1. 可以从语音质量，合成速度以及模型复杂度多个角度去入手
2. 能不能实现一个完全端到端的，直接输出波形，而不是中间会有谱