Dispatching Sound Level Indicators in Conferences (Problem Statement)
draft-ivov-dispatch-slic-ps-00

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on November 23, 2009.

Copyright Notice

Copyright (c) 2009 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents in effect on the date of publication of this document (http://trustee.ietf.org/license-info). Please review these documents carefully, as they describe your rights and restrictions with respect to this document.
Abstract

The Conferencing Framework described in RFC 4353 defines the semantics necessary for conducting conference calls with the session initiation protocol. It also introduces a mixer entity responsible for combining all media streams and delivering them to the participants of the call. This document presents the lack of a standardized way for such mixers to deliver information about the audio activity (sound level) of participants in a conference call. The document describes the problem and discusses a few possible ways of transporting such information.

Table of Contents

1. The Problem .......................................................... 3
2. Possible Approaches .................................................. 5
   2.1. An Extension to the Conference State Event Package for SIP .................................. 5
   2.2. Various RTP Etensions ............................ 5
   2.3. Extending the Role of the CSRC Identifiers in RTP ........ 6
3. Security Considerations ............................................. 9
4. Informative References ............................................... 10
Authors’ Addresses .................................................. 11
1. The Problem

The Framework for Conferencing with the Session Initiation Protocol defined in RFC 4353 [RFC4353] presents an overall architecture for multi-party conferencing. Among others, the framework borrows from RTP [RFC3550] and extends the concept of a mixer entity "responsible for combining the media streams that make up a conference, and generating one or more output streams that are delivered to recipients". Every participant would hence receive, in a flat single stream, media originating from the others.

Using such centralized mixer-based architectures simplifies support for conference calls on the client side since they would hardly differ from one-to-one conversations. However, the method also introduces a few limitations. The flat nature of the streams that a mixer would output and send to participants makes it difficult for users to identify the original source of what they are hearing.

The IETF has already defined mechanisms (e.g. the CSRC fields in RTP [RFC3550]) that allow the mixer to send to participants cues on current speakers, but they only work for speaking/silent binary indications. In other words, there are still a number of use cases where one would require more detailed information. Possible examples include the presence of background chat/noise/music/typing, someone breathing noisily in their microphone, or other cases where identifying the source of the disturbance would make it easy to remove it (e.g. by sending a private IM to the concerned party asking them to mute their microphone).

One way of presenting such information in a user friendly manner would be for a conferencing client to attach sound level indicators to the corresponding participant related components in the user interface as displayed in Figure 1.
Delivering detailed speaker information to the user by displaying sound level for every participant.

Figure 1

Implementing a user interface like the above on the client side, however, would be quite delicate (if at all possible) since, as we have already mentioned, conference participants are generally receiving a single, flat audio stream and have therefore no immediate way of determining sound level based solely on the media. With today’s common conferencing solutions a mixer is the only party aware of such information. It therefore seems like a logical next step to determine what would be the best way to allow a mixer to deliver such information to conference participants.

The rest of this document investigates existing IETF mechanisms that could be extended in order to allow for a way to transport sound level information.
2. Possible Approaches

This section dwells on various existing mechanisms and their use for transporting participant sound level indicators.

2.1. An Extension to the Conference State Event Package for SIP

RFC 4575 [RFC4575] defines a conference event package for tightly coupled conferences using the Session Initiation Protocol (SIP) events framework. It allows for the delivery of various conference related details such as conference descriptions, participant count and identity. The document also provides a way of indicating who the speakers are at any given moment by specifying a mechanism for mapping conference participants to RTP SSRC/CSRC identifiers. All these details are dispatched in an asynchronous manner using the SIP events framework, or, in other words, through NOTIFY SIP requests following an initial SUBSCRIBE from a participant. It may therefore seem logical to try and extend the framework by adding the syntax necessary to convey sound levels.

Further thought on the subject, however, raises numerous issues with such an approach. Sound level in human speech is obviously a very time sensitive characteristic which would require frequent updates (i.e. approximately once every 50-100 ms). In order for the update of the user interface to appear "natural" to the user, sound level information would probably have to be delivered after every one or two RTP packets. Using RFC 4575 [RFC4575] or SIP in general for this would generate traffic on the (often low-bandwidth) signalling path comparable to, if not exceeding, the media itself.

It is probably also worth mentioning that the use of RFC 4575 [RFC4575] for such a feature would make the mechanism incompatible with non-SIP signaling protocols like, for example, XMPP [RFC3920] and its Jingle extensions.

2.2. Various RTP Extensions

The sound levels of different human voices in a conversation are one kind of particularly fast changing information RTP seems to be well suited for. Additionally, RTP syntax, through the CSRC list in the RTP packet header and one or more SDES RTCP packets, already allows a mixer to specify the identities of the users whose voices were aggregated in a mixed stream. It seems thus straightforward to consider an extension to RTP as a possible approach for carrying such information.

A first option for extending RTP is to define an RTP header extension as specified in RFC 3550 [RFC3550] that would allow encoding sound
level indicators for each element of the CSRC list. The main advantage of such an approach would consist of the very little impact it would have in terms of bandwidth overhead; however, the RTP header extension mechanism was initially meant only for experimentation and its use for specifying new features is explicitly discouraged.

A possible workaround for such a limitation could be the definition of that extension in a new RTP profile, in turn defined as an extension of the Audio/Video profile specified in RFC 3551 [RFC3551]. However, the complexity introduced in the profile negotiation process, especially when done with ICE [I-D.ietf-mmusic-ice], makes the approach an overkill for the goal it tries to achieve.

Alternatively, the syntax needed for encoding sound level indicators for the participants in an audio conference can be specified as a new payload type for the RTP Audio/Video profile defined in RFC 3551 [RFC3551]. The drawback of such an approach resides in the significant increase of RTP packets it would generate; in fact, even if the amount of additional information would be very small, encoding it in a new payload would require a separate RTP packet for each update (that, for a decent user experience, should happen several times per second).

2.3. Extending the Role of the CSRC Identifiers in RTP

The RTP [RFC3550] specification defines a Synchronization Source (SSRC) identifier. SSRCs are used by every RTP source (e.g. every participant in a conference call) and they are meant to be globally unique within a particular RTP Session. Again, according to the specification, mixers are expected to record the SSRC identifiers of all contributing streams as a list of CSRC identifiers in the RTP packets transporting the resulting combined stream. In the case of a conference call this would mean that if the mixer is respecting the above, every participant would receive the SSRC identifier of every other active participant.

RFC 4575 [RFC4575] then defines a way of mapping an SSRC identifier to an actual conference participant through the <src-id> tag. The mapping provides a way of determining which are the currently active (i.e. speaking) conference call participants.

A very simple way for a mixer to use the CSRC fields as a transport means for sound level indication would be to extend their meaning over a series of packets rather than a single one. This way it could be specified that the sound-level of a particular participant, represented on a zero to ten scale, corresponds to the number of occurrences of its CSRC identifier in the ten most recent RTP packets.
received from the mixer.

For example, consider a conference call with four participants: Alice, Bob, Carol, and Dave. At a certain point in time Alice has a sound level of 6/10, Bob 1/10, Carol is silent or in other words 0/10 and Dave has a level of 3/10. In order to describe this state the mixer could have sent the last ten RTP packets with the following CSRC configuration:

```
+-------+----+-----+-----+-----+-----+-----+-----+-----+-----+------+
|       | P1 | P2  | P3  | P4  | P5  | P6  | P7  | P8  | P9  | P10  |
+-------+----+-----+-----+-----+-----+-----+-----+-----+-----+------+
| Alice | +  | +   | +   | +   | +   | +   |     |     |     |      |
|       |    |     |     |     |     |     |     |     |     |      |
| Bob   |    | +   |     |     |     |     |     |     |     |      |
|       |    |     |     |     |     |     |     |     |     |      |
| Carol |    |     |     |     |     |     |     |     |     |      |
|       |    |     |     |     |     |     |     |     |     |      |
| Dave  |    |     |     |     |     |     | +   | +   | +    |
+-------+----+-----+-----+-----+-----+-----+-----+-----+-----+------+
```

A possible representation of a particular sound level configuration through the presence/absence of CSRC identifiers in subsequent RTP packets.

Table 1

The graphical interface of a user agent involved in such a conference (like the one sketched in Figure 1) would then display correct sound levels just showing for each participant as many ticks as were the occurrences of the respective CSRC in the previous ten RTP packets.

The algorithm for encoding sound level information this way is relatively simple. In order to determine whether or not to include a particular CSRC a mixer should:

- include the CSRC if the sound level of the participant in the current packet is greater than the number of occurrences of that same CSRC in the nine previous packets;
- omit the CSRC if the sound level of the participant in the current packet is lower than or equal to the number of occurrences of that same CSRC in the nine previous packets.

There are several advantages to using this approach, the most obvious being its simplicity as well as the fact that sound level information is transported together with the parts of the audio stream that it actually concerns which should make synchronization straightforward.
The technique would also work with other signaling protocols using RTP such as XMPP’s [RFC3920] Jingle extensions for example.

One of the first disadvantages that come to mind with this approach is the fact that mixer would not be able to indicate level in a single packet but would have to distribute it over a succession of up to ten packets which would reduce the reactivity of the representation.

It is probably worth mentioning, however, that a granularity that allows switching from a level of zero to ten and back to zero again in an instant manner is not of much use anyway since such UI updates would be barely perceptible to the user. Still, this is a UI decision and making it on a protocol level may bring some inconveniences.

Another possible problem would come from implementations using CSRC presence in a binary way to determine current speaker. When running against a mixer that supports sound level indication such implementations may appear to be jumpy as the participants that they are designating as active may be changing status too rapidly.
3. Security Considerations

1. A MITM could modify sound level indicators and make participants believe that someone is saying something when they actually aren’t ...

2. Should use some authentication method to resolve this?

3. Could break compatibility with SRTP?
4. Informative References

[I-D.ietf-mmusic-ice]


Authors’ Addresses

Emil Ivov
SIP Communicator
Strasbourg 67000
France

Email: emcho@sip-communicator.org

Enrico Marocco
Telecom Italia
Via G. Reiss Romoli, 274
Turin 10148
Italy

Email: enrico.marocco@telecomitalia.it