**多声调语言语音合成技术综述**

**摘要：**本文系统梳理并评述了多声调语言（如汉语、粤语、越南语、泰语等）语音合成技术的发展现状与关键方法。与非声调语言相比，多声调语言TTS系统在声调建模、韵律表达和模型对齐等方面面临更为复杂的挑战。文章回顾了从传统统计参数模型（如HMM、GMM）到主流深度学习端到端模型（如Tacotron、FastSpeech、VITS、Glow-TTS、Diff-TTS等）的技术演进，深入分析了不同模型在语音自然度、声调准确性、多语言适应能力及低资源场景下的表现。同时，汇总了常用的开源数据集和多维度评估指标，指出了当前TTS系统在声调处理、数据资源不足以及评测规范等方面存在的主要瓶颈。最后，对多声调语言语音合成问题的可行研究方向进行了探讨与展望，以期为后续相关领域的研究与技术进步提供参考和借鉴。

关键词：多声调；语音合成；声调建模；端到端模型

**A review of speech synthesis techniques for multi-tone languages**

**Abstract:** This paper systematically reviews the development status and key methods of speech synthesis technology in multi-tone languages (such as Chinese, Cantonese, Vietnamese, Thai, etc.). Compared with non-tonal languages, multi-tone language TTS systems face more complex challenges in terms of tonal modeling, prosodic expression, and model alignment. This paper reviews the technological evolution from traditional statistical parameter models (such as HMM and GMM) to mainstream deep learning end-to-end models (such as Tacotron, FastSpeech, VITS, Glow-TTS, Diff-TTS, etc.), and deeply analyzes the performance of different models in speech naturalness, tone accuracy, multi-language adaptability and low-resource scenarios. At the same time, the commonly used open-source datasets and multi-dimensional evaluation indicators are summarized, and the main bottlenecks of the current TTS system in terms of tone processing, insufficient data resources and evaluation specifications are pointed out. Finally, the feasible research directions of speech synthesis in multi-tone languages are discussed and prospected, in order to provide reference and reference for subsequent research and technological progress in related fields.

**Keywords:** multi-tone; speech synthesis; tonal modeling; End-to-end model

1.引言部分

1.1研究背景与意义

世界上的语言可分两类，即声调语言（tone language）和语调语言(intonation language)[1]。声调语言广泛分布于亚洲、非洲和美洲等地区，其特点是在只发同一个语音时，不同长短、不同调值的声调会构成不同的语义。据统计，在亚洲地区汉藏语系的约30多种语言中，除个别语言和方言外，绝大多数语言都有声调，典型的声调语言有普通话、粤语、越南语、壮语、泰语、老挝语等。与语调语言相比，声调语言在音系结构中对音高变化的依赖程度更高，而不同的声调又决定着不同的字义、语义，这对语音合成技术（Text to Speech,简称“TTS”，从文本转换为语音）提出了更高的要求[2]。

随着深度学习和神经网络模型的发展，基于序列到序列（Sequence-to-Sequence）框架的语音合成技术在自然性和流畅性方面取得了显著进展。当前主流的语音合成研究多以英语、日语等非声调语言为主，对声调语言而言，虽对汉语的研究也已相当成熟，但对其他具有丰富声调系统的语言，尤其是轮廓复杂、调型多样的东南亚及南方汉语方言的研究仍然较为匮乏。因此开展对多声调语言语音合成的研究不仅是对国家“语保”政策的响应也是对民族特色语言文化的传承。

本文从语言自身差异的角度阐述多声调语言语音合成的关键技术和发展现状。不同于现存较多的笼统地概括语音合成和转换的技术发展[3]、或是侧重于情感表达呈现类的情感语音合成[4]、亦或是基于深度学习的语音合成[5]、语音伪造[6,7]、语音克隆等综述文章，本文以语言结构差异的角度为出发点，系统的梳理并总结了传统语音合成技术以及主流的端到端（End to End）的系列语音合成模型在多声调语言上的应用，同时归纳整理了用于多声调语言语音合成的数据集和评价指标供相关研究者参考，并希望为后续多语言TTS系统研究提供借鉴与启发。

1.2声调语言的相关介绍

每一种语言都有一定的语调，音节的高低升降具有辨别词汇意义、语法意义功能的称为有声调的语言[8]。声调是通过音高的变化来区分词义的韵律特征，在声学上主要体现为基频（F0）的变化。不同的声调对应着不同的字义，如汉语中的“mā（阴平）、麻 má（阳平）、马 mǎ（上声）、骂 mà（去声）、吗 mɑ（轻声不标调）”，四个音调分别代表不同的意思。同时文献[8]中预测亚非地区的声调语言有着不断增多的倾向。与声调语言对应的语调语言是指音高不具有区分词义但具有在短语和句子层面语用功能的语言[9]，例如英语。英语不同于汉语，英语的单词，声调不能改变它的词义[1]。比如：cat这个词，无论你念成平调的[kæt]，升调的[kæt]，降升调的[kæt]，或是降调的[kæt]，它的意思仍然是“猫、猫科动物”，不会是别的意思。（声调和韵律的关系）

国内外常见的多声调语言主要集中在四大核心语系，各语系的调类数量与包含语言各有特征：汉藏语系的多声调语言调类普遍丰富，涵盖 2-10 个调类，具体包括汉语及汉语方言中的闽语、粤语、吴语、客家语等，壮侗语族的壮语、侗语、仫佬语、毛南语、拉伽语、泰语、老挝语（柬埔寨高棉语标准语声调不显著，其周边小众分支如 “斯丁语”“布劳语” 为多声调语言），苗瑶语族的苗语、瑶语、畲语等，以及藏缅语族的藏语（部分方言）、彝语、哈尼语、傈僳语、拉祜语、缅语等；尼日尔 - 刚果语系的多声调语言调类多为 2-5 个，主要包含班图语支的斯瓦希里语、祖鲁语、科萨语、恩德贝莱语、卢干达语、奔巴语等，以及大西洋 - 刚果语支的约鲁巴语、伊博语、埃维语、丰语、阿肯语等；南岛语系的多声调语言调类以 2-5 个为主，集中在台湾南岛语族的泰雅语、排湾语、鲁凯语、阿美语等，及菲律宾南岛语族的伊洛卡诺语、邦板牙语等；南亚语系的多声调语言调类多为 3-5 个，主要有越芒语支的越南语、芒语、赫蒙语等，卡西语支的卡西语、加罗语等；此外，其他语系中也存在少量多声调语言，如纳瓦霍语、玛雅语族语言等。常见的多声调语言的详细信息汇总如表1所示。(后续表格下插入脚注《中国少数民族语言研究六十年》（戴庆夏）)

表1 声调语言信息汇总表（部分）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **语系** | **语族/支** | **语种** | **常见调值** | **分布地区** |
| 汉藏语系 | 汉语族 | 普通话 | 4 | 全球 |
| 粤语 | 9/6 | 广州、香港、澳门[9] |
| 壮侗语族 | 壮语 | 9/6 | 广西、云南、越南北部等[10,11] |
| 泰语 | 5 | 泰国、老挝、缅甸等 |
| 老挝语 | 6 | 老挝、云南等[12] |
| 苗瑶语族 | 苗语 | 8 | 湖南、贵州、云南等 |
| 瑶语 | 8 | 广西、广东等 |
| 藏缅语族 | 藏语 | 4 | 西藏、贵州等 |
| 缅语 | 4 | 缅甸、孟加拉国等 |
| 南亚语系 | 越芒语支 | 越南语 | 8/6 | 越南、柬埔寨等 |

1.3 多声调语言TTS研究现状  
 多声调语言语音合成（TTS）是指将文本转换为具有自然音调和语调的语音的过程，其核心目标是生成符合语言规则、发音自然、带有情感和韵律的语音信号。近年来，多声调语言 TTS 研究虽取得一定进展，但因声调语言本身复杂性，仍面临诸多挑战。从研究对象看，汉语 TTS 研究相对成熟，已广泛应用于智能语音助手、有声读物等领域。如科大讯飞的汉语语音合成系统(添加文献引用)，在普通话合成自然度和准确性上表现出色，能精准还原汉语四个声调及轻声发音，为用户提供清晰、自然语音服务。但对于东南亚多声调语言及南方汉语方言，相关研究与应用较为匮乏。以粤语为例[13–15]，虽有研究关注其语音合成，粤语九声六调系统复杂，不同调型、调值组合使准确模拟粤语发音颇具难度，当前粤语 TTS 系统在自然度和流畅度上与普通话 TTS 系统仍有差距。

在技术方法上，传统语音合成技术如基于规则的合成[16,17]、基于单元挑选和波形拼接的合成[18]在多声调语言 TTS 中应用广泛。基于规则的合成通过制定声调规则生成音高轮廓，在一些资源匮乏的多声调语言 TTS 研究中有一定应用，但需大量人工编写规则，且规则难以覆盖所有声调变化情况，合成语音自然度受限。基于单元挑选和波形拼接的合成从语音库中挑选合适语音单元拼接成目标语音，在多声调语言中能保留原始语音自然特性，但对语音库规模和质量要求高，构建大规模、高质量多声调语言语音库成本高、难度大，尤其对稀有或濒危多声调语言，难以获取足够语音数据构建语音库。

随着深度学习发展，端到端语音合成模型在多声调语言 TTS 中逐渐崭露头角。如 Tacotron 系列模型[19,20]、Fast speech系列模型[21,22]、VITS[23]等，能直接将文本映射为语音波形，无需复杂中间处理步骤，在多声调语言 TTS 中可学习声调与语音之间复杂映射关系，提升合成语音自然度和准确性。但这些模型训练需大量标注数据，多声调语言数据标注难度大，不仅要标注文本对应的音标，还需标注准确声调信息，标注成本高、效率低，限制端到端模型在多声调语言 TTS 中的广泛应用 。此外，多声调语言 TTS 研究在数据集和评价指标方面也存在不足。现有多声调语言 TTS 数据集规模小、覆盖语言种类有限，难以满足深度学习模型训练需求。评价指标上，传统语音合成评价指标如自然度、可懂度等，不能充分反映多声调语言中声调合成准确性和自然度，缺乏专门针对多声调语言 TTS 的有效评价指标体系。

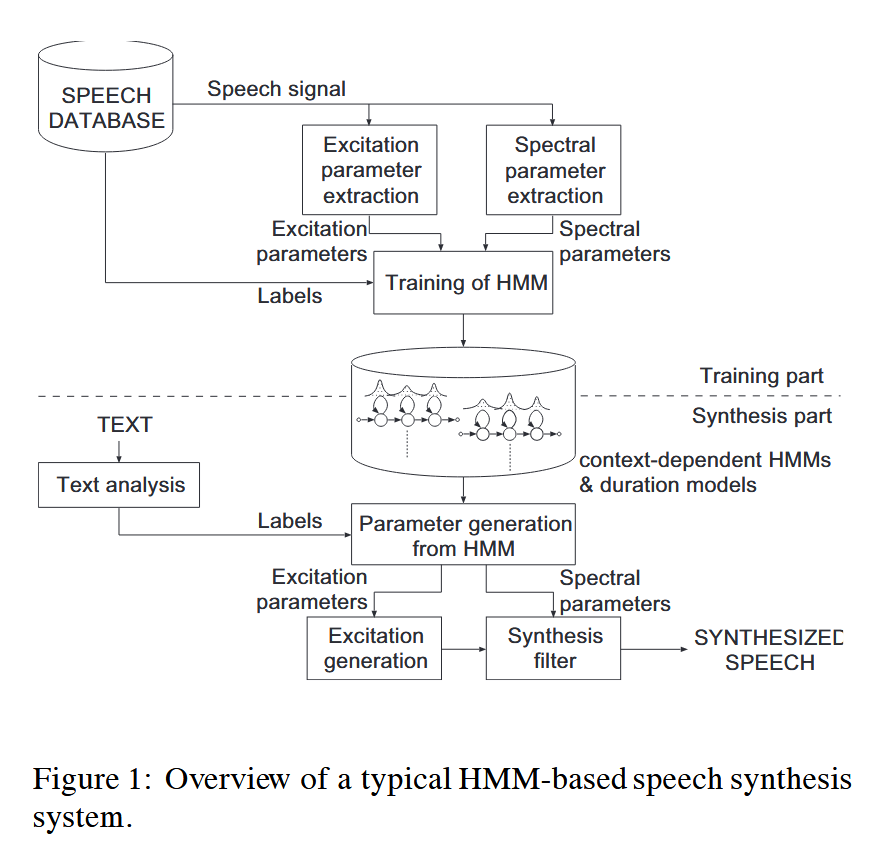
2.多声调语言语音合成建模方法的演进与范式变迁

2.1传统统计参数模型“：声调的解构与独立建模”

核心论点： HMM/GMM时代，语音被分解为声学特征（频谱、F0等）。声调（主要体现在F0轮廓上）被独立于频谱进行建模。

早期的语音合成方法主要包含拼接法和参数法[24]两大类。拼接法往往依赖于大规模高质量的语料数据，而对于大多数声调语言，特别是粤语、壮语、苗语等典型多声调语言而言，使用人数有限、资深的语言学专家数量更是稀少，这就导致了语料采集难度大、成本高、语料数据规模小等问题，并且经拼接法生成的语音质量不高、缺乏各种语音特征，因此本节对拼接法TTS不做过多赘述，主要围绕早期传统统计参数模型如隐马尔可夫模型[25]（Hidden Markov Model,以下简称HMM）在声调语言语音合成上的应用进行归纳和综述。

图1是一个典型的基于HMM的语音合成系统[26]。（图例后续重绘为中文）该系统由训练和合成两部分组成。在训练部分，从语音数据库中提取频谱（梅尔倒谱系数[27]及其动态特征）和激励（对数基频logF0及其动态特征）参数，并通过上下文相关的隐马尔可夫模型进行建模，为了恰当地对具有非浊音区域的可变维参数序列（如logF0）进行建模，使用了多空间概率分布（MSD-HMM）[28]。每个HMM都具有状态持续时间概率密度函数（PDF），以捕捉语音的时间结构。因此，该系统在统一的HMM框架中对频谱、由声调上下文决定的激励参数以及持续时间进行建模。在合成部分，系统先将待合成文本转换为上下文相关的标签序列，根据该标签序列拼接对应的HMMs，并基于状态持续时间概率密度函数确定HMMs的状态持续时间，随后通过参数生成算法[29]生成使输出概率最大化的频谱和激励参数序列，最后使用相应的语音合成滤波器（MLSA合成滤波器），直接从生成的频谱和激励参数中合成语音波形。

与非声调语言不同，对于汉语、粤语等多声调语言，合成质量的关键在于声调信息的建模。为此，系统在训练阶段将F0作为独立的激励参数流，利用MSD-HMM对有声段和无声段的基频分布进行建模，并在上下文特征中显式编码调类、调型及其前后关系，以增强模型对声调特征的辨识能力；在合成阶段则通过最大似然参数生成（MLPG）算法平滑输出F0曲线，得到与文本中声调标记相匹配的连续音高轮廓。通过这种机制，HMM才能在统一框架下保持声调的自然性和稳定性。

早在2012年Wang等[14]首次将HMM引入到粤语语音合成系统（HTS），选用Crystal（如图1所示）为基本架构，通过从语料库中训练上下文相关的音节HMM模型，再使用基于决策树聚类训练上下文相关的HMM模型以更精准的提取声学参数，替代传统基于语料库的波形拼接法，生成的音频频谱过渡更平滑、自然度更高。

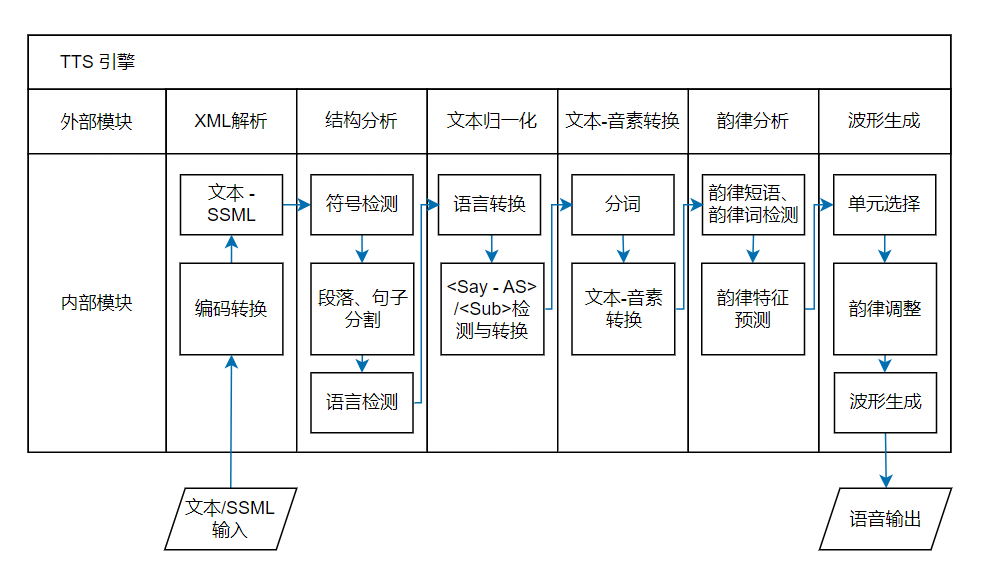


图1 Crystal架构图[30]——TTS引擎统一框架的模块划分与流程图

SHER[31]等分析现代文字台湾语（MLT）的子音节特征并应用HMM工具包（HTK）识别声调（八声调）音素信息构建出了高鲁棒性的台湾声调音素集，基于该音素集和HMM开发出高质量的台湾TTS系统。

虽然基于HMM合成的语音质量已经远远优于拼接法，但比起真实的自然语音仍略显不足，自然语音中包含着丰富的情感色彩和韵律特征，尤其是针对小语料库，韵律在语义区分、句法结构标记、情感表达等方面起着关键作用，Chen[32]等提出了韵律词(PW)级别的韵律调整框架，融合发音特征(AFs)与残差补偿策略从有限语料中挖掘潜在的韵律单元，在音素/PW级分层优化单元序列，并基于分类树对异常音节进行HMM参数与模板匹配双重调整。

然而，基于统计参数的语音合成往往建模能力有限，实际的语音特征分布要比建模时复杂得多，无法更有效的捕捉语音中的长距离依赖和上下文信息，导致合成语音的自然度、流畅度仍存在一定缺陷，除了传统统计参数模型外，李建文[33]等注意到多数研究都忽视语音声调对情感表达的影响，尝试从数学角度分析汉语四种声调特征参数及变换，采用多项式函数拟合方式来重构声调，构建音高和音长变化可控、所需参数更少且适应各种发音的声调变换模型用于语音合成。

2.2基于深度学习的端到端模型“序列到序列模型：声调的隐式端到端学习”

核心论点： 以Tacotron为代表，模型不再显式分解F0。声调信息被隐式地编码在生成的梅尔频谱中，由注意力机制自动学习文本与声调模式之间的映射。

随着深度学习技术近年来的蓬勃发展，语音合成经历了从传统统计参数模型到深度神经网络模型的重大转变。传统方法（如HMM）依赖于在语音合成过程中需要对声学模型和声码器独立建模，这就导致了模型缺少联合建模能力，且对于长文本语料数据建模不足，这种不足主要体现在对语音中的长时依赖建模困难、对声调等韵律信息建模不足[34]，致使语音的连贯性、自然度受到极大限制。

基于深度学习的语音合成模型主要分为两种：在传统语音合成的基础上引入深度学习方法，对其部分模块进行改进，进而提升传统TTS的整体性能，例如百度的Deep Voice-1和Deep Voice-2等；基于端到端的语音合成系统，是当前TTS的主流方法其发展历程如图2所示，所谓端到端是不同于传统TTS方法，这种方法摒弃了对声学模型和声码器进行分离建模，而是联合建模输入文本到语音波形的整体过程。这种联合建模方式显著地简化了模型结构，同时有助于捕捉语音中的长时依赖信息，更清晰地建模语义、语法和韵律之间的联系，这对于声调语义高度依赖且富有声调对立如粤语、壮语的语音合成尤为重要。

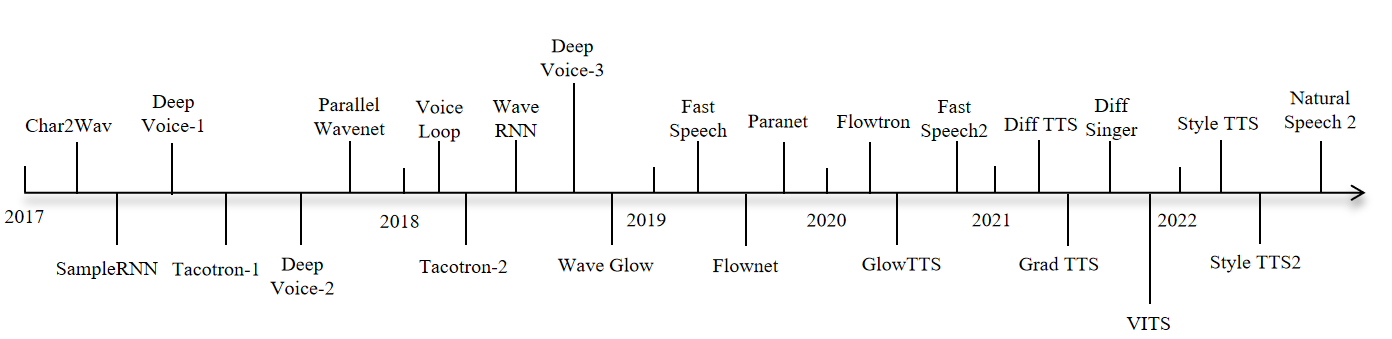


图2 基于深度学习的的语音合成技术发展历程

2.2.1 Tacotron系列

Tacotron系列模型是序列到序列语音合成中最为典型且富有重大创新性的一系列模型。最初由Google团队在2017年提出[19]，旨在实现从文本到语谱（Mel-Spectrogram）的端到端联合建模，模型结构如图3所示。Tacotron摒弃了传统统计参数方法中对声学模型和声码器进行解耦建模的方式，而是借助注意力机制（Attention Mechanism），实现输入文本序列到输出语谱序列的联合建模。而在2018年Tacotron2基于 Tacotron 又进行了功能优化，解决了原模型在对齐稳定性、语音质量和训练效率等方面存在的问题。

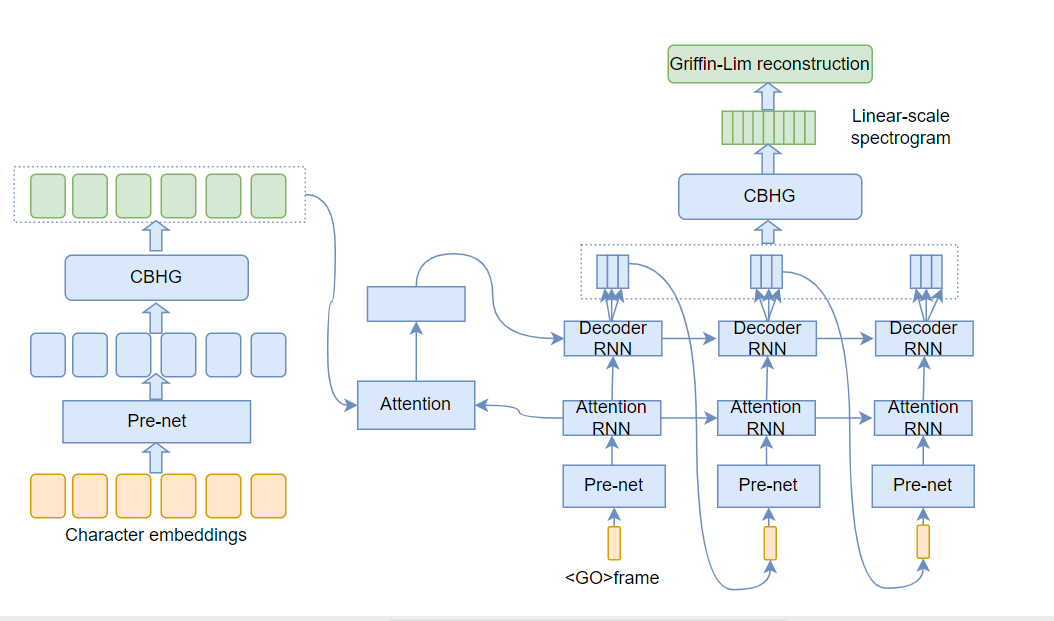


图3 Tacotron结构示意图

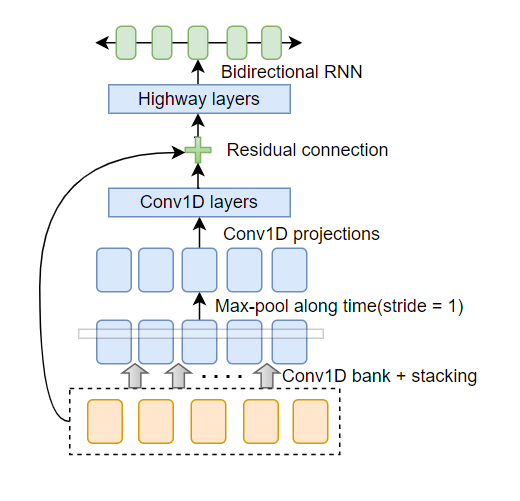


图4 CBHG模块示意图

Tacotron 2 延续了从字符到 Mel 频谱图的端到端建模思路，但摒弃了 Tacotron 中结构复杂且训练不稳定的 CBHG 模块如图4所示，转而采用基于卷积层和双向 RNN 的简洁编码器架构，有效提升了训练效率和模型稳定性。此外，Tacotron 2[20] 引入了位置感知的注意力机制（Location-Sensitive Attention），建模特征如（3）(4)所示，在解码过程中实现更加稳定和连续的文本-语音对齐，显著减少了跳词或重复的现象。在声码器部分，Tacotron 2 不再使用传统的 Griffin-Lim 算法，而是将生成的 Mel 频谱图输入到一个单独训练的 WaveNet 神经声码器中，利用其强大的建模能力进行高质量语音波形重建。

 （3）

 (4)

表示编码器第i步与解码器第j步之间的相关性得分；

表示归一化后的注意力权重，反映解码器j时刻“关注”输入i的程度。

这样的两阶段设计，即“端到端文本到 Mel 谱图 + 独立神经声码器”，不仅有效规避了原 Tacotron 在语音重建质量上的瓶颈，还增强了模块替换与升级的灵活性，使 Tacotron 2 在多个标准语音合成任务中取得了接近人类自然语音的表现。

虽然Tacotron及其改进版本在普通话等主流语言的语音合成中表现优越，但在结构复杂、声调丰富,尤其在语调和声调信息深度绑定的少数民族语言中仍面临诸多挑战。例如，在粤语中，句末升调常用于表达疑问语气，而 Tacotron2 难以准确合成这类“陈述式疑问句”（如“你去学校？”）的上扬语调，主要由于模型仅基于字符或音素进行建模，缺乏语义层面的理解能力，导致对语用语调的感知不足。为解决这一问题，Bai [35]等人提出将 Tacotron2 与 BERT 句子分类器相结合，以显式识别句子类型（陈述/疑问），辅助模型进行语调控制。该方法在 CanTTS 粤语语料库上的实验显示，在 MOS、F0 Frame Error 等指标上显著优于原始 Tacotron2，尤其在 declarative question 类型句子的升调合成方面效果突出，验证了语义引导机制在 Tacotron 框架中的潜在价值。

除了语调感知问题，多声调语言还普遍存在声调密集、同音异调等带来的语音歧义难题。在此背景下，Tacotron2 在如苗语这类结构紧凑、声调信息承担词义区分功能的语言上亦存在性能瓶颈。为此，蔡姗[36]等人提出了基于 Tacotron2 架构的 子音节建模与单调对齐优化方法（SRHTTS），以“声母 + 韵母+声调”的子音节为输入建模单位，显著提升模型对不同声调下发音差异的辨识能力。该方法还引入单调对齐损失（Monotonic Alignment Loss），强化注意力机制的对齐准确性，有效缓解了模型在处理苗语这类单音节语言时容易出现的跳词、重复现象。

从粤语的语调控制到苗语的声调歧义处理，这两项研究表明 Tacotron2 虽作为通用端到端架构，在多声调语言中仍需结合语言结构特异性进行定制优化，未来也可为如壮语、瑶语、越南语等复杂声调语言的语音合成提供启示。

2.2.2 Fast Speech系列2.3 非自回归模型：声调与时长的解耦与显式预测

核心论点： 以FastSpeech为代表，为了解决速度和鲁棒性问题，模型重新将时长和音高（声调）解耦，并通过独立的预测器进行显式预测。

随着Tacotron系列模型在语音合成任务中展现出的优异表现，训练和推理过程中的高速且稳定地生产语音成为又一个重大挑战。为此，Fast Speech系列应运而生[21,22]，旨在克服Tacotron中序列到序列模型训练困难（对齐不稳定、训练缓慢）、推理时需要自回归（导致推理时延增加）等不足，为语音合成提供高速且稳定的非自回归方案。

Fast Speech最重大且最基础的创新在于非自回归（Non-Autoregressive）序列到序列结构。它通过时长预测器（Duration Predictor）预测每个输入（比如每个音素）所对应的时长信息，进而实现对输入和输出序列进行对齐。这种对齐方式消除了对注意力机制的依赖，显著地简化训练过程，同时也消除了解码时的序列依赖，显著地提升推理速度。而Fast Speech2[22]是在Fast Speech基础上进行了一系列改进，旨在弥补非自回归模型中对语义信息建模不足导致的语音缺少活力和鲁棒性的不足。

FastSpeech2 凭借其高效的非自回归结构和对韵律特征的灵活建模能力，已成为多语言、低资源语音合成的主流技术平台。例如，Qiao[37]等人以 FastSpeech2 为基础，针对多声调、低资源的越南语语音合成场景，提出了基于多特征跨语言迁移学习的方法：首先在大规模普通话数据上预训练模型，通过利用国际音标（IPA）映射，实现对中越两种语言音素的统一表示，然后将模型迁移并微调至越南语，从而有效应对数据稀缺和过拟合难题。同时，该系统在 FastSpeech2 结构中融合了丰富的韵律和语言学特征（如音节长度、音高、能量、分词、词性、韵律边界），并设计了新的韵律嵌入层与 BiLSTM 模块，使模型即使在缺乏人工标注的情况下也能生成自然流畅、韵律丰富的合成语音。实验结果显示，该方法显著提升了越南语合成语音的自然度和韵律表现，揭示了 FastSpeech2 在低资源、多声调语言场景中的技术扩展潜力与创新能力。

2.2.3 VITS与语音合成-音色建模融合方法2.4 新兴统一生成模型：声调的联合建模与精细化生成

核心论点： 以VITS、Glow-TTS、Diff-TTS等为代表，模型追求更深层次的端到端，将声学模型和声码器统一建模，或者利用强大的生成范式（流模型、扩散模型）直接从文本生成高质量波形。

在 Fast Speech 系列通过非自回归建模显著提升语音合成效率的基础上，研究者开始进一步关注语音合成的自然度与音色表达能力，即如何在高效的同时实现更真实、更具表现力的语音生成。此背景下，如何实现声学模型与声码器（vocoder）的深度融合建模，成为推动端到端语音合成发展的关键方向。

VITS（Variational Inference Text-to-Speech）模型[23]针对这一问题提出了一种创新的端到端联合建模方案。不同于传统两阶段TTS系统中“先生成 Mel 频谱图，再由独立声码器还原波形”的模式，VITS 在统一架构中同时完成文本到语音的建模任务，实现了声学模型与声码器的深度耦合与协同优化。该模型融合了变分自编码器（VAE）、正向传播（Flow-based）建模与 GAN 声码器的优势，在提高训练稳定性与语音质量的同时，显著简化了模型训练与推理流程。

虽然 VITS 模型通过端到端联合建模实现了高质量语音生成，但在多声调语言（如普通话、越南语、苗语）中，仍可能出现声调偏移问题，即合成语音的音节正确但声调错误，影响可懂度和语义识别。为此，Tao[38]等人进一步探索了将 VITS 与声调感知离散单元建模结合，以提升语音合成在声调语言中的表现力。他提出了Tone Unit 框架，一种通过引入CTC声调监督机制与有限标量量化（FSQ）方式对离散语音单元进行显式建模框架，有效避免了传统 VQ 方法中的代码本崩溃问题，并提升对声调信息的捕捉能力。Tone Unit框架先利用 SPIRAL 编码器提取连续语音表示，再通过 FSQ 实现高效、可控的离散量化，最后借助 CTC 声调标签监督训练，使生成的离散单元更准确地对齐音节和声调分布。最终，这些离散单元可输入到 VITS 解码器中作为条件，构成完整的声调感知 TTS 系统。

除了从语音建模机制本身进行结构性优化外，也有学者从语言建模单元设计与语音数据预处理策略等角度出发，进一步提升 VITS 在具体声调语言中的适应性与合成质量。Wisetpaitoon [39]等人以泰语语音合成为研究对象，构建了一套完整的端到端 TTS 系统，并系统比较了三种语言建模单元（音素、音节、词）在 VITS、Tacotron2-DDC 和 Your TTS 等主流模型中的表现差异，该研究首次在声调语言语音合成中引入了语音修剪处理（静音段剔除）与双重评估机制，以全面提升声调保持能力与发音准确性。

2.2.4 其他模型

随着 VITS 将语音合成与音色建模有效统一，端到端建模范式逐渐走向任务融合与表达增强。然而，除了 VAE-GAN 与 Flow 融合架构外，学术界仍在积极探索其他更具可控性、更易训练或更具多样性的语音生成模型。为此，研究者从不同建模机制出发，提出了Flow-TTS、Glow-TTS 、 Diff-TTS 与Style-TTS、等一系列创新方法，旨在进一步提高语音自然度、风格迁移能力以及生成过程的透明性与稳定性。

归一化流模型（normalizing flows）由于其在建模表达能力和生成灵活性方面的优势，逐渐被引入TTS任务中。Flow TTS[40]是首个将流模型应用于 TTS 的非自回归方法之一，其通过最大似然估计训练声学模型，并实现了快速并行的语音生成。然而，Flow TTS 依赖外部对齐器来提供音素与帧之间的映射，这不仅引入额外复杂度，也限制了模型在多语言、特别是多声调语言如汉语、越南语、泰语等中的灵活性与扩展性。

为解决上述问题，Glow TTS[41]在 Flow TTS 的基础上引入了更加结构化的 Glow 模型，并创新性地提出了 Monotonic Alignment Search（MAS） 方法，使模型能够在训练阶段自动学习音素与帧之间的隐式对齐关系，从而实现真正意义上的端到端建模。Glow TTS 不仅在生成语音的自然度、稳定性和训练效率上优于 Flow TTS，更重要的是，其对齐模块的可微分设计为多声调语言建模提供了更好的支持。多声调语言往往存在声调与音素、声调与韵律之间的耦合关系，而 Glow TTS 中显式可控的生成机制和灵活的建模能力，使其更易于适配此类语言的音高与时长特征，从而提高语音自然度与语言一致性。

除基于归一化流的模型外，扩散概率模型（Diffusion Models）也被引入语音合成领域。 **Diff-TTS**[42]将反向扩散过程用于声学建模，从固定高斯噪声逐步还原出梅尔频谱，有效捕捉语音的细节与变化。Li[43]等人针对缅甸语提出了一种低资源扩散模型TTS方法，创新性地结合了视觉文本嵌入与音素编码，有效提升了对复杂字符和声调的建模能力。该方法利用预训练的RoBERTa模型加强语义理解，并通过引入OT-CFM训练策略，在保证合成语音有较高自然度的前提下实现推理速度提升54%。

Style TTS[44]提出一种结合条件扩散模型与可学习风格嵌入的非自回归方法，通过显式建模说话风格来提升语音的自然度与个性化表达。该方法使用一个先验风格表示空间，引导扩散过程生成具备目标风格的梅尔频谱，并通过联合训练实现风格与内容的解耦。在此基础上，Style TTS2[45]进一步引入统一的先验模型架构，以提升风格一致性和推理稳定性。该方法不仅在多个公开语音合成任务中实现了接近甚至超过 GT（Ground Truth）的主观自然度评分（MOS），同时也支持零样本语音合成（zero-shot TTS）与高质量音色迁移，在语音风格控制方面表现出强大的灵活性与可控性。相较于 Diff-TTS 等纯扩散路径，Style TTS 系列强调风格建模与语义表达解耦，更适用于多说话人、多风格或情感 TTS 场景。

综上所述，近年来在端到端语音合成范式不断演进的背景下，学术界围绕生成建模机制展开了多样化探索。从 Flow TTS 与 Glow TTS 等归一化流模型，到 Diff-TTS、Style TTS 系列所代表的扩散生成路径，不同方法在建模能力、生成效率、风格控制及语言适配性等方面各具优势。特别是在多声调语言和多风格合成任务中，这些新兴模型凭借其对对齐、风格、声调等特征的显式建模能力，展示出超越传统架构的潜力。未来，这类具备可控性强、对齐机制灵活、风格表达丰富的建模框架，有望进一步推动语音合成系统向高保真、多语言、多风格和低资源适应性方向发展。

3.开源多声调语言TTS语料概览

针对多声调语言的语音合成与识别研究，丰富且多样的开源数据集是基础资源。以下介绍几种典型多声调语言的主要数据集如表2所示，，涵盖普通话、粤语、越南语、闽南语、泰语及老挝语等。

表2 声调语言开源数据集汇总

|  |  |  |
| --- | --- | --- |
| 数据集 | 简介 | 资源链接 |
| 标贝(baker)CSMSC | 12h女声平均16字普通话 | https://www.databaker.com/open\_source.html |
| 希尔贝壳Aishell-3 | 85h 88035句218位说话人普通话 | https://sos1sos2sixteen.github.io/aishell3/ |
| Zhvoice | 900h 113万条文本3200位说话人普通话 | https://github.com/fighting41love/zhvoice |
| PhoAudiobook | 1494h710位说话人越南语 | https://huggingface.co/datasets/thivux/phoaudiobook |
| KeSpeech | 1542h27237位说话人普通话及8个次方言 | <https://github.com/KeSpeech/KeSpeech> |
| MAGICDATA | 755h1080位说话者普通话 | https://openslr.org/68/ |
| TTS-SCCantFSC | 3.22h女声1位说话人粤语 | https://magichub.com/datasets/guangzhou-cantonese-speech-corpus-for-tts/ |
| ASR-SCCantCabSC | 5h6219句话语10位说话人粤语 | https://magichub.com/datasets/guangzhou-cantonese-scripted-speech-corpus-in-the-vehicle/ |
| HmongSpeech | 4650个音频片段(自建语料库) | http://sxjxsf.gzmu.edu.cn/info |

4.评估指标与系统性能对比

4.1 TTS系统主流评价指标

为全面衡量文本到语音系统的合成效果，当前研究中广泛采用主观评价与客观评价两类指标，从不同维度评估模型生成语音的自然度、音质、可懂度与风格表达能力等关键性能。

4.1.1 主观评价指标

主观评价是语音合成系统评估中最具代表性的方法之一，通常依赖人工听感打分，其结果更贴近人类真实感知。主要包括：

MOS（Mean Opinion Score）：即平均主观评分，测试者对语音样本在自然度、清晰度或音色相似度等方面进行1~5分打分，是衡量语音合成质量的最常用指标。

CMOS（Comparison MOS）：即成对比较主观评分，评估者需判断两个系统输出中哪一个更优，并给出评分差值，适合评估多个模型之间细微的质量差异。

SMOS（Similarity MOS）：用于评价合成语音与目标说话人语音在音色上的相似度，常用于多说话人或音色克隆系统。

XAB/ABX 测试：用于音色或风格迁移任务，评估者需在两个样本中选择更接近参考样本的合成语音，用于判断模型在说话人特征或风格方面的保真性。

4.1.2客观评价指标

客观指标依赖自动计算，适合大规模模型评估或训练过程中的性能监控。常见指标包括：

Mel Cepstral Distortion（MCD）：衡量预测梅尔倒谱系数与参考语音之间的差异，值越小表示音质越接近真实语音。MCD的计算公式如（5）所示：

（5）

D是梅尔倒谱系数的维度（通常为13或25）；

表示参考音频在第t帧的第d维MCC；

表示合成语音在相应帧的MCC；

F0 RMSE / F0 Correlation：用于评估基频预测精度，尤其在语调建模、多声调语言中具有重要意义。

WER（Word Error Rate） / CER（Character Error Rate）：通过自动语音识别（ASR）系统对合成语音转写并与参考文本比对，间接衡量语音可懂度。

Duration Error / Alignment Score：评估生成语音与文本对齐的准确性，特别适用于非自回归或扩散模型的对齐机制分析。

4.2 系统对比分析

多声调语言 TTS 系统的性能差异本质上源于模型架构对声调特征的建模能力。以下从架构设计、建模单元、对齐机制三个维度，对主流模型进行系统性对比：

表3主流模型信息表（部分）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **模型类型** | **代表模型** | **架构核心** | **多声调语言适配设计** | **数据需求** |
| **序列到序列模型** | Tacotron2 | 编码器-解码器 + 注意力机制+ Wave Net 声码器 | 引入位置敏感注意力优化对齐+ BERT 识别句子类型控制语调[35] | 大规模标注语料（≥100 小时） |
| **非自回归模型** | FastSpeech2 | 时长预测器 + 声学特征生成器 | 跨语言迁移学习（普通话→越南语），融合韵律嵌入层处理声调时序特征[37] | 中等规模语料（50-100 小时） |
| **端到端联合模型** | VITS | 变分自编码器 + 流模型 + GAN 声码器 | 引入 Tone Unit 框架显式建模声调单元+ CTC （联结主义时序分类）监督提升声调对齐准确率[38] | 标注精细的声调语料（需声调标签） |
| **流模型** | Glow-TTS | 归一化流架构+Monotonic Alignment Search（MAS）实现隐式对齐 | 显式可控的音高生成机制，适配泰语等调型复杂语言的音高轮廓建模[39] | 需对齐预训练数据（额外标注成本） |
| **扩散模型** | Diff-TTS | 反向扩散生成梅尔频谱+视觉文本嵌入增强语义理解 | 缅甸语低资源场景中引入 OT-CFM 策略[43] | 少量标注数据（可结合预训练模型） |

如表3所示，，不同模型在架构设计上各有侧重，特别是在声调特征的建模方式及对齐机制上的差异，是决定多声调语言合成性能的关键因素。

基于公开数据集的实验结果进一步验证了上述模型在多声调语言中的实际表现：

1. 汉语普通话（CSMSC 数据集，12 小时）

表4 普通话 TTS 模型评价指标对比表（CSMSC 数据集）

|  |  |  |  |
| --- | --- | --- | --- |
| **评价指标** | **Tacotron2** | **FastSpeech2** | **VITS** |
| MOS（自然度） | 4.2±0.3 | 4.3±0.2 | 4.5±0.1 |
| F0 RMSE（Hz） | 12.5 | 11.8 | 10.2 |
| MCD（dB） | 4.8 | 4.5 | 4.1 |
| 声调错误率（%） | 5.7 | 4.9 | 3.2 |

如表4所示，VITS 在声调错误率方面明显优于其他模型，归因于其端到端联合建模策略直接对音高轮廓进行学习，从而提升了声调表达的准确度[20,22,23]。

1. 粤语（CSS10-Cantonese 数据集，3.22 小时）

表5 粤语 TTS 模型评价指标对比表（CSS10-Cantonese数据集）

|  |  |  |  |
| --- | --- | --- | --- |
| **评价指标** | **Tacotron2** | **Tacotron2+BERT** | **VITS+Tone Unit** |
| MOS（疑问语调） | 3.8±0.4 | 4.3±0.3 | 4.4±0.2 |
| F0 帧错误率（%） | 18.2 | 12.7 | 9.5 |
| 九声识别准确率（%） | 82.3 | 87.6 | 91.4 |

如表5所示，结合 BERT 的语义引导有效提升了 Tacotron2 在粤语疑问句调合成的表现，而 VITS 通过离散声调单元的显式建模，进一步优化了复杂调类的识别和合成效果[35,38]。

1. 越南语（VLSP 数据集，低资源场景）

表6 越南语 TTS 模型评价指标对比表（VLSP数据集）

|  |  |  |  |
| --- | --- | --- | --- |
| **评价指标** | **传统 HMM** | **FastSpeech2** | **FastSpeech2 + 迁移学习** |
| MOS（自然度） | 2.9±0.2 | 3.5±0.3 | 3.8±0.2 |
| 声调连贯度（%） | 65.7 | 78.3 | 85.6 |
| 推理速度（句 / 秒） | 0.8 | 5.2 | 5.2 |

如表6所示，跨语言迁移学习（普通话→越南语）显著提升了 FastSpeech2 在低资源环境下的自然度和声调连贯性，实现了自然度提升31%且保持了较快的推理速度[37]。

综上所述，针对多声调语言的 TTS 系统，端到端联合模型（如 VITS）通过显式声调单元和联合训练机制，在声调准确率和自然度上表现突出；非自回归模型（FastSpeech2）结合迁移学习，在中等资源及低资源条件下展现了良好适应性；序列到序列模型（Tacotron2）依赖大量数据和语义引导，仍具备竞争力。未来多声调语言 TTS 的提升方向，将侧重于模型架构对声调特征更细粒度的建模能力与对齐机制的进一步优化。

5.研究挑战与发展趋势

5.1研究挑战

尽管基于深度学习的端到端语音合成模型已在多种语言场景中实现了语音自然度和生成效率的显著提升，但在多声调语言的建模中仍面临诸多挑战。首先，声调作为语义区分的基本手段，其音高轮廓不仅与音素发音强耦合，还受韵律、语境与语言风格等多因素影响，这种复杂性使得当前模型在声调识别与生成上仍存在一定的模糊性和不稳定性。现有系统多数通过梅尔频谱间接表达声调信息，缺乏显式建模路径，导致在连续语流、弱读或跨声调边界情况下的合成语音可能出现声调错配、调类漂移或语义偏差，特别是在高调类语言如越南语、泰语以及多民族语言中更为明显。

此外，多声调语言中的数据资源稀缺性进一步加剧了模型训练的难度。除汉语外，诸如壮语、瑶语、苗语等少数民族语言普遍面临数据量有限、音素系统未标准化、标注质量不一致等问题，限制了深层神经网络对其声调特征的充分学习和泛化能力。即便引入迁移学习或跨语种建模机制，仍难以有效捕捉其语言特有的声调—音素—语义之间的深层耦合关系。

建模流程中的对齐机制也是影响合成性能的关键因素。传统注意力机制在端到端系统中的应用虽然提高了建模效率，但对齐的不稳定性依然存在，在某些语音段落中可能产生跳对齐、延迟或重复现象。这一问题在声调显著依赖时序的语言中尤为突出，进一步影响了语音的连贯性与可懂度。虽然如GlowTTS等模型尝试引入可微分对齐算法（如MAS），实现自动学习音素与帧的对应关系，但在复杂语音场景中仍存在鲁棒性瓶颈。而StyleTTS 系列虽引入显式风格表示与条件扩散生成机制，有效提升语音的多样性和可控性，但对声调调控与跨说话人一致性的支持仍有提升空间。尤其是在零样本或低资源场景下，语音生成易出现调型模糊、语气失真或音色漂移，影响最终语音质量与表达一致性。

目前面向多声调语言语音合成的评估体系尚不健全。主流的评价方法多依赖于主观听感评分（如MOS）或通用客观指标（如MCD、F0 RMSE），缺乏针对声调还原、调类准确性、音高连续性等多声调特有维度的系统性指标。这在模型评估和跨系统对比中造成一定限制，也不利于推动新型建模方法的精准验证与迭代优化。

5.2发展趋势

面向未来，融合语言学知识与建模机制的协同设计将是提升语音合成系统性能的关键方向。将声调标注、词法特征、音系结构等语言先验信息纳入端到端建模流程，有望增强模型对多声调语言内部规律的结构化建模能力，从而提升语音的自然度与表达一致性。在建模范式方面，归一化流（normalizing flows）与扩散概率模型（diffusion models）等新兴生成机制仍将是研究热点，尤其在提升合成过程的可控性、多样性以及训练的稳定性方面具有显著潜力。未来的TTS系统有望通过条件建模机制融合语义、语境、说话人和风格控制信号，形成更为通用、可扩展的语音生成框架。

在跨语言与低资源语种场景中，构建具备跨语种泛化能力的预训练模型将成为主流趋势。基于多语言联合训练和迁移学习策略，结合共享音素空间与统一建模架构，有望提升系统在多声调语言（如汉语、越南语、泰语、壮语等）中的适应能力。同时，引入多模态辅助信息（如文本语义嵌入、口型图像、发音视频等）也将成为提升建模表达力的重要途径，有助于解决小样本条件下语音质量不稳定的问题。

评估体系方面，构建针对多声调语言特性的综合指标将是未来工作的关键环节。除了传统的自然度与音质评价，还需开发更具诊断性的量化指标，如声调还原准确率、音高连续性度量、调类边界识别率等，以提升模型优化的针对性和评估的可信度。同时，推动面向多声调语言的标准化语料建设和开源评测平台搭建，将有助于形成可比性强、开放共享的研究生态，为该领域的持续发展与实际落地奠定基础。

6.总结

本文系统回顾了近年来基于深度学习的多声调语言语音合成技术的发展脉络，重点分析了从早期基于串行建模的 Tacotron 系列、FastSpeech 系列，到融合对齐学习与生成建模的 VITS，再到引入归一化流与扩散模型等新兴路径的发展演进。针对多声调语言在音高、时长、音素-声调耦合等方面的建模挑战，文中梳理了各类模型在音素对齐机制、声调感知建模、语音风格控制等方面的创新设计，并讨论了其在提高合成自然度、音色多样性与语音可控性等方面的表现与局限。

此外，本文总结了当前语音合成系统的主流评价指标体系，指出在面向多声调语言建模时，应结合语言学特征建立更具诊断性的评估框架，以提升系统优化效率和评估可信度。在此基础上，进一步探讨了多声调语音合成面临的数据稀缺、对齐学习困难、语言适配性差、风格迁移能力不足等研究挑战。

展望未来，构建融合语言学知识的多层次建模机制，发展更加可控、稳健、泛化能力强的生成模型，将成为多声调语言TTS系统的关键突破方向。与此同时，推动低资源语言语料库建设、多语言联合建模与多模态协同增强策略，将进一步拓宽语音合成技术的应用边界，为多语言人机交互、智能语音助理、数字文化传播等领域提供核心支撑。

参考文献

[1] 薛永. 论声调语言和语调语言[J]. 彭城职业大学学报, 2001(4): 30-31.

[2] LI J, HASEGAWA-JOHNSON M. Autosegmental Neural Nets 2.0: An Extensive Study of Training Synchronous and Asynchronous Phones and Tones for Under-Resourced Tonal Languages[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 1918-1926.

[3] 潘孝勤, 芦天亮, 杜彦辉, 等. 基于深度学习的语音合成与转换技术综述[J]. 计算机科学, 2021, 48(8): 200-208.

[4] 施昊翔, 张旭龙, 王健宗, 等. 情感语音合成综述[J]. 大数据, 2024, 10(5): 56-73.

[5] 张小峰, 谢钧, 罗健欣, 等. 深度学习语音合成技术综述[J]. 计算机工程与应用, 2021, 57(9): 50-59.

[6] 魏方达, 刘淼, 孙毅, 等. 音视频深度伪造与鉴伪综述[J/OL]. 计算机工程, 2025[2025-05-17]. https://doi.org/10.19678/j.issn.1000-3428.00EC0070700.

[7] 许裕雄, 李斌, 谭舜泉, 等. 语音深度伪造及其检测技术研究进展[J]. 中国图象图形学报, 2024, 29(8): 2236-2268.

[8] 李锦芳. 汉藏语言声调研究的回顾与展望[J]. 西藏民族学院学报（社会科学版）, 1993(2): 58-65, 71.

[9] 马秋武. 汉语语调音系研究中的几个问题[J]. 当代外语研究, 2014(1): 6-11, 77.

[10] 黄行. 中国民族语言声调类型的实证研究[J]. 东方语言学, 2019(1): 18-27.

[11] 李锦芳. 壮语和越南语汉借词语音差异及成因[J]. 中央民族大学学报（哲学社会科学版）, 2023, 50(4): 130-135.

[12] 许秀娜. 老挝语声调特点及老挝学习者汉语声调习得研究[D/OL]. [2025][2025-05-18]. https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFDTEMP&filename=1024609461.nh.

[13] JIAN-QING WANG, KA-HO WONG, PHENG-ANN PHENG, 等. A real-time cantonese text-to-audiovisual speech synthesizer[C/OL]//2004 IEEE International Conference on Acoustics, Speech, and Signal Processing: 卷 1. Montreal, Que., Canada: IEEE, 2004: I-653-656[2025-05-13]. http://ieeexplore.ieee.org/document/1326070/.

[14] WANG X, WU Z. An HMM-based Cantonese speech synthesis system[C/OL]//2012 IEEE Global High Tech Congress on Electronics. Shenzhen, China: IEEE, 2012: 141-142[2025-05-26]. http://ieeexplore.ieee.org/document/6490141/.

[15] BAI Q, KO T, ZHANG Y. A Study of Modeling Rising Intonation in Cantonese Neural Speech Synthesis[EB/OL]. arXiv, 2022[2025-06-23]. http://arxiv.org/abs/2208.02189.

[16] DO T T, TAKARA T. Vietnamese Text-To-Speech system with precise tone generation[J]. Acoustical Science and Technology, 2004, 25(5): 347-353.

[17] WIN K Y, TAKARA T. Myanmar text-to-speech system with rule-based tone synthesis[J]. Acoustical Science and Technology, 2011, 32(5): 174-181.

[18] 才让卓玛, 才智杰. 基于语料库的藏语语音合成单元选择算法[J]. 中文信息学报, 2017, 31(5): 59-63.

[19] WANG Y, SKERRY-RYAN R J, STANTON D, 等. Tacotron: Towards End-to-End Speech Synthesis[EB/OL]. arXiv, 2017[2025-06-16]. http://arxiv.org/abs/1703.10135.

[20] SHEN J, PANG R, WEISS R J, 等. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions[EB/OL]. arXiv, 2018[2025-06-23]. http://arxiv.org/abs/1712.05884.

[21] REN Y, RUAN Y, TAN X, 等. FastSpeech: Fast, Robust and Controllable Text to Speech[EB/OL]. arXiv, 2019[2025-06-16]. http://arxiv.org/abs/1905.09263.

[22] REN Y, HU C, TAN X, 等. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech[EB/OL]. arXiv, 2022[2025-06-16]. http://arxiv.org/abs/2006.04558.

[23] KIM J, KONG J, SON J. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech[EB/OL]. arXiv, 2021[2025-06-16]. http://arxiv.org/abs/2106.06103.

[24] BLACK A W, ZEN H, TOKUDA K. Statistical Parametric Speech Synthesis[C/OL]//2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP ’07: 卷 4. 2007: IV-1229-IV-1232[2025-09-27]. https://ieeexplore.ieee.org/document/4218329.

[25] MASUKO T, TOKUDA K, KOBAYASHI T, 等. Speech synthesis using HMMs with dynamic features[C/OL]//1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings: 卷 1. 1996: 389-392 卷 1[2025-09-27]. https://ieeexplore.ieee.org/document/541114/.

[26] ZEN H, NOSE T, YAMAGISHI J, 等. The HMM-based Speech Synthesis System (HTS) V ersion 2.0[J]. 2007.

[27] FUKADA T, TOKUDA K, KOBAYASHI T, 等. An adaptive algorithm for mel-cepstral analysis of speech[C/OL]//[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing: 卷 1. 1992: 137-140 卷1[2025-09-28]. https://ieeexplore.ieee.org/document/225953/.

[28] TOKUDA K, MASUKO T, MIYAZAKI N, 等. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling[C/OL]//1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258): 卷 1. 1999: 229-232 卷1[2025-09-28]. https://ieeexplore.ieee.org/document/758104/.

[29] TOKUDA K, YOSHIMURA T, MASUKO T, 等. Speech parameter generation algorithms for HMM-based speech synthesis[C/OL]//2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100): 卷 3. 2000: 1315-1318 卷3[2025-09-28]. https://ieeexplore.ieee.org/document/861820/.

[30] ZHIYONG W, GUANGQI C, MENG M H, 等. A unified framework for multilingual text-to-speech synthesis with SSML specification as interface[J]. Tsinghua Science and Technology, 2009, 14(5): 623-630.

[31] YUNG-JI SHER, MING-CHUN HSU, YU-HSIEN CHIU, 等. Development of a Taiwanese Speech Synthesis System Using Hidden Markov Models and a Robust Tonal Phoneme Corpus[J]. Journal of Information Science and Engineering, 2024, 40(2).

[32] CHEN Y Y, WU C H, HUANG Y C, 等. Candidate Expansion and Prosody Adjustment for Natural Speech Synthesis Using a Small Corpus[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(6): 1052-1065.

[33] 李建文, 王咿卜. 多项式函数拟合实现汉语声调的语音合成[J]. 西安科技大学学报, 2021, 41(3): 506-515.

[34] YANG S, XIE L, CHEN X, 等. Statistical Parametric Speech Synthesis Using Generative Adversarial Networks Under A Multi-task Learning Framework[EB/OL]. arXiv, 2017[2025-06-16]. http://arxiv.org/abs/1707.01670.

[35] BAI Q, KO T, ZHANG Y. A Study of Modeling Rising Intonation in Cantonese Neural Speech Synthesis[EB/OL]. arXiv, 2022[2025-06-23]. http://arxiv.org/abs/2208.02189.

[36] 蔡姗, 王林, 谭棉, 等. 基于子音节表征的苗语语音合成方法[J]. 科学技术与工程, 2024, 24(19): 8176-8185.

[37] QIAO Z, YANG J, WANG Z. Multi-Feature Cross-Lingual Transfer Learning Approach for Low-Resource Vietnamese Speech Synthesis[C/OL]//Proceedings of the 2023 3rd International Conference on Artificial Intelligence, Automation and Algorithms. Beijing China: ACM, 2023: 175-180[2025-06-23]. https://dl.acm.org/doi/10.1145/3611450.3611476.

[38] TAO D, TAN D, YEUNG Y T, 等. ToneUnit: A Speech Discretization Approach for Tonal Language Speech Synthesis[EB/OL]. arXiv, 2024[2025-06-23]. http://arxiv.org/abs/2406.08989.

[39] WISETPAITOON K, SINGKUL S, SAKDEJAYONT T, 等. End-to-End Thai Text-to-Speech with Linguistic Unit[C/OL]//Proceedings of the 2024 International Conference on Multimedia Retrieval. Phuket Thailand: ACM, 2024: 951-959[2025-06-23]. https://dl.acm.org/doi/10.1145/3652583.3658029.

[40] VALLE R, SHIH K, PRENGER R, 等. Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis[EB/OL]. arXiv, 2020[2025-07-03]. http://arxiv.org/abs/2005.05957.

[41] KIM J, KIM S, KONG J, 等. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search[EB/OL]. arXiv, 2020[2025-07-03]. http://arxiv.org/abs/2005.11129.

[42] JEONG M, KIM H, CHEON S J, 等. Diff-TTS: A Denoising Diffusion Model for Text-to-Speech[EB/OL]. arXiv, 2021[2025-07-03]. http://arxiv.org/abs/2104.01409.

[43] LI Y, JIAN Y. LOW-RESOURCE BURMESE SPEECH SYNTHESIS BASED ON VISUAL TEXT EMBEDDING AND DIFFUSION MODEL[C/OL]//Proceedings of the 2024 4th International Conference on Artificial Intelligence, Automation and Algorithms. Shanghai China: ACM, 2024: 34-40[2025-07-03]. https://dl.acm.org/doi/10.1145/3700523.3700531.

[44] LI Y A, HAN C, MESGARANI N. StyleTTS: A Style-Based Generative Model for Natural and Diverse Text-to-Speech Synthesis[EB/OL]. arXiv, 2023[2025-07-03]. http://arxiv.org/abs/2205.15439.

[45] LI Y A, HAN C, RAGHAVAN V S, 等. StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models[EB/OL]. arXiv, 2023[2025-07-03]. http://arxiv.org/abs/2306.07691.