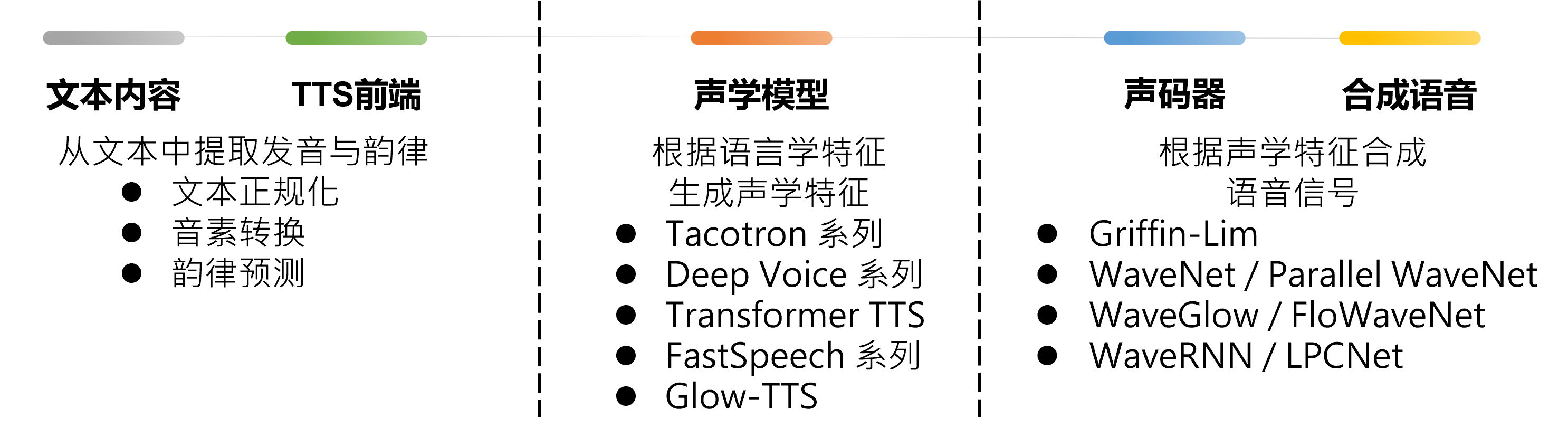
# Text to Speech自回归模型与实际应用

## 0 前言

### 0.1 TTS介绍



上述流程中，文本由TTS前端转换为音素，随后音素序列通过声学模型转化为Mel频谱，将Mel频谱输入到声码器中，最后由声码器将其转换为语音。其中Mel频谱由时间轴（横轴）与频率轴（纵轴）两部分组成，是一个二维数据。人听到的声音本质是一段波形，是由大量采样点形成的一维数据。

从Mel频谱到波形的过程中，需要声码器来完成频谱转换。

### 0.2 早期TTS方法

拼接法与参数法(不是很重要，放在前面简单介绍一下就行了)

拼接法（Concatenative synthesis）是指将预先录制或提取出来的语音片段（如音素、音节、单词等）进行拼接，以生成新的语音。拼接法的优点是生成的语音自然度高，因为它使用真实语音单位来构建语音输出。然而，缺点是需要耗费大量人力时间录制和标注语音片段，并且不能很好地适应新的单词或说话人。

参数法（Parametric synthesis）则是通过建立声学模型来合成语音。该模型通常基于语音科学知识和数学模型，并使用一组相关参数来描述语音信号的产生过程。根据不同的参数估计方法，参数法可以分为基频线性预测（Pitch-synchronous Linear Prediction, PSOLA）、隐马尔可夫模型（Hidden Markov Model, HMM）、神经网络模型等多种类型。与拼接法相比，参数法的优点是可适用于新的单词和说话人，并且可以使用较少的数据进行训练。但是，由于参数法在建模语音信号的时候经常涉及到一些简化假设，所以生成的语音可能会失去一些自然度。

这些传统的TTS技术存在着许多限制和不足，无法达到与人类自然语音相媲美的效果，缺乏自然度和表现力。这时候，深度学习技术的出现为TTS带来了新的突破。

深度学习TTS可以通过模仿人类的发音方式设计相应的神经网络模型，利用大量真实语音数据进行训练，从而生成更加逼真、自然、富有表现力的语音输出。与传统的拼接法和参数法相比，深度学习TTS具有更强的普适性和鲁棒性，能够提供更加精准、个性化的语音合成服务。

## 1 基于自回归模型的声码器

### 1.1 介绍

自回归模型是较早研究的模型，基于自回归模型的声码器是一种语音合成技术，旨在使用神经网络自回归模型来生成高质量的语音波形。

该声码器通常由两个主要组件组成：一个用于转换输入文本为语音特征的前置模型，以及一个使用自回归模型从这些特征生成对应语音波形的后置模型。

在训练阶段，该模型的输入文本和相应的波形序列用于计算损失函数，并优化模型参数。实际生成语音时，输入文本被送入前置模型，生成相应的语音特征，然后被送入后置模型，以生成对应的语音波形。

基于自回归模型的声码器是一种逐帧生成语音波形的方法，每个时间步都依赖于前一个时间步的输出，这种特点导致该类语音合成模型的合成速度较慢，在对速度与实时性要求较高的场景下难以应用。从语音质量的角度来看，基于自回归模型的声码器生成的语音自然、流畅，但这需要消耗较大的计算资源和计算时间。

由于自回归模型的文本段与Mel频谱端之间的联系主要由Attention机制(本质在于使机器自身去学习文本与语音之间的映射关系)去保证，但Attention机制的映射存在不稳定性，基于自回归模型的声码器还存在重复吐词或漏词现象，不适用于商业环境。

### 1.2 应用广泛的自回归模型声码器

#### 1.2.1 WaveNet：A GENERATIVE MODEL FOR RAW AUDIO

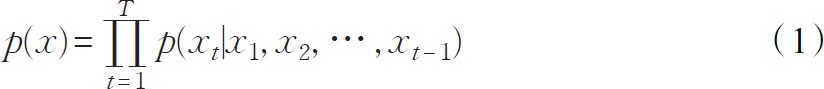
WaveNet是由Google DeepMind于2016年推出的自回归模型，是一种用于生成原始音频的深度神经网络模型，其主模型基于卷积神经网络PixelCNN架构实现。模型使用语音记录进行训练，并通过学习音频波形的概率分布来生成新的语音波形。具体地说，WaveNet采用了一种逐步增量的采样策略，按顺序预测每个时刻的输出，直到生成整个波形。

WaveNet一秒钟生成的语音：

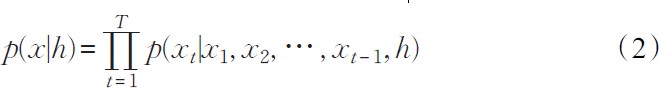


##### 1.2.1.1训练方法

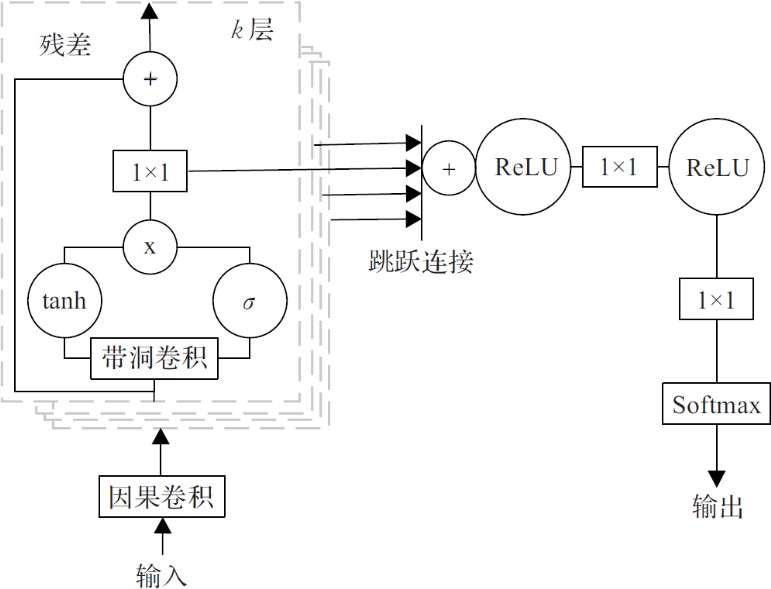
展开而言，在训练时，模型首先将语音序列联合概率 x = {x1 ,x2 ,…,xT } 分解为各时刻条件概率的乘积，如公式（1） 所示，其中x 是语音波形值序列，xt 是一个时刻波形值：



下一步，需要根据公式 2 对声码器进行建模：

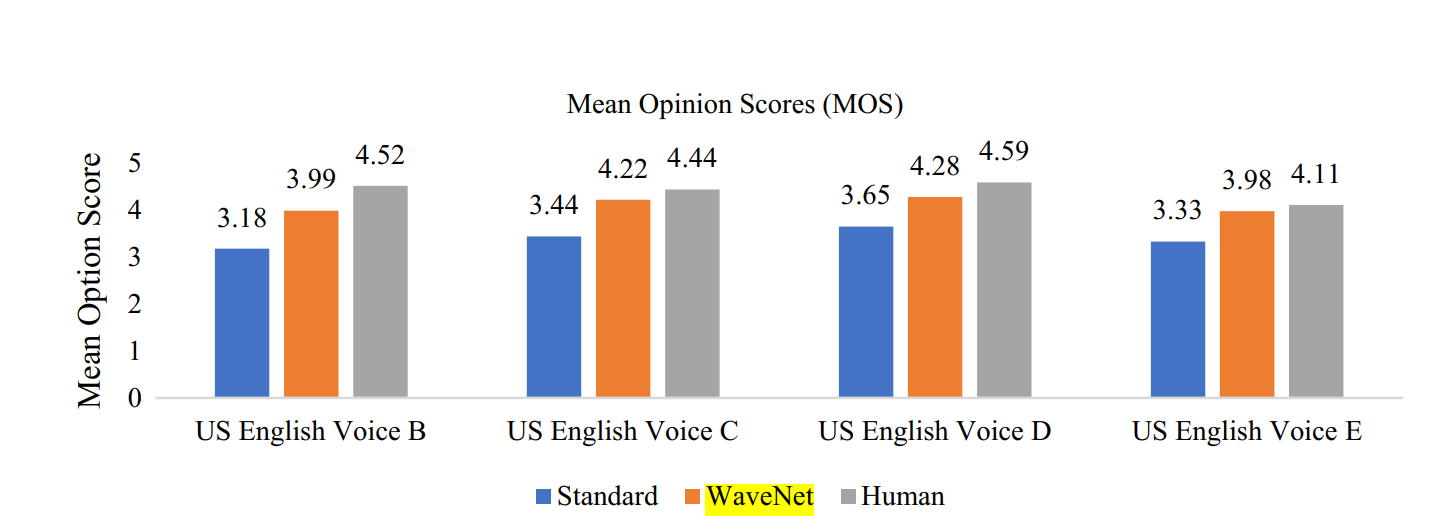


在建模时，Wavenet有两种方法：全局方法和局部方法。全局方法是指模型接收单额外输入条件h ，该条件在所有的时间点上影响模型的输出；局部条件建模方法是指模型有第二种时间序列 ht ，通过对原始数据的低采样率获得。

图 1：Wavenet 模型

##### 1.2..1.2 WaveNet评价

综合而言，WaveNet是一种自回归模型，训练时间和计算资源消耗较大，需要大量的语音样本和高性能计算机，因此生成语音的速度较慢，每秒只能生成24,000个样本；然而它的优势在于生成的合成语音自然而真实，比传统的文本转语音系统更加逼真。在使用WaveNet生成的不同声音的平均选项分数（MOS）评级中，其生成的美式英语语音的平均MOS评级为1-5分的4.1，比标准语音提高了20%。这将自然人类语音和合成语音之间的差距缩小了70％



##### 1.2.1.3 遇到的问题及解决方法

为了处理原始音频所需的长期时间依赖性，开发了基于扩展因果卷积的新架构。(文章原文：In order to deal with long-range temporal dependencies needed for raw audio generation, we develop new architectures based on dilated causal convolutions, which exhibit very large receptive fields.)

基于扩张因果卷积（dilated causal convolution）的卷积结构，该结构同时支持因果卷积（causal convolution）和扩张卷积，可以让模型很好地捕捉长时序信号的依赖性。因果卷积只保留前面的时刻信息，避免了时间上的信息泄露，而扩张卷积则扩大了卷积核的感受野，为模型提供了更多的历史上下文信息。因此，基于扩张因果卷积的结构使得WaveNet可以处理更长的时间序列，并获得更准确和自然的音频生成结果。

在WaveNet声码器中，由于卷积神经网络难以捕捉语音信号的动态信息，可能会导致合成语音中出现噪音和失真等问题。文献[2](Tachibana K, Toda T, Shiga Y, et al. An investigation of noise shaping with perceptual weighting for WaveNet-based speech generation[C]. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018: 5664-5668.)通过引入基于感知权重的噪声整形技术来解决这一问题。具体地，该方法利用预测残差的梅尔频谱特征进行感知权重建模，并将感知权重应用于形状参数的预测中，从而产生更加自然流畅的语音样本。

##### 1.2.1.4 解决的问题以及技术突破

相较于拼接法和参数法，WaveNet生成了更加自然流畅、饱满的语音信号。

WaveNet使用卷积神经网络自动学习特征，不再需要通过人工选择声学特征检测语音的结构周期，避免了引入失真。

遇到的问题、出现原因：

相较于以前的技术(拼接法和参数法)，WaveNet的出现主要是为了解决语音合成中声音不自然、语调不连贯等问题。

WaveNet介绍：

WaveNet是一种深度神经网络，可以生成原始声音波形。该模型是全概率模型和自回归模型，每个音频样本的预测分布都是基于之前所有音频样本的条件的。尽管如此，实验证明它可以高效地处理每秒数万个音频样本的数据集。

特点：

WaveNet在文本到语音领域表现非常出色，人类听众评价其发音更加自然，而且英语和普通话都表现优异，超过了最好的参数化和拼接语音系统。一个单独的WaveNet可以以相同精度捕捉许多不同说话人的特征，并能够通过对说话人身份的条件限制来在不同说话人之间进行切换。在训练模型生成音乐时，我们发现它能够生成新颖且高度逼真的乐曲片段。此外，该模型还能被用作辨别性模型，返回对音素识别有希望的结果。

核心：

相比于传统的拼接法、参数法，使用了PixelNet模型。

效果：

WaveNet的出现解决了语音合成中的声音不自然、语调不连贯等问题，并提出了一种全概率模型和自回归模型，实现更加高效的训练。此外，WaveNet还支持多说话人语音合成，并可用于生成音乐片段。他们的实验表明，WaveNet的表现非常出色，可以在多个领域实现优异的性能。

[1609.03499.pdf (arxiv.org)](https://arxiv.org/pdf/1609.03499.pdf)

#### 1.2.2 WaveRNN

遇到的问题，出现原因：

与WaveNet相比，WaveRNN的出现是为了解决神经网络架构复杂，采样过程中时间效率低下的问题。WaveRNN主要研究了文本到语音合成的采样效率优化。

WaveRNN介绍：

WaveNet是一个十分复杂且深度的模型，这个性质限制了WaveNet的应用场景，WaveRNN最初的设计目的就是保持WaveNet的高速序列生成，它使用简化模型、矩阵稀疏化、并行序列等技术显著提升了序列生成速度。WaveRNN采用了双路 softmax 层，在保持语音合成质量的同时，其单层循环神经网络相比 WaveNet 更加紧凑，可以在 GPU 上 4 倍快速地生成 24 kHz 16 位音频。其次，通过权重剪枝技术，降低了 WaveRNN 中权重的数量。实验结果显示，对于固定数量的参数，大型稀疏网络比小型稠密网络表现更好，这种关系在稀疏度超过 96% 时仍然成立。稀疏 WaveRNN 的权重数量较少，能够在移动 CPU 上实时采样高保真音频。最后，提出了基于子缩放方法的新的生成方案，将长序列折叠成多个较短的序列，可以一次性生成多个样本。子缩放 WaveRNN 可以每次生成 16 个样本，且没有质量损失，提供了一种增加采样效率的另一种方法。

特点：

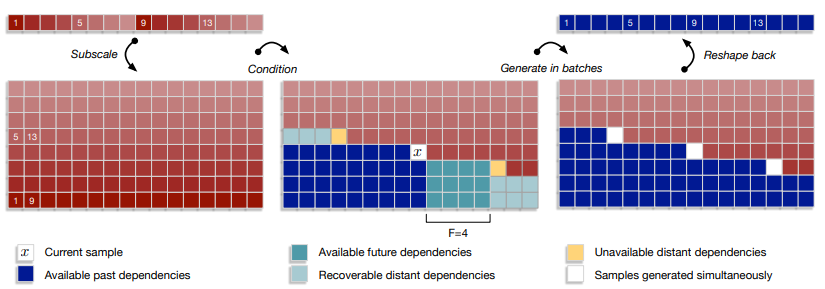
WaveRNN是基于循环神经网络PixelRNN的自回归模型。

核心：

WaveRNN对WaveNet进行了改进，用单GRU来处理语音信号的时序，通过矩阵稀疏化和拆分策略实现加速。

效果：

WaveRNN的网络模型较为简单，时间复杂度较低，推理需要的时间相比WaveNet更短，应用良好，甚至可以在CPU上实现实时语音合成，即使是手机的CPU都可以应用。



[Efficient Neural Audio Synthesis (arxiv.org)](https://arxiv.org/pdf/1802.08435.pdf)

#### 1.2.3 LPCNet

遇到的问题，出现原因：

神经语音合成模型在最近取得了高质量语音合成成果，但这些新模型通常需要强大的GPU才能实现实时操作，因此需要减少它们的复杂度以开辟更多应用领域。

LPCNet介绍：

LPCNet模型是一种结合了线性预测和递归神经网络的WaveRNN变体，并可以显著提高语音合成的效率。通过实验证明LPCNet在相同规模的网络下，比WaveRNN具有更高的语音质量，并且高质量的LPCNet语音合成只需要不到3 GFLOPS的运算复杂度。这使得神经语音合成应用可以轻松部署在低功耗设备上，如嵌入式系统和手机。

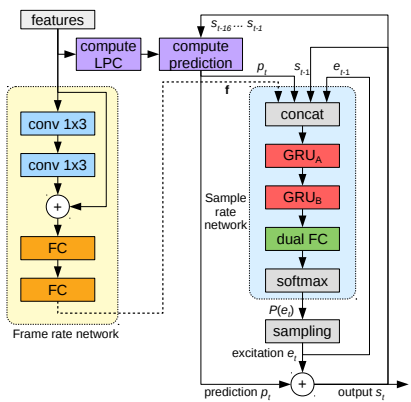
特点:

通过LPC filter预测语音的线性部分，以bfcc作为local condition通过GRU预测激励信号。

核心：预测激励信号，LPC

效果：

LPCNet的出现解决了神经语音合成模型需要强大计算资源的问题，并通过采用线性预测和递归神经网络的结合方式来提高语音合成的效率。LPCNet的效果表现出了更高的语音质量，同时运算复杂度较低，因此更易于在低功耗设备上部署实现。



[LPCNet:ImprovingNeuralSpeechSynthesisThroughLinearPrediction (jmvalin.ca)](https://jmvalin.ca/papers/lpcnet_icassp2019.pdf)

## 2 实际应用

### 2.1 应用介绍

TTS，即文本到语音合成，可以将任意文本内容转化为语音输出，因此具有广泛的应用前景，包括但不限于以下几个方面：

语音辅助技术：TTS可以帮助视力受损人士获取文字信息，并且对于老人、残障人士等群体也具有很大的帮助作用。

教育和培训领域：TTS可以将教材、课件等内容转化为语音，实现自动朗读，更好地服务于学生和教师。

电子书和有声读物：TTS技术可以将图书、报纸、杂志等内容实时转换为语音阅读，使得用户可以不受时间和场所的限制而轻松阅读。

智能客服和语音交互：TTS还可以被应用于智能客服、语音助手等领域，实现更加便捷的语音交互，提高用户体验。

影视配音和虚拟主播：在影视配音和虚拟主播领域，TTS技术可以为电影、动画等节目制作语音配音，同时也可以生成虚拟角色的语音和情感表达。

总之，TTS的应用场景非常广泛，其能够将文字信息转化为语音输出，为用户带来更加便捷和自由的体验，同时也具有很大的商业价值。

### 2.2 transpeech

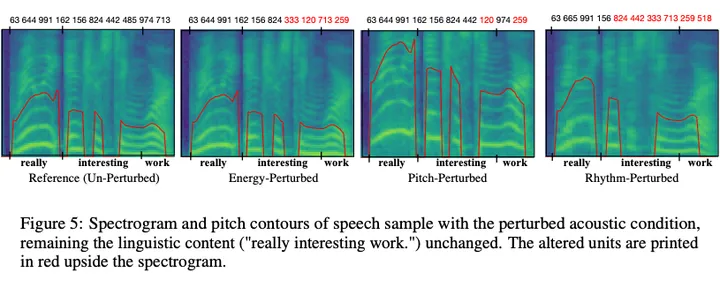
TranSpeech是一个语音到语音的翻译模型，可以将语音信号翻译成另一种语言的语音信号。其中，TTS（Text-to-Speech）技术被广泛应用于该模型中，用于将生成的文本转化为声音。

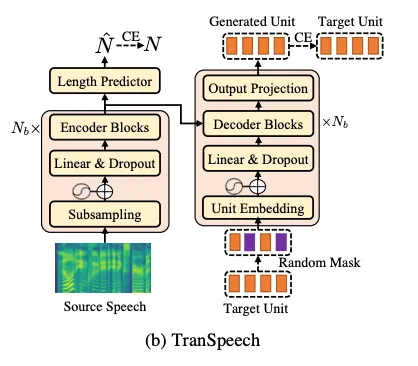
TranSpeech中TTS的应用主要体现在以下两个方面：

语音合成：通过TTS技术，TranSpeech可以将生成的文本信息转化为语音信号输出，从而实现语音到语音的翻译功能。

信息增强：TranSpeech还使用了双向扰动（BiP）技术来增强语音信息的表达。BiP技术包括样式归一化和信息增强阶段，旨在学习语音样本中的语言信息，并生成更确定的语音表示形式。

在跨语言交流、文化交流以及旅游等方面，TranSpeech可以提供实时的语音翻译服务，方便人们交流和沟通；在教育领域，TranSpeech可以实现自动语音听写和语音朗读，帮助学生更好地学习和理解课程内容；在智能客服和虚拟助手方面，TranSpeech也可以提供更加自然、流畅的语音交互体验，提高用户满意度。





[2205.12523.pdf (arxiv.org)](https://arxiv.org/pdf/2205.12523.pdf)

### 2.3 diffsinger

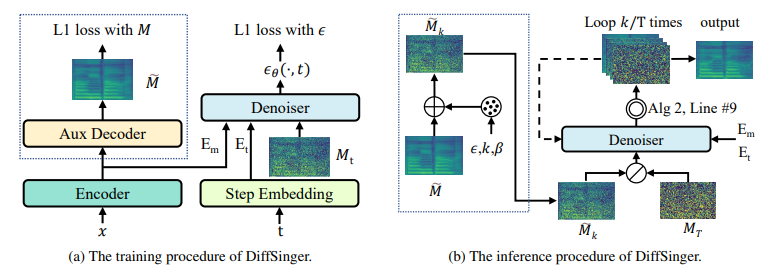
DiffSinger是一种基于扩散概率模型的歌声合成系统，其目标是合成高质量和富有表现力的歌唱声音。TTS技术在其中被用于将文本信息转化为音频输出。

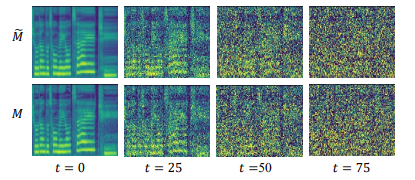
DiffSinger中TTS的应用主要体现在以下两个方面：

文本输入：DiffSinger接受音乐谱面和歌词文本作为输入，并利用TTS技术将歌词转化为对应的音频信息。

声音合成：通过Diffusion概率模型，DiffSinger可以生成高质量、自然的歌唱声音。同时，DiffSinger引入了浅扩散机制来加速推理速度，并采用边界预测方法自适应地确定浅层步数。

在音乐娱乐方面，DiffSinger可以帮助音乐人和制作人合成更加自然、富有表现力的歌唱声音，从而提高音乐的艺术价值和市场竞争力；在教育领域，DiffSinger可以为音乐教学提供更加清晰、直观的演示方式，提高学生的学习效果和兴趣；在虚拟主播和游戏领域，DiffSinger也可以为角色和场景配音提供更加丰富、真实的声音效果，从而增强用户的沉浸感和游戏体验。



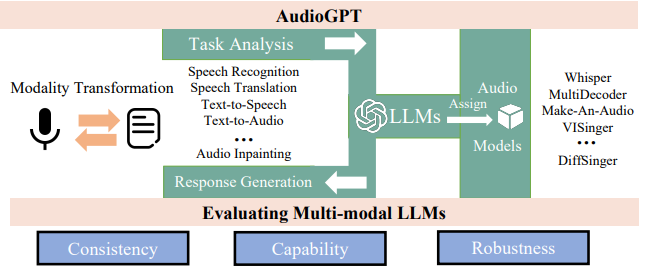


[2105.02446v2.pdf (arxiv.org)](https://arxiv.org/pdf/2105.02446v2.pdf)

### 2.4 audiogpt

AudioGPT是由一组中国和美国研究人员在2023年4月发布的研究项目，是一个多模态人工智能系统，旨在为当前的大型语言模型（LLMs）添加处理复杂音频信息和实现语音对话的功能。其中，TTS技术也被应用于该系统中以支持自然的语音交互。它可以在类似ChatGPT的聊天机器人界面中使用。

Transpeech是AudioGPT的核心之一，利用基于深度学习的TTS技术，可以将文本信息自动转化为逼真、自然的语音输出。通过Transpeech的应用，用户可以直接通过语音来与AudioGPT进行交互，从而更加自然、高效地使用多模态人工智能系统进行各种任务的完成，如查询、聊天、翻译等等。



[2304.12995.pdf (arxiv.org)](https://arxiv.org/pdf/2304.12995.pdf)