IPMR Speech Codec
draft-spiritdsp-ipmr-01.txt

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with
the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering
Task Force (IETF), its areas, and its working groups. Note that
other groups may also distribute working documents as Internet-
Drafts.

Internet-Drafts are draft documents valid for a maximum of six
months and may be updated, replaced, or obsoleted by other documents
at any time. It is inappropriate to use Internet-Drafts as
reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at

This Internet-Draft will expire on February 09, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the
document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal
Provisions Relating to IETF Documents
(http://trustee.ietf.org/license-info) in effect on the date of
publication of this document. Please review these documents
carefully, as they describe your rights and restrictions with
respect to this document.
Abstract

This document describes IPMR, a scalable variable adaptive multi-rate speech and audio codec designed for use in IP based networks. This codec is suitable for real time communications such as telephony, voice&video conferencing. Four different sampling frequencies are supported for encoding the audio input signal. Adaptation to network characteristics is provided through control of bitrate, packet rate, packet loss resilience and use of discontinuous transmission (DTX).

IP-MR support different profiles for input signal content which should be specified during codec initialization. It can be in Speech, Audio or Auto-detection mode. In Auto-detection mode codec recognizes type of input content automatically and switch to appropriate Speech or Audio mode automatically.

Table of Contents

1. Introduction .................................................... 3
2. Technical Requirements ......................................... 4
   2.1. Voice/Audio Quality .................................... 4
   2.2. Sampling Rate .......................................... 4
   2.3. Adaptive Multi Rate .................................... 4
   2.4. Bitrate Scalability ...................................... 4
   2.5. Packet Loss Resilience ................................ 4
   2.6. Delay .................................................. 4
   2.7. DTX ................................................... 5
3. IP-MR Codec Description ....................................... 5
4. Algorithm Overview ............................................ 8
   4.1. Coding profiles ......................................... 8
   4.2. Mixed CELP/MDCT codec ................................ 9
   4.3. Scalable CELP-based encoder ........................... 11
   4.4. Scalable CELP-based decoder ............................ 13
   4.5. Scalable MDCT-based encoder ........................... 14
   4.6. Scalable MDCT-based decoder ............................ 16
5. Security Considerations ....................................... 19
6. Informative References ....................................... 20
7. IANA Considerations ......................................... 21
Authors’ Addresses ............................................. 22
1. Introduction

To ensure high-quality IP audio transmitting the codec has to overcome a set of problems and obstacles. The best codec should be able to work at a wide range of bitrates with relatively small delay, should deliver high quality speech even in case of packet losses and poor network connection and should be able to provide wideband quality (which is a must for today’s biz-level communication) and ultra wideband quality for next-generation applications. This document describes the IP-MR codec which is scalable variable adaptive multi-rate speech and audio codec designed for use in IP based networks.
2. Technical Requirements

We agree with some technical requirements described in [SILK] and include them into this section. The Internet Wideband Speech/Audio Codec must be optimized towards real-time communications over the Internet, and must have the flexibility to adjust to the environment it operates in. Below is a list of main requirements for the codec.

2.1. Voice/Audio Quality
The codec should provide a quality/bitrate trade-off that is competitive with other state-of-the-art codecs. At low bitrates it should deliver good quality of speech in any language. At high bitrates the quality should be excellent for any audio signal, including music, at standard conditions.

2.2. Sampling Rate
Audio bandwidth is determined by the codec sampling frequency - 8 kHz for narrowband voice (PSTN) and 16 kHz for wideband. Obviously, wideband speech is much more natural and comfortable and wideband codecs are more convenient to use in IP communication. However, sometimes there isn’t enough bandwidth to allow 16 kHz sampling frequency, and codec must be able to switch to 8 kHz. Moreover, codec should support ultra wide band (20 kHz and more) for next-generation high-end quality.

2.3. Adaptive Multi Rate
The codec should have a set of bitrates with needed granularities to fit into different channels capacities. The bitrates should be adjustable in real-time. The codec should be capable of running at bitrates starting from 6 kbps.

2.4. Bitrate Scalability
Codec should have bitrate scalability feature (embedded or layered structure of bitstream) to enable reduce voice traffic during transition without re-encoding. This is necessity for dynamic congestion control, multicast and conferencing applications. From the other hand the payment for scalability is less compression efficiency and more computational complexity at the same bitrate. Because of that it will be good if scalability feature can be switched-off when it’s not needed.

2.5. Packet Loss Resilience
The codec should 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 should have a packet loss resilience that is adjustable in real-time, where a lower packet loss resilience setting improves the quality/bitrate trade-off.

2.6. Delay
For comfort conversation the codec must have algorithmic delay not more than 50 ms.
2.7. DTX
The codec should be capable of using Discontinuous Transmission (DTX) where packets are sent at a reduced rate when the input signal contains only background noise.

3. IP-MR Codec Description
The IP-MR codec is scalable variable adaptive multi-rate speech and audio codec designed for use in IP based networks. This codec is suitable for real time communications such as telephony, voice&video conferencing.

Sampling rate
IP-MR support three sampling rate modes: 8, 16 and 32 kHz

Speech/Audio modes
IP-MR support different profiles for input signal content which should be specified during codec initialization. It can be in Speech, Audio or Auto-detection mode. In Auto-detection mode codec recognizes type of input content automatically and switch to appropriate Speech or Audio mode automatically.

Voice Quality
The Mean Opinion Score (MOS) of this speech codec’s speech quality is about 3.7-4.4 (for clean speech) and it’s depends on current mode and average bit rate. At higher bitrates codec achieves FM quality on generic audio content.

Algorithmic delay
The frame length is 20 ms. Algorithmic delay varies from 35 to 50 ms depending of coding profile.

Adaptive Multi Rate
Depending of sampling rate IP-MR has 8 or 10 bitrate modes between 6 and 120 kbps which can be changed in real time in compliance with the current network conditions.
<table>
<thead>
<tr>
<th>Sampling Rate</th>
<th>Coding profile</th>
<th>Frame size</th>
<th>Algorithm. Delay</th>
<th>Number of Rates</th>
<th>Avg. Bit Rates for active speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 kHz</td>
<td>Speech/ Auto- -detection with short delay</td>
<td>20 ms</td>
<td>35 ms</td>
<td>8</td>
<td>6 - 50 kbps</td>
</tr>
<tr>
<td></td>
<td>Audio/ Auto- -detection with long delay</td>
<td>ms</td>
<td>50 ms</td>
<td>10</td>
<td>6 - 70 kbps</td>
</tr>
<tr>
<td>16 kHz</td>
<td>Speech/ Auto- -detection with short delay</td>
<td>20 ms</td>
<td>36.875 ms</td>
<td>10</td>
<td>6 - 70 kbps</td>
</tr>
<tr>
<td></td>
<td>Audio/ Auto- -detection with long delay</td>
<td>ms</td>
<td>50 ms</td>
<td>10</td>
<td>6 - 120 kbps</td>
</tr>
<tr>
<td>32 kHz</td>
<td>Speech/ Auto- -detection with short delay</td>
<td>20 ms</td>
<td>37.8125 ms</td>
<td>10</td>
<td>6 - 120 kbps</td>
</tr>
<tr>
<td></td>
<td>Audio/ Auto- -detection with long delay</td>
<td>ms</td>
<td>50 ms</td>
<td>10</td>
<td>6 - 120 kbps</td>
</tr>
</tbody>
</table>

Variable Bit Rate
Encoder's bit rate is constantly varying in accordance with the actual speech content (voiced/unvoiced, pauses, stationary/non-stationary voiced, etc.). IP-MR codec optimizes and reduces traffic while keeping the efficiency, as the encoding is adaptive to the actual characteristics of speech. All average bitrates are specified for active speech without consideration of inter-speech (silence) regions.
Bitrate Scalability

The coded frame has layered (embedded) structure. It consists of multiple coding layers - base (or core) layer and several enhancement layers which are coded independently. Only the core layer is mandatory to decode understandable speech and upper layers provide quality enhancement. These enhancement layers may be omitted and remaining base layer can be meaningfully decoded without notable artifacts. This making the bit stream scalable and allows reduce bit rate during transmission without re-encoding.

Bitrate scalability provides additional possibilities for congestion control. Some intermediate network node may modify the IP-MR codec’s payload by dropping some of the layers during transmission to meet the available bandwidth requirements. In case the payload is forwarded with modified content at least the base layer must be preserved in the payload which is being delivered to receiving side guarantees meaningful speech decoding without packet loss concealment procedure.

But because of the scalable nature of IP-MR codec there is no need to duplicate the whole previous frame - only the core layer may be retransmitted. This reduces redundancy overhead while keeping efficiency.

Moreover, the speech bits encoded in core layer are divided on six classes (from A to F) of perceptual sensitivity to errors. Class A contains most perceptually significant bits. This class’s bits should be delivered to Decoder to exclude fully "error propagation". Class F contains less significant bits. Sum of all classes from A to F contains all encoded parameters of the first (core) encoding layer. These parameters are sufficient to synthesize speech with near "toll quality".

Using these classes as introduced redundancy make possible to smoothly adjust trade-off between overhead and robustness against packet loss.

DTX

IP-MR codec support Discontinuous Transmission mode for silence compression. During silence intervals the codec bitrate can be reduced to 0.3 kbps.
4. Algorithm overview

4.1. Coding profiles
IP-MR support different profiles for type of input signal content. It can be Speech, Audio or Auto-detection modes. In Auto-detection mode codec recognizes type of input content automatically and switch to appropriate Speech or Audio mode automatically. At high level encoder consists of three basic modules (see Figure 1).

- Speech/Music detector - automatically classify type of input content as speech or music to enable appropriate coding model.
- CELP-based speech coder - implements source-filter model, speech content oriented.
- MDCT-based audio coder - for general audio coding purpose.

```
+-------------------+
|Predefined Speech/ |
|       Audio       |
|      Profile      |
+----------+--------+

|           |
+----------+-------+
input signal | Speech/|
---------------+  Music detector |
+---+---------+----+
|   |        M|
| S |        u|
| P |        s|
| e |        i|
| e |        c|
| c |        c|
| h |         |
+---------+-------+----+

+............+............+............+
 .         \/          \/ coder .
 . +----------+----------+ .
 . | CELP/MDCT | MDCT | .
 +----------+----------+ .
+............+............+............+
 |         |         |        |
 \/        \/       |
 +-------------------+
| Bitstream +--->
+-------------------+

Figure 1 High level encoder structure
Depending of type of input signal (speech/music) different coding models are used. The type of input signal can be detected automatically in ‘Autodetection’ mode or specified as predefined setting during codec initialization. The speech content is coded by mixed CELP/MDCT based model. General audio content is coded by pure MDCT-based model.

The decoder does backward operations. First, compressed frame goes to CELP-decoder; it extracts core and extension layers. Then, both the rest of bitstream and reconstructed signal go to MDCT-decoder which restores residue and generates joint output.

![Decoder Diagram](image)

Figure 2 High level decoder structure

In fact CELP and MDCT are two different decoders and thus, they can work simultaneously. Parallel processing requires only two modules to be carried out of decoder structure (see Figure 1) they are - bitstream demultiplexing and signal mixing.

![Decoder Diagram](image)

Figure 2 High level decoder structure (parallel)

Note, that demultiplexing is simple to implement because of the size of CELP stream portion can be calculated without decoding.

4.2. Mixed CELP/MDCT codec

The mixed CELP/MDCT Codec is composed from two independent codecs - CELP and MDCT-based. The first one processes source signal and feeds the residue to the second. In order to provide flexible and transparent coupling between codecs, corresponding sampling rate conversion and frame synchronization procedures are applied.
The resulting bitstream naturally constructed from two continues regions belong to CELP and MDCT codecs correspondingly. The CELP-codec bitstream has a layer structure (core + extensions) while the MDCT-codec generates byte-scalable stream.

The next figure provides an example of 16 kHz source material encoding if CELP-base encoder operates at 8 kHz sampling rate.

Figure 3 Structural block diagram of mixed CELP/MDCT encoder (16kHz mode)

First, input signal is down-sampled to 8 kHz and encoded by Scalable CELP-based encoder which packs quantized parameters in layered bitstream. The difference between up-sampled synthesized signal and original source goes to Scalable MDCT-based encoder which forms the rest of bitstream.

Below CELP and MDCT-based codecs are considered in more details.
4.3. Scalable CELP-based encoder

Scalable CELP-based coder applied to speech coding consists of the core (base layer) encoder and three enhancement encoders. In Figure 4 the structure of core encoder is shown.

Core Encoder codes speech in a "base frequency bandwidth" (up to 4 kHz) with speech quality near to "Toll Quality" and forms a coded bit stream at minimum average bit rate (about 6.0 kbps). Current bit rate is driven by information content of input speech and can vary in range from 4.3 kbps up to 10.35 kbps.

The Core Encoder performs LPC analysis and pitch detection, estimates parameters of the pitch-predictor and excitation by the "analysis-by-synthesis" method on the "subframe-by-subframe" base. The subframe length is 5 ms.

Encoded parameters and bits are separated to 6 sensitivity classes from: Class A to Class F to provide a possibility of the additional protection them against packet losses.

Class A contains most perceptually significant bits. This class’s bits should be delivered to Decoder to exclude fully "error propagation".

Class F contains less significant bits. Sum of all classes from A to F contains all encoded parameters of the first (core) encoding layer. These parameters are sufficient to synthesize speech with "toll quality".
Figure 4 Structural block diagram of CELP-based Core Encoder
The difference between input speech and synthesized speech (by Core Encoder) is delivered to extension coding. Each next Extension Encoder codes the residual (delivered from previous layer) and forms own additional coded bit stream. Therefore, full bit stream contains a sum of the base and extension bit streams. The number of layers, which is used at coding and corresponded to number of the bit streams in the sum on the encoder’s output, can be changed "on the fly".

Each CELP Extension Encoder uses results of previous layer’s encoding and estimates additional excitation by the "analysis-by-synthesis" method on the "subframe-by-subframe" base (Figure 5). There are total 3 CELP Extension Encoders.

4.4. Scalable CELP-based decoder

The decoder dequantizes parameters of each encoding layer, reconstructs total excitation by sum of adaptive codebook and fixed codebooks (core and enhancement) and synthesizes speech using LPC-filter. Reconstructed speech is post-filtered and output to the 160 samples buffer (20 ms at 8 kHz). In Figure 6 the structure of CELP-based decoder is presented.
Decoder has ability to conceal of the lost frames (PLC-like function) by partial reconstruction of speech, using speech parameters of the last received frames. However, to provide highest robustness to packet loss, classes of the most significant parameters only should be protected.

4.5. Scalable MDCT-based encoder

Scalable MDCT-based encoder operates on a frame basis in a domain of MDCT spectrum. Quantized spectrum samples are written into the bitstream.

```
--Input signal-->+ MDCT -->+ Quantizer -->+ Bitstream -->Scalable
                 +---------+-----------+ formatter | bitstream-->
                 +---------+-----------+
```

Figure 7 Scalable MDCT-based Encoder
This approach is widely used in modern audio coding algorithms. The main advantage of developed compression scheme is a bitstream formatter unit. It constructs stream in a way that any initial part of the compressed data can be decoded and used for reconstruction. In other words, each initial part of compressed frame carries self-sufficient information about band-limited signal with a given level of accuracy.

The bitstream formatter unit operates on a band basis, each eight samples long. Coding loop iterates over all bands and transmits update for a given band. Loop ends if all spectrum bands are fully transmitted.

```plaintext
+-----------+
/ Spectrum /
+-----------+
      |
      \|
+-----------+ +----------------+
| Start +--------->/ numCodedBands=0 /
+-----------+ +----------------+
      |
      \|
+-----------+ no +----------------+ yes +-----+
        | chooseCodedBand()------>/ isAllBandsCoded()------>/ End |
+--------+ +----------------+ +--------+ +----------------+ +--------+
| yes \/ | no \/|
+--------+ +----------------+ +--------+ +----------------+ +--------+
| updateBand()<--- startNewBand()<--- numCodedBands++ |
+--------+ +----------------+ +--------+ +----------------+ +--------+
| .
+-----------+ +----------------+
| applyCompressionModel() |
+--------+ +----------------+
      |
      \|
+--------+ +----------------+
        | rangeCodec()------>/ bits/sample |
+--------+ +----------------+
      |
      \|
+--------+ +----------------+
| Compressed frame |
+----------------+
```

Figure 8 Spectrum encoding loop
Bandwidth expansion (coding band increment) is based on actual bit/samples ratio known for both encoder and decoder. Coding band increment only occurs if compression rate exceed some fixed threshold or all available bands are already fully encoded. Practical experiments show that if compression ratio exceeds $1.7 - 2$ bits/sample than it is reasonable to expand bandwidth rather than update existing bands.

Band update procedure is based on a bit-planes data representation. One bit-plane issues per band at time. In terms of binary planes it means that each update carries one bit of mantissa for each band sample. Current implementation uses ternary planes instead of conventional binary planes. This allows encoder to reduce the amount of noise introduced if only top plane is transmitted.

The sign and sample presence flag together form a top plane for particular band which transmitted first than on band coding start. Encoder keeps a track of transmitted planes for each band and chooses the highest non transmitted plane to update.

Encoder applies different statistic models and compression schemes for different planes and bands. Actually only several top planes (following by sign/flag plane) are well suited for compression, whereas all others tend to have random distribution and in fact can’t be compressed at all. After compression scheme is applied, raw data and chosen statistic model go to range codec(1) which writes it into a bitstream.

### 4.6. Scalable MDCT-based decoder

Decoder performs all the same operations as encoder does, but in backward manner. First bitstream reader reconstructs quantized spectrum samples from compressed frame, than inverse quantized reconstructs MDCT spectrum and inverse MDCT transforms signal back from frequency to time domain.

```
+-----------+    +-----------+    +---------+
 Scalable | Bitstream -->+ Inverse | | Inverse --Reconstructed
-.bitstream->+ reader | | Quantizer -->+ MDCT | | signal -->
 +-----------+    +-----------+    +---------+
```

Figure 9 Scalable MDCT-based Decoder

(1) Range codec is a sort of arithmetic codec providing byte stream granularity.
The resulting signal accuracy and bandwidth dependent on the amount of available input data. Codec introduces no inter frame data dependency except 50% time domain overlapping required for MDCT transform. In practice, it means that signal can't be correctly reconstructed from a first successfully received compressed frame, but the second frame will be reconstructed correctly.

The bitstream reader decompress input stream using inverse range coder. Because of encoder and decoder operate synchronously, each time decoder runs inverse range codec it uses exactly the same context as were used by encoder during compression. Stream parsing ends if no more data available for compressed frame. The following figure demonstrates spectrum decoding loop.
In spite of codec has no lower bitrate limit, the compression scheme used provides artificial reconstructed signal if transmission rate is low than 16-24 kbps. For low bitrates presented audio codec is used in a bunch with speech codec and processes the speech codec residue.
5. Security Considerations

To Be Defined.
6. Informative References

action=AttachFile&do=get&target=draft-vos-silk-00.txt
7. IANA Considerations

This document has no actions for IANA
Authors’ Addresses

Vladimir Sviridenko  
SPIRIT DSP  
Solzhenitsina 27  
Moscow 109004  
Russia  
Phone: +7 495 661 2178  
Email: vladimirs@spiritdsp.com

Sergey Ikonin  
SPIRIT DSP  
Solzhenitsina 27  
Moscow 109004  
Russia  
Phone: +7 495 661 2178  
Email: s.ikonin@gmail.com

Dmitry Yudin  
SPIRIT DSP  
Solzhenitsina 27  
Moscow 109004  
Russia  
Phone: +7 495 661 2178  
Email: yudin@spiritdsp.com

Person & email address to contact for further information:  
Yury Morzeev  
morzeev@spiritdsp.com