AVT
Internet Draft
Intended status: Informational
Expires: February 2010

Christian Hoene
University of Tuebingen
August 17, 2009

Requirements of an Audio Communication System (ACS)
draft-hoene-avt-acs-requirements-00.txt

Status of this Memo

Copyright Notice

Abstract

This document describes the requirements of an audio communication system (ACS) for acoustic content, especially speech and music. The ACS consists of all components above the IP layer and below a digital PCM audio interface. These include codec, jitter buffer, and transport.

The goal of the ACS is to provide a bidirectional acoustic communication between any two Internet hosts at a good quality, constrained only by the available resources at the hosts and the characteristics of the transmission path between both hosts.

The intention of the document is to provide the requirements for a codec that is solely intended for the Internet, to provide the requirements for the codec's payload specification, and to define the requirements on the transport protocol.

Table of Contents

1. Introduction

This document is based mainly on the discussions on the Codec BOF mailing list, which took place in 2009. It also based on the internal requirement documents of ITU-T G.718 [SG16 314-WP3], on the ITU-T G.719 standard, on the 3GPP document [TS26.114-830], and on existing IETF codec drafts.

It is intended as basis of a requirement document that should lead to the design of an audio codec for the Internet. However, this document

address the requirements of the entire system not only of a single component because we want to ensure that the system as a hole works well not only some parts of it.

We introduce the term audio communication system to describe the parts of an IP based telephone which are care for the bidirectional transmission of acoustic content between two Internet hosts. These include the encoder, the payload encapsulation, guide lines on how to use transport protocols (RTP, UDP, TCP, DCCP), the playout buffer, the decoder, the concealment of packet loss, time adjustments, changes of encoding parameter, and various mechanisms to manage, control and monitor the acoustic transmission.

The ACS is intended mainly for the use on the public Internet and should be as easily distributable as most other Internet protocols that run on virtually all kind of devices and on all kind of communication links. Also, the ACS shall be affordable by all humans that have Internet access. If possible, it should be royalty free and available as open-source software. If these requirements are given, then the ACS can fulfill its goal of providing acoustic transmission between _any_ two Internet hosts.

1.1. Basics Architectural Guidelines of the Public Internet

The ACS is intended for the public Internet and follows similar architectural design guidelines as those which are valid for other Internet protocols, too. These include:

- End-to-end semantics saying that transport protocol units are transmitted from one end (an Internet host) to the other end without any intermediate changes.

- Network neutrality.

- Best effort service that try to transmit packets as good as possible but that cannot guaranty any minimal transmission bandwidth or maximal transmission delay. Instead one has to cope with any end-to-end transmission quality that is provided.

- Congestion control to prevent congestion collapse of the Internet (such as TCP or DCCP). Typically, TCP controls the number of packets that are sent during periods of congestion. Thus, one has to consider that the number of packets per second might be an important constraining limitation and not only the bits per second.

- Internet protocols are scalable to wide degree. They work on links having a very low bandwidth (in the order of bits per second) and with very high bandwidth (in the order of gigabits per second). The transmission latency can range from microsecond up to seconds. Also, the Internet hosts might have very low processing and memory capabilities (such as an 8-bit micro controller). However, even then they can communicate with any other hosts. Flow control (such as in TCP) is used to cope with hosts that have limited resources.

- Functions to help monitoring the communication (such as the features provided ICMP)

- The most important Internet protocols can be used without paying royalties.

- The public Internet allows global communication between any two hosts connected to the public Internet. Typically, the user only has to pay for getting access to the public Internet not for the distance that the IP packets have to travel.

- Internet standards should be as simple as possible (but no simpler).

1.2. Problem Statement

The ACS should enable an acoustic communication between any two Internet hosts considering the features of the Internet as described above. We see the need for designing the ACS because we see the following weaknesses in the existing codec and VoIP designs.

- Many standardized speech and audio codecs require the payment of royalty fees. Only codecs such as G.711, G.722, G722.1, and G.722.1C that have mediocre performances can be used license free. Thus, one cannot ensure that a good codec can be afforded by all owners of all Internet hosts.

- All known codecs have a small operational range, in addition they do not adapt to a wide range of bandwidth. For example, AMR support bit rates between 4.75 and 12.2 kbps and ITU G.719 support rates between 32 kbps and 128 kbps.

- An acoustic communication at superb transmission quality is not supported. Especially, if the latency is very low and the bandwidth is very high, we do not have a standardized codec that support hifi quality at ultra low delays. Only the SBC audio codec standardized by Bluetooth SIG [A2DPV10] can be considered for the usage scenario.
  Ultra-low delay transmissions at hifi quality are especially useful for distributed ensemble performances or distributed choruses.

- Similar, if the transmission quality is very bad, no standardized audio codec supports a grateful degradation. If the loss rate becomes too high then all speech and audio codecs become useless. However, in those cases one can use half-duplex, push-to-talk like transmission of short audio segments that would still allow a very slow communication at very low bitrates.

- Frequently, a PSTN call needs to be transcoded. Transcoding reduces the speech quality and increases latency. Thus, most codecs are designed to work well in conditions of transcoding. However, in case end-to-end IP transmission, the need for transcoding vanishes. It might only be needed for teleconferencing applications or for connecting to the PSTN network.

- The quality of a PSTN call has hardly increased during the last decade. Often, it is even worse because of IP based interconnections and support of cellular networks. Even those support of wideband speech transmission system have been developed, the lack of the willingness of users to pay more has limited the introduction of wideband speech. Also on the Internet we do not expect users to pay more for high quality phone calls. However, we believe that they will be delighted if they can communicate at nearly perfect quality.

- Neither a single standardized codec nor its RTP payload RFC specifies how to cope with time varying bandwidth and latency nor this is considered as required feature. This hinders the wide spread use of an adaptive coding mode selection and thus reduces the quality of many Internet phone calls.

- Not a single standardized codec supports varying complexities to support devices with low resources.

- Standardized codecs do not support any functionality for self-observation and self-monitoring. Also, they do not provide information about how well they encoded and decoded the audio content under a given set of coding parameters and packet loss rates. However, this information is important for the transport in order to rightly adapt the codec's transmission parameters.

- Packet losses occur in the Internet, the transmission time of packets and the playout time varies and the coding mode is changed in response to changed available transmission bandwidths. All these things cause the audio stream to be temporally distorted. The codec shall support concealment algorithms to limit the perceptual distortion. However, none existing standardized codec support the concealment of the adjustment of the playout time. Also, standardized PLC work on extrapolation of previous audio segments and do not support the interpolation. Lastly, often one cannot distinguish between delaying the playout time and packet loss because the missing packet might still arrive. Thus, an algorithm that uses the same extrapolation for packet loss concealment and time stretching might be beneficial.

- None of the standardized interactive speech and audio codec supports mechanisms to decrease the packet rate. Usually, packet rates are reduced by putting multiple speech frames into on RTP packet. However, the codecs do not take advantage of the high algorithmic delay that can be utilized then. Thus, they work less efficient in situations of congestion.

2. Usage Scenarios

The ACS should be optimized towards real-time communications over the Internet. It should support applications like collaborative network music performance, high-quality teleconferencing, wireless audio equipment, low-delay links for broadcast applications, network sound servers for using multimedia applications remotely, telepresence (enterprise) and the digital living room (consumer), and other.

The ACS shall be general enough to support multiple and quite diverse network conditions. For example, if network latency is low and bandwidth is plenty, it can be used for quasi-simultaneous music transmissions allowing distributed ensemble performances. It is also applicable interactive hifi quality audio transmission. If the network connection worsens, the transmission quality degrades to (wide-band) interactive speech transmission. As a last resort, it emulates a high-delay, half-duplex push-to-talk like communication service.

In the following, we enlist four main scenarios and describe their quality requirements.

2.1. Scenario 1: Person-to-person calls (VoIP)

The classic scenario is that of the phone usage to which we will refer in this document as Voice over IP (VoIP). Human speech is transmitted interactively between two Internet hosts. Typically, beside speech some background noise is present, too.

The quality of a telephone call is traditionally judged with subjective tests such as those describe in [ITU-T P.800]. The ACR scale used in MOS-LQS sometimes might not be very suitable for high quality, then - for example - the MUSHRA [ITU-T BS.1534-1] rating can be applied.

A telephone call is considered good if it has a maximal mouth-to-ear delay of 150ms [ITU-T G.107] and a speech quality of MOS-LQS 4 or above. However, interhuman communication is still possible if the delay is much larger.

This scenario does not include the use case of using a VoIP-PSTN gateway to connected to legacy telephone systems. In those cases, the gateway would make an audio conversion from broadband Internet voice to the frugal 1930's 3.1 kHz audio bandwidth. Interconnections to the PSTN will most likely stick with its legacy codecs to avoid transcoding.

2.2. Scenario 2: High quality interactive audio transmissions (AoIP)

In this first scenario we consider a telephone call having a very good audio quality at modest acoustic one-way latencies ranging from 50 and 150 ms [ITU-T G.107], so that music can be listened over the telephone while two persons talk interactively.

The Absolute Category Rating (ACR) (refer to ITU-T P.800) can be used, too. However, it might be more efficient to measure quality with the MUSHRA tests given in [ITU-T BS.1534-1], which is intended for intermediate audio qualities.

Also, for today's teleconferencing and videoconferencing systems there is a strong and increasing demand for audio coding providing the full human auditory bandwidth of 20 Hz to 20 kHz. This rising demand for high quality audio is due to the following:

- Conferencing systems are increasingly used for more elaborate presentations, often including music and sound effects which occupy a wider audio bandwidth than that of speech. For example, Web conferences such as WebEx, GoToMeeting, Adobe Acrobat Connect are based on an IP based transmission and benefit from a IP optimized ACS.

- The new "Telepresence" video conferencing systems, providing High Definition video and audio quality to the user, are giving the experience of being in the same room by introducing high quality media delivery (such as from Cisco).

- The emerging Digital Living Rooms will likely be interconnected and might require a constant acoustic transmission at high qualities.

2.3. Scenario 3: Ensembles performing over a network (MMoIP)

In some usage scenarios, users want to act simultaneously and not just interactively. For example, if persons sing in a chorus, if musicians jam, or if e-sportsmen play computer games in a team together they need to acoustically communicate. We call it the Make Music Over IP (MMoIP) scenario.

In this scenario, the latency requirements are much harder than for interactive usages. For example, if two musicians are placed more than 10 meters apart, they can hardly keep synchronized. Empirical studies [Gurevich2004] have shown that if ensembles playing over networks, the optimal acoustic latency is around 11.5 ms with targeted range from 10 to 25 ms.

In addition to the MUSHRA tests, the recommendation [ITU-R BS.1116] can be used for audio transmissions that just have minor impairments.

2.4. Scenario 4: Push-to-talk like service (PTT)

In spite of the development of broadband access (xDSL), a lot of users would only have service access via PSTN modems or mobile links. Also, on these links the available bandwidth might be shared among multiple flows and is subjected to congestion. Then, even low coding rates at about 8 kbps are too high.

If transmission capacity hardly exists, one still can degrade the quality of a telephone call to something like a push-to-talk (PTT) like service having very high latencies. Technically, this scenario takes advantage of bandwidth gains due to disruptive transmission

(DTX) modes and very large packets containing multiple speech frames causing a very low packetization overhead.

The quality requirements of a push to talk like service have been hardly studied. The OMA lists as a requirement of a Push To Talk over Cellular service a transmission delay of 1.6 s and a MOS values of above 3.0 that typically should be kept [OMAPoCReq]. However, as long as an understandable transmission of speech is possible, the delay can be even higher. For example, [OMAPoCReq] allows a delay of typically up 4s for the first talk-burst.

Also, [OMAPoCReq] describes a maximum duration of speaking. If a participant speaking reaches the time limit, the participant's right-to-speak shall be automatically revoked.

If the quality of a telephone call is very low, then instead of listening-only speech quality the degree of understandability can be chosen as performance metric. For example, objective tests of the understandability use automatic speech recognition (ASR) systems and measure the amount of correctly detected words.

In any case, the participant shall be informed about the quality of connection, the presence of high delays, the half-duplex style of communication, and its (limited) right-to-speak. For example this can be achieved by a simulated talker echo.

3. High-Level Requirements

Based on the four scenarios, we list the following high-level requirements that the ACS should fulfill.

3.1. Low cost and licensing free

The codec shall be affordable by all humans having Internet access.

Thus, one of the key requirements is patent/licensing free technology. However, it cannot be seen as "legally binding requirement" but rather as a desired working goal. Typically, one cannot verify 100% whether a codec is totally free of unknown IPRs. Some patents may be overlooked. It can also be assured that the known IPRs are "license-free" and "free from the need to sign licensing agreement(s) before use" (The ability for any user to get the codec and use it without signing any paperwork).

If one is practicing potentially patented technologies, there is no real mechanism to protect oneself from a patent troll at claims license fee for a standardized ACS. We have to assume that there is a

certain probability that the designed ACS is covered by patents what the IETF is not aware of. Thus, one has to define proper procedures on how to cope with IPR claims even if the ACS is already standardized.

Because of the lack of financial income, the codecs design, testing and standardization process must be cost effective, too. A cheap approach is needed to characterize the ACS, which might include tests having volunteer participants. For example, codecs can be provided to thousands of users in public to test them. Also, potential performance comparisons must not be as precise and proven as beyond any doubts because nobody wins or loses IPR fees if one solution wins or fails.

3.2. Reliable on the Internet

The ACS must be optimized towards acoustic real-time communications over the Internet, and must have the flexibility to adjust to the environment it operates in. Based on the quality of the end-to-end speech packet transmission, the codec should adapt its quality and delay to achieve an optimal benefit for the user.

As most Internet transport, it should be used with a wide range of condition allowing a high reliability regardless the networking condition. The reliability of the audio transmission should be high, even in cases of low and varying bandwidth. This implies that the codec is used on top of a transport protocol that implements a congestion control algorithm and that the ACS adapts to changes of available bandwidth. For example, if the available transmission bandwidth is too low to allow the codec to transmit audio at a high quality, the application can lower the sampling, bit or frame rate of the stream at the cost of higher algorithmic delay or a degraded audio quality.

3.3. Quality

The ACS must provide a quality/bitrate trade-off that is competitive with other state-of-the-art codecs. Also, the codec must have a very low algorithmic delay so that it can support the typical requirements of its users.

The speech and audio quality of the ACS should not be significantly worse than existing standardized codecs, if measures on the ACR scale.

4. Technical Requirements

4.1. Audio content

At all bitrates the ACS must deliver speech in any language at good quality. The ACS must be tested for different speakers and at least with two languages and should support tonal languages as well.

Frequently, speech needs to be transmitted not only without background noise but also at conditions including car, office and street noise. Background signals shall be considered not as the noise but as a part of the signals that convey information. Background signal can include background music at a SNR of 25 dB, office noise at a SNR of 20 dB, car noise at a SNR of 15 dB, babble Noise at a SNR of 25 dB, interfering talker at a SNR of 15 dB and street noise at a SNR of 20 dB.

At high bitrates the quality must be excellent for any audio signal, especially music. Stereo is considered as a must. Also, for high quality audio conferencing, reverberant input signals should be considered for testing the modes.

The speech and audio signals might have varying loudness. The transmission shall support a wide range of dynamics. The nominal input level of -36 dB, -26 dB and -16dB with respect to the overlapping bandwidth limit (OVL) point (-20 dBm0).

4.2. Quality

At a given operational mode, the ACS must not have perfect quality and must not perform better than any other standardized codec. However, considering the most common network conditions, the ACS shall perform better than any combination of existing codecs most of the time.

4.3. Reliability and congestion control

The acoustic transmission should be reliable and robust. The ACS shall be not only robust against packet losses but also for periods of low bandwidth.

The mean availability of the audio transmissions, calculated over all users, might be one of the metrics for assessing the performance of an Internet audio codec.

The ACS should adapt to the current network situation. Also, the codecs of ACS themselves must be adaptable, because switching among

multiple codecs is difficult to negotiate and unlikely to work well in situations of inter-operation.

Responding to congestion is a more complex issue and out of the scope of this document. However, it shall be defined on how to use existing congestion control protocols like DCCP and TCP. The ACS shall provide the mechanisms that congestion control requires from the codec (i.e. bitrate/framerate adaptability).

Because of the interactive nature of the acoustic transmission, the bidirectional transmission of audio content can be used for transmitting the required feedback and implementing a control loop. As such, it can be considered as a requirement that the acoustic transmission should be always bidirectional--even if the backward channel just sends "compressed silence".

4.4. Coding bit rate

The ACS must be capable of running at bitrates below 10 kbps. At low bitrates it must deliver good quality for clean, noisy or hands-free speech in any language. At high bitrates the quality must be excellent for any audio signal, including music. The bitrate must be adjustable in real-time. The bit rate can go up to 128 kbit/s per channel or more. The bitrate must be adjustable in real-time and at a fine granularity.

Variable bit rates depending on the content should be supported.

4.5. Sampling rate

The codec must support multiple sampling rates, ranging from 8 kHz to full band. Switching between sampling rates must be carried out in real-time.

4.6. Complexity

The ACS should have a complexity that is adjustable in real-time, where a higher complexity setting improves the quality/bitrate trade-off.

As a lower limit, the ACS shall run on hosts that common in developing countries. These may include OLPC XO-1s or other low-end (refurbished) computers (refer to Computer Aid International) and smart phones like those based on Texas Instruments Open Multimedia Application Platform (OMAP), which include both a host ARM CPU and one or more DSP.

On those devices, the ACS must not be capable of running at highest quality but at least at 8 kHz sampling rate.

4.7. Latency

To maintain a good quality of services requiring interactivity, it is necessary to maintain the overall delay as low as possible. But the delay requirement tends to have less importance in applications involving VoIP, possibly combined with other media and/or in heterogeneous network environment. A trade-off must be found between low delays and flexibility (scalability, ability to operate in various conditions with many types of signals etc.).

In interactive scenarios, the codec should be capable of running with an algorithmic delay of no more than 30 milliseconds.

For the making music scenario, the algorithmic delay must be between 3 to 9 ms. Still, given the speed of light as the fundamental limit of speed of information exchange, distributed ensembles can perform only regionally if latency budget of 25 ms must be kept. Typically, an optical fiber has a refractive index of 1.46 and thus in an optical fiber bits travel about 5136 km one-way in 25 ms.

The total codec delay consists of the algorithmic delay and the processing delay. Algorithmic delay includes the frame size delay plus any other delays inherent in the algorithm (look-ahead, noise suppression and error correcting codes for algorithm purposes and any algorithmic decoding delay). Processing delay is the additional delay caused by implementation with a finite speed processor.

4.8. Packet rate

The ACS must support a variable and dynamic changeable packet rate. Putting several frames into one packet is useful for packet grouping, which in turn is very useful for bandwidth adaptation and network usage efficiency.

This is because of the fact that a lot of bandwidth is used for protocol packet headers like those of Ethernet, IP, UDP, and RTP and thus to overhead at the MAC layer. If even IP header compression is applied, still many layer 2 protocols introduce an additional overhead that is not compressed [Hoene2005].

Classically, it is usually specified in the RTP payload specification, not in the codec specification itself. In general, a codec can take advantage of a larger frame size. This is especially true for a transform codec, where a larger frame means better

frequency resolution. The gain is somewhat smaller time-domain codec especially for > 20 ms frames. However, in larger packets the inter-frame dependencies can be adjusted on the fly to choose a trade-off between bitrate and amount of error propagation. It may even be possible to just make use of more inter-frame correlation for frames 2...N in a packet of N frames and get most of the benefits it would get from a larger frame size. Thus, the ACS codec should support large frame sizes (up to a MTU).

4.9. Packet loss resilience

The codec must be capable of running with little error propagation, meaning that the decoded signal after one or more packet losses is close to the decoded signal without packet losses after no more than two additional packets. The codec must have a packet loss resilience that is adjustable in real-time, where a lower packet loss resilience setting improves the quality/bitrate trade-off.

Also, the codec may add inter-frame redundancies to achieve better loss robustness.

4.10. Frame erasure concealment

The ACS must have a packet loss concealment algorithm. The PLC must be standardized to know how well the decoder can cope with packet losses in cases when the transmission parameters must be adjusted. However, the ACS may implement a PLC that performs better than the standardized PLC.

The purpose of standardizing the PLC (and the other concealment algorithms) is to guarantee a certain quality level over a range of conditions. For good results, a PLC operates on decoder-internal parameters and states, which requires tight algorithmic integration. So the PLC is as much part of a decoder as any other decoder module. The above also applies to time compression/stretching methods for handling network jitter and other kind of concealment algorithms (as mentioned below).

4.11. Jitter compensation and playout buffer

The ACS must cope with jitter. It must be able to receive the out of order de-packetized frames and present them in order for decoder consumption. It must be able to receive duplicate speech frames and only present unique speech frames for decoder. It must be able to handle clock drift between the encoding and decoding end-points.

The playout buffer should minimize the buffering time at all times while still conforming to the minimum performance requirements. If the limit of jitter induced concealment operations cannot be met, it is always preferred to increase the buffering time in order to avoid growing jitter induced concealment operations.

4.12. Playout adjustments

The ACS should support time scale modifications especially for jitter compensations such as time stretching and time shrinking because on the Internet jitter is the norm not a special case.

Because the operations going on in time scale modification algorithms are similar as those for the PLC, these operations should be combined into a single algorithm.

Also, the ACS shall be able to determine a desired length of a time scale modification (so it can e.g. leave out or add one or more pitch periods), to keep a 'backup' decoder state of the previous frame or to add one more frame length of decoding latency - otherwise you can not compress the voice of the previous packet and for stretching its suboptimal.

In general, the use of a high-quality time scaling algorithm is recommended. The amount of scaling should be as low as possible, scaling should be applied as infrequently as possible, and oscillating behavior is not allowed.

4.13. Concealment of mode switches

The ACS should also support the concealment of distortions caused by switching coding modes [Hoene2005]. Also, the negative effect of switching the coding mode shall be low.

For example, the transmission and coding mode might change several times (up to 5Hz) per second after getting feedback from the decoder.

4.14. Extrapolation

Sometimes, it is not possible to distinguish between a packet that arrives too late and packet that is lost and needs to be concealed. The decision on whether to conceal the loss or whether to conduct time stretching cannot be made yet. Thus, the ACS should support a general extrapolation of the audio signal which allows for late decision on whether to play out a delayed packet or whether to use a loss concealment operation

4.15. Interpolation

If a packet n has not arrived but the previous packet n-1 and the following packet n+1, when the packet n shall be interpolated using the frame of the previous and following packets.

4.16. DTX

The codec must be capable of using Discontinuous Transmission (DTX) where packets are sent at a reduced rate when the input signal contains only background noise.

4.17. Testing

The testing of ACS and the quality characterization shall be performed with real network profiles such as with [TIA-921] or those given in the appendix [TS.26114-830], not with fixed set of "average distributed errors and losses". Later do not clearly reflect the Internet nature.

Also, test vectors might be provided to check the correctness of the implementations.

4.18. Licensing and source code

The usage of ACS should not require paying royalties and signing NDA. At the time of standardization it should be available for royalty free (RF) and at reasonable and non-discriminatory terms (RAND). The codec should be available as open source allowing implementation under BSD, LGPL and/or GPL.

The codec specification description and implementation shall be based on a bit-exact fixed-point modular ANSI-C code using basic operators set provided in the ITU-T Software Tool Library to follow. In addition, an interoperable floating-point implementation can be provided.

The source code shall be normative because of a number of reasons. One is ease of implementation (either using the reference code directly, or being able to use it to validate the ported code). Another is that it assures that the characterization tests actually measure the standard's performance. Even if it is not officially normative, readily available reference code becomes de facto normative, since most implementers will simply use the code and ignore the text in the RFC.

4.19. Versioning and software updates

In order to cope with changes in the bitstream format, which might be required due to errors in the specification or - more important - due to newly claimed IPR, it must be possible to update the ACS online.

Also, it must be indicated, which bitstream format is going to be used.

4.20. RFC Type

It should become a standard, not an experimental RFC.

4.21. Side channel

Congestion control should be must for all Internet applications also for the ACS. [RFC3550] suggests in Chapter 10 somewhere that the RTP profile should care for rate adaptation. Thus, the ACS should take advantage of a feedback loop for variable coding parameter control in order to allow a wide range of operation and to adapt to the the current available bandwidth and processing power.

Congestion control per se is outside the review of this group, but providing the hooks for a congestion-control mechanism to interact with the codec is quite important. For example, running this codec on a TFRC-enabled or DCCP RTP stream - TFRC and DCCP need to be able to adjust (via the application) the bitrate of the codec in order to implement congestion control and perhaps adjust packetization periods/packet-rates.

A side channel for adaptation can be added. This would make sense because in usage scenarios audio is always transmitted in both directions. Adding a control channel would give a real advantage to existing codec designs. Alternatively, such as side channel can be also added with alternative solutions, such as handling that communication in SIP/SDP and in RTP/RTCP.

4.22. Layered coding

The ACS can support a layered encoding like in G.729.1 and G.718. Layered coding can be seen as a method for computationally efficient transcoding. Layered coding make sense in the conferencing environment as such stripping should be done at the sender after

encoding. Then, for all receivers the encoding has to be done only once.

However, for bidirectional transmissions, you do not need layered encoding as most codecs now are VBR, its enough already to adapt codec (at the source) to the bandwidth. Also, layered coding comes at additional cost (about 10% of the coding rate)

4.23. Interoperability with PSTN

The ACS might be developed to be interoperability to existing PSTN systems. Especially interoperability with 2G and 3G mobile radio systems is desirable. Also, the interoperability with G.722.2 @ 12,65 kb/s and with G.722 (for DECT devices) are of particular interest.

4.24. Conferencing and speech recognition

A teleconference server should be able to mix the audio signals at lower complexity than decoding + encoding. The ACS shall be capable of support automatic speech recognition.

4.25. Self-testing support

ACS should support means of testing the quality of a connection by feedback loops and quality feedbacks.

4.26. Self-awareness

The ACS should be aware on how well it can transmit acoustic content at various coding parameters and packet loss rates.

5. Out of scope

5.1. Multichannel

5.1 is worth supporting but that would most likely be through multiple independent channels/pairs, so that's probably not that much of an issue.

5.2. Repacketization

The ACS needs not to support repacketization in a network because this would violate the end-to-end semantic of the Internet.

5.3. Support for circuit-switched transmissions

The ACS needs not to support circuit-switched transmission.

5.4. Support of packet networks other than the Internet

The ACS needs not to support other packet networks (VoATM, private networks) beside the Internet.

5.5. Support of streaming

The ACS needs not to support multimedia streaming (e.g. video + audio involving bit-rate tradeoff), multicast content distribution (offline/online) and message retrieval systems.

5.6. Random packet losses

The usage of random packet losses to measure the concealment performance is meaningless because it does not reflect the nature of the Internet. Thus, the codec needs not be optimized nor tested using these criteria. Instead, real packet loss and delay traces should be considered. Also, short and long bursts of packet losses, which occur during due to handoffs, fast fading, congestion events, and route changes, should be considered.

5.7. Packet loss differentiation

The ACS cannot assume that the quality of packet transmission changes one per packet basis. For example, in layered coding the core layers cannot expect to be less subjected to packet losses than enhancement layers.

5.8. Robustness against bit errors

The ACS needs not to support bit errors because they are quite seldom on top of Ethernet. This is especially true as long as UDP-Lite is not supported widely.

5.9. IRS and other kind of bandwidth filters

The ACS must not consider bandwidth filters like the IRS because they are based on the traditions of circuit-switched connections.

5.10. Support of voice band data, fax and DTMF

The ACS needs not to support voice band data such as fax or DTMF. Instead, alternative ways of communication or other RTP payload format should be considered.

5.11. Idle noise

The generation of idle channel noise should not be used to indicate that the call is still active. Instead, in case of transmission problems an acoustic notification can be given.

5.12. Tandem coding

The ACS needs not to be optimized for tandem coding conditions because one can assume an end-to-end transmission of IP packets.

Tandem coding might only be used for PSTN gateways and for conference bridges.

5.13. FEC

RTP support of Forward Error Correction (FEC) needs not to be considered. Also, support of adding "redundant speech frames", which have been transmitted in preceding packets, in a RTP packet is not required. Instead, the redundancy can be added by the encoder which does this in a more efficient way.

6. Security Considerations

To do.

7. IANA Considerations

To do.

8. References

8.1. Normative References

[ITU-T BS.1534-1] "BS.1534 : Method for the subjective assessment of intermediate quality levels of coding systems", ITU-T Recommendation BS.1534-1 (01/03).

[ITU-T G.107] "G.107 : The E-model, a computational model for use in transmission planning", ITU-T Recommendation G.107 (04/09).

[ITU-T P.800] "P.800 : Methods for subjective determination of transmission quality", ITU-T Recommendation P.800 (08/96).

[ITU-R BS.1116] "BS.1116 : Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", ITU-R Recommendation BS.1116 (10/97).

[OMAPoCReq] "Push to talk over Cellular Requirements", Open Mobile Alliance, Approved Version 1.0, 09 Jun 2006, OMA-RD-PoC-V1_0-20060609-A.pdf

[TIA-921] TIA-921-A Document Information: "Network Model for Evaluating Multimedia Transmission Performance Over Internet Protocol", Publisher: Telecommunications Industry Association, Publication Date: Jun 18, 2008

[TS26.114-830] 3GPP TS 26.114 V8.3.0, "IP Multimedia Subsystem (IMS); Multimedia telephony; Media handling and interaction", Rapporteur: Per Froejdh, Version 8.3.0, 2009-06-12, RTS/TSGS-0426114v830.

8.2. Informative References

[A2DPV10] Bluetooth SIG, "Advanced Audio Distribution Profile", Audio Video WG, adopted specification, revision V1.0, May 22th, 2003.

[celt-draft] J-M. Valin, T. Terriberry, G. Maxwell, C. Montgomery, "Constrained-Energy Lapped Transform (CELT) Codec", Internet draft, draft-valin-celt-codec-01, work in progress, July 13, 2009.

[Gurevich2004] Gurevich, M., Chafe, C., Leslie, G., and Tyan, S., "Simulation of Networked Ensemble Performance with Varying Time Delays: Characterization of Ensemble Accuracy", Proceedings of the 2004 International Computer Music Conference, Miami, USA, 2004.

[Hoene2005] Hoene, C., and Karl, H., and Wolisz, A., "A perceptual quality model intended for adaptive VoIP applications", International Journal of Communication Systems, Wiley, August 2005.

[SG16 314-WP3] ITU-T SG16, "Agenda and list of documents for Q9/16", Temporary Document 314-WP3, Received on 2008-04-22 From Rapporteur Q9/16.

[silk-draft] K. Vos, S. Jensen, K. Soerensen, "SILK Speech Codec", Internet draft, draft-vos-silk-00.txt, work in progress, July 6, 2009.

[RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, July 2003.

9. Acknowledgments

The authors like to thank the various contributors taking part at the discussion on the Codec BOF mailing list in the period till September 2009. Also, this document is based on the SILK [silk-draft] and CELT drafts, the internal requirement documents of ITU-T G.718 [SG16 314-WP3] and the 3GPP document [TS26.114-830].

The author likes to thank Henry Sinnreich for his valuable feedback and support.

Funding for this draft has been provided by the University of Tuebingen within the "Projektfoerderung fuer Nachwuchswissen-schaftler".

This document was prepared using 2-Word-v2.0.template.dot.

Author's Address

Christian Hoene
University of Tuebingen
WSI-RI
Sand 13
72076 Tuebingen
Germany

Phone: +49 7071 2970532
Email: hoene@ieee.org