VoIP 音视频质量评估方案

1. 概述

音视频质量是 VoIP 系统最核心的指标，直接关系到用户体验，因此需要一套行之有效的评估方案，一是用来横向比对自研产品、竞品及第三方合作伙伴的质量优劣，二可以纵向体现整体系统的改进结果，对后面的持续优化具有非常重要的指导意义。在具体执行上，音频的测试评估手段有客观声学测试、客观外场路测、主观测试、机型适配测试、群聊模式测试、性能测试等；视频方面主要有主观测试、机型适配测试、群聊模式测试、性能测试等。

2. 音频测试

2.1 客观声学测试

2.1.1 测试目的及标准

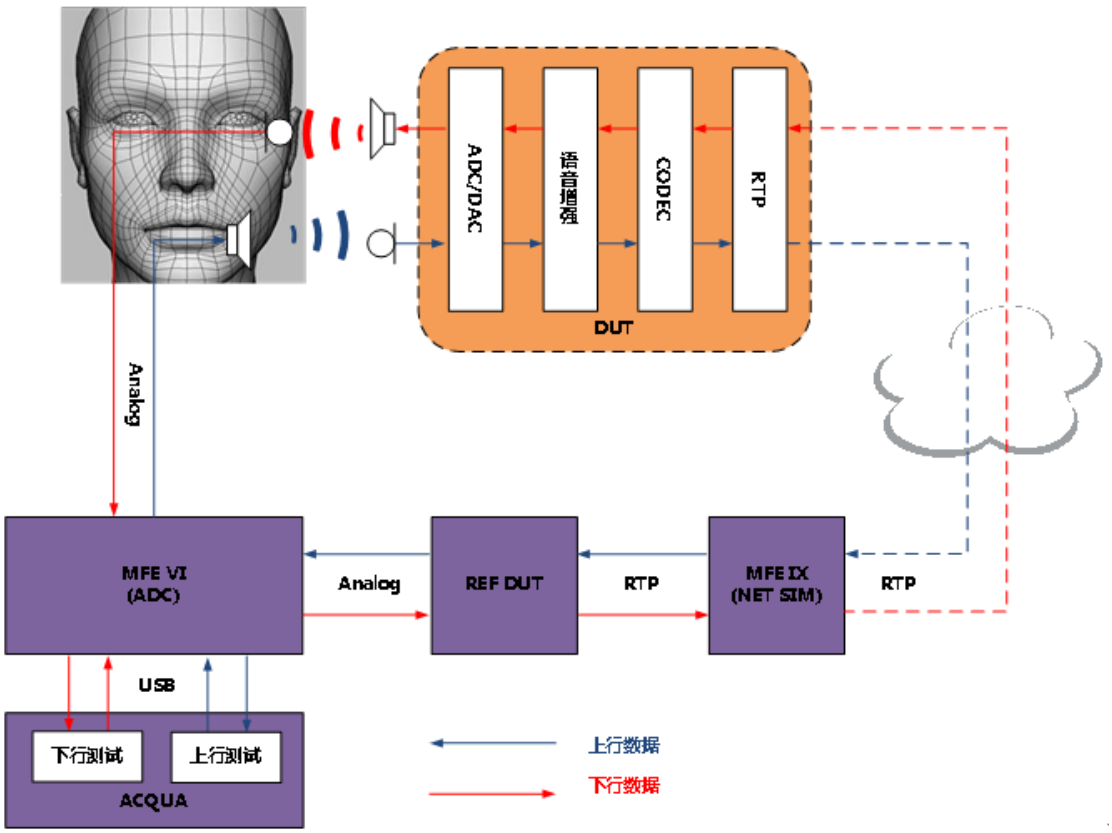
在实验室环境下按 ITU 标准测试客观声学指标，是业内衡量声音质量的最权威方法，所采用标准如下：

- POLQA – Perceptual objective listening quality assessment (ITU-T P.863)
- 3QUEST

需要说明的是，此项测试由于需要使用专业环境及设备，测试的灵活度和测试项目受到较大限制，测试场景相对单一，因而并不能完全覆盖移动端复杂多变的使用环境。

2.1.2 测试系统搭建

需要在特定的声学屏蔽室（最好为全消）利用专业声学测试系统进行测试，下图为基于 HEAD ACOUSTICS 设备和 ACQUA 软件搭建的测试环境。



说明：

1，测试下行部分，ACQUA 中的数字语音信号给 MFE VI 转为模拟信号后给 Ref DUT，然后打包通过网络经过 MFE IX 加网损传给 DUT，人公头的人耳收到 DUT 发出的声音并将其录下来给 MFE VI 后转为数字信号给到 ACQUA；

2，测试上行部分，ACQUA 将数字语音信号给 MFE VI 转为模拟信号后给人工头的人工嘴，发出此声音后 DUT 的 MIC 接受此信号后通过网络经过 MFE XI 的网损后给 Ref DUT，然后给到 MFE VI 将此信号转为数字信号给到 ACQUA 做分析。

3，此配置中，语音信号会在 Ref DUT 和 DUT 中做两次处理，所以测试的是一对 APP 打电话的语音通话性能；

4，MFE IX 可以人为的设定一定的网络损耗，但是由于测试时候是接入外网，这部分的网络损耗是不可控制的，所以测试时候的网络损耗必然比设定的 MFE IX 的网络损耗大；

2.1.3 声学测试项目

声学测试内容			解释
听筒模式	**主要指标**	语音延迟（MS)	端到端的延迟统计
		噪声环境语音质量评分	对噪声环境下通话语音的综合评分
		安静环境语音质量评分	对安静环境下通话语音的综合评分
		回声抑制（dB）	对回声信号的抑制
		双讲衰减（dB）	双讲场景下回波算法对语音信号的衰减损伤
	次要指标	静默噪声（dB）	静默环境下底噪音检测
		声音响度（dB）	声音响度测量
		频率响应（dB）	语音频率响应测量
		频率失真	语音频率失真分析
外放模式	**主要指标**	语音延迟（MS)	端到端的延迟统计
		噪声环境语音质量评分	对噪声环境下通话语音的综合评分
		安静环境语音质量评分	对安静环境下通话语音的综合评分
		回声抑制（dB）	对回声信号的抑制
		双讲衰减（dB）	双讲场景下回波算法对语音信号的衰减损伤
	次要指标	静默噪声（dB）	静默环境下底噪音检测
		声音响度（dB）	声音响度测量
		频率响应（dB）	语音频率响应测量
		频率失真	语音频率失真分析

2.1.4 网损测试项目

采用网络模拟仪或模拟软件尽量拟合 ITU 对网络模型的定义，在多种网络环境下得到 MOS 分测试结果并参照实际用户环境进行加权处理。

Impairment type	Units	Profile A, well-managed range (min to max)	Profile B, partially-managed range (min to max)	Profile C, unmanaged range (min to max)
One-way latency	ms	20 to 100 (regional) 90 to 300 (intercontinental)	20 to 100 (regional) 90 to 400 (intercontinental)	20 to 500
Jitter (peak-to-peak)	ms	0 to 50	0 to 150	0 to 500
Sequential packet loss	ms	Random loss only	40 to 200	40 to 10'000
Rate of sequential loss	s^{-1}	Random loss only	$< 10^{-3}$	$< 10^{-1}$
Random packet loss	%	0 to 0.05	0 to 2	0 to 20
Reordered packets	%	0 to 0.001	0 to 0.01	0 to 0.1

2.2 客观外场路测

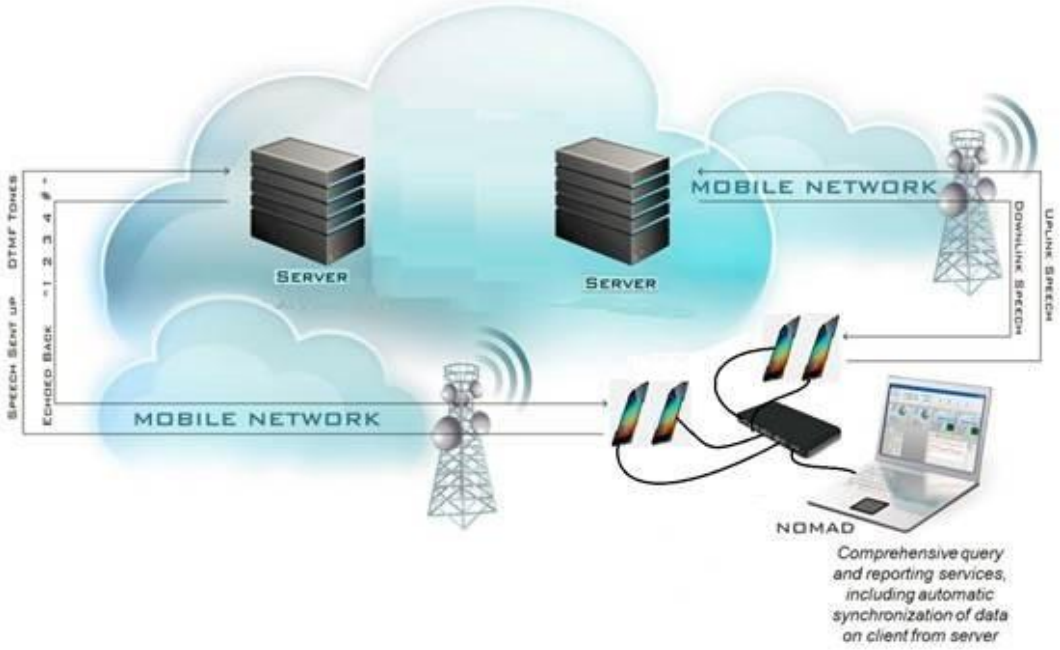
2.2.1 测试目的

通过便携式的语音质量测试仪器，以路测的方式实现实际网络环境下的对比测试，可以大大提高测试环境的覆盖率。该方法的局限在于一是只能测试耳机模式，二是测试项一般只有 MOS 分和延迟。

2.2.2 测试环境搭建

以 Spirent Nomad HD 为例，该系统使用了最新的 POLQA（ITU-P.863）客观评价方法，可以便捷地对语音质量进行度量。
Nomad 系统有 4 路 HD 语音接口，最多连接 4 部手机，两两一组，可以同时对 2 种网络或 2 种技术方案的语音进行实时的采集、对比和评价。

Nomad 测试环境如下图所示：



2.2.3 测试项目及标准

MOS 评价标准：

MOS	3.5~5.0	3.0 以下	3.5~5.0
语音质量描述	优	差	优

注：非正式而言，MOS 分值差在 0.5~0.8 为差异较大且能明显感知，

0.2~0.5 差异稍大且能感知但尚可接受，0.1~0.2 为差异较小但有感知

Delay 评价标准：

Delay	<400ms	>600ms
单向时延时间 ms	优	差

2.3 声学主观测试

2.3.1 测试目的

客观测试结果不能完全代表主观体验，因而主观测试必不可少，二者互为补充，作为最终体验的衡量手段。我们参考 ITU-T P.800/P.830 主观 MOS 测试规范设计了一套涵盖大部分主观体验要素的测试标准，为 VoIP 系统的声学质量评判提供最直接的参考，同时也可以作为系统改进的重要依据。

2.3.2 测试项目和标准

主观测试标准（满分 147 分，88 分以上及格，117 分以上良好，132 分以上优）

主观测试内容			判断标准
听筒模式	主要指标	语音延迟	1.整个通话中延迟大到无法正常通话 2.有明显延迟(>2s) 3.有延迟(>1s) 4.可以感知到有延时(<1s) 5.没有感知延时对通话的影响
		噪声控制	1.噪声吵到听不清对方说话 2.噪声很大，通话效果很差 3.噪声偏大，对通话造成一点干扰 4.有一点噪声，但是不影响通话 5.噪声很小，不会引起注意
		通话音质	1.严重失真，不似人声 2.有明显失真，无法判断对方是谁

			3.有些许失真，但是能判断对方是谁 4.清晰明亮真实 5.感觉像面对面说话
		回声	1.整个通话中一直有很大的回声,且对正常通话产生严重干扰 2.经常会出现回声，且声音大，明显影响通话 3.经常会出现持续回声,但轻微,对通话影响不大 4.偶尔出现过几次，但很快没了 5.整个通话完全没有回声
		流畅性	1.整个通话中长时间人声信息丢失,基本不能沟通 2. 经常出现大量人声信息丢失，沟通不够顺畅 3.经常出现少量人声信息丢失，基本沟通无碍 4.偶尔会出现几个字，听不到的情况 5.整个通话中没有感知到人声有丢失
		双讲	1.两方都听不清对方说话 2.只有一方可以听清楚 3.双方说话内容丢失大量字 4.双方说话内容丢失少量字 5.两边完全能听清楚，对方说的所有话
	次要指	音量	1.音量小/刺耳

	标		2.音量适中 3.音量大
		杂音/颤音/抖音/电音	1.大部分时间会出现杂音 ,对整体通话的干扰明显 2.偶尔出现过几次，但很快没了 3.整个通话完全没有杂音
		静默噪声/舒适噪声	1.双方不说话时不能感知环境音 2.双方不说话时能感知，环境音过大/过小 3.双方不说话时能感知，适中音量环境音
		噪声控制	1.有明显失真，无法判断对方是谁 2.有些许失真，但是能判断对方是谁 3.没有失真
		平稳度/声音起伏	1.声音一直都忽大忽小，很不稳定，影响通话 2.偶尔会出现几个字，声音变强或变弱 3.每句话的头中尾的声音，强弱很稳定
外放模式	主要指标	增益控制	1.有声但听不清楚/声音大有爆音 2.仔细听能听见 3.能听到但是声音小 4.完全能听清楚且声音适中 5.完全能听清楚且声音大

			1.听不到对方声音 2.有声但听不清楚 3.仔细听能听见 4.能听到但是声音小 5.完全能听清楚
			1.听不到对方声音 2.有声但听不清楚 3.仔细听能听见 4.能听到但是声音小 5.完全能听清楚
		语音延迟	1.整个通话中延迟大到无法正常通话 2.有明显延迟(>2s) 3.有延迟(>1s) 4.可以感知到有延时(<1s) 5.没有感知延时对通话的影响
		噪声控制	1.噪声吵到听不清对方说话 2.噪声很大，通话效果很差 3.噪声偏大，对通话造成一点干扰 4.有一点噪声，但是不影响通话 5.噪声很小，不会引起注意

		通话音质	1.严重失真，不似人声 2.有明显失真，无法判断对方是谁 3.有些许失真，但是能判断对方是谁 4.清晰明亮真实 5.感觉像面对面说话
		回声	1.整个通话中一直有很大的回声,且对正常通话产生严重干扰 2.经常会出现回声，且声音大，明显影响通话 3.经常会出现持续回声,但轻微,对通话影响不大 4.偶尔出现过几次，但很快没了 5.整个通话完全没有回声
		流畅性	1.整个通话中长时间人声信息丢失,基本不能沟通 2. 经常出现大量人声信息丢失，沟通不够顺畅 3.经常出现少量人声信息丢失，基本沟通无碍 4.偶尔会出现几个字，听不到的情况 5.整个通话中没有感知到人声有丢失
		双讲	1.两方都听不清对方说话 2.只有一方可以听清楚 3.双方说话内容丢失大量字 4.双方说话内容丢失少量字

			5.两边完全能听清楚，对方说的所有话
	次要指标	杂音/颤音/抖音/电音	1.大部分时间会出现杂音 ,对整体通话的干扰明显 2.偶尔出现过几次，但很快没了 3.整个通话完全没有杂音
		噪声控制	1.有明显失真，无法判断对方是谁 2.有些许失真，但是能判断对方是谁 3.没有失真
		静默噪声/舒适噪声	1.双方不说话时不能感知环境音 2.双方不说话时能感知，环境音过大/过小 3.双方不说话时能感知，适中音量环境音
		平稳度/声音起伏	1.声音一直都忽大忽小，很不稳定，影响通话 2.偶尔会出现几个字，声音变强或变弱 3.每句话的头中尾的声音，强弱很稳定
耳机模式	主要指标	语音延迟	1.整个通话中延迟大到无法正常通话 2.有明显延迟(>2s) 3.有延迟(>1s) 4.可以感知到有延时(<1s) 5.没有感知延时对通话的影响
		噪声控制	1.噪声吵到听不清对方说话

			2.噪声很大，通话效果很差 3.噪声偏大，对通话造成一点干扰 4.有一点噪声，但是不影响通话 5.噪声很小，不会引起注意
		通话音质	1.严重失真，不似人声 2.有明显失真，无法判断对方是谁 3.有些许失真，但是能判断对方是谁 4.清晰明亮真实 5.感觉像面对面说话
		回声	1.整个通话中一直有很大的回声,且对正常通话产生严重干扰 2.经常会出现回声，且声音大，明显影响通话 3.经常会出现持续回声,但轻微,对通话影响不大 4.偶尔出现过几次，但很快没了 5.整个通话完全没有回声
		流畅性	1.整个通话中长时间人声信息丢失,基本不能沟通 2. 经常出现大量人声信息丢失，沟通不够顺畅 3.经常出现少量人声信息丢失，基本沟通无碍 4.偶尔会出现几个字，听不到的情况 5.整个通话中没有感知到人声有丢失

		双讲	1.两方都听不清对方说话 2.只有一方可以听清楚 3.双方说话内容丢失大量字 4.双方说话内容丢失少量字 5.两边完全能听清楚，对方说的所有话
	次要指标	音量	1.音量小/刺耳 2.音量适中 3.音量大
		杂音/颤音/抖音/电音	1.大部分时间会出现杂音 ,对整体通话的干扰明显 2.偶尔出现过几次，但很快没了 3.整个通话完全没有杂音
		静默噪声/舒适噪声	1.双方不说话时不能感知环境音 2.双方不说话时能感知，环境音过大/过小 3.双方不说话时能感知，适中音量环境音
		噪声控制	1.有明显失真，无法判断对方是谁 2.有些许失真，但是能判断对方是谁 3.没有失真
		平稳度/声音起伏	1.声音一直都忽大忽小，很不稳定，影响通话 2.偶尔会出现几个字，声音变强或变弱 3.每句话的头中尾的声音，强弱很稳定

2.3 机型适配测试

2.3.1 测试目的

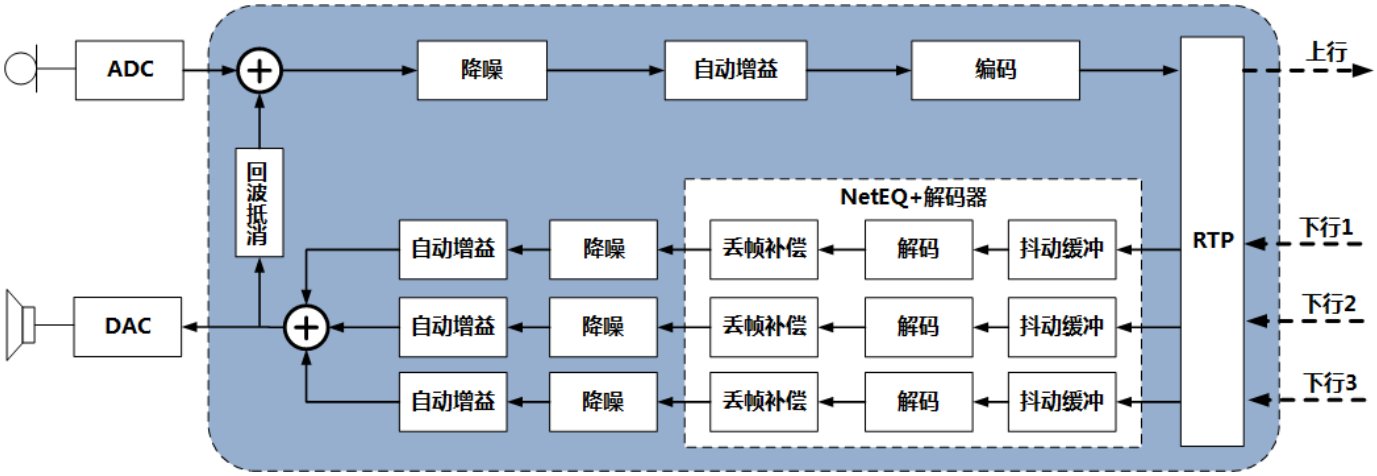
每款智能手机尤其 Android 手机的音频系统实现都存在或大或小的差异，声腔设计、底层算法方案、算法 tuning、BSP 配置的不同都会造成声音质量的差异，在上层不可能有一种算法通吃所有机型，因此要根据手机的底层音频特性调整应用层算法的参数，并通过测试手段加以验证。

2.3.2 测试项目和标准

按用户机型分布情况确定适配及测试优先级，测试项目可参考主观测试标准并进行适度裁剪。
社交类 TOC 业务可参考友盟机型分布数据：
http://www.umindex.com/?spm=0.0.0.0.V2lwDJ

2.4 群聊模式测试

群聊数据处理流程比 1 对 1 单聊复杂，在客户端侧为了降低系统负载和带宽消耗往往会采取妥协策略，可参考主观标准对群聊模式加以验证。



2.5 性能测试

包括 CPU、内存、电量、流量、温度等性能指标测试，需要注意的是要覆盖多种网络环境、手机状态，再配合相关音频质量结果即可得出被测系统在不同条件下的性能与质量的均衡策略，结合本身业务场景进而可以判断此种策略是否为最优选择。