用于语音情感识别的生成式情感人工智能：合成情感语音增强的必要性

Abdullah Shahid[a]、SiddiqueLatif[b]、* 和Junaid Qadir[c]

[a]信息技术大学（*ITU*），巴基斯坦旁遮普省[b]南昆士兰大学，澳大利亚[c]卡塔尔大学，多哈，卡塔尔

文章信息

关键词：Tacotron、WaveRNN、语音合成、文本转语音、情感语音合成、语音情感识别

摘要

尽管深度学习取得了进展，但由于缺乏语音情感数据集，当前最先进的语音情感识别（SER）系统的性能仍然较差。本文提出使用基于扩展 Tacotron 架构的端到端文本转语音（TTS）系统生成的合成情感语音来增强 SER 系统。所提出的 TTS 系统包括用于说话人嵌入和情感嵌入的编码器、用于生成梅尔频谱图的序列到序列文本生成器，以及用于从梅尔频谱图生成音频的 WaveRNN。大量实验表明，与基线相比，生成的情感语音具有更高的平均意见得分（MOS），其质量能够显著提升多个数据集上的 SER 性能。生成的样本同样能够有效增强 SER 性能。

arXiv:2301.03751v1 [cs.SD] 10 Jan 2023

1. 引言

语音情感识别（SER）是一个快速发展的领域，在医疗保健、客户服务、媒体、教育和法证学等领域有着广泛应用。尽管深度学习（DL）在开发 SER 系统方面展现出良好前景，但其性能仍受到情感数据集稀缺的限制 [20, 24]。现有语音情感识别语料库规模较小，因为创建情感数据的过程成本高且耗时，多名标注人员必须手动听取并标注材料 [26, 31]。为扩大数据规模，一些研究使用了多个语料库，但标准基准数据集的数量同样有限，阻碍了 SER 系统的发展 [19]。

研究人员长期以来一直致力于打造听起来自然的文本转语音系统。文本转语音技术已取得长足进步：早期系统通常根据输入文本，将预先录制的波形拼接起来 [11]。这类系统容易出现边界伪影问题；后来，人们开发了统计技术，为声码器生成平滑的音频特征，以合成语音 [37, 44]。近来，人们提出了基于神经网络的端到端方法，能够合成听起来更加自然的人类语音 [3, 34]。当前最先进的文本转语音系统采用深度学习算法，以端到端 方式进行训练；常见模型包括 Tacotron [41], Deepvoice [3],FastSpeech [33, 34], FastPitch [18],等。

与传统系统不同，端到端 TTS 模型能够直接从文本学习生成频谱图，无需进行任何复杂的预处理。然而，这些模型目前只能合成自然语音。使用生成式深度学习技术（如生成对抗网络（GAN））[8] 进行情感语音合成也颇具挑战，因为这需要大量时间对齐数据，其中同一位说话人以不同情感朗读相同内容，并且还需要复杂的方程来指导模型利用音频特征转换情感，。一些研究使用 TTS 模型在单说话人情感语音合成方面取得了良好的结果 [17], ，但合成语音用于扩充语音情感识别数据时的质量尚未得到评估。

本文提出一种利用情感文本转语音（TTS）系统增强语音情感识别系统的方法，并作出两项主要贡献。首先，我们开发了一种端到端多说话人情感文本转语音系统，无需对音频文件进行对齐即可转换情感，也无需对输入数据进行复杂的预处理。受端到端文本转语音模型成功经验的启发，我们采用了类似 Tacotron 的架构。我们提出使用条件编码器来控制输出语音中的说话人音色和情感。我们利用编码器网络生成说话人音色特征向量，并将这些特征向量与一种编码后的情感特征表示进行调制。随后，利用调制后的特征向量对 Tacotron 进行条件控制，使其能够以不同说话人音色和情感合成语音。主观评估任务表明，我们提出的模型提升了可控性，并能成功合成情感语音。其次，我们利用合成的情感语音扩充语音情感识别系统，并开展多项实验，对生成的数据进行定量评估。结果表明，在语料库内和跨语料库两种设置下，合成数据都有助于提升语音情感识别性能。

本文其余部分安排如下。在第2节中，我们简要介绍改变音频不同特征的相关研究。第3节介绍模型架构、损失函数以及架构的流程。第4节说明数据集、模型训练所采用的实验条件及超参数。第5节报告结果。最后，本文在第6节作总结。

*通讯作者开放研究者与贡献者身份识别码：

2. 先前工作

本节回顾以下方面的相关文献：（1）Tacotron 在文本转语音（TTS）中的应用；（2）情感语音合成；以及（3）语音情感识别（SER）的增强过程。

2.1. 基于 Tacotron 的文本转语音系统

许多近期研究都集中于修改 Tacotron 模型，以便更好地控制文本转语音系统的输出。例如，[13]提出了一种基于 Tacotron 的模型，通过以说话人验证模型生成的说话人语音嵌入作为条件，使 Tacotron 合成多说话人语音 [40]。 [42]提出了一种 Tacotron 变体，该变体通过学习不同的风格并将其保存为向量或标记来改变说话风格。这些标记通过对相似口音进行聚类，并用平均值表示每个聚类获得。在合成过程中，以其中一个标记作为 Tacotron 的条件，从而生成具有特定风格的语音。 [35] 提出了一种多说话人 Tacotron，可以改变口音（例如美式、印度式和英式口音）。他们的模型将两个编码器网络与 Tacotron 结合使用，并要求输入两个音频样本（一个用于口音，另一个用于说话人的声音），以生成所需的输出。 [36] 提出了一种 Tacotron 模型，该模型以变分自编码器编码后的输出音频作为输入进行训练。这不仅提升了 Tacotron 的多说话人性能，还通过变分自编码器的均值-方差属性，实现了对生成音频能量的控制。 [43] 开发了一种 Tacotron 模型，通过使用 Tacotron 中的自注意力机制，学习与不同口音中音高相关的复杂依赖关系，从而学习更复杂的发声。他们声称，对于具有更强音高依赖性口音的语言（如日语），其模型优于传统的端到端方法。我们提出的模型同样能够在多说话人场景下生成语音，并且能够进一步控制输出中的情感。

2.2. 基于 Tacotron 的情感文本转语音系统

此前已有多项研究尝试使用文本转语音系统生成情感语音。例如， [38] 开发了一种基于 GST-Tacotron 网络的文本转语音系统情感控制方法 [35], , 并证明该方法在韩语单说话人场景下合成情感语音的有效性。 [27] 也评估了一个基于 Tacotron 的韩语情感语音合成器，并发现其生成的单说话人语音质量有所提升。其他研究，例如 [15, 17], ，也提出了控制情感语音合成的方法，但这些方法只能合成单说话人的情感语音

。相比之下，我们提出的方法能够控制多说话人文本转语音中的情感语音合成，并且我们还评估了合成数据的质量，以增强语音情感识别系统。

2.3. 语音情感识别的数据增强技术

速度扰动 [16] 是一种广受欢迎的数据增强技术，已在不同情境中得到广泛研究 [2, 23]。研究发现，通过生成具有不同速度效果的输入数据副本，该方法能够提升语音情感识别（SER）的性能。Mixup [45]是另一种数据增强技术，它将输入数据中的原始样本进行线性组合，以生成增强样本。多项研究证明了 Mixup 在 SER 中的有效性，包括 Latif 等人 [25], 使用该技术增强语音情感识别系统，并提升了其性能和鲁棒性。另一种名为 SpecAugment [30],的方法最初是为自动语音识别提出的，也已应用于 SER [4]。在这项研究中，作者将语音情感识别系统的样本数量加倍，并发现 SpecAugment 提升了模型性能。其他研究 [2, 21, 23]也通过采用基于输入扰动的数据增强技术来增加训练数据，取得了性能提升。

还需要开展进一步研究，探索数据驱动的方法来增加语音情感识别的训练数据。本文提出探索一种基于文本转语音的数据增强方法，通过改变不同情感下的说话人和性别声音，研究训练数据的不同变化。

3. 提出的框架

我们提出使用基于 Tacotron 的情感文本转语音系统生成合成语音，并使用合成语音数据来增强语音情感分类器。下文将介绍情感文本转语音系统和分类器的具体细节。

3.1. 情感语音合成

我们的模型由一个编码器构成，该编码器为 Tacotron 提供条件（如图 1 所示），以改变输出语音中的说话人声音和情感。Tacotron 根据给定的文本和嵌入向量生成梅尔频谱图，而基于 Wave-RNN 的声码器则用于根据梅尔频谱图生成音频信号

3.1.1. 条件编码器

我们提出使用条件编码器来生成同时体现说话人身份和情感的嵌入。为此，我们采用 [40], 中提出的说话人识别模型，该模型通过任意长度语音信号计算出的梅尔频谱图序列，生成一个固定维度的嵌入，称为 d 向量 [10, 39], 。我们使用端到端说话人验证损失来训练该模型：最大化同一说话人的语句之间的余弦相似度，同时最小化不同说话人的语句之间的余弦相似度。我们在情感语料库上对该网络进行微调，以生成情感嵌入。因此，条件编码器经过优化，能够最大化同一说话人在不同情感下的嵌入之间的余弦相似度，并最小化不同情感及不同说话人之间的相似度。通过这种方式，模型学会生成同时包含情感和说话人身份信息的特征向量。说话人的语音音频和情感音频经过嵌入

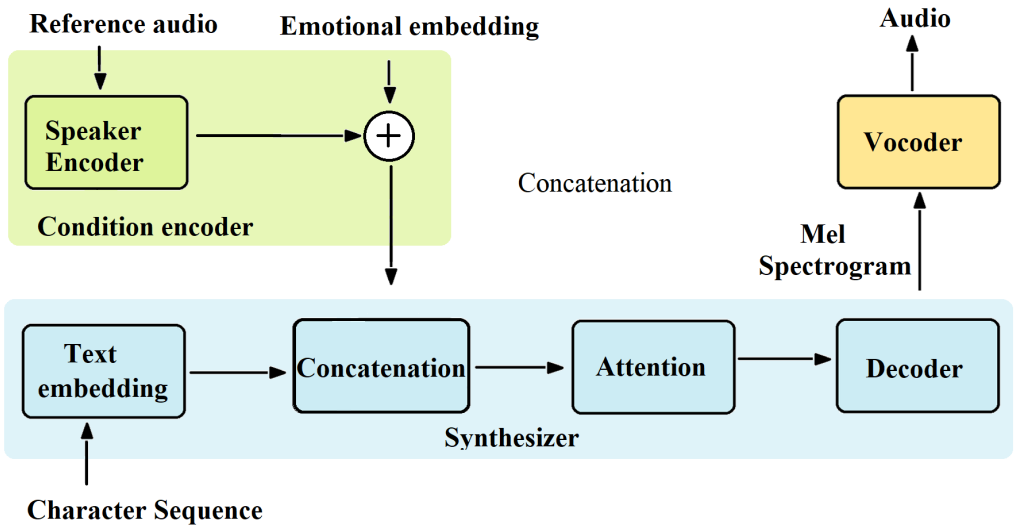

图 1： 系统架构流程图。如模型架构所述，首先对参考说话人的音色进行编码，然后将其调制为所需的情感。随后，将输出与文本嵌入一同传入 Tacotron 解码器，以合成梅尔频谱图。

条件编码器处理后组合生成最终嵌入，并以此作为合成器的条件，使其输出具有所选情感和说话人音色的语音。

对于数据集中每位说话人的每种独特情感，通过对每位独特说话人的每种独特情感对应的嵌入取平均值，计算质心 c_k。当嵌入与质心 c_k具有相同的说话人和情感时，嵌入 e_i 的损失计算如下：

$$\mathcal{L}(e_i, c_k) = -1 \times \sigma(\cos(e_i, c_k)) \tag{1}$$

当 e_i 具有不同的情感或不同的说话人时，对于质心 c_k，损失计算如下：

$$\mathcal{L}(e_i, c_k) = \sigma(\cos(e_i, c_k)) \tag{2}$$

$$\mathcal{L}_G(S) = \sum_{i,k} L(e_i, c_k) \tag{3}$$

公式 1 最大化相同说话人音色和相同情感对应嵌入之间的余弦相似度。公式 2 表示嵌入与质心在说话人音色不同、情感不同或二者均不同时的余弦相似度。公式 3 表示所有嵌入的最终损失，该损失通过对每个嵌入与每个质心之间的损失求和计算得出。

条件编码器由三个 LSTM 层组成，每层包含 768 个单元，最后接一个长度为 256 的全连接层。模型的输入是从参考说话人的音频样本中的语音语句生成的梅尔频谱图，输出则是一个长度为 256 的嵌入向量，用于表示说话人的身份。训练好模型后，我们使用它从给定音频中提取说话人和情感信息。为了将情感与说话人的声音分离开来，我们使用训练好的条件编码器，分别为同一说话人的中性语音和情感语音生成嵌入向量，从而得到只包含情感信息的向量。随后根据公式（4），从情感嵌入向量中减去中性嵌入向量，得到仅包含情感信息的向量。推理时可使用该向量控制合成音频的情感。

$$emb_{\text{em}} = (emb_{\text{en}} - emb_{\text{neu}}) \tag{4}$$

其中，emb_{en} 表示从说话人的情感语音中生成、包含情感和语音信息的嵌入；emb_{neu} 由同一说话人的中性音频生成，emb_{em} 则表示仅包含情感信息的嵌入。在推理过程中，将参考音频的嵌入（即我们希望输出样本采用的语音）与情感嵌入相加，生成最终的嵌入向量。

$$emb_{\text{final}} = emb_{\text{ref}} + emb_{\text{em}} \tag{5}$$

最后，将调制后的嵌入向量和文本输入 Tacotron，由其生成梅尔频谱图。随后，使用声码器将这些梅尔频谱图转换到时域，从而得到音频信号。

3.1.2. 合成器架构

该合成器是 Tacotron [41], 的一种变体。Tacotron 是一种序列到序列模型，能够根据输入逐帧生成输出。此外，我们还让该合成器以条件编码器生成的嵌入向量为条件，该向量包含有关所需输出情感和说话人声音的信息。条件嵌入与合成器的文本嵌入进行拼接，然后通过解码器进行合成

输出梅尔频谱图。该合成器使用 80 通道梅尔频谱图进行训练，窗口大小为 50 ms，跳步长度为 12.5 ms。该合成器通过三个卷积层将输入字符编码为隐藏表示，这些卷积层能够学习类似 n 元语法的长期上下文。随后，这些卷积层的输出被传递至一个包含 256 个单元的双向长短期记忆层，该层从这些类似 n 元语法的特征中学习时间依赖关系。长短期记忆层返回一个能够完整表示输入文本序列的编码向量。该向量与来自编码器的情感和说话人嵌入向量进行拼接。

值得注意的是，在此阶段，编码器已经训练完成，其权重不会更新。文本、说话人和情感嵌入的组合被传递给解码器，以生成梅尔频谱图。解码器架构包含一种位置敏感注意力机制，可将输入嵌入转换为固定长度的向量。前一步的输出帧经过两个全连接层后，与嵌入向量拼接，以确保生成的序列不会出现时间伪影。随后，该向量被传递到两个 LSTM 层，并通过线性变换生成梅尔频谱图的下一帧。该 LSTM 的输出也会被投影为一个标量，用作停止标记，指示何时停止生成后续帧。梅尔频谱图生成后，会被传递到一个称为 PostNet 的 5 层卷积网络中，以改善整体重建效果。

3.1.3. 声码器

传统上，Griffin-Lim 算法 [9] 用于从频谱图生成时域音频，但速度较慢，且输出语音缺乏自然感。为解决这一问题，我们采用基于 WaveRNN 架构的声码器[14], 该架构是一种速度更快、性能更强的循环网络，适用于高保真音频的序列建模。它利用残差卷积和门控循环单元层，从梅尔频谱图逐帧生成时域音频信号

3.2. 情感分类器

为了评估合成的情感，我们训练了一个用于语音情感识别的深度神经网络（DNN）。我们实现了一个基于卷积神经网络（CNN）的分类器，该分类器由一个卷积层、一个批归一化层和一个位于 Softmax 层之前的稠密层组成。分类器以梅尔频率倒谱系数（MFCCs）作为输入。CNN 层从输入特征中学习高层特征，然后由稠密层将其转换到更具判别性的空间，以便在经过归一化层后更好地进行情感分类。

4. 实验方案

本节介绍数据集、输入特征和模型训练的详细信息。

4.1. 数据集

我们使用Librispeech数据集 [29]训练文本转语音模型。该数据集包含来自不同说话人的1000小时语音数据，采样率为16 kHz。对于情感嵌入，我们使用了情感语音数据库（EVD） [1] 和多伦多情感语音集（TESS）, [7], 其中包含六位不同的说话人，他们分别以不同的情感朗读不同的句子。我们进行了多项实验，以评估模型的性能。在情感分类实验中，我们使用了瑞尔森情感语音与歌曲视听数据库（RAVDESS） [28] 和TESS。对于跨语料库情感分类，我们使用了CREMA-D [6], SAVEE [12], EmoDB [5], 以及合成音频。这些数据集的详细信息见表1。为确定平均意见分，我们使用了一位来自Librispeech的说话人，以及EVD和TESS中的所有说话人，并选取了两个未包含在训练集中的样本。在情感分类实验中，我们采用说话人独立的情感分类方式。我们随机选取CREMA-D中70%的数据用于训练、10%用于验证、20%用于测试。在情感分类实验中，完整语料库（包括RAVDESS和EmoDB）用作测试集；在跨语料库情感分类实验中，SAVEE数据集用作测试集。

4.2. 输入特征

Tacotron 接收文本字符串作为输入，文本字符串由字符序列组成。每个字符都会被编码为独热编码向量，并嵌入连续向量中。Tacotron 的另一个输入是条件嵌入向量，其中包含说话人和情感信息。该向量由编码器生成；编码器接收说话人音频作为输入，并将其转换为梅尔频率倒谱系数（MFCCs）。这些 MFCCs 包含 40 个对数滤波器组和 80 帧，且不使用重叠窗口。为了生成 t-分布随机邻域嵌入

（t-SNE）合成音频的图，我们使用 [13] 中介绍的模型对合成音频进行编码。该模型的输入同样是包含 40 个对数滤波器组和 80 帧、且不使用重叠窗口的 MFCCs，最终得到一个 80x40 维的特征向量。该模型还用于评估说话人验证中的等错误率（EER）。在情感分类和跨语料库情感分类中，我们使用包含 40 个对数滤波器组且跳步长度为 64 毫秒的 MFCCs。按照先前工作 [32]中的方法，我们对 MFCC 数组进行转置，并沿其水平轴计算算术平均值。

4.3. 语音合成模型训练

首先，在 Librispeech 数据集上训练编码器，使其学会生成能够区分不同说话人的说话人嵌入。编码器以梅尔频谱图为输入，输出一个大小为 256 的嵌入向量。根据这些嵌入向量构建相似度矩阵，使每一列都包含同一位说话人的嵌入向量，并使列中所有单元格的余弦相似度最大化、行中所有单元格的余弦相似度最小化。列中的余弦相似度之所以最大化，是因为列中包含同一个人的音频嵌入，而行中的余弦相似度则最小化

表1所考虑的所有数据集的描述。

Name	数量 说话人	数量 语句
CREMA-D	91	7,442
EmoDB	10	535
EVD	5	7,590
Librispeech	2484	281,241
REVDESS	24	7,356
SAVEE	4	480
TESS	2	2,800

是因为行中包含不同人的音频嵌入。这样，同一个人的嵌入彼此相似，而不同人的嵌入则彼此不同。

在 Librispeech 数据上训练编码器后，再使用 EVD 和 TESS 数据集对其进行微调，以生成能够区分不同情感的嵌入向量。此次构建相似度矩阵时，同一说话人的同一种情感所生成的嵌入向量放在同一列，其他情感则放在其他列。对所有说话人都执行此操作，然后最大化同一列内的余弦相似度，并最小化不同列之间的余弦相似度。这样做是为了增大同一个人的不同情感之间的距离，因此通过累加不同列之间的余弦相似度，而非同一列内的相似度，来使余弦相似度最小化。我们使用的批量大小为 30，学习率为 10^{-4}。

训练期间，首先使用 Librispeech 数据训练合成器模型，使其能够学习根据多种多样的文本生成不同说话人的音频。这是因为 EVD 和 TESS 数据集合起来只有六位说话人。当合成器经过充分训练，能够生成听起来像参考说话人的音频后，我们再使用 EVD 和 TESS 数据集对其进行微调，通过训练使其生成不同情感的梅尔频谱图。训练合成器时，我们使用的学习率为 10^3，并按指数规律衰减至 10^{-5}，批量大小为 30。优化器采用 Adam优化器，参数为$\beta_1 = 0.9$、$\beta_2 = 0.999$和 $\epsilon = 10^{-6}$。教师强制率设为 1（即向模型输入原始的前一序列，以预测下一序列）。通过最小化预测梅尔频谱图的均方误差进行优化。

5. 结果

在本节中，我们从合成说话人的相似度和合成情感的细粒度程度两方面评估所提出模型的性能。

5.1. 评估合成语音质量

为评估合成语音的质量，我们进行了多项实验。以下介绍这些实验的具体内容。

表 2不同合成器的说话人验证等错误率。

	样本数	EER
情感 + 语音转换文本转语音	100	**0.16**
基线情感转换文本转语音	100	0.24
语音转换文本转语音	100	0.10
真实音频	100	0.04

表 3平均意见分（MOS）及95%置信区间。

情感	愤怒	快乐	Sad	中性	总体
录制	4.6	4.50	4.50	4.60	4.55
基线	2.80	3.10	2.70	4.20	3.20
所提方法	3.60	3.70	3.80	4.10	3.80

5.1.1. 说话人验证

我们使用说话人验证评估了合成音频与真实语音之间的说话人相似度，并按照 [13] 测量等错误率（EER）。EER 用于衡量说话人验证系统的性能，方法是在不同灵敏度水平下比较误拒率（FRR）和误接受率（FAR）。FRR 与 FAR 相等时的数值即为 EER。计算 EER 时，我们使用了 100 个音频样本，其中 40 个为合成样本。我们仅在系统中注册合成说话人，并计算 EER。通过使用多说话人 Tacotron 模型进行语音转换，我们实现了 0.10% 的 EER [13]。我们还使用基础模型生成了情感音频样本，结果说话人验证模型在这些合成音频上的 EER 为 0.24%。相比之下，使用所提出的模型同时进行情感转换和语音转换时，我们实现了 0.16% 的 EER。使用 [13] 中的方法处理真实样本时，EER 为 0.04%。我们在表 2 中比较了这些模型的 EER。

5.1.2. 听力实验

我们进行了平均意见分（MOS）评估，以衡量合成语音的质量。我们请接受过研究生教育的受试者聆听音频后，根据以下标准打分： 1 = 差；2 = 较差；3 =一般； 4 =良好；以及 5 =优秀。表3所示的结果表明，与基线模型相比，所提出的模型能够合成高质量的情感语音。与基线相比，所提出的模型显著提高了愤怒、悲伤和快乐等情感的 MOS 得分。不过，在自然语音方面，其 MOS 得分略低于基线。这可能是因为基线模型专为生成自然语音而设计，因此在中性语音方面表现更好。尽管如此，我们提出的模型在所有情感类别上都表现出色。读者可以通过此网址收听生成语音的样本[1]。

[1] https://emotaco.github.io/Emotional_Tacotron/

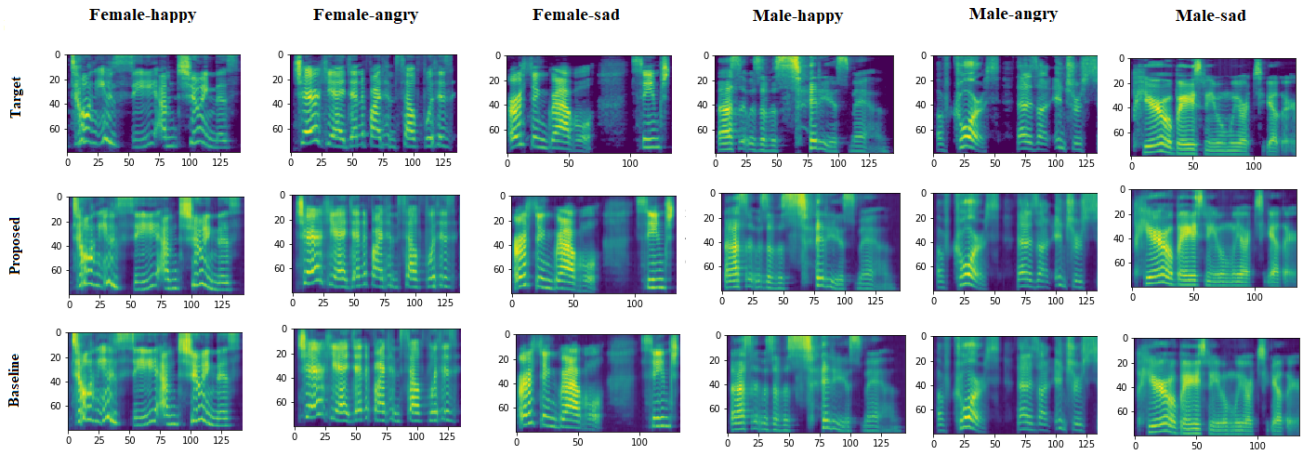

图 2： 比较　　男性和女性音频中不同情感的目标梅尔频谱图与合成梅尔频谱图的比较　　s.

5.1.3. 说话人和情感可视化

在这项实验中，我们没有使用教师强制，而是按照推理部分所述生成音频。我们绘制了基线模型和所提出模型合成的不同情感的梅尔频谱图，如图2所示，并将结果与目标梅尔频谱图进行了比较。与基线模型不同，我们提出的模型没有对生成的梅尔频谱图进行平滑处理，而这种处理有助于通过 WaveRNN 声码器生成质量更好的情感语音。

为进行评估，我们展示了 t-SNE 图。该图通过使用说话人验证模型作为编码器，对利用合成器输出样本生成的嵌入向量进行处理而生成。需要注意的是，说话人编码器并未使用合成器进行训练，因此并未针对合成器输出进行优化。我们使用基础论文中的模型生成了情感音频合成结果的 t-SNE 图，并将结果与所提出的模型进行了比较。男性和女性声音的合成语音 t-SNE 图分别如图 3 和 4所示。这些图表明，与基础模型相比，我们的模型能够合成具有明显差异的情感。可以观察到，不同情感彼此分离，而相似情感聚集在一起，表明情感之间具有相似性。

由于愤怒情绪相比悲伤和快乐情绪具有更强的表现力，而后两者属于语调变化，因此愤怒情绪的聚类与快乐情绪的聚类距离更远。我们还在图5中使用所提出的模型展示了中性语音中多个说话人的 t-SNE 图。图中不同说话人形成了明显的聚类，表明该模型能够有效学习多个说话人的嵌入。

.

5.2. 增强语音情感识别（SER）

在本节中，我们使用合成语音来增强语音情感识别系统。我们使用语料库

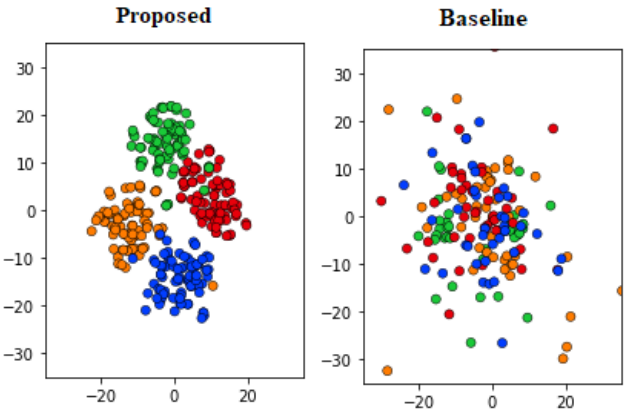

图 3： 使用基线模型和我们提出的模型对不同情感的男性音频进行 t-SNE 可视化对比，结果表明，我们的模型在情感识别方面表现更佳。

和跨语料库设置进行了评估。接下来将介绍这些实验的结果。

5.2.1. 语料库内评估

我们使用 RAVDESS 和 TESS 数据集进行评估。我们合并两个数据集，然后按 70:10:20 的比例随机划分数据，分别用于训练、验证和测试。模型训练了 45 个训练轮次。图6比较了真实语音和合成语音上的说话人识别结果。合成语音的准确率达到 80%，而真实语音测试集的准确率为 92.4%。这表明我们的模型能够合成输出语音的情感特征。我们还使用合成数据扩充分类器，并用真实语音和合成语音数据进行训练。准确率达到 94.6%，优于仅使用真实数据训练的分类器。这项实验表明，我们的模型还可用于

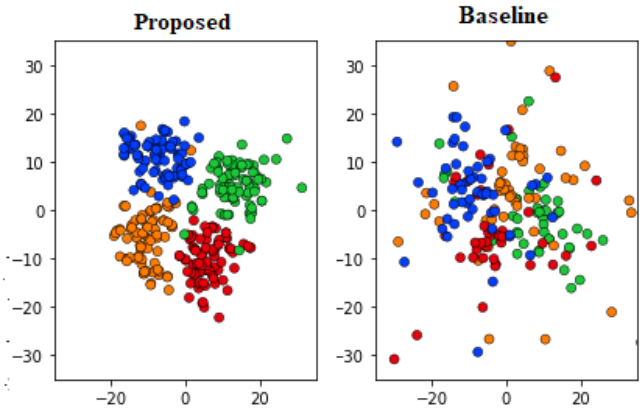

图4： 使用基线模型和我们提出的模型对不同情感的女性音频进行 t-SNE 可视化对比，结果表明，我们的模型在情感表现方面更出色。

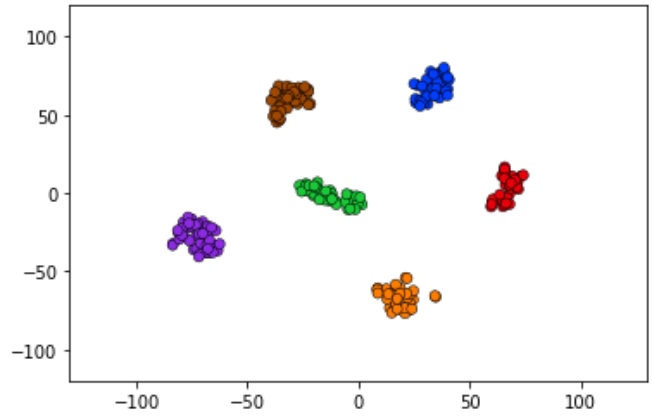

图5： 合成结果中说话人语音的 t-SNE 图显示，各个说话人的语音明显聚类在一起。

生成额外的音频数据，用于扩充说话人识别系统并提升其性能。

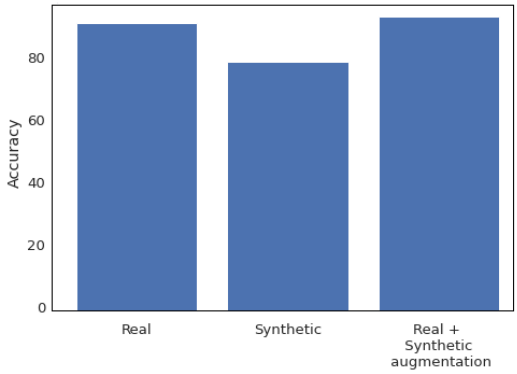

图6： 柱状图显示，我们合成音频的情感与真实音频的情感被分类模型划分得几乎相同。

我们还在图中绘制了混淆矩阵 7 ，用于训练集分别采用真实音频、合成音频以及真实数据与合成数据的组合时的情感分类。混淆矩阵显示，使用合成数据增强的模型能够更好地对语音情感进行分类。其他情感的准确率也有所提高，但最显著的提升体现在快乐情绪的分类上。

5.2.2. 跨语料库语料库评估

我们还通过开展跨语料库情感分类，评估了使用合成数据进行增强的效果。为此，我们实现了一个分类器，其中包含一个长短期记忆网络层、三个全连接层和一个 Softmax 层，用于情感分类。我们还在全连接层之间使用了两个 dropout 层，以学习更具泛化能力的表征。我们根据先前研究的发现选择了模型架构 [22, 23]。我们在从输入音频中提取的梅尔频率倒谱系数特征上训练分类器。模型采用稀疏分类交叉熵损失和 Adam 优化器进行训练，共训练 100 轮。模型使用 CREMA-D数据集以及经合成数据增强的 CREMA-D数据集进行训练，并在 CREMA-D、SAVEE 和 EMODB数据集上进行评估。结果如图8所示，加入合成数据不仅在不对模型进行微调的情况下提高了 SAVEE 和 EMODB数据集上的准确率，也提高了 CREMA-D 测试集上的准确率。

5.2.3. 改变性别和说话人分布

在本实验中，我们比较了使用给定语料库中不存在的新说话人音色进行数据增强的结果。例如，SAVEE 语料库包含四名男性说话人，既可以使用这四名男性说话人的音色创建合成数据，也可以使用其他男性和女性说话人的音色创建合成数据，从而提高数据的多样性并增强语音情感分类效果。我们在表 4 中展示了结果。我们将结果与基线模型进行了比较，该模型未经任何增强训练；同时也比较了对训练数据应用速度扰动的结果。我们遵循 [19]，使用速度扰动数据增强技术创建了两份增强样本副本。我们发现，与基线模型以及广泛使用的速度扰动数据增强技术相比，使用不同说话人音色增强数据有助于提升性能。

6. 结论

本文提出利用情感文本转语音（TTS）系统来增强语音情感识别（SER）系统。我们介绍了一种基于 Tacotron 的多说话人情感 TTS 系统，用于以不同说话人音色生成合成语音，并将其用于语音情感识别中的数据增强，以提升性能。结果表明，所提出的 TTS 系统能够生成高质量且具有情感区分性的样本。当我们使用这些增强样本来增强语音情感识别系统时，我们发现，与语音情感识别中广泛使用的语音数据增强技术相比，使用不同情感语音中的合成数据

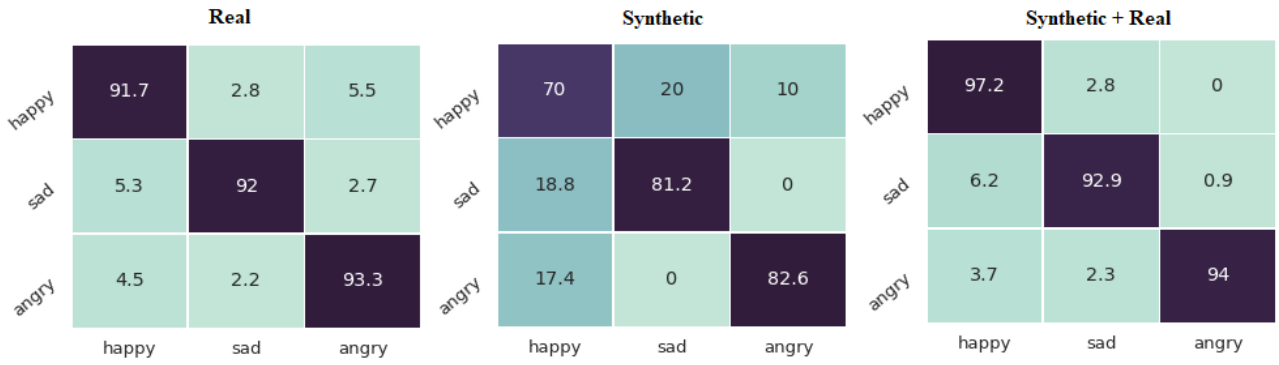

图7：真实、合成及合成与真实音频组合的测试集混淆矩阵

数据可提升情感分类

或真实、合成以及合成与真实音频组合的测试集。添加合成曾道人效果。

c

表4采用不同合成数据分布时，各说话人和性别的结果

数据集	准确率（%）				
	基线	速度扰动增强	男性说话者合成数据	女性说话人合成数据	女性和男性合成数据
SAVEE	65.4	66.8	68.2	69.4	72.3
CREMA-D	68.3	70.1	72.7	72.9	74.3

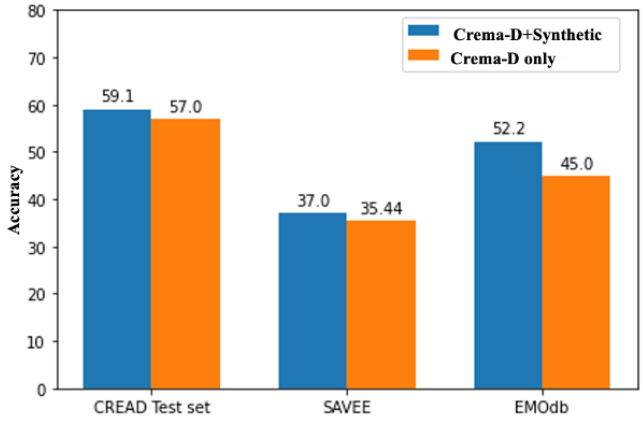

图 8： 跨语料库场景下的测试结果，表明在模型中加入合成数据后性能有所提升。

有助于提升性能。未来的工作将着重研究如何学习一种统一的嵌入表示，以便为所有人控制风格和情感，而不受年龄、背景和性别的影响。

参考文献

[1] Adigwe, A., Tits, N., Haddad, K.E., Ostadabbas, S., Dutoit, T., 2018年。《情感语音数据库：实现对语音生成系统中情感维度的控制》。arXiv 预印本 arXiv:1806.09514。

[2] Aldeneh, Z., Provost, E.M., 2017年. 使用区域显著性进行语音情感识别，载于：2017年 IEEE 声学、语音与信号处理国际会议（ICASSP），IEEE。第 2741–2745 页。[3] Arik, S.O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., 等，2017年. Deep voice：实时神经文本转语音。arXiv 预印本 arXiv:1702.07825。[4] Baird, A., Amiriparian, S., Milling, M., Schuller, B.W., 2021年. 在公开演讲场景中利用注意力机制和长短期记忆循环神经网络方法进行情感识别，载于：2021年 IEEE 口语语言技术研讨会（SLT），IEEE。第 397–402 页。[5] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., 2005年. 德语情感语音数据库，载于：第九届欧洲语音通信与技术会议。[6] Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R., 2014年. CREMA-D：众包情感多模态演员数据集。《IEEE情感计算汇刊》5，377–390。[7] Dupuis, K., Pichora-Fuller, M.K., 2010年. 多伦多情感语音集（TESS）——年轻说话人_快乐。[8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014年. 生成对抗网络。《神经信息处理系统进展》27。[9] Griffin, D., Lim, J., 1984年. 基于修正短时傅里叶变换的信号估计。《IEEE声学、语音与信号处理汇刊》32，236–243。[10] Heigold, G., Moreno, I., Bengio, S., Shazeer, N., 2016年. 端到端文本相关说话人验证，载于：2016 年 IEEE 声学、语音与信号处理国际会议（ICASSP），IEEE。第 5115–5119 页。[11] Hunt, A.J., Black, A.W., 1996年. 在大型语音数据库中使用单元选择的拼接式语音合成系统，载于：1996年 IEEE 声学、语音与信号处理国际会议论文集，IEEE。第 373–376 页。[12] Jackson, P., Haq, S., 2014年. 萨里情感视听语音数据库（SAVEE）。萨里大学：英国吉尔福德。

[13] Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Nguyen, P., Pang, R., Moreno, I.L., Wu, Y., 等，2018年. 从说话人验证迁移学习到多说话人文本转语音合成，载于：《神经信息处理系统进展》，第 4480–4490 页。 [14] Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., Oord, A.v.d., Dieleman, S., Kavukcuoglu, K., 2018年. 高效神经音频合成。arXiv 预印本 arXiv:1802.08435。 [15] Kim, T.H., Cho, S., Choi, S., Park, S., Lee, S.Y., 2020年. 利用文本转语音和多任务学习进行情感语音转换，载于：ICASSP 2020-2020 IEEE 声学、语音与信号处理国际会议（ICASSP），IEEE。第 7774–7778 页。 [16] Ko, T., Peddinti, V., Povey, D., Khudanpur, S., 2015年. 用于语音识别的音频数据增强，载于：国际语音通信协会第十六届年会。 [17] Kwon, O., Jang, I., Ahn, C., Kang, H.G., 2019年. 一种用于端到端情感语音合成的有效风格标记权重控制技术。《IEEE信号处理快报》26，1383–1387。 [18] Łańcucki, A., 2021年. FastPitch：结合音高预测的并行文本转语音，载于：ICASSP 2021-2021 IEEE 声学、语音与信号处理国际会议（ICASSP），IEEE。第 6588–6592 页。[19] Latif, S., 2020年. 用于提升语音情感识别的深度表征学习。博士生论坛，Interspeech 2020。 [20] Latif, S., Cuayáhuitl, H., Pervez, F., Shamshad, F., Ali, H.S., Cambria, E., 2022a. 音频应用深度强化学习综述。《人工智能评论》，1–48。 [21] Latif, S., Khalifa, S., Rana, R., Jurdak, R., 2020a. 面向语音情感识别应用的联邦学习，载于：2020年第19届ACM/IEEE传感器网络信息处理国际会议（IPSN），IEEE。第 341–342 页。 [22] Latif, S., Qadir, J., Bilal, M., 2019a. 用于跨语言语音情感识别的无监督对抗域自适应，载于：2019年第八届情感计算与智能交互国际会议（ACII），IEEE。第 732–737 页。 [23] Latif, S., Rana, R., Khalifa, S., Jurdak, R., Epps, J., 2019b. 直接从原始语音建模语音情感，载于：Interspeech 2019 论文集，第 3920–3924 页。网址：http://dx.doi.org/10.21437/Interspeech.2019-3252，doi：10.21437/Interspeech.2019-3252.[24] Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., Schuller, B.W., 2021年. 语音情感识别深度表征学习综述。《IEEE情感计算汇刊》。[25] Latif, S., Rana, R., Khalifa, S., Jurdak, R., Schuller, B.W., 2020b. 提升语音情感识别抗噪鲁棒性、抵御对抗攻击并适用于跨语料库场景的深度架构。Interspeech 2020 论文集，2327–2331。[26] Latif, S., Rana, R., Khalifa, S., Jurdak, R., Schuller, B.W., 2022b. 利用增强辅助数据进行多任务学习，以提升语音情感识别性能。《IEEE情感计算汇刊》。[27] Lee, Y., Rabiee, A., Lee, S.Y., 2017年. 情感端到端神经语音合成器。arXiv 预印本 arXiv:1711.05447。[28] Livingstone, S.R., Russo, F.A., 2018年. 瑞尔森情感语音与歌曲视听数据库（RAVDESS）：一套用于北美英语的动态多模态面部与声音表情数据集。《公共科学图书馆·综合》13，e0196391。[29] Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015年. LibriSpeech：基于公有领域有声读物的自动语音识别语料库，载于：2015年 IEEE 声学、语音与信号处理国际会议（ICASSP），IEEE。第 5206–5210 页。[30] Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D., Le, Q.V., 2019年. SpecAugment：一种用于自动语音识别的简易数据增强方法。Interspeech 2019 论文集，2613–2617。[31] Parthasarathy, S., Busso, C., 2020年. 利用阶梯网络进行半监督语音情感识别。《IEEE/ACM音频、语音与语言处理汇刊》28，2697–2709。[32] de Pinto, M.G., Polignano, M., Lops, P., Semeraro, G., 2020年. 情感-使用深度神经网络和梅尔频率倒谱系数从口语中提取理解模型，载于：2020年IEEE演化与自适应智能系统会议（EAIS），IEEE。第1–5页。[33] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.Y., 2020. FastSpeech 2：快速、高质量的端到端文本转语音，载于：国际学习表征会议。[34] Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.Y., 2019. FastSpeech：快速、稳健且可控的文本转语音，载于：神经信息处理系统进展，第3171–3180页。[35] Skerry-Ryan, R., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R.J., Clark, R., Saurous, R.A., 2018. 利用Tacotron实现表现性语音合成的端到端韵律迁移。arXiv预印本 arXiv:1803.09047。[36] Sun, G., Zhang, Y., Weiss, R.J., Cao, Y., Zen, H., Wu, Y., 2020. 用于可解释语音合成的全分层细粒度韵律建模，载于：ICASSP 2020——2020年IEEE声学、语音与信号处理国际会议（ICASSP），IEEE。第6264–6268页。[37] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., 2000. 基于HMM的语音合成中的语音参数生成算法，载于：2000年IEEE声学、语音与信号处理国际会议论文集（目录号00CH37100），IEEE。第1315–1318页。[38] Um, S.Y., Oh, S., Byun, K., Jang, I., Ahn, C., Kang, H.G., 2020. 具有丰富且细粒度控制的情感语音合成，载于：ICASSP 2020——2020年IEEE声学、语音与信号处理国际会议（ICASSP），IEEE。第7254–7258页。[39] Variani, E., Lei, X., McDermott, E., Moreno, I.L., Gonzalez-Dominguez, J., 2014. 用于小型文本相关说话人验证的深度神经网络，载于：2014年IEEE声学、语音与信号处理国际会议（ICASSP），IEEE。第4052–4056页。[40] Wan, L., Wang, Q., Papir, A., Moreno, I.L., 2018. 用于说话人验证的广义端到端损失，载于：2018年IEEE声学、语音与信号处理国际会议（ICASSP），IEEE。第4879–4883页。[41] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., 等，2017. Tacotron：迈向端到端语音合成。arXiv预印本 arXiv:1703.10135。[42] Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., Saurous, R.A., 2018. 风格标记：端到端语音合成中的无监督风格建模、控制与迁移。arXiv预印本 arXiv:1803.09017。[43] Yasuda, Y., Wang, X., Takaki, S., Yamagishi, J., 2019. 探究采用自注意力机制的增强型Tacotron文本转语音系统在音高重音语言中的应用，载于：ICASSP 2019——2019年IEEE声学、语音与信号处理国际会议（ICASSP），IEEE。第6905–6909页。[44] Zen, H., Tokuda, K., Black, A.W., 2009. 统计参数语音合成。《语音通信》51，1039–1064。[45] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2018. Mixup：超越经验风险最小化，载于：国际学习表征会议。