International Journal of Security and Its Applications
Vol. 4, No. 1, January, 2010

Speaker Dependent Coefficients for Speaker Recognition

Filip Orság

Brno University of Technology
Faculty of Information Technology
Bozetechova 2
612 66 Brno, Czech Republic
orsag@fit.vutbr.cz

Abstract

This work aims at speaker recognition based upon a new set of features. Feature extraction is a crucial phase of the recognition process and a proper feature set dramatically influences the speaker recognition. Many well-known features are not suitable for the speaker recognition as those merge the specifics of the individual voices. Therefore, we need features accentuating the individual differences of our voices to be able to recognise speakers reliably. This work introduces new, speaker dependent features called Speaker Dependent Frequency Cepstrum Coefficients (SDFCC), created for the speaker recognition purposes only. Experimental results show their performance in comparison to the well-known features. According to the test results, the SDFCC are, for the speaker recognition, very useful and promising.

Keywords: *speaker, recognition, verification, identification, features*

1. Introduction

Speaker recognition based on a speech signal is one of the most exciting technologies of the human recognition. This is because it is, for men, the most natural form of identification. Scientists are trying to make it natural for a machine as well. However, the task of the speaker recognition is not an easy one.

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?"*. Thus, in case of the speaker identification we want to identify an unknown voice among other voices, which can be e.g. stored in a database, and in case of the speaker verification, we want to determine similarity of two speakers. One of them is known – their voice features are stored in the database – the other one is unknown.

There are many applications for the speaker verification and identification. Example of the speaker identification application is an investigation of a crime. The speaker verification is (relatively) easier to perform than the identification. The main difference is that the speaker being verified usually wants to cooperate and wants to be positively verified, which need not to be true in case of the speaker identification (the criminals usually do not want to be identified and arrested, hence they are not cooperating). Generally, the speaker recognition is

applicable where a restricted access to some facilities or services is required. Result of the speaker recognition is either true (in case a valid user tries to use the protected object) or false (in case an intruder tries to break in). The speaker recognition can be accomplished in many ways, but the most popular is the method based on the Hidden Markov Models with the Gaussian Mixtures (HMM-GM) [1].

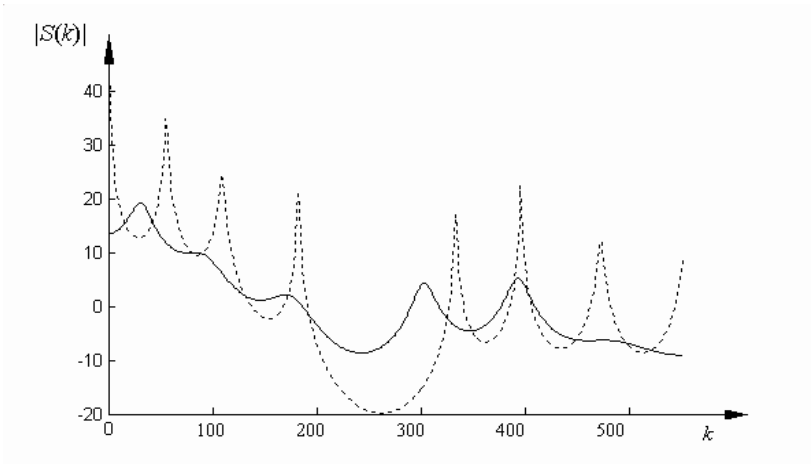

Figure 1. Effect of the signal normalization by a long-term spectrum. Solid line - original LPC spectrum, dotted line – normalized LPC spectrum.

2. Speaker Dependent Feature Extraction

Speaker dependent features are speaker recognition oriented features. These features emphasise speaker individuality and are not usable for any other purpose than the speaker recognition. These features are based on an algorithm used to calculate well-known Mel-Frequency Cepstrum Coefficients (MFCC [2, 3]). The main difference lies in design of the filter bank used in the calculations.

2.1 Filters

For the purpose of the speaker recognition, three filter shapes were chosen: a triangle, Gaussian and Tukey filter. These filters are frequency domain filters. Hence, they are applied to the discrete frequency spectrum of a signal. Suppose length of a discrete frequency spectrum to be N samples, which implies the original signal to be $2N$ samples long. The following filters are defined in the discrete frequency domain. Hence, given F_S is sampling frequency, all frequencies must be multiplied by $(N \cdot F_S)^{-1}$ to transform them to the discrete frequency domain. The original frequencies should lie within an interval $<0;0.5 \cdot F_S>$. Consider $0 \leq f_{low} \leq 0.5 \cdot F_S$ to be a frequency in the frequency domain. Then, it corresponds to the discrete frequency $F_{low} = f_{low} \cdot (N \cdot F_S)^{-1}$.

Triangular filter is the easiest one to implement. and is defined in the discrete frequency domain as

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

where $k = 0,1,\ldots,N-1$ is a discrete frequency, $0 \le F_{low} < F_{centre} < F_{high} < N$ are basic filter frequencies, A denotes an amplitude of the filter, and length of the filter is N samples.

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

$$H_{Gauss}(k, F_{low}, F_{centre}, F_{high}) = A \cdot e^{-\left(\frac{k - F_{centre}}{0.25 \cdot (F_{high} - F_{low})}\right)^2} \quad (2)$$

where $k = 0,1,\ldots,N-1$ is a discrete frequency, $0 \le F_{low} < F_{centre} < F_{high} < N$ are basic filter frequencies and A denotes an amplitude of the filter. Again, length of the filter (signal, or frame) is N samples.

Tukey filter is, in fact, a window, which is usually used to process frames of a signal. It is a combination of the rectangular window and the Hann window [3], i.e. it is a cosine-tapered window and is defined as follows

$$H_{Tukey}(k, F_{low}, F_{centre}, F_{high}) = \begin{cases} \dfrac{A}{2} \cdot \left(1 - \cos\left(2\pi \cdot \dfrac{k - F_{low}}{N_{Hann}}\right)\right), & k = F_{low}, F_{low} + 1, \ldots, F_{low} + \dfrac{N_{Hann}}{2} \\ A, & k = F_{low} + \dfrac{N_{Hann}}{2} + 1, \ldots, F_{high} - \dfrac{N_{Hann}}{2} - 1 \\ \dfrac{A}{2} \cdot \left(1 - \cos\left(2\pi \cdot \dfrac{k - N_{Rect}}{N_{Hann}}\right)\right), & k = F_{high} - \dfrac{N_{Hann}}{2}, \ldots, F_{high} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $k = 0,1,\ldots,N-1$ is a discrete frequency, $0 \le F_{low} < F_{centre} < F_{high} < N$ are basic filter frequencies and A denotes an amplitude of the filter, which is N samples long. N_{Hann} is defined as

$$N_{Hann} = (1 - \alpha) \cdot (F_{high} - F_{low} + 1) \quad (4)$$

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 filter. N_{Rect} in the Eq. (3) is a complement to the N_{Hann} and is defined as

$$N_{Rect} = \alpha \cdot (F_{high} - F_{low} + 1). \quad (5)$$

In Figure 2 you can see shapes of all the filters.

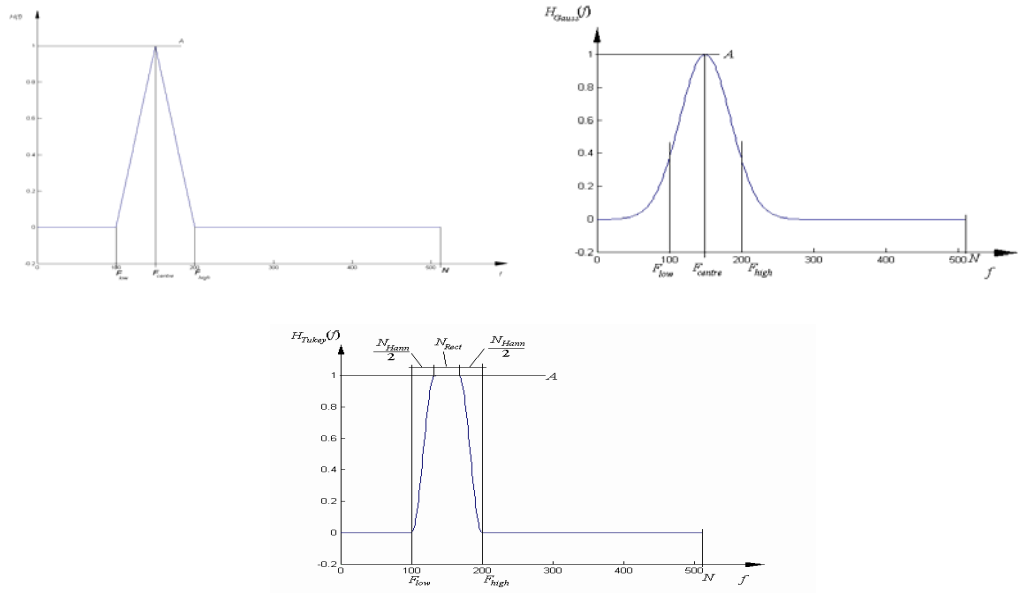

Figure 2. Triangular filter (top left), Gaussian filter (top right), Tukey filter (bottom). All the filters are *N* samples long.

2.2 Speaker Dependent Frequency Filter Bank (SDFFB)

Speaker Dependent Frequency Filter Bank (SDFFB) is a new approach to the filter bank construction. It is based on an average long-term LPC (*Linear Prediction Coefficients*) spectrum [4, 5]. The SDFFB is much like the mel-frequency based bank of triangular filters [2, 3, 6], but the SDFFB differs from the mel-frequency based filter bank in distribution of the centres, in amplitude, and in shape of the individual filters.

Basic idea results from the dissimilarity of human's vocal tracts. Shape of the vocal tracts differs obviously in some important details among people [5]. When a speech is being recognised, it is useful to extract the features, which are speaker independent. This does not hold true in case of the speaker recognition. In this case, it is necessary to extract *speaker dependent features*. However, most known systems use the common features like LPC coefficients, mel-cepstrum coefficients and the like, which are rather general.

Differences of the vocal tracts give us opportunity to use this fact to create speaker dependent filters based on the vocal tract or its model. For this purpose we can use normalised long-time spectrum [4]. In Figure 1 you can see effects of the normalisation of the autocorrelation coefficients by the average long-term LPC coefficients. In Figure 3 there are examples of two normalised LPC spectra. The spectra belong to two male speakers, both in the age of nineteen. A difference in the long-term spectra is obvious. Thus, it should be relatively easy to distinguish one speaker from the other. The peaks in the spectrum are obvious and they are speaker dependent. Hence, if we used them to build the bank of filters, we would be able to create a unique bank of filters for each individual.

The maxima and the minima become central frequencies F_{centre} of the individual filters in terms of the previous filter definitions. First, we define an ordered set of the maxima F_{max} and minima F_{min} as a series of the individual maxima and minima of the long-term normalised

LPC spectrum. Experimentally was found that if the order of the LPCs used for the calculation of the LPC spectrum equals 22, there are at least 10 peaks in the spectrum. We decided to use the first eight extremes for the filter design ($L = 8$).

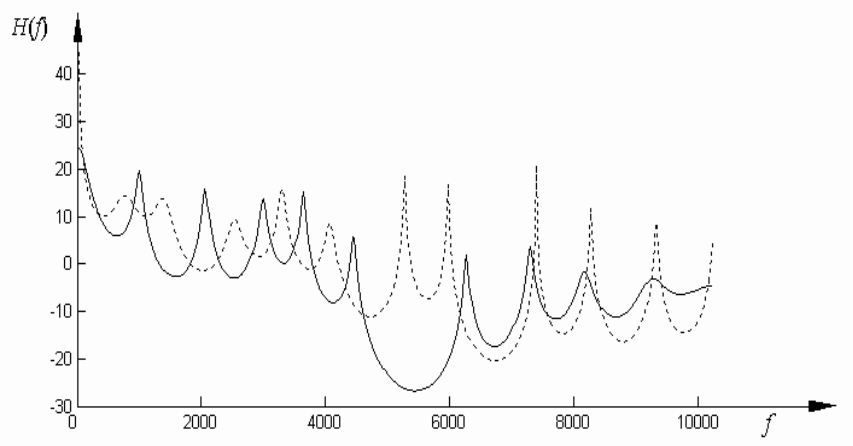

Figure 3. Example of two normalized long-term LPC spectra, prediction order is $M = 22$. The original signal was normalized by 4 average long-term LPC coefficients.

The maximal and minimal values are algorithmically easy to acquire – the algorithm just searches for the position of the local extremes. The positions of the extremes are members of two sorted sets

$$F_{\max}(l) = \left\{ f_l^{\max} \right\} = \left\{ f_0^{\max}, f_1^{\max}, f_2^{\max}, \ldots, f_L^{\max} \right\}, \quad l = 0,1,\ldots,L \tag{6}$$

$$F_{\min}(l) = \left\{ f_l^{\min} \right\} = \left\{ f_1^{\min}, f_2^{\min}, \ldots, f_I^{\min}, f_{L+1}^{\min} \right\}, \quad l = 1,2,\ldots,L,L+1 \tag{7}$$

where $f_l^{\max}$ is position (frequency) of the l-th maximum and $f_l^{\min}$ is position of the l-th minimum. Note that there is one maximum at the position $i = 0$ and one minimum at the position $l = L+1$. It is useful to merge the two sets. When the sets of frequencies were built up using the given instructions, a condition

$$f_0^{\max} < f_1^{\min} < f_1^{\max} < f_2^{\min} < \cdots < f_L^{\max} < f_{L+1}^{\min} \tag{8}$$

should be met. We can merge set $F_{\max}(l)$ and $F_{\min}(l)$ into one ordered set

$$F(i) = \left\{ f_0^{\max}, f_1^{\min}, \ldots, f_L^{\max}, f_{L+1}^{\min} \right\}, \quad i = 0,1,\ldots,I+1 \tag{9}$$

where $I = 2L$ is total number of the filters in the filter bank. Given the set $F(i)$ of $2L+2$ frequencies, we can define the filter bank itself. The filter bank consists of $I = 16$ filters (other counts are possible as well, but this count was chosen as an optimal value), we need at least $L = 8$ extremes plus one more maximum and one more minimum. Generally, the filter bank is defined as

$$H_{SDFB}(i,k) = H_{SDF}\left(k, F(i-1), F(i), F(i+1)\right), \quad i = 1,2,\ldots,I \tag{10}$$

where I is a total count of the Speaker Dependent Filters $H_{SDF}\left(k, F_{low}, F_{centre}, F_{high}\right)$ – given

H_{SDF} is one of the filters defined above – the triangular filter, the Gaussian filter, or the Tukey filter. The discrete frequency $k = 0,1,\ldots,N-1$ and N is length of the signal or frame.

In the previous filter definitions (Equation 1, 2, and 3) there is used an amplitude *A*. Two types of the Speaker Dependent Frequency Filter Banks were proposed – type I and type II. The Speaker Dependent Frequency Filter Bank of the type I (SDFFB-I) assumes the amplitude $A = 1$. We can call this type filter bank *constant amplitude filter bank*. The SDFFB of the type II (SDFFB-II) assumes the amplitudes of the individual filters equal to the values of the long-term LPC spectrum at the positions of F_{centre}.We can call this type of the filter bank *variable amplitude filter bank*. In Figure 4, can you see a comparison of three filter banks of type II (SDFFB-II).

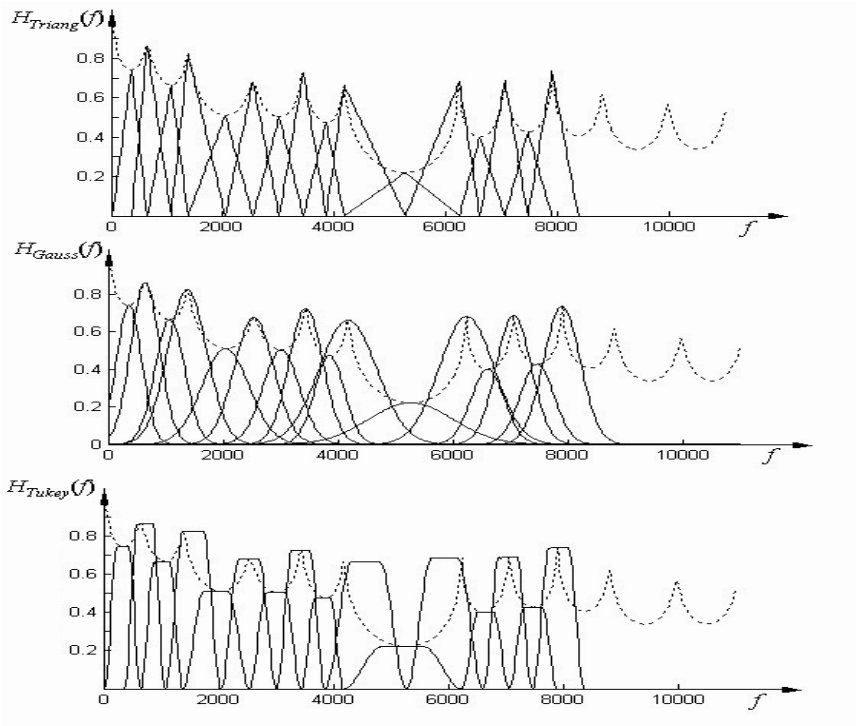

Figure 4. Comparison of filter banks of type II, with a variable amplitude *A* based on the triangular (top), Gaussian (middle), and Tukey (bottom) filter.

2.3 Speaker Dependent Frequency Cepstrum Coefficients (SDFCC)

Speaker Dependent Frequency Cepstrum Coefficients (SDFCC) are much like the mel-frequency cepstrum coefficients in terms of the general definition [3]. Nevertheless, there is an essential difference between the both of them. It differs in the filter bank used for the computation of the coefficients. The SDFCC are calculated using the SDFFB described in the previous chapter.

The process of the calculation of the SDFCC is same as in the case of the MFCC [2, 3]. The first difference is that instead of the triangular mel-frequency based filter bank $H(i,k)$, we use the speaker dependent filter bank $H_{SDFB}(i, k)$ given by the Equation 10. We can

express the log-energy at the output of the speaker dependent filters as

$$C_{SDFB}(i) = \ln\left(\sum_{k=0}^{N-1} H_{SDF}(i,k) \cdot |S(k)|^2 \right), \quad i = 1,2,\ldots,I \tag{11}$$

where I is total number of the filters in the filter bank, $S(k)$ is the a *Fourier Transformation* (FT) of the signal $s(n)$, which is N samples long. We use these coefficients to calculate the cepstral coefficients $C_{SDFB}(i)$, which results in

$$c_{SDFB}(j) = \sum_{i=0}^{I-1} C_{SDFB}(i) \cdot \cos\left(\pi m \frac{i-1}{2I} \right), \quad j = 0,1,\ldots,I \tag{12}$$

Though the SDFCC are calculated the same way as the MFCC, they are not same at all. Difference in the filter banks and filter shapes causes very different results of the speaker recognition. The SDFCC cannot be used in the universal speech recognition systems, since it emphasise the speaker influence too much. However, it can be used in a speech recognition system or in a single user speech recognition system.

3. Experimental Results

Some experiments were performed to test quality of the proposed features. For the purpose of the experiments, we must have created a voice database in order to test performance of the proposed features. The voice database used for all the performed experiments was created in cooperation with a group of the first term students of the Faculty of Information Technology. This gave us a set of voices that are very similar one to another, which makes it a very difficult environment for the algorithms.

The voice database consists of 125 speakers, two of which belonged to females and the rest of them to males. All speakers were students in the age from 19 to 21 years and all of them recorded 11 samples. Six of the utterances were stated as training samples and remaining five utterances became testing samples. These samples were used to test quality of the proposed features. 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 22050 Hz and the precision 16 bits per sample.

3.1 Feature Sets

For testing were used LPC coefficients, MFCC with delta coefficients, cepstral coefficients and the new speaker dependent coefficients proposed in this work. The first feature sets consisted of the 12 and 24 LPC coefficients [6, 7]. The following sets consisted of the MFCC, which are very widely used in the speech recognition [2, 3, 8]. These sets were 13 MFCC, 13 MFCC with 13 first order delta coefficients (MFCC+D), and 13 MFCC with 13 MFCC+D and 13 second order delta coefficients (MFCC+DD). Last common coefficients were the cepstral coefficients – 12 and 24 coefficients of the cepstrum. The common features were tested against the newly proposed features represented by a group of six SDFCC-based sets. All the SDFCC-based sets consisted of 16 coefficients. The first group of sets was calculated using the filter bank of type I with the triangular, Gaussian and the Tukey filter. The second group of the SDFCC-based sets was calculated using the filter bank of type II with the triangular, Gaussian, and the Tukey filter. The individual feature sets are referred by abbreviations given by Table 1.

Table 1. List of abbreviations of the feature sets used for the experiments.

Abbreviation	Description
LPC12	12 LPC coefficients
LPC24	24 LPC coefficients
MFCC	13 MFCC
MFCC+D	13 MFCC with 13 first order delta coefficients
MFCC+DD	13 MFCC with 13 first order and 13 second order delta coefficients
CEP12	12 cepstral coefficients
CEP24	24 cepstral coefficients
SDFCC-I-Gauss	16 SDFCC – filter bank type I, Gaussian filter
SDFCC-I-Triang	16 SDFCC – filter bank type I, triangular filter
SDFCC-I-Tukey	16 SDFCC – filter bank type I, Tukey filter
SDFCC-II-Gauss	16 SDFCC – filter bank type II, Gaussian filter
SDFCC-II-Triang	16 SDFCC – filter bank type II, triangular filter
SDFCC-II-Tukey	16 SDFCC – filter bank type II, Tukey filter

The experiments consisted of two approaches – the **speaker verification** and **speaker identification** approach to the speaker recognition [8, 9, 10]. The experiments were performed upon an HMM-GM with 3, 5, and 7 states [1, 7]. The count of states was chosen according to the given voice password. Often a HMM-GM with a lower number of states performs better than another one with a higher number of states, so in the test there are the HMM-GM with 3, 5 and 7 states. All the feature sets were tested in three categories of samples. The first one was a combination of six training samples and five unknown samples. This category shows an average accuracy of the recognition. The second category contained the six training samples. Thus, this group was supposed to be the best performer, but it was not valid for the comparison of the performance and was included for illustration purposes only. The third group contained five unknown samples, which should result in the worst performance, but it is very close to the reality, since in the real world application there are only the unknown samples provided to the system.

3.2 Speaker Verification Approach

In Figure 5, you can see comparison of the relation of the FAR (*False Acceptance Rate*) and FRR (*False Rejection Rate*) [7,10] as functions of a threshold when using the common features extracted from the combined testing samples. You can see that the curves differ for each of the feature sets. Almost ideal seems to be the curves given by the MFCC feature set, however the EER (*Equal Error Rate*, which is a measure and should be as low as possible) is rather high in comparison to the LPC12. Similarly, Figure 6 shows a comparison of FRR and FAR as a function of a threshold T, but for the SDFCC-based feature sets in this case. In all cases now, the curves are almost ideal. The curves of the FAR and the FRR for the other count of HMM-GM states and other testing samples are not present, because they are much too similar to the presented ones to show any significant difference.

Overall, the best solution was based on the SDFCC-I-Gauss and SDFCC-II-Gauss feature sets with the overall EER of 1.98 % in case of the HMM-GM with three states. These results were reached with the combined samples. The results of the unknown samples should be compared to obtain real-life-like results. Taking this in account, the winner was the HMM-

GM with three states and SDFCC-II-Gauss with the EER = 3.90 %. The number of states has only a little 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 accurately (at least when using the SDFCC based features). You can see that the LPC12 performs very well and, in case of the HMM-GM with 7 states, is even better than some of the SDFCC.

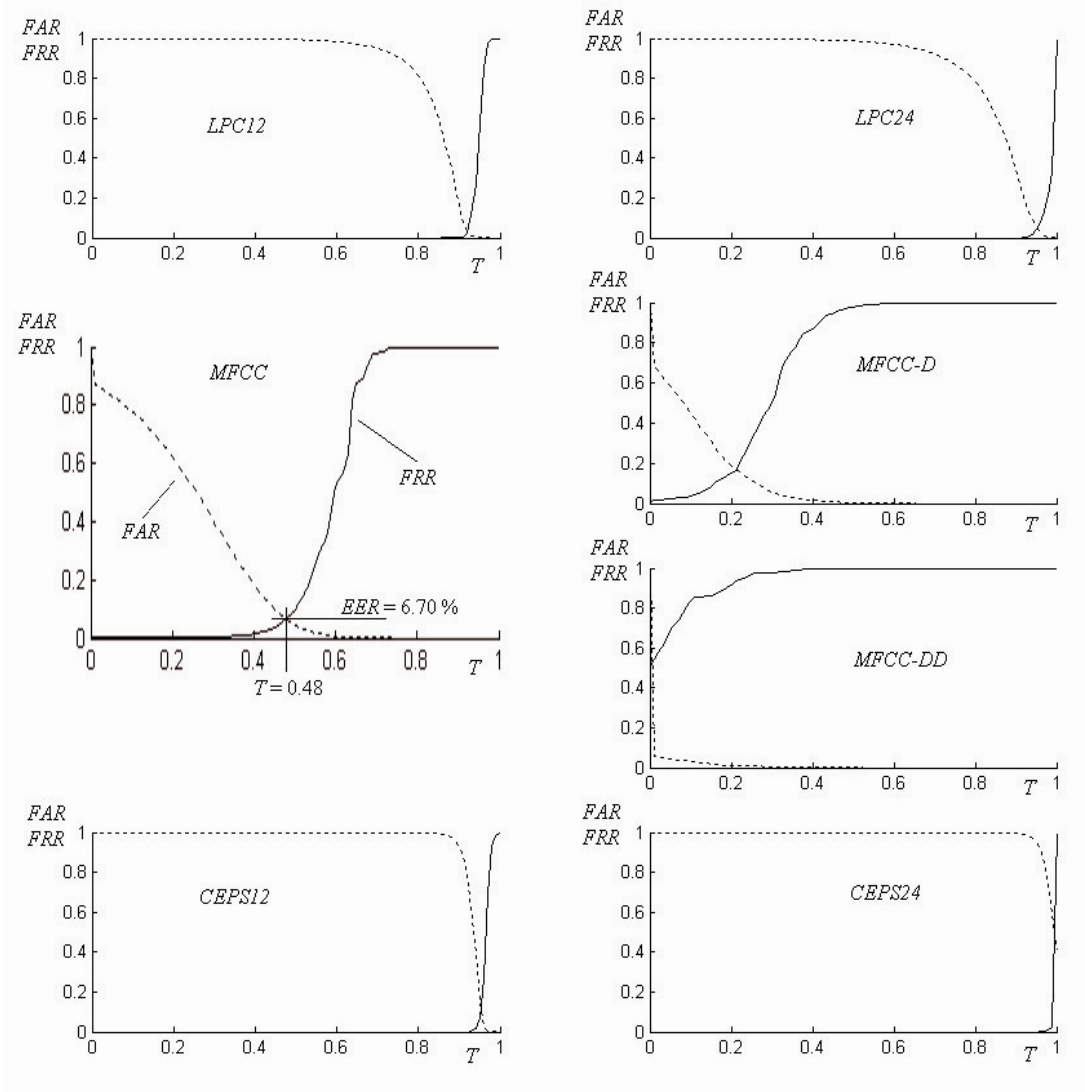

Figure 5. Speaker verification approach – comparison of the FRR (solid line) and FAR (dotted line) as a function of a threshold T using HMM-GM with 3 states and combined set of samples of LPCC, MFCC and cepstrum coefficients based feature sets.

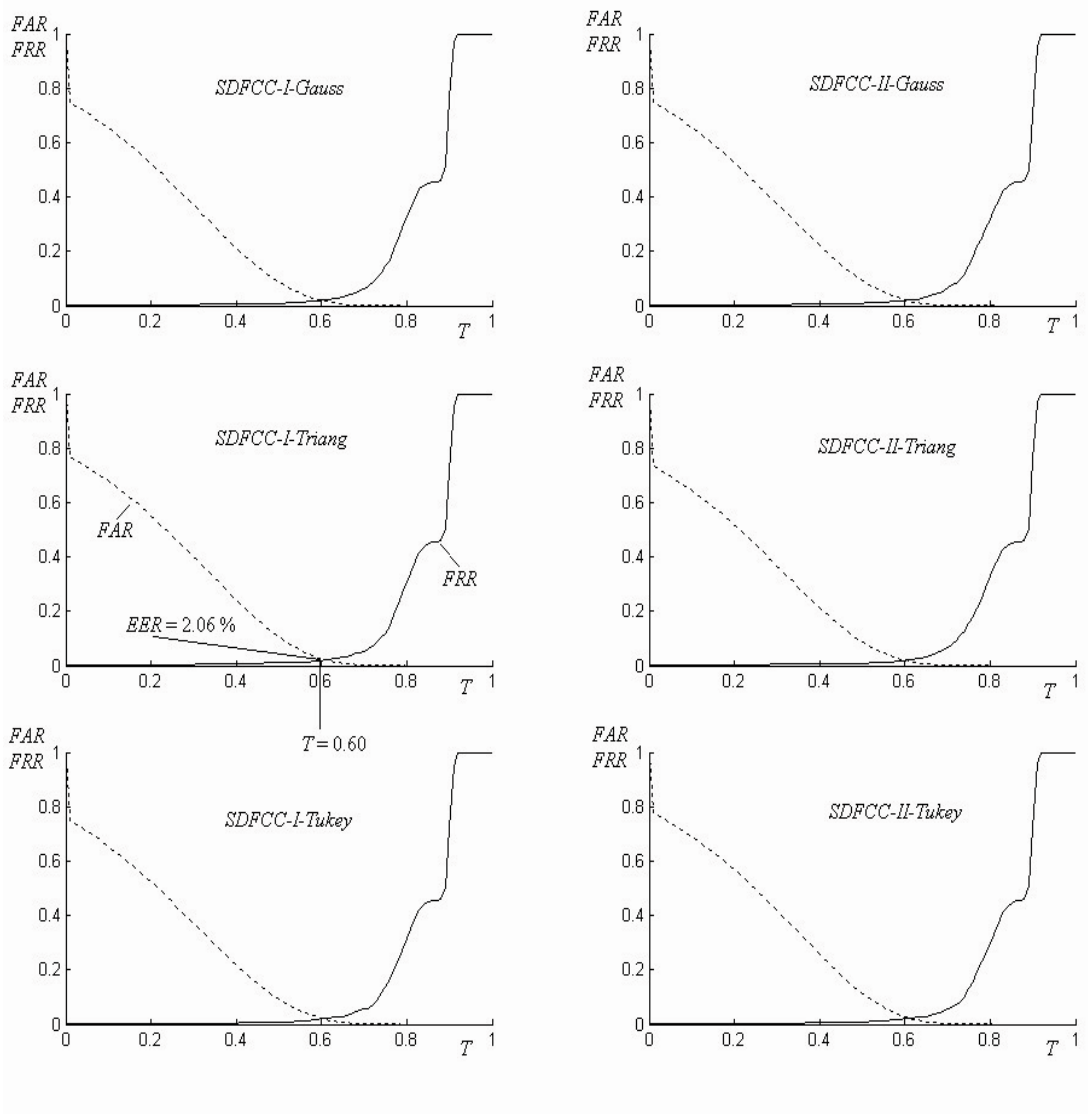

Figure 6. Speaker verification approach – comparison of FRR (solid line) and FAR (dotted line) as a function of a threshold T using HMM-GM with 3 states and combined set of samples of SDFCC coefficients-based feature sets.

In Table 2, there are summarised all the results. There are three groups of columns representing the HMM-GM with 3, 5, and 7 states, respectively. There are the EERs of the verification given in percents. Results of three types of samples are noted in the table, the first one containing combination of six training and five unknown samples, the second one containing five unknown samples, and the last one containing six training samples. The first group should be an average performer, the second one should be the worst performer, and the results of the last group should be the best ones, but they are not valid as a performance measure. The results of the second group are close to the reality, since in the real world

applications there are only the unknown samples present.

Table 2. Resume of the experimental results of the speaker verification.

Resume of the experimental results of the speaker verification									
HMM-GM states	**3**	**5**	**7**	**3**	**5**	**7**	**3**	**5**	**7**
Testing samples	**combined samples**			**unknown samples**			**training samples**		
Features	**EER [%]**	**EER [%]**	**EER [%]**	**EER [%]**	**EER [%]**	**EER [%]**	**EER [%]**	**EER [%]**	**EER [%]**
LPC12	2.88	2.97	2.86	4.80	4.95	4.78	0.53	0.55	0.56
LPC24	5.05	4.74	4.98	7.98	7.24	8.04	1.29	1.18	0.85
MFCC	6.70	11.80	12.11	11.53	18.80	19.27	1.34	4.71	5.25
MFCC+D	16.71	18.41	17.76	24.17	31.53	31.13	10.50	10.35	7.34
MFCC+DD	30.17	26.22	22.65	83.92	85.12	87.04	22.92	15.30	7.66
CEPS12	10.81	10.88	10.84	12.46	12.81	12.67	9.60	8.25	6.92
CEPS24	30.72	30.94	30.42	29.89	30.17	29.54	72.31	72.54	72.18
SDFCC-I-Gauss	**1.98**	2.25	**2.43**	4.05	4.03	4.89	**0.00**	**0.00**	**0.00**
SDFCC-I-Triang	2.06	**2.02**	2.77	4.10	**3.91**	5.00	**0.00**	**0.00**	**0.00**
SDFCC-I-Tukey	2.05	2.40	2.61	4.13	4.84	5.12	**0.00**	**0.00**	**0.00**
SDFCC-II-Gauss	**1.98**	2.27	2.68	**3.90**	4.46	5.14	**0.00**	**0.00**	**0.00**
SDFCC-II-Triang	2.12	2.51	2.77	4.13	4.57	5.13	**0.00**	**0.00**	**0.00**
SDFCC-II-Tukey	2.01	2.68	2.68	4.38	4.77	**4.80**	**0.00**	**0.00**	**0.00**

3.4. Speaker Identification Approach

The experimental results of the speaker identification are organised the same way as the experimental results of the speaker verification. In Figure 7, you can compare FAR and FRR shown as the functions of the threshold when using the common features and the combined testing samples. 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-D and the MFCC-DD feature sets, however the EER is very high compared to the values of the winning feature set – LPC12. Similarly, the same can you see in Figure 8, but for the SDFCC based feature sets in this case. FAR and the FRR for the other counts of HMM-GM states and the other testing samples are not present, because they are much too similar to the presented ones.

In all cases, the curves are nearly ideal. However, very important is the shape of the FAR curve. The FAR curve is nearly constant for the low values of the threshold. This is because of the properties of the identification process. When the threshold $T = 0$, all users should be accepted, i.e. the FAR should be equal to 1. This is true in case of the verification process, but it is not true, when using the identification approach. When recognised, the changes of the threshold cannot influence the acceptance rate, only the rejection rate 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. It is clear, that 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. That is why the values of the FAR curve for the low threshold values are almost constant.

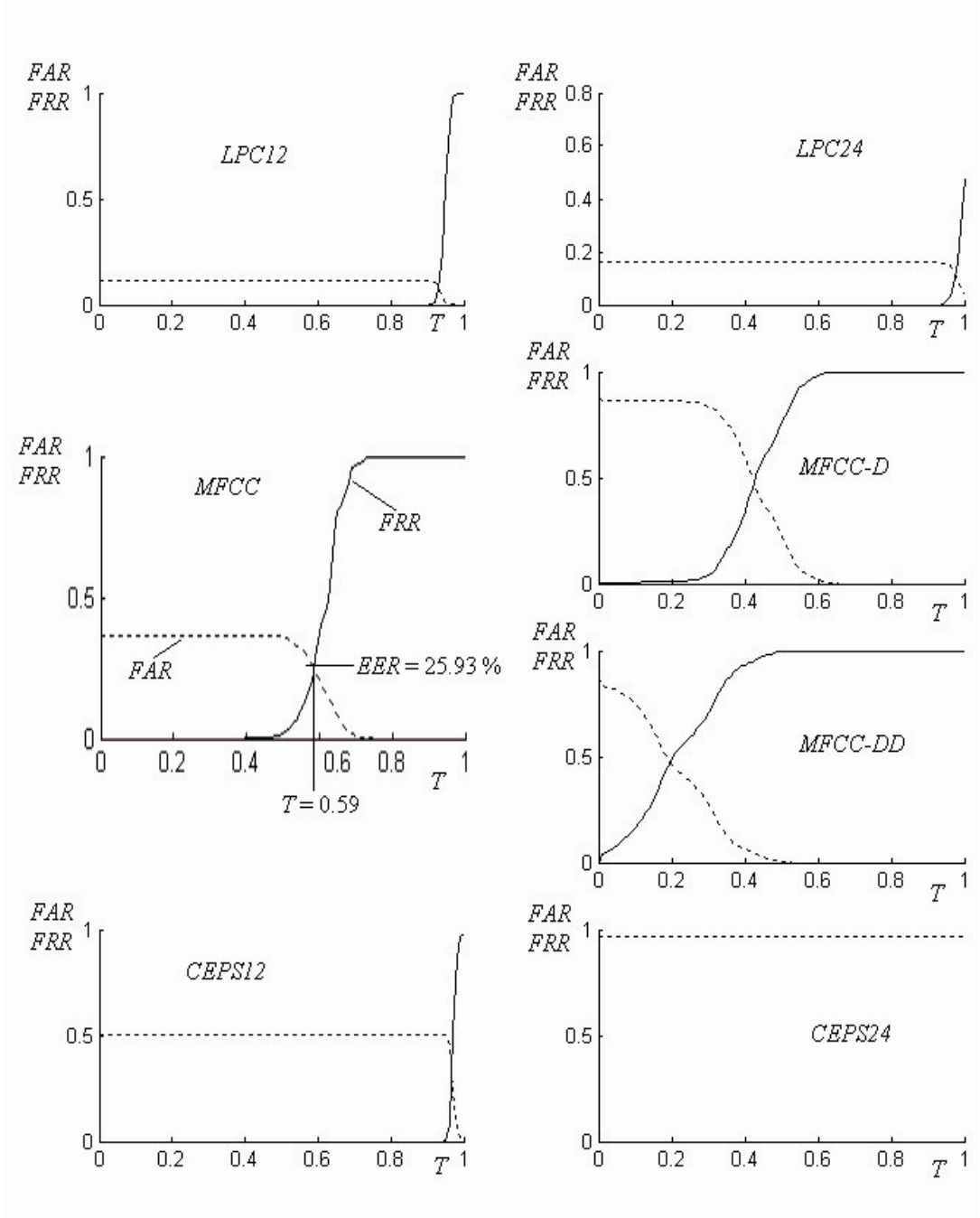

Figure 7. Speaker identification approach – comparison of the FRR (solid line) and FAR (dotted line) as a function of a threshold T using HMM-GM with 3 states and combined set of samples of LPCC, MFCC and cepstrum coefficients based feature sets.

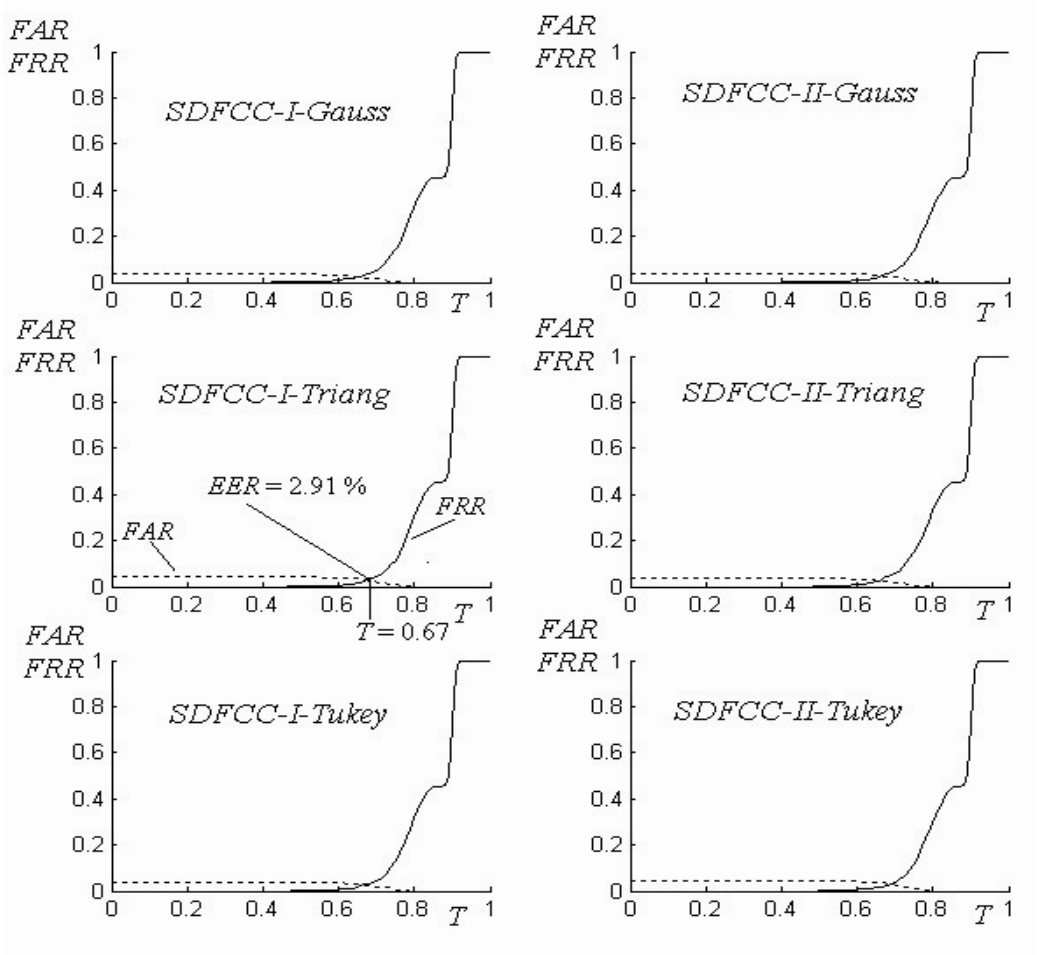

Figure 8. Speaker verification approach – comparison of FRR (solid line) and FAR (dotted line) as a function of a threshold T using HMM-GM with 3 states and combined set of samples of SDFCC coefficients-based feature sets.

In Table 3, there is a resume of the experimental results of the identification approach. It is easier to compare the influence of the number of the HMM-GM states to the equal error rate and differences among the various testing samples. The experimental results of the testing samples were expected to have the lowest values of the EER and it was true except for the CEPS24 features, which performed better in case of testing of the unknown samples than in case of testing of the training and combined samples testing. The unknown testing samples should have been the worst ones, which holds true (except for the CEPS24). The results of testing of the combination lie between the results of the other two groups (except for the CEPS24). Testing of the unknown samples is closest to real-life application behaviour. This is why these results are taken as a measure of the quality.

The EER depends on the number of the HMM-GM states. The winner of the common feature sets 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.

Table 3. Resume of the experimental results of the speaker identification.

Resume of the experimental results of the speaker identification									
HMM-GM states	**3**	**5**	**7**	**3**	**5**	**7**	**3**	**5**	**7**
Testing samples	**combined samples**			**unknown samples**			**training samples**		
Features	**EER [%]**	**EER [%]**	**EER [%]**	**EER [%]**	**EER [%]**	**EER [%]**	**EER [%]**	**EER [%]**	**EER [%]**
LPC12	7.71	7.27	7.24	16.24	15.44	15.52	0.60	0.47	0.33
LPC24	10.18	12.47	12.00	20.00	23.60	19.04	2.87	3.20	2.47
MFCC	25.93	37.71	44.40	42.64	47.12	47.84	12.00	30.93	39.13
MFCC+D	47.45	45.96	44.36	48.64	48.80	48.64	45.20	42.93	37.67
MFCC+DD	47.56	44.80	39.27	49.12	48.88	49.04	44.60	39.93	27.40
CEPS12	37.38	34.69	33.67	41.76	40.08	39.60	33.73	30.20	28.73
CEPS24	48.25	48.22	48.25	47.92	47.84	47.84	48.53	48.53	48.60
SDFCC-I-Gauss	**2.29**	3.09	2.98	**5.04**	6.80	6.56	**0.00**	**0.00**	**0.00**
SDFCC-I-Triang	2.91	2.91	**2.76**	6.40	6.40	**6.08**	**0.00**	**0.00**	**0.00**
SDFCC-I-Tukey	2.87	3.05	3.27	6.32	6.72	7.20	**0.00**	**0.00**	**0.00**
SDFCC-II-Gauss	2.62	**2.69**	2.80	5.76	**5.92**	6.16	**0.00**	**0.00**	**0.00**
SDFCC-II-Triang	2.40	3.45	2.95	5.28	7.60	6.48	**0.00**	**0.00**	**0.00**
SDFCC-II-Tukey	3.13	3.02	3.16	6.88	6.64	15.52	**0.00**	**0.00**	**0.00**

In Figure 9, there is a comparison of the cross-likelihood among a set of reference models and a set of stored models. There are 125 different speakers, thus, there are 125 different models. Each model is compared to all the other models and this is done with all stored models. The highest value should lie on the diagonal, since the reference model should be primarily similar to itself. There are compared two different samples – one of them (on the left side of the figure) is an unknown sample and one of them (on the right side) is a training sample. In the first and third row, you can see intensity images of the likelihood and, in the second and fourth row, there are corresponding 3D representations. Generally, we want to see a black diagonal line and very light grey pixels around the diagonal in the intensity images. The higher contrast in the image the higher the distance of the reference model from all other models. This is very important mainly in case of the speaker identification process. Similarly, the higher the diagonal peaks in the 3D graph the higher the distance of the reference model form the others. In the upper part (upper quaternion of images) you can see four images representing results provided by the SDFCC-I-Gauss and in the lower part there are images representing results provided by the LPC12, which are the best common features from the experiments. The diagonals in the intensity images are not clearly visible and the peaks in the 3D equivalents are not as high as in the case of the SDFCC based features. This difference shows that the SDFCC based features are better in the task of the speaker recognition.

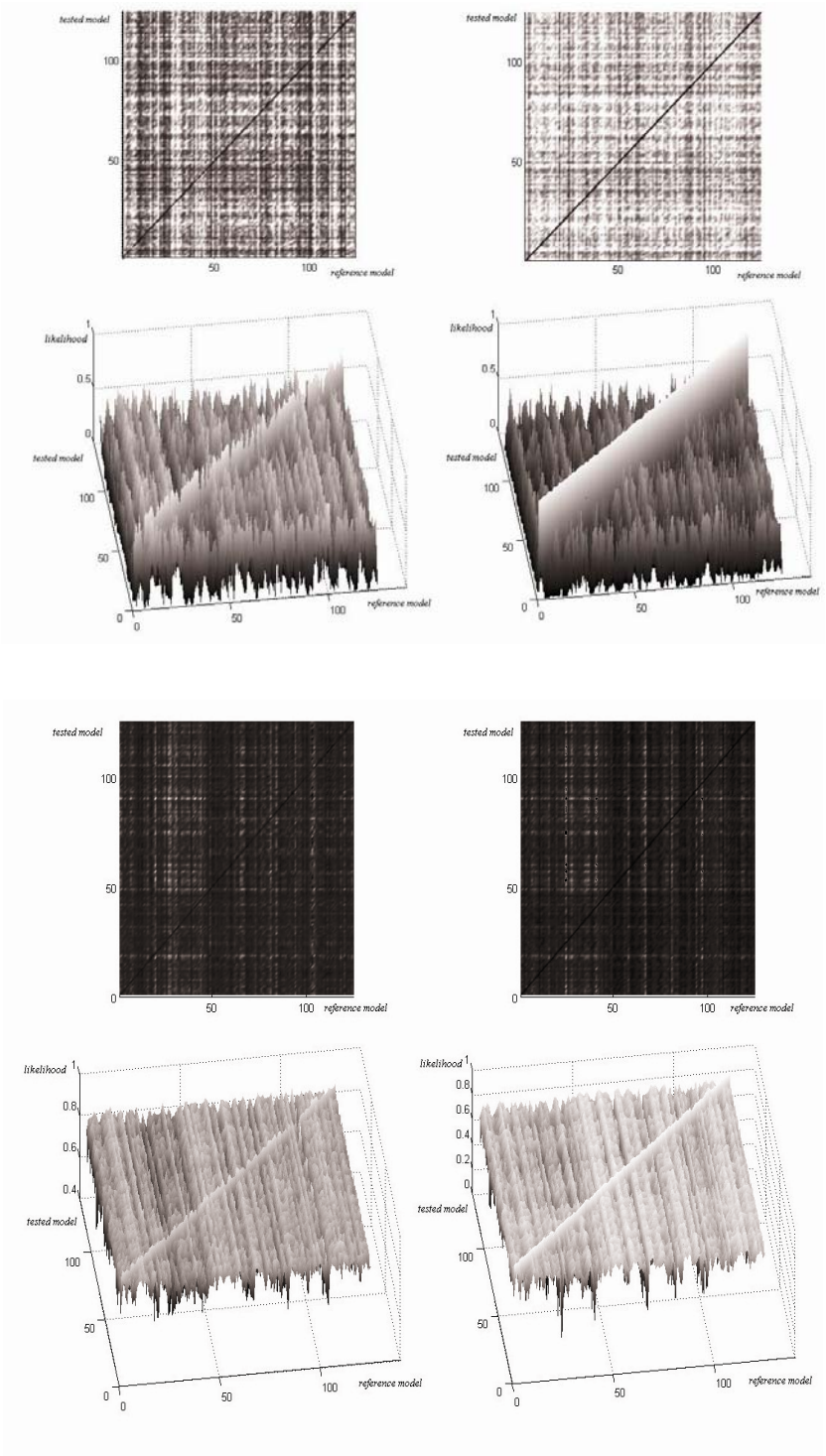

Figure 9. Cross-comparison of similarity among reference samples and other samples when using the SDFCC-I-Gauss based features (upper quaternion) and LPC12 based features (lower quaternion)

3.4. Summary

The experimental results of the speaker recognition using the verification and the identification approaches proves there is some influence of the count of the HMM-GM states on the EER. If we compare both, the speaker verification approach and the speaker identification approach we can say, generally, that the speaker verification approach performs better than the identification approach, which is clear and we had not been expecting anything else. Reason for this is given by the algorithm. Assume N models stored in the security system. When verifying, only one-to-one comparison is being done and chance of making a mistake is 1:1. When identifying, a comparison one-to-many is being done, chance of making a mistake is (N-1):1 in case the user being identified is an authorised user and N:1 otherwise, i.e. you can make a mistake (N-1)-times instead of one possible mistake when verifying.

The features suitable for both, the identification and the verification, are the SDFCC based ones. The overall EER of the identification and the verification lies in range from 3.5 % to 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. Some of them proved not to be applicable at all (CEPS24, or surprisingly the MFCC-DD). The best choice for the speaker recognition purposes proved to be the LPC coefficients. Generally, the HMM-GM recognisers with lower number of features and lower number of states performed better than the other ones.

The testing itself was very difficult for the algorithms, since we were using the same voice password for all the speakers. All the speakers were approximately at the same age, which made the task even more difficult. The quality of the SDFCC based features must be tested on a larger database, since 125 samples is not enough to draw a statistically correct conclusion. However, even with a database like this, we can show some essential qualities of the proposed features.

4. Conclusion

This work deals with the speaker recognition technology based on the HMM and a new feature set designed especially for the purposes of the speaker recognition – Speaker Dependent Frequency Cepstrum Coefficients (SDFCC). These coefficients aim at the speaker recognition and, in some special cases, are also usable for the speech recognition. The experimental results prove their qualities. Test of the speaker verification proved capability of the SDFCC for this task. The Equal Error Rate (EER), which was lower than 5% in nearly all cases with the best rate at 3.9%, is excellent taking the conditions given by voice database into account. Results of the speaker identification are not shameful too. The EER that is overall lower than 8% with the best EER f 5.04% is not bad, either. Before all, the SDFCC outperformed the common features, which aim mainly at the speech recognition. 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, which was expected and proved many times before.

The SDFC coefficients are not the only solution to the speaker dependent features. In this field, it would be possible to find more functional solutions aimed at the speaker recognition. The SDFCC can be created in many ways (think of all variables – the order of the LPC coefficients used to the long-term LPC spectrum creation, total number of maxima and minima, type of the filter used for this purpose, and even the classifier itself), which gives

many possible solutions – they only have to be experimentally verified. 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.

Acknowledgements

This research has been done under support of Ministry of Education of the Czech Republic by a grant: "Security-Oriented Research in Information Technology", MSM0021630528 (CZ).

References

[1] Baggenstoss, P.M.: *Hidden Markov Models Toolbox*. Naval Undersea Warfare Centre, Newport, RI, 2001.

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

[3] Oppenheim, A.V., Schafer, R.W., Buck, J.R.: *Discrete-Time Signal Processing*. 2nd ed., Upper Saddle River, NJ, Prentice Hall, 1999.

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

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

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

[7] Orsag, F.: *Biometric Security Systems – Speaker Recognition Technology*. Dissertation. Brno, CZ, 2004.

[8] Xafopoulos, A.: *Speaker Verification*. Tampere International Center for Signal Processing, TUT, Tampere, Finland , 2001.

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

[10] 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.