Network Working Group
Request for Comments: 4348
Category: Standards Track

S. Ahmadi

January 2006

# Real-Time Transport Protocol (RTP) Payload Format for the Variable-Rate Multimode Wideband (VMR-WB) Audio Codec

## 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. A media type registration is included for VMR-WB RTP payload format.

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 document 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 is described in Section 6. The VMR-WB file format (i.e., for transport of VMR-WB speech data in storage mode applications such as email) is specified in [7]. In Section 9, a media type registration for VMR-WB RTP payload format 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 format 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:

```
3GPP   - The Third Generation Partnership Project
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 Transport Protocol
RTCP   - RTP Control Protocol
MIME   - Multipurpose Internet Mail Extension
SDP    - Session Description Protocol
VoIP   - Voice-over-IP
```

The term "interoperable mode" in this document refers to VMR-WB mode 3, which is interoperable with AMR-WB codec modes 0, 1, and 2.

The term "non-interoperable modes" in this document refers to VMR-WB modes 0, 1, and 2.

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 [1]. 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 (as shown in Table 2) is chosen based on the traffic condition of the network and the desired quality of service. The desired average data rate (ADR) in each mode is obtained by encoding speech frames at permissible rates (as shown in Tables 1 and 3) compliant with CDMA2000 system, 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 [4,12] 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.

Note that the concept of mode in VMR-WB is different from that of AMR-WB where each fixed-rate AMR-WB codec mode is adapted to prevailing channel conditions by a tradeoff between the total number of source-coding and channel-coding bits.

VMR-WB is able to transition between various modes with no degradation in voice quality that is attributable to the mode switching itself. The operating mode of the VMR-WB encoder may be switched seamlessly without prior knowledge of the decoder. Any non-interoperable mode (i.e., VMR-WB modes 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 between VMR-WB modes is not allowed because there is only one AMR-WB interoperable mode in VMR-WB. Since the AMR-WB codec may request a mode change, depending on channel conditions, 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 VMR-WB frame types 0, 1, and 2 in VMR-WB mode 3 (corresponding to AMR-WB codec modes 0, 1, or 2).

As mentioned earlier, VMR-WB is compliant with CDMA2000 system 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 | 6.2 |
| Quarter-Rate | 54 | 2.7 |
| Eighth-Rate | 20 | 1.0 |
| Blank | 0 | 0 |
| Erasure | 0 | 0 |

Table 1: CDMA2000 system permissible frame types and their associated encoding rates

VMR-WB is robust to high percentage of frame loss and frames 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 frames 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 either 16000-Hz or 8000-Hz sampled input/output speech signals in all modes of operation [1]. The VMR-WB decoder does not require a priori knowledge about the sampling rate of the original media (i.e., speech/audio signals sampled at 8 or 16 kHz) at the input of the encoder. The VMR-WB decoder, by default, generates 16000-Hz wideband output regardless of the encoder input sampling frequency. Depending on the application, the decoder can be configured to generate 8000-Hz output, as well.

Therefore, while this specification defines a 16000-Hz RTP clock rate for VMR-WB codec, the injection and processing of 8000-Hz narrowband media during a session is also allowed; however, a 16000-Hz RTP clock rate MUST always be used.

The choice of VMR-WB output sampling frequency depends on the implementation and the audio acoustic capabilities of the receiving side.

### 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 (e.g., VoIP), 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].

### 3.3. Support for Multi-Channel Session

The octet-aligned RTP payload format defined in this document supports 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 encode and decode each of the individual channels separately.

To transport 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 (see Section 9.2 for more details), 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 [6].

## 4. Robustness against Packet Loss

The octet-aligned payload format described in this document (see Section 6 for more details) 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 [8], 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.

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.

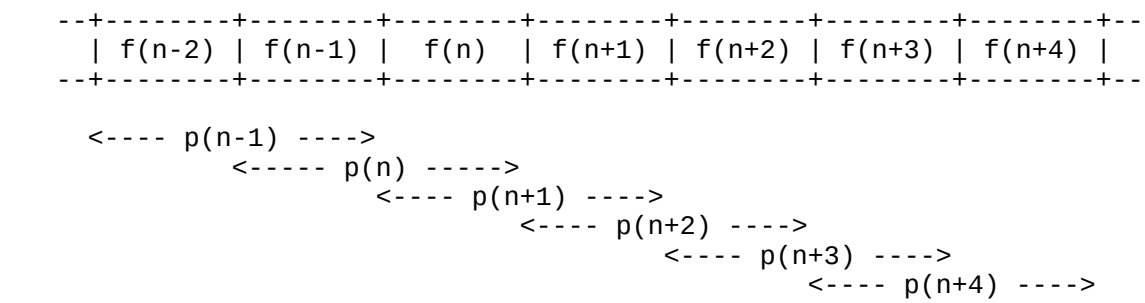



Figure 1: An example of redundant transmission

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 that 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 non-IP information. The sender is also responsible for avoiding congestion, which may be aggravated by redundant transmission (see Section 7).

### 4.2. Frame Interleaving and Multi-Frame Encapsulation

To decrease protocol overhead, the octet-aligned payload format, described in Section 6, allows several speech frame-blocks to be encapsulated into a single RTP packet. One of the drawbacks of this approach is that in case of packet loss several consecutive speech frame-blocks are lost, 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 9).

## 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. Nevertheless, this scenario can be generalized to an interoperable interconnection between VMR-WB-enabled and AMR-WB-enabled IP terminals using the offer-answer model described in Section 9.3. This payload format is expected to be useful for both conversational and streaming services.

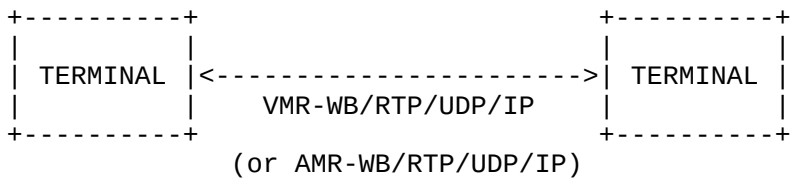



Figure 2: IP terminal to IP terminal

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.

For VMR-WB-enabled IP terminals at both ends, depending on the implementation, all modes of the VMR-WB codec can be used in this scenario. Also, both header-free and octet-aligned payload formats (see Section 6 for details) can be utilized. For the interoperable interconnection between VMR-WB and AMR-WB, only VMR-WB mode 3 is used, and all restrictions described in Section 9.3 apply.

## 5.2. 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 IP terminal, and/or vice versa, as depicted in Figure 3.

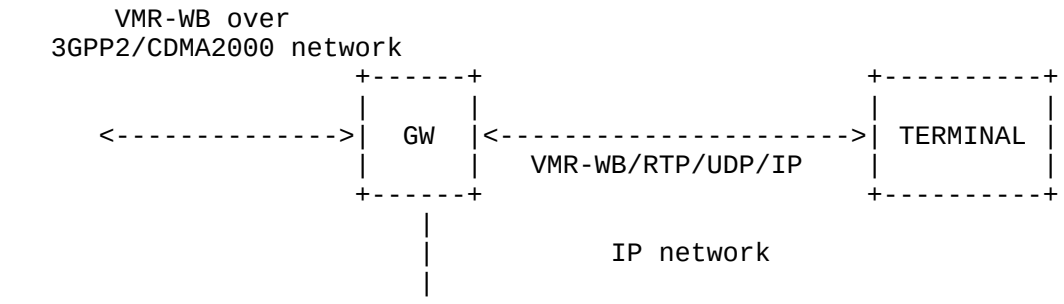



Figure 3: GW to VoIP terminal scenario

VMR-WB's capability to switch seamlessly 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 and to perform congestion control on the IP network.

## 5.3. GW to GW (between VMR-WB- and AMR-WB-Enabled Terminals)

A third 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 4. This is the most likely scenario for an interoperable interconnection between 3GPP/(GSM-WCDMA)/AMR-WB and 3GPP2/CDMA2000/VMR-WB-enabled mobile stations. In this scenario, the VMR-WB-enabled terminal also declares itself capable of AMR-WB with restricted mode set as described in Section 9.3. 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 three values: (1) the CMR value it receives on the IP side; (2) a CMR value it may choose for congestion control of transmission on the IP side; and (3) the CMR value based on its estimate of reception quality on the non-IP side. The details of the traffic control algorithm are left to the implementation.

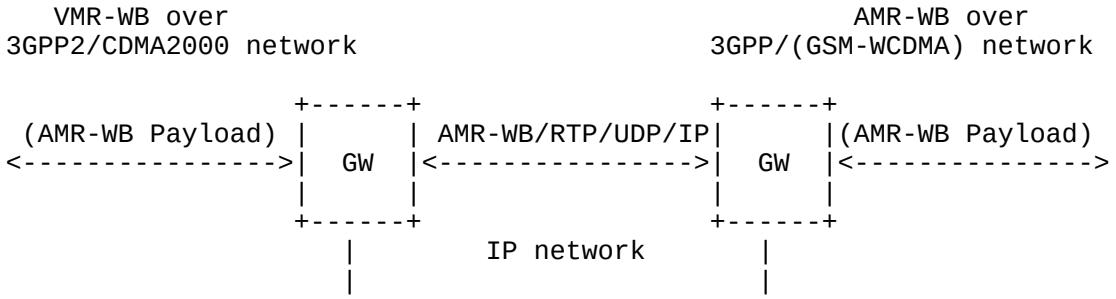



Figure 4: GW to GW scenario (AMR-WB <-> VMR-WB interoperable interconnection)

During and upon initiation of an interoperable interconnection between VMR-WB and AMR-WB, only VMR-WB mode 3 can 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 needs to be limited to 0, 1, and 2.

## 5.4. GW to GW (between Two VMR-WB-Enabled Terminals)

The fourth example VoIP scenario is composed of 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 5. 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.

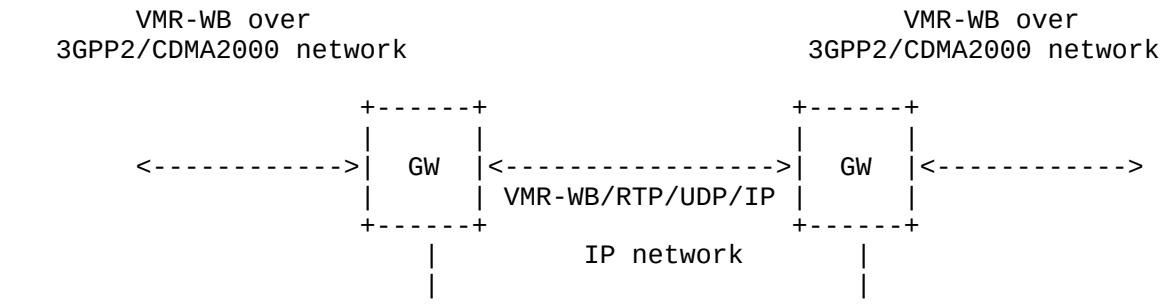



Figure 5: GW to GW scenario (a CDMA2000 MS-to-MS VoIP scenario)

## 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 packet. None of the payload header fields or table of contents (ToC) entries is present (the same consideration is also made in [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).

### 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 default sampling frequency (i.e., 16 kHz), 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 media 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.

The VMR-WB codec is capable of processing speech/audio signals sampled at 8 kHz. By default, the VMR-WB decoder output sampling frequency is 16 kHz. Depending on the application, the decoder can be configured to generate 8-kHz output sampling frequency, as well. Since the VMR-WB RTP payload formats for the 8- and 16-kHz sampled media are identical and the VMR-WB decoder does not need a priori knowledge about the encoder input sampling frequency, a fixed RTP clock rate of 16000 Hz is defined for VMR-WB codec. This would allow injection or processing of 8-kHz sampled speech/audio media without having to change the RTP clock rate during a session. Note that the timestamp is incremented by 320 per frame-block for 8-kHz sampled media, as well.

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 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 instead of breaking a multi-frame-block packet into two, as explained in Section 6.3.2.

No matter which payload format is used, the RTP 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].

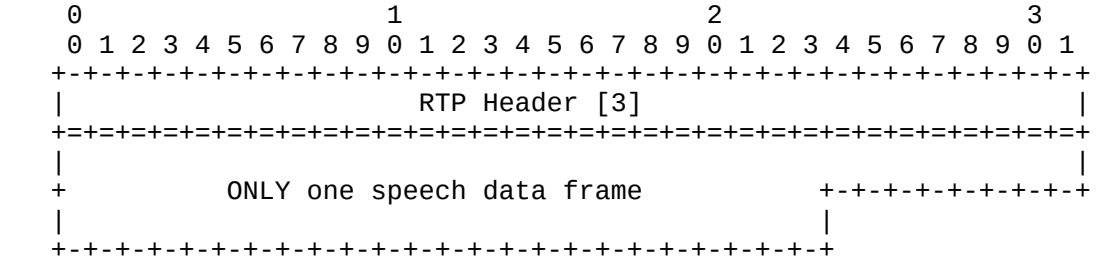
The RTP header marker bit (M) SHALL be always set to 0 if the VMR-WB codec operates in continuous transmission. When operating in discontinuous transmission (DTX), the RTP header marker bit SHALL be set to 1 if the first frame-block carried in the packet contains a speech frame, which is the first in 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 payload 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 (see Section 9).

## 6.2. Header-Free Payload Format

The header-free payload format is designed for maximum bandwidth efficiency, simplicity, and minimum delay. Only one speech data frame presents in each header-free payload format packet. 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.

The 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 RTP payload, only non-interoperable modes of VMR-WB SHALL be used with this payload format. That is, FT=0, 1, 2, and 9 SHALL NOT be used with header-free payload format.

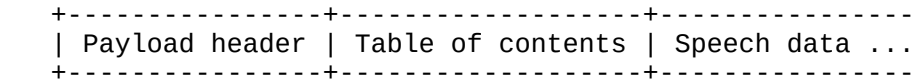



Note that the mode of operation, using this payload format, is decided by the transmitting (encoder) site. The default mode of operation for VMR-WB encoder is mode 0 [1]. The mode change request MAY also be sent through non-RTP means, which is out of the scope of this specification.

## 6.3. Octet-Aligned Payload Format

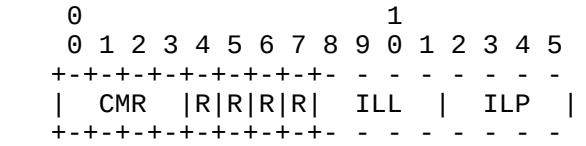
### 6.3.1. Payload Structure

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:




### 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.




CMR (4 bits): This indicates a codec mode request sent to the speech encoder at the site of the receiver of this payload. CMR value 15 indicates that no mode request is present, and other unused values are reserved for future use.

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 | VMR-WB mode 2 with maximum half-rate encoding |
| 7-14 | (reserved) |
| 15 | No Preference (no mode request is present) |

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

R: This 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., until 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 a payload is received with a CMR value that 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.

There are two factors that affect the VMR-WB mode selection: (i) the performance of any CDMA link connected via a gateway (e.g., in a GW to IP terminal scenario), and (ii) the congestion state of an IP network. The CDMA link performance is signaled via the CMR field, which is not used by IP-only end-points. The IP network state is monitored using, for example, RTCP. A sender needs to select the operating mode to satisfy both these constraints (see Section 7).

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, interleaving MUST be performed on a frame-block basis, as opposed 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.

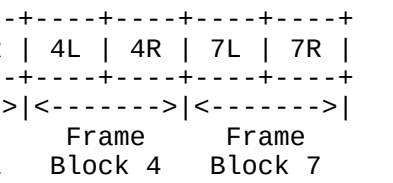
When performing interleaving, the sender MUST use a proper number of frame-blocks per payload (N) and ILL so that the resulting size of an interleave group is less than or equal to I, i.e., N*(L+1)<=I.

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.

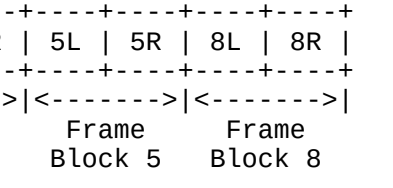
Packet #1
---

ILL=2, ILP=0:



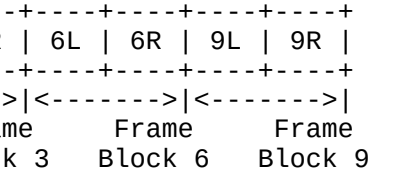

Packet #2
---

ILL=2, ILP=1:




Packet #3
---

ILL=2, ILP=2:




### 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.

```
+---------------------+
| list of ToC entries |
+---------------------+
```

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

```
 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+
|F|  FT   |Q|P|P|
+-+-+-+-+-+-+-+-+
```

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 Table 3.

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]).

If a payload with an invalid FT value is received, the payload MUST be discarded. Note that for ToC entries with FT=14 or 15, there will be no corresponding speech frame in the payload.

Depending on the application and the mode of operation of VMR-WB, any combination of the permissible frame types (FT) shown in Table 3 MAY be used.

Q (1 bit): Frame quality indicator. If set to 0, indicates that 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. The VMR-WB decoder may ignore the Q bit.

P bits: Padding bits MUST be set to zero and MUST be ignored by a receiver.

| 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) | 40 |
| 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 transport

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

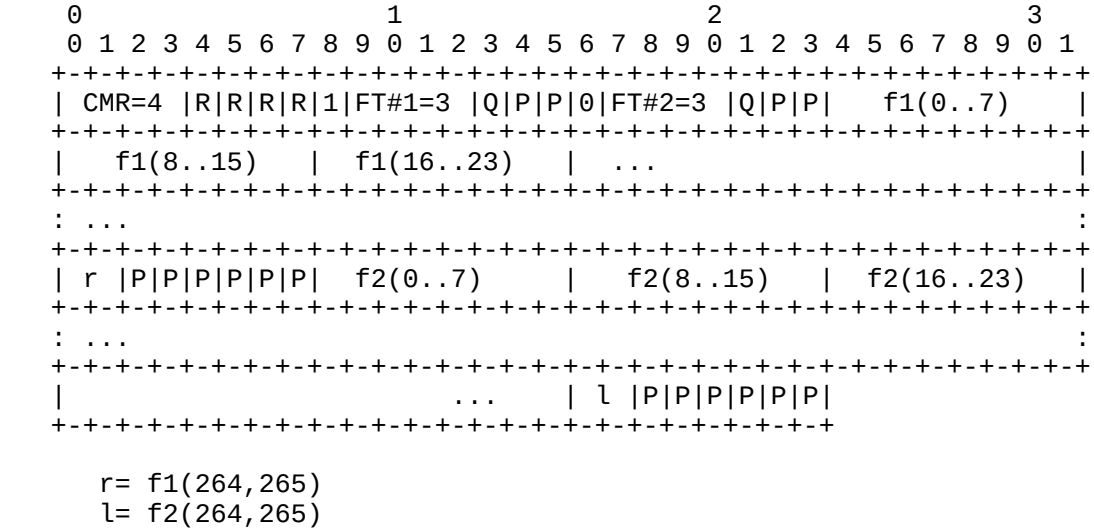
The order and numbering notation of the speech data bits are as specified in the VMR-WB standard specification [1].

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 the purpose of packetization, all 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 that the encoder at the receiver's side use VMR-WB mode 1. No interleaving is used. Note that in the example below the last octet in both speech frames is padded with zeros to make them octet aligned.




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

### 6.4. Implementation Considerations

An application implementing this payload format MUST understand all the payload parameters. Any mapping of the parameters to a signaling protocol MUST support all parameters. Therefore, an implementation of this payload format in an application using SDP is required to understand all the payload parameters in their SDP-mapped form. This requirement ensures that an implementation always can decide whether it is capable of communicating.

To enable efficient interoperable interconnection with AMR-WB and to ensure that a VMR-WB terminal appropriately declares itself as a AMR-WB-capable terminal (see Section 9.3), it is also RECOMMENDED that a VMR-WB RTP payload implementation understand relevant AMR-WB signaling.

To further ensure interoperability between various implementations of VMR-WB, implementations SHALL support both header-free and octet-aligned payload formats. Support of interleaving is optional.

#### 6.4.1. 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, does 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., the unavailability of a packet when it is needed for decoding at the receiver) should 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 available frame (sequential order). In other words, the unavailability of a packet in an interleave group at a certain time should not invalidate the other packets within that interleave group that may arrive later.

## 7. 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 codec may provide an advantage over other payload formats for controlling congestion since the bandwidth demand can be adjusted by selecting a different operating mode.

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.

Congestion control for RTP SHALL be used in accordance with RFC 3550 [3] and any applicable RTP profile, for example, RFC 3551 [6]. This means that congestion control is required for any transmission over unmanaged best-effort networks.

Congestion on the IP network is managed by the IP sender. 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 selecting an appropriate operating mode, but also includes adjusting the level of redundancy or number of frames per packet.

## 8. Security Considerations

RTP packets using the payload format defined in this specification are subject to the general security considerations discussed in RTP [3] and any applicable profile such as AVP [9] or SAVP [10].

As this format transports encoded audio, the main security issues include confidentiality, integrity protection, and data origin authentication of the audio itself. The payload format itself does not have any built-in security mechanisms. Any suitable external mechanisms, such as SRTP [10], MAY be used.

This payload format and the VMR-WB decoder do not exhibit any significant non-uniformity in the receiver-side computational complexity for packet processing; thus, they are unlikely to pose a denial-of-service threat due to the receipt of pathological data.

### 8.1. Confidentiality

In order to ensure confidentiality of the encoded audio, all audio data bits MUST be encrypted. There is less need to encrypt the payload header or the table of contents since they only carry information about the frame type. This information could also be useful to a third party, for example, for quality monitoring.

The use of interleaving in conjunction with encryption can have a negative impact on the confidentiality for a short period of time. Consider the following packets (in brackets) containing frame numbers as indicated: {10, 14, 18}, {13, 17, 21}, {16, 20, 24} (a typical continuous diagonal interleaving pattern). The originator wishes to deny some participants the ability to hear material starting at time 16. Simply changing the key on the packet with the timestamp at or after 16, and denying the new key to those participants, does not achieve this; frames 17, 18, and 21 have been supplied in prior packets under the prior key, and error concealment may make the audio intelligible at least as far as frame 18 or 19, and possibly further.

### 8.2. Authentication and Integrity

To authenticate the sender of the speech, an external mechanism MUST be used. It is RECOMMENDED that such a mechanism protects both the complete RTP header and the payload (speech and data bits).

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

## 9. Payload Format Parameters

This section defines the parameters that may be used to select optional features in the VMR-WB RTP payload formats.

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. In control protocols that do not use MIME or SDP, the media type parameters must be mapped to the appropriate format used with that control protocol.

### 9.1. VMR-WB RTP Payload 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 in VoIP applications. This MIME registration only covers real-time transfers via RTP.

Note, the receiver MUST ignore any unspecified parameter and use the default values instead. Also note that if no input parameters are defined, the default values will be used.

Media Type name: audio

Media subtype name: VMR-WB

Required parameters: none

Furthermore, if the interleaving parameter is present, the parameter "octet-align=1" MUST also be present.

OPTIONAL parameters:

| | |
|---|---|
| mode-set: | Requested VMR-WB operating mode set. Restricts the active operating modes to a subset of all modes. Possible values are a comma-separated list of integer values. Currently, this list includes modes 0, 1, 2, and 3 [1], but MAY be extended in the future. If such mode-set is specified during session initiation, the encoder MUST NOT use modes outside of the subset. If not present, all operating modes in the set 0 to 3 are allowed for the session. |
| channels: | The number of audio channels. The possible values and their respective channel order is specified in Section 4.1 in [6]. If omitted, it has the default value of 1. |
| octet-align: | RTP 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: | See RFC 3267 [4] |

interleaving: Indicates that frame-block level interleaving SHALL be used for the session. 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.

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 only defined for transfer of VMR-WB-encoded data via RTP (RFC 3550) using the payload formats specified in Section 6 of RFC 4348.

Security considerations:

See Section 8 of RFC 4348.

Public specification:

The VMR-WB speech codec is specified in 3GPP2 specifications C.S0052-0 version 1.0. Transfer methods are specified in RFC 4348.

Additional information:

Person & email address to contact for further information:

Sassan Ahmadi, Ph.D. sassan.ahmadi@ieee.org

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:

IETF Audio/Video Transport working group delegated from the IESG

## 9.2. 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 media type ("audio") goes in SDP "m=" as the media name.
- The media 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.
- The parameter "channels" (number of channels) MUST be either 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 [6]. The parameter "channels", if present, is specified subsequent to the MIME subtype and RTP clock rate as an encoding parameter in the "a=rtpmap" attribute.
- 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 examples of SDP session descriptions utilizing VMR-WB encodings follow.

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 octet-align=1
```

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 octet-align=1; interleaving=30
a=maxptime:100
```

### 9.3. Offer-Answer Model Considerations

To achieve good interoperability for the VMR-WB RTP payload in an Offer-Answer negotiation usage in SDP [13], the following considerations are made:

- The rate, channel, and payload configuration parameters (octet-align and interleaving) SHALL be used symmetrically, i.e., offer and answer must use the same values. The maximum size of the interleaving buffer is, however, declarative, and each agent specifies the value it supports to receive for recvonly and sendrecv streams. For sendonly streams, the value indicates what the agent desires to use.

- To maintain interoperability among all implementations of VMR-WB that may or may not support all the codec’s modes of operation, the operational modes that are supported by an implementation MAY be identified at session initiation. The mode-set parameter is declarative, and only operating modes that have been indicated to be supported by both ends SHALL be used. If the answerer is not supporting any of the operating modes provided in the offer, the complete payload type declaration SHOULD be rejected by removing it from the answer.

- The remaining parameters are all declarative; i.e., for sendonly streams they provide parameters that the agent desires to use, while for recvonly and sendrecv streams they declare the parameters that it accepts to receive. The dtx parameter is used to indicate DTX support and capability, while the media sender is only RECOMMENDED to send using the DTX in these cases. If DTX is not supported by the media sender, it will send media without DTX; this will not affect interoperability only the resource consumption.

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

- 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 of octet-align mode of operation.

Example:

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

An example of offer-answer exchange for the VoIP scenario described in Section 5.3 is as follows:

CDMA2000 terminal -> WCDMA terminal Offer:

```
m=audio 49120 RTP/AVP 98 97
a=rtpmap:98 VMR-WB/16000
a=fmtp:98 octet-align=1
a=rtpmap:97 AMR-WB/16000
a=fmtp:97 mode-set=0,1,2; octet-align=1
```

WCDMA terminal -> CDMA2000 terminal Answer:

```
m=audio 49120 RTP/AVP 97
a=rtpmap:97 AMR-WB/16000
a=fmtp:97 mode-set=0,1,2; octet-align=1;
```

For declarative use of SDP such as in SAP [14] and RTSP [15], all parameters are declarative and provide the parameters that SHALL be used when receiving and/or sending the configured stream.

## 10. IANA Considerations

The IANA has registered one new MIME subtype (audio/VMR-WB); see Section 9.

## 11. 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.

## 12. References

### 12.1. Normative References

[1] 3GPP2 C.S0052-0 v1.0 "Source-Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB) Service Option 62 for Spread Spectrum Systems", 3GPP2 Technical Specification, July 2004.

[2] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

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

[4] Sjoberg, J., Westerlund, M., Lakaniemi, A., and Q. Xie, "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", RFC 3267, June 2002.

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

[6] Schulzrinne, H. and S. Casner, "RTP Profile for Audio and Video Conferences with Minimal Control", STD 65, RFC 3551, July 2003.

### 12.2. Informative References

[7] 3GPP2 C.S0050-A v1.0 "3GPP2 File Formats for Multimedia Services", 3GPP2 Technical Specification, September 2005.

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

[9] Baugher, M., McGrew, D., Naslund, M., Carrara, E., and K. Norrman, "The Secure Real-time Transport Protocol (SRTP)", RFC 3711, March 2004.

[10] Perkins, C., Kouvelas, I., Hodson, O., Hardman, V., Handley, M., Bolot, J., Vega-Garcia, A., and S. Fosse-Parisis, "RTP Payload for Redundant Audio Data", RFC 2198, September 1997.

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

[13] Rosenberg, J. and H. Schulzrinne, "An Offer/Answer Model with Session Description Protocol (SDP)", RFC 3264, June 2002.

[14] Handley, M., Perkins, C., and E. Whelan, "Session Announcement Protocol", RFC 2974, October 2000.

[15] Schulzrinne, H., Rao, A., and R. Lanphier, "Real Time Streaming Protocol (RTSP)", RFC 2326, April 1998.

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

## Author's Address

Dr. Sassan Ahmadi
EMail: sassan.ahmadi@ieee.org

## Full Copyright Statement



## Intellectual Property



## Acknowledgement

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).