Text to Speech

- **前端模块**：将任意文本转为语言学特征，通常包括文本正则化，分词，词性预测，多音字消歧、韵律估计等子模块。文本正则化可以将一些书面表达转为口语表达，如1%转为“百分之一”，1kg转为“一千克”等。分词和词性预测是韵律估计的基础，韵律词和韵律短语会在分词和词性信息的基础上生成。比较知名的开源分词工具有海量分词和结巴分词。
- **后端模块**：将前端提供的语言学特征经过算法生成声学特征。声学特征可以是mfcc，也可以是f0，sp和ap这些声码器特征。后端模块有统计参数建模，单元挑选与拼接，神经网络模型三个技术主线。
    - 参数合成：参数语音合成系统的特点是，在语音分析阶段，需要根据语音生成的特点，将语音波形(speech waves) 通过声码器转换成频谱，基频，时长等语音或者韵律参数。在建模阶段对语音参数进行建模。并且在语音合成阶段，通过声码器从预测出来的语音参数还原出时域语音信号。参数语音合成系统的优势在于模型大小较小，模型参数调整方便（说话人转换，升降掉），而且合成语音比较稳定。缺点在于合成语音音质由于经过参数化，所以和原始录音相比有一定的损失。
    - 拼接语音：拼接语音合成系统的特点是，不会对原始录音进行参数化，而会将原始录音剪切成一个一个基本单元存储下来。在合成过程中，通过一些算法或者模型计算每个单元的目标代价和连接代价，最后通过Viterbi算法并且通过PSOLA(Pitch Synchronized Overlap-Add)或者WSOLA(Waveform Similarity based Overlap-Add)等信号处理的方法“拼接”出合成语音。因此，拼接语音合成的优势在于，音质好，不受语音单元参数化的音质损失。但是在数据库小的情况下，由于有时挑选不到合适的语音单元，导致合成语音会有Glitch 或者韵律、发音不够稳定。而且需要的存储空间大。
    - 神经网络：WaveNet 波形统计语音合成是Deep Mind 首先提出的一种结构，主要的单元是 Dilated CNN （空洞卷积神经网络）。这种方法的特点是不会对语音信号进行参数化，而是用神经网络直接在时域预测合成语音波形的每一个采样点。优势是音质比参数合成系统好，略差于拼接合成。但是较拼接合成系统更稳定。缺点在于，由于需要预测每一个采样点，需要很大的运算量，合成时间慢。WaveNet 证明了语音信号可以在时域上进

行预测，这一点以前没有方法做到。现阶段WaveNet是一个研究热点。

- **声码器模块**：将声学特征转为语音波形，有相位恢复算法Griffin Lim，传统声码器WORLD和STRAIGHT，神经声码器WAVENET，WAVERNN，SAMPLERNN和WAVEGLOW。

Vocoder

- 声码器：复现声音信号，重现声音细节

Acoustic Model