A LOW COST AND OPEN-SOURCE SOLUTION FOR END-TO-END SECURE CALLS OVER VOLTE

Sebastian CIORNEI[1], Ion BOGDAN[2], Luminita SCRIPCARIU[3], Mihai CALIN [4]

Voice calls on cellular networks, including LTE standards and upgrades are encrypted only over the wireless link between the Mobile Equipment and the Base Station. A secure end-to-end voice and data transfer is left under question. This paper brings this end-to-end encryption of voice over VoLTE audio channel, namely HDVoice, without depending on data channels and their associated disadvantages. A step-by-step methodology on how to use the HDVoice channel to create secure calls is detailed. A realistic, open source and easy deployable solution is being presented, along its proof of concept using the appropriate codecs of the HDVoice, namely AMR-WB.

Keywords: HD voice, secure end-to-end voice communication, VoLTE, modem, LTE, wireless communication

1. Introduction

Wireless data communication technologies became the most common medium for voice communication in the recent decade with over 6 billion subscribers worldwide. The world’s first worldwide adopted cellular phone standard, LTE, was lacking the voice channels, requiring the operators to have two networks, one for data, and another one for the voice. VoLTE not only overcomes this drawback, but also adopts HDVoice as the voice standard. These voice calls, just like the previous 2G-3G standards, are encrypted only over a section of their path, which is the wireless link between the Mobile Equipment (ME) and the Base Station (BTS/NB/eNodeB). As per the 3GPP standards, the path from the caller’s BTS to the destination’s BTS is not required to enforce encryption. Moreover, it has been proven that even the ME-BTS link cannot be considered secure in some cases [1, 2].

[1] Faculty of Electronics, Telecommunications and Information Technology, Technical University ‘Gh. Asachi’ of Iasi, Romania, e-mail:sciornei@ieee.org

[2] Faculty of Electronics, Telecommunications and Information Technology, Technical University ‘Gh. Asachi’ of Iasi, Romania

[3] Faculty of Electronics, Telecommunications and Information Technology, Technical University ‘Gh. Asachi’ of Iasi, Romania

[4] Electrical Engineering Faculty, University POLITEHNICA of Bucharest, Romania, e-mail: calinmihai@ieee.org

Where there is a strong need for privacy, an end-to-end (ME to ME) encryption is the only solution, especially when the calls are to or from different network operators in a country or area with security not at an acceptable level [3, 4]. Most of the current secure voice communication solutions depend on a data connection and on fulfilment of network parameters required for a voice communication, such as: reliable, with quick and smooth handover between cells, with Forward Error Correction (FEC) and, most important, low latency. The data channels have been designed for other scopes/needs and do not provide these desired parameters. Even if there would be such a data communication channel, the network operator would have to decrease its parameters due to business reasons (the subscribers would not use voice subscription and might go instead for much cheaper over-the-top options (e.g. Viber/Skype/Line/Gtalk) [5-7].

Examples of previous trials for end-to-end secure communication have been focused on GSM networks [8, 9]. In [8] the authors modulated encrypted data into speech like waveforms, but the whole processes adds a high delay, about 135 ms, on top of the GSM delay and the used modulation techniques and codecs are proprietary and unavailable. In [9] the authors present a more general approach that can be adapted to various codecs by using genetic algorithms to generate the waveforms. However, one of the main drawbacks of this solution relates to its complexity, while its deployment for different processing architectures needs a re-implementation of the required genetic algorithm libraries, making this option unapproachable for a regular mobile.

The solution proposed in this paper differentiates from the previous approaches by taking advantage of the already available networks using AMR-WB codec [10], while ensuring a realistic, implementable and easy deployable option on available smart phones. The software implementation uses mostly already available open libraries. It is also shown how the technological advances in the past years from AMRNB to AMRWB make a big difference.

The rest of the paper is structured as follows: Part 2 presents a detailed description of the proposed modem solution; Part 3 describes the implementation methodology; Part 4 underlines the performance characteristics of the solution through simulation results; Part 5 concludes the paper and discusses the future research directions.

2. VoLTE Modem and transmission flow

The focus of the work is on the transmission of voice signals. This imposes a series of hard and soft (relaxed) constraints related to: latency, transmission errors and dropped frames.

Hard constraints are:

1) *The maximum mouth to ear (end-to-end) latency should not exceed 300ms*, ideally up to 200ms, especially for highly compressed vocoders[11]. This maximum latency is calculated as the sum of the delays made by the transmission chain, D_{tx}, by the channel, D_{ch}, and by the receiving chain, D_{rx}:

$$D_{tx} = D_{A/D} + D_{cod} + D_{enc} + D_{mad} \tag{2.1}$$

$$D_{ch} = D_{AMRWB} + D_{ntw} \tag{2.2}$$

$$D_{rx} = D_{dmad} + D_{denc} + D_{dcod} + D_{D/A} \tag{2.3}$$

where, $D_{A/D}$ is the delay of the analog to digital convertor; D_{cod} is the delay of the inner-codec; D_{enc} is the delay of the encryption; D_{mod} is the delay of the modulator; D_{AMRWB} is the delay of the VoLTE standard vocoder, AMRWB; D_{ntw} is the delay introduced by the network; D_{dmod}, D_{denc}, D_{dcod} and $D_{D/A}$ are the corresponding delays introduced by the parts of the receiving chain.

2) *The error corrections cannot be done with retransmission* due to the latency constraint;

3) Due to the latency constraint *the FEC has to be either removed or done with small packets sizes*;

4) *A proper balance should be set between the length of the inner-codec frame and the bit rate of its output* (e.g. shorter frames introduce lower delay, but decrease the compression rate).

<u>Soft/relaxed constraints</u> are:

1) *Errors up to a threshold where the voice remains intelligible are accepted*;

2) *The threshold for the maximum number of dropped frames is set up such that the voice remains intelligible.*

Transmission Chain

Let us consider a call to be made and to represent the analog voice as a signal $s(t)$ which is to be transmitted over a wireless communication channel. This signal is expected to be in the bandwidth of 300-3400Hz, the telecommunication standard for the audio bandwidth required to make most of the human voice intelligible. Fig. 2.1 gives the flow of the transmission chain.

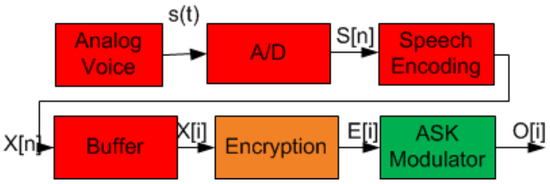

Fig. 2.1. Transmission Chain

where,

- **A/D** states for the analog to digital converter, which samples the continuous time signal $s(t)$ into a discrete signal $\boldsymbol{S}[n]$. The A/D converter has a predefined and codec dependent sampling rate (e.g. of 8kHz and the quantization of 16 bits/sample which are the preconditions for the input in Codec2 [12]).
- $\boldsymbol{S}[n]$ is the output of the A/D converter and has a bit rate of 128000 bps with $n \in \boldsymbol{N}$**,** where $\boldsymbol{N}$ is the total number of frames coming out of the A/D converter. Each frame has a pre-defined length imposed by the upcoming inner codec.
- **Speech Encoder (Inner codec)** is a high compression vocoder to be used for compressing the voice samples.
- $\boldsymbol{X}[n]$ is the encoded frame corresponding to $\boldsymbol{S}[n]$. Note that the length in bits of $\boldsymbol{X}[n]$ is usually different than the length of $\boldsymbol{S}[n]$ and the bit rate compression ratio is defined as

$$R_{codec,\mathrm{mod}e} = \frac{length(S[n])}{length(x[n])} \tag{2.4}$$

In the case of Codec2 the ratio is between 64 and 107.

- **Buffer** block is used to transform the inner-codec output frame, $\boldsymbol{X}[n]$, into a bitstream, $\boldsymbol{X}[i]$, to be used by the next block (Encryption).

Encryption block is a stream cipher that makes bit by bit XOR between the output of the buffer block (plaintext streaming) and a keystream (pseudorandom cipher stream). The output of the encryption block (ciphertext), $\boldsymbol{E}[i]$, is given in (2.5), where, $\boldsymbol{K}[i]$, is the keystream.

$$E[i] = X[i] \oplus K[i] \tag{2.5}$$

While the modem does not enforce a specific encryption method, the stream cipher was selected as the best choice due to that fact that it was designed for continuous flows, it is designed to work on communication channels with high probability of errors, and due to the fact that it does not add additional delay.

Modulator block

Let us consider a symbol coding alphabet A_N, with N the number of symbols. For an efficient coding, N is selected as a multiple of 2. The incoming ciphertext bitstream is divided in groups of length $N_{\text{bit}}=\log_2(N)$ by the Serial-to-Parallel block. The bit error rate of the receiver increases with the number of symbols in the constellation. When a high bit rate is desired, error correction codes are also required. Therefore, the increase of the bit rate introduces additional delay.

The modulation techniques which have been studied were: FSK (frequency shift keying), PSK (phase shift keying), ASK (amplitude shift keying) and OOK (on off keying). Being a vocoder, the entire signal is being regenerated fully at the reception, and it is expected that AMRWB will neither maintain the phase nor the frequency of the input signal. For this reason, the main focus has been taken into the ASK and OOK. After the simulations performed, it has been identified that for the AMRWB the best option is ASK with a0=0.3 and a1=0.91.

As shown in Fig. 2.1, the voice signal at the output of the ASK is a bit by bit *O*[*i*], and is calculated as:

$$O[i] = \sum_{m=0}^{N_{spb}} M_{ASK}(E[i,m]) \cdot \sin[2\pi f_c \cdot (t - m \cdot T_s)] \qquad (2.6)$$

where, $M_{ASK}(\cdot)$ is the ASK modulator, f_c is the carrier frequency; *m* is the sampling index; T_s is the sampling period, with $T_s = \frac{1}{F_s}$, and F_s is the sampling frequency; $N_{spb} = \frac{T_b}{T_s}$, with T_b being the bit period.

Starting from the Nyquist criterion, we will use:,

$$F_s > 2 \cdot f_c + BW \qquad (2.7)$$

where, *BW* is the bandwidth required for the base band of the modulated signal.

Communication Channel

HD voice term took ground in the recent years as an umbrella referring to technologies that enhance the quality of an audio call, or the audio portion of a video call. In order to sustain HD voice communication, the audio channel for the base stations needs to be widened such that to include the AMRWB (Adaptive Multi-Rate Wideband) codec [13-15]. This codec has an audio bandwidth of 50-7000 Hz, 14 bit samples, and internal sampling rate of 12.8 kHz [16].

The decoded signal received by the destination Mobile Station (MS) is affected by the transfer function of the channel, with a general form, *T*(·):

$$Y_t[k] = T(O[k], Enc_{made}, \Psi, Er_{net}, D_{ntw}) \qquad (2.8)$$

where,

- $Y_t[k]$ is the output of the AMRWB decoder (at the receiver) and Ψ is the space of internal states of the vocoder;
- *O*[k] is the input in the vocoder, at the transmitter, containing all the bits within a period of the frame *k*. This is to explain that each bit transmitted through the vocoder is dependent on the bits passed in the last 20ms as well as the bits in the 5ms after (“look-ahead” feature).
- Enc_{mode} is the vocoder’ s mode (defined by the desired bit rate)*;*
- Er_{net} is the rate of errors introduced by the cellular network.
- D_{ntw} is the delay introduced by the cellular network .

The cellular network might also drop some of the frames. For the analytical form of the output the drastic situation of frames being dropped is only temporary, until the network will switch to a vocoder mode with lower bit rate. Assuming that both the delay (D_{net}) and the bit error rate (Er_{net}) introduced by the network can be considered outside of the transfer function of the audio channel, $T(\cdot)$ can be identified with the transfer function of the vocoder, $T_v(\cdot)$. Note that in this assumption the delay will be added back after the detection block, while performing a recoding (encoding followed by a decoding). Therefore, one may write,

$$Y_v[k] = T_v(O[k], \Psi, Enc_{made}) \tag{2.9}$$

Note that the vocoder is a lossy codec, the output of such a process will be of lower quality than the input, and therefore the decoding function is not the inverse of the encoding.

Reception channel

Reception channel is the mirror of the transmission channel and with similar components such as: demodulator/detector, decryption block, inner codec decoding and the D/A block. Their operation is similar to their mirror blocks as transmission channel presented above; therefore, they will not be detailed in this section. To be noted however, for the *Demodulation & Detection block* the appropriate method has been designed according to the modulation technique. Both ASK and OOK require an adaptive threshold or automatic gain control (AGC) in order to ensure an optimal threshold setting. Based on the modulation selected the detection can be changed. The block performs the following steps:

- From the samples of each bit, it drops the first ***J*** samples (with ***J*** being between one and five). These are being dropped because in the transition cases they bring too much information from the previous bit, affecting negatively the detection.
- From the rest of the samples an average is being computed.
- The average is being compared to a threshold in order to decide between the symbols sent (in both binary ASK and OOK the symbols are 0 and 1).

3. Implementation

The high level architecture of the proposed and implemented solution is described in Fig. 3.1. This architecture allows sending compressed voice. The current implementation uses Codec2 for compressing the voice, without encapsulation in a VoIP container (there are no headers). This is important because repeating headers, in combination with a problematic ciphering scheme, would render an insecure design, as it did in the initial 802.11 standard [17].

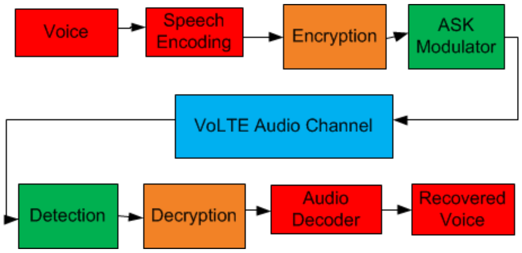

Fig. 3.1. High level architecture

The proposed modem has two modes: one at the beginning of the call, the first few seconds for protocol negociation between the two devices engaged in the call; and the main one with the voice transmission itself.

As our target is to prepare the prerequisites of a reliable encrypted voice communication and not a high quality of voice. Therefore, the selection criteria for the audio codec were based on low delay, robustness (both error tolerance and no artefact on background noise), availability for review – open source (ideally royalty free) and a good voice quality at very low bit rates (1200-2000 bps).

From authors' research on available codecs it resulted that the codec that perfectly balanced all these conditions was the free & open source Codec2, developed by David Rowe [12]. This is a Linear Predictive Vocoder, with the following tech specifications: 40 ms frame size and a total latency of 70 ms at 1200 bps, and 25 ms frame size and 50 ms at 2000/2400 bps. According to the Mean Opinion Score (MOS) tests performed, and the total codec delay calculations, Codec 2 has similar or better quality and delay compared to the military grade codecs like MELPe [18, 19] or similar [20].

A number of modulation techniques, such as FSK, PSK, ASK and OOK (on/off keying) have been tested before opting for the ASK as mentioned in sub-Section 2.1. Due to the way vocoders work, PSK and FSK signals suffer a big level of distortions after passing through the AMRWB codec. An example for the PSK case can be seen in Fig. 3.2.:.

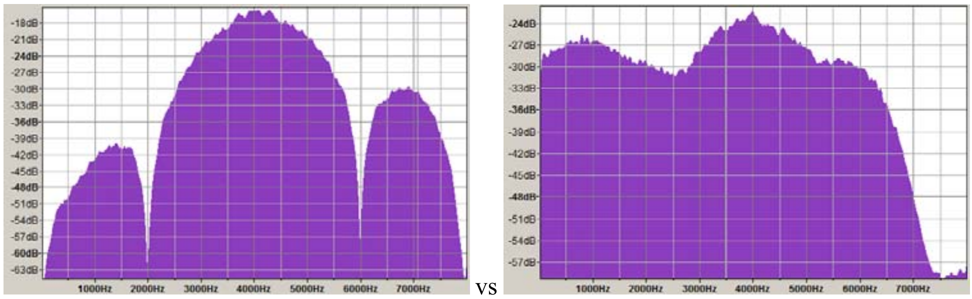

Fig. 3.2. PSK Spectrum at Transmission (left) vs Reception (right)

Binary ASK choice was led by reasons such as keeping the error rate introduced by the vocoder in the limits by using a modulation with the highest possible separation between the symbols after passing the AMRWB.

A snapshot showing the detection, with modem set on a bit rate of 4000bps, carrier frequency=4000 Hz, AMRWB mode 23850 and threshold of 0.04 can be seen in

Fig. 3.. In the figure, the red binary signal represents the transmitted bit-stream, the received audio signal is depicted with blue in background, and the received audio signal averaged per bit period is depicted in black, while detection threshold is the horizontal line.

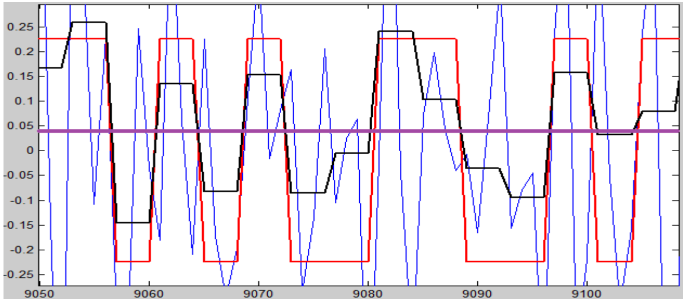

Fig. 3.3. Matlab Simulation Snapshot

The HD Voice Audio Channel used for testing the proposed solution is a virtual “software” channel that was deployed, on top of the radio channels. The codec recommended by 3GPP for IP Multimedia Subsystems IMS (MTSI) and HD Voice in general, is AMR-WB[13, 14]. It has an audio bandwidth of 50-7000 Hz, 14 bit samples, and internal sampling rate of 12.8 kHz[16].

In order to pass through the AMRWB codec with the lowest attenuation possible [21] the carrier frequency was set to 6kHz. The current implementation used the FFMpeg project to simulate the AMRWB software channel. The project uses VisualOn and OpenCORE standard implementations of AMRWB on various platforms, including Android phones. FFMpeg is being used also for the upsampling/downsampling in the tests where the wav formatted data from Matlab has a higher sampling rate than the 16kHz required by the AMRWB encoder.

For the demodulation/detection block the number of dropped samples ***J*** was empirically chosen to one.

The calculation of delays takes into account both, the proposed system as well as the transport over the LTE communication channel. The LTE network from the same operator, has a delay of about 23 ms (standard states up to max of 50-70ms [10]). In the LTE real scenario, the network latency data taken into account has been retrieved from [22]. As an example, the delay budget when there is no-FEC over the inner audio channel on a LTE network summarizes as follows: *Codec2* 2000/2400 total delay of 50ms, the AMRWB delay of 25ms and the 4G

RTT network delay of 23ms, thus totalling 86.5ms (and a maximum of 100-110 ms).

The Operating System used was a Windows7 64 bit, Matlab 2013a 64 bit, Audacity 2.0.5, Cygwin 1.7.9 64 bit, FFMpeg versions 0.9-2.2 64 bit. During the implementation a series of ***C*** files and scripts have been created as well in order to call FFMpeg from Matlab. These files have been released by us as an open source project called ffmpeg_matlab on github. This project also takes the challenge to extend windows Matlab's capability to handle audio codecs/formats from 5-10 (as of 2014) to hundreds, by creating mex files for interfacing FFMpeg from Matlab with highest performance and flexibility.

4. Simulation Analysis

The system proposed has been tested with various parameters, like different AMRWB modes, different Codec2 bit rates, different lengths of the call, different sampling rates and network delays. The results are being classified in 3 categories: BER, Delay calculations and Threshold values.

Bit Error Rates (BER)

Low computation requirements, robust and simplicity of the implementation have been the guidelines of this work. For this reason the followings have been used as viable selections in the proposed end-to-end secure voice communication solution: NRZ line coding, binary ASK modulation, low computation detection. It has been demonstrated that by using the above led to BERs good for secure voice transmission, as can be seen in Fig. 4.1.:

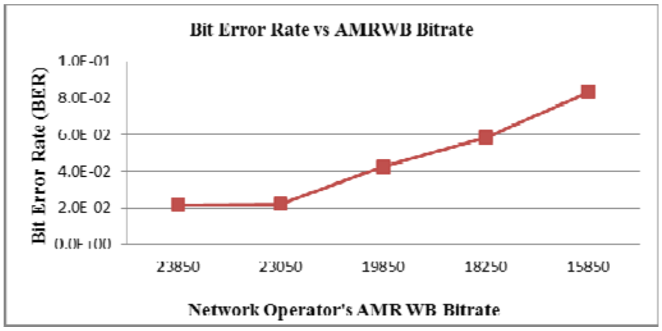

Fig. 4.1. BER vs AMRWB Bit rate

During the tests, the payload to be sent via the modem has been created by encoding audio samples from the ITU standard set of samples[23] for both modes (1200 and 2000 bps) of Codec2. As per the tests made, the BER rates on the AMRWB modes quality modes (23850 bps, 23050 bps or even 19850 bps) have kept the audio samples at good quality, within Codec2's power to cope with errors, in both 1200 and 2000bps modes. The difference up to modem's capacity of 4000bps can be used either for higher quality codecs or for data exchange required for live synchronizations of the encryption protocols during the call.

In order to test the BER's stability, it has been tested in calls of various durations. Fig. 4.2 shows that the solution achieves 4000bps, double from OOK solution in [15], while keeping the BER rates stable, around the value $2.3 \bullet 10^{-2}$, a value tested and is acceptable for Codec2.

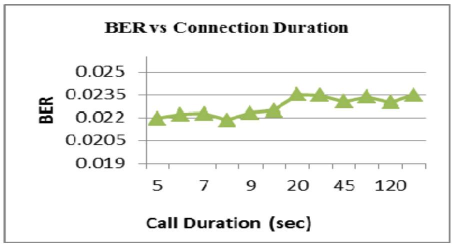

Fig. 4.2. BER vs Connection Duration

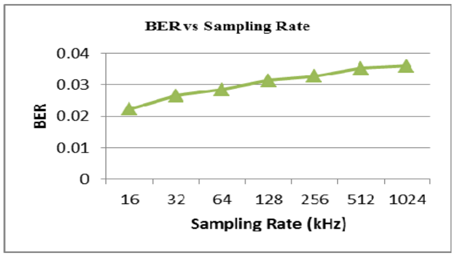

Fig. 4.3. BER vs Sampling Rate

Further tests have been done to check the best sampling rate at which the audio signal to be passed to the outer codded should be processed. Fig. 4.3 shows that the BER increases with the sampling rate. This is due to the fact that the AMRWB codec requires 16 kHz as input and when the data is transferred from modulator (simulated in Matlab) to the AMRWB encoder, a down sampling is required. As there have been good result with 16kHz sampling, and the fact that additional filtering is expensive from both delay and computation point of view, it was concluded that the 16kHz is the best choice for processing the modulated signal.

Delay Calculations

According to ITU-T G.114 standard, for a good speech communication, the one-way delay needs up to 150 ms and maximum 200 ms[11]. As can be seen in Fig. 4.4, in all the cases tested the delay of the proposed solution ranges from 85 ms to 130 ms in real world LTE scenarios, therefore within the range of the recommendation.

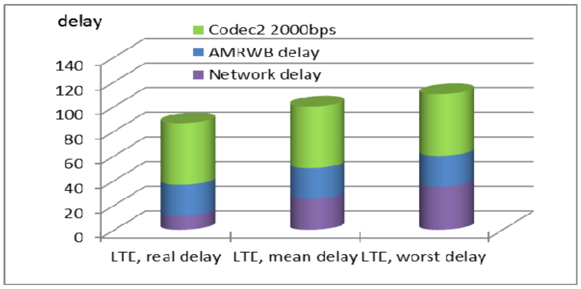

Fig. 4.4. Delay evaluations

In the situation the Codec2 1200bps is selected, the delay increases with another 20ms, required for a stronger compression. Even in this situation, the delay remains within the limits recommended by ITU standard.

Detection Threshold

As per Fig. 4.5, with the modem at 4000bps, the threshold value for detection remains stable along the tests.

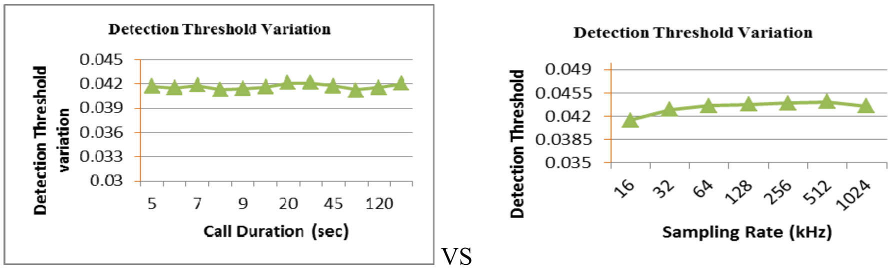

Fig. 4.5. Detection Threshold vs Sampling Rate and Call Duration

This suggests that the detection can even be simplified by using a constant value for codec/mode case, currently: 0.040.

5. Conclusions

This paper presented an innovative, open-source based implementation of a modem over VoLTE audio channels. Such modem allows the transmission of encrypted voice calls, even in the situations where there is no data connection. The codecs used, the libraries and the associated 3rd party tools are all open source making the implementation of this modem available without limitations. This paper contributes with the followings: (a) overcome standard's lack of secure end-to-end data transfer solution (across network operators), even when only voice channels are available; (b) allow an ubiquitous encrypted end-to-end voice communication; (c) have a realistic and easy implementation and deployable solution by relying only on software installable on over the counter smart phones; and (d) be future proof, by relying on voice calls instead of data connections and by keeping network operator's businesses safe (profitability of network operator's is ensured). Using the proposed ASK parameters, the speed rate has been doubled, compared to OOK, while keeping the same BER. Next steps of our research are towards further increase of the data rates to allow even higher voice quality.

Acknowledgments

The authors would like to thank Dr. David Rowe for providing the open and freely available Codec2, as well the open source communities of FFMpeg, Gnumex, Audacity, VisualOn, OpenCORE, MinGW and Cygwin.

REFERENCES

[1]. *S. Ciornei and I. Bogdan*, GSM security-attacks and protection methods. Part II. , PhD 3rd Report: Technical University of Iasi, Dept. of Electronics and Telecommunication, Oct. 2008.

[2]. *C. Blanchard*, "Security for the Third Generation (3G) Mobile System," Network Systems & Security Technologies, www.isrc.rhul.ac.uk/useca/OtherPublications/3G_UMTS%20Security.pdf, Sept. 2000].

[3]. *L. Tae-Ho and C. Taesang*, "Self-powered wireless communication platform for disaster relief," 13th Asia-Pacific Network Operations and Management Symposium (APNOMS), Taipei, Taiwan, 21-23 Sept. 2011, pp. 1-3.

[4]. *A. Anand, V. Pejovic, E. M. Belding, and D. L. Johnson*, "VillageCell: cost effective cellular connectivity in rural areas," Fifth International Conference on Information and Communication Technologies and Development, Atlanta, Georgia, 12 March 2012, pp. 180-189.

[5]. *EurActive.com*. "EU regulators say telecoms block Skype," www.euractiv.com.

[6]. -. "Mobile operators seek to 'block' Skype in Sweden," www.thelocal.se/39938/20120328/.

[7]. *A. C. a. M. Authority*, Communications report, ACMA, 2013.

[8]. *N. Katugampala, V. Al-Naimi, S. Villette, and A. Kondoz*, " Real time end to end secure voice communications over GSM voice channel," 13th European Signal Processing Conference, Antalya, Turkey, 4-8 Sept. 2005, pp. 1-4.

[9]. *V. V. Sapozhnykov and K. S. Fienberg*, “A Low-rate Data Transfer Technique for Compressed Voice Channels,” Journal of Signal Processing Systems, vol. 68, no. 2, 2012, pp. 151-170.

[10]. *L. Norell, E. Parsons, and P. Synnergren*, Telephony services over LTE end-to-end, vol. 1, Ericsson Review, 2010.

[11]. *R. ITU-T and I. Recommend*, “G. 114,” One-way transmission time, vol. 18, 2000.

[12]. *D. Rowe*, "Codec 2 - Open Source speech coding at 2400 bit/s and below," www.tapr.org, 2011].

[13]. *GSMAssociation*, "Adaptive Multirate Wide Band, IR.36," 2011.

[14]. *3GPP*, "TS 26.114," IP Multimedia Subsystem (IMS), 2012.

[15]. *S. Ciornei, I. Bogdan, and L. Scripcariu*, "HD voice modem for end to end secure call," Telecommunications Forum (TELFOR), 2012 20th, 2012, pp. 580-583.

[16]. *3GPP*, "3GPP TS 26.171 " Adaptive Multi-Rate - Wideband (AMR-WB) speech codec, 2011.

[17]. *S. Ciornei, I. Bogdan, and L. Scripcariu*, "ON 802.11 Standard and the WiFi Network Security," Third European Conference on the Use of Modern Information and Communication Technologies ECUMICT Gent, Belgium, 13-14 Mar. 2008, pp. 89 – 98.

[18]. "MELPe - Enhanced Mixed-Excitation Linear Predictive Vocoder," www.compandent.com.

[19]. "MELPe vocoder," www.dspini.com/dspini_melpe.htm.

[20]. *I. Cable Television Laboratories*, "Codec and Media Specification," 2012.

[21]. *3GPP*, 26.976 - Performance characterization of the Adaptive Multi-Rate Wideband (AMR-WB) speech codec, 2012.

[22]. *EmpitiroLtd*, LTE ‘Real World’ Performance Study. Broadband and Voice over LTE (VoLTE) Quality Analysis, Report: TeliaSonera, Turku, Finland, Mar. 2011.

[23]. *ITU*, "ITU-T P Series: Telephone transmission quality, telephone installations, local line networks," Recommendation ITU-T P.501: Test signals for use in telephonometry.