视频会议中基于延迟感应的丢包隐藏

华云生
香港中文大学

关键词：视频会议　丢包隐藏

IP网络上的视频会议技术

对一个视频会议系统来说，在互联网协议层面上，信息的发送端会先后把影音信号封装成帧和数据包来将其传送到接收端。在信息传送中，数据包可能会丢失或“迟到”，从而导致信号中断。在传输过程中产生的网络延迟（network delay）及缓存时延（jitter buffer delay）被并称为端到端延迟（end-to-end delay）。在实时系统中，每个数据包都有一个播放的时限（play out deadline）。此时限之前接收到的数据包可以正常播放；此时限后接收到的则算作迟到，在实时系统中无法再被利用及播放，影音的质量也会因此下降（见图1）。

针对数据包的“迟到”，我们的对策是播放调度策略（play-out scheduling，POS），即通过一定方法控制播放时限。这里涉及一个重要概念，称为谈话时延（mouth-to-ear delay，MED），即声音从说话者口中传到聆听者耳中所需要的时间。之所以说它重要，是因为它既关系着影音的质量，又关系着人们交流的便捷。

正如前面提到的，如果网络上有丢包或延迟，影音质量会有所下降。我们有办法改善此状况下的影音质量。这可以通过在数据包流层面（packet stream layer）和编解码层面（codec-layer）上做丢包隐藏来实现。然而，这些方法都需要一定的缓存时间，需要通过播放调度策略增长谈话时延来满足。

但是，谈话时延不可能被无限延长。在面对面对话中是

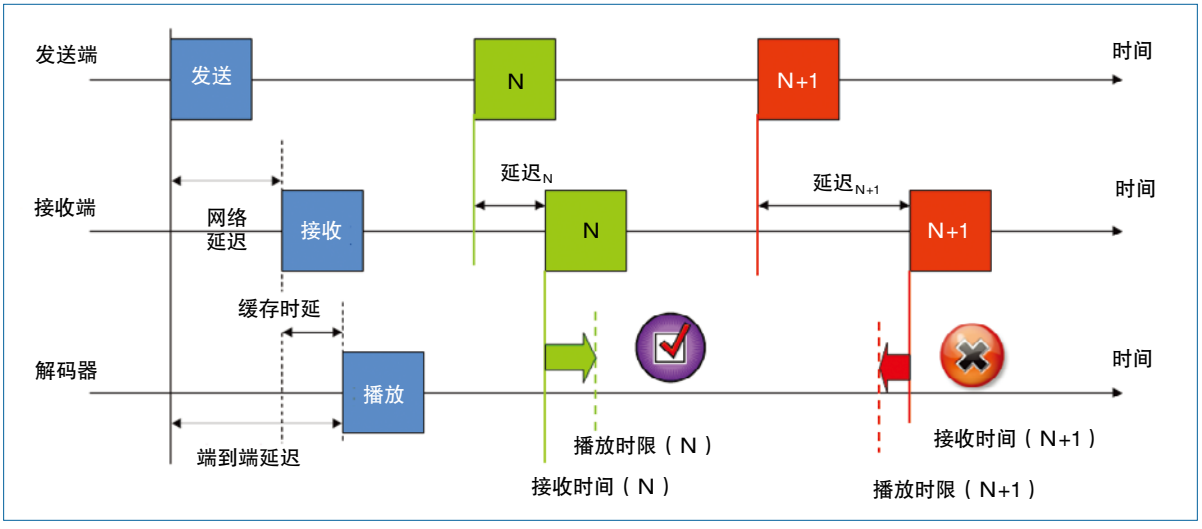

图1　缓存时延与播放时限

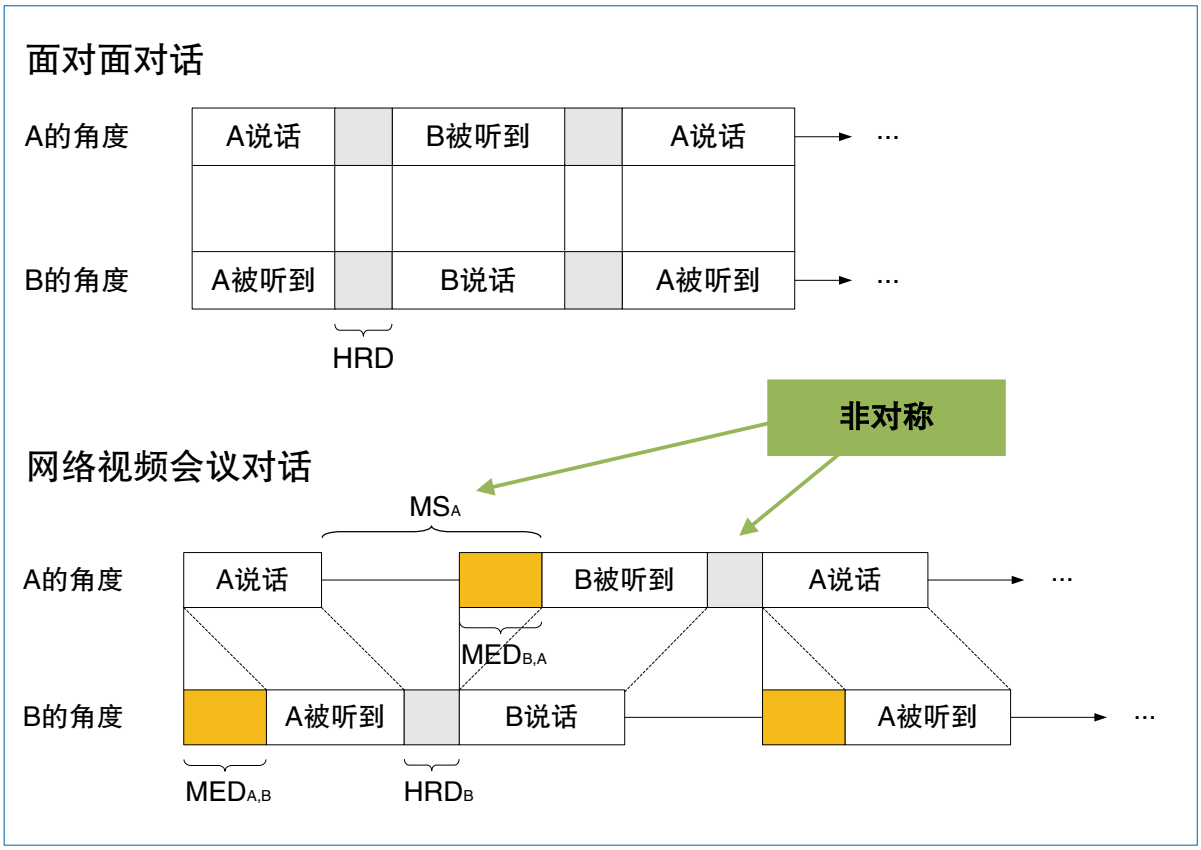

图2　网络视频会议的对称性

说话语速	平均语段时间	平均思考时间	切换数	总时间
慢(例：日常对话)	3,737 毫秒	729 毫秒	7	35 秒
快(例：商业对话)	1,670 毫秒	552 毫秒	15	35 秒

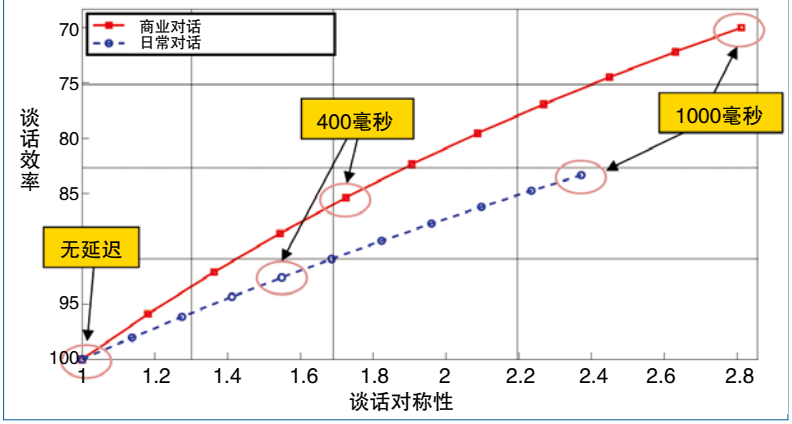

图3　不同时延对谈话平衡和效率的影响

没有谈话时延的，在这种情况下双方能顺利而自然地进行交流。但是，有了网络延迟，状况就不同了，双方听到的对方声音都会有较明显的延迟。如图2所示，A和B两个人进行对话。在面对面对话中，A一说完话，B会立刻听到。在B经过一段思考时间（human response delay，HRD）后，他可以立刻给A回复。B对A的讲话也是这样。可是，在有网络延迟的情况下，A的话会经过一段谈话时延才能传到B那里。这会使得A和B对双方谈话的节奏有不同的认识。A会觉得B要经过很长一段时间才

能回复他的话，而不是像平常那样只经过一小段思考时间。这是因为中间多了两段谈话时延（$MED_{A,B}$和$MED_{B,A}$），使得静默时间（mutual silence，MS）变长了。A在对比B回复之前的静默时间（MS_A）和他自己的思考时间（HRD_A）后，会得到一个结论：他跟B的思考时间明显不对称。A在对比他在有网络延迟情况下的谈话时间和他平常的谈话时间后，会得到另一个结论：耗费的时间久了。我们把这两方面因素定义为谈话对称性（Conversational Symmetry，CS）及效率。它们是人们能间接察觉到网络延迟的重要原因。研究发现，在不同的对话节奏下，人们对谈话对称及效率有不同的要求。在图3上面我们可以看到有蓝色和红色的两条曲线，分别代表两个不同场景的对话：蓝色的是比较慢的日常对话，红色代表比较快的商务对话。图中显示即便有相同的谈话时延（400毫秒或1000毫秒），这两种对话的对称性跟效率都会不同。

综上所述，谈话时延对影音质量和人们交流的便捷度都有很大影响。它们对谈话时延的要求是相反的。前者需要更多的谈话时延来提高影音质量，后者需要更短的谈话时延来提高谈话便捷性。这里存在着取舍（trade-off）。

谈话时延增长，视频通话的影音质量就会提高，但相应的交流便捷度会降低；谈话时延缩短，视频通话质量就会降低，对话效率则会得到提高。在这中间有一个最优的结合点，也就是最佳平衡点，在其上可以得到最好的质量和便捷度。不同的对话情形会有不同的最优结合点。如何在实时网络对话中得到最佳点呢？这就需要设计播放调度策略。

这个策略包含多重度量准则，如影音质量、谈话对称性及效率。例如，在Skype平台上，如果增加了Skype的延迟，影音的质量会得到提高，但是谈话对称性及效率会受到影响。同样在MSN上，如果把延迟增长，影音质量都能相应的提高，而谈话便捷性会受影响。网络中的延迟变化很大，虽然通常只有20~30毫秒，但是有时候会飙升到250毫秒，怎样能在这种网络状态下获得好的质量是个值得研究的问题。

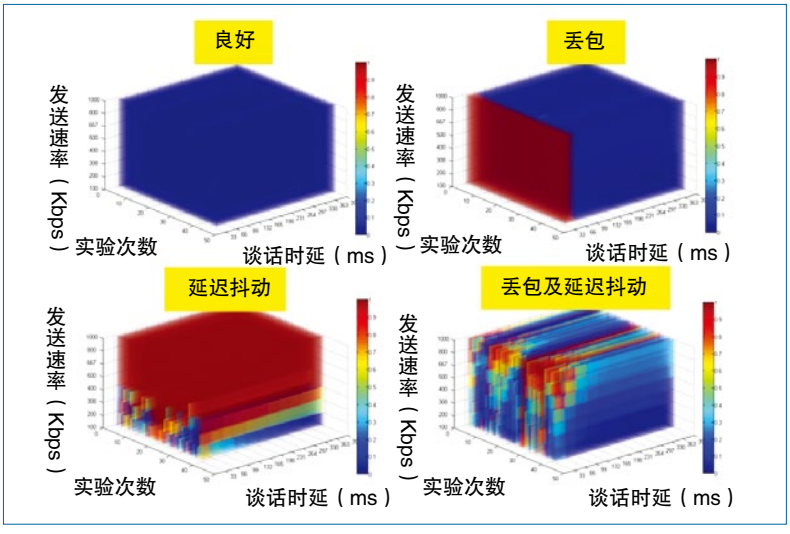

图4　四种网络状态下的效果图

互联网上的网络行为

互联网每天都在改变。十年前在网络上做的实验，现在再做获得的结果已经不同，今年的实验结果又跟去年的不同。互联网的网络质量也有很大差异，例如在无线网络上，通话质量会受到很大影响。网络上信号传过来后，延迟有时会突然增加，导致数据包迟到。有时甚至会直接导致丢包。

延迟抖动（delay jitters） 网络延迟通常是20~30毫秒，但是中间有时会突然达到500毫秒，然后再降回20~30毫秒。我们称之为延迟抖动，它对影音质量会有很大影响。我们在视频聊天时，有时候信号会突然停下来，就是因为受到延迟抖动的影响。这些失去的信号无法恢复。

丢包（loss） 是指数据包在网络上的丢失。有时候网络延迟不大且稳定，但中间有很多数据包丢失了，视频跟音频会中

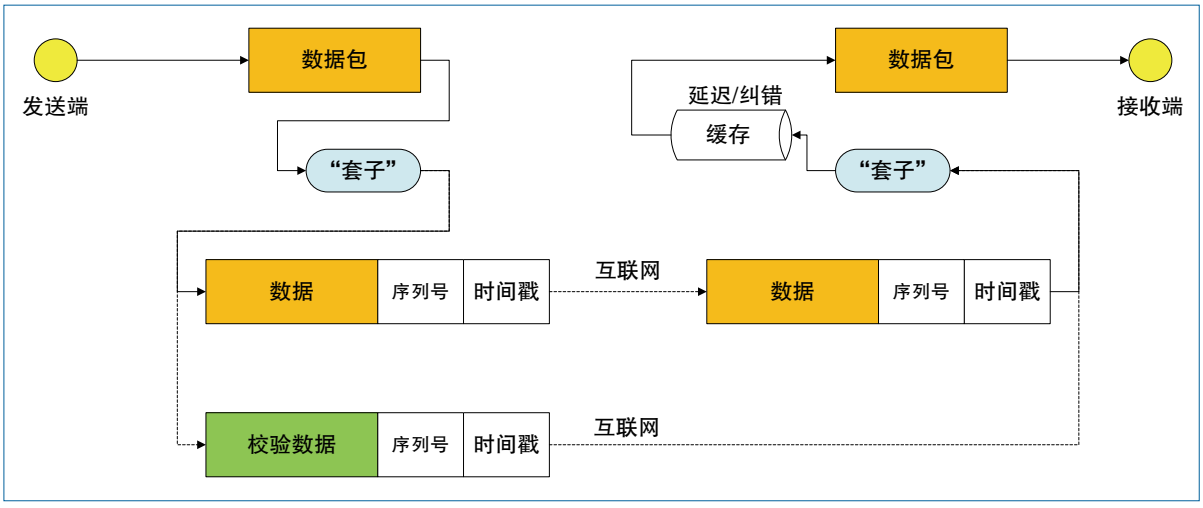

图5 Windows平台的“套子”

途停下或出现花屏与噪音，人们看起来会很不舒服。

解决此类问题的方法有两个。一是视频源层面方法（source-level methods）。我们可以把每一个数据包之间的相互关联减少，这样少量的丢包并不会引起连串数据的损失。然而，这可能导致码率增高，在相同带宽上视频质量下降。另外一个是通道层面方法（channel-level methods），也就是把缓存时间调高，使得我们可以做更好的纠错。传统的传输控制协议（TCP）重传方法并不适合实时系统，因为数据包重传的耗时太长。我们使用的纠错方式主要是延后播放时限以及添加数据冗余。之前已经提到，这些方法都会增加谈话时延。关键问题是，人们为了提高影音质量，可以容忍的增加的谈话时延最大是多少。是50毫秒，100毫秒还是200毫秒？这是问题的关键所在。通过实验我们可以看到250毫秒的延迟是可以接受的；在增加到431毫秒的时候，也没有太大的区别；但再长就不能接受了，因为随着时延的增加，虽然影音质量会增加，但谈话的便捷性和互

谈话时延= 网络平均延迟+50 毫秒

视频 PSNR[dB]

路径	Akiyo		Mother&Daughter		Foreman		Outdoor		Business	
	参照	结果	参照	结果	参照	结果	参照	结果	参照	结果
1	41.86	42.03	38.95	39.99	30.93	32.04	39.09	40.38	38.69	38.70
2	41.00	41.90	38.99	39.95	30.81	31.83	38.35	40.12	38.16	38.54
3	39.89	41.11	37.78	38.86	28.82	30.64	36.59	39.78	36.88	38.15
4	42.18	42.18	39.70	40.44	31.77	32.82	39.43	40.54	39.01	38.71

音频 PESQ[MOS]

路径	Outdoor		Business	
	参照	结果	参照	结果
1	3.19	3.33	3.33	3.49
2	3.34	3.56	3.30	3.43
3	2.93	3.32	2.89	3.23
4	3.26	3.16	3.64	3.64

图6 编解码器对音视频质量的改善

动性（interactivity）会降低。

我们在2012年9月的一个关于当前互联网研究的课题中，收集到10000多条数据传输样本（traces），记录着各种传输的状况。我们把所有的样本进行分类，发现大部分状况良好，丢包率很低，延迟也很短（最长的延迟是153毫秒）。同时，也有部分状况不太好，或有丢包，或有延迟抖动，或两者均有，比例大约为23%，这些是我们需要重点改善质量的状况。

改善方法：

方法一：改变数据包的传输率（packet transmission rate）。通过这种方法，丢包和延迟跟抖动有时会得到改善，但有时没有明显效果。因此单独控制速率没有太大的作用。

方法二：改变缓存时间（buffering time）。这个已经在上文中叙述。

结合这两种方法，可以改善大部分状况下的网络质量。但并不是在所有的网络上都能达到这样的效果。

情形一：良好的网络。质量在各种传输率下都能保持得很好。情形二：有丢包的网络。如果谈话时延非常短，所有数据包都会丢失，但如果谈话时延增加到一定程度，所有的数据包都可以恢复。情形三：有延迟抖动的网络。谈话时延短的时候，效果很不好。如果发送速率高的话也很不稳定。唯一稳定的情形是谈话时延要非常的长，同时发送速率很低，这样才能接收到大部分的数据包。情形四：同时有丢包和延迟抖动的网络。在谈话时延较长，发送速率也不是太高的状态下，可以恢复大部分的数据包（见图4）。

改进Skype和Windows Live Messenger的方法

要改进Skype和Windows Live Messenger，我们提出使用“临界感知差异（Just Noticeable Difference，JND）”这个心理物理学的概念，它指的是普通人无法感知到的细小变化。把Skype和Windows Live Messenger的谈话时延调高来改善影音质量，同时又使其不被察觉，就是我们的目标。

按照心理物理学的传统方法，我们通过统计的方式来定义“无法感知”这一概念。只要有75%的受测者感觉不到差异，我们就认为这个差异无法被感知。事实上，心理物理学也有使用50%作为度量标准的。然而，50%的准确性是可以通过随机猜测来得到的，其区别并不显著，因此不被我们采用。在理想的网络情形下，谈话时延增加的时候，临界感知差异也相对的增加。例如，我们的实验结果显示，谈话时延为100毫秒的时候，临界感知差异也是100毫秒。但是谈话时延达到300毫秒的时候，临界感知差异可达200毫秒。也就是说，此时即便我们把谈话时延从300毫秒增长到500毫秒，受测者并没有留意到有什么不同。

我们还发现，在有丢包的网络上，临界感知差异增长得比较慢。在谈话时延为300毫秒的时候临界感知差异大概是150毫秒，此时受测者没办法留意到有什么不同。另一方面，在对话语速较慢的情形下，在谈话时延为300毫秒的时候，临界感知差异可达250毫秒。

由于Skype和Windows Live Messenger的源代码不开放，我们无法直接改变源代码来测试我们的播放调度策略。所以唯一的办法只能是在这些软件外面做一个“套子”，等于把额外的缓存时延放在Skype和Windows Live Messenger上面去（如图5所示）。

我们还部署了一个测试平台（testbed），它能够模拟真实网络的通信状况。在外面加了一个套子后，把延迟调高，质量相对就会好一些。经过对受测者的询问，75%的人都没有发现延迟有什么不同。这说明把延迟调高后，受测者是不会感知到区别的。与此同时，影音质量却得到了很大的提升。这样用户的综合体验会得到改善。

改善稳定性的新编解码器

我们研究如何设计一个新的编解码器（codec）来改善视频会议中影音的质量。在H263、H264视频格式下，发送网络包的时候，很多情况下需要发送一个

大的帧，但在30~50毫秒内收到它是非常困难的。我们认为，在做编解码时候，可以把一些帧的使用时间调后，比如在300毫秒后才用它，这样传输跟纠错都会容易很多。

通过图6可以看出，在不同的网络环境下效果改善都很明显。视频通常会有1.5个数值（dB）的提升，音频则有0.7~0.8的提升。正因为我们把重要的帧用较多的时间进行传送，影音质量才有这样的提升。

谈话时延和影音质量之间的平衡

在300毫秒的延迟下，谈话的便捷性跟互动性会较高，但是影音质量会比较低；相反，在1000毫秒的延迟下，谈话的互动性没有之前高，但影音质量提高了。如前所述，在讲话时间比较长的情形下人们能承受较强的谈话时延。从300毫秒增加到350毫秒甚至450毫秒，通话双方根本听不出有什么区别。当然，如果超过1000毫秒，人们还是会感觉等待时间太长了。

我们工作的重点就是要调节谈话时延以便达到最好的取舍。我们采用了以下解决方案来达到这个目标：在不同的通话场景和网络环境性下，通过高效的主观测试去搜集人们通话的偏好，然后再根据实时的网络环境因素和线下搜集的通话偏好，实时调整谈话时延。

总结

本文的研究过程使用了整体化分析方法（holistic approach）。首先通过分析互联网的网络特性，设计出针对性的解决方案；使用了临界感知差异的概念，在不影响视听觉感受的前提下提高谈话时延；通过重新设计编解码器优先传输关键信息；并使用丢包隐藏策略去恢复丢失或延迟的信息，提高实时编解码及传输的影音质量。实验结果表明，这样的综合设计方案取得了明显的改进效果。■

（本文根据2012中国计算机大会（CCF CNCC 2012）报告整理而成）

华云生

2012CCF海外杰出贡献奖获得者。香港中文大学常务副校长、伟伦计算机科学与工程学讲座教授。主要研究方向为非线性规划、多媒体信号处理。bwah@cuhk.edu.hk

CCF将协办IEEE Big Data 2013

应IEEE邀请，CCF将协办于2013年10月6~9日在美国硅谷举行的IEEE BigData 2013（The IEEE International Conference on Big Data 2013）。中国科学院计算技术研究所研究员、CCF大数据专家委员会秘书长程学旗将代表CCF参加该会的指导委员会（Steering Committee）。根据CCF和IEEE双方的合作备忘录，CCF会员参加该会议时，注册费享受和IEEE会员一样的优惠。