情感语音合成综述

施昊翔[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
作为语音领域一个重要的研究方向，语音合成致力于将文本转化为语音。随着深度学习技术的快速发展，语音合成的目的早已不仅仅是合成一段“能听懂”的音频这么简单，情感的加入往往能使语音变得更加具有表现力。基于此，情感语音合成在语音中加入不同的情感并对情感进行调控，以生成灵活且准确的情感语音。从情感语音合成中的几个关键科学问题出发，分别对近几年来基于情感迁移、情感强度控制和情绪混合的发展进行了总结分析，并介绍了情感语音合成的相关数据集和评价指标，最后对情感语音合成进行了展望。

关键词
情感语音合成；情感迁移；情感强度；深度学习

中图分类号：TP391 文献标志码：A doi: 10.11959/j.issn.2096-0271.2024014

A survey of emotional speech synthesis

SHI Haoxiang[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
As a significant research area in the field of speech technology, speech synthesis is dedicated to converting text into speech. With the rapid development of deep learning technology, the objective of speech synthesis has evolved beyond merely producing "understandable" audio. The incorporation of emotion often enhances the expressiveness of synthesized speech. Consequently, emotional speech synthesis aims to combine speech with different emotions and regulate these emotions to generate flexible and precise emotional speech. Starting from several key issues in emotional speech synthesis, this paper summarizes and analyzes the development based on emotion transfer, emotion intensity control and emotion mixing in recent years, and introduces the relevant data sets and evaluation indicators of emotion speech synthesis. Finally, the emotional speech synthesis is prospected.

Key words
emotional speech synthesis, emotion transfer, emotion intensity, deep learning

0 引言

语音合成是一种将文本转换为语音的技术，是人工智能的子领域之一，赋予机器像人一样自如说话能力的技术，是人机语音交互中重要的一环。随着深度学习的发展，基于神经网络的语音合成有了蓬勃的发展，近几年来大量研究工作集中在语音合成的不同方面，例如优化语音合成系统组件、解决语音合成的前沿热点问题等，这一系列的研究使近年来合成语音的质量有了很大的提高[1-5]。

随着语音合成技术的蓬勃发展，现在基于序列到序列的模型可以产生比较自然的声音，这些模型从大型语音数据集中学习中性说话风格，但人类的语音本质上是表达性的，从文本中传递准确、可控的表达性语音在人机交互和音频内容生成方面有着广泛的应用，适当的表达渲染可以影响整体的语音感知，因此人们对如何使用这些模型来传递富有表现力的语音越来越感兴趣[6-10]。近年来备受关注的是情感语音合成领域[11-16]，该领域的侧重点是情感表达的呈现。情感语音合成的核心在于利用深度学习模型和神经网络来模拟人类语音的特征和情感表达方式。传统的语音合成系统通常只能生成单调机械的声音，缺乏情感色彩。相比之下，情感语音合成通过语音数据训练，使合成的声音更自然、流畅，同时具备了丰富的情感维度。这使得我们能够为声音赋予多种情感，例如喜悦、悲伤、愤怒等。这种特性让合成的声音更能与人产生共鸣，从而提升了交流的真实感和情感连接。另外，情感语音合成还具备个性化定制的能力，能根据特定场景或用户需求进行声音定制，使合成声音更具针对性。这在诸如虚拟助手、客服系统等应用场景中尤为重要，情感语音能够提供更符合用户需求的交互体验。总之，情感语音合成旨在提取并学习语音中蕴含的情绪信息，比如情绪类别和情感强度，这对于合成具有丰富表现力的语音至关重要。

目前情感语音合成系统主要关注两个问题，一个是生成情感的准确性，另一个是情感控制的灵活性。具体来说，前者需要生成语音所传达的情感是可辨别的，听者应该很容易理解而不会混淆。而后者需要在保证情感准确性的基础上，通过强度调控生成不同情感强度的语音。例如，要生成“我今天出去玩了”这段话，以往的语音合成系统会机械地生成中性语音，而情感语音合成会将情感信息加入进来，表达出或喜悦或惊讶或悲伤的情感，同时，还可以控制情感强度，是非常喜悦还是一般喜悦，这可以极大地提升生成语音的多样性。

1 情感语音合成

目前情感语音合成系统基本采用序列到序列模型，起初带有注意力机制的序列到序列模型首先在机器翻译中得到研究[17]，后来在语音合成中被发现是有效的[18]。这是因为序列到序列模型在对单词、短语和话语等不同时间层次的长期依赖关系进行建模方面更有效[19]。通过学习注意对齐，序列到序列模型可以捕获话语中的动态韵律变化。它们还允许在运行时预测语音持续时间，这是语音情绪的关键韵律因素。基于序列到序列的模型一般采用两种方法来模拟语音情绪：基于显性标签的方法[20-21]和基于参考音频的方法[22-23]。使用显性情绪标签来表征情绪是最直接的方法，模型学习将标签与情绪风格相关联。Lee等人[20]基于注意力的解码器采用情绪

标签向量来产生所需的情绪，Tits等人[21]则使用带有少量情感标签的模型实现了低资源的情感文本到语音的转换。

由于情感语音往往缺乏大量的情感风格标签进行学习，因此，情感语音合成最广泛的方法是利用参考音频对其情感风格进行迁移[22-28]。Wang等人[8]在参考编码器方法的基础上，提出了全局风格令牌（global style tokens，GST），以无监督的方式学习可解释的风格嵌入。利用GST，该模型可以模仿参考音频的风格，并通过选择特定的标记来控制合成语音的风格，但是GST对于超出训练范围的风格缺乏泛化能力，这可能导致在对训练数据之外的参考音频进行迁移时表现不佳。Li等人[23]以Tacotron[1]框架为基础，提出了一种基于参考的情感语音合成方法，采用参考编码器学习一个情感嵌入空间，进而指导Tacotron[1]进行情感语音合成。该模型通过在参考编码器后连接一个情感分类器来增强情感空间的情感识别能力，同时使用风格损失来约束生成梅尔频谱和参考梅尔频谱之间的风格差异。其他一些研究使用变分自编码器（variational auto-encoder，VAE）[29]，通过学习、缩放或组合解耦表示来控制语音风格[30-31]。

除此之外，在进行情感风格迁移的同时，控制情感强度近年来也受到了广泛关注。情感强度被认为与所有有助于言语情绪的声音线索相关[33]，这使得它更加主观，并且具有挑战性。一些研究使用发声、不发声和沉默状态（voiced，unvoiced and silence，VUS）等辅助特征[33]、注意权重或显著性图[34]来控制情绪强度，其他研究则通过插值[15]、缩放[35]或基于距离的量化[36]来操纵内部情绪表征。一些工作[37-38]则引入了相对属性[39]来学习情绪强度表示，通过相对属性排序方法自动学习低水平声学特征与不同强度的高水平情感表达之间的关系，用得到的排序结果代表情感样本的情感强度，进而指导相关的语音合成模型（例如FastSpeech 2[40]等）合成不同情感强度的语音。

受限于情感数据集的情感类别，大多数情感语音合成模型只能生成几个数据集中现有的情感，例如快乐、惊讶、悲伤等，但人类的情绪类别是多种多样的，这极大程度上限制了生成语音的情感多样性，受此启发，Zhou等人[41]基于情绪论理论提出情绪混合来丰富语音合成的情感类别，Tang等人[42]利用扩散模型生成多种情绪的混合语音，同样丰富了情感类别。

本节从情感迁移、情感强度控制和情绪混合3个角度来详细介绍情感语音合成的发展，同时根据不同角度的情感语音合成方法，见**表1**。

1.1 情感迁移

随着序列到序列建模体系结构的快速发展，基于参考的风格迁移已经成为一种具有巨大潜力的解决方案。情感语音合成模型可以利用参考音频中的情感信息来控制生成语音的情感风格。通过无监督学习表达性音频样本构建风格嵌入空间，并利用风格嵌入调节语音合成模型，即使参考音频和合成音频的说话人不同，也可以生成与参考音频韵律相匹配的合成音频。最近基于情感迁移模型范式已经有了很多的研究[22,43]，例如全局风格令牌（GST）[8,15]、变分自动编码器（VAE）[8]及其变体[44-45]。然而，在合成音频中传递的情感往往是单调的，很难选择一个适当的参考来传递期望的情感。为了解决准确传递情感的问题，Lei等人[23]提出了一种基于Tacotron2[2]框架的基于参考的情感语音合成方法。具体来说，该模型采用参考编码器学习一个情感嵌入空间，利用Tacotron2[2]在情感嵌入

表1 不同技术路线的情感语音合成方法

合成方法分类	文献作者	时间	技术路线
基于显性标签	Lee等人[20]	2017年	基于注意力的解码器采用情绪标签向量来产生所需的情绪
	Noé Tits等人[21]	2019年	通过使用小型情感数据集微调中性TTS模型来生成情感语音
基于参考音频	Wang等人[8]	2018年	以无监督的方式学习可解释的全局风格嵌入
	Skerry-Ryan等人[22]	2018年	学习韵律的潜在嵌入空间，该空间自包含所需韵律的参考声学表示
	Li 等人[23]	2021年	采用参考编码器学习一个情感嵌入空间
	Lei等人[24]	2022年	引入了一种新颖的多参考结构，并提出了交叉训练方法，确保了多参考编码器的每个子编码器独立解耦并控制特定的风格，以生成准确的情感风格语音
情感强度控制	Matsumoto等人[33]	2020年	之前的工作使用情感ID进行语音合成，本文使用有声状态、无声状态和沉默状态（VUS）作为辅助特征。生成3种类型的情感言语，即中性、愤怒和快乐
	Schnell等人[34]	2021年	情绪识别器能够通过注意力或显著性来测量情绪强度，并标记情感强度用于后续训练
	Se-Yun Um等人[26]	2020年	引入了一种有效的插值技术，使目标情绪的强度逐渐转变为中性语音的强度
	Chae-Bin Im等人[36]	2022年	通过基于距离的强度量化来呈现，无须人工标记
	Zhu等人[37]	2019年	引入相对属性的方法，自监督提取语音中的情感强度信息，进而用于情感语音合成
	Wang等人[38]	2023年	利用中性情感样本和非中性情感样本进行混合排序，提取类内和类间强度信息，提高了情感细粒度，同时利用参考音频和候选池机制控制情感语音合成
情绪混合	Zhou等人[41]	2022年	提出了一个新的算法来衡量不同情绪的语音样本之间的相对差异，通过手动定义情感属性向量来控制模型以产生所需的混合情感
	Tang等人[42]	2023年	基于扩散概率模型和语音情感识别模型，可以生成具有指定强度或混合情绪的情感语音

的条件下进行情感迁移。模型结构如**图1**所示。该模型首先将输入的中文文本转换为字符序列。编码器由两个完全连接层的前置网络和CBHG[46]模块组成。CBHG模块将预计输出转换为最终编码器表示，然后经过注意力机制模块。解码器是一种自回归递归神经网络，其中一堆具有垂直残差连接的循环门控单元在每个解码器中输出音频特征进入注意力机制模块。最后使用基于CBHG的后置网络将梅尔尺度谱图转换为线性谱图，通过多波段WaveRNN重构波形[4]。同时设计了一个网络来构建

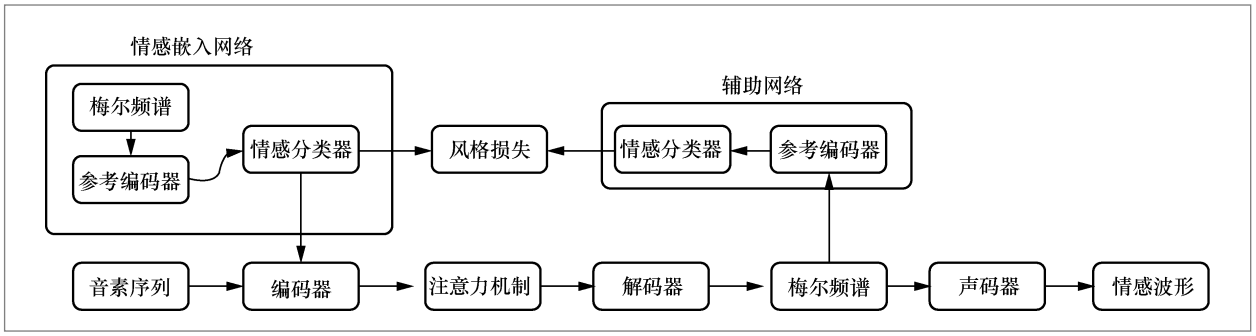

图1 基于风格损失的情感迁移模型

一个情感嵌入空间，该空间从参考音频样本中学习，并在推理过程中进行情感迁移。该模型主要参考了风格损失来测量生成语音和参考语音之间的情感风格差异，以保证前后情感风格的一致性。风格损失[14]最初是在计算机视觉领域中提出的，使用卷积神经网络（convolutional neural networks，CNN）生成的特征映射出Gram矩阵来捕捉图像的艺术风格，其中Gram矩阵以位置不变的方式计算特征统计，例如纹理。后来Gram矩阵被用于测量音频信号的梅尔谱图[44]，目的是捕捉音频信号在频域的局部统计量。人们认为，Gram矩阵能够表示语音的低水平特征，如响度、重音、速度、音高等，这些特征与情绪表达高度相关。该模型中的情感嵌入是CNN输出序列的集合，可以很自然地将其视为梅尔谱图的特征映射。特征映射的每个值都来自特定滤波器在目标位置的卷积，其本质是特征的提取和量化，因此每个值都可以看作情感相关特征的强度。

1.2 情感强度控制

目前还没有带情感强度标签的情感语音数据集，这意味着我们无法直接从数据集中利用强度标签进行强度监督学习。在保证情感传递准确性的同时，情感控制的灵活性依然是一个很大的挑战。研究表明，情绪在本质上是相对的，对情绪的注释和分析应遵循顺序路径[47-49]。排序方法不是给出一个绝对分数或一个情绪类别，而是通过比较来描述情绪（例如，第一句话比第二句更快乐吗?），序数方法表现出了显著的性能，特别是在语音情感识别方面[50-52]。排序方法的关键思想是根据给定的标准学习排序，类似偏好学习[53]，其任务是建立样本之间的偏好。一旦确定了偏好，对样本进行排序[54-56]就很简单了。其他基于排序的方法[57-59]也显示了对语音情感识别的影响建模的有效性。在情绪语音合成方面，研究人员也探索了情绪的序数性质，尝试对情绪强度进行序数建模，其中Zhu等人[37]将情绪强度视为中性样本与情绪样本之间的相对差异，实现了语音合成中情感强度的控制。近两年情感强度控制方向有很多新的方法，比如广泛使用的相对属性排序方法[37]，以及利用mixup数据增强进行对比式学习排序顺序[38]。

1.2.1 相对属性排序

情感强度控制语音合成的目标是在一个特定的情绪类别（如快乐）中合成具有不同感知强度的情绪言语。但在每个类别中只有小部分情感语音样本，手头没有强度标签，同时还有一组中性情绪的样本。为每个情感样本分配一个强度标签（或根据强度对每个样本进行排序），并训练一个提取情感强度信息的语音合成模型是一个可行的办法。Zhu等人[37]提出了一种方法，运用相对属性[39]的原理，即利用中性和情绪化样本之间的差异，从少量具有属性相对排序的成对示例中学习排序函数。在情感强度控制中，该属性为言语情感，如快乐。然后使用学习到的排序函数来估计一个分数作为未见语音的情感强度。这种排序函数的学习过程可以表述为一个最大边际优化问题。假设有一个在音频特征空间中的训练集 T ， $T=N\cup H$， N 和 H 分别为中性和快乐情绪样本集。针对提取的语音特征 x_t ，构造排序函数 $r(x_t)=\boldsymbol{w}x_t$ 对 x_t 进行加权，并返回一个加权和分数，该分数代表话语 t 的中快乐的情绪强度，其中 $\boldsymbol{w}$ 为需要学习的排序权重。为了学习排序函数， $\boldsymbol{w}$ 需要满足以下相对约束条件：

$$\forall(i,j)\in O:\boldsymbol{w}x_i>\boldsymbol{w}x_j \ ; \ \forall(i,j)\in M:\boldsymbol{w}x_i=\boldsymbol{w}x_j \quad (1)$$

其中，O 和 M 分别为有序和相似成对样本组成的集合，x_i 和 x_j 分别为 T 中第 i 和第 j 个样本的声学特征向量。有序集合 O 由每对 $\{i,j\}$ 中强度不同的话语组成，在 H 中选取一个样本，在 N 中选取另一个样本，形成有序对 $\{i,j\}$，其中 $i\in H$，$j\in N$；相对属性的方法认为快乐语音样本中的快乐情绪强度大于中性语音样本中的强度。另外，相似集合 M 中包含了情感强度相似的话语，即 $\{i,j\}$ 取自同一个集合 N 或 H。据此，可以通过学习每个情绪类别的排序函数，给训练集中每个样本赋一个分数来表示对应的情感强度。为了方便进行强度量化调整，情感强度被归一化为[0,1]。在模型训练时，在编码器输出中加入情感强度和情感类别嵌入，引导模型学习细粒度的情感信息。在推理过程中，可以直接指定目标情绪类型和强度，例如{快乐，0.8}，控制语音合成模型合成想要的情绪语音。

1.2.2 Mixup对比排序

使用如上的相对属性排序算法提取情绪强度信息，其假设是：①来自同一情绪类的语音样本具有相似的排名，②中性情绪的强度最弱，其他所有情绪的排名都高于中性。尽管生成了具有不同情感强度水平的可识别语音样本，但这些模型忽略了类内的差异。在训练过程中，同一情感类别的样本（例如，最强和最弱的快乐）被视为相同。然而，在实践中，将中等强度的语音与强或弱强度的语音进行比较时可能会引起混淆。为解决这种混淆问题，Wang等人[38]提出了一个基于新型排序模型的语音合成模型，同时考虑了类内和类间的距离。该模型利用两个通过混合增强的样本，而不是将非中性样本和中性样本进行排序。每个经过混合增强的样本都是来自相同的非中性和中性语音的混合，通过对非中性和中性语音施加不同的权重，一种混合样本比另一种混合样本包含更多的非中性成分。通过对比式的学习对这两种混合样本集合进行排序，此时排序模型不仅需要确定情绪类别（类间距离），还需要捕获混合语音中存在的非中性情绪的数量，即非中性情绪的强度（类内距离），此时混合的比例值即为非中性情绪的情感强度。基于混合样本的排序模型如**图2**所示。该模型输入是梅尔频谱、音高轮廓和能量的综合表示。X_{neu} 表示来自中性类样本特征的输入，而 X_{emo} 表示来自其他非中性情感类样本特征的输入，对两种输入进行mixup[59]数据增强，并使用强度提取器提取强度表示。该模型采用FastSpeech 2[40]中相同的前馈Transformer结构（feed-forward transformer，FFT）来处理输入，并在FFT的输出中添加情感嵌入，以产生 I^i_{mix} 和 I^j_{mix} 的强度表示，接着将这两个强度表示序列平均为两个向量 $\boldsymbol{h}^i_{mix}$ 和 $\boldsymbol{h}^j_{mix}$ 以计算mixup的损失函数，最后经过全连接层映射得到标量对（r^i_{mix}，r^j_{mix}），这个标量对代表对应混合样本的情感强度。这种排序的方式损失迫使模型对两个都包含非中性情绪的样本进行正确排序。如此一来，通过混合中性情绪和非中性情绪来实现无监督

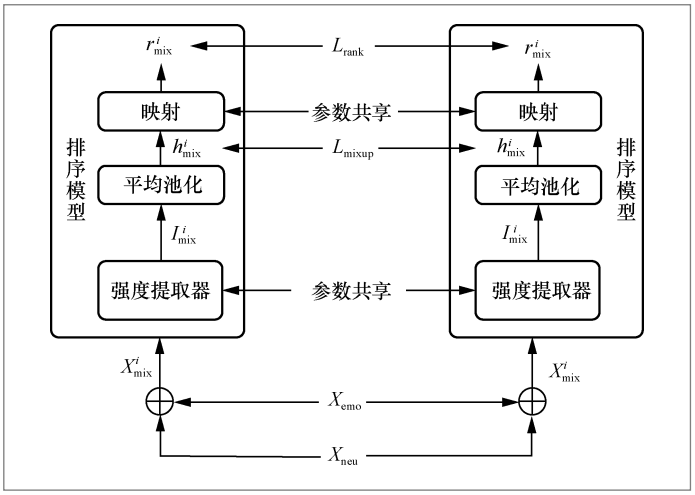

图 2 混合样本排序模型

强度排序可以感知到情感类内信息，该方法相比基于相对属性的方法更关注类内信息，极大地提高了情感控制的粒度。

1.3 情绪混合

研究表明，人类可以经历大约34 000种不同的情绪[60-61]。然而人们很难理解所有的情绪，Plutchik等人[60]将这些情绪总结成了8种基本情绪——愤怒、恐惧、悲伤、厌恶、惊诧、期待、信任和喜悦，并将它们排列在一个情绪轮中，如**图3**所示。所有其他情绪都可以被视为这些主要情绪的混合或衍生状态。根据情绪轮理论，情绪强度的变化可以产生我们所能感受到的不同数量的情绪。此外，原始情绪的叠加可以产生新的情绪类型。例如，欢乐可以通过喜悦和惊喜结合而产生[62-63]。尽管研究人员在心理学上做出了这些努力，但在之前的语音合成工作中，没有人尝试过对混合情绪进行建模。因此，受情绪轮理论的启发，有的工作认为情绪可以用范围[64]或维度[65-66]来表征情绪。基于维度的方法试图用维度表征来模拟言语情感的物理特性[66]。例如Russell的环形模型[65]，该模型将情绪分布在二维圆形空间中，包含唤醒和价态维度，混合情绪可以通过仔细调整每个维度来合成。然而只有少数情感语音数据库[67-68]提供了这样的标注，并且这些维度注释是主观的，收集起来代价高昂。Zhou等人[41]提

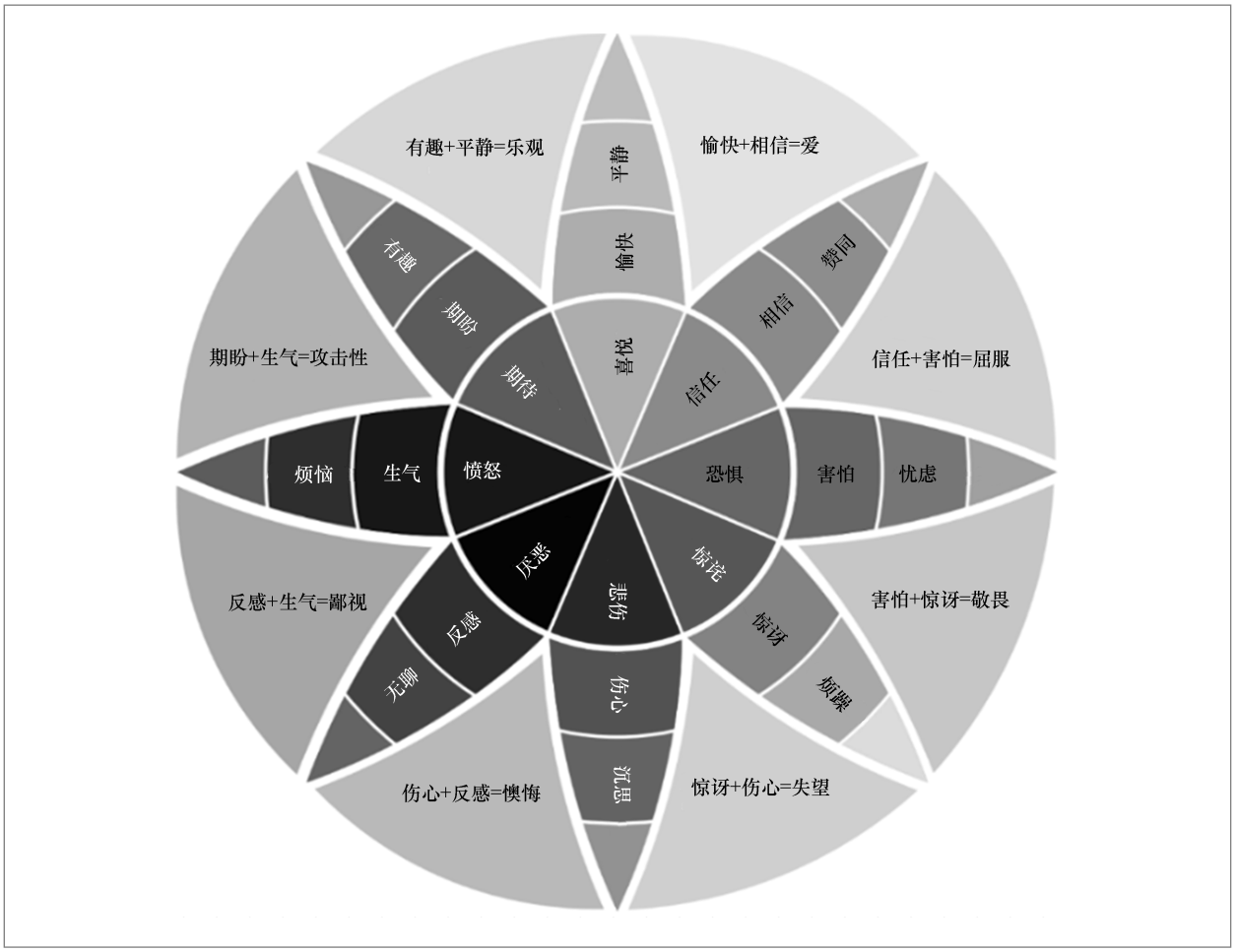

图 3 情绪轮[60]

出了混合多种不同情绪风格的情感语音合成模型Mixed Emotion，使手动操纵合成情绪成为可能。在推理时，模型框架通过文本输入将参考情感转换为新的情感语音。Mixed Emotion模型只使用大多数数据库中可用的离散情感标签。首先根据情绪轮理论做一个假设：混合情绪的特征是主要情绪的组合、混合或化合物。如何量化不同情绪之间的联系或相互作用是合成混合情绪的一个挑战。受情感的有序性的启发，根据情绪轮的理论做两个假设：①所有情绪都有一定程度的关联；②每种情绪都有刻板印象风格。Mixed Emotion提出基于等级的相对方案来量化不同情绪类型语音记录之间的相对差异，并通过调整与其他情绪类型的相对差异来表征混合情绪。该方法不仅描述了每种情绪的可识别风格，而且还量化了不同情绪风格之间的相似性，提出了一种基于等级的方法来测量情绪类别之间的相对差异，这种方法可以提供更多信息的描述，从而更接近人类的感觉[69]。

除了Mixed Emotion的融合方法，Tang等人[42]还参考图像中的混合生成模型[70]，提出一个基于扩散模型的情绪混合语音合成模型，如**图4**所示。该模型基于GradTTS[71]打造，只是预测的持续时间取决于情绪和说话者。其中的隐藏表示包括输入文本、情感嵌入e和说话人嵌入s的语言内容。因此，频谱去噪通过迭代方式将μ细化为参考音频中具有目标主要情感和说话人的梅尔频谱图。此外，该模型还采用情感识别模型进行辅助，进而使生成的情感与原始情感尽可能相同。在混合情感生成中，Emomix扩展了以单一情绪为条件训练后的概率扩散模型的逆向过程，用于合成混合情绪。通过在推理过程中替换采样步骤$K_{\max}$之后的条件向量，从而实现两种不同情绪的情绪混合，目的是用混合情绪覆盖基础的情绪细节。

2 数据集和评价指标

2.1 数据集

情感语音合成数据集的语音样本包含了多种情感或情绪，通常涵盖不同说话人和情感类别。这些数据集大多从专业实验室录制而来，例如ESD（emotional speech dataset）[72]，针对不同的语种，

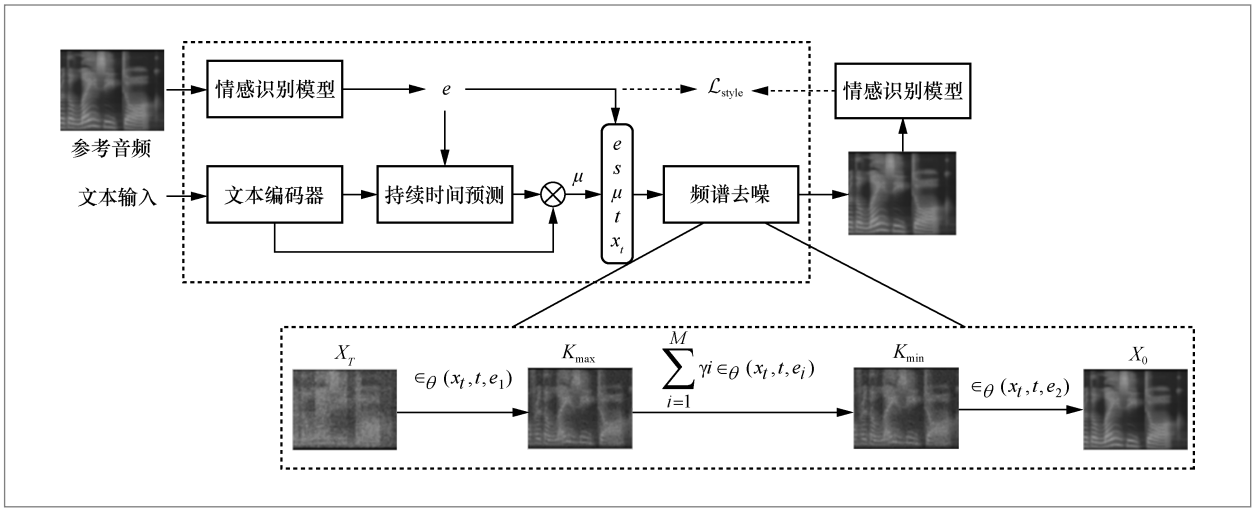

图 4 Emomix 模型结构图 [43]

情感语音合成数据集也分为汉语、英语和德语等。如**表2**所示，笔者收集并罗列了一些情感数据集。CASIA汉语情感语音数据集是由中国科学院自动化研究所（Chinese Academy of Sciences Institute of Automation，CASIA）提出的情感语音数据集，其中包含4个专业发音人、6种情绪、300句相同文本和100句不同文本。CHEAVD[73]（CASIA中文自然情感视听数据库）是同单位提出的多模态情感分析数据集，其中包含从电影、电视剧和脱口秀中提取的140 min的情感片段，一共238位说话人，年龄覆盖儿童到老人。该数据集标记了6种基础情感大类，并从中引申出26种细化情感类别。英文情感语音数据集使用最广泛的主要为ESD[72]和IEMOCAP[67]，其中ESD为双语数据集，包含了10位英语说话者和10位汉语说话者，共计29 h的语音。每位说话者分别对应有愤怒、高兴、悲伤、惊讶以及中性5种情感类别，在话语数量方面，每个说话者每种情感都有350句话语。IEMOCAP则为纯英文数据集，该数据集收录的是情感场景的即兴表演，或者是戏剧剧本的表演，为男女混合的双人情感对话录音，其中包括音频、文本转录、视频和动作捕捉记录，是一个多任务的综合数据集，可以提供情感之外的一些信息。此外，Liu等人[74]提出了一个情感综合数据集BEAT，该数据集包含30位说话者以8种不同的情绪和4种不同的语言对话中捕获的76 h高质量多模态数据，其丰富的帧级情绪和语音相关性注释可以用于高质量情感数字人生成。综上，可以看出汉语情感语音数据集目前体量比较小，希望今后会有情感类别更加丰富、数据量更大的汉语情感语音数据集，以促进汉语情感语音合成的发展。

2.2 评价指标

为了全面评估合成语音的情感表达效果，情感语音合成中的评价指标涵盖了自然度、流畅性、情绪准确性和情绪感染力等方面。目前广泛使用的评价指标主要分为主观和客观两个方面，主观评价指标依赖人类听感进行打分或者排序比较，客观评价指标从频谱特征等角度对合成语音进行计算打分，**表3**中总结了一些主客观评价指标。

表2 情感语音合成数据集

语言	数据集	情感种类
汉语	CASIA	中性、快乐、伤心、生气、害怕、惊讶
	CHEAVD[73]	生气、期盼、尴尬、快乐、伤心、惊讶、担心、紧张、厌恶
	Emovie[75]	中性、欢乐、愤怒、厌恶、困倦
	CH-SIMS[76]	消极、弱消极、弱积极、积极
汉英双语	ESD[72]	中性、伤心、快乐、生气、惊讶
英汉日西4种语言	BEAT[74]	中性、生气、快乐、害怕、厌恶、伤心、鄙视、惊讶
英语	IEMOCAP[67]	中性、愤怒、快乐、兴奋、悲伤、沮丧、恐惧、惊讶
	EmoV-DB[77]	中性、欢乐、愤怒、厌恶、困倦
	RAVDESS[78]	中性、平静、快乐、悲伤、愤怒、恐惧、惊讶、厌恶
德语	EMODB[79]	中性、无聊、厌恶、麻烦、恐惧、欢乐、悲伤

表 3 情感语音合成评价指标

分类	评价指标
主观	MOS、SMOS、情感强度排序和偏好性测试等
客观	MCD[80]、PESQ[81]、POLQA[82]、ANIQUE+[83]、E-Model[84]、AutoMOS[85]、QualityNet[86]、NISQA[87]和MOSNet[88]等

2.2.1 主观评价指标

主观评价通过人类对语音进行打分，包括平均意见得分（mean opinion score，MOS）和ABX测试等方式。MOS评测可以评估语音的不同方面，如自然度和相似度。然而，人类给出的评分受到很多因素的影响。比如，在语音合成评估中，对于较长的语音片段，提供的上下文信息会显著影响参与者的打分结果。目前主流使用的是由国际电信联盟提出的绝对等级评分（absolute category rating，ACR）方法，被试者需要对整体语音质量进行打分，通常为1到5分，见**表4**。ACR得分高于4表示大多数被试者认为语音质量较好，低于3则表示语音存在较大缺陷，大部分被试者不太满意。此外，为了评估情感迁移方法中生成语音与参考语音之间的相似性，从MOS评分标准中引申出了相似度平均意见得分（similarity mean opinion score，SMOS），该分数主要用来衡量生成语音和参考语音的情感相似度。由于一直以来没有情感语音合成数据集带有情感强度标签，针对情感强度的准确性评估，一般采用主观排序的方法进行，根据参评者的排序结果和真实情感强度进行对比，生成混淆矩阵，如**图5**所示。以从0到1衡量情感强度为例，混淆矩阵中每个子图的垂直和水平坐标分别表示实际情感强度的排序和参评者排序得到的感知情感强度顺序，图中的颜色代表选择率，选择率越高，颜色越深。如果参评者做出的排序和实际排序是一致的，那么混淆矩阵的斜对角线应该都为1，故在混淆矩阵中，斜对角线的数值越大越好，其他位置的数值越小越好。对于情感的表现力，大多数的方法则直接采取不同模型之间的主观偏好性测试（A/B测试），让参评者从不同模型生成的情感语音中挑选一个情感表现力最强的语音，这些情感语音由相同输入文本或参考语音生成。

表 4 平均意见得分标准

音频级别	意见得分	评价标准
优秀	5	听感很清楚，说话流畅
良好	4	听感清楚，比较流畅，有点杂音
中等	3	听不太清，表达稍有磕绊
较差	2	听不清，表达含糊，杂音很多
极差	1	完全听不清

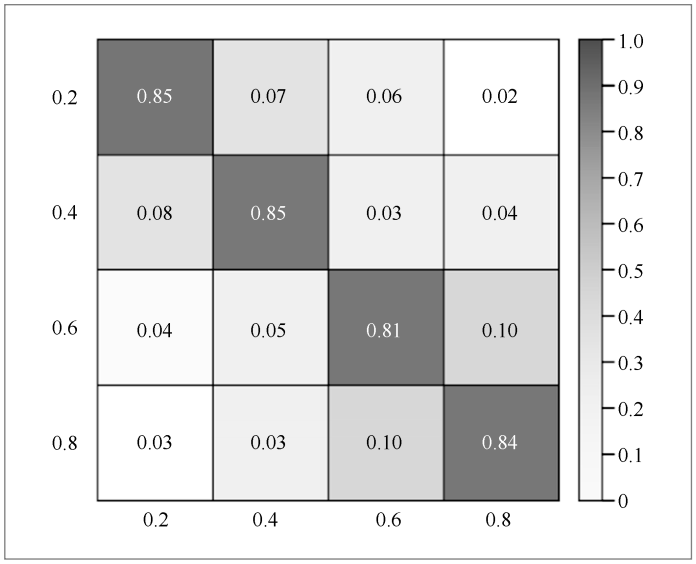

图 5 情感强度排序混淆矩阵

2.2.2 客观评价指标

语音合成中的客观评价指标是通过

计算机自动评价语音质量的指标，通常用于量化和评价合成语音的声音质量。目前大多数将频谱细节、梅尔倒谱失真（Mel cepstral distortion，MCD）[80]等指标作为评价标准。其中MCD为使用最广泛的客观评价指标，通过提取参考语音和生成语音的梅尔频谱，根据帧的步长（一般为5 ms），将帧序列表示为 $\boldsymbol{v}_d(t)$，d代表维度，t代表帧索引，则生成语音用 $\boldsymbol{v}^{\text{targ}}$ 表示，参考语音用 $\boldsymbol{v}^{\text{ref}}$ 表示，MCD的计算式如下：

$$\text{MCD}\boldsymbol{v}^{\text{targ}},\boldsymbol{v}^{\text{ref}} = \frac{\alpha}{T'}\sum_{t=0}^{T-1}\sqrt{\sum_{d=s}^{\text{D}}\left(\boldsymbol{v}_d^{\text{targ}}(t)-\boldsymbol{v}_d^{\text{ref}}(t)\right)^2} \tag{2}$$

其中，$\alpha=\frac{10\sqrt{2}}{\ln 10}=6.14185$，$T=\min\left(\left|\boldsymbol{v}^{\text{targ}}\right|,\left|\boldsymbol{v}^{\text{ref}}\right|\right)$，$T'$ 为非静默帧数，s为计算的起始维度，取值0或1，代表是否计算第0维梅尔倒谱，D为总维度。除了MCD还有语音质量的感知评估（perceptual evaluation of speech quality，PESQ）[81]的方法，PESQ根据主观听觉测试的结果，利用信号处理和模型计算的方法，对合成语音与参考语音之间的差异进行量化评估。

通过参考音频的有无，客观评价指标可以分为以下两类：①有参考音频，此时需要一个标准音质的参考信号，上面的MCD和PESQ就属于有参考的方法，此外还有ITU-T P.863（POLQA）[82]等；②无参考音频，此时直接根据待测音频进行评分，常见的有基于信号（ITU-T P.563和ANIQUE+[83]）和基于参数ITU-T G.107（E-Model）[84]的方法，其中E-Model基于网络的基本参数，如网络的丢包率、延迟、抖动等参数来计算语音通信的整体质量得分。近年来，深度学习技术也逐渐应用于无参考质量评价，如AutoMOS[85]、QualityNet[86]、NISQA[87]和MOSNet[88]。这些评价方法在评价语音质量和合成技术的发展中起着重要作用。

3 总结与展望

本文从情感语音合成的情感准确性、情感控制灵活性与情感类别扩展性3个研究方向进行了分析与总结，探究了基于深度学习的情感语音合成近几年来的发展，根据以上分析，可以看出情感语音合成目前仍然存在一些问题，同时也为后续的研究提出一些方向。

- 情感标记和上下文理解：情感语音合成系统通常需要从输入的文本中理解情感的含义，并相应地调整声音的表达方式。然而，在某些情况下，系统可能无法准确地理解情感标记或无法适应复杂的上下文，导致生成的声音与用户期望的情感不一致。
- 情感过渡和平滑度：在某些情况下，情感语音合成系统可能在不同情感之间的过渡上存在一定程度的不自然，使得在情感转换时声音的变化不够平滑。
- 长文本的处理：对于较长的文本，一些情感语音合成系统可能会面临语义理解的挑战，导致在长篇文本合成时出现不连贯或不准确的情况。
- 语言和口音的适应性：情感语音合成系统可能在不同语言和口音的处理上存在一定的局限性，导致在多语言环境下表现不尽如人意。
- 数据隐私和安全：随着情感语音合成技术的发展，涉及大量个人语音数据的隐私保护和安全问题也变得尤为重要。

参考文献：

[1] WANG Y X, SKERRY-RYAN R J, STANTON D, et al. Tacotron: towards

end-to-end speech synthesis[C]// Proceedings of Interspeech 2017. [S.l.]: ISCA, 2017: 4006-4010.

[2] 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.

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

[4] YU C Z, LU H, HU N, et al. DurIAN: duration informed attention network for speech synthesis[C]//Proceedings of Interspeech 2020. [S.l.]: ISCA, 2020: 2027-2031.

[5] LING Z H, KANG S Y, ZEN H G, et al. Deep learning for acoustic modeling in parametric speech generation: a systematic review of existing techniques and future trends[J]. IEEE Signal Processing Magazine, 2015, 32(3): 35-52.

[6] 唐浩彬, 张旭龙, 王健宗, 等. 表现性语音合成综述[J]. 大数据, 2023, 9(6): 53-71.
TANG H B, ZHANG X L, WANG J Z, et al. A survey of expressive speech synthesis[J]. Big Data Research, 2023, 9(6): 53-71.

[7] KWON O, JANG I, AHN C, et al. Emotional speech synthesis based on style embedded Tacotron2 framework[C]// Proceedings of 2019 34th International Technical Conference on Circuits/ Systems, Computers and Communications. Piscataway: IEEE Press, 2019: 1-4.

[8] 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 ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2019: 6945-6949.

[9] WANG Y, SKERRY-RYAN R, XIAO Y, et al. Uncovering latent style factors for expressive speech synthesis[EB]. arXiv preprint, 2017, arXiv: 1711.00520.

[10] LI T, WANG X S, XIE Q C, et al. Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 30: 1448-1460.

[11] OHTANI Y, YU N S, MORITA M, et al. Emotional transplant in statistical speech synthesis based on emotion additive model[C]//Proceedings of Interspeech 2015. [S.l.]: ISCA, 2015: 274-278.

[12] INOUE K, HARA S, ABE M, et al. An investigation to transplant emotional expressions in DNN-based TTS synthesis[C]//Proceedings of 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Piscataway: IEEE Press, 2018: 1253-1258.

[13] CHOI H, PARK S, PARK J, et al. Multi-speaker emotional acoustic modeling for CNN-based speech synthesis[C]// Proceedings of ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2019: 6950-6954.

[14] JOHNSON J, ALAHI A, LI F F. Perceptual losses for real-time style transfer and super-resolution[M]//Computer Vision - ECCV 2016. Cham: Springer, 2016: 694-711.

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

Press, 2020: 7254–7258.
[16] 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(6): 526–533.
[17] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB]. arXiv preprint, 2014, arXiv: 1409.0473.
[18] SOTELO J, MEHRI S, KUMAR K, et al. Char2wav: end-to-end speech synthesis[C]//Proceedings of International Conference on Learning Representations. [S.l.:s.n.], 2017.
[19] SCHULLER D M, SCHULLER B W. A review on five recent and near-future developments in computational processing of emotion in the human voice[J]. Emotion Review, 2021, 13(1): 44–50.
[20] LEE Y, RABIEE A, LEE S Y. Emotional end-to-end neural speech synthesizer [EB]. arXiv preprint, 2017, arXiv: 1711.05447.
[21] TITS N, EL HADDAD K, DUTOIT T. Exploring transfer learning for low resource emotional TTS[C]//Proceedings of SAI Intelligent Systems Conference. Cham: Springer, 2020: 52–60.
[22] 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.
[23] 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.
[24] LEI Y, YANG S, WANG X S, et al. MsEmoTTS: multi-scale emotion transfer, prediction, and control for emotional speech synthesis[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 30: 853–864.
[25] BIAN Y, CHEN C, KANG Y, et al. Multi-reference tacotron by intercross training for style disentangling, transfer and control in speech synthesis[EB]. arXiv preprint, 2019, arXiv: 1904.02373.
[26] UM S Y, OH S, BYUN K, et al. Emotional speech synthesis with rich and granularized control[C]//Proceedings of ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2020: 7254–7258.
[27] LEE Y, KIM T. Robust and fine-grained prosody control of end-to-end speech synthesis[C]//Proceedings of ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2019: 5911–5915.
[28] LORENZO-TRUEBA J, HENTER G E, TAKAKI S, et al. Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis[J]. Speech Communication, 2018, 99: 135–143.
[29] KINGMA D P, WELLING M. Auto-encoding variational bayes[EB]. arXiv preprint, 2013, arXiv: 1312.6114.
[30] 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 ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2019: 6945–6949.
[31] KENTER T, WAN V, CHAN C A, et al. CHiVE: varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network[C]// Proceedings of the 36th International Conference on Machine Learning. [S.l.:s.n.], 2019: 3331–3340.

[32] FRIJDA N H, ORTONY A, SONNEMANS J, et al. The complexity of intensity: Issues concerning the structure of emotion intensity[J]. Emotion, 1992: 60-89.

[33] MATSUMOTO K, HARA S, ABE M. Controlling the strength of emotions in speech-like emotional sound generated by WaveNet[C]//Proceedings of Interspeech 2020. [S.l.]: ISCA, 2020: 3421-3425.

[34] SCHNELL B, GARNER P N. Improving emotional TTS with an emotion intensity input from unsupervised extraction[C]// Proceedings of 11th ISCA Speech Synthesis Workshop. [S.l.]: ISCA, 2021: 60-65.

[35] CHOI H, HAHN M. Sequence-to-sequence emotional voice conversion with strength control[J]. IEEE Access, 2021, 9: 42674-42687.

[36] IM C B, LEE S H, KIM S B, et al. EMOQ-TTS: emotion intensity quantization for fine-grained controllable emotional text-to-speech[C]//Proceedings of ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2022: 6317-6321.

[37] ZHU X L, YANG S, YANG G, et al. Controlling emotion strength with relative attribute for end-to-end speech synthesis[C]//Proceedings of 2019 IEEE Automatic Speech Recognition and Understanding Workshop. Piscataway: IEEE Press, 2020: 192-199.

[38] WANG S J, GUÐNASON J, BORTH D. Fine-grained emotional control of text-to-speech: learning to rank inter- and intra-class emotion intensities[C]// Proceedings of ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2023: 1-5.

[39] PARIKH D, GRAUMAN K. Relative attributes[C]//Proceedings of 2011 International Conference on Computer Vision. Piscataway: IEEE Press, 2012: 503-510.

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

[41] ZHOU K, SISMAN B, RANA R, et al. Speech synthesis with mixed emotions[J]. IEEE Transactions on Affective Computing, 2022, 14(4): 3120-3134.

[42] TANG H B, ZHANG X L, WANG J Z, et al. EmoMix: emotion mixing via diffusion models for emotional speech synthesis[C]// Proceedings of INTERSPEECH 2023. ISCA: ISCA, 2023.

[43] WU X X, SUN L F, KANG S Y, et al. Feature based adaptation for speaking style synthesis[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2018: 5304-5308.

[44] MA S, MCDUFF D, SONG Y. Neural TTS stylization with adversarial and collaborative games[C]//Proceedings of International Conference on Learning Representations. [S.l.:s.n.], 2018.

[45] WHITEHILL M, MA S, MCDUFF D, et al. Multi-reference neural TTS stylization with adversarial cycle consistency[C]// Proceedings of Interspeech 2020. ISCA: ISCA, 2020: 4442-4446.

[46] LEE J, CHO K, HOFMANN T. Fully character-level neural machine translation without explicit segmentation[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 365-378.

[47] YANNAKAKIS G N, COWIE R, BUSSO C. The ordinal nature of emotions[C]// Proceedings of 2017 Seventh International Conference on Affective Computing and Intelligent Interaction. Piscataway: IEEE Press, 2017: 248-255.

[48] HARVILL J, LEEM S G, ABDELWAHAB M, et al. Quantifying emotional similarity in speech[J]. IEEE Transactions on Affective Computing, 2021, 14(2): 1376–1390.

[49] CAO H W, VERMA R, NENKOVA A. Speaker–sensitive emotion recognition via ranking: studies on acted and spontaneous speech[J]. Computer Speech & Language, 2015, 29(1): 186–202.

[50] HARVILL J, ABDEL WAHAB M, LOTFIAN R, et al. Retrieving speech samples with similar emotional content using a triplet loss function[C]// Proceedings of ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2019: 7400–7404.

[51] FÜRNKRANZ J, HÜLLERMEIER E. Pairwise Preference Learning and Ranking[C]//Proceedings of European Conference on Machine Learning. Heidelberg: Springer, 2003: 145–156.

[52] LOTFIAN R, BUSSO C. Retrieving categorical emotions using a probabilistic framework to define preference learning samples[C]//Proceedings of Interspeech 2016. [S.l.]: ISCA, 2016: 490–494.

[53] PARTHASARATHY S, COWIE R, BUSSO C. Using agreement on direction of change to build rank–based emotion classifiers[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(11): 2108–2121.

[54] MARTÍNEZ H P, YANNAKAKIS G N, HALLAM J. Don't classify ratings of affect; rank them![J]. IEEE Transactions on Affective Computing, 2014, 5(3): 314–326.

[55] YANNAKAKIS G N, MARTÍNEZ H P. Grounding truth via ordinal annotation[C]//Proceedings of 2015 International Conference on Affective Computing and Intelligent Interaction. Piscataway: IEEE Press, 2015: 574–580.

[56] HUANG J, LI Y, TAO J H, et al. Speech emotion recognition from variable–length inputs with triplet loss function[C]// Proceedings of Interspeech 2018. ISCA: ISCA, 2018: 3673–3677.

[57] FENG K X, CHASPARI T. A Siamese neural network with modified distance loss for transfer learning in speech emotion recognition[C]// Proceedings of the 3rd Workshop on Affective Content Analysis. [S.l.:s.n.], 2020: 29–35.

[58] ZHOU K, SISMAN B, RANA R, et al. Emotion intensity and its control for emotional voice conversion[J]. IEEE Transactions on Affective Computing, 2022, 14(1): 31–48.

[59] ZHANG H, CISSE M, DAUPHIN Y N, et al. Mixup: beyond empirical risk minimization [EB]. arXiv preprint, 2017, arXiv: 1710.09412.

[60] PLUTCHIK R. The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice[J]. American Scientist, 2001, 89(4): 344–350.

[61] MOORS A, ELLSWORTH P C, SCHERER K R, et al. Appraisal theories of emotion: state of the art and future development[J]. Emotion Review, 2013, 5(2): 119–124.

[62] CROSS M, HANRAHAN C. Changing minds: the go–to guide to mental health for family and friends[M]. Perth: Harper Collins Australia, 2016.

[63] WHISSELL C M. The dictionary of affect in language[M]//The Measurement of Emotions. Amsterdam: Elsevier, 1989: 113–131.

[64] EKMAN P. An argument for basic emotions[J]. Cognition and Emotion, 1992, 6(3/4): 169–200.

[65] RUSSELL J A. A circumplex model of affect[J]. Journal of Personality and Social Psychology, 1980, 39(6): 1161–1178.

[66] SCHRODER M. Expressing degree of activation in synthetic speech[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(4): 1128–1136.

[67] BUSSO C, BULUT M, LEE C C, et al. IEMOCAP: interactive emotional dyadic motion capture database[J]. Language Resources and Evaluation, 2008, 42(4): 335–359.

[68] BUSSO C, PARTHASARATHY S, BURMANIA A, et al. MSP-IMPROV: an acted corpus of dyadic interactions to study emotion perception[J]. IEEE Transactions on Affective Computing, 2017, 8(1): 67–80.

[69] CHAPELLE O. Training a support vector machine in the primal[J]. Neural Computation, 2007, 19(5): 1155–1178.

[70] LIEW J H, YAN H, ZHOU D, et al. MagicMix: semantic mixing with diffusion models [EB]. arXiv preprint, 2022, arXiv: 2210.16056.

[71] POPOV V, VOVK I, GOGORYAN V, et al. Grad-TTS: a diffusion probabilistic model for text-to-speech [EB]. arXiv preprint, 2021, arXiv: 2105.06337.

[72] ZHOU K, SISMAN B, LIU R, et al. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset[C]//Proceedings of ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2021: 920–924.

[73] LI Y, TAO J H, CHAO L L, et al. CHEAVD: a Chinese natural emotional audio-visual database[J]. Journal of Ambient Intelligence and Humanized Computing, 2017, 8(6): 913–924.

[74] LIU H Y, ZHU Z H, IWAMOTO N, et al. BEAT: a large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2022: 612–630.

[75] CUI C Y, REN Y, LIU J L, et al. EMOVIE: a mandarin emotion speech dataset with a simple emotional text-to-speech model[C]//Proceedings of Interspeech 2021. [S.l.]: ISCA, 2021: 2766–2770.

[76] YU W M, XU H, MENG F Y, et al. CH-SIMS: a Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 3718–3727.

[77] ADIGWE A, TITS N, HADDAD K E, et al. The emotional voices database: towards controlling the emotion dimension in voice generation systems[EB]. arXiv preprint, 2018, arXiv: 1806.09514.

[78] LIVINGSTONE S R, RUSSO F A. The Ryerson audio-visual database of emotional speech and Song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English[J]. PLoS One, 2018, 13(5): e0196391.

[79] BURKHARDT F, PAESCHKE A, ROLFES M, et al. A database of German emotional speech[C]//Proceedings of Interspeech 2005. ISCA: ISCA, 2005: 1517–1520.

[80] KOMINEK J, SCHULTZ T, BLACK A. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion[C]//Proceedings of International Workshop on Spoken Languages Technologies for Under. [S.l.:s.n.], 2008.

[81] RIX A W, BEERENDS J G, HOLLIER M P, et al. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs[C]//Proceedings of 2001 IEEE International Conference on Acoustics,

Speech, and Signal Processing. Proceedings. Piscataway: IEEE Press, 2002: 749-752.
[82] BEERENDS J, SCHMIDMER C, BERGER J, et al. Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part I-temporal alignment[J]. Journal of the Audio Engineering Society, 2013, 61: 366-384.
[83] KIM D S, TARRAF A. ANIQUE+: a new American national standard for non-intrusive estimation of narrowband speech quality[J]. Bell Labs Technical Journal, 2007, 12(1): 221-236.
[84] DING L, GOUBRAN R A. Speech quality prediction in VoIP using the extended E-model[C]//Proceedings of IEEE Global Telecommunications Conference. Piscataway: IEEE Press, 2004: 3974-3978.
[85] PATTON B, AGIOMYRGIANNAKIS Y, TERRY M, et al. AutoMOS: learning a non-intrusive assessor of naturalness-of-speech[EB]. arXiv preprint, 2016, arXiv: 1611.09207.
[86] FU S W, TSAO Y, HWANG H T, et al. Quality-net: an end-to-end non-intrusive speech quality assessment model based on BLSTM[C]//Proceedings of Interspeech 2018. [S.l.] : ISCA, 2018: 1873-1877.
[87] MITTAG G, NADERI B, CHEHADI A, et al. NISQA: a deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets[C]//Proceedings of Interspeech 2021. [S.l.]: ISCA, 2021: 2127-2131.
[88] LO C C, FU S W, HUANG W C, et al. MOSNet: deep learning-based objective assessment for voice conversion[C]// Proceedings of Interspeech 2019. [S.l.]: ISCA, 2019: 1541-1545.

作者简介

施昊翔(2000-),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,复旦大学计算机理学博士,主要研究方向为语音合成、语音转换、音乐信息检索以及机器学习和深度学习方法在人工智能领域应用。担任清华大学深圳研究院以及中国科学技术大学先进技术研究院校外导师,目前是IEEE、中国自动化学会以及中国计算机学会会员,担任联邦数据与联邦智能专委会委员,2023年入选上海市东方英才计划青年项目。

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,智能金融前沿技术研究院院长。美国佛罗里达大学人工智能博士后,美国莱斯大学和华中科技大学联合培养博士,中国计算机学会资深会员,中国计算机学会大数据专家委员会委员,中国自动化学会联邦数据和联邦智能专业委员会副主任。主要研究方向为大模型、联邦学习和深度学习等。

程宁（1981– ），男，博士，平安科技（深圳）有限公司高级人工智能专家、中国科学院自动化所博士。专注于人工智能算法研究以及其在语音处理和自然语言处理领域的应用。目前在大数据、机器学习、人工智能国际顶会或期刊上发表学术论文50余篇，发明专利申请100余项。

肖京（1972– ），男，美国卡耐基梅隆大学博士，IEEE Fellow，国家特聘专家。国家新一代普惠金融人工智能开放创新平台技术负责人、深圳市政协委员、深圳市决策咨询委员会委员，兼中国计算机学会深圳分部副主席、广东省人工智能与机器人学会副理事长、深圳市人工智能行业协会会长、深圳市人工智能学会副理事长， 清华大学、上海交通大学、同济大学等客座教授。长期从事人工智能与大数据分析挖掘相关领域研究，先后在爱普生美国研究院及美国微软公司担任高级研发管理职务，现任平安集团首席科学家，负责人工智能技术研发及其在金融、医疗、智慧城市等领域的应用，带领团队树立了多项传统行业智能化经营的标杆。已发表学术论文249篇，美国授权专利101项，中国发明专利155项，参与及承担国家级项目8项。凭借在技术创新及应用的杰出贡献，先后获得2018年中国专利奖、2019年吴文俊人工智能杰出贡献奖、2020年吴文俊人工智能科技进步一等奖、2020年上海市科技进步奖一等奖、2020年中国人工智能十大风云人物、2021年深圳市五一劳动奖章、2022年深圳市最美科技工作者等荣誉。

收稿日期： 2023-09-26
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)