                       RTP Payload Format for
European Telecommunications Standards Institute (ETSI) European Standard
        ES 201 108 Distributed Speech Recognition Encoding

Status of this Memo

Copyright Notice

Abstract

   This document specifies an RTP payload format for encapsulating
   European Telecommunications Standards Institute (ETSI) European
   Standard (ES) 201 108 front-end signal processing feature streams for
   distributed speech recognition (DSR) systems.

Table of Contents

1. 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].

   The following acronyms are used in this document:

   DSR - Distributed Speech Recognition

   ETSI - the European Telecommunications Standards Institute

   FP   - Frame Pair

   DTX - Discontinuous Transmission

2. Introduction

   Motivated by technology advances in the field of speech recognition,
   voice interfaces to services (such as airline information systems,
   unified messaging) are becoming more prevalent. In parallel, the
   popularity of mobile devices has also increased dramatically.

However, the voice codecs typically employed in mobile devices were
designed to optimize audible voice quality and not speech recognition
accuracy, and using these codecs with speech recognizers can result
in poor recognition performance. For systems that can be accessed
from heterogeneous networks using multiple speech codecs, recognition
system designers are further challenged to accommodate the
characteristics of these differences in a robust manner. Channel
errors and lost data packets in these networks result in further
degradation of the speech signal.

In traditional systems as described above, the entire speech
recognizer lies on the server. It is forced to use incoming speech
in whatever condition it arrives after the network decodes the
vocoded speech. To address this problem, we use a distributed speech
recognition (DSR) architecture. In such a system, the remote device
acts as a thin client, also known as the front-end, in communication
with a speech recognition server, also called a speech engine. The
remote device processes the speech, compresses the data, and adds
error protection to the bitstream in a manner optimal for speech
recognition. The speech engine then uses this representation
directly, minimizing the signal processing necessary and benefiting
from enhanced error concealment.

To achieve interoperability with different client devices and speech
engines, a common format is needed. Within the "Aurora" DSR working
group of the European Telecommunications Standards Institute (ETSI),
a payload has been defined and was published as a standard [ES201108]
in February 2000.

For voice dialogues between a caller and a voice service, low latency
is a high priority along with accurate speech recognition. While
jitter in the speech recognizer input is not particularly important,
many issues related to speech interaction over an IP-based connection
are still relevant. Therefore, it is desirable to use the DSR
payload in an RTP-based session.

2.1 ETSI ES 201 108 DSR Front-end Codec

The ETSI Standard ES 201 108 for DSR [ES201108] defines a signal
processing front-end and compression scheme for speech input to a
speech recognition system. Some relevant characteristics of this
ETSI DSR front-end codec are summarized below.

The coding algorithm, a standard mel-cepstral technique common to
many speech recognition systems, supports three raw sampling rates: 8
kHz, 11 kHz, and 16 kHz. The mel-cepstral calculation is a frame-
based scheme that produces an output vector every 10 ms.

After calculation of the mel-cepstral representation, the
representation is first quantized via split-vector quantization to
reduce the data rate of the encoded stream. Then, the quantized
vectors from two consecutive frames are put into an FP, as described
in more detail in Section 4.1.

2.2 Typical Scenarios for Using DSR Payload Format

The diagrams in Figure 1 show some typical use scenarios of the ES
201 108 DSR RTP payload format.
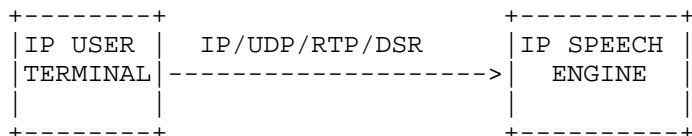
```
+--------+                        +----------+
|IP USER |   IP/UDP/RTP/DSR       |IP SPEECH |
|TERMINAL|-------------------->|   ENGINE  |
|        |                        |          |
+--------+                        +----------+

   a) IP user terminal to IP speech engine


+--------+ DSR over        +-------+                  +----------+
| Non-IP |  Circuit link   |       |  IP/UDP/RTP/DSR  |IP SPEECH |
|  USER  |:::::::::::::::>|GATEWAY|--------------->|   ENGINE  |
|TERMINAL|  ETSI payload   |       |                  |          |
+--------+  format         +-------+                  +----------+

   b) non-IP user terminal to IP speech engine via a gateway


+--------+                  +-------+  DSR over        +----------+
|IP USER |   IP/UDP/RTP/DSR |       |  circuit link    |  Non-IP  |
|TERMINAL|---------------->|GATEWAY|:::::::::::::::>|  SPEECH  |
|        |                  |       |  ETSI payload    |  ENGINE  |
+--------+                  +-------+  format          +----------+

   c) IP user terminal to non-IP speech engine via a gateway
```

Figure 1: Typical Scenarios for Using DSR Payload Format.

For the different scenarios in Figure 1, the speech recognizer always
resides in the speech engine. A DSR front-end encoder inside the
User Terminal performs front-end speech processing and sends the
resultant data to the speech engine in the form of "frame pairs"
(FPs). Each FP contains two sets of encoded speech vectors
representing 20ms of original speech.

3. ES 201 108 DSR RTP Payload Format

   An ES 201 108 DSR RTP payload datagram consists of a standard RTP
   header [RFC3550] followed by a DSR payload. The DSR payload itself
   is formed by concatenating a series of ES 201 108 DSR FPs (defined in
   Section 4).

   FPs are always packed bit-contiguously into the payload octets
   beginning with the most significant bit. For ES 201 108 front-end,
   the size of each FP is 96 bits or 12 octets (see Sections 4.1 and
   4.2). This ensures that a DSR payload will always end on an octet
   boundary.

   The following example shows a DSR RTP datagram carrying a DSR payload
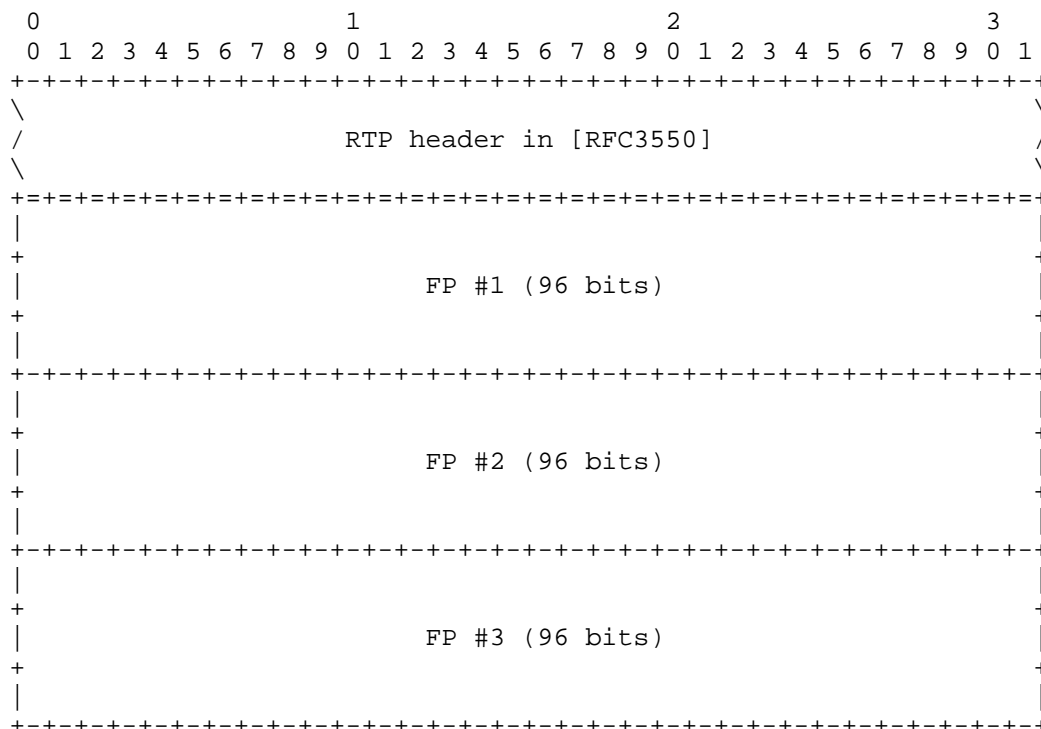   containing three 96-bit-long FPs (bit 0 is the MSB):

```
 0                   1                   2                   3
 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 in [RFC3550]                     /
\                                                               \
+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
|                                                               |
+                                                               +
|                       FP #1 (96 bits)                         |
+                                                               +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
+                                                               +
|                       FP #2 (96 bits)                         |
+                                                               +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
+                                                               +
|                       FP #3 (96 bits)                         |
+                                                               +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

         Figure 2. An example of an ES 201 108 DSR RTP payload.

3.1 Consideration on Number of FPs in Each RTP Packet

   The number of FPs per payload packet should be determined by the
   latency and bandwidth requirements of the DSR application using this
   payload format. In particular, using a smaller number of FPs per
   payload packet in a session will result in lowered bandwidth
   efficiency due to the RTP/UDP/IP header overhead, while using a
   larger number of FPs per packet will cause longer end-to-end delay
   and hence increased recognition latency. Furthermore, carrying a
   larger number of FPs per packet will increase the possibility of
   catastrophic packet loss; the loss of a large number of consecutive
   FPs is a situation most speech recognizers have difficulty dealing
   with.

   It is therefore RECOMMENDED that the number of FPs per DSR payload
   packet be minimized, subject to meeting the application's
   requirements on network bandwidth efficiency. RTP header compression
   techniques, such as those defined in [RFC2508] and [RFC3095], should
   be considered to improve network bandwidth efficiency.

3.2 Support for Discontinuous Transmission

   The DSR RTP payloads may be used to support discontinuous
   transmission (DTX) of speech, which allows that DSR FPs are sent only
   when speech has been detected at the terminal equipment.

   In DTX a set of DSR frames coding an unbroken speech segment
   transmitted from the terminal to the server is called a transmission
   segment. A DSR frame inside such a transmission segment can be
   either a speech frame or a non-speech frame, depending on the nature
   of the section of the speech signal it represents.

   The end of a transmission segment is determined at the sending end
   equipment when the number of consecutive non-speech frames exceeds a
   pre-set threshold, called the hangover time. A typical value used
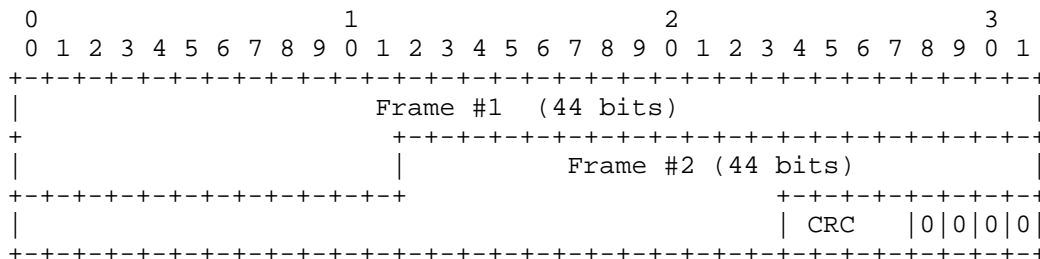   for the hangover time is 1.5 seconds.

   After all FPs in a transmission segment are sent, the front-end
   SHOULD indicate the end of the current transmission segment by
   sending one or more Null FPs (defined in Section 4.2).

4. Frame Pair Formats

4.1 Format of Speech and Non-speech FPs

   The following mel-cepstral frame MUST be used, as defined in
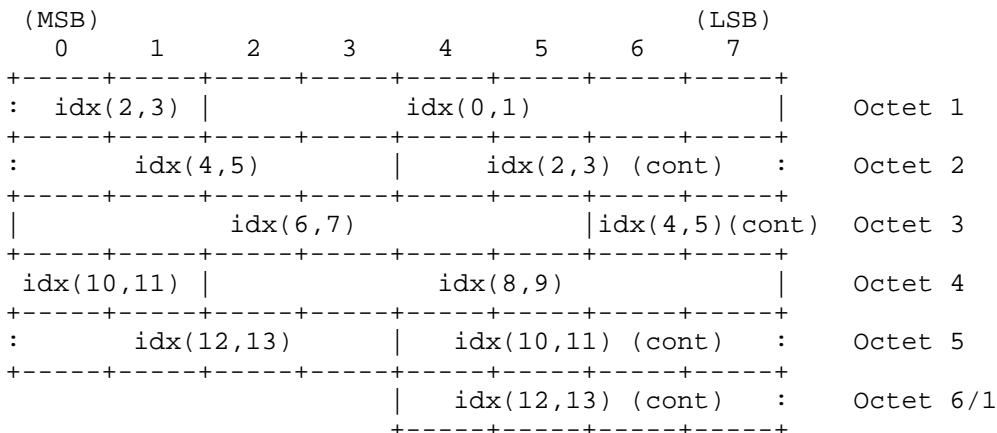   [ES201108]:

   As defined in [ES201108], pairs of the quantized 10ms mel-cepstral
   frames MUST be grouped together and protected with a 4-bit CRC,
   forming a 92-bit long FP:

```
 0                   1                   2                   3
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Frame #1  (44 bits)                       |
+               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               |             Frame #2 (44 bits)          |
+-+-+-+-+-+-+-+-+-+-+-+-+                   +-+-+-+-+-+-+-+-+
|                                           | CRC   |0|0|0|0|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

   The length of each frame is 44 bits representing 10ms of voice. The
   following mel-cepstral frame formats MUST be used when forming an FP:

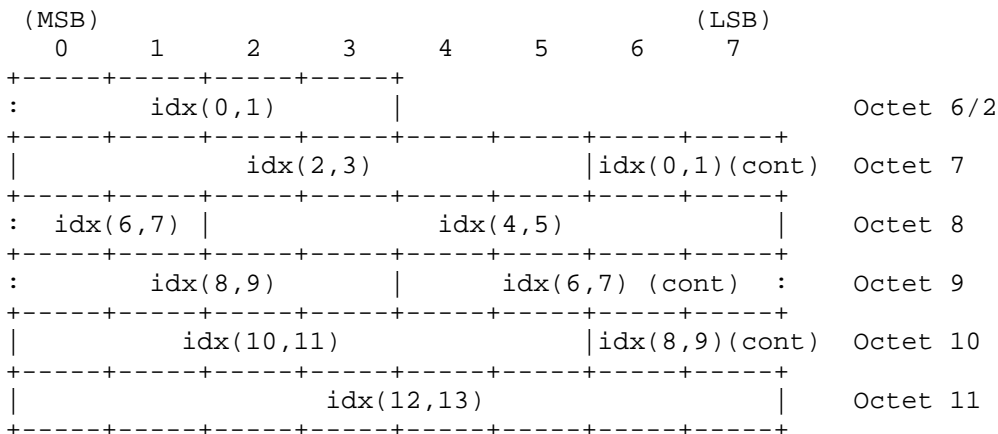   Frame #1 in FP:
   ===============

```
    (MSB)                                     (LSB)
      0     1     2     3     4     5     6     7
   +-----+-----+-----+-----+-----+-----+-----+-----+
   : idx(2,3) |            idx(0,1)              |   Octet 1
   +-----+-----+-----+-----+-----+-----+-----+-----+
   :       idx(4,5)        |     idx(2,3) (cont)    :   Octet 2
   +-----+-----+-----+-----+-----+-----+-----+-----+
   |             idx(6,7)              |idx(4,5)(cont)  Octet 3
   +-----+-----+-----+-----+-----+-----+-----+-----+
    idx(10,11) |            idx(8,9)              |   Octet 4
   +-----+-----+-----+-----+-----+-----+-----+-----+
   :       idx(12,13)      |   idx(10,11) (cont)   :   Octet 5
   +-----+-----+-----+-----+-----+-----+-----+-----+
                           |   idx(12,13) (cont)   :   Octet 6/1
                           +-----+-----+-----+-----+
```

```
   Frame #2 in FP:
   ===============
      (MSB)                                          (LSB)
        0     1     2     3     4     5     6     7
      +-----+-----+-----+-----+
      :          idx(0,1)      |                           Octet 6/2
      +-----+-----+-----+-----+-----+-----+-----+-----+
      |          idx(2,3)           |idx(0,1)(cont)  Octet 7
      +-----+-----+-----+-----+-----+-----+-----+-----+
      :  idx(6,7) |         idx(4,5)              |     Octet 8
      +-----+-----+-----+-----+-----+-----+-----+-----+
      :          idx(8,9)      |    idx(6,7) (cont)  :   Octet 9
      +-----+-----+-----+-----+-----+-----+-----+-----+
      |          idx(10,11)         |idx(8,9)(cont)  Octet 10
      +-----+-----+-----+-----+-----+-----+-----+-----+
      |               idx(12,13)                   |   Octet 11
      +-----+-----+-----+-----+-----+-----+-----+-----+
```

Therefore, each FP represents 20ms of original speech. Note, as
shown above, each FP MUST be padded with 4 zeros to the end in order
to make it aligned to the 32-bit word boundary. This makes the size
of an FP 96 bits, or 12 octets. Note, this padding is separate from
padding indicated by the P bit in the RTP header.

The 4-bit CRC MUST be calculated using the formula defined in 6.2.4
in [ES201108]. The definition of the indices and the determination of
their value are also described in [ES201108].

4.2 Format of Null FP

A Null FP for the ES 201 108 front-end codec is defined by setting
the content of the first and second frame in the FP to null (i.e.,
filling the first 88 bits of the FP with 0's). The 4-bit CRC MUST be
calculated the same way as described in 6.2.4 in [ES201108], and 4
zeros MUST be padded to the end of the Null FP to make it 32-bit word
aligned.

4.3 RTP header usage

The format of the RTP header is specified in [RFC3550]. 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 FP in the packet. The timestamp clock
frequency is the same as the sampling frequency, so the timestamp
unit is in samples.

As defined by ES 201 108 front-end codec, the duration of one FP is
20 ms, corresponding to 160, 220, or 320 encoded samples with
sampling rate of 8, 11, or 16 kHz being used at the front-end,
respectively. Thus, the timestamp is increased by 160, 220, or 320
for each consecutive FP, respectively.

The DSR payload for ES 201 108 front-end codes is always an integral
number of octets. If additional padding is required for some other
purpose, then the P bit in the RTP in the header may be set and
padding appended as specified in [RFC3550].

The RTP header marker bit (M) should be set following the general
rules defined in [RFC3551].

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.

5. IANA Considerations

One new MIME subtype registration is required for this payload type,
as defined below.

This section also defines the optional parameters that may be used to
describe a DSR session. The parameters are defined here as part of
the MIME subtype registration. A mapping of the parameters into the
Session Description Protocol (SDP) [RFC2327] is also provided in 5.1
for those applications that use SDP.

Media Type name: audio

Media subtype name: dsr-es201108

Required parameters: none

Optional parameters:

rate: Indicates the sample rate of the speech. Valid values include:
   8000, 11000, and 16000. If this parameter is not present, 8000
   sample rate is assumed.

maxptime: The maximum amount of media which can be encapsulated in
   each packet, expressed as time in milliseconds. The time shall be
   calculated as the sum of the time the media present in the packet
   represents. The time SHOULD be a multiple of the frame pair size
   (i.e., one FP <-> 20ms).

If this parameter is not present, maxptime is assumed to be 80ms.

Note, since the performance of most speech recognizers are extremely sensitive to consecutive FP losses, if the user of the payload format expects a high packet loss ratio for the session, it MAY consider to explicitly choose a maxptime value for the session that is shorter than the default value.

ptime: see RFC2327 [RFC2327].

Encoding considerations : This type is defined for transfer via RTP [RFC3550] as described in Sections 3 and 4 of RFC 3557.

Security considerations : See Section 6 of RFC 3557.

Person & email address to contact for further information:
   Qiaobing.Xie@motorola.com

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

Author/Change controller:
   Qiaobing.Xie@motorola.com
   IETF Audio/Video transport working group

5.1 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) [RFC2327], which is commonly used to describe RTP sessions. When SDP is used to specify sessions employing ES 201 018 DSR codec, the mapping is as follows:

o The MIME type ("audio") goes in SDP "m=" as the media name.

o The MIME subtype ("dsr-es201108") goes in SDP "a=rtpmap" as the encoding name.

o The optional parameter "rate" also goes in "a=rtpmap" as clock rate.

o The optional parameters "ptime" and "maxptime" go in the SDP "a=ptime" and "a=maxptime" attributes, respectively.

   Example of usage of ES 201 108 DSR:

      m=audio 49120 RTP/AVP 101
      a=rtpmap:101 dsr-es201108/8000
      a=maxptime:40

6. Security Considerations

   Implementations using the payload defined in this specification are
   subject to the security considerations discussed in the RTP
   specification [RFC3550] and the RTP profile [RFC3551]. This payload
   does not specify any different security services.

7. Contributors

   The following individuals contributed to the design of this payload
   format and the writing of this document: Q. Xie (Motorola), D. Pearce
   (Motorola), S. Balasuriya (Motorola), Y. Kim (VerbalTek), S. H. Maes
   (IBM), and, Hari Garudadri (Qualcomm).

8. Acknowledgments

   The design presented here benefits greatly from an earlier work on
   DSR RTP payload design by Jeff Meunier and Priscilla Walther. The
   authors also wish to thank Brian Eberman, John Lazzaro, Magnus
   Westerlund, Rainu Pierce, Priscilla Walther, and others for their
   review and valuable comments on this document.

9. References

9.1 Normative References

   [ES201108]   European Telecommunications Standards Institute (ETSI)
                 Standard ES 201 108, "Speech Processing, Transmission
                 and Quality Aspects (STQ); Distributed Speech
                 Recognition; Front-end Feature Extraction Algorithm;
                 Compression Algorithms," Ver. 1.1.2, April 11, 2000.

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

   [RFC2026]   Bradner, S., "The Internet Standards Process -- Revision
                 3", BCP 9, RFC 2026, October 1996.

   [RFC2119]   Bradner, S., "Key words for use in RFCs to Indicate
                 Requirement Levels", BCP 14, RFC 2119, March 1997.

   [RFC2327]   Handley, M. and V. Jacobson, "SDP: Session Description
                 Protocol", RFC 2327, April 1998.

9.2 Informative References

   [RFC3551]   Schulzrinne, H. and S. Casner, "RTP Profile for Audio
                 and Video Conferences with Minimal Control", RFC 3551,
                 July 2003.

   [RFC2508]   Casner, S. and V. Jacobson, "Compressing IP/UDP/RTP
                 Headers for Low-Speed Serial Links", RFC 2508, February
                 1999.

   [RFC3095]   Bormann, C., Burmeister, C., Degermark, M., Fukushima,
                 H., Hannu, H., Jonsson, L-E, Hakenberg, R., Koren, T.,
                 Le, K., Liu, Z., Martensson, A., Miyazaki, A., Svanbro,
                 K., Wiebke, T., Yoshimura, T. and H. Zheng, "RObust
                 Header Compression (ROHC): Framework and four profiles",
                 RFC 3095, July 2001.

10. IPR Notices

   The IETF takes no position regarding the validity or scope of any
   intellectual property or other rights that might be claimed to
   pertain to the implementation or use of the technology described in
   this document or the extent to which any license under such rights
   might or might not be available; neither does it represent that it
   has made any effort to identify any such rights. Information on the
   IETF's procedures with respect to rights in standards-track and
   standards-related documentation can be found in BCP-11. Copies of
   claims of rights made be made available, or the result of an attempt
   made to obtain a general license or permission for the use of such
   proprietary rights by implementors or users of this specification can
   be obtained from the IETF Secretariat.

   The IETF invites any interested party to bring to its attention any
   copyrights, patents or patent applications, or other proprietary
   rights which may cover technology that may be required to practice
   this standard. Please address the information to the IETF Executive
   Director.

11. Authors' Addresses

David Pearce
Motorola Labs
UK Research Laboratory
Jays Close
Viables Industrial Estate
Basingstoke, HANTS, RG22 4PD

Phone: +44 (0)1256 484 436
EMail: bdp003@motorola.com

Senaka Balasuriya
Motorola, Inc.
600 U.S Highway 45
Libertyville, IL 60048, USA

Phone: +1-847-523-0440
EMail: Senaka.Balasuriya@motorola.com

Yoon Kim
VerbalTek, Inc.
2921 Copper Rd.
Santa Clara, CA 95051

Phone: +1-408-768-4974
EMail: yoonie@verbaltek.com

Stephane H. Maes, PhD,
Oracle
500 Oracle Parkway, M/S 4op634
Redwood City, CA 94065 USA

Phone: +1-650-607-6296.
EMail: stephane.maes@oracle.com

Hari Garudadri
Qualcomm Inc.
5775, Morehouse Dr.
San Diego, CA 92121-1714, USA

Phone: +1-858-651-6383
EMail: hgarudad@qualcomm.com

12. Editor's Address

     Qiaobing Xie
     Motorola, Inc.
     1501 W. Shure Drive, 2-F9
     Arlington Heights, IL 60004, USA

     Phone: +1-847-632-3028
     EMail: Qiaobing.Xie@motorola.com

13. Full Copyright Statement

Acknowledgement