一种改进的单声道混合语音分离方法

李 鹏 [1] 关 勇 [2] 刘文举 [2] 徐 波 [1,2]

摘 要 在回顾了基于语音客观质量评估和计算听觉场景分析的单声道混合语音分离方法的基础上，针对该方法所采用的 ITU-T P.563 语音客观质量评估标准存在的使用限制以及计算量大的缺点, 提出了一种采用基于时域包络表示的语音客观质量评估算法来替代 P.563 算法的单声道混合语音分离方法. 该方法在几乎不降低原方法分离性能的前提下, 大大节约了算法运行所需的时间和资源消耗.

关键词 语音分离, 语音客观质量评估, 计算听觉场景分析, 信噪比, 时域包络
中图分类号 TP391

A Modified Monaural Mixture Speech Separation Method

LI Peng[1] GUAN Yong[2] LIU Wen-Ju[2] XU Bo[1,2]

Abstract We firstly review the monaural speech separation method based on objective quality assessment of speech (OQAS) and computational auditory scene analysis (CASA). Considering of the defects of application limitations and time consuming of the employed ITU-T P.563 algorithm, we then propose an alternative method which combines CASA with the temporal envelope representations based OQAS algorithm. The proposed method greatly reduces the operation time and resource requirement, yet almost does not decrease the performance of separation.

Key words Speech separation, objective quality assessment of speech, computational auditory scene analysis, signal to noise ratio (SNR), temporal envelope

在语音信号处理中, 一个重要的问题就是如何从混合语音信号中分离出感兴趣的语音. 这方面的研究在语音识别、多媒体检索、语音增强等领域都有着重要的意义[1]. 一些方法通常被用来进行语音分离, 例如盲源分离[2]、空间滤波[3] 等. 这些方法需要多路语音信号输入. 但是, 许多实际应用需要提供单声道的语音信号分离解决方案. 由于该情况下只有一个传感器信号可以利用, 因此单声道语音分离问题是一个非常具有挑战性的课题, 至今仍是研究人员重点关注的对象.

虽然单声道语音分离仍然是一个充满挑战性的课题, 但是人类的听觉系统还是展现出了出众的单声道语音分离能力. 人类听觉系统的这一能力也促使研究人员进一步加深了对人类听觉感知机理的研究. 1990 年, Bregman 首先提出了听觉场景分析 (Auditory scene analysis, ASA) 的概念[4], 为单声道语音分离问题提供了一个新的思路, 同时也带动了计算听觉场景分析 (Computational auditory scene analysis, CASA) 研究的发展. 许多基于 CASA 的单声道语音分离系统相继被提出[5−11].

最初的 CASA 研究主要采用原始的数据驱动 (Primitive data-driven) 的方法. 近十年来, CASA 的研究重点转向基于知识的图式驱动 (Knowledge-based schema-driven) 方法. 越来越多的高层知识, 例如语音识别中使用的声学模型、声源特性以及声源方位等[7, 9, 12−13], 被引入到原始的 CASA 系统中来指导分离. 但是, 有关语音感知质量方面的知识仍旧未被 CASA 系统所有效利用. 另一方面, 在大多数 CASA 系统中, 对系统性能的评估往往基于信噪比 (Signal to noise ratio, SNR). 然而, 信号的信噪比越高并不一定意味着其感知质量越好. 因此, 寻找一种有效的语音感知质量测量方法并将其以一种适当的方式结合到 CASA 系统中来改善分离语音的信噪比和感知质量, 将极大地推动 CASA 研究的发展.

本文的结构安排如下: 第 1 节回顾了基于语音客观质量评估 (Objective quality assessment of speech, OQAS) 和计算听觉场景分析相结合的单声道语音分离方法; 在此基础上, 第 2 节指出了该方法所采用的 ITU-T P.563 语音客观质量评估算

收稿日期 2007-05-17 收修改稿日期 2008-11-20
Received May 17, 2007; in revised form November 20, 2008
国家重点基础研究发展规划 (973 计划) (2004CB318105), 国家高科技研究发展计划 (863 计划) (2006AA010103, 2006AA01Z194) 资助
Supported by National Basic Research Program of China (973 Program) (2004CB318105), and National High Technology Research and Development Program of China (863 Program) (2006AA010103, 2006AA01Z194)
1. 中国科学院自动化研究所数字内容技术研究中心 北京 100190 2. 中国科学院自动化研究所模式识别国家重点实验室 北京 100190
1. Digital Content Technique Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100190 2. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190
DOI: 10.3724/SP.J.1004.2009.01087

法[14] 在使用条件和计算消耗方面的缺点, 并提出了一种采用基于时域包络表示的语音客观质量评估算法[15] 替代 ITU-T P.563 算法构建 OQAS 与 CASA 结合系统的改进方法; 第 3 节对提出的改进方法进行了评估, 并比较了改进方法与其他分离方法之间的性能, 第 4 节对全文进行了总结.

1 基于 OQAS 与 CASA 相结合的单声道语音分离方法

事实上, 语音质量是一种主观意见, 它主要基于听者自身对实际听到的语音信号的反应, 因此对语音质量的评估以主观的评估方法最为可靠. 但是主观评估方法需要耗费大量的时间和金钱, 并不适合大量频繁、快速的应用, 因此实际应用中多采用客观评估的方法来反映语音信号的主观评分等级. 一般而言, 语音质量的客观评估方法可以分为侵入式 (Intrusive) 和非侵入式 (Non-intrusive) 两种[16]. 侵入式方法依靠参考语音和测试语音之间某种形式的距离特性来预测主观平均观点得分 (Mean option score, MOS). 非侵入式方法则仅依据测试语音来预测语音的质量, 因而更加具有挑战性. 在语音分离应用中, 由于无法获得参考语音信号, 因此只能使用非侵入的方法来评估语音的感知质量.

基于上述分析, 文献 [11] 中提出了一种基于计算听觉场景分析和语音客观质量评估相结合的多阶段分离系统框架 (如图 1 所示). 在该框架下, 选择了 ITU-T P.563 标准作为其中 OQAS 算法的具体实现, 将其与 HuWang 系统[10] 这一代表性的 CASA 系统在初始分离阶段和最终分离阶段进行了结合, 实现了将语音质量信息应用到分离过程中的目标, 并取得了良好的分离效果.

具体而言, 在分解和特征提取阶段, 一组听觉滤波器被用来在连续的时间帧上分析输入混合信号. 经过这一处理, 输入信号被分解为二维的时-频图. 图中的每一个单元被称为一个时频单元, 分别对应某个滤波器的某个时间帧. 之后提取出滤波器响应的自相关、滤波器响应的包络自相关、互通道相关以及每个时间帧的粗略的主导基音等特征, 这些特征将为后续阶段所使用.

在初始分离阶段, 系统将产生一些片段. 片段是比时频单元更大的听觉场景组成成分, 它由时频单元在空间上相邻的区域组成. 这种片段结构包含了人类听觉场景分析所作用于时域和频域的基本的邻近法则. 根据前一阶段提取出来的粗略的基音轮廓, 这些片段接下来被分组到分别对应于目标语音和干扰的初始前景流和背景流中. 由于干扰的存在, 粗略的主导基音可能不够精确, 因此前景流中可能遗失了一些目标语音且包括了一些干扰. 为了减弱不准确主导基音的影响, 这里引入了 ITU-T P.563 语音客观质量评估算法, 其输出被作为标准来判断分配到前景中的片段的分配准确性以获得更准确的分类. 经过该处理, 仍旧保留在前景中的那些片段将更接近于来自于目标声源, 它们将在后续处理中被用来跟踪更加准确的主导基音.

在基音跟踪和时频单元标记阶段, 从前景流中估计出来的目标语音的基音被用来标记时频单元是语音占主导还是干扰占主导.

在最终分离阶段, 初始分离阶段产生的片段根据时频单元的标记被重新组织到前景和背景中. 一些因为不准确的主导基音引起的初始分组错误在此得到了修正. 此外, 还组成了一些由对应于目标语音不确定谐波的时频单元结合成的片段, 这些片段被分配到前景流中. 这里, 与初始分离阶段相同, ITU-T P.563 语音客观质量评估算法被引入到 CASA 系统中, 并作为标准来判断那些由无法分配到前景中的单元组成的片段是否可以被分组到背景中. 接下来, 前景流逐渐扩展以包括那些被标记为语音主导的邻近的时频单元.

最后的再合成阶段, 使用 Weintraub[17] 提出的方法, 从前景流中合成分离后的语音波形. 合成过程可以看作是一个二值掩蔽过程. 其中, 目标语音占主导的单元被标记为 1, 而干扰占主导的单元被标记为 0. 然后, 混合信号中对应于掩蔽值 1 的声学能量被保留下来, 对应于掩蔽值 0 的声学能量被拒绝. 该处理的详细细节可参考文献 [6, 8, 17].

2 改进的基于 OQAS 与 CASA 相结合的单声道混合语音分离方法

2.1 OQAS 与 CASA 结合方法存在的问题

值得注意的是, 在前述 OQAS 与 CASA 相结合的方法中, 用来检验分离阶段中前景流和背景流的划分是否正确的客观语音质量评估算法 — ITU-T P.563 算法虽然能够提供较为准确的语音客观质量估计, 但是该算法在使用中对测试语音具有较多的限制. 测试语音不仅需要满足激活语音的最小长度为 3 秒、最大信号长度为 20 秒的要求, 还要满足最小语音激活率为 25 %、最大语音激活率为 75 % 等要求[14]. 为此, 很多语音在评估前不得不进行相应的处理 (例如文献 [11] 中对被测信号进行三遍延拓处理) 才能够被评估, 因而大大降低了系统的使用灵活性. 此外, 由于 ITU-T P.563 算法在评估语音客观质量时, 共需要计算 51 个表征信号特性的参数, 因此算法本身的计算量较大. 这一点在需要评估多个片段对语音质量影响的情况下, 无疑增加了算法

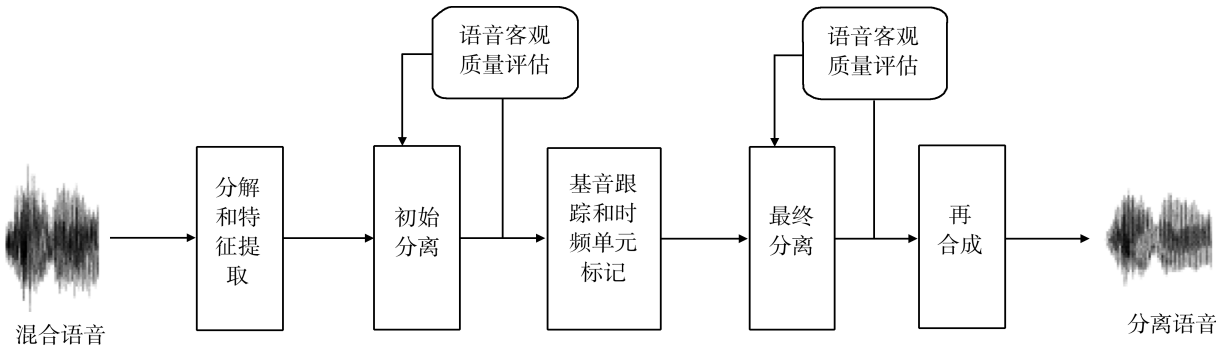

图 1　基于 OQAS 和 CASA 相结合的单声道混合语音分离系统框架

Fig. 1　Diagram of the monaural speech separation system based on OQAS and CASA

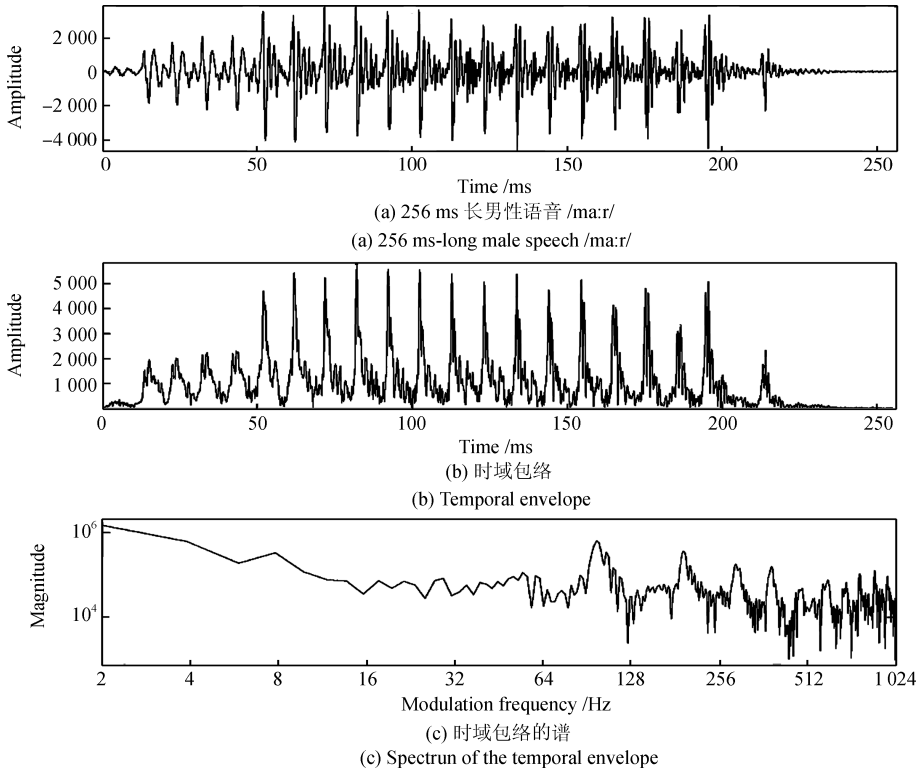

图 2　语音信号及其时域包络

Fig. 2　An example of speech signal and its temporal envelope

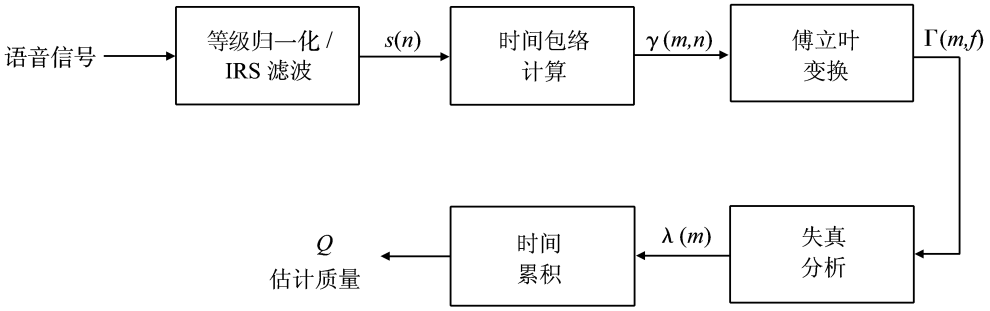

图 3　基于时域包络表示的客观语音质量评估算法结构图

Fig. 3　Block diagram of the speech quality estimation algorithm based on temporal envelope representations

的处理时间, 降低了算法实际应用的可能性.

2.2 改进措施

为了解决上述问题, 本文提出了一种在上述 CASA 与 OQAS 的相结合的系统框架中采用基于时域包络表示的语音客观质量评估算法[15] 作为 OQAS 算法替代 ITU-T P.563 算法来提高分离系统的灵活性和快速性的新方法. 所使用的基于时域包络表示的语音客观质量评估算法不仅克服了 ITU-T P.563 算法对测试语音要求较多的缺点, 而且大大减少了系统在计算时间和资源方面的消耗. 使用该算法对信号进行分离的结果表明, 算法具有很高的实用价值, 能够给出较为满意的分离结果.

2.3 基于时域包络表示的语音客观质量评估

2004 年, 朗讯实验室的 Kim 指出, 语音时域包络表示反映了人类听觉系统和语音产生系统的感知特性, 为非侵入的语音客观质量评估提供了有用的线索[15]. 在此基础上, Kim 提出了一种基于时域包络表示的语音客观质量评估算法[15]. 算法的具体原理如下文所述.

2.3.1 时域包络和调制谱

有限带宽信号 $s(n)$ 通常可以用它的时域包络和载波信号来表示:

$$s(n) = \gamma(n)\cos\phi(n) \tag{1}$$

其中,

$$\gamma(n) = \sqrt{s^2(n) + \hat{s}^2(n)} \tag{2}$$

$$\phi(n) = \arctan\frac{\hat{s}(n)}{s(n)} \tag{3}$$

这里, $\hat{s}(n)$ 是信号 $s(n)$ 的 Hilbert 变换, $\gamma(n)$ 为时域包络, $\phi(n)$ 为瞬时相位.

图 2 给出了语音时域包络的实例. 其中, 图 2 (a) 是男性语音/ma:r/的语音片段, 图 2 (b) 是该段语音的时域包络, 图 2 (c) 是该段语音时域包络的谱. 从图中可以看出, 时域包络中移除了语音的好的载波结构, 而只描绘了包络的波动情况. 它清晰地反映了 100 Hz 附近的声门激励产生的调制成分信息以及在更低频率上由人的声道产生的调制成分信息.

事实上, 时域包络是语音信号的一个非常有趣的表现. 在某种意义上, 它被认为与语音的可懂度和质量等许多感知属性直接相关[18−19]. 在心理生理学中, 时域调制变换函数 (Temporal modulation transfer function, TMTF) 是一个被广泛接受的概念, 它被用来刻画人类对激励信号时域包络的敏感性. 对人类进行测试的实验结果表明, 人类对调制的敏感性可以用一个截止频率约为 50 Hz 的低通滤波器来模拟. 这意味着人类的听觉系统在感知高调制频率信号变化方面是比较迟缓的.

2.3.2 算法介绍

图 3 给出了基于时域包络表示的非侵入 (Non-intrusive) 式客观语音质量评估算法的结构图. 算法首先使用 ITU-T P.56 标准[20] 将语音信号归一化到 −26dBov, 然后对归一化后的语音信号进行 IRS 滤波[21]. 在完成上述处理后, 式 (2) 被用来计算信号的时域包络 $\gamma(n)$. 计算出信号的时域包络后, 再将其按照帧长 128 ms、帧移 64 ms 进行分帧, 并对分帧后的每帧信号加 128 ms 的 Hamming 窗, 以获得相应第 m 帧的时域包络 $\gamma(m;n)$. 这里, 窗长是根据经验确定的. 之所以使用比较长的窗长是为了获取合适的频率分辨率.

有了每一帧的时域包络, 可以通过傅立叶变换得到第 m 帧的调制谱 $\Gamma(m,f)$:

$$\Gamma(m,f) = |\mathcal{F}\{\gamma(m;n)\}| \tag{4}$$

其中, f 代表调制频率.

对人类听觉系统的研究表明, 人类很可能利用调制谱来确定语音的质量; 而且, 位于特定调制频率区域内的谱的成分信息比其他频率更易受到影响. 基于上述情况, 这里将这种与失真有关的频率区域设定在 30 Hz ∼ 50 Hz 范围内. 之所以选择这一范围, 是因为:

1) 人类发声系统机械运动的速度在 2 Hz ∼ 30 Hz 范围内;

2) 人的调制检测呈现出截止频率约为 50 Hz 的低通特性.

如果用 F_D 表示上述与失真有关的调制频率区域, 那么测试语音第 m 帧的感知失真可以定义为:

$$\varepsilon(m) = \int_{F_D}\frac{\Gamma^2(m,f)}{\Gamma^2(m,0)}\,\mathrm{d}f \tag{5}$$

它是区域 F_D 内归一化的调制谱能量. 接下来, 借助语音质量能够由负的失真程度来估计的想法, 可以定义对数尺度下的质量指标:

$$\lambda(m) = -\log[1+\varepsilon(m)] \tag{6}$$

计算出每帧的质量指标 $\lambda(m)$ 后, 需要对其在语音的全部时间帧上进行累加以获得与主观 MOS 评分相对应的客观评估结果. 在算法中, 语音客观质量评估结果是通过对调制能量谱中具有较高直流分量且超过一定门限的帧使用 L_3 范数得到的. 其计算公式如下:

$$Q = \left[\frac{1}{T_s}\sum_{\substack{m \\ P(m)>P_{TH}}}^{\text{s.t.}}\lambda^3(m)\right]^{\frac{1}{3}} \tag{7}$$

其中, $P(m) = \log\Gamma(m, 0)$ 是调制能量谱的直流分量. P_{TH} 是用于确定能听到的帧的门限, T_S 是能听到的帧的数量. 这里, P_{TH} 的值依据经验设定. 经过式 (7) 处理后, 就可以得到估计出的语音客观质量 Q.

对基于时域包络表示的语音客观质量评估算法的实验评估结果表明, 算法的性能达到了带参考语音的语音客观质量评估标准 ITU-T P.862 性能的 96.3 %[15], 具有很好的预测准确性.

3　评估与比较

本部分使用标准数据集从信噪比 (SNR) 及语音感知质量两方面对提出的方法进行了评估, 同时比较了所提出的方法与其他语音分离方法的性能.

3.1　评估

与文献 [11] 相同, 这里仍使用被 CASA 评估所广泛采用的英国谢菲尔德大学 Cooke 搜集的由 10 句浊音句子与 10 种不同干扰噪声组成的 100 句混合语音组成的数据集[5](事先降采样到 8 kHz), 从信噪比及语音感知质量两方面对基于时域包络表示的语音质量评估算法在语音分离系统中的实际应用效果进行评估.

表 1 给出了不同干扰情况下分离语音以及原始混合语音的信噪比. 其中, Mixture 为原始混合语音的结果, HuWang为 HuWang 系统的结果, System 1 为基于 ITU-T P.563 标准的系统结果, System 2 为基于时域包络表示算法的系统结果. 表中最后一行列出了各种噪声条件下的平均信噪比. 从表 1 中可以看出, 处理后的分离语音的信噪比相比原始混合语音在所有干扰条件下均得到了改善, 平均信噪比提高约为 9.3 dB. 与文献 [11] 中提出的方法类似, 本文提出的方法对于那些与目标语音谱没有很大交叠的干扰类型 (例如 N5), 能够获得较高的信噪比改善; 而对于与目标语音谱具有明显交叠的干扰类型 (例如 N8 和 N9), 信噪比的改善较低.

表 1　SNR 结果

Table 1　SNR results

SNR	Mixture	HuWang	System 1	System 2
N0	−7.380	10.330	11.129	7.215
N1	−8.269	3.346	3.507	2.975
N2	5.474	14.251	14.411	14.871
N3	0.803	5.094	5.218	4.251
N4	0.679	1.095	6.669	6.901
N5	−9.999	12.869	12.933	13.003
N6	−1.609	15.213	14.662	13.158
N7	3.842	9.040	9.391	9.735
N8	9.526	12.556	11.506	13.210
N9	2.749	5.100	3.964	3.692
平均	−0.418	8.889	9.339	8.901

为了评估分离语音的感知质量, 这里依旧采用了 ITU-T P.862 和 MOS 这两种客观和主观评价标准. 评估结果如表 2 和表 3 所示. 从表中可以看出, 本文提出的分离方法能够明显改善分离语音的感知质量. 特别是在表 3 中, 在所有干扰情况下通过主观测量方法得到的 MOS 值均得到了一定程度的提高.

表 2　P.862 结果

Table 2　P.862 results

P.862	Mixture	HuWang	System 1	System 2
N0	2.239	2.630	2.673	2.413
N1	1.288	0.583	0.699	0.947
N2	1.548	2.266	2.247	2.282
N3	1.654	0.922	1.087	1.079
N4	1.507	0.910	1.162	1.452
N5	0.307	2.271	2.255	2.301
N6	2.056	2.459	2.372	2.191
N7	2.010	1.652	1.761	1.779
N8	2.329	2.000	1.886	2.106
N9	2.263	1.673	1.306	1.516
平均	1.720	1.737	1.745	1.807

表 3　MOS 结果

Table 3　MOS results

MOS	Mixture	HuWang	System1	System2
N0	1.59	3.26	3.36	3.17
N1	1.03	1.41	1.50	1.46
N2	1.67	2.95	3.13	3.16
N3	1.28	1.90	2.11	2.13
N4	1.19	1.41	2.02	2.06
N5	1.36	2.89	3.04	3.14
N6	1.33	2.86	3.01	2.93
N7	1.25	2.34	2.49	2.65
N8	1.30	2.39	2.35	2.48
N9	1.26	2.16	1.96	2.10
平均	1.326	2.357	2.497	2.528

3.2　比较

由于本文研究的目标是在几乎不影响基于 CASA 与 OQAS 相结合的分离方法性能的前提下, 提高分离的灵活性和执行速度, 因此在分离效果评估的基础上, 进一步比较了本文提出的改进方法与文献 [11] 中提出的方法以及 HuWang 方法之间的分离效果. 详细的评估结果如表 1 ∼ 3 所示. 从表中可以看出, 相比于使用 P.563 标准, 使用基于时间包络表达的语音客观质量评估算法使得分离信号的

平均 SNR 从 9.339 dB 变为 8.901 dB; 平均客观语音质量 (P.862 估计得到) 由 1.745 上升为 1.807; 而 MOS 值则由 2.497 上升为 2.528. 由此可以得出如下结论: 使用基于时域包络表示的客观语音质量评估算法代替 ITU-T P.563 算法后, 分离方法的整体分离性能无论在信噪比还是感知质量方面均未发生明显变化.

为了说明改进方法在速度方面的优势, 我们分别对使用 P.563 标准和使用基于时间包络表达的语音客观质量评估算法的分离方法完成全部分离任务所需的时间进行了统计. 统计的结果如表 4 所示, 其中 System 1 为基于 ITU-T P.563 算法的系统, System 2 为本文提出基于时域包络表示算法的系统. 从表 4 中可以看出: 使用基于时域包络表示的语音客观评估算法后, 完成全部分离任务所需的的时间大约相当于使用基于 P.563 算法的 16.5 %, 极大地提高了算法的执行速度和系统实际应用的可能性.

表 4 算法运行时间比较

Table 4 Comparison of the operation time

	System 1	System 2
平均每句处理时间 (min)	3.95	0.65

4 结论

本文提出了一种在基于 CASA 与 OQAS 相结合的单声道混合语音分离方法中采用基于时域包络表示的语音客观质量评估算法替代 ITU-T P.563 标准来提高分离方法的灵活性和快速性的新方法. 所使用的基于时域包络表示的语音客观质量评估算法不仅克服了 ITU-T P.563 算法对测试语音要求较多的缺点, 而且大大减少了分离方法在资源方面的消耗. 使用相同的混合语音数据集进行分离实验的评估结果表明, 相比于文献 [11] 中提出的采用 ITU-T P.563 标准的 CASA 与 OQAS 相结合的方法, 本文提出的改进方法不仅能够获得与之相当的分离性能, 而且大大降低了算法运行所需的时间, 因而具有更高的实用价值.

References

1 Wang D L, Brown G J. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. New Jersey: Wiley, 2006

2 Kokkinakis K, Nandi A K. Multichannel blind deconvolution for source separation in convolutive mixtures of speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, **14**(1): 200−212

3 Leukimmiatis S, Dimitriadis D, Maragos P. An optimum microphone array post-filter for speech applications. In: Proceedings of the 9th International Conference on Spoken Language Processing. Pittsburgh, USA: IEEE, 2006. 2142−2145

4 Bregman A S. *Auditory Scene Analysis*. Cambridge: The MIT Press, 1990

5 Cooke M P. Modeling Auditory Processing and Organization [Ph. D. dissertation], University of Sheffield, UK, 1991

6 Brown G J, Cooke M. Computational auditory scene analysis. *Computer Speech and Language*, 1994, **8**(4): 297−336

7 Ellis D P W. Prediction-Driven Computational Auditory Scene Analysis [Ph. D. dissertation], Massachusetts Institute of Technology, USA, 1996

8 Wang D L, Brown G J. Separation of speech form interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks*, 1999, **10**(3): 684−697

9 Hu G N, Wang D L. Auditory segmentation based on onset and offset analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, **15**(2): 396−405

10 Hu G N, Wang D L. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on Neural Networks*, 2004, **15**(5): 1135−1150

11 Li P, Guan Y, Xu B, Liu W J. Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, **14**(6): 2014−2023

12 Godsmark D, Brown G J. A blackboard architecture for computational auditory scene analysis. *Speech Communication*, 1999, **27**(3-4): 351−366

13 Roman N, Srinivasan S, Wang D L. Binaural segregation in multisource reverberant environments. *Journal of the Acoustical Society of America*, 2006, **120**(6): 4040−4051

14 Single Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications, ITU-T Recommendation P.563, International Telecommunication Union, 2004

15 Kim D S. A cue for objective speech quality estimation in temporal envelope representations. *IEEE Signal Processing Letters*, 2004, **11**(10): 849−852

16 Rix A W, Beerends J G, Kim D S, Kroon P, Ghitza O. Objective assessment of speech and audio quality — technology and applications. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, **14**(6): 1890−1901

17 Weintraub M. A Theory and Computational Model of Monaural Auditory Sound Separation [Ph. D. dissertation], Stanford University, USA, 1985

18 Drullman R, Festen J M, Plomp R. Effect of temporal envelope smearing on speech reception. *Journal of the Acoustical Society of America*, 1994, **95**(2): 1053−1064

19 Ghitza O. On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. *Journal of the Acoustical Society of America*, 2001, **110**(3): 1628−1640

20 Objective Measurement of Active Speech Level, ITU-T Recommendation P.56, International Telecommunication Union, 1993

21 Specification for An Intermediate Reference System, ITUT Recommendation P.48, International Telecommunication Union, 1988

李 鹏 中国科学院自动化研究所助理研究员. 分别于 2000 年和 2003 年在天津大学获得学士和硕士学位, 并于 2007 年在中国科学院自动化研究所获得博士学位. 主要研究方向为计算听觉场景分析、语音分离、语音增强、语音识别. 本文通信作者.
E-mail: pengli@hitic.ia.ac.cn
(**LI Peng** Assistant research fellow at the Institute of Automation, Chinese Academy of Sciences. He received his B. S. and M. S. degrees from Tianjin University in 2000 and 2003, respectively, and his Ph. D. degree from the Institute of Automation, Chinese Academy of Sciences in 2007. His research interest covers computational auditory scene analysis, speech segregation, speech enhancement, and speech recognition. Corresponding author of this paper.)

关 勇 诺基亚 (中国) 研究中心博士后. 2002 年在清华大学获得学士学位, 并于 2008 年在中国科学院自动化研究所获得博士学位. 主要研究方向为计算听觉场景分析、语音分离、说话人识别、语音合成. E-mail: ext-yong.guan@nokia.com
(**GUAN Yong** Postdoctor in Nokia (China) Research Center. He received his B. S. degree from Tsinghua University in 2002 and Ph. D. degree from the Institute of Automation, Chinese Academy of Sciences in 2008, respectively. His research interest covers computational auditory scene analysis, speech segregation, speaker recognition, and speech synthesis.)

刘文举 中国科学院自动化研究所副研究员. 分别于 1983 年、1989 年、1993 年在北京大学、北京邮电大学、清华大学获得学士、硕士、博士学位. 主要研究方向为语音识别、语音合成、说话人识别、语音转换、计算听觉场景分析.
E-mail: lwj@nlpr.ia.ac.cn
(**LIU Wen-Ju** Associate professor at the Institute of Automation, Chinese Academy of Sciences. He received his B. S., M. S., and Ph. D. degrees from Peking University, Beijing University of Post and Telecommunication, and Tsinghua University in 1983, 1989, and 1993, respectively. His research interest covers speech recognition, speech synthesis, speaker recognition, voice conversion, and computational auditory scene analysis.)

徐 波 中国科学院自动化研究所研究员. 1988 年在浙江大学获得学士学位, 并于 1992 年和 1997 年在中国科学院自动化研究所获得硕士和博士学位. 主要研究方向为数字内容管理、语音识别、语音翻译. E-mail: xubo@hitic.ia.ac.cn
(**XU Bo** Professor at the Institute of Automation, Chinese Academy of Sciences. He received his B. S. degree from Zhejiang University in 1988 and his M. S. and Ph. D. degrees from the Institute of Automation, Chinese Academy of Sciences in 1992 and 1997, respectively. His research interest covers digital content management, speech recognition, and speech to speech translation.)