一种端到端语音合成中的高效解码自注意力网络

赵伟[1,2]，许力[1,2]
[1]浙江大学电气工程学院，中国杭州市，310027
[2]浙江大学机器人研究院，中国余姚市，315400

摘要：自注意力网络由于其并行结构和强大的序列建模能力，被广泛应用于语音合成（TTS）领域。然而，当使用自回归解码方法进行端到端语音合成时，由于序列长度的二次复杂性，其推理速度相对较慢。当部署设备未配备图形处理器（GPU）时，该效率问题更加严重。为解决该问题，提出一种高效解码自注意力网络（EDSA）作为替代。通过一个动态规划解码过程，有效加速 TTS 模型推理，使其具有线性计算复杂度。基于普通话和英文数据集的实验结果表明，所提 EDSA 模型在中央处理器（CPU）和 GPU 上的推理速度分别提高 720%和 50%，而性能几乎相同。因此，在 GPU 资源有限的情况下，该方法可使此类模型的部署更加容易。此外，所提模型在域外语言处理上可能比基线 Transformer TTS 性能更好。

关键词：高效解码；端到端；自注意力网络；语音合成
https://doi.org/10.1631/FITEE.2100501