Computer Science and Application 计算机科学与应用, 2025, 15(11), 85-93
Published Online November 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.1511286

Hans汉斯

BVTS：基于BERT增强的高质量语音合成模型

尹鹏飞[1]，刘雪晴[2]

[1]北京印刷学院信息工程学院，北京
[2]广东外语外贸大学马克思主义学院，广东 广州

收稿日期：2025年10月2日；录用日期：2025年10月31日；发布日期：2025年11月10日

摘　要

近年来，语音合成(Text-to-Speech, TTS)技术在端到端建模、音质优化等方面取得显著进展，合成语音的清晰度与流畅度大幅提升，但在逼近人类真实语音质感方面仍存挑战，主要瓶颈在韵律建模、语义理解适配方面欠缺。本文提出一种基于BERT (Bidirectional Encoder Representations from Transformers)模型增强的语音合成框架——BVTS (BERT-Integrated-VITS2)，模型以VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech)为框架，引入多模态文本编码器，在BERT特征嵌入的引导下，通过特征级融合语音学及韵律特征，同时采用双向可逆流模型与随机时长预测器，实现对语音节奏与语速的细粒度控制。在LJ Speech数据集与自制游戏数据集上的实验结果表明，相较于当前主流模型，BVTS的平均意见得分(MOS)整体提升明显，且字符错误率(CER)更低，此模型明显提升了合成语音的表现力、自然度与可懂度。

关键词

语音合成，BERT模型，VITS

BVTS: A High-Quality Speech Synthesis Model Enhanced by BERT

Pengfei Yin[1], Xueqing Liu[2]

[1]School of Information Engineering, Beijing Institute of Graphic Communication, Beijing
[2]School of Marxism, Guangdong University of Foreign Studies, Guangzhou Guangdong

Received: October 2, 2025; accepted: October 31, 2025; published: November 10, 2025

Abstract

In recent years, Text-to-Speech (TTS) technology has achieved significant progress in end-to-end modeling and sound quality optimization, with the clarity and fluency of synthesized speech improved

文章引用: 尹鹏飞, 刘雪晴. BVTS: 基于 BERT 增强的高质量语音合成模型[J]. 计算机科学与应用, 2025, 15(11): 85-93.
DOI: 10.12677/csa.2025.1511286

substantially. However, challenges remain in approaching the texture of human real speech, and the main bottlenecks lie in the insufficient prosody modeling and semantic understanding adaptation. This paper proposes a BERT (Bidirectional Encoder Representations from Transformers)-enhanced speech synthesis framework named BVTS (BERT-Integrated-VITS2). Based on the VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) framework, the model introduces a multimodal text encoder. Guided by BERT feature embedding, it fuses phonetic and prosodic features at the feature level. Meanwhile, it adopts a bidirectional reversible flow model and a random duration predictor to achieve fine-grained control over speech rhythm and speed. Experimental results on the LJ Speech dataset and the self-constructed game dataset show that compared with current mainstream models, BVTS achieves a significant overall improvement in Mean Opinion Score (MOS) and a lower Character Error Rate (CER). This model significantly enhances the expressiveness, naturalness and intelligibility of synthesized speech.

Keywords

Speech Synthesis, BERT Model, VITS

1. 引言

近年来，语音合成技术已成为支撑众多实际应用的基础技术，广泛应用于人机交互、虚拟助手、有声读物、智能设备及辅助功能工具等领域。传统的基于规则与参数的语音合成方法[1]，例如基于隐马尔可夫模型(HMM) [2]的合成技术和基音同步叠加(PSOLA) [3]算法，虽为实现可控的语音生成奠定了基础，却存在韵律机械生硬、表现力匮乏的问题，难以满足实际应用中对自然语音的需求。

随着深度学习技术的兴起，Tacotron [4]和 Parallel Tacotron 2 [5]等神经网络架构应运而生，它们引入端到端的框架设计，大幅提升了合成语音的流畅度与自然度。这类模型借助注意力机制实现文本与声学表征之间的对齐，但在对齐稳定性与韵律控制能力方面仍面临挑战，制约了合成语音质量的进一步提升。

近年来，语音合成(TTS)模型的研究取得诸多新进展，部分上述问题得到一定解决。VITS [6]模型及其改进版本 VITS2 [7]融合变分自编码器(VAE)、对抗训练与归一化流技术，不仅实现了高质量的语音波形生成，还集成了时长建模功能；NaturalSpeech 3 [8]创新性地引入因子化编解码器与扩散模型，突破了现有技术在语音保真度与泛化能力上的局限；VoiceCraft [9]采用灵活的基于令牌(token)的架构，成功实现零样本语音合成与编辑；HAM-TTS [10]则借助分层声学建模方法，支持多语言合成与说话人规模扩展；ZSE-VITS [11]展现出强大的语音克隆性能。与此同时，Parallel WaveGAN [12]、Glow-TTS [13]等系统也分别在高效语音波形生成与单调对齐策略方面做出了重要贡献。尽管这些模型在众多基准测试中表现优异，但仍存在共性局限：语义理解深度不足、韵律变化不够丰富，且采用的确定性时长建模方式限制了语音节奏的表现力，难以生成富有情感与变化的自然语音。

为攻克上述难题，本文提出一种基于 BERT [14]模型增强 VITS 的表现力语音合成框架——BVTS。与传统基于音素的编码器不同，该框架集成预训练的 BERT 特征，能够捕捉深度的语境语义信息。同时，创新性地引入多模态特征融合策略，在特征层面实现语义、语音学与韵律线索的融合。基于可逆流技术 FLOW [15]构建随机时长预测器，可实现灵活的时间对齐与富有表现力的节奏建模，性能优于传统的确定性时长建模方法。在 LJ Speech 数据集与自定义游戏内数据集上的实验结果表明，相较于 VITS、VITS2、

NaturalSpeech 3 等性能强劲的基准模型，BVTS 在平均意见得分(MOS)、字符错误率(CER)、韵律指标及合成速度等方面均表现更优。实验结果证实，在语音合成系统中融入深度语义编码与多模态韵律建模技术，能够有效提升合成语音的质量。

2. 相关工作

传统文本语音合成系统高度依赖基于规则或参数统计的方法，这类方法灵活性有限，无法有效泛化到未见过的语言结构。由于采用人工设计特征与简化的声学模型，早期系统合成的语音往往存在单调、不自然的问题，难以满足实际应用对语音自然度的需求。

端到端神经架构的出现标志着 TTS 技术的重大突破。通过注意力机制与变分推理联合建模文本与声学映射关系，在提升语音自然度与流畅度的同时，省去了复杂的中间表征环节，简化了传统系统的模块化设计流程。但此类模型对浅层文本编码器的依赖，使其语义理解能力受到限制，尤其在处理多义词与长距离依赖关系时，易出现语义表征不精准的问题，进而影响合成语音的连贯性。

近年来，通过架构创新与数据规模扩展，致力于提升 TTS 模型的韵律表现与情感表达能力。NaturalSpeech 3 采用因子化编解码器结合扩散模型，在零样本场景下实现了高保真语音合成；VoiceCraft 借助令牌级(token-level)对齐与迁移机制，支持开放域语音编辑功能；HAM-TTS 通过分层声学建模，实现了跨语言与多说话人的模型扩展；PiCo-VITS 则融入音高轮廓特征，更精准地捕捉情感韵律信息。尽管这些模型各具优势，但普遍将文本视为令牌(token)或音素序列进行处理，未充分整合丰富的语义上下文，导致在表现力强或多语言场景中，易出现发音模糊、韵律平淡及语篇连贯性不足等问题，制约了模型的实际应用范围。

BERT 能够编码深度语境语义与长距离依赖关系，在解决多义性、建模复杂句法结构方面具备天然优势。为解决上述语义建模缺陷，探索将 BERT 等预训练语言模型融入 TTS 流程，探索将语义信息与韵律、语音学线索有效融合的机制，本文提出的 BVTS 框架通过将 BERT 引导的语义编码、特征级多模态融合与随机时长建模相结合，填补了现有研究的空白。架构旨在提升合成语音在多样化语言场景与情感语境下的可懂度与表现力，使模型能以精细的时间粒度联合建模语义与节奏特征，为解决当前 TTS 技术的核心局限提供了新的技术路径。

3. 方法

3.1. BVTS 架构概述

BVTS 模型在 VITS 框架的基础上进行了改进，旨在提升语音合成过程中的上下文理解能力、韵律表现力以及节奏可控性。整体架构如图 1 所示，主要由四个核心模块组成：多模态文本编码器、后验编码器、基于流的对齐模块以及神经声码器。

多模态文本编码器引入了预训练的 BERT 嵌入，以从输入文本中提取丰富的语义特征，从而克服传统基于音素的编码器在处理复杂语言与上下文关系时的局限性。这些语义特征与音素信息和韵律信息进行融合，形成统一的潜在表示，从而实现更加自然、富有表现力的语音合成。在训练阶段，后验编码器进一步提炼语音特征，以辅助波形重建。同时，基于流的对齐模块保证文本特征与语音特征之间的时序精确对应。声码器将潜在表示解码为高质量波形，确保语音的可懂度与自然度。

图 1 展示了训练阶段与推理阶段的区别。在训练阶段(虚线箭头)，模型联合优化对齐与波形生成，文本特征由多模态文本编码器处理，后验编码器则辅助潜在语音特征的重建。而在推理阶段(实线箭头)后验编码器被省略，以简化流程。文本特征可直接通过对齐模块与神经声码器处理，生成自然流畅的语音。

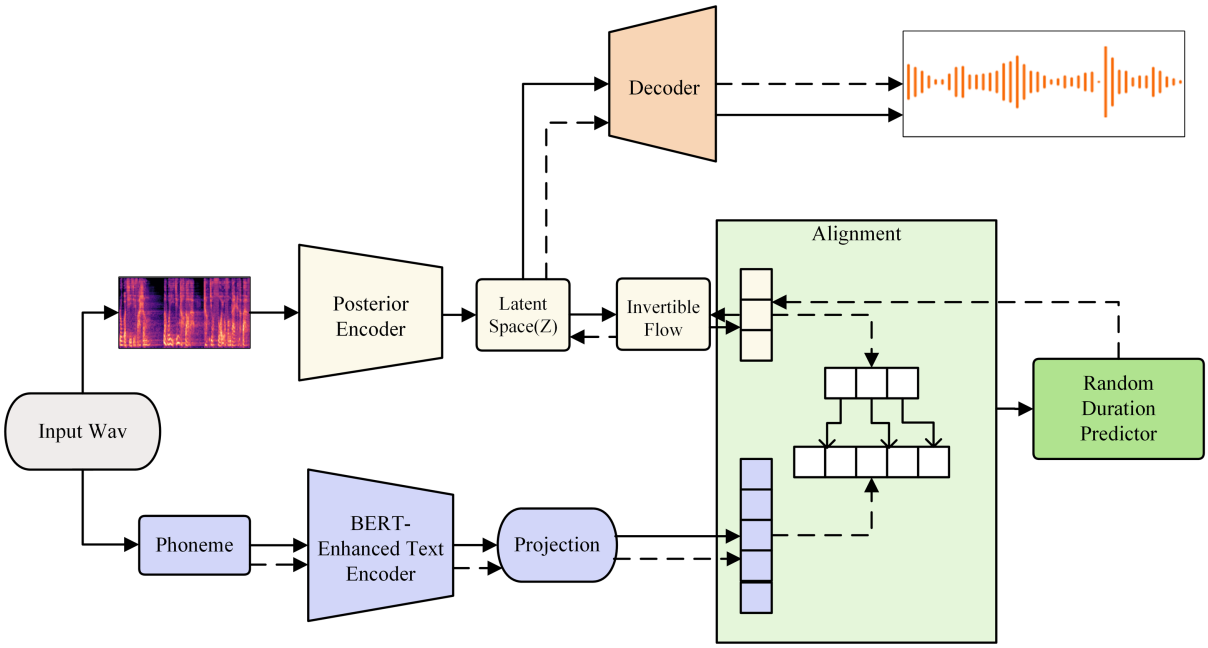

Figure 1. Overall structure
图 1. 整体架构图

3.2. 基于 BERT 的文本编码器

图 2 展示了 BVTS 框架的核心模块——基于 BERT 的文本编码器。该编码器旨在弥补传统语音合成系统在语义建模方面的不足，通过将深层语义理解与韵律和音素建模相结合，实现对复杂语言结构的建模能力。编码器利用预训练的 BERT 嵌入提取上下文中的语言特征，从而捕捉文本中的复杂语义关系。通过特征变换与多模态融合，该模块为生成自然、富有表现力且连贯的语音奠定了坚实基础。

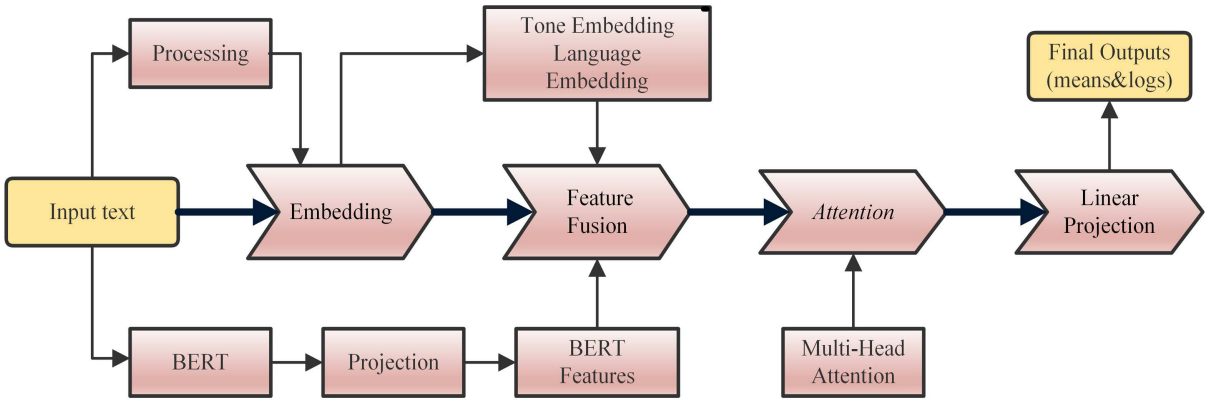

Figure 2. BERT-enhanced text encoder
图 2. 基于 BERT 的文本编码器示意图

3.3. 随机时长预测器

随机时长预测器(Stochastic Duration Predictor)是一种结合流模型(flow-based models)与卷积神经网络(CNN)的神经网络结构，如图 3 所示。为了实现对语音节奏和语速的精细控制，本研究采用基于归一化流(normalizing flows)和卷积层的随机时长预测器。预测器能够在文本条件上下文嵌入的引导下，从高斯分

布噪声中生成真实感的时长预测值。在推理阶段如图 4 所示，预测的时长用于调节语音输出的长度与对齐方式，从而保证语音韵律的自然性，尤其是在表现力和节奏敏感的语句中效果显著。

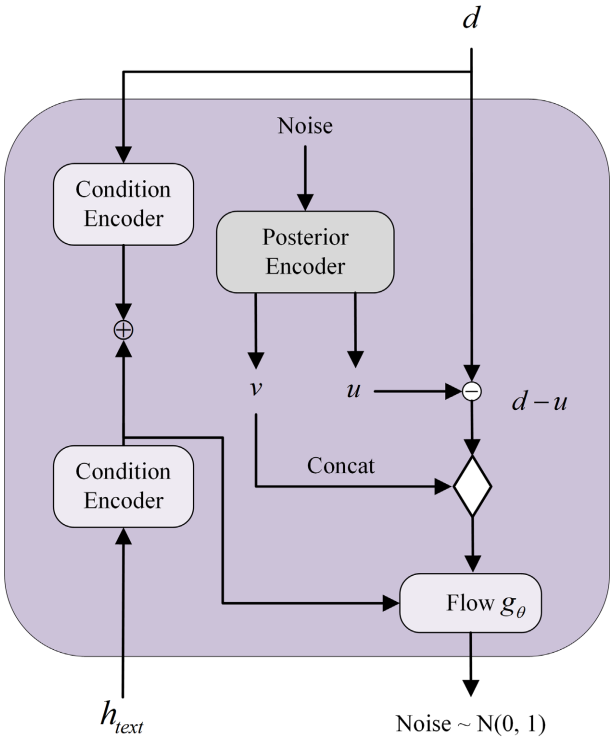

Figure 3. Duration predictor architecture
图 3. 时长预测器架构图

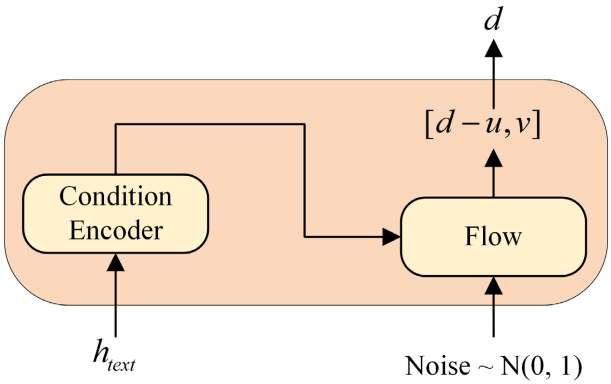

Figure 4. Inference phase of the duration predictor
图 4. 时长预测器推理阶段示意图

在该模块中，输入由后验编码器生成，记作 z，其为从标准正态分布中采样得到的样本，即均值为 0、方差为 1 的高斯分布随机变量。该样本经由流模型变换为随机分布 μ，随后通过 Sigmoid 函数进行归一化，使其满足式(1)并被限制在区间[0, 1]内：

$$u = \text{sigmoid}(z) = \frac{1}{1+\text{e}^{-z}} \tag{1}$$

在训练过程中，参数 d 表示时长(周期)，其通过两个条件编码器进行内部建模：输入为 d 与文本嵌入

h_{text}，输出结合后形成条件向量c_{text}。由于μ被约束在[0, 1]区间内，因此在推理阶段，利用差值$d-u$作为实际帧数，并将结果取整，如图 4 所示。

考虑到时长变量d的离散性(其为整数)，BVTS 模型采用连续近似方法来解决建模问题。这一方法避免了离散模型在测试时遇到未见数据时的似然估计困难。模型利用连续变量μ来近似建模时长d，并通过连续模型p_θ建立对$d-u$的概率建模。训练目标为最大化期望似然，如式(2)所示，从而保证模型能够精确预测时长：

$$\log p_\theta\left(d\middle|c_{\text{text}}\right)\geq Eq_{\varnothing(u,v|d,c_{text})}\left[\log\frac{p_\theta\left(d-u,v\middle|c_{\text{text}}\right)}{q_\varnothing\left(u,v\middle|d,c_{\text{text}}\right)}\right] \tag{2}$$

4. 实验与结果

4.1. 实验数据集

为了全面评估 BVTS 模型的性能与泛化能力，本文在公开数据集与自建数据集上进行了实验：

LJ Speech 数据集：该数据集包含 13,100 条由一名女性说话人录制的短英语语音片段。每条语音均配有对应的文本转录，并采用单声道 16 位 WAV 格式存储。该数据集是语音合成领域常用的基准数据集，用于与现有 TTS 模型进行对比实验。

自建游戏数据集：该数据集来源于游戏中不同角色的对话语音，涵盖了从简短的台词到较长的叙事性语音等多种富有表现力的发声形式。其中约 90%的语音片段时长介于 5 至 12 秒之间，剩余部分为角色技能释放时的语音。数据集为模型在多样化韵律条件下的表现提供了具有挑战性的测试环境。

4.2. 实验结果分析

4.2.1. 自然度与相似度评价

实验采用主观听感评价方法，包括平均意见得分(Mean Opinion Score, MOS)与比较平均意见得分(Comparative MOS, CMOS)，MOS 基于语音清晰度、流畅度与自然度进行评分，分值范围为 1 至 5 分。CMOS 基于合成语音与参考语音的差异进行对比，采用−3 至+3 的七分制。对各模型的语音自然度与相似度进行评估。共邀请 50 名来自不同语言背景与专业领域的评价者，在安静环境下使用专业音频设备，对每个系统随机选取的 20 条音频样本进行评分。

Table 1. MOS and CMOS scores of different models on the LJ Speech dataset
表 1. LJ Speech 数据集上不同模型的 MOS 与 CMOS 对比

Model	MOS (CI)	CMOS
Ground Truth	4.48 (±0.05)	0
Tacotron2 + HiFi-GAN	3.75 (±0.07)	−0.83
VITS	3.86 (±0.07)	−0.82
NaturalSpeech 3	4.15 (±0.06)	−0.33
HAM-TTS	4.32 (±0.07)	−0.26
VITS2	4.30 (±0.07)	−0.28
BVTS (DDP)	4.35 (±0.07)	−0.23
BVTS	**4.46 (±0.07)**	**0.11**

实验结果如表 1 所示。BVTS 系统在 MOS 与 CMOS 两项指标上均取得最优成绩，其 MOS 分数接近真实人声。采用确定性时长预测器(Deterministic Duration Predictor, DDP)的 BVTS 变体在各系统中排名第二，进一步验证了 BVTS 框架的优势。结果表明：1) 引入多模态信息的文本编码器显著提升了语音自然度；2) 即使在时长预测器架构相似的情况下，BVTS 仍优于现有 TTS 模型。

4.2.2. 语义准确率

为评估合成语音的语义保真度，本文采用谷歌对合成语音进行转写，并计算其字符错误率(Character Error Rate, CER)。结果如表 2 所示。

Table 2. Speech intelligibility test on the LJ Speech dataset
表 2. LJ Speech 数据集上的语音可懂度测试

Model	CER
Ground Truth	1.85
NaturalSpeech 3	2.13
VITS	2.33
VITS2	2.04
BVTS	**1.78**

BVTS 取得了最低的 CER 值，甚至优于真实语音。这表明 BERT 引导的语义建模与随机时长预测器的精细化控制，使模型在快速或节奏复杂的语句中也能保持准确的语义表达。

4.2.3. Mel 谱图分析

为了对比合成语音的频谱特性，我们选取了相同的测试句并绘制了 Mel 频谱图。VITS 模型生成的频谱如图 5 所示存在能量分布不均、谱线突变及噪声干扰等问题，导致谐波结构模糊、元音与辅音过渡不连贯，整体语音质量受限。相比之下，BVTS 模型生成的频谱表现更加平滑自然，如图 6 所示，谐波与共振峰清晰可见，辅音到元音的过渡流畅且频率变化连续，不存在明显的谱断裂或异常跳变。BVTS 生成的语音能量分布均衡，既避免了局部能量过高导致的失真，也避免了低能量区过大造成的模糊，从而更好地还原了自然语音特性。该结果与客观指标和主观评价实验相互印证，进一步验证了 BVTS 模型在语音合成质量上的优势。

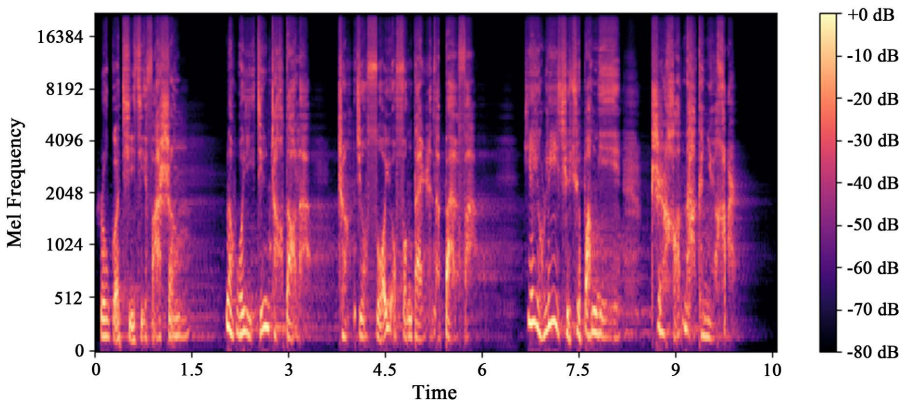

Figure 5. Spectrogram of synthesized speech by VITS model
图 5. VITS 模型生成语音的频谱图

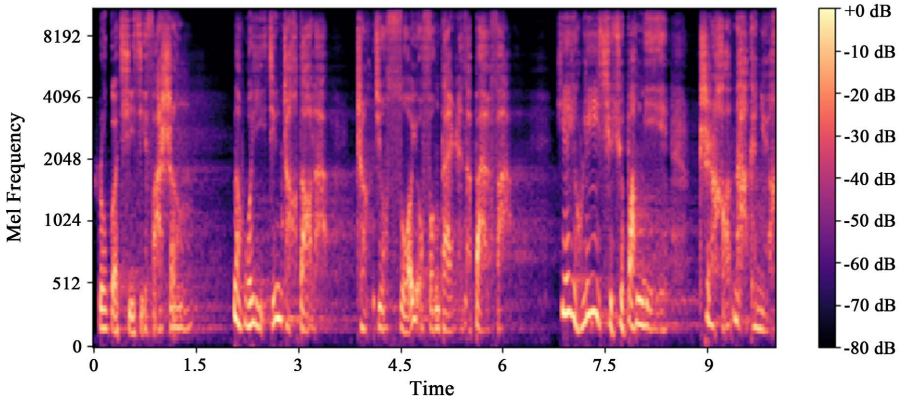

Figure 6. Spectrogram of synthesized speech by BVTS model
图 6. BVTS 模型生成语音的频谱图

4.2.4. 消融实验

为了评估 BVTS 架构中各个组件的独立贡献，我们进行了消融实验，通过修改或移除部分模块构建了三种模型变体。BVTS w/o BERT 即为将基于 BERT 的编码器替换为标准嵌入层，从而移除上下文语义建模能力。BVTS w/o Prosodic Fusion 是去除韵律和音素特征，仅保留由 BERT 提取的语义表示。为客观评价则引入梅尔倒谱失真(Mel-Cepstral Distortion, MCD)以量化合成语音与目标语音的频谱差异。这些模型在 LJ Speech 数据集上使用 MOS、CER 和 MCD 进行评估。结果如表 3 所示。

Table 3. System resulting data of standard experiment
表 3. 标准试验系统结果数据

Model Variant	MOS	CER	MCD
BVTS w/o BERT	4.08 ± 0.06	2.51	4.40
BVTS w/o Prosodic Fusion	4.18 ± 0.07	2.02	4.27
BVTS	**4.45 ± 0.07**	**1.78**	**3.95**

实验结果清楚地表明了各个组件的重要性。移除 BERT 会导致各项指标出现最大幅度的下降，CER 增加了 41%，MCD 上升了 0.35，凸显了上下文语义编码在模型中的关键作用。去除韵律融合会降低语音的可理解性和频谱质量，进一步验证了韵律与音素信息融合的必要性。将随机时长预测器替换为确定性预测器则在一定程度上影响了语音的自然度和时序表现，表现为 MOS 分数下降和 MCD 的中等幅度上升。所以语义建模、韵律融合以及随机时长建模这三部分均对 BVTS 的整体性能具有重要贡献。

5. 总结

本研究提出一种新型单阶段语音合成框架 BVTS，整合基于 BERT 的语义编码、多模态特征融合、可逆流建模及随机时长预测四大核心模块，多模态文本编码器可有效融合语义、韵律及语言学信息，不仅实现精准的韵律建模，还显著提升模型的多语言适配能力。在公开数据集与专有数据集上的实验结果表明，BVTS 在 MOS、CER 及合成效率等关键指标上，均优于当前主流先进模型。

参考文献

[1] Li, N., Liu, S., Liu, Y., Zhao, S. and Liu, M. (2019) Neural Speech Synthesis with Transformer Network. *Proceedings*

of the AAAI Conference on Artificial Intelligence, **33**, 6706-6713. https://doi.org/10.1609/aaai.v33i01.33016706

[2] Tokuda, K., Zen, H. and Black, A.W. (2002) An HMM-Based Speech Synthesis System Applied to English. *IEEE Speech Synthesis Workshop*, Santa Monica, 13 September 2002, 227-230.

[3] Schnell, N., Peeters, G., Lemouton, S., *et al*. (2000) Synthesizing a Choir in Real-Time Using Pitch Synchronous Overlap Add (PSOLA).

[4] Wang, Y., Skerry-Ryan, R.J., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., *et al*. (2017) Tacotron: Towards End-to-End Speech Synthesis. *Interspeech* 2017, Stockholm, 20-24 August 2017, 4006-4010. https://doi.org/10.21437/interspeech.2017-1452

[5] Elias, I., Zen, H., Shen, J., Zhang, Y., Jia, Y., Skerry-Ryan, R.J., *et al*. (2021) Parallel Tacotron 2: A Non-Autoregressive Neural TTS Model with Differentiable Duration Modeling. *Interspeech* 2021, Brno, 30 August-3 September 2021, 141-145. https://doi.org/10.21437/interspeech.2021-1461

[6] Kim, J., Kong, J. and Son, J. (2021) Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. *International Conference on Machine Learning*, *PMLR*, 18-24 July 2021, 5530-5540.

[7] Kong, J., Park, J., Kim, B., Kim, J., Kong, D. and Kim, S. (2023) VITS2: Improving Quality and Efficiency of Single-Stage Text-to-Speech with Adversarial Learning and Architecture Design. *Interspeech* 2023, Dublin, 20-24 August 2023, 4374-4378. https://doi.org/10.21437/interspeech.2023-534

[8] Ju, Z., Wang, Y., Shen, K., *et al*. (2024) Naturalspeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models.

[9] Peng, P., Huang, P., Li, S., Mohamed, A. and Harwath, D. (2024) Voicecraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild. *Proceedings of the* 62*nd Annual Meeting of the Association for Computational Linguistics*, Vol. 1, 12442-12462. https://doi.org/10.18653/v1/2024.acl-long.673

[10] Wang, K., Zhang, G., Zhou, Z., *et al*. (2025) A Comprehensive Survey in LLM (-Agent) Full Stack Safety: Data, Training and Deployment.

[11] Li, J. and Zhang, L. (2023) ZSE-VITS: A Zero-Shot Expressive Voice Cloning Method Based on Vits. *Electronics*, **12**, Article No. 820. https://doi.org/10.3390/electronics12040820

[12] Yamamoto, R., Song, E. and Kim, J. (2020) Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram. 2020 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Barcelona, 4-8 May 2020, 6199-6203. https://doi.org/10.1109/icassp40776.2020.9053795

[13] Kim, J., Kim, S., Kong, J., *et al*. (2020) Glow-tts: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. *Annual Conference on Neural Information Processing Systems* 2020, 6-12 December 2020, 8067-8077.

[14] Koroteev, M.V. (2021) BERT: A Review of Applications in Natural Language Processing and Understanding.

[15] Csikszentmihalyi, M., Abuhamdeh, S. and Nakamura, J. (2014) Flow. In: Csikszentmihalyi, M., Ed., *Flow and the Foundations of Positive Psychology*: *The Collected Works of Mihaly Csikszentmihalyi*, Springer, 227-238.