Internet Engineering Task Force (IETF)
Request for Comments: 6562
Category: Standards Track
ISSN: 2070-1721

C. Perkins
University of Glasgow
JM. Valin
Mozilla Corporation

March 2012

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

Abstract

This memo discusses potential security issues that arise when using variable bit rate (VBR) audio with the secure RTP profile. Guidelines to mitigate these issues are suggested.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6562.

Table of Contents

1. Introduction

The Secure RTP (SRTP) framework [RFC3711] is a widely used framework for securing RTP sessions [RFC3550]. SRTP provides the ability to encrypt the payload of an RTP packet, and optionally add an authentication tag, while leaving the RTP header and any header extension in the clear. A range of encryption transforms can be used with SRTP, but none of the predefined encryption transforms use any padding; the RTP and SRTP payload sizes match exactly.

When using SRTP with voice streams compressed using variable bit rate (VBR) codecs, the length of the compressed packets will depend on the characteristics of the speech signal. This variation in packet size will leak a small amount of information about the contents of the speech signal. This is potentially a security risk for some applications. For example, [spot-me] shows that known phrases in an encrypted call using the Speex codec in VBR mode can be recognized with high accuracy in certain circumstances, and [fon-iks] shows that approximate transcripts of encrypted VBR calls can be derived for some codecs without breaking the encryption. How significant these results are, and how they generalize to other codecs, is still an open question. This memo discusses ways in which such traffic analysis risks may be mitigated.

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 RFC 2119 [RFC2119].

2. Scenario-Dependent Risk

Whether the information leaks and attacks discussed in [spot-me], [fon-iks], and similar works are significant is highly dependent on the application and use scenario. In the worst case, using the rate information to recognize a prerecorded message knowing the set of all possible messages would lead to near-perfect accuracy. Even when the

audio is not prerecorded, there is a real possibility of being able to recognize contents from encrypted audio when the dialog is highly structured (e.g., when the eavesdropper knows that only a handful of possible sentences are possible), and thus contain only little information. Recognizing unconstrained conversational speech from the rate information alone is unreliable and computationally expensive at present, but does appear possible in some circumstances. These attacks are only likely to improve over time.

In practical SRTP scenarios, how significant the information leak is when compared to other SRTP-related information must be considered, such as the fact that the source and destination IP addresses are available.

3. Guidelines for Use of VBR Audio with SRTP

It is the responsibility of the application designer to determine the appropriate trade-off between security and bandwidth overhead. As a general rule, VBR codecs should be considered safe in the context of low-value encrypted unstructured calls. However, applications that make use of prerecorded messages where the contents of such prerecorded messages may be of any value to an eavesdropper (i.e., messages beyond standard greeting messages) SHOULD NOT use codecs in VBR mode. Interactive voice response (IVR) applications would be particularly vulnerable since an eavesdropper could easily use the rate information to recognize the prompts being played out. Applications conveying highly sensitive unstructured information SHOULD NOT use codecs in VBR mode.

It is safe to use variable rate coding to adapt the output of a voice codec to match characteristics of a network channel, provided this adaptation is done in a way that does not expose any information on the speech signal. For example, VBR audio can be used for congestion control purposes, where the variation is driven by the available network bandwidth, not by the input speech (i.e., the packet sizes and spacing are constant unless the network conditions change). VBR speech codecs can safely be used in this fashion with SRTP while avoiding leaking information on the contents of the speech signal that might be useful for traffic analysis.

4. Guidelines for Use of Voice Activity Detection with SRTP

Many speech codecs employ some form of voice activity detection (VAD) to either suppress output frames, or generate some form of lower-rate comfort noise frames, during periods when the speaker is not active. If VAD is used on an encrypted speech signal, then some information

about the characteristics of that speech signal can be determined by watching the patterns of voice activity. This information leakage is less than with VBR coding since there are only two rates possible.

The information leakage due to VAD in SRTP audio sessions can be much reduced if the sender adds an unpredictable "overhang" period to the end of active speech intervals, obscuring their actual length. An RTP sender using VAD with encrypted SRTP audio SHOULD insert such an overhang period at the end of each talkspurt, delaying the start of the silence/comfort noise by a random interval. The length of the overhang applied to each talkspurt must be randomly chosen in such a way that it is computationally infeasible for an attacker to reliably estimate the length of that talkspurt. This may be more important for short talkspurts, since it seems easier to distinguish between different single word responses based on the exact word length, than to glean meaning from the duration of a longer phrase. The audio data comprising the overhang period must be packetized and transmitted in RTP packets in a manner that is indistinguishable from the other data in the talkspurt.

The overhang period SHOULD have an exponentially decreasing probability distribution function. This ensures a long tail, while being easy to compute. It is RECOMMENDED to use an overhang with a "half life" of a few hundred milliseconds (this should be sufficient to obscure the presence of interword pauses and the lengths of single words spoken in isolation, for example, the 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). Despite the overhang (and no matter what the duration is), there is still a small amount of information leaked about the start time of the talkspurt due to the fact that we cannot apply an overhang to the start of a talkspurt without unacceptably affecting intelligibility. For that reason, VAD SHOULD NOT be used in encrypted IVR applications where the content of prerecorded messages may be of any value to an eavesdropper.

The application of a random overhang period to each talkspurt will reduce the effectiveness of VAD in SRTP sessions when compared to non-SRTP sessions. However, it is still expected that the use of VAD will provide significant bandwidth savings for many encrypted sessions.

5. Padding the Output of VBR Codecs

For scenarios where VBR is considered unsafe, a constant bit rate (CBR) codec SHOULD be negotiated and used instead, or the VBR codec SHOULD be operated in a CBR mode. However, if the codec does not support CBR, RTP padding SHOULD be used to reduce the information

leak to an insignificant level. Packets may be padded to a constant size or to a small range of sizes ([spot-me] achieves good results by padding to the next multiple of 16 octets, but the amount of padding needed to hide the variation in packet size will depend on the codec and the sophistication of the attacker) or may be padded to a size that varies with time. The most secure and RECOMMENDED option is to pad all packets throughout the call to the same size.

In the case where the size of the padded packets varies in time, the same concerns as for VAD apply. That is, the padding SHOULD NOT be reduced without waiting for a certain (random) time. The RECOMMENDED "hold time" is the same as the one for VAD.

Note that SRTP encrypts the count of the number of octets of padding added to a packet, but not the bit in the RTP header that indicates that the packet has been padded. For this reason, it is RECOMMENDED to add at least one octet of padding to all packets in a media stream, so an attacker cannot tell which packets needed padding.

6. Security Considerations

This entire memo is about security. The security considerations of [RFC3711] also apply.

7. Acknowledgements

ZRTP [RFC6189] contains similar recommendations; the purpose of this memo is to highlight these issues to a wider audience, since they are not specific to ZRTP. Thanks are due to Phil Zimmermann, Stefan Doehla, Mats Naslund, Gregory Maxwell, David McGrew, Mark Baugher, Koen Vos, Ingemar Johansson, and Stephen Farrell for their comments and feedback on this memo.

8. References

8.1. Normative References

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

[RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, July 2003.

[RFC3711] Baugher, M., McGrew, D., Naslund, M., Carrara, E., and K. Norrman, "The Secure Real-time Transport Protocol (SRTP)", RFC 3711, March 2004.

8.2. Informative References

[RFC6189] Zimmermann, P., Johnston, A., and J. Callas, "ZRTP: Media Path Key Agreement for Unicast Secure RTP", RFC 6189, April 2011.

[fon-iks] White, A., Matthews, A., Snow, K., and F. Monrose, "Phonotactic Reconstruction of Encrypted VoIP Conversations: Hookt on fon-iks", Proceedings of the IEEE Symposium on Security and Privacy 2011, May 2011.

[spot-me] Wright, C., Ballard, L., Coull, S., Monrose, F., and G. Masson, "Spot me if you can: Uncovering spoken phrases in encrypted VoIP conversation", Proceedings of the IEEE Symposium on Security and Privacy 2008, May 2008.

Authors' Addresses

Colin Perkins
University of Glasgow
School of Computing Science
Glasgow G12 8QQ
UK

EMail: csp@csperkins.org

Jean-Marc Valin
Mozilla Corporation
650 Castro Street
Mountain View, CA 94041
USA

Phone: +1 650 903-0800
EMail: jmvalin@jmvalin.ca