Computing and Informatics, Vol. 25, 2006, 369–391

SPEAKER RECOGNITION IN THE BIOMETRIC SECURITY SYSTEMS

Filip Orság

Faculty of Information Technology
Institut of Intelligent Systems
Božetěchova 2
612 66 Brno, Czech Republic
e-mail: orsag@fit.vutbr.cz

Manuscript received 19 January 2005
Communicated by Milan Rusko

Abstract. At present, the importance of the biometric security increases a lot in cotext of the events in the world. Development of the individual biometric technologies such as the fingerprint recognition, iris or retina recognition or speaker recognition has been considered very important. However, it comes to be true that only one biometric technology is not sufficient enough. One of the most prosperous solutions might be a combination of more such technologies. This article aims at the technology of the speaker recognition and proposes a solution of its integration into a more complex biometric security system. Herein a design of the complex biometric security system is introduced based on the speaker recognition and the fingerprint authentication. A method of acquisition of a unique vector from speaker specific features is introduced as well.

Keywords: Biometry, security systems, speaker, recognition

1 DESIGN OF THE BIOMETRIC SECURITY SYSTEMS

It has been quite a long time since the first *Biometric Security Systems* (BSS) were introduced. However, until now, they have not become widely used. This is usual in case of a new, not well-tested, and unverified technology, among which the biometric systems shall be included. Scepticism, of course, has its place in our lives and is necessary to improve our work. Such scepticism is doubled in case of the security

systems. Imperfections in the security systems are not tolerated and this is right. A precise research is necessary because of the increased sensitivity to the quality of the security systems. This is a reason for absence of the BSS in the real-world applications.

1.1 Task of the BSS

Biometric security system is a security system, whose protection is based on the biometric features. The BSS can be applied in many ways and situations. Usually, we want to protect a device or a service. Typically, we want to enable the authorized users to access to some network resources such as a printer, a file server, or a database, as a counterpart, we want to disable the unauthorized users to access them. This task can be called biometric login. Another typical task can be a protection of objects such as buildings or other facilities, which can be called biometric doorkeeper. These applications require a trustworthy and reliable protection method, since in these cases the price of the protected object is higher than in case of the network resources.

The first thing to consider prior to installing a BSS is its price. How valuable is the protected object and how strong should be the protection? There is a dependency between the price of the protected object and the strength of the protection. Nowadays, the doorkeeper task is performed by a key or a chip card and the login task is accomplished by a login name with an appropriate password. These solutions are very cheap, but the protection level is not that high. This holds true, since the key could be easily lost and the login name with the password could be stolen (because of lack of the user's caution). In case the protected object is not too valuable, these solutions could fulfil this task. However, i.e. in case of a bank vault we need very strong protection.

We can ask another question. How can we increase the strength of the protection? There are many possibilities and one of them is the biometric security system. The BSSs are not widely used, because people do not trust them yet. The security systems based on the biometry are relatively new and people usually do not trust a new technology. Since present, only the one-level or single-biometric security systems have been used [16, 17, 18]. The multi-biometric systems are rare, but there are some commercial solutions based on multiple biometric features [19] and some books on this topic [7].

How much money do we want to spend to protect the object then? Answer to this question depends directly on the price of the protected object. The more expensive the object is, the more money we are willing to spend on the security system. From the technical (hardware) point of view, the BSS are not very expensive! This should be said, since to check a speaker we need only a microphone, which does not cost much, and a computer. The price of a fingerprint scanner is not high as well. The most expensive portion of the BSS is the software. The research and the development of the biometric systems account for most of the costs. Testing and validating of the proposed algorithms is not free either.

Thus, having these questions in mind we can decide whether to apply a BSS or not.

1.2 Advantages and Disadvantages of the BSS

When talking of the BSSs, we shall mention their advantages and disadvantages. The biometric features are relatively universal. Most people have their fingers, they can speak, and their cells contain the DNA. Of course, there are some exceptions. Dumb people cannot speak. People, who lost their finger or arm, cannot use fingerprint-based systems etc. However, all of us have got the DNA for example.

The biometric features are extracted from various parts of the human body (DNA, fingerprint, and others). The biometric security systems are supposed to be reliable, because the biometric features used in the authentication process of the biometric security systems are unique. It is not valid in every case, but it holds mostly true.

The biometric security systems are very safe. It is not possible to deceive the protected devices and services thanks to their biometric protection. Most biometric features are possible to acquire by the authorised person only, so that his/her presence at the point, where the authorisation device is placed, is necessary. However, it could be possible to fool the security system, i.e. you can cut off a finger and try to persuade the system to accept it. Still, the security system designers are able to prevent the system from accepting such a sham. Nowadays it is possible to use a fingerprint scanner able to measure blood flow in veins in the scanned finger. This scanner proves aliveness of the human, who is being identified.

The individual biometrical technologies can be stacked and can be built in a multilevel authentication system. Such multilevel system includes a standard login, a voice login, a fingerprint login, and some others. Stacking of the separate technologies in one complex unit increases the security of the whole system. If one level of the multilevel system is broken or cheated, the others have to decrease a break-through possibility. The break-through possibility decreases with the count and strength of each of the login levels. However, we must be careful when increasing the count of the login levels, because a large number of them could result in worse overall security than a set of two strong, well-designed login levels.

Some biometric features are permanent and some are not. Human voice is not constant during the whole life. This effect is obvious mostly in the changes of teenagers' voices. The influence of illnesses or psychological condition should not be neglected. This can be the greatest difficulty. Human voice is not the only one unstable feature, even many other biometric features change during human' life. Counterpart to the inconstancy of some biometric features is their permanency. An example of the permanent feature set is a DNA-based feature set. DNA is permanent and does not change. The same can be said of the fingerprints. Another relatively permanent biometric feature is the retina image. Other biometric features are not permanent – they are short-term (relatively to the length of a human's life).

Some of the biometric features can be very exacting to acquire. Among them could be included e.g. the DNA. Though the DNA is unique for all of us (apart from the monovular twins), it is very difficult to extract and to analyse it quickly enough. Besides, the price of the analysis is not low. This excludes the DNA from the real time applications – it is not worth the advantages it can bring these days.

Not to be forgotten is the cooperation unwillingness. Some humans are not happy with acquiring their biometrical features. Results of the recent public inquiry show that most people dislike scanning their retinas. Many of them dislike scanning their fingerprints and faces and the least of them dislike voice recording. It is clear then that it would be very useful to develop a reliable technology based upon speech processing. Most of us are ready to let the machine analyse our voices rather than anything else.

Mass use of the biometric security systems is not breaking by difficult implementation, cooperation unwillingness, an inconstancy, or because they are too demanding. Another obstacle is inability of some people to pass an enrolment. Still, this could be solved exactly by a multi-biometric security system. In such system, one authentication level can be skipped for some people to enable them to use biometrically protected devices or services.

1.3 Single-Biometric Security System (SBBS)

Usually, only a one-level biometric security system [3, 4, 5, 6] is applied to provide the security services and protection. The one-level biometric system will be called Single-Biometric Security System (SBBS). It consists typically of two hardware components – an input device and a processing unit.

The input devices include: scanners (a fingerprint scanner, a palm scanner, a retina scanner, or an iris scanner), microphone, special sensors (an odour sensor or a thermal sensor), and many other devices. These devices serve the acquisition of the physical biometric features. The physical biometric features are a fingerprint scan, thermal scan, or speech signal. Data acquired from the input devices are sent to the processing unit.

The processing unit can be a built-in device or an external device, which is responsible for the further processing of the input data and for the final decision. The input data are obtained from the input devices and the final decision is usually an answer to the question: "Is the unknown individual really the one, who he/she is claiming to be?" The processing unit can be, for example, a computer with an appropriate application or a built-in processor of a smart card. Both the external and built-in solutions have their advantages and disadvantages. The external solution is cheaper but it provides weaker protection, because the communication channel between the input device and the processing unit could be wiretapped and the data could be misused. The built-in solution is more expensive but the protection is better. The decision between these two solutions depends on the requirements of the real application. In Figure 1 schematic comparison of the external and internal solutions is shown. The external solution can be hazardous for the data. However,

badly designed internal solution could be as hazardous for the data as the external solution.

Though the protection strength of the external solution is not high, this solution used more frequently than the internal one. The weakness of the external solution can be solved in several ways. One possible solution is to encipher the information sent through the communication channel using a symmetric key generated from the biometric features. Another difficulty could be the storage question, i.e. in the asymmetric cryptography, you get a private key, which must be stored somewhere (usually in a storage device). If the storage device were stolen, the private key would be misused. This can be solved as the first problem – by enciphering the private information using a symmetric biometric key.

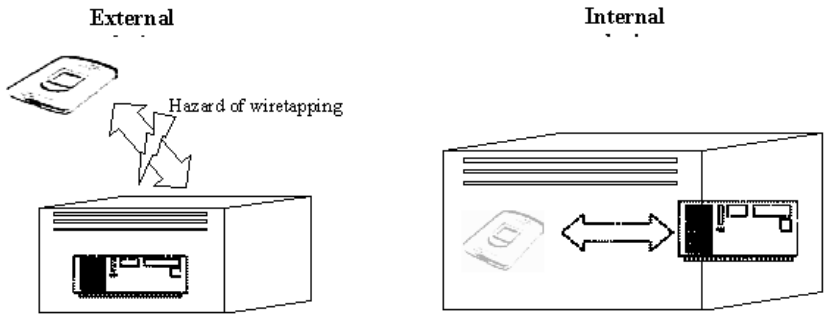

Fig. 1. Schematic diagram of the external solution and the internal solution of the SBSS

Thus, the main disadvantage of the SBSS solution is its solitariness. If the recognition process failed in terms of false acceptance, access to the protected object would be achieved by an unauthorised user. Sometimes this is not so crucial. Usually, it is a question of the priorities and even the one-level system can be set up so that it is able to reject all impostors. Price for this is higher false rejection rate.

To design a balanced system, in which the FAR and FRR rates were very low, is the goal of BSS designers. Solution to this could be a multi-biometric security system.

1.4 Multi-Biometric Security System (MBSS)

Multi-Biometric Security System (MBSS) is a biometric system based on a combination of more than one biometric technologies [3, 4, 5]. The MBSS is counterpart to the one-level system. Its main advantage over the SBSS is its complexity that makes the system more robust to the FAR, before all. The security system administrators

usually design the BSS so that the FAR is as low as possible even for the price of the higher FRR. This limitation can be solved using an MBSS.

Generally, authorisation using multiple biometrics is reduced to a fusion problem (Figure 2), which utilises results of multiple biometric technologies to increase the fault-tolerance capability, to reduce uncertainty, to reduce noise, and to overcome the limitations of the SBBS. Well-designed MBSS can increase the reliability of the final decision. Multiple biometrics used in the MBSS enable some user to be identified even if they are not able to provide all biometric features. However, when permitting exceptions, details of the exceptions must be specified when designing a MBSS. It is not possible to make general exceptions, since this would cancel the advantages of the MBSS.

In Figure 2 you can see a schema of the integration of the multiple biometric features into one complex. The most important part is the block of the decision fusion and the partial decision blocks (here for the fingerprint recognition, the speaker verification, and the iris recognition).

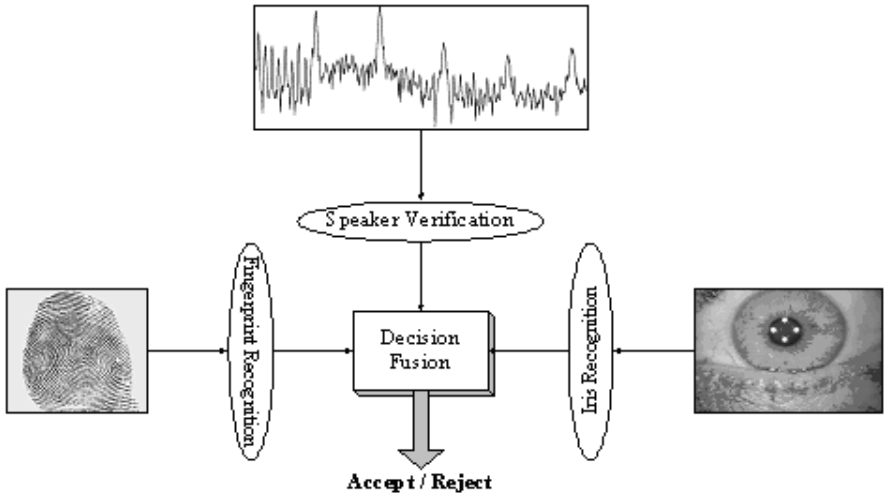

Fig. 2. Integration of multiple biometrics

As each of the individual biometric subsystems in a MBSS has very different characteristics and pattern matching scheme, it is useful to integrate them at the decision level.

Let $\Phi = \{\phi^1, \phi^2, \ldots, \phi^I\}$ denote a set of templates ϕ^i representing each of I users of the MBSS. Each user has exactly one multi-biometric template, which consists of one biometric template for each biometric feature used in the system. Template $\phi^i = \{\phi^i_1, \ldots, \phi^i_M\}$ of the i-th user consists of biometric templates, hence the MBSS consists of M various biometric technologies ($M = 3$ in case of the situation in Figure 2). There are two classes of users denoted ω_{true} (an authorised user, whose

template ϕ is one from the set Φ) and ω_{false} (an unauthorised user). Given a classifier $C(\phi_1, \phi_2)$ we can say

$$C(\phi^i, \phi^i) = \omega_{true}, i = 1, 2, \ldots, I. \tag{1}$$

When verifying an unknown user, whose template is denoted $\overline{\phi}$ and who claims to be one of the authorised users $\phi^i, i \in \{1, 2, \ldots, I\}$, we evaluate distance (or decision) $D(\overline{\phi}, \phi^i)$ as a function of the templates $\overline{\phi}$ and ϕ^i. Given a threshold T, the unknown user is classified as

$$C(\overline{\phi}, \phi^i) = \begin{cases} \omega_{true}, & D(\overline{\phi}, \phi^i) \geq T \\ \omega_{false}, & otherwise \end{cases}, \ 0 \leq T \leq 1 \tag{2}$$

where $D(\phi^a, \phi^b)$ is the distance between the template ϕ^a and ϕ^b satisfying

$$0 \leq D(\phi^a, \phi^b) \leq 1 \tag{3}$$

and defined as

$$D(\phi^a, \phi^b) = F\left(d_1(\phi_1^a, \phi_1^b), \ldots, d_M(\phi_M^a, \phi_M^b)\right) \tag{4}$$

where $F\left(d_1(\phi_1^a, \phi_1^b), \ldots, d_M(\phi_M^a, \phi_M^b)\right)$ is a decision fusion of all M partial decisions $d_j(\phi_j^a, \phi_j^b)$. A partial decision can be i.e. a result of the speaker verification of or the fingerprint recognition. The overall distance (decision) $D(\phi^a, \phi^b)$ and the partial distances (decisions) $d_j(\phi_j^a, \phi_j^b)$ can be a probability or a measure of similarity of the two templates.

The overall decision $D(\phi^a, \phi^b)$ is based on the integration of the decisions made by the individual biometric modules. The value of the overall decision can be based on the theory of probability. Based on a Bayes' decision rule, we can determine the overall decision as

$$F\left(d_1(\phi_1^a, \phi_1^b), \ldots, d_M(\phi_M^a, \phi_M^b)\right) = \frac{p\left(d_1(\phi_1^a, \phi_1^b), \ldots, d_M(\phi_M^a, \phi_M^b)\right) | \omega_{true}}{p\left(d_1(\phi_1^a, \phi_1^b), \ldots, d_M(\phi_M^a, \phi_M^b)\right) | \omega_{false}} \tag{5}$$

when knowing the class-conditional probability density functions $p\left(d_1(\phi_1^a, \phi_1^b), \ldots, d_M(\phi_M^a, \phi_M^b)\right) | \omega_{true}$ and $p\left(d_1(\phi_1^a, \phi_1^b), \ldots, d_M(\phi_M^a, \phi_M^b)\right) | \omega_{false}$. The value of the threshold T as needed for the final classification using the Equation (2) can be expressed as

$$T = \frac{P(\omega_{false})}{P(\omega_{true})}. \tag{6}$$

The probabilities $P(\omega_{true})$ and $P(\omega_{false})$ are the prior probabilities reflecting the prior values of the probabilities corresponding to the classes ω_{true} and ω_{false}.

This is one possible approach to the fusion decision making. Other possibilities can be the fusion mechanism based on a neural network classifier or a fuzzy classifier. Then, the overall decision $D(\phi^a, \phi^b)$ is replaced by the chosen classifier. There must be one classifier for each user. Hence, not only the template ϕ_i need to be stored but also the congruent network.

2 UNIQUE VECTOR GENERATING

Biometric features can be used in cryptography to generate a symmetric key [3]. The key is given as a combination of vectors acquired using the individual biometric technologies. Each of the biometric technologies should be able to provide at least one unique vector, which will be then considered as a part of the biometric cryptographic key. To be able to generate the unique vector we have to design an algorithm producing unique and re-estimable vector when provided with a biometric feature set.

If the vector were not unique, there would be the possibility of false acceptance of the biometric features of some other (possibly unauthorised) user. This one might be granted access to some private resources and information, which is not desired. Speech technology can provide only one unique vector that is very difficult to re-gain. The fingerprint technology can produce many possible keys, but these keys are not the same each time a fingerprint is scanned and analysed. This is because some fingerprint minutiae need not to be recognised and some new ones could be recognised. However, fingerprint offers a better set of features than speech.When using more than one biometric technology, we can supply ambiguity of the vectors between two different users given by one technology by the other biometric technology. Both of them can supply each other.

Having a symmetric key produced using the biometric features, we can encrypt all private information using this key and store it securely anywhere. We can even generate a private key for an asymmetric ciphering, encipher it using the symmetric biometric key, and store it safely in a public storage device. An example of the encryption process using the fingerprint and speech technologies for the symmetric ciphering is illustrated in Figure 3.

We can use such key for the authentication purposes as well. When a user is about to login, he/she claims his/her identity and then must provide the biometric features needed to be authenticated. The possibility of losing private data decreases with growing number of the biometric technologies used for this purpose. However, when increasing the number of technologies, the possibility of errors grows as well. Again, an optimal solution must be chosen. Sometimes it can be impossible to reconstruct the key, which can be caused by the inconstancy.

The process of unique vector generating from the speech signal is rather challenging. As the speech features are so unstable, it is not easy to fulfil the preconditions given – the uniqueness and the re-estimableness. The speaker's vocal tract characteristics change during his/her whole life, which aggravates the process, since the system must be held up to date and the parameters of the estimator must be properly updated.

2.1 Signal Processing Mathematical Background

Most of the signal processing math used for the unique vector generating is well known and can be found i.e. in [9, 11, 12, 13]. For the purposes of the unique

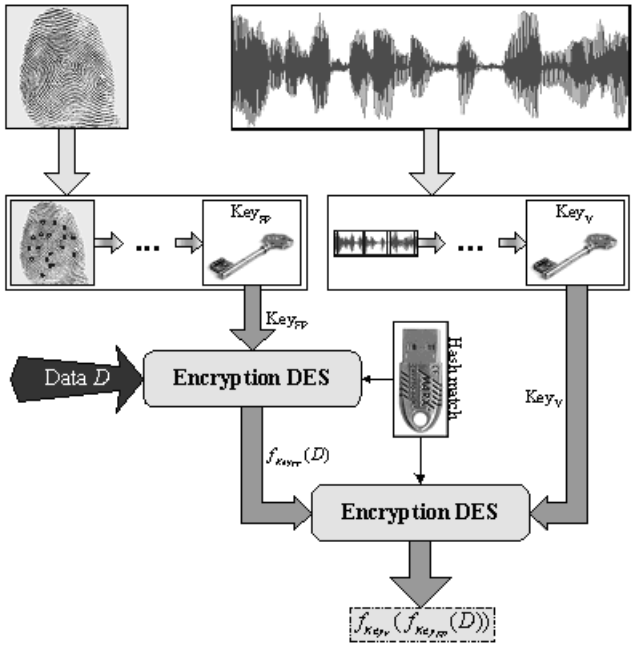

Fig. 3. Encryption using the biometric keys

vector generating will be used autocorellation coefficients and Linear Prediction Coefficients (LPC) in this case. Autocorrelation and the LPC coefficients are very useful for various tasks in speech processing. It can be used for speech recognition as well as for speaker recognitinon.

LPC coefficients can be used to calculate an LPC frequency spectrum. The main difference from the common frequency spectrum is its smoothness. In the smoothed spectrum, there are obvious resonance frequencies of the formants. These frequencies are very useful both in speech recongition and speaker verification. The LPC spectrum is defined as

$$S_{LPC}(f) = \left| 1 - \sum_{m=1}^{M} a(m) \cdot z^{-m} \right|^{-2} \tag{7}$$

where M is number of the LPC coefficients (prediction order), $a(m)$ are the LPC coefficients themselves and f is frequency. Now we can substitute $z = e^{2\pi j \frac{f}{F_S}}$ which results in

$$S_{LPC}(f) = \left| 1 - a(1) \cdot e^{-2\pi j \frac{f}{F_S}} - a(2) \cdot e^{-4\pi j \frac{f}{F_S}} - \ldots - a(M) \cdot e^{-2M\pi j \frac{f}{F_S}} \right| \tag{8}$$

where F_S denotes the sampling frequency. As we work with the finite discrete signal, it is useful to transform the Equation (8) to the following form:

$$S_{LPC}(k) = \left| 1 - \sum_{m=1}^{M} a(m) \cdot e^{-2m\pi j \frac{k}{N}} \right|^{-2} \tag{9}$$

where N is length of the signal $s(n)$.

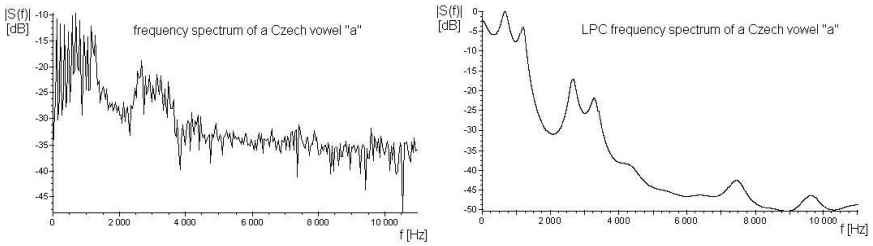

Fig. 4. Frequency spectrum (left) and LPC frequency spectrum (right) of a vowel /ah/ as pronounced in the word "but"

Comparison of a frequency spectrum and an LPC frequency spectrum is shown in Figure 4. You can obviously see the difference – on the left side you can see the frequency spectrum as obtained by an FFT, and on the right side there is the LPC frequency spectrum. In the LPC spectrum the main formants are clear, which is very useful.

Long-term statistics of many various features are used often to recognise speech [10, 14]. However, not only the area of the speech recognition is a domain of the long-term statistics. The average long-term LPC spectrum is applicable to the speaker verification as well.

The long-term LPC spectrum is estimated using the average autocorrelation coefficients [10]. These are estimated over all frames of the given signal $s(n)$. Assume the signal $s(n)$ divided into total count of J frames N samples long. Then, the average autocorrelation coefficients are defined as

$$\overline{R}(k) = \frac{1}{J} \cdot \sum_{j=1}^{J} R(j,k) \tag{10}$$

where $R(j,k)$ is the autocorrelation of the j^{th} frame, the index j should be $1 \leq j \leq J$. The average autocorrelation coefficients $\overline{R}(k)$ can be used to derive the average LPC coefficients $\overline{a}(i)$ from using the Durbin recursive procedure. In the redefined equation the autocorrelation coefficients $R(k)$ are substituted by the average autocorrelation coefficients $\overline{R}(k)$, the result of which are the average LPC coefficients $\overline{a}(i)$. The average LPC coefficients $\overline{a}(i)$ can be used then to estimate the long-term LPC spectrum derived from Equation (8). The LPC coefficients in this equation are replaced by the average LPC coefficients, and thus

$$S_{LPC}(f) = \left|1 - \overline{a}(1)\cdot e^{-2\pi j\frac{f}{F_S}} - \overline{a}(2)\cdot e^{-4\pi j\frac{f}{F_S}} - \cdots - \overline{a}(M)\cdot e^{-2M\pi j\frac{f}{F_S}}\right| \qquad (11)$$

In Figure 5, there is a sample of a long-term LPC spectrum derived from the average LPC coefficients of the orders $M \in \{4, 8, 22\}$. The original signal is sampled at the sampling frequency $F_S = 16\,000$ Hz, length of the signal is $N = 512$ samples and the LPC spectra belong to one speaker only.

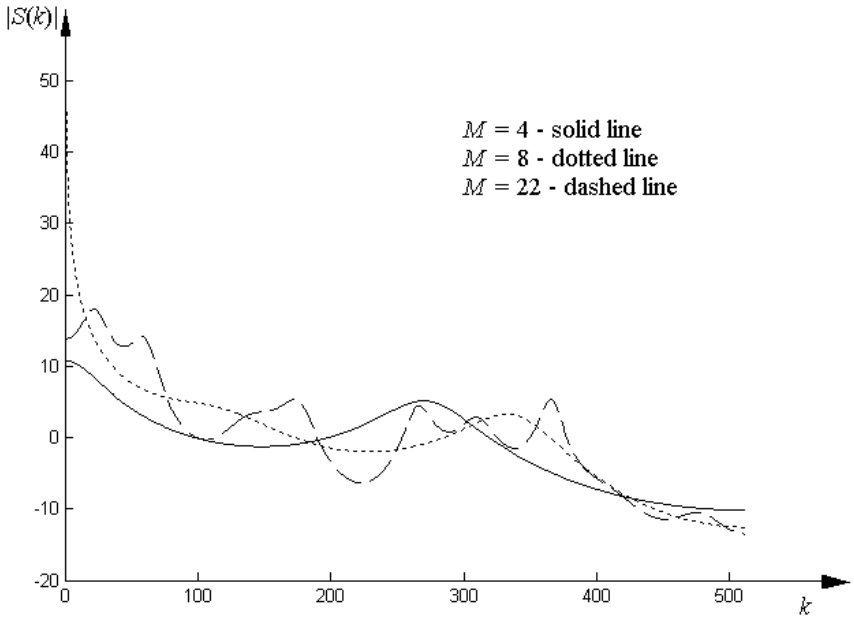

Fig. 5. Sample of a long-term LPC spectrum derived from the average LPC coefficients of the orders of $M \in \{4, 8, 22\}$. The original signal is sampled at sampling frequency $F_S = 16\,000$ Hz, and length of the signal is $N = 512$ samples.

Another very useful tool is signal normalisation by a long-term spectrum. This is applicable primarily to speaker-independent speech recognition [15], but here it will be used for unique vector generating purposes. Given a framed signal, the autocorrelation coefficients $R(j, k)$ of the j^{th} frame and the order $k = 0, 1, \ldots, M$, the normalised autocorrelation coefficients are defined as with

$$R_{norm}(j, k) = R_a(0)\cdot R(j, 0) + \sum_{m=1}^{M} R_a(m)\cdot [R(j, |k - m|) + R(j, |k + m|)] \qquad (12)$$

with

$$R_a(k) = \sum_{i=0}^{M-k} \overline{a}(i) \cdot \overline{a}(i+k) \tag{13}$$

where $\overline{a}(i)$ are the average LPC coefficients. The effect of the signal normalisation is shown in Figure 6. You can see an original LPC spectrum (solid line) that is normalised by a long-term spectrum. The normalisation (dotted line) emphasises formant peaks. It is useful for speech recognition because it improves its results.

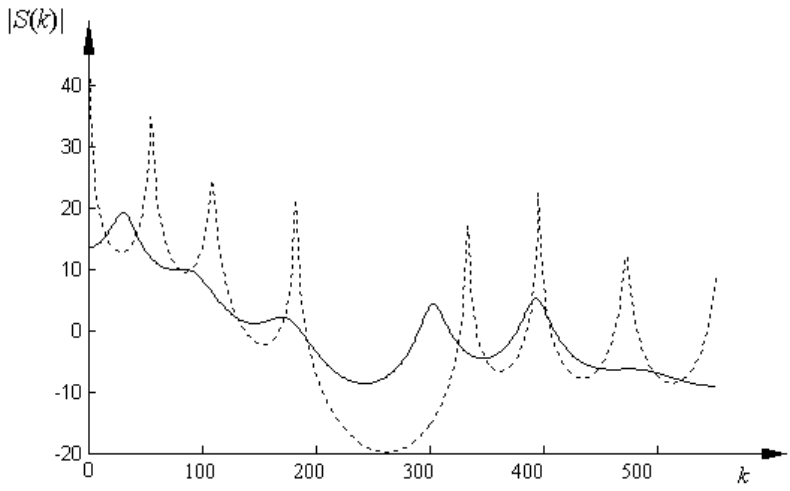

Fig. 6. Effect of the signal normalization by a long-term spectrum. The original LPC spectrum is drawn by a solid line, the normalized LPC spectrum by a dotted line.

2.2 Unique Vector Generating from the Speech Signal

To generate a unique vector, a special method must be designed, since the features usable to get a unique vector are not precisely re-estimable. A tolerance must be defined within which the features can vary. This process is very similar to sampling of a signal or to quantization. We define a step – a sampling period or a quantisation step – and then we sample the features along one dimension. The most difficult task in this process is the step estimation. One possible solution to this results from the statistical measures.

Consider a long-term LPC spectrum derived from the LPC coefficients of the 22$^{\text{nd}}$ order. There is a set of frequencies (positions of the maxima, see Figure 7)

$$F_{max}(l) = \{f_l^{max}\} = \{f_1^{max}, f_2^{max}, \ldots, f_L^{max}\}, l = 1, \ldots, L \tag{14}$$

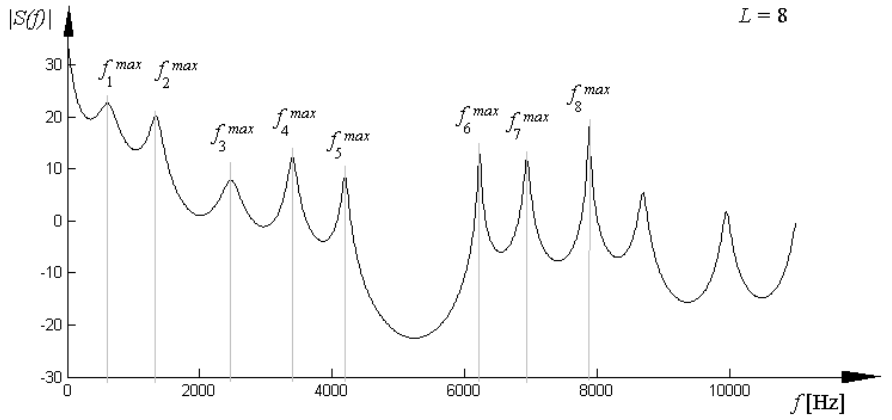

Fig. 7. Example of a long-term LPC spectrum with $L = 8$ emphasized and marked maxima

omitting just the maximum at the position $l = 0$. This set is the basis for the unique vector. We chose a group of $L = 8$ extremes (maxima). Eight maxima is the experimentally determined value. All tested speech signals contained at least eight maxima. Hence, we can have eight unique values. When we scale them so that they range from 0 to 255, we can use them to create an array of 16 bytes, i.e. there are eight numbers consisting two hex-digits, which means we can have a 128 bits long vector. How should be the maxima scaled, so that the same values are re-estimated when analysing a new sample of the same speaker? To define the quantisation steps q_l for each group $l = 1, 2, \ldots, L$ of maxima $f_{l,n}^{max}$, we need more than one training sample. Having N training samples we get N sets of maxima

$$F_{max}^{n}(l) = \left\{ f_{l,n}^{max} \right\} = \left\{ f_{1,n}^{max}, f_{2,n}^{max}, \ldots, f_{L,n}^{max} \right\}, l = 1, \ldots, L \wedge n = 1, \ldots, N. \quad (15)$$

Upon these sets, we can base estimation of the quantisation step. There is a desired range from 0 to 255 and there is a real range of the frequencies representing positions of maxima of the long-term LPC spectrum. Now, we can unite these ranges. The quantisation step q_1 of the 1st maximum equals the range of the 1st group of frequencies (see Figure 8 for illustration). Generally, the quantisation step of the l^{th} generated value is defined as

$$q_l = \max_{n=1,\ldots,N} (f_{l,n}^{max}) - \min_{n=1,\ldots,N} (f_{l,n}^{max}), l = 1, 2, \ldots, L. \quad (16)$$

Except from the quantisation step, a value of the initial shift must be defined. Prior to get the quantised value using the quantisation step q_l, we have to subtract the initial shift s_l to get proper results. The initial shift is defined as

$$s_l = \min_{n=1,\ldots,N} (f_{l,n}^{max}) - q_l \,\mathrm{int}\left(\frac{\min\limits_{n=1,\ldots,N} (f_{l,n}^{max})}{q_l} + 1 \right), l = 1, 2, \ldots, L \qquad (17)$$

where the function int(x) returns the integer part of the x.

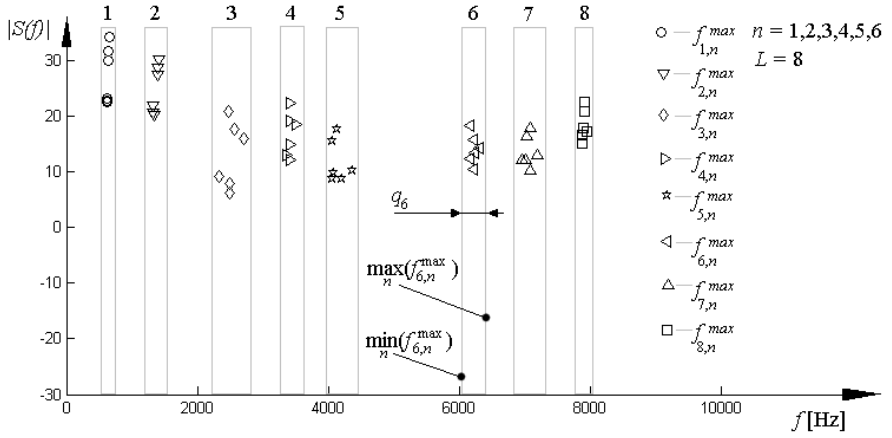

Fig. 8. Illustration of $L = 8$ groups of maxima extracted from a group of six speech samples using the long-term LPC spectrum. The quantisation step, and the maximal and minimal frequencies from the sixth group of maxima ($n = 6$) are marked

This is illustrated in Figure 8, where you can see $L = 8$ groups of maxima extracted from six speech samples using the normalised long-term LPC spectrum (like the one in Figure 7). You can see markings of the quantisation step q_6, $\max\limits_n \left(f_{6,n}^{max}\right)$, and $\min\limits_n \left(f_{6,n}^{max}\right)$ calculated from the sixth group of maxima extracted from six normalised long-term spectra.

Now, we can generate a unique vector. Given the quantisation step q_l and corresponding initial shift s_l, the quantized value $\widetilde{v}_l(x)$ is defined as

$$\widetilde{v}_l(x) = \mathrm{int}\left(\frac{x - s_l}{q_l} \right). \qquad (18)$$

It is clear that the quantisation step cannot include all frequencies, which the speaker is able to produce in the given band. Hence, we have to define percentage tolerance, which enlarges the accepted range slightly whereby the quantisation step increases as well. The more the training samples, the lower can be the tolerance and the final error. Thus, given a normalized tolerance factor t, the quantisation step with the tolerance is defined as

$$\widehat{q}_l = q_l + t{\cdot}q_l, l = 1, 2, \ldots, L \wedge 0{\leq}t{\leq}1 \qquad (19)$$

and, correspondingly, the initial shift must be redefined as

$$\widehat{s}_l = \min_{n=1,\ldots,N}(f_{l,n}^{max}) - \frac{t \cdot q_l}{2} - \widehat{q}_l \text{int}\left(\frac{\min\limits_{n=1,\ldots,N}(f_{l,n}^{max}) - \frac{t \cdot q_l}{2}}{q_l} + 1\right), l = 1, 2, \ldots, L \quad (20)$$

which gives us an equation for the quantised value $\widehat{v}_l(x)$ as

$$\widehat{v}_l(x) = \text{int}\left(\frac{x - \widehat{s}_l}{\widehat{q}_l}\right). \quad (21)$$

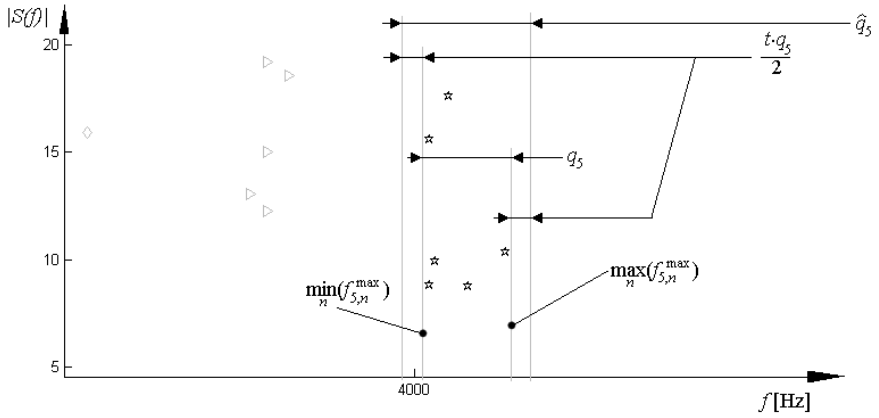

Fig. 9. Detail of the fifth group ($n = 5$) of maxima. The most important variables are marked.

In Figure 9, you can see detail of Figure 8. The fifth set of maxima is emphasized and the important variables are marked. The only value which cannot be marked is the initial shift, which is applied later when calculating the quantised values. The initial shift is illustrated in Figure 10, in which the third group of maxima is emphasized and the congruent initial shift and quantisation step are marked.

There is one more problem to solve. In case the quantised value exceeds the interval $\langle 0, 255 \rangle$, we have to correct it. The easiest way to do that is using the remainder of the division by 256, which results in the quantised value defined as

$$v_l(x) = \text{rem}\left(\widehat{v}_l(x), 256\right) \quad (22)$$

where $\text{rem}(x, y)$ returns the remainder of the division x/y. Having this, we can create a unique vector for each speaker in the voice database. The unique vector can be defined as

$$V = \left(v_1(\overline{f}_1^{max}), v_2(\overline{f}_2^{max}), \ldots, v_L(\overline{f}_L^{max})\right) \quad (23)$$

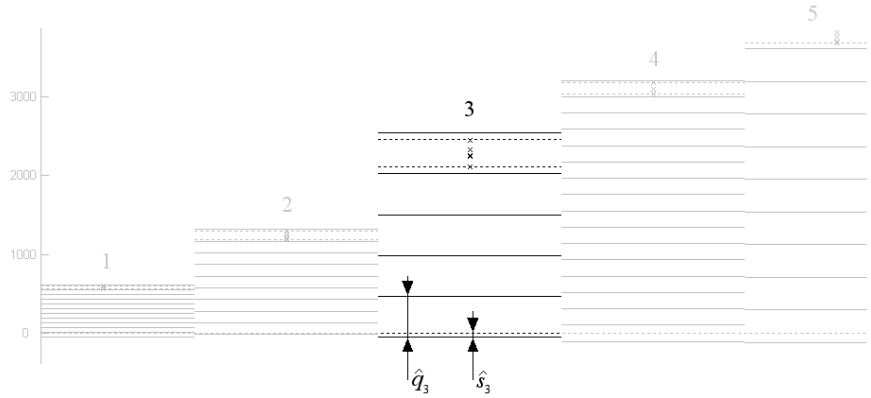

Fig. 10. Detail of the third group ($n = 3$) of maxima with the initial shift and the quantisation step marked

whereby

$$\overline{f}_1^{max} = \frac{1}{N} \sum_{n=1}^{N} f_{l,n}^{max} \tag{24}$$

and N is total number of the training samples (see the definitions above). Given the unique vector V it is possible to transform it to a hexadecimal string, which is defined as a concatenation of the individual components of the unique vector transformed to the hexadecimal form. It can be defined as

$$V_{hex} = \text{hex}(V_1)\text{hex}(V_2)\ldots\text{hex}(V_L) = v_{1,1}^{hex} v_{1,2}^{hex} v_{2,1}^{hex} v_{2,2}^{hex} \ldots v_{L,1}^{hex} v_{L,2}^{hex} \tag{25}$$

whereby the function $\text{hex}(x)$ returns a hexadecimal representation of the integer value x and $V_i = v_i\left(\overline{(}f)_i^{max}\right), i = 1, 2, \ldots, L$ are the components of the unique vector defined by Equation (23). The returned hexadecimal value given by the function $\text{hex}(x)$ ranges from 00 to FF, since the values of the unique vector components range from 0 to 255. This allows to write the unique vector as an array of $2L$ hexadecimal digits denoted in Equation (25) by $v_{l,i}^{hex}$, $l = 1, 2, \ldots, L$ and $i = 1, 2$. This helps define a measure $d\,(V^1, V^2)$, which expresses a distance between two unique vectors as

$$d\,(V^m, V^n) = \sum_{l=1}^{L} \sum_{i=1}^{2} \left| v_{l,i}^{m,hex} - v_{l,i}^{n,hex} \right|, m, n = 1, 2, \ldots, M \tag{26}$$

where M is total number of the users stored in the voice database. Another possible measure of the distance can be defined as

$$d'\,(V^m, V^n) = \sum_{l=1}^{L} |V_l^m - V_l^n|, m, n = 1, 2, \ldots, M \tag{27}$$

where $V_i = v_i\left(\overline{f}_i^{max}\right), i = 1, 2, \ldots, L$ are the components of the unique vector defined by Equation (23) above.

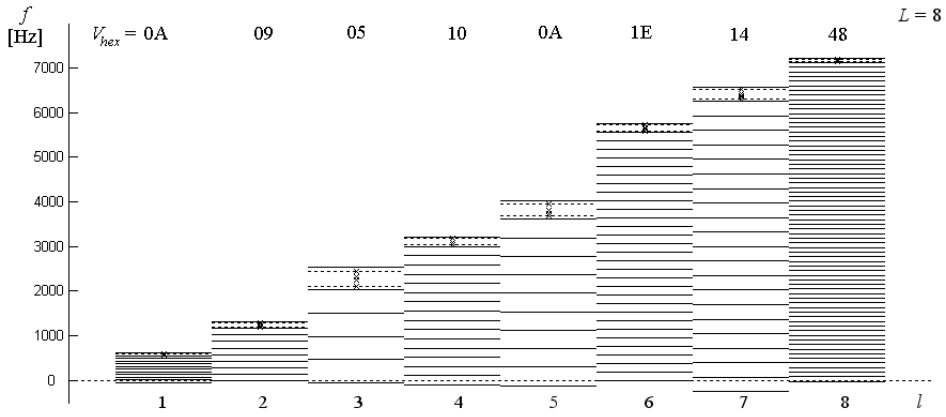

Fig. 11. Example of the scaling for all eight maxima with a congruent example of the unique vector given by Equation (25)

The experimental results of the creation of the unique vector are summarised in Section 3.2. If this vector proved unique, it would be used for the cryptographic applications even for the purposes of the identification and verification. Since this vector is usable for the purposes of the verification and identification, the measures of the FRR and FAR can be used for the performance measures.

3 EXPERIMENTAL RESULTS

Some experiments had to be made to prove validity of the algorithms and to test quality of the proposed features. The experiments were made on a voice database created at the university in cooperation with the students of the 1st term. In the following sections the results of the experiments are described.

3.1 Voice Database

A voice database is necessary to test all suggested features and algorithms. The voice database used for all the performed experiments was built in cooperation with a group of the first term students of the Faculty of Information Technology. Finally, there were 125 students willing to cooperate. Two of them were females and the rest of them were males. All students were in the age from 19 to 21 years. This gives us a set of voices that are very similar one to another. Such testing set is deadly for the speaker recognition algorithms.

The voice database consists of 125 speaker samples. Each speaker was asked to say eleven times the Czech word 'Emanuel', which is a name phonetically good for

recognition purposes. Six of the utterances were signed as training samples and the remaining five utterances became testing samples. These samples were used to test the quality of the speaker recognition process. As all speakers said the same word, in case of speech recognition it is wanted to recognise the word, but in case of speaker recognition, we want to distinguish one speaker from another. Hence, we have 1 word and 125 various speakers. Good algorithm with a proper set of features must be able to recognise the speakers well. Given these conditions, this task is not easy.

The utterances were recorded using a common microphone with a low signal-to-noise ratio, which should test the quality of the algorithms and the chosen features. The sampling frequency of the recordings was 22 050 Hz and the precision 16 bits per sample. The sampling frequency was higher than usually. However, you can downsample it easily, but you cannot resample when a better sampling frequency is needed.

3.2 Unique Vector Generating

In Section 2.2 an algorithm is proposed that creates a unique vector from the speech signal using the long-term LPC spectrum. To test the quality of the generated vector, some experiments were made. The quantisation performed on the groups of frequencies influences the final unique vector. The higher is the value of the tolerance as used in Equations (19) and (20), the worse is the quality of the vector.

For all tests the first eight maxima were chosen to generate a unique vector. In Figure 12, you can see influence of the size of the tolerance to the FAR and the FRR when testing using a) the training samples, and b) the unknown samples. As expected, when using the training samples the FRR equals zero, since the vector is constructed so that it is not possible to reject a training vector. Of course, the increasing tolerance increases the FAR, which is not desired. However, even when the tolerance equals 1, the FAR does not exceed 4 %, which is very good result. The results of the unknown samples testing are compared with testing of the training samples. When testing the unknown samples, the FRR ranges from 35 % to 85 %, which is bad result. Nevertheless, the FAR values are below 4 %, which is again very good result, since it proves the vectors are far one from another in terms of the distance defined by Equation (26).

In Table 1, unique vectors of the training sample of a valid user can be compared to the corresponding unique vectors of another sample of the same user and of a sample of an unauthorized user. The congruent values of the distances between the individual vectors given by Equation (27) are marked. It is clear that some tolerance is necessary, since even the sample of the same user does not generate the same unique vector when the tolerance is low. Positive is the fact that even when the tolerance is high, no one of the vectors of the sample of the unauthorized user is. To compare the distances given by Equation (26) and (27), see Figure 13. In the tests when using Equation (26), the minimal distance of all the unique vectors generated for each valid user was 23 and, when using Equation (27), the minimal distance was 74, which is rather good result.

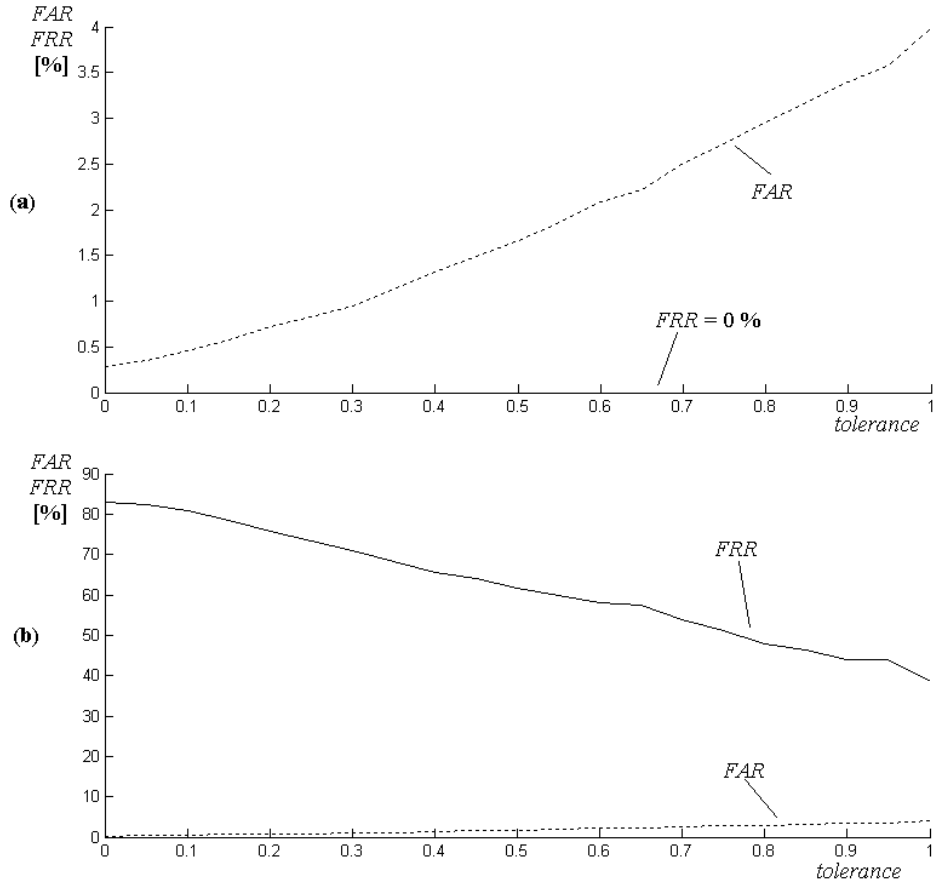

Fig. 12. Comparison of the FAR and the FRR as functions of the tolerance a) when testing the training samples, and b) when testing unknown samples

In Figure 14, the distances given by Equation (26) as a function of a sample of the model and a sample of the reference model are shown. Samples used for this test were the unknown ones.

The distinguishable diagonal is desired, since it represents distance of the reference models, which should equal zero; or, in this case, it should be represented by the highest values (the Z-axis is reversed to show the diagonal). You can also see that sometimes the distance of the reference model and corresponding model is not the lowest (or the highest) one. This is because the unknown samples were used to testing and this behaviour illustrates the difficulty of re-estimableness of the unique vector.

t	unique vector of the training sample of a user	unique vector of the unknown sample of the same user	d	unique vector of the unknown sample of another user	d
0.00	0F0D07160E2C1E68	0F0D07160E2B1D68	2	0E0D07150D2D1E68	4
0.05	0E0C06150D2A1C63	0E0C06150D291B63	2	0D0C06140C2B1C63	4
0.10	0D0B06140D281B5E	0D0B06140D271B5E	1	0C0B06130C291B5E	4
0.15	0D0B06140C261A5A	0D0B06140C251A5A	1	0C0B06130B271A5A	4
0.20	0C0A06130C251957	0C0A06130C241957	1	0B0A06120B261957	4
0.25	0C0A05120B231853	0C0A05120B221853	1	0B0A05110A241853	4
0.30	0B0A05110B221750	0B0A05110B211750	1	0A0A05110A231750	3
0.35	0B0905110A21164D	0B0905110A20164D	1	0A0905110922164D	3
0.40	0A0905100A20154A	0A0905100A1F154A	1	090905100921154A	3
0.45	0A0905100A1E1448	0A0905100A1E1448	0	09090510091F1448	3
0.50	0A08040F091D1445	0A08040F091D1445	0	0908040F081E1445	3
0.55	0908040F091C1343	0908040F091C1343	0	0808040F081D1343	3
0.60	0908040E091C1241	0908040E091C1241	0	0808040E081D1241	3
0.65	0908040E081B123F	0908040E081B123F	0	0808040E071C123F	3
0.70	0907040D081A113D	0907040D081A113D	0	0807040D071B113D	3
0.75	0807040D0819113B	0807040D0819113B	0	0707040D071A113B	3
0.80	0807040C0819103A	0807040C0819103A	0	0707040C0719103A	2
0.85	0807040C07181038	0807040C07181038	0	0707040C06181038	2
0.90	0807030C07171037	0807030C07171037	0	0707030C06171037	2
0.95	0706030C07170F35	0706030C07170F35	0	0606030C06170F35	2
1.00	0706030B07160F34	0706030B07160F34	0	0606030B06160F34	2

Table 1. Comparison of the unique vectors of the training sample of a user, the unknown sample of the same user, and the unknown sample of another user together with the distance given by Equation (27)

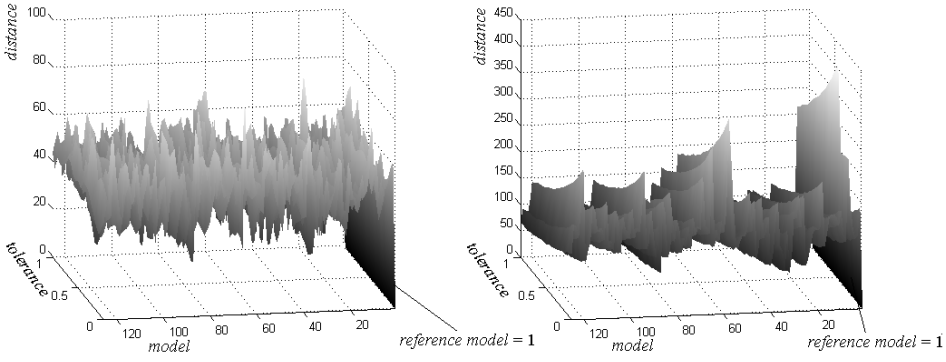

Fig. 13. Comparison of the distance measures. The first one was chosen as the reference model. On the left there is the distance given by Equation (26), and on the right there is the distance given by Equation (27).

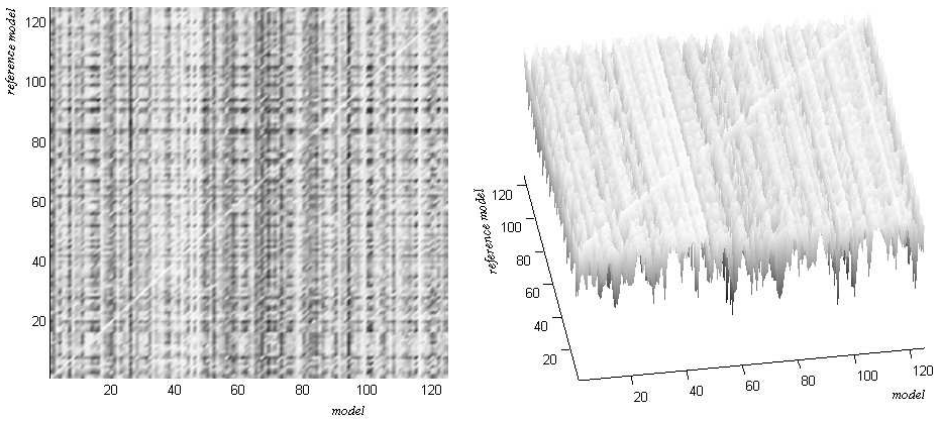

Fig. 14. Distance given by the Equation (26) as function of a model and a reference model, Z-axis reversed. You can clearly see the diagonal. Low (in this case high due to the reversed Z-axis) diagonal values express high distance of the reference model from the other models. In this case, a set of five unknown samples from each of 125 models (speakers) was tested.

In any case, the experimental results prove this way as a possible one. The results are not as good as necessary for a real application. However, some principles proposed here can be improved so that our voice could be used as a biometric key. When using the features as proposed, there are e.g. some possibilities of improving the quantisation step estimation – instead of the linear scale a non-linear one could be used, or another mechanism of the estimation of the quantisation step. It is even not necessary to quantise the frequencies, because there are surely other possibilities.

4 CONCLUSIONS

The designed MBSS was presented in conferences and was accepted by many experts. However, the final realisation of this MBSS is a future work. The vectors that are generated using the introduced algorithm are unique in the given set of voice samples. The distance of the nearest pair of the unique vectors was 23 (74, using another distance measure). There are some serious problems when generating the unique vectors. Though they are unique for the given set, they are not easily re-estimable. This is very limiting factor, since the FAR was very low, which is desired, but the FRR was too high. In many cases this is not a problem, but in the real application is it a big trouble for the user leading to uninstalling of such security system. As the users are indolent, they do not like to repeat the voice password many times before it is accepted, at last.

Biometric security systems are advancing and their importance is growing. An example of this can be the way of use of the biometric technologies in Germany.

Nowadays, they are testing quality of the fingerprint scanners to use them in the near future. The German government prepares ID cards containing fingerprints, which is a great step ahead. Nevertheless, the fingerprints have been many times verified and proved unique, and they are a respected biometric feature. Voice based technology is still waiting for such respect – it has its strengths, i.e. there is a source of information, but there is no suitable tool to extract them and use them yet.

This work offers many suggestions for further research. In the BSS field, there are many ways of possible research topics, too. Remember we used only two biometric technologies to design a MBSS. This can be improved and some methods of the decision fusion must be designed, though there are many existing solutions, some new ones could be developed especially for this task. A more complex way of the unique vector generating would be welcomed. When talking of the unique vector, more topics of further investigation can be named, namely the basis for unique vector specification, the quantisation step estimation, the overall algorithm specification, and many others.

The Biometric Security Systems can offer much space for further work. Reward for the development of the first-quality BSS is a higher reliability and users' satisfaction with the functionality and absence of all the passwords, keys, and other tools used nowadays.

REFERENCES

[1] Orság, F.: Vision für die Zukunft, Biometrie. Kreutztal, DE, b-Quadrat, 2004, pp. 131–145, ISBN 3-933609-02-X.

[2] Drahanský, M.—Orság, F.: Biometric Security Systems: Robustness of the Fingerprint and Speech Technologies. In: BT 2004 – International Workshop on Biometric Technologies, Calgary, CA, 2004, p. 99–103.

[3] Drahanský, M.—Orság, F.—Zbořil, F.: Biometry in Security Applications. In: Proceedings of 38th International Conference MOSIS '04, Ostrava, CZ, MARQ, 2004, p. 6, ISBN 80-85988-98-4.

[4] Drahanský, M.—Orság, F.—Smolík, L.: Biometric Security Systems. In: Proceedings of Mikulášská Kryptobesídka 2003, Praha, CZ, ECOM, 2003, p. 10.

[5] Drahanský, M.—Orság, F.—Smolík, L.: Design of a Biometric Security System. Bonn, DE, BSI, 2003, p. 4.

[6] Drahanský, M.—Orság, F.: Biometric Security Systems: Fingerprint and Speech Technology. In: Proceedings of the 1st Indian International Conference on Artificial Intelligence, Tallahassee, US, IICAI, 2003, p. 703–711, ISBN 0-9727412-0-8.

[7] Jain, A.—Hong, L.—Kulkarni, Y.: A Multimodal Biometric System Using Fingerprint, Face, and Speech. Michigan State University, USA, 2001.

[8] Orság, F.: Rozpoznávání Hlasů. In: Sborník prací studentů a doktorandů, Brno, CZ, CERM, 2000, p. 216–218, ISBN 80-7204-155-X.

[9] Rodman, D. R.: Computer Speech Technology. Boston, Mass.: Artech House, 1999.

[10] Sigmund, M.: Speaker Recognition – Identifying People by their Voices. Conferment Thesis FEE BUT, Brno, 2000, ISBN 80-214-1590-8.

[11] Jan, J.: Cislicova Filtrace, Analyza a Restaurace Signalu. BUT, Brno, CZ, 1997.

[12] Markel, J. D.—Gray, A. H.: Linear Prediction of Speech. Springer Verlag, New York, 1976.

[13] Schukat-Talamazzini, E. G.: Automatische Spracherkennung. Braunschweig/ Wiesbaden: Vieweg, 1995.

[14] Sigmund, M.: Estimation of Vocal Tract Long-Time Spectrum. In: Proceedings of Elektronische Sprachsignalverarbeitung, Vol. 9, Dresden, 1998, pp. 190–192.

[15] Sigmund, M.: Speaker Normalization by Long-Time Spectrum. In: Proceedings of Radioelektronika'96, Brno, CZ, 1996, pp. 144–147.

[16] Woodward, J. D., Orlans, N. M., Higgins, P. T.: Biometrics: Identity Assurance in the Information Age. McGraw-Hill/Osborne, Berkley, USA, 2003, ISBN 0-07-222227-1.

[17] Ashbourn, J.: Biometrics: Advanced Identity Verification. Springer-Verlag, London, GB, 2000, ISBN 1-85233-243-3.

[18] Tistarelli, M.—Jain, A. K.: Biometric Authentication. Proceedings of the International ECCV 2002 Workshop, Springer-Verlag, Berlin, DE, 2002, ISBN 3-540-43723-1.

[19] HumanScan: BioID? Technology. 2004, `http://www.bioid.com`.

Filip Orság graduated at FEI BUT in 2001. During the years 2000–2001 he stayed at FH Wiesbaden in Germany where he finished his master thesis. In 2001 he started his doctoral studies at FIT BUT and finished it in 2004. Currently he is an assistant professor at FIT BUT. His research interests include speaker recognition, biometry and biometric security.