Hans Journal of Wireless Communications 无线通信, 2013, 3, 149-154
http://dx.doi.org/10.12677/hjwc.2013.36024 Published Online December 2013 (http://www.hanspub.org/journal/hjwc.html)

Study on Objective Speech Quality Assessment Algorithm

Leilei Xiao, Weiwei Zhang

WT & T, Beijing University of Posts and Telecommunications, Beijing
Email: bernabeu_147@foxmail.com

Received: Nov. 13th, 2013; revised: Nov. 15th, 2013; accepted: Nov. 18th, 2013

Abstract: The aim of this paper is to investigate the performance of the latest objective speech quality assessment algorithm. In the communication system, the quality of speech is an important factor to evaluate the performance of the system. In order to achieve speech quality assessment, it is necessary to develop effective speech quality assessment method. The perceptive speech quality objective assessment algorithm is the most useful and convenient method. In this paper we will discuss the PESQ (Perceptual Evaluation of Speech Quality) and POLQA (Perceptual objective listening quality assessment) algorithm, which are the latest ITU standard for evaluating speech quality for communication systems and networks. POLQA is a technology upgrade covering the latest speech coding and network transport technology, with higher accuracy for 3G, 4G/LTE and VoIP networks. We analyze the performance of POLQA, and make a comparison between POLQA and PESQ. From the result of the experiment data, we can conclude that the POLQA performance is better than the PESQ, and the POLQA will replace the PESQ in the future.

Keywords: Objective Speech Quality Assessment Algorithm; PESQ Algorithm; POLQA Algorithm

客观语音质量评估算法的研究

肖累累，张伟伟

北京邮电大学无线理论与技术实验室，北京
Email: bernabeu_147@foxmail.com

收稿日期：2013 年 11 月 13 日；修回日期：2013 年 11 月 15 日；录用日期：2013 年 11 月 18 日

摘　要：本文的目的是研究最新客观语音质量评估算法的表现。在通信系统中，语音的质量对于评估系统的表现是一个主要的因素。为了达到评估语音质量的目的，开发有效的语音质量评估算法是必须的。感知的语音质量客观评估算法是最有用和最便捷的方法。在这篇论文里，我们将要讨论评估通信系统和网路中的语音质量的最新的 ITU 标准 PESQ (语音质量的感知评估)和 POLQA (感知客观语音质量评估)算法。POLQA 是一个技术升级，它能够覆盖最新的语音编码和网络传输技术，对于 3G，4G/LTE 和 VoIP 网络有了更高的准确度。我们分析了 POLQA 的表现，并且将 POLQA 和 PESQ 作了对比。从实验数据的结论来看，我们得出了如下结论：POLQA 相较于 PESQ 有更好的表现，并且即将替代 PESQ。

关键词：客观语音质量评估算法；PESQ 算法；POLQA 算法

1. 引言

随着通信技术的发展，现代通信网络提供了大量的语音服务。语音通信成为了现代生活中最重要的部分之一[1]。由于技术和语音服务的快速发展，通信系统的传输特性的评估和优化变得越来越重要[2]。服务供应商面临着提供高质量的语音通信系统[3]。系统表现的有效的评估变得关键。发展可靠的、便捷的、灵活的语音评估系统成为了一个共同的目标。

语音通信评有两种方法，即主观评估方法和客观评估方法[4]。主观评估是通过主观语音测试获得的。这些测试通常昂贵、耗时并且需要大量的语音测试。所以它不适合实时通信[5]。客观质量评估替代了主观方法[6][7]。并且，它已经成为了主要的质量评估方法。

PESQ 算法是一种应用于通信系统和语音编码的端对端语音质量评估的客观语音质量评估方法。它被核准为 ITU-T Rec. P.862[8]。PESQ 是一个出名的用于语音评估的客观语音质量评估方法。它对于通信延迟和环境噪音具有较好的鲁棒性[9]。

但是，PESQ 算法对于语音质量评估具有一定的局限性[10]。为了结局 PESQ 的局限性，ITU 发展了新的 POLQA 标准，并核准为 ITU-T Rec. P.863[11]。POLQA 是下一代移动语音质量评估标准并且被发展应用于超宽带高清语音，3G，VoLTE (4G)，VoHSPA 和 VoIP。

本文，我们首先介绍了主观语音质量评估方法。并且给出了主观 MOS (平均意见得分)的计算过程和 PESQ 算法的过程。然后我们给出了 POLQA 算法的一个概览。最后，我们通过实验数据分析了 POLQA 算法的表现。

2. 语音质量评估方法

2.1. 主观语音质量评估

MOS (平均意见得分)是应用最广的评估语音质量的一种度量。它是 ITU (国际电信联盟)推荐的。ITU 逐步地提出了一些音视频服务的主观评估方法。比如 ITU-T Rec. P.800[12]，ITU-T Rec.P.830 并且 ITU-T Rec. P.835 给出了语音服务的主观评估方法。

ITU-T Rec.P.800 是最为流行的主观语音质量评估方法。我们简要地解释 ITU-T Rec.P.800 的主观 ACR(绝对等级分)测试方法。ACR 测试方法分为四部分。

第一部分是录制音源。它包括录音环境，录音系统、发送系统，语音材料，录音过程和录音者选择。第二部分是条件选择。这部分包括语音输入和参考条件的选择。第三部分是实验的设计。最后一部分是语音测试过程。它包括语音环境，语音系统，听者选择，意见分标准，数据分析和结果报告。在一个主观测试，测试者试听每个语音样本[13]。之后，测试者根据图 1 对语音样本进行打分。

Score (W_i)	Quality Level	Description
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Figure 1. The score grade of MOS
图 1. MOS 的分数等级

最终的 MOS_{LQS} (Mean Opinion Score Listening Quality Subjective)分数是由下式计算得出：

$$MOS_{LQS} = \frac{1}{N}\sum_{i=1}^{p} W_i N_i \quad (1)$$

其中 N 是总票数，N_i 是一个特定分的数量，W_i 是每个投票的得分，i 是每个等级的得分，p 是总的得分等级，p 的值为 5。

主观评估方法可以直接、准确地反映出用户的主观感受。但是它要求考虑许多因素，实施步骤也较为复杂，而且它耗时、昂贵。近年来，客观质量评估代替了主观质量评估，并且编委了主要的质量评估方法。下面，我们将要讨论客观评估方法。

2.2. 客观语音质量评估

现在有各种各样的客观语音质量评估方法。但是感知域的评估方法是最为成功的客观语音质量方法。

典型的感知评估方法有 PSQM (Perceptual Speech Quality Measure)，PAMS (Perceptual Assessment of Speech Quality)，PESQ (Perceptual Evaluation of Speech Quality)和应用于通信系统和网络中评估语音质量的最新的 ITU 标准 POLQA (Perceptual objective listening quality assessment)。

PESQ 可以用于不同类型通信网络的评估。它考虑了网络延迟，并且应用了听觉和认知建模技术。PESQ 的结构如图 2 所示。

P.862 提供的原始 PESQ 得分为 0.5 到 4.5 分。为了获得可以与 MOS 分值相比较的得分，需要将原始得分映射为 MOS-LQO (MOS-Listening Quality Objective)。映射公式如下[14]：

$$y = 0.999 + \frac{4}{1+\mathrm{e}^{-1.4945x+4.6607}} \quad (2)$$

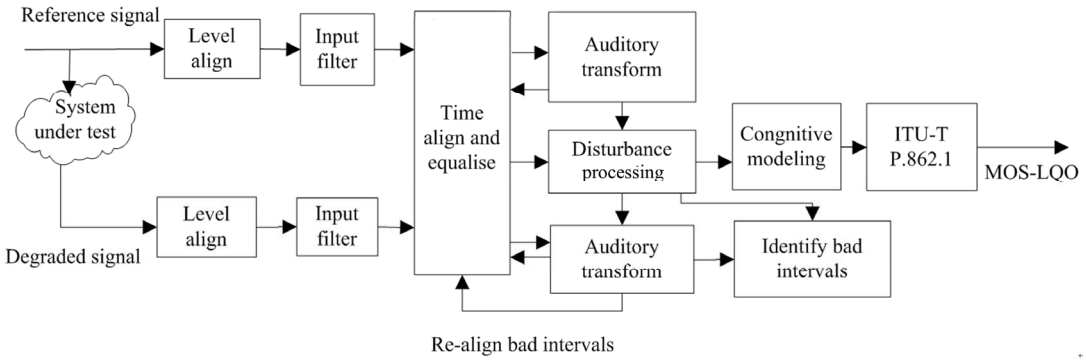

Figure 2. The structure of the PESQ algorithm
图 2. PESQ 算法的结构

PESQ 算法在许多情况下有缺陷。它应用于 CDMA 编码(如 EVRC)时不够准确并且在特定的 GSM/WCDMA 网络条件下过于敏感。PESQ 应用于语音处理(增强)设备(降噪自动增益控制)时同样具有局限性。

此外，语音通信从窄带到宽带甚至是超宽带，PESQ 不能处理超宽带语音信号。为了解决 PESQ 的这些局限性，ITU-T 的 12 研究组自 2006 开始发展新的 POLQA 标准。

POLQA 将提供一个决定移动网络服务的语音质量的新的基准等级。POLQA 已被出版为 ITU-T Rec.P.863。POLQA 是可以覆盖最新的语音编码和网络传输技术的技术升级，在用于 3G，4G/LTE 和 VoIP 网络时具有更高的准确性，并且支持传输高质量语音的网络，而这些语音是之前的电信网络所不能传输。

3. POLQA 算法概览

POLQA 算法概览如图 3 所示。有两个输入信号，即参考信号和衰减信号。二者均为 16 比特 PCM 样本。POLQA 处理包括三个步骤：时间对齐，采样率预估和感知模型。

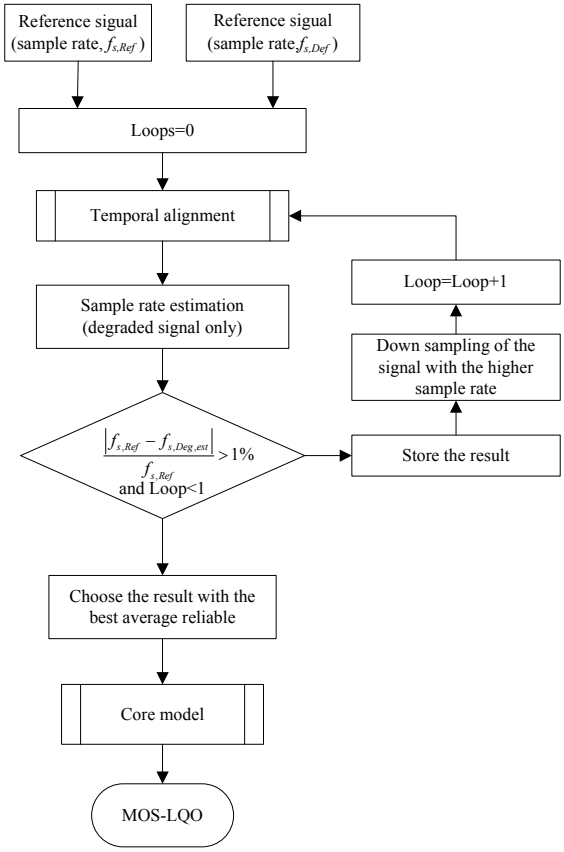

Figure 3. The overview of the POLQA algorithm
图 3. POLQA 算法概览

3.1. 第一步：时间对齐

时间对齐的目的是将信号分割为帧，以计算每一个帧对的时延。时间对齐包括 5 个模块，即滤波，预对齐，粗略对齐，精确对齐和部分组合。

滤波：参考和衰减信号都需要带通滤波。滤波形状取决于工作模式是窄带模式还是宽带模式。

预对齐：衰减信号需要和参考信号对齐。首先，确定延迟极限，即整体延迟搜索范围的一些合理的上限和下限。第二步是估计整体延迟和标识解析点。第三步是获得每个解析点的初始延迟并且计算每个宏帧的信息。

粗对齐：粗对齐是基于每帧得出的。第一部是细分每个信号为特征帧，并且计算每部分的特征。特征帧的长度是独立于宏观帧的长度。粗对准的结果是一个包含了用样本表示的每个宏帧延迟的向量。

细对齐：细对准直接在参考信号和衰减信号可能

的最高分辨率上进行并且它确定每一样本帧的准确的延迟。细对齐的结果是每个宏帧的精确样本延迟值。

部分组合：在这一步，所有具有相同延迟的部分将结合，这意味着整个部分的一套信息(延迟、可靠性、启动、停止、语音活动)被存储。由此产生的信息将被传递给心理声学模型。

3.2. 第二步：采样率估计

由于失真，采样率是不同。这可能导致延迟变化的分离。因此需要估计采样率以补偿播放的参考信号和衰减信号的感知无关差异。采样比率的检测是基于每帧向量和每个语音信号中探测到的活跃部分的延迟所得出的。

3.3. 第三步：感知模型

在感知模型参考信号和衰减信号都转换为内置表示。处理过程中感知模型的细节如下：

常数设置的预计算：参考信号和衰减信号用一个由采样频率决定窗口长度的 FFT 变换到特定的频率域。在转换频率轴到巴克范围后，音高功率密度的峰值振幅就通过乘以一个功率比例因子而被归一化到一个 10^4 级的功率值。

音高功率密度的计算：将赫兹域的频率刻度映射到巴克域内的音高刻度的曲线函数近似于文献中给出的值。作为结果的参考信号和衰减信号分别为音高频率密度 $PPX(f)_n$ 和 $PPY(f)_n$，其中 f 是巴克域频率，脚标 n 代表了帧序。

参考信号的比例调整：参考信号现在处于一个理想的标准而衰减信号代表了回放标准。参考信号向衰减信号按比例调整以补偿标准不同所带来的影响。

噪音补偿：为了解决引入了无声线性频率响应失真的测试系统中的滤波，参考信号在音高功率密度域已被部分过滤。为了进一步纠正线性失真比非线性失真影响较小这一事实，参考信号现在在音高响度域被部分过滤。

最终扰动密度的计算：最终扰动和附加扰动密度每一帧都是整合的，在每一帧通过音高轴得到了两个不同的扰动，一个由扰动得出，一个由附加扰动得出。

最终 MOS-LQO POLQA 分数的计算：原始 POLQA 分数由类似 MOS 的运用了 4 种不同补偿的中间指标得出来的。然后，原始 POLQA 分数通过运用一个优化 ITU-T P.863 数据库组的三阶多项式映射到 MOS-LQO 分数。

4. POLQA 算法性能分析

4.1. POLQA 性能指标

客观准则的准确性通过运用被称为均方根误差的 $rmse^*$ 标准来评估[15]。均方根误差考虑了每个 MOS 分数的置信区间。$rmse^*$ 值的计算只考虑了和目标值附近 ξ 宽带有关的差异。这里的 ξ 定义为主观 MOS 值的 95% 置信区间。$rmse^*$ 值通过预测误差 *Perror* 计算得出。

$$Perror(i) = \max\left(0, \left|\mathrm{MOS}_{\mathrm{LQS}}(i) - \mathrm{MOS}_{\mathrm{LQO}}(i)\right| - ci_{95}(i)\right) \tag{3}$$

其中脚标 i 表示测试环境或者语音样本，ci_{95} 是置信区间。它是基于所有评分都是针对同一文件或者测试环境这一考虑计算得出的。

$$ci_{95} = t(0.05, M)\frac{\sigma}{\sqrt{M}} \tag{4}$$

标准偏差 σ 和单个分数的数量 M 决定了置信区间。对于给定的 M，建议使用精确的 T 值。

最终修正的 $rmse^*$ 值照例计算，但是计算时需通过下面的公式基于 *Perror* 得出。

$$rmse^* = \sqrt{\left(\frac{1}{N-d}\sum_{N} Perror(i)^2\right)} \tag{5}$$

计算每个数据库的 $rmse^*$ 值，并给出预测误差如何超出了 ci_{95}。

这篇论文里，我们使用 $rmse^*$ 值来评估 POLQA 算法的性能。

4.2. POLQA 和 PESQ 性能对比

ITU-T 实施了 POLQA 和 PESQ 算法的性能测试。实验数据从 ITU-T POLQA 中提取[11]。语音数据库包括了窄带语音和宽带语音。$rmse^*$ 值反映了 POLQA 和 PESQ 算法在每种情况下的性能。当 $rmse^*$ 值增加时，预测准确度下降。

ITU-T P.862 提出的 PESQ 和 ITU-T P.863 提出的 POLQA 的 $rmse^*$ 值如图 4，图 5 和图 6 所示。图 4 和图 5 的实验数据是基于窄带信号的，图 6 的实验数据则是基于宽带信号。

为了更清楚地展示结果，PESQ 和 POLQA 的 rmse*值展示于图7，8，9。由图，我们可以发现，POLQA 的 rmse*值小于 PESQ 的 rmse*值。这就是说，POLQA 算法的性能要比 PESQ 算法的性能更为准确。

Set A (narrowband)	ITU-T P.862.1	ITU-T P.863
NB_BT_P862_BGN_ENG	0.1490	0.0981
NB_BT_P862_PROP	0.1603	0.1658
NB_DT_P862_1st	0.1916	0.1473
NB_DT_P862_BGN_GER	0.0973	0.1112
NB_DT_P862_Share	0.1263	0.0895
NB_ERIC_AMR_4B	0.0918	0.1356
NB_ERIC_P862_NW_MEAS	0.2214	0.1767
NB_TNO_P862_KPN_KIT97	0.2967	0.1891
NB_TNO_P862_NW_EMU	0.3017	0.1530
NB_TNO_P862_NW_MEAS	0.2493	0.1654
NB_ITU_SUPPL23_EXP1a	0.1342	0.1184
NB_ITU_SUPPL23_EXP1d	0.0780	0.0676
NB_ITU_SUPPL23_EXP1o	0.1091	0.1096
NB_ITU_SUPPL23_EXP3a	0.1939	0.1660
NB_ITU_SUPPL23_EXP3c	0.1370	0.0862
NB_ITU_SUPPL23_EXP3d	0.1258	0.0585
NB_ITU_SUPPL23_EXP3o	0.1537	0.0569
NB_FT_P563_PROP	0.1139	0.0662
NB_LUC_P563_PROP	0.0632	0.0926
NB_OPT_P563_PROP	0.1150	0.1198
NB_PSY_P563_PROP	0.1623	0.1736
NB_SQ_P563_PROP	0.1915	0.1701
Average	0.1574	0.1235

Figure 4. Performance of PESQ compared to POLQA, NB1
图 4. PESQ 和 POLQA 的性能对比，窄带 1

Database	rmse*3rd	
Set B (narrowband)	ITU-T P.862.1	ITU-T P.863
NB_ATT_iLBC	0.2268	0.1937
NB_ERIC_Field_GSM_EU	0.2401	0.1546
NB_ERIC_Field_GSM_US	0.1986	0.1454
NB_GIPS_EXP1	0.2943	0.1019
NB_QUALCOMM_EXP1b	0.1588	0.1206
NB_QUALCOMM_EXP2b	0.1826	0.1491
NB_QUALCOMM_EXP3w	0.1546	0.0956
NB_8kHz104_ERICSSON	0.3570	0.2840
NB_48kHz404_PSYTECHNICS	0.3260	0.1614
NB_8kHz504_SWISSQUAL	0.4203	0.2311
NB_8kHz_NTT_PTEST_1	0.1073	0.0872
NB_QUALCOMM_EXP4	0.1730	0.1254
NB_QUALCOMM_EXP6a	0.2480	0.2164
NB_QUALCOMM_EXP6b	0.1491	0.1191
NB_16kHz_HUAWEI_1	0.1719	0.1317
Average	0.2272	0.1545

Figure 5. Performance of PESQ compared to POLQA, NB2
图 5. PESQ 和 POLQA 的性能对比，窄带 2

5. 结论

本文，我们研究了客观语音质量评估算法。我们

Set C (wideband)	ITU-T P.862.2	ITU-T P.863
WB_48kHz102_ERICSSON	0.4521	0.1936
WB_16kHz402_PSYTECHNICS	0.3245	0.1839
WB_GIPS_EXP3	0.3467	0.1341
WB_QUALCOMM_EXP1w	0.2606	0.1269
WB_QUALCOMM_EXP3w	0.2852	0.0708
WB_16kHz204_FT_DT	0.4221	0.2319
WB_QUALCOMM_EXP5	0.3236	0.1100
Average	0.3450	0.1502

Figure 6. Performance of PESQ compared to POLQA, WB
图 6. PESQ 和 POLQA 的性能对比，宽带

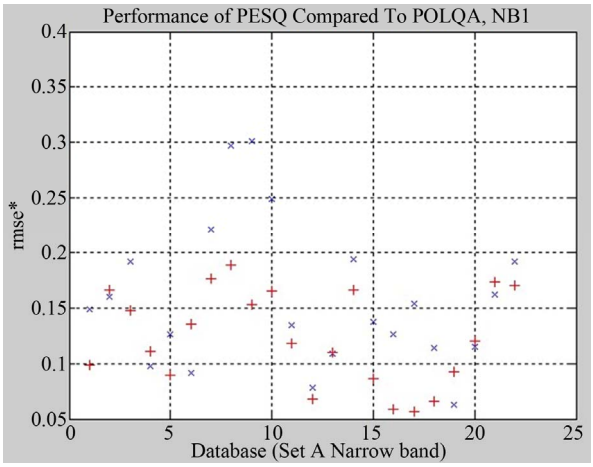

Figure 7. rmse* of PESQ compared to POLQA, NB1
图 7. PESQ 和 POLQA 的 rmse*对比，窄带 1

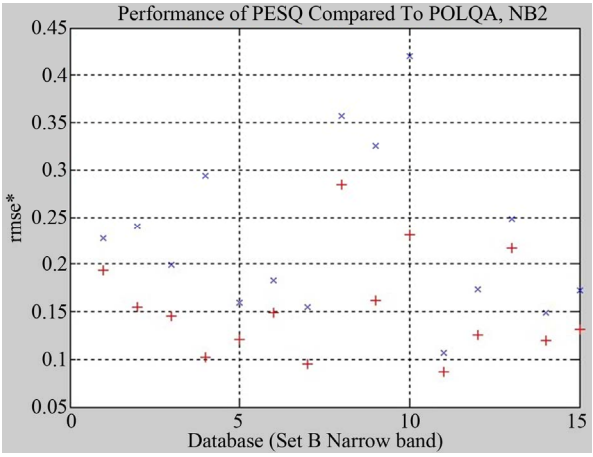

Figure 8. rmse* of PESQ compared to POLQA, NB2
图 8. PESQ 和 POLQA 的 rmse*对比，窄带 2

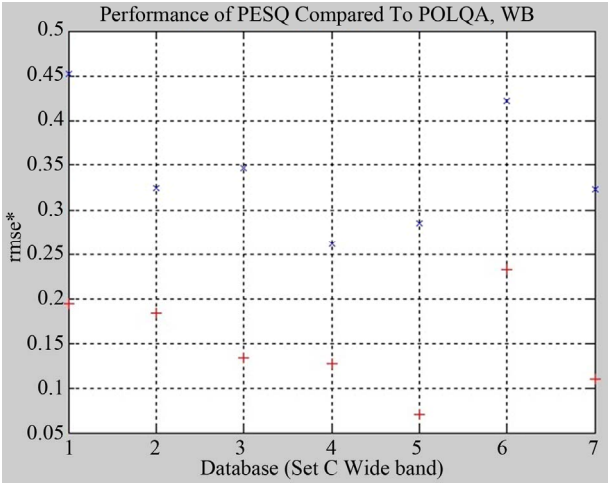

Figure 9. rmse* of PESQ compared to POLQA, WB
图 9. PESQ 和 POLQA 的 rmse*对比，宽带

概览了最新一代的客观算法 POLQA 算法，它是被设计用于克服之前诸如 PESQ 等算法中存在的缺陷的最新算法。然后我们运用了被称为均方根误差的 rmse* 标准测试了 POLQA 的性能。为了知道 POLQA 算法的准确性，我们比较了 POLQA 和 PESQ 算法的 rmse*。从实验数据的结果来看，我们可以得出结论，即 POLQA 的性能比 PESQ 更好，并且 POLQA 即将取代 PESQ。

参考文献 (References)

[1] Shaikh, J., Fiedler, M. and Collange, D. (2010) Quality of Experience from user and network perspectives. *Annals of Telecommunications*, **65**, 47-57.

[2] Jelassi, S., Rubino, G., Melvin, H., Youssef, H. and Pujolle, G. (2012) Assessing the quality of voice communications over internet backbones. *IEEE Communications Surveys & Tutorials*, **14**, 1.

[3] Taal, C.H., Hendriks, R.C., Heusdens, R. and Jensen, J. (2011) An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Audio*, *Speech*, *and Language Processing*, **19**, 2125-2136.

[4] Mowlaee, P., Saeidi, R., Christensen, M.G. and Martin, R. (2012) Subjective and objective quality assessment of single-channel speech separation algorithms. *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Kyoto, 25-30 March 2012, 69-72.

[5] Rix, A.W., Beerends, J.G., Kim, D.-S., Kroon, P. and Ghitza, O. (2006) Objective assessment of speech and audio quality—Technology and applications. *IEEE Transactions on Audio*, *Speech*, *and Language Processing*, **14**, 1890-1901.

[6] Ma, J. and Loizou, P.C. (2011) SNR loss: A new objective measure for predicting the intelligibility of noise-suppressed speech. *Speech Communication*, **53**, 340-354.

[7] Brooks, P. and Hestnes, B. (2010) User measures of quality of experience: Why being objective and quantitative is important. *IEEE Network*, **24**, 8-13.

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

[9] Chen, W.-E. (2011) Real-time VoIP quality measurement for mobile devices. *IEEE Systems Journal*, **5**, 538-544.

[10] ITU-T Study Group 12 (2008) PESQ limitations for EVRC family of narrowband and wideband speech codecs. Qualcomm Inc., San Diego.

[11] ITU-T Recommendation P.863 (2011) Perceptual objective listening quality assessment (POLQA). International Telecommunication Union, Geneva.

[12] ITU-T Recommendation P.800 (1996) Methods for subjective determination of transmission quality. International Telecommunication Union, Geneva.

[13] Zhang, W., Chang, Y., Liu, Y., et al. (2013) A new method of objective speech quality assessment in communication system. *Journal of Multimedia*, **8**, 291-298.

[14] ITU-T Rec. P.862.1 (2003) Mapping function for transforming P.862 raw result scores to MOS-LQO. International Telecommunication Union, Geneva.

[15] ITU-T Recommendation P.1401 (2012) Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. International Telecommunication Union, Geneva.