Improving of speaker identification from mobile telephone calls

Radosław Weychan, Agnieszka Stankiewicz,
Tomasz Marciniak, and Adam Dabrowski

Poznan University of Technology, Chair of Control and Systems Engineering,
Division of Signal Processing and Electronic Systems,
ul. Piotrowo 3a, 60-965 Poznań, Poland
{radoslaw.weychan,agnieszka.stankiewicz,
tomasz.marciniak,adam.dabrowski}@put.poznan.pl
http://www.dsp.org.pl

Abstract. The paper examines issues related to proper selection of models used for quick speaker recognition based on short recordings of mobile telephone conversations. A knowledge of the encoder type used during the transmission of speech allows to apply an appropriate model that takes specific characteristics of the encoder into account: full rate (FR), half rate (HR), enhanced full rate (EFR) and adaptive multi-rate (AMR). We analyse both proper model selection and automatic silence removal. Analysis of time of processing is also a part of this study.

Keywords: speaker identification, GSM, Gaussian mixture models

1 Introduction

Automatic speaker recognition with contemporary computer systems is an attractive functionality, which can be used e.g. in various types of call-center systems to verify identity of speakers. In case of the public switched telephone network (PSTN) and the typical pulse-code modulation (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 % [10]. An inseparable use of the speech codecs applied in mobile networks decreases efficiency of the speaker identification [7, 8, 13]. Direct application of the encoder parameters seems to be impractical according to [7], since such approach can be realized only in the systems implemented by the operators of the cellular network.

In case of crossing the signal through several base stations and base station controllers in the cellular network the GSM-type coding of speech may be applied to the signal multiple times. Such occurrence is referred in this article as the effect of tandeming on speech signal.

Based on our research [8] we propose speaker verification system presented in Fig. 1. First step of signal processing is automatic removal of silence in the speech signal based on voice-activity detection (VAD) algorithm as discussed in [3, 10]. For each of the speakers we have developed models incorporating different types

of speech coders. The model that is the best match for identification process can be used in the phase of speaker verification. The modeling algorithm used during this stage is the Gaussian mixture model (GMM).

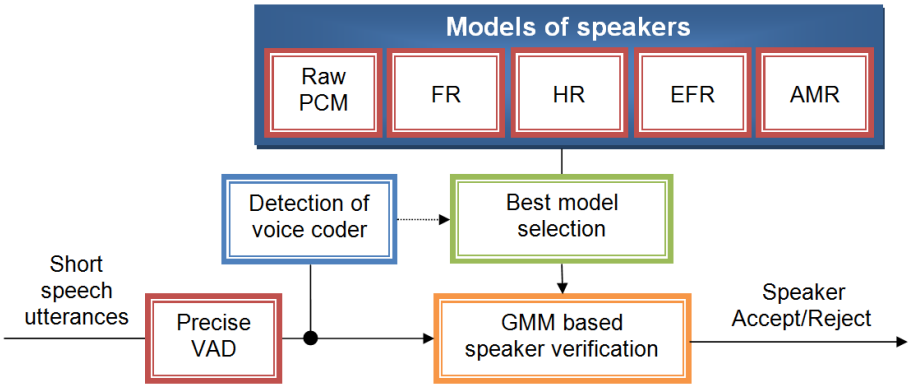

Fig. 1. Speaker verification with automatic selection of speaker models

Short discussion of encoders used in the experimental studies is given in [8]. Detection of the encoder type based on the mel–frequency cepstral coefficient (MFCC) parameters and the mean square error (MSE) [2]. In [11] an extended 85 dimensional feature vector was proposed to detect the encoder type, which includes also LPC coefficients, energy per frame, zero-crossing rate, Hilbert envelope, its variance and dynamic range, importance weighted signal to noise ratio (ISNR), pitch period estimate and estimate of the long term acoustic channel by blind channel estimation (BCE) algorithm. A classification and regression tree (CART) method was used. Results obtained in [2, 11] show that detection of the transcoded speech, encoder type and even the bitrate can be used as the first step of the GSM-based speaker recognition.

In our application two end-point detection (EPD) algorithms were used for VAD and futher for silence removal. Applied method are: energy analysis and Jang HOD algorithm, both analyzed in [8]. In this study we propose the enhanced verions of those method, by detection and removal of silence in the beginning, middle, and at the end of each sentence. Therefore we are calling them “middle energy” and “middle Jang HOD” algorithms.

2 Related work

Research on the analysis of the effects of speech coding on the quality of speech and speaker recognition are conducted for about 20 years. Such publications began to appear primarily at the development of the mobile communication techniques.

In the paper [9] it is shown that the speech signal coding reduces quality of the speech recognition (only the GSM full rate (FR) coder was investigated and the speaker recognition was not tested). An important issue is the analysis of multiple encoding (called tandeming). While in case of the adaptive differential pulse-code modulation (ADPCM) the encoders do not affect the recognition effectiveness, the GSM codecs reduce speech recognition performance significantly.

The tests related to matched and mismatched conditions can be found in paper [5]. The experiments were performed using the Gaussian mixture model – universal background model (GMM-UBM) speaker verification. Identification of the speakers was studied for the speech coded with G.729 (8 kbps), GSM (12.2 kbps), and G.723.1 (5.3 kbps) codecs. No experiments related to the half rate (HR), enhanced full rate (EFR), and adaptive multi-rate (AMR) encoders were made.

However, the above-mentioned encoders, i.e., FR, HR and EFR (except AMR) have been tested in the paper [7]. During this research the TIMIT database and its 8 ksps downsampled version was used.

An importance of the fitted model is shown in the paper [12]. Its authors compared linear predictive cepstral coefficients (LPCC), mel-frequency cepstral coefficients (MFCC), real cepstral coefficients (RCC). Results for the PCM training and the matched training only, were reported.

Studies on the influence of speech coding on the speaker recognition, reported in the literature, are so far quite limited. For example in the paper [4] the database ARADIGIT was tested only the matched conditions were only examined.

3 TIMIT Database

During the experiments the standard TIMIT database [6] was used. It contains recordings of the speech signal of 630 speakers presenting eight main dialects of English, each of them uttering 10 sequences. This gives 6300 speech files in summary. The average time of each sentence is about 3 s. This database is used primarily to test the efficiency of speech recognition algorithms. The sequences of speech were recorded with a resolution of 16 bits and sampling rate 16000 samples/second. In our experiments, the sampling rate was converted to the value of 8000 samples/second with the use of the same processing technique as the previous one.

In order to achieve the proper transcoding with the use of the GSM encoders, all files have been normalized to 0.9 of their maximum amplitude.

4 SNR after GSM speech coding

The degradation of signal quality can be measured by calculating the signal to noise ratio (SNR). In [9] the authors proposed a method that used only linear-prediction based GSM algorithms, making the SNR calculation much easier. We applied this coefficient for short utterances as described in [8].

The SNR values have been computed for the TIMIT database. This 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.

It can be noticed that the 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 ones (the use of the correlation function gives better results than those presented in [9]).

Fig. 2 presents the transcoding effects that occur during the FR coding. The SNR values for subsequent tandems (multiple subsequent transcodings) are presented in Table 1 and Fig. 3, which contains SNR values for subsequent tandemings in case of the TIMIT database. Subsequent tandems affect the signal less and less but the SNR values decrease nonlinearly. The transcoding operation most degrades the speech in the case of AMR encoder, and for FR coder / decoder the SNR coefficient decreases the least, as could be expected. For GSM encoders that use CELP algorithm, the length of the file seems to be very important in case of comparison between subsequent recordings.

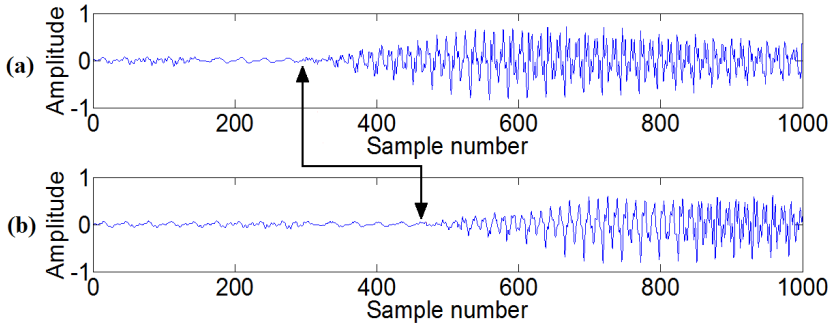

Fig. 2. Subset of frames of original speech (a) and once transcoded by FR encoder (b) (black line with arrows describes a delay between the encoder input and output — about 160 samples/20 ms)

5 Speaker recognition from GSM-coded speech

In order to determine the influence of GSM coders on speaker recognition accuracy, we tested four GSM coders (AMR, EFR, FR, and HR) and the unprocessed by EPD (raw) speech in the 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.

Table 1. SNR values computed for tandems in case of TIMIT database

Encoder type	SNR [dB]			
	1 coding	2 codings	3 codings	4 codings
FR	27.211	27.1127	27.0195	26.8455
AMR	22.3673	19.2768	17.1698	15.6038
EFR	34.7623	31.5058	29.5246	28.3428
HR	25.2108	23.6957	22.6965	21.9322

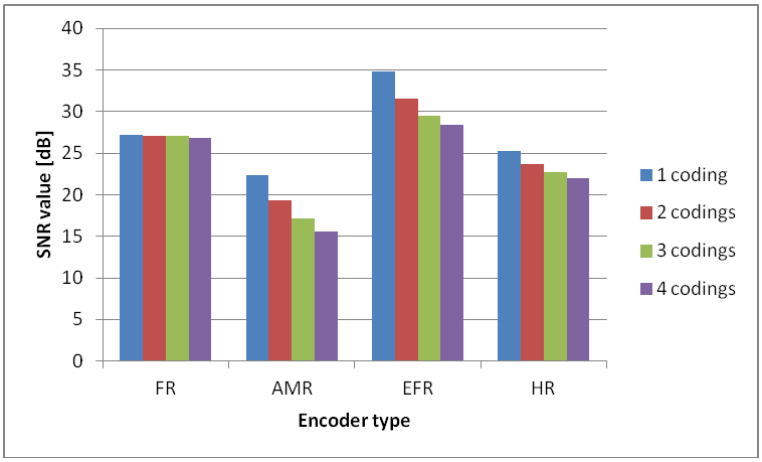

Fig. 3. SNR values computed in tandems for the TIMIT database

As it can be seen in Fig. 4 and 5, coding of speech decreases accuracy of the speaker recognition, which was expected. The coders in the order from the best to the worst result are: FR, EFR, AMR, and HR. The equal error rate (EER) values for this experiment are presented in Table 2 in bold.

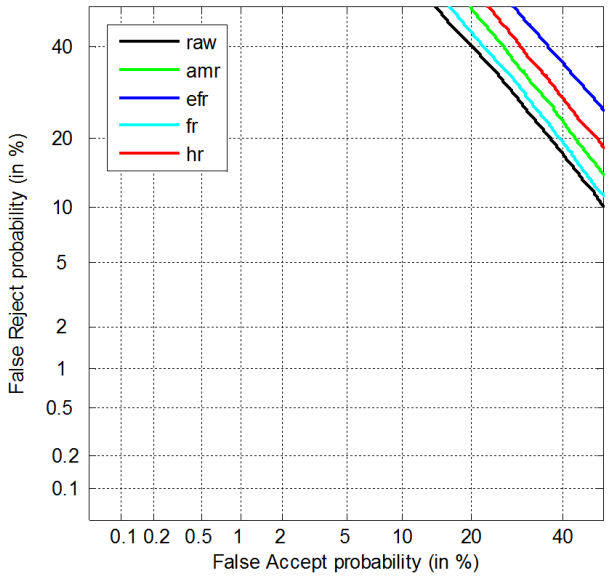

Fig. 4. Influence of GSM coding on speaker recognition for matched conditions in case of the TIMIT database

6 Analysis of tandeming effect

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 an 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 2 show 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.

In case of TIMIT database, for single transcoding two models give the best result: raw speech and FR encoded. For four transcodings, the proper choice of model is much more significant to increase speaker recognition accuracy.

Fig. 5 present bar chart of values included in Table 2. The abscissa axis is described with the tested sequence transcoded by a specific encoder, while the

Table 2. EER of speaker recognition for transcoded speech (in %) in case of TIMIT database

Test vs. Model	Raw speech	Speech transcoded 1 time				Speech transcoded 4 times			
		AMR	EFR	FR	HR	AMR	EFR	FR	HR
Raw speech	**28.6**	31.1	34.3	32.9	35.0	37.3	40.6	34.5	41.9
AMR	32.6	**31.9**	38.0	33.2	34.6	37.3	44.6	34.7	41.4
EFR	35.2	36.0	**37.9**	35.8	37.9	39.5	41.7	36.1	42.4
FR	34.6	34.0	35.9	**30.0**	33.9	38.1	42.4	30.3	41.4
HR	38.0	36.3	41.2	35.4	**34.4**	39.1	45.8	35.7	40.2

ordinate axis shows the EER value. Colors of the bars correspond to the used models. As it can be observed, the full rate encoder gives better results even than the raw data in case of TIMIT database.

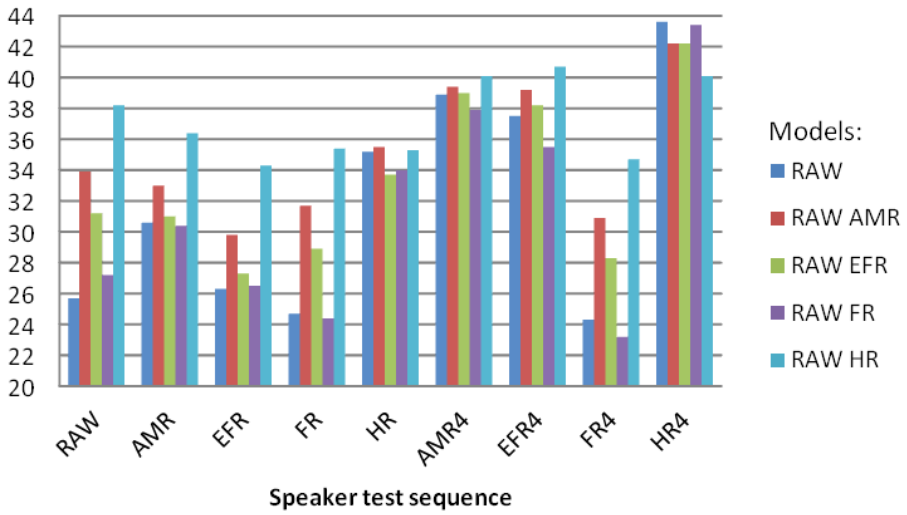

Fig. 5. EER of speaker recognition system for transcoded speech in case of TIMIT database

7 Analysis of silence removal

A comparison of the Tables 3 and 4 shows that removing silence has a significant influence on the speaker recognition system. In case of TIMIT database, the result is better (up to 6 %), but amount of speakers uttering various sentences once and different recording conditions make EER values higher.

7.1 Middle energy algorithm

Table 3 show results for matched and mismatched conditions with the use of four GSM coders and middle energy algorithm.

Table 3. EER of speaker recognition for transcoded speech with removed silence by middle energy method (in %) in case of TIMIT database

Test	Raw	Speech transcoded 1 time				Speech transcoded 4 times			
vs. Model	speech	AMR	EFR	FR	HR	AMR	EFR	FR	HR
Raw speech	29.9	31.2	30.9	35.0	36.2	36.5	32.9	37.3	43.3
AMR	31.4	30.2	29.8	32.2	32.5	32.4	30.3	34.5	39.5
EFR	32.7	31.0	29.9	31.6	32.3	33.6	30.8	33.5	40.2
FR	35.7	34.1	32.2	29.2	32.5	36.9	33.7	29.5	40.4
HR	37.8	35.3	33.9	32.8	32.0	34.7	32.8	33.3	36.1

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 regardless of coder and model applied.

In case of TIMIT database, best EER values trend to be diagonal. Almost the same results for AMR and EFR encoders are dictated by used type of CELP algorithm. In both cases, single and four times transcodings, speech encoded with FR encoder as a model and for test stage give best results.

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

Testing the TIMIT database, model speech encoded with encoder related to test stage gives a result which can be better distinguished from any other in case of single transcoding.

7.2 Middle Jang HOD algorithm

It can be observed based on Table 4 and Fig. 7, that 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).

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

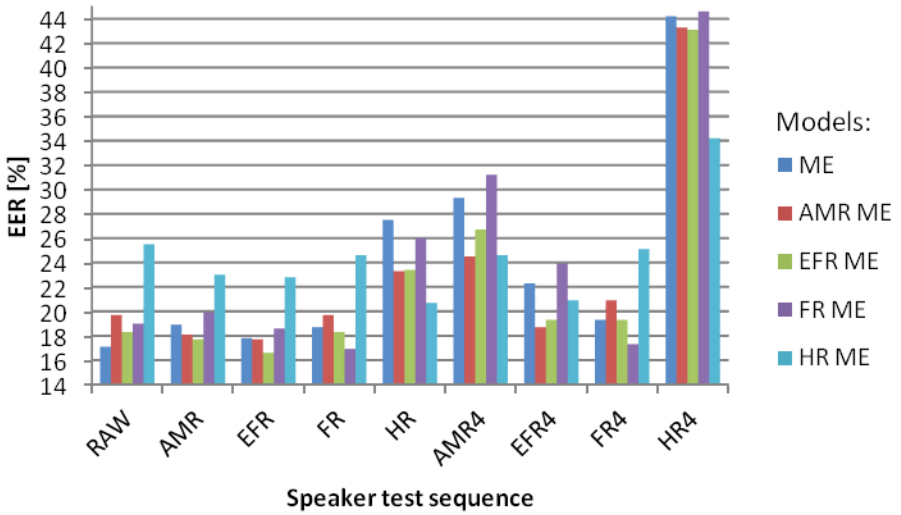

Fig. 6. EER of speaker recognition system for transcoded speech with removed silence with use of middle energy method in case of TIMIT database

Table 4. EER of speaker recognition for transcoded speech with removed silence by Jang HOD method (in %) for TIMIT database

Test vs. Model	Raw speech	Speech transcoded 1 time				Speech transcoded 4 times			
		AMR	EFR	FR	HR	AMR	EFR	FR	HR
Raw speech	30.8	31.7	31.7	35.1	35.9	35.7	32.3	36.5	42.2
AMR	32.9	31.4	30.9	33.5	32.9	32.7	30.5	35.2	39.8
EFR	33.5	31.7	30.7	32.3	32.4	33.4	30.8	34.1	39.9
FR	37.0	35.2	33.6	30.6	33.1	37.3	34.1	30.5	40.5
HR	38.4	36.1	34.9	34.1	32.9	35.1	33.2	34.4	36.2

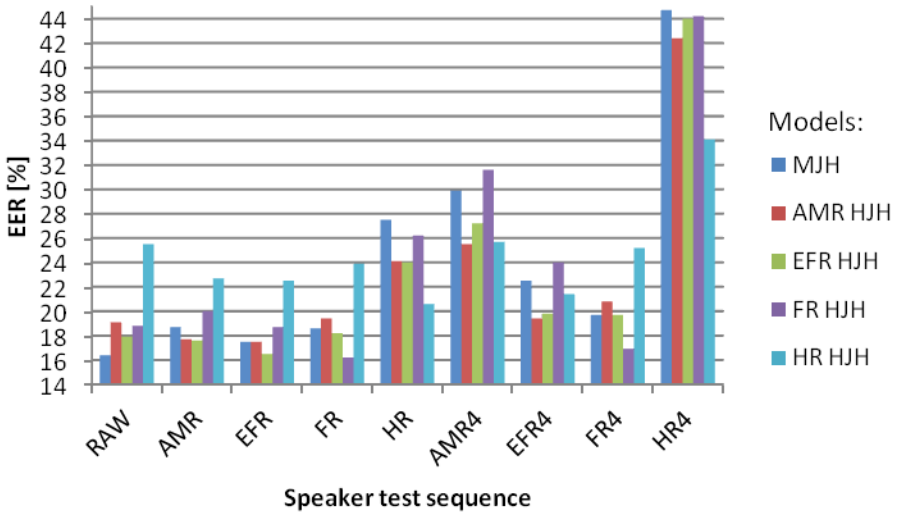

Fig. 7. EER of speaker recognition system for transcoded speech with removed silence with use of middle Jang HOD method in case of TIMIT database

8 Time of processing

The average times of computations for each part of the speaker recognition system are shown in Table 5 and Fig. 8. The tested stages are: silence removal, feature extraction, creation of model, and verification of one particular speaker with one model.

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

Even though detection of silence is time consuming, speech recordings containing only voiced parts enables to create the speaker model faster (because of 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. When the middle Jang HOD method is used the computational time is longer.

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

9 Conclusions

In this paper we analysed speaker verification based on short utterances of phone conversations. We examined model selection for FR, HR, EFR and AMR encoderd and effect of tandeming. Results shows that selection of the appropriate

Table 5. Time of processing of TIMIT database (in [ms])

	Raw speech	Speech with removed silence	
		Middle energy	Middle Jang HOD
Silence removal	—	27.5	158.4
Feature extraction	5.1	2.8	2.8
Model creation	71	42.5	42.1
Verification of speaker	2.2	1.3	1.3
Overall	78.3	74.1	204.6

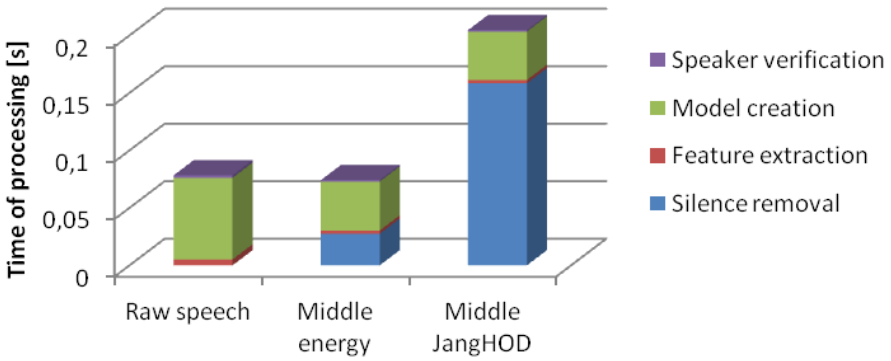

Fig. 8. Time of processing raw speech and speech with removed silence in case of TIMIT database

model improves verification of the speaker (up to 6 % of EER). Verification with the use of silence removal algorithms improves results up to 9–16 % of the EER, but it is closely linked with the dedicated model. The computation time for cutting silence by the middle energy method is shorter than for the unprocessed speech. The low complexity of proposed system – voice activity detection, encoder detection and also simplified GMM algorithm, give the possibility of an implementation in embedded systems, which are not able to work with the use of effective but complex and time-consuming algorithm like Alize [1].

References

1. Bonastre, J.F., Wils, F., Meignier, S.: Alize, a free toolkit for speaker recognition. In: Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on. Volume 1. (March 2005) 737–740
2. Dabrowski, A., Drgas, S., Marciniak, T.: Detection of gsm speech coding for telephone call classification and automatic speaker recognition. ICSES '08. International Conference on Signals and Electronic Systems (Sept. 14–17 2008) 415–418
3. Dabrowski, A., Marciniak, T., Krzykowska, A., Weychan, R.: Influence of silence removal on speaker recognition based on short polish sequences. Proc. Of SIGNAL PROCESSING SPA (September, 29-30th 2011) 159–163
4. Debyeche, M., Krobba, A., Amrouche, A.: Effect of gsm speech coding on the performance of speaker recognition system. (2010) 137–140
5. Dunn, R.B., Quatieri, T.F., Reynolds, D.A., Campbell, J.P.: Speaker recognition from coded speech in matched and mismatched conditions. A Speaker Odyssey-The Speaker Recognition Workshop (2001)
6. Garofolo, J.S., et al.: Timit acoustic-phonetic continuous speech corpus (1993) Linguistic Data Consortium, Philadelphia, http://catalog.ldc.upenn.edu/LDC93S1.
7. 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) (September 4–8 2000) 437–440
8. Krzykowska, A., Marciniak, T., Weychan, R., Dabrowski, A.: Influence of gsm coding on speaker recognition using short polish sequences. Proc. Of SIGNAL PROCESSING SPA'2012 (2012) 197–202
9. Lilly, B., Paliwal, K.: Effect of speech coders on speech recognition performance. ICSLP 96. Fourth International Conference on Spoken Language Proceedings **4** (Oct 3-6 1996) 2344–2347
10. Reynolds, D.: Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Trans. Speech Audio Proc. **3**(1) (1995) 72–83
11. Sharma, D., Naylor, P., Gaubitch, N., Brookes, M.: Non intrusive codec identification algorithm. Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on (2012) 4477–4480 doi 10.1109/ICASSP.2012.6288914.
12. Vuppala, A.K., Sreenivasa Rao, K., Chakrabarti, S.: Effect of speech coding on speaker identification. India Conference (INDICON), 2010 Annual IEEE (2010) 1–4
13. Vuppala, A., Sreenivasa Rao, K., Chakrabarti, S.: Effect of speech coding on speaker identification. Annual IEEE India Conference (INDICON) (2010) 1–4