Influence of GSM coding on speaker recognition using short Polish sequences

Agnieszka Krzykowska, Tomasz Marciniak, Radosław Weychan, Adam Dąbrowski
Division of Signal Processing and Electronic Systems
Chair of Control and Systems Engineering
Poznań University of Technology
Poznań, Poland
tomasz.marciniak@put.poznan.pl

Abstract— **The paper presents a comparison of speaker models used for fast speaker identification in short recordings of telephone conversations. The knowledge of the encoder type used during the transmission of speech allows to apply a model that takes specific characteristics of the encoder into account. This improves efficiency of the speaker recognition process. The influence of the following GSM encoders was tested: FR, HR, EFR and AMR. During the experimental research we used our database of short voice phrases that usually occur during emergency calls. This paper is based on our studies related to techniques for the GSM encoding detection and to algorithms for removing silence in the voice recordings.**

Keywords-GSM, GMM, speaker identification

I. Introduction

Speaker recognition a is a new and attractive functionality, which often occurs e.g. in various types of call-center systems. An important factor influencing effectiveness of the speaker recognition (verification / identification) is e.g. the quality of transmission / recording of the speech signal. In case of the public switched telephone network (PSTN) and the typical PCM bitstream of 64 kbit/s (8-bit quantization with sampling rate equal to 8000 samples per second) the speaker verification performance is about 95 %. An intrinsic use of the speech codecs applied in mobile networks decreases efficiency of the speaker identification [1, 2]. It can currently be observed that most of the calls are performed using the mobile network. For example, in 2010, the conventional telephone network density (a number of lines per 100 inhabitants) in Poland was 21.6 only and is declining since 2004. In the same time the mobile network penetration rate (a number of SIM cards per 100 people) increased up to 123.4 [3].

The problem of building a correct model of a person to be verified by a biometric system is an issue that requires consideration of the specific acquisition and transformation of the signal. Knowing properties of the signal under test, we can compare it with patterns held in the dedicated database. A general idea of our speaker verification system is shown in Fig. 1. For each of the speakers we have developed models incorporating different types of speech coders.

Discussion of encoders used in the experimental studies is given in Chapter 2 followed by description of the detection of the encoder type, based, according to our previous studies [4], on the mel-frequency cepstral coefficient (MFCC) parameters and the mean square error (MSE).

A selection of the speaker during the verification stage is realized by means of the Gaussian mixture model (GMM) approach. The speech preprocessing includes an algorithm for automatic removal of silence in the speech signal sequence as discussed in [5].

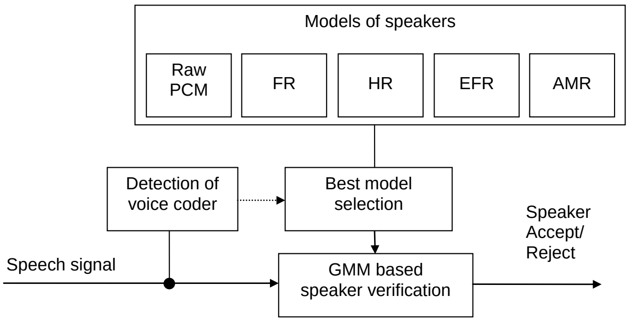

Figure 1. Speaker verification with automatic selection of speaker models

We assume that the analysis is based on the resynthesized speech. Direct application of the encoder parameters seems to be impractical according to [2], since such approach can be realized only in the systems implemented by the operators of the cellular network. Another assumption of our experiments is the use of the database of short speech utterances. Thus we assume that the detection of people should be very fast [6].

II. Speech coders

A. *GSM speech encoders overview*

As mentioned in the introduction, an important element of the speaker verification systems is selection of the proper comparison model, depending on the speech coding technique. The following encoders used in the mobile telephony were tested during our experimental studies:

- full rate (FR) encoder [7],
- enhanced full rate (EFR) encoder [8],
- half rate (HR) encoder [9],
- adaptive multi-rate (AMR) encoder [10].

These encoders can be divided into two groups, which use:

- full rate linear-prediction based on the analysis-synthesis with the RPE-LTP algorithm (*regular pulse excitation - long term prediction*) that generates a 13 kbps bitstream with 8000 samples per second sampling rate; every 40 samples are computed with the use of the previous 120 and split into frame, which consists of 160 samples, being equivalent to 20 ms of speech in the time domain
- CELP (*code excited linear prediction*) algorithm – enhanced full rate / adaptive multi-rate or half rate; these encoders use ACELP (*algebraic-code-excited linear predictive*) algorithm (both EFR and AMR) or VSELP (*vector-sum excited linear prediction*), respectively, generating bitstream of 4,75-12,2 kbps.

Implementation of AMR [11], EFR [12], and HR [13] encoders have been based on the official ANSI C code, adapted to the authors' for batch processing. An interesting fact is that the EFR implementation [12] cuts about 1024 final samples from every speech file – in this case all sequences have to be extended with additional 1024 zero values before the next transcoding. Implementation of the FR encoder has been realized with Matlab with the use of equations described in [7].

B. *SNR after GSM speech coding*

It is obvious that speech coders degrade signal quality. Errors between original and processed speech files can be estimated calculating SNR (signal to noise ratio). This value also gives us possibility to check how subsequent transcodings influence unprocessed speech. In [14] the authors proposed the following equation

$$SNR = 10\log\frac{\sum(S_{ORG})^2}{\sum(S_{ORG}-S_{CODED})^2} \quad (1)$$

This equation assumes that we can directly compare samples of the original and encoded sequence. However, it is impossible to calculate SNR in this way in case of a delay between the unprocessed and processed speech. In order to obtain proper SNR results, we have used the correlation function to find the mentioned delay and shift of the coded sequence. In [14] the authors used only linear-prediction based GSM algorithms, making SNR calculation much easier.

SNR values have been computed for our database of short sequences. It consists of 7200 speech files recorded with 22050 samples per second and downsampled to 8000 samples per second with the use of high order polyphase filters. The database has been transcoded four times by four main GSM encoders: full rate (FR), enhanced full rate (EFR), half rate (HR), and adaptive multi-rate (AMR). The last three encoders use adaptive and fixed codebooks. In this case, it is impossible to calculate SNR value even with the use of correlation function.

Fig. 2 shows the corresponding set of samples (after fitting with the correlation function) of the original speech and the transcoded once by the AMR coder. As it can be noticed, calculation of the SNR values sample by sample may give incorrect results if large differences occur between the corresponding samples. The EFR and HR encoders give very similar results. In the case of the FR encoder, which uses predictive algorithms, the corresponding transcoded frames are more similar to the original (using the correlation function gives better results than those presented in [14]).

Fig. 3 shows the transcoding effects that occur during the FR coding. The SNR values for subsequent tandems (multiple subsequent transcodings) are presented in Table I. Subsequent tandems affect the signal less and less but the SNR values decrease nonlinearly, as it was expected.

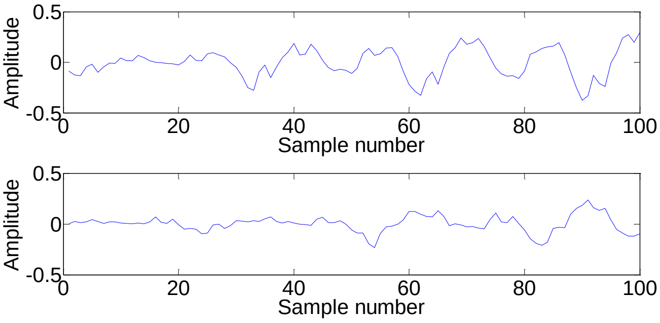

Figure 2. Corresponding subset of frames of original speech and transcoded once by AMR ecnoder

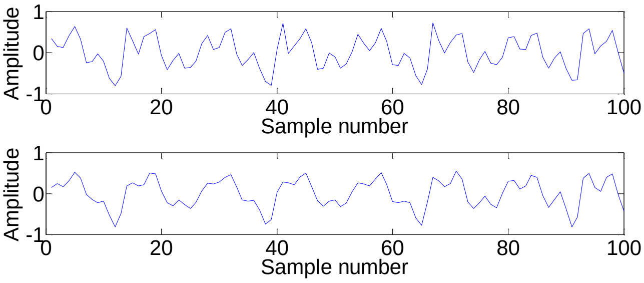

Figure 3. Corresponding subset of frames of original speech and transcoded once by FR ecnoder

TABLE I. SNR VALUES COMPUTED FOR TANDEMS IN CASE OF FULL RATE ENCODER

Encoder type	**SNR [dB]**			
	1 coding	2 codings	3 codings	4 codings
Full rate	31.7701	29.9596	28.4174	27.0173

III. ALGORITHMS AND DATABASE

A. *Speaker modeling method*

An extraction of the speaker features from a particular speech sequence starts with the division of the sampled signal into blocks of the length equivalent to 16 ms. Next, each block is multiplied by the Hamming window function, and the DFT (*discrete Fourier transform*) is calculated to finally receive 12 MFCC (*mel frequency cepstral coefficients*) by mel-scaling. Obtained feature vectors are used to train the model with the GMM. The expectation-maximization (EM) algorithm is used during this step.

B. End-point detection – removing silence from speech

Experiments described in [5, 15] proved that removing silence from speech significantly improves speaker recognition even if the modeled speech contains only several words. From four of the presented EPD (*end-point detection*) algorithms (both in their basic and extended forms) we chose two for further tests: the energy analysis and the Jang HOD (*high-order differences*) method (presented by Roger Jang in [16]). Table II summarizes the main ideas of the mentioned methods.

TABLE II. EPD METHODS USED FOR SREAKER RECOGNITION

Method	Short Description
Energy analysis	Calculation of energy
Jang HOD	Use of volume and high-order differences

In our application the described algorithms were enhanced by detection and removal of silence in the beginning, middle, and at the end of each sentence (extended version), therefore we are calling them "middle energy" and "middle Jang HOD" algorithms.

1) Middle energy algorithm

This EPD algorithm is based on the analysis of the signal energy, which can be computed from equation

$$E_i = \sum_{n=k_{ip}}^{k_{ik}} x^2(n) \tag{2}$$

where *i* stands for the number of the window of the signal *x*, k_{ip} is the first sample of the *i*-th window, and k_{ik} is the last one. Number of samples in one window used to count the energy is 80, and the length of each window is 0.01 s. The offset of the window is equal to 0.001 s.

Selection of threshold parameters were tested on wave sequence: "Chciałbym zgłosić wypadek" (in English: "I would like to report an accident") which has 16 383 samples and lasts for 2.05 seconds. A mean value of the energy in this wave is 1.7578 with the minimal value very close to zero (0.00006) and the maximum value of 15.384. The energy threshold was chosen experimentally and is equal to 0.1. After removing the silence from the tested sentence the length of raw speech decreased to 1.04 second i.e. to 8335 samples. Mean energies of the voiced and silence parts are 3.6 and 0.016, respectively.

2) Middle Jang HOD algorithm

This end-point detection algorithm uses high-order differences of the speech signal to detect speech. Individual steps of the algorithm are as follows:

a) Computing volume (*V*) with the use of equation (3) and the absolute value of the sum of the *j*-order difference (*H*) with equation (4):

$$V_i = \sum_{n=k_{ip}}^{k_{ik}} \left| x(n) \right| \tag{3}$$

$$H_i = \sum_{n=k_{ip}}^{k_{ik}} \left| \frac{\Delta^j x(n)}{\Delta n^j} \right| , \tag{4}$$

where *x* describes the signal, *n* – the sample number, *i* stands for the number of the time window, k_{ip} and k_{ik} are the first and the last sample of the *i*-th window, respectively, Δ represents the differential. Next, the values of *V* and *H* are normalized.

b) Computing the *VH* curve from volume (*V*) and sum of obtained differentials (*H*):

$$VH = (V + H)/2 \tag{5}$$

c) Computing threshold t for *VH* in order to determine the end-points. The threshold is defined as

$$t = VH_{\min} + (VH_{\max} - VH_{\min}) \times r \ . \tag{6}$$

A default value of *r* = 0.125. A length of the time window to compute the volume is 0,016 s. A number of samples in a window is 128. Adjacent windows are not overlapping.

A number of differences (in our case *j* = 4) was chosen experimentally for the sentence mentioned in the description of the previous method. In this example the mean value of *VH* was 0.1757, while the range of values fluctuated between 0.0015 to 0.9611. The threshold value computed from equation (6) equals 0.0922. In this case the length of the voiced signal was 0.75 second i.e. 6022 samples. The mean *VH* values of the speech samples and the removed silence are 0.3762 and 0.0274, respectively.

C. Database

The database (described in [5]) used in our experiment consists of 6 sentences repeated 30 times by each of 40 speakers (males and females between 20 and 55 years of age). An average length of the recorded sentences is 1.6 second.

Recordings were realized during three sessions. Every speaker repeated each sentence 10 times at once. Time break between sessions was 1 to 6 weeks.

Each of 7200 samples was recorded in an anechoic chamber with the use of an omnidirectional condenser microphone. The sampling rate of the recorded samples was set to 22050 samples/second and to 16 bit resolution. Next, the recorded phrases were downsampled to 8000 samples/second to compare with the telephone quality speech. An average duration of speech sentences is about 1 s, thus during the experiment the features obtained from 5 random sequences were combined to create a model of each speaker.

IV. INFLUENCE OF GSM CODERS ON SPEAKER RECOGNITION

To determine how each coder influences the speaker recognition accuracy, we tested four GSM coders (AMR, EFR, FR, and HR) and the unprocessed (raw) speech in matched conditions. This means that the speaker model and the speaker sentence used for tests were taken from the same database, thus coded with the same coder.

As it can be seen in Fig. 4, coding of speech decreases accuracy of the speaker recognition, which was expected. However the results show that the FR coder apparently extracts some crucial speaker features because effectiveness of the speaker recognition system for speech encoded with this coder is increased in comparison to raw (unprocessed) speech. Other coders in the order from the best to the worst result are: EFR, AMR, and HR. EER (*equal error rate*) values for this experiment are presented in Table III in bold.

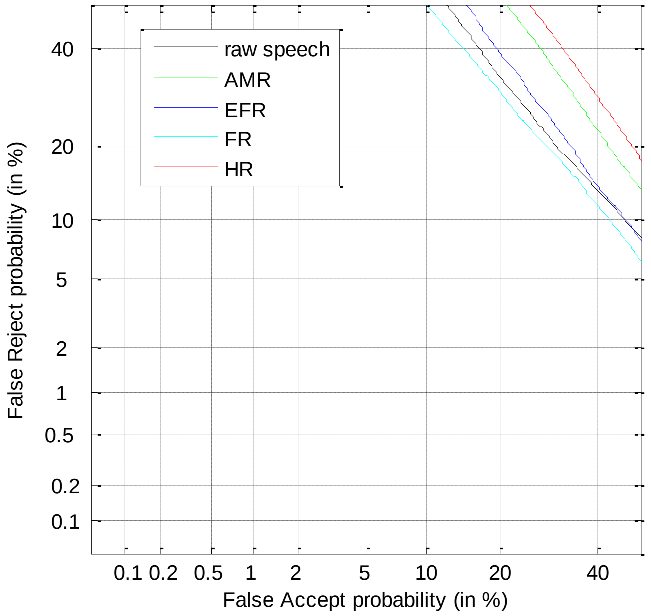

Figure 4. Influence of GSM coding on speaker recognition for matched conditions

V. Influence of Tandeming and Silence Removal on Speaker Recognition

In the next step of our experiment we tested the effect of multiple coding and decoding of speech on speaker recognition. For that, we used speaker models coded once with one of the coders, and test sequences coded one and four times with each coder.

Furthermore, we tested the influence of silence removal algorithms on already transcoded speech. Two methods described earlier were used and the effects are described below. Tables III to V show the results of the performed experiments both in matched and in mismatched conditions. Gray highlight shows the lowest value of EER for a given test sequence.

A. Tandeming

Every cellular phone call comes through the base station and the base station controller. There can be several switches between the sender and the receiver. Every operation of coding and decoding has impact on the signal quality. An idea of tandeming (transcoding the speech several times) is to check how it influences the speaker recognition accuracy.

Table III shows EER values for matched and mismatched conditions for four GSM coders. Speech samples used for the test and the training part were transcoded 1 and 4 times. It can be observed that in cases of AMR, EFR, and HR encoders EER value increases with the number of tandeming. An interesting fact is that for the FR encoder, recognition accuracy is higher than for the raw data even when transcoded 4 times. It can be observed in the highlighted fields in Table III, whose structure tends to be diagonal.

TABLE III. EER of Speaker Recognition for Transcoded Speech (in %)

Test / Model	Raw speech	Speech transcoded 1 time				Speech transcoded 4 times			
		AMR	EFR	FR	HR	AMR	EFR	FR	HR
Raw speech	**25.7**	30.6	26.3	24.7	35.2	38.9	37.5	24.3	43.6
AMR	33.9	**33.0**	29.8	31.7	35.5	39.4	39.2	30.9	42.2
EFR	31.2	31.0	**27.3**	28.9	33.7	39.0	38.2	28.3	42.2
FR	27.2	30.4	26.5	**24.4**	34.0	37.9	35.5	23.2	43.4
HR	38.2	36.4	34.3	35.4	**35.3**	40.1	40.7	34.7	40.1

Fig. 5 presents plotted lines of values included in Table III. The abscissa axis is described with the tested sequence transcoded by a specific encoder, while the ordinate axis shows the EER value. Colors of the plots correspond to the used models. As it can be observed, the full rate encoder gives better results even than the raw data.

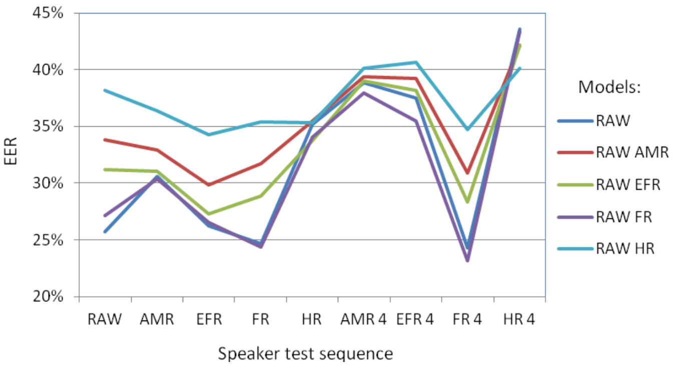

Figure 5. EER of speaker recognition system for transcoded speech

B. End-point detection

As it can be inferred from comparing Tables III, IV, and V removing silence has a significant influence on the speaker recognition system. Even for raw speech the results are better (an improvement from 26% to 17% and 16% for the middle energy and the middle Jang HOD algorithms, respectively). Though, both algorithms give quite similar accuracy.

1) Middle energy algorithm

Table IV shows results for matched and mismatched conditions with the use of four GSM coders.

Removing silence, both from the model and the tested speech, improved the recognition efficiency. The best results were obtained for the test speech coded only once with any coder and the model, which was adequate to the coder.

In case of tandeming for the tested speech coded with the EFR coder, the best results were acquired for the speaker models coded with the AMR and EFR coder. When voice of the verified speaker were coded 4 times with the FR coder the best accuracy can be observed for the speaker modeled after coding with the FR encoder. However, good results are also received for the unprocessed speech and the EFR model.

TABLE IV. EER OF SPEAKER RECOGNITION FOR TRANSCODED SPEECH WITH REMOVED SILENCE BY MIDDLE ENERGY METHOD (IN %)

Test / Model	Raw speech	Speech transcoded 1 time				Speech transcoded 4 times			
		AMR	**EFR**	**FR**	**HR**	**AMR**	**EFR**	**FR**	**HR**
Raw speech	17.1	18.9	17.8	18.7	27.5	29.3	22.3	19.3	44.2
AMR	19.7	18.1	17.7	19.7	23.3	24.5	18.7	20.9	43.3
EFR	18.3	17.7	16.6	18.3	23.4	26.7	19.3	19.3	43.1
FR	19.0	19.9	18.6	16.9	26.0	31.2	23.9	17.3	44.6
HR	25.5	23.0	22.8	24.6	20.7	24.6	20.9	25.1	34.2

Fig. 6 illustrates the obtained EER values. Designation ME stands for the model processed with the middle energy algorithm, and each line represents speaker model tested against the speaker test sequence listed on the horizontal axis.

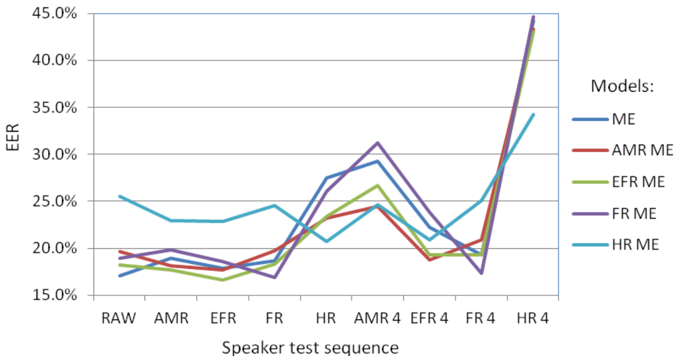

Figure 6. EER of speaker recognition system for transcoded speech with removed silence with use of middle energy method

Every model (apart from the one coded with the HR coder) gives similar results for testing speech coded only once. HR coded test speech stands out here (the only exception is when the model is coded also with the HR encoder). For testing the speech transcoded 4 times with the HR coder all results are worse, though the HR coded model is better in this case.

2) *Middle Jang HOD algorithm*

As it can be inferred from Table V and Fig. 7, the middle Jang HOD algorithm, used for removing unvoiced parts of speech, shows significant improvement of the speaker recognition accuracy. It brings a very similar effect to that of the previously studied middle energy method. The differences between them are almost unnoticeable.

It can be also observed that in case when the tested speech is transcoded 4 times with the HR coder no EPD algorithm can improve recognition accuracy (except when the model is also coded with the HR encoder).

As in the previous method, the best results can be observed when the model and the speech of the verified speaker are coded with the same coder (this conclusion is certainly valid in both cases of 1 and of 4 transcodings).

In Fig. 7 the abbreviated designation MJH stands for the speaker model processed with the middle Jang HOD algorithm. The corresponding graph shows recognition accuracy (EER) in reference to the transcoded test speech sequence.

TABLE V. EER OF SPEAKER RECOGNITION FOR TRANSCODED SPEECH WITH REMOVED SILENCE BY MIDDLE JANG HOD METHOD (IN %)

Test / Model	Raw speech	Speech transcoded 1 time				Speech transcoded 4 times			
		AMR	**EFR**	**FR**	**HR**	**AMR**	**EFR**	**FR**	**HR**
Raw speech	16.4	18.7	17.5	18.6	27.5	29.9	22.5	19.7	44.7
AMR	19.1	17.7	17.5	19.4	24.1	25.5	19.4	20.8	42.4
EFR	17.9	17.6	16.5	18.2	24.0	27.2	19.8	19.7	44.0
FR	18.8	20.0	18.7	16.2	26.2	31.6	24.0	16.9	44.2
HR	25.5	22.7	22.5	23.9	20.6	25.7	21.4	25.2	34.1

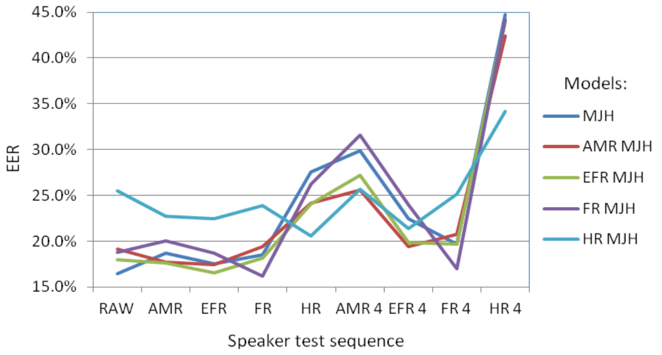

Figure 7. EER of speaker recognition system for transcoded speech with removed silence with use of middle Jang HOD method

VI. TIME OF PROCESSING

Presented experiments were performed in the Matlab environment on the Linux operating system. The hardware used for tests involved Xenon Quad Core E5405 2.0 GHz CPU and 2 GB RAM.

Table VI and Fig. 8 present average time of computations for each part of the speaker recognition system, i.e.: silence removal, feature extraction, creation of model, and verification of one particular speaker with one model. The time unit is milliseconds.

TABLE VI. PROCESSING TIME (IN [MS])

	Raw speech	Speech with removed silence	
		Middle energy	Middle Jang HOD
Silence removal	---	54.0	236.2
Feature extraction	2.7	1.7	1.6
Model creation	181.7	65.6	61.4
Verification of speaker	1.8	1.3	1.2
Overall	**186.2**	**122.6**	**300.4**

The tested algorithms are of various complexities, thus the time consumption of their performance is adequate – removing silence with middle Jang HOD algorithm takes more time than computation of the signal energy.

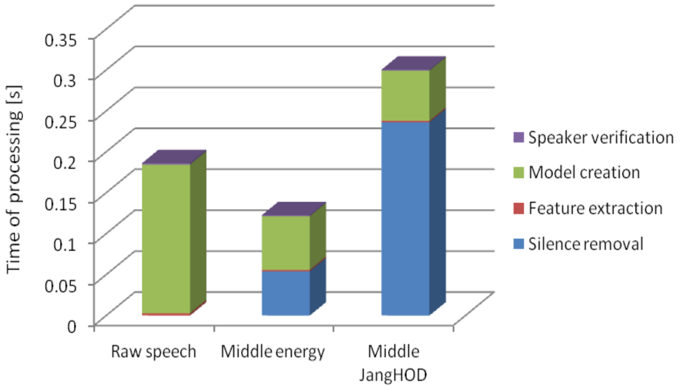

Figure 8. Time of processing raw speech and speech with removed silence

Even though detection of silence is time consuming, speech containing only voiced parts enables to create the speaker model faster (because of the much smaller amount of the samples left). Therefore, in case of the middle energy algorithm the sum of the time spent on detecting silence and creating the speaker model is shorter than for the unprocessed speech. Only when the middle Jang HOD method is used the computational time is longer.

VII. CONCLUSIONS

The article shows that selection of the appropriate model significantly improves verification of the speaker. In general, the studied verification is the most plausible when the base models of the speakers include the FR encoder parameters. Such a situation occurs if the EPD algorithms are not used.

Removal of silence improves verification, but it is closely linked with the dedicated model. Removing of silence from the transcoded speech leaves the voiced speech parts only, which is significant for coding features of the particular speaker. The computation time for cutting silence by the middle energy method is shorter than for the unprocessed speech. Thus, applying this algorithm greatly enhances the speaker recognition correctness and can be used for creating better speaker identification algorithms.

REFERENCES

[1] Vuppala, A.K.; Sreenivasa Rao, K.; Chakrabarti, S.; , "Effect of speech coding on speaker identification," *India Conference (INDICON), 2010 Annual IEEE* , pp.1-4, 17-19 Dec. 2010

[2] Grassi, S.; Besacier, L.; Dufaux, A.; Ansorge, M.; Pellandini, F., "Influence of GSM Speech Coding on the Performance of Text-Independent Speaker Recognition", Proc. Of European Signal Processing Conference (EUSIPCO) 2000, pp. 437-440, Tampere, Finland,September 4-8, 2000

[3] COMMUNICATION - ACTIVITY RESULTS IN 2010, Central Statistical Office, Warsaw, Poland, 2011

[4] Dabrowski, A.; Drgas, S.; Marciniak, T.; , "Detection of GSM speech coding for telephone call classification and automatic speaker recognition," *Signals and Electronic Systems, 2008. ICSES '08. International Conference on* , pp.415-418, 14-17 Sept. 2008

[5] Dąbrowski, A.; Marciniak, T.; Krzykowska, A.; Weychan, R., "Influence of silence removal on speaker recognition based on short Polish sequences", *Proc. Of SIGNAL PROCESSING SPA'2011*, Poland Section, Chapter Circuits and Systems IEEE, pp. 159-163, September, 29-30th 2011

[6] Marciniak, T.; Weychan, R.; Drgas, S.; Dabrowski, A.; Krzykowska, A.; "Speaker recognition based on short polish sequences," *Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference Proceedings (SPA), 2010* , pp. 95-98, 23-25 Sept. 2010

[7] FR description http://www.3gpp.org/ftp/Specs/archive/06_series/06.10/0610-820.zip

[8] EFR description http://www.3gpp.org/ftp/Specs/archive/06_series/06.60/0660-801.zip

[9] HR description http://www.3gpp.org/ftp/Specs/archive/06_series/06.20/0620-801.zip

[10] AMR description http://www.3gpp.org/ftp/Specs/archive/06_series/06.90/0690-721.zip

[11] AMR ANSI C source http://www.3gpp.org/ftp/Specs/html-info/26073.htm

[12] HR ANSI C source http://www.3gpp.org/ftp/Specs/html-info/46006.htm

[13] EFR ANSI C source http://www.3gpp.org/ftp/Specs/html-info/46053.htm

[14] Lilly, B.T.; Paliwal, K.K.;, "Effect of speech coders on speech recognition performance," *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on* , vol.4, pp. 2344-2347 vol.4, 3-6 Oct 1996

[15] Marciniak, T.; Krzykowska, A.; Weychan, R., "Speaker recognition based on telephone quality short Polish sequences with removed silence", *Przegląd Elektrotechniczny*, vol. 06/2012, pp. 42-46, June 2012

[16] Jyh-Shing Roger Jang, "Audio and Signal Processing and Recognition", available at the links for on-line courses at the author's homepage at http://www.cs.nthu.edu.tw/~jang

This work was prepared within the INDECT and DS projects.