CLUE
Internet-Draft
Intended status: Standards Track
Expires: December 2, 2012

A. Romanow
R. Hansen
Cisco Systems
A. Pepperell
Silverflare
B. Baldino
Cisco Systems

May 31, 2012

The need for audio rendering tag mechanism in the CLUE Framework
draft-romanow-clue-audio-rendering-tag-00

Abstract

The purpose of this draft is for discussion in the CLUE working group.

It proposes adding an audio rendering tag to the CLUE framework [I-D.ietf-clue-framework], which makes it possible for the consumer to correctly render audio with respect to video in a multistream video conference. The solution proposed is in partial response to CLUE Task #10, Does Framework provide sufficient info for receiver?

Status of this Memo

Copyright Notice

Table of Contents

1. Motivation- the issue

A goal for CLUE audio is that listeners perceive the direction of a sound source to be the same as that of the visual image of the source; this is referred to as directional audio. In some situations the existing clue mechanisms are adequate. The consumer can use the spatial information to correctly place the audio when the provider advertisement includes spatial information (point of origin and capture area) giving a static relationship between both video and associated audio captures.

However, in some circumstances, for different reasons, the audio and/or video spatial information is not sent in the provider advertisement. For instance, the case of a three-screen system advertising three video captures and one switched audio capture, where the audio is switched from the loudest of three microphones. In this case, how will the consumer know how to associate the audio with the correct video so it can be played out in the correct location?

Here we suggest a simple mechanism -- audio rendering tagging.

When audio and video cannot be matched through provider advertisement spatial information, we would like the ability to play out audio on multiple speakers matching the position of the speaker in the original scene. Also, the audio may be assigned to a speaker in real-time. It may need to be mixe locally and played out on any speaker. For example, if the consumer wants to hear the top 3 speakers, regardless of where they are located remotely, if all 3 top speakers are coming from the left, then the 3 speakers need to be mixed, perhaps locally, and played out on the left.

Note: Several typical scenarios are described at the end of this note in section titled Use Case.

2. Terminology

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] and indicate requirement levels for compliant implementations.

3. Audio Rendering Tag Mechanism

We propose an audio tagging mechanism In order to cope with a changing mapping of the most significant audio and video participants

(i.e., normal MCU operations in the presence of more participants' media streams that can be rendered simultaneously) and to get audio played out correctly to multiple speakers. A consumer optionally tells the provider an audio tag value corresponding to each of its chosen video captures which enables received audio to be associated with the correct video stream, even when the set of audible participants changes. This information is included with the consumer request so there is no need for additional CLUE message exchanges (specifically, no additional provider capture advertisements or consumer requests).

The audio tags are defined in the consumer request as opposed to in a capture advertised by a producer. The reason for this is that it is valid for a consumer to request a capture multiple times (with different encodings, for example) and hence a method is required for differentiating between these streams.

When the consumer configures the provider, saying which captures it wants, it also optionally includes an audio tag with each capture request. For example, VC1, ATag1; VC2, ATag2. When the provider sends audio packets to the consumer, it includes the appropriate audio tag in an RTP header extension. For example, if the provider is sending audio packets that are associated with VC1, it tags the packets with ATag1. The consumer can then play out the audio in a position appropriate for video from VC1.

Suppose that several audio streams need to be played out through the same speaker - for example, the 3 audio streams (AC1, AC2, AC3) need to be played out at the speaker associated with VC1. The provider would send:

```
AC1  ATag1
AC2  ATag1
AC3  ATag1
```

AC1, AC2 and AC3 are all played out on the same speaker, the audio output associated with VC1. This takes care of the issue of dynamic audio output - assigning the right speaker to audio streams.

Figure 1 illustrates an example showing 3 screens, each with a main video and 3 PIPs. Below each screen is a list of the video captures, VCs with the associated Audio Tag.

```
          ----------------------3 Screens ---------------------
         |------------------+- -----------------+------------------Y
         |                  |                   |                  |
         |    VC1           |    VC2            |    VC3           |
         |                  |                   |                  |
         |                  |                   |                  |
         |                  |                   |                  |
         | ''''|'''''''''|  |  ''''|'''''|'''|  |  '''''|''''|''''||
         | |VC4|.VC5.|VC6|  |  |VC7|.VC8.|VC9|  |  |VC10|VC11|VC12||
         '------------------+-------------------+------------------
           VC1                  VC2                  VC3
           VC4  Audio Tag 1     VC7  Audio tag 2     VC10 Audio tag 3
           VC5                  VC8                  VC11
           VC6                  VC9                  VC12
```

Figure 1: Audio rendering tags for 3 screen example

The provider may choose not to include the extension header in an audio packet, signaling that there is no association between the current audio and current video (i.e., an audio-only participant). It may also include more than one audio tag in the extension header, signaling that this audio is associated with multiple current video participants, due perhaps to a capture being received multiple times at different resolutions, or two video captures that both include the current speaker.

This mechanism also allows multiple audio streams to be associated with a single video stream (i.e. for a composed video stream); this simply requires the appropriate audio packets to be tagged with the same id.

4. Use of the RTP header extension

We propose that audio tags are integer numbers between 0 and 255 optionally set by the consumer per requested capture. This allows up to 16 tags to be included in a one-byte RTP header extension [RFC 5285]. An example header extension for an audio packet with one tag follows. The audio tag extension is ID1. The example includes another header extension (ID0) to show how the proposal would interact with [I-D.lennox-clue-rtp-usage]:

```
   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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |       0xBE    |    0xDE       |           length=1            |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

```
   |  ID0   | L=0  |     data      |  ID1  |  L=0  |     Tag       |
   -+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

RTP ext headers for audio rendering tag and capture ID

The lack of the RTP header extension in a packet means that the audio packet is not associated with any of the requested video streams that included audio tags.

5. Use case note

- An endpoint can receive multiple video and audio streams and render complex layouts locally.
- It may have a wide display area so directional audio is important.
- It may have one loudspeaker per display, or perhaps some entirely different multi-loudspeaker setup known only to the endpoint itself.
- The endpoint may therefore have the capability of playing back audio from a wide range of positions.
- Either from a few fixed zones or with fine granularity.
- Either by routing a sound source to a single loudspeaker, by panning between pairs of loudspeakers, or by some other advanced distribution scheme involving several or even all loudspeakers.

6. Security Considerations

TBD

7. Acknowledgements

Thanks to Johan Nielsen for discussions and adding the Use case note.cuss

8. IANA Considerations

TBD

9. References

9.1. Normative References

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

9.2. Informative References

[I-D.ietf-clue-framework]
Romanow, A., Duckworth, M., Pepperell, A., and B. Baldino, "Framework for Telepresence Multi-Streams", draft-ietf-clue-framework-05 (work in progress), May 2012.

[I-D.lennox-clue-rtp-usage]
Lennox, J., Witty, P., and A. Romanow, "Real-Time Transport Protocol (RTP) Usage for Telepresence Sessions", draft-lennox-clue-rtp-usage-03 (work in progress), March 2012.

Authors' Addresses

Allyn Romanow
Cisco Systems
San Jose, CA 95134
USA

Email: allyn@cisco.com

Robert Hansen
Cisco Systems
Langley,
UK

Email: rohanse2@cisco.com

Andy Pepperell
Silverflare

Email: andy.pepperell@silverflare.com

Brian Baldino
Cisco Systems
San Jose, CA 95134
USA

Email: bbaldino@cisco.com