University of Glasgow | School of Computing Science

Guidelines for the use of Variable Bit Rate Audio with Secure RTP

draft-ietf-avtcore-srtp-vbr-audio-03

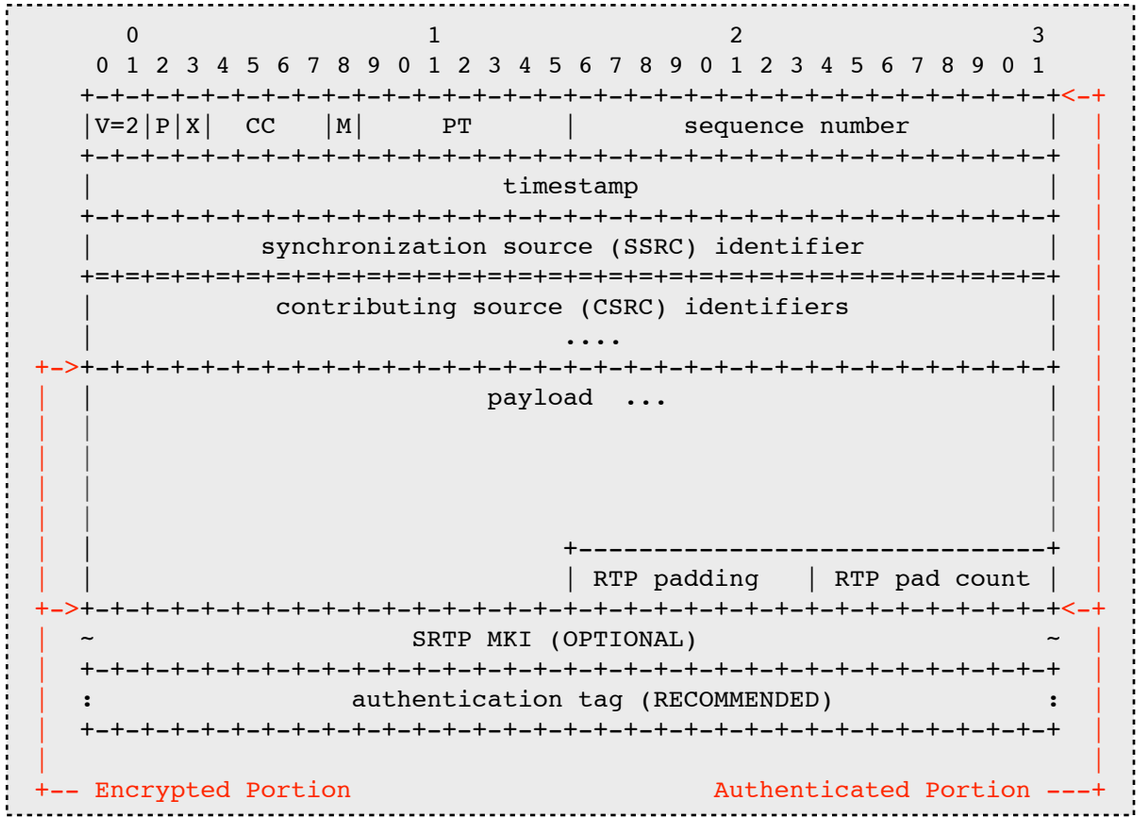
Colin Perkins – University of Glasgow
Jean-Marc Valin – Octasic, Inc.

Talk Outline

- Introduction to SRTP and VBR audio
- Problem – information leakage
- Recommendations

Secure RTP (SRTP)

- Real-time Transport Protocol (RTP) [RFC3550] is a standard framework for multimedia transport
  - Augmented by numerous *payload formats* that define how to map particular audio and video codecs onto RTP framing
- Secure RTP is an extension that gives confidentiality, message authentication, and replay protection [RFC3711]
  - AES counter mode encryption with HMAC-SHA1 authentication and integrity protection by default; other algorithms have been defined
  - Encrypts the payload, leaving the headers in the clear; adds authentication tag as a trailer
  - RTP has a padding mechanism, but if not used, SRTP packets using AES counter mode reflect the size of the payload data
- Keying mechanisms defined separately



Variable Bit Rate Audio

- Variable bite rate (VBR) coding
  - Some audio codecs produce fixed size output (e.g., GSM compresses 20ms of speech into 33 octets)
  - Others are variable bit rate, where the size of the output depends on the characteristics of the speech being encoded
  - These VBR codecs are desirable, because they tend to generate smaller output on average → save bandwidth
- Voice activity detection (VAD) is also used, where the codec suppresses the silence periods between words, phrases, etc.
  - Most codecs send “comfort noise” to fill the gap – a heavily compressed version of the background noise
  - Again, can save a significant fraction of bandwidth

The Problem

- The size of the RTP packets produced by VBR audio codecs, and the presence of gaps due to VAD, leaks some information about the speech
  - It has been shown that known phrases in an encrypted call using the Speex codec in VBR mode can be recognised with high accuracy in certain circumstances, without breaking the encryption (and it seems unlikely that the problem is specific to Speex)
  - Other work has shown that the language spoken in encrypted conversations can also be recognised
  - The known attacks are likely to increase – there is much ongoing work in this area – this draft gives guidelines for mitigation

Wright *et al.*, "Spot me if you can: Uncovering spoken phrases in encrypted VoIP conversation", Proc. IEEE Symposium on Security and Privacy 2008, May 2008 http://www.cs.jhu.edu/~cwright/oakland08.pdf

Recommendations (1)

- Guidelines for use of VBR audio with secure RTP
  - As a general guideline, VBR codecs should be considered safe in the context of encrypted unstructured calls
  - However, structured calls and applications that make use of pre-recorded messages, where the contents of such pre-recorded messages may be of any value to an eavesdropper, SHOULD NOT use VBR coding
    - Or should use RTP padding to hide speech packet lengths, padding to simulate a constant rate codec (the amount of padding needed will depend on the codec)
    - This will increase the bandwidth use of the speech call, compared to using VBR coding

- It is safe to use VBR coding to adapt to the characteristics of a network channel, e.g., for congestion control, provided this is done in a way that does not expose any information on the speech signal

Recommendations (2)

- Guidelines for use of VAD with secure RTP
  - Disabling VAD is secure, but has a significant impact on bandwidth usage
  - Instead, recommend that SRTP senders using VAD SHOULD insert an overhang period at the end of each talk spurt, delaying the start of the silence/comfort noise by a random interval
    - During the overhead period, SRTP audio packets must be generated that are indistinguishable from regular speech packets
  - The length of the overhang applied to each talk spurt must be randomly chosen in such a way that it is computationally infeasible for an attacker to reliably estimate the length of that talk spurt
    - The overhang period SHOULD have an exponentially-decreasing probability distribution to ensure a long tail, while being easy to compute.
    - RECOMMENDED to use an overhang with a “half life” of a few hundred milliseconds (to obscure the presence of inter-word pauses and the lengths of single words spoken in isolation, e.g., digits of a credit card number clearly enunciated for an automated system, but not so long as to significantly reduce the effectiveness of VAD for detecting listening pauses)
    - Still leaks some information, so SHOULD NOT be used in sensitive applications (e.g., IVR systems with known prerecorded messages that may be of interest to the attacker)

Discussion

- Draft is in IETF last call
  - draft-ietf-avtcore-srtp-vbr-audio-03
- Feedback on the recommendations is solicited