Brno University of Technology
Faculty of Information Technology
Department of Intelligent Systems

Ing. Filip Orság

BIOMETRIC SECURITY SYSTEMS

SPEAKER RECOGNITION TECHNOLOGY

BIOMETRICKÉ BEZPEČNOSTNÍ SYSTÉMY

TECHNOLOGIE ROZPOZNÁVÁNÍ MLUVČÍCH

Short version of Ph.D. Thesis

Study field:	Information Technology
Supervisor:	Doc. Ing. František Zbořil, CSc.
Opponents:	Doc. Dr. Ing. Jana Klečková
	Doc. Ing. Milan Sigmund, CSc.
Presentation date:	October 27, 2004

Keywords: speaker recognition, speaker verification, speaker identification, biometric security system, hidden Markov model

Klíčová slova: rozpoznání mluvčího, verifikace mluvčího,, identifikace mluvčího, biometrické bezpečnostní systémy, skryté Markovovy modely

The original of the dissertation is available in the library of the Faculty of Information Technology of the Brno University of Technology.

ISBN 80-214-2771-X
ISSN 1213-4198

CONTENTS

1 INTRODUCTION

It has been quite a long time since the first Biometric System was 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 could be the biometric systems included. Scepticism, of course, has its place in our live and is necessary to improve our work. Such scepticism is doubled in case of the security systems. However, even if the popularity of the biometric systems is not high yet, they are supposed to be very important soon.

1.1 State of Art

Nowadays, there are various applications based on the speech technology. The speech technology can be divided into three main parts: *speech recognition*, *speaker recognition*, and *other recognition*. Emphasis is laid on the speech recognition more than on the other ones. The most widely used techniques in the speech technology are the hidden Markov models, or neural networks experiencing their renaissance.

In the speech recognition, there is a machine trying to answer the question "*What was said by a speaker?*". Thus, the content of the speech is recognised. Result of this process should be a transcription of a text said by the speaker, accomplishment of a speech command, or another action. There are many ways of usage of the speech recognition technology [9][17][21][23][24][30]. Before all, phone banking, device controlling, or just "simple" speech-to-text transcription can be mentioned.

The speaker recognition can be divided into two main groups: *speaker identification* and *speaker verification*. The process of the speaker identification answers a question "*Who is speaking*?" On the other hand, the speaker verification answers a question "*Is the one, who is speaking, really the one, who he is claiming to be*?". I.e. in case of the identification, we are trying to identify an unknown voice among other voices, and in case of the verification, we would like to determine a similarity of two voices, one of which is known and the other one unknown.

The other recognition includes everything other that is not a speech or speaker recognition. Among the other recognition may be included *stress recognition* or *alcohol detection.* These sorts of recognition are based on a fact that some features of the voice change under the influence of stress. Likewise, the alcohol changes voice characteristics even when there are only few thousandths per mille in blood. There are many other sorts of the recognition using voice. Among them may be named e.g. *mental condition recognition*, *gender recognition*, or *age recognition*.

The range of use of the speech technology in the biometry is wide. The most common one is the speaker authentication using his/her voice. In some cases, another speech technology is applicable, e.g. in case of the phone banking can be the system fully automatic, which is realised by combination of the speech recognition, speaker recognition, and speech synthesis.

The speaker recognition is becoming very important in the biometry. Nowadays, only the fingerprint technology is fully accepted so that many real applications have been created on the base of this technology. The speech technology is in the frame to

be as successful as the fingerprint technology, but there are some limiting factors, which make this technology not as popular as it could be. There are only few commercial speaker verification based security systems now, e.g. Apple's voice login for its iMac®, Sprint's Voice FONCARD®, or the BioID by HumanScan [47], etc. However, more systems are supposed to come soon.

1.2 Goals of the Dissertation

The main goal of the dissertation is to research some new ways of the speech signal processing and its application in the field of the Biometric Security Systems (BSS). The speech recognition process consists of speech signal processing, feature extraction, and classification. The last task is integration of the speech technology to the BSS. The goals of the dissertation can be divided into three main topics described in the following chapters.

1.2.1 Speech Signal Processing

Speech signal processing focused on the speaker recognition is the first topic of the dissertation. The speech signal processing consists of some elementary steps (Chapter 2.1): a *recording*, *voice activity detection*, a *pre-emphasis*, a *framing*, and a *windowing*. The first goal is to improve a quality of the processing with respect to the speaker recognition. In this stage of the speaker recognition, the aim became improvement of the Voice Activity Detection (VAD).

1.2.2 Speaker Dependent Feature Extraction

Definition of speaker dependent features suitable for the speaker recognition is the next topic. There are many possible usable features – *LPC coefficients*, *MFCC*, and many others [17][18][20][21][23][24][30]. The second goal is to extract some new speaker dependent features from the speech signal so that it was possible to use them for the speaker recognition. A new group of speaker dependent features (Chapter 3) based upon a speaker dependent frequency filter bank was introduced. The new features are the *Speaker Dependent Frequency Cepstrum Coefficients*.

1.2.3 Design of Biometric Security System (BSS)

Design of Biometric Security System (BSS) including the speaker recognition technology is the last goal of the dissertation. The speaker recognition may be integrated in many ways. The BSS will be formed from the following security elements: common password, fingerprint authentication, and a voice-based authentication. The designed BSS should be able to accept the basic security elements. Besides, it should provide tools for generating a cryptographic key applicable to the cryptographic services. The BSS is designed only theoretically.

The biometric security systems have their advantages and disadvantages [3][4][5][6]. The advantages are *universality*, *uniqueness*, *low circumvention, scalability* and, in some cases, *permanence*. The disadvantages are *exactingness*, *difficult implementation*, *cooperation unwillingness*, and, in some cases, *inconstancy*, which is a counterpart to the permanence of some of them.

2 SPEECH SIGNAL PROCESSING

In this chapter, there is some theory of the speech signal processing. In the first part, the digitising, the recording, and the pre-processing are described. Then, some common features are described. The last part of this chapter engages in the process of the voice activity detection.

2.1 Recording, Digitising and Pre-processing

Recording of a signal is the first stage in the whole process of the speaker recognition. Recording and digitising is important and can influence the speaker recognition very much [19][34][39]. The analogue speech signal is recorded using a microphone. Subsequently to the recording, the analogue signal is sampled and quantised.

Framing is next important step in the signal processing process. The recorded discrete signal $s(n)$ is always of a finite length N_{total}, but is usually not processed as a whole. The signal is framed. Length of a frame is $N << N_{total}$ samples. Total count of the frames is defined as integer part of the quotient of the total signal length and the frame length. Usually, an overlapping of the individual frames is defined. The overlapping is used to increase precision of the recognition process.

Before a further processing, the individual frames are *windowed*. The frame itself is implicitly windowed by a rectangular window. However, spectral characteristic of the rectangular window is unsuitable. This is why other windows are applied. There are many types of windows using for this purpose. The most frequently used one is a Hamming window [19][36].

Pre-emphasis [14][17][23] is a processing of the input signal by a low order digital FIR filter. It is used to flatten spectrally the input signal in favour of vocal tract parameters. It makes the signal less susceptible to later finite precision effects [31]. Given a signal $s(n)$, the pre-emphasised signal is

$$s_p(n) = s(n) - \lambda s(n-1) \tag{2.1}$$

where λ is a pre-emphasis coefficient from the interval $\langle 0.9; 1 \rangle$.

2.2 Common Features

Feature extraction is a crucial phase of the speaker verification process. Well-chosen feature set may result in quality recognition as well as wrongly chosen feature set may result in a poor recognition.

2.2.1 Energy and Zero-Crossing Rate

Energy of a signal [14][17][30] expresses strength of the signal. Its value is usable to voice activity detection, because the energy of a voice is higher than the energy of a noise (which is not valid in case of consonants – they are much more similar to the noise). The energy of the *j*-th frame $s(j,n)$, *N* samples long, is

$$E(j) = \sum_{n=1}^{N} s^2(j,n), \quad j = 1,2,\ldots,J \tag{2.2}$$

where J is total number of frames. Energy of the background noise depends on the quality of the microphone used to record the signal.

Zero-crossing rate (ZCR) expresses how many times cross the signal value of zero. The ZCR of the noise and some consonants is usually much higher than the ZCR of the vowels. The ZCR of the j-th frame is mathematically formulate [14] as

$$Z(j)=\frac{1}{2}\sum_{n=1}^{N-1}\left|\text{sign}\left(s(j,n)\right)-\text{sign}\left(s(j,n+1)\right)\right|,\quad j=1,2,\ldots,J \tag{2.3}$$

where N is length of the j-th frame of a signal. The function $\text{sign}(s(n))$ is defined as

$$\text{sign}\left(s(n)\right)=\begin{cases}+1, \text{ when } s(n)>0 \text{ or } s(n)=0 \text{ and } s(n-1)>0\\ -1, \text{ when } s(n)<0 \text{ or } s(n)=0 \text{ and } s(n-1)<0\end{cases} \tag{2.4}$$

2.2.2 Linear Predictive Coding (LPC)

An approximation of a speech signal may be calculated by a linear combination of the Linear Prediction Coding (LPC) coefficients [17][22] and M previous samples of the original signal. This approximation is called linear prediction of the signal. The LPC coefficients $a(m)$ are solution of the set of the M Yule-Walker equations

$$\sum_{m=1}^{M}a(m)\phi(i,m)=\phi(i,0),\quad i=1,2,\ldots,M \tag{2.5}$$

The Yule-Walker equations can be solved [17][19][20] by the covariance method or by the autocorrelation method using the Levinson-Durbin recursion. Signal calculated using the LPC coefficients is usually smoothed. Disruptive, speaker specific influence is suppressed. The LPC coefficients of can be used to calculate an LPC frequency spectrum, which is smoother than the FFT spectrum. In the smoothed spectrum, there are obvious resonance frequencies of the formants. The discrete LPC spectrum is defined as

$$S_{LPC}(k)=\left|1-\sum_{m=1}^{M}a(m)\cdot e^{-2m\pi j\frac{k}{N}}\right|^{-2},\quad k=0,1,\ldots,N \tag{2.6}$$

where M is number of the LPC coefficients (prediction order), a_m are the LPC coefficients, N is length of the signal $s(n)$ and k is the discrete frequency.

2.2.3 Discrete Transforms

Discrete transforms are widely used in the speech processing. The first one is the Fourier Transform. It is a very useful one, because it uses complex exponentials as its basis functions [19][20][34]. Given a periodic signal $s(n)=S(n+N)$ with a period N we can define the Discrete Fourier Transform (DFT) as

$$S(k)=\sum_{n=0}^{N-1}s(n)e^{-j2\pi\frac{nk}{N}},\quad k=0,1,\ldots,N-1 \tag{2.7}$$

The second useful transform is the Discrete Cosine Transform (DCT) [20][38]. It is so often used because of its energy compaction, which results in its coefficients being more concentrated at low indices than the coefficients of the DFT. This allows

us to approximate a signal using fewer coefficients [39]. There are several definitions of the DCT. Coefficients $C(k)$ of one of them, DCT-II, are defined as

$$C(k)=\sum_{n=0}^{N-1}s(n)\cos\left(\pi k\frac{n+1}{2N}\right),\quad k=0,1,\ldots,N-1 \tag{2.8}$$

where $s(n)$ is a real signal. Both the DFT and DCT have their inverses [19].

2.2.4 Mel-Frequency Cepstrum Coefficients (MFCC)

Some psychoacoustic experiments were undertaken to derive scales attempting to model the natural response of the human perceptual system, since the cochlea of the inner ear act as a spectrum analyser [20]. Fletcher's work [41] pointed to the existence of critical bands in the cochlear response, which results in a choice defining a frequency scale called mel-frequency scale [41], which is linear below 1 kHz, and logarithmic above, with equal numbers of samples taken below and above 1 kHz. The mel-scale is based on experiments with simple sinusoidal tones [20][41].

The Mel-Frequency Cepstrum Coefficients (MFCC) are defined as the real cepstrum of a windowed short-term signal derived from the FFT of that signal. It differs from the real cepstrum in the nonlinear frequency scale used for approximation of the behaviour of the auditory system. For the windowing of the spectrum, a filter bank is used. The mel-frequency cepstrum is the DCT of the I filter outputs

$$c(j)=\sum_{i=0}^{I-1}\ln\left(\sum_{k=0}^{N-1}H(i,k)\cdot|S(k)|^2\right)\cdot\cos\left(\pi n\frac{i-1}{2I}\right),\quad j=0,1,\ldots,I \tag{2.9}$$

where $|S(k)|$ is the magnitude of the FFT of the signal $s(n)$, which is N samples long, and $H(i,k)$ is the i-th filter from total number of I triangular filters in the filter bank. The count of the filters varies from 24 to 40 in different applications. In the speech recognition, only the first 13 MFCC are used together with the first 13 delta-MFCC and the first 13 delta-delta-MFCC.

2.2.5 Average Long-Term LPC Spectrum

The long-term LPC spectrum [33] is calculated using the average autocorrelation coefficients [18]. These are estimated over all frames of the given signal $s(n)$. The average autocorrelation coefficients can be used to derive the average LPC coefficients $\bar{a}(i)$ from Eq. (2.5) using the Durbin recursive procedure and corresponding LPC spectrum is calculated using the Eq. (2.6) substituting the LPC coefficients by the average LPC coefficients.

The signal normalization by a long-term LPC spectrum is a tool useful for speech signal normalization and applicable to speaker-independent speech recognition [40]. A framed signal is given. The autocorrelation coefficients $R(j,k)$ of the j-th frame are of the order $k=0,1,\ldots,M$. Given the average LPC coefficients $\bar{a}(i)$ the normalized autocorrelation coefficients are defined as

$$R_n(j,k)=R(j,0)\sum_{i=0}^{M}\bar{a}^2(i)+\sum_{m=1}^{M}\left[R(j,|k-m|)+R(j,|k+m|)\right]\sum_{i=0}^{M-m}\bar{a}(i)\cdot\bar{a}(i+m) \tag{2.10}$$

2.3 Voice Activity Detection (VAD)

Voice Activity Detection (VAD) is important topic in many speech-processing systems. The system has to distinguish noise from a voice. The systems processing single words have to locate beginning and end of the words [48]. In the real-world cases, the noise interferes with some phonemes too much to recognise the beginning and the end correctly. Such phonemes are [34]: unvoiced fricatives or voiced fricatives that become unvoiced at the end of a word, unvoiced stops – plosives, nasals at the end, and trailing vowels at the end.

2.3.1 Common VAD Methods

The first VAD method is determination of the noise threshold. In this case, the threshold is determined from the noise. This method is disadvantageous, since in noisy environment with low-class signal source is it difficult to determine the threshold. It is because the unvoiced phonemes are much like noise, which makes their recognition more difficult. This method is using because of its simplicity. In fact, it is usable in case there is good signal source with a high signal-to-noise ratio.

Another widely used method is a method based on a comparison of the energy and the zero-crossing rate of all frames. The zero-crossing rate is higher in case a frame represents an unvoiced region or in case it does not belong to the voice at all. The value of the energy is high in case a frame is in a voiced region of the speech signal. Knowing this it is possible to find approximate endpoints of a word. This method is based on searching of thresholds of the ZCR and energy [14].

The envelope tracking method is based on the tracking and comparing of the envelope of the speech signal with an adaptive threshold determined as a mean of the envelope of the background noise. Two independent signals are needed to estimate the threshold – a noise and speech signal, which is limiting. Advantage of this method is simplicity and possibility of use of an analogue device to the VAD.

2.3.2 VAD Using Neural Network

Neural Networks (NN) are well known for their ability to separate vectors and to classify them. This was the main reason for their application in the task of the VAD [7]. From a large amount of NN, the forward feed network was chosen as the best one for this task. The NN suitable for this task contains either one or two hidden layers, one output neuron and input neuron count same as length of the input vector.

There are many feature sets usable for the separation of the noise and the voice, but in the tasks like this, it holds true that a simple feature set is better than a complex one. One unusual choice is a combination of the ZCR together with a Sum of the Magnitudes of the Fourier Spectrum (SMFS). The SMFS is defined as

$$SMFS(j) = \sum_{k=1}^{N/2} |S(j,k)|, \quad j = 1,2,\ldots,J \tag{2.11}$$

where N is the length of the j-th frame, J is the total number of the frames, and $S(j,k)$ is given by the Eq. (2.7).

3 SPEAKER DEPENDENT FEATURE EXTRACTION

Speaker dependent features are speaker recognition oriented features. These features emphasise speaker individuality and are not usable for other purposes than the speaker recognition.

3.1 Speaker Dependent Frequency Filter Bank (SDFFB)

Three filter shapes were chosen: a triangular shape, a Gaussian curve shape and a shape of the Tukey window [36]. Let's assume A an amplitude, N denotes length, $0 \le F_{low} < F_{centre} < F_{high} < N$ denote basic discrete frequencies of the filters, and $k = 0,1,\ldots,N-1$ is a discrete frequency.

The triangular filter is defined in the discrete frequency domain as

$$H_{Triang}(k, F_{low}, F_{centre}, F_{high}) = \begin{cases} A \cdot \dfrac{k - F_{low}}{F_{centre} - F_{low}}, & k = F_{low}, F_{low}+1, \ldots, F_{centre} \\ A \cdot \left(1 - \dfrac{k + F_{centre}}{F_{high} - F_{centre}}\right) & k = F_{centre}+1, F_{centre}+2, \ldots, F_{high} \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

The Gaussian curve filter is based on the Gaussian (normal) probability density function and is defined as

$$H_{Gauss}(k, F_{low}, F_{centre}, F_{high}) = A \cdot \exp\left[-\left(\frac{k - F_{centre}}{0.25(F_{high} - F_{low})} \right)^2 \right] \tag{3.2}$$

The Tukey filter is a combination of a rectangular window and a Hann window, i.e. it is a cosine-tapered window. The original equation [36] was slightly modified for the filtering purposes. The final equation is

$$H_{Tukey}(k, F_{low}, F_{centre}, F_{high}) = \begin{cases} \dfrac{A}{2} \cdot \left(1 - \cos\left(2\pi \cdot \dfrac{k - F_{low}}{N_{Hann}}\right)\right), & k = F_{low}, F_{low}+1, \ldots, F_{low} + \dfrac{N_{Hann}}{2} \\ A, & k = F_{low} + \dfrac{N_{Hann}}{2} + 1, \ldots, F_{high} - \dfrac{N_{Hann}}{2} - 1 \\ \dfrac{A}{2} \cdot \left(1 - \cos\left(2\pi \cdot \dfrac{k - N_{Rect}}{N_{Hann}}\right)\right), & k = F_{high} - \dfrac{N_{Hann}}{2}, \ldots, F_{high} \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

In this case play a role some other variables. N_{Hann} is defined as

$$N_{Hann} = (1 - \alpha) \cdot (F_{high} - F_{low} + 1) \tag{3.4}$$

where α is a ratio of taper to constant section and $0 \le \alpha \le 1$. When $\alpha = 0$, the filter corresponds to a rectangular filter. When $\alpha = 1$, the filter corresponds to a Hann window (filter). N_{Rect} in the Eq. (3.3) is complement to the N_{Hann} and is defined as

$$N_{Rect} = \alpha \cdot (F_{high} - F_{low} + 1). \tag{3.5}$$

Now, we can describe Speaker Dependent Frequency Filter Bank (SDFFB), which is a new approach to a filter bank definition. It is based on an average long-term LPC

spectrum. The SDFFB differs from the mel-frequency based bank in a distribution of filter centres, in amplitude of the filters, and in the shape of the filters.

Basic idea results from the fact of dissimilarity of human vocal tracts. Shape of the vocal tract differs obviously in some important details. The positions of the maxima and minima in a long-term LPC spectrum become central frequencies F_{centre} of the individual filters. The positions are members of two sorted sets

$$F_{\max}(l) = \left\{ f_l^{\max} \right\} = \left\{ f_0^{\max}, f_1^{\max}, f_2^{\max}, \ldots, f_L^{\max} \right\}, \quad l = 0,1,\ldots,L \tag{3.6}$$

$$F_{\min}(l) = \left\{ f_l^{\min} \right\} = \left\{ f_1^{\min}, f_2^{\min}, \ldots, f_I^{\min}, f_{L+1}^{\min} \right\}, \quad i = 1,2,\ldots,L,L+1 \tag{3.7}$$

where $f_l^{\max}$ is the position of the l-th maximum and $f_l^{\min}$ is the position of the l-th minimum. The frequencies should satisfy

$$f_0^{\max} < f_1^{\min} < f_1^{\max} < f_2^{\min} < \cdots < f_L^{\max} < f_{L+1}^{\min} \tag{3.8}$$

Now, we can merge the sets $F_{\max}(l)$ and $F_{\min}(l)$ into one ordered set

$$F(i) = \left\{ f_0^{\max}, f_1^{\min}, \ldots, f_L^{\max}, f_{L+1}^{\min} \right\}, \quad i = 0,1,\ldots,I+1 \tag{3.9}$$

where $I = 2L$ is total number of the filters in the filter bank. As $L = 8$, the filter bank consists of $I = 16$ filters. Thus, we need at least eight extremes plus one more maximum and one more minimum. It was experimentally proved, that the number of the extremes is even higher than eight. Generally, the filter bank is defined as

$$H_{SDFB}(i,k) = H_{SDF}(k, F(i-1), F(i), F(i+1)), \quad i = 1,2,\ldots,I \tag{3.10}$$

where there are I Speaker Dependent Filters $H_{SDF}(k, F_{low}, F_{centre}, F_{high})$ – one of the filters defined in the previous chapter. The discrete frequency is $k = 0,1,\ldots,N-1$ and N is length of the signal or frame.

The amplitude A offers two types of the SDFFB – type I and type II. The type I (SDFFB-I) assumes the amplitude $A = 1$ ~ *constant amplitude filter bank*. The SDFFB of the type II (SDFFB-II) assumes the amplitudes equal to the values of the long-term LPC spectrum at the positions F_{centre} ~ *variable amplitude filter bank*.

3.2 Speaker Dependent Frequency Cepstrum Coefficients

The Speaker Dependent Frequency Cepstrum Coefficients (SDFCC) are much like the mel-frequency cepstrum coefficients in terms of the overall algorithm. It differs in the filter bank used for the computation of the coefficients. The SDFCC are computed using the SDFFB defined in the previous chapter. Instead of the triangular mel-frequency based filter bank, we use a SDFFB given by the Eq. (3.10). Then in the Eq. (2.9), the triangular filter H is replaced by other one H_{SDF} defined in the previous chapters. Then, the SDFCC are

$$c_{SDFB}(j) = \sum_{i=0}^{I-1} \ln\left(\sum_{k=0}^{N-1} H_{SDF}(i,k) \cdot |S(k)|^2 \right) \cdot \cos\left(\pi n \frac{i-1}{2I} \right), \quad j = 0,1,\ldots,I \tag{3.11}$$

where I is total number of the filters in the filter bank, $S(k)$ is the FFT of a signal $s(n)$, which is N samples long. Though the SDFCC are computed the same way as the MFCC, they are not same at all. The difference in the filter banks and filter shapes issues in very different results in case of the speaker recognition.

4 SPEAKER RECOGNITION

The speaker recognition divides into two groups: *speaker identification* and *speaker verification* [17][18][30][31][43][45][46]. The speaker identification aims at answering the question "*Who is speaking?*". On the other hand, the speaker verification tries to answer the question "*Is the one, who is speaking, really the one, who he claims to be?*". Thus, in case of the speaker identification we try to identify an unknown voice among other voices, which may be e.g. stored in a database. In case of the speaker verification, we would like to determine a similarity of two speakers. One of them is known – his voice features are stored in a database. The other one is unknown and he/she claims to be the known one.

For a BSS, the speaker verification is a reasonable choice, but the speaker identification can be used too in the way the verification is used. There are many ways of accomplishing the recognition, but the most popular is the method based on the Hidden Markov Models with the Gaussian Mixtures. Usually, the process of the speaker recognition consists of the *recording and pre-processing*, *feature extraction*, *pattern matching*, and *decision making*.

The recording and the pre-processing of the input signal are described in Chapter 2.1. The feature extraction is described in Chapters 2.2 and 3. The last two steps should be explained more.

4.1 Hidden Markov Models with Gaussian Mixtures

After the feature extraction, it is possible to come to the next phase of the speaker recognition – the pattern matching. It can be accomplished in many ways, there are – nonparametric techniques, methods based on maximum-likelihood and Bayesian parameter, linear discriminant functions, neural networks, or stochastic methods [16][17][35]. For this task was chosen the Hidden Markov Models with the Gaussian Mixtures (HMM-GM).

In the Markov processes generally, there are considered states, which correspond to an observable (physical) event. However, this model is too restrictive [48] to be applicable to many problems of interest, so that the HMM are defined. The HMM is defined by the following variables: total number N of states, total number M of distinct observation symbols per state, state transition probability distribution A, observation symbol probability distribution B, and initial state distribution π. The observation sequence is noted as $O = o_1, o_2, \ldots, o_T$ and the state sequence is referred to as $X = x_1, x_2, \ldots, x_T$.

There are three main problems of the HMM. Given a HMM $\lambda = (A, B, \pi)$ how can be determined probability $P(O \mid \lambda)$ of an observation sequence O in the model λ? There is a direct solution to this calculating all combinations of the state sequences of the length T. However, there are N^T possible state sequences. For the large T, the direct solution to this problem is unusable. For this purposes, a Forward Procedure and a Backward Procedure were developed [17][28][49]. These procedures are induction procedures and both decrease total number of calculations.

The second problem is summarised as follows. Given a HMM $\lambda = (A, B, \pi)$ how can we choose a state sequence X so that the probability of the occurrence of the observation sequence O would be maximal? Possible solution to this can be the Viterbi algorithm [11][16][25]. Result of the algorithm is an optimal state sequence. The algorithm is similar in implementation to the forward procedure.

The third and the most difficult problem of the HMM is to determine a method to adjust the model parameters A, B, and π. Given the model $\lambda = (A, B, \pi)$ how to choose the parameters to maximise the probability $P(O, X \mid \lambda)$, given an observation sequence O and the state sequence X? There are some methods of adjustment of the HMM parameters [16][25][27][28][49]. One of the algorithms – segmental k-means algorithm – algorithm calculates parameters of the HMM $\lambda = (A, B, \pi)$ so that the probability $P(O, X \mid \lambda)$ was maximal and at the same time the state sequence X was optimal state sequence to the given observation sequence. Another solutions can be the Baum-Welch re-estimation algorithm or the maximum likelihood algorithm. These algorithms are general ones. They creates a general complete HMM based on the training observation sequences. Nevertheless, in the speech processing we do not need a complete model. The models are reduced to the left-to-right models, which models the forward behaviour of the speech.

There are some practical issues in using the HMM and some limitations. Some of them are: problem of the initial estimation, model topology, training criteria, duration modelling etc [11][16][17][20][23][24][28][49].

In case the observations do not come from a finite set, but from a continuous space, the discrete output distribution discussed before needs to be modified [20]. The main difference lies in a different form of output probability functions. As the continuous output probability density functions $b_j(\bar{o})$, where $\bar{o}$ is an n-dimensional observation symbol, were chosen the multivariate Gaussian mixture density functions, which can approximate any continuous probability density function.

4.2 Decision-Making

Decision-making is the final step of the speaker recognition [43][45][46]. From the previous steps, we have the result of the pattern matching. Now, we have to decide, whether the result of the matching is positive or negative. The decision-making can be done in many ways, but every-time it finishes by a simple threshold based (linear) classification.

Accuracy of the speaker recognition process is an important aspect of the security systems. There is a pair of measures used to determine accuracy of the BSS – the False Acceptance Rate (FAR) and the False Recognition Rate (FRR) [43]. The FAR is also known as a False Match Rate or Type II error. It expresses how many times someone is inaccurately positively matched. Hence, the higher the FAR, the more intruders may be positively matched. The FRR is also known as a False Non-Match Rate or Type I error. It denotes the number of times someone is inaccurately rejected. It means, the higher the FRR, the more times must someone who should be

accepted complete the enrolment procedure. The combination of these two measures helps to determine quality or accuracy of the biometric security system.

Threshold estimation is a serious problem too. First, we must know the purpose of the designed BSS. Then, we must decide, how strong should be the protection. The FRR and FAR can help us to choose proper threshold value. If we want all authorised users to be accepted without problems, we have to choose a threshold according to the low values of the FRR. However, when we choose such threshold, then some unauthorised users may be accepted as well. On the other hand, if we choose a higher threshold value, then more authorised users may be rejected. Thus, choice of the threshold is a compromise between the FAR and the FRR.

The balanced threshold is the one lying on the crossing of the FAR and FRR curves. The value of the crossing is called Equal Error Rate (EER), i.e. at this point the FAR and the FRR should be of the same value. The algorithm searching for the EER gets an absolute value of the subtraction of the values of the FAR and the FRR for each given threshold and searches for the minimum. It returns the approximate value of the threshold.

Another usable threshold can be searched by calculating the Optimal Error Rate (OER). This algorithm finds minimum of the sum of the FAR and the FRR. When the OER is chosen for threshold estimation, then the overall recognition error rate is the lowest possible. The OER cannot be performance measure, since it is a special case optimisation and is incomparable due to dissimilarities of the models.

4.3 Approaches to the Speaker Recognition

There are two different approaches to the speaker recognition. The first one is based on the speaker verification and the second one is based on the speaker identification. When verifying a speaker, there are two possible results – the speaker is (‘yes’) or is not (‘no’) the one, who he/she claims to be. When identifying a speaker, there are other two possible results. If the unknown speaker were found in the voice database, the answer would be an identification number of the speaker. Hence, the answer would be ‘yes’, since the user was found in the database. If the unknown speaker were not found in the voice database, the answer would be unknown (unauthorised) user, i.e. the answer would be ‘no’.

Main difference between the two approaches is that in case of the speaker verification the user has to tell the system, who he/she claims to be. Then, the system compares the stored pattern of the claimed user with the new pattern of the unknown user. Hence, a one-to-one comparison is performed. In case of the speaker identification, the system itself determines the user. Thus, a one-to-many comparison is performed. If the user’s voice were not found in the database, the answer would be ‘no’, otherwise the answer would be ‘yes’.

5 DESIGN OF BIOMETRIC SECURITY SYSTEMS

The third goal of the dissertation is to design a Biometric Security System (BSS) utilising at least two biometric technologies. The theory presented in this chapter is supported by the experimental results of the speaker recognition as presented in Chapter 6.3, and the overall concept of the BSS is supported by results of testing of the unique vector presented in Chapter 6.4. Our concepts of the BSS were accepted by many experts when presenting in conferences [3][4][5][6][8] and workshops [2].

5.1 Single- and Multi-Biometric Security System

Usually, a one-level biometric security system [3][4][5] is applied to provide the security services. The one-level biometric system or *Single-Biometric Security System* (SBBS) consists typically of an input device and a processing unit. Among the input devices belong: scanners, microphone, and many other devices. These devices serve the acquisition of the physical biometric features. Data acquired from the input devices are processed by be a built-in or an external device. Both the external and the built-in solution have their pros and contras. The external solution is cheaper but it provides weaker protection. The built-in solution is more expensive but the protection is better.

Multi-Biometric Security System (MBSS) is a biometric system based on a combination of more than one biometric technology [3][4][5]. Main advantage over the SBSS is complexity that makes the system more robust to the FAR. Generally, authorisation using multiple biometrics reduces to a fusion problem, 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.

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. A decision fusion of all partial decisions must be estimated. One partial decision can be based on the result of the speaker verification as described in Chapter 4.3. The overall decision is based on a fusion of the decisions made by the individual biometric modules and can be based on theory of probability, on a neural network, or on a fuzzy classification.

5.2 Task of the BSS

Biometric security system is a security system, which protection is based on the biometric features. The BSS can be applied in many ways and situations. Usually, we want to protect an object or a service – we want to enable the authorised users to access to some network resources (a biometric login). Another typical task can be a protection of the objects like buildings or other facilities (a biometric doorkeeper). These applications require a reliable protection method, since in these cases the price of the protected object is higher than in case of the network resources.

There is a dependency between the price of the protected object and the strength of the protection. In case the protected object is not too valuable, can the common solutions fulfil requirements of the protected object owner.

There are many possibilities of strengthening of the security. One of them lies in the BSS, which are not widely used, because people do not trust them yet. Since present, only the SBSS have been used [43][45][46]. The MBSS are rare, but some commercial solutions [47] and some books on this topic [10] are on hand.

5.3 Biometry in Cryptography

The 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 generate at least one unique vector, which will be then considered as a part of the biometric cryptographic key.

Two preconditions must be fulfilled. The first one is the *uniqueness*. The vector used as the cryptographic key must be unique. If the vector were not unique, there would be possibility of false acceptance of the biometric features of someone other (possibly unauthorised) user. The second precondition is the ability of *re-estimableness* of the unique vector, e.g. the speech technology can provide only one unique vector difficult to re-gain.

However, the uniqueness is not such a big trouble. When using more than one biometric technology, we can supply ambiguity of the vectors given by one technology by the other biometric technology. Both of them can supply each other.

5.4 Unique Vector Generating from the Speech Signal

The process of unique vector generating from the speech signal is rather challenging. As the speech features are so unstable, the uniqueness and the re-estimableness are not fulfilled. The speaker vocal tract characteristics change during whole life, which aggravates the process, since the system must be held up to date. To generate a unique vector, a special method must be designed, since the features possible to use to get a unique vector are not precisely re-estimable. A tolerance must be defined, within which the features may vary. This process is very similar to the quantisation. We define a quantisation step and then we sample the features along one dimension.

Consider a long-term LPC spectrum derived from the LPC coefficients of the 22^{nd} order. We can use the same spectrum as the one used in case of the SDFFB. There is a set of frequencies like in the Eq. (3.6) omitting just the first maximum. This set is a base for the unique vector. Like in the case of the SDFFB, we chose a group of $L = 8$ maxima, which implies we can have 8 unique values. When we scale them to the range from 0 to 255, we can use them to create an array of 16 bytes.

To define the quantisation steps q_l for each group $l = 1,2,...,L$ of maxima, we need more than 1 training sample. Having N training samples we get N sets of maxima. Upon these sets, we can base estimation of the quantisation step. Except from the quantisation step, there is a value of the initial shift. Prior to get the quantised value using the step q_l, we have to subtract the initial shift s_l to get proper results.

As the training samples do not cover all frequencies, 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 the tolerance and the lower the final error can be. Thus, given a normalised tolerance factor t and the quantisation step q_l, the quantisation step with the tolerance is defined as

$$\hat{q}_l = q_l + t \cdot q_l, \quad l = 1,2,\ldots,L \wedge 0 \le t \le 1 \tag{5.1}$$

and, given the initial shift s_l, the initial shift with tolerance is defined as

$$\hat{s}_l = \min_{n=1,\ldots,N}\left(f_{l,n}^{\max}\right) - \frac{t \cdot q_l}{2} - \hat{q}_l \operatorname{int}\left(\frac{1}{\hat{q}_l}\left(\min_{n=1,\ldots,N}\left(f_{l,n}^{\max}\right) - \frac{t \cdot q_l}{2}\right) + 1\right), \quad l = 1,2,\ldots,L \tag{5.2}$$

where $f_{l,n}^{\max}$ is l-th group of maxima from N training samples. Given function int(x) returning integer part of x, the quantised value $\hat{v}_l(x)$ can be defined as

$$\hat{v}_l(x) = \operatorname{int}\left((x - \hat{s}_l) \cdot \hat{q}_l^{-1}\right) \tag{5.3}$$

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 as the l-th value v_l. Now, we can create a unique vector for each speaker in the database defined as

$$V = \left(v_1\left(\bar{f}_1^{\max}\right), v_2\left(\bar{f}_2^{\max}\right), \ldots, v_L\left(\bar{f}_L^{\max}\right)\right) \tag{5.4}$$

whereby $\bar{f}_l^{\max}$ is mean of the l-th group of maxima consisting of N frequencies, i.e. there are N training samples. Given the unique vector V it is possible to transform it to a hexadecimal string containing $2L$ hexadecimal digits. Upon that a measure, which expresses a distance between two unique vectors, can be defined.

5.5 Robustness of the BSS

When generating a biometric key, there must be enough information in the biometric features usable to the key generating [2]. In the speech technology, possibility of generating a key is limited. We must define a new method to generate a unique re-construable vector. There are typical speech features usable for this purpose. Some of them are the *formant frequencies* corresponding approximately to the resonance frequencies of the vocal tract cavities. We suppose various speakers have various long-term characteristics of the formant frequencies.

Given the unique vector generated as described in the previous chapter, we can compute the factors of the vector. The factor equals to the number of bits needed to store the vector. Thus, given $L = 8$ values transformed so that they range from 0 to 255, we need 64 bits to store the vector, which results in a factor of 2^{64}, i.e. there are 2^{64} possible combinations.

The length of 64 bits is however not enough to satisfy nowadays needs in the field of the cryptography. The factor may be even higher, because it strongly depends on the chosen features. Nevertheless, we can say that 64 bits are enough in a MBSS and the missing bits can be supplied by the other biometric technologies. For example, the fingerprint technology can offer even a factor of 2^{214} [2], which provides together with the speech technology a factor of 2^{278}, which is enough.

6 EXPERIMENTAL RESULTS

Some experiments had to be done to prove validity of the algorithms and to test quality of the proposed features.

6.1 Voice Database and Feature Sets

The voice database used for testing was built in the cooperation with a group of the first term students of BUT FIT. There were 125 students willing to cooperate; all of them in the age from 19 to 21 years and 2 of them females. This gives us a set of voices similar one to another. Such testing set is deadly for recognition algorithms.

The voice database consists of 125 speakers, eleven samples for each speaker. Six samples were for training and the remaining five for testing. The utterances were recorded using a common microphone with low signal-to-noise ratio. The sampling frequency was 22050 Hz and precision was 16 bits per sample.

For the test, some typical feature sets were provided. The first group of features consisted of the common features (see Chapter 2.2). The first feature set consisted of 12 and 24 LPC coefficients noted as *LPC12* and *LPC24*. The next sets consisted of the Mel-Frequency Cepstrum Coefficients. These sets were 13 MFCC, 13 MFCC with 13 first order delta coefficients, and 13 MFCC with 13 first order and 13 second order delta coefficients. These ones were noted as *MFCC*, *MFCC-D*, and *MFCC-DD*. Last common coefficients were 12 and 24 cepstrum coefficients noted as *CEPS12* and *CEPS24*. Next, there was a group of six Speaker Dependent Frequency Cepstrum Coefficients based sets (see Chapter 3.2). All the SDFCC-based sets consisted of 16 coefficients. The first group of sets was calculated using the filter bank type I and the triangular filter noted as *SDFCC-I-Triang*, the Gaussian filter noted as *SDFCC-I-Gauss*, and the Tukey filter noted as *SDFCC-I-Tukey*. The second group of sets was calculated using the filter bank type II and the triangular filter noted as *SDFCC-II-Triang*, the Gaussian filter noted as *SDFCC-II-Gauss*, and the Tukey filter noted as *SDFCC-II-Tukey*.

6.2 VAD Using Neural Network

The VAD using neural network was experimentally tested [7]. Performance of the VAD was related to the manually found beginnings and ends of a group of utterances chosen from the voice database (Chapter 6.1).

The first tested network contained two hidden layers, both of them with 10 neurons. The second one was same with only five neurons in the hidden layer. Next three tested configurations consisted only of one hidden layer with 10, 5, and 2 neurons. The NN was trained using the Levenberg-Marquardt back-propagation method, goal performance was set to 10^{-10}, maximum number of epochs was set to 500, and minimum gradient was 10^{-12}. The tangential sigmoid function was chosen as the transfer function of all the neurons.

Results of all tests are summarized in Table 6.1. The results show that it is not useful to enlarge the neural network, because, in many cases, performance of the smaller networks was better than performance of the larger ones. The smaller the

network is the better is the possibility of its usage in the real time applications (because of its speed). In addition, the features that are usually used in the speech processing proved to be improper for the task of the voice activity detection using neural network. As the best choice showed combination of the ZCR with the SMFS, and the MFSC.

Table 6.1
Error rate of the endpoints detection using a forward feed neural network.

Features	E ZCR	SMFS ZCR	LPCC				MFSC 12	MFCC 12
			4	8	16	24		
***N* = 1024 (i.e. 46 ms), *O* = 512**								
10x10x1	6,7	1,1	1,7	4,2	7,6	11,2	3,4	16,0
5x10x1	4,2	0,6	4,2	8,4	10,1	2,5	0,8	10,1
10x1	1,7	0,8	2,5	3,4	9,2	9,2	1,7	33,6
5x1	2,5	0,3	3,4	7,6	4,2	5,4	2,5	37,0
2x1	7,6	0,3	8,4	7,6	9,2	3,3	2,9	3,4
***N* = 512 (i.e. 23 ms), *O* = 256**								
10x10x1	17,6	56,3	2,5	7,9	3,8	4,2	11,0	10,8
5x10x1	5,8	1,3	2,1	8,3	6,3	3,4	2,5	7,1
10x1	5,9	1,7	2,5	4,6	6,7	4,2	2,1	8,8
5x1	6,7	0,8	4,6	6,3	5,8	2,5	3,5	12,5
2x1	4,6	0,8	2,5	5,6	5,0	7,1	1,7	2,1
***N* = 256 (i.e. 11 ms), *O* = 128**								
10x10x1	2,0	1,7	2,3	4,2	8,9	6,0	2,3	11,2
5x10x1	1,9	0,8	3,1	4,8	5,2	3,1	2,1	7,7
10x1	1,5	1,6	4,6	7,3	6,4	9,1	3,1	15,8
5x1	3,5	0,8	1,7	11,0	4,6	8,5	1,0	4,4
2x1	1,2	0,6	2,9	9,2	4,0	5,4	2,1	10,8

6.3 Speaker Recognition Experiments

The speaker recognition experiments consist of two main groups of the speaker recognition – the speaker verification approach and the speaker identification approach. The experiments were done upon a HMM-GM classifier with 3, 5, and 7 states. These numbers of states were chosen because of the given voice password. It consists of seven phones, so that seven should be enough. Often a HMM-GM with a lower number of states performs better than another one with a higher number of states, so that it was decided to test the HMM-GM even with 3 and 5 states.

In the Figure 6.1, you can compare a relationship of the FAR and the FRR with the threshold when using the common features, the SDFCC-I-Gauss, and SDFCC-II-Gauss features. You can see that the curves differ for each of the feature sets. Almost ideal seems to be the curves given by using the MFCC feature set, however

the EER is rather high in comparison with the LPC12. The FAR/FRR curves for the other numbers of HMM-GM states and other testing samples are similar.

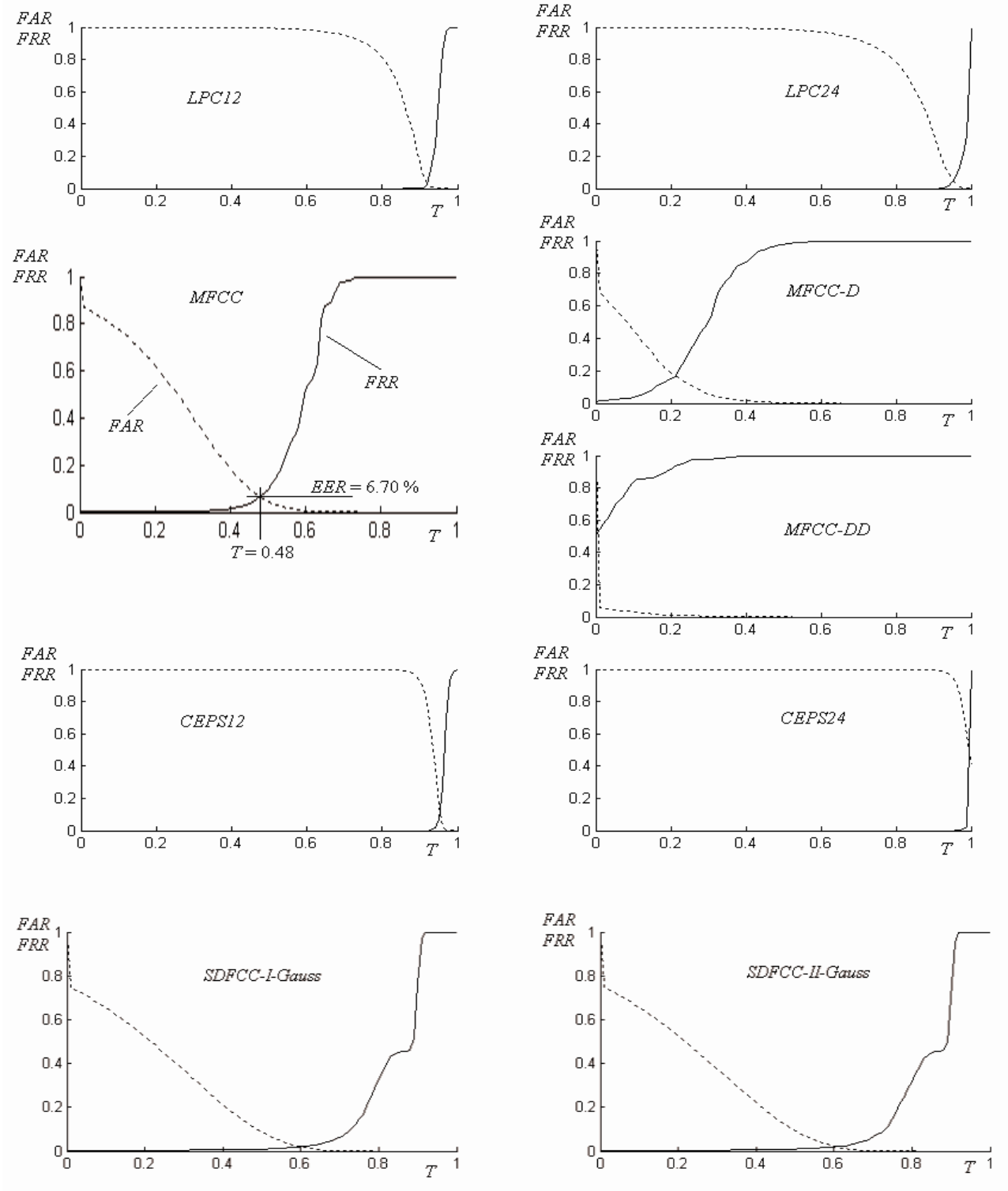

Figure 6.1*FRR (solid line) and the FAR (dotted line) as the functions of the threshold T, when using the HMM-GM with 3 states and combined set of samples (6 training samples + 5 unknown samples for every speaker). Here, the commonly used feature sets are compared – LPCC, MFCC and cepstrum coefficients. In case of the MFCC can you see a marking of the EER and corresponding threshold T. The values of the FAR and of the FRR are normalised.*

In the table Table 6.1, there are summarised experimental results of the speaker verification and identification. There are the Equal Error Rates of the recognition in percents. Results of two types of samples are noted in the table, the first one containing six training samples and the second one containing unknown samples. The first group should be the best performer, but is not valid as a performance comparison. The results of the second group are the worst ones, but they are nearest the reality, since in the real world application there are only unknown samples.

Table 6.1
EER [%] of the speaker identification and verification

		identification						**verification**					
States		**3**	**5**	**7**	**3**	**5**	**7**	**3**	**5**	**7**	**3**	**5**	**7**
Features		**unknown**			**training**			**unknown**			**training**		
LPC12		16.2	5.4	5.5	0.6	0.5	0.3	4.8	5.0	4.8	0.5	0.6	0.6
LPC24		20.0	3.6	9.0	2.9	3.2	2.5	8.0	7.2	8.0	1.3	1.2	0.9
MFCC		42.6	7.1	7.8	12.0	0.9	39.1	11.5	18.8	19.3	1.3	4.7	5.3
MFCC+D		48.6	8.8	8.6	45.2	2.9	37.7	24.2	31.5	31.1	10.5	10.4	7.3
MFCC+DD		49.1	48.9	49.0	44.6	39.9	27.4	83.9	85.1	87.0	22.9	15.3	7.7
CEPS12		41.8	0.1	9.6	33.7	0.2	28.7	12.5	12.8	12.7	9.6	8.3	6.9
CEPS24		47.9	7.8	7.8	48.5	8.5	48.6	29.9	30.2	29.5	72.3	72.5	72.2
SDFCC-I	Gauss	**5.0**	6.8	6.6	**0.0**	**0.0**	**0.0**	4.1	4.0	4.9	**0.0**	**0.0**	**0.0**
	Triang	6.4	6.4	**6.1**	**0.0**	**0.0**	**0.0**	4.1	**3.9**	5.0	**0.0**	**0.0**	**0.0**
	Tukey	6.3	6.7	7.2	**0.0**	**0.0**	**0.0**	4.1	4.8	5.1	**0.0**	**0.0**	**0.0**
SDFCC-II	Gauss	5.8	**5.9**	6.2	**0.0**	**0.0**	**0.0**	**3.9**	4.5	5.1	**0.0**	**0.0**	**0.0**
	Triang	5.3	7.6	6.5	**0.0**	**0.0**	**0.0**	4.1	4.6	5.1	**0.0**	**0.0**	**0.0**
	Tukey	6.9	6.6	5.5	**0.0**	**0.0**	**0.0**	4.4	4.8	**4.8**	**0.0**	**0.0**	**0.0**

The overall best verification solution seems to be the SDFCC-II-Gauss based feature set, with the ERR of 3.9 % in case of the HMM-GM with 3 states. The number of states has some influence on the recognition error. In case of the commonly used features increases the EER sometimes and sometimes it decreases. There seems not to be any dependency. In case of the SDFCC based feature sets, the EER increases with the increasing number of states. The number of three states proves to be enough to verify a speaker rather accurately.

The FAR/FRR dependency on the threshold is similar to the shapes presented in Figure 6.1 with one important difference in the shape of the FAR curve, which is

nearly constant for the low values of the threshold. This is because of the properties of the identification process. When $T = 0$, FAR should equal to 1. This is true in case of the verification, but it is not true, when identifying a user. When recognised, the changes of the threshold cannot influence the FAR, only the FRR changes. In other words, we check the unknown sample against all models stored in the database. Then, if the maximal likelihood exceeds the threshold, we declare the unknown sample belongs to the model with the maximal likelihood. When, in the event, the sample does not belong to the winning model, the FAR increases. When correctly found, the FAR cannot further increase, because the change of the threshold does not influence falseness of the recognition, since there is only one winner every time.

The EER of the speaker identification depends on the number of the HMM-GM states. The identification approach winner when using the common features is the LPC12 based set with no exception. The situation is not as clear when using the SDFCC based features. The number of the states influences the accuracy much (relatively to the results of the other SDFCC based features). When using the HMM-GM with 3 states, the best is the SDFCC-I-Gauss feature set, in case of the 5 states is the best one the SDFCC-II-Gauss and, in case of the 7 states, is the winner the SDFCC-I-Triang.

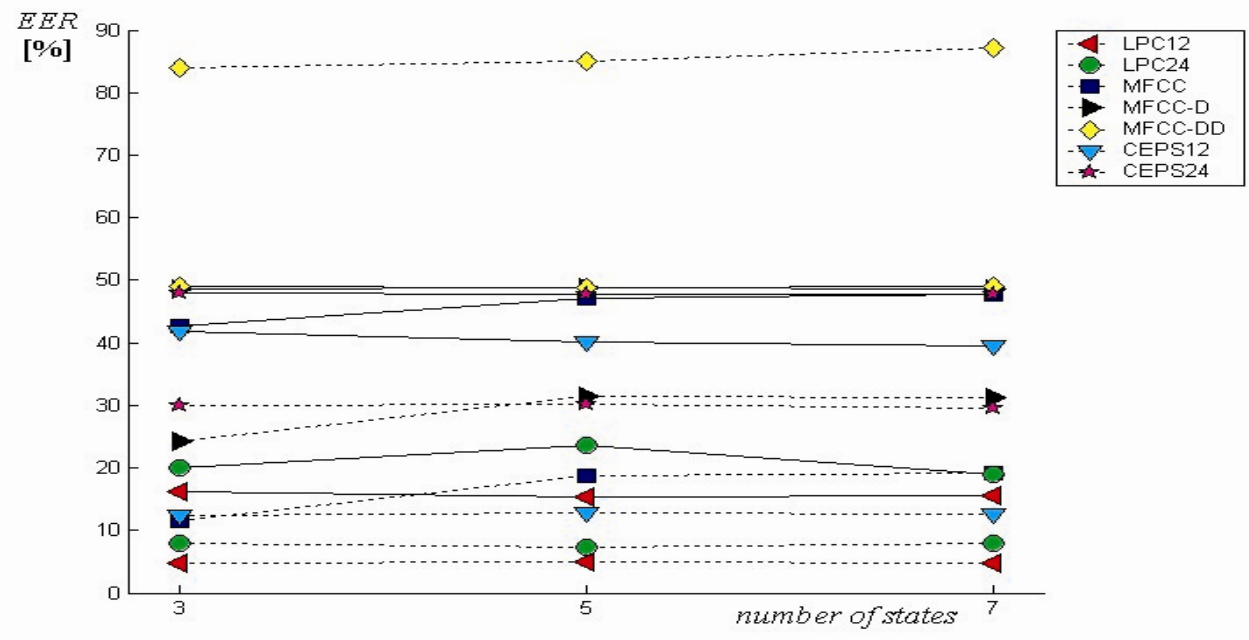

Figure 6.2*Comparison of the EER when using the speaker verification approach (dotted line) and the speaker identification approach (solid line).*

The Figure 6.2 and the Figure 6.3 illustrates influence of the number of the HMM-GM states on the EER. Generally, the speaker verification approach is better than the identification approach. Reason for these results is given by the used algorithm. The features suitable for both, the identification and the verification, are the SDFCC based one. The overall EER of the identification and the verification lies in between 3.5 % and 8 %, which is, in comparison to the common features, very good result. As the common features are not primarily designed for such task, their

results are not so impressive. Generally, the HMM-GM with lower number of features and lower number of states perform better than the others do.

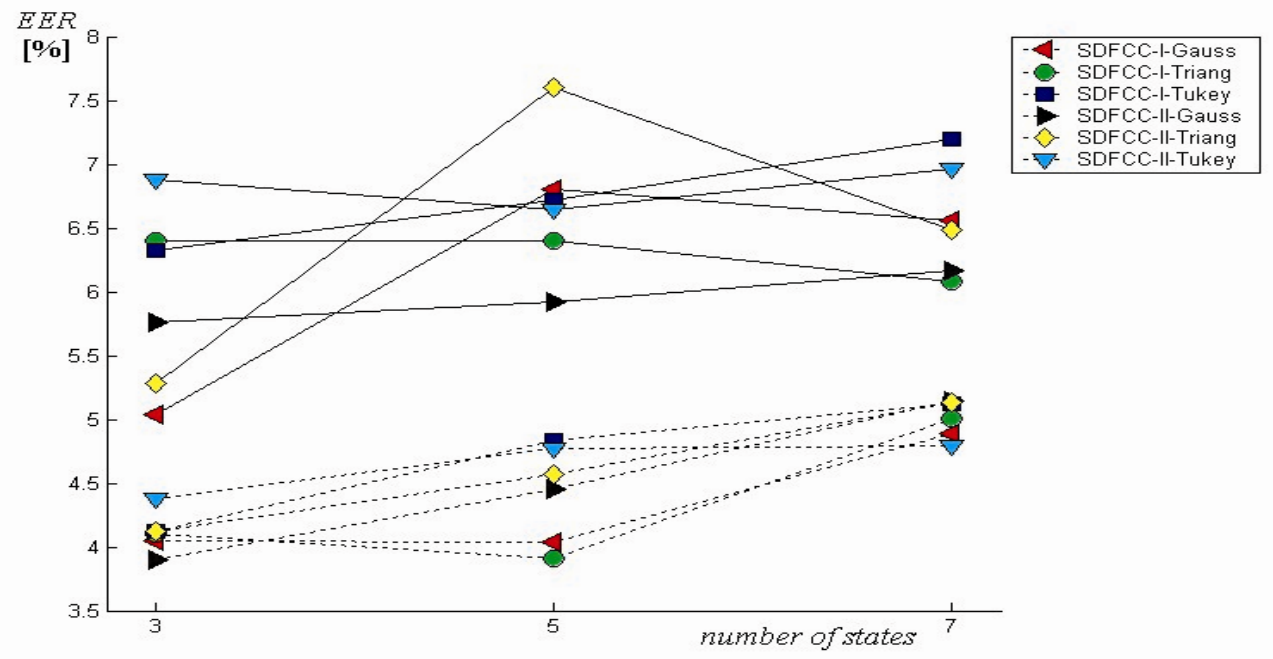

Figure 6.3*Comparison of the EER when using the speaker verification approach (dotted line) and the speaker identification approach (solid line).*

6.4 Unique Vector Generating

In Chapter 5.4, there is proposed a method of unique vector generating from the speech signal using the long-term LPC spectrum. The quantisation performed on the groups of frequencies influences the final unique vector. The higher the value of the tolerance as used in the Eqs. (5.1) and (5.2), the worse is the quality of the vector.

In the Figure 6.4, 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, the FRR equals zero when using the training samples, since the vector is constructed so that it is not possible to reject a training vector. Increasing tolerance increases the FAR, which is not desired. However, even when the tolerance is high, the FAR does not exceed 4%, which is good result. When testing the unknown samples, the FRR ranges from 35% to 85%, which is bad result. The FAR lie below 4%, which is good result.

In Table 6.2, you can compare unique vectors of the training sample of a valid user to the corresponding unique vectors of another sample of the same user and of a sample of an unauthorised user. It is clear that some tolerance is necessary, since even the sample of the same user do not generate the same vector when the tolerance is low.

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.

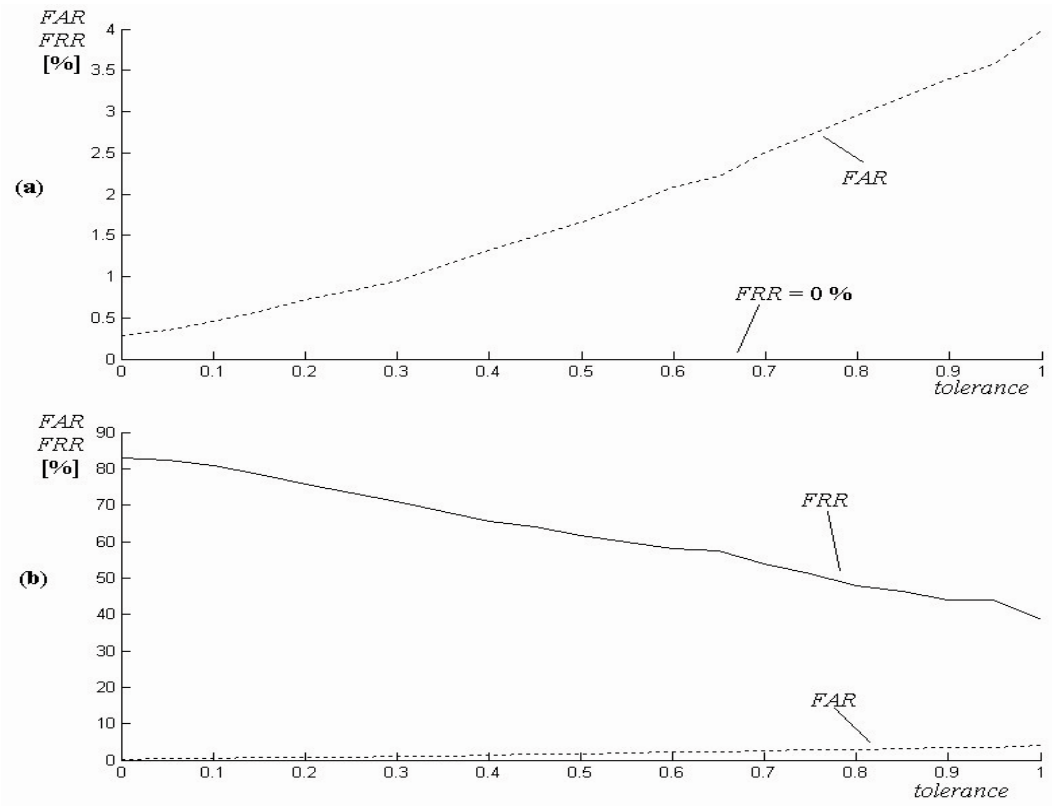

Figure 6.4*Comparison of the FAR and the FRR as functions of the tolerance, (a) training samples results, and (b) unknown samples results.*

Table 6.2

Comparison of the unique vectors

tolerance	unique vector of the training sample of an authorised user	unique vector of the unknown sample of the authorised user	distance	unique vector of the unknown sample of an unauthorised user	distance
0.00	0F0D07160E2C1E68	0F0D07160E2B1D68	2	0E0D07150D2D1E68	4
0.10	0D0B06140D281B5E	0D0B06140D271B5E	1	0C0B06130C291B5E	4
0.20	0C0A06130C251957	0C0A06130C241957	1	0B0A06120B261957	4
0.30	0B0A05110B221750	0B0A05110B211750	1	0A0A05110A231750	3
0.40	0A0905100A20154A	0A0905100A1F154A	1	090905100921154A	3
0.50	0A08040F091D1445	0A08040F091D1445	0	0908040F081E1445	3
0.60	0908040E091C1241	0908040E091C1241	0	0808040E081D1241	3
0.70	0907040D081A113D	0907040D081A113D	0	0807040D071B113D	3
0.80	0807040C0819103A	0807040C0819103A	0	0707040C0719103A	2
0.90	0807030C07171037	0807030C07171037	0	0707030C06171037	2
1.00	0706030B07160F34	0706030B07160F34	0	0606030B06160F34	2

7 CONCLUSIONS

The three goals of the dissertation were accomplished. The first goal aimed at the speech signal processing focused on the speaker recognition was accomplished by the proposal of the voice activity detection using a neural network. Experimental results noted in Chapter 6.2 show that the NN is very usable for this purpose. The VAD with an error lower than 1% is good result.

The second goal – definition of speaker dependent features suitable for the speaker recognition – was accomplished by defining a new set of speaker dependent features – the Speaker Dependent Frequency Cepstrum Coefficients (SDFCC). These coefficients aim at the speaker recognition and are not usable for the speech recognition. The experimental results noted in Chapter 6.3 show their qualities. The Equal Error Rate (EER) of the speaker verification lower than 5% in nearly all cases with the minimum at 3.9%, is excellent taking the conditions given by voice database into account. The EER of the speaker identification lower than 8% with the best EER of 5.04% is good as well. The SDFCC outperformed the common features, which aim mainly at the speech recognition. The good performance of the speaker recognition is result of the dependence of the new features on the speaker, because it strengthens speaker's individual voice characteristics. Generally, the verification approach transpired to perform better than the identification approach.

The third goal of design of BSS including speaker recognition technology was accomplished by proposing a design of a multi-biometric security system based on the speech and fingerprint technologies and proposing an algorithm of unique vector generating. The vectors 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 (respectively 74) using different distance measures. There are some problems when generating the unique vectors. Though unique, they are not easily well re-estimable. This topic showed there is a way of solution of this problem.

This work offers many suggestions for the further research. The SDFC coefficients are not the only solution to the speaker dependent features. The SDFCC can be created in many ways, which gives many possible solutions. When talking of the new speaker dependent features we cannot forget shape of the human's vocal tract, which is supposed to be unique, as the base for the further investigation. Alternatively, the shape of the excitation signal could be very useful in this field.

In the field of the BSS, there are many ways of possible research topics. 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. A more complex way of the unique vector generating would be welcomed. Improvement of the base for the unique vector specification, the quantisation step estimation, the overall algorithm specification, and many others are needed.

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.

8 REFERENCES

[1] Orság, F.: *Vision für die Zukunft*, Biometrie, Kreutztal, DE, b-Quadrat, 2004, p. 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] Orság, F., Zbořil, F.: *Endpoint Detection in the Continuous Speech Using the Neural Networks*, In: Proceedings of 37th International Conference MOSIS´03 Modelling and Simulation of Systems, Ostrava, CZ, MARQ, 2003, p. 7, ISBN 80-85988-86-0

[8] Drahanský, M., Orság, F.: *Fingerprints and Speech Recognition as parts of the biometry*, In: Proceedings of 36th International Conference MOSIS '02, Ostrava, CZ, MARQ, 2002, p. 177-183, ISBN 80-85988-71-2

[9] Orság, F.: *Some Basic Techniques of the Speech Recognition*, In: Proceedings of 8th Conference STUDENT EEICT 2002, Brno, CZ, FEKT VUT, 2002, p. 5, ISBN 80-214-2116

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

[11] Desai, U.B., Dugad, R.: *A Tutorial On Hidden Markov Models*, Technical Report No.: SPANN-96.1, Bombay, IN, 1996

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

[13] Orság, F.: *Válcový model hlasového traktu*, In: Nové smery v spracovaní signálov V., Liptovský Mikuláš, SK, neznámá, 2000, p. 101-106, ISBN 80-8040-071-125X

[14] Sigmund, M.: *Analýza řečových signálů*, textbook, FEI VUT Brno, Brno, 2000, ISBN 80-214-1783-8

[15] Kohonen, T., Hynninen, J., Kangas, J., Laaksonen, J.: *SOM_PAK: The self-organizing map program package*, Report A31, Helsinki University of

Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996
[16] Rabiner, L.R., Juang, B.H.: *An introduction to hidden Markov models*, IEEE ASSP Mag., pp 4-16, 1986
[17] Rodman, D.R.: *Computer Speech Technology*, Boston, Mass.: Artech House, 1999
[18] Sigmund, M.: *Speaker Recognition – Identifying People by their Voices*, conferment thesis FEE BUT, Brno, 2000, ISBN 80-214-1590-8
[19] Jan, J.: *Číslicová filtrace, analýza a restaurace signálů*, BUT, CZ, 1997
[20] Huang, X., Acero, A., Hon, HW.: *Spoken Language Processing*, New Jersey, USA, Prentice Hall, 2001, ISBN 0-13-022616-5
[21] Gold, B., Morgan, N.: *Speech and Audio Signal Processing*, New York, USA, John Wiley & sons, inc., 2000, ISBN 0-471-35154-7
[22] Markel, J.D., Gray, A.H.: *Linear Prediction of Spe*ech, Springer Verlag, New York, 1976
[23] Eppinger B., Herter E.: *Sprachverarbeitung*, Wien: Hanser, 1993
[24] Fellbaum K.: *Sprachverarbeitung und Sprachübertragung*, Berlin, Heidelberg, New York, Tokyo: Springer, 1984
[25] Forney jr.,G.D.: *The Viterbi Algorithm*, Proc. IEEE, vol. 61, no. 3, pp. 263-278, 1973
[26] Kundu A.., Yang H.: *2-D shape classification using HMMs*, IEEE Trans. Patt. Anal. Machine Intell., vol. 13, no. 11, pp. 1172-1184, 1991
[27] Juang B.H., Rabiner L.R.: *The segmental k-means algorithm for estimating the parameters of hidden Markov models*, IEEE Trans.Accoust., Speech, Signal Processing, vol. ASSP-38, no. 9, pp. 1639-1641, 1990
[28] Rabiner L.R., Juang B.H.: *An introduction to hidden Markov models*, IEEE ASSP Mag., pp 4-16, 1986
[29] Hung W., Wang H.: *On the Use of Weighted Filter Bank Analysis for the Derivation of Robust MFCCs*, IEEE Signal Processing Letters, vol. 8, no. 3, 2001
[30] Schukat-Talamazzini E.G.: *Automatische Spracherkennung*, Braunschweig/Wiesbaden: Vieweg, 1995
[31] Xafopoulos, A.: *Speaker Verification*, Tampere International Center for Signal Processing, TUT, Tampere, Finland , 2001
[32] Hui-Ling, L.: *Toward a high-quality singing synthesizer with vocal texture control*, PhD thesis, 2002
[33] Sigmund, M.: *Estimation of Vocal Tract Long-Time Spectrum*, In: Proceedings of Elektronische Sprachsignalverarbeitung, Dresden, Vol. 9, 1998, pp.190-192
[34] Deller, J.R., Hansen, J.H.L., Proakis, J.G.: *Discrete-Time Processing of Speech Signals*, New York, USA, IEEE Press, 2000, ISBN 0-7803-5386-2
[35] Duda, O.R., Hart, P.E., Stork, D.G.: *Pattern Classification*, A Wiley-Interscience Publication, New York, USA, 2001, ISBN 0-471-05669-3

[36] Harris, F. J.: *On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform*, Proceedings of the IEEE. Vol. 66, 1978, pp. 66-67
[37] Burrus, C.S., Parks T.W.: *DFT-FFT and Convolution Algorithms: Theory and Implementation*, New York, John Wiley, 1985
[38] Rao, K.R., Yip, P.: Discrete *Cosine Transform: Algorithms, Advantages and Applications*, Academic Press, San Diego, CA, 1990
[39] Oppenheim, A.V., Schafer, R.W., Buck, J.R.: Discrete-Time Signal Processing, 2nd ed., Upper Saddle River, NJ, Prentice Hall, 1999
[40] Sigmund, M.: *Speaker Normalization by Long-Time Spectrum*, In: Proceedings of Radioelektronika'96, Brno, CZ, 1996, pp. 144-147
[41] Fletcher, H.: *Auditory Patterns*, Rev.Mod.Phys., 1940, pp. 47-65
[42] Baggenstoss, P.M.: *Hidden Markov Models Toolbox*, Naval Undersea Warfare Centre, Newport, RI, 2001
[43] 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
[44] Schroeder, R.M.: *Computer Speech: Recognition, Compression, Synthesis*, Springer-Verlag, Berlin, DE, 1999, ISBN 3-540-64397-4
[45] Ashbourn, J.: *Biometrics: Advanced Identity Verification*, Springer-Verlag, London, GB, 2000, ISBN 1-85233-243-3
[46] Tistarelli, M., Jain, A.K.: *Biometric Authentication*, Proceedings of the International ECCV 2002 Workshop, Springer-Verlag, Berlin, DE, 2002, ISBN 3-540-43723-1
[47] HumanScan: BioID® Technology, 2004, http://www.bioid.com
[48] Rabiner, L.R., Sambur, M.R.: *An algorithm for determining the endpoints of isolated utterances*, Bell System Technical Journal, vol. 54, pp. 297-315, 1975
[49] Rabiner, L.R.: *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, In: Proceedings of the IEEE, Vol. 77, No. 2, 1989

AUTHOR'S CURRICULUM VITAE

PERSONAL DATA

Name:	Ing. **Filip Orság**
Born:	5.12.1977, Brno, Czech Republic
E-mail:	orsag@fit.vutbr.cz
WWW:	http://www.fit.vutbr.cz/~orsag

EDUCATION

2001 – 2004	doctoral (PhD) studies, FIT, Brno University of Technology, CZ
1996 – 2001	MSc. in computer science, FEECS, Brno University of Technology, CZ
2000 – 2001	one year study, Fachhochschule Wiesbaden, Germany
1992 – 1996	grammar school aimed at mathematics and physics, Brno, CZ

LANGUAGES

English:	intermediate (FCE level, grade A)
German:	intermediate (MittelstufeII/III level)

INTERESTS

Technical:	biometric security systems, speech technology, robotics
Personal:	computer technology, riding

PRIZES

09/2002	AFCEA prize for the diploma thesis
12/2001	Siemens prize for the diploma thesis

ABSTRACT

It has been quite a long time since the first biometric systems were introduced and, until present, they have not become widely used. In spite of this, the importance of the biometric security systems is growing now. Many research groups are working on the development of the individual biometric technologies such as the fingerprint recognition, iris or retina recognition or the speaker recognition. However, not many of them combine these technologies together. Future of the biometric security systems is in the combination of more technologies. This work aims at the technology of the speaker recognition and proposes a solution of its integration into a more complex biometric security system.

There are many speaker recognition systems but none of them is reliable and stable enough to be applied as a standalone security system. In the first part of the dissertation there are described the goals of this works and the state of art of the biometry. Then, description of some methods of the common signal processing and techniques of the pre-processing phase together with feature extraction follows. The feature extraction is a crucial phase of the recognition process. In the next part, some new features focused on the speaker recognition are introduced. The following part aims at the speaker recognition itself. The speaker may be recognised in many ways, but the most suitable one for this task seems to be a method based on the Hidden Markov Models. The next part introduces a design of a complex biometric security system based on the speaker recognition and the fingerprint authentication. A method of acquisition of a unique vector from speaker specific features is outlined as well. If the vector proved to be unique, it would be possible to use it with success in cryptography. In the last part, there are summarised experimental results. Some experiments were performed, their results were summarised and interpreted.

Keywords: speaker recognition, speaker verification, speaker identification, biometric security system, hidden Markov model

ABSTRAKT

První biometrické systémy byly představeny již před dlouhou dobou a dosud se nestaly široce využívanými. Přesto důležitost těchto systémů nyní stoupá. Mnoho výzkumných skupin pracuje na vývoji individuálních biometrických technologií jako je například rozpoznávání otisků prstů, rozpoznávání sítnice nebo duhovky, nebo rozpoznávání mluvčích. Nicméně ne mnoho z nich slučuje jednotlivé technologie do jednoho celku. Budoucnost biometrických systémů však leží především v kombinaci více biometrických technologií. Tato práce se zaměřuje na technologii rozpoznání mluvčích a navrhuje řešení její integrace do komplexního biometrického systému.

Existuje již mnoho systémů pro rozpoznávání mluvčích, ale žádný z nich není dostatečně spolehlivý a stabilní, aby bylo možné ho použít jako samostatný bezpečnostní systém. V první části disertace jsou popsány běžné metody zpracování signálů a techniky předzpracování signálů za účelem rozpoznávání mluvčích. Další část práce se zabývá příznaky a jejich extrakcí. Extrakce příznaků je rozhodující a velmi významná fáze procesu rozpoznávání. Jsou zde popsány běžně používané příznaky. V následující části je popsána skupina nových příznaků zaměřených speciálně na oblast rozpoznávání mluvčích. Následuje popis problematiky samotného rozpoznávání mluvčích. Mluvčí mohou být rozpoznáni mnoha způsoby, ale nejvhodnější pro tento úkol se jeví metoda založená na skrytých Markovových řetězcích. V další části práce je představen návrh komplexního biometrického bezpečnostního systému založeného na rozpoznávání mluvčích a otisků prstů. Je zde také uvedena metoda výpočtu unikátního vektoru z řečového signálu. Pokud se prokáže unikátnost tohoto vektoru, bude možné ho úspěšně využít v kryptografii. V poslední části jsou shrnuty výsledky experimentů. Byly provedeny pokusy jejichž výsledky jsou zde shrnuty a vyhodnoceny.

Klíčová slova: rozpoznání mluvčího, verifikace mluvčího, identifikace mluvčího, biometrické bezpečnostní systémy, skryté Markovovy modely