coefficient was appended. This coefficient was obtained by averaging the energy coefficients over a timespan of approximately 2 seconds, centered around the window from which the first timescale coefficients were calculated. First and second order derivatives were appended. This way, a significant improvement was gained compared to our single timescale baseline system.

4 CONCLUSIONS AND OUTLOOK

In this article, the multiple timescale feature combination approach was investigated. It was shown to significantly increase robustness of ASR systems in the case of additive noise, while yielding a performance comparable to state-of-the-art systems for clean speech.

However, some more parameter tuning might lead to an even better recognition performance. This includes a higher number of timescales, changing lower and upper cut-off frequencies of the frequency bands used for the different timescales as well as varying the length of the analysis (or averaging) window. For some settings, LDA might proof to be advantageous. For instance, it might replace the averaging technique for the calculation of the second timescale, which was applied for the experiments described in this paper. Different feature extraction or preprocessing techniques might be used for the various timescales. Likelihood combination [11] may be used instead of (or in addition to) feature combination, pointing towards a more general multi-stream framework, as recently described in [3].

The multiple timescale feature combination approach in combination with posterior combination (employed in the same spirit as the likelihood combination in [11]) is currently investigated at our institute in the framework of hybrid (HMM/ANN) systems. Furthermore, research is conducted towards the incorporation of these techniques into the “Full Combination” approach to multiband speech recognition [10].

4 ACKNOWLEDGEMENTS

This work was supported by grant FN 2100-50742.97/1 from the Swiss National Science Foundation.

5 REFERENCES

[1] H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach.* Kluwer Academic Publishers, 1993.

[2] R. A. Cole, M. Noel, T. Lander, and T. Durham. New Telephone Speech Corpora at CSLU. *Proc. Eurospeech*, 1:821-824, September 1995.

[3] Daniel P. W. Ellis. Stream Combination before and/or after the acoustic Model. *ICSI Technical Report TR-00-007*, 2000.

[4] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press. New York, 1972.

[5] R. Haeb-Umbach and H. Ney. Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition. *Proc. ICASSP*, I:13-16, March 1992.

[6] H. Hermansky and S. Sharma. Temporal Patterns (TRAPS) in ASR of Noisy Speech. *Proc. ICASSP*, I:289-292, March 1999.

[7] M. J. Hunt and S. M. Richardson. Use of Linear Discriminant Analysis in a Speech Recognizer. *Speech Tech Worldwide*, pp. 87-93, 1990.

[8] P. McCourt, S. Vaseghi, and N. Harte. Multi-Resolution Cepstral Features for Phoneme Recognition across Speech Sub-Bands. *Proc. ICASSP*, I:557-560, May 1998.

[9] Nelson Morgan. Temporal Signal Processing for ASR. *Proceedings of Automatic Speech Recognition and Understanding (ASRU) workshop*, December 1999.

[10] A. Morris, A. Hagen, H. Glotin, and H. Bourlard. Multi-Stream Adaptive Evidence Combination to Noise Robust ASR. *Speech Communication* (to appear), 2000.

[11] S. Okawa, E. Bocchieri, and A. Potamianos. Multi-Band Speech Recognition in Noisy Environments. *Proc. ICASSP*, II:641-644, May 1998.

[12] S. Wu, B. Kingsbury, N. Morgan, and S. Greenberg. Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition. *Proc. ICASSP*, II:721-724, May 1998.

[13] D. Pearce. Experimental Framework for the Performance Evaluation of Distributed Speech Recognition Front-Ends. Aurora document number AU/120/98, Sept. 1998

[14] A. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones. The Noisex-92 Study on the Effect of Additive Noise on Automatic Speech Recognition. *Tech. Rep. DRA Speech Research Unit*, 1992.

[15] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book*. Cambridge University, 1995.