PC-Based System for Robust Speaker Recognition[1]

Stefan Hadjitodorov and Boyan Boyanov

Central Laboratory of Biomedical Engineering, Bulgarian Academy of Sciences, Sofia, Bulgaria

A PC-based system for robust speaker recognition is proposed. It includes three one level recognition methods and a two level classifier. New procedures for voice analysis are proposed: a) Robust periodicity/aperiodicity separation by neural networks; b) Robust pitch period detection; c) Analysis of the temporal, spectral and cepstral speech characteristics. Several pattern recognition methods are implemented, because they allow analysis of different static and dynamic characteristics of the speech parameters:

1) Prototype distribution maps (*PDM*). The *PDM* is used because: a) weight vectors of *PDM*'s neurons try to imitate the probability density function — pdf (whatever complex the form of the pdf is) and less significant *PDM*'s neurons are eliminated by filtering.

2) AR-vector models (ARVM). The ARVM are used because they model the evolution of speech parameters.

3) The covariance approach combined with the arithmetic-harmonic sphericity measure, because this method performs effective speaker recognition over noisy signals.

4) Two level classifier, incorporating the discriminant capabilities and classification power of the multilayer perceptron (MLP) with the pdf's estimating, statistical modeling and compressing power of the *PDM*. The first level consists of several *PDM*s and the second — of MLP networks.

The experiments show that the proposed system is an efficient and useful tool for speaker recognition over clean and noisy signals.

Keywords: Speaker identification, Neural networks, Self-organizing map, MLP network, Two-level classifier.

1. Introduction

Over the last years speaker recognition has been becoming an important tool for: access control in high security areas, access and realization of banking operation via telephone lines, forensic applications-automatic computerized phonoscopy, etc. The speaker recognition could be speaker identification or speaker verification. The purpose of speaker identification is to identify a speaker among a set of speakers, based on the individual's utterance. Speaker identification systems may be closed-set and open-set. If the speaker is *a priori* known to be a member of a set of M speakers, the system is closed-set. Open-set speaker identification includes an additional possibility — existence of a speaker not belonging to the set of M known speakers. Another feature which is classically used to specify a speaker recognition system is whether it is text-dependent or text-independent. Text- dependent systems require that a specific speech utterance pronounced by a test speaker be identical with the speech material used for training. The text-independent systems identify the speaker regardless of his utterance. The objective of the speaker verification is to verify the person's claimed identity. This paper focuses on the text-independent closed-set speaker identification.

There are many methods and approaches (Bennani, 1991; Bimbot, 1992; Bimbot, 1995; Farrell,1994; Furui, 1981; Hadjitodorov, 1994; Hadjitodorov, 1997b; Montacie, 1992; Morgan, 1991; Matsui, 1992; Reynolds, 1995) for

[1] Work supported by National Fund "Scientific Research" — contract TN 469/94.

pattern recognition used for speaker recognition. However every approach or method is characterized by some advantages and inconveniences. Generally, most of the systems and researches are based on one method/approach. There are only few publications devoted to multilevel classifiers applied for speaker recognition. In (Castellano,1996) a multilevel fuzzy classifier is proposed and in (Naik, 1994) a combination between Hidden Markov Models and multilayer perceptron is implemented. Here we propose a flexible tool for solving the task of speaker recognition in different practical situations characterized by:

a) different level of the required accuracy (banking, forensic or general applications);

b) different types of the speech signal — clean, noisy or over telephone channel;

c) different requirements for the task performance- on line or off line.

In various conditions a specific method or methods will be optimal concerning the accuracy and the performance. In order to allow broader application of the proposed system two accurate and well known methods (AR-vector models (ARVM) and the covariance approach combined with the arithmetic-harmonic sphericity measure) are implemented in it and two are proposed by the authors (Prototype distribution maps (*PDM*) and a two-level classifier). These methods use different statistical characteristics, as well as different static and dynamic information of the speech parameters. Accuracy, efficiency and advantages of these methods have been proven by several experimental researches (Bimbot, 1992; Montacie, 1992; Hadjitodorov, 1997a; Hadjitodorov, 1997b; Bimbot, 1995; Reynolds, 1995;Hadjitodorov, 1994). Any user of the proposed system has the opportunity to decide which of the implemented methods to apply in a given condition.

In addition, the proposed system is built around a low cost personal computer (PC) with its standard sound (audio) card like "Sound Blaster", allowing realization of a low cost speaker recognition.

The system is a specialized software package based on the above mentioned methods, algorithms and approaches for speaker recognition. For more accurate evaluation of the speech parameters new speech analysis procedure is proposed:

a) Robust periodicity/aperiodicity separation by means of neural networks;

b) Robust pitch period (To) detection;

c) Evaluation and analysis of the temporal, spectral and cepstral characteristics of the speech;

d) Evaluation of the group delay function by different procedures for low and for high pitched voices.

2. Evaluation of the Speech Parameters

To minimize the errors during the speech parameters evaluation, the following procedure for speech analysis is proposed and used:

2.1. Segmentation of the speech signal

The quantized signal is divided into segments with length three To by means of a Hamming window. The duration of the segments is dynamically adapted to 3 To (using To from the previous segment), because our experiments (Hadjitodorov, 1997a) have shown that such segment's length is optimal for To evaluation. Overlapping between the segments is two pitch periods in order to analyze the dynamics of the speech parameters. The length of the first segment is 30 ms in order to assure that the window contains at least 2 To.

2.2. Periodicity/aperiodicity separation

The periodicity/aperiodicity separation (PAS) is very important for correct To detection (because errors in PAS will produce drastic errors in To) and for correct evaluation of the speech parameters. To minimize the number of errors in PAS the detector proposed in (Boyanov, 1997) is implemented, because it is characterized by:

1. Parallel analysis of the speech in time, spectral and cepstral domains. In this way different characteristics of the signal in these domains are used and the signal is analyzed more completely and from different view points.

2. Realization of robust PAS by means of multilayer perceptron (MLP) neural network. As a result the accuracy is improved, because the

MLP are characterized by good discriminant capabilities and high classification power.

In order to minimize the influence of the noisy components the aperiodic segments are eliminated.

2.3. Pitch period (To) evaluation

In order to evaluate To correctly the robust hybrid pitch detector (Boyanov, 1993) is used.

This method is implemented, because it has the following useful properties:

a) rejects practically most of the segments, where To is wrongly evaluated (experimental research over 200 speakers (Hadjitodorov, 1997a)) when the signal is preprocessed by the PAS detector (Boyanov, 1997). However it eliminates up to 1% of the voiced segments. The loss of these segments may be tolerated, because for all the speakers (in our data base) the analyzed sentences are relatively longer and contain always more than 200 segments.

b) evaluates To correctly from clean, noisy and telephone speech;

c) realizes parallel analysis of the speech signal in temporal, spectral and cepstral domains;

d) evaluates the pitch period by means of logical analysis of the results from these three domains. For every p-th segment (containing 3 To) the mean pitch period (Tom(p)) is calculated.

2.4. Cepstral analysis (over the voiced segments)

Many experimental researches (Atal, 1974; Assaleh, 1994; Furui, 1981; Farrell, 1994) have shown that the LP-derived cepstral coefficients $(c(n))$ are very informative for speaker recognition. In the proposed approach the $c(n)$ are calculated for voiced segments in order to minimize the influence of the noisy components. The cepstral analysis is carried out by means of the standard LPC analysis procedure by means of the autocorrelation method (Rabiner, 1978) and then the first 16 LPCC coefficients are calculated. The number of $c(n)$ is 16, because our experimental research (Hadjitodorov, 1997a) has shown that the first 16 coefficients are the most informative for speaker recognition for our data base.

2.5. Evaluation of the group delay function (for the voiced segments)

It is known (Hollien, 1990; Hadjitodorov, 1997a) that the analysis of spectrograms and sonograms (representing the formant structure) is very useful for speaker recognition. However, estimation of the formants is a difficult problem not yet completely solved. That is why the formant structure is analyzed by evaluation and analysis of the group delay function (GDF) (GDF is the negative derivative of the phase spectrum). The GDF is used for approximation of the formant structure, because the GDF has the following useful properties (Murthy, 1989):

1. The GDF is proportional to the squared magnitude response near resonances (formants) and approaches zero asymptotically for frequencies away from the frequency of the resonator. In this way the formants are represented by distinct and sharp peaks in the GDF.

2. The vocal tract may be represented by a cascade of resonators. The GDF of such system is the sum of GDFs of these resonators. As a result the influence of one resonator on another is minimized — even closely spaced formants are represented in the GDF by separated peaks.

The main problems in calculation of the GDF are:

1. The phase function is wrapped by the presence of zeros near the unit circle and the signal windowing prior the spectral analysis (Murthy, 1989).
2. Most of the methods for phase unwrapping do not yield satisfactory results (Nashi, 1989).
3. The spectral resolution in the GDF for medium and high pitched voices (To < 5 ms and Fo > 200 Hz), when GDF is calculated from $c(n)$, is decreased. This is due to the short To minimizing the number of $c(n)$ and the number of GDF coefficients $(gdf(i))$ respectively. For To, shorter than 5 ms, the spectral resolution will be less than 262 Hz, because in our experiments the sampling rate is 21 KHz.
4. Distortions (represented in most of the cases by extra peaks) caused by the influence of the glottal source.

In order to solve these problems and to guarantee spectral resolution higher than in the standard wide band sonogram used for formant analysis (here 262 Hz), the following approach for GDF calculation is proposed and used:

a) Analysis of low pitched voices

For such voices the GDF is calculated indirectly (without phase calculation and unwrapping) from $c(n)$ — detailed proof is given in (Murthy, 1989).

In order to increase the robustness of speaker recognition (especially the impostor elimination) the information for the vocal tract (represented by the cepstral coefficients $c_v(n)$ corresponding to the vocal tract) is used. This is performed because characteristics of the vocal tract are practically very difficult to be imitated in contrast to the glottal source characteristics (many actors are able to imitate the voice of known persons — i.e. the glottal source characteristics). The cv(n) are separated by means of liftering with a lifter having length (L) shorter than the pitch period (detailed description of the lifters is given in (Rabiner, 1978)) in order to eliminate the influence of the glottal source. In the proposed system the value of L = 0.8To, because our experiments (Hadjitodorov, 1997a) have shown, that such length is sufficient for suppression of the glottal source influence for low pitched voices. The GDF is calculated by means of the formulae described in(Murthy, 1989):

$$GDF(i) = \frac{2\pi}{N} \sum_{n=0}^{N-1} nc(n) \cos\left(\frac{2\pi ni}{N}\right) \qquad (1)$$

where: N — number of cepstral coefficients, used for GDF calculation.

In our experiments, the value of N is limited to 80, because the shortest To used for calculation of the $gdf(i)$ from $c(n)$ is 5 ms and the sampling rate is 21 KHz.

b) Analysis of high pitched voices

For such voices the GDF is calculated by means of the following procedure described in (Duncan, 1989):

a) transformation of the voiced speech into minimum phase signal;

b) direct calculation of the phase spectrum from this minimum phase signal;

c) calculation of the first derivative of the phase spectrum — the GDF.

Unfortunately for the high pitched voices the influence of the glottal source is not suppressed.

However, in most of the practical cases of speaker recognition, the male voices analyzed have appeared to be generally characterized by low values of the pitch period.

The first S GDF coefficients — $gdf(i)$ ($i = 1, \ldots, S$) are used as feature vectors representing the formant structure. The value of S is determined on the basis of the pass band (300-3000 Hz) of the phone lines, because one of the main applications of this system might be speaker recognition over phone lines. To cover this spectral range the value of S is 12, because in our experiments the resolution in the GDF is 262 Hz.

The following input vectors are formed for every p-th (for $p = 1, \ldots, P$) voiced segment: mean pitch period ($\text{To}_m(p)$); first 16 $c(i)$ and the first 12 $gdf(i)$.

3. Pattern Recognition Methods Used in the System

3.1. The prototype distribution map (*PDM*)

Detailed description of this method is given in (Hadjitodorov, 1994, Hadjitodorov, 1997a; Hadjitodorov, 1997b). The *PDM* is based on the Self-organizing feature map (SOFM) of Kohonen (Kohonen, 1990; Kohonen, 1984).

3.1.1. SOFM formation

The n-dimensional input vectors $y(t)$, $t = 0, 1, \ldots, N$ representing the speech of M different classes(speakers) are projected as neurons on a two-dimensional ($q \times q$) square map. Each neuron is defined by a n-dimensional model vector $\{m_{i,j}\}$, where $m_{i,j}$ corresponds to the (i, j)-th neuron. At the beginning of self-organization process the algorithm is initialized with random values for the $m_{i,j}(0)$. Placements of the input vectors on the map are optimized by iterative corrective steps(Kohonen, 1990). At each step the model vector $m_{a,b}(t)$ that is closest to $y(t)$ is determined according to the Euclidean distance. The best matching model vector $m_{a,b}(t)$

and the model vectors of the surrounding locations(neurons) are corrected by a specific rule (Kohonen,1990).

3.1.2. Formation of the prototype distribution map (*PDM*)

The feature vectors of each speaker(class) s, $s = 1, \dots, M$ are passed again through the already joint (commonly) trained SOFM. As a result for each class at each neuron in SOFM the frequency of activation ($f_{i,j}$; $i, j = 1, \dots, q$), i.e. the number of input vectors activating this neuron (according to the minimum of Euclidean distance), is obtained. The values of $f_{i,j}$ are normalized by dividing by the number of all speaker's vectors. Thus for each speaker a new map, containing the frequencies of activation, is formed. This map is named prototype distribution map — *PDM*. The *PDM* is used because(Hadjitodorov, 1994; Hadjitodorov, 1997a; Hadjitodorov, 1997b):

1. The *PDM*'s neurons try to imitate the probability density function (pdf) of the input signals, whatever complex the form of the pdf is. This property is due to the fact that the *PDM* is formed on the basis of the already trained SOFM.

2. The *PDM* allows dimensionality reduction — a two-dimensional SOFM with n-dimensional weight vectors is transformed into a two-dimensional map with one dimensional weight vectors (frequencies of activation).

3. Less significant neurons in the *PDM* can be eliminated by filtering of the map according to the following rule:

$$\begin{aligned}&\text{if } 0 \le f_{i,j} < k \cdot f_{\max} \text{ then } f_{i,j} = 0;\\ &\text{if } k \cdot f_{\max} \le f_{i,j} \le f_{\max} \text{ then } f_{i,j} = f_{i,j},\end{aligned} \tag{2}$$

where $0 < k < 1$ is a filter coefficient experimentally adjusted; $f_{\max}$ is the maximal value of $f_{i,j}$.

3.1.3. The *PDM* classifier

Components of the speaker's input vectors are normalized (by dividing by the norm of the vector) before being subject to the training and decision making stages.

Training the PDM: Using the n-dimensional (here $n = 29$) training input vectors of the speakers a joint two-dimensional SOFM is trained. On the basis of the SOFM for each of the M speakers from the data base the filtered $PDM(s)$, ($s = 1, \dots, M$; M-number of the reference speakers) is formed.

Speaker recognition: The following classification procedure is used:

a) formation and filtering by means of procedure (2) of the PDM^U of the unknown speaker (speaker U);

b) calculation of the similarities ($Sim(s)$) between the PDM^U and each of the $PDM(s)$ by means of the following cross-correlation type measure proposed in (Hadjitodorov, 1994):

$$Sim(U, S) = \sum_{i=1}^{q} \sum_{i=1}^{q} (f_{i,j}^U \cdot f_{i,j}^S)(d + |f_{i,j}^U - f_{i,j}^S|)^{-1} \tag{3}$$

where: $\{f_{i,j}^U\}$ — frequency of activation (value of a neuron) in PDM^U;

$\{f_{i,j}^S\}$ — frequency of activation (value of a neuron) in $PDM(s)$;

$0 < d < 1$ — experimentally determined constant (here $d = 0.02$).

c) the speaker U is identified as the speaker with the maximal value of $Sim(u, s)$.

3.2. The Auto Regressive Vector models (ARVM) classifier

Detailed description of ARVM is given in (Bimbot, 1992; Montacie, 1992). Standard ARVM is a generalization to the vectorial case of the widely used scalar autoregressive modeling technique. For each speaker its ARVM is determined. In fact these models are series of prediction coefficients matrices (A_k — for the forward model and $\mathring{A}_k$ for the backward model). They are computed by solving an equation system containing the block-Toeplitz matrix obtained from the lag covariance matrix of the input vectors. The ARVM are used because they allow modeling of the evolution of speech parameters (Bimbot, 1992; Montacie, 1992). Classification is done by calculating the forward-backward symetrized (FBS) Itakura distances between ARVM(U) and each of the reference

ARVM(s). This distance has four terms of similar type . The first one is :

$$\mu(U, S) = \log\left(\mathrm{tr}\,\frac{\mathbf{A}_s[\mathcal{T}]\mathbf{A}_s^T}{\mathbf{B}[\mathcal{T}]\mathbf{B}^T}\right), \qquad (4)$$

where: $\mathbf{A}_s$ is the matrix containing the matrix series for the forward model of the s-th speaker;
$\mathbf{A}_s^T$ is the transpose of $\mathbf{A}_s$;
$\mathbf{B}$ is the matrix containing the matrix series for the forward model of the unknown speaker U;
$\mathbf{B}^T$ is the transpose of $\mathbf{B}$;
$\mathcal{T}$ is the block-Toeplitz matrix of the unknown speaker U.

Changing the matrices for the forward model with these for the backward model and changing the places of **A** and **B** together with replacement of the unknown speaker's block-Toeplitz matrix $\mathcal{T}$ with the s-th speaker's block-Toeplitz matrix **X**, the other terms of the FBS Itakura distance are obtained.

Training the ARVM: During the training for all the reference speakers their ARVM(s) are formed using the procedure (Bimbot, 1992).

Speaker recognition: The classification is done by formation the ARVM of the unknown speaker-ARVM(U) and computation of the FBS Itakura distances between ARVM(U) and each of the ARVM(s). The unknown speaker U is classified as the speaker with the minimal FBS Itakura distance $\mu(U, S)$.

3.3. The Gaussian speaker's models combined with the arithmetic-harmonic sphericity measure (GMAHSM)

Detailed description is given in (Bimbot, 1995; Reynolds, 1995).The Gaussian speaker's models are used because:

a) These models allow robust speaker recognition when noisy and telephone speech signals are analyzed (Reynolds, 1995).

b) The Gaussian speaker's models may be combined with the arithmetic-harmonic sphericity measure (Bimbot, 1995; Bimbot,1993). This measure is symmetric and based on the eigenvalues λ_p of the matrix $\mathbf{X}_u\mathbf{X}^{-1}$s , where $\mathbf{X}_u$ is the covariance matrix of the unknown speaker and $\mathbf{X}^{-1}$s is the inverted covariance matrix of the s-th reference speaker. The measure is as follows:

$$\mu(S, U) = \log\left(\frac{A}{H}\right), \qquad (5)$$

where

$$A(\lambda_1, \lambda_2, \ldots, \lambda_p) = \frac{1}{p}\sum_{i=1}^{p}\lambda_i \qquad (6)$$

$$H(\lambda_1, \lambda_2, \ldots, \lambda_p) = p\left(\sum_{i=1}^{p}\frac{1}{\lambda_i}\right)^{-1} \qquad (7)$$

are respectively the arithmetic and harmonic means of the eigenvalues. (Bimbot, 1993) describes an efficient procedure which does not require the explicit extraction of the eigenvalues for the computation of the measure (5).

Training: The covariance matrices (COV(s)) for each of the known speakers are calculated.

Speaker recognition: The covariance matrix (COV(U)) of the unknown speaker (U) is calculated. The arithmetic-harmonic sphericity measure(AHSM) between each of the reference speakers and the unknown speaker is evaluated. The speaker U is classified as the speaker with the minimal AHSM $\mu(S, U)$.

3.4. The two level classifier (2LC)

Detailed description of the 2LC is given in (Hadjitodorov, 1997a; Hadjitodorov, 1997b).The applied two level classifier shown in Fig 1. incorporates the pdf's estimating, statistical modeling and compressing powers of the *PDM* (see Section 3.1) with the discriminant capabilities and classification power of the multilayer perceptron (MLP) nets. In fact, on the first level a preprocessing (compression and transformation) is done and on the second level the final classification is carried out. As a result, the classifier is better than either *PDM* or MLP used separately.

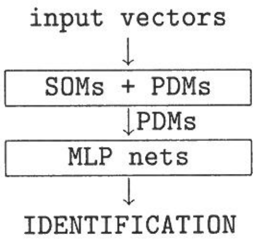

Fig. 1. The two-level clasifier.

A. The first level of the classifier consists of several ($T, T > 1$) ***PDM*s.** The reason for building T number of *PDM*s is that the back-propagation(BP) training algorithm provides "good" estimates of Bayes *a posteriori* probability functions only if the MLP has enough flexibility to closely approximate the Bayes functions and if there is sufficient training data. The sufficient amount of training data is provided by using several *PDM*s for every speaker. Several *PDM*s means above all several SOFMs. Each *PDM* is obtained after passing of the input feature vectors of a speaker's utterance through a SOFM. Each of these T number of SOFMs is trained using a corresponding feature vectors subset. Included in a subset are input vectors for a part of each speaker's training utterances of all the speakers. These subsets are nonoverlapping.

B. The second level of the classifier consists of MLPs. For each reference speaker a MLP is trained using the above mentioned *PDM*s. The MLPs are trained under supervision using the BP algorithm, which minimizes the squared error between the actual outputs of the network and the desired outputs.

Training: At the first classifier level during this stage T number of SOFM are trained and on their base for each speaker's utterance his T number of *PDM*s is determined. These *PDM*s are input feature vectors for the second level. When the MLP net of a given speaker is trained, his *PDM*s are labeled as "one" and *PDM*s of the remaining speakers as "zero".

Speaker recognition: At the first level during this stage for the unknown speaker U his T number of PDM^us are obtained. These PDM^us (the test vectors for an unknown speaker) are applied to each MLP. The outputs of the MLP of every speaker are accumulated. The unknown speaker is classified as the speaker whose MLP is with the maximum accumulated output.

Architecture of the used MLP networks is with two hidden layers and one output and the nodes have sigmoid nonlinearities.

4. Experimental Research

In order to make a comparison between the implemented methods in the system and their application to different speech conditions, experiments with clean and noisy speech data have been carried out.

4.1. The speech data base and recording conditions

Speakers: The speech (clean and over telephone lines) of 92 speakers (48 males and 44 females) has been analyzed. The speakers' age was within the range 19–67.

Speech material: The training set consists of 3 sentences: My name is *first name, second name, family name*; My code is *six digits*; I am a(n) *profession*. The test sets consist of other 3 sentences: I am *two digits* years old; My mother's name is first name; My hobby is *up to three words*.

Recording conditions: These sentences were uttered once (in Bulgarian) by each speaker in six separate sessions in a silent room. The same sentences were uttered once in other six sessions over telephone line (two local lines were used). The noisy and clean signals were digitized with sampling rate of 21 KHz and 16 bits per sample using the standard sound (audio) board type "Sound Blaster" for PC.

a) clean speech — the signals were quantized by means of the "Sound Blaster" board with its standard electret microphone and saved directly into the computer's memory in order to avoid any distortions caused by the tape recorders. The description of this board and its microphone is given in the technical reference provided with the board by Creative Inc.

b) phone speech — the signals were firstly recorded by means of a "SONY" tape recorder. Then they were quantized using the "Sound Blaster" board by its line input.

PC architecture: IBM/PC 386, Math co-processor 80387, 16 MB RAM, 800 MB HD, MS DOS 6.20, Sound Card "Sound Blaster" of Creative Inc. with its standard electret microphone.

4.2. Practical implementation of the methods

For the *PDM* method we used the optimal values (values which give the best recognition results for the described data base) for the filter coefficient (k) and for the size (Q) the *PDM*s

Identification method	*MLP*	*GMAHSM*	*ARVM*	*Two-level classifier*
error rate [%] (abs. number of errors)	1.81% (5)	1.45% (4)	1.81% (5)	1.09% (3)

Table 1. Identification results from the experimented methods for clean signals.

found in the experimental researches (Hadjitodorov, 1994, Hadjitodorov, 1997a): $k = 0.1$ and $Q = 20$.

For the ARVM method the order of AR-Vector Models was 2, because it is shown that for orders greater than 2, the prediction error doesn't decrease significantly (Montacie, 1992; Montacie, 1993; Le Floch,1994).

For the 2LC we used the optimal values for the number of $PDMs(T)$, k and for Q found in the experimental researches(Hadjitodorov, 1997a, Hadjitodorov, 1997b): $k = 0.1$, $T = 10$ and $Q = 4$. The architecture of the MLP for each speaker at the second level was: first hidden layer — 30 neurons, second — 2 and output layer — 1 neuron.

4.3. Results and discussion

The results with clean data are shown in Table 1 and for noisy in Table 2, where the absolute number of errors is given in brackets. As it was mentioned above, the test sets consist of three sentences pronunciations by the 92 speakers (clean and over telephone lines). From these results we can conclude that the two-level classifier performs better than other methods especially when the speech material is noise corrupted. The tests with the "paired t-test" show that:

1. The error rate reduction for noisy (phone) speech is statistically significant.

2. The error rate reduction for clean speech is not statistically significant.

The results show that the system could be very useful for robust speaker recognition, because never all the methods fail together and at least one of the described methods could be applied to given requirements and conditions. Another possibility is to combine the decisions of the four methods by a majority voting procedure. If the final classification is made when the majority of the methods (= >2) takes the same decision, then the accuracy is slightly improved — for clean signals - 0.73%, and for noisy signals- 1.45%. That means that the combination of the decisions could further enhance the classification accuracy. Finally, efficient use of the system for impostor detection is under investigation. Preliminary results are very encouraging.

References

1] B. ATAL, Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *J. Acoust. Soc. Amer.* **55** (1974), 1304–1312.

[2] K. ASSALEH, R. MAMMONE, Robust cepstral features for speaker identification, Presented at the *Proc. of the IEEE Int. Conf. on ASSP*, (1994), San Francisco, USA, 129–132.

[3] Y. BENNANI AND P. GALLINARI, On the use of TDNN-extended features information in talker identification, Presented at the *Proc. of the IEEE Int. Conf. on ASSP*, (1991), San Diego, USA, 385–388.

[4] F. BIMBOT, L. MATHAN, DE LIMA, AND G. CHOLLET, Standard and Target Driven AR-Vector Models for Speech Analysis and Speaker Recognition, Presented at the *Proc. of the IEEE Int. Conf. on ASSP*, (1992), San Francisco, USA, 5–8.

[5] F. BIMBOT, I. MAGRIN, L. MATHAN, Second-order statistics for text-independent speaker identification, *Speech Communication*, **17** (1995), 177–192.

[6] F. BIMBOT, L. MATHAN, Text-Free Speaker Recognition Using an Arithmetic-Harmonic Sphericity Measure, Presented at the *Proc. of the EUROSPEECH'93*, (1993), pp. 169–172.

[7] B. BOYANOV, S. HADJITODOROV, T. IVANOV, G. CHOLLET, Robust Hybrid Pitch Detector, *Electronic Letters*, **29** (1993), 1924–1926.

Identification method	*MLP*	*GMAHSM*	*ARVM*	*Two-level classifier*
error rate [%] (abs. number of errors)	6.15% (17)	5.43% (15)	5.79% (16)	2.17% (6)

Table 2. Identification results from the experimented methods for noisy (phone) signals.

[8] B. BOYANOV, S. HADJITODOROV, G. CHOLLET, "Robust periodicity/aperiodicity detector", *Ann. of Bulgarian Academy of Sciences*, **50** (1997), 43–46.

[9] P. CASTELLANO, S. SRIDHARAN, A Two Stage Fuzzy Decision Classifier for Speaker Identification, *Speech Communication* **18** (1996), pp. 139–149.

[10] K. FARRELL, R. MAMMONE AND K. ASSALEH, Speaker recognition using neural networks and conventional classifiers, *IEEE Trans. on Speech and Audio Processing*, **2**, (1994) 194–205.

[11] S. FURUI, Cepstral analysis technique for automatic speaker verification, *IEEE Trans. on Acoust., Speech and Signal Processing* **29**, (1981) 254–272.

[12] S. HADJITODOROV, B. BOYANOV, T. IVANOV, N. DALAKCHIEVA, Text-independent speaker identification using neural nets and AR-models, *Electronics Letters*, **30** (1994), 838–840.

[13] S. HADJITODOROV, B. BOYANOV, SPEAKER RECOGNITION, Research report for the National Scientific Fund, Bulgaria, (1997a),(in Bulgarian).

[14] S. HADJITODOROV, B. BOYANOV, N. DALAKCHIEVA, A two-level classifier for text-independent speaker identification, *Speech Communication*, **21** (1997b), pp. 209–217.

[15] H. HOLLIEN, *The acoustics of crime*, Plemun Press, New York, 1990.

[16] T. KOHONEN, "*Self-Organization and Associative Memory*", (1984), Berlin, Springer Verlag.

[17] T. KOHONEN, "The Self-Organizing Map", *Proc. IEEE*, Vol. 78, No 9,(1990), pp. 1464–1480.

[18] R. LIPPMANN, Pattern classification using neural networks, *IEEE Communications Magazine*, **5** (1989), 47–64.

[19] T. MATSUI AND S. FURUI, Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs, Presented at the *Proc. of the IEEE Int. Conf. on ASSP*, (1992), San Francisco, USA, 157–160

[20] C. MONTACIE AND J. FLOCH, AR-Vector Models for Free Text Speaker Recognition, Presented at the *Proc. Int. Conf. on Spoken Language Processing*, (1992), Banff, Alberta, Canada, 475–478.

[21] D. P. MORGAN AND C. L. SCOFIELD, *Neural Networks and Speech Processing, Kluwer Academic Publishers*, Norwell, MA, 1991.

[22] K. MURTHY, MURTHY, B. YEGNANARAYANA, Formant extraction from phase using weighted group delay function, *Electronics Letters*, **25** (1989), 1609–1611.

[23] J. NAIK, D. LUBENSKY, A Hybrid HMM-MLP Speaker Verification Algorithm for Telephone Speech, Presented at the *Proc. ICASSP*, (1994), pp. I-153 — I-156.

[24] M. NASHI, Phase unwrapping of digital signals, *IEEE Trans. on Acoust., Speech and Signal Processing*, **37** (1989), 1693–1702.

[25] L. RABINER, R. SCHAFFER, *Digital processing of speech signals*, Prentice Hall, New Jersey, 1978.

[26] D. REYNOLDS, Speaker identification and verification using Gaussian mixture speaker models, *Speech Communication*, **17** (1995), 91–108.

Received: June, 1997
Revised: November, 1997
Accepted: June, 1998

Contact address:
Stefan Hadjitodorov and Boyan Boyanov
Central Laboratory of Biomedical Engineering
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Block 105
1113 Sofia
Bulgaria
e-mail: sthadj@argo.bas.bg

S. HADJITODOROV graduated in 1978 from Technical University (Sofia, Bulgaria). He received a Ph.D. degree in 1983. He was an engineer at Technical University from 1978 to 1983. From 1983 to 1989 he was a Research Fellow and since 1989 he has been a Senior Research Fellow in the Center on Biomedical Engineering (CBME), Bulgarian Academy of Sciences. He is the head of Department of Biomedical Informatics in CBME. Dr. Hadjitodorov is a member of the International Association for Cybernetics, Bulgarian Society of Pattern Recognition, Balkan Federation for Fuzzy Systems and the IEEE Engineering in Medecine and Biology Society. He is author of many publications in international journals. His research interests are in the fields of data analysis, pattern recognition and classification, neural networks, statistical and fuzzy methods, voice analysis.

B. BOYANOV graduated in 1982 from Technical University (Sofia, Bulgaria). He received a Ph.D. degree in 1985. In 1984 he started as a Research Fellow and since 1995 he has been a Senior Research Fellow in the Center on Biomedical Engineering (CBME), Bulgarian Academy of Sciences. Dr. Boyanov is head of the Dept. SPEECH in CBME. Dr. Boyanov is a member of the Bulgarian Societies of Pattern Recognition and of Biomedical Engineering. He is author and co-author of many publications in international journals. His research interests are in the fields of signal processing, speech analysis, laryngeal pathology detection, neural networks, singer voice analysis, application of the wavelet transform and higher-order spectra for speech analysis.