文章编号:1009-3087(2013)03-0103-05

引入内容特性分析的包层语音质量评价模型

江亮亮,李雪敏,杨付正,杨 旭
(西安电子科技大学 综合业务网理论及关键技术国家重点实验室,陕西 西安 710071)

摘 要:为了实现对网络语音质量的实时监控,提出一种包层语音质量评价模型。该模型无需介入数据包的载荷部分,只利用数据包的头信息评价语音质量。首先通过分析包头信息区分出语音段和静音段,获取语音段的编码参数和丢包参数,然后根据语音段的编码参数预测编码失真,在此基础上利用语音段的丢包参数评价丢包引起的失真,从而得到语音流的总质量。实验结果表明,相比于国际标准G.107中的E-model,提出的模型得到的语音质量评分与PESQ算法评分的皮尔森相关系数平均提高0.041 2,均方根误差平均降低0.045 1。

关键词:语音编码;语音传输;服务质量;丢包
中图分类号:TN912.3 **文献标志码**:A

A Packet-layer Model for Speech Quality Assessment Introducing the Analysis of Content Feature

JIANG Liang-liang, LI Xue-min, YANG Fu-zheng, YANG Xu
(State Key Lab. of Integrated Service Networks, Xidian Univ. Xi'an 710071, China)

Abstract: A packet-layer model for speech quality assessment was proposed to monitor the quality of networked speech. Without resorting to any media-related payload information, the proposed model predicted the speech quality only in terms of the information provided by packet headers. First, the analysis of content feature was performed to locate the voiced segments and silence segments by analyzing the information from packet headers, and the coding and packet loss parameters of voiced segments were obtained. Then the coding distortion was estimated according to the coding parameters of voiced segments, based on which the overall speech quality was further evaluated by taking account of the impact of packet loss. Experimental results showed that the proposed model could get an increment about 0.041 2 in Pearson correlation coefficient (PCC) and a decrement about 0.045 1 in root mean squared error (RMSE) compared with the E-model proposed in ITU-T recommendation G.107.

Key words: speech coding; speech transmission; quality of service; packet loss

近年来,VoIP(voice over internet protocol)凭借低成本等优点迅速兴起。然而由于VoIP提供尽力而为的服务,通话质量容易受到网络环境的影响。为了提高VoIP系统的服务质量(quality of service, QoS)和保证用户的体验质量(quality of experience, QoE),通常需要实时监测语音质量来优化网络资源分配。语音质量评价方法包括主观方法和客观方法2种。主观方法费时费力,不宜用于网络语音质量的实时评价[1]。所以,如何对网络语音质量进行客观评价成为一个亟待解决的问题。

根据输入信息类型以及对码流的介入程度,客观语音质量评价方法可以分为:参数规划模型、包层模型、比特流层模型、媒体层模型以及混合模型[2]。其中,包层模型只允许利用数据包的头信息预测语音质量,计算复杂度较低,适用于网络节点对大量语音流进行质量监控。目前已有一些文献针对包层模型进行研究[3-5]。文献[3]提出根据丢包率和丢包的突发程度评价丢包造成的语音质量下降。日本电报电话公司(Nippon Telegraph and Telephone, NTT)提出的包层评价模型利用码率、丢包次数以及平均突发长度评价语音质量[4]。目前,ITU-T正在针对包层评价模型制定新的国际标准P. NAMS[5]。另外,国际标准G.107提出的E-Model虽然属于参数规划模型,但是它的输入参数能够通过包层信息获得,常用作包层评价语音质量[6]。

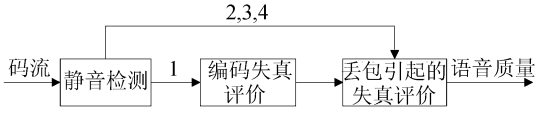
收稿日期:2012-11-22
基金项目:国家自然科学基金资助项目(60902081);中央高校基本科研业务费专项资金资助项目(72115612);高等学校学科创新引智计划资助项目(B08038)
作者简介:江亮亮(1988—),男,博士生. 研究方向:多媒体通信;音频质量估计. E-mail: lljiang@stu.xidian.edu.cn

然而已有的包层模型都未对语音流的内容特性进行分析,这会影响语音质量评价的准确性。因为语音流通常包含语音段和静音段,其中,语音段携带重要信息,是影响系统 QoS 和用户 QoE 的主要部分。而通过对语音数据的包头信息进行分析,可以预测出数据包是静音包还是语音包,在此基础上能够对语音质量进行更加准确地评价。因此,作者提出一种引入内容特性分析的包层语音质量评价模型,首先利用包层信息判断数据包的类型,然后根据内容的重要性分别对不同类型的数据包进行处理,区分不同类型数据包对语音质量的不同影响。

1 引入内容特性分析的包层语音质量评价模型

语音信号在到达客户端之前需要经过语音编码和语音传输。语音编码通常采用有损编码方式,导致无法完全重建原始语音信号,造成语音质量的下降。而为了满足实时性的要求,语音传输通常会选择 RTP/UDP/IP 作为传输协议,容易引起数据包丢失。所以,引起网络语音失真的主要因素是语音编码和数据包丢失。那么,如何有效评价语音编码和数据包丢失对语音质量的影响就成为建立网络语音质量评价模型的关键环节。

作者提出的模型框架如图 1 所示。首先通过分析包头信息区分出语音段和静音段,获取用于质量评价的语音段信息,包括语音段的平均码率、丢包次数、平均突发长度和数据包数。然后根据语音段的平均码率预测编码失真,在此基础上利用其余参数评价丢包引起的失真,从而得到语音流的总质量。



1. 语音段平均码率; 2. 总的语音包数;
3. 语音包的丢包次数; 4. 语音包的平均突发长度。

图 1 引入内容特性分析的包层语音质量评价模型

Fig. 1 Packet-layer modelfor speech quality assessment introducing the analysis of content feature

作者以自适应多速率(adaptive multi-rate, AMR)语音编码标准为例给出各个模块的详细算法。另外,由于 PESQ(perceptual evaluation of speech quality)算法得到的语音序列评分能够很好地反映序列的主观质量[7],建立无参考评价模型及评价无参考模型性能时,PESQ 模型评分经常作为主观评分使用[8-9]。因此,在回归模型参数及评价所提出模型性能时也采用 PESQ 模型评分代替主观评分。

1.1 静音检测

静音检测方法通过分析接收到数据包的头信息预测所有数据包的类型,在此基础上可以统计语音部分的压缩码率和丢包参数。对于接收到的数据包,通过分析包头信息可以得到帧的编码字节数,然后根据编码字节数区分语音帧和静音帧。因为 AMR 在静音时段采用了话音激活检测和舒适噪声生成技术,静音帧的编码字节数仅为 6 或 1,远远小于语音帧的编码字节数[10]。以常用的 RTP/UDP/IP 协议栈为例,UDP 头信息中的长度域标识了 UDP 头及其有效载荷部分的总长度,有效载荷大小可以由 UDP 包的总长度减去 UDP 和 RTP 头信息长度得到。RTP 头中的时间戳表示 RTP 包有效载荷的第一个字节的采样时刻,前后 2 个相邻数据包时间戳的差值表示前一个数据包有效载荷的持续时间,又因为 AMR 每帧的持续时间固定,因而可以得到数据包包含的帧数。因此当接收到的帧字节数为 6 或 1 时可判断该帧为静音帧,否则判断为语音帧。

对于丢失的数据包,无法获取帧的编码字节数,因而不能直接预测数据包的类型。但是语音信号具有短时相关性,即静音段和语音段通常会持续多个帧周期,因此可以利用相邻未丢失包的类型来判断丢失包的类型[11]。如果相邻的 2 个未丢失包都是语音包,就将丢失包判断为语音包。如果相邻的 2 个未丢失包都是静音包,就将丢失包判断为静音包。当相邻的 2 个未丢失包的类型不一致时,丢失包可能是语音包或静音包,作者将该丢失包判断为语音包,因为语音包在语音流中占的比例较高。数据包是否丢失可以根据 RTP 头中序列号是否连续进行判断,按照 RTP 协议,无丢包情况下的序列号是连续的,每发送一个 RTP 包,序列号就增加 1。

1.2 编码失真评价

码率是反映语音流编码失真的重要参数,通常码率越高,编码失真越小。对于语音段,AMR 支持 8 种编码速率,分别为 4.75、5.15、5.9、6.7、7.4、7.95、10.2 以及 12.2 kb/s。从国际标准 ITU-T P-series 的语音数据库 supplement 23 中随机挑选 20 个男声和 20 个女声语音序列[12],评价这些语音序列在各码率下失真语音序列的质量,将各语音序列在同一码率下的语音质量取平均作为该码率下的失真语音质量,结果如图 2 所示。

已有的包层模型通常根据整个语音流的平均码率预测编码失真[4]。然而由于静音段对语音质量基本没有影响,且 AMR 语音段的码率可能根据实

时信道条件自适应地发生变化,作者提出利用语音段的平均码率预测编码失真。

根据图 2 中的数据,按照最小二乘准则拟合出只考虑编码失真的语音质量与语音段平均码率之间的关系为:

$$Q_c = m_1 \exp\left(-\frac{m_2}{br}\right) \tag{1}$$

其中,Q_c 为只考虑编码失真的语音质量,br 为语音段平均码率,m_1、m_2 为待定常数。根据图 2 中的数据,得到最小二乘意义下 m_1 和 m_2 的最优值。

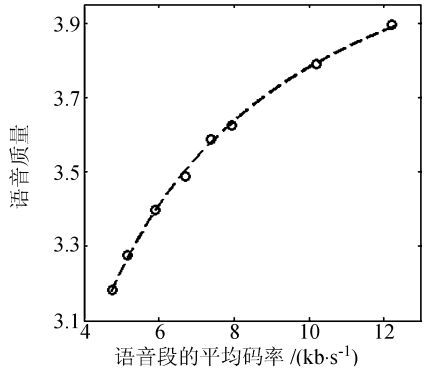


图 2　只考虑编码失真的语音质量与语音段平均码率的关系

Fig. 2　Relationship between the quality regarding coding distortion and the average bit-rate of voiced segments

1.3　丢包引起的失真评价

丢包是影响网络语音质量的另一个重要因素。早期的算法通常只使用丢包率评价丢包引起的失真[6],研究发现丢包引起的失真不仅取决于丢包率,还与丢包的分布情况有关[3]。改进的 E-model 引入突发比来表征丢包的突发程度[13],文献[14]在突发丢包和随机丢包之间建立一种等价的转换关系。虽然这 2 种方法考虑了丢包的突发程度对语音质量的影响,但是效果并不明显。文献[4]提出的包层评价模型利用丢包次数和平均突发长度来评估丢包引起的失真,综合考虑了丢包的次数、个数以及突发程度等因素对语音质量的影响,评价性能较好,作者称该模型为 NTT 模型。这里的丢包次数是指丢包分组的个数,平均突发长度是指丢包分组包含的平均包数,而丢包分组是指一组连续发生的丢包。

图 3 给出了丢包次数对语音质量的影响,选取 1.2 节的 40 个原始语音序列(时长均为 8 s),将编码速率设为 5.9 kb/s, 平均突发长度设为 1,并将丢包次数分别设置为 1、2、4、8。

根据图 3 中的实验数据,拟合出二者的关系为:

$$Q = \left[(1-m_3)\exp\left(-\frac{T}{m_4}\right) + m_3\exp\left(-\frac{T}{m_5}\right)\right] \times (Q_c - 1) + 1 \tag{2}$$

其中,Q 为语音流的总质量,T 为丢包次数,m_3、m_4 和 m_5 为待定常数。根据图3 的数据,利用最小二乘拟合得到 m_3、m_4 和 m_5 的值。

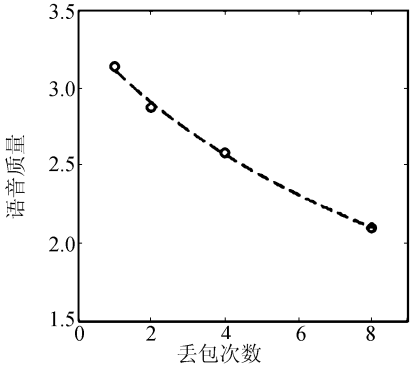


图 3　语音质量与丢包次数的关系

Fig. 3　Relationship between the speech quality and the times of packet loss

为了分析平均突发长度对语音质量的影响,实验将丢包次数设定为 1,将平均突发长度分别设置为 2、3、4、6、8,然后评价各平均突发长度下的语音质量。对于平均突发长度 L_{ab}($L_{ab} > 1$),构造一个对应的虚拟丢包次数 T_v[4], 使丢包次数为 1、平均突发长度为 L_{ab} 和丢包次数为 T_v、平均突发长度为 1 这 2 种丢包情况下的语音质量相等。根据上述定义,将实验中每一个平均突发长度对应的语音质量赋值给式(2) 中的 Q,解出的 T 值即为该平均突发长度对应的 T_v 值。虚拟丢包次数 T_v 与平均突发长度 L_{ab} 的对应关系如图 4 所示,二者呈线性关系:

$$T_v = m_6(L_{ab} - 1) + 1 \tag{3}$$

其中,m_6 为待定常数,通过对图 4 的数据进行最小二乘拟合得到。

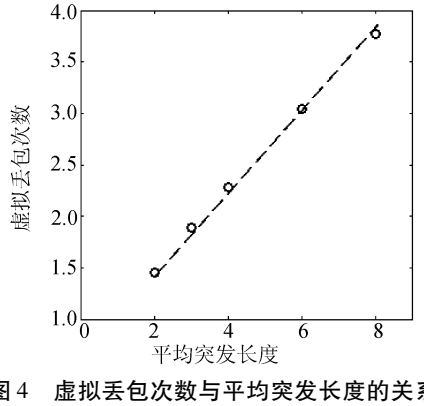


图 4　虚拟丢包次数与平均突发长度的关系

Fig. 3　Relationship between the virtual times of packet loss and the average burst length

在 $L_{ab} > 1$ 的情况下,利用式(3)可以将丢包次数为1、平均突发长度为 L_{ab} 的丢包等价转换为丢包次数为 T_v、平均突发长度为1的丢包。依此类推,丢包次数为 T、平均突发长度为 L_{ab} 的丢包可以等价转换为丢包次数为 T_vT、平均突发长度为1的丢包,这样就可以用式(2)评价转换后的语音质量,即将 T_vT 代入式(2),结果如式(4)所示:

$$Q = \left[(1-m_3)\exp\left(-\frac{T_vT}{m_4}\right)+m_3\exp\left(-\frac{T_vT}{m_5}\right)\right]\times (Q_c-1)+1 \tag{4}$$

另外,如果将 $L_{ab}=1$ 代入式(3),可以得到 $T_v=1$,此时式(4)和(2)是等价的,可见式(4)同样适用于 $L_{ab}=1$ 的情况。所以,不论 L_{ab} 值为多大,均可使用式(3)和(4)评价语音流质量。

由于携带信息的重要性不同,语音包和静音包丢失对语音质量的损伤程度差别较大。语音包携带了重要的信息,如果发生丢失,不仅造成本身信息的缺失,还会导致后续语音的质量下降,而静音包发生丢失对语音质量基本没有影响,可见语音包丢失是造成语音质量下降的主要因素。因此,作者提出采用语音包的丢失次数和平均突发长度评价丢包引起的失真。此外,提出的模型还引入了总的语音包数这个参数。除了语音包的丢包次数和平均突发长度,丢包引起的失真还与总的语音包数有关。在语音包的丢包次数和平均突发长度一定的情况下,总的语音包数越大,语音包的丢包率越小,丢包对语音质量造成的损伤就越小。基于以上分析,在式(3)和(4)的基础上,采用式(5)和(6)评价语音流的质量:

$$Q = \left[(1-m_7)\exp\left(-\frac{T_{vo}T_o}{m_8N_o}\right)+m_7\exp\left(-\frac{T_{vo}T_o}{m_9N_o}\right)\right]\times (Q_c-1)+1 \tag{5}$$

$$T_{vo} = m_{10}(L_{abo}-1)+1 \tag{6}$$

其中,T_o 为语音包的丢包次数,L_{abo} 为语音包的平均突发长度,T_{vo} 为 L_{abo} 对应的虚拟丢包次数,N_o 为总的语音包数,m_7、m_8、m_9 和 m_{10} 为待定常数,由参数拟合实验得到。具体的拟合方法如下:

1)选取1.2节的40个原始语音序列,将编码速率分别设为4.75、5.15、5.9、6.7、7.4、7.95、10.2以及12.2 kb/s,L_{abo} 设定为1,T_o 分别设为1、2、4、8,然后利用PESQ算法评价失真语音的质量。

2)根据步骤1)的实验数据得到最小二乘意义下 m_7、m_8 和 m_9 的最优值。

3)在各编码速率下,将 T_o 设定为1,L_{abo} 分别设为2、3、4、6、8,然后将每个 L_{abo} 对应的失真语音质量赋值给式(5)中的 Q,得到每个 L_{abo} 对应的 T_{vo} 值。

4)根据步骤3)得到的 L_{abo} 值及其对应的 T_{vo} 值,计算最小二乘意义下 m_{10} 的最优值。

2 实验结果

为了全面测试提出模型的性能,从国际标准ITU-T P-series的语音数据库supplement 23中随机挑选10个男声和10个女声语音序列[12]作为原始测试序列,这些序列与1.2节选取的原始训练序列完全不同。首先使用AMR编码器对原始测试序列进行压缩,压缩码率分别为:4.75、5.15、5.9、6.7、7.4、7.95、10.2和12.2 kb/s,然后将压缩码流打包并采用4状态马尔可夫模型模拟丢包[15],实验选用的丢包率分别为0%、1%、3%、5%和10%。最后,生成800个失真语音序列用于模型性能测试。

作者所提出模型的参数值如表1所示。

表1 参数列表

Tab.1 Parameter values

m_1	m_2	m_7	m_8	m_9	m_{10}
4.416	1.555	0.044	0.151	0.01	0.385

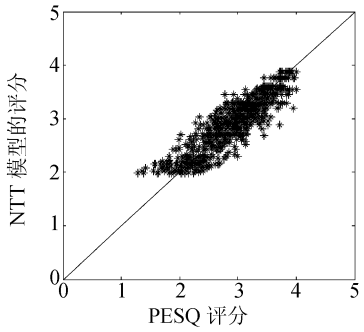
为了验证提出模型的性能,将其与E-model[6]、NTT模型[4]进行比较,选取皮尔森相关系数(Pearson correlation coefficient,PCC)和均方根误差(root mean square error,RMSE)2个指标作为衡量模型性能的标准。表2给出了3个比较模型的具体性能,可以看到提出的模型在性能上明显优于E-model和NTT模型。与E-model和NTT模型相比,提出的模型在PCC上分别平均提高0.041 2和0.037 4,同时在RMSE上分别平均降低0.045 1和0.041 1。

表2 模型性能比较

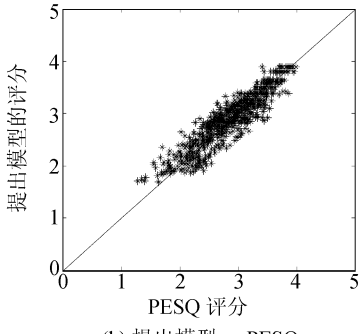
Tab.2 Performance comparison

模 型	PCC	RMSE
E-model	0.888 9	0.252 4
NTT模型	0.892 7	0.248 4
提出的模型	0.930 1	0.207 3

为了更加直观地比较模型性能,给出了NTT模型的评分和提出模型的评分与PESQ评分的对比散点图,如图5所示,其中,图5(a)为NTT模型的评分与PESQ评分的对比,图5(b)为提出模型的评分与PESQ评分对比。通过观察可以发现,与图5(a)相比,图5(b)中的散点更加集中在对角线附近。由此可见,提出模型的评分与PESQ评分具有更好的一致性。



(a)NTT 模型 vs PESQ



(b) 提出模型 vs PESQ

图 5　PESQ 评分与 NTT 模型、提出的模型评分的对比散点图

Fig. 5　Scatter plots of the scores acquired by PESQ vs the scores predicted by the NTT and proposed model

3　结　论

作者提出一种低复杂度的包层语音质量评价模型，通过分析数据包的内容特性，提取出对语音质量评价起关键作用的语音包，并根据语音包的码率和丢包参数预测语音质量，有效地提高了评价的准确性。实验结果表明，提出的模型在性能上要明显优于 E-model 和 NTT 模型。

提出的模型主要适用于语音信号，对音乐等其它音频则不适用。另外，虽然提出的模型是针对 AMR 编码，不能直接用于其他的编码标准，但是对其他编码标准下的语音质量评价仍有一定的参考意义，如宽带自适应多速率（AMR-WB）编码器在编码时采用了话音激活检测和舒适噪声生成技术，同样可以在语音质量评价时引入内容特性分析，区分语音段和静音段，从而提高语音质量评价的准确性。

参考文献：

[1] Huang Li, Zhou Jie, Ma Hong, et al. Object speech quality assessment by neural network with distance measure as inputs[J]. Journal of Sichuan University: Natural Science Edition, 2007, 40(6): 1210 – 1214. [黄丽，周杰，马洪，等. 以测度作为神经网络输入的客观音质评价研究[J]. 四川大学学报：自然科学版，2007，40(6)：1210 – 1214.]

[2] Takahashi A, Yoshino H, Kitawaki N. Perceptual QoS assessment technologies for VoIP[J]. IEEE Communications Magazine, 2004, 42(7): 28 – 34.

[3] Clark A. Modeling the effects of burst packet loss and recency on subjective voice quality[C]//Proceeding of the 2nd IP Telephony Workshop. New York, USA, 2001: 123 – 127.

[4] Egi N, Hayashi T, Takahashi A. Parametric packet-layer model for evaluation audio quality in multimedia streaming services[J]. IEICE Transactions on Communications, 2010, E93-B(6): 1359 – 1366.

[5] ITU-T SG12 Temporary Document TD 297　Updated draft terms of reference for P. NAMS[S]. Geneva, Switzerland: ITU-T, 2010.

[6] ITU-T Recommendation G. 107　The E-model, a computational model for use in transmission planning[S]. Geneva, Switzerland: ITU-T, 2002.

[7] ITU-T Recommendation P. 862　Perceptual evaluation of speech quality (PESQ), an objective method for end to end speech quality assessment of narrowband telephone networks and speech codecs[S]. Geneva, Switzerland: ITU-T, 2001.

[8] Radhakrishnan K, Larijani H, Buggy T. A non-intrusive method to assess voice quality over internet[C]//Proceeding of the 2010 International Symposium on Performance Evaluation of Computer and Telecommunication Systems. Ottawa, Canada: IEEE, 2010: 380 – 386.

[9] Roychoudhuri L, Al-Shaer E S. Real-time audio quality evaluation for adaptive multimedia protocols[C]//Proceeding of the 8th IFIP/IEEE International Conference on Management of Multimedia Networks and Services. Barcelona, Spain: IFIP/IEEE, 2005: 133 – 144.

[10] 3GPP TS 26. 101 v6. 0. 0　Adaptive multi-rate (AMR) speech codec frame structure[S]. Valbonne, France: 3GPP, 2004.

[11] Yang Fuzheng, Jiang Liangliang, Li Xiao. Real-time quality assessment for voice over IP[J]. Concurrency and Computation: Practice and Experience, 2012, 24(11): 1192 – 1199.

[12] ITU-T Recommendation P-series supplement 23　ITU-T coded-speech database[S]. Geneva, Switzerland: ITU-T, 1998.

[13] ITU-T Recommendation G. 107　The E-model, a computational model for use in transmission planning[S]. Geneva, Switzerland: ITU-T, 2005.

[14] Jelassi S, Rubino G. A comparison study of automatic speech quality assessors sensitive to packet loss burstiness[C]//Proceeding of the 8th Annual IEEE Consumer Communications and Networking Conference on Multimedia & Entertainment Networking and Services. Las Vegas, USA: IEEE, 2011: 415 – 420.

[15] ITU Rep. COM12-D97-E　Packet loss distributions and packet loss models[S]. Geneva, Switzerland: ITU-T Study Group 12, 2003.

（编辑　赵　婧）