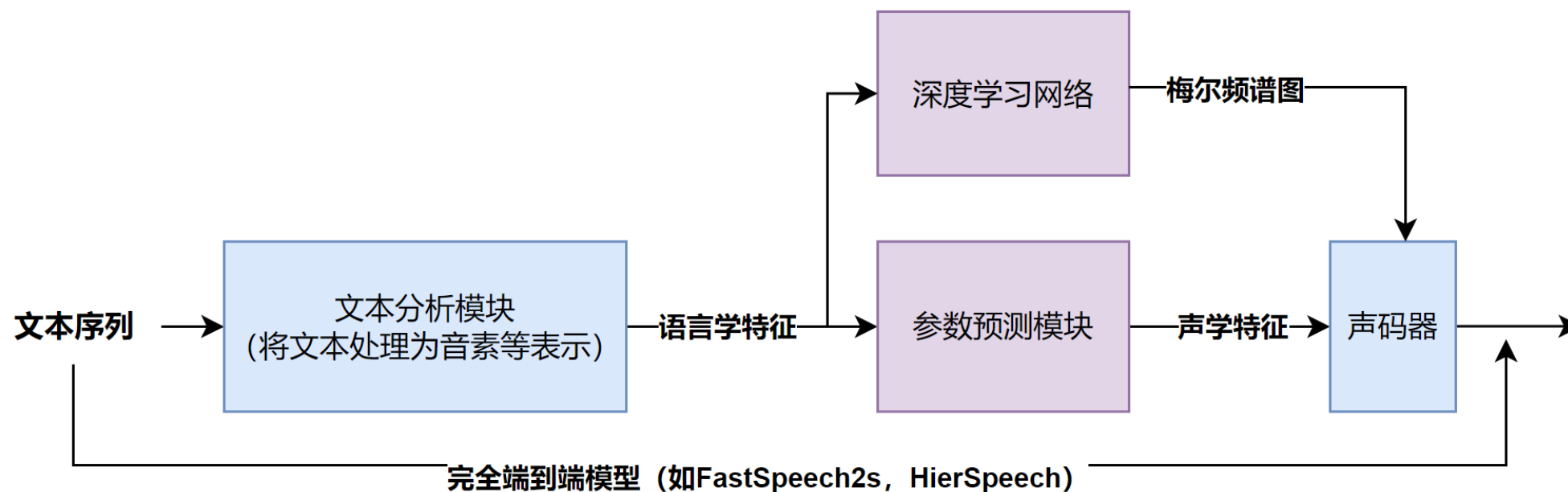


毕业设计 “可控语音合成方法及系统” 答辩

刘文韬

课题背景

语音合成任务的主要目的是将文本转化为语音，使计算机具有发声能力。目前，本领域内的相关工作基本聚焦于基于统计参数的语音合成方法（SPSS）。其基本思想是先生成一些声学参数，然后通过一些算法从中预测语音。



课题目标及研究内容

本课题的目标是研究可控语音合成技术，克服传统方法在细节调控方面的限制，实现对声音音色、音调和情感进行控制的语音合成系统。

音色可控

在多说话人数据集上训练模型，实现基于标签的说话人音色控制能力

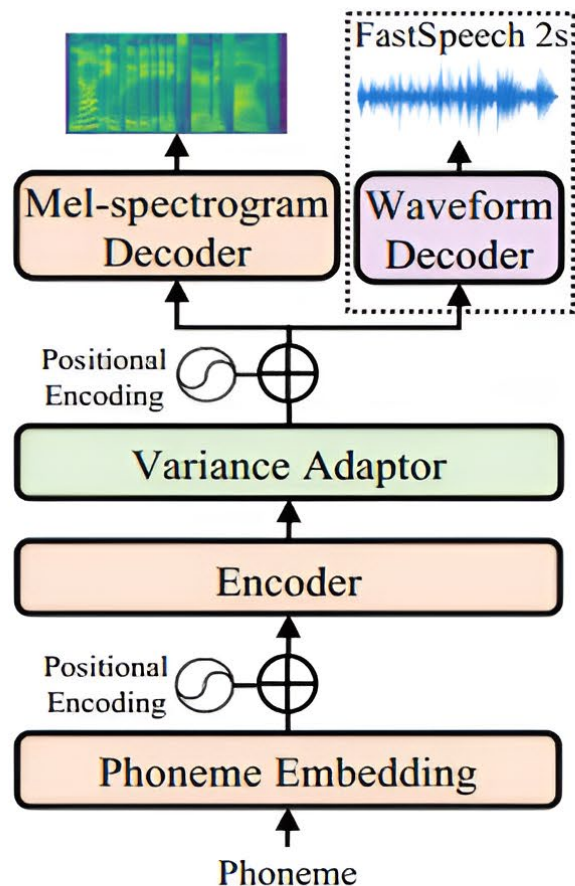
音调可控

通过参数调节，实现对合成语音整体的音高，韵律，以及音强高低的控制

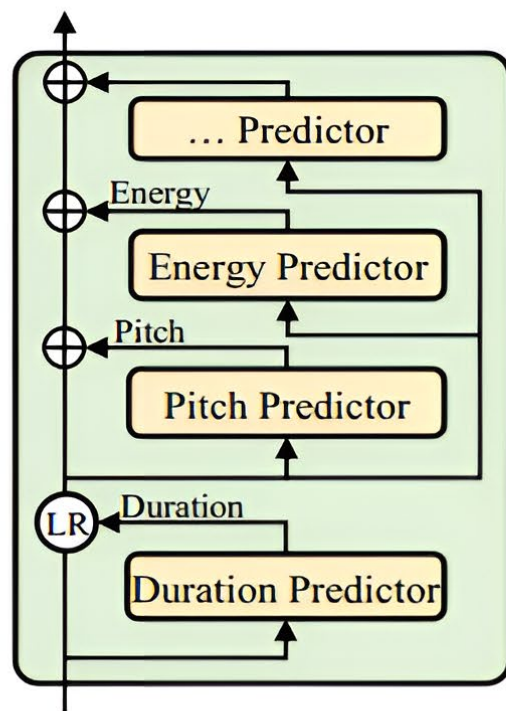
情感可控

在多情感数据集上训练模型，实现基于标签的合成语音情感控制能力

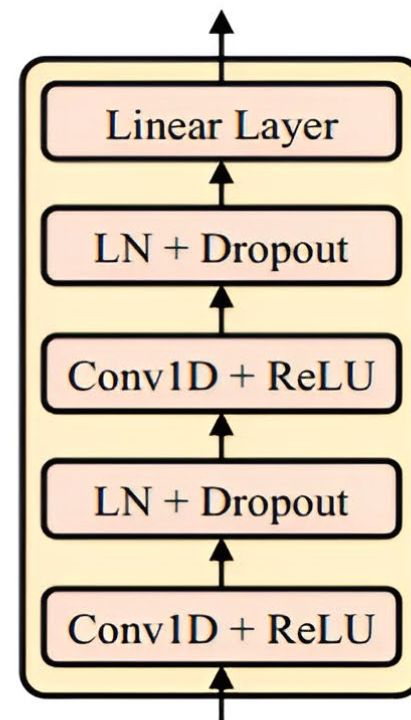
Backbone: FastSpeech2



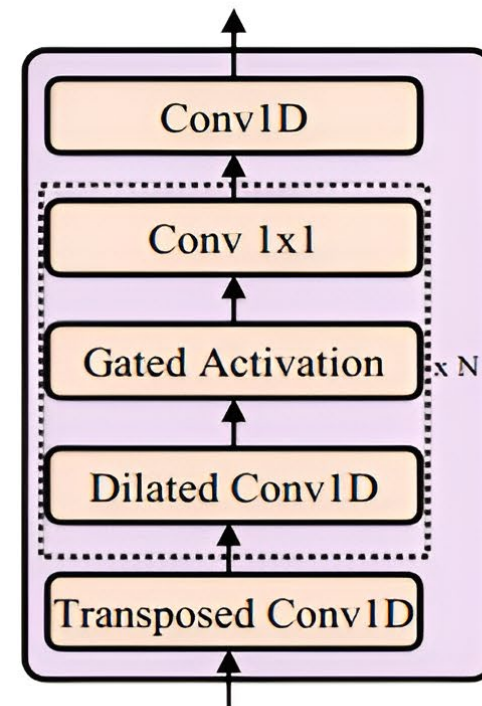
(a) FastSpeech 2



(b) Variance adaptor

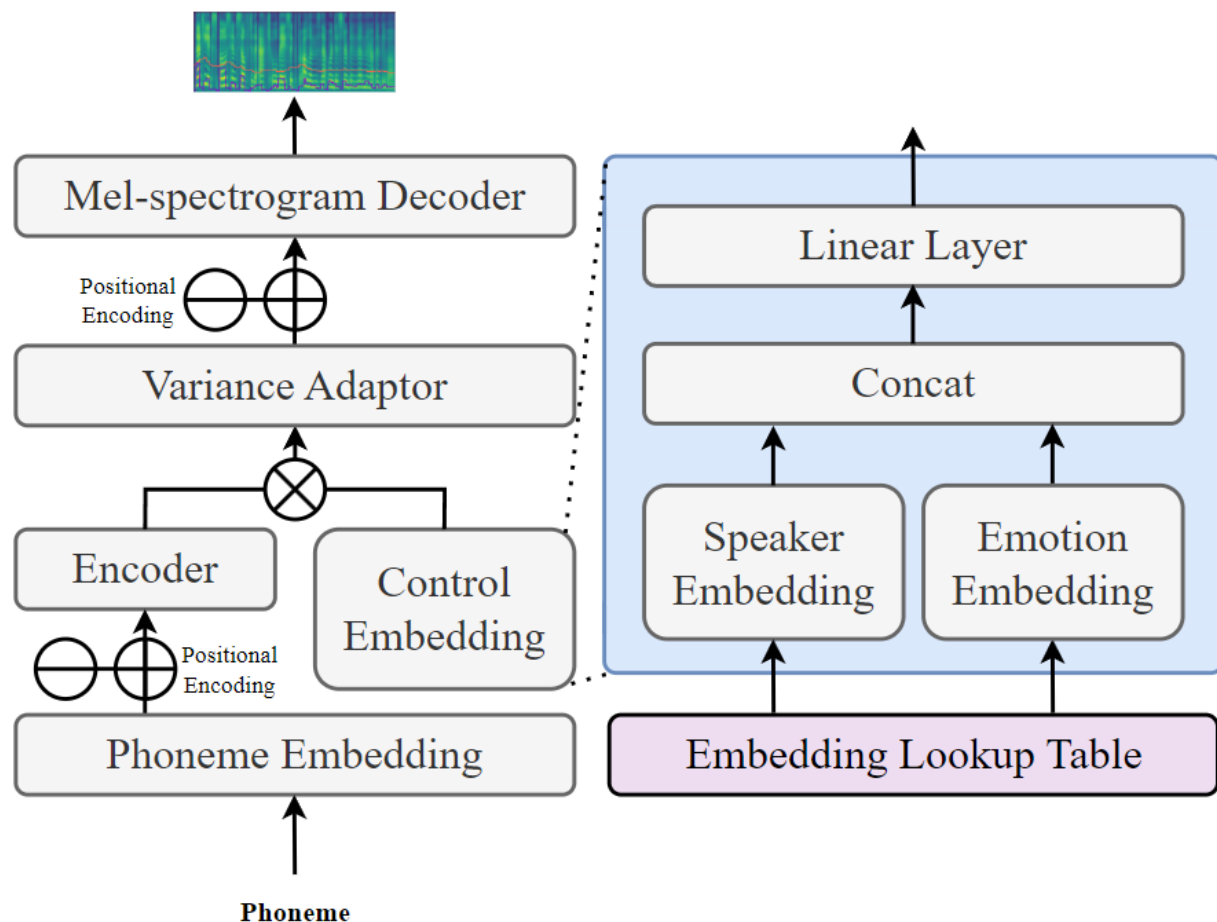


(c)
Duration/pitch/energy
predictor



(d) Waveform decoder

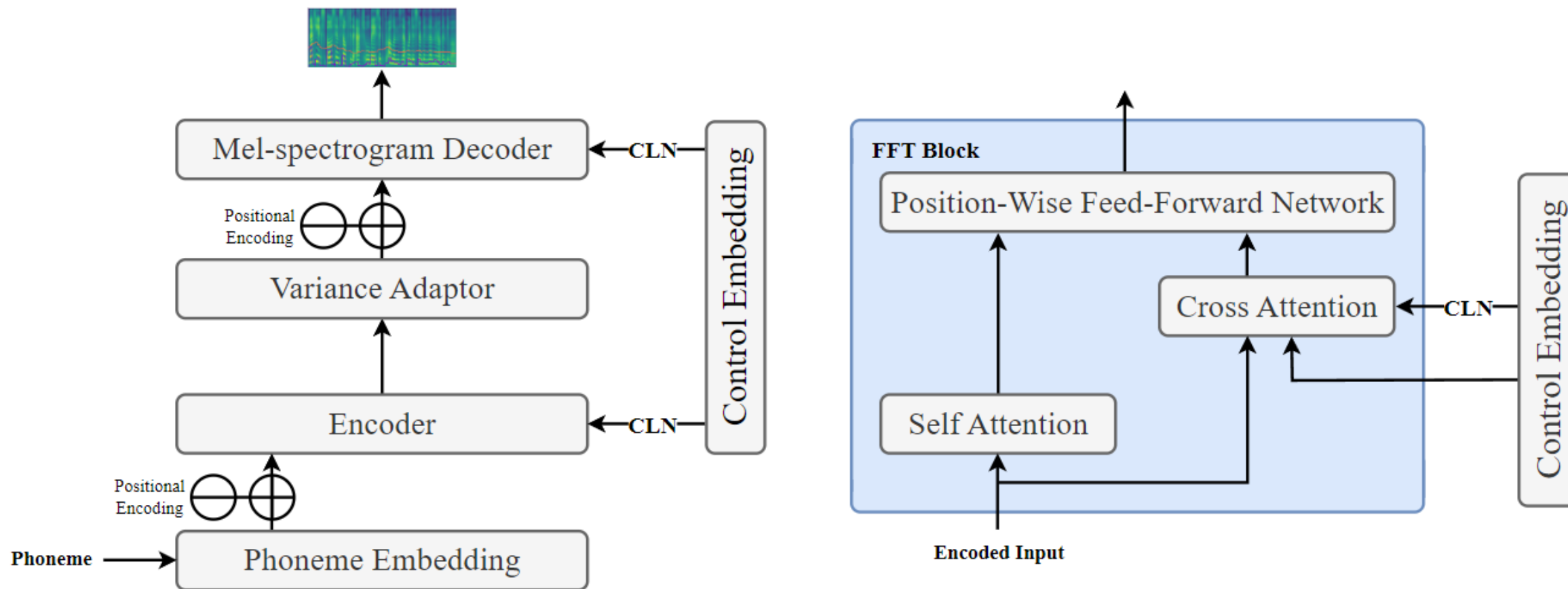
基于多元嵌入的可控语音合成模型 (Baseline Model)



通过将说话人和情感嵌入叠加到隐藏序列中，构建了一个具有音色和情感可控性的基准模型。后称此模型为Baseline模型。

该模型的主要目的是验证控制方法的可行性，并用作后续正式模型的评价基准，测试其他方法下模型的性能

基于条件层归一化的可控语音合成模型 (CLN Model)



条件层归一化

标准的层归一化（LN）旨在减少不同输入分布对模型训练的影响，有助于稳定训练过程并加速模型的收敛：

$$LN(x) = \gamma \left(\frac{x - \mu}{\sigma} \right) + \beta$$

条件层归一化扩展了LN的概念，允许模型根据外部提供的条件信息动态调整归一化参数：

$$CLN(x) = \gamma(c) \left(\frac{x - \mu}{\sigma} \right) + \beta(c)$$

在本课题的模型实现中， γ 以及 β 均为全连接层

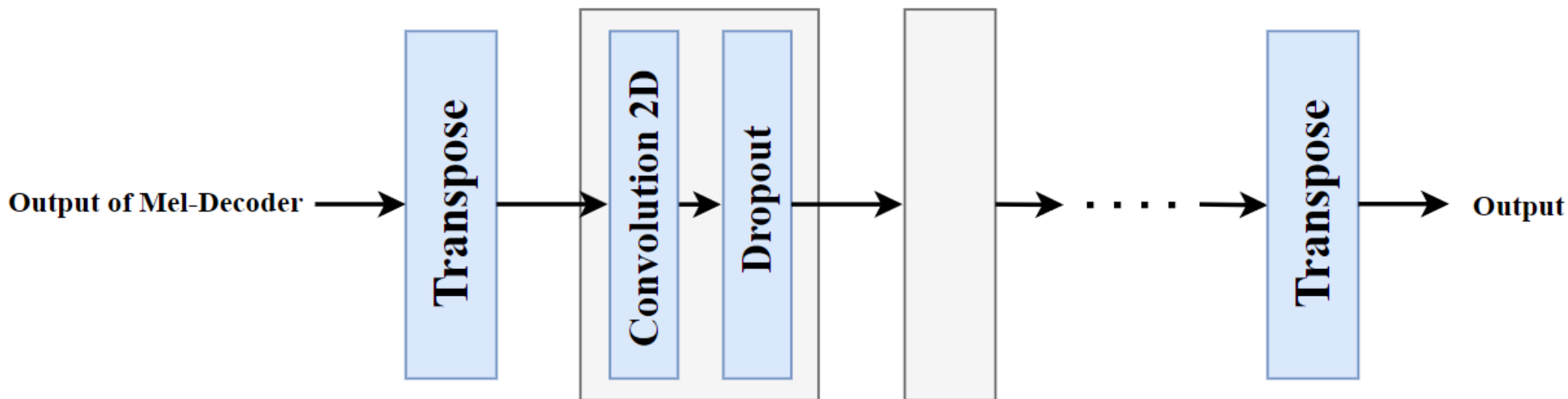
交叉注意力机制

在交叉注意力中，模型计算一个序列（查询序列）中的元素如何关注另一个序列中的元素。交叉注意力层通过在自注意力机制的基础上融合条件嵌入信息，允许模型在处理输入序列时考虑额外的上下文信息。在本课题的模型实现中，将嵌入向量作为k,v。使控制向量对模型输入产生影响。

$$\textit{Output} = \textit{SelfAttention}(\textit{input}, \textit{CLNEmbedding}, \textit{CLNEmbedding})$$

后处理声学网络

语音合成是典型的一对多问题，因此提升模型的泛化性能十分重要。为提升模型语音生成的泛化性能以及生成适合声码器处理的特征表示，在Mel-Decoder 之后加入后处理网络，该后处理网络是堆叠的卷积层序列。



数据集选择

LJSpeech

LJSpeech数据集专为文本到语音系统的开发而设计，由单个女性说话人录制，包含约24小时的朗读语音，涵盖了大约13,100个来自7本真实书籍的句子。

ESD

Emotion Speech Database (ESD) 数据集则是用于音频对话研究的语音数据集，由10名英语母语者和10名中文母语者录制，共包含350条文本音频。每条文本覆盖了5种不同的情感类别：中性、快乐、愤怒、悲伤或惊讶

量化评测指标

Mel Cepstral Distortion: MCD通过比较合成语音和参考（原始）语音在梅尔频率倒谱系数（MFCCs）上的差异来评估合成语音的还原程度。

Similarity: 基于声纹识别网络EACPA-TDNN计算的语音还原相似度指标。使用ECAPA-TDNN模型预训练的说话人识别编码器（Speaker Recognition）将合成音频和原始音频转换为特征向量，并计算这两组向量之间的余弦相似度，以反映合成音频的音色与原始音频的相似程度。

Similarity(Emotion based): 由于EACPA-TDNN只能反映音色的相似度，因此在不同的情感验证集上单独计算相似度，以间接评价模型的情感还原能力。

对比模型

FastSpeech2: 原始 FastSpeech2 模型

Baseline Model: 基于多元嵌入的可控语音合成模型

HierSpeech: 一种基于分层条件自分编码器（VAE）的高质量端到端文本到语音系统。该模型利用自监督语音表示作为额外的语音信息，以弥合文本和语音之间的信息差距，并采用分层条件VAE连接这些信息。在可控性上，HierSpeech 允许语音 Prompt 输入，因此提供了对语音风格（包括音色，语气，情感）的可控性。在模型评价中，使用<https://github.com/sh-lee-prml/HierSpeechpp>中基于LibriTTS数据集的预训练模型 “hierspeechpp_lt960_ckpt”

模型性能评价：参数量

下表为各模型的参数量

Method	Parameters (Inference)
FastSpeech2	34M
Baseline Model	34M
CLN Model	36M
HierSpeech	64M

可知CLN模型在为模型引入较少参数量增加的同时，实现了多音色与多情感语音合成。且所有基于FastSpeech2的模型参数量均远低于HierSpeech 模型，使得这些模型训练算力要求更低，推理效率更高。

模型性能评价：MCD & Similarity

下表为各模型的梅尔频谱失真，音频相似度以及推理时间

Method	MCD	Similarity(EACPA-TDNN)	Inference Speed(s)
FastSpeech2	160.58	0.93	0.3024
Baseline Model	143.69	0.92	0.3040
CLN Model	157.98	0.94	0.2948
HierSpeech	263.99	0.96	1.6162

推理速度：在对比模型中，FastSpeech2，Baseline Model以及CLN模型均在Nvidia GTX1080 8G GPU上进行推理。HierSpeech模型在Nvidia a30 20G GPU 上进行推理。均取验证集上500 条数据推理耗时的平均值作为最终评价结果。

模型性能评价：Similarity(Emotion based)

下表为各模型在不同情感下的音频相似度

Method	Similarity(EACPA-TDNN)				
Emotion	Neutral	Ang	Sad	Hap	Sur
FastSpeech2	0.94	0.95	0.90	0.93	0.92
Baseline Model	0.94	0.93	0.88	0.93	0.90
CLN Model	0.96	0.95	0.90	0.95	0.94
HierSpeech	0.97	0.97	0.94	0.96	0.98

CTTs: Controllable Text To Speech

Controllable Text-To-Speech: FastSpeech2 with speaker and emotion embedding, bring FastSpeech2 to the next level of its controllability while remaining its train/inference efficiency.

Input Text

Clear are your eyes and bright your breath.

Speaker

speaker08

Emotion

Neutral

Angry

Happy

Sad

Surprise

pitch

0.91

duration

0.88

energy

1.1

Checkpoint file

ESD_LJSpeech_10000_eng.pth.tar

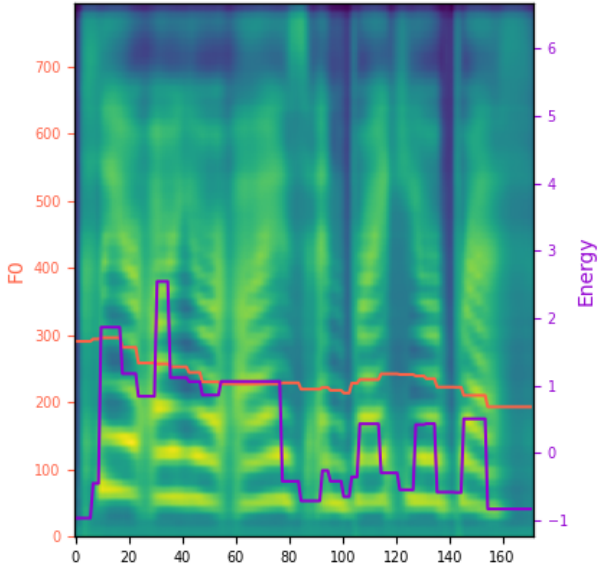
☒ Strict Mode

Clear

Submit

Mel Spectrogram

Synthetized Spectrogram



Model Output Audio

0:00 / 0:01

Inference_time

0.30699849128723145

☰ Examples

	duration	energy	Checkpoint file	Strict Mode	16
Clear are your eyes and bright your breath.	speaker08	Neutral	0.91	0.88	1.1
	ESD_LJSpeech_10000_eng.pth.tar	true			

Demo

 **Evry wird is spelled incorectly but stil reedable.**

 **I did go, and made many prisoners.**

 **Come on my jack in the boxes.**

 **I say I will be emperor.**

 **Both side were softly curved.**

结论

本课题的主要工作：

1. 构建了基于多元嵌入的可控语音合成模型
2. 构建了基于条件层归一化的可控语音合成模型
3. 引入后处理网络，提高模型泛化性能
4. 构建用户友好的交互界面

本课题依然存在的局限性：

1. 支持的最大文本序列长度为1000。且在执行超长文本处理时，在音频的后半段会出现显著的质量下降
2. 算力资源无法支持 CLN 模型在更大的数据集（如 LibriTTS）上预训练

感谢观看



哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN