arXiv:2305.02269v1 [cs.SD] 3 May 2023

M²-CTTS：端到端多尺度多模态对话式文本转语音合成

薛金龙[1], 邓雅月[1], 王凤萍[1],李雅[1,*], 高英明[1],陶建华[2],孙建清[3],梁佳恩[3]

[1]中国北京，北京邮电大学[2]中国科学院自动化研究所模式识别国家重点实验室（NLPR），中国北京[3]云知声智能科技股份有限公司，中国北京

摘要

对话式文本转语音合成（TTS）旨在根据历史对话合成具有恰当回复韵律的语音。然而，全面建模对话仍是一项挑战，而且大多数对话式 TTS 系统只关注提取全局信息，忽略了局部韵律特征，而这些特征包含关键词和强调等重要的细粒度信息。此外，仅考虑文本特征是不够的，声学特征也包含各种韵律信息。因此，我们提出了 M²-CTTS，这是一种端到端多尺度、多模态的对话式文本转语音系统，旨在全面利用历史对话并增强韵律表达。具体而言，我们设计了文本上下文模块和声学上下文模块，分别进行粗粒度和细粒度建模。实验结果表明，融合细粒度上下文信息并额外考虑声学特征的模型，在成对比较平均意见分测试中展现出更好的韵律表现和自然度。

索引词— 语音合成、对话式文本转语音、韵律、多粒度、多模态

1. 引言

过去几年，基于深度学习的文本转语音（TTS）模型发展迅速， [1, 2] 能够生成自然流畅、达到人类水平且具有旁白般高品质的语音。此外，通过在大规模情感语音语料库上训练，人们提出了许多表现性语音合成系统， [3, 4, 5, 6, 7] 能够生成韵律丰富且富有表现力的语音。作为人机交互（HCI）领域中一种更复杂的技术，会话式 TTS 旨在根据历史对话合成韵律恰当的语音。如果将这些表现性语音合成方法直接应用于会话式场景，则无法取得良好效果。

* Ya Li 为本文的通讯作者。

TTS [8], 因为对话式文本转语音高度依赖历史对话，因此对这类对话信息进行建模至关重要。

近期提出了一些对话建模方法。Guo 等人 [9]提出了一种上下文编码器，以强化历史上下文中的话语级信息（文本方面），并使用额外的辅助编码器提取有用的文本统计特征，如语义和句法特征。Cong 等人 [10] 引入了额外的声学上下文编码器，以从先前的语音中获取声学嵌入。此外，Li 等人 [8] 提出了一种基于图的多模态上下文建模方法，用于建模对话中的说话人间和说话人内影响。

然而，这些对话式文本转语音研究仅关注在单一模态的全局层面预测韵律嵌入，忽略了不同方面的局部特征，而这些特征对于反映真实的人类行为 [11] （如关键词强调或韵律变化）至关重要。因此，我们采用一种层次化建模方法，涵盖文本和声学方面的话语级与音素级模块。此外，有些方法 [9, 10]仅使用拼接操作和门控循环单元进行建模，其描述当前话语与过去话语之间关系的能力有限。因此，我们在工作中通过多头跨模态注意力机制增强这种联系。此外，我们进一步应用下一项预测器模块以提升韵律相似度，并使用风格自适应层归一化（SALN） [5] 注入韵律信息。我们在对话语音语料库 DailyTalk [12] 上进行了实验，结果表明，我们的方法在成对比较平均意见分测试方面优于其他方法。尤其是，细粒度表现力得到了显著增强。

我们的主要贡献如下：（1）我们不仅提取全局特征，还利用局部信息来增强表现力；（2）我们开发了一种同时利用文本和声学对话信息的方法；（3）我们通过应用 SALN 和韵律预测模块，更好地利用韵律嵌入。

本文其余部分的组织结构如下。第 2 节

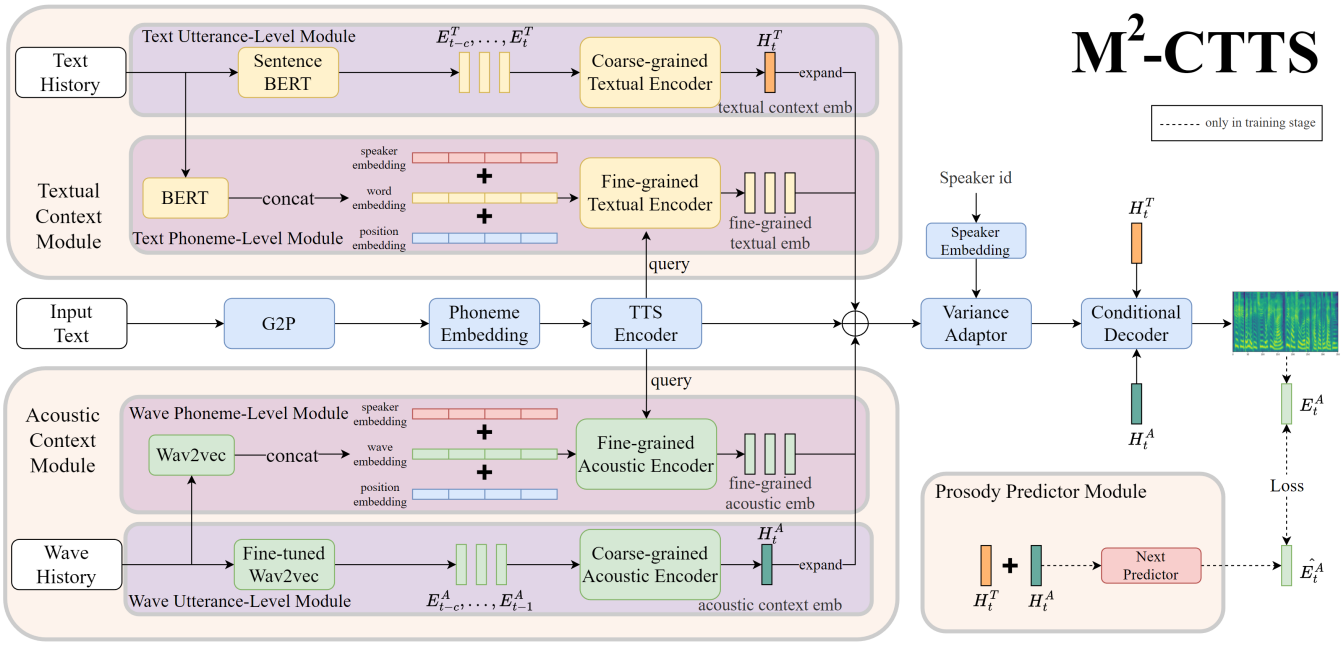


图 1。我们提出的 M²-CTTS 模型概览。

详细介绍我们提出的对话式文本转语音系统。第 3 节展示我们的实验结果与讨论。本文在第 4 节总结。

2. M²-CTTS 模型

M²-CTTS 的整体模型架构如图 1 所示。M²-CTTS 基于 FastSpeech2 [2]，由编码器、解码器和方差适配器组成。为了从历史对话中捕获文本和声学信息，我们设计了文本上下文模块（TCM）和声学上下文模块（ACM）。此外，为了同时考虑分层历史上下文，我们在 TCM 中设计了文本话语级模块（TUM）和文本音素级模块（TPM），并在 ACM 中设计了波形话语级模块（WUM）和波形音素级模块（WPM）。此外，我们采用条件解码器和韵律预测模块（PPM）来增强生成语音的表现力。具体而言，我们在解码器中采用风格自适应层归一化（SALN） [5] 作为条件解码器，以增强表现力并更好地利用历史上下文。

2.1. 多尺度对话建模

虽然情绪或意图等全局信息有助于模型理解对话，但关键词或强调等包含强烈或隐含含义的细节信息也应得到凸显。此外，话语内部风格不一致或混杂可能导致韵律不自然。我们还通过实验证实，固定长度的对话嵌入效果不佳，并导致平均韵律。因此，我们进一步结合粗粒度与细粒度的对话建模，并采用多头注意力来获取全面的历史信息。我们将 t 时间戳处的对话历史表示为：

$$Conversation = \{A_{t-c}, B_{t-c+1}, \ldots, A_{t-1}, B_t\} \quad (1)$$

其中，c 是记忆容量参数，用于确定说话者 A 和 B 之间会考虑多少轮对话。

粗粒度上下文建模。在粗粒度上下文建模方面，我们利用文本话语级模块（TUM）和波形话语级模块（WUM）来获取全局上下文信息。

对于 TUM，我们遵循并修改了 [9]提出的模块，以建模文本话语级信息。该模块由基于 Sentence BERT [13] 的文本话语级特征提取器和粗粒度文本编码器组成。由于 Sentence BERT 能够捕捉语义信息，我们采用它从 $Conversation$ 中提取包含最近 c 轮对话的历史句子嵌入 $E^T_{t-c:t-1}$，以及当前话语嵌入 E^T_t。对于粗粒度文本编码器，我们首先使用一个 GRU 层对历史嵌入 $E^T_{t-c:t-1}$ 进行编码，以考虑累积的过往对话信息。然后，我们将 GRU 的最终隐状态与当前话语嵌入 E^T_t 拼接，并将其输入带有额外注意力机制的线性层，以计算加权的全局文本上下文嵌入 H^T_t，如图 1 所示。

For WUM，我们采用了类似的架构来扩展。 ract the

全局声学上下文嵌入 H_t^A，但特征提取器除外，而且我们不使用当前话语的语音B_t。我们没有采用使用基于频谱的特征和参考编码器或全局风格令牌（GST） [3], 的传统方法，而是采用在下游语音情感识别（SER）任务中微调的 Wav2vec 2.0 [14] ，直接从原始音频中提取声学话语级韵律信息 E_i^A。这种方法有两方面的优势。与缺少相位信息、所含韵律数据少于原始音频的基于频谱的特征相比，预训练模型能够捕捉更全面的声学特征。此外，使用预训练模型可以降低复杂模型的训练难度，并获得稳健的性能。

细粒度上下文建模。 虽然粗粒度上下文建模模块能够捕捉整段话语或话语之间联系中的信息，但让每个音素都采用相同的韵律嵌入并不自然，这会导致生成语音的韵律平淡。考虑现实生活中的对话，人们会对交流中他人说的特定词语或短语作出反应。因此，为了模拟这些情形，我们开发了文本音素级模块（TPM）和波形音素级模块（WPM），用于建模细粒度的历史信息。

在 TPM 和 WPM 中，分别从 BERT [15] 和 Wav2vec 2.0 [14] 获取第 i 个句子或话语的隐藏特征序列 P_i 。随后，利用多头跨模态注意力机制，识别对上下文理解和表现力有重要影响的关键短语或帧。下面将作更详细的介绍。

2.2. 多模态上下文建模

除了语义和句法特征外，副语言信息（如声学韵律线索）在上下文理解中也发挥着重要作用，因为即使在相同内容或情境下，人们也会采用不同的语气。因此，对这两种模态进行建模意义重大。

文本上下文建模。如图 1 所示，文本上下文模块包含 TPM 和 TUM。TPM 由特征提取器和细粒度文本编码器组成。我们采用 BERT [15] 获取第 i 个句子的隐藏特征序列 P_i，并将最近 c 轮对话聚合为一个长序列。随后，我们添加说话人标识以保留说话人身份，并使用正弦位置编码表示对话顺序。对于细粒度文本编码器，首先采用一维卷积层进行上下文化处理。然后，将 TTS编码器的输出作为多头交叉注意力模块的查询，并将历史细粒度表示序列同时作为键和值。得到的加权上下文表示将通过残差设计加回 TTS编码器的输出。TUM 的结构如第 2.1 节所示。因此，重要的文本方面的关键词或短语会被纳入考虑。

声学上下文建模。如图 1 所示，声学上下文模块包括 WPM 和 WUM。WPM 由特征提取器和细粒度声学编码器组成。不同于使用梅尔频谱图作为声学输入，我们采用 Wave2vec 2.0 [14] 从原始音频中提取隐藏特征序列。其余处理与上文所述的 TPM 类似。因此，该模块能够关注声学层面的局部表达或强调。

2.3. 韵律约束

与 [10], 类似，我们还使用韵律预测模块（PPM），根据多模态全局上下文嵌入H_t^T和 H_t^A预测当前话语 E_t^A 的韵律嵌入。如公式 2 所示，我们采用额外的 MSE 损失，约束粗粒度上下文模块根据历史信息预测当前话语的韵律表达。

$$\hat{E}_t^A = PPM(H_t^T, H_t^A),\ Loss = ||\hat{E}_t^A, E_t^A|| \qquad (2)$$

其中，$\hat{}E_t^A$表示预测的韵律嵌入。 E_t^A是通过微调后的 Wav2vec 模型提取的真实韵律嵌入。需要注意的是，韵律预测器模块和损失计算仅在训练期间使用。

3. 实验

3.1. 实验设置

我们使用公开的英语语料库 DailyTalk [16] 进行实验。DailyTalk 数据集源自 DailyDialog 数据集 [16], 包含由一名男性和一名女性说话者演绎的 2,541 段对话。DailyTalk 中的每段对话都超过五轮。为方便使用，我们按对话轮次将其切分为 23,773 个音频片段，总时长为 20 小时。我们留出 128 段对话作为验证集，其余作为训练集。在实验中，所有话语均下采样至 22050 Hz，并用于提取 80 维梅尔频谱图。

我们使用 FastSpeech2 作为 TTS 主干模型，并采用无监督时长建模 [17] ，而非使用带有外部对齐器的有监督时长模型，例如 Montreal Forced Aligner [18], , 因为外部对齐器存在分布外问题的风险，而软对齐在表现力方面更灵活。我们使用预训练的 HiFi-GAN [19] 模型作为声码器，将 80 维梅尔频谱图转换为 22050 Hz 音频文件。我们的对话式文本转语音模型均在 GeForce RTX 3090 上以 16 的批量大小训练 400K 步。优化器等训练超参数沿用原始配置。

对于文本特征提取器，我们在 TUM 中使用预训练的 Sentence BERT [1] ，在 TPM 中使用原始 BERT。对于

[1]https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

表1。对比方法及 MOS 评分。

方法	TUM	WUM	TPM	WPM	MOS 分数
M0	根据真实值重建				4.38 ± 0.06
M1					3.64 ± 0.07
M2	✓				3.71 ± 0.06
M3		✓			3.65 ± 0.07
M4	✓		✓		3.64 ± 0.07
M5		✓		✓	3.70 ± 0.07
M6	✓	✓			3.73 ± 0.06
M7	✓	✓	✓	✓	3.74 ± 0.07

对于声学特征提取器，我们在 WUM 中采用经过 IEMOCAP [20] 情感训练数据微调的预训练 Wav2vec 2.0[2] ，并在 WPM 中使用原始 Wav2vec 2.0。

为研究我们提出的方法的性能，我们采用 M0 至 M7 作为对比方法进行评估。详细组合见表 1。M1 是 FastSpeech2 基线模型，而 M2 至 M7 均采用条件解码器和韵律预测模块来增强表现力。

3.2. 评估

我们在测试集上进行对比平均意见分（CMOS）测试，以评估每组设计配对（A，B）的性能。我们随机选取第 i 轮的 20 条合成语句 U_i，并提供 4 个包含文本和语音的真实历史上下文 $U_{i-4:i-1}$ 作为参考，以比较自然度和表现力。具体而言，参与者需要判断韵律是否更具表现力，以及是否更符合历史对话，并根据标准以 1 分为间隔，在 -3（A 模型好得多）到 3（B 模型好得多）的范围内评分。绝对值表示程度。

CMOS结果如表2所示。我们发现，分别从文本方面或声学方面采用TUM或WUM以话语级建模对话历史，都能显著提升表现力。我们还发现，在声学方面结合细粒度特征，对于合成韵律丰富的语音有显著益处。然而，我们在文本方面并未发现类似的结论。我们认为，局部声学特征包含更多隐式信息，因为人们在不同情境下说同一个词或短语时，韵律和情感会有所不同。“M4与M5”的比较考察了来自不同方面的多尺度信息，也证实声学特征更为有效。此外，我们还研究了同时采用文本和声学模态的表现。“M2与M6”的CMOS分数表明，在话语级结合这两个方面只能略微提升性能。随后，我们将注意力可视化，发现TUM和WUM在大多数情况下都关注相同的话语，尤其是最新轮次。因此

[2]https://huggingface.co/speechbrain/emotion-recognition-wav2vec2-IEMOCAP

表 2。成对比较平均意见分测试的结果。CMOS描述 A 和 B 之间的偏好程度（记作 A vs. B）。偏好根据 CMOS分数计算。

		偏好率（%）		
	CMOS	Left	中性	正确
M1 与 M2	0.25	18.9	46.1	35.0
M1 与 M3	0.19	14.4	53.9	31.7
M2 与 M4	0.03	25.6	45.5	28.9
M3 与 M5	0.21	17.2	46.1	36.7
M4 与 M5	0.19	18.9	45.5	35.6
M2 与 M6	0.05	25.6	43.8	30.6
M2 与 M7	0.34	17.8	38.3	43.9
M6 与 M7	0.32	16.7	41.1	42.2

，提取到的信息相似度很高。测试案例“M2与M7”以及“M6与M7”表明，我们的方法优于早期研究中分别提出的其他方法 [9, 10]。M2-CTTS对历史对话的理解更好，并显著增强了细粒度韵律表达。

我们还进行了五级平均意见分（MOS）测试，以评估语音的质量和自然度。分数越高，质量越好。表1中的结果表明，从不同方面对韵律信息进行建模，可以缓解平均韵律生成问题。我们还注意到，尽管我们的方法包含许多模块，但得益于预训练模型提供的先验知识，这并不一定会损害语音的质量和自然度。合成语音示例可在项目页面 [3]中查看。

4. 结论

本文提出了一种端到端、多尺度、多模态的对话式语音合成系统（M^2-CTTS），能够对从声学和文本特征中提取的多粒度上下文信息进行建模。实验结果表明，同时考虑声学和文本模态可以提升合成语音的韵律和自然度。此外，将粗粒度特征与细粒度特征相结合，能够进一步提升表现力，尤其是从声学特征中提取的细粒度韵律。未来，我们将研究具有可控性的对话式文本转语音细粒度建模。

5. 致谢

本研究得到国家自然科学基金（NSFC）（项目编号：62271083）和模式识别国家重点实验室开放课题计划（NLPR）（项目编号：202200042）的资助。

[3]音频样例：https://happylittlecat2333.github.io/icass

6. 参考文献

[1] RJ Skerry-Ryan、Eric Battenberg、Ying Xiao、Yuxuan Wang、Daisy Stanton、Joel Shor、Ron Weiss、Rob Clark 和 RifASaurous，《利用 Tacotron 实现表现性语音合成的端到端韵律迁移》，载于国际机器学习会议。PMLR，2018 年，第 4693–4702 页。

[2] Yi Ren、Chenxu Hu、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 和 Tie-Yan Liu，《FastSpeech 2：快速、高质量的端到端文本转语音》，*arXiv* 预印本 *arXiv:2006.04558*，2020 年。

[3] Yuxuan Wang、Daisy Stanton、Yu Zhang、RJ-Skerry Ryan、Eric Battenberg、Joel Shor、Ying Xiao、Ye Jia、Fei Ren 和 Rif A Saurous，《风格令牌：端到端语音合成中的无监督风格建模、控制与迁移》，载于国际机器学习会议。PMLR，2018年，第5180–5189页。[4]Younggun Lee 和 Taesu Kim，《端到端语音合成的稳健细粒度韵律控制》，载于 *ICASSP 2019-2019 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2019年，第5911–5915页。[5]Dongchan Min、Dong Bok Lee、Eunho Yang 和 Sung Ju Hwang，《Meta-StyleSpeech：多说话人自适应文本转语音生成》，载于国际机器学习会议。PMLR，2021年，第7748–7759页。[6] Pengfei Wu、Junjie Pan、Chenchang Xu、Junhui Zhang、Lin Wu、Xiang Yin 和 Zejun Ma，《基于说话人条件层归一化和半监督训练的文本转语音跨说话人情感迁移》，*arXiv* 预印本 *arXiv:2110.04153*，2021年。[7] Yi Lei、Shan Yang、Xinsheng Wang 和 Lei Xie，《Msemotts：情感语音合成的多尺度情感迁移、预测与控制》，*IEEE ACM* 音频、语音与语言处理汇刊，第30卷，第853–864页，2022年。[8] Jingbei Li、Yi Meng、Chenyi Li、Zhiyong Wu、Helen Meng、Chao Weng 和 Dan Su，《通过基于图的多模态上下文建模增强对话式文本转语音合成中的说话风格》，载于*ICASSP*。2022年，第7917–7921页，IEEE。[9] Haohan Guo、Shaofei Zhang、Frank K. Soong、Lei He 和 Lei Xie，《面向语音代理的端到端对话式文本转语音》，载于*IEEE* 口语技术研讨会，*SLT 2021*，中国深圳，*2021*年*1*月*19*日至*22*日。2021年，第403–409页，IEEE。

[10] Jian Cong、Shan Yang、Na Hu、Guangzhi Li、Lei Xie 和 Dan Su，《可控的上下文感知对话语音合成》，载于*Interspeech*。2021年，第4658–4662页，ISCA。[11] Xin Xie、Andr´es Bux´o-Lugo 和 Chigusa Kurumada，《通过结构化的语调语音韵律变异对意义进行编码与解码》，*Cognition*，第211卷，第104619页，2021年。[12] Keon Lee、Kyumin Park 和 Daeyoung Kim，《Dailytalk：用于对话式文本转语音的口语对话数据集》，*arXiv* 预印本 *arXiv:2207.01063*，2022年。[13] Nils Reimers 和 Iryna Gurevych，《Sentence-BERT：使用孪生 BERT 网络生成句子嵌入》，*arXiv* 预印本 *arXiv:1908.10084*，2019年。[14] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli，《wav2vec 2.0：语音表征自监督学习框架》，神经信息处理系统进展，第33卷，第12449–12460页，2020年。[15]Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova，《BERT：用于语言理解的深度双向 Transformer 预训练》，*arXiv* 预印本 *arXiv:1810.04805*，2018年。[16] Yanran Li、Hui Su、Xiaoyu Shen、Wenjie Li、Ziqiang Cao 和 Shuzi Niu，《Dailydialog：人工标注的多轮对话数据集》，*arXiv* 预印本 *arXiv:1710.03957*，2017年。[17] Rohan Badlani、Adrian La´ncucki、Kevin J Shih、Rafael Valle、Wei Ping 和 Bryan Catanzaro，《一种通用的 TTS 对齐方法》，载于*ICASSP 2022-2022 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2022年，第6092–6096页。[18] Michael McAuliffe、Michaela Socolof、Sarah Mihuc、Michael Wagner 和 Morgan Sonderegger，《蒙特利尔强制对齐器：使用 Kaldi 训练文本与语音对齐》，载于*Interspeech*，2017年，第2017卷，第498–502页。[19] Jungil Kong、Jaehyeon Kim 和 Jaekyoung Bae，《HiFi-GAN：用于高效、高保真语音合成的生成对抗网络》，神经信息处理系统进展，第33卷，第17022–17033页，2020年。[20]Carlos Busso、Murtaza Bulut、Chi-Chun Lee、Abe Kazemzadeh、Emily Mower、Samuel Kim、Jeannette N Chang、Sungbok Lee 和 Shrikanth S Narayanan，《IEMOCAP：交互式情感双人动作捕捉数据库》，语言资源与评估，第42卷，第4期，第335–359页，2008年。