STATISTICAL TECHNIQUES FOR AUTOMATIC SPEAKER RECOGNITION

E. Bunge*)
U. Höfker, P. Jesorsky, B. Kriener, D. Wesseling**)

*) Philips GmbH Forschungslaboratorium Hamburg, Vogt-Kölln-Str. 30, 2000 Hamburg 54
**) Heinrich-Hertz-Inst. für Nachrichtentechnik, Berlin

ABSTRACT

Within a government sponsored research program various methods of speech analysis techniques and pattern recognition methods have been applied to the speaker identification and verification problem. For this purpose a modular speaker recognition system has been developed to be used for comparative studies. Real-time speech signal analysis, mainly based on two-stage statistical measurements in combination with minimum risk classifiers allows code-word related as well as text-independent speaker verification and identification, both with very high accuracy for male and female voices.
This paper describes the structure and modules of the speaker recognition system, results of comparative experiments are being discussed.

1. THE SPEAKER RECOGNITION SYSTEM AUROS

In automatic speaker recognition, speaker verification, the two-class problem, has to be distinguished from speaker identification, a multiclass problem of pattern recognition. While verification can be used for introducing "acoustical voice passports" for banking and security systems, speaker identification is of interest for law enforcement and crime investigation. Both aspects are worked on in a joint research project by Philips Research Laboratories, Hamburg, and Heinrich Hertz Institute, Berlin.

There is a large number of feature extraction techniques and pattern recognition algorithms that can be combined for speaker recognition [1,2,3,4]. But since there is no a priori knowledge of which combination is best suited for given requirements different methods have to be investigated and compared. For this purpose, a large modular speaker recognition system for identification and verification, AUROS, has been set up to serve as a flexible instrument for designing dedicated subsystems for specified applications [5].

Fig. 1 shows the structure of the AUROS system. It consists of a general purpose computer for the pattern recognition modules, a set of real-time feature extraction processors and a dedicated computer for off-line signal analysis and simulation purpose. Most of the processors perform two stage statistical analysis in real-time: in the time domain as well as in the frequency domain, groups of events are detected segmentwise and are described by mean value vectors or frequency distributions. Feature vectors obtained by these processors are suited for the difficult task of text independent speaker recognition, assuming at least an utterance length of 10 s. An example for this kind of analysis technique, for instance, is the statistical description of "local features" like relative maxima and relative minima in short term spectra. It is registered, how often this "local feature" occured in each spectral channel.

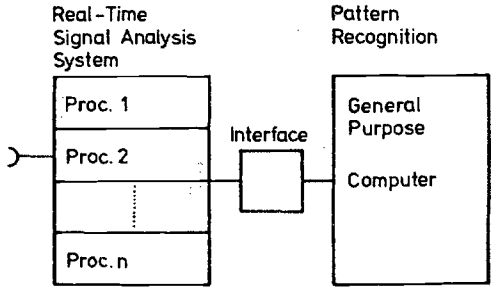

Fig. 1. Principle of the modular speaker recognition system AUROS

Fig. 2 shows the histogram of local maxima and local minima for a 10 s utterance. Another 10 statistical feature extraction methods are incorporated in AUROS.

A second class of analysis procedures based on linear transformations like FFT, Walsh Hadamard transform and inverse filtering are used to evaluate off-line cepstrum, formant frequencies, formant bandwidths, pitch and long term Walsh spectra. These analysis methods are based on short term analysis and need preceding segmentation. A novel feature

vector has been added to AUROS recently, the modified standard deviation profile, MSP, which automatically eliminates the disturbing influence of varying telephone line transmission characteristics. One component of the MSP vector is given by equation (1)

$$S_{bj} = \frac{\sum_{i=1}^{I} |a_j X_{ij} - \frac{1}{I} \sum_{i=1}^{I} a_j X_{ij}|}{\frac{1}{I} \sum_{i=1}^{I} a_j X_{ij}} \qquad (1)$$

where X_{ij} is the j-th component of the i-th short term spectrum and a_j is the unknown weighting factor of the superimposed transfer function of the actually dialed telephone line. First experiments with the MSP vectors yielded promising results [5].

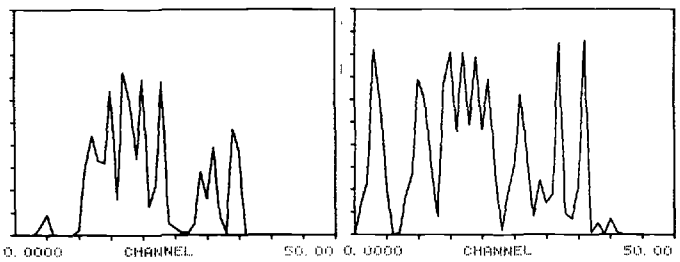

Fig. 2. Statistical speech signal analysis
a) distribution of relative minimum in short term spectra
b) distribution of relative maximum in short term spectra

The feature vectors obtained by the above mentioned techniques are fed to the general purpose computer for statistical description (evaluation of correlation matrices, variability coefficients, similarity matrices etc.) and automatic classification.

For pattern recognition, a set of different classification methods is available: linear classificator, piecewise linear classificators, minimum risk classificators, distribution-free tolerance region classificators, Mahalanobis classificator and linear regression classificator. Different distance measures and different algorithms for probability density approximation can be chosen, thus providing a large variety of methods.

A supervising program allows to select a special analysis procedure and a pattern recognition method and to combine them for a special speaker identification or speaker verification experiment. An own programming language PATSY was developed for controlling the AUROS system.

2. RESULTS

According to the boundary conditions for a practical system different methods are optimal. Therefore known algorithms have been compared with respect to their effectiveness and new methods have been investigated.

2.1. COMPARISON OF PATTERN MATCHING TECHNIQUES FOR TIME NORMALIZATION

For short time feature extraction, voices are described by the analysis coefficients of a specified 20-50 ms time interval. This special time interval has to be found in an utterance by segmentation algorithms. Pattern matching methods provide reliable information about segment boundaries of certain key points which can additionally be used to evaluate warping functions for time normalization [3].

In the frequency domain, the desired speech event is described by its sequence of spectra (reference pattern). Sliding this reference pattern along the test utterance, the distance between reference and spectra to be tested provides a measure of dissimilarity [2]:

$$A_r = \sum_{k=1}^{L} \sum_{i=1}^{M} |a(n+k,i) - a_r(k,i)| \qquad (2)$$

a(n,i) = Intensity of the i-th channel in the n-th spectrum
L = Number of spectra in the reference pattern
M = Number of spectral channels.

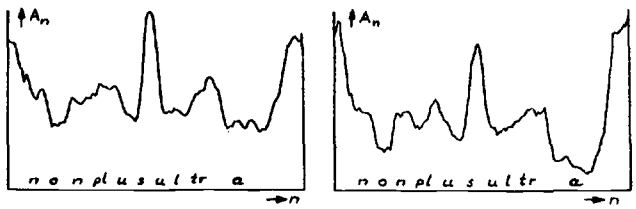

Fig. 3. Dissimilarity function of the utterance "non plus ultra" evaluated in different frequency ranges
a) 3 cs - 6000 cs
b) 3 cs - 1000 cs

Comparing linear and logarithmic scaling of the spectral energies, the latter was superior in segment detection. In further experiments, the number of frequency channels to be compared for evaluation of equation (2) was reduced. Detection of segment boundaries is more reliable and even faster when using only

a part of the spectrum. This effect is shown by the two dissimilarity functions in fig. 3, evaluated for detecting the "a" in the utterance "non plus ultra". While the desired segment boundary can easily be found when using the band-limited spectrum, ambiguity occurs for the complete spectrum.

2.2 COMPARISON OF SHORT TERM FEATURE EXTRACTION TECHNIQUES

From utterances of 12 speakers 42 ms segments were isolated and analysed using various linear transformations. The coefficients of the short term analysis describe properties of the vocal tract anatomy, and therefore are well suited for code-word related speaker recognition. Six different analysis methods have been compared with respect to computation efficiency and obtained recognition rate. From former investigations it is known, that the phonemes have different suitability for the speaker recognition task. To eliminate phoneme dependences for the different analysis algorithms, the recognition rate was evaluated separately for 10 phonemes with the highest rang or order in the list of phonemes (n, ŋ, m, z, ð, l, i, j, r, e). Then the average of the 10 recognition experiments was used for comparing the efficiency of the analysis procedures.

Fig. 4 shows the analysis coefficients for the six different methods that have been investigated. Specifications of the algorithms can be found in [8]. The feature vector dimensionality was 17 for every experiment. For classification, minimum distance classifiers were used with two kinds of distance measures

a) Euclidean distance

$$d_E^2(\underline{X},\underline{M}) = (\underline{X}-\underline{M})\cdot(\underline{X}-\underline{M})^T$$

b) Mahalanobis distance

$$d_M^2(\underline{X},\underline{M}) = (\underline{X}-\underline{M})\cdot\Sigma^{-1}\cdot(\underline{X}-\underline{M})^T$$

Σ being the pooled covariance matrix, averaged over all speaker classes.

For the experiment, the speech signals were preemphasized (1st order high-pass filter) and weighted by a Hamming window.

The average rates obtained for the six analysis methods are shown in fig. 5. Logarithmic power spectrum as well as the cepstrum seem to be superior to the other methods. Taking into account the computation time, however, the most effective analysis method is the cepstrum obtained by inverse filtering.

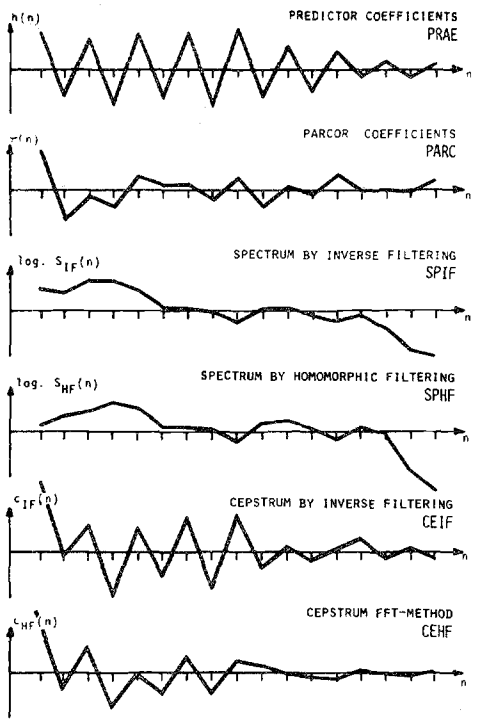

Fig. 4. 6 Examples for short term feature extraction (vowel "a")

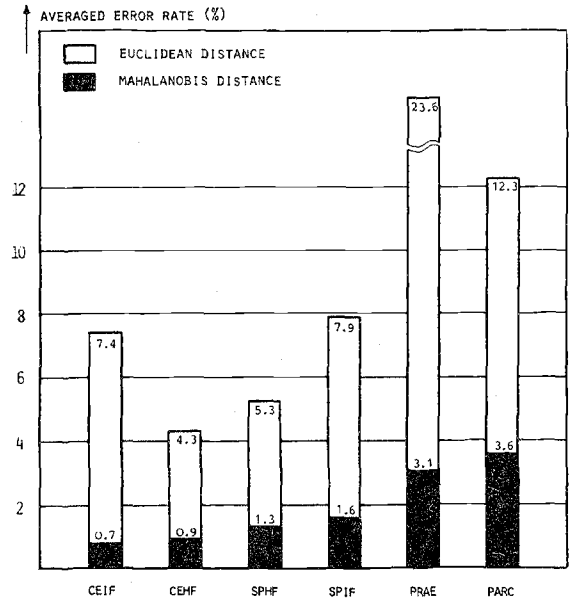

Fig. 5. Error rates for six different short term feature extraction methods averaged ofer 10 phonemes

3. SELECTION OF EFFECTIVE PATTERN RECOGNITION ALGORITHMS

Effectiveness of a pattern recognition algorithm for a practicable speaker recognition system can be evaluated by relating the recognition rate to the storage size and computation time needed for obtaining it. Details of the pattern recognition algorithm that were compared are described in [5]. In general, for identification, three basic types of classificators were compared:

a) linear classificator,
b) piecewise linear classificator, both using 6 different distance measures
c) minimum risk classificators using three kinds of multivariate probability density approximation.

The best result for identification of 2500 utterances was error free identification without rejection for the Mahalanobis classificator. But due to the large storage size needed for storing the covariance matrices and due to the large computation time, a special "economic" version of the minimum risk classificator [9] was chosen for further experiments:

Minimum risk classification is defined by equation (3)

$$R_{\underline{x}}(i) \gtrless R_{\underline{x}}(S_j) \rightarrow \begin{cases} \underline{x} \rightarrow S_j \\ \underline{x} \rightarrow S_i \end{cases} \qquad (3)$$

The vector $\underline{x}$ is classified into that class S_i with the smallest classification risk $R_{\underline{x}}(S_i)$.

Defining a cost matrix

$$C(S_i|S_j) \text{ with the elements } C(S_i|S_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad (4)$$

with equal costs for each error, and applying Baye's rule lead to the simplified classification rule:

$$p(\underline{x}|S_i) \gtrless p(\underline{x}|S_j) \rightarrow \begin{cases} \underline{x} \rightarrow S_i \\ \underline{x} \rightarrow S_j \end{cases} \qquad (5)$$

For using this classification rule, the knowledge of the conditional multivariate probability density functions $p(\underline{x}|S_i)$ is necessary. Different methods for approximating these functions have been compared: histogram techniques and Parzen window estimation. Rectangular and Gaussian shaped windows were used. Parzen window estimation seemed to be superior to histogram techniques with respect to the recognition rates. But since the latter method was 20 times faster, effort was made to optimize the histogram technique by varying the number of registration intervals and by adapting the histogram boundaries to the minimum and maximum valued samples. These optimations and additional normalization of the feature vectors to equal pattern energy increased the recognition rate for this "economic" version of a minimum risk classificator considerably. Using this algorithm, recognition rates of 99% were obtained codeword related as well as text-independent for two voice data bases with 2500 utterances of 50 speakers each. For feature extraction, statistical methods like long term averaged spectra and distribution of "local features" were used. Speaker verification yielded error rates $\leq 1\%$ for false acceptance and false rejection for the same voice data bases. Details can be found in [5].

REFERENCES

[1] Rosenberg, A.: "Automatic Speaker Verification: A Review", Proc. of the IEEE, Vol. 64, No. 4, 475-487, April 1976

[2] Atal, B.S.: "Automatic Recognition of Speakers from their Voices", Proc. of the IEEE, Vol. 64, No. 4, 460-475, April 1976

[3] Doddington, G.: "A Method of Speaker Verification", Thesis, University of Wisconsin, 1971, 71-16, 071

[4] Bunge, E.: "Automatic Speaker Recognition by Computers", Proc. of 9-th Ann. Carnahan Conf. on Electronic Crime Countermeasures, Lexington, Ken., May 1975

[5] Bunge, E.: "Comparative Investigations on Automatic Identification and Verification of Cooperative Speakers", Dissertation, Technical University of Darmstadt, to be published 1977

[6] Kriener, B.: "Vergleich verschiedener Segmentierverfahren bei der Sprechererkennung", Proc. of the 5-th IITB Kolloquium, Karlsruhe 1976

[7] Höfker, U.: "Die Eignung verschiedener Sprachlaute für die automatische Sprechererkennung", Proc. of the 5-th IITB Kolloquium, Karlsruhe 1976

[8] Jesorsky, P.: "Merkmalsgewinnung als Teilaufgabe der automatischen Sprechererkennung", NTZ, Dec. 1976

[9] Bunge, E.: "Statistical Techniques for Speaker Recognition", Proc. of the 5-th Meeting of the Deutsche Arbeitsgemeinschaft für Akustik", DAGA 76, Heidelberg, Sept. 1976.