P1：

各位同学老师下午好，我们是xx组，今天我们展示的主题是TTS的发展与应用

P2：

我们将从以下四个角度展开，我们为什么需要TTS，TTS的经典模型与发展，TTS当前的应用；TTS的不足与未来展望。

P3：

TTS技术在我们的身边几乎无处不在，从常见的文本阅读（Reading texts）、指导助手（voices assistant）、到目前教育界常见的双语朗读翻译（Bilingual translation），娱乐界出现的虚拟歌姬（virtual singer）、音频还原（例如用周董的声线合成一首浙大的校歌），这些都是TTS的重要应用代表。

P4：

因此，作为AIGC领域的重要一份子，TTS技术的研究有其必要性。然而，目前市面上常见的TTS技术往往有：语句间转折较为模糊，粒度与音频质量较差；往往只支持单一语种；语音、翻译、歌声等模型不一，缺少统一的模型等问题，这些都推动着我们去研究TTS技术，并将其广泛运用到实践中去。

P5：

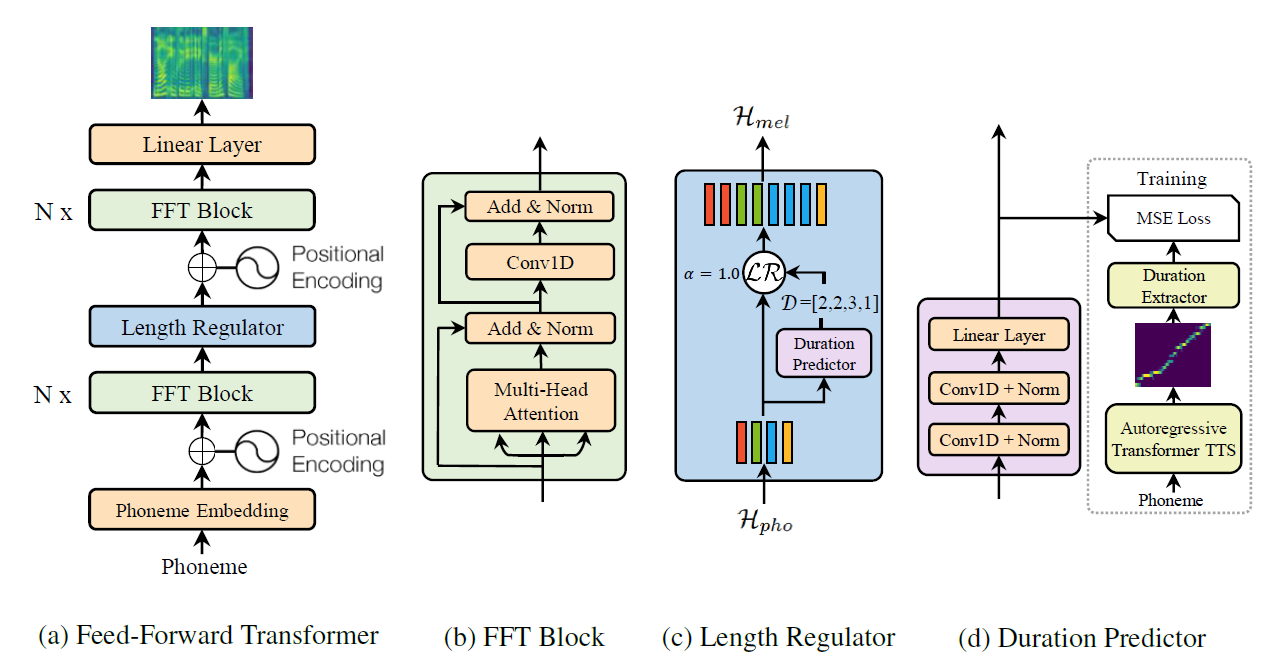
下面让我们简要看一下TTS技术在机器学习中的演变过程，早期的TTS技术大多使用

Concatenative synthesis（拼接法）与Parametric synthesis（参数法），其在建模语音信号的时候经常涉及到一些简化假设，效果较差；而后TTS陆续出现了一批优秀的自回归模型，其基于一个很直观的假设见解：语音中后一个发音总是依赖于前一个词语发音的。此时由Google基于PixelCNN架构开发的WaveNet变应运而生，缓解了前后音频衔接模糊的问题；而为了解决WaveNet高深度、运行长时间的问题，使用矩阵稀疏化、并行序列等技术的WaveRNN被研发而出；进一步，为了提高音频质量，结合了线性预测和递归神经网络的WaveRNN变体——LPCNet出现，有效提高了音频合成的质量。

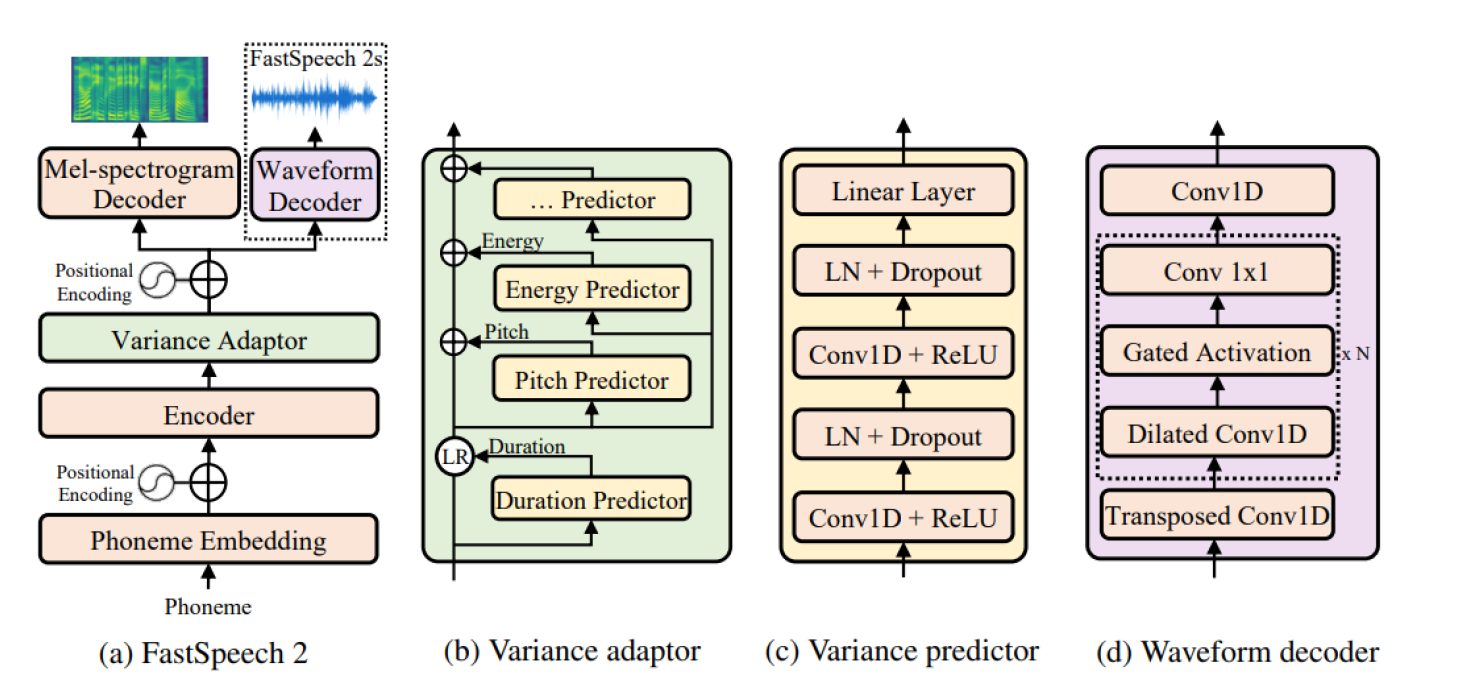
P6：

然而，自回归模型的最大通病就是速度慢，无法最大化利用并行资源，难以适应实际运用中求快的目的。非自回归模型则通过将序列转化为一组相互独立的隐变量，实现了并行化生成的目的，从而大大提高了生成速度。

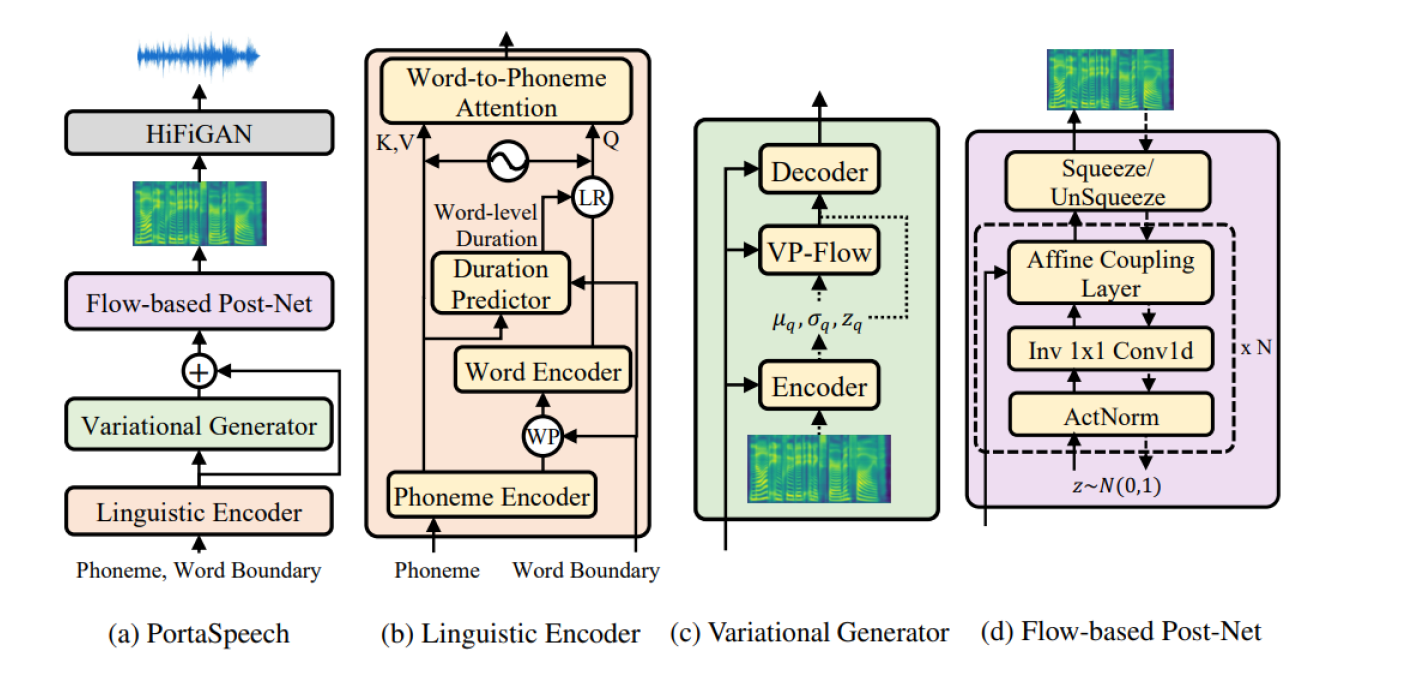
首先是利用Transformer多头注意力并行机制的FastSpeech模型，首次实现了非自回归TTS模型的突破，比之前的TTS模型有了三十倍速度的提升；



而FastSpeech从一个基于编码器-解码器的教师模型中提取注意力，常会遇到蒸馏中pipeline过于复杂，提取信息有缺失的状况，这使得音频质量有所下降，因此FastSpeech2则是在保证速度的同时提高质量，提取更多例如音高、能量、持续时间等信息，用更多的音频信息来提高准确度。

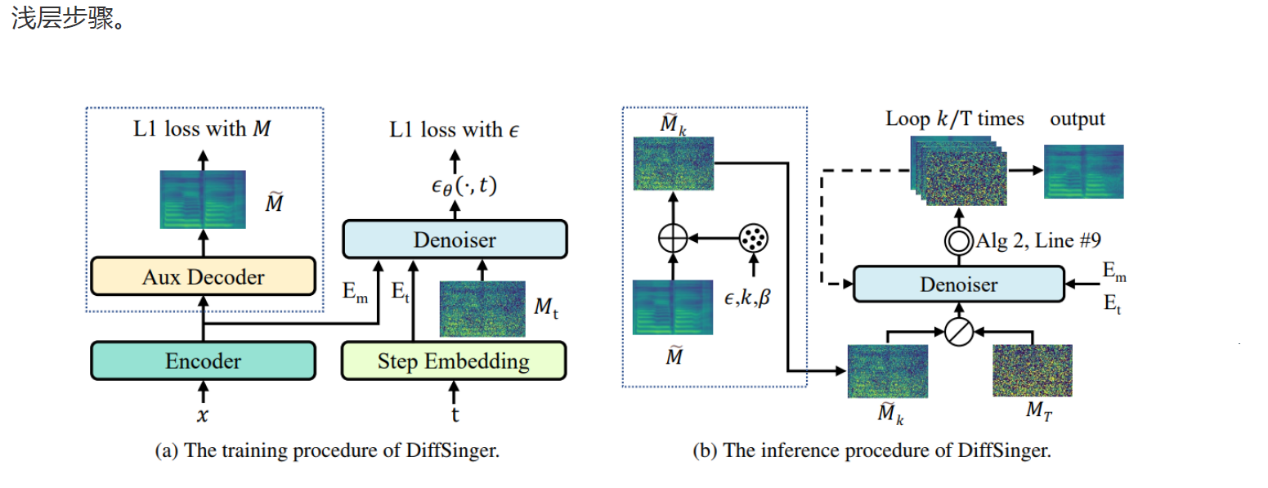


而后为了让TTS能够保持更小的轻量级，部署在各种设备之中，出现了PortaSpeech架构，其结合了Flow与VAE的优点，用flow去复杂化轻量级VAE的分布，提高质量的同时保持了十分小的模型结构。



P7：

当然，上面是TTS技术的缩影，TTS也可以在不同的领域运用，TTS – singer领域有diffspeech声码器，它是一个基于扩散概率模型的歌声声学模型，借助一个参数化马尔可夫（Markov）链，根据音乐乐谱迭代转换为梅尔频谱图，达到了较好的效果。



P8：

TTS-双语翻译中，采用基于离散表示的 S2ST 技术和双向扰动技术(BiP)的Transpeech，解决了双语翻译中常见的准确率和延迟等问题，从而使得双语TTS在实用性和性能方面有了很大的提升。

P9：

结合如上介绍，我们不难看出TTS的各种应用，比如TranSpeech可以提供实时的双语语音翻译服务，可以缓解同声传译的稀少困难；DiffSpeech可以应用到歌声合成之中，形成我们开头讲的AI歌姬；而PortaSpeech、FastSppech则是可以快速生成质量高的文本朗读。

P10：

但是TTS技术依旧存在一些不足，比如说，尽管共同以音频作模态为输出，语音、歌声等各种应用却都独立研究，统一的多任务音频合成模型成为未来研究重点方向；又比如说，TTS模型的输入较为复杂，比如歌声合成，往往要输入每一个音高、持续时间、停顿等，这给它的推广显然造成了困难。

未来的TTS技术，希望可以向统一模型、简易输入的方向进一步发展，也需要有更个性化的服务提供。