

一种改进的语音质量感知评估算法

黄石磊^{1,2}, 刘 轶², 程 刚²

(1. 北京大学信息科学技术学院, 北京 100871; 2. 深港产学研基地智能媒体和语音实验室, 广东 深圳 518057)

摘 要: 为提高语音质量客观评估的性能, 提出一种改进的语音质量感知评估(PESQ)算法。该算法利用音节稳定性检测和清浊静音分类的方法, 通过音节的帧间稳定性和损伤参数来描述语音听觉感知所受到的影响, 这些参数对不同的语音段, 如清音、浊音和静音具有不同的特性。实验结果表明, 该算法能在窄带语音上提高 PESQ 得分与主观平均意见分的相关性。

关键词: 语音质量评估; 平均意见分; 语音质量客观评估; 语音质量感知评估; 语音编码; 清浊静音分类

A Modified Algorithm for Perceptual Evaluation of Speech Quality

HUANG Shi-lei^{1,2}, LIU Yi², CHENG Gang²

(1. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China;

2. Intelligent Media and Speech Lab, PKU-HKUST ShenZhen-HongKong Institution, Shenzhen 518057, China)

【Abstract】 This paper proposes a modified Perceptual Evaluation of Speech Quality(PESQ) algorithm to improve the performance of objective speech quality evaluation. Syllable stability detection and Unvoiced/Voiced/Silence(UVS) classification are used in the proposed method. Parameters of stability of frames and syllables distortions are used to describe the effect to hearing perception, and these parameters are different to variant speech segments, especially to unvoiced, voiced and silence segments. Experimental results show that the proposed algorithm for PESQ is able to improve correlation results for narrowband speech and can help to improve other speech quality evaluation.

【Key words】 speech quality evaluation; Mean Opinion Score(MOS); objective evaluation of speech quality; Perceptual Evaluation of Speech Quality(PESQ); speech coding; Unvoiced/Voiced/Silence(UVS) classification

DOI: 10.3969/j.issn.1000-3428.2012.18.005

1 概述

在语音处理和通信领域中, 语音质量是衡量不同语音传输系统性能的重要指标, 因此, 语音质量评估具有重要意义。本文中的语音质量是指语音接收端的信号质量, 在语音传输系统中类似服务质量(Quality of Service, QoS), 语音质量强调信息的受体也就是人的感受, 是一个复杂的生理-心理过程。语音质量的下降(也称为损伤)一般由语音编码造成的失真、信道错误以及其他等因素导致。语音质量的评估一般有2类方法: (1)主观评测方法, 如ITU-T P.800系列标准、平均意见分(Mean Opinion Score, MOS); (2)客观评测方法, 如PSQM、PSQM+、语音质量感知评估(Perceptual Evaluation of Speech Quality, PESQ)等。在ITU-T P.800标准中定义的语音质量MOS主观评测方法^[1], 是评测语音质量的最直接和有效的方法, 虽然其费时费力, 但仍然是其他客观评测标准参考的目标。ITU-T P.862标准中的PESQ方法是ITU-T推荐的语音质量客观评测标准, 它被广泛用于语音编码器、语音传输系统和语音传输设备的语音质量评测之中^[2-3]。

ITU-T P.862的标准PESQ主要用于普通窄带电话线路和窄带(8 kHz 采样)语音编码器。它的核心是把原始输入信号和通过系统(编码器或者通信系统)的信号进行比较, 用一个听觉感知模型计算总体上的失真, 根据文献[4]的结果, PESQ算法的结果和标准MOS测试的相关性可以达到0.935。为了得到和MOS测试更为接近的结果, ITU-T P.862.1提出了MOS-LQO方法^[5], 通过一个基于ACR主观测听的结果得到的映射函数来改进P.862的效果。继而ITU-T又提出P.862.2, 把PESQ的方法扩展到宽带音频系统(通带50 Hz~7 000 Hz)^[6]。近年来对语音质量客观评估的研究, 包括研究主观可懂度(subjective intelligibility)和客观得分的关系^[7]、在VoIP特别是移动互联网的VoIP上的一些改进^[8], 以及将PESQ扩展到更大带宽的音频系统等^[9]。

本文利用语音听觉感知所具有的重要意义的信息, 主要是音节的稳定性和音节损伤, 来对标准的语音质量感知评估(PESQ)算法进行改进。该算法是一种客观的评测方法。

基金项目: 深圳基础研究基金资助重点项目“面向互联网的智能语音搜索和监控关键技术的研究”(JC201005260245A)

作者简介: 黄石磊(1979—), 男, 在站博士后, 主研方向: 语音信号处理, 语音识别, 音频检索; 刘 轶, 副研究员、博士; 程 刚, 工程师、硕士

收稿日期: 2011-12-22

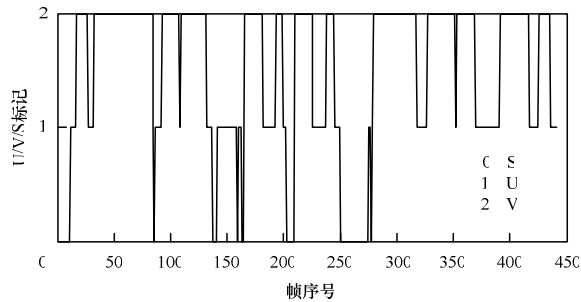
修回日期: 2012-02-02

E-mail: shilei.huang@imsl.org.cn

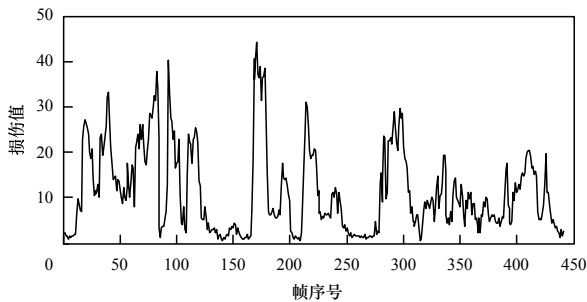
2 系统结构

2.1 基本原理

PESQ 算法的核心思想就是希望通过一种模型描述影响主观测听实验的各种因素在语音质量上的影响。从测听实验的角度出发,对于语音流中的清音(Unvoiced, U)、浊音(Voiced, V)和静音(Silence, S)的听觉感知存在着差异。同时可以看到,即使在同样的传输或者信号处理条件下,以上 U/V/S 类型中每一类的损伤(Disturbance)是不同的。图 1 是一段语音的例子,可以看出,浊音段的损伤明显较大,清音段和静音段部分相对较小。整个语音段的质量可以认为是由不同语音段的感知综合起来决定的。在本文中, U/V/S 判决用来把语音划分为不同的类别,然后分别计算不同帧的损伤,并使用非对称帧损伤值(asymmetrical frame disturbance)来对每一类进行综合。比较而言,基本的 PESQ 算法使用 2 个参数计算最终的总损伤值,而本文算法考虑 3 种语音类型,一共使用了 6 个参数计算最终的损伤值。



(a) U/V/S 分类结果



(b) PESQ 损伤值

图 1 U/V/S 和对应每帧语音的 PESQ 损伤值

对不同损伤程度的语音,听觉感知也存在差异。在一段语音中,具有较大听觉损伤程度的语音往往决定了其邻近相当大范围内语音的感知。PESQ 通过综合时频域上的损伤密度和非对称损伤密度来度量这种损伤。因此,通过 2 层综合来描述这 2 类损伤,首先在音节(依照 P.800 标准中定义,非语言学定义)长度范围内(20 帧,对应大约 320 ms)综合,然后再在长时范围内(10 s)分别进行综合。在一个音节范围内的语音质量损伤反映了一个音节音质的稳定性,并影响最终的听觉感知。本文使用帧间损伤值的标准差来衡量一个音节音质的稳定性,对于稳定和非稳定的音节使用不同的加权对分割的秒内损伤值进行综合。图 2 显示了 U/V/S 中每一类基于音节的帧损伤标准差,其

值越大则音节越不稳定。在图 2 中,通过在数据库中找到一部分纯粹 U/V/S 的“音节”,并统计其音节损伤的标准差来确定恰当的阈值。

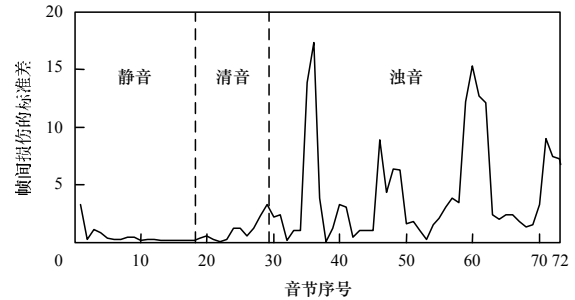


图 2 U/V/S 每一类对应的标准差(例子)

2.2 改进算法的总体框架

本文使用 2 种机制,包括 U/V/S 分类和加权的综合来改进 PESQ 算法。改进算法的总体框架见图 3。

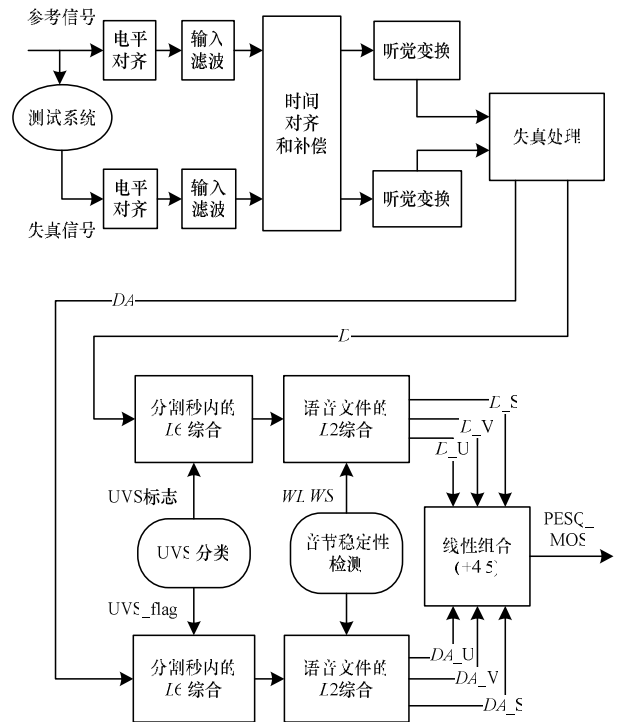


图 3 本文 PESQ 算法的框图

3 算法实现

3.1 基于 U/V/S 分类的 PESQ 计算

考虑到计算复杂度,在分割的时间间隔内使用一个简单的值来进行 U/V/S 分类。U/V/S 分类通过将短时对数能量(E)、短时过零率(Z)与给定的统计门限相比较来进行。

标准的帧损伤综合不区分这些类别,因此,简单的 U/V/S 分类就可以改进 PESQ 的性能,即使这种判决存在少量的错误,也不会影响到算法性能。根据 U/V/S 标志,损伤密度(D)和非对称损伤密度(DA)对于每一类进行综合,首先基于 320 ms 音节内的 20 个帧使用 L6 范数,然后对于音节使用 L2 范数,由下式实现:

$$D_{\text{indicator}} = \sqrt[p]{\frac{1}{N} \sum_{n=1,2,\dots,N} (D_n)^p} \quad (1)$$

其中, p 是 Lp 范数的阶数, 值为 2 或 6; N 是被综合的帧或者音节的个数, 对于每一类有 50% 的交叠。

最后的 PESQ 分数(PESQ_MOS)是每一类的平均损伤值($D_{\text{indicator}}$)和平均非对称损伤值($DA_{\text{indicator}}$)的线性组合:

$$\begin{aligned} MOS_{\text{PESQ}} = & 4.5 - D_{\text{weight}(S)} \times D_{\text{indicator}(S)} - A_{\text{weight}(S)} \times DA_{\text{indicator}(S)} - \\ & D_{\text{weight}(U)} \times D_{\text{indicator}(U)} - A_{\text{weight}(U)} \times DA_{\text{indicator}(U)} - \\ & D_{\text{weight}(V)} \times D_{\text{indicator}(V)} - A_{\text{weight}(V)} \times DA_{\text{indicator}(V)} \end{aligned} \quad (2)$$

每一类的 D 或者 DA 的加权通过 6 个因子的回归分析得到, 且加权系数事先进行训练。

3.2 基于音节稳定性的损伤加权综合

在 U/V/S 分类的基础上, 每一类的音节被分为稳定和稳定的, 语音文件激活部分的综合加权值根据音节的稳定性进行改变, 并根据帧损伤值得标准差进行测量。如果标准差大于门限值(通过实验设定为 5), 则音节是稳定, 否则是非稳定的。

通过这种方法, 当对分割的秒内损伤值执行 $L2$ 范数时, 式(1)变为:

$$D_{\text{indicator}} = \sqrt{\frac{WL \times \sum_{n=1,2,\dots,N_L} (D_n)^2 + WS \times \sum_{n=1,2,\dots,N_S} (D_n)^2}{N_L + N_S}} \quad (3)$$

其中, WL 和 WS 是非稳定和稳定音节的综合加权值; N_L 和 N_S 是非稳定和稳定音节的个数。

3.3 加权值的优化

式(2)中每一类 D 和 DA 的加权值以及式(3)中的综合加权值 WL 和 WS 应当通过叠代的训练和测试来获得。训练数据库和测试数据库分别取自大量不同的 ACR 主观听音测试分数和样本。在叠代训练过程中, WL 和 WS 范围分别为 0.1~0.9 以及 0.9~0.1, 计算 10 组 D 和 DA 的加权值。然后对于每一组加权值, 伯松公式的相关系数由叠代测试计算。本文加权值用客观和主观分数间的最大相关系数来优化。

从表 1 中可以看到, 在最大相关系数的情况下, WL 和 WS 应当分别设置为 0.8 和 0.2, 这样式(2)中相应的加权值要有所改变。

表 1 不同综合加权值的叠代搜索结果

WL	WS	相关系数
0.1	0.9	0.944 8
0.2	0.8	0.941 6
0.3	0.7	0.943 3
0.4	0.6	0.946 9
0.5	0.5	0.950 8
0.6	0.4	0.954 5
0.7	0.3	0.957 6
0.8	0.2	0.958 9
0.9	0.1	0.950 8

4 实验结果与分析

将标准的 PESQ 算法(记为 Original)、基于 U/V/S 分类的改进的算法(记为 Method I)以及基于 U/V/S 分类和音节稳定性的改进算法(记为 Method II)使用相关系数和残

余误差分布来进行比较, 并且使用影射函数 MOS_LQO^[4]来与 PESQ_MOS 进行比较。窄带语音的测试数据库和主观分数取自于合作单位执行的汉语的 ACR 主观听音测试, 包含各种测试条件例如不同的语音编码器、背景噪声条件, 以及由不同的 MNRU 值设置的损伤语音。为了使 PESQ 有效, 客观分数被用一个单调的三阶多项式回归^[3]映射为主观分数, 这样可以在 MOS 域内进行比较。

4.1 相关系数的结果

PESQ 和主观分数的拟合接近度由 P.862 中伯松公式定义的相关系数来衡量:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (4)$$

其中, x_i 是条件 i 下的 MOS; \bar{x} 是条件 MOS 的均值; y_i 是每个条件 i 下的映射条件均值 PESQ 分数; \bar{y} 是预测的条件 MOS 分数 y_i 的平均值。

从表 2 中可以看出, 当 U/V/S 分类用于改进 PESQ 算法时, 相关系数有明显的提升。当使用音节稳定性来改进基于 U/V/S 分类的 PESQ 算法时, 相关系数继续表现较好。本文算法也可以被用来改进 MOS_LQO 分数。ITU-T P.862.1 中 MOS_LQO 的映射函数要比 ITU-T P.862 中标准的 PESQ MOS 的表现好。

表 2 不同相关系数的实验结果

方法	和主观 MOS 的相关系数	
	无 MOS_LQO	有 MOS_LQO
Original	0.828 7	0.853 1
Method I	0.949 2	0.950 2
Method II	0.958 9	0.952 0

4.2 残余误差分布的表现

主观分数和客观分数的残余误差分布用来作为客观和主观接近程度另一种表示, 绝对残余误差定义为:

$$|e_i| = |x_i - y_i| \quad (5)$$

从表 3 中的 PESQ_MOS 来看, 残余误差分布显示了当使用改进算法时, 绝对误差小于 0.25 MOS(5 分制下)的比例占 39.58%, 绝对误差小于 0.5 MOS(5 分制下)的比例占 76.04%。残余误差分布表明, 在采用改进方法的情况下, 更多的得分分布在和主观分数较接近的区间里。本文算法比标准的 PESQ 算法表现更好, 同样可以改进 MOS_LQO。

表 3 残余误差分布的表现

方法	残余误差范围	出现误差的比例/(%)	
		无 MOS_LQO	有 MOS_LQO
Original	<0.25	30.21	33.85
	<0.50	58.85	59.90
Method I	<0.25	39.06	40.10
	<0.50	75.00	75.25
Method II	<0.25	39.58	39.58
	<0.50	76.04	76.56

(下转第 25 页)