

RealComm：把音频鉴伪放进真实通话链路

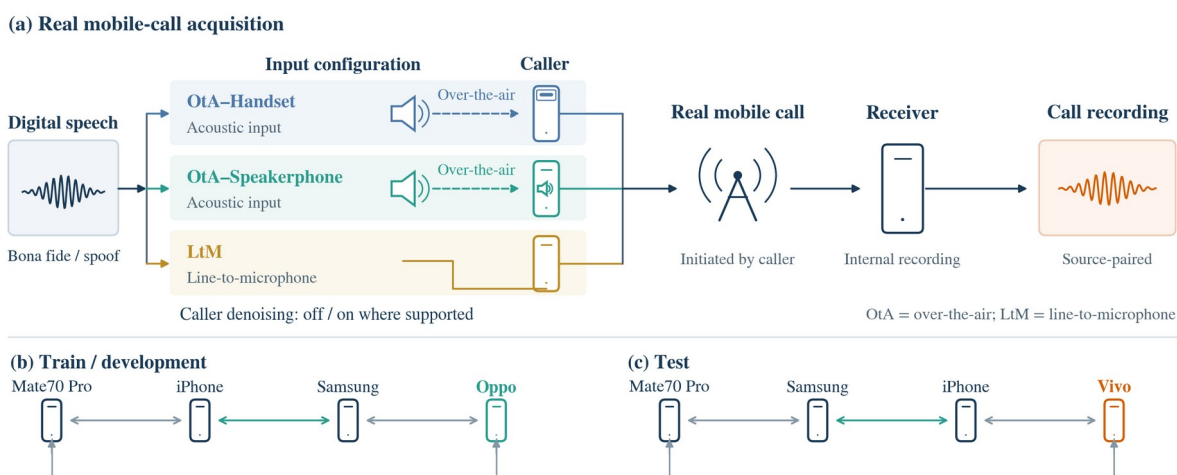
同一句语音，打过电话后，检测结果会怎样变化？

一段合成语音在音频文件中容易被识别，并不意味着它通过电话传到接听方后，仍然容易被识别。外部播放、麦克风采集、手机处理和通话传输会共同改变信号。RealComm 研究的就是这段从源文件到接听端的过程。

用同一句语音，比较通话前后

RealComm 将真实语音和七种生成系统的合成语音送入实际手机通话，保留每条录音对应的数字源文件。拨出方可以通过外部扬声器接收语音，也可以通过线缆直接注入；录制同时覆盖手持、免提、不同设备组合，以及降噪开关。

数据包含中文和英文。RealCommTrain 提供 12,320 条训练与开发录音，RealCommBench 提供 23,520 条测试录音。同一源语音的不同通话版本始终保留在同一划分中。这样，通话前后的比较不需要更换文本、说话人或生成器，更容易看清通信路径带来的影响。



论文图 1：同一源语音通过声学或有线方式进入拨出方，经真实通话传到接听方。训练与测试使用部分重叠的设备组合，测试中保留未见设备。

六个模型，都出现明显的通话域退化

六个已有检测模型在同一批数字文件及其通话录音上测试。通话后的等错误率（EER，越低越好）均升高，增幅为 11.73—37.95 个百分点。例如，WavLM-L 从数字域的 4.64% 升至通话域的 40.00%。这些结果说明，数字域的低错误率不能直接代表实际通话中的检测能力。

Model	Digital	Call	Δ EER
WavLM-L	4.64	40.00	+35.36
Whisper	15.71	47.88	+32.17
AASIST	7.50	40.19	+32.69
Teffic-Audio	0.00	37.95	+37.95
XLSR-SLS	38.57	50.31	+11.73
DF Arena	7.86	38.77	+30.91

论文表 II：六个模型的数字域与真实通话 EER，以及通话后增加的百分点。六个模型均尚未进行 RealComm 适配。

退化也与输入路径有关：六个模型在外部播放、麦克风采集的两类声学路径上，都比有线注入更难检测。质量分析进一步发现，音质下降较大的通话条件通常更难鉴伪；但高频衰减最多的有线路径，检测效果反而较好。通话域偏移涉及多个环节，不能只用带宽变化解释。

模拟增强和真实录音，可以一起帮助适配

RealComm-Aug 按照声学环境、设备处理、文件编码和通话传输组织增强。噪声、混响等常见扰动可以单独输出，也可以与后续阶段组合；文件转码与通话编码采用不同分支。训练时，真实与伪造语音使用相同的增强分布，并保留部分原样输入。

我们在三个模型上比较了四种设置：原模型 F、仅模拟增强 A、加入真实通话训练数据 T，以及二者结合的 A+T。各适配设置均从对应的 F 权重开始。

Model	Domain	F	A	T	A+T
WavLM-L	Digital	4.64	4.29	3.21	3.21
	Call	40.00	38.71	35.48	34.16
AASIST	Digital	7.50	8.93	6.07	8.21
	Call	40.19	38.49	37.96	36.90
Teffic-Audio	Digital	0.00	0.36	0.36	0.00
	Call	37.94	34.92	23.84	22.37

论文表 IV：三个模型在 F、A、T 和 A+T 下的 EER。粗体为每个模型在对应数据域上的最优结果。

Teffic-Audio 的通话 EER 从 37.94% 降至 22.37%；WavLM-L 和 AASIST 在 A+T 下分别达到 34.16% 和 36.90%。真实通话训练带来的改善大于单独模拟增强，而在真实训练基础上加入增强，三个模型又进一步改善了 1.06—1.46 个百分点。

改善仍有边界：AASIST 在部分子集与数字域上出现回退，声学注入仍比有线注入更难。RealComm 让这些差异能够在同源语音上配对分析。