7.0 Adaptive classification structure

In the following we show an example to establish the procedure of the adaptive calculation of the classifier. First we use the principal componet analysis PCA again to show their limits.
As an example we use spoken numerals.

The aim of the voice recognition can be the classification of isolated spoken words or the recognition of interrelated spoken language, whereas the efficiency of the process should be preferably independent of the speaker – but actually it is not possible.
To achieve a good recognition rate and independent speaker system most of the devices are adaptive i.e. after a training phase they adapt to a certain speaker.

[Schürmann78] starts the speech pattern recognition with the signal of a microphone, which works up 14 channels (200 Hz to 5 kHz) with a channel vocoder and then sampling in 10 ms which is quantified in 10 bit per sample.
The developing temporally unlimited language frame is segmented in word sections and recesses.
One receives for each spoken command a language frame with a variable number of time sampling values, which are related to the length of the spoken word.

From centring, time standardisation, amplitude standardisation results a language frame with 280 components (20 time standardisation and 14 vocoder channels).

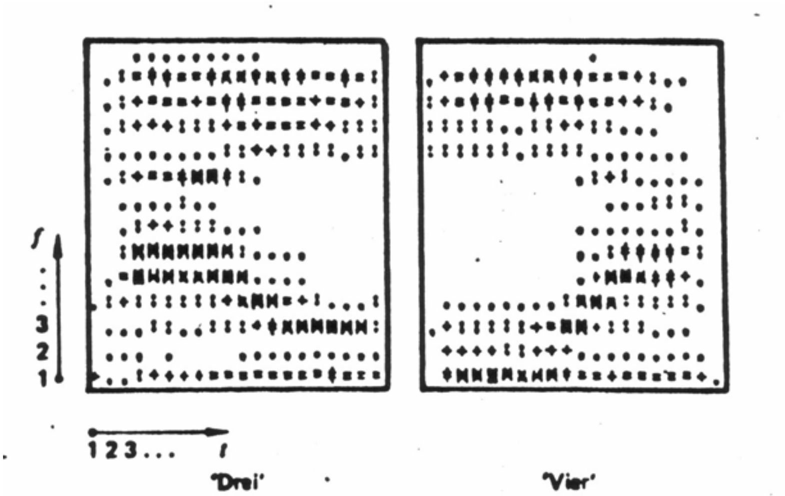

Figure 7.1: Example für 14 to 20 language frames in greyscale. The vocoder channel index is vertical from bottom to top, the time index is horizontal from left to right.

The least square method classifier is used, whereby the experiments were performed with a direct and an indirect polynomial approach.
For the direct polynomial approach the original feature-vector is used. For the indirect polynominal approach the feature-vector x is first mapped in a transformed feature-vector w and then a complete least square method classifier is calculated.

Feature transformation = principal component analysis PCA

$$w = B^T(x - \mu)$$

The principal component analysis – a linear transformation -- is characterised by n < N is the smallest possible reconstruction error regarding the reconstruction.
n = 40 was used for all experiments, which means the 280 dimensions were reduced to 40.

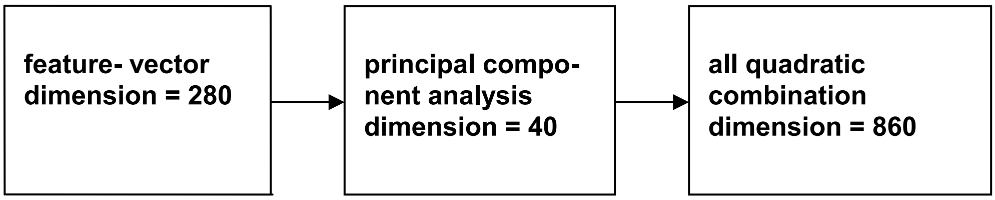

Figure 7.02: Principal component analysis.
Reduction of 280 dimension to 40 dimensions, then all quadratic combinations are calcuated to $N = m + \frac{m(m+1)}{2} = 860$ dimensions

From each speaker we have altogether 1500 words, which were spoken to 250 words each on six different days. For the adaptation the 1500 words were divided into a learning set with 1000 words and a test set with 500 words.
In order to seize the influence of the number of speakers, the available group of all ten speakers with a variable separation was divided into two groups:

- Learn speaker are involved in the adaption
- External speaker are not involved in the adaption

The borderline is shifted in steps for two speakers.The learning patterns was taken from the learning part and from the test part the test set. In addition each reclassification behavior was measured against the external patterns - formed from all words of the external speakers.
Figure 7.03 shows the change to the error rate ε related to the number of speakers. The three curves are displayed, which are based on the learning-, test- and external set. The increase of the number of speakers is a restriction of the reclassification exercise in the case of measurement of the learning set and test set, but a saturation behaviour will occur. On the other hand the error rate of the external speakers decreases with increasing number of speakers.

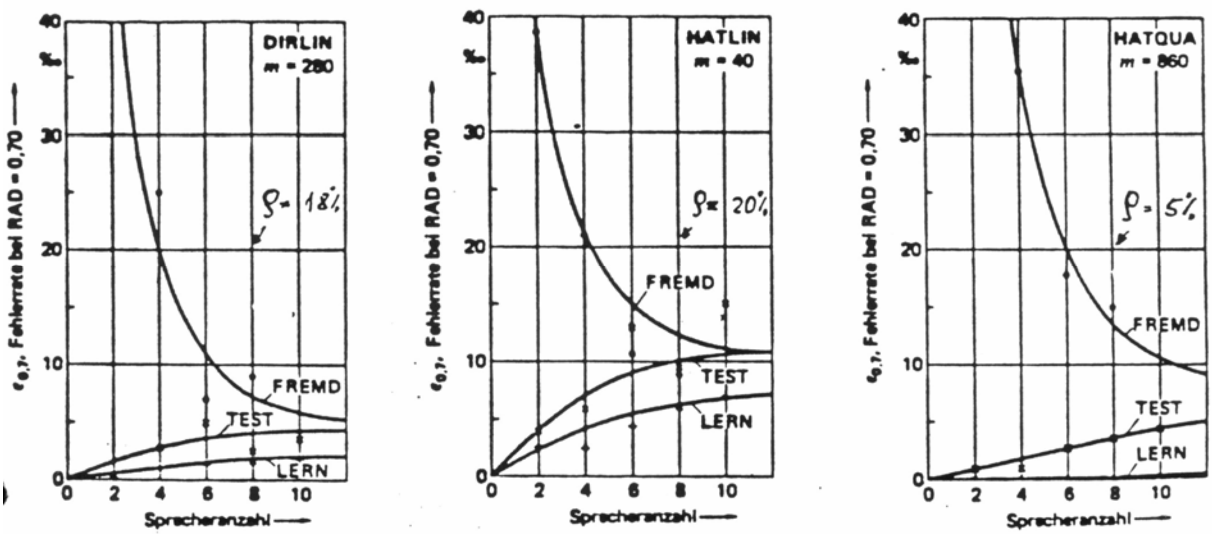

Figure. 7.03: Influence of the number of speakers on th error rate at the least square method classifier

If one takes the characteristics directly and calculates a linear classifier, the following errors will occur **(DIRLIN)**

	time-invariant 8 speaker classifier **m = 280**		
	U	**F**	**ε in %**
learn-set	8000	120	1,5
test-set	4000	320	3,5
external-set	3000	510	17,0

The result will be more badly if you only take 40 fetures of the principal component analysis. The following errors will occur (PCALIN). The following errors will occur (PCALIN).

	time-invariant 8 speaker classifier **m = 40**		
	U	**F**	**ε in %**
learn-set	8000	480	6
test-set	4000	400	10
external-set	3000	600	20

The efficiency of the completely quadratic least square method classifier shows the following table. The following errors will occur (PCAQUA).

	time-invariant 8 speaker classifier **m = 860**		
	U	**F**	**ε in %**
learn-set	8000	4	0,05
learn-set	4000	21	0,53
external-set	3000	109	3,63

U = all pattern
F = error without rejection
ε = F/U error-rate

The high error rate of the foreign sample led to correct the classifier data as a function of the manner of speaking of the speakers after each spoken word and this gives a speaker-specific classifier, which has the ability to change its classifier data.

7.1 Derivation of the adaptive classification correction formula

The technique of the recursive average value estimation

$$\mu_N = \frac{1}{N}\sum_{i=1}^{N} X_i \quad \text{time invariant mean value}$$

which gives the last pattern by separation

$$\mu_N = \frac{1}{N}\left(\sum_{i=1}^{N-1} X_i + X_N\right) = \frac{1}{N}\left(\frac{N-1}{N-1}\sum_{i=1}^{N-1} X_i + X_N\right)$$

$$\mu_N = \frac{1}{N}(N-1)\mu_{N-1} + \frac{1}{N}X_N \quad \text{ap}-\text{to}-\text{date feature}-\text{vector}$$

$$\mu_N = (1-\frac{1}{N})\mu_{N-1} + \frac{1}{N}X_N \qquad \text{generalizing}$$

$$\mu_N = (1-\alpha)\mu_{N-1} + \alpha X_N \qquad \alpha = \text{adaption constant}$$

The calculation will be applied on both mean values, which are necessary to calculated the known equation (chapter 5)., to find a new solution for the matrix A (classifier data)

$$A = E\left\{x \cdot x^T\right\}^{-1} E\left\{x \cdot y^T\right\} \Rightarrow A_N = \overline{x \cdot x^T}^{-1} \cdot \overline{x \cdot y_N^T}$$

$$\overline{xx_N^T} = (1-\alpha)\overline{xx_{N-1}^T} + \alpha x_N x_N^T$$
$$\overline{xy_N^T} = (1-\alpha)\overline{xy}_{N-1}^T + \alpha x_N y_N^T$$

A real time long-term average value is formed with $\alpha = \frac{1}{N}$. A short time average value is formed with with α= const (α = adaptation constant). With the initial equation after longer calculation

$$A_N = A_{N-1} + \overline{\alpha xx_N^T}^{-1} \cdot \underset{\substack{\text{up}-\text{to}-\text{date}\\ \text{pattern}}}{x_N} \cdot \underset{\substack{\text{estimation error}\\ \text{for } A_{N-1}}}{\left(y_N - A_{N-1}^T x_N\right)^T}$$

The weakening sequence can be replace by the unit matrix ("quick and dirty recursion") which gives the used Klassifikator correction equation [Schürmann77]

$$\underset{\text{classifier}}{\underset{\text{old}}{A_N = A_{N-1}}} + \underset{\text{correction} - \text{term}}{\alpha x_N \left(y_N - A_{N-1}^T x_N \right)^T}$$

Starting with the 8-speaker classifier (mentioned above) all classifier data A_N will be corrected after each spoken word.
The correction equation shows, that the correction is controlled by the performance of the classifier A_{N-1}.
If the target-vector is matched exactly, the classifier data are not changed.

The test set of the adaptive classifiier contains only one sample and the size of the learning sample is guided by the adaption constant α.

spoken words	errors speaker 9	errors speaker 10
1 - 50	0	0
51 - 100	1	0
101 - 200	1	1
201 - 400	2	1
401 - 600	1	1
601 - 1000	4	1
sum	**9**	**4**

A_{N-1} = time invariant classifier N = 0
$\alpha = 1*10^{-4}$

The error rate after 1000 spoken words of speaker 9 is 9 ‰ and of speaker 10 4 ‰. Without any correction the error-rate was 3.6 %. See above table (time-invariant 8 speaker classifier **m = 860)**

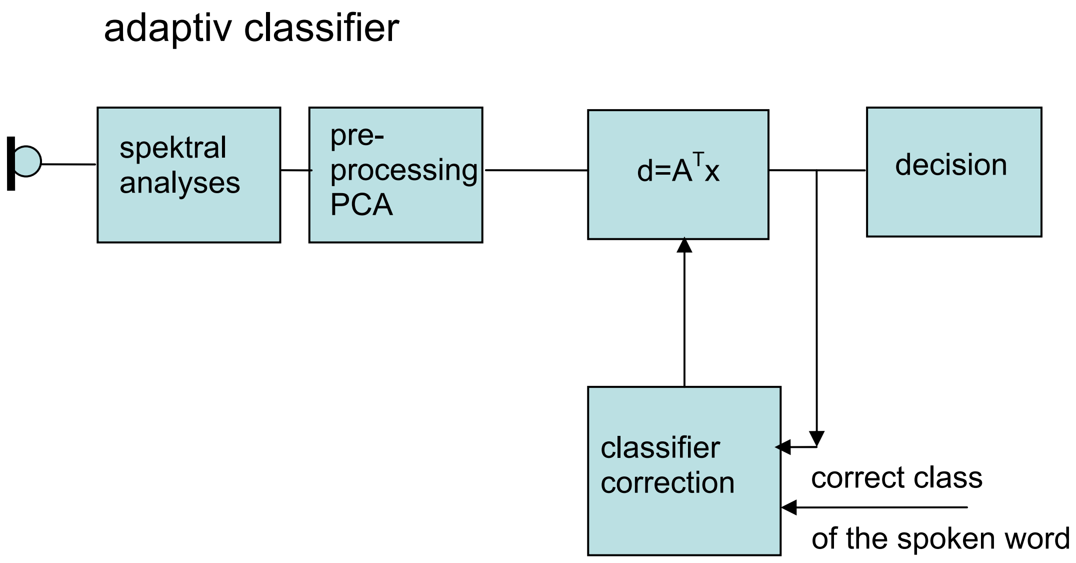

Figure 7.04: Adaptiv classifier-structure