表现性语音合成综述

唐浩彬[1,2]，张旭龙[1]，王健宗[1]，程宁[1]，肖京[1]
1. 平安科技（深圳）有限公司，广东 深圳 518063；
2. 中国科学技术大学，安徽 合肥 230026

摘要
语音合成是语音、语言和机器学习领域的一个热门研究课题，旨在合成给定文本的可理解和自然的语音，在工业中有广泛的应用。语音合成的目标之一是合成自然的语音，而目前的语音合成在情感、韵律等方面还有很大的改进空间。对表现性语音合成进行了全面的调查，旨在更好地了解当前的研究现状和未来的趋势。对近年来基于情感及韵律的表现性语音合成进行了全面的总结、比较和分析。首先介绍了普通语音合成的传统实现方式及瓶颈；然后引入表现性语音合成并描述表现性语音合成在情感、韵律等方面为语音合成自然化带来的增益；最后对表现性语音合成进行了展望和总结。

关键词
语音合成；表现性语音合成；机器学习

中图分类号：TP391 文献标志码：A doi: 10.11959/j.issn.2096-0271.2022082

A survey of expressive speech synthesis

TANG Haobin[1,2], ZHANG Xulong[1], WANG Jianzong[1], CHENG Ning[1], XIAO Jing[1]
1. Ping An Technology (Shenzhen) Co., Ltd., Shenzhen 518063, China
2. University of Science and Technology of China, Hefei 230026, China

Abstract
Speech synthesis is a hot research topic in the field of speech, language and machine learning, which aims to synthesize understandable and natural speech for a given text. It has a wide range of applications in industry. One of the goals of speech synthesis is to make the synthesized speech natural, and there is still a lot of room for improvement in emotion, prosody and other aspects of speech synthesis. A comprehensive survey of expressive speech synthesis was conducted with the aim of better understanding current research status and future trends. A comprehensive summary, comparison and analysis of emotion-based and prosodic speech synthesis in recent years were given. Firstly the traditional way and bottleneck of common speech synthesis were introduced, then expressive speech synthesis was introduced and the benefits of expressive speech synthesis in the aspects of emotion and prosody were described. Finally, the prospect and summary of expressive speech synthesis were presented.

Key words
speech synthesis, expressive speech synthesis, machine learning

0 引言

语音合成旨在从文本合成可理解且自然的语音，在人类通信中有广泛的应用，长期以来一直是人工智能、自然语言和语音处理的研究课题。语音合成的研究历史可追溯至18世纪，从早期的基于规则的机械式、电子式语音合成器[1]，发展到基于波形拼接[2-5]、统计参数的语音合成[6-10]。近年来，基于深度学习和神经网络的建模方法在机器学习领域各个任务上取得了快速的发展，语音合成技术也在此基础上得到了显著的提升。随着信息技术及人工智能技术的发展，各种应用场景对语音合成的效果有了越来越高的要求。

一个好的语音合成系统应该产生自然且可理解的语音，大量的语音合成研究工作旨在提高语音合成的可理解性和自然度。自然度在很大程度上取决于合成语音的表现力，而表现力由内容、音色、韵律、情感和风格等多种特征共同决定。目前的语音合成模型合成的语音往往采用机械、木讷、单一的方式进行表达，仅仅保证了合成语音内容的正确性，在自然度方面十分欠缺。为了弥补自然度方面的缺陷，表现性语音合成应运而生。表现性语音合成旨在从音色、韵律、情感和风格等多方面提升合成语音自然度的语音合成，是目前语音合成领域中比较活跃的方向。表现性语音合成和单纯语音合成的区别是，它更关注合成声音的自然度，包括风格（如新闻播报、讲故事、解说）、情感（如生气、兴奋、悲伤）、韵律（如重读、强调）等。其中韵律是指在去除了语音、说话人身份和通道效应（即录音环境）引起的变化后剩余的语音信号的变化[11-12]。表现性语音合成发展的问题在于处理一对多映射问题，这是指在持续时间、音调、音量、说话者风格、情感等方面，存在与同一文本相对应的多个语音表达。为了解决这个问题，表现性语音合成系统必须隐式或显式地输入许多在简单文本输入中没有给出的因素，如韵律中的语调、重音、节奏等因素。表现性语音合成的关键技术在于如何高效地利用数据集中关于这些因素的显式标签或显式音频信号作为额外输入以增添合成语音中的表现力，或者如何隐式地对音频中的这些因素建模并对其进行控制。以语音中每个音素的持续时间为例，在不带任何韵律风格的语音中该持续时间分布大多在0~25 ms，而当说话者自发地延长或缩短音素以表达多样的韵律提升话语的表现力时，音素的持续时间分布将扩展到0~40 ms[13]，而如何控制该持续时间的因素并没有在输入的文本中提供。表现性语音合成希望通过对类似音素持续时间变化这种文本中没有的因素建模，从而实现在合成的语音中通过自发的改变音素持续时间以表达多样的韵律。这是因为话语表达的意义本质上是文本所不明确的。例如简单的语句“他坐在树底下”可以用很多不同的方式来表达。如果这句话是对“他在哪里”问题的回答，说话者可能会强调“他”一词，以表明它是问题的答案。说话者可能会决定用上升的音调来回答以表达知识的不确定性。这些句子的语调带有文本内容未指明的语境和上下文含义，一般来说言语中存在许多这样的细微差别，它们传递的信息超出了文本内容，而表现性语音合成的任务就是隐式或显式地对这些细微差别建模，补充语音合成难以体现的细微差别，以达到使合成语音更加自然、更具有表现性的目的。

本文对语音合成及表现性语音合成进行了概述。首先阐述语音合成，而后引入更加关注语音自然度的表现性语音合成，

最终目标是在不破坏语音合成质量的前提下，探索最大限度地提高合成语音表现力的方法。

1 语音合成

1.1 语音合成发展

第一个基于计算机的语音合成系统出现在20世纪。早期基于计算机的语音合成方法包括发音合成、共振峰合成和级联合成。后来，随着统计机器学习的发展，有人提出了统计参数语音合成，用于预测语音合成的频谱、基频和持续时间等参数。随着计算机科学技术的发展，基于神经网络的语音合成逐渐成为主流方法。最开始出现的发音合成通过模拟人类发音器的行为产生语音，如嘴唇、舌头、声门和移动的声道。之后，基于共振峰的方法[14-17]和基于单元选择的波形拼接方法出现，再到基于隐马尔可夫模型（hidden Markov model，HMM）的统计参数语音合成（statistical parametric speech synthesis，SPSS）方法[6-10]。统计参数语音合成方法[18]的基本思想是首先生成语音所需的声学参数[19]，然后使用一些算法[20]从生成的声学参数中恢复语音。与以前的语音合成系统相比，SPSS方法有几个优点，具体如下。①自然：音频更自然。②灵活性：便于修改参数来控制生成语音。③数据成本低：比级联合成需要更少的记录。然而，SPSS方法也有其缺点，具体如下。①由于低沉、嗡嗡声或嘈杂音频等伪影，生成的语音比较难理解。②生成的语音比较机械，是可很容易区分的机器人声音。

最近，端到端语音合成系统已经取得了显著的进步，显示出几乎与人类相似的语音质量。在端到端语音合成系统中，Tacotron[21]、Tacotron2[22]和Transformer TTS[23]等自回归（autoregression，AR）模型首次利用注意机制显示出最先进的性能。Tacotron模型是首个真正意义上的端到端语音合成深度神经网络模型。与传统语音合成相比，它没有复杂的语音学和声学特征模块，而是仅用<文本序列，语音声谱>配对数据集对神经网络进行训练，模型如**图1**所示。

然而，由于模型的生成速度慢，且由于注意失败而缺乏稳定性，近年来有人提出了非自回归（not-autoregression，NAR）模型，如FastSpeech[24]、FastSpeech2[25]等。FastSpeech提供了一种基于Transformer的前馈网络，用于并行生成语音合成的梅尔频谱，通过并行梅尔频谱生成，FastSpeech在合成语音方面比之前具有类似质量的自回归模型速度快得多。FastSpeech2模型如**图2**所示，在FastSpeech基础上提高了语音合成速度并提出了方差适配器，旨在向音素隐藏序列中添加方差信息（如持续时间、基音、能量等），从而为语音合成中的

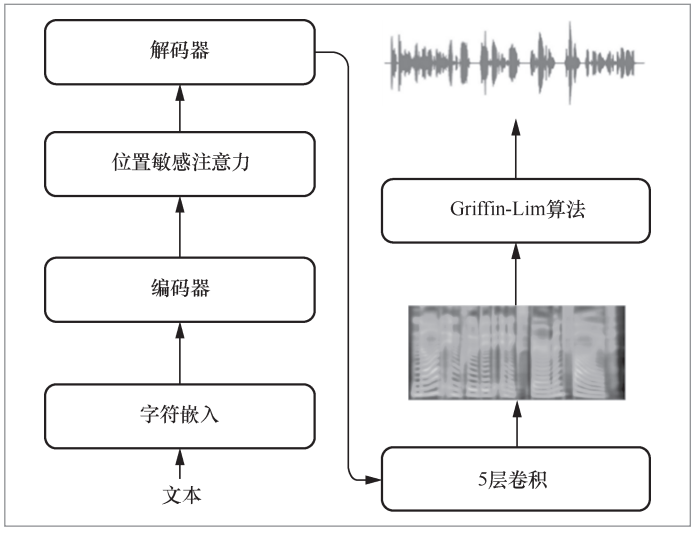

图1 自回归语音合成模型

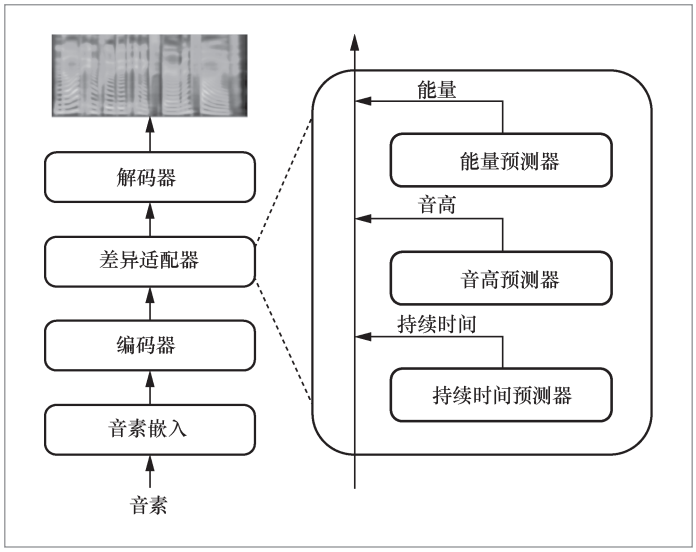

图 2 非自回归语音合成 FastSpeech2 模型

一对多映射问题提供足够的信息来预测变换的语音。FastSpeech2模型作为目前最好的非自回归模型，也是许多非自回归表现性语音合成的基础模型。

1.2 神经网络语音合成组成

神经网络语音合成主要由文本分析前端、声学模型和声码器三部分组成，如**图3**所示。首先，文本前端将文本转换为标准输入。然后，声学模型将标准输入转换为中间声学特征，用于建模语音的长期结构。最常见的中间声学特征是频谱图、声码器特征或语言特征。最后，使用声码器填充低电平信号细节，并将声学特征转换为时域波形样本。

文本到语音转换过程中有几种数据表示，具体如下。①字符：文本的原始格式。②通过文本分析获得的语言特征，包含丰富的语音和韵律语境信息。音素是语言特征中重要的元素之一，在基于神经网络的语音合成模型中，音素通常单独用来表示文本。③声学特征：是语音波形的抽象表示。④波形：语音的最终格式。在统计参数语音合成中，线谱对（line spectral pairs，LSP）[26]、梅尔频率倒谱系数（Mel frequency cepstral coefficients，MFCC）[27]、梅尔广义系数（Mel-generalized coefficients，MGC）[28]、基频和频带非周期性（band a periodicities，BAP）[29-30]被用作声学特征，这些特征可以通过STRAIGHT[31]和WORLD[32]等声码器轻松转换为波形。在基于神经网络的端到端语音合成模型中，通常使用梅尔频谱或线性谱图作为声学特征，并使用基于神经网络的声码器将其转换为波形。

2 表现性语音合成

深度学习方法合成的语音音调平滑，没有节奏感和表现力，因此往往与真实的人声有一定的差距。为了实现表现性语音合成的目标，即提升合成语音的自然度，需要考虑3个部分："说什么""谁说""如何说"。"说什么"由输入文本和文本前端控制，实现"说什么"是语音合

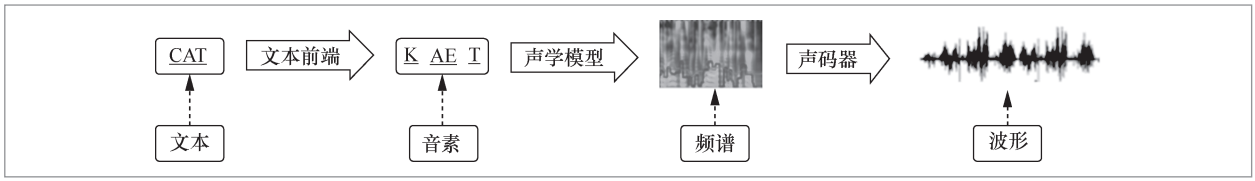

图 3 神经网络语音合成模型组成

成的基本任务。“谁说”和“如何说”则是表现性语音合成在不影响语音合成完成“说什么”的基础上需要重点关注的两个问题。“谁说”可以通过收集一个人的大量语音数据，然后训练模型来学习模仿说话人的声音来控制，衍生为表现性语音合成中的多说话人任务，将在后续内容具体介绍。“如何说”由合成语音的音调、语速和情感等韵律信息控制，可以利用表现性语音合成通过显式或隐式地对这些信息进行建模并加以控制。

语音合成在实现文本到语音的过程中涉及很多变化的信息。最基本的文本信息可以是字符或音素，代表合成语音的内容（即“说什么”）。为了产生真实自然的语音，文本到语音系统必须隐式或显式地输入许多在简单文本输入中没有给出的因素，这些因素控制着语音的自然度。表现性语音合成需要解决的问题就是如何控制、分离并在合成的语音中加入这些因素。用于控制语音的一些属性有：代表说话人的信息（即“谁说”），即一些多说话人语音合成系统通过说话人查找表或说话人编码器明确建模说话人表示；韵律、风格和情感信息，包括语调、重音和节奏等，代表如何说出文本。韵律/风格/情感是提高言语表达能力的关键信息，绝大多数关于表达性语音合成的工作专注于提高言语的韵律/风格/情感；录音设备或噪声环境是传递语音的渠道，与语音的内容/说话人/韵律无关，但会影响语音质量。该领域的研究工作主要集中在语音合成的信息分离、控制和去噪。建模这些信息的方式主要有显式和隐式两种。如果有每个属性的标签，即显式信息，将标签作为模型训练的输入让模型学习这些属性，并使用相应的标签在推理中显式地控制合成语音。然而，当没有标签可用时，如何分离和控制这些属性是一个挑战，在这一过程中需要对这些变化的属性进行隐式的建模，并以此实现分离和控制。可以根据建模的信息类型对模型进行分类，具体见**表1**。其中，显式信息可以显式获得这些变化信息的标签，隐式信息只能隐式获得这些变化信息。

2.1 显式信息

上述说话人信息、语言信息等往往可以作为显式信息，直接使用它们作为输入，以增强表现性语音合成的模型。对于这些显式信息，可以通过约束波形的韵律特征直观地进行控制。

首先可以从标签数据中获取语言ID、说话人ID、风格和韵律等显式特征。例如，韵律信息可以根据注释模式进行标记，自基于HMM的语音合成研究以来，韵律和说话风格建模一直在研究中。例如，Eyben F等人[46]提出了一种系统，该系统首先对训练集进行聚类，然后执行基于

表1 表现性语音合成信息的显式及隐式建模方法

信息种类	描述	模型
显式	语言ID、说话人ID	SRM2TTS[33]、Ye H等人[34]、Arik S O等人[35]
	音高/持续时间/能量	FastSpeech[24]、FastSpeech2[25]、Fastpitch[36]
隐式	参考编码器	Skerry-Ryan R J等人[37]、Gururani S等人[38]
	全局风格标记（global style token，GST）	GST-Tacotron[39]、Mellotron[40]
	变分自编码器（variational auto-encoder，VAE）	CHiVE[41]、GMVAE-Tacotron[42]
	生成对抗网络（generative adversarial network，GAN）	WaveGAN[43]、GAN-TTS[44]、MelGAN[45]

HMM的聚类自适应训练[47]。Rosenberg A[48]使用AuToBI标签改进基于HMM的合成。Morrison M等人[49]提出了一种用户可控、上下文感知的神经韵律生成器，该生成器允许输入特定时间帧的基频轮廓，并根据输入文本和上下文韵律生成剩余时间帧。汉语语音合成系统也对韵律进行预测，典型的韵律边界标签由韵律词（prosodic word，PW）、韵律短语（prosodic phrase，PPH）和语调短语（intonational phrase，IPH）组成，它们构成了一个3层韵律结构树，**图4**[50]展示了韵律结构树的示例。

除标签数据外，还可以从语音中直接提取音调和能量信息，并从成对的文本和语音数据中提取持续时间。FastSpeech2在FastSpeech的基础上提出差异适配器，旨在向音素隐藏序列中添加3种差异信息：音素持续时间，表示语音的持续时间；音高，传达情感的关键特征，对语音韵律有很大影响；能量，表示梅尔频谱的帧级量级，直接影响语音的音量和韵律。FastPitch[36]将基音预测网络添加到FastSpeech中，以控制基音。与FastSpeech和FastPitch相比，FastSpeech2引入了更多的风格特征，如基音、能量和更精确的持续时间，作为条件输入构建差异适配器，并使用经过训练的能量、基音和持续时间预测器合成具有特定风格的语音。

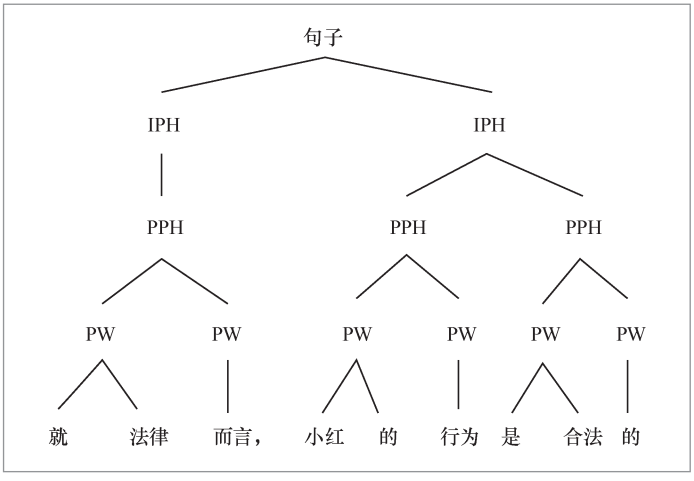

图 4 韵律结构树示例

2.2 隐式建模

在大多数情况下，语音数据集中并没有合适的可用的显式标签，一句话中的抑扬顿挫难以进行人为标记，同时显式地为庞大的语音数据集中的每一句话添加相应的标签会带来大量的人力成本，并且这些标签无法覆盖特定或细粒度的变化信息。因此，表现性语音合成的另一重要任务就是从数据中隐式地建模变化信息，以实现对这些难以标记信息的控制和分离。笔者将对不同的隐式建模方法进行相关模型的综述，主要包含基于参考编码器[37]、全局风格标记[40]、变分自编码器[41]、生成对抗网络等[43]等。

2.2.1 参考编码器

通过添加参考编码器合成表现性语音，可以引入风格信息。主要有两种基于参考编码器的方法可以用来合成具有特定风格的语音。第一种方法是使用经过训练的参考编码器直接控制各种语音风格参数，例如音调、响度和情绪。第二种方法是将参考音频输入参考编码器，并使用参考编码器编码的风格参数在参考语音和目标语音之间传输语音风格特征。人们提出了不同的方法和模型分离不同的风格特征信息，这样每个风格特征都可以很容易地单独控制，从而合成具有目标风格的语音。后文将介绍这些方法和模型。Skerry-Ryan R J等人[37]将语音特征分为3个部分：文本、说话人和韵律。在Tacotron中加入参

考编码器，从特定风格的参考语音中提取韵律嵌入，并使用说话人嵌入查找表获得说话人嵌入。然后将韵律嵌入、说话人嵌入和文本嵌入相结合，输入解码器合成具有参考语音风格的语音。参考编码器可以从参考语音中分离得到韵律嵌入，实现对韵律的整体建模，但是不能分离出韵律嵌入中具有代表性的韵律内容，如音调、持续时间等。Gururani S等人[38]在Skerry-Ryan R J等人的基础上对模型进行了改进，将语音的风格特征分为音调和响度，并选择两个时间序列分别对参考语音的基频和响度进行建模。日常对话中往往包含了很多情感信息，为了更准确地传递参考语音中的情感特征，Li T等人[51]在参考编码器和解码器之后分别添加了两个情感分类器，以增强情感空间中的情感分类能力。此外，他们采用了风格损失[52-53]测量生成的和参考梅尔频谱[54]之间的风格差异以实现将参考语音中的情感传递至生成的语音中。

2.2.2 全局风格标记

为了分离语音中的不同风格特征，并达到单独控制从参考编码器获得的韵律嵌入中的每个特征的目的，Wang Y等人[39]在Tacotron中引入了一个全局风格标记网络，如**图5**所示，该网络起到了聚类的作用。

当GST网络使用不同风格的语音数据进行训练时，可以获得多个有意义且可解释的标记，这些标记（A、B、C、D）就被称为全局风格标记。这些标记加权求和后被用作风格嵌入来控制和传递语音的风格特征。但GST方法的缺陷在于难以解释学习到的每个风格标记并赋予其实际意义，即无法分辨每个标记代表的具体风格。对于标记权重的选择，Kwon O等人[55]提出了一种基于控制权重的方法，通过研究情绪向量空间中每种情绪的分布定义权重值。Um

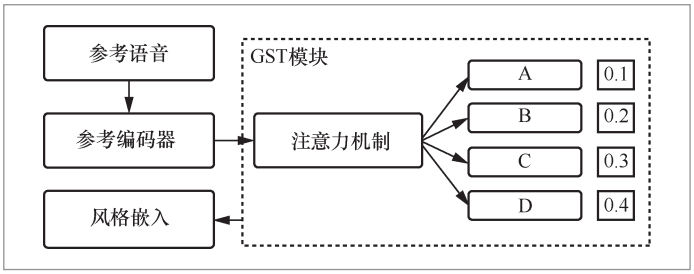

图 5 GST 模型

S Y等人[56]提出改进方法，简单地平均属于每个情感类别的风格嵌入向量[57]，通过最大化类别间距离与类别内距离的比率确定代表性权重向量，并提出用感知扩散的方法改变情绪强度，而不是简单地基于线性插值的方法。该方法通过对情绪强度建模控制实现了对情感的细粒度控制，可以对情绪的强度进行手动控制。Mellotron方法[40]还引入了基频信息，并将文本、说话人、基频、注意力映射和GST作为合成语音的条件，其中说话人代表音色，基频代表音高，注意力映射代表节奏，GST代表韵律。由于GST Tacotron仅使用成对的输入文本和参考语音进行训练，因此在合成过程中输入未配对的文本和语音将导致生成的声音变得模糊。导致这种情况的原因可能是参考编码器未能完全分离韵律信息而引入了一些文本信息。由于韵律迁移模型往往在训练时使用与输入文本具有相同文本的参考语音，而在推理中不同，训练和推理之间产生了差距。Liu D R等人[58]为了解决这个问题利用双重学习的思想，提出用不成对的文本和语音训练GST-Tacotron，并将输出的梅尔频谱输入语音识别模型以预测输入文本，从而防止参考编码器编码任何文本信息。为了更灵活地控制合成语音的多种风格特征，可以使用多参考编码器分别提取多参考语音的不同风格特征。例如，Bian Y[59]等人使用基于GST网络的多个参考编码器分离不

同的风格特征，并提出交叉训练技术，通过在每个编码器提取的风格之间引入正交约束分离风格潜在空间。然而，这种交叉训练方案并不能保证在训练过程中看到每一种风格类别的组合，从而错过了在不相交的数据集上学习风格的分离表示和次优结果的机会。Whitehill M等人[60]使用对抗性循环一致性训练方案，确保使用所有风格维度的信息，以应对Bian Y等人方法无法解决的不相交数据集上多参考风格转换的挑战。参考编码器及全局风格标记建模模型对比见**表2**。

2.2.3 变分自编码器

变分自编码器[61]最早在计算机视觉中被提出，从潜在变量的分布中采样生成具有特定特征的样本。隐变量是连续的，可以插值，类似于语音中的隐式风格特征。变分自编码器以无监督的方式学习的语音风格特征可以很容易地分离、缩放和组合。因此，有许多任务使用变分自编码器来控制合成语音的风格。变分自编码器在未观察到的连续随机潜在变量和观察到的数据集之间构建了一个关系，从而利用中间的潜在变量Z实现对X的建模。无法直接求得真正的后验密度 $p_\theta(Z|X)$ 导致了不可微的边缘似然 $p_\theta(X)$。为了解决这个问题，引入了 $q_\phi(Z|X)$ 近似无法直接求得的后验 $p_\theta(Z|X)$。根据变分原理，$\log p_\theta(X)$ 可改写为式（1），其中笔者希望优化变分下界 $L(\theta,\phi;X)$。

$$\log p\theta(x) = KL[q_\phi(Z|X)\,||\,p_\theta(Z|X)] + L(\theta,\phi;X) \quad (1)$$

$$\log p\theta(x) \geqslant L(\theta,\phi;X) = E_{q_\phi(Z|X)}[-\log p_\theta(X|Z) + \log p_\theta(X,Z)] \quad (2)$$

$$L(\theta,\phi;X) = E_{q_\phi(Z|X)}[\log p_\theta(X|Z)] - KL[q_\phi(Z|X)||p_\theta(Z)] \quad (3)$$

潜变量的先验 $p_\theta(X)$ 假设服从中心各向同性多变量高斯分布 $\mathcal{N}(Z;0,I)$，其中$\boldsymbol{I}$是单位矩阵。$q_\phi(Z|X)$ 的一般取值

表 2 参考编码器及全局风格标记建模模型对比

建模方法分类	具体建模方法	描述
参考编码器	Skerry-Ryan R J等人[37]	利用参考编码器实现对韵律信息的完整建模，缺点是不能实现韵律中信息的分离控制
	Gururani S等人[38]	将Skerry-Ryan R J等人整体建模的韵律嵌入分为音调和响度，分别建模，实现对音调和响度的控制
	Li T等人[51]	在Skerry-Ryan R J等人的基础上引入两个情感分类器进行对抗训练，以实现从参考语音中情感迁移
全局风格标记	GST-Tacotron[39]	将学习到的多个有意义且可解释的全局风格标记的加权作为风格嵌入，缺点是不能了解每个风格标记代表的具体风格
	Kwon O等人[55]	将GST-Tacotron中的风格标记加权映射到情感空间，利用加权和表示特定情感
	Um S Y等人[56]	在控制情感类别的基础上采用感知扩散方法控制每个情感的强度
	Mellotron[40]	引入了基频信息，并将文本、说话人、基频、注意力映射和GST作为合成语音的条件
	Liu D R等人[58]	存在在韵律迁移训练中使用相同文本的参考语音而推理时使用不同文本语音的问题，利用语音识别损失减小产生训练和推理之间的差距
	Bian Y等[59]	引入多参考编码器和正交约束实现多种韵律信息的分离解耦和分别建模，缺点是不能在不相交数据集上学习风格分离
	Whitehill M等人[60]	在Bian Y等人的基础上引入循环一致性损失，实现在不相交数据集上的说话人和情感迁移

为 $\mathcal{N}\left(Z;\mu(X),\sigma^2(X)I\right)$，从而闭合计算 $KL\left[q_\phi(Z|X)||p_\theta(Z)\right]$。在实际应用中，$\mu(X)$ 和 $\sigma^2(X)$ 是通过神经网络从数据集中学习的，此处的神经网络可以看作一个编码器。式(1)中的第一项也即期望项起到了解码器的作用，解码器对潜在变量进行解码来重构 X。如果解码器的输出在X和Z的多个样本上取平均，解码器就可以产生期望的重构。$E_{q_\phi(Z|X)}\left[\log p_\theta(X|Z)\right]$ 也被称为重构损失，$KL\left[q_\phi(Z|X)||p_\theta(Z)\right]$ 被称为KL损失。

当多种风格或韵律信息纠缠在一起时，要想更好地进行表达性语音合成和控制，有必要在训练过程中对它们进行分离。例如，CHiVE[41]是一个具有层次结构的条件VAE模型，它可以生成适合声码器使用的韵律特征，如基频、能量和持续时间，并生成一个韵律空间，从中可以对有意义的韵律特征进行采样。为了有效地捕获语言输入（单词、音节）的层次性，自动编码器的编码器和解码器部分都是层次的，与语言结构一致，各层都以各自的速率动态计时。Zhang Y J等人[62]在Tacotron2中添加了一个变分自编码器网络，以学习代表语音风格的潜在变量，潜在变量的每个维度代表不同的风格特征。为了进一步理清语音的各种风格特征，基于高斯混合变分自编码器网络的GMVAE-Tacotron[42]具有两个层次的潜在变量。第一个层次是一个离散的潜在变量，代表某种类型的风格（例如说话人ID、干净/嘈杂）；第二个层次是由多元高斯分布近似的连续潜变量。每个分量代表第一级类别下特征的度（例如噪声级、说话速率、音调）。该模型能有效地分解和独立控制语音信号的潜在属性。然而，这些方法只对语音的整体风格特征进行建模，没有考虑音素和单词层面的韵律控制。为了在不同分辨率下对声学特征进行建模，Sun G等人[63]除了对全局语音特征（如噪声和通道数）进行建模外，还对单词级和音素级韵律特征（如基频、能量和持续时间）进行了建模，使用具有自回归结构的条件变分自编码器来进行，每一层的韵律特征都更具可解释性，并在所有潜在维度上施加等级制约。

2.2.4 生成对抗网络

生成对抗网络（GAN）[64]已广泛用于数据生成任务，如图像生成[65]、文本生成[66]和音频生成[43]。GAN由一个用于数据生成的生成器和一个用于判断生成数据真实性的判别器组成，通过生成器和判别器的不断博弈提高建模能力。GAN模型如**图6**所示。

在语音领域，GAN的生成器主动生成梅尔频谱，生成虚假频谱“欺骗”判别器，而判别器需要不断提高判别能力甄别生成结果的真伪，在对抗过程中改善模型生成效果。GAN可以用于风格语音合成。

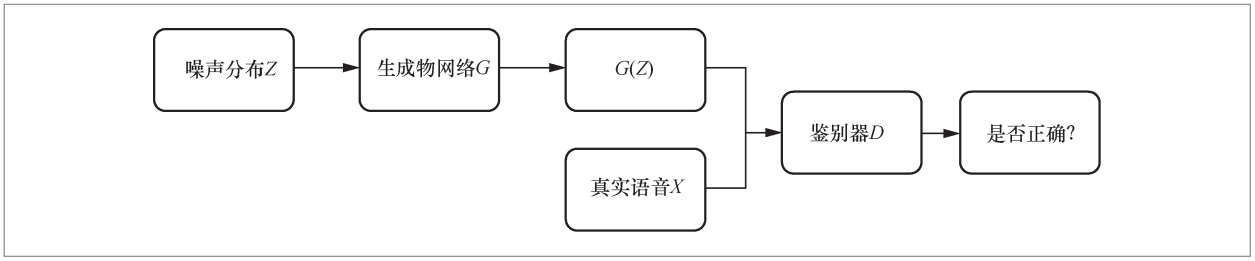

图 6 GAN 模型

例如，Ma S等人[67]将成对训练过程、对抗性游戏和协作性游戏组合到一个训练方案中，增强了模型的内容风格分离能力和可控性。对抗博弈集中真实的数据分布，协作博弈最小化原始空间和潜在空间中真实样本和生成样本之间的距离。由于单独的对抗性反馈不足以训练生成器，当前模型仍然需要比较直接生成的梅尔频谱和真值所得到的重建损失。Multi-SpectroGAN[68]可以通过将生成器的自监督隐藏表示条件化为条件鉴别器来训练仅具有对抗性反馈的多说话人模型，为训练生成器提供更好的指导。此外还提出了对抗式风格组合，以更好地概括数据集中未包括的说话风格和文本，它可以从多个梅尔频谱中学习组合风格嵌入的潜在表示。Multi-SpectroGAN通过对抗式风格组合和特征匹配进行训练，通过控制和混合不同的说话风格（如持续时间、音调和能量），合成高度多样性的梅尔图谱。GANtron利用生成对抗网络将情感作为文本到语音模型的输入，考虑6种不同的情绪（即愤怒、厌恶、恐惧、幸福、悲伤和中性），并提出了一种新的基于引导注意丢失的训练策略。同时指出未来可以对训练中的损失函数进行修改，在训练循环中集成一个情感分类器作为损失计算的一部分，类似于在计算机视觉领域对知觉损失所做的工作可能会带来潜在的改进。

2.2.5 其他网络

基于流的模型也被应用于表现性语音合成中，例如Flowtron[69]是一种基于自回归流的梅尔频谱生成模型，Flow TTS[70]和Glow TTS[71]利用生成流进行非自回归梅尔频谱生成。规范化流是一种生成模型，它用一系列可逆映射变换概率密度。可以通过基于变量变化规则的可逆映射序列得到标准的、规范化的概率分布（如高斯分布），这种基于流的生成模型被称为规范化流。在采样期间，它通过这些变换的逆运算从标准概率分布生成数据。Flowtron将规范化流应用于Tacotron，通过学习存储非文本信息的潜在空间来控制语音变化和风格转换。An X等人[72]采用逆自回归流（inverse autoregressive flow，IAF）改进变分推理和学习风格表示，以分离说话人及风格信息。

除了基于流的模型外，Diff TTS[73]是基于扩散模型的表现性语音合成的基础模型。其基本思想如**图7**所示，通过扩散过程和反向过程描述数据与潜在分布之间的映射，即在扩散过程中，波形数据样本逐渐加入一些随机噪声，梅尔频谱逐渐被高斯噪声破坏并转化为潜变量，最终成为高斯噪声。设 $x_1,\cdots,x_T$ 是相同维数的变量序列，其中 $t=0,1,\cdots,T$ 为扩散时间步长指数。然

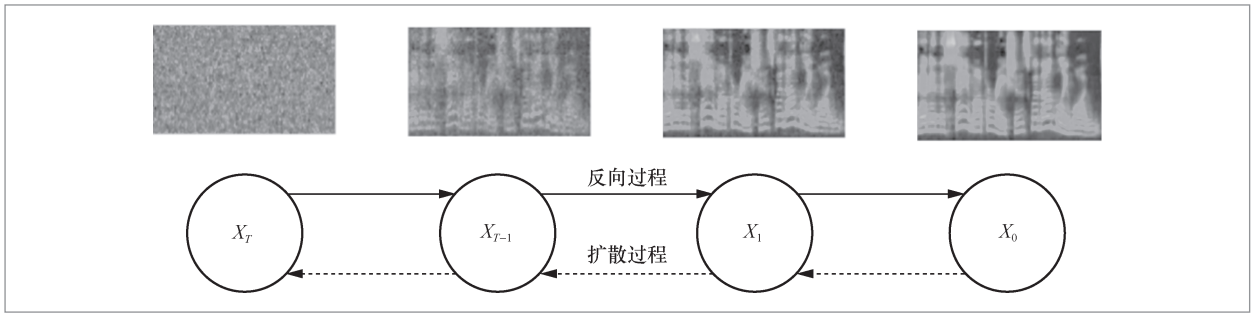

图 7 扩散及反向过程

后，扩散过程通过一系列马尔可夫变换将梅尔谱图 x_0 转化为高斯噪声 x_T。每个过渡步骤都预先定义了方差计划 $\beta_1,\beta_2,\cdots,\beta_T$，每一次变换都是按照假定独立于文本 c 的马尔可夫跃迁概率 $q(xt\,|\,xt-1,c)$ 进行的，其定义如式（4）所示：

$$q(x_t|x_t-1,c)=N\left(x_t;\sqrt{1-\beta_t}x_t-1,\beta_t I\right) \quad (4)$$

整个扩散过程 $q(x_1:T|x_0,c)$ 为马尔可夫过程，可分解为式（5）：

$$q(x_1,\cdots,x_T|x_0,c)=\prod_{t=1}^{T}q(x_t|x_t-1) \quad (5)$$

反向过程是一个梅尔频谱生成过程，正好是扩散过程的反向过程。与扩散过程不同，反向过程的目标是从高斯噪声中恢复出梅尔谱图。反向过程定义为条件分布 $p_\theta(x_0:T-1|x_T,c)$，可根据马尔可夫链性质分解为多个变换，如式（6）所示：

$$p_\theta(x_1,\cdots,x_{T-1}|x_T,c)=\prod_{t=1}^{T}p_0(x_{t-1}|x_t-1) \quad (6)$$

Diff-TTS学习了一个反向过程得到的模型分布 $p_\theta(x_0|c)$。设 $q(x_0|c)$ 为梅尔频谱分布，为了使模型能很好地近似 $q(x_0|c)$，反向过程的目的是使梅尔频谱的对数似然 $E_{\log q(x_0|c)}\left[\log p_0(x_0|c)\right]$ 最大化。由于 $p_0(x_0|c)$ 是难以求得的，可以利用参数化技巧来计算对数似然的封闭形式的变分下界。设 $\alpha_t=1-\beta_t$，$\overline{\alpha}_t=\prod_{t'=1}^{t}\alpha'_t$。Diff TTS的训练目标为最小化模型输出 $\epsilon_\theta(\cdot)$ 和高斯噪声 ϵ 之间的L1损失，如式（7）所示：

$$\min L(\theta)=E_{x_0,\epsilon,t}\left\|\epsilon-\epsilon_\theta\left(\sqrt{\overline{\alpha}_t}x_0+\sqrt{1-\overline{\alpha}_t}\epsilon,t,c\right)\right\|_1 \quad (7)$$

这是第一次将去噪扩散概率模型应用于非自回归语音合成。Diff TTS可以在不受模型结构约束的情况下进行稳定训练，在仅使用Tacotron2和Glow-TTS一半参数的情况下合成高质量的语音。高质量且高效的基于去噪扩散的概率模型可能在未来成为表现性语音合成的研究重点，会将韵律风格等特征与去噪扩散相结合以增添语音的表现力。

3 表现性语音合成任务

第2节中提到表现性语音合成需要考虑“谁说”和“如何说”。“谁说”和“如何说”分别对应表现性语音合成中的多说话人和韵律控制任务，本节进行具体介绍。此外，虽然表现性语音合成任务已经得到了广泛的研究，但目前现有的大多数表现性语料资源相对匮乏，无法有效地用于训练基于深度学习的表现性语音合成模型，需要采用语音增强等技术充分利用有限的表现性语料资源以实现低资源语音合成。

3.1 多说话人语音合成

多说话人语音合成是表现性语音合成的一项重要任务。合成多个说话人声音的一种简单方法是在输入中添加一个说话人嵌入向量[74]。说话人嵌入向量可以通过额外训练参考编码器获得。例如，Ye J等人[34]、Arik S O等人[35]分别在Tacotron2、Deep Voice 3[75]中引入说话人编码器，将参考语音中的说话人信息编码到固定维说话人嵌入向量中。嵌入向量只能从目标说话人的少量语音片段中提取。用于训练说话人编码器的语音数据语料库只需要包含大量说话人的录音，但不要求高质量。即使

训练数据中含有少量噪声，也不会影响音色特征的提取。然而这些方法在合成具有未包含在数据集中的未知说话人的语音时并不是很有效。为了解决这个问题，Cooper E等人[76]在Ye J等人的基础上，使用可学习字典编码（learnable dictionary encoding，LDE）提取说话人信息，并将嵌入Prenet层和Tacotron2注意网络中的说话人作为附加信息。在训练说话人编码器时，Nachmani E[77]等人除了使用均方误差（mean square error，MSE）损失外，还引入了对比度损失项和循环损失项，使模型能够用少量音频合成新说话人的声音。在训练说话人编码器时，除了MSE损失外，Nachmani E等人还提出了对比损失项和循环损失项，使模型能够用少量音频合成新说话人的声音。此外，一个说话者无论如何不可能模仿所有的说话风格并录制足够的录音。SRM2TTS[33]旨在通过将一个说话人的任何说话风格与另一个说话人的音色相结合，为了实现此任务，提出了一种基于显式韵律特征的风格建模方法。该方法以Tacotron2为主干，采用细粒度文本韵律预测模块和说话人控制器。该方法可以避开对单个说话人多风格语料库的依赖，解决了一个说话人无法表达所有说话风格的问题。

3.2 低资源表现性语音合成

近年来，表现性语音合成系统的合成效果非常好，但它们通常需要目标说话人以所需的说话风格进行大量的记录工作。Huybrechts G等人[78]提出了一种新的3步方法，以避免记录大量目标数据的昂贵操作，只需15 min的记录就可以构建出富有表现力的声音。首先通过语音转换增加数据，利用其他说话人以所需的说话风格录制的录音；然后在可用记录的基础上使用该合成数据训练TTS模型；最后对该模型进行微调，进一步提高质量。利用语音转换的数据增强已成功应用于低资源表现性语音合成。Terashima R等人[79]提出了一种结合基音偏移和语音转换技术的新的数据增强方法。由于基音偏移数据增强能够覆盖各种基音动态，因此它极大地稳定了语音转换和语音合成模型的训练。Ribeiro M S等人[80]通过语音转换的数据增强解决从文本到语音的跨说话人风格传输问题，该方法假设有一个来自目标说话人的中性的非表现性数据语料库，包含来自不同说话人的表现性数据。首先从不同说话人的表现性数据集中利用语音转换生成目标说话人的高质量数据；然后将语音转换数据与目标说话人的自然数据合并，用于训练单说话人多风格TTS系统。Shah R等人[81]在采用语音转换数据增强的基础上，将基于注意的自回归语音合成模型改为非自回归模型，用外部持续时间模型代替注意，并且增加基于条件生成对抗网络的微调步骤。Lajszczak M等人[82]通过解析文本和音频中的树成分替换创建新的训练样本，以解决扩充样本分布不均的问题。为语音合成引入一种新的数据增强技术，显著增加模型中文本条件的多样性，这是分布增强技术在基于神经网络的语音合成中的首次应用，该方法同时减少了模型对输入文本的过度拟合。

4 表现性语音合成研究方向展望

基于以上对表现性语音合成方法的介绍和总结，可以预测未来语音合成领域至少会有如下发展方向。

（1）在表现性语音合成中以精确、精细的方式控制合成语音的风格。在谈话中，情绪、语调和节奏等讲话风格经常发生变化。然而，目前的神经语音合成系统无法单独精确地控制语音的这些风格特征。如何在词级和短语级上实现语音的细粒度风格控制是未来语音合成研究的重点。

（2）数据高效的语音合成。在表现性语音合成中由于情感语音数据难以记录和标注，如何有效地利用数量和质量有限的情感语音数据来训练语音合成模型，使其能够学习语音中各种风格特征的表示方法，也是语音合成领域亟待解决的问题。此外，许多低资源语言缺乏训练数据。如何利用无监督/半监督学习和跨语言迁移学习来帮助低资源语言是一个有趣的方向。

（3）强大的生成模型。语音合成是一项生成波形和/或声学特征的生成任务，表现性语音合成以这些生成模型为基础。功能强大的生成模型可以更好地处理这些特征。尽管基于VAE、GAN的高级生成模型已被用于声学模型、声码器和完全端到端模型，但对更强大、更高效的生成模型的研究工作正在吸引人们进一步提高合成语音的质量。

（4）语音合成和图像生成有很大的相似性，可以将其他任务中使用的深度学习方法应用到语音合成中作为生成任务。语音合成中使用的许多方法受到图像生成方法的启发，生成具有特定风格的图像和语音的方法也非常相似。其次，由于识别和生成是双重任务，可以采用多任务学习将识别和生成模型结合起来，以相互改进，减少训练过程中对标记数据的需求。除了结合语音合成与语音识别之外，还可以将说话人识别与多说话人语音合成相结合，并将语音情感识别与情感语音合成相结合用于双重训练。

参考文献:

[1] COKER C H. A model of articulatory dynamics and control[J]. Proceedings of the IEEE, 1976, 64(4): 452-460.

[2] CAPES T, COLES P, CONKIE A, et al. Siri on-device deep learning-guided unit selection text-to-speech system[C]// Proceedings of Interspeech 2017. [S.l.:s.n.], 2017.

[3] GONZALVO X, TAZARI S, CHAN C A, et al. Recent advances in google real-time HMM-driven unit selection synthesizer[C]//Proceedings of Interspeech 2016. [S.l.:s.n.], 2016.

[4] HUNT A J, BLACK A W. Unit selection in a concatenative speech synthesis system using a large speech database[C]// Proceedings of 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. Piscataway: IEEE Press, 2002: 373-376.

[5] MOULINES E, CHARPENTIER F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones[J]. Speech Communication, 1990, 9(5/6): 453-467.

[6] ZEN H, NOSE T, YAMAGISHI J, et al. The HMM-based speech synthesis system (HTS)[J]. SSW, 2007, 6: 294-299.

[7] SAITO Y, TAKAMICHI S, SARUWATARI H. Statistical parametric speech synthesis incorporating generative adversarial networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 26(1): 84-96.

[8] NOSE T, NOSE T, NOSE T. Efficient implementation of global variance compensation for parametric speech synthesis[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing,

2016, 24(10): 1694–1704.

[9] KAWAHARA H, MORISE M, TAKAHASHI T, et al. Tandem-STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation[C]//Proceedings of 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2008: 3933–3936.

[10] CHEN L H, RAITIO T, VALENTINI-BOTINHAO C, et al. A deep generative architecture for postfiltering in statistical parametric speech synthesis[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(11): 2003–2014.

[11] LADD D R. Intonational phonology[M]. Cambridge: Cambridge University Press, 1996.

[12] WATSON D G, WAGNER M, GIBSON E. Experimental and theoretical advances in prosody: a special issue of language and cognitive processes[M]. [S.l.:s.n.],2012.

[13] YAN Y, TAN X, LI B, et al. AdaSpeech 3: adaptive text to speech for spontaneous style[EB]. arXiv preprint, 2021, arXiv: 2107.02530.

[14] HONG Y W, CHO S J, KIM J M, et al. Formant synthesis of Haegeum sounds using Cepstral envelope[J]. The Journal of the Acoustical Society of Korea, 2009, 28: 526–533.

[15] KHORINPHAN C, PHANSAMDAENG S, SAIYOD S. Thai speech synthesis with emotional tone: based on Formant synthesis for Home Robot[C]//Proceedings of 2014 3rd ICT International Student Project Conference. Piscataway: IEEE Press, 2014: 111–114.

[16] KLATT D H. Software for a cascade/parallel formant synthesizer[J]. The Journal of the Acoustical Society of America, 1980, 67(3): 971–995.

[17] VOGTEN L, BERENDSEN E. From text to speech: the MITalk system[J]. Journal of Phonetics, 1988, 16(3): 371–375.

[18] YOSHIMURA T, TOKUDA K, MASUKO T, et al. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis[C]//Proceedings of 6th European Conference on Speech Communication and Technology. [S.l.:s.n.], 1999.

[19] FUKADA T, TOKUDA K, KOBAYASHI T, et al. An adaptive algorithm for mel-cepstral analysis of speech[C]//Proceedings of 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE Press, 2002: 137–140.

[20] IMAI S, SUMITA K, FURUICHI C. Mel log spectrum approximation (MLSA) filter for speech synthesis[J]. Electronics and Communications in Japan (Part I: Communications), 1983, 66(2): 10–18.

[21] WANG Y X, SKERRY-RYAN R J, STANTON D, et al. Tacotron: towards end-to-end speech synthesis[C]//Proceedings of 2017 Interspeech. [S.l.:s.n.], 2017.

[22] SHEN J, PANG R M, WEISS R J, et al. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2018: 4779–4783.

[23] LI N H, LIU S J, LIU Y Q, et al. Neural speech synthesis with transformer network[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 6706–6713.

[24] REN Y, RUAN Y J, TAN X, et al. FastSpeech: fast, robust and controllable text to speech[EB]. arXiv preprint, 2019, arXiv: 1905.09263.

[25] REN Y, HU C X, TAN X, et al. FastSpeech 2: fast and high-quality end-to-end text to speech[EB]. arXiv preprint, 2020, arXiv: 2006.04558.

[26] ITAKURA F. Line spectrum representation of linear predictor coefficients of speech signals[J]. The Journal of the Acoustical Society of America, 1975, 57(S1): S35.

[27] FUKADA T, TOKUDA K, KOBAYASHI T, et al. An adaptive algorithm for mel-cepstral analysis of speech[C]// Proceedings of 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE Press, 2002: 137-140.

[28] TOKUDA K, KOBAYASHI T, MASUKO T, et al. Mel-generalized cepstral analysis - a unified approach to speech spectral estimation[C]//Proceedings of 3rd International Conference on Spoken Language Processing. [S.l.:s.n.], 1994.

[29] KAWAHARA H, MASUDA-KATSUSE I, DE CHEVEIGNÉ A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds[J]. Speech Communication, 1999, 27(3/4): 187-207.

[30] KAWAHARA H, ESTILL J, FUJIMURA O. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT[J]. Second International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, 2001.

[31] KAWAHARA H. STRAIGHT, exploitation of the other aspect of VOCODER: perceptually isomorphic decomposition of speech sounds[J]. Acoustical Science and Technology, 2006, 27(6): 349-353.

[32] MORISE M, YOKOMORI F, OZAWA K. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications[J]. IEICE Transactions on Information and Systems, 2016, E99.D(7): 1877-1884.

[33] XIE Q C, LI T, WANG X S, et al. Multi-speaker multi-style text-to-speech synthesis with single-speaker single-style training data scenarios[C]// Proceedings of 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP). Piscataway: IEEE Press, 2023: 66-70.

[34] JIA Y, ZHANG Y, WEISS R J, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM, 2018: 4485-4495.

[35] ARIK S O, CHEN J, PENG K, et al. Neural voice cloning with a few samples[EB]. arXiv preprint, 2018, arXiv: 1802.06006.

[36] ŁAŃCUCKI A. Fastpitch: parallel text-to-speech with pitch prediction[C]// Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2021: 6588-6592.

[37] SKERRY-RYAN R, BATTENBERG E, XIAO Y, et al. Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron[EB]. arXiv preprint, 2018, arXiv: 1803.09047.

[38] GURURANI S, GUPTA K, SHAH D, et al. Prosody transfer in neural text to speech using global pitch and loudness features[EB]. arXiv preprint, 2019, arXiv: 1911.09645.

[39] WANG Y, STANTON D, ZHANG Y, et al. Style tokens: unsupervised style modeling, control and transfer in end-to-end speech synthesis[J]. arXiv preprint, 2018, arXiv:

1803.09017.

[40] VALLE R, LI J, PRENGER R, et al. Mellotron: multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens[C]//Proceedings of2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2020: 6189-6193.

[41] WAN V, CHAN C, KENTER T, et al. CHiVE: varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network[EB]. arXiv preprint, 2019, arXiv: 1905.07195.

[42] HSU W N, ZHANG Y, WEISS R J, et al. Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization[C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2019: 5901-5905.

[43] DONAHUE C, MCAULEY J, PUCKETTE M. Adversarial audio synthesis[EB]. arXiv preprint, 2018, arXiv: 1802.04208.

[44] BIŃKOWSKI M, DONAHUE J, DIELEMAN S, et al. High fidelity speech synthesis with adversarial networks[EB]. arXiv preprint, 2019, arXiv: 1909.11646.

[45] KUMAR K, KUMAR R, DE BOISSIERE T, et al. MelGAN: generative adversarial networks for conditional waveform synthesis[EB]. arXiv preprint, 2019, arXiv: 1910.06711.

[46] EYBEN F, BUCHHOLZ S, BRAUNSCHWEILER N, et al. Unsupervised clustering of emotion and voice styles for expressive TTS[C]// Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2012: 4009-4012.

[47] ZEN H, NOSE T, YAMAGISHI J, et al. The HMM-based speech synthesis system[J]. IEICE Technical Report Natural Language Understanding & Models of Communication, 2007.

[48] ROSENBERG A. AutoBI - a tool for automatic toBI annotation[C]// Proceedings of the 2010 Interspeech. [S.l.:s.n.], 2010.

[49] MORRISON M, JIN Z, SALAMON J, et al. Controllable Neural Prosody Synthesis[C]//Proceedings of the 2020 Interspeech. [S.l.:s.n.], 2020.

[50] SUN J W, YANG J, ZHANG J P, et al. Chinese prosody structure prediction based on conditional random fields[C]// Proceedings of 2009 5th International Conference on Natural Computation. Piscataway: IEEE Press, 2009: 602-606.

[51] LI T, YANG S, XUE L M, et al. Controllable emotion transfer for end-to-end speech synthesis[C]//Proceedings of 2021 12th International Symposium on Chinese Spoken Language Processing. Piscataway: IEEE Press, 2021: 1-5.

[52] JOHNSON J, ALAHI A, LI F F. Perceptual losses for real-time style transfer and super-resolution[EB]. arXiv preprint, 2016, arXiv: 1603.08155.

[53] GATYS L, ECKER A, BETHGE M. A neural algorithm of artistic style[J]. Journal of Vision, 2016, 16(12).

[54] GATYS L A, ECKER A S, BETHGE M. Image style transfer using convolutional neural networks[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2414-2423.

[55] KWON O, JANG I, AHN C, et al. An effective style token weight control technique for end-to-end emotional speech synthesis[J]. IEEE Signal Processing Letters, 2019, 26(9): 1383-1387.

[56] UM S Y, OH S, BYUN K, et al. Emotional speech synthesis with rich and granularized control[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2020: 7254-7258.

[57] KWON O, SONG E, KIM J M, et al. Effective parameter estimation methods for an ExcitNet model in generative text-to-speech systems[EB]. arXiv preprint, 2019, arXiv: 1905.08486.

[58] LIU D R, YANG C Y, WU S L, et al. Improving Unsupervised Style Transfer in end-to-end Speech Synthesis with end-to-end Speech Recognition[C]//Proceedings of 2018 IEEE Spoken Language Technology Workshop. Piscataway: IEEE Press, 2019: 640-647.

[59] BIAN Y Y, CHEN C B, KANG Y G, et al. Multi-reference Tacotron by intercross training for style disentangling, transfer and control in speech synthesis[EB]. arXiv preprint, 2019, arXiv: 1904.02373.

[60] WHITEHILL M, MA S, MCDUFF D, et al. Multi-reference neural TTS stylization with adversarial cycle consistency[C]//Proceedings of Interspeech 2020. [S.l.:s.n.], 2020.

[61] KINGMA D P, WELLING M. Auto-encoding variational bayes[EB]. arXiv preprint, 2014, arXiv: 1312.6114.

[62] ZHANG Y J, PAN S F, HE L, et al. Learning latent representations for style control and transfer in end-to-end speech synthesis[C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2019: 6945-6949.

[63] SUN G Z, ZHANG Y, WEISS R J, et al. Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2020: 6264-6268.

[64] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. New York: ACM Press, 2014: 2672-2680.

[65] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2242-2251.

[66] YU L T, ZHANG W N, WANG J, et al. SeqGAN: sequence generative adversarial nets with policy gradient[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1).

[67] MA S, MCDUFF D, SONG Y L. Neural TTS stylization with adversarial and collaborative games[C]//Proceedings of 2019 International Conference on Learning Representations, [S.l.:s.n.], 2019.

[68] LEE S H, YOON H W, NOH H R, et al. Multi-SpectroGAN: high-diversity and high-fidelity spectrogram generation with adversarial style combination for speech synthesis[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(14): 13198-13206.

[69] VALLE R, SHIH K, PRENGER R, et al. Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis[EB]. arXiv preprint, 2020, arXiv: 2005.05957.

[70] MIAO C F, LIANG S, CHEN M C, et al. Flow-TTS: a non-autoregressive network for text to speech based on flow[C]//Proceedings of2020 IEEE International Conference on Acoustics, Speech and

Signal Processing. Piscataway: IEEE Press, 2020: 7209-7213.

[71] KIM J, KIM S, KONG J, et al. Glow-TTS: a generative flow for text-to-speech via monotonic alignment search[EB]. arXiv preprint, 2020, arXiv: 2005.11129.

[72] AN X C, SOONG F K, XIE L. Disentangling style and speaker attributes for TTS style transfer[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2022, 30: 646-658.

[73] JEONG M, KIM H, CHEON S J, et al. Diff-TTS: a denoising diffusion model for text-to-speech[C]//Proceedings of Interspeech 2021. [S.l.:s.n.], 2021.

[74] ARIK S Ö, DIAMOS G, GIBIANSKY A, et al. Deep voice 2: multi-speaker neural text-to-speech[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 2966-2974.

[75] WEI P, PENG K N, GIBIANSKY A, et al. Deep Voice 3: 2000-speaker neural text-to-speech[EB]. arXiv preprint, 2017, arXiv:1710.07654.

[76] COOPER E, LAI C I, YASUDA Y, et al. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2020: 6184-6188.

[77] NACHMANI E, POLYAK A, TAIGMAN Y, et al. Fitting new speakers based on a short untranscribed sample[EB]. arXiv preprint, 2018, arXiv: 1802.06984.

[78] HUYBRECHTS G, MERRITT T, COMINI G, et al. Low-resource expressive text-to-speech using data augmentation[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2021: 6593-6597.

[79] TERASHIMA R, YAMAMOTO R, SONG E, et al. Cross-speaker emotion transfer for low-resource text-to-speech using non-parallel voice conversion with pitch-shift data augmentation[C]//Proceedings of Interspeech 2022. [S.l.:s.n.], 2022.

[80] SAM RIBEIRO M, ROTH J, COMINI G, et al. Cross-speaker style transfer for text-to-speech using data augmentation[C]//Proceedings of ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2022: 6797-6801.

[81] SHAH R, POKORA K, EZZERG A, et al. Non-autoregressive TTS with explicit duration modelling for low-resource highly expressive speech[C]//Proceedings of 11th ISCA Speech Synthesis Workshop. [S.l.:s.n.], 2021.

[82] LAJSZCZAK M, PRASAD A, VAN KORLAAR A, et al. Distribution augmentation for low-resource expressive text-to-speech[C]//Proceedings of ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2022: 8307-8311.

作者简介

唐浩彬（1999- ），男，中国科学技术大学硕士生，平安科技（深圳）有限公司算法工程师，主要研究方向为人工智能、语音识别和语音合成等。

张旭龙（1988– ），男，博士，平安科技（深圳）有限公司高级算法研究员，主要研究方向为语音合成、语音转换、音乐信息检索、机器学习和深度学习方法在人工智能领域应用。

王健宗（1983– ），男，博士，平安科技（深圳）有限公司副总工程师，资深人工智能总监，联邦学习技术部总经理。美国佛罗里达大学人工智能博士后，中国计算机学会高级会员，中国计算机学会大数据专家委员会委员，主要研究方向为联邦学习和人工智能等。

程宁（1981– ），男，博士，平安科技高级专家算法研究员，中国科学院软件所高级工程师，主要研究方向为语音识别、语音合成、自然语言处理等。

肖京（1972– ），男，博士，中国平安集团首席科学家，2019年吴文俊人工智能杰出贡献奖获得者，中国计算机学会深圳分部副主席，主要研究方向为计算机图形学学科、自动驾驶、3D显示、医疗诊断、联邦学习等。

收稿日期： 2022-11-29
通信作者： 王健宗，jzwang@188.com
基金项目： 广东省重点领域研发计划“新一代人工智能”重大专项（No.2021B0101400003）
Foundation Item: The Key Research and Development Program of Guangdong Province (No.2021B0101400003)