Adaptive VoIP with Audio Watermarking for Improved Call Quality and Security

Wojciech Mazurczyk[1], Zbigniew Kotulski[1,2]

[1]Warsaw University of Technology, Faculty of Electronics and Information
Technology, Institute of Telecommunications 15/19 Nowowiejska Str.
00-665 Warszawa, Poland
{wmazurczyk, zkotulsk}@tele.pw.edu.pl

[2]Polish Academy of Sciences, Institute of Fundamental Technological Research
zkotulsk@ippt.gov.pl

***Abstract*:** In this paper we describe a novel adaptive method of speech quality control which may be used to adjust three call parameters: speech codec configuration, playout buffer size, and amount of FEC (Forward Error Correction) mechanism information during VoIP (Voice over Internet Protocol) call under changing network conditions. This solution differs from previously proposed because it utilizes audio watermarking techniques as a communication channel between calling parties to send information about the change in quality of the call. Assessment of the call quality is performed with non-intrusive objective methods like E-Model [9, 10] and expressed in MOS scale. Moreover we show how to use the proposed speech control mechanism in conjunction with adaptation of security measures applied for VoIP traffic. Thus, we gain a tradeoff between the quality and security of the call which is very a important and still unsolved issue for IP telephony.

***Keywords*:** Adaptive Voice over IP, Audio Watermarking, Speech Quality Control, IP Telephony Security.

1. Introduction

VoIP is a real-time service that enables conversation through IP networks. It is very popular and currently VoIP providers play important role in the telecom market. However, two unsolved problems still exist for IP telephony. One is providing security of the traffic that is exchanged between calling parties and the other is providing reasonable quality of the call for end-users. The latter is most important issue because if the quality of the call is insufficient then the participants of the conversation may not be able to communicate at all. Both above-mentioned problems are related as security mechanisms affect QoS (Quality of Service) parameters, e.g., by introducing additional delays and increasing protocols overheads. If too many security mechanisms are applied for VoIP service then the quality of the call may be degraded. So, when network congestion occurs then applied security measures may make this call impossible to continue, while without them the conversation could be potentially continued. Reasonable tradeoff between security and providing quality is always necessary for real-time services like VoIP.

That is why in this paper we address both described aspects: providing quality and security simultaneously by introducing a novel method of AVoIP (Adaptive VoIP) system. We consider end-to-end based adaptation approach where QoS control is implemented at application level of TCP/IP model. While using this solution voice applications adapt certain call parameters to the changing network conditions (in a way that introduces least session disruption possible) in order to achieve better QoS perceived by end-users. The proposed method may be used to complement existing, traditional network QoS mechanisms like Intserv [1] or Diffserv [2]. These mechanisms are mostly suitable for wired, fixed network topology (and fixed network resources); their main goal is to reserve/assure certain network resources so the packets which come from real-time traffic sources are handled better then other present in the network. That is why, it may be important to implement adaptive VoIP solutions especially in heterogeneous networks that may change dynamically (e.g., wireless or mobile). Additionally, such an approach is reaffirmed by the characteristic nature of multimedia applications as they allow to adjust the traffic flow and quality to be able to fit (to some extent) requirements of end-users, applications and network.

By utilizing AVoIP calls configuration parameters like: speech codec configuration (output rate, size of the voice frame, etc.), playout buffer size and amount of information used for FEC mechanism, generated traffic may be adapted to the current state of the network. So, if a congestion occurs in the network, the bandwidth of a VoIP audio stream is lowered, as well as other mentioned parameters. In this situation the probability of further packet losses and excessive delay due to network conditions is decreased. Thus, we are able to reduce the load of the traffic in the network when congestion occurs (when all VoIP sources in the network use such a method). For effective network congestion control for real-time services like VoIP, application-layer algorithms, and lower-layer rate control schemes should cooperate.

In proposed here solution there are four important contributions. First, our AVoIP system uses all possible call parameters that may be changed during the VoIP conversation. We combine adaptation parameters that were

Received September 29, 2007.

proposed earlier but were used separately in previous works. Here, adjusting different parameters affect the call in more complex way and may improve its quality.

Next, the proposed AVoIP system utilizes audio watermarking techniques to transfer the information about network conditions to the receiving side. Data describing network status is embedded into the audio stream, then sent and finally retrieved at the receiver. Currently, existing solutions to achieve the same goal use usually RTCP protocol [3]. The main disadvantage of RTCP is that its messages consume additional bandwidth (they are sent separately from the audio stream) that should be utilized for voice packets. In a case of congestion those messages may likely be lost or they will make it harder for audio stream to get through the network and to reach the receiving side. Moreover, when RTCP messages are lost, any adaptation mechanism that is implemented and uses them may not function properly.

The third contribution is the following: because the audio watermarking covert channel that is created in audio stream has limited capacity we propose to exchange only the score that characterizes completely quality of the call (e.g., in MOS scale) and is acquired as a result of the operation of call assessment algorithms.

Finally, the AVoIP technology enables adaptation also for security mechanisms applied to VoIP system. The better network conditions, more resources may be used for security measures. If the network is congested not only the providing security fails but also the call may impossible to continue because of the drop in quality. So, by using AVoIP in the proposed way we gain the tradeoff between providing security services and expected quality of the call (QoS).

The paper is organized as follows: Section 2 provides basic information about VoIP service and general overview of adaptive methods operations are provided. We also present there fundamentals about call quality assessment algorithms and audio watermarking techniques. Next, in Section 3 available VoIP call parameters are presented which may be subjected to adaptation process during the call. Then, in Section 4 proposed adaptive VoIP solution is described in details. In Section 5 relation between quality of the call and security measures applied for VoIP call are characterized. We also propose how to adjust security mechanisms while the network is congested by utilizing lightweight security mechanisms. Finally, we summarize the obtained results and circumscribe potential future work in Section 6.

2. Backgrounds

For adaptive VoIP service two types of QoS (Quality of Service) must be considered (as introduced in [4]). First, the most important one, is **perceived quality**, which refers to humans evaluation of the quality of the phone call that they participate in. Second is **networking quality** which relays on the estimation of the network conditions (by measuring parameters like, e.g., jitter, delay, and packet losses). Additionally, perceived quality depends on the network impairments like delay and jitter, packet loss (networking QoS) and other source characteristic parameters like: type of speech codec used or size of the voice frame. So, besides the fact that perceived quality of the call is the most important parameter for end-users, it may be also used to assess network conditions. If the network is not congested then this value is high and while under congestion it may decrease significantly.

2.1 VoIP Traffic Flow

VoIP is a real-time service that enables voice conversations through IP networks. It is possible to offer IP telephony due to four main groups of protocols:

- **Signalling protocols** that allow to create, modify, and terminate connections between the calling parties; currently the most popular are SIP [5], H.323 [6], and H.248/Megaco [7],
- **Transport protocols** from which the most important one is RTP [3], and it provides end-to-end network transport functions suitable for applications transmitting real-time audio. RTP is often used in conjunction with UDP (or rarely TCP) for transport of digital voice stream,
- **Speech codecs** (e.g., G.711, G.729, G.723.1) that allow to compress/decompress digitalized human voice and prepare it for transmitting in IP networks.
- Other **supplementary protocols** like RTCP [3], SDP, or RSVP that complement VoIP functionality. For purposes of this paper we explain the role only of RTCP protocol. RTCP is a control protocol for RTP and it is designed to monitor the Quality of Service and to convey information about the participants in an on-going session. Most of the existing adaptation VoIP solutions utilize information from RTCP messages to assess network status.

Generally, IP telephony connection consists of two phases: a signalling phase and a conversation phase. In both phases certain types of traffic are exchanged between calling parties. After the signalling messages are exchanged between the caller and callee, and the connection is successful, the conversation takes place (in form of audio streams which are sent bidirectional).

Besides utilizing the protocols mentioned above VoIP systems are able to provide calls in IP networks, despite the negative effects like: delay, packet loss and jitter. This can be achieved due to the utilization of the mechanisms like playout buffer or FEC mechanisms that help to alleviate these negative effects.

Nevertheless, parameters like: delay, packet loss, and jitter effect (delay variations) affect perceived quality and when their values are exceeded the call may be unable to continue due to quality degradation. It is also worth mentioning that for IP telephony, we consider a packet is lost when:

- It does not reach the destination point,
- It is delayed excessive amount of time (so, it is no longer valid), it can be used no more for current voice reconstruction in the receiver at the arrival time.

2.2 General Adaptive Call Quality Control Methods Operations

Currently there are different approaches used in existing adaptation mechanisms. The main difference is what information should be processed in order to the adaptation to happen and what parameters of the call will be adjusted.

Generally, measures which may be used for network status evaluation are packet loss ratio, delay, jitter. They are usually exchanged with use of RTCP protocols reports: SR (Sender Report) and RR (Receiver Report). Based on these messages certain values are calculated and the adaptation decision is made. General adaptation scenario is presented in Fig. 1.

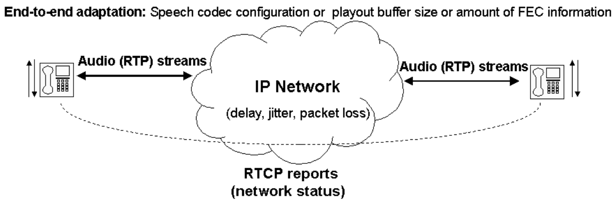
End-to-end adaptation: Speech codec configuration or playout buffer size or amount of FEC information
Audio (RTP) streams
Audio (RTP) streams
IP Network
(delay, jitter, packet loss)
RTCP reports
(network status)

Figure 1. General adaptation mechanism scenario

What we propose in this paper is to send simple but complete metric for VoIP service quality and that is perceived quality score achieved from real-time quality models like ITU-T E-Model [8] or model like presented in [4]. As mentioned earlier all of existing adaptation methods use RTCP protocol messages (reports) to exchange network status data. Main drawback of this solution is that this protocols messages consume additional bandwidth. So, in case of congestion it is likely that those packets will be lost or even make the network conditions worse. What we are proposing in this paper is to utilize a steganographic channel created with audio watermarking techniques so the necessary information will be sent inside the audio stream (conversation). Thus, we save the network resources that in time of congestion may be needed for voice packets. And as described earlier, perceived quality model score will be transmitted with use of audio watermarking inside the audio stream.

2.3 Call Quality Assessment Methods

Call quality assessment methods can be divided into two broad groups: subjective and objective. Subjective measurements of the QoS are carried out by a test subjects (people) by e.g. listening tests. They are performed usually in the special environment e.g. rooms where background noises and other factors that can influence quality are kept under control during test execution. However, subjective methods have certain limitations like that they are impracticable and quite expensive to run.

To overcome these shortcomings objective methods were developed. They allow to calculate certain values (scores) that represent different factors of the network that affect call quality. The output result should be close (in ideal situation equal) to subjective method score. Moreover, objective methods can be divided into intrusive and non-intrusive algorithms [24]. Generally the difference is that the first one is based on comparison between reference signal (not distorted, original) and the same signal but after traversing a network. Examples of these methods include: PSQM (Perceptual Speech Quality Measurement Method), PAMS (Perceptual Analysis Measurement System) or PESQ (Perceptual Evaluation of Speech Quality) algorithms. On the other hand non-intrusive speech quality measurement may be utilized for real-time services like VoIP, because the original signal (reference) may not be known. Examples of these methods include: E-Model or PSOM (Perceptual Single Ended Objective Measure).

For our purpose we will utilize non-intrusive methods as presented above because they are most suitable for real-time quality assessment, which we need. As mentioned earlier the most important measure for calling parties, that participate in VoIP call, is perceived quality of the call. That is why in proposed solution we will send the information about the call quality (e.g. in form of the MOS score) from the transmitter to the receiver by embedding this data (in a form of digital watermark) into voice samples of the audio stream. Thus, participants of the call may influence the ongoing conversation (transparent human interaction). Then after retrieving watermark, based on speech codec set available for calling parties the specialized algorithm is executed to decide whether to adapt certain parameters for ongoing call or not. The adjustment of call configuration depends on the network conditions, and may be increased or decreased as necessary.

2.4 Audio Watermarking Techniques for VoIP Improvement

So far primary application of audio watermarking was to preserve copyrights and/or intellectual properties sometimes called DRM (Digital Right Management). However, it may be also utilized to create effective covert channel inside a digital content (in our case it is voice). Each of audio watermarking algorithms consists of two phases: embedding of the digital watermark into the voice at the source and then its extraction at the destination. In IP telephony we can distinguish those phases too; as soon as the conversation begins, certain information is embedded into the voice samples and sent through the communication channel. Then, the digital watermark is extracted from a voice stream before it reaches a callee.

Currently, we can point out a number of audio watermarking algorithms which may be exploited in proposed AVoIP system. The most popular techniques that are applicable in real-time communication for VoIP service, include: LSB (Least Significant Bit), QIM (Quantization Index Modulation), Echo Hiding, DSSS (Direct Sequence Spread Spectrum), and FHSS (Frequency Hopping Spread Spectrum). For these methods the bandwidth of available covert channels depends mainly on the sampling rate and the type of audio material being encoded. Research results in [22] have shown that for LSB technique communication rate is 1 kbps per 1 kHz (e.g. for 8 kHz sampling rate the capacity is 8 kbps), echo hiding around 16 bps, while DSSS achieved 4 bps. Other experiments in [23] have shown that DSSS method's bandwidth is about 22.5 bps, FHSS 20.2 bps, echo hiding 22.3 bps and LSB around 4 kbps.

As mentioned earlier in this paper we will utilize audio watermarking as a covert communication channel to transfer information about perceived quality of the call. What must take under consideration is that covert channel possesses limited capacity and that is why the data sent must be compact with size kept to minimum.

3. Adaptive parameters for VoIP call

Adaptive call quality control mechanism may be based on different parameters that are dynamically modified during VoIP connection, as mentioned in Section 2.2. What must be circumscribed first is which parameters may be adjusted during the VoIP call and in result which of them may be used in adaptation process. These parameters are listed below and we can classify them according whether they are modified at the source or at the receiver. At the source these adaptive parameters are:

- Modifying **speech codec configuration** or alternatively **codec switching**. If only one speech codec is available at endpoints, for some codecs' output rate and/or frame size may be adaptively modified (e.g., for AMR [11]). If it is not a case, but more than one speech codec may be used for VoIP connection then current speech codec can be switched to different one (e.g. with lower output rate). The output rate of the speech encoder (whether it is the same speech codec with different output rate or changed from one to another) is adjusted to match the current characteristics of the network: speech will be coded with low bit rates when the network is congested and with higher otherwise. Adaptive methods that function as presented above are described for example in [12], [13] and [21].
- Amount of **FEC** (Forward Error Correction) **information.** FEC is a mechanism that is used for lost packets recovery (as long as the following packets are received successfully). It adds redundancy to the transmission and also introduces additional delay. What is important is that amount of information which is used for voice reconstruction can be adjusted dynamically. For example, such a method is proposed in [14].

At the receiver the following parameter may be adaptively changed to match current network conditions:

- **Playout buffer size** (also called de-jitter, jitter buffer or playout scheduler), which is used to properly reconstruct speech signal by temporally storing each packet so they can be played out in a timely manner. Adaptive playout buffer algorithms should be characterized by the ability to provide the buffering delay as short as possible and on contrary minimizing the number of packets that arrive too late to be played out (when this happens they are considered as lost). Adaptive methods that modifies playout buffer size are presented in [15], [16], [17], [18].

All abovementioned parameters which may be used for AVoIP systems are relaying on the information about the current network conditions. Whether we utilize variable speech codec rate, adaptive playout buffer size or FEC two problems arise:

- How to estimate current network status,
- How to transfer this data to the sending or receiving side when e.g. congestion occurs.

Every effective AVoIP system must possess efficient solution to these problems.

4. Adaptive VoIP Solution Description

Proposed in this paper AVoIP system should be implemented in end-users applications. Additionally, it should run automatically without users intervention and its operation should be transparent for them. General mechanism scenario is presented in Fig. 2.

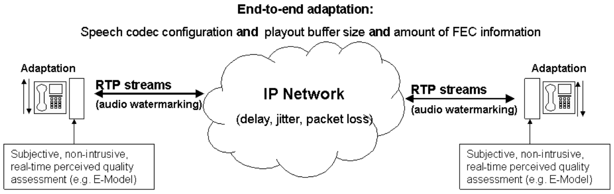
End-to-end adaptation:
Speech codec configuration and playout buffer size and amount of FEC information
Adaptation
RTP streams
(audio watermarking)
IP Network
(delay, jitter, packet loss)
RTP streams
(audio watermarking)
Adaptation
Subjective, non-intrusive, real-time perceived quality assessment (e.g. E-Model)
Subjective, non-intrusive, real-time perceived quality assessment (e.g. E-Model)

Figure 2. General AVoIP mechanism scenario operation

Using proposed Adaptive VoIP method to implement VoIP service allows more efficient use of network resources. AVoIP by adapting to potential network congestions, not only improve perceived quality of the call but also leave more available bandwidth to e.g. signaling and/or critical in-band flow management messages or for security measures. AVoIP changes the size of the playout buffer, speech codec parameters (or switches codecs) and the amount of information for forward error correction mechanism in order to maximize the perceived VoIP quality, which can be expressed as a following function (***F***):

$$QoS_{Perceived} = F(Ns, Pbs, Cc, FEC) \tag{1}$$

where:

QoS

Ns denotes network conditions which includes parameters like: delay, jitter and packet loss ratio,

Pb is playout buffer size,

Cc denotes speech codec configuration parameters like output rate and voice frame size,

FEC is amount of information that is used for forward error correction mechanism.

By adapting these parameters we provide only the bandwidth that the network is capable of carrying, at the certain moment, with a given perceived quality of service. As mentioned before, change in perceived quality measure (expressed in MOS score) triggers AVoIP adaptation process.

To properly characterize proposed adaptive VoIP system the following aspects must be addressed:

- The problem of **estimating the state of the network** and the way to send these data. Because IP service model does not offer congestion notification we have to define how to evaluate network conditions and how to send these information between calling parties.
- A speech codec configuration / playout buffer / FEC

control algorithm. There must an algorithm exist that will define how VoIP application will react when the congestion occurs.

- Which is more efficient and suitable mode of speech codec operation: changing output rate of the codec during the call (Variable Bit Rate Speech Coding) or codec switching (e.g. from G.711 to G.729 and reverse).

Proposed AVoIP solution will be characterized based on above aspects in the next sections.

4.1 Non-intrusive, Real-time Quality Assessment Methods

For purposes of AVoIP any perceived quality assessment model that is non-intrusive and calculates its score in real-time may be utilized. The examples of such models are: E-Model [9, 10] or the one presented by Hoene in [4].

We assume that the quality score is expressed in MOS scale (almost all popular quality assessment models' results can be expressed as MOS score). Based on this value adaptive parameters of VoIP call, which were mentioned earlier, are modified.

In AVoIP (see Fig. 2), the quality assessment model is used twice: first when the audio stream leaves transmitter and second before it reaches the receiver. Firstly, at the transmitter, the speech that is transferred in output audio stream is evaluated ($\boldsymbol{MOS_T}$ value is obtained). Next, result value is embedded into transmitted audio stream. Notice that audio watermarking algorithm also affects the perceived quality of the call in a certain way ($\boldsymbol{MOS_W}$). Then, when the packets with audio watermark reach receiver it is again subjected to quality assessment ($\boldsymbol{MOS_R}$). So, the loss in call quality introduced by network ($\boldsymbol{MOS_{Loss}}$) can be expressed as follows:

$$MOS_{Loss} = MOS_W - MOS_R \tag{2}$$

Similarly loss in perceived quality introduced by audio watermarking technique ($\boldsymbol{MOS_{WLoss}}$) can be expressed as:

$$MOS_{WLoss} = MOS_T - MOS_W \tag{3}$$

where:

$\boldsymbol{MOS_T}$ is quality assessment score which is calculated after coding processing,

$\boldsymbol{MOS_W}$ denotes quality assessment score which is obtained after the output audio stream is subjected to audio watermarking technique. This value depends on the $\boldsymbol{MOS_T}$ score ($\boldsymbol{MOS_W < MOS_T}$),

$\boldsymbol{MOS_R}$ is quality assessment score that is evaluated before the audio stream reaches receiver ($\boldsymbol{MOS_R{<}MOS_W <MOS_T}$).

4.2 AVoIP Estimation of Network Status and the Method for Exchanging These Information

As mentioned earlier, in most adaptive VoIP systems the estimation of the network conditions is realized based on RTCP reports. They use parameters like delay, packet loss ratio and jitter to evaluate network status.

We present other approach by utilizing perceived quality of the call as a measure based on which the adaptation will occur. For obtaining this meaasure, as mentioned earlier, we propose to use a real-time, non-intrusive, subjective quality assessment models like E-Model or the one proposed in [4]. The score that is an output result of quality model operation is continually send between end users. When network congestion occurs this value decreases and based on this information adaptive parameters of the VoIP call are adequately modified. When the network conditions, after congestion, are back to normal again adaptive parameters are also changed to match current network status (and perhaps to offer better perceived quality to the user).

In described AVoIP system we also propose novel way of transporting obtained call quality assessment scores. They are embedded in a form of digital watermark into the audio streams exchanged between calling parties (into conversation itself). That means that each voice packet will carry whole or part of the voice quality score. This process is repeated continuously as the information is exchanged during the whole conversation. At the receiving side the digital watermark is retrieved and its value is used to estimate network conditions.

Additionally, apart from presented methods marking of the important speech frames may be also implemented like presented in [12].

4.3 AVoIP Configuration Parameters Control Algorithm

For AVoIP control algorithm the most important parameter that triggers adaptation process is **threshold value**, so it is vital to choose it right. While it is misfited it may influence perceived QoS significantly. That is why it is vital to determine this value appropriately. If the adaptation threshold value is too low then e.g. speech codec (or codec's rate) may be switched to a lower perceived quality of the call even if it does not provide any improvements. The same situation is in reverse scenario if the codec is changed from lower rate one to higher too early. This may lead to a negative effects that adaptation mechanism may introduce in providing perceived QoS.

Additionally, the control algorithm should be immune to very short periods of quality degradation. If it reacts to every change in voice perceived quality (even very short ones) then the switching of the speech codec (or its output rate) will be too often (flipping) which may lead to drop in conversation quality.

That is why we propose control algorithm which may be expressed in the pseudo-code as presented in Algorithm 1. We present an algorithm for scenario where two speech codecs (C1, C2) are available and C1 is characterized by higher output rate and MOS score than C2 codec.

Algorithm 1. Control algorithm for call quality drop scenario.

```
i = 0

Do{
   MOS_W = ExtractWM(Received_Audio)
(5)      MOS_R = CalculateMOS(Received_Audio)
(6)
(7)      If (MOS_R) <= (Threshold)
(8)       StartTimer(Threshold_Timer)
(9)
(10)     MosTable[i] = MOS_R
(11)
(12)     If (Threshold_Timer) >= (Max_Timer)
(13)      {
(14)       AvLoss = CalcAvMOS(MosTable[])
(15)       If (AvLoss) <= (Treshold)
(16)        {
(17)        StopTimer(Treshold_Timer)
(18)        AdjustOutputAudioRate(C1,C2)
(19)        AdjustPlayoutBuffer(size)
(20)        AdjustFEC(size)
(21)        i = -1
(22)        Free(MosTable[])
(23)       }
(24)      }
(25)    i = i + 1
(26)   }while (conversation_lasts)
```

The proposed solution works as follows: when the quality of the call drops to the chosen threshold (see line 7: let us assume that its value is set to the MOS score of the second speech codec measured at the transmitter output after audio watermark is embedded) special timer (*Threshold_Timer*) is started (line 8). Then, for the given period of time measured MOS values of the incoming audio are stored in *MosTable[]* table. Next, when the timer expires the average MOS value for the timer period is calculated. Then, if this value is still below the selected threshold voice call parameters are adjusted accordingly.

The algorithm that controls situations when the MOS values rises (network recovers from congestion) works in analogous way as presented in Algorithm 1.

Example 1

Lets consider the following scenario where both communication sides utilize AVoIP and have available two speech codecs C1 and C2. This AVoIP system is characterized with:

- MOS score at the transmitter output (without watermark embedded) for C1 is 4.2 and for C2 is 3.7 (MOS_T),
- MOS score after the digital watermark is embedded into the audio stream for C1 is 4.1 and for C2 is 3.6 (MOS_W),
- The threshold value for quality adaptation is set at 3.6 (MOS_W of the C2, where the adaptation decision will be made).

As soon as the conversation phase of the call begins the incoming audio stream (which traversed through network) at the receiver is subjected to quality assessment model (MOS_R). Lets assume that at one moment of the call MOS_R score while using C1 codec drops to 3.5. According to the control algorithm because MOS_R is less than 3.6 (the threshold value) the timer is started and the MOS_R values are stored for further calculations. Lets assume that after the timer expires the average MOS_R is 3.2, which points out that most likely the congestion happened. That is why the C1 speech codec is switched to C2 (the one with lower output rate) and also playout buffer size and the amount of FEC information (if used) are modified.

By using these algorithm we gain better network congestion situation handling – without adaptation the perceived quality of the call may degrade to the point that it will be not possible to continue it. With AVoIP when the congestion occurs dropping quality of the call triggers voice configuration change. In result, under the same network conditions perceived quality of the call may improve.

4.4 Changing Codec Rate vs. Switching Codecs

During the call end-users may renegotiate audio sessions. That is why, by utilizing this fact, two possibilities are available to modify the output rate of the audio stream:

- If more than one codec with different output rates is available for the caller and callee then the adaptation may happen by switching between these codecs,
- If only one codec is available but it is capable of providing different output rates (e.g. AMR codec) then the adaptation may be also used in the same way.

First solution is especially important because it enables AVoIP system for IP telephony even if specialized multi-rate codec is unavailable (or licensed). As to the second case, research in [20] showed that such codecs like AMR or M^3R are effective solutions for adaptive VoIP applications.

5. AVoIP Importance for VoIP Security

As outlined in Section 1 providing security services for IP telephony traffic is still an open issue. Security and QoS parameters of the call are, in general, antagonists as security mechanisms impose overheads and add delays which affects performance. This situation causes end-users to switch off security measures in order to get better perceived quality or even to be able to make a call at all.

Generally, for audio watermarking techniques, the higher codec rate, the more data can be embedded into the voice samples and also the perceived quality is better. That means that more information can be transferred in covert channel created with audio watermarking algorithm.

What we propose is to tie those two aspects: providing security and perceived quality within AVoIP system. Based on abovementioned facts, we want to apply adaptive mechanism not only for speech quality control (as presented in this paper) but also to adjust the level of provided security for VoIP. As mentioned earlier, too high security level, under network congestion, may lead to sooner call degradation than if we do not use any security mechanisms. But we do not consider here VoIP systems that provides no security services at all. Instead we adjust security measures to current network conditions: if there is no congestion in the network more complex (these which introduce higher delays and impose overheads but cryptographically stronger) security mechanism may be used. If congestion occurs, lightweight security solutions that involves less resources utilization should be exploited. In result, we can provide better perceived quality for end-users and simultaneously we do not resign from security measures.

5.1 VoIP Security Services

For IP telephony the most important security services are: authentication (with integrity) and confidentiality. These services should be provided for all types of VoIP traffic that are exchanged between calling parties. That includes: signalling messages and packets from audio streams (conversation). For each of these traffics certain protocol-specific security mechanism exists e.g. for RTP (Real-Time Transport Protocol), which is the most popular transport protocol for audio streams, it is SRTP (Secure RTP) which provides authentication and confidentiality services for VoIP calls.

While network is under congestion we may resign from complex and time-consuming security mechanisms and utilize lightweight security solutions. Thus, we may not be able to provide all security services but at least one of them (authentication or confidentiality) must be provided.

5.2 Lightweight Security Mechanism Based on Network Steganography and Digital Watermarking

Lightweight security mechanism for real-time service as VoIP should be characterized by low bandwidth consumption, low complexity and should has minimum effect on perceived quality of the call. Such a solution was presented by authors of this paper in [19]. The main idea is to utilize special steganographic protocol which header is transferred with use of network steganography (in free or unused fields of IP/UDP/RTP protocols) and the payload is sent in the covert channel created with audio watermarking techniques (the same covert channel that is used for transmitting MOS score between calling parties in AVoIP). Described situation is presented in Fig. 3.

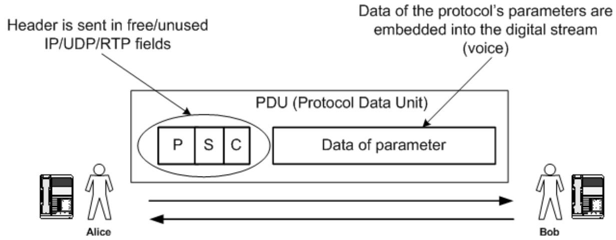
Header is sent in free/unused IP/UDP/RTP fields
Data of the protocol's parameters are embedded into the digital stream (voice)
PDU (Protocol Data Unit)
P
S
C
Data of parameter
Alice
Bob

Figure 3. General lightweight security mechanism idea

The payload of this protocol, when it transfers security parameter (value that is used for security procedures), is called a token and it is formed based on the result of hash function on voice samples (VF_N), shared password (PASS), identifier of the sending side (ID_A), timestamp (TS), random number (R). For token calculation also potentially may be taken signalling messages (SM_N) that were exchanged in the signalling phase of the call. So, the token may be expressed in the following way:

$$TokenA_N = H\left(H(VF_N)) \,\|\, H(SM_N) \,\|\, \begin{pmatrix} TS \\ PASS \\ ID_A \end{pmatrix} \|\, R \right) \|\, R \qquad (4)$$

Two security payloads for this solution are available:

- One is used to provide the authentication and integrity of the voice, its source and signalling protocol that is used in a particular VoIP system (the token as presented above),
- Second is to authenticate exchanged protocol parameters that were sent earlier (parameters chaining for security reason). Its form is different from one presented in Eq. (4).

As mentioned above, the second type security payload is a special purpose security parameter that is used internally for improving protocols self security. Without such counter measurement, every parameter (token or MOS score) that is transmitted inside covert channel may be susceptible to, e.g., modifications or other types of attacks. To prevent such situations every *n-th* parameter is used to authenticate and provide integrity of *n-1* parameters that were transferred earlier. The general idea of its calculation is presented in Fig. 4.

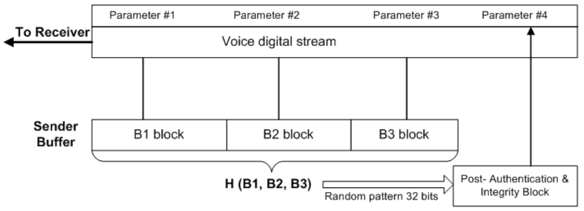

Figure 4. Example of authentication and integrity mechanism for transmitted parameters

In our AVoIP system we will transfer mainly MOS scores between calling parties. Besides that security parameters may be transferred if the network conditions allow it. So, the presented in Fig. 4 mechanism may be used also to secure MOS scores that they will be not tampered during transmission.

To summarize, by utilizing abovementioned lightweight security solution we gain from the security perspective:

- Authentication of the data source (one can be sure of the identity of the caller),
- Authentication of the signalling messages (one can prove that the caller is the source of the signalling messages that were exchanged during the signalling phase of the call),
- Signalling messages integrity (one knows that the signalling messages were not modified during the transmission through the communication channel)
- Data authentication – integrity (one can be sure that the audio comes from the caller and it has not been tampered).

5.3 Security Adaptation in AVoIP

We will now present how the security level may be adjusted according to the network conditions and values of other adaptively changed parameters of the VoIP call.

First of all, we want to emphasis that in AVoIP we do not allow to communicate with VoIP service without any security measures. Nowadays, many commercial VoIP products do not provide any security mechanisms at all as they degrade quality of the call and increase the cost of the hardware or software implementations.

In AVoIP we begin the call with the highest level of security that is available for end-users (highest means cryptographically strong). If the congestion in network occurs we first try to adapt VoIP call parameters (output rate of speech codec, playout buffer size and FEC information) to stabilize quality. Thus, we may modify the Eq. (1) as security mechanisms affects perceived quality:

$$QoS_{Perceived} = F(Ns, Pbs, Cc, FEC, SEC) \qquad (4)$$

where
SEC denotes security mechanisms applied to VoIP traffic.

When the network conditions still decrease quality of the call and the adaptation of the call parameters do not improve it then switching to lightweight security solutions (like presented in Section 5.2) takes place. This way we always gives up security of the call at the end that is after we tried to affect the quality of the call by adjusting call parameters. That is why, the control algorithm (Algorithm 1) may be modified as follows:

Algorithm 2. Control algorithm for call quality drop scenario with security adaptation.

```
(1)    i = 0
(2)
(3)    Do{
(4)      MOS_W = ExtractWM(Received_Audio)
(5)      MOS_R = CalculateMOS(Received_Audio)
(6)
(7)      If (MOS_R) <= (Threshold)
(8)       StartTimer(Threshold_Timer)
(9)
(10)     MosTable[i] = MOS_R
(11)
(12)     If (Threshold_Timer) >= (Max_Timer)
(13)      {
(14)       AvLoss = CalcAvMOS(MosTable[])
(15)       If (AvLoss) <= (Threshold)
(16)        {
(17)        StopTimer(Threshold_Timer)
(18)        AdjustOutputAudioRate(C1,C2)
(19)        AdjustPlayoutBuffer(size)
(20)        AdjustFEC(size)
(21)        i = -1
(22)        Free(MosTable[])
(23)       }
(24)      }
(25)    If (Codec == C2) and (PlayoutBuffer ==
(26)   min)     and (FEC_info == min)
(27)    SwitchToLightWeightSecurity()
(28)
(29)    i = i + 1
(30)   }while (conversation_lasts)
```

As one can see in lines 25-27, if the limits of VoIP call parameters adaptation are reached then the AVoIP system changes from previous security mechanism to lightweight one. This change is the last chance to save perceived quality of the call. When this operation does not improve quality and the network conditions deteriorate then the call will likely be broken.

6. Conclusions and future work

In this paper we presented AVoIP system which may be utilized to adjust call parameters during the conversation to improve the perceived quality for end-users. The parameters that may be adjusted include: output rate of the speech codec, playout buffer size and amount of information for FEC mechanism. We also proposed the speech control algorithm which is based on the perceived quality score (expressed in MOS scale). Additionally, information about call quality is sent inside the audio stream with a use of audio watermarking techniques. Finally, we combined speech quality control and providing security into one solution. We showed that adapting VoIP parameters may not be enough if the security mechanisms are used which add excessive delay and impose overheads. Providing perceived quality and security are related and should be addressed simultaneously.

Future work will be focused on experimental confirmation of certain values for proposed AVoIP solution: especially control mechanism's parameters: threshold and the threshold timer values should be evaluated.

References

[1] R. Baden et al., "Integrated services in the Internet architecture: An overview,", Tech. Rep. IETF RFC. 1633, June 1994.

[2] S. Blake et al., "An architecture for differentiated services,", Tech. Rep. IETF RFC. 2475, December 1998.

[3] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson. "RTP: A Transport Protocol for Real-Time Applications", IETF, RFC 3550, July 2003.

[4] C. Hoene, H. Karl, A. Wolisz. „A perceptual quality model intended adaptive VoIP applications". *International Journal of Communication Systems*, Wiley, August 2005

[5] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston - SIP: Session Initiation Protocol", IETF, RFC 3261, June 2002

[6] ITU-T Recommendation H.323: Infrastructure of audiovisual services – Systems and terminal equipment for audiovisual services - Packet-based multimedia communications systems version 6, Telecommunication Standardization Sector, ITU-T, June 2006

[7] F. Cuervo, N. Greene, A. Rayhan, C. Huitema, B. Rosen, J. Segers - Megaco Protocol Version 1.0, IETF, RFC 3015, November 2000.

[8] ITU-T Recommendation P.862, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs.

[9] ITU-T Recommendation G.107, "The E-model, a computational model for use in transmission planning."

[10] ITU-T Recommendation G.108, "Application of the E-model: A planning guide."

[11] ETSI EN 301 704 V7.2.1 (2000-04), Digital cellular telecommunications system (Phase 2+); "Adaptive Multi-Rate (AMR) speech transcoding" (GSM 06.90 version 7.2.1 Release 1998)

[12] Z. Qiao, L. Sun, N. Heilemann, E. Ifeachor, "A new method for VoIP quality of service control use combined adaptive sender rate and priority marking", *IEEE International Conference on Communication*, Vol. 3, Page(s): 1473 – 1477, 20-24 June 2004

[13] C. Casetti, J. C. De Martin, and M. Meo, "A framework for the analysis of adaptive voice over IP,"

presented at the ICC2000, New Orleans, LA, June 18–22, 2000.

[14] C. Padhye, K. Christensen, W. Moreno, "A New Adaptive FEC Loss Control Algorithm for Voice Over IP Applications", *IEEE Computing and Communications Conference*, IPCCC '00. pp. 307-313, 2000

[15] M. Narbutt and L. Murphy, "VoIP Playout Buffer Adjustment Using Adaptive Estimation of Network Delays", *In Proc. 18th Int. Teletraffic Congress* (ITC-18), Elsevier, 2003, pp. 1171–1180

[16] R. Ramjee, J. Kurose, D. Towsley, H. Schulzrinne. "Adaptive playout mechanisms for packetized audio applications in wide-area networks", *In Proceedings of IEEE Infocom*, pp. 680–688, Toronto, Canada, 1994.

[17] J. Pinto, K. Christensen. "An algorithm for playout of packet voice based on adaptive adjustment of talkspurt silence periods", *In Proceedings of the IEEE 24th Conference on Local Computer Networks (LCN)*, pp. 224–231, Lowell, MA, USA, 1999.

[18] S. Moon, J. Kurose, D. Towsley. "Packet audio playout delay adjustments: performance bounds and algorithms". *ACM/Springer Multimedia Systems*, 27(3):17–28, 1998

[19] W. Mazurczyk, Z. Kotulski, "New VoIP traffic security scheme with digital watermarking", *In Proceedings of The 25-th International Conference on Computer Safety, Reliability and Security SafeComp 2006*, Lecture Notes in Computer Science 4166, pp. 170 - 181, Springer-Verlag, Heidelberg 2006, ISBN 978-3-540-45762-6

[20] F. Beritelli, S. Casale, G. Ruggeri, "Performance Comparison Between VBR Speech Coders for Adaptive VoIP Applications", *IEEE Communications Letters*, Vol. 5, No. 10, October 2001

[21] N. See Leng, H. Simon, D. Singh, "Effectiveness of adaptive codec switching VOIP application over Heterogeneous Networks", *IEEE 2nd International Conference on Mobile Technology, Applications and Systems*, pp. 1-7, November 2005

[22] W. Bender, D. Gruhl, N. Morimoto, A. Lu. "Techniques for data hiding", *IBM. System Journal*,. vol. 35, Nos. 3&4. pp 313-336, 1996.

[23] T. Takahashi, W. Lee, "An Assessment of VoIP Covert Channel Threats", *In Proceedings of The 3rd International Conference on Security and Privacy in Communication Networks (SecureComm'07)*, Nice (France) 17-21 September 2007

[24] F. De Rango, M. Tropea, P. Fazio, S. Marano, "Overview on VoIP: Subjective and Objective Measurement Methods", *IJCSNS International Journal of Computer Science and Network Security*, Vol.6 No.1B, January 2006

Author Biographies

Wojciech Mazurczyk received the B.S. and M.S. degrees in Telecommunication from Warsaw University of Technology, Faculty of Electronics and Information, Institute of Telecommunication in 2003 and 2004, respectively. Currently he is a Research Assistant at Warsaw University of Technology and is finishing his thesis about evaluating information hiding techniques (digital watermarking and steganography) for improving Voice over IP service security. Member of Network Security Group at Department of Electronics and Information Technology of Warsaw University of Technology, Poland.

Zbigniew Kotulski received his M.Sc. in applied mathematics from Warsaw University of Technology and Ph.D. and D.Sc. degrees from Institute of Fundamental Technological Research (IPPT PAN). He is currently professor at Institute of Fundamental Technological Research of the Polish Academy of Sciences and professor and head of Security Research Group at Department of Electronics and Information Technology of Warsaw University of Technology, Poland.