Audio Video Transport WG
INTERNET-DRAFT
Category: Standards Track
Expires: November 17, 2004

Sassan Ahmadi
Nokia Inc.

May 17, 2004

Real-Time Transport Protocol (RTP) Payload and File Storage Formats for the Variable-Rate Multimode Wideband (VMR-WB) Audio Codec

<draft-ahmadi-avt-rtp-vmr-wb-02.txt>

Status of this Memo

Copyright Notice

Abstract

This document specifies a real-time transport protocol (RTP) payload format to be used for the Variable-Rate Multimode Wideband (VMR-WB) speech codec. The payload format is designed to be able to interoperate with existing VMR-WB transport formats on non-IP networks. In addition, a file format is specified for transport of VMR-WB speech data in storage mode applications such as email. A MIME type registration is included, for VMR-WB, specifying use of

both the RTP payload and the storage formats

VMR-WB is a variable-rate multimode wideband speech codec

that has a number of operating modes, one of which is interoperable with AMR-WB (i.e., RFC 3267) audio codec at certain rates. Therefore, provisions have been made in this draft to facilitate and simplify data packet exchange between VMR-WB and AMR-WB in the interoperable mode with no transcoding function involved.

Table of Contents

1. Introduction

This document specifies the payload format for packetization of VMR-WB encoded speech signals into the Real-time Transport Protocol (RTP) [3]. The VMR-WB payload formats support transmission of single and multiple channels, frame interleaving, multiple frames per payload, header-free payload, the use of mode switching, and interoperation with existing VMR-WB transport formats on non-IP networks, as described in Section 3.

The payload format itself is specified in Section 6. A related file format is specified in Section 7 for transport of VMR-WB speech data in storage mode applications such as email. In Section 10, a MIME type registration for VMR-WB is provided.

Since VMR-WB is interoperable with AMR-WB at certain rates, an attempt has been made throughout this document to maximize the similarities with RFC 3267 while optimizing the payload and storage formats for the non-interoperable modes of the VMR-WB codec.

2. Conventions and Acronyms

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC2119 [2].

The following acronyms are used in this document:

```
3GPP2  - The Third Generation Partnership Project 2
CDMA   - Code Division Multiple Access
WCDMA  - Wideband Code Division Multiple Access
GSM    - Global System for Mobile Communications
AMR-WB - Adaptive Multi-Rate Wideband Codec
VMR-WB - Variable-Rate Multimode Wideband Codec
CMR    - Codec Mode Request
GW     - Gateway
DTX    - Discontinuous Transmission
FEC    - Forward Error Correction
SID    - Silence Descriptor
TrFO   - Transcoder-Free Operation
UDP    - User Datagram Protocol
```

RTP - Real-Time Transfer Protocol
RTCP - Real-Time Control Protocol
MIME - Multipurpose Internet Mail Extension
SDP - Session Description Protocol

SIP - Session Initiation Protocol

The term "frame-block" is used in this document to describe the time-synchronized set of speech frames in a multi-channel VMR-WB session. In particular, in an N-channel session, a frame-block will contain N speech frames, one from each of the channels, and all N speech frames represent exactly the same time period.

3. The Variable-Rate Multimode Wideband (VMR-WB) Speech Codec

VMR-WB is the wideband speech-coding standard developed by Third Generation Partnership Project 2 (3GPP2) for encoding/decoding wideband/narrowband speech content in multimedia services in 3G CDMA cellular systems. VMR-WB is a source-controlled variable-rate multimode wideband speech codec. It has a number of operating modes, where each mode is a tradeoff between voice quality and average data rate. The operating mode in VMR-WB is chosen based on the traffic condition of the network and the desired quality of service [1]. The desired average data rate (ADR) in each mode is obtained by encoding speech frames at different rates compliant with CDMA Rate-Set II depending on the instantaneous characteristics of input speech and the maximum and minimum rate constraints imposed by the network operator. While VMR-WB is a native CDMA codec complying with all CDMA system requirements, it is further interoperable with AMR-WB [4] at 12.65, 8.85, and 6.60 kbps. This is due to the fact that VMR-WB and AMR-WB share the same core technology. This feature enables Transcoder Free (TrFO) interconnections between VMR-WB and AMR-WB across different wireless/wireline systems (e.g., GSM/WCDMA and CDMA2000) without use of unnecessary complex media format conversion.

VMR-WB is able to transition between various modes with no degradation in voice quality that is attributable to the mode switching itself. The operation mode of the VMR-WB encoder may be switched seamlessly without prior knowledge of the decoder. Any non-interoperable mode (i.e., mode 0, 1, or 2) can be chosen depending on the traffic conditions (e.g., network congestion) and the desired quality of service.

While in the interoperable mode (i.e., VMR-WB mode 3), mode switching is not allowed. There is only one AMR-WB interoperable mode in VMR-WB. Since AMR-WB codec depending on channel conditions may request a mode change, in-band data included in VMR-WB frame structure (see Section 8 of [1] for

more details), is used during an interoperable interconnection to switch between AMR-WB codec modes 0, 1, or 2.

As mentioned earlier, VMR-WB is compliant with CDMA Rate-Set

II (see Section 2 of [1]) with the permissible encoding rates shown in Table 1.

Frame Type	Bits per Packet (Frame Size)	Encoding Rate (kbps)
Full-Rate	266	13.3
Half-Rate	124	7.2
Quarter-Rate	54	2.7
Eighth-Rate	20	1.0
Blank	0	-
Erasure	0	-

Table 1: CDMA Rate-Set II frame types and their associated encoding rates

VMR-WB is robust to high percentage of packet loss and packets with corrupted rate information. The reception of an Erasure (SPEECH_LOST) frame type at decoder invokes the built-in frame error concealment mechanism. The built-in frame error concealment mechanism in VMR-WB conceals the effect of lost packets by exploiting in-band data and the information available in the previous frames.

3.1. Narrowband Speech Processing

VMR-WB has the capability to operate with 8000 Hz sampled input/output speech signals in all modes of operation [1]. Mode switching can be utilized to change the mode of operation while processing narrowband speech signals. However, during a session, transition between narrowband and wideband processing is not RECOMMENDED due to different timestamps and other likely synchronization problems.

3.2. Continuous vs. Discontinuous Transmission

The circuit-switched operation of VMR-WB within a CDMA network requires continuous transmission of the speech data during a conversation. The intrinsic source-controlled variable-rate feature of the CDMA speech codecs is required for optimal operation of the CDMA system and interference control. However, VMR-WB has the capability to operate in a discontinuous transmission mode for some packet-switched applications over IP networks, where the number of transmitted bits and packets during silence period are reduced to a minimum. The VMR-WB DTX operation is similar to

that of AMR-WB [4,12].

4. Support for Multi-Channel Session

Both the octet-aligned RTP payload format and the storage format defined in this document support multi-channel audio content (e.g., a stereophonic speech session).

Although VMR-WB codec itself does not support encoding of multi-channel audio content into a single bit stream, it can be used to separately encode and decode each of the individual channels.

To transport (or store) the separately encoded multi-channel content, the speech frames for all channels that are framed and encoded for the same 20 ms periods are logically collected in a frame-block.

At the session setup, out-of-band signaling must be used to indicate the number of channels in the session and the order of the speech frames from different channels in each frame-block. When using SDP for signaling, the number of channels is specified in the rtpmap attribute and the order of channels carried in each frame-block is implied by the number of channels as specified in Section 4.1 in [10].

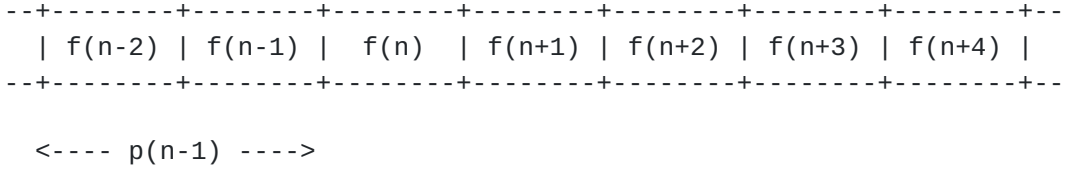
4. Robustness against Packet Loss

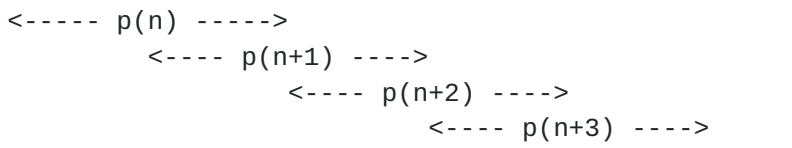
The octet-aligned payload format, described in this document, supports several features including forward error correction (FEC) and frame interleaving in order to increase robustness against lost packets.

4.1. Forward Error Correction (FEC)

The simple scheme of repetition of previously sent data is one way of achieving FEC. Another possible scheme, which is more bandwidth efficient is to use payload external FEC,e.g., RFC2733 [5], which generates extra packets containing repair data.

The repetition method involves the simple retransmission of previously transmitted frame-blocks together with the current frame-block(s). This is done by using a sliding window to group the speech frame-blocks to send in each payload. Figure 1 illustrates an example.




<----- p(n) ----->
<---- p(n+1) ---->
<---- p(n+2) ---->
<---- p(n+3) ---->

<---- p(n+4) ---->

Figure 1: An example of redundant transmission.

In this example each frame-block is retransmitted one time in the following RTP payload packet. Here, f(n-2)..f(n+4) denotes a sequence of speech frame-blocks and p(n-1)..p(n+4) a sequence of payload packets.

The use of this approach does not require signaling at the session setup. In other words, the speech sender can choose to use this scheme without consulting the receiver. This is because a packet containing redundant frames will not look different from a packet with only new frames. The receiver may receive multiple copies or versions of a frame for a certain timestamp if no packet is lost. If multiple versions of the same speech frame are received, it is RECOMMENDED that the highest rate be used by the speech decoder.

This redundancy scheme provides the same functionality as the one described in RFC 2198 "RTP Payload for Redundant Audio Data" [10]. In most cases the mechanism in this payload format is more efficient and simpler than requiring both endpoints to support RFC 2198. If the spread in time required between the primary and redundant encodings is larger than 5 frame times, the bandwidth overhead of RFC 2198 will be lower.

The sender is responsible for selecting an appropriate amount of redundancy based on feedback about the channel, e.g., in RTCP receiver reports, or network traffic. A sender should not base selection of FEC on the CMR, as this parameter most probably was set based on none-IP information. The sender is also responsible for avoiding congestion, which may be aggravated by redundant transmission.

4.2. Frame Interleaving and Multi-Frame Encapsulation

To decrease protocol overhead, the octet-aligned payload format allows several speech frame-blocks to be encapsulated into a single RTP packet. One of the drawbacks of such approach is that in case of packet loss this means loss of several consecutive speech frame-blocks, which usually causes clearly audible distortion in the reconstructed speech. Interleaving of frame-blocks can improve the speech quality in such cases by distributing the consecutive losses into a series of single frame-block losses. However, interleaving and bundling several frame-blocks per payload will also

increase end-to-end delay and is therefore not appropriate for all types of applications. Streaming applications will most likely be able to exploit interleaving to improve speech quality in lossy transmission conditions.

The octet-aligned payload format supports the use of frame interleaving as an option. For the encoder (speech sender) to use frame interleaving in its outbound RTP packets for a given session, the decoder (speech receiver) needs to indicate its support via out-of-band means (see Section 10).

5. VMR-WB Voice over IP Scenarios

5.1 IP Terminal to IP Terminal

The primary scenario for this payload format is IP end-to-end between two terminals incorporating VMR-WB codec, as shown in Figure 2. This payload format is expected to be useful for both conversational and streaming services.

Figure 2: IP terminal to IP terminal scenario

A conversational service puts requirements on the payload format. Low delay is a very important factor, i.e. fewer speech frame-blocks per payload packet. Low overhead is also required when the payload format traverses across low bandwidth links, especially if the frequency of packets will be high.

Streaming service has less strict real-time requirements and therefore can use a larger number of frame-blocks per packet than conversational service. This reduces the overhead from IP, UDP, and RTP headers. However, including several frame-blocks per packet makes the transmission more vulnerable to packet loss, so interleaving may be used to reduce the effect of packet loss on speech quality. A streaming server handling a large number of clients also needs a payload format that requires as few resources as possible when doing packetization.

Note that all modes of the VMR-WB codec can be used in this scenario. Also both header-free and octet-aligned payload formats can be utilized.

5.2 IP Terminal to GW to IP Terminal

A second scenario for this payload format is IP end-to-end

(Through a gateway) between two terminals, one with AMR-WB codec and the other one with VMR-WB codec using the interoperable mode of VMR-WB, as shown in Figure 3. This payload format is expected to be useful for both

conversational and streaming services.

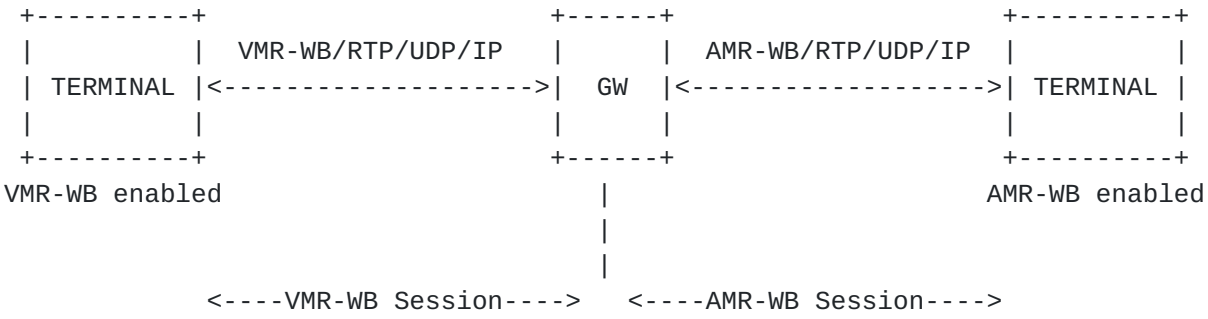


Figure 3: IP terminal to GW to IP terminal scenario
(AMR-WB <-> VMR-WB interoperable interconnection)

The VMR-WB mode 3 and octet-aligned payload format SHALL be used for this scenario. Moreover, to avoid signaling conflicts in the IP network, two sessions SHALL be established using SIP/SDP, one between the VMR-WB enabled terminal and the gateway and another session between the gateway and the AMR-WB enabled terminal. Note that no transcoding is involved since the VMR-WB payload is identical to that of AMR-WB.

5.3 GW to IP Terminal

Another scenario occurs when VMR-WB encoded speech will be transmitted from a non-IP system (e.g., 3GPP2/CDMA2000 network) to an RTP/UDP/IP VoIP terminal, and/or vice versa, as depicted in Figure 4.

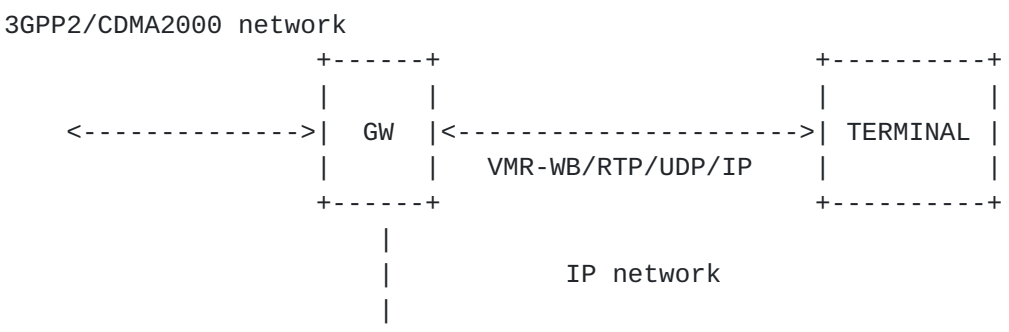


Figure 4: GW to VoIP terminal scenario

VMR-WB's capability to seamlessly switch between operational modes is exploited in CDMA (non-IP) networks to optimize speech quality for a given traffic condition. To preserve this functionality in scenarios including a gateway to an IP network using the octet-aligned payload format, a codec mode

request (CMR) field is considered. The gateway will be responsible for forwarding the CMR between the non-IP and IP parts in both directions. The IP terminal should follow the CMR forwarded by the gateway to optimize speech quality going

to the non-IP decoder. The mode control algorithm in the gateway SHOULD accommodate the delay imposed by the IP network on the response to CMR by the IP terminal.

The IP terminal should not set the CMR (see Section 6.3.2), but the gateway can set the CMR value on frames going toward the encoder in the non-IP part to optimize speech quality from that encoder to the gateway. The gateway can alternatively set a different CMR value, if desired, as one means to control congestion on the IP network.

5.4 GW to GW (Between VMR-WB and AMR-WB Enabled Terminals)

A fourth likely scenario is that RTP/UDP/IP is used as transport between two non-IP systems, i.e., IP is originated and terminated in gateways on both sides of the IP transport, as illustrated in Figure 5. This is the most likely scenario for an interoperable interconnection between 3GPP/(GSM,WCDMA)/AMR-WB and 3GPP2/CDMA2000/VMR-WB.

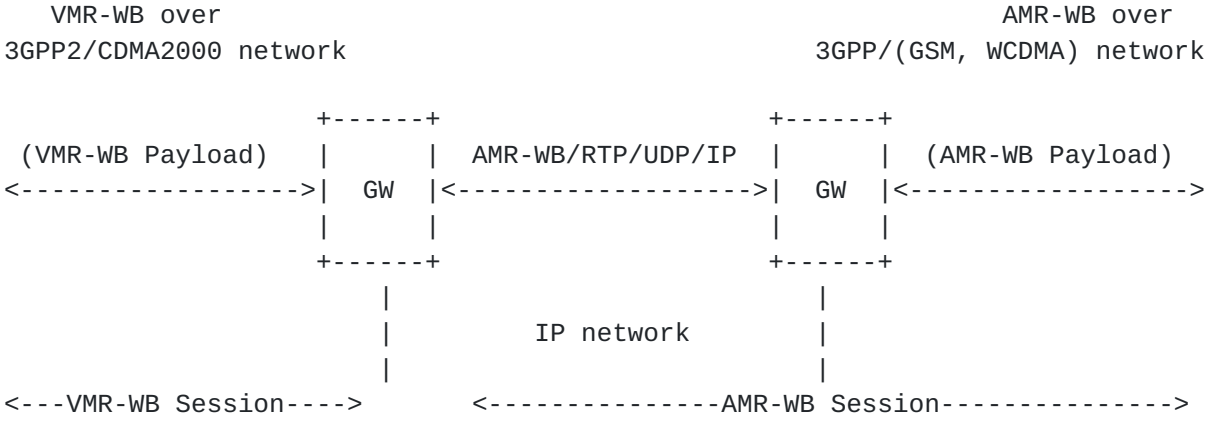


Figure 5: GW to GW scenario (AMR-WB <-> VMR-WB interoperable interconnection)

The VMR-WB mode 3 and octet-aligned payload format SHALL be used for this scenario. Moreover, to avoid signaling conflicts in the IP network, two sessions SHALL be established using SIP/SDP, one between the VMR-WB enabled terminal and the gateway and another session between the gateway and the AMR-WB enabled terminal. Note that no transcoding is involved since the VMR-WB payload is identical to that of AMR-WB.

The CMR value may be set in packets received by the gateways on the IP network side. The gateway should forward to the non-IP side a CMR value that is the minimum of two values (1) the CMR value it receives on the IP side; and (2) a CMR value

it may choose for congestion control of transmission on the IP side.

The details of the traffic control algorithm are left to the

implementation.

During and upon initiation of an interoperable interconnection between VMR-WB and AMR-WB, only VMR-WB mode 3 SHALL be used. There are three Frame Types (i.e., FT=0, 1, or 2 see Table 3) within this mode that are compatible with AMR-WB codec modes 0, 1, and 2, respectively.

If the AMR-WB codec is engaged in an interoperable interconnection with VMR-WB, the active AMR-WB codec mode set SHALL be limited to 0, 1, and 2.

5.5 GW to GW (Between two VMR-WB Enabled Terminals)

The fifth example VoIP scenario comprises a RTP/UDP/IP transport between two non-IP systems, i.e., IP is originated and terminated in gateways on both sides of the IP transport, as illustrated in Figure 6. This is the most likely scenario for Mobile Station-to-Mobile Station (MS-to-MS) Transcoder-Free (TrFO) interconnection between two 3GPP2/CDMA2000 terminals that both use VMR-WB codec.

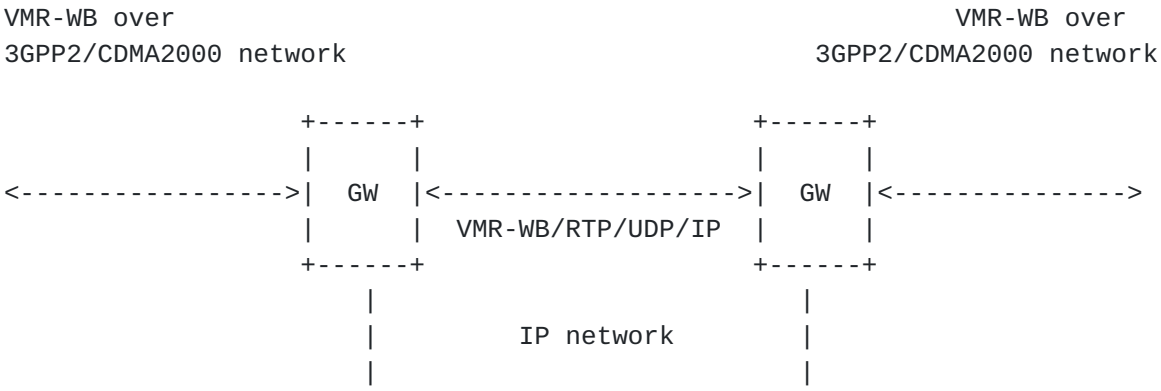


Figure 6: GW to GW scenario (a CDMA2000 MS-to-MS voice over IP scenario)

6. VMR-WB RTP Payload Formats

For a given session, the payload format can be either header free or octet-aligned, depending on the mode of operation that is established for the session via out-of-band means and the application.

The header-free payload format is designed for maximum bandwidth efficiency, simplicity, and low latency. Only one codec data frame can be sent in each header-free payload format. None of the payload header fields or ToC entries is

present [11].

In the octet-aligned payload format, all the fields in a payload, including payload header, table of contents entries,

and speech frames themselves, are individually aligned to octet boundaries to make implementations efficient.

Note that octet alignment of a field or payload means that the last octet is padded with zeroes in the least significant bits to fill the octet. Also note that this padding is separate from padding indicated by the P bit in the RTP header.

Between the two payload formats, only the octet-aligned format has the capability to use the interleaving to make the speech transport robust to packet loss.

The VMR-WB octet-aligned payload format in the interoperable mode is identical to that of AMR-WB (i.e., RFC 3267).

Implementations SHOULD support both header-free and octet-aligned payload formats to increase interoperability.

6.1. RTP Header Usage

The format of the RTP header is specified in [3]. This payload format uses the fields of the header in a manner consistent with that specification.

The RTP timestamp corresponds to the sampling instant of the first sample encoded for the first frame-block in the packet. The timestamp clock frequency is the same as the sampling frequency, so the timestamp unit is in samples.

The duration of one speech frame-block is 20 ms for VMR-WB. For normal wideband operation of VMR-WB, the input/output sampling frequency is 16 kHz, corresponding to 320 samples per frame from each channel. Thus, the timestamp is increased by 320 for VMR-WB for each consecutive frame-block.

For narrowband operation of VMR-WB, the input/output sampling frequency is 8 kHz, corresponding to 160 encoded speech samples per frame from each channel. Thus, the timestamp is increased by 160 for VMR-WB for each consecutive frame-block while processing narrowband input/output speech signals. The choice of sampling frequency MUST be indicated in the beginning of a session (see section 10). The default input/output sampling rate is 16 kHz. Note that during a session, the change of sampling rate is not RECOMMENDED.

A packet may contain multiple frame-blocks of encoded speech or comfort noise parameters. If interleaving is employed, the frame-blocks encapsulated into a payload are picked according

to the interleaving rules as defined in Section 6.3.2. Otherwise, each packet covers a period of one or more contiguous 20 ms frame-block intervals. In case the data from all the channels for a particular frame-block in the period

is missing, for example at a gateway from some other transport format, it is possible to indicate that no data is present for that frame-block rather than breaking a multi-frame-block packet into two, as explained in Section 6.3.2.

The payload is always made an integral number of octets long by padding with zero bits if necessary. If additional padding is required to bring the payload length to a larger multiple of octets or for some other purpose, then the P bit in the RTP header MAY be set and padding appended as specified in [3].

The RTP header marker bit (M) SHALL be always set to 0 if the VMR-WB codec operates in continuous transmission. When operating in discontinuous transmission (DTX), the RTP header marker bit SHALL be set to 1 if the first frame-block carried in the packet contains a speech frame, which is the first in a talkspurt. For all other packets the marker bit SHALL be set to zero (M=0).

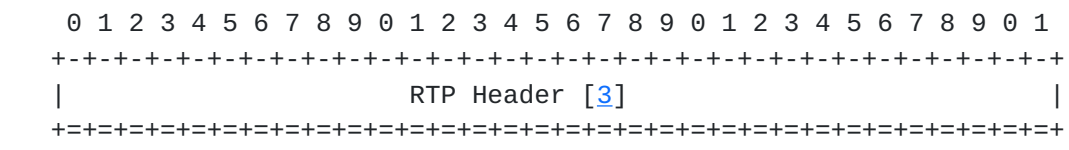
The assignment of an RTP payload type for this new packet format is outside the scope of this document, and will not be specified here. It is expected that the RTP profile under which this payload format is being used will assign a payload type for this encoding or specify that the payload type is to be bound dynamically.

6.2. Header-Free Payload Format

The header-free Packet payload format is designed for maximum bandwidth efficiency, simplicity, and minimum delay. Only one speech data frame can be sent in each header-free payload format. None of the payload header fields or ToC entries is present. The encoding rate for the speech frame can be determined from the length of the speech data frame, since there is only one speech data frame in each header-free payload format.

Use of the RTP header fields for header-free payload format is the same as the corresponding one for the octet-aligned payload format. The detailed bit mapping of speech data packets permissible for this payload format is described in Section 8 of [1].

Since the header-free payload format is not compatible with AMR-WB, it is RECOMMENDED that only VMR-WB modes 0, 1, and 2 be used with this payload format.


0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
RTP Header [3]

```
|                                                                 |
+          ONLY one speech data frame           +-+-+-+-+-+-+-+-+
|                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Note that the mode of operation, using this payload format, is decided by the transmitting (encoder) site. The default mode of operation for VMR-WB encoder is mode 0 [1]. The mode change request MAY also be sent through non-RTP means, which is out of the scope of this specification.

6.3. Octet-Aligned Payload Format

6.3.1 Payload Structure

The complete payload consists of a payload header, a payload table of contents, and speech data representing one or more speech frame-blocks. The following diagram shows the general payload format layout:

```
+----------------+-------------------+----------------
| Payload header | Table of contents | Speech data ...
+----------------+-------------------+----------------
```

6.3.2. The Payload Header

In octet-aligned payload format the payload header consists of a 4-bit CMR, 4 reserved bits, and optionally, an 8 bit-interleaving header, as shown below

```
 0                   1
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-+-+-+-+-+-+-+-+- - - - - - - -
|  CMR  |R|R|R|R|  ILL  |  ILP  |
+-+-+-+-+-+-+-+-+- - - - - - - -
```

CMR (4 bits): Indicates a codec mode request sent to the speech encoder at the site of the receiver of this payload, provided that the network allows the use of the requested mode.

The value of the CMR field is set according to the following Table

```
+-------+-----------------------------------------------------------+
| CMR   |                VMR-WB Operating Modes                     |
+-------+-----------------------------------------------------------+
|   0   | VMR-WB mode 3 (AMR-WB interoperable mode at 6.60 kbps)    |
```

1	VMR-WB mode 3 (AMR-WB interoperable mode at 8.85 kbps)
2	VMR-WB mode 3 (AMR-WB interoperable mode at 12.65 kbps)
3	VMR-WB mode 2
4	VMR-WB mode 1

5	VMR-WB mode 0
6-14	(reserved)
15	No Preference (Operating mode SHOULD be set by the network)

Table 2: List of valid CMR values and their associated VMR-WB operating modes.

R: is a reserved bit that MUST be set to zero. The receiver MUST ignore all R bits.

ILL (4 bits, unsigned integer): This is an OPTIONAL field that is present only if interleaving is signaled out-of-band for the session. ILL=L indicates to the receiver that the interleaving length is L+1, in number of frame-blocks.

ILP (4 bits, unsigned integer): This is an OPTIONAL field that is present only if interleaving is signaled. ILP MUST take a value between 0 and ILL, inclusive, indicating the interleaving index for frame-blocks in this payload in the interleave group. If the value of ILP is found greater than ILL, the payload SHOULD be discarded.

ILL and ILP fields MUST be present in each packet in a session if interleaving is signaled for the session.

The mode request received in the CMR field is valid until the next CMR is received, i.e. a newly received CMR value overrides the previous one. Therefore, if a terminal continuously wishes to receive frames in the same mode x, it needs to set CMR=x for all its outbound payloads, and if a terminal has no preference in which mode to receive, it SHOULD set CMR=15 in all its outbound payloads.

If receiving a payload with a CMR value, which is not valid, the CMR MUST be ignored by the receiver.

In a multi-channel session, CMR SHOULD be interpreted by the receiver of the payload as the desired encoding mode for all the channels in the session, if the network allows.

An IP end-point SHOULD NOT set the CMR based on packet losses or other congestion indications, for several reasons

- The other end of the IP path may be a gateway to a non-IP network (such as a radio link) that needs to set the CMR field to optimize performance on that network.

- Congestion on the IP network is managed by the IP sender, in this case at the other end of the IP path. Feedback about congestion SHOULD be provided to that IP sender

through RTCP or other means, and then the sender can choose to avoid congestion using the most appropriate mechanism. That may include adjusting the codec mode, but also includes adjusting the level of redundancy or number

of frames per packet.

The encoder SHOULD follow a received mode request, but MAY change to a different mode if the network necessitates it, for example to control congestion.

The CMR field MUST be set to 15 for packets sent to a multicast group. The encoder in the speech sender SHOULD ignore mode requests when sending speech to a multicast session but MAY use RTCP feedback information as a hint that a mode change is needed.

If interleaving option is utilized, It MUST be performed on a frame-block basis as oppose to a frame basis in a multi-channel session.

The following example illustrates the arrangement of speech frame-blocks in an interleave group during an interleave session. Here we assume ILL=L for the interleave group that starts at speech frame-block n. We also assume that the first payload packet of the interleave group is s and the number of speech frame-blocks carried in each payload is N. Then we will have

Payload s (the first packet of this interleave group):
ILL=L, ILP=0,
Carry frame-blocks: n, n+(L+1), n+2*(L+1),..., n+(N-1)*(L+1)

Payload s+1 (the second packet of this interleave group):
ILL=L, ILP=1,
Carry frame-blocks: n+1, n+1+(L+1), n+1+2*(L+1),..., n+1+(N-1)*(L+1)

...

Payload s+L (the last packet of this interleave group):
ILL=L, ILP=L,
Carry frame-blocks: n+L, n+L+(L+1), n+L+2*(L+1), ..., n+L+(N-1)*(L+1)

The next interleave group will start at frame-block n+N*(L+1).

There will be no interleaving effect unless the number of frame-blocks per packet (N) is at least 2. Moreover, the number of frame-blocks per payload (N) and the value of ILL MUST NOT be changed inside an interleave group. In other words, all payloads in an interleave group MUST have the same ILL and MUST contain the same number of speech frame-blocks.

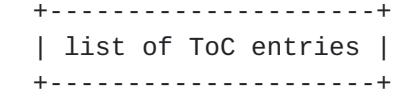
The sender of the payload MUST only apply interleaving if the receiver has signaled its use through out-of-band means. Since interleaving will increase buffering requirements at

the receiver, the receiver uses MIME parameter "interleaving=I" to set the maximum number of frame-blocks allowed in an interleaving group to I.

When performing interleaving the sender MUST use a proper number of frame-blocks per payload (N) and ILL so that the resulting size of an interleave group is less than or equal to I, i.e., N*(L+1)<=I.

6.3.3. The Payload Table of Contents

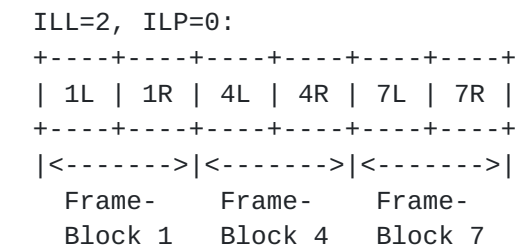
The table of contents (ToC) in octet-aligned payload format consists of a list of ToC entries where each entry corresponds to a speech frame carried in the payload, i.e.,



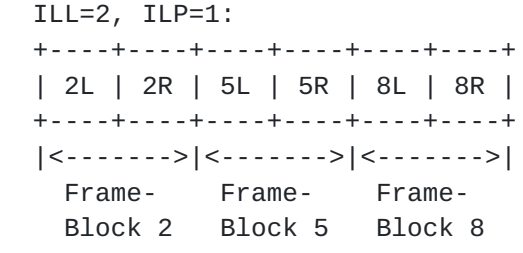
When interleaving is used, the frame-blocks in the ToC will almost never be placed consecutive in time. Instead, the presence and order of the frame-blocks in a packet will follow the pattern described in 6.3.2.

The following example shows the ToC of three consecutive packets, each carrying 3 frame-blocks, in an interleaved two channel session. Here, the two channels are left (L) and right (R) with L coming before R, and the interleaving length is 3 (i.e., ILL=2). This makes the interleave group 9 frame-blocks large.

Packet #1

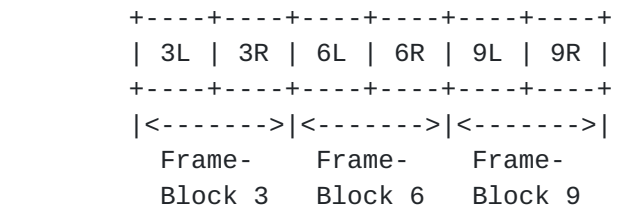


Packet #2

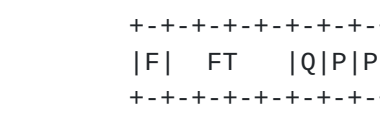


Packet #3

ILL=2, ILP=2:



A ToC entry for the octet-aligned payload format is as follows:



The table of contents (ToC) consists of a list of ToC entries, each representing a speech frame.

F (1 bit): If set to 1, indicates that this frame is followed by another speech frame in this payload; if set to 0, indicates that this frame is the last frame in this payload.

FT (4 bits): Frame type index whose value is chosen according to the following Table.

FT	Encoding Rate	Frame Size (Bits)
0	Interoperable Full-Rate (AMR-WB 6.60 kbps)	132
1	Interoperable Full-Rate (AMR-WB 8.85 kbps)	177
2	Interoperable Full-Rate (AMR-WB 12.65 kbps)	253
3	Full-Rate 13.3 kbps	266
4	Half-Rate 6.2 kbps	124
5	Quarter-Rate 2.7 kbps	54
6	Eighth-Rate 1.0 kbps	20
7	(reserved)	
8	(reserved)	
9	CNG (AMR-WB SID)	35
10	(reserved)	
11	(reserved)	
12	(reserved)	
13	(reserved)	
14	Erasure (AMR-WB SPEECH_LOST)	0
15	Blank (AMR-WB NO_DATA)	0

Table 3:VMR-WB payload frame types for real-time (or non real-time) transport and storage

During the interoperable mode, FT=14 (SPEECH_LOST) and FT=15

(NO_DATA) are used to indicate frames that are either lost or not being transmitted in this payload, respectively. FT=14 or 15 MAY be used in the non-interoperable modes to indicate frame erasure or blank frame, respectively (see Section 2.1 of [1]).

Note that for ToC entries with FT=14 or 15, there will be no corresponding speech frame in the payload.

Q (1 bit): Frame quality indicator. If set to 0, indicates the corresponding frame is corrupted. During the interoperable mode, the receiver side (with AMR-WB codec) should set the RX_TYPE to either SPEECH_BAD or SID_BAD depending on the frame type (FT), if Q=0. The VMR-WB encoder always sets Q bit to 1.

P bits: Padding bits MUST be set to zero.

For multi-channel sessions, the ToC entries of all frames from a frame-block are placed in the ToC in consecutive.

Therefore, with N channels and K speech frame-blocks in a packet, there MUST be N*K entries in the ToC, and the first N entries will be from the first frame-block, the second N entries will be from the second frame-block, and so on.

6.3.4. Speech Data

Speech data of a payload contains one or more speech as described in the ToC of the payload.

Each speech frame represents 20 ms of speech encoded in one of the available encoding rates depending on the operation mode. The length of the speech frame is defined by the frame type in the FT field with the following considerations:

- The last octet of each speech frame MUST be padded with zeroes at the end if not all bits in the octet are used. In other words, each speech frame MUST be octet-aligned.

- When multiple speech frames are present in the speech data, the speech frames MUST be arranged one whole frame after another.

The order and numbering notation of the speech data bits are as specified in the VMR-WB standard specification [1].

The payload begins with the payload header of one octet or two if frame interleaving is selected. The payload header is followed by the table of contents consisting of a list of one-octet ToC entries.

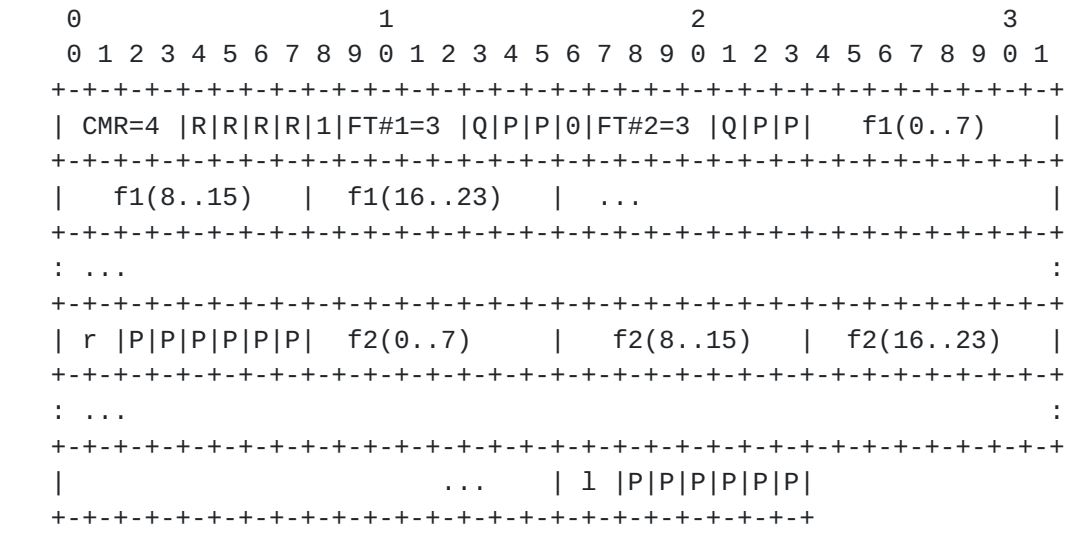
The speech data follows the table of contents. For packetization in the normal order, all of the octets comprising a speech frame are appended to the payload as a

unit. The speech frames are packed in the same order as their corresponding ToC entries are arranged in the ToC list, with the exception that if a given frame has a ToC entry with FT=14 or 15, there will be no data octets present for that

frame.

6.3.5. Payload Example: Basic Single Channel Payload Carrying Multiple Frames

The following diagram shows an octet-aligned payload format from a single channel session that carries two VMR-WB Full-Rate frames (FT=3). In the payload, a codec mode request is sent (e.g., CMR=4), requesting the encoder at the receiver's side to use VMR-WB mode 1. No interleaving is used.



r= f1(264,265)
l= f2(264,265)

Note, in above example the last octet in both speech frames is padded with zeros to make them octet-aligned.

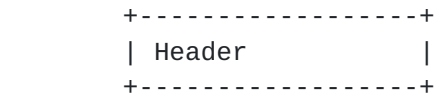
6.4. Implementation Considerations

An application implementing this payload format MUST understand all the payload parameters in the out-of-band signaling used. For example, if an application uses SDP, all the SDP and MIME parameters in this document MUST be understood. This requirement ensures that an implementation always can decide if it is capable or not of communicating.

7. VMR-WB Storage Format

The storage format is used for storing VMR-WB encoded speech frames in a file or as an e-mail attachment. Multiple channel content is also supported.

In general, VMR-WB file has the following structure:


Header

```
| Speech frame 1   |
+------------------+
: ...              :
+------------------+
| Speech frame n   |
+------------------+
```

7.1. Single channel Header

A single channel VMR-WB file header contains only a magic number.

The magic number for single channel VMR-WB files containing speech data generated in the non-interoperable modes; i.e., VMR-WB modes 0, 1, or 2, MUST consist of ASCII character string

"#!VMR-WB\n"
(or 0x2321564d522d57420a in hexadecimal).

Note, the "\n" is an important part of the magic numbers and MUST be included in the comparison; otherwise, the single channel magic number above will become indistinguishable from that of the multi-channel file defined in the next section.

The magic number for single channel VMR-WB files containing speech data generated in the interoperable mode; i.e., VMR-WB mode 3, MUST consist of ASCII character string

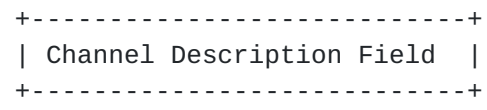
"#!VMR-WB_I\n"
(or 0x2321564d522d57425F490a in hexadecimal).

In the interoperable mode, a file generated by VMR-WB is decodable with AMR-WB (with the exception of different magic numbers). However, to ensure compatibility and because VMR-WB can only decode AMR-WB codec modes 0, 1, or 2, AMR-WB codec SHOULD be instructed not to generate the modes that are not in common so that files generated by AMR-WB can be decoded by VMR-WB.

7.2. Multi-channel Header

The multi-channel header consists of a magic number followed by a 32-bit channel description field, giving the multi-channel header the following structure:

```
+----------------------------+
|        Magic Number        |
```


Channel Description Field

The magic number for multi-channel VMR-WB files containing speech data generated in the non-interoperable modes; i.e., VMR-WB modes 0, 1, or 2, MUST consist of the ASCII character string

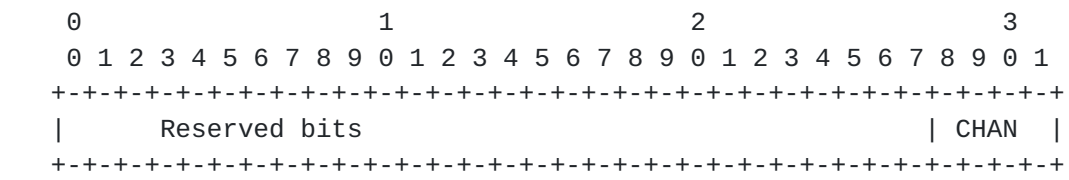
"#!VMR-WB_MC1.0\n"
(or 0x2321564d522d57425F4D43312E300a in hexadecimal).

The version number in the magic numbers refers to the version of the file format.

The magic number for multi-channel VMR-WB files containing speech data generated in the interoperable mode; i.e., VMR-WB mode 3, MUST consist of the ASCII character string

"#!VMR-WB_MCI1.0\n"
(or 0x2321564d522d57425F4D4349312E300a in hexadecimal).

The 32-bit channel description field is defined as



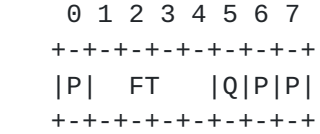
Reserved bits: MUST be set to 0 when written, and a reader MUST ignore them.

CHAN (4 bit unsigned integer): Indicates the number of audio channels contained in this storage file. The valid values and the order of the channels within a frame-block are specified in Section 4.1 in [10].

7.3. Speech Frames

After the file header, speech frame-blocks consecutive in time are stored in the file. Each frame-block contains a number of octet-aligned speech frames equal to the number of channels, and stored in increasing order, starting with channel 1.

Each stored speech frame starts with a one-octet frame header with the following format:

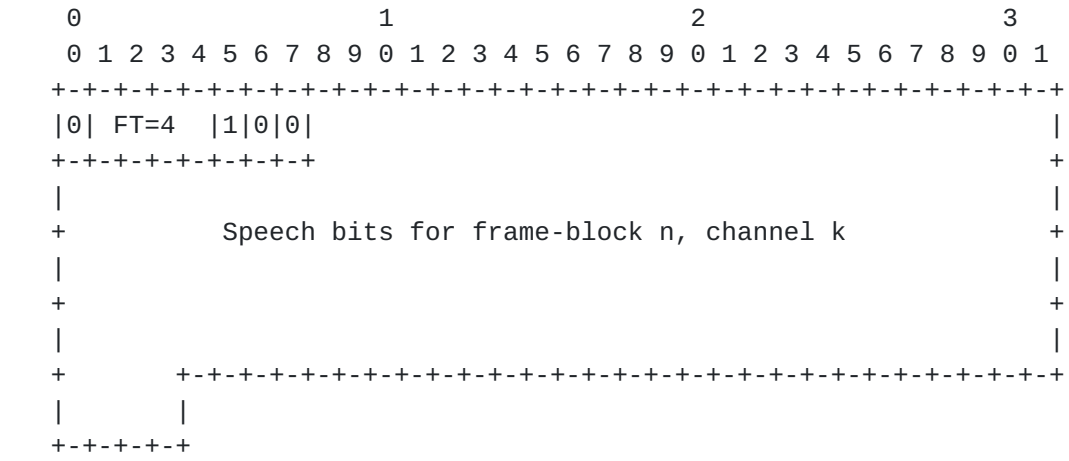


The FT field is defined as shown in Table 3. The P bits are padding and MUST be set to 0.

Q (1 bit): Frame quality indicator. If set to 0, indicates the corresponding frame is corrupted. The VMR-WB encoder always sets Q bit to 1.

Following this one octet header, the speech bits are placed as defined in 6.3.4. The last octet of each frame is padded with zeroes, if needed, to achieve octet alignment.

The following example shows a VMR-WB speech frame encoded at Half-Rate (with 124 speech bits) in the storage format.



Frame-blocks or speech frames that are lost in transmission and thereby not received MUST be stored as Blank/NO_DATA frames (FT=15) or Erasure/SPEECH_LOST (FT=14) in complete frame-blocks to keep synchronization with the original media.

8. Congestion Control

The general congestion control considerations for transporting RTP data apply to VMR-WB speech over RTP as well. However, the multimode capability of VMR-WB speech coding may provide an advantage over other payload formats for controlling congestion since the bandwidth demand can be adjusted by selecting a different operating mode (i.e., mode switching).

Another parameter that may impact the bandwidth demand for VMR-WB is the number of frame-blocks that are encapsulated in each RTP payload. Packing more frame-blocks in each RTP payload can reduce the number of packets sent and hence the overhead from RTP/UDP/IP headers, at the expense of increased delay.

If forward error correction (FEC) is used to alleviate the packet loss, the amount of redundancy added by FEC will need

to be regulated so that the use of FEC itself does not cause a congestion problem.

It is RECOMMENDED that VMR-WB applications using this payload

format employ congestion control. The actual mechanism for congestion control is not specified but should be suitable for real-time transport of datagrams.

9. Security Considerations

RTP packets using the payload formats defined in this specification are subject to the general security considerations discussed in [3].

As this format transports encoded speech, the main security issues include confidentiality and authentication of the speech itself. The payload format itself does not have any built-in security mechanisms. External mechanisms, such as SRTP [8], MAY be used.

This payload format does not exhibit any significant non-uniformity in the receiver side computational complexity for packet processing and thus is unlikely to pose a denial-of-service threat due to the receipt of pathological/corrupted data.

9.1. Confidentiality

To achieve confidentiality of the encoded VMR-WB speech, all speech data bits MAY be encrypted. There is no need to encrypt the payload header or the table of contents due to the following reasons:

1) They only carry information about the requested speech mode, frame type, and frame quality

2) This information could be useful to some third party, e.g., quality monitoring.

As long as the VMR-WB payload is only packed and unpacked at either end, encryption may be performed after packet encapsulation so that there is no conflict between the two operations.

Interleaving may affect encryption. Depending on the encryption scheme used, there may be restrictions on, for example, the time when keys can be changed. Specifically, the key change may need to occur at the boundary between interleave groups.

The type of encryption method used may impact the error robustness of the payload data. The error robustness may be

severely reduced when the data is encrypted unless an encryption method without error-propagation is used, e.g. a stream cipher.

9.2. Authentication

To authenticate the sender of the speech, an external mechanism MUST be used. It is RECOMMENDED that such a mechanism protect all the speech data bits.

Data tampering by a man-in-the-middle attacker could result in erroneous depacketization/decoding that could lower the speech quality. For example, tampering with the CMR field may result in speech in a different quality than desired.

To prevent a man-in-the-middle attacker from tampering with the payload packets, some additional information besides the speech bits SHOULD be protected.

This may include the payload header, ToC, RTP timestamp, RTP sequence number, and the RTP marker bit.

9.3. Decoding Validation and Provision for Lost or Late Packets

When processing a received payload packet, if the receiver finds that the calculated payload length, based on the information of the session and the values found in the payload header fields, do not match the size of the received packet, the receiver SHOULD discard the packet to avoid potential degradation of speech quality and to invoke the VMR-WB built-in frame error concealment mechanism. Therefore, invalid packets SHALL be treated as lost packets.

Late packets (i.e., unavailability of a packet when needed for decoding at the receiver) SHALL be treated as lost packets. Furthermore, if the late packet is part of an interleave group, depending upon the availability of the other packets in that interleave group, decoding MUST be resumed from the next (sequential order) available packet. In other words, the unavailability of a packet in an interleave group at certain time SHOULD not invalidate the other packets within that interleave group that MAY arrive later.

10. Payload Format Parameters

This section defines the parameters that may be used to select optional features in the VMR-WB payload. The parameters are defined here as part of the MIME subtype registration for the VMR-WB speech codec. A mapping of the parameters into the Session Description Protocol (SDP) [5] is also provided for those applications that use SDP. Equivalent parameters could be defined elsewhere for use with control

protocols that do not use MIME or SDP.

The data format and parameters are specified for both real-time transport in RTP and for storage type applications such as e-mail

attachments.

10.1. VMR-WB MIME Registration

The MIME subtype for the Variable-Rate Multimode Wideband (VMR-WB) audio codec is allocated from the IETF tree since VMR-WB is expected to be a widely used speech codec in multimedia streaming and messaging as well as VoIP applications. This MIME registration covers both real-time transfer via RTP and non-real-time transfers via stored files.

Note, the receiver MUST ignore any unspecified parameter and use the default values instead.

Media Type name: audio

Media subtype name: VMR-WB

Required parameters: none

Note that if no input parameters are defined, the default values will be used.

Also note that "crc" and "robust-sorting" parameters from RFC 3267 [4] are not applicable to VMR-WB RTP payload and storage file formats. To ensure compatibility between VMR-WB and AMR-WB in the interoperable sessions, one SHOULD make sure that AMR-WB does not utilize crc and robust-sorting (i.e., these options are deactivated in the session initiation).

OPTIONAL parameters:
These parameters apply to RTP transfer only.

payload_format: Permissible values are 0 and 1. If 1, octet-aligned payload format SHALL be used. If 0 or if not present, header-free payload format is employed (default).

maxptime: The maximum amount of media, which can be encapsulated in a payload packet, expressed as time in milliseconds. The time is calculated as the sum of the time the media present in the packet represents. The time SHALL be an integer multiple of the frame size. If this parameter is not present, the sender MAY encapsulate any number of speech frames into one RTP packet.

interleaving: Indicates that frame-block level interleaving SHALL be used for the session and its value defines the maximum number of frame-blocks allowed in an interleaving

group (see Section 6.3.1). If this parameter is not present, interleaving SHALL not be used. The presence of this parameter also implies automatically that octet-aligned operation SHALL be used.

ptime: see RFC2327 [5]. It SHALL be at least one frame size for VMR-WB.

channels: The number of audio channels. The possible values and their respective channel order is specified in section 4.1 in [10]. If omitted it has the default value of 1.

These parameters apply to both real-time and non-real-time transfers

dtx: Permissible values are 0 and 1. The default is 0 (i.e., No DTX) where VMR-WB normally operates as a continuous variable-rate codec. If dtx=1, the VMR-WB codec will operate in discontinuous transmission mode where silence descriptor (SID) frames are sent by the VMR-WB encoder during silence intervals with an adjustable update frequency. The selection of the SID update-rate depends on the implementation and other network considerations that are beyond the scope of this specification.

Encoding considerations:
This type is defined for transfer via both RTP (RFC 3550) and stored-file methods as described in Sections 6 and 7, respectively, of RFC XXXX. Audio data is binary data, and must be encoded for non-binary transport; the Base64 encoding is suitable for Email.

Security considerations:
See Section 9 of RFC XXXX.

Public specification:
The VMR-WB speech codec is specified in following 3GPP2 specifications C.S0052-0 version 1.0.
Transfer methods are specified in RFC XXXX.

Additional information:
The following applies to stored-file transfer methods:

Magic numbers:
Single channel (for the non-interoperable modes)

ASCII character string "#!VMR-WB\n"
(or 0x2321564d522d57420a in hexadecimal)

Single channel (for the interoperable mode)

```
          ASCII character string "#!VMR-WB_I\n"
          (or 0x2321564d522d57425F490a in hexadecimal)

          Multi-channel (for the non-interoperable modes)
          ASCII character string "#!VMR-WB_MC1.0\n"
          (or 0x2321564d522d57425F4D43312E300a in hexadecimal)

           Multi-channel (for the interoperable mode)
          ASCII character string "#!VMR-WB_MCI1.0\n"
          (or 0x2321564d522d57425F4D4349312E300a in hexadecimal)

   File extensions for the non-interoperable modes: vmr, VMR
                         Macintosh file type code: none
                         Object identifier or OID: none

   File extensions for the interoperable mode: vmi, VMI
                    Macintosh file type code: none
                    Object identifier or OID: none

   Person & email address to contact for further information:
                 Sassan Ahmadi, Ph.D.   Nokia Inc. USA
                 sassan.ahmadi@nokia.com

   Intended usage: COMMON.
     It is expected that many VoIP, multimedia messaging and
     streaming applications (as well as mobile applications)
     will use this type.

   Author/Change controller:
                 Sassan Ahmadi, Ph.D.   Nokia Inc. USA
                 sassan.ahmadi@nokia.com
               IETF Audio/Video Transport Working Group
```

10.2. Mapping MIME Parameters into SDP

The information carried in the MIME media type specification has a specific mapping to fields in the Session Description Protocol (SDP) [5], which is commonly used to describe RTP sessions. When SDP is used to specify sessions employing the VMR-WB codec, the mapping is as follows:

- The MIME type ("audio") goes in SDP "m=" as the media name.
- The MIME subtype (payload format name) goes in SDP "a=rtpmap" as the encoding name. The RTP clock rate in "a=rtpmap" MUST be 16000 for VMR-WB (Note that 8000 is also supported by VMR-WB for narrowband I/O processing), and the encoding parameters (number of channels) MUST either be explicitly set to N or omitted, implying a

default value of 1. The values of N that are allowed is specified in Section 4.1 in [10].

- The parameters "ptime" and "maxptime" go in the SDP

"a=ptime" and "a=maxptime" attributes, respectively.
- Any remaining parameters go in the SDP "a=fmtp" attribute by copying them directly from the MIME media type string as a semicolon separated list of parameter=value pairs.

Some example SDP session descriptions utilizing VMR-WB encodings follow. In these examples, long a=fmtp lines are folded to meet the column width constraints of this document; the backslash ("\") at the end of a line and the carriage return that follows it should be ignored.

Example of usage of VMR-WB in a possible VoIP scenario (wideband audio):

```
m=audio 49120 RTP/AVP 98
a=rtpmap:98 VMR-WB/16000
a=fmtp:98 payload_format=1
```

Example of usage of VMR-WB in a possible VoIP scenario (narrowband audio):

```
m=audio 49120 RTP/AVP 98
a=rtpmap:98 VMR-WB/8000
a=fmtp:98
```

Example of usage of VMR-WB in a possible streaming scenario (two channel stereo):

```
m=audio 49120 RTP/AVP 99
a=rtpmap:99 VMR-WB/16000/2
a=fmtp:99 interleaving=30
a=maxptime:100 payload_format=1
```

10.3. Offer-Answer Model Considerations

To achieve good interoperability for the VMR-WB RTP payload in an Offer-Answer negotiation usage in SDP the following considerations SHOULD be made:

- Both header-free and octet-aligned payload formats MAY be offered by a VMR-WB enabled terminal. However, for an interoperable interconnection with AMR-WB only octet-aligned payload format SHALL be used.

- The parameters "maxptime" and "ptime" should in most cases not affect the interoperability, however the setting of the parameters can affect the performance of the application.

- To maintain interoperability with AMR-WB in cases where negotiation is possible using the VMR-WB interoperable mode, a

VMR-WB enabled terminal SHOULD also declare itself capable of AMR-WB with limited mode set (i.e., only AMR-WB codec modes 0, 1, and 2 are allowed) and octet-align mode of operation. Example:

```
m=audio 49120 RTP/AVP 98 99
a=rtpmap:98 VMR-WB/16000/1
a=rtpmap:99 AMR-WB/16000/1
a=fmtp:99 octet-align=1; mode-set=0,1,2
```

11. IANA Considerations

The new attributes "dtx" and "payload_format" need to be registered. The definition of the "maxptime" attribute used in this specification is consistent with the corresponding parameter in RFC 3267.

12. Acknowledgements

The author would like to thank Redwan Salami of VoiceAge Corporation, Ari Lakaniemi of Nokia Inc., and IETF/AVT chairs Colin Perkins and Magnus Westerlund for their technical comments to improve this document.

Also, the author would like to acknowledge that some parts of RFC 3267 [4] and RFC 3558 [11] have been used in this document.

References

Normative References

[1] 3GPP2 C.S0052-0 "Source-Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB) Service Option 62 for Wideband Spread Spectrum Communication Systems", 3GPP2 Technical Specification, June 2004.

[2] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels", IETF RFC 2119, March 1997.

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

[4] J. Sjoberg, et al., "Real-Time Transport Protocol (RTP) Payload Format and File Storage Format for the Adaptive Multi-Rate (AMR) and Adaptive Multi-Rate Wideband (AMR-WB) Audio Codecs", IETF RFC 3267, June 2002.

[5] M. Handley and V. Jacobson, "SDP: Session Description Protocol", IETF RFC 2327, April 1998.

[6] M. Handley, S. Floyd, J. Padhye, J. Widmer, "TCP Friendly Rate Control (TFRC): Protocol Specification",

IETF RFC 3448, January 2003.

[7] J. Rosenberg, and H. Schulzrinne, "An RTP Payload Format for Generic Forward Error Correction", IETF RFC 2733, December 1999.

[8] Baugher, et al., "The Secure Real Time Transport Protocol", IETF Draft (Work in Progress), November 2001.

[9] C. Perkins, et al., "RTP Payload for Redundant Audio Data", IETF RFC 2198, September 1997.

[10] H. Schulzrinne, "RTP Profile for Audio and Video Conferences with Minimal Control" IETF RFC 3551, July 2003.

[11] A. Li, "RTP Payload Format for Enhanced Variable Rate Codecs (EVRC) and Selectable Mode Vocoders (SMV)", IETF RFC 3558, July 2003.

[12] 3GPP TS 26.193 "AMR Wideband Speech Codec; Source Controlled Rate operation", version 5.0.0 (2001-03), 3rd Generation Partnership Project (3GPP).

Any 3GPP2 document can be downloaded from the 3GPP2 web server, "http://www.3gpp2.org/", see specifications.

Author's Address

The editor will serve as the point of contact for all technical matters related to this document.

Dr. Sassan Ahmadi
Phone: 1 (858) 831-5916
Fax: 1 (858) 831-4174
Email: sassan.ahmadi@nokia.com

Nokia Inc.
12278 Scripps Summit Dr.
San Diego, CA 92131 USA

This Internet-Draft expires in six months from May 17, 2004.

Full Copyright Statement