案例研究

第三代英特尔® 至强® 可扩展处理器
语音分析合成

定制化声码器优化方案，提升实时语音合成性能

全新一代英特尔® 至强® 可扩展处理器，为腾讯云小微智能语音与视频服务接入平台注入强劲动力。

腾讯云小微

“通过不断提升的吞吐量以及更高的实时性，云小微平台能够成功为企业级应用提供高质量智能化语音服务。得益于英特尔先进的软硬件技术支撑，基于第三代英特尔® 至强® 可扩展处理器的定制化解决方案，使云小微平台的语音合成性能得以更充分释放。”

田乔
高级研究员
腾讯云

智能语音应用正获得前所未有的突破与增长，预测表明，中国智能语音市场规模到 2021 年底或可达 194.8 亿元[1]。为了更好地赋能智能语音硬件厂商，一直致力于人工智能研发和互联网创新的腾讯，正全力打造云小微智能语音与视频服务接入平台，以基于神经网络声码器模型 (Neural Based Vocoder) 的从文本到语音 (Text To Speech, TTS) 合成技术为核心能力，通过端到端声学模型，实现文本到语音的高质量转化与表达。

经典的语音合成声码器模型，如 WaveNet，虽可生成高保真的语音质量，但其模型复杂，所需计算量非常大，导致语音合成时间较长，难以满足实际生产中对实时性的高要求，且大量设备的持续接入也对云小微平台的吞吐量提出了更高要求，若一味扩容服务器势必会带来高昂的建设成本。为此，腾讯采用更前沿的声码器模型对云小微平台进行了深度优化，通过与英特尔密切合作，共同构建定制化 Parallel WaveNet 声码器模型 (以下简称 pWaveNet) 解决方案以及定制化 WaveRNN 声码器模型解决方案，为平台提供了优异的语音合成性能，并有效降低了用户总拥有成本 (Total Cost of Ownership, TCO)。

方案采用全新第三代英特尔® 至强® 可扩展处理器作为核心算力引擎。新一代处理器不仅以更多的内核与线程为平台提供强大算力支撑，且内置的 BF16 指令及英特尔® 高级矢量扩展 512 (英特尔® AVX-512) 指令，大大减少了内存访问量，在英特尔® oneAPI 深度神经网络库 (Intel® oneAPI Deep Neural Network Library, oneDNN) 的配合下，有效支撑硬件加速。同时，该处理器配备的更大缓存提高了缓存命中率，可有效提升处理效率。融合以上先进英特尔软硬件技术的定制化解决方案，有力地助力腾讯云小微平台为更多企业及设备厂商提供一流的语音合成服务能力，获得了市场的良好反馈。

凭借更优的定制化声码器模型解决方案，腾讯云小微平台获得：

更快响应-定制化 pWaveNet 声码器模型通过简化网络结构，并与第三代英特尔® 至强® 可扩展处理器平台一起，形成并行计算优势，在保证语音质量的同时，能有效地提升语音合成速度。经测算，新方案在平均主观意见分（Mean Opinion Score，MOS）值为 4.4 的条件下，实现了 0.036 的语音合成实时率[2]；

更高性能-定制化 WaveRNN 声码器模型凭借简单模型架构，结合线性处理、子代划分、稀疏化等技术，有力降低计算量，与第三代英特尔® 至强® 可扩展处理器平台相配合，在提升语音合成速度的同时，可承载更大的工作负载。经测算，新方案单核心为 100 个实例提供服务时，可达到与为一个实例提供服务时相近的性能[3]；

更强算力-全新一代英特尔® 至强® 可扩展处理器的内置硬件加速技术，配合强劲核心和更大缓存，帮助云小微平台获得更高效能，使其能为更多企业提供服务，并打造优质智能生态，推进人工智能创新发展。

随着人工智能在各行业的落地，各新兴智能产品企业正基于智能语音合成技术，研发语音导航、有声读物、智能客服、智能语音输入与识别等应用，来打通人机交互的闭环。但在基于这些创新功能，获得生活便利的同时，人们也发现很多智能语音产品的 AI 能力参差不齐，操作方式也各异，用户体验尚待提升。究其原因，是因为这些应用的研发往往基于不同平台，既不能通过数据和技术优势形成高品质的智能语音服务，也无法通过设备间的互联互通建立有效工作协同。

为应对这一窘境，腾讯推出云小微智能语音与视频服务接入平台，通过结合全栈语音语义 AI 能力和腾讯云服务，在为用户输出高品质 AI 平台能力的同时，依托腾讯丰富的产品线和大数据能力，帮助用户获得整合腾讯中台能力的丰富场景应用方案。以智慧酒店为例，接入云小微的硬件，不仅能使智能问询、客房控制等产品快速具备丰富的听觉和视觉感知能力，也可让这些产品与微信、微信地图、微信音乐等常见手机 APP 形成联动，让最终用户轻松上手，使用体验也更为流畅轻快。而在智能交通领域，平台在助力车企提供车载语音导航等交互能力外，也可接入腾讯音乐、腾讯新闻等提供的海量娱乐内容，大幅提升行车舒适体验。同样，在教育、金融、传媒等各个领域中，腾讯云小微平台也都获得了广泛的应用。

在与用户联手打造良好产品生态的同时，腾讯也正不断优化声码器模型，持续提升平台的核心 TTS 语音合成能力，从而进一步提升最终用户的使用体验。TTS 语音合成技术可将外部输入的文本或计算机自己产生的信息，通过自然流畅的语音表达出来。该过程主要通过声码器模型进行计算分析以输出语音波形，而不同声码器模型的选用对合成效果具有重要的影响。较为典型的声码器，如 WaveNet 模型，是基于卷积神经网络的深度自回归模型，即将上一层的输出结果放到输入层末尾进行卷积迭代，生成的语音质量之高几乎可以接近自然人声。但在实际运用中，传统 WaveNet 模型还有以下两个方面的不足：

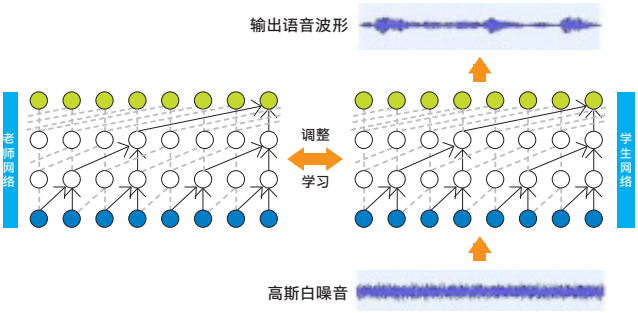
- 首先，WaveNet 模型结构较为复杂，不仅对计算力要求较高，合成速度也不尽如人意，在实时性要求较高的智能语音交互场景中，无法提供令人满意的用户体验；
- 其次，智能语音应用场景的普及，使云小微平台需要对更多设备提供支持，随之带来更多的工作负载 (吞吐量)，需要声码器模型具备更高效的语音合成工作效能。

为此，腾讯亟需寻求高效的语音合成方案，应对实时性与吞吐量的巨大挑战。腾讯与英特尔一直保持着深入的合作，面对以上需求，腾讯再次选择与英特尔强强联手，共同构建了定制化 pWaveNet 声码器以及定制化 WaveRNN 声码器这两套语音合成解决方案，将平台性能推向更优。

定制化 Parallel WaveNet 声码器解决方案

在定制化 pWaveNet 声码器解决方案中，之所以选用 pWaveNet 模型进行语音合成，是因为该模型不仅具有更轻的量级，而且它在 WaveNet 模型的基础上，引入了“概率密度蒸馏”技术，即用一个提前训练好的 WaveNet 模型作为“老师”，来指导真正实施生产的“学生”网络进行预测。该“学生”网络体量更小，采

用随机白噪声作为输入，通过学习“老师”的概率分布并不断调整，来减小与“老师”的差距，产生理想输出。区别于 WaveNet 必须依赖于先前已生成的点作为输入条件的顺序生成模式，pWaveNet 利用“学生”网络，直接学习“老师”的每一个音频采样点，不依赖于“学生”自身网络任何先前的输出节点，使并行计算成为可能，可以一次性生成整个序列的输出采样点，大幅减少语音合成时间。



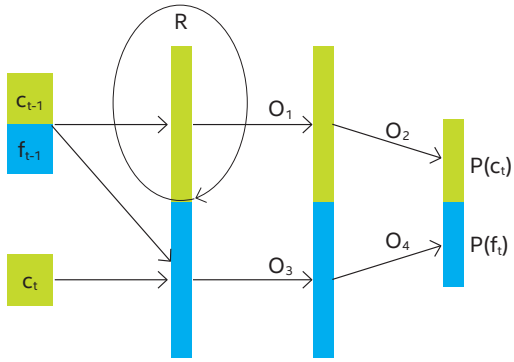
图一 Parallel WaveNet 模型架构图

但是，pWaveNet 模型中的“学生”网络依旧是以卷积神经网络为基础的网络架构，虽然规模较小，但是众所周知，卷积操作相较于普通的加减乘除运算要耗费更大的计算量。为此，腾讯在 pWaveNet 模型的基础上进行定制化开发，将网络中一维卷积运算转换为几个通用矩阵相乘的操作，以简化网络拓扑并减少计算量，同时引入 Open-MP 并行机制，充分发挥 pWaveNet 模型中的并行计算优势，使得该定制化模型在不影响语音质量的同时，有效提高了语音合成速度。

定制化 WaveRNN 声码器解决方案

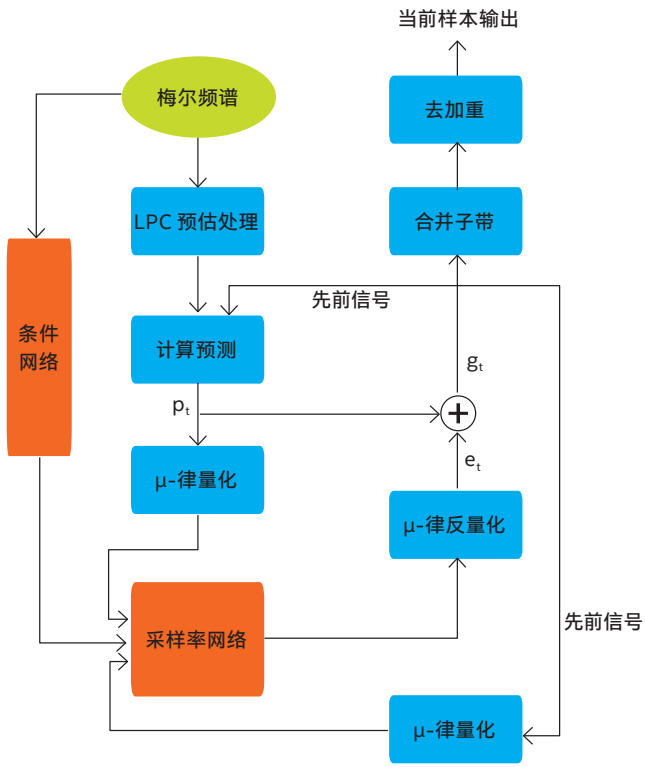
除了对语音合成速度的不断追求以外，云小微平台还面对着越来越多设备的接入压力，随之而来的是对整体吞吐量的严苛要求。即在面对大量的实例运算时，单核心所服务的实例数越多越好，而提升单核吞吐量最直接的方法是进一步降低计算量。

针对这一问题，腾讯选用了先进的 WaveRNN 模型，在其基础上构建高性能的 WaveRNN 语音合成方案。WaveRNN 模型的基本结构是一个具有双 softmax 层的单循环网络，将 16 位样本序列分为高 8 位的粗动 (coarse) 部分和低 8 位的精细 (fine) 部分，采用 GRU 门控循环单元分别进行样本预测。由于该结构只有单层循环，每预测一个 16 位样本序列仅需要 5 步操作计算，远远小于 WaveNet 深度神经网络结构所需的计算操作数。



图二 WaveRNN 模型架构图

除了 WaveRNN 模型本身结构方面的优势外，腾讯还在该模型基础上进行定制化开发，以进一步降低计算量并提升合成速度。定制化 WaveRNN 模型的主体部分——采样率网络，依旧是一个具有双 softmax 层的单循环网络，不同的是，方案将该网络原始输入中的线性部分分离出来，预先进行了 LPC 预估处理，以大幅降低网络处理难度，并将样本序列划分成多个子带，在前一个子带生成开始不久后即启动下一个子带的计算，有效提高整体计算速度，同时方案还引入了稀疏化技术，减少带宽占用，降低网络整体计算时间，并且在多核环境中，大型稀疏模型能更好地平衡计算力，比小型密集模型性能更好。



图三 定制化 WaveRNN 声码器模型架构图

英特尔助力语音合成解决方案大幅提升性能

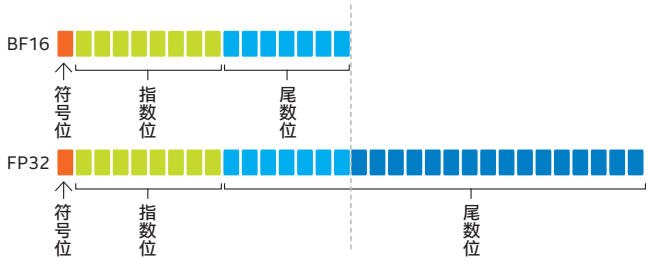
“提升语音合成速度的关键，一在于加快数据在内存上的读写时间，二在于加速数据的执行效率。全新第三代英特尔® 至强® 可扩展处理器所内置的 BF16 指令和英特尔® AVX-512 指令集，帮助我们的定制化模型有效达成了以上两项目标，令平台中的定制化 pWaveNet 声码器在 MOS 值为 4.4 的条件下，实现 0.036 的语音合成实时率；而定制化 WaveRNN 声码器也在提升语音合成速度的同时，具备了更强的工作负载处理能力。”

田乔
高级研究员
腾讯云

在优异的模型架构敲定后，腾讯选择采用先进的英特尔硬件作为底层支撑，为整个方案达到最佳性能增光添彩。定制化 pWaveNet 声码器模型与定制化 WaveRNN 模型解决方案都采用了第三代英特尔® 至强® 可扩展处理器，该处理器具有高达 28 核的强劲内核，在提升计算力的同时，也很好满足了云小微平台对吞吐量的需求，内置的 BF16 指令集在整个方案中起到了十分关键的作用，可有效提升内存利用率，同时与英特尔® AVX-512 指令一起，在英特尔® oneAPI 深度神经网络库的配合下，加速硬件效率，配合以新一代处理器的超大缓存，能够有效提升处理性能，为语音合成速度的提升做出卓越贡献。

英特尔® BF16 指令减少内存读写时间

BFloat16 浮点数是一种新型数据格式，由 1 位符号位、8 位指数位与 7 位尾数位组成，相当于在 FP32 浮点数基础上截断后 16 位尾数位。此种格式同 FP32 浮点数格式具有相同的指数位，即具有近似的动态范围，从而可达到与 FP32 格式相似的模型精度，但由于尾数位的减少，大大降低了计算量并提升了内存存储与读取性能。在以上模型优化方案中采用 BF16 数据格式可达到与 FP32 格式同等的语音质量，却可大为缩短语音合成时间。



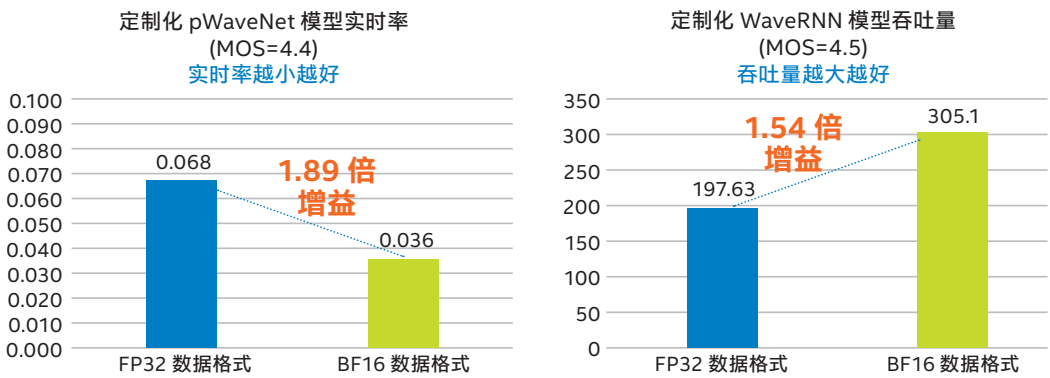
图四 BF16 与 FP32 浮点数据格式结构

英特尔® AVX-512 指令提高执行效率

英特尔® AVX-512 是在处理器上执行单指令多数据 (Single Instruction Multiple Data, SIMD) 运算的指令集，即通过一个处理器控制多个处理微元实现并行操作多段数据以提升处理效率。该指令集将指令宽度扩展到 512 位，使每个时钟周期内可打包更多运算。且该指令集支持三操作 (3-Operand)，即通过创建复杂高级指令来代替多个简单单独指令，增强指令灵活性，减少内存访问数量，从而实现单核执行效率最大化。

超大处理器缓存提升处理性能

介于处理器与内存之间的缓存被用来存储经常需要访问的数据内容。由于处理器的处理速度远远大于内存读写速度，缓存的重要性就在于提供一个比内存更快的临时中转存储，以减少处理器等待数据的时间。当处理器读取数据时，首先从位置更近的缓存中查找，若没查到再去内存中查找，英特尔超大处理器缓存可有效提升缓存命中率，从而提升处理器性能。



图五 定制化解决方案性能增益

解决方案性能测试验证

为验证英特尔产品为以上定制化解决方案带来的性能增益，腾讯与英特尔一起基于第三代英特尔® 至强® 可扩展处理器，分别采用 BF16 数据格式及 FP32 数据格式运行计算，对语音合成的实时率与吞吐量进行测算，为云小微平台的后续扩展提供数据支撑。

其中，定制化 pWaveNet 模型在保证同等语音合成质量，即 MOS 为 4.4 的条件下，达到了 **0.036** 的实时率，且在采用 BF16 数据格式优化后，比 FP32 格式提升了 **1.89** 倍的性能增益[4]。定制化 WaveRNN 模型也表现出了非常优异的性能，测试表明，使用该模型方案在单核上为 100 个实例提供服务与只为一个实例提供服务的性能差异很小，且在保证同等语音合成质量，即 MOS 为 4.5 的条件下，达到了 **305.1** 的吞吐量，采用 BF16 数据格式优化后，比 FP32 格式提升了 **1.54** 倍的性能增益[5]。

展望

腾讯与英特尔的合作已经成就了许多先进平台与系统。得益于第三代英特尔® 至强® 可扩展处理器的支持，定制化解决方案在 TTS 语音合成应用场景中展现出了十分优异的性能。接下来，英特尔与腾讯将继续开展深度合作，结合英特尔更多先进软硬件技术，拓展新的业务场景，赋能各行各业智能化发展进程，在语音识别、声纹识别以及其他 AI 重要领域实现新的价值，进一步完善软硬一体化的智能生态。

在云小微平台之外，腾讯与英特尔也将继续携手，借助于全新一代英特尔® 至强® 平台提供的良好基础设施能力，在架构云、数据上云、AI、高性能计算以及安全等领域中，为用户提供更为敏捷、高效、可靠和多样化的创新云服务，使用户在降低系统管理与维护成本，提升业务上线部署敏捷度的同时，更专注于业务创新，从而在激烈的市场竞争中占得先机。

如欲了解更多信息，请访问如下链接：

第三代英特尔® 至强® 可扩展处理器：https://www.intel.cn/content/www/cn/zh/products/docs/processors/xeon/3rd-gen-xeon-scalable-processors-brief.html

腾讯云小微平台：https://xiaowei.qcloud.com/

[1] 数据援引自相关媒体报道《赛迪数据 | 2019-2021 年中国智能语音市场预测与展望数据》：http://www.cena.com.cn/industrynews/20200109/104168.html

[2, 4] 测试配置：pWaveNet 模型测试配置：FP32 解决方案配置为：单节点第三代英特尔® 至强® 可扩展处理器平台；4 路第三代英特尔® 至强® 可扩展处理器 CPX ES2(QU3H)；内核/线程：26/52；睿频开启；超线程开启；BIOS版本：WCCCPX6.RPB.0018.2020.0410.1316；内存：DDR4 2933MHz 16GB*24；存储：英特尔® SSDPE2KX010T7；网络接口控制器：以太网控制器 10G X550T*2；操作系统：CentOS 8.1；系统内核：4.18.0-147.5.1.el8_1.x86_64；数据分析加速库版本：1.3；精度：FP32；OMP_NUM_THREADS设为1；BF32 解决方案配置为：单节点 WhiteCloudCity4S 平台；4 路第三代英特尔® 至强® 可扩展处理器 CPX ES2(QU3H)；内核/线程：26/52；睿频开启；超线程开启；BIOS 版本：WCCCPX6.RPB.0018.2020.0410.1316；内存：DDR4 2933MHz 16GB*24；存储：英特尔® SSDPE2KX010T7；网络接口控制器：以太网控制器 10G X550T*2；操作系统：CentOS 8.1；系统内核：4.18.0-147.5.1.el8_1.x86_64；数据分析加速库版本：1.3；精度：BF16；OMP_NUM_THREADS 设为 1。

[3, 5] WaveRNN 模型测试配置：FP32 解决方案配置为：单节点第三代英特尔® 至强® 可扩展处理器平台；4 路第三代英特尔® 至强® 可扩展处理器 CPX ES2(QU3H)；内核/线程：26/52；睿频开启；超线程开启；BIOS 版本：WCCCPX6.RPB.0018.2020.0410.1316；内存：DDR4 2933MHz 16GB*24；存储：英特尔® SSDPE2KX010T7；网络接口控制器：以太网控制器 10G X550T*2；操作系统：CentOS 8.1；系统内核：4.18.0-147.5.1.el8_1.x86_64；数据分析加速库版本：1.3；精度：FP32；OMP_NUM_THREADS设为1；BF32 解决方案配置为：单节点 WhiteCloudCity4S 平台；4 路第三代英特尔® 至强® 可扩展处理器 CPX ES2(QU3H)；内核/线程：26/52；睿频开启；超线程开启；BIOS 版本：WCCCPX6.RPB.0018.2020.0410.1316；内存：DDR4 2933MHz 16GB*24；存储：英特尔® SSDPE2KX010T7；网络接口控制器：以太网控制器 10G X550T*2；操作系统：CentOS 8.1；系统内核：4.18.0-147.5.1.el8_1.x86_64；数据分析加速库版本：1.3；精度：BF16；OMP_NUM_THREADS 设为 1。