基于流特性和真值程度的 VoIP 语音质量单端客观评价[1]

成卫青 [1,2] 龚俭 [1] 丁伟 [1]
(东南大学 计算机科学与工程学院 南京 210096)
(南京邮电大学 计算机学院 南京 210003)

摘要：提出了一个仅利用 IP 网络测量技术，基于流特性客观评价 VoIP 感知服务质量(PQoS)的非侵入单端方法——FSPAV，其中最关键的是定义了 3 个与用户感知相关的流特性测度：平均语音包大小、语音包平均到达间隔、平均到达间隔抖动。FSPAV 不需要监测语音信号，而是监测用户主机接收到的包含对端用户语音数据的 IP 分组，因为不需要同步时钟或解析应用协议，3 个测度的测量成本低廉；FSPAV 使用个体真值程度度量，将通话片段的 3 个测度测量值映射成一个通话质量客观评价值，在计算过程中还能得出有助于语音质量降级因素分析的每个特性的优劣程度。为验证本文方法的效果，利用流行的 VoIP 软件 QQ 和 Skype 进行多次语音通话，由自主开发的实现了 FSPAV 的 VoIP PQoS 测量工具监测通话流量并预测**通话质量**，并由通话用户进行通话质量的主观评价。实验结果显示客观评价值与主观评价值之间具有相当高的相关性（相关系数高达 0.9677），表明了本文方法的有效性。

关键词：VoIP; 语音质量客观评价; 流特性; 真值程度度量

Flow features and truth scale based single-ended objective assessment of perceived quality of VoIP service

Cheng Wei-qing, Gong Jian, Ding Wei
(School of Computer Science and Engineering, Southeast University, Nanjing 210096)

Abstract: This paper proposes a flow features-based, nonintrusive, single-ended objective speech quality assessment method called “FSPAV”, which only utilizes IP measurement techniques to assess perceived quality of VoIP service (PQoS). The cores of the method are three PQoS related metrics of flow features: average length of speech packets, average inter-arrival time of speech packets, and average inter-arrival jitter of speech packets. The FSPAV needs to monitor no speech signal but received IP packets containing voice data from the peer user by a local end user. It costs lowly to measure the metrics due to no need for clock synchronization or parsing application protocols. A measure of the individual truth grad is used to map the measurement values of the three metrics on a call segment into a single speech quality prediction. Besides, during the calculation, the goodness grad for each flow feature that makes for factor analysis of quality degradation can also be derived. To verify the effectiveness of the method, the popular VoIP software QQ and Skype are used to create VoIP talks, our VoIP PQoS measurement tool implementing the FSPAV performs traffic analysis on a flow basis and the objective speech quality assessment, and talk users provide their subjective evaluation of speech quality. The experimental results show that the objective assessments correlate well with subjective assessments since the correlation coefficient between the predicted scores and the subjective scores is up to 0.9677. The results indicate that the proposed nonintrusive, objective quality assessment method for VoIP speech performs well.

Keywords: VoIP; objective assessment of speech quality; flow feature; measure of truth grad

1 引言

语音质量评价可分为主观评价和客观评价两大类[1-8]。主观评价方法有绝对等级评定(ACR)、比较等级评定(CCR)等，主观质量可分为**收听质量**和**会话质量**[3,4]**。**语音质量评定结果常用平均意见得分(Mean Opinion Score, MOS)表示[4]。主观评价最能够准确测量用户感知(指感知服务质量,PQoS)，但需要多人参加，准备和执行既费时又费力。客观评价指用机器自动(非人工)地预测主观质量测试结果，相对容易执行且节约时间。客观评价可分为侵入式(intrusive)和非侵入(nonintrusive)式两种方式[1,5-8]。侵入式客观评价要求原始的输入信号和失真的输出信号都可用，以原始信号为参考，通过比较二者之间的失真预测语音质量，也称基于输入-输出的方法。非侵入式方法不需要原始信号作参考，仅根据输出信号客观估计语音质量，也称基于输出的方法。根据估计失真的域，语音质量的客观评价可分为时域、频域、感知域三类[8]，其中基于人类听觉系统的心理声学模型的感知域测量最能预测感知语音质

[1]本课题得到国家 973 重点基础研究发展规划项目基金(No. 2003CB314804)和南京邮电大学攀登项目(NY206010)的支持和资助。

量。文献[6]根据目标和测量规程等将语音质量客观评价分为**意见模型、基于语音信号的客观模型、基于分组的客观模型**三类。ITU-T 定义了 2 个可估计普通用户会话质量的意见模型：E-model 和 CCI[6]。用 E-model(1998)预测感知语音质量仅适用于提供经电话听筒的窄带(3.1kHz)实时电话的端到端连接，并且 E-model 仅建议作为一个传输规划工具，并不适合精确估算主观质量[9]。CCI 模型使用 INMD(运行中非侵入测量装置)[10]测量一些话音级参数预测普通用户的会话质量，在这方面比规划模型 E-model 更健壮些[11]。但它们(即使 2005 版本的 E-model)都仅适用于 PSTN(公共交换电话网)窄带电话连接的会话质量预测，其中即使涉及到 IP 网络，IP 网络也只是充当传输元素[9]，或者假设 IP 损伤可以忽略[11]。基于语音信号的客观模型的典型代表是 ITU-T P.862 中提出的 PESQ[12]和 P.563 中提出的单端非侵入方法。PESQ(2001)是一种侵入式的属于感知域的客观模型。侵入评估技术需要访问网络的两端并实现输入输出信号之间的同步，在很多情形下都不可行。P.563(2004)方法无需参考信号，仅根据接收到的受损信号分析失真，与 PESQ 相同的是都适于预测窄带电话收听质量[13][14]。

长期以来，基于语音信号特征和人类听觉系统特征，ITU-T 及学术界主要研究采用时分复用(TDM)技术的 PSTN 中窄带电话应用语音质量的客观评价[5]-[18]。VoIP(本文专指主机到主机的 Voice over IP)是一种在采用分组交换技术的互联网上提供的对传统电话极具冲击力的新型语音通信服务，随着它的大力发展，对于其感知服务质量的研究已逐渐成为研究热点[10][19]-[25]。在分组网络上传输语音，带来了新形式的质量降级，因此 ITU-T SG 12 近几年在致力于包括 VoIP 感知质量在内的有关 IP 的性能/QoS 测量研究。P.VTQ(2004,草案)是 ITU-T 提出的第 1 个完全基于 IP 分组信息而不是负载中的语音的客观模型，用于实时监测并估计在 IP 分组网上传输的语音的收听质量[6]。P.VTQ 使用了可从 RTP、RTCP、RTCP-XR[19]分组得到的分组丢失率、分组丢失模式和时延抖动等质量参数[6]。G.1020(2006)定义了一组能较好反映感知语音质量的分组网络和终端性能参数，包括时延、时延抖动、丢包、编码等相关参数，其中的丢包相关参数必须依赖对 RTCP-XR 等分组的解析[20]。目前有关 VoIP PQoS 评估的研究存在一些问题：定义的参数比较精细，适于 VoIP 软件的优化，而不是用户主观意见的预测；不少参数的测量依赖于对 RTP/RTCP/RTCP-XR 协议分组的解析，而这在现实中常常不可行，因为流行的 VoIP 软件都是 P2P 的，协议往往不公开，例如 Skype[21]；未能充分利用 VoIP 流量特征和已有的 IP 网络测量研究成果。为此，本文基于对 VoIP 流量特征的分析，已有的 VoIP PQoS 相关研究成果和已有的 IP 流测量技术[26]，提出了一个全新的基于流特性的 VoIP PQoS 单端客观评价方法——FSPAV(Flow features-based Single-ended PQoS Assessment of VoIP)。其中最关键的是提出了三个能够较好反映用户感知，又无需协议解析且可以单点测量的 IP 流特性测度，可大大降低测量成本并提高可扩展性；另一独特之处是采用中介真值程度度量方法[27][28]将测得的多个测度值映射成一个通话质量(指会话质量)客观评价值，该方法与经常使用的回归方法相比，优点是能够结合专业知识得到各测度性能的单独评价，有助于语音质量降级因素分析。

2 中介真值程度

自然界中的事物并不都是非此即彼，而是还存在着大量“亦此亦彼”的现象，如半导体可看作是由此(导体)至彼(绝缘体)或由彼至此的中介过渡。鉴于此，20 世纪 80 年代学者朱梧槚和肖奚安提出了中介原则，并以自创的中介逻辑演算系统作为推理工具，建立了以中介公理集合论为主要内容的中介数学系统。中介数学系统贯彻始终的中介原则是：存在着谓词

P 和对象 x，使得 $P(x)$和$\neg P(x)$（$\neg$ 为反对对立否定词）都部分地为真。

为了使中介数学广泛应用于科学研究和工程实践，文献[27][28]提出了中介真值程度的数值化度量，并引入了超态概念，用符号+表示“更”，例如^{+}P 表示比 P 更 P，如此，有相应的真值集合 $Truth$={$^{+}T, T, M, F, ^{+}F$}。为方便理解，介绍其中的一些基本定义。

定义 1[27]　给定非空对象集合 X，称任何一个映射 f：$X\to\boldsymbol{R}^n$ 是对象集合 X 的(n 维)数值化映射。

定义 2[28]　若对于一维数值化映射 f，存在映射 Q：$\boldsymbol{R}\to Truth$，使得谓词 $P=Q\circ f$，即对任何 $x\in X$，$P(x)=Q(f(x))$，则称 P 是与数值化映射 f 对应的谓词。

定义 3[27]　f 是非空对象集合 X 的一维数值化映射，即 f：$X\to\boldsymbol{R}$，与谓词 P 的真值对应的数值区域是闭区间$[a-e,a+e]$，则称a为 P 的e标准度。设谓词 P or ~P or $\neg P$ or $\neg\,^{+}P$ or ^{+}P 成立，在 P 的“真数值区域”，a_T是 P 的e_T标准度；在 P 的“假数值区域”，a_F是$\neg P$ 的 e_F标准度。若$a_F<a_T$，称 P 是正谓词；若$a_T<a_F$，称 P 是负谓词。

设 P 是正谓词，则定义 3 中所刻画的数值区域与谓词的真值对应关系如图 1 所示。

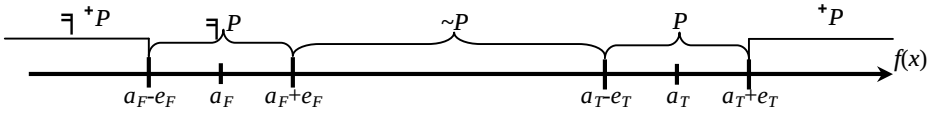

图 1 数值区域与谓词的对应关系[27]

定义 4[27]　集合 X={$x_1, x_2,\ldots, x_n,$}，设 f：$X\to\boldsymbol{R}$，则称 $g_T=x_T\circ h_T\circ f$：$X\to\boldsymbol{R}$ 为相对于 P 的真值程度函数，其中的 h_T为相对于 P 的距离比率函数。

f 为 X 的 n 维数值化映射时的个体真值程度度量请参见文献[27]。

3　语音 IP 流的流特性测度

文献[23]分析了 IP 网络上语音质量降级的各种原因：采样、数字化、编/解码、分组丢失和时延抖动，发现其中时延抖动引起的语音信号的失真和分组丢失引起的部分信号丢失对语音质量影响最大。此外还发现若没有设置去抖动缓存，则时延抖动对语音质量有破坏性的影响；如果设置缓存，则时延抖动带来的降级类似于由于分组丢失引起的降级。ITU-T P.561 提到了单点 IP 时延变化、IP 分组丢失率以及 IP 往返时延、IP 分组到达失序统计、丢失分布、话音编码的类型与配置、IP 分组中发送的数据量、分组到达描述信息(RTP 序号等)等 VoIP 话音参数 [10]。基于以上研究成果，以及我们对 VoIP 的 IP 流特性的研究，本节将定义 3 个与用户感知的 VoIP 服务质量相关性较高的语音流的流特性测度。这 3 个测度，均是单点测度，不要求测量两点测度所必需的时钟同步；并且不依赖协议解析获得测度值，可突破很多文献并不是很现实的假定：VoIP 基于 RTP/UDP/IP 协议[10][20][24][25]；因而测量成本低且扩展性较好。下面先解决语音流的识别问题。

3.1 语音流的识别

定义 5　通过一个观测点并具有一组共同属性且满足流超时时间约束的 IP 分组的集合称为一个 IP 流(flow[26])，本文设流标识为 5 元组(local IP, local port, remote IP, remote port, protocol)，其中(local IP, local port, protocol)表示本地端点，(remote IP, remote port, protocol)表示远端端点。protocol 为 IP 分组头中协议字段的值，值为 6 和 17 的 IP 流分别称为 TCP

流和 UDP 流。

这样，一个 IP 流由观测点观测到的往返于特定本地和远端端点之间的 IP 分组序列组成，并且其中任意相邻到达的分组通过观测点的时间间隔小于流超时时间。新 IP 流的状态为**活跃的**，该状态一直不变，直到超过流超时时间没有观测到属于该流的新分组时，该流状态由活跃的变为**终止的**。新观测到的分组若不属于已有的某个活跃 IP 流，则属于一个新 IP 流。

定义 6 称传送话音数据的 IP 流为语音流。

本文采用扩展性好、适应性强的基于非内容特征的方法识别语音流。鉴于 VoIP 基本上都基于 UDP 协议，因此本文仅考虑 UDP 流。为了能够从所有 UDP 流中识别出语音流，一条**流记录**[26]除关键字段外，还需包括流起始时间、最近一个 IP 分组的到达时间、本地到远端的 IP 分组数及字节数(IP 分组总长度之和)、远端到本地的 IP 分组数及字节数等流属性；对于语音流，为预测用户感知，还需包括下小节定义的 3 个语音质量相关的流特性测度。

目前国内常用的 VoIP 工具是 QQ 和 Skype，它们均采用 P2P 技术支持 VoIP，一对用户通话一般基于一个 UDP 流。在 VoIP 中，语音信号是通过定时采样、数字化、压缩编码、装入 UDP 报文，再封装在 IP 包中发向接收方的。因此一个语音流的主要特征是持续时间较长，立足于某一 VoIP 端主机观测 VoIP 流量，可以观测到发送方向上(本地到远端)有以相对固定间隔[2]发送的载有语音信息的 IP 分组序列，而接收方向上为经过网络传送后的到达间隔与发送间隔相比有所变形的 IP 分组序列。

基于对大量实际 VoIP 流量的监测与分析，本文提出一个简单有效，适用于非抽样监测，观测点可设在被监测主机或主机所在端网络处的语音流识别算法。针对目前流行的各版本的 QQ 和 Skype，已做过大量测试，测试结果是：本识别方法准确率为 100%，语音流一般持续不到 10s 即可被成功识别[3]。

语音流识别算法（将另行文详述） 对被监测本地主机收发的 UDP 分组按定义 5 进行组流，对每个分组或者生成一个新流记录并初始化流属性值，或者仅更新流属性值，再适时(例如每隔 500*ms*)根据判别准则 1 识别语音流。

判别准则 1 UDP 流为语音流的充分条件是：该流是活跃的，且每个方向至少已有 100 个 IP 包，且每个方向平均每秒至少 10 个包，且每个方向平均 IP 包大小小于 300 字节。

3.2 三个应用层测度

为实现基于输出(即基于接收到的语音包)的 VoIP PQoS 客观评价，本小节定义 3 个与通话质量密切相关又容易测量的应用层流特性测度。鉴于语音流一般持续时间较长，为降低测量代价，规定在语音流被识别后第 3 个语音包开始计算 3 个测度的值，以后每隔一定时间间隔（通话质量测量间隔 $\boldsymbol{T}$ s）输出前一时间段($\boldsymbol{T}$ s) 的 3 个测度值，并准备开始新一轮的计算。

3.2.1 平均语音包大小

平均语音包大小(*avgAppPktLen*,单位:byte)定义为语音流中远端到本地方向的各个含有语音数据的 IP 分组(简称语音包)所包含的应用层协议数据单元的平均大小。在每个 $\boldsymbol{T}$ s 间隔内对属于给定语音流的每个新到达语音包按下式计算最新平均语音包大小：

[2] Skype 不支持静音抑制[21]，据观察 QQ 也是如此。

[3] 对于故意仿 VoIP 伪造的流量不能保证正确。

$$avgAppPktLen^{(k)} = ((k-1)/k)avgAppPktLen^{(k-1)} + AppPktLen^{(k)}/k,\quad k = 1,2,3,... \qquad (1)$$

其中，$AppPktLen^{(k)}$是新到达语音包(该语音流本间隔内第 k 个语音包) 中 UDP 有效载荷的字节数，$avgAppPktLen^{(k)}$为该语音流本间隔当前平均语音包大小，$avgAppPktLen^{(k-1)}$为新语音包到达前的平均语音包大小。

3.2.2 语音包平均到达间隔

语音包平均到达间隔(*avgArrivalInterval*,单位:ms)定义为语音流中远端到本地方向相邻语音包的平均到达时间差。在每个 T s 间隔内对属于给定语音流的每个新到达语音包按下式计算最新语音包平均到达间隔：

$$avgArrivalInterval^{(k)} = ((k-1)/k)avgArrivalInterval^{(k-1)} + ArrInterval^{(k)}/k,\quad k = 1,2,3,... \qquad (2)$$

其中，$ArrInterval^{(k)}$为该语音流新到达语音包(本间隔内第 k 个语音包)与前一个语音包的到达时间之差，$avgArrivalInterval^{(k)}$为该语音流本间隔当前语音包平均到达间隔，$avgArrivalInterval^{(k-1)}$为新语音包到达前的语音包平均到达间隔。

3.2.3 语音包平均到达间隔抖动

语音包平均到达间隔抖动(*avgArrivalJitter*,单位:ms)定义为语音流中远端到本地方向相邻 3 个语音包中前 2 个语音包到达间隔与后 2 个语音包到达间隔之绝对差值的平均值。在每个间隔内对属于给定语音流的每个新到达语音包按下式计算最新语音包平均到达间隔抖动：

$$avgArrivalJitter^{(k)} = \frac{(k-1)avgArrivalJitter^{(k-1)}}{k} + \frac{\left|ArrInterval^{(k)} - ArrInterval^{(k-1)}\right|}{k},\ k = 1,2,3,... \qquad (3)$$

其中，$ArrInterval^{(k)}$为给定语音流新到达语音包(本间隔内第 k 个语音包)与前一个语音包的到达时间差，$ArrInterval^{(k-1)}$为新到达语音包的前一个语音包和再前一个语音包的到达时间差，$avgArrivalJitter^{(k)}$为该语音流本间隔当前语音包平均到达间隔抖动，$avgArrivalJitter^{(k-1)}$为新语音包到达前的语音包平均到达间隔抖动。

以上测度中，平均语音包大小和语音包平均到达间隔与所用的语音压缩编码标准有关；语音包平均到达间隔和平均到达间隔抖动两测度受网络服务质量影响，两测度与丢包率成正比，后者还与 IP 包传送时延和时延变化成正比。此外，经过统计分析我们发现语音包平均到达间隔抖动与语音质量相关性最高，其次是语音包平均到达间隔，再次是平均语音包大小。

4 基于中介真值程度度量的语音质量评价

基于上节提出的语音 IP 流特性测度和文献[27]提出的中介真值程度度量，本文提出一种新颖的 VoIP 会话质量单端客观评价方法——FSPAV。具体方法描述如下。

任意点到点的 VoIP 通话可划分成一个个 T 秒（亦即 3.2 节中的通话质量测量间隔）为单位时长的通话片段。设所有时长为 T 秒的 peer-to-peer VoIP 通话片段(个体)构成论域 D。例如特定用户 a 和 b 之间的 15*T 秒时长的通话，每个用户可将其看作包括 15 个个体。

记 X 是某特定用户(记用户 A)与另一用户一次通话期间产生的到 A 方向的所有通话片段的集合。每个通话片段的平均语音包大小、语音包平均到达间隔、语音包平均到达间隔抖动

分别由 f_1, f_2, f_3 映射表示，即 $f_i : X \to R(i = 1,2,3)$。设通话片段会话质量的**数值化映射为**

$$f : X \to R^3, \text{且有} f(x) = (f_1(x), f_2(x), f_3(x)) \tag{4}$$

显然，$f_i(x)$是 $f(x)$的分支映射，也即通话片段的会话质量由平均语音包大小、语音包平均到达间隔、语音包平均到达间隔抖动三个因素表示。

记谓词 $P(*)$表示“$*$是好的”，即用户感觉到通话片段$*$的质量是好的。若 f 的每个分支映射 f_i 都存在对应的谓词 P_i $(i = 1,2,3)$，使得

$$P(x)=\mathrm{G}(P_1(x),\ P_2(x),\ P_3(x)) \tag{5}$$

其中 $G : Truth^3 \to Truth$ 是一个 3 元命题联结词，则 P_i 是 P 的关于 f 的分支谓词。在分支谓词中，我们约定 $P_1(x)$: x 的平均语音包大小特性是好的(大为好)，$P_2(x)$: x 的语音包平均到达间隔特性是好的(小为好)，$P_3(x)$: x 的语音包平均到达间隔抖动特性是好的(小为好)。

设谓词 P 的真数值区域是 $n(n=3)$维超立方体$[a_{\mathbf{T}}\text{-}e_{\mathbf{T}},\ a_{\mathbf{T}}+e_{\mathbf{T}}]$，假数值区域是 n 维超立方体$[a_{\mathbf{F}}\text{-}e_{\mathbf{F}},\ a_{\mathbf{F}}+e_{\mathbf{F}}]$，其中 P 的$e_{\mathbf{T}}$标准度向量$a_{\mathbf{T}}=(a_{1\mathrm{T}}, a_{2\mathrm{T}}, a_{3\mathrm{T}})$，$e_{\mathbf{T}}=(e_{1\mathrm{T}}, e_{2\mathrm{T}}, e_{3\mathrm{T}})$，$\neg P$ 的$e_{\mathbf{F}}$标准度向量$a_{\mathbf{F}}=(a_{1\mathrm{F}}, a_{2\mathrm{F}}, a_{3\mathrm{F}})$，$e_{\mathbf{F}}=(e_{1\mathrm{F}}, e_{2\mathrm{F}}, e_{3\mathrm{F}})$。根据语义易知分支谓词 P_1 是正谓词(即$a_{1\mathrm{F}}<a_{1\mathrm{T}}$)，$P_2$ 和 P_3 是负谓词(即$a_{2\mathrm{F}}>a_{2\mathrm{T}}$, $a_{3\mathrm{F}}>a_{3\mathrm{T}}$)。

定义 7　对于正谓词 P_k，通话片段 x_i 相对于 P_k 的距离比率函数 h_T: $f_k(X)\to\boldsymbol{R}$，当取 $y_{ik}=f_k(x_i)\in f_k(X)$时有[27]：

$$h_T(y_{ik})=\begin{cases}\dfrac{-d(y_{ik},a_{kF}-e_{kF})}{d(a_{kT}-e_{kT},a_{kF}-e_{kF})} & y_{ik}<a_{kF}-e_{kF}\\ 0 & a_{kF}-e_{kF}\le y_{ik}\le a_{kF}+e_{kF}\\ \dfrac{d(y_{ik},a_{kF}+e_{kF})}{d(a_{kT}-e_{kT},a_{kF}+e_{kF})} & a_{kF}+e_{kF}<y_{ik}<a_{kT}-e_{kT}\\ 1 & a_{kT}-e_{kT}\le y_{ik}\le a_{kT}+e_{kT}\\ \dfrac{d(y_{ik},a_{kF}+e_{kF})}{d(a_{kT}+e_{kT},a_{kF}+e_{kF})} & y_{ik}>a_{kT}+e_{kT}\end{cases} \tag{6}$$

对于负谓词 P_k，通话片段 x_i 相对于 P_k 的距离比率函数 h_T: $f_k(X)\to\boldsymbol{R}$，当取 $y_{ik}=f_k(x_i)\in f_k(X)$时有：

$$h_T(y_{ik})=\begin{cases}\dfrac{-d(y_{ik},a_{kF}+e_{kF})}{d(a_{kT}+e_{kT},a_{kF}+e_{kF})} & y_{ik}>a_{kF}+e_{kF}\\ 0 & a_{kF}-e_{kF}\le y_{ik}\le a_{kF}+e_{kF}\\ \dfrac{d(y_{ik},a_{kF}-e_{kF})}{d(a_{kT}+e_{kT},a_{kF}-e_{kF})} & a_{kF}-e_{kF}>y_{ik}>a_{kT}+e_{kT}\\ 1 & a_{kT}-e_{kT}\le y_{ik}\le a_{kT}+e_{kT}\\ \dfrac{d(y_{ik},a_{kF}-e_{kF})}{d(a_{kT}-e_{kT},a_{kF}-e_{kF})} & y_{ik}<a_{kT}-e_{kT}\end{cases} \tag{7}$$

式(6)和式(7)中的 $d(a,b)$表示 a 和 b 的欧氏距离。

定义 8　通话片段 x_i 相对于 P 的真值程度加权平均函数 $g_{nT-W}=h_{nT-W}\circ h_T\circ f : X\to\boldsymbol{R}$ 为：

$$g_{nT-W}(x_i)=h_{nT-W}(h_T(f_1(x_i)),h_T(f_2(x_i)),h_T(f_3(x_i)))=\frac{1}{w}\sum_{k=1}^{3}w_k h_T(f_k(x_i)),\ \text{其中}\sum_{k=1}^{3}w_k=w \tag{8}$$

式(8)即为本文提出的基于中介真值程度度量的 VoIP PQoS(会话质量)客观评价计算公

式，“质量好”的真值程度域为(-∞,+∞)，评价值 1 表示质量好，大于 1 表示比好还好，0 表示质量差，小于 0 表示比差更差。

下面采用层次分析法(AHP)[29]确定式(8)中的权值，即确定平均语音包大小(B1)、语音包平均到达间隔(B2)和语音包平均到达间隔抖动(B3)3 个因素（指标层）相对于通话质量(V)（最高目标层）的重要性权重。为此，首先比较各因素之间的相对重要性，并采用 1-9 标度方法构造判断矩阵。通过进行大量的实际 VoIP 通话质量测试，定性地分析比较通话质量的主观评价值与实际通话流量的 3 个测度值之间的相关性，得到各因素之间的相对重要性，判断矩阵 A 见表 1。再按照方根法计算判断矩阵的最大特征根及其对应的特征向量。计算步骤为[29]：(1)计算判断矩阵每一行元素的乘积 M_i；(2)计算 M_i 的 n（n=3）次方根 $\overline{W_i}$；(3)对向量 $\overline{W}=[\overline{W_1},\overline{W_2},\overline{W_3}]^T$ 正规化，即 $W_i=\overline{W_i}\Big/\sum_{j=1}^{n}\overline{W_j}$，求得特征向量 $W=[W_1,W_2,W_3]^T$，表 2 列出了计算结果；(4)计算判断矩阵的最大特征根 $l_{\max}=\sum_{i=1}^{n}((AW)_i/(nW_i))$ =3.0246；(5)最后检验判断矩阵的一致性。已知 3 阶判断矩阵的平均随机一致性指标 *RI*=0.58[29]，矩阵 A 的一致性指标 *CI*=($l_{\max}$-*n*)/(n-1)=0.0123，随机一致性比率 *CR*=*CI*/*RI*=0.0212<0.1，即认为判断矩阵 A 具有满意的一致性。综上，三个因素的权重应分别设为：0.117,0.2,0.683。

表 1 判断矩阵 A

V	B1	B2	B3
B1	1	1/2	1/5
B2	2	1	1/4
B3	5	4	1

表 2 权重计算过程

M	$\overline{W}$	*W*
0.1	0.464	**0.117**
0.5	0.794	**0.2**
20	2.71	**0.683**

5 实验结果

本节通过实际测试验证 FSPAV 的效果。为此我们基于 Winpcap[30]开发了一个 VoIP PQoS 测量工具，该工具能够监听主机流量并进行组流，识别语音流，计算语音流的 3 个应用层测度，以及按照式(8)计算通话质量的客观评价值。

5.1 实验设置

本文实验利用当前流行的 VoIP 软件 QQ 2006 或 Skype v1.4.0.84 进行通话产生 VoIP 流量。测试地点分别设在以局域网方式接入某校园网的主机、通过 modem 拨号接入该校园网的主机、宽带接入南京联通的住宅主机、宽带接入淮安电信的住宅主机等处，观测点就设在测试主机上。由于客观评价实质上是对主观评价值的一种预测，因此可以通过计算主、客观评价值的相关系数来衡量客观评价方法的通话性能[1,12]。为此，实验中对每次通话除了使用自己开发的 VoIP PQoS 测量工具监测语音流量，还要同时进行主观评价。评价值 1 表示质量好，大约相当于 MOS 为 4.2，即声音清晰，通话双方响应及时，不亚于长话质量，大于 1 表示比好还好。评价值 0 表示质量差，相当于 MOS 为 1，即声音不清晰，响应严重滞后，无法交流，小于 0 表示比差更差。介于 0 和 1 之间的评价值表示质量介于好和差之间，例如 0.4 大约相当于 MOS 为 2.5。

通话质量测量间隔 ***T*** 设为 20s[10]，即对语音流每隔 20s 输出前一时间段平均语音包大小、语音包平均到达间隔和平均到达间隔抖动 3 个测度值，以及通话质量客观评价值。基于中介真值程度的通话质量评价需要设置 3 个特性好和差时测度的取值范围。基于对语音 IP 流中

语音包发出时和到达接收方时 3 个测度值的分析，将特性好差时测度的取值范围确定如下：

表 3 好差等级的测度取值范围

测 度	差	好
平均语音包大小（byte）	23-40	80-130
语音包平均到达间隔（ms）	110-90	30-20
语音包平均到达间隔抖动（ms）	90-75	8-0

采用相对于好的真值程度进行衡量，根据表 3 有：a_T-e_T = (80, 20,0)，a_T+e_T=(130,30,8)，a_F-e_F=(23,90,75)，a_F+e_F=(40,110,90)，相应的有 $\boldsymbol{d}(a_F+e_F, a_T+e_T)$ =(90,80,82)。

5.2 实验结果

表 4 列出了 6 组实验结果，每组包含一次通话中连续的 10 个通话片段的测试结果，其中每行分别是各通话片段(20s 时长)的 3 个测度值、PQoS 主观评价值、按照式(8)计算得到的 PQoS 客观评价值。主客观评价值的对比还可参见图 2。

表 4 基于中介真值程度度量的语音质量客观评价实例

平均语音包大小(byte)	语音包平均到达间隔(ms)	平均到达间隔抖动(ms)	PQoS 主观评价	PQoS 客观评价
145	54.6	28.28	0.8	0.731
64	29.4	29.8		0.732
88	37.5	29.82		0.752
155	57.6	23.86		0.779
156	58.3	23.62		0.78
157	57.9	24.14		0.778
138	50.3	23.67		0.783
65	29.2	30.12		0.729
64	29.5	30.18		0.728
70	32.1	30.6		0.734
159	57	13.15	0.9	0.895
121	44.7	7.56		0.951
68	30.7	5.88		0.963
148	57.9	13.8		0.871
159	58.3	11.51		0.907
147	58.7	14.26		0.863
132	54.9	13.73		0.862
64	29.5	5.5		0.953
111	46.8	9.92		0.924
64	29.7	6.04		0.954
83	59.9	47.83	0.4	0.494
78	61.9	57.57		0.383
83	60.5	50.12		0.469
82	62.3	54.18		0.422
84	63.5	58.79		0.37
83	67.8	63.34		0.31
82	60.1	50.79		0.464
83	69.6	61.94		0.318
82	61.8	51.02		0.456
83	65.8	59.46		0.356

平均语音包大小(byte)	语音包平均到达间隔(ms)	平均到达间隔抖动(ms)	PQoS 主观评价	PQoS 客观评价
136	30.5	6.88	0.95	1.006
131	30.1	4.69		1.001
106	30.1	6.36		1
108	30.1	5.78		1
108	30.1	5.77		1
112	30.1	5.54		1
110	30.1	6.58		1
101	30.1	6.74		1
95	30.1	4.82		1
103	30.1	5.38		1
133	30.4	4.81	1	1.002
140	30.4	4.08		1.012
190	30.4	4.8		1.077
130	30.1	4.22		1
130	30.4	4.84		0.999
129	31.3	6.65		0.996
266	50.5	7.85		1.108
254	60.9	12.47		1.013
238	59.8	8.62		1.035
216	59.6	10		0.993
155	58.4	8.43	0.85	0.933
140	40.6	5.93		0.977
125	30.4	5.03		0.999
141	37.9	5.58		0.988
139	58.8	8.09		0.915
140	58.4	7.71		0.918
138	58.9	9.51		0.899
140	58.4	6.9		0.919
135	58.8	9.37		0.897
137	57.9	8.05		0.916

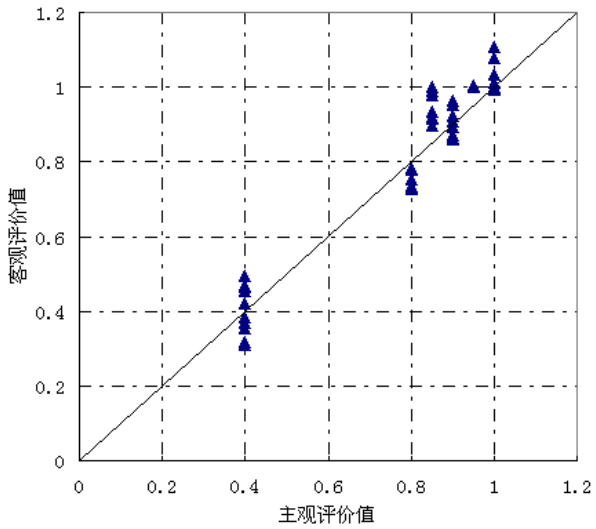

图 2 主观评价值与基于中介真值程度度量的客观评价值

对表 4 中的观察数据（前 4 列）也可以建立多元线性回归模型，利用 MATLAB 解得回归系数的最小二乘估计值，得到回归方程：

$$\hat{y} = 0.9823 + 0.0012x_1 - 0.0039x_2 - 0.0072x_3 \tag{9}$$

通过检验发现回归方程在水平 α =0.01 下有显著意义，即回归效果高度显著。

对于本实例，基于线性回归及基于中介真值程度度量所得到的语音质量客观评价值和主观评价值之间的相关系数分别达到了 0.9685 和 0.9677，由此说明本文定义的 3 个测度能够反映 VoIP PQoS。表 5 总结了两种映射方法的主要区别。

表 5 两种映射方法的对比

对比的方面	基于线性回归方法	基于中介真值程度度量方法
先验知识	不需要	需要(需要先确定相关测度好、差等级的取值范围)
语音质量降级因素分析	不支持	支持(能够给出每个特性的优劣程度)

6 相关工作比较

ITU-T 对语音质量的客观测量做了大量的研究。ITU-T P.862[12]中提出的 PESQ 是一种用于预测窄带电话网络和窄带语音编码的端到端语音质量客观评价方法，PESQ 算法仅测量单向语音失真和噪声对语音质量的影响，并没有提供对传输质量的综合评价，响度损耗、时延、侧音、回声及其他与双向交互有关的损伤的影响没有反映在 PESQ 得分中，因此不能预测会话质量。PESQ 首先要将输入信号(原始)和输出(受损)信号进行同步对齐，再使用感知模型将输入和输出信号分别转换为人类听觉感知表示，然后比较二者之间的距离，再经由认知模型计算得到客观**收听**质量 MOS。PESQ 不支持服务在线测量。ITU-T P.563(2004 年)中提出的语音质量评价模型是第 1 个被 ITU 接受并标准化的单端模型。该模型基于话音产生和感知模型，首先将降级语音信号进行预处理，再利用语音重构算法重构出准原始信号，并经由基于输入输出的感知模型估计失真，再经由感知映射得到收听语音质量的客观估计[13]。

ITU-T G.107 提出了一个以网络传输规划为目的的估计感知语音质量的计算模型：E-model，它将不同的传输参数转变为不同的损伤因素，使规划者获得预期的普通用户所感知的语音传输质量的估计[9]，其中损伤因素包括与话音信号同时产生的损伤、时延损伤和设

备损伤，传输参数有发送/接收响度额定值(SLR/RLR)、发话人回声响度额定值(TELR)、回声通道时延、量化噪声、室内噪声、信息包丢失概率等。E-model 中使用的部分参数 SLR、噪声、TELR 和回声通道时延可以由 INMD 测得的参数映射得到[11]。ITU-T P.562 中描述的预测普通用户意见的意见模型 CCI，使用 INMD 测得的话音级参数，如话音电平、噪声电平、回声损耗和回声通道时延，结合对有关网络和两端用户的假设，预测到达每个用户耳朵的信号；然后将这些预计的信号，根据人类听觉系统知识，变换为用户感知到的本呼叫收听和会话质量的主观意见预测。

以上 ITU-T 的 4 种方法都可以用于预测用户对语音质量的主观意见，但 PESQ 和 P.563 只考虑收听质量，而 E 模型目的是用于网络规划，不太适合具体评测。最主要的是它们全都设计用于评定 PSTN 网络中 3.1kHz 窄带电话连接的语音质量，不适于主机到主机的 VoIP 会话质量的测量。

学术界对 VoIP 语音质量做了一些相关研究。例如以分析网络性能参数对端到端感知语音质量的影响为主旨的文献[22]和[23]。[22]利用分组丢失仿真程序产生网络损伤，遵照 P.800 进行语音质量主观评价实验，对男声和女声的实验数据分别进行二维线性回归，再对结果进行平均得到一个 MOS 预测模型：$Predicted\ MOS = 4.0 - 0.7ln(loss) - 0.1ln((M - hsize) / drate)$，其中 *loss* 为分组丢失率，*M* 为 IP 分组大小，*hsize* 为 RTP/UDP/IP 首部大小，*drate* 为编码数据率。[23]专门搭建了一个测试环境，在收发主机之间设置一台配置成路由器并能够产生指定的网络损伤的 Linux PC，并利用一个通过比较输入和输出模拟语音信号之间的差异来预测语音质量的数字话音电平分析器(DSLA)工具，以及一个 MOS 预测模型来分析采样、编解码、分组丢失、时延抖动对语音质量的影响，MOS 预测模型为：$MOS_{pred} = MOS_{opt} - C*ln(loss+1) - D*ln(size/size_{min})$，其中 MOS_{opt} 为某种编码在没有任何损伤时的最佳 MOS，*loss* 是分组丢失率，*C* 和 *D* 是常系数，对分组丢失较敏感的编码，*C* 大约为 0.4，对丢失质量降级较小的编码 *C* 约为 0.25，*size* 为使用的分组大小(以 ms 为单位)，$size_{min}$ 为某种编码可使用的最小分组大小(一般为 10 ms)，*D* 一般约为 0.15。[22]和[23]中的丢失率不容易测量，测量成本较高，无论是通过在通话双方分别设置测量点进行测量，还是通过协议解析得到。此外，Lingfen Sun 等在[24]中提出采用非线性回归模型，以网络参数分组丢失率和端到端时延作为自变量，预测**会话**语音质量，其中参数值通过分析 RTP 包得到，并对每一种编码分别建立回归模型。[25]提出了一种基于语音信号频率特性和网络传输条件的评价方法，其中网络传输参数往返时延和抖动均依赖解析 RTCP 包获得。

与上述方法、P.VTQ 以及 G.1020 相比，本文定义的测度均是单点测度，而且较容易测量。因为既不需要时钟同步，对时钟频率的精准度要求也不高，又不需要协议解析，前者极大降低了测量成本，后者也很重要，因为并不是所有 VoIP 系统都采用 RTP/UDP 模式传送话音数据，VoIP 一般都是基于 P2P 的，完全可以基于自定义的应用协议格式(例如 RTP 的变种)，甚至采用加密方式传送话音数据。

7 结论

在深入研究已有的语音质量评价方法基础上，针对互联网上 VoIP 应用的特点，以及已有 VoIP 语音质量客观评价研究中过于依赖 RTP/RTCP 协议解析的局限性，本文提出了一个基于语音流流(flow)特性和中介真值程度模型的 VoIP 会话质量客观评价方法——FSPAV，其

中定义了 3 个流特性测度。该方法属于非侵入单端点测量方法，仅根据接收到的语音包预测会话质量。3 个测度的测量成本低廉，不需要监测语音信号，不需要时钟同步，也不需要解析应用协议。基于这 3 个测度和中介真值程度度量的 PQoS 评价也很容易计算，而且还能得出每个特性的优劣程度，有助于语音质量降级因素分析。实验结果显示本文方法得到的客观评价值与主观评价值具有相当高的相关系数(0.9677)，表明 FSPAV 效果不错。不过，显然 FSPAV 不能够反映语音设备导致的 PQoS 降级，如果依据 FSPAV 得出的 PQoS 评价与实际用户主观感知不符，则很有可能是语音设备设置有问题，而网络和端主机提供了较好的语音传输服务，因此用户借助本文的 PQoS 预测功能还有助于及时发现语音设备问题或故障。

本文实验使用了当今最流行的 VoIP 软件 QQ 和 Skype，它们都采用 GIPS 公司推出的某种音频编码器：iSAC, iLBC, iPCM[21][31][32]，不同编码情况下流特性好差时测度如何取值、甚至是否还需要补充其他测度值得进一步研究。测量点位置和接收端主机行为对本文方法适用性的影响也有待进一步分析。单向时延(包括网络和终端中的时延)是影响通话质量的重要因素。单向时延从几十毫秒开始造成语音质量降级并逐渐加重，会引起回声甚至声耦合；达到几百毫秒时能觉察到通话对方响应的延迟，影响交谈[33]。本文定义的语音包平均到达间隔及抖动两测度均受网络时延、终端时延的影响，但本文没有给出理论或严格的实验证明，这也值得进一步研究。

参考文献：

[1] 陈国,胡修林,张蕴玉,朱耀庭. 语音质量客观评价方法研究进展[J]. 电子学报, 2001, 29(4): 548-552.

[2] 吴淑珍,赵朝阳. 基于听觉模型的客观音质评价方法研究[J]. 电子学报, 1999, 27(7): 92-94.

[3] Methods for subjective determination of transmission quality. ITU-T Rec. P.800, 1996.

[4] Mean Opinion Score (MOS) terminology. ITU-T Rec. P.800.1, 2003.

[5] Grancharov V., Zhao D.Y., Lindblom J., Kleijn W.B. Low-Complexity, Nonintrusive Speech Quality Assessment [J]. IEEE transactions on audio, speech, and language processing, Vol. 14(6), 2006, Page(s): 1948-1956.

[6] Takahashi A., Yoshino H., Kitawaki N. Perceptual QoS assessment technologies for VoIP [J]. IEEE Communications Magazine, Volume 42, Issue 7, July 2004 Page(s):28-34.

[7] Rix A.W. Perceptual speech quality assessment - a review [C]. ICASSP '04. Volume 3, 17-21 May 2004 Page(s):iii - 1056-9 vol.3.

[8] Branko Somek, Josip Herceg, Mladen Maletic. Speech quality assessment [C]. 46th International Symposium Electronics in Marine. 2004, Page(s):307-312.

[9] Application of the E-model: A planning guide. ITU-T Rec. G.108, 1999.

[10] In-service non-intrusive measurement device Voice service measurements. ITU-T Rec. P.561, 2002.

[11] Analysis and interpretation of INMD voice service measurements. ITU-T Rec. P.562, 2004.

[12] Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. ITU-T Rec. P.862, 2001.

[13] Malfait L., Berger J., Kastner M. P.563—The ITU-T Standard for Single-Ended Speech Quality Assessment [J]. IEEE Transactions on Audio, Speech and Language Processing, Volume 14, Issue 6, Nov. 2006 Page(s): 1924-1934.

[14] Single ended method for objective speech quality assessment in narrow-band telephony applications. ITU-T Rec. P.563, 2004.

[15] Picovici D., Mahdi A.E. Output-based objective speech quality measure using self-organizing map [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).

Volume 1, 6-10 April 2003 Page(s):I-476 - I-479.

[16] Doh-Suk Kim, Tarraf A. Perceptual model for non-intrusive speech quality assessment [C]. ICASSP '04, Volume 3, 17-21 May 2004 Page(s):III-1060- III-1063.

[17] Doh-Suk Kim, Tarraf A. Enhanced Perceptual Model For Non-Intrusive Speech Quality Assessment [C]. ICASSP 2006 Proceedings, Volume 1, Page(s):I829- I832.

[18] Dimolitsas S. Objective speech distortion measures and their relevance to speech quality assessments [J]. IEE Proceedings I Communications, Speech and Vision, Vol. 136(5), Part 1, Oct 1989 Page(s): 317-324.

[19] T. Friedman, R. Caceres, A. Clark. RTP Control Protocol Extended Reports (RTCP XR) [S]. IETF RFC 3611, 2003.

[20] Performance parameter definitions for quality of speech and other voiceband applications utilizing IP networks. ITU-T Rec. G.1020, 2006.

[21] An Analysis of the Skype Peer-to-Peer Internet Telephony Protocol [C]. IEEE Infocom 2006, 2006 Page(s): 1-11.

[22] L.Yamamoto, J. Beerends. Impact of network performance parameters on the end-to-end perceived quality. Expert ATM Traffic Symposium, 1997.

[23] B. Duysburgh , S. Vanhastel , B. Vreese , C. Petrisor and P. Demeester. On the influence of best-effort network conditions on the perceived speech quality of VoIP connections [C]. Proc. of ICCCN'01, 2001: 334-349.

[24] Lingfen Sun and Emmanuel C. Ifeachor. Voice Quality Prediction Models and Their Application in VoIP Networks [J]. IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 8, NO. 4, AUGUST 2006, 809-820.

[25] 张军,张德运. 分组网络环境下的实时语音质量客观评价[J]. 西安交通大学学报, 2006, 40(8): 936-939.

[26] G. Sadasivan, N. Brownlee, B. Claise, et al. Architecture for IP Flow Information Export [EB/OL]. http://www.ietf.org/internet-drafts/draft-ietf-ipfix-architecture-12.txt. 2006-9-6.

[27] 洪龙, 肖奚安, 朱梧槚. 中介真值程度的度量及其应用(I) [J]. 计算机学报, 2006, 29(12): 2186-2193.

[28] 洪龙. 中介真值程度的度量及其在计算机系统结构研究中的应用[博士学位论文]. 南京航空航天大学, 2006.

[29] 赵焕臣,许树柏,和金生. 层次分析法[M]. 科学出版社, 1986, 23-32.

[30] http://www.winpcap.org/

[31] Batu Sat and Benjamin W. Wah. Analysis and evaluation of the skype and google-talk VoIP systems[C]. IEEE international conference on MULTIMEDIA and EXPO., 2006.

[32] Duric, A. and S. Andersen. Real-time Transport Protocol (RTP) Payload Format for Internet Low Bit Rate Codec (iLBC) Speech, RFC 3952, December 2004.

[33] ITU-T Rec. G. 1010. End-user multimedia QoS categories [S]. Nov 2001.

文章项目资助：本课题得到国家973重点基础研究发展规划项目基金(No. 2003CB314804)和南京邮电大学攀登项目(NY206010)的支持和资助。

文章主要创新说明：

（1） 提出了三个与 VoIP 通话质量相关性较高的流(flow)特性测度：平均语音包大小、语音包平均到达间隔、平均到达间隔抖动，请见 3.2 节和第 6 节最后一段。

（2） 提出了一个仅利用 IP 网络测量技术，基于语音流(通常是 UDP 流)流特性测度和中介真值程度度量的，非侵入单端的 VoIP 通话质量客观评价方法——FSPAV，实验表明该方法相当有效，客观评价值与主观评价值之间的相关系数高达 0.9677，请见第 4 和第 5 节。

（3） 提出了一个简单有效的语音流识别方法，请见 3.1 节。