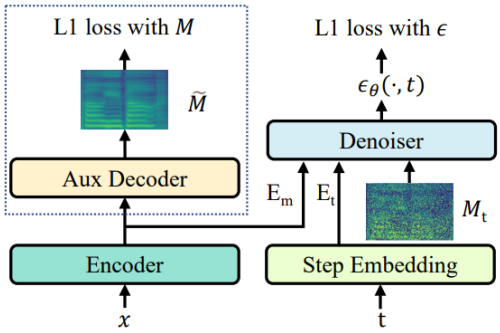
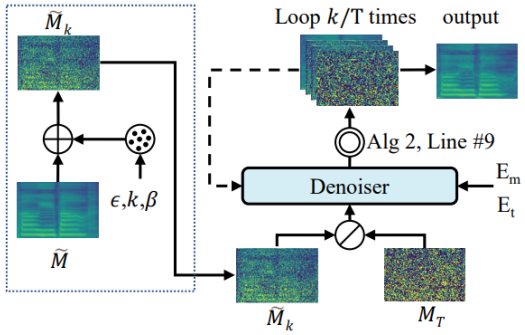
diffspeech

为了解决什么问题：解决之前的歌唱声学模型采用简单损失（simple loss）或生成对抗网络(GAN)时因为过度平滑和不稳定训练限制了合成歌唱的自然度的问题。

技术突破点：它是基于扩散概率模型的歌声声学模型，是一个参数化马尔可夫（Markov）链，根据音乐乐谱将噪声迭代转换为mel-spectrogram。采用通过隐式地优化变分下界，使得输出逼真。引入了浅层扩散机制，以更好地利用简单损失所学习的先验知识。提出了边界预测方法，自适应地定位交点并确定浅层步骤。



(a) The training procedure of DiffSinger.



(b) The inference procedure of DiffSinger.

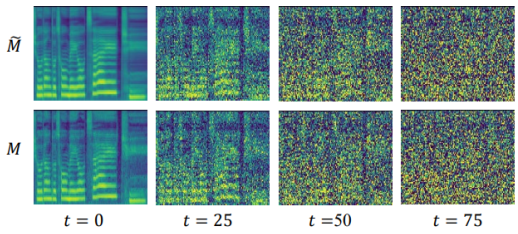


Figure 3: The mel-spectrograms at different steps in the diffusion process. The first line shows the diffusion process of mel-spectorgrams $\widetilde{M}$ generated by a simple decoder trained with L1 loss; the second line shows that of ground truth mel-spectrograms.

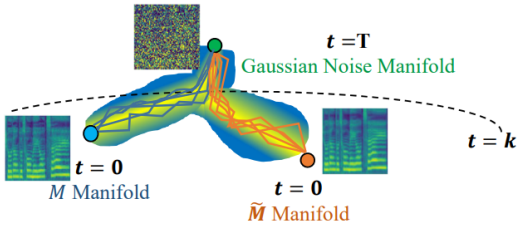
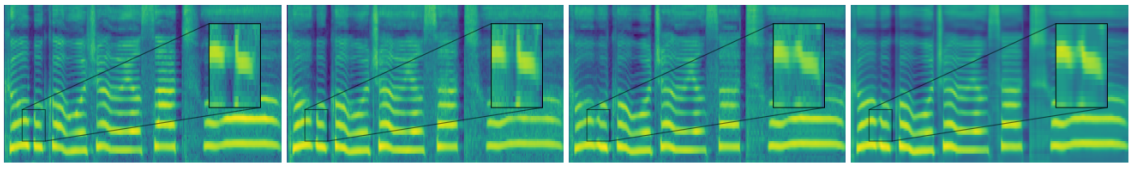


Figure 4: The diffusion trajectories of M and $\widetilde{M}$. Two distributions $q(M_t|M_0)$ and $q(\widetilde{M}_t|\widetilde{M}_0)$ become closer as t increases.



(a) *GT* (b) *Diffsinger* (c) *GAN-singer* (d) *FFT-Singer*

Figure 5: Visualizations of mel-spectrograms in four systems: GT, DiffSinger, GAN-Singer and FFT-Singer.

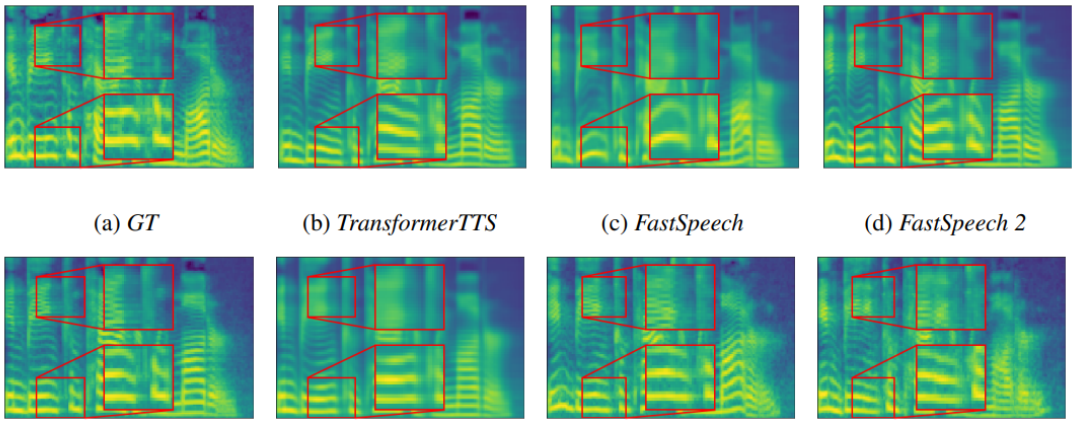
protaspeech

为了解决什么问题：1、解决VAE模型尺寸过小时模糊和不自然的现象。2、解决标准化流模型参数受限时表现不佳的问题。3、为了使用轻量级架构生成具有自然细节和丰富韵律的多样化语音

技术突破点：1、为了准确地建模韵律和mel频谱细节，我们采用一个带有增强先验的轻量级VAE，后跟具有强大条件输入的流式后置网络作为主要架构

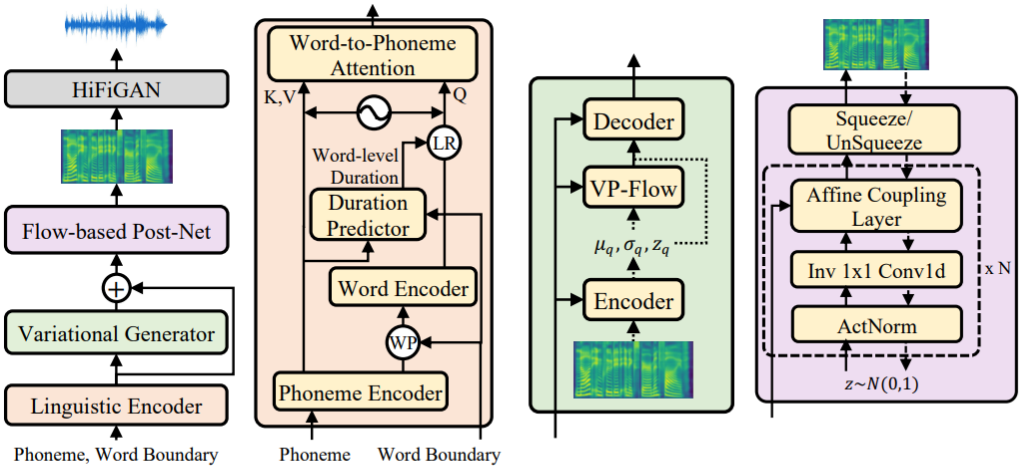
2、为了进一步压缩模型尺寸和内存占用，我们在后置网络的仿射耦合层中引入了分组参数共享机制

3、为了提高合成语音的表现能力和减少对文本和语音之间精确细粒度对齐的依赖性，我们提出了一种具有混合对齐的语言编码器，结合硬词级对齐和软音素级对齐，明确提取词级语义信息



(a) *GT* (b) *TransformerTTS* (c) *FastSpeech* (d) *FastSpeech 2*

(e) *Glow-TTS* (f) *BVAE-TTS* (g) *PortaSpeech (normal)* (h) *PortaSpeech (small)*

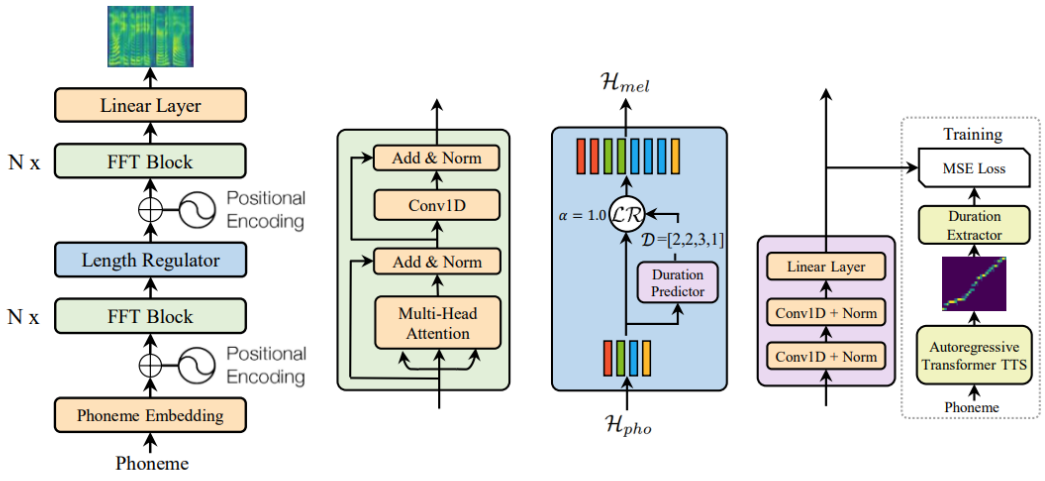


(a) PortaSpeech (b) Linguistic Encoder (c) Variational Generator (d) Flow-based Post-Net

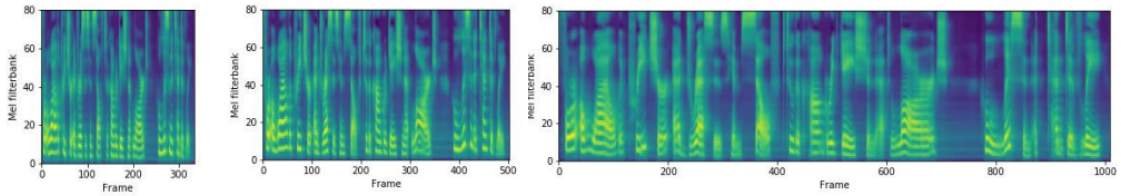
fastspeech

为了解决什么问题：为了解决基于神经网络的端到端模型遭受慢速推理（slow inference speed）的问题。且合成语音通常不够鲁棒性（即有些单词被跳过或重复）且缺乏可控性（如语速或声调控制）

技术突破点：提出了一种基于变压器的前馈网络，用于并行生成TTS的mel-spectrogram。从一个编码器-解码器的教师模型中提取注意力对齐，用于预测音素持续时间，并由长度调节器将源音素序列扩展以匹配目标mel-spectrogram序列的长度，以进行并行的mel-spectrogram生成。结果：模型在语音质量上与自回归模型相匹配，在特别难处理的情况下几乎消除了单词跳过和重复的问题，并且可以平滑地调整语速。与自回归变压器TTS相比，我们的模型将mel-spectrogram的生成速度提高了270倍，端到端语音合成速度提高了38倍。



(a) Feed-Forward Transformer (b) FFT Block (c) Length Regulator (d) Duration Predictor

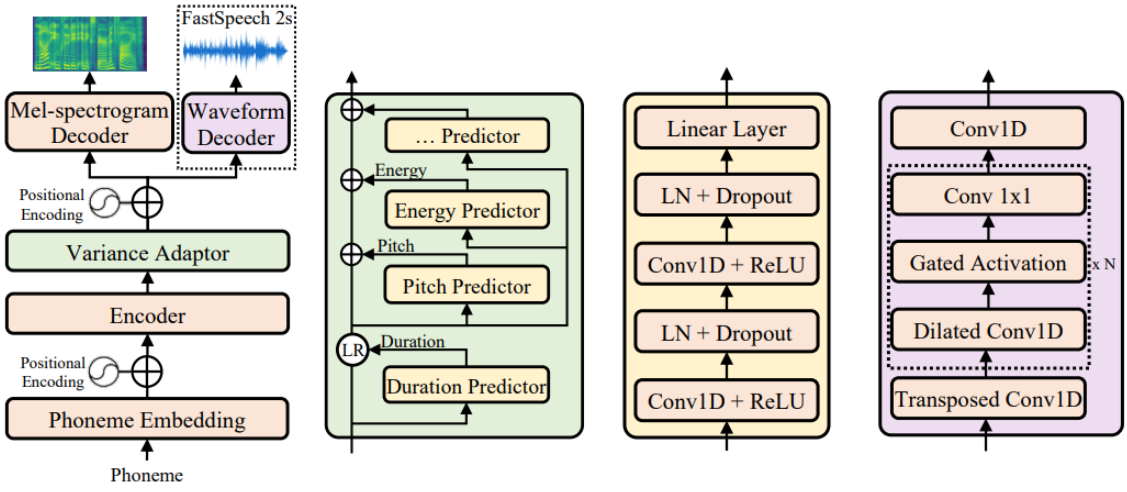


(a) 1.5x Voice Speed (b) 1.0x Voice Speed (c) 0.5x Voice Speed

fastspeech2

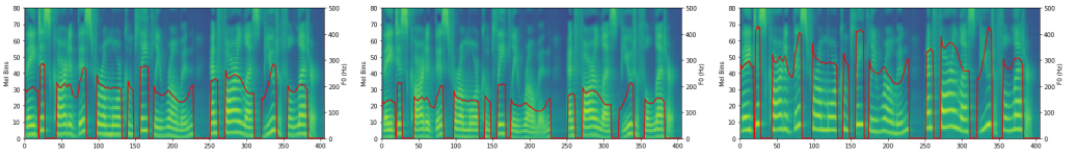
为了解决什么问题：FastSpeech模型的训练依赖于一个自回归的教师模型来进行持续时间预测（为了提供更多的信息作为输入）和知识蒸馏（为了简化输出的数据分布），但是它有如下缺点：1）教师-学生蒸馏管道过于复杂，2）从教师模型提取的持续时间不够准确，并且从教师模型蒸馏得到的目标mel-spectrogram会因为数据简化而失去信息，这都限制了语音质量。FastSpeech 2解决了上述问题，更好地解决TTS中的一对多映射问题。

技术突破点：1、直接使用真实目标进行模型训练，而不是使用教师的简化输出。2、引入更多的语音变化信息作为条件输入（如音高，能量和更精确的持续时间）。 具体而言，从语音波形中提取持续时间，音高和能量，并直接在训练过程中将其作为条件输入，在推理过程中使用预测值。3、进一步设计了FastSpeech 2s，首次尝试从文本中并行直接生成语音波形，享受完全端到端的培训好处



(a) FastSpeech 2 (b) Variance adaptor (c) Variance predictor (d) Waveform decoder

Figure 1: The overall architecture for FastSpeech 2 and 2s. LR in subfigure (b) denotes the length regulator operation proposed in FastSpeech. LN in subfigure (c) denotes layer normalization. Variance predictor represents duration/pitch/energy predictor.



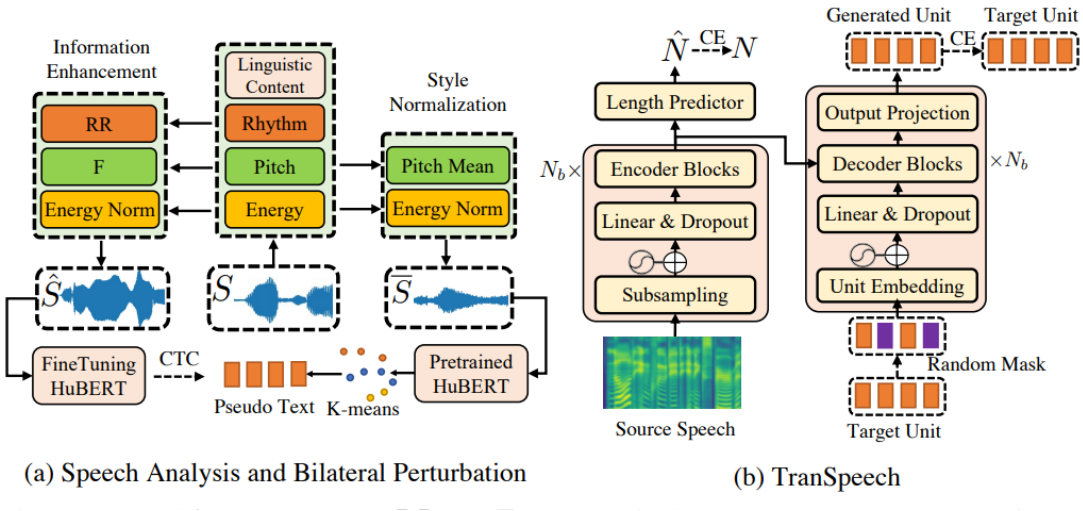
(a) $\hat{F}_0 = F_0$ (b) $\hat{F}_0 = 0.75F_0$ (c) $\hat{F}_0 = 1.50F_0$

Figure 3: The mel-spectrograms of the voice with different $\hat{F}_0$. F_0 is the fundamental frequency of original audio. The red curves denote $\hat{F}_0$ contours. The input text is "*They discarded this for a more completely Roman and far less beautiful letter.*"

transpeech

为了解决什么问题：S2ST仍然有以下挑战：1、声学多模性：由于声学属性（如节奏、音高和能量），来自相同内容的语音的离散单元可能是不确定的，这会导致翻译精度下降；2）高延迟：目前的S2ST系统利用自回归模型，在预测每个单元时都受先前生成序列的条件限制，无法充分利用并行处理的优势。该模型解决了这些内容

技术突破点：提出了双向扰动(BiP)来缓解声学多模式的问题。随着多模态性的减少，作者率先建立了非自回归S2ST，该技术重复掩盖和预测单元选择，并在仅几个周期内产生高精度结果。结果：BiP在平均BLEU上提高了2.9。此外，并行解码显示推理延迟显著降低，使推理速度比自回归技术提高了21.4倍。



(a) Speech Analysis and Bilateral Perturbation

(b) TranSpeech