doi:10.3969/j.issn.1001-2400.2013.03.003

利用两级时域联合的包层语音质量评价模型

江亮亮，杨付正，任光亮
(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：针对相同丢包率下不同丢包模式对应的语音质量存在差异的情况,提出了一种能够反映丢包模式对语音质量影响的包层语音质量评价模型.首先通过分析数据包头获取编码速率和丢包位置等信息,在此基础上,结合静音检测技术及误码传播结果预测每一帧的质量;然后根据人的感知特性将语音序列自由划分为变长帧组,并联合各帧的质量得到帧组质量;最后,综合各帧组的质量得到语音序列的总质量.提出的模型在两级时域联合过程中,为失真严重的区域分配更大的权重,从而有效反映丢包模式对语音质量的影响.实验结果表明,相比于国际标准 G.107 中的 E-model,所提模型的评分与语音质量感知评估方法的评分相比,皮尔森相关系数平均提高了 0.012 9,同时均方根误差平均降低了 0.023 4.

关键词：语音质量评价;时域联合;丢包;服务质量
中图分类号:TN912.3 **文献标识码**:A **文章编号**:1001-2400(2013)03-0014-06

Packet-layer model for voice quality assessment using two-level temporal pooling scheme

JIANG Liangliang, *YANG Fuzheng*, *REN Guangliang*
(State Key Lab. of Integrated Service Networks, Xidian Univ., Xi'an 710071, China)

Abstract: Aiming at the problem that the voice qualities corresponding to different packet loss patterns show significant differences at the same packet loss rate, a packet-layer model for voice quality assessment, which well reflects the effect of the packet loss patterns on the voice quality, is presented. First, the information about the coding bit-rate and packet loss is obtained by analyzing the packet header, on the basis of which the frame quality is measured with the further information about silence detection and error propagation. Then the voice sequence is divided into groups of frames (GOFs) with a variable length and a short-term temporal pooling method is employed to obtain the GOF quality. Finally, the overall voice quality is determined by the long-term temporal pooling of the GOF qualities. The proposed two-level temporal pooling scheme well describes the effect of different packet loss patterns on the voice quality since the strongest impairments are predominately emphasized. Experimental results show that the presented model can lead to an increment of about 0.012 9 in the Pearson Correlation coefficient (PCC) and a decrement of about 0.023 4 in the Root Mean Squared Error (RMSE) compared with the E-model in ITU-T recommendation G.107.

Key Words: voice quality assessment; temporal pooling; packet loss; quality of service

近年来,网络电话(Voice over Internet Protocol,VoIP)在国内外获得了飞速的发展.与传统电话相比,VoIP 具有占用网络资源少、成本低等优势.然而 IP 网络只提供尽力而为的服务,时变的网络特性会影响网络语音服务的质量[1].通过对网络语音质量的监控和反馈,可以调整压缩或传输参数,改善网络语音的质量.因此,如何对网络语音质量进行实时和准确评价,成为一个亟待解决的问题.语音质量评价方法分为主观和

收稿日期:2012-10-08 **网络出版时间**:2013-02-25
基金项目:国家自然科学基金资助项目(60902081,60902052);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,西安电子科技大学博士研究生,E-mail: lljiang@stu.xidian.edu.cn.
网络出版地址:http://www.cnki.net/kcms/detail/61.1076.TN.20130225.1050.201303.19_003.html

客观两种,主观质量评价能够准确反映语音质量,但是实现起来步骤复杂,实时性不好,不宜用于实时话音通信中的质量评价[2];客观质量评价方法对于网络语音业务更为适用,目前是语音质量评价研究领域的热点.

根据是否需要原始信号,客观评价方法分为全参考和无参考两类[3].全参考评价方法是通过比较原始信号和失真信号之间的差别来判别语音质量的好坏的,语音质量感知测量算法(Perceptual Speech Quality Measure, PSQM)引入认知模型来描述原始语音信号和失真语音信号在听觉变换过程中产生的区别,能够较准确地预测语音质量,成为了国际电信联盟(ITU)的语音质量客观评价标准 P.861[4].2001 年,ITU-T 推出了新一代语音质量评价标准 P.862[5],提出的语音质量感知评估算法(Perceptual Evaluation of Speech Quality, PESQ)与主观评价的相关性高,被广泛应用.全参考模型需要原始语音信号,极大地限制了其在网络环境中的应用[6-7],而无参考评价方法只根据失真信号或统计网络参数来评价语音质量,可以适用于不同的网络环境.

根据模型输入信息类型以及对码流的介入程度,评价方法还可以分为: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[8].包层模型只允许利用数据包的头信息预测语音质量[9],计算复杂度低,适用于网络节点对语音质量的实时监控,且该模型适用于媒体相关的有效载荷加密的情况.文中研究的是包层的无参考语音质量评估模型.目前,包层无参考模型正受到广泛关注,ITU-T 正在针对包层参数模型制订新的国际标准 P.NAMS[10].

影响网络语音流质量的因素主要为语音压缩和数据包丢失.有效评价数据包丢失对网络语言质量的影响是评价网络语音质量的关键模块,由于包层参数模型无法介入载荷信息,常用丢包率来反映丢包引起的失真,如 E-Model[11]和 VQmon 模型[9].然而,网络丢包往往具有突发性,研究表明,突发丢包的长度越长,对语音质量的损伤就越严重[12].新 E-Model 引入了突发比这一参量来表征丢包的突发性.文献[12]中的 Q-Model 则是采用一定的方式将突发丢包率映射为等价的随机丢包率.然而丢包对语音质量的影响与具体的丢包位置相关[13],丢包率和突发比都是统计参数,它们无法准确表征具体的丢包模式对语音质量的影响.为了准确反映丢包模式对语音质量的影响,笔者提出了基于时域联合的语音质量评估框架[13],并在此基础上,引入了人的感知特性,提出了一种包层语音质量评价模型.

图 1 给出了文中提出的包层语音质量评价模型框图.首先根据具体的传输协议对语音流的数据包头部进行分析,获得每个语音帧的编码速率和每个语音帧是否丢失等信息.以实时传输协议/用户数据报协议/因特网协议(RTP/UDP/IP)传输协议栈为例,丢包可以通过分析 RTP 包头部的序列号得到,每个语音帧的码率可以通过 UDP 包头部的长度域及 RTP 头的时间戳得到.丢包内容检测模块通过相邻帧的特性,预测丢失帧是语音帧还是静音帧,以区分不同内容丢失对语音质量的影响.在此基础上,根据编码速率及丢失帧的位置,预测每个语音帧的质量,最后使用两级时域联合模块得到语音序列质量.

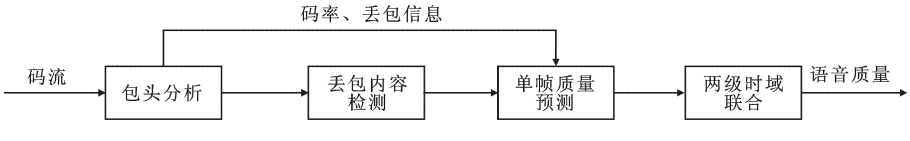


图 1　利用两级时域联合的包层语音质量评价模型

为了验证算法的有效性,采用无线网络环境广泛使用的宽带自适应多速率(AMR-WB)编码标准,语音源序列的采样频率均为 16 000 Hz.由于 PESQ 算法与主观评价的强相关性,使用 PESQ 算法评分代替语音序列的主观评分,既可以省去费时费力的主观测试,又能够避免主观实验结果的不稳定性带来的影响,这种方法常用于评价无参考语音质量评估模型的性能[14].

1　语音帧质量

引起网络语音流失真的因素主要为语音压缩和数据丢失.每个语音帧都受到有损压缩而引起失真,压缩失真主要与采用的压缩标准和编码速率相关.数据包丢失会导致相应语音帧无法正常恢复,时域预测技术还会导致后续的语音帧质量下降.根据受丢包的影响程度,将语音帧分为:正常帧、丢失帧以及受误码扩散帧.

正常帧是指没有受到丢包影响的帧,正常帧质量直接由压缩失真决定;丢失帧是指由于数据包丢失而无法解码的帧,通常丢失帧采用误码掩盖方法进行弥补;受误码扩散帧是指受丢失帧影响的帧,一般为丢失帧的后续帧.由于不同类型的帧受丢包的影响程度不同,这里将依此预测每种类型语音帧的质量.

1.1 帧内容检测

语音序列根据内容通常可以分为语音帧和静音帧,语音帧携带重要的语音信息,而静音帧不包含任何有用的信息.语音帧的丢失会引起该帧数据无法恢复,造成失真;而静音帧丢失则几乎不影响语音的质量.因此,为了提高语音质量评价模型的性能,有必要判断丢失帧是否为语音帧.如图 1 所示,帧内容检测位于单帧质量预测之前,后续的预测帧质量部分只考虑语音帧丢失的情况,对于丢失的静音帧当作未丢失处理.

AMR-WB 编码器对静音部分采用话音激活检测和舒适噪声生成技术,使得每个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,因而可以直接根据帧的编码长度判断帧的内容.由于丢失帧的数据已经丢失,无法直接利用丢失帧的数据判断丢失帧的类型,而语音信号具有很强的短时相关性,可以利用相邻未丢失帧的类型判断当前丢失帧的类型[15].如果相邻的两个未丢失帧均为静音帧,则将丢失帧判断为静音帧;在其他的情况下,则将丢失帧判断为语音帧.在相邻的两个未丢失帧类型不一致的情况下,丢失帧可能为语音帧或为静音帧,这里将其判断为语音帧,因为语音帧被误判为静音帧往往会引起较大的预测质量偏差.

1.2 正常帧质量

正常帧的质量只与压缩失真有关,而包层参数模型又无法介入载荷信息,所以正常帧的质量只能根据编码方式和编码速率来预测.对于确定的压缩算法,正常帧质量由其压缩速率决定.AMR-WB 共支持 9 种编码速率,分别为 6.60 kb/s、8.85 kb/s、12.65 kb/s、14.25 kb/s、15.85 kb/s、18.25 kb/s、19.85 kb/s、23.05 kb/s 以及 23.85 kb/s.为了得到 AMR-WB 各速率下的正常帧质量,从 ITU-T 语音数据库中选取 8 个不同内容不同人声录制的语音序列作为训练序列,分别为:A_eng_f1、A_eng_f3、A_eng_f4、A_eng_f7、A_eng_m2、A_eng_m3、A_eng_m5 以及 A_eng_m8.首先对训练序列进行 AMR-WB 编码和解码,然后利用 PESQ 算法评价重建语音信号的质量,如图 2 所示.在无丢包的情况下,重建语音信号的各帧均为正常帧,因而将重建语音信号的质量视为该编码速率下正常帧的质量.同时为了消除内容差异带来的影响,将同一编码速率下各训练序列正常帧质量的平均值作为该编码速率下的正常帧质量.实验得到的 AMR-WB 各编码速率下正常帧质量如表 1 所示.

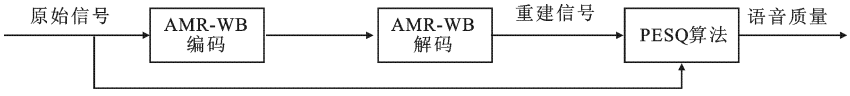


图 2 PESQ 算法评价语音质量的流程

表 1 AMR-WB 各编码速率下的正常帧质量

速率/(kb·s^{-1})	6.60	8.85	12.65	14.25	15.85	18.25	19.85	23.05	23.85
正常帧质量	3.368	3.653	3.906	3.951	3.990	4.053	4.064	4.100	4.107

1.3 丢失帧和受误码扩散帧的质量

采用 1.2 节的 8 个语音序列作为训练序列,对编码码流随机挑选 9 个不同的语音帧进行丢弃,解码过程使用静音对丢包区域进行掩盖,采用 PESQ 算法对各个丢失帧进行质量评价,并将各丢失帧质量的均值作为丢失帧的质量.实验表明,各个编码速率下的丢失帧质量近似,丢失帧的质量与编码速率无关,丢失帧的质量如表 2 所示.

编码算法采用的时域预测技术导致丢失帧的失真向后续帧传播,为了确定对后续帧质量的影响程度,采用 PESQ 算法对丢失帧的后续帧进行质量评价.实验表明,受丢失帧影响的范围与编码算法相关,AMR-WB 的受丢失帧影响的范围大约为 5 帧.另外,受误码扩散帧的质量与距离丢失帧的距离密切相关,距离丢包位置越近,受误码扩散帧的质量越差.表 2 中列出了 AMR-WB 各个编码速率下受误码扩散帧的质量.

表 2　AMR-WB 各编码速率下丢失帧及受误码扩散帧的质量

编码速率/(kb·s^{-1})	丢失帧质量	丢失帧后续各帧的质量				
		第 1 帧	第 2 帧	第 3 帧	第 4 帧	第 5 帧
6.60	0.69	2.528	2.533	2.556	2.723	2.962
8.85	0.69	2.546	2.590	2.650	2.862	3.117
12.65	0.69	2.607	2.620	2.670	3.021	3.294
14.25	0.69	2.606	2.694	2.717	3.008	3.296
15.85	0.69	2.611	2.660	2.663	3.013	3.307
18.25	0.69	2.612	2.620	2.750	3.076	3.335
19.85	0.69	2.590	2.691	2.707	3.073	3.397
23.05	0.69	2.628	2.654	2.716	3.092	3.369
23.85	0.69	2.646	2.698	2.711	3.093	3.412

2　两级时域联合

时域联合是指把语音序列中每个帧的质量组合得到语音序列的质量,已有的研究表明,时域联合算法可以有效反映不同丢包模式对语音序列质量的影响[13]. 每个语音帧的时长很短,如 AMR-WB 一个语音帧的时长为 20 ms,在实际的听觉过程中,人耳无法准确地感受每一帧的质量,人耳可以感受的语音段包含多个帧. 基于此,笔者提出了基于自由分段的两级时域联合方法,该方法包括短时时域联合和长时时域联合两个阶段,在短时时域联合阶段,将语音序列划分为变长的帧组,联合各帧的质量得到帧组的质量;长时时域联合则将各帧组的质量联合起来得到语音序列的质量.

2.1　短时时域联合

由于人耳对严重失真的区域较敏感,导致低质量帧对语音质量的影响较大,Minkowski 联合方法[16]可以有效反映这一规律. 这里采用 Minkowski 时域联合方法估计帧组的质量,即

$$D_g=\left(\frac{1}{T}\sum_{t=1}^{T}D_t^{p_1}\right)^{1/p_1}\quad,\tag{1}$$

其中,T 为帧组包含的帧数;D_t 为各帧的失真值;p_1 为 Minkowski 指数($p_1>1$);D_g 为帧组的失真值. 可以看出,这种方法给不同质量的帧分配了不同的权重. 在 p_1 值固定的情况下,D_t 越大的帧,分配到的加权系数越大,突出了严重失真的区域对语音质量的影响. Minkowski 指数 p_1 控制着严重失真的区域对语音质量的影响程度,p_1 越大,严重失真的区域对语音质量的影响就越大.

2.2　帧组的划分

由于人耳往往被低质量的语音部分所吸引,且语音部分几乎承载着全部的信息. 另外研究表明,人耳能够正确感知语音段的时长约为 300 ms[17]. 基于此,提出了以语音段的严重失真区域为中心,将语音序列自由划分为帧组,具体算法如下:

(1) 根据各数据包的载荷长度,区分出语音序列中的语音段和静音段,语音段只包含语音帧,静音段只包含静音帧.

(2) 将静音段的内容平均分配到相邻的两个语音段中,如图 3 所示,插入静音后的语音段依次记为 S_1,S_2,…,S_n,其中 n 为语音段的个数.

(3) 设置一个 320 ms 的时间窗,从第 i(i 初始值为 1) 段 S_i 的起始帧开始,以 1 帧为步长,滑动时间窗,时间窗不能超出 S_i 的范围.

(4) 将时间窗内的帧设为可能的帧组,并根据式(1)计算每个帧组的失真值.

(5) 挑选一个失真最大的帧组作为目标帧组. 如果候选的帧组不止一个,就分别计算每个候选帧组的 W_g 值,选择 W_g 值最小的帧组作为目标帧组. W_g 值的计算方法为

$$W_g=\sum_{t=1}^{T}D_t\left|C_t-C_g\right|\quad,\tag{2}$$

其中，T 是帧组包含的帧数；D_t 是各帧的失真值；C_t 表示各帧的中心时刻；C_g 表示帧组的中心时刻.

(6) 将步骤(5)中的目标帧组设为初始帧组.

(7) 设置一个 320 ms 的时间窗，以前一个划分的帧组的右边界为起点，以 1 帧为步长，向右滑动 100 ms，重复步骤(4)和(5)来确定当前的帧组. 如果当前的帧组与前一个划分的帧组之间存在一定的间隔，就将间隔里的帧平均分配到这两个帧组中. 另外，如果剩余的帧不够填充一个时间窗，就将其划分为一个帧组.

(8) 重复步骤(7)，直到初始帧组右边的所有帧都划入帧组.

(9) 设置一个 320 ms 的时间窗，以前一个划分的帧组的左边界为起点(首次进入步骤(9)，则以初始帧组的左边界为起点)，以 1 帧为步长，向左滑动 100 ms，重复步骤(4)和(5)来确定当前的帧组. 如果当前的帧组与前一个划分的帧组之间存在一定的间隔，就将间隔里的帧平均分配到这两个帧组中. 另外，如果剩余的帧不够填充一个时间窗，就将其划分为一个帧组.

(10) 重复步骤(9)，直到初始帧组左边的所有帧都划入帧组.

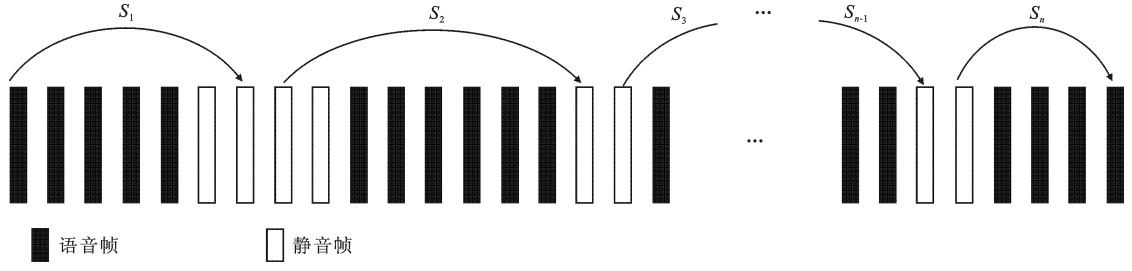


图 3 插入静音后的语音段示意图

2.3 长时时域联合

帧组划分结束后，可以根据式(1)计算出各帧组的质量. 同样为了体现严重失真段对序列影响较大的现象，长时时域联合也采用 Minkowski 方法，即

$$D=\left(\frac{1}{G}\sum_{g=1}^{G}D_g^{p_2}\right)^{1/p_2}, \tag{3}$$

其中，G 为帧组的个数；D_g 是帧组的失真值；p_2 是 Minkowski 指数，且 $p_2>1$；D 是整个语音序列的失真值. 虽然长时时域联合和短时时域联合都是采用 Minkowski 方法，但是，不同阶段的 Minkowski 指数不同. 在[2,10]内遍历 p_1 和 p_2 的值，p_1 和 p_2 分别为 4 和 2 时，文中提出模型的性能最优.

3 实验结果

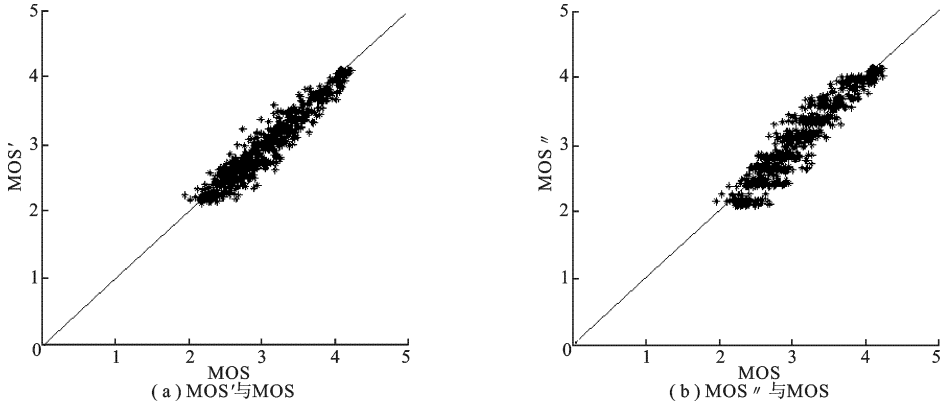
从 ITU-T 语音库中选取 8 个不同内容特性的语音序列：A_eng_f2、A_eng_f5、A_eng_f6、A_eng_f8、A_eng_m1、A_eng_m4、A_eng_m6 以及 A_eng_m7. 这 8 个测试序列与 1.2 节的训练序列是完全不同的. 首先对测试序列进行 AMR-WB 编码，然后将编码码流封装为 RTP/UDP/IP 包，并采用四状态马尔科夫模型[18]模拟各种丢包率下的网络丢包，丢包率分别为 1%、3%、5%、7%、10%、12%、15%以及 20%.

为了验证文中模型的性能，将其与已有的模型 E-Model 和 Q-Model 进行比较，将文中的模型、E-Model、Q-Model 分别记为模型Ⅰ、Ⅱ、Ⅲ. 表 3 列出了这 3 种模型的预测分值与 PESQ 算法的评价分值之间的皮尔森相关系数(Pearson Correlation Coefficient，PCC)和均方根误差(Root Mean Square Error，RMSE). 由表 3 可知，模型Ⅰ的评价性能要优于模型Ⅱ和模型Ⅲ. 其中，PCC 值分别提高了 0.012 9 和 0.012 8，RMSE 值分别下降了 0.023 4 和 0.035 4.

为了更直观地比较模型的性能，图 4 给出了模型Ⅰ和Ⅱ的预测分值与 PESQ 算法的评价分值(MOS)之间的对比散点图，其中图 4(a)表示的是模型Ⅰ的预测分值(MOS′)和 MOS 的相对分布，图 4(b)表示的是模型Ⅱ的预测分值(MOS″)与 MOS 的相对分布. 比较图 4(a)和(b)可以看出，图 4(a)中的分值更加集中在对角线附近，可见模型Ⅰ的预测分值更接近于 PESQ 算法的评价分值.

表 3　模型评价性能比较

语音序列	模型Ⅰ		模型Ⅱ		模型Ⅲ	
	PCC	RMSE	PCC	RMSE	PCC	RMSE
A_eng_f2	0.963 0	0.170 1	0.953 3	0.183 4	0.951 2	0.189 4
A_eng_f5	0.974 8	0.125 5	0.965 8	0.164 1	0.969 8	0.149 3
A_eng_f6	0.968 8	0.142 3	0.965 5	0.160 7	0.953 0	0.195 6
A_eng_f8	0.946 5	0.196 4	0.925 5	0.246 3	0.934 1	0.220 9
A_eng_m1	0.968 7	0.199 5	0.958 9	0.223 5	0.951 3	0.255 5
A_eng_m4	0.962 0	0.223 2	0.951 6	0.216 4	0.952 9	0.243 5
A_eng_m6	0.977 2	0.148 8	0.963 4	0.176 0	0.967 1	0.184 5
A_eng_m7	0.962 4	0.216 3	0.952 4	0.243 6	0.958 2	0.266 0
全部	0.957 0	0.181 0	0.944 1	0.204 4	0.944 2	0.216 4



(a) MOS′与MOS　　(b) MOS″与MOS

图 4　模型Ⅰ、Ⅱ与 PESQ 模型的对比

4　结 束 语

引入人的感知特性,提出了一种能够反映丢包模式对语音质量影响的包层语音质量评价模型.该模型首先通过分析数据包头部,获得编码速率和丢包位置等信息,在此基础上,结合丢包内容检测的结果估计各帧的质量.最后,采用两级时域联合的方法将各帧质量组合得到语音序列的质量.实验结果表明,相比于 E-model,文中提出的模型能够更准确地预测语音质量.

参考文献:

[1] Daengsi T, Preechayasomboon A, Wutiwiwatchai C, et al. A Study of VoIP Quality Evaluation: User Perception of Voice Quality from G.729, G.711 and G.722 [C]//Proc of the 9th Annual IEEE Consumer Communications and Networking Conference. Piscataway: IEEE Computer Society, 2012: 342-345.

[2] Jelassi S, Rubino G, Melvin H, et al. Quality of Experience of VoIP Service: a Survey of Assessment Approaches and Open Issues [J]. IEEE Communications Surveys & Tutorials, 2012, 14(2): 491-513.

[3] Rix A W, Beerends J G, Kim D S, et al. Objective Assessment of Speech and Audio Quality: Technology and Applications [J]. IEEE Trans on Audio, Speech, and Language Processing, 2006, 14(6): 1890-1901.

[4] ITU-T Recommendation P.861. Objective Quality Measurement of Telephone-band (300-3400 Hz) Speech Codecs [S]. Geneva: Telecommunication Standardization Sector, 1996.

[5] 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: Telecommunication Standardization Sector, 2001.

(下转第 94 页)