DISPATCH WG
Internet-Draft
Intended status: Informational
Expires: January 13, 2011

A. Romanow
Cisco
S. Botzko
Polycom
July 12, 2010

Problem Statement for Telepresence Multi-streams
draft-romanow-dispatch-telepresence-prob-statement-01.txt

Abstract

Telepresence systems create a "being there" conferencing experience. A number of issues need to be solved largely by manipulating multiple audio and video streams. Different systems take different approaches, employ different techniques, and convey information by using different vocabularies, making interoperability extremely challenging. This problem statement describes the typical issues that must be solved and uses examples to illustrate the kind of diversity that makes interworking problematic.

Table of Contents

1. Introduction

In a Telepresence conference, the idea is to create a feeling of presence - that you are in the same room with the remote parties. In order to create the "being there" or telepresence experience, a number of technical issues need to be solved. These issues are addressed by manipulating multiple media streams, video and audio - by describing them, controlling them, and signaling about them. The fundamental features of telepresence require handling multiple streams of media, and considering additional characteristics of those streams beyond those normally specified in existing videoconferencing standards.

Different telepresence systems approach solving the basic issues differently. They use disparate techniques, and they describe, control and signal media in dissimilar fashions. Such diversity creates an interoperability problem. The same issues are solved in different ways by different systems, so that they are not directly interoperable. This makes interworking difficult at best and sometimes impossible.

Some degree of interworking is possible through transcoding and translation. This requires additional devices, which are expensive and not entirely automatic. Specialized knowledge is required to operate a telepresence conference where the endpoints use different equipment and a transcoding and translating device is employed for interoperability. Often such conferences are interrupted by difficulties that arise.

The general problem that needs to be solved is this. The transmitting side sends audio and video streams based upon a model for rendering a realistic depiction from this information. If the receiving side belongs to the same vendor, it works with the same model and renders the information according to that shared model. However, if the receiver and the sender are from different vendors, the models they each have for rendering presence differ.

It is as if Alice and Bob are at different sites. Alice needs to tell Bob information about what her camera and sound equipment see at her site so that Bob’s receiver can create a display that will capture the important characteristics of her site. Alice and Bob need to agree on what the salient characteristics are as well as how to represent and communicate them. The telepresence multi-steam work seeks to describe the sender situation in a way that allows the receiver to render it realistically though it may have a different rendering model than the sender.

This problem statement identifies the fundamental issues that need to

be addressed to provide telepresence in typical use case scenarios. We show how different approaches to solving the problems and different techniques for handling multiple media create a challenge for interoperability.

This document describes some of the problems that arise, it is not an complete list, but rather it is more illustrative than exhaustive. Requirements, use cases and solutions are discussed in other documents.

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].

3. Fundamental Issues for Telepresence

The fundamental issues that must be handled to produce a typical telepresence conference, either point to point or multipoint include:

1. Participant display
    A. Placement of video
    B. Size
    C. Angle
    D. Overlap
    E. Display technology
2. Audio
    A. Placement, emanating from right place
    B. Type of audio
3. Different number of screens on sender and receiver sides
4. Participant display for multipoint
    A. Placement of video

B. Continuous presence

C. Control of display, how does it change? - automatic, user

5. Maintaining eye contact and gaze connection

6. Panoramic view for site switching

7. Mismatches between media characteristics between sender and receiver, such as:

A. aspect ratio

B. format

C. frame rate

D. resolution

8. Presentation

A. What methodology?

9. Security

A. SRTP?

B. Key methodology

4. Manipulating Media Streams

In addressing the fundamental issues, multiple media streams are handled in the following ways:

1. Sender and receiver understand each others capabilities

A. Number of video, audio and presentation streams that can be sent/received simultaneously

B. What media signaling protocol being used (SDP, proprietary, etc.)

2. Streaming control

3. Feedback mechanisms

4. Signaling about RTP payload

5. Media control signaling

   A. Video refresh

   B. Flow control

6. Signaling media formats and media capabilities

7. Signaling content type

8. Signaling device type

9. Signaling network characteristics per stream

10. Floor control signaling

5. Examples of Interworking Issues

This section describes several examples that illustrate the kinds of incompatibilities that arise when different systems take different approaches to an issue.

5.1. Designating Roles and Positions for transmitted streams

Senders and receivers need to have the same vocabulary and understanding of stream roles and positions in order to place them appropriately. For example one system may define roles as: center, left, right, legacy center, legacy right, legacy left, auxiliary 1/5 fps and auxiliary 30 fps positions. These roles as defined are a combination of "input devices" + "codec type/format" for transmission positions, and a combination of "stream decoders/output devices" + "codec type/format" for receive positions. Another system will not have the exact same vocabulary and meaning, though it still has to accomplish the same placement task.

How the cameras and encoders are wired determines how the local scene is displayed on the remote screen. In many systems right and left need to be exchanged to be seen properly, but this depends on the way the equipment is wired.

In describing how to display the local scene, the language can be misleading if there is no agreed upon reference for right and left. [for example, more]

Although often the video is displayed on separate monitors, it is

also possible to use projectors to create a video wall. In this case, there may be an overlap region between cameras which allows for projector blending. Also, although cameras are generally arranged to create a seamless panoramic view of the participants, it is also possible for there to be gaps between cameras (and corresponding gaps between displays).

There is also no reference for image size. Some rooms use proportionally larger displays, and set the camera field of view to show participants either standing or sitting at life size. Others use smaller displays, and set the field of view for sitting participants (cropping off heads when people stand). In order to preserve full size display when these systems interoperate, both systems must rescale their video.

5.2. Multipoint

Multipoint conferences, where there are more than two endpoints, create a wealth of technical issues to be solved. The primary one is which participants to display on each screen at each site. If the number of sites is greater than can be shown on the number of displays at a site, this adds to the complexity. There are, of course, almost unlimited ways this can be handled. We discuss the common approaches and how they differ.

The local screens can show all the camera image from the a particular remote site (site switching); or each local screen can show a participant or two from each of the remote sites (segment switching); or local displays can show a composite of remote camera shots (continuous presence). The choice of who to display on a screen can be determined by users, or, more often, automated according to voice activity level.

[Add user-controlled personal telepresence scenario.]

Policies are created and implemented in many ways. They tend to be based on some combination of what H.323 defines as centralized and decentralized. One of the challenges is that the endpoints in the conference may have different number of cameras and displays from each other so a common mode on the number of streams and their priority is required. Also, the various endpoints might have different bandwidth constraints and support different codec profiles.

A centralized multipoint conference is one in which all participating endpoints communicate in a point-to-point fashion with an MCU. The endpoints transmit their control, audio, video, and/or data streams to the MCU. The MCUA centrally manages the conference, processes the audio, video and/or data streams, and returns the processed streams

to each endpoint. In this mode, the MCU will mix the audio streams; and if using centralized video, will either use voice activated video switch, where everyone will see the active speaker and the speaker will see the previous speaker, or will use continuous presence mode, where the MCU will create a video stream with sub windows for each of the participants. MCUs can support multiple video layouts and they can be created automatically based on the number of participants or by a conference management application.

There are three methods commonly used for video stream distribution in centralized multipoint conferences. The three conference policies above can be implemented using any of these technologies.

Simple video switching (forwarding) has the advantage of low latency and low complexity. It can be used if all systems are capable of receiving the encodings used by the sending endpoints (including both the video codec and the image resolution/aspect ratio). In some situations it can be wasteful of bandwidth.

Full video transcoding usually has higher latency than switching It does not require system to be capable of receiving identical encodings, and different sites can connect with different bandwidths.

Layered video encoding combines some of the benefits of video switching and video transcoding. It is more complex than video switching, but less complex than video transcoding. Bandwidth and resolution can be reduced for each site. Since this is done by filtering out layers of the original encoding, the available bandwidths and resolutions are not as fine-grained as full video transcoding.

In decentralized mode or full mesh mode each endpoint creates its display mode. This requires each endpoint to receive multiple streams and send its video and audio to all participants, using multicast of unicast.

In practice, multicast is not now being used in commercial systems, so the size of a strictly decentralized multipoint conference is limited.

There are analogous issues for audio. Like video, the audio is rotated, so there is no clarity on the meaning of left and right. Since the number of streams, microphones, and speakers are not matched, the systems need to re-process the received audio in order to create the correct sound field for their respective rooms.

There are two ways in which the audio might be handled in this use case:

- o A single stereo audio stream is sent to the remote site, just as in standard videoconferencing.

- o Three monaural audio streams are sent to the remote site, with proprietary signaling to associate each audio stream with a video stream.

Microphones and speakers positions vary; and there is no agreed upon way to describe their placement. There is no agreed upon reference for audio level. In addition, audio may be sent as an independent stream from each microphone or as a multi-channel channel stream.

5.3. Capability Negotiation

Call setup for the telepresence conference will start with a single call establishing one video media stream. After the connection is established, a proprietary capability negotiation takes place that will enable both sides to identify that they are telepresence applications and capable of having two more video sessions and provide the connectivity information. The result is that two or more video sessions are established. The system may use two new SIP call legs or just add the two new video streams to the existing dialog.

[more to be added]

5.4. Differences in Media Characteristics

Media characteristics such as video format, aspect ratio, and visual scale can be handled differently at different sites creating incompatibility. To interwork, an adaptive strategy is necessary. Although differences in media characteristic must also be handled in a typical video conference, the problem is made more complex in Telepresence due to the multiple screens, cameras and streams.

Two examples - aspect ratio and visual scale are described here.

5.4.1. Aspect Ratio

If the aspect ratios in different sites are not the same, some technique needs to be applied to adjust for the difference. Although the same situation arises in normal video conferencing, multiple streams in telepresence conferencing causes more difficulties.

For simplicity let us assume a point to point case - two conference room on a point to point call. Both rooms have 3 screens and 3 cameras, as in 4.1 above. Both rooms have identical visual scale - the display width and distance between the participants and the displays are identical in both rooms. However the equipment -

cameras and displays - in each room has a different aspect ratio, 16:9 in one room and 4:3 in the other.

Although 4:3 is usually associated with standard definition TV and 16:9 with HDTV, telepresence systems may choose the aspect ratio to obtain a particular field of view. Projecting images in the 16:9 aspect ratio offers a wider presentation angle that shows fine details well (the pixel density is greater than a 4:3 system of the same resolution and scale). In the room with 16:9 media characteristic, people are shown at full size when they are seated. However, when they stand up the height of the display results in their image being cropped so that their heads are not shown. The other room uses projectors to display HD images with 4:3 aspect ratios. This results in an increased image height - the vertical field of view is 33% greater than the 16:9 system. The increased height allows most of the population to be shown full size whether they are standing or sitting.

Some strategy is necessary to deal with the case of the two sites having a point to point call. In order to convert formats of unequal ratios a variety of techniques can be used, such as: zooming (enlarging) and cropping (removing), letterboxing (adding horizontal bars), pillarboxing (adding vertical bars) to retain the original format’s aspect ratio, or scaling (which distorts) in a variety of ways.

For the video sent from the 4:3 room to the 16:9 room, several techniques can be used:

1. The 16:9 system might simply crop the top 1/4 of each 4:3 image. This will result in full size display, eye contact, and gaze awareness for the individuals who are seated. However, the standing presenter’s head will be cropped.

2. The 16:9 system might stretch each to the 4:3 images to fully fit the 16:9 display. This would reduce image height (creating geometric distortion) and create eye-contact error. Continuity of the panoramic image would be preserved.

3. The 16:9 system could pillarbox each of the 4:3 images, placing horizontal borders on the three displays. This results in reducing the image size to less than full size. It also destroys the continuity of the panoramic image, and introduces additional error in eye contact and gaze awareness.

4. The 16:9 system could pillarbox only the center display. This reduces the size of the presenter who is the focus of the meeting.

5. The 16:9 system could also crop the bottom of the center display. Visually this reduces the height of the presenter, but maintains full size. There is a vertical discontinuity in the panoramic image. Whether this is objectionable or not depends on the room layout.

Strategies 4 and 5 could be accomplished in response to a user command or automatically. The details will be discussed in more detail in future documents.

For the video sent from the 16:9 room to the 4:3 room, the receiving system simply letterboxes the video displays. Since the scales are identical, this full size image displays in the 4:3 room.

For the video sent from the 16:9 room to the 4:3 room, the common techniques are:

1. The 4:3 system places the border above the image. This maintains eye contact for those who are seated, but cannot maintain eye contact for the presenter.

2. The 4:3 system places the border below the images. If the 16:9 system crops the bottom of the center display then this will maintain eye contact for the presenter and the remote site.

3. The 4:3 system centers the images. Eye contact suffers for everyone, but the worst case eye contact error is better controlled.

In this use case, negotiation between the systems is not strictly necessary, no matter which scheme is used. However, the best user experience is obtained if both systems have knowledge about apect ratios being used and which participants are standing and which are sitting so they can adjust optimally.

5.4.2. Visual Scale

The visual scale of displays may differ between sites. Again, let us use the point to point case as a simple example. Assume two conference rooms in a point to point call. One room is designed for 6 participants, and has three 16:9 screens and 3 cameras. This room is designed to show participants at their normal size when seated (2 participants per camera/display). It does not have adequate display height to capture those who are standing. The second room is also designed for 6 participants, but shows 3 participants per camera/ display also at their full size. Therefore, it only needs two 16:9 cameras/display pairs. Since the field of view in both the vertical and horizontal is increased by 50%, it also shows those who are

standing without cropping.

For the video sent from the 2 screen (larger scale) room to the 3 screen (smaller scale) room, two approaches can be used:

1. The 3 screen system might simply show the participants on two of its displays. Participants will be shown at 67% of their full size. Eye contact and gaze awareness will be lost.

2. The 3 screen system might construct and display a vertically cropped 3-screen view, showing 2 participants on each screen. Participants will be shown at full size, with preservation of eye contact and gaze awareness.

For the video sent from the 3 screen to the 2 screen room, there are two analogous approaches:

1. The 2 screen system selects 2 streams and simply shows them on its displays. Participants will be shown at 150% of their normal size. Eye contact and gaze awareness will be lost, and some of the remote site is lost.

2. The 2 screen system might construct and display a 2 screen view (with a vertical border on the top) which shows 3 participants on each screen. Participants will be shown at full size, with preservation of eye contact and gaze awareness.

Although there is no need for negotiation between the systems, the best user experience is obtained if both systems have knowledge of the visual scale, and where individuals are seated, and can then choose the best manner of display.

6. IANA Considerations

This document contains no IANA considerations.

7. Security Considerations

While there are likely to be security considerations for any solution for telepresence interoperability, this document has no security considerations.

8. Acknowledgements

The draft has benefitted from input from a number of people including

Roni Even, Jim Cole, Nermeen Ismail, Nathan Buckles.

9. Informative References

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

Authors' Addresses

Allyn Romanow
Cisco
San Jose, CA 95134
US

Email: allyn@cisco.com

Stephen Botzko
Polycom
Andover, MA 01810
US

Email: stephen.botzko@polycom.com