计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
42(12):2149~2154, 2005

一种自适应 IP 语音缓冲算法的研究与应用

苟先太　金炜东　靳　蕃
(西南交通大学电气工程学院　成都　610031)
(gouxiantai@sohu.com)

An Adaptive Jitter Buffering Algorithm for Voice over IP Networks

Gou Xiantai, Jin Weidong, and Jin Fan
(*School of Electrical Engineering, Southwest Jiaotong University, Chengdu* 610031)

Abstract The continuous playout of voice packets in the presence of variable network delays is often achieved by buffering the received voice packets for sufficient time. Basic jitter buffering algorithms can work well only when the delay does not spike in the IP networks. In this work, an adaptive jitter buffering algorithm based on the detecting and studying the spike status of the networks, is presented to promote the quality of voice communication. It timely adjusts the minimal and maximal depth of buffer queue according to the control target of end-to-end delay and packet loss rate. The algorithm can much more easily achieve the continuous playout because it plays voice packet at a fixed inter-play time in the most time of a talk-spurt. The control target of packet loss rate can be extended to 20%. However, the basic algorithms can only bear 5%~10% of the packet loss rate. Perceptual evaluation of speech quality (PESQ) is applied to assess the speech quality in the simulation. It is shown that the algorithm can obviously promote the quality of voice communication in IP networks with spike delay. The practical application in voice gateway can also prove the effects of voice quality promotion.

Key words voice quality; spike of delay jitter; end-to-end delay; packet losses; buffering; PESQ

摘　要　当 IP 语音包的网络时延抖动较小时，一般的语音缓冲算法可以得到较好的语音质量. 当网络中存在突发大时延时，就会出现极大丢包率或极大端到端时延，从而难以获得好的语音质量. 为此，提出针对突发大时延下的自适应语音缓冲算法. 通过估算网络平均时延和学习语音包经过的网络路径上的状态，来确定需要控制端到端时延大小和语音包的丢包率，动态调整 Jitter Buffer 队列的最小深度和最大深度，从而可以尽量减小语音裂缝(gap)的出现. 通过基于听觉模型的客观音质评价(PESQ)仿真计算以及在实际语音网关设备中的应用表明算法对语音通信质量有一定的改善作用.

关键词　语音质量;突发大时延;端到端时延;丢包率;缓冲;PESQ

中图法分类号　TP18;TP391;TP393

1　引　　言

IP 语音从讲话端产生到接听者听到之间存在着时延. 时延包括语音编解码时延、网络传输时延、接收端调度时延(或者叫播放时延). 其中，因为网络传输时延存在抖动，造成语音数据包不能以固定的时间间隔到达接听者的播放设备，如图 1 所示. 由于到达的时延间隔存在抖动，为了获得连续、较少裂缝(gap)的语音质量，需要语音缓冲(jitter buffer)

收稿日期:2004-06-14;修回日期:2005-01-17
基金项目:教育部高等学校骨干教师支持计划基金项目(教技司[200065])

以尽量减少时延抖动产生的影响.由于人的讲话分为话音区(talk-spurt)和静音(silence)区,通过活动语音检测(VAD)功能,可以适当压缩静音区的长度,让端到端时延减少.

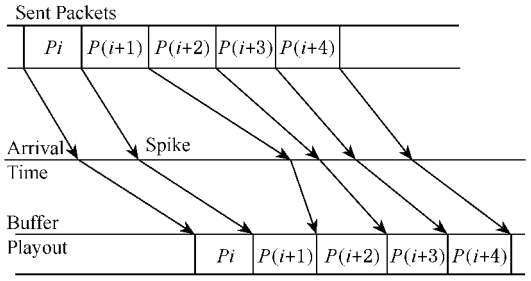

Fig. 1 IP voice packets buffering and playout.

图1 IP语音的缓冲和播放

IP语音包从发送端是以等间隔值 T_int 发出去.间隔值 T_int 为一个语音包中语音信息的编码时间长度.当两个相邻包到达的时间间隔值大于阈值时,就判定出现了突发大时延(spike delay).判断公式为

$$D\_spike(i) = A[i] - A[i-1] \geqslant SPIKE\_THEOLD.$$

网络传输时延由网络固定传播时延和网络设备中的排队时延组成.当网络负载情况(traffic load)较大时,排队时延就容易变大,从而就易产生突发大时延.文献[1]给出了网络时延概率密度和网络负载的关系,当网络负载超过网络本身负载能力的93%以上时,存在突发大时延的概率就很大.为此,本文给出了自适应IP语音缓冲算法(AAIVB).算法根据网络状态及时调整jitter buffer的队列深度和丢包率.AAIVB是面向实际设备的算法.它并不直接计算每一个包的播放时延,而是根据DSP对IP语音包进行播放的特点,通过对jitter buffer队列的控制,来控制其真正的播放时延.本文使用PESQ感知性客观音质评价方法进行了仿真计算,验证了本文的算法.

2 相关工作

现有工作[2~7],基本都是在考虑历史语音包播放时延的基础上进行过滤和预测当前到达的语音包的播放时延.文献[3,5,6]给出的仿真结果表明其在网络传输时延比较稳定(steady)的情况下有好的性能表现,但是对突发大时延的情况却没有考虑.文献[4]考虑了突发时延,但是其允许的最大端到端时延为260ms,其突发时延不大.

由于现有工作在确定算法时受到最大丢包率或最大端到端(end-to-end)时延条件限制,当突发大时延出现时,就会因单一保证较小丢包率而出现极大的端到端时延;或者单一保证端到端时延时而出现较大的丢包率.文献[8]给出分组语音允许的最大丢包率为5%.文献[9]给出电信网络中一个完美的对话(human conversation)需要满足的最大端到端时延为250ms.文献[8,9]给出了最理想情况下的选择.文献[3]讨论最大丢包率为5%时的缓冲算法,其结果显示其最大播放(调度)时延会超过1000ms.文献[7]对丢包容忍度进行了一定的放宽,允许丢包率达到10%,但是这种扩展还是不够.文献[6]讨论最大端到端时延为400ms下的缓冲算法,同样这些算法在很多实际网络环境中会受到局限.由于突发大时延客观存在,所以必须在丢包率与最大端到端时延之间寻找一个平衡点.

3 时延和丢包率分析

丢包率和端到端时延的大小直接影响交谈(conversation)的语音质量.语音质量的评定可以采用平均意见得分(MOS)、感知语音质量测度(PSQM)以及基于听觉模型的客观音质评价(PESQ)等方法.MOS的评价存在一定的主观性.PSQM是ITU在1996年推出的P.861建议,当出现丢包率、时延变化与噪音等情况时会产生不精确得分.PESQ是2001年由ITU推出的P.862建议,它更能适应通过分组网络传输的语音质量评价.这几种评价方法都把语音质量分为5级,MOS得分范围为0~5.0,PESQ的得分范围为-0.5~4.5,相当于MOS得分的0.5~4.5[10].5级音质的用户满意度为:很好、稍差、还可以、勉强、极差.

3.1 时延分析

根据ITU-T的G.114建议[11],要求语音通信单向端到端时延 $D_{e\text{-}e}$ 小于400ms,网络时延抖动小于80ms.在网络环境受局限时,往往达不到这个要求.实际上,国际长途的IP语音单向时延大大高于这个400ms.实际应用说明,适当大的时延($D_{e\text{-}e}$ > 600ms)下,只要语音平滑、连续,用户仍然可以在一定满意度范围内完成IP语音通信.

网络的端到端时延 $D_{e\text{-}e}$ 计算公式如下:

$$D_{e\text{-}e} = T\_code + T\_tran[i] + D\_shed[i], \quad (1)$$

$D\_shed[i]$ 为第 i 个语音包在jitter buffer中等待调

度的时延. T_code 为编解码消耗的时延,为一个固定值. G.729 为 20ms,G.723 为 30ms. $T\_tran[i]$ 为网络固定传播延迟和数据包的排队时延.

AAIVB 算法通过控制 jitter buffer 队列最大深度来控制最大调度时延 Max_shed,从而控制语音包的最大端到端时延 Max_delay. 最大调度时延.

$$Max\_shed = Max\_delay - T\_code - A\_trans, \quad (2)$$

A_trans 是在没有突发时延出现时,通过对历史数据的检测,而获得的网络平均传输时延. 本文采用 RFC3611 里面的 RTP 控制协议扩展报告(RTCP XR)[12] 来获得 A_trans. 文献[2~7]由于只是基于仿真模型上的研究,相比之下获得 A_trans 值要容易得多.

3.2 丢包率分析

ITU-T 的 G.114 建议[11] 要求:网络丢包率 $R\_drop<10\%$,话音质量在 3 级以上. 当出现突发大时延时,必须在时延和丢包率之间寻找一个平衡点:根据人说话的语言特点,本文将出现突发大时延时允许的最大丢包率扩大到 20%,让 MOS 得分大于 2.5. 此时的用户满意度恰好界于"勉强"和"可以"之间,称之为"勉强可以".

人对语言流中可以自然辨别出来的最小语音单位为音节[13]. 普通人的正常语速虽然对不同语言有小的差别,不过都平均在 5~7 个音节/秒. 也就是每个音节的长度在 142~200ms 之间. 如果采用 20ms 长的 G.729 的编码,一个音节的语音信息包含在 7~10个语音包里. 20% 的丢包效率意味着每个音节平均会被丢弃 1 到 2 个语音包. 由于语音信号显示出大量的短期自相似性,所以大脑可以重构音节里面丢失的音素[14]. 而丢包率高于 20% 以后,MOS 得分很快下降到 2 以后,大脑重构音丢失音素变得较为困难. 由后面的仿真和实际应用结果显示,20% 丢包率是优化的折中选择.

3.3 时延和丢包率控制策略

AAIVB 算法对丢包率 R_drop 和端到端时延 Max_delay 的基本控制策略如下:

① $Max_delay<200$ms 时,控制丢包率 5%.

② 端到端时延在 200~800ms 之间时,丢包率保持固定斜率 S1 增长;$S1=(20-5)/(800-200)=0.025$,如图 2 所示.

当 $D_{e\text{-}e}$ 超过 800ms 时,对于通话一方采用"倾听式"的方式参与还可以. 但是此时只允许再加大时延,不允许再施加任何过滤丢包.

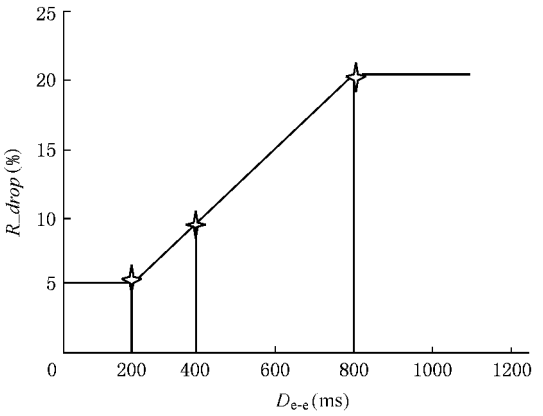

Fig. 2 The packet drop rate and delay control strategy.

图 2 AAIVB 算法的丢包率和时延控制策略

4 AAIVB 算法

4.1 接收端对 IP 语音的播放策略

播放策略为 DSP 周期性的产生中断,中断服务程序 ISP 到 jitter buffer 队列取出包送给 DSP 播放. 如果 ISP 取不到包,播放声音则会产生裂缝.

一个语音包在 jitter buffer 队列中的位置决定了它的调度延迟和整个端到端时延. 其调度延时

$$D\_shed[i] = m \times T\_int. \quad (3)$$

相邻语音包播放时间期望存在关系:

$$V_i = V_{i-1} + T\_int. \quad (4)$$

现存的消除抖动的方式很少基于 IP 语音的播放方式上进行讨论[2~7],虽然其仿真结果比较好,但语音在实际设备播放时,两个相邻语音包的播放很难满足式(4)所描述的正常情况. 结果会产生拥塞播放和 DSP 欠载(under-run),从而产生语音裂缝.所以本文采用对 jitter buffer 队列的控制来获得准确实用的语音播放.

4.2 AAIVB 算法

算法包括队列控制和播放两部分. 队列控制确定队列的最小深度 Q_min 和最大深度 Q_max. 对每一个话音区,从第 1 个包开始排队,当队列深度达到 Q_min 时,将 DSP 的播放标记设为 1,DSP 开始播放第 1 个语音包. 然后,在后续时间里,需要能够很连续地播放语音包. Q_min 为该语音区每个包增加调度延迟. Q_max 确定了语音包的最大端到端延迟. Q_min 取值太小会引起 DSP 欠载而产生语音裂缝. Q_min 取值太大,端到端的平均延迟也会被加大.

AAIVB 算法如下:

① 一路语音呼叫过程完成后,设置初始值:

$R\_drop=0$;

根据经验 $Q\_min=7$,$Q\_max=9$.

丢包个数 $N\_drop=0$;

最大抖动时延 $Max\_jit=0$;

收到该语音区包的总个数 $N\_uttrance=0$.

② 通过 RTCP XR 获得平均网络传输时延 A_trans. 然后计算 Max_delay:

$Max\_delay=A\_trans+Q\_max\times T\_int+D\_code$;

然后根据图 2 计算最大丢包率:

if($Max\_delay\leqslant 200$) then

$R\_drop=5\%$;

else

$R\_drop=(5+Max\_delay\times S1)/100$.

③ 第 i 个包到达时:

if($(A[i]-A[i-1])>SPIKE\_THEOLD$) then {

$Max\_jit=(A[i]-A[i-1])$;

$R\_drop=R\_drop+(((A[i]-A[i-1])+A\_trans+D\_code-200)\times S1)/100$;

$Q\_max=round((1-R\_drop)\times(A[i]-A[i-1])/T\_int)$;

$Max\_delay=A\_trans+(Q\_max\times T\_int)+D\_code$;

}

$N\_uttrance=N\_uttrance+1$;

if($N\_uttrance\geqslant Q\_min$)

$Playout\_flag=1$;

if $Queue\_Is\_Full()$ then

if $((N\_drop/N\_uttrance)<R\_drop)$ {

$flag=drop\_at\_rate(R\_drop)$;

if($flag==1$)

/* 如果丢的不是刚到达的包 */

$Add\_into\_queue()$;

$N\_drop=N\_drop+1$;

}

else {

$Max\_delay=Max\_delay+T\_int$;

$R\_drop=R\_drop+(T\_int\times S1)/100$;

$Q\_max=Q\_max+1$;

}

else

$Add\_into\_queue()$. /* 插入队列 */

④ 直到收到静音包之前,一直重复③.

⑤ 一个话音区结束后,调整 Q_min 的值,

$Q\_min=round(Max\_jit/T\_int\ Q\_max)-1$;

$Q\_max=Q\_min+1$;

然后从②开始新的话音区.

⑥ 重复②~⑤知道通话结束.

⑦ 通话中的语音播放:DSP 每隔 T_int 发出中断,激活中断服务程序从队列里面取包给 DSP 进行播放. 如果队列为空,而上次又取包成功,则重放上次取的包,否则 DSP 不放音.

算法中,$drop\_at\_rate(R\_drop)$ 采用等概率过滤丢包法:从队列头向队列尾,直到第 i 个刚收包,以 R_drop 的概率进行过滤丢包. 过滤丢包还有一种是简单的尾部丢包法,如文献[7,15]. 本文的等概率丢包法的语音质量好于尾部丢包法.

5 实验和仿真

本文使用仿真,用 PESQ 法对算法进行客观的评价. 仿真使用两台电脑. 发送端将 8KHz 采样率 16b 的.wav 文件作为源信号,进行 G.729 编码,并通过网络延迟模拟程序,直接发送给接收端电脑. 接收端用 AVVID 算法进行缓冲和解码"播放",写入同样格式的.wav 文件. 由 PESQ 将发送端的源信号文件和接收端的失真信号文件进行计算,求出失真信号的得分值. 作为源信号的.wav 文件从 ITU-T 的 P.862 附件里面选取,本文选取 u_am1s01.wav 作为源信号[10],如图 3 所示. 网络延迟模拟程序可以给出最大 800ms 的时延. 为了便于比较,突发延迟分别固定在第 1 个话音区的第 200ms 和第 2 个话音区的第 400ms 处开始施加. PESQ 计算失真信号得分时,将丢弃的部分按削峰处理进行. 仿真分别计算各种延迟下的得分值. 仿真结果见表 1.

从仿真结果可以看出,本文的 AAVIB 自适应算法的得分好于 Cisco 3600 这类的固定队列长度算法的得分[15]. 说明本文的算法能很好适应突发大时延出现的网络环境中.

该算法在实际语音网关设备中进行了应用,应用表明算法对突发大时延存在的网络中的语音通信质量有明显改善作用.

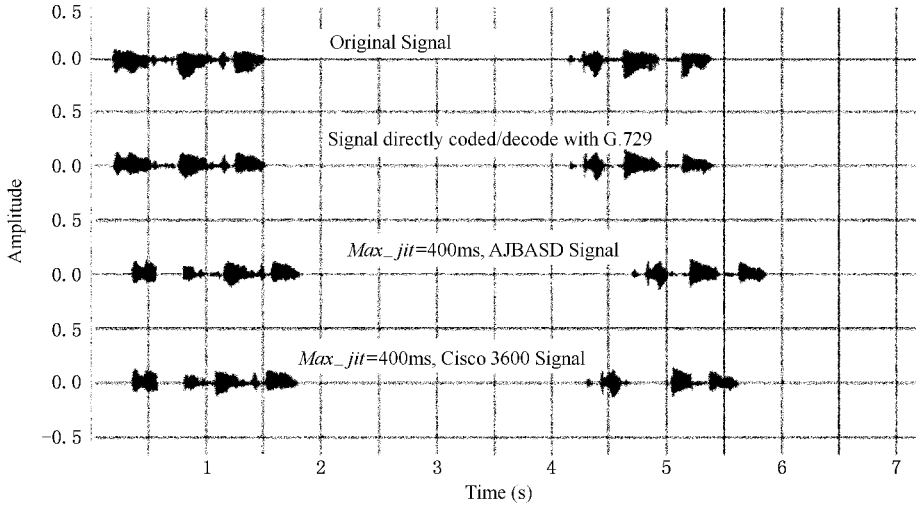

Fig. 3 Some signals in the simulation.

图 3 仿真所用信号和部分仿真结果信号

Table 1 PESQ Scores under Various Delays and Delay Jitters

表 1 各种时延情况下的 PESQ 得分值表

Delay (ms)	Spike Delay (ms)	Score of AAIVB			Cisco 3600 Score
		Talk-Spurt 1	Talk-Spurt 2	Total Score	
100	150	3.636	3.636	3.636	3.636
100	250	2.791	3.477	3.283	2.909
100	300	2.731	3.426	3.204	2.643
150	350	2.833	3.177	3.058	2.262
150	400	1.813	3.244	2.774	1.853
200	450	1.841	3.234	2.767	1.841
200	500	1.848	3.166	2.605	1.777
250	550	1.703	3.159	2.566	1.649

6 结 论

本文采用自适应语音缓冲技术解决突发大时延下的语音质量问题. 算法通过感知和学习其网络传输状况,动态调整 jitter buffer 队列的最小和最大深度. 仿真结果和实际应用情况表明,本文的算法在改善语音通信质量方面有一定的效果.

本文使用 PESQ 来评估语音通信的质量,PESQ 和 MOS 主观评价之间的相关度为 0.935[10]. 从仿真结果表 1 可以看出,PESQ 的得分在一些情况下还不是很准确,在时间对齐方面还可以做深入的研究和改进.

致谢 感谢迈普(四川)公司对本项目的支持!

参 考 文 献

1 Li Zheng, Liren Zhang, Dong Xu. Characteristics of network delay and delay jitter and its effect on voice over IP. 2001 IEEE Int'l Conf. Communications, Helsinki, 2001

2 R. Ramjee, Jim Kurose, Don Towsley, *et al*. Adaptive playout mechanisms for packetized audio applications in wide area networks. IEEE INFOCOMM 1994, Toronto, 1994

3 Aman Kansal, Abhay Karandikar. Adaptive delay estimation for low jitter audio over Internet. IEEE GLOBECOM '01, San Antonio, TX, 2001

4 Liu Fang, Kim JongWon, C. -C. J Kuo. Adaptive delay concealment for Internet voice applications with packet based time-scale modification. 2001 IEEE Int'l Conf. Acoustics, Speech and Signal Processing, Salt Lake City, UT, 2001

5 A. K. Anandakumar, A. McCree, E. Paksoy. An adaptive voice playout method for VOP applications. IEEE GLOBECOM '01, San Antonio, TX, 2001

6 M. Benaissa, V. Lecuire, F. Lepage. An algorithm for playout delay adjustment for interactive audio applications in mobile ad hoc networks. The 7th Int'l Symposium on Computers and Communications, Vandoeuvre-les-Nancy, France, 2002

7 J. Pinto, K. J. Christensen. An algorithm for playout of packet voice based on adaptive adjustment of talkspurt silence periods.

LCN '99 Conf. Local Computer Networks, Lowell, MA, 1999
8 NS Jayant. Effects of packet loss on waveform coded speech. The 5th Int'l Conf. Computer Communications, Atlanta, GA, 1980
9 N. Kitawaki, K. Itoh. Pure delay effects on speech quality in telecommunications. IEEE Journal on Selected Areas in Communications, 1991, 9(4): 586~593
10 ITU-T Recommendation P.862. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. http://www.itu.org, 2001
11 Recommendation G.114. Transmission systems and media general characteristics of Int'l telephone connections and Int'l telephone circuits one-way transmission time. http://www.pesq.org, 1996
12 T. Friedman, R. Caceres, A. Clark. RTP control protocol extended reports (RTCP XR). RFC 3611. http://www.ietf.org, 2003
13 Du Limin, Hou Ziqiang. Research issues on Chinese speech recognition. Acta Electronica Sinica, 1995, 123(10): 110~116 (in Chinese)
(杜利民,候自强. 汉语语音识别面临的一些科学问题. 电子学报, 1995, 123(10): 110~116)
14 Zhang Jialu, Hu Xinghui. A comparative study of F_0 patterns between Chinese and foreign languages. Acta Acoustics, 1994, 120(1): 66~71 (in Chinese)
(张家禄,胡兴慧. 汉语和外语的基频模式的对比研究. 声学学报, 1994, 120(1): 66~71)
15 Cisco Systems Inc. Playout Delay Enhancements. http://www.cisco.com/en/US/products/sw/iosswrel/ps1834/products_feature_guide09186a008008033c.htm, 2004-06

Gou Xiantai, born in 1971. Ph. D. candidate. His current research interests include data communication technology and agent technology.
苟先太,1971 年生,博士研究生,主要研究方向为数据通信、Agent 技术.

Jin Weidong, born in 1959. Professor and Ph. D. supervisor. His current research interests are intelligent information processing and system simulation.
金炜东,1959 年生,教授,博士生导师,主要研究方向智能信息处理、系统仿真.

Jin Fan, born in 1934. He is a professor and Ph. D. supervisor. His current research interests are computer coding and neural network.
靳蕃,1934 年生,教授,博士生导师,主要研究方向为计算机编码、神经网络.

Research Background

In 2003, a test was carried out for voice communication between an Intranet of the Ministry of Transportation in Beijing and an enterprise network in New York, and the round-trip average delay was found to be 1713 ms. The official IP voice service network organized in this environment still enjoys a sound development. The actual application demonstrates that given the condition of appropriately size of end-to-end delay, as long as the voice is smooth and continuous, users can still complete the IP voice communication with certain satisfaction degree. However, the spike delay jitters cause the quality of voice. In this work, an adaptive jitter buffering algorithm based on the detection and study of the spike status of the networks is presented to promote the quality of voice communication. It timely adjusts the minimal and maximal depth of buffer queue according to the control target of end-to-end delay and packet loss rate. The algorithm can much more easily achieve the continuous playout because it plays voice packet at a fixed inter-play time in the most time of a talk-spurt. The control target of packet loss rate can be extended to 20%. However, the basic algorithms can only bear 5%~10% of the packet loss rate. Perceptual evaluation of speech quality (PESQ) is applied to assess the speech quality in the simulation. It is shown that the algorithm can obviously promote the quality of voice communication in IP networks with spike delay. The practical application in the voice gateway Mypower VG2000 can also prove the effects of voice quality promotion. Our research is supported by Maipu Data Communication Co., Ltd. The quality of voice communication with spike delay will be a serious problem in the next generation networks (NGN). More works should be done in the future.