机器人音视频框架

胡剑锋

hujianfeng@yongyida.com

2016-01-28

实时音视频通信的应用一般都是经过采集、编码、打包、传输、拆包、解码、播放、录制等过程。其中网络传输还包括网络打通、建立会话、端口协商、数据收发等一系列过程。

整个视频应用可以分为音视频编解码模块、RTP 打包和传输模块、P2P 模块、业务交互模块。

一 音视频编码模块

由于采集后的原始音视频数据都很大，如果直接发送到网络，需占用大量的带宽，所以在传送之前需要进行压缩编码，在接收之后再进行解码还原。

音视频编解码一般分为硬件编解码和软件编解码。硬件编解码故名思意就是由硬件完成编解码。硬件编解码是一种广泛使用的硬件加速技术，一般由特定的 DSP 来完成。软件编解码由 CPU 来完成大量的编解码运算，所以比较占用 CPU 资源。

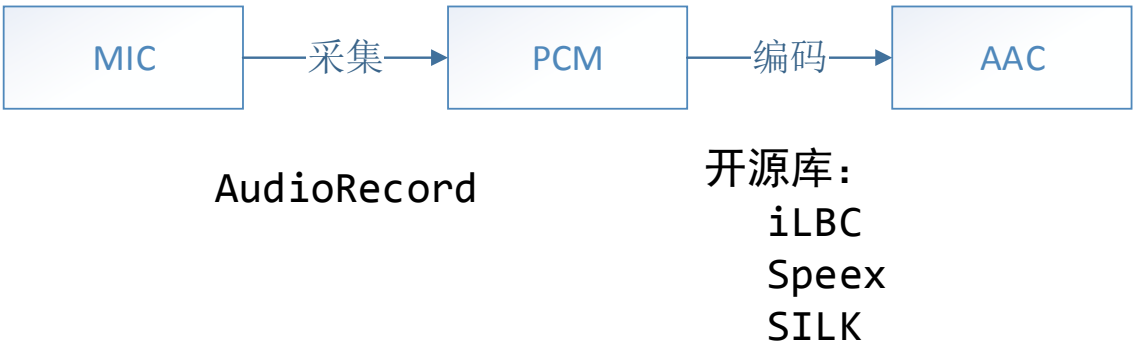

图 1. 音频编码过程

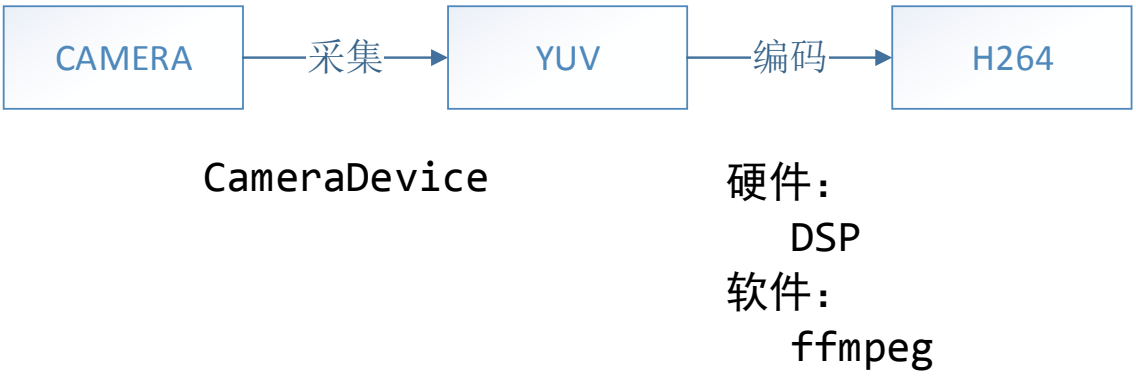

图 2. 视频编码过程

因为硬件编解码有很大的优势，加上机器人上的 MTK 芯片方案支持视频硬编解码，所以确定使用硬件进行视频编解码，使用 iLBC 进行音频编解码。

手机端 IOS 都支持硬编解码，ANDROID 系统大部分可以使用硬件编解码，如果不支持的可以使用 ffmpeg 进行软件视频编解码。

二 RTP打包和传输模块

实时传输协议 RTP（Real-time Transport Protocol）是一个网络传输协议，它是由 IETF 的多媒体传输工作小组 1996 年在 RFC 1889 中公布的，后在 RFC3550 中进行更新。RTP 协议详细说明了在互联网上传递音频和视频的标准数据包格式。RTP 用来为端到端的实时传输提供时间信息和流同步。

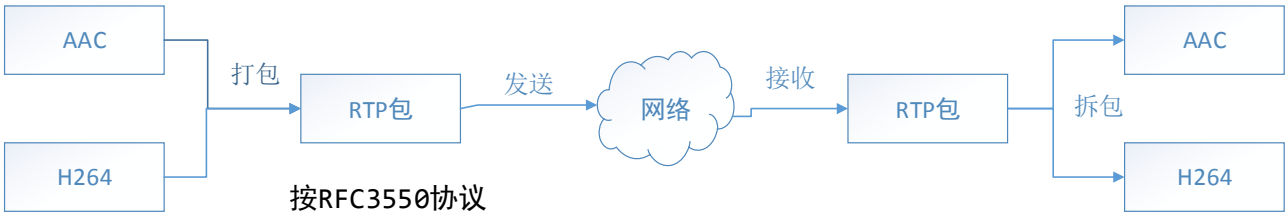

图 3. RTP 打包传输过程

RTP 头格式大致如下：

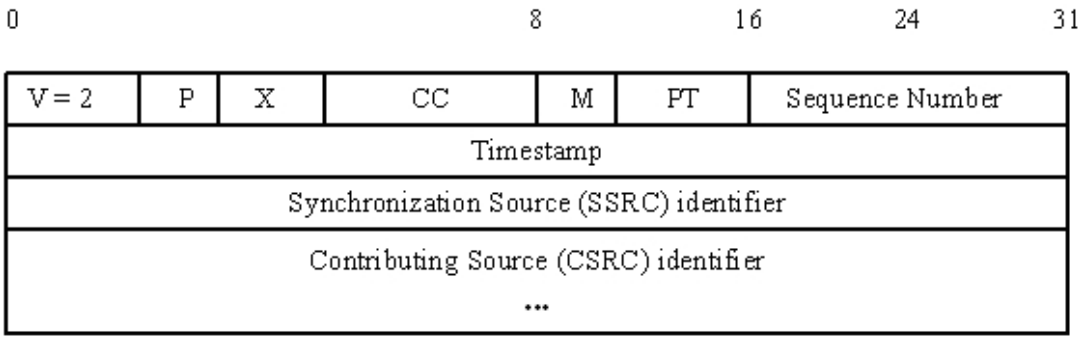

图 4. RTP 头结构

版本号（V）：2 位，表示 RTP 版本号。协议初始版本为 0，RFC3550 中规定的版本号为 2。

填充标志（P）：1 位，如果 P=1，则在该报文的尾部填充一个或多个额外的八位组，它们不是有效载荷的一部分。

扩展标志（X）：1 位，如果 X=1，则在 RTP 头后跟有一个扩展头。

CSRC 计数器（CC）：4 位，表示 CSRC 标识符的个数。

标记位（M）：1 位，标记 RTP 流中的重要事件，不同的有效载荷有不同的含义，对于视频，标记一帧的结束；对于音频，标记会话的开始。

载荷类型（PT）：7 位，用来指出 RTP 负载的具体格式。在 RFC3551 中，对常用的音视频格式的 RTP 传输载荷类型做了默认的取值规定，例如，类型 2 表明该 RTP 数据包中承载的是用 ITU G.721 算法编码的语音数据，采用频率为 8000HZ，并且采用单声道。

序号：16 位，每发送一个 RTP 数据包，序号加 1。接受者可以用它来检测分组丢失和恢

复分组顺序。

时间戳：32 位，时间戳表示 RTP 数据分组中第一个字节的采样时间，反映出各 RTP 包相对于时间戳初始值的偏差。对于 RTP 发送端而言，采样时间必须来源于一个线性单调递增的时钟。

同步信源(SSRC)标识符：32 位，用于标识同步信源。该标识符是随机选择的，参加同一视频会议的两个同步信源不能有相同的 SSRC。

特约信源(CSRC)标识符：32 位，可以有 0～15 个。每个 CSRC 标识了包含在该 RTP 报文有效载荷中的所有特约信源。

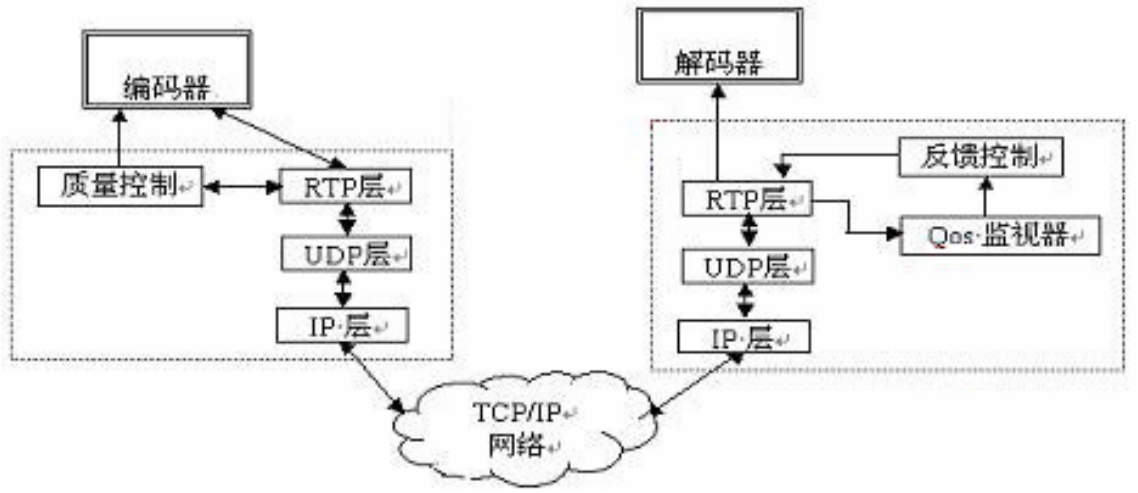

图 5. RTP 数据在网络中传输过程

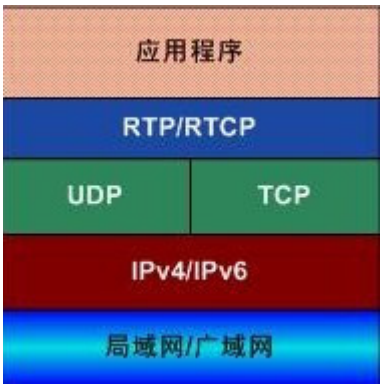

图 6.　RTP 协议与其他协议的关系

在一次视频会议中同时使用了音频和视频，两种媒体将分别在不同的 RTP 会话中传送，每一个会话使用不同的传输地址（IP 地址＋端口）。

RTP 一般由 UDP 进行承载传输，但也可以由 TCP 进行传输。基于性能考虑，这里选用 UDP 进行传输。

三 音视频解码模块

音视频解码就是编码的反过程，把编码后的数据解码为原始数据，再在输出设备上输出。

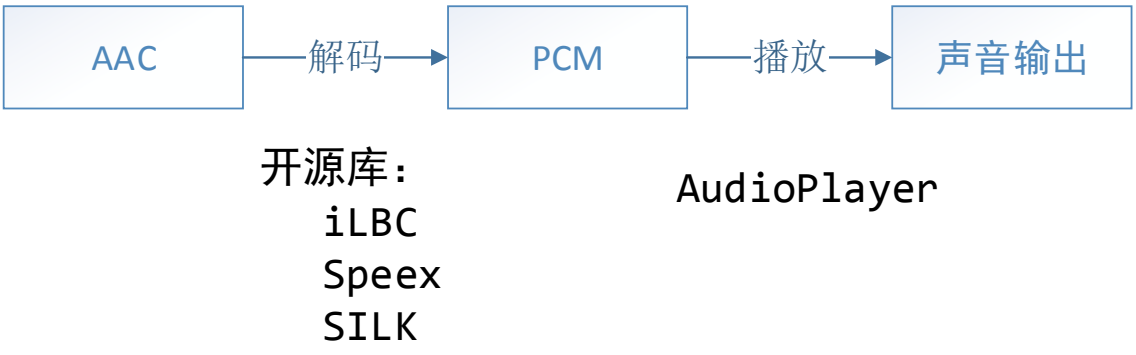

图 7. 音频解码过程

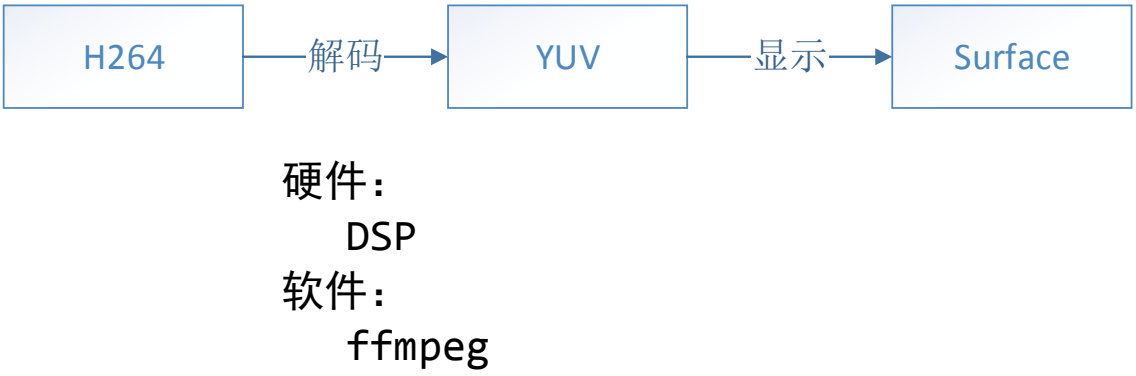

图 8. 视频解码过程

四 P2P模块

P2P 模块主要完成客户端和服务器打洞前的消息交互，以及客户端之间的打洞过程。

本模块后续再完成。

五 业务交互模块

要完成一个视频会议通信，必须有用户注册、登录、获取好友列表、建立房间、视频邀请、传送数据等一系列过程，所有这些都必须视频双方或多方和服务器进行信息交互。

这些交互信息分为两类：

- 客户端和服务器的交互
- 客户端和客户端的交互

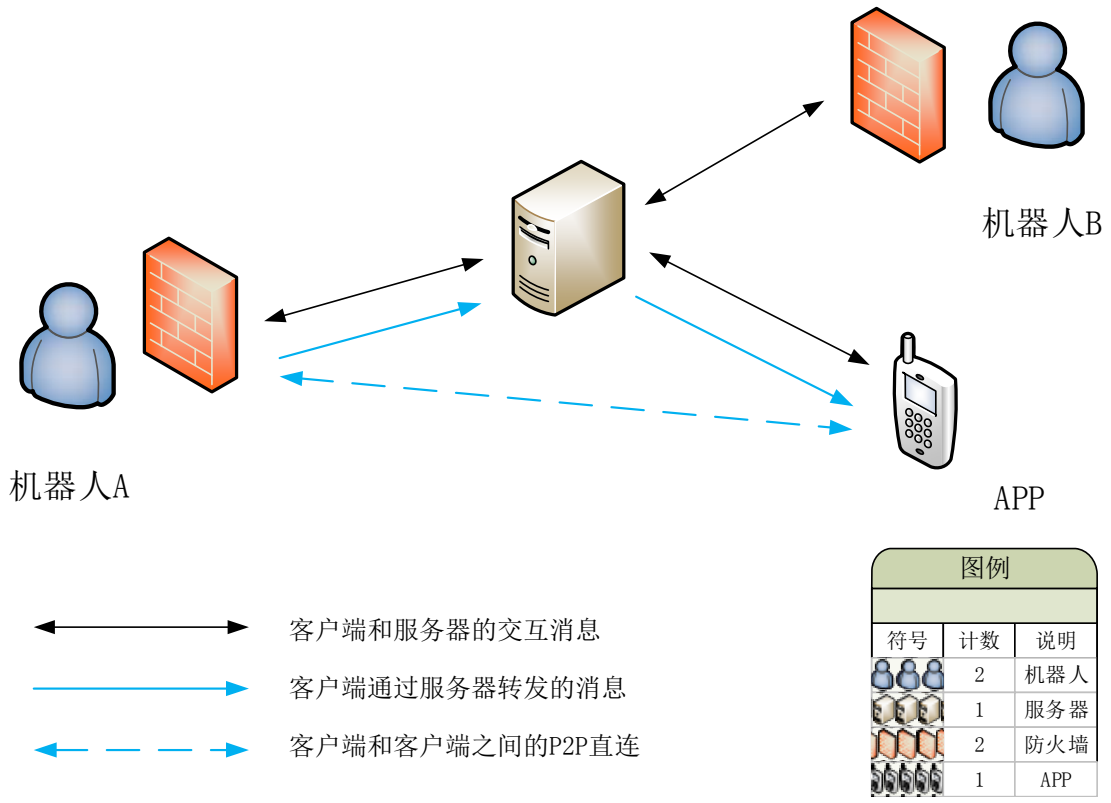

图 9. 客户端和服务器端消息交互过程

交互过程中的数据也分为两类：

- 音视频数据
- 其它消息数据

基于 TCP 和 UDP 传输特点和传输数据对可靠性的要求，规定音视频数据使用 UDP 进行传输，其它消息数据使用 TCP 进行传输。

在未实现 P2P 功能前，客户端和客户端不可能进行直接通讯，所以所有客户端发送到其它客户端的消息均通过服务器进行中转。