www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

PAN ARCHIVES OF ACOUSTICS
Vol. **40**, No. 3, pp. 407–417 (2015)

DOI: 10.1515/aoa-2015-0042

The Effect of Voice over IP Transmission Degradations on MAP-EM-GMM Speaker Verification Performance

Waldemar MACIEJKO

Forensic Bureau, Internal Security Agency
Rakowiecka 2A, 00-993 Warsaw, Poland; e-mail: w.maciejko@abw.gov.pl

(*received February 19, 2015; accepted August 7, 2015*)

Despite the growing importance of packet switching systems, there is still a shortage of thorough analyses of VoIP transmission effect on speech and speaker recognition performance. Voice over IP transmission systems use packet switching. There is no guarantee of delivery. The main disadvantage of VoIP is a packet loss which has a major impact on the performance experienced by the users of the network. There are several techniques to mask the effects of a packet loss, referred to as packet loss concealment.

In this study, the effect of voice transmission over IP on automatic speaker verification system performance was investigated. The analyzed system was based on MAP-EM-GMM modelling methods. Four various speech codecs of H.323 standard were investigated with special emphasis placed on the packet loss phenomenon and various packet loss concealment techniques.

Keywords: automatic speaker verification, packet loss, speech compression, voice over IP.

1. Introduction

The popularity of automatic speaker recognition as one of the methods of biometric human identification, is constantly growing. The natural consumer of this type of technology is a bank sector. The voice biometrics is becoming supplemental to traditional security methods like password authentication. Another potential beneficiary of voice biometrics are forensic sciences. Forensic speaker identification experts may use not only aural-perceptual methods but also unbiased parametric analysis based on automatic verification (Rose, 2002; Maciejko, 2012).

VoIP technology is an effect of a widespread access to the internet and continued development of IP technologies, including wireless communication systems based on UMTS/HSPA, LTE and LTE Advanced standards. The most important advantages of VoIP are: lower costs of telephone conversations and possibility of parallel non-audio data transmission. According to telecommunication market predictions, there will be 348.5 million VoIP users by the end of 2020 and the global VoIP services market will generate revenue worth US$136.76 billion (Transparency Market Research, 2014). These market data show that the role of VoIP is still growing.

The effect of GSM and PSTN transmission degradation on speaker verification performance was a subject of many previous studies. The majority of investigations focused at evaluating the influence of additive noise like white and colored noise as well as the influence of non-linear spectral distortions (Reynolds *et al.*, 1995; Reynolds, 1996). The results showed that this type of distortions caused the degradation of speaker verification performance. Other researchers evaluated the influence of audio GSM codecs such as GSM 06.10, GSM 06.20 and GSM 06.60 (Besacier *et al.*, 2000; Byrne *et al.*, 2004). All GSM and PSTN speech transmission phenomena lead to the decline of the speaker verification performance. The degree of degradation depends on transmission technology.

There is a number of reports on automatic speaker verification systems under voice over IP conditions (Besacier *et al.*, 2004; Jelassi, Rubino, 2011), but there are no available reports describing experiments with different codecs in connection with various packet loss rate and different packet loss concealment methods. Some experiments on the automatic speaker identification system for forensic purposes under packet switching conditions have been recently carried out by this author (Maciejko, 2014). The present study investigated the effect of factors related to voice trans-

mission over packets switching system on automatic speaker verification system. The investigated factors include: the packet loss rate, the packet loss concealment methods, and the different voice coding methods of the ITU-T H.323 standard.

2. Voice transmission over IP

Each telecommunication system makes up a collection of transmission and switching devices. The older telephone systems, like public switched telephone network, used the circuit switching in which two network nodes established a communication channel. When the connection was established, the full bandwidth was guaranteed and the communication path remained busy for the entire session duration (Jajszczyk, 2009).

Alternatively, the VoIP systems use the packet switching. The transmission unit in a packet-switching network is a data block called packet (Peinado, Segura, 2006). VoIP provides the capability to break up speech signal into small pieces (known as samples) and place them into packets which circulate between the terminals as a series of bits (Davidson, Peters, 2000). The packets consist of two parts: the header and the body. The header contains, among others, the control information such as terminals' IP, the number of packets into which the message has been divided and synchronization information. The body is a part comprising the coded and divided speech signal. Each packet is individually transmitted through the network. Unlike in the case of circuit switching, the communication path remains busy only during packet transmission. When the packet transmission is finished, the path becomes immediately accessible (Jajszczyk, 2009).

2.1. Packet loss and packet loss concealment

The IP transmission may cause many different packet errors. Among others, the packets can be lost, damaged or delayed. Most of the IP traffic is under control of TCP protocol, which provides solution to packet retransmission. However, the VoIP technology uses UDP protocol that doesn't provide any recovery method. Figure 1 presents the effect of packet loss.

The investigation of packet loss phenomenon reveals that voice packets are lost in bursts. The packet loss is described by a two-state simple Gilbert model defined by Sanneck (2000) and Jelassi and Rubino (2011), which allows using only two parameters to describe the loss process and provides good approximation of this phenomenon (Mohamed *et al.*; 2004, Staroniewicz, 2006). One of the states (state B) represents a packet loss and the other state (state G) represents the case where packets are correctly transmitted. Figure 2 presents example of Gilbert model based on Gilbert (1960).

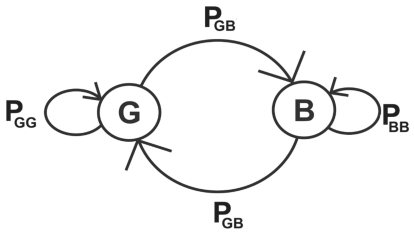

Fig. 2. Gilbert model (source: self elaboration based on Sanneck (2000)).

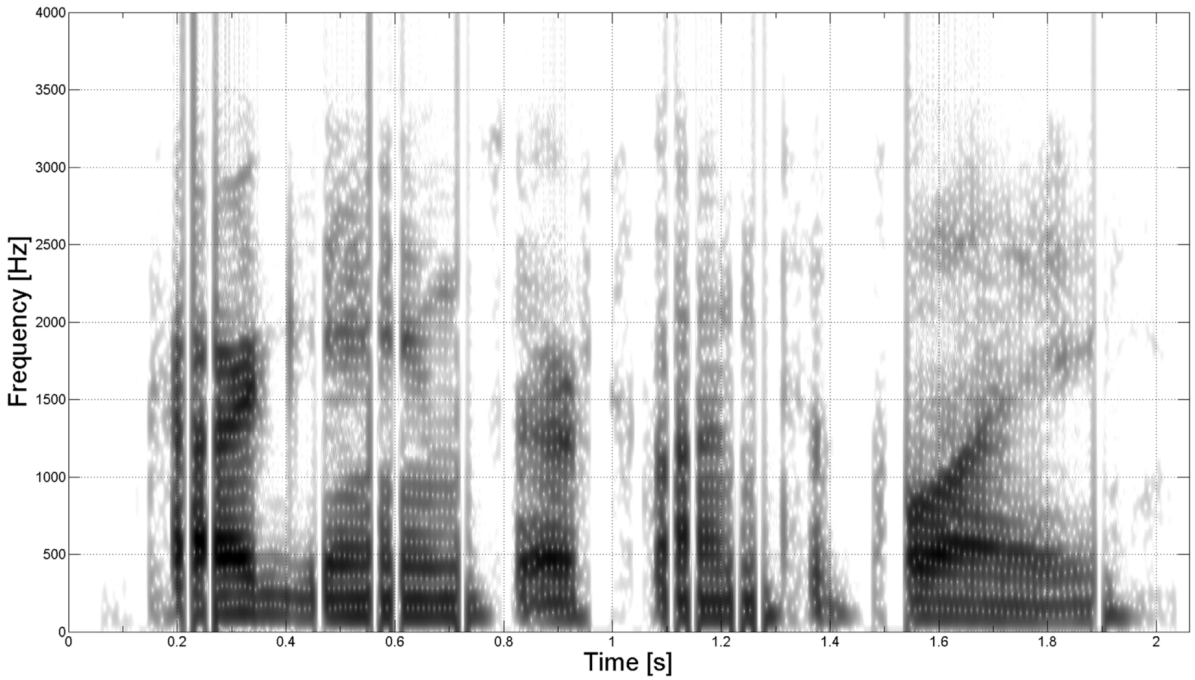

Fig. 1. The utterance "car games are fun to play", audio was compressed by ITU-T G.729 6.4 kbps, probability of packets loss equals 25%, packet duration equals 20 ms (source: self elaboration).

The probability P_{BB} stands for *conditional loss probability*. *Mean burst loss length* (D_{burst}) based on Sanneck (2000) is computed as:

$$D_{\text{burst}} = \frac{1}{1 - P_{BB}} . \tag{1}$$

Finally, based on Eq. (1), P_{BB} can be computed as shown below (Sanneck, 2000; Jelassi, Rubino, 2011):

$$P_{BB} = 1 - \frac{1}{D_{\text{burst}}} . \tag{2}$$

The probability of being in state B, representing the mean loss, is denoted as *unconditional loss probability* (P_Z):

$$P_Z = P_G P_{GB} + P_B P_{BB} , \tag{3}$$

where P_G and P_B are the stationary probabilities $P_G + P_B = 1$ and $P_Z = P_B$. Consequently, P_{GB} can be expressed as:

$$P_{GB} = \frac{P_Z(1 - P_{BB})}{1 - P_Z} . \tag{4}$$

Gilbert model based on Eqs. (4) and (2) is described by Mohamed *et al.*, (2004):

$$P_{GB} = \frac{P_Z}{D_{\text{burst}}(1 - P_Z)} . \tag{5}$$

The packets loss is a serious problem in signal quality assessment. Therefore, numerous techniques of packets loss concealment (PLC) have been developed. Some of them are shortly explained below and presented in Fig. 3:

- repetition – the lost packet is replaced by a copy of last received packet (Mayorga, 2003),
- simple linear interpolation – the lost packet is replaced by last received packet and connection point is the averaged of the two surrounding values so the connection between the last before the lost and the first after the lost packets is linearly interpolated (Peinado, Segura, 2006; IETF, 2004),
- white noise – the lost packet is replaced by the white noise.

In practical applications, there are many other solutions, some of them being much more complex. In this paper, three PLC methods presented in Fig. 3 are being investigated.

2.2. Speech signal compression

The final effect of speech signal compression (quality and compression ratio) depends on applied codec. Usually, higher compression causes higher quality loss and allows using more narrow band to voice transmission.

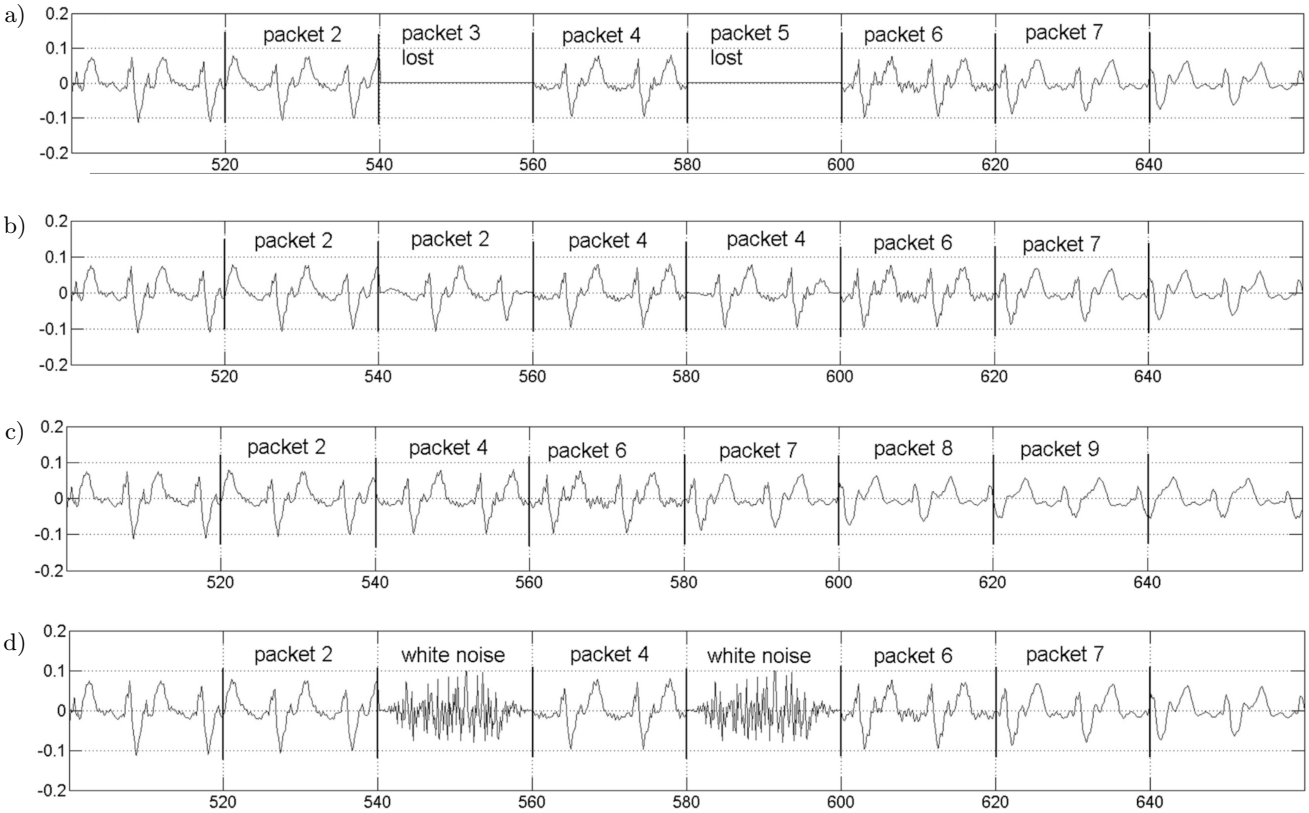

Fig. 3. Three alternative methods of packets loss concealment: a) packets loss effect without concealment, b) packets concealment by repetition, c) packets loss concealment by simple interpolation, d) concealment by replaced by white noise (source: self elaboration based on (Mayorga, 2003)).

The first standard of audio-video transmission by IP network was defined by H.323 (Recommendation ITU-T H.323, 2009). This standard uses many audio codecs, of which: G.711 (codec with the highest bit rate), G.729 (codec with medium bit rate) and G.723.1 (codec with the lowest bit rate) were investigated in this study and are summarized in Table 1.

Table 1. Selected H.323 standard audio coders (Recommendation ITU-T H.323, 2009).

Standard	Compression algorithm	Frame (ms)	Compression ratio	Bit rate kbps
G.711	PCM	0.125	1:1	64.0
G.729	+CS-ACELP	10	9:1	6.4
G.729	CS-ACELP	10	8:1	11.8
G.723.1	MP-MLQ	30	10:1	6.3

3. Automatic speaker verification

The task of automatic speaker verification is to determine if a hypothesized test utterance Y was spoken by a hypothesized speaker M. The general approach to this task is to test the likelihood ratio. In automatic systems, the similarity between the test utterance Y and the voice of target speaker M is quantified by similarity score, according to Bayes' defined by:

$$\text{match score} = \frac{P(Y|M)}{P(Y|M_{UBM})}. \tag{6}$$

The likelihood ratio match score is compared to a threshold δ and in case of an excess the speaker is accepted ($\{\delta > \text{match score}\} \rightarrow$ accept). The model represents the target speaker M which can be adapted by using training speech. The model M_{UBM}, called universal background model (UBM) represents the entire spectrum of possible alternatives to the hypothesized speaker and cannot be estimated with maximum accuracy (Reynolds *et al.*, 2000). The criteria for speaker selection to the alternative population are for example: quality of speech, gender and language. Figure 4 presents the exemplary application of the automatic speaker verification based on UBM.

The main objective of an automatic verification model is to determine the likelihood functions in the nominator and the denominator shown on Eq. (6). These functions depend, among others, on the spectral features used. The current system extracts 16 MFCC coefficients, 16 Δ MFCC first order derivatives and the signal log energy. The core of the mel-frequency cepstral coefficients is the short-term Fourier spectrum. The mel-frequency analysis uses filters spaced linearly at low frequencies and logarithmically at high frequencies. Mel-scale is related to human pitch sensation. Such signal representation captures individually and phonetically important characteristics of speech (Davis, Mermelstein, 1980). The MFCC signal representation underlies the subsequent speech analysis.

The next step is the cepstral mean variance-normalization (CMVN). This technique permits to use recognizer even in case of a mismatch between training and testing environment (Furui, 1981). The applied algorithm uses the voice activity detection (VAD) and the CMVN normalization is conducted only on speech, non-zero frames. The VAD method is based on a 2-nd order GMM of log energy distribution. The assumption of this method is a difference between the speech energy and the energy of non-speech frames (Margin-Chagnolleau *et al.*, 2001).

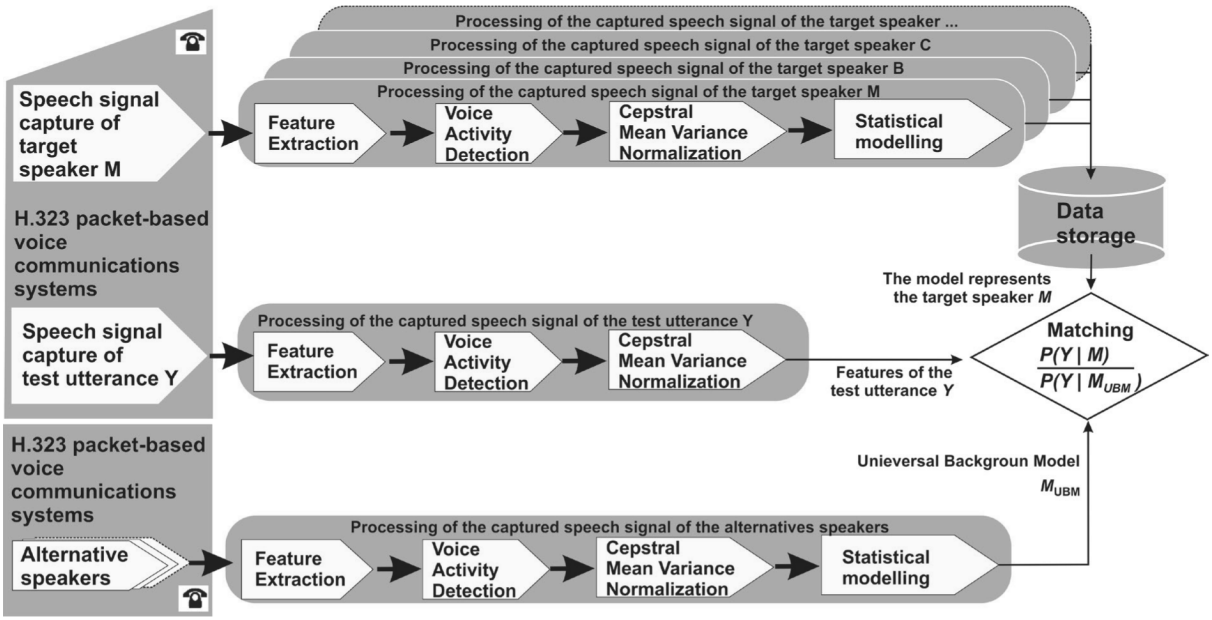

Fig. 4. Basic building blocks of speaker verification system (source: self elaboration).

In text-independent speaker recognition there is no previous knowledge of what the speaker will say. Over the past dozen years, Gaussian mixture models (GMM) have been remaining the most successful method for modelling in text-independent speaker recognition applications (Reynolds *et al.*, 2000). The GMM is based on assumption that the probability density function of some features, being defined in a multidimensional space, can be estimated with function of combination of R unimodal densities components. Each component represents some high level phonetic sound (Reynolds, 2000). For a D-dimensional feature vector $\mathbf{y}$, the Gaussian mixture model is a weighted convex combination of R Gaussian densities $p_i(y)$, defined as:

$$P(y|M) = \sum_{i=1}^{R} \boldsymbol{\omega}_i p_i(y), \tag{7}$$

where R is the model order, $\boldsymbol{\omega}_i$ is the $R \times 1$ vector of weights of each $p_i(y)$. Each Gaussian density component is parameterized by a mean $D \times 1$ vector ($\boldsymbol{\mu}_i$), and a $D \times D$ covariance matrix ($\boldsymbol{\Sigma}_i$). Because of an assumption that MFCC feature vectors are independent, the likelihood of appearance of a sequence of MFCC vectors $\mathbf{Y} = \{y_1, y_2, ..., y_T\}$ in the model represents the target speaker M is defined as:

$$\log P(Y|M) = \sum_{t=1}^{T} \log P(y_t|M), \tag{8}$$

where model is described by statistical parameters $\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\omega}_i\}$, where $i = 1, \ldots, R$.

To estimate the universal background model (M_{UBM}), the expectation maximization (EM) algorithm described by Reynolds *et al.* (2000) and Reynolds *et al.* (1995) with model order 128 is used. The EM algorithm is an iterative algorithm for optimization of the model parameters. To estimate the model M represents the target speaker, maximum *a posteriori* algorithm is used where relevance factor r, according to literature data – equals 14 (Reynolds *et al.*, 2000; 1995). MAP procedure allows to adapt the parameters of the universal background model using the speaker's training speech. The relevance factor is a way of controlling of how much data used to build UBM should be observed in a mixture which represents the target speaker model (Reynolds *et al.*, 2000).

4. Speaker verification over IP networks experiments

4.1. Speaker database description

The verification experiments were conducted on a speaker database containing recordings of 38 Polish language speakers performed with a high quality condenser microphone in PCM format with 44.1 kHz sampling frequency and 16 bit resolution. Both the speakers' training and the test utterance were performed on approximately 30-second long, spontaneous, phonetically rich sentences. Only test samples were interrupted with the packet loss phenomenon. The UBM model was trained from voice speaker database containing 30-seconds long, phonetically rich sentences of 36 Polish language speakers.

The speaker database was transcoded with coders described in Subsec. 2.2. The simulation of packet loss was performed according to Gilbert model (described in Subsec. 2.1). The packet loss phenomenon analyzed at varying degrees of packet loss (between 0–25%) and with constant value of mean burst loss length (D_{burst}) equaled 2, which yielded the best match between expected and obtained value of loss rate, according to 30 second long speech samples. The data were bundling into 20 milliseconds long packets. There was no synchronization between bundling into packets and framing for MFCC calculation.

4.2. Results and discussion

In the automatic speaker verification system, two types of errors are possible, namely: false alarm also known as false acceptance and miss detection also known as false rejection. The rate of both errors depends on the threshold value which is set during verification process as shown in Fig. 4. This means that the decision threshold is adjustable. For example, some kind of applications may need to operate at the high level security – possibly low level of false acceptance rate. In such applications, the threshold needs to be higher than, inter alia, in forensic applications where false acceptance rate and false rejection rate should be balanced. Therefore, automatic speaker verification systems have many possible threshold values, the so called operating points. Hence, a single performance value is insufficient to represent the capability of the system under various conditions (Martin, 1997). A very popular method of presenting the performance is using well known detection error tradeoff (DET) curves, showing the performance on the test set for various thresholds. The DET curve represents the system performance as a miss detection rate (miss probability) in the function of false alarm rate (false alarm probability). The following are the major features of DET:

- plot error rates shown on both axes, giving uniform treatment to both types of error,
- non-linear deviate scale shown for both axes, providing better resolution of plots in the critical operating region,
- functions decreasing monotonically, usually close to linear,
- improvements in performance is shown by functions moving closer to the lower left hand corner of the plot.

DET plots are easily readable and comparable for various experiments under different conditions. A number of special points may be included on DET curve (Martin, 1997). One of them is the equal error rate (EER) which is a measure summarizing the performance. This corresponds to the operating point where FAR is equal to FRR. In Figs. 5–8 and 10, EER points are indicated by ◯.

Figures 5–9 present the results of the performed tests in the form of detection error tradeoff curves.

Figures 5–8 show detection error tradeoff characteristics which represent an effect of voice transmission over IP on text-independent speaker verification performance. The following codecs were investigated: G.711 (Fig. 5), G.723.1 (Fig. 6), G.729 operating at 11.8 kbps (Fig. 7) and G.729 operating at 6.4 kbps (Fig. 8). Another phenomenon examined in the study was the influence of packets loss and the packet loss concealment method using repetition. The results obtained for various codecs may differ according to EER, but it is possible to set some general rules that govern all codecs subject to investigation.

The packets loss influenced verification results and moved the critical operating region of detection error tradeoff curves out of the left corner. This phenomenon was apparent for all codecs examined (Figs. 5–8). For packet loss probability P_Z not exceeding 0.01, the increase of EER was fractional, i.e. below 1%. In the remaining cases EER increase was significant, i.e. above 1%, as compared to no packet loss. For the packet loss probability $P_Z = 0.1$, the highest increase occurred when G.729 operating at 6.4 kbps was applied (Fig. 8). The uppermost value of packet loss probability investigated in the study was 0.25. Under such conditions G.711 and two variants of G.729 (operating at 11.8 kbps and 6.4 kbps) yielded EER values of approx. 15% (Figs. 5, 7, 8). The result for G.723 varied significantly, reaching 18.32% (Fig. 6). G.723 codec operates at 6.3 kbps but the highest obtained EER value was probably not the effect of a low bit rate but it was related to the unpredictable packet loss.

For all codecs, packet loss concealment using repetition improved speaker verification performance, under the condition that packet loss probability was higher than 0.01. In other cases, PLC may facilitate degradation of the results, as seen in Figs. 5 and 8 – PLC applied for P_z=0.01 caused EER increase to 0.01 and 0.02, respectively.

Figure 9 shows the influence of various concealment methods on speaker verification performance for codec G.729 operating on 6.4 kbps and constants packet loss probability $P_Z = 0.25$. The following PLC methods were investigated: no concealment, repetition, simple interpolation, white noise.

It can be seen that EER differs significantly depending on the PLC method. The best result was obtained for PLC applied by interpolation, yielded EER value of approx. 11%.

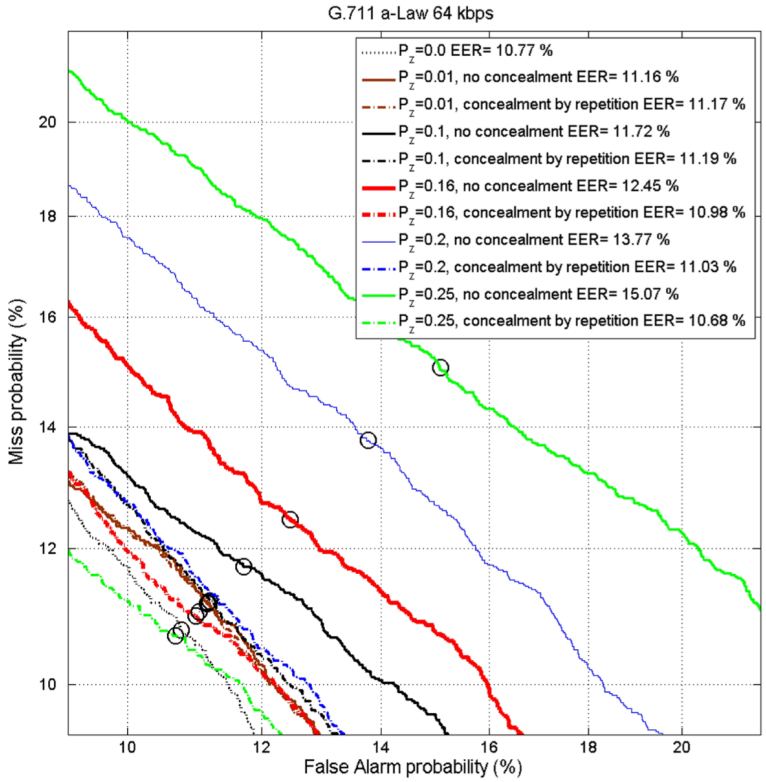

Fig. 5. DET curves for automatic speaker verification with G.711 speech coding for varying degrees of packet loss, from 0 up to 25%, and packets loss concealment method using repetition.

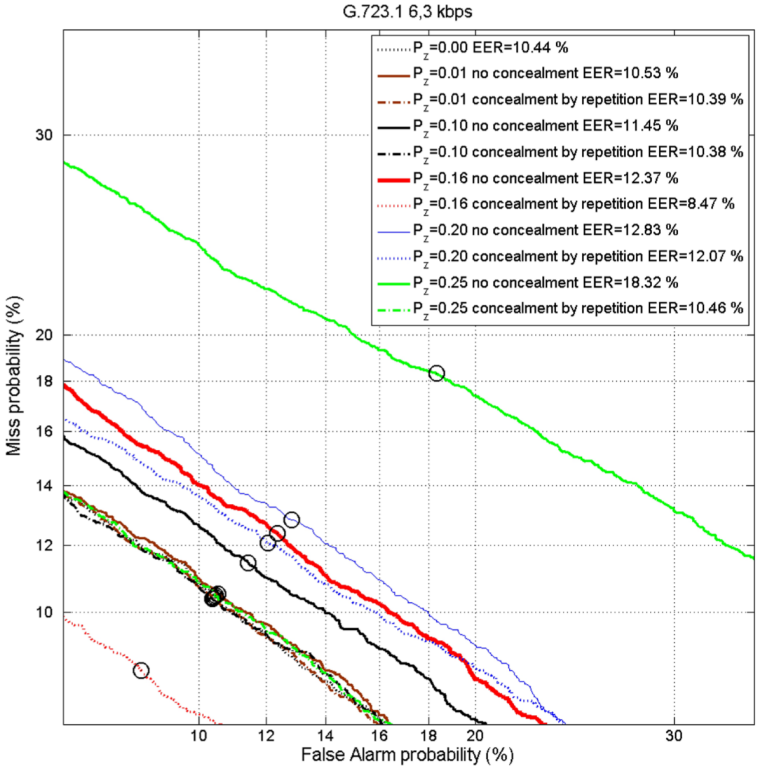

Fig. 6. DET curves for automatic speaker verification with G.723.1 6.3 kbps speech coding for varying degrees of packet loss, from 0 up to 25%, and packets loss concealment method using repetition.

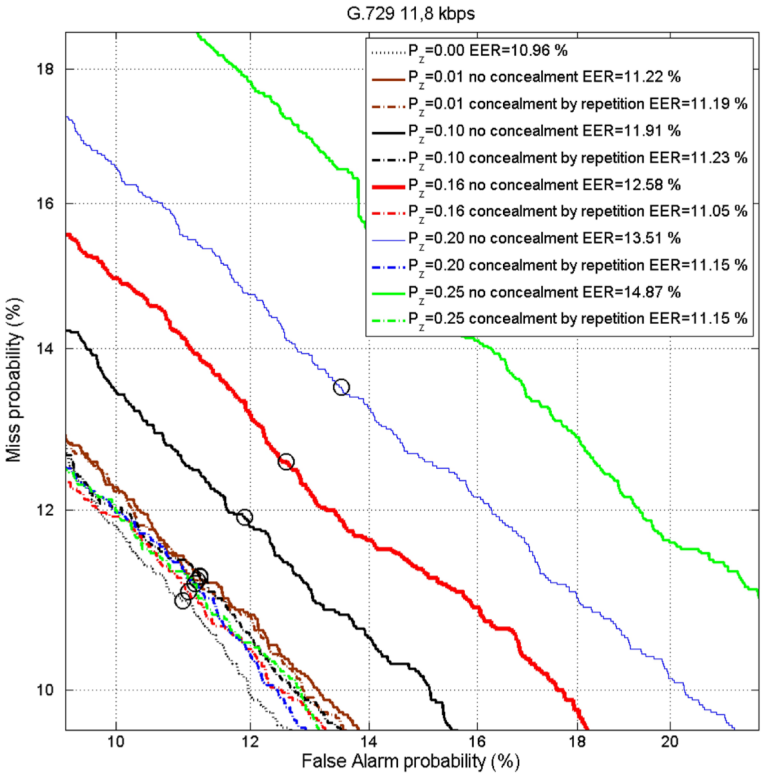

Fig. 7. DET curves for automatic speaker verification with G.729 11.8 kbps speech coding for varying degrees of packet loss, from 0 up to 25%, and packets loss concealment method using repetition.

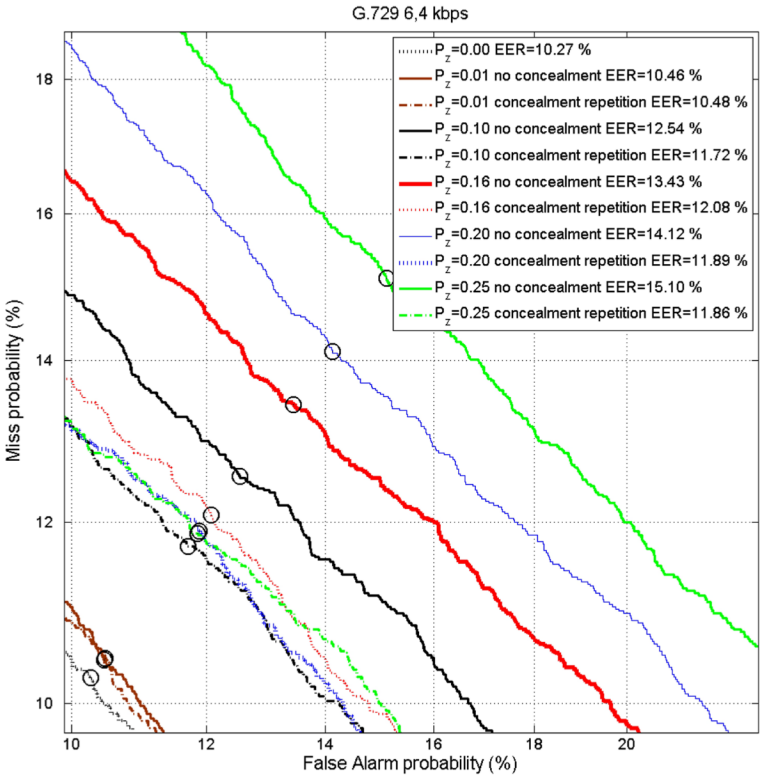

Fig. 8. DET curves for automatic speaker verification with G.729 6.4 kbps speech coding for varying degrees of packet loss, from 0 up to 25%, and packets loss concealment method using repetition.

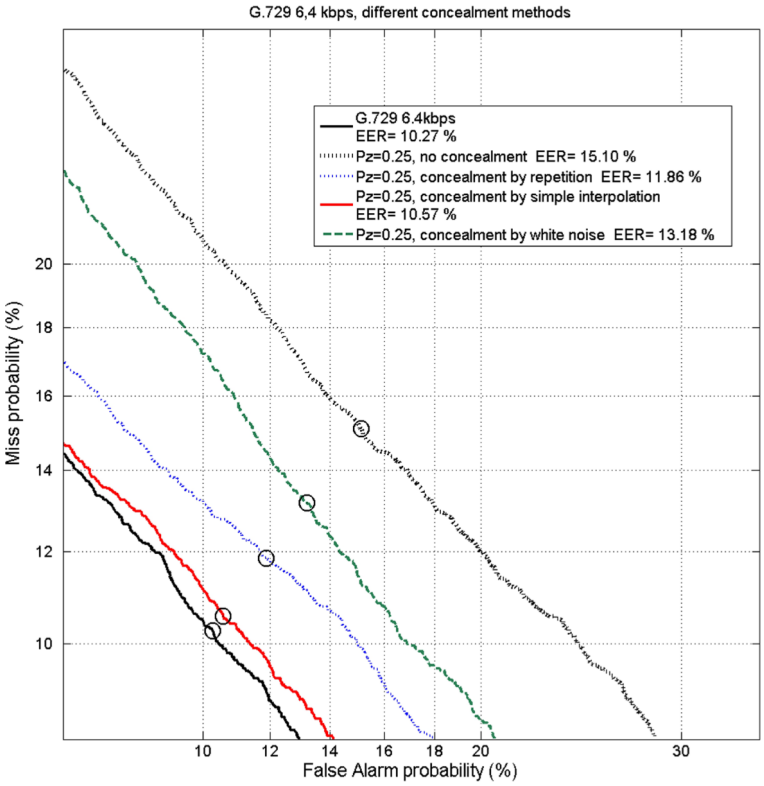

Fig. 9. DET curves for automatic speaker verification with G.729, 6.4 kbps speech coding for constants packet loss equaling 25% and various packets loss concealment methods.

5. Summary

Speaker verification performance is summarized in Fig. 10. Presented results show that there are cases where EER can be lower for codecs with lower bitrate – for example for ITU-T G.723 (bitrate equals 6.3 kbps), compared for example to ITU-T G.711 operates at 6.3 kbps. The source of those phenomenon lies in the degradation introduced by the speech coding process. Other study indicates that bitrate is an important attribute of the speech signal codecs in the context of speaker recognition process (Besacier *et al.*, 2004; Staroniczwicz, 2007). However, the effect of low bitrates on speaker verification performance can be analyzed using only a set of target and test utterances encoded with the same codec but operating at various bitrates.

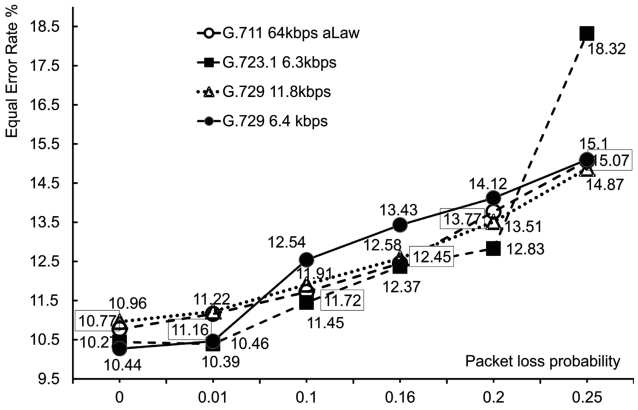

Fig. 10. Equal error rate as a function of the probability packet loss for various methods of speech coding.

The main factor of the speaker verification performance degradation is a packet loss. The dependence between a packet loss probability and an equal error rate is almost linear, regardless of applied codec which is shown on Fig. 10. On the other hand, the highest degree of the packet loss ensuring acceptable voice quality must not exceed 1%. At this rate, the EER factor has increased by no more than 1%, as compared to no packet loss.

Figures 5–8 also present the effect of the packets loss concealment by repetition on the speaker verification performance. The results indicate that PLC improves performance, yet it is difficult to determine the relation between the number of recovered packets and equal error rate, due to the unpredictability of the packet loss and the packet loss concealment phenomena.

Making a comparison between the influence of various PLC methods (no concealment, interpolation, repetition, white noise) on the speaker verification performance, the best result was obtained for PLC using interpolation (Fig. 9). According to Fig. 3a, the packets lost from the speech signal are represented as "gaps". The packet loss concealment by simple interpolation consists of discarding these "gaps" by interpolating using the packets after and before the lost packet (Fig. 3c). As it was mentioned in Section 4.1, there was no synchronization between signal bundling during packet formation and framing during parametrization. Therefore "gaps" in the speech signal may not be detected in a proper way during voice activity detection. Thus, it was concluded, that detection and removing of "gaps" representing lost packets from the speech signal may be the method of improvement in parametrization of automatic speaker verification system, which compensates for packet loss distortion effects.

As it was mentioned in Sec. 3, each of the GMM components represents high level phonetic units (Reynolds, 2000). Using signal distorted by packet loss to estimate a statistical GMM model may lead to modification of statistical parameters $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\omega}\}$. It means that representation of the high level phonetic units changes under the packet loss influence.

Figure 11 shows examples of means $\mu(c_n)$ of bimodal EM-GMM, estimated using first and second

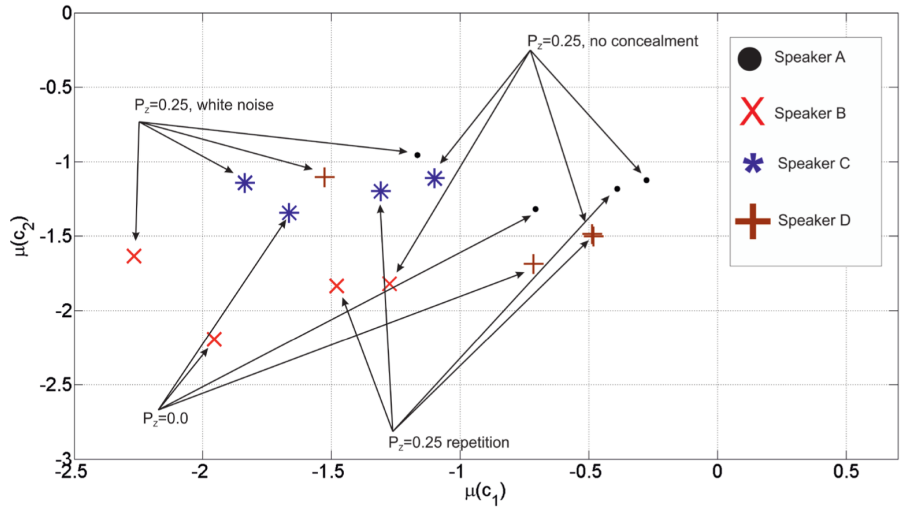

Fig. 11. The influence of packet loss phenomenon on the mean of GMM parameters.

MFC coefficients for four various speakers: A, B, C, D and for various transmission conditions: no distorted signal ($P = 0.0$), distorted signal by packet loss with no concealment ($P = 0.25$), distorted with concealment by repetition ($P = 0.25$, repetition) and distorted with concealment by white noise ($P = 0.25$, white noise). The means of GMM of distorted signal by the packet loss has shifted towards zero values compared to the means of GMM of no distorted signal (Fig. 11). The packet loss concealment by repetition has led to another shift of means but in opposite direction, back to the original values of GMM means of no distorted signal. This phenomenon is the reason of the improvement of speaker verification performance after PLC by repetition applied. Opposite to PLC by repetition packet loss concealment by white noise provides equivocal results. Means for only two speakers (Fig. 11 speakers B, C) for four analyzed, has shifted closer to the original values of means. This observation is correlated with the results shown in Fig. 9 where PLC by white noise has provided better speaker verification performance compared to no concealment case but worse to remaining PLC methods like simple interpolation and repetition.

Acknowledgment

This work was supported by The Polish National Center for Research and Development, the Defense and Security Program, project no. 0023/R/ID3/2012.

References

1. Besacier L., Ariyaeeinia A.M., Mason J.S., Bonastre J.F., Mayorga P., Fredouille C., Meignier S., Siau J., Evans N.W.D., Auckenthaler R., Stapert R. (2004), *Voice biometrics over the internet in the framework of COST action 275*, EURASIP Journal on Applied Signal Processing 2004:4, 466–479, Hindawi Publishing Corporation.
2. Besacier L., Grassi S., Dufaux A., Ansorge M., Pellandini F. (2000), *GSM speech coding and speaker recognition*, ICASSP.
3. Byrne C., Foulkes P. (2004), *The mobile phone effect on vowel formants*, International Journal of Speech Language and the Law, **11**, 1.
4. Davidson J., Peters J. (2000), *Voice over IP fundamentals. A systematics approach to understanding the basics of Voice over IP*, CISCO Press, Indianapolis.
5. Davis S.B., Mermelstein P. (1980), *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, IEEE Transactions on Acoustics, Speech and Signal Processing, **28**, 4, 357–366.
6. Furui S. (1981), *Cepstral analysis technique for automatic speaker verification*, IEEE Transactions Acoustics, Speech, Signal Processing, ASSP, **29**, 254–272.
7. Gilbert E.N. (1960), *Capacity of a burst-noise channel*, The Bell System Technical Journal, September.
8. IETF (2004), *The Effect of Packet Loss on Voice Quality for TDM over Pseudowires*, Internet Draft, October 20.
9. Jajszczyk A. (2009), *Introduction to telecomunication*, [in Polish: *Wstęp do telekomunikacji*], Podręczniki akademickie WNT, Warszawa.
10. Jelassi S., Rubino G.A. (2011), *A study of artificial speech quality assessor of VoIP calls subject to limited bursty packet losses*, EURASIP Journal on Image and Video Processing, 2011:9.
11. Maciejko W. (2012), *Biometric speaker recognition in forensic science*, [in Polish: *Biometryczne rozpoznawanie mówców w kryminalistyce*], Problemy Kryminalistyki 275, Warszawa.
12. Maciejko W. (2014), *Impact of telephone transmission VoIP on forensic automatic speaker identification system based on EM-UBM-MAP algorithms*, [in Polish: *Wpływ transmisji głosu z wykorzystaniem telefonii internetowej VoIP na skuteczność automatycznego systemu kryminalistycznej identyfikacji mówców opartego na metodzie EM-UBM-MAP*].
13. Margin-Chagnolleau I., Gravier G., Blouet R. (2001), *Overview of the 2000-2001 ELISA consortium research activities*, ISCA A speaker Odyssey The Speaker Recognition Workshop Crete.
14. Martin A., Doddington G., Kamm T., Ordowski M., Przybocki M. (1997), *The DET curve in assessment of detection task performance*, Proc. Eurospeech '97, pp 1895–1898, Rhodes, Greece.
15. Mayorga P., Besacier L., Lamy R., Serignat J.-F. (2003), *Audio packet loss over IP and speech recognition*, Automatic Speech Recognition and Understanding, ASRU '03.2003 IEEE Workshop on 30 Nov.-3 Dec., 607–612.
16. Mohamed S., Rubino G., Varela M. (2004), *Performance evaluation of real-time speech through a packet network: a random networks-based approach*, Performance evaluation. An international Journal, **57**, 141–161.
17. Peinado A.M., Segura F.C. (2006), *Speech recognition over digital channels. Robustness and Standards*, John Wiley & Sons, Ltd.
18. Sanneck H. (2000), *Packet loss recovery and control for voice transmission over the internet*, Ph.D. Thesis Technischen Universität Berlin, unpublished.
19. Staroniewicz P. (2006), *Influence of specific VoIP transmission conditions on speaker recognition problem*, Archives of Acoustics, **31**, 4 (Supplement), 197–203.
20. Staroniczwicz P. (2007), *Tests of robustness of GMM speaker verification in VoIP telephony*, Archives of Acoustics, **32**, 4 (Supplement), 187–192.
21. Recommendation ITU-T H.323 (2009), *ITU-T H.323 Series H: Audiovisual and multimedia systems. Infrastructure of audiovisual services – Systems and terminal equipment for audiovisual services. Packet-based multimedia communications systems*, International Telecommunication Union 12/2009.

22. Reynolds D.A., Quatieri T.F., Dunn R.B. (2000), *Speaker verification using adapted gaussian mixture models*, Digital Signal Processing, **10**, 19–41.

23. Reynolds D.A. (1996), *The effects of handset variability on speaker recognition performance: experiments on the switchboard corpus*, Acoustics, Speech and Signal Processing, ICASSP-96, Conference Proceedings.

24. Reynolds D.A., Rose R.C. (1995), *Robust text-independent speaker identification using gaussian mixture speaker models*, IEEE Transactions on Speech and Audio Processing, **3**, 1, 72–83.

25. Reynolds D.A., Zissman M.A., Quatieri T.F., O'Leary G.C., Carlson B.A. (1995), *The effect of telephone transmission degradation on speaker recognition performance*, Acoustics, Speech, and Signal Processing, 1995, ICASSP-95.

26. Rose P. (2002), *Forensic speaker identification*, Taylor & Francis, New York.

27. Transnexus, Inc. (2013), *Four VoIP Trends to Watch for in 2013*, Retrieved May 2-nd, 2013 from Transnexus, Inc. Newsletter Issue 5, January 2013, http://www.transnexus.com/index.php/issue-5-january-2013/four-voip-trends-to-watch-for-in-2013.

28. Viikki O., Laurila K. (1998), *Cepstral domain segmental feature vector normalization for noise robust speech recognition*, Speech Communication, **25**, 133–147.

29. Young S., Evermann G., Gales M., Hain T., Kershaw D., Liu X., Moore G., Odell J., Ollason D, Povey D., Valtchev V., Woodland P. (2009), *The HTK book v3.4*, Cambridge.