TTS 端到端语音合成

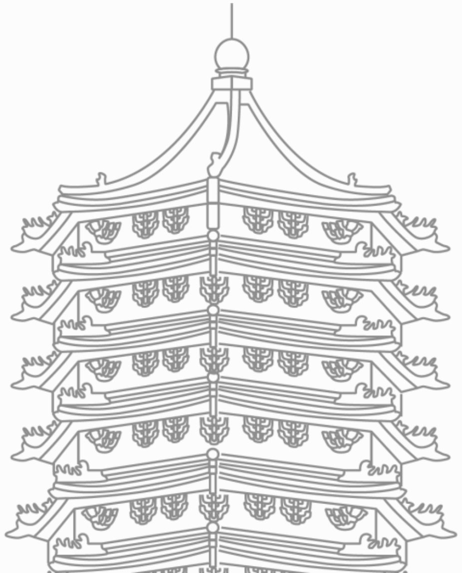
框 架 结 构　/　结 果 分 析　/　系 统 展 示

Neural network based end-to-end speech synthesis / Text To Speech

小组汇报

CONTENTS

Based on tacotron2, SV2TTS, tensorflowTTS, Real-time Voice Cloning and mockingbird

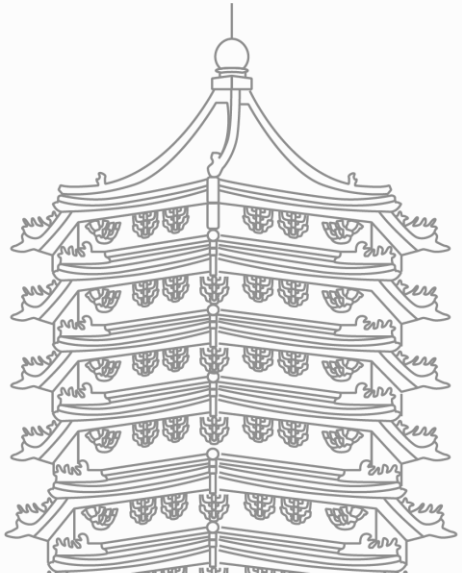


PART 01

系统原理

Natural TTS Synthesis by Conditioning WaveNet on Mel-Spectrogram Predictions

Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis



系统原理-Tacotron2

Natural TTS Synthesis by Conditioning WaveNet on Mel-Spectrogram Predictions

端到端神经网络语音合成

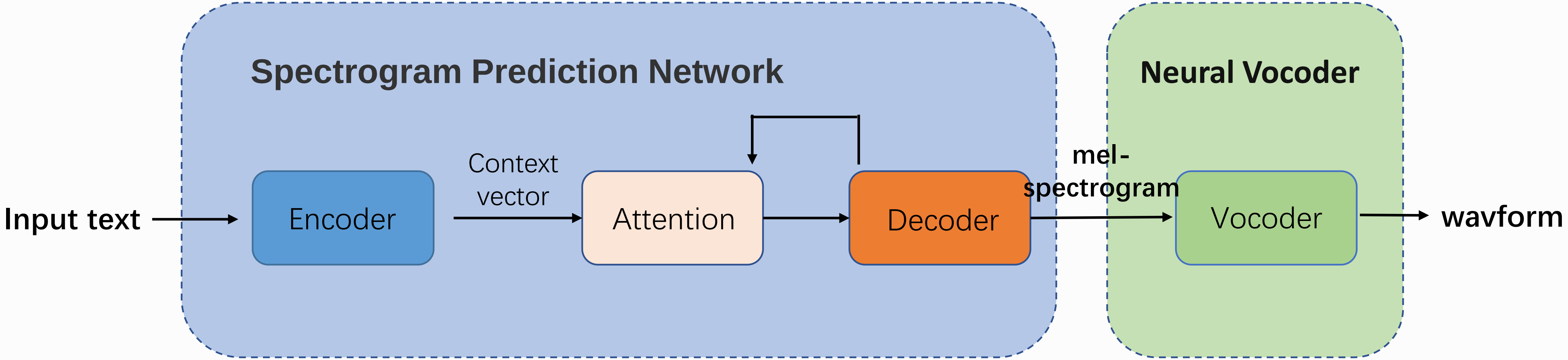
- **声谱预测网络**：文本序列 → 帧级语音特征（以梅尔频谱表示）
- **神经声码器**：帧级语音特征 → 语音波形

系统原理-Tacotron2

Natural TTS Synthesis by Conditioning WaveNet on Mel-Spectrogram Predictions

端到端神经网络语音合成

- **声谱预测网络**：文本序列 ⟶ 帧级语音特征（以梅尔频谱表示）
- **神经声码器**：帧级语音特征 ⟶ 语音波形



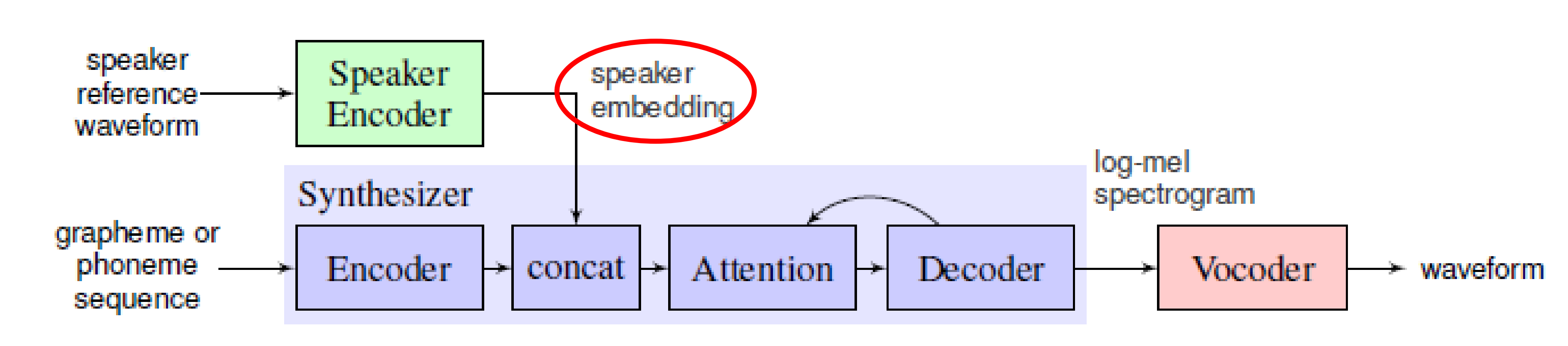
01 系统原理-SV2TTS

Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis

基于Google 2017年发布的论文
《Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis(SV2TTS)》

模型分为三个独立训练的组件:

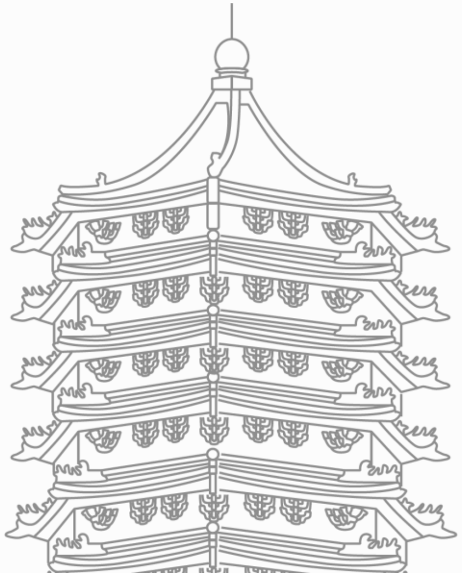
- 编码器(encoder):生成代表说话人音色的向量
- 合成器 (synthesizer):将文本转换成梅尔频谱图
- 声码器(vocoder):将梅尔频谱图转换成waveform



PART 02

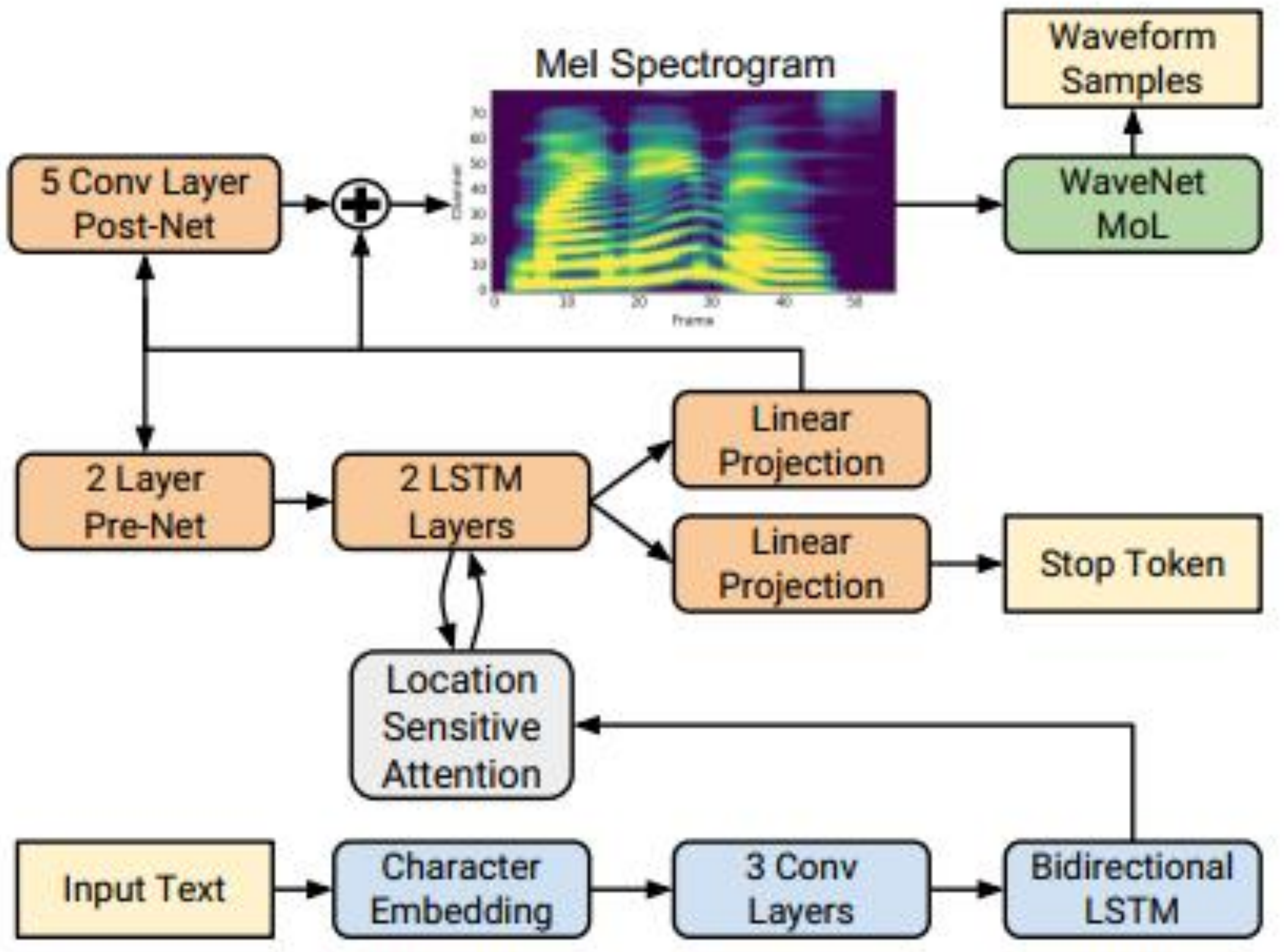
框架结构

print the presentation and make it
into a film to be used in a wider field
The user can demonstrate on a
projector or computer



02 框架结构：Tacotron2

Natural TTS Synthesis by Conditioning WaveNet on Mel-Spectrogram Predictions



Vocoder (WaveNet MoL)

Decoder

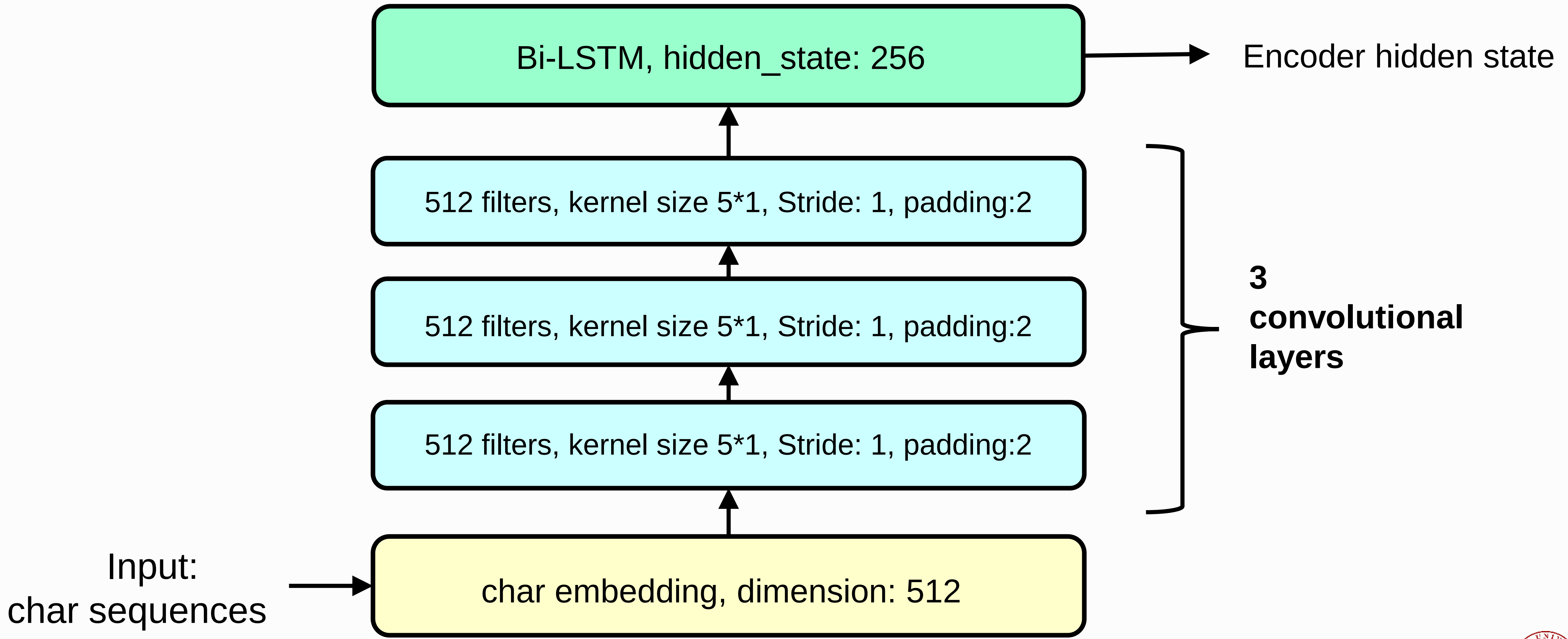
Attention (Location-sensitive)

Encoder

2.1 Tacotron2: Encoder

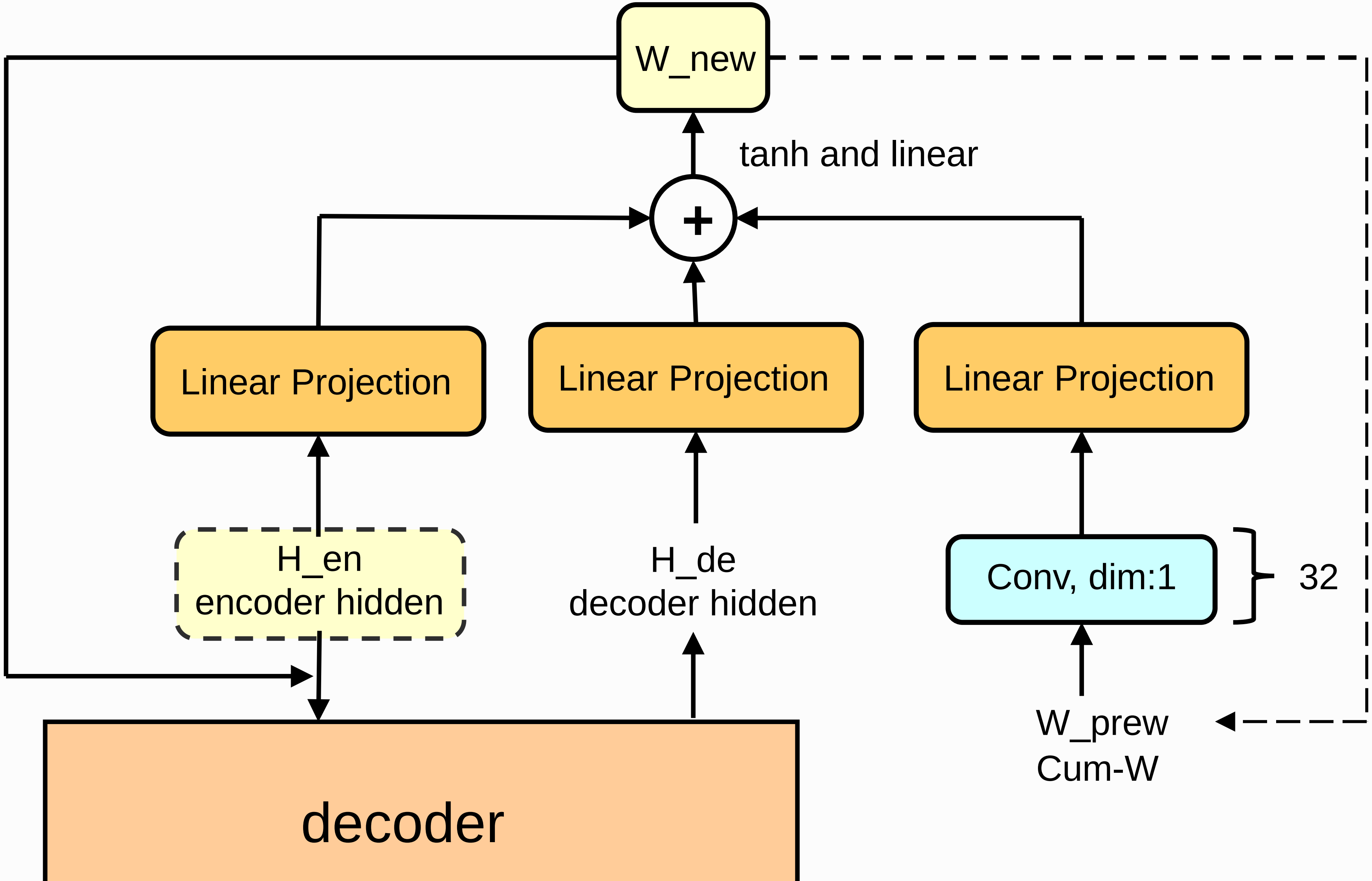
Natural TTS Synthesis by Conditioning WaveNet on Mel-Spectrogram Predictions: encoder

Structure: Embedding + 3*Conv_layers + Bidirectional_LSTM



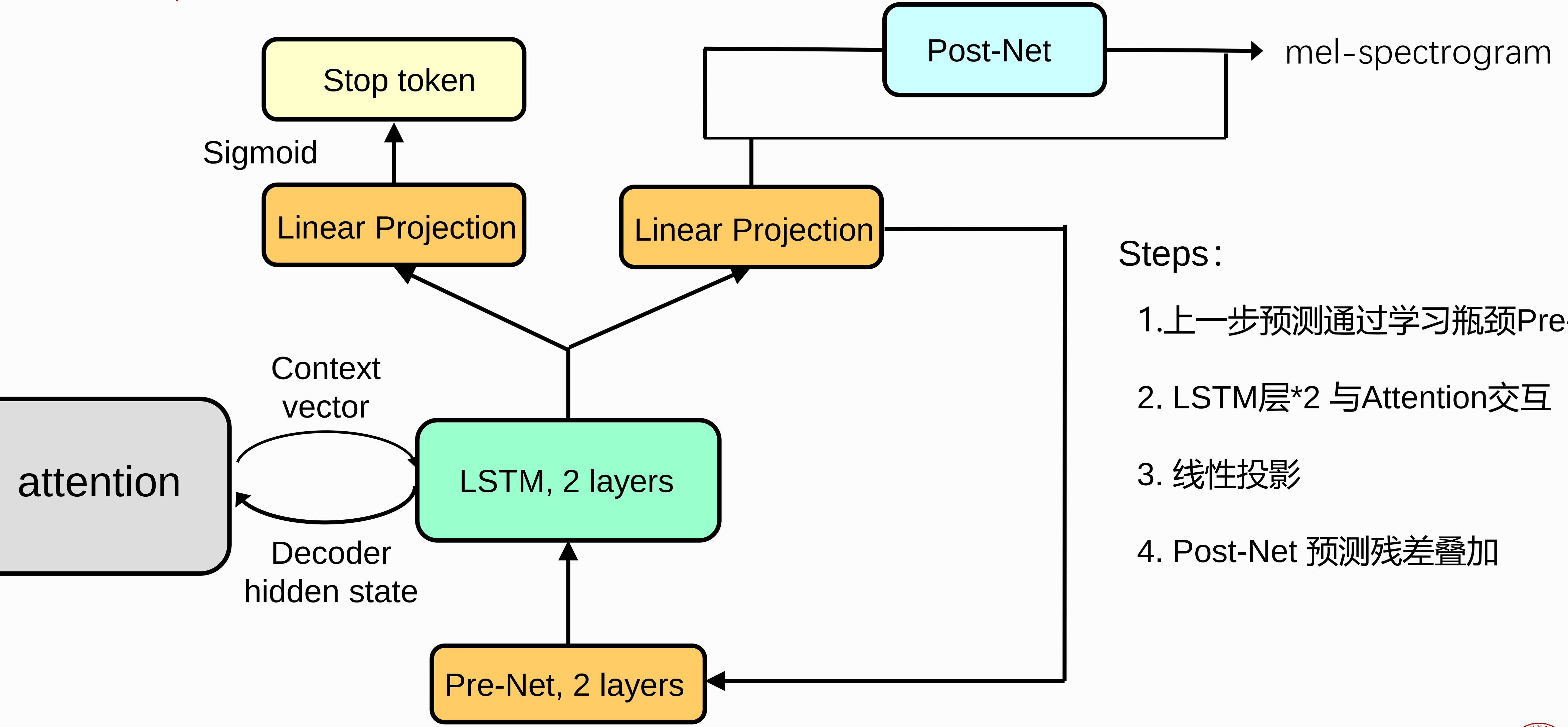
Tacotron2: Attention

Natural TTS Synthesis by Conditioning WaveNet on Mel-Spectrogram Predictions: attention



Tacotron2： Decoder

Natural TTS Synthesis by Conditioning WaveNet on Mel-Spectrogram Predictions: decoder



Steps：

1.上一步预测通过学习瓶颈Pre-Net

2. LSTM层*2 与Attention交互

3. 线性投影

4. Post-Net 预测残差叠加

Tacotron2： Vocoder

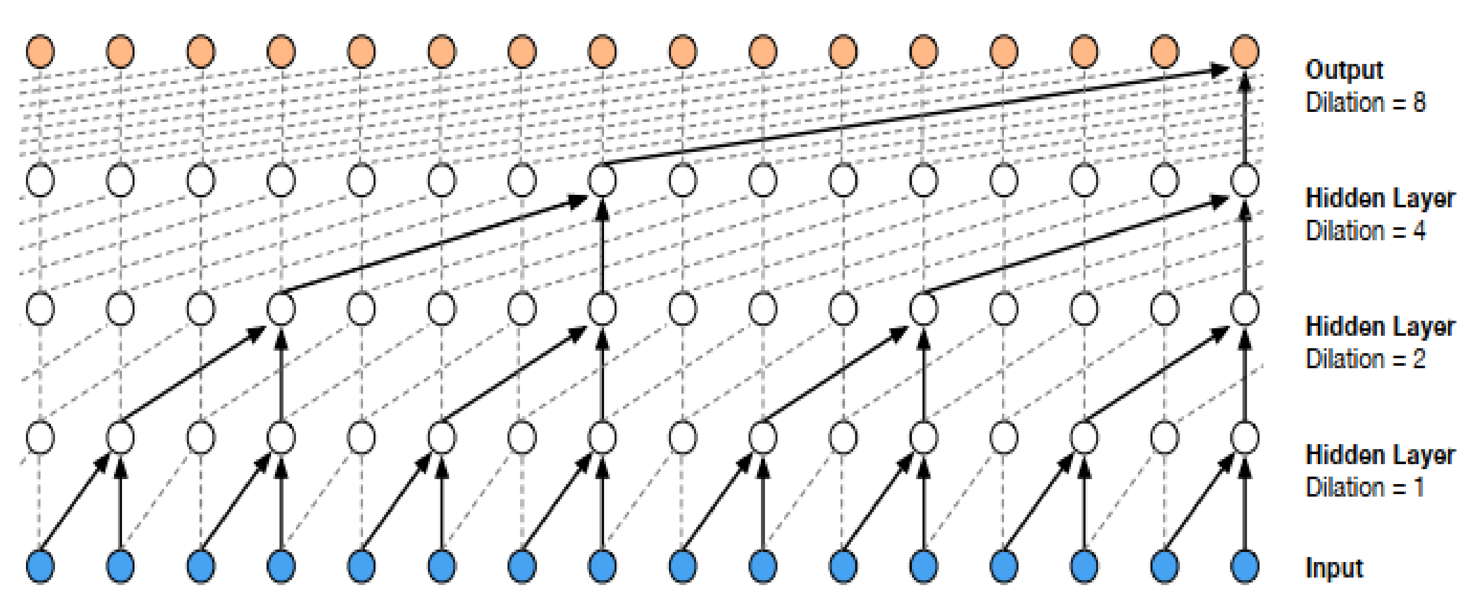
Natural TTS Synthesis by Conditioning WaveNet on Mel-Spectrogram Predictions: vocoder

Table 3: Mean Opinion Scores

Model	MOS	95% CI
Griffin Lim	1.57	±0.04
WaveGlow	4.11	±0.05
WaveNet	4.05	±0.05
MelGAN	3.61	±0.06
Original	**4.52**	**± 0.04**

在ljspeech上声码器主观意见分数比较

WaveNet 使用的膨胀因果卷积示意图



SV2TTS：Synthesizer & Vocoder

Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis

合成器：

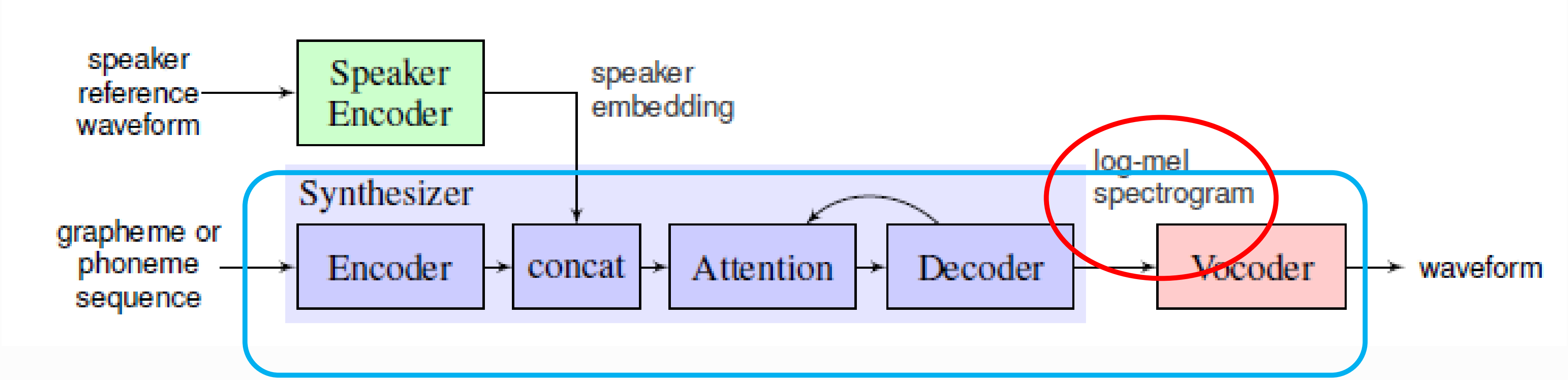
基于Tacotron2；

输入文本，生成对应的梅尔频谱图

声码器：

论文中采用WaveNet，项目实现用了HiFiGAN；

将梅尔频谱图转换为最终的语音

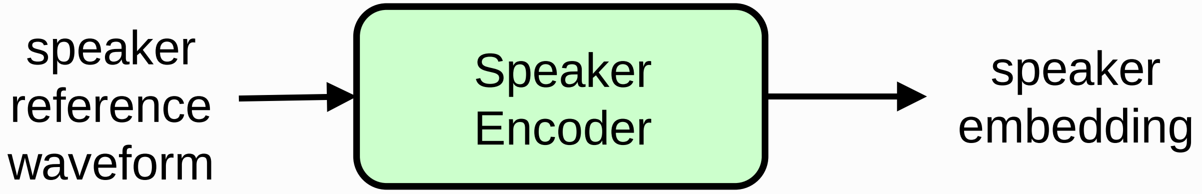


2.2 SV2TTS：Speaker Encoder

Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis

数据集：

- 来自多个说话者(>18K人，约36M条语句)
- 嘈杂
- 短暂(只有几秒)
- 无文本



训练： 采用GE2E损失函数进行说话者验证

生成： speaker embedding (嵌入向量)

2.2 SV2TTS：GE2E

generalized end-to-end algorithm

即广义端到端(generalized end-to-end，GE2E)损失

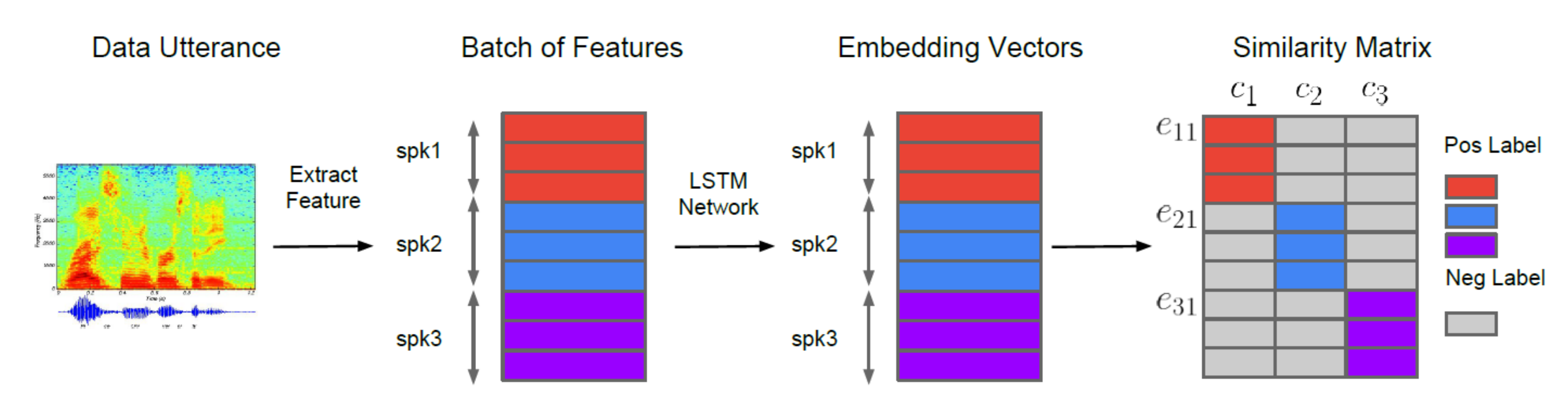
训练：

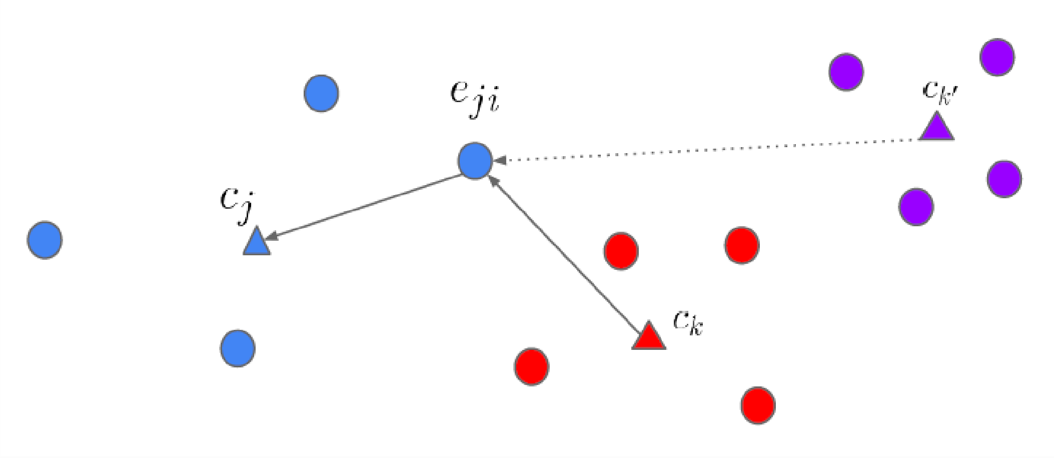
- 一个batch中包含N个说话者，平均每个说话者提供M条语句，共N*M条语句；以x_ji表示从说话人j第i条语句中提取的特征向量
- 通过LSTM网络得到输出结果f(x_ji;w)，正则化后则有 $\mathbf{e}_{ji} = \frac{f(\mathbf{x}_{ji};\mathbf{w})}{||f(\mathbf{x}_{ji};\mathbf{w})||_2}$ ，是为说话人j第i条语句的embedding向量
- 第j个说话者的embedding向量的中心定义为 $\mathbf{c}_k = \mathbb{E}_m[\mathbf{e}_{km}] = \frac{1}{M}\sum_{m=1}^{M}\mathbf{e}_{km}$
- 定义相似矩阵 Sji,k = w*cos(e_ji; c_k) + b 为每个embedding向量e_ji与所有中心c_k之间的缩放余弦相似性

2.2 SV2TTS：GE2E

generalized end-to-end algorithm

训练时，我们希望每条语句的embedding向量与各自说话者的中心相似，同时远离其他说话者的中心





SV2TTS：GE2E

generalized end-to-end algorithm

两种实现方式： 两种损失

- Softmax

  在S_ji,k上设置softmax，使输出在k=j时等于1，否等于0；则每个embedding

  向量e_ji上的损失可以定义为 $$L(\mathbf{e}_{ji}) = -\mathbf{S}_{ji,j} + \log \sum_{k=1}^{N} \exp(\mathbf{S}_{ji,k})$$

- Contrast(对比度损失) $$L(\mathbf{e}_{ji}) = 1 - \sigma(\mathbf{S}_{ji,j}) + \max_{\substack{1 \leq k \leq N \\ k \neq j}} \sigma(\mathbf{S}_{ji,k})$$

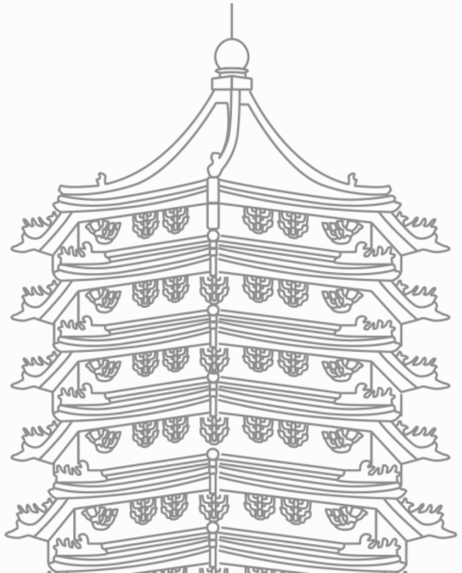
最终GE2E损失LG 是相似矩阵上所有损失的总和

$$L_G(\mathbf{x}; \mathbf{w}) = L_G(\mathbf{S}) = \sum_{j,i} L(\mathbf{e}_{ji})$$

PART 03

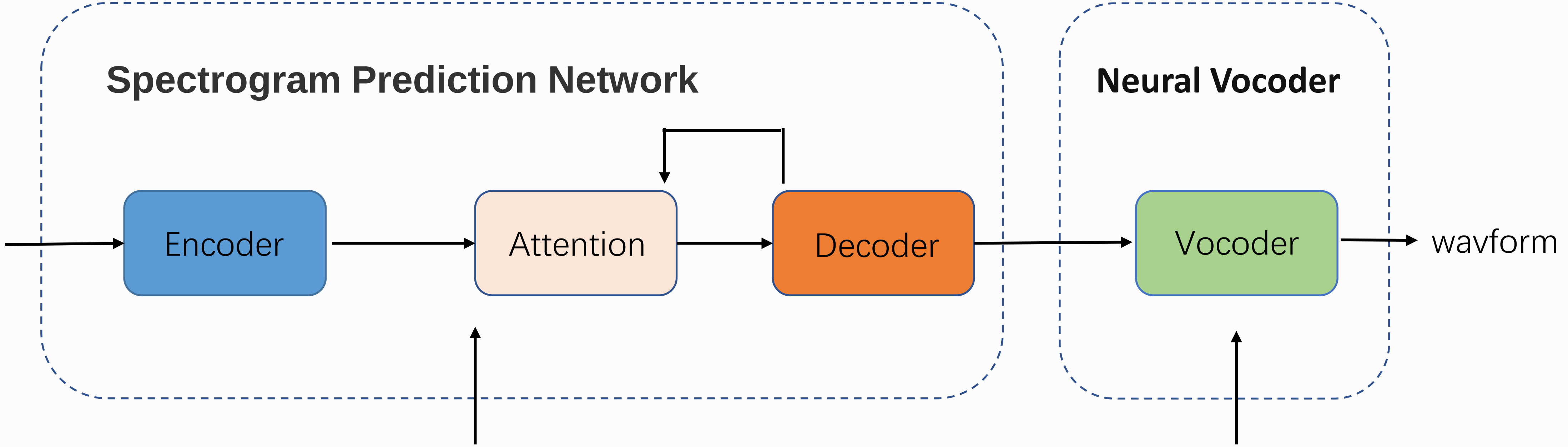
结果分析

Result analysis & auditory feedback
& improvement measurements



03 模型训练

Natural TTS Synthesis by Conditioning WaveNet on Mel-Spectrogram Predictions: train

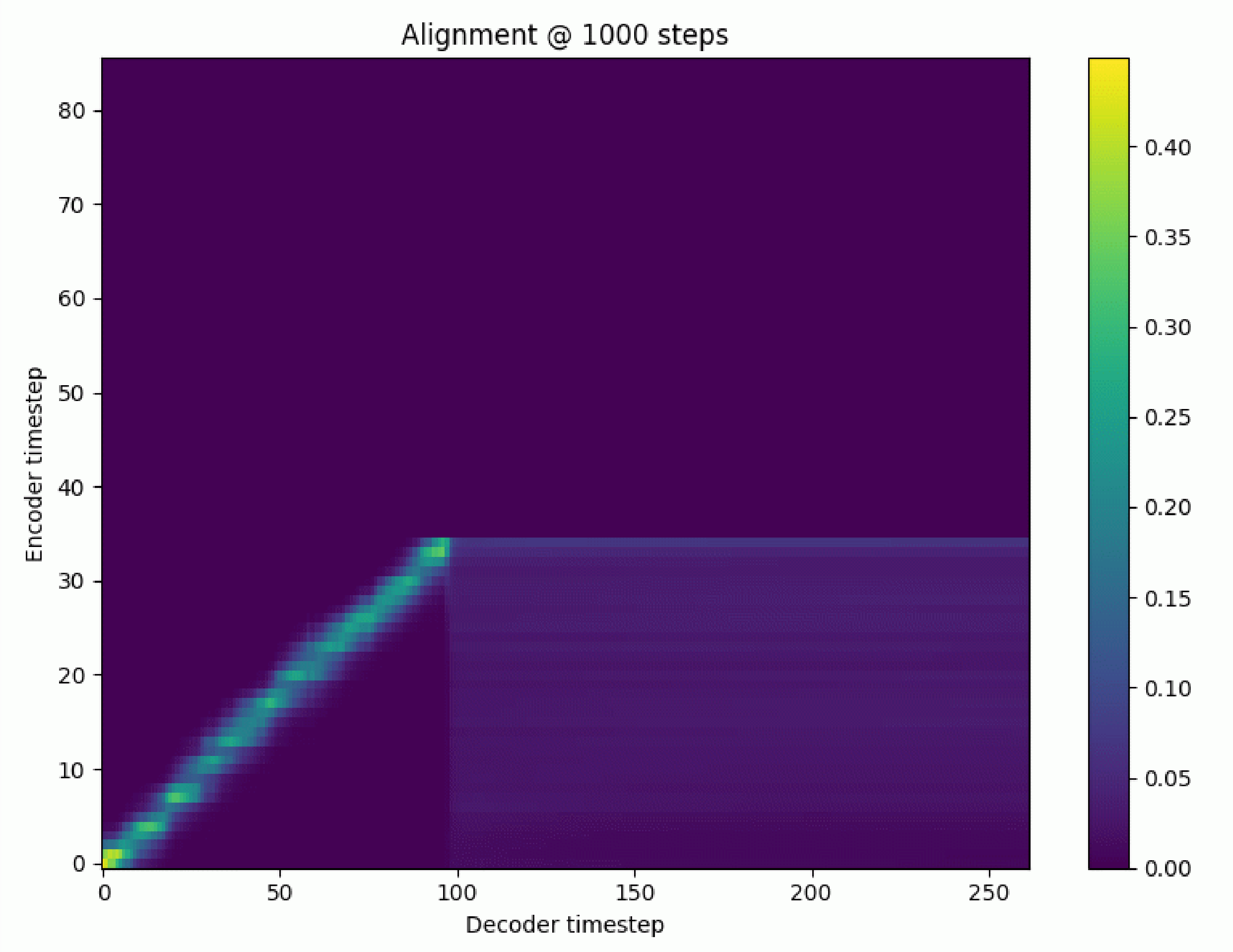


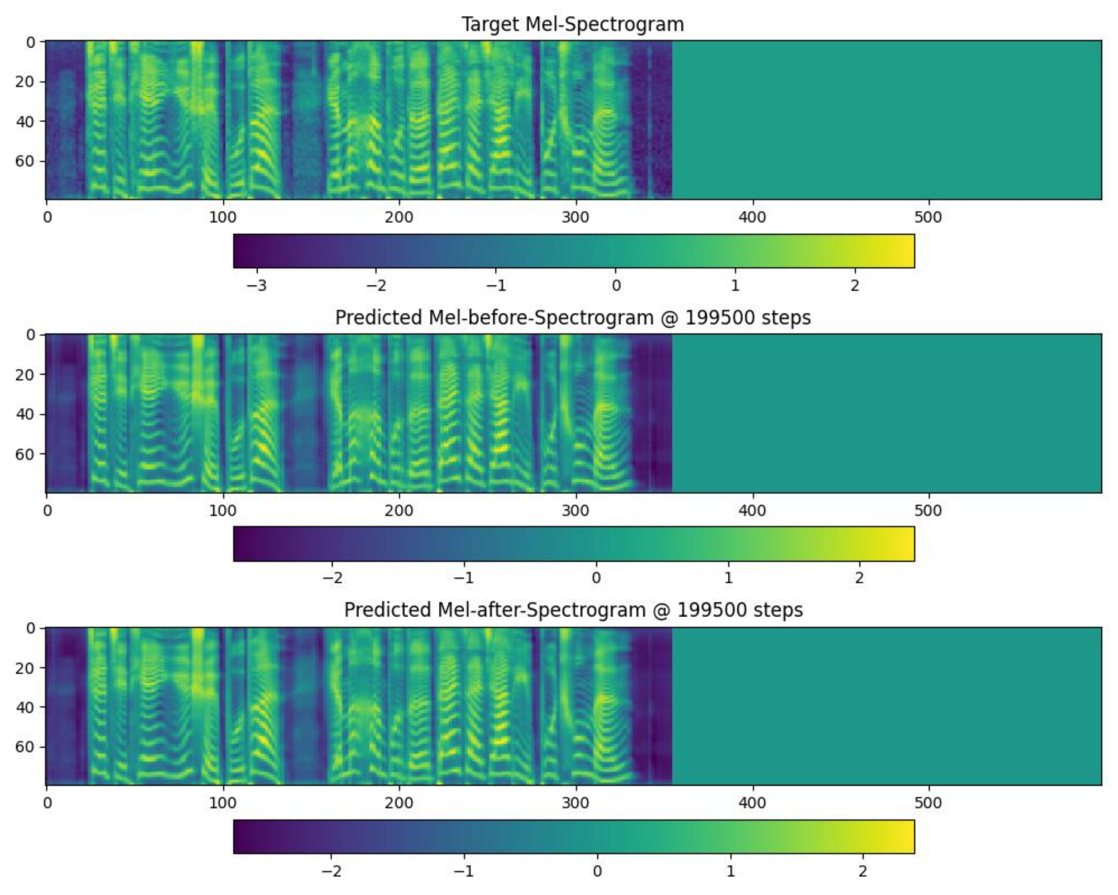
Chinese Standard Mandarin Speech Copus
(10000 Sentences)
尝试自制数据集，如罗翔、trump、bidden

ljspeech

03 结果分析：Tacotron2

Natural TTS Synthesis by Conditioning WaveNet on Mel-Spectrogram Predictions: train





结果分析：SV2TTS

Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis

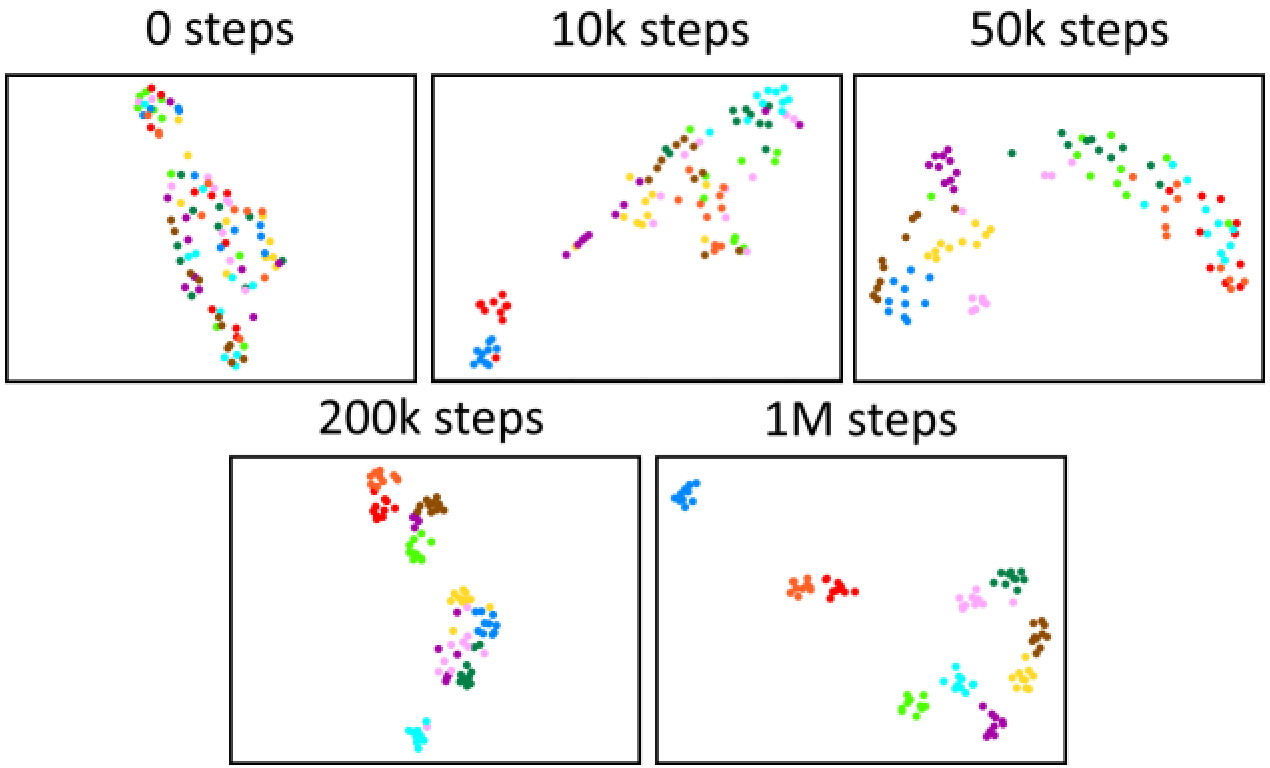


Figure 12: UMAP projections of utterance embeddings from randomly selected batches from the train set at different iterations of our model. Utterances from the same speaker are represented by a dot of the same color. We specifically omit to pass labels to UMAP, so the clustering is entirely done by the model.

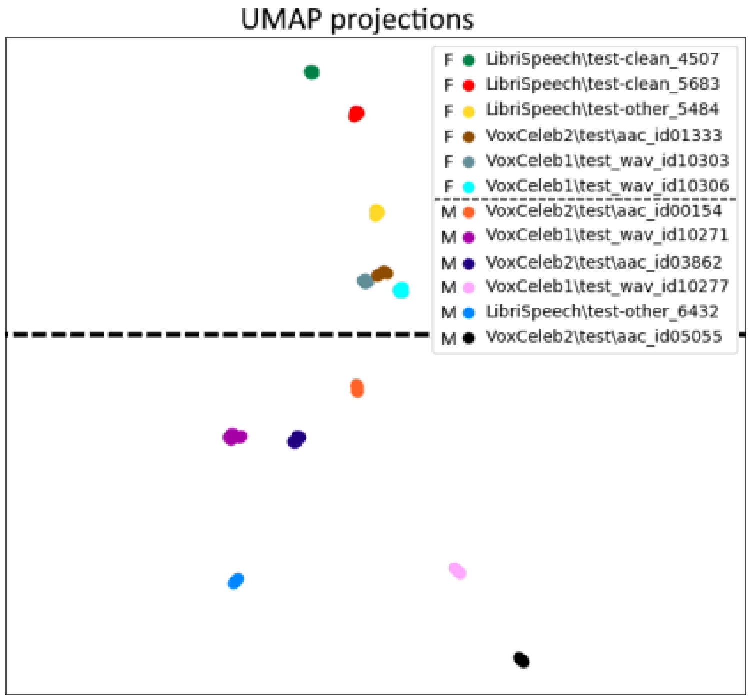
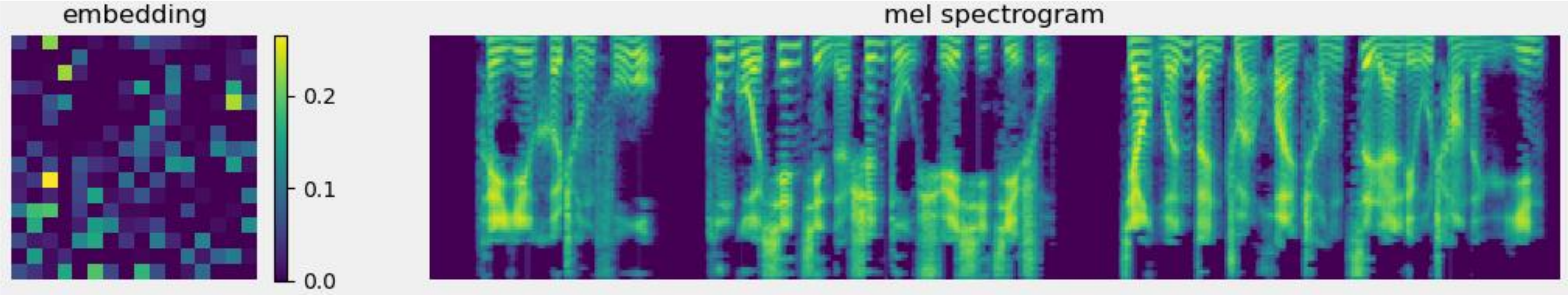
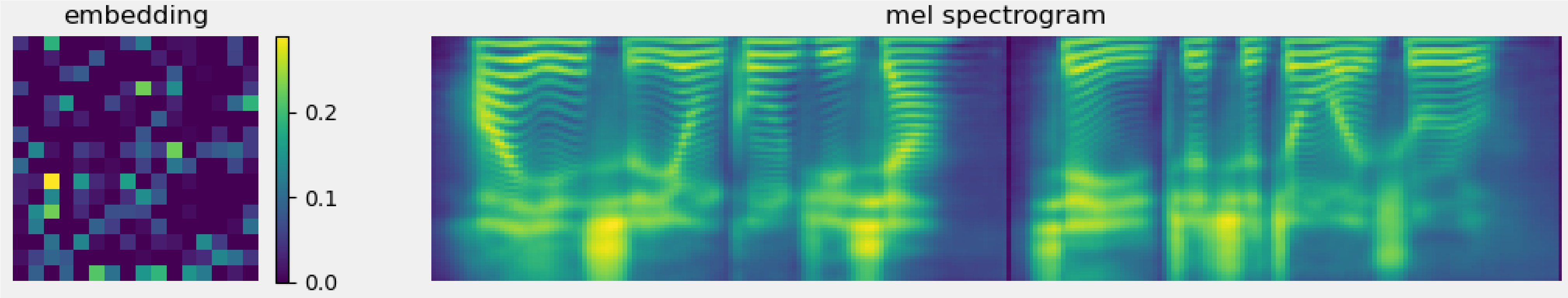


Figure 13: UMAP projections of 120 embeddings, 10 for each of the 12 speakers. Six male and six female speakers are selected at random from the test sets

03 结果分析：SV2TTS

Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis





03 听感反馈 & 改进措施

Result analysis & auditory feedback & improvement measurements

输出语音含有噪声

The generated speech is mixed with noise. which make it sound hoarse, affecting the sense of hearing

模型正则化/防止过拟合

Regularize the model to prevent over-fitting
Improve vocoder

输入语音质量要求高

Input voice must be noise free and loud, otherwise the output is mixed with noise

降噪、语音增强/含噪样本上的训练

noise reduction and Speech enhancement
train the model on more samples with noise

4

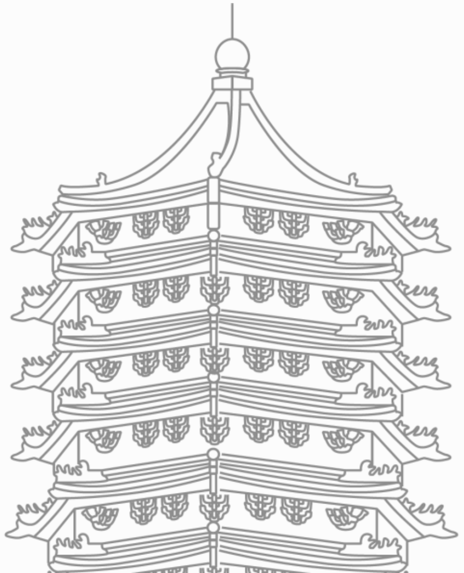
PART 04

系统演示

System demonstration

We will demonstrate to you the functions of our system.

Let’s try it!



05

参考文献

bibliography & reference

[1] https://arxiv.org/abs/1712.05884

[2] https://arxiv.org/abs/1308.0850

[3] https://arxiv.org/abs/1710.10467

[4] https://arxiv.org/abs/1806.04558

[5] https://arxiv.org/abs/1802.08435

[6] https://arxiv.org/abs/1609.03499

[7] https://github.com/CorentinJ/Real-Time-Voice-Cloning

[8] https://github.com/babysor/MockingBird

2022

框 架 结 构　/　结 果 分 析　/　系 统 展 示

感谢您的观看

Thanks for watching! There are still many shortcomings QAQ. We welcome criticism and correction

汇报人：……