分类号：

学校代码：10112

密　　级：

太原理工大学

硕士学位论文

（学术学位）

基于说话人特征双阶段迁移学习的情感语音克隆技术研究

姓　　名：李因

学　　号：2020510428

培养单位：计算机科学与技术学院（大数据学院）

学　　科：计算机科学与技术

研究方向：智能信息处理、语音合成

指导教师：相洁　教授

论文提交日期：2023 年 6 月

A Dissertation Submitted to

Taiyuan University of Technology

In partial fulfillment of the requirement

For the degree of Master

Research on Emotional Voice Cloning Technology Based on Two-Stage Transfer Learning of Speaker Features

By

Nan Li

College of Computer Science and Technology (College of BigData)

June 2023

学位论文答辩信息表

论文题目	基于说话人特征双阶段迁移学习的情感语音克隆技术研究		
答辩日期	2023 年 5 月 26 日	答辩秘书	李颖
学位论文答辩委员会成员			
姓名	职称	工作单位	备注
李茹	教授	山西大学	主席
王彬	教授	太原理工大学	委员
邓红霞	副教授	太原理工大学	委员

摘 要

智能语音技术是人工智能领域的重要组成部分，其研究内容主要涉及到语音识别技术和语音合成技术。尽管当前的语音合成技术已经取得了一定的成果，但由于缺乏完备的目标说话人数据集，其结果并不能完全适应于所有的目标说话人。

语音克隆是一种将文本转换为目标说话人语音的技术。该技术主要采用两种方法，即说话人自适应和说话人编码。说话人自适应方法是指利用少量目标说话人的语音数据微调训练好的多说话人语音合成模型。然而，此方法通常需要执行上千步的计算才能达到高质量的效果。说话人编码是利用说话人识别技术中的说话人编码器提取目标音频的说话人向量，并使用该向量来调节语音的生成过程。虽然这种方法可以适应大多数的实时场景，但由于在训练模型的过程中未考虑音频的情感韵律特征，因此生成的克隆语音存在自然度较低以及缺乏表达能力的缺点。

为解决上述提及到的目前已有的语音克隆模型中所存在的问题，本研究主要进行了以下工作：

第一，鉴于传统的语音合成方法存在模型结构和训练方式的限制，难以生成任意目标说话人的语音。针对此问题，本研究提出了一种基于单阶段迁移学习的语音克隆模型。该模型通过利用在说话人识别框架下预训练的说话人编码器，提取目标音频中的说话人声纹特征，并利用此特征对语音的生成过程进行调节。具体而言，该模型使用语音合成框架中的合成器生成相应的梅尔频谱图，并对其做出改进，将说话人特征与字符特征向量拼接，对梅尔频谱图的特征进行调整，从而使生成的语音具有目标说话人的信息。最后，通过声码器对梅尔频谱图进行处理，生成克隆语音。

第二，在基于单阶段迁移学习的语音克隆模型中，说话人编码器在说话人识别框架下进行训练，训练过程中仅针对说话人身份特征的提取进行学习，并未考虑到音频的情感韵律特征，因此生成的克隆语音自然度较差，缺乏较好的表达能力。针对此问题，本研究提出了对说话人编码器进行说话人训练以及情感训练的双阶段训练方式。在音色克隆阶段的训练中，使用说话人识别语料库进行说话人训练，学习提取身份特征；在情感克隆阶段的训练中，使用情感语料库并以情感作为标签，学习对音频中情感特征的提取。经过说话人训练以及情感训练后，说话人编码器可以提取说话人情感特征，进而提升克隆语音的表达能力。

第三，在基于双阶段迁移学习的情感语音克隆方法中，进行情感克隆阶段的训练时，对目标音频进行特征提取时仅考虑情感特征，可能会影响克隆语音的音色相似性，并且结果与目标情感存在差距。为解决此问题，本研究对现有的说话人编码方式进行了研究，同时对音频的相关情感特征进行分析，提出了 e-vector 说话人模型。与传统说

话人编码器不同的是，e-vector 在根据目标音频提取相关特征时，不再是单一的对某个特征进行学习，而是通过多特征融合的方式同时学习说话人身份特征以及情感特征，避免了模型先后学习不同特征间产生的影响。最后，使用 e-vector 说话人编码器对情感语音克隆模型进行了改进。

研究分别基于以上三个工作内容进行了实验设计，并与传统的语音克隆方法进行了比较。由实验结果可知，e-vector 较传统说话人模型，音色相似度最高提升了 0.26，情感相似度最高提升了 0.48。由此可以得出，本研究提出的情感语音克隆方法，可以有效解决传统语音克隆模型生成的克隆语音自然度较低的问题，更好地提升了克隆语音的表达能力。

本研究的工作受到了国家自然科学基金（61876124，61873178）以及山西省科技厅基础研究项目（20210302123129，20210302124166，20210302123099）的支持。

关键词：说话人特征；双阶段；迁移学习；情感语音克隆；语音合成

ABSTRACT

Intelligent speech technology is a crucial aspect of the artificial intelligence field, with research primarily concentrating on speech recognition and synthesis. Although the current speech synthesis technology has made some achievements, the lack of a comprehensive dataset for each individual speaker means that the results may not be adaptable to all target speakers.

Voice cloning is a technique that converts text into target speaker's speech, with two methods: speaker adaptation and speaker encoding. Speaker adaptation involves fine-tuning a multi-speaker speech synthesis model using a small amount of target speaker speech data. However, this method usually requires thousands of computations to achieve high-quality results. Speaker encoding uses a speaker encoder from speaker recognition to extract the speaker vector of the target audio and adjust the generation process of speech using this vector. Although this method can be adapted to most real-time scenarios, the emotional rhythmic features in the audio are not taken into account during the training of the model, which results in low naturalness and limited expression ability of the generated cloned speech.

To address the problems with existing voice cloning models mentioned above, the main work of thesis includes:

First, due to the limitations of the structure of the model and the training process, the traditional text-to-speech cannot generate the speech of all target speakers. To address this problem, a transfer learning-based voice cloning model is constructed in thesis. The model extracts speaker embeddings from the target audio by using a speaker encoder pre-trained in a speaker recognition framework and uses this embedding to regulate the speech generation process. Specifically, the model utilizes the synthesizer within the speech synthesis framework to produce the relevant Mel Spectrogram and further enhances it by combining the speaker embedding with the character feature vector and adjusting the resulting features. This results in the generated speech having the desired characteristics of the target speaker. Lastly, the vocoder processes the Mel spectrogram to produce the cloned speech.

Second, the model of voice cloning based on single-stage transfer learning involves training the speaker encoder under the speaker recognition framework. However, this process only focuses on extracting the speaker's identity features and does not consider the emotional rhythmic features present in the audio. As a result, the synthesized cloned speech lacks naturalness and expressiveness. To overcome this issue, the thesis proposes a two-stage training method for speaker encoders that incorporates both speaker training and emotion training.

During timbre cloning, a speaker recognition corpus is used to learn how to extract identity features, while during emotion cloning, an emotion corpus is utilized to learn the extraction of emotion features from audio by using emotions as labels. By combining speaker and emotion training, the proposed method allows the speaker encoders to extract both speaker identity and emotion features, thereby improving the expressiveness of the cloned speech.

Third, the method of emotional voice cloning based on two-stage transfer learning only considers emotional features during the training of the emotional cloning stage. This approach has an impact on the timbre similarity of the cloned speech, and there may be deviations from the target emotion. To address this issue, this thesis investigates existing speaker encoding methods and proposes an e-vector speaker model after analyzing the audio's emotional features. Unlike traditional speaker encoders, the e-vector model simultaneously learns both speaker identity features and emotion features through multi-feature fusion, avoiding any impact between different features learned by the model successively. Finally, the emotional voice cloning model is improved using the e-vector speaker encoder.

In the thesis, experiments were conducted to verify the three elements of the proposed emotional voice cloning approach and compared with traditional voice cloning methods. From the experimental results, it can be seen that e-vector improves the speaker similarity of the generated cloned speech by up to 0.26 and the emotion similarity by up to 0.48 compared to the traditional speaker model. As a result, it can be concluded that the proposed emotion speech cloning method effectively addresses the problem of low naturalness in cloned speech generated by traditional speech cloning models and enhances the expressive ability of cloned speech.

The thesis was supported by research grants from the National Natural Science Foundation of China (61876124, 61873178), Shanxi Provincial Department of Science and Technology Basic Research Project (20210302123129, 20210302124166, 20210302123099).

Keywords: speaker features; two-stage; transfer learning; emotional voice cloning; text-to-speech

目　录

图索引

表索引

第 1 章 绪论

传统的语音合成方法受限于模型与训练方式，难以将合成语音扩展至所有说话人。因此，研究者们开始关注语音克隆技术，它能合成具有目标说话人音色的语音。但由于其在训练过程中仅关注说话人，生成的克隆语音缺乏丰富的表达能力。本章节主要介绍了情感克隆语音的研究背景、相关技术的国内外研究现状以及本研究的创新点。

1.1 研究背景与意义

语言是人类最重要的交际工具，是人们进行沟通交流的主要表达方式。其显著优势在于使得交往过程变得更加自然、便捷，并且有助于传递准确无误的信息。随着工业化的进步，机器在社会中的重要性不断提升。特别是人工智能等技术的广泛应用，为人类与机器之间的交流方式带来了显著的改善和升级。因此，如何改善人类与机器之间的关系，以提升机器的操作效率并使其成为人类生产和社会活动的有力助手，已经成为一个更为重要的问题。随着智能语音技术的普及和发展，人们逐渐认识到通过语言与机器进行交流已经成为最便捷和高效的方式。

语音是语言的一种声学表现形式。语音信号处理的发展最早可以追溯到 19 世纪 70 年代，当时电话的发明使得人类首次能够通过声电转换实现语音的远距离传输[1]。近年来，随着新兴自然语言技术的涌现和已有技术的不断成熟，智能语音技术已经进入了成熟阶段，并在多个领域迅速发展：在智能驾驶场景中，车载语音技术提高了驾驶员的行车体验；金融行业中，智能语音客服的研究和应用不断深入；智慧且高效的学习环境推动了信息化教育的快速发展；此外还进一步提升了智能家居的便利性，增强了人们的幸福感等。智能语音技术的研究主要集中在语音合成技术与语音识别技术上。其中，语音合成技术是指通过分析、翻译、合成等过程，把文本转化为语音波形，并模拟人类语音表达方式的技术[2]。目前主要的语音合成技术包括基于规则的方法、基于数据的方法、基于统计的方法以及基于深度学习的方法[2]。由于深度学习技术的快速发展，使得感知机、卷积神经网络、循环神经网络、长短时记忆网络等网络结构成功地应用到语音合成技术中，提升了生成语音的质量和自然度，与传统语音合成技术相比展现出了更好的性能，得到了越来越广泛的关注。但是，由于目前的语音合成技术在训练过程中使用的是一定说话人的语料，生成的语音只能针对于非特定说话人[3]。如果要合成特定说话人的音频，就要使用到语音克隆技术。

语音克隆是近年来在语音合成技术基础上衍生的相对新颖的研究方向，它是一种利用机器合成出特定说话人音频的语音合成方法[4]，在生成个性化语音的领域中发挥了

重要作用。目前，语音克隆技术已经取得了较为理想的结果，能够通过给定任意目标说话人的音频和文本，合成具有目标说话人音色的语音。这一技术按照训练中使用目标说话人语音的数量可以分为两种方法。具体而言，第一种方法是需要使用大量目标说话人的语料来训练语音合成模型[5]，以便生成与特定说话人相似的语音。据前人研究表明，为达到良好的效果，此方法需要数十分钟的目标说话人语料进行训练[6]。然而，在实际操作中，收集大量目标说话人的音频是一个非常困难的任务，因此这种方法很少被使用。第二种方法是采用少量目标说话人语料进行训练，该方法有两种实现方式，一是说话人自适应[7]，使用目标说话人的少量音频，在训练好的多说话人语音合成模型的基础上，利用自适应算法对模型参数进行微调，使得模型生成的语音具有目标说话人特征。二是利用迁移学习的方法，在说话人识别技术（Speaker Recognition，简称 SR）基础之上，通过说话人编码器提取目标音频中能代表说话人身份的特征向量，并将此特征向量作为语音合成模型的输入，参与语音的合成[8]。此方法的优越之处在于当模型训练完成之后，不需要对模型进行任何调整，同时对目标说话人的语料数量要求较低，更符合实际应用中的场景要求。

随着研究的深入，语音克隆技术快速发展，并广泛应用于多个领域。具体地，在医疗领域中可以辅助有语言障碍的人进行康复运动的练习。此外有研究表示，在与精神疾病患者进行交流诊断时，采用智能数字替身技术与患者进行互动，可以提升诊断的效率以及质量[9]。近年来，随着线上直播教学的飞速发展，智慧教育得到了极大的推动，同时也对教师工作提出了新的要求和挑战。其原因在于这种教学方式增加了教师的工作强度和压力。为此，语音克隆技术可以有效地辅助老师完成重复性的工作，提高工作效率。与此同时，随着智能终端的快速发展，人们对于个性化语音助手、车载导航定制播报以及智能家居等方面提出了更多的需求。此外，语音克隆技术还可以应用于短片视频、有声电子书和动画配音等领域，为这些领域的发展带来了更多的机遇和挑战。

虽然语音克隆技术已经得到了广泛应用，但其仍存在一些问题亟需解决。目前的语音克隆模型在合成语音中可以保留目标说话人的音色，但是它不提供对除了文本、特定于说话人嵌入的语音以外的其他方面的控制，例如表达能力。音频的表达能力包含多个方面，例如语气、语速、重音和情感的变化等。对于大多数应用而言，对克隆语音的表达能力进行显式控制是必要的。如果未考虑情感因素的变化，生成的克隆语音会显得平淡且缺乏自然性，同时可能导致机器音的出现，从而促使使用者产生不适或排斥的情绪。在配音领域中，合成语音需要根据文本内容所包含的情感变化做出必要的调整，才能实现更好的效果。因此，在语音克隆技术中，合理地处理语音表达的各个方面，特别是情感因素的变化，对于提高语音合成的准确性和自然度非常重要。

综上所述，语音克隆技术的研究是目前语音合成研究领域中非常重要的一个方向。

作为实现个性化语音合成的重要手段，语音克隆技术具有广泛的应用领域和重要的使用价值。在这些领域内，丰富的表达能力是促使其持续良好发展的关键。通过增加情感特征，可以提高合成语音的自然度和适应性，从而推动语音克隆技术的快速发展，并将其应用范围扩大至更多的领域。近年来，研究人员通过在语音克隆模型的基础上添加风格迁移的方法，以提高克隆语音的表达能力[10]。虽然该方法在一定程度上改善了语音的表达能力，但在合成过程中需要输入目标说话人的音频和文本，以及风格参考音频，这增加了输入参数的复杂性。此外，该方法未能完全实现对合成语音明确的情感控制。

为进一步弥补此类研究中的不足，本研究首先在说话人识别技术基础之上，利用迁移学习的方法，构建了由说话人编码器、合成器以及声码器组成的基于单阶段迁移学习的语音克隆模型。然后，为了使得克隆语音具有丰富的表达能力，对说话人编码器的训练方式做出改进。在音色克隆阶段，使用说话人标签对该编码器进行训练，并引入了情感克隆阶段来学习情感特征的表达。对实验结果分析发现，克隆语音的表达能力得到明显改善，但是与目标情感的相似度还有一定差距，同时音色相似度的提升也不显著。对传统说话人编码器的训练过程进行分析，发现其只学习音频中单一方面的特征，因此在音色克隆阶段与情感克隆阶段中分别只考虑了说话人特征与情感特征，而忽略了特征之间可能会产生的影响。为了解决上述问题，本研究提出了 e-vector 说话人编码方式。该方法采用多特征融合的学习方法，并对双阶段训练方式进行改进，以提升合成语音的情感表达能力，同时改善其与目标说话人的音色相似性。

1.2 国内外研究现状

1.2.1 语音合成技术

语音合成技术是智能语音技术领域的一个重要分支。近年来，基于神经网络的语音合成方法取得了迅速发展，现已涌现出多种模型，可以根据输入的目标文本生成自然流畅的合成语音，其平均意见得分（Mean Opinion Score，简称 MOS）已接近于真实语音[11]。

其中，Graves 等人提出了 Tacotron 模型，该模型使用序列到序列模型作为基础架构，并结合注意力机制根据输入文本生成对应频谱，同时使用泛化模型卷积银行与高速公路网络和门控循环单元（Convolutional Bank with Highway Networks and Gated Recurrent Unit，简称 CBHG）从序列中提取高层次特征[12]。基于 Tacotron，Shen 等人进一步改进该模型，提出了 Tacotron2。Tacotron2 使用普通的长短时记忆网络以及卷积层代替泛化模型，成功减小了丢音的概率[11]。由于循环神经网络难以处理长序列依赖关系，Tacotron2 模型在训练速度上存在缺陷。为了解决此类问题，提高并行效率和减

少计算量，Vaswani等人提出了Transformer模型，该模型不再使用循环神经网络或卷积神经网络，而是仅使用自注意力机制，对编解码器模型进行改进，并且不会对实验结果产生负面影响[13]。基于此，Li等人将Transformer模型应用于语音合成中，并根据语音数据的特点进行相应调整，提出了TransformerTTS模型来有效解决Tacotron2模型中所存在的长序列依赖关系难以建模的问题[14]。然而，由于TransformerTTS是自回归模型，对训练数据的语音文本对齐度要求较高，这会影响到合成语音的质量。为了解决这个问题，Ren等人提出了FastSpeech模型，该模型采用Transformer的前馈网络并行生成梅尔频谱图，并引入了持续时长预测器来确保音素序列与频谱序列的长度匹配，从而提高了语音合成的效率[15]。在基于特定说话人模型的基础上，Park等人提出了适用于多说话人的语音合成系统，并且在实验中取得了显著的效果[16]。此外，科大讯飞公司推出了在线语音合成平台，支持多语种、多方言，广泛应用于多种场景。

总体而言，语音合成技术正在不断地发展和创新，在不同的研究方向中都取得了显著的进展。早期的研究中主要关注于降低语音合成模型的训练复杂度和提升生成语音的自然度，并已经取得了令人瞩目的进展。这些研究改善了模型的效率，并使生成的语音质量逐渐接近真实语音。同时，研究者尝试将语音合成模型推广到多说话人情境，但这些模型仍然无法有效地适应于未出现在训练集中的新说话人。传统语音合成模型中缺乏目标说话人的信息，因此无法对生成过程进行调整，这是导致该问题的主要原因。

1.2.2 说话人模型

在语音克隆模型中，使用说话人模型从目标音频中提取说话人特征，该特征决定了合成语音与目标说话人之间的相似度。因此，说话人模型是构建语音克隆模型中重要的组成部分。目前，常见构建说话人模型的方法可以分为基于统计分析和基于深度学习两种。

因子分析（Factor Analysis，简称FA）是一种用于从多个变量中提取具有代表性因素的统计学分析技术，最初应用于社会学和心理学领域。在2004年，Kenny将因子分析方法引入到音频分析中，用于说话人识别任务[17]。然而，随着训练任务难度的增加和音频数据的扩充，简单的因子分析方法已不能满足需求，为此研究人员提出了联合因子分析方法（Joint Factor Analysis，简称JFA）[18]。JFA通过统计方法来估计说话人空间和信道空间，并通过去除信道空间而保留说话人空间的方法，提高语音识别任务中的说话人识别准确性。然而，在JFA建模后，由于统计分析方法的不确定性，信道空间可能仍然包含与说话人相关的信息。为了解决这个问题，Dehak等提出了一种新的方法—i-vector，该方法使用一个总变化空间（Total Variability Space，简称TVS）来描述语音信息，其中定义了说话人因子，并使用该因子来表征说话人的身份特征，应用于说话人识别任务领域[19]。

深度学习是一种机器学习技术，它利用多层神经网络进行自动特征提取和分类。基于深度学习的方法通常具有良好的泛化性能，并且可以从大量的数据中自动地学习特征表示。随着深度学习技术的不断发展和完善，越来越多的研究者开始将其应用于说话人确认领域，以提高识别的准确率和效率。Variani 等人提出一种基于深度神经网络（Deep Neural Network，简称 DNN）的说话人模型，该方法采用 DNN 作为训练的模型结构，极大地推动了神经网络在说话人识别技术中的发展[20]。在训练过程中，利用深度神经网络对输入音频每一帧的说话人特征向量进行处理，将网络最后一层的输出进行累加并计算平均值，用此向量表示说话人模型，并将其命名为 d-vector。在此基础上，为了得到语句级别的特征，Snyder 等人提出了一种基于时延神经网络（Time-Delay Neural Network，简称 TDNN）的说话人识别方法。该方法利用时延神经网络来学习语音信号帧之间的关系，然后通过统计池化层计算帧级别特征，最后利用统计层以及全连接层来构建音频句子级别的特征，将其作为能够代表说话人身份的特征向量，即 x-vector[21]。

总结上述内容，说话人模型的研究主要集中在如何从目标语音中提取能够代表说话人身份的特征，并利用该特征对语音和说话人的所属关系进行判断。虽然早期的研究通过多种方法对说话人模型进行改进，提高了识别的准确率[22]，但是由于在训练过程中只针对音频中的说话人身份特征向量，未考虑相关情感特征，因此不适应于情感语音克隆的目标。

1.2.3 语音克隆技术

语音克隆技术是生成个性化语音的重要手段。Kons 等人提出了一种基于模块化体系结构的神经语音合成方法，该方法以 LPCNet[23]为基础，将深度神经网络与输出的中间信号处理相结合，并使用目标说话人的少量数据对模型进行全部微调，从而解决说话人适应性问题[7]。通过实验，研究人员发现使用五分钟的目标说话人音频可以得到更高质量的结果。此外，Chen 等人提出一种以 WaveNet[24]为基础，使用说话人嵌入层为语音合成模型中的说话人身份建模，并借用元学习的思想提出了三种语音克隆方法。第一种方法是固定模型的其他参数，只更新说话人的嵌入向量。第二种方法是对模型的所有参数进行微调。第三种方法是使用经过训练的神经网络编码器来预测说话人的嵌入。实验结果表明，用第二种方法合成的语音具有最高的自然度[25]。同时，Jia 等人提出了一种基于神经网络的语音合成模型，包括说话人编码器网络、基于 Tacotron2 的合成器网络以及声码器网络。该模型使用迁移学习的方法对说话人编码器进行训练，使得该编码器可以提取目标音频中的说话人嵌入向量，最后利用此向量对语音的生成过程进行调节。此训练方式对数据集要求较低，并且在相对短的时间内即可得到较好的结果[8]。在此基础上，Chen 等人完成了一个跨语言且适应于多说话人的端到端语音合成框架，并通过实验证明了利用中文音频生成带有目标说话人信息的英语音频的可

行性[26]。最后，Jemine 对 Jia[8]等人提出的语音克隆框架进行了详细分析，提出采用各个模块独立训练的方式以提升训练质量，同时对此模型进行了代码复现并开源[27]。

虽然使用说话人编码器作为说话人嵌入层可以提高语音合成模型对多个说话人的适应性，然而，当面临未在该编码器训练集中出现过的说话人时，模型的效果仍然存在差距。为此，Cooper 等人对多说话人语音合成方法进行了研究，并探究了不同类型的说话人模型对合成效果的影响。研究者通过对比实验得出了基于可学习的字典编码（Learnable Dictionary Encoding-based，简称 LDE）[28]在未参与训练的说话人中比 x-vector 的表现更好，并且 LDE 对合成语音的自然度也有一定改善[29]。但是，在测试结果中，未参与训练的说话人与参与训练的说话人之间相似度仍有一定差距。为解决此问题，Arik 等人研究了语音克隆的两种方法：说话人适应和说话人编码，并分析证明了采用少量数据对预训练的多说话人模型进行微调的优势。同时，提出了一种新的说话人编码方式，以减少克隆语音生成所需的时间。在实验中，Arik 等人还提出了一种基于神经网络的说话人分类和说话人验证方法，以评价克隆语音模型的质量[30]。此外，Huang 等人提出将元学习技术应用到语音克隆模型中，通过模型不可知元学习（Model Agnostic Meta-Learning，简称 MAML）算法，来训练多说话人语音合成模型。通过这种训练方式，使得模型可以找到一个合适的元初始状态，以便其快速地适应新的说话人，并且改善了在说话人自适应方法中需要至少上千万步微调才能取得高质量结果的问题[31]。

总之，语音克隆方法的相关研究取得了较大进展，主要集中在改进模型以提高克隆语音与目标说话人之间的相似度，并缩短训练模型所需的时间。尽管这些研究已获得了较好的音色相似度，但是由于在生成语音的过程中未考虑到情感韵律特征，导致克隆语音的自然度较低，缺乏一定的表达能力，但是这样的能力对提高克隆语音的质量来说是必要的。

1.2.4 情感语音合成技术

近年来，情感语音合成技术已经逐渐应用于虚拟人物、机器人、语音助手、自动应答系统等相关领域，以提高用户交互体验的真实性和自然性。该技术在数字化时代具有重要意义，是人工智能领域中不可或缺的一部分。最新的研究也表明，情感语音合成技术正得到不断改进和拓展。

情感语音合成技术中，一种比较直观的方法是将显式的风格标签作为语音合成技术模型的输入条件，以此指导系统生成与标签相匹配的语音。Lee 等人提出了一个基于 Tacotron 的端到端情感语音合成模型，利用上下文向量以及递归神经网络中的残差连接，对 Tacotron 中存在的曝光误差以及注意力排列不规则进行改进，将字符序列以及目标情感标签作为模型的输入生成相应的情感语音[32]。同时，该研究通过实验证明了生成语音的质量与注意力对齐的性能高度相关。然而，利用情感标签只能学习到一种“平均”

的情感表达方式，无法捕捉到语音数据中的细微情感变化。为了解决这个问题，Hu 等人提出了一种根据输入文本和风格向量来生成情感语音的方法，从参考语音中学习提取风格向量，因此不需要相关的情感标签[33]。在训练过程中，该研究通过使用基于互信息的风格内容分离方法，明确估计最小化风格与内容之间的互信息，并使用对抗性训练来解决传统无监督方法训练过程中存在的“内容泄露”问题，保证了合成语音中信息的完整性。尽管利用参考音频实现情感语音合成的风格迁移方法可以提升结果的表现力，但是此种方法所能实现的仅限于特定方面的语音风格迁移和改变。为了解决此问题，Lei 等人提出了一种基于注意力机制的序列到序列的多尺度情感语音合成框架，即 MsEmoTTS[34]。该模型利用多个层级对语音的情感表达能力进行建模，并在中文情感语料库上进行了大量实验，证明了所提出的方法优于以参考音频为基础和基于文本的情感语音合成方法。这些研究虽然提升了情感语音的自然度和表达能力，但是在合成过程中不能对信号的相关特征进行有目的地调整，同时训练时间和计算复杂度也较高。为了解决此类问题，Pamisetty 等人提出了一种考虑韵律参数的端到端情感语音合成模型，即 ProsodyTTS[35]。该方法通过利用神经网络模型学习大量的语音数据，提升泛化能力与自适应能力，并结合统计参数模型以控制基音频率与音素的持续时长，从而调整音调与韵律，提高合成语音的自然度。

从总体上来看，情感语音合成技术的相关研究已经有了显著进展。目前，该技术主要分为三种方法：第一种方法是先生成中性语音，再通过修改其声学特征参数来转换成目标情感语音。虽然这种方法能够更好地控制情感效果，但需要人工干预，因此合成语音的自然度相对较低；第二种方法则是基于情感语料库生成目标情感语音，生成的结果更加自然，但情感表达不够准确；第三种方法在前两者的基础上，对情感语音的声学特征进行参数调整，以提高情感表达的准确性。因此，在本研究中，通过在克隆语音的生成过程中加入情感特征，并对该过程进行调整，以提高克隆语音的表达能力。

1.3 本文主要研究内容

本研究是基于随着新兴自然语言处理技术的涌现以及已有技术的不断成熟，智能语音技术得到了快速发展，国内外对智能语音技术的研究关注度不断上升，生成个性化语音的研究方向又是智能语音技术一个极为重要的部分的大背景下，重点研究了基于说话人特征双阶段迁移学习的情感语音克隆技术。首先，构建了基于单阶段迁移学习的语音克隆模型。同时，为了提升克隆语音的自然度，使其具有丰富的表达能力，对模型中说话人特征提取的方法进行了改进，并采用主客观评测对方法的有效性进行了验证。

本研究包含以下几项工作：

第一，传统的语音合成方法由于模型结构和训练方式的限制，无法将合成语音扩展到任意说话人。最近，随着说话人识别技术的快速发展，传统语音合成方法得到了改进。本研究首先构建了一种基于单阶段迁移学习的语音克隆模型。该方法在说话人识别技术的基础之上，通过迁移学习的方法，利用说话人编码器提取目标音频中的说话人特征，并将此特征作为语音合成框架的输入以调整语音的生成过程。具体而言，研究中采用去除了 WaveNet 的 Tacotron2 模型作为合成器，并对其架构做出了改进，将说话人编码器提取的特征与输入文本所产生的特征向量进行拼接，以获得具有目标特征的梅尔频谱图，并经声码器处理生成最终的克隆语音。

第二，在基于单阶段迁移学习的语音克隆模型中，说话人编码器的作用是针对输入的目标音频提取符合需求的说话人特征，在训练过程中使用说话人识别语料库，采用说话人标签对其进行音色克隆阶段的训练。由此过程可以发现，该编码器仅针对说话人进行训练，并未考虑到音频的情感特征，导致合成语音缺乏较好的表达能力。因此，针对传统说话人编码器不能提取具有情感的说话人特征的问题，本研究提出了一种对该编码器进行说话人训练以及情感训练的双阶段训练方式。在这种方法中，首先对说话人编码器进行说话人训练，使用说话人识别语料库完成对音频中声纹特征的提取；然后，通过利用第一阶段训练的结果对其进行参数初始化，并在此基础之上进行情感克隆阶段的训练，使用情感语料库以情感作为标签对其进行训练，从而学习音频中情感特征的提取。经过该双阶段训练后，说话人编码器可以提取目标音频中相应的说话人情感特征，使得生成的语音不仅具有目标说话人的音色，同时也具有相应的情感表达能力。

第三，本研究采用对说话人编码器进行双阶段训练的方式，成功地改善了克隆语音的情感表达能力。然而，主客观评测结果表明，生成的语音与目标情感仍存在一定差距，并且音色相似度的提升不够显著。在对特征提取的过程进行分析后发现，音色克隆阶段的训练中，说话人编码器根据目标音频提取能够反映说话人身份的动态梅尔频率倒谱系数等特征向量，并对其进行处理以获得最终的说话人特征嵌入。在情感克隆阶段的训练中，说话人编码器根据目标音频进行特征提取时仅关注情感特征。即使在使用音色克隆阶段训练的结果作为初始化参数模型的情况下，如果在训练过程中未对说话人进行考虑，仅关注情感特征，也会对合成语音的音色相似性产生一定的影响。为解决此问题，本研究对现有的说话人编码方式的实现过程进行了研究，同时对音频的情感特征进行分析，提出了在根据情感语料库进行情感阶段的训练时，不仅关注音频中的情感特征，同时提取与说话人身份相关的声纹特征，之后通过多特征融合的方式，得到新的说话人编码方式，将其命名为 e-vector。同时，本研究结合 e-vector 说话人编码方式的特征，对双阶段训练方式做出相应的改进，使得其在提升克隆语音的情

感表达能力的同时，减少对音色相似度产生的影响。

1.4 本文章节安排

本文共分为六个章节来对研究内容进行阐述，如图 1-1 所示。

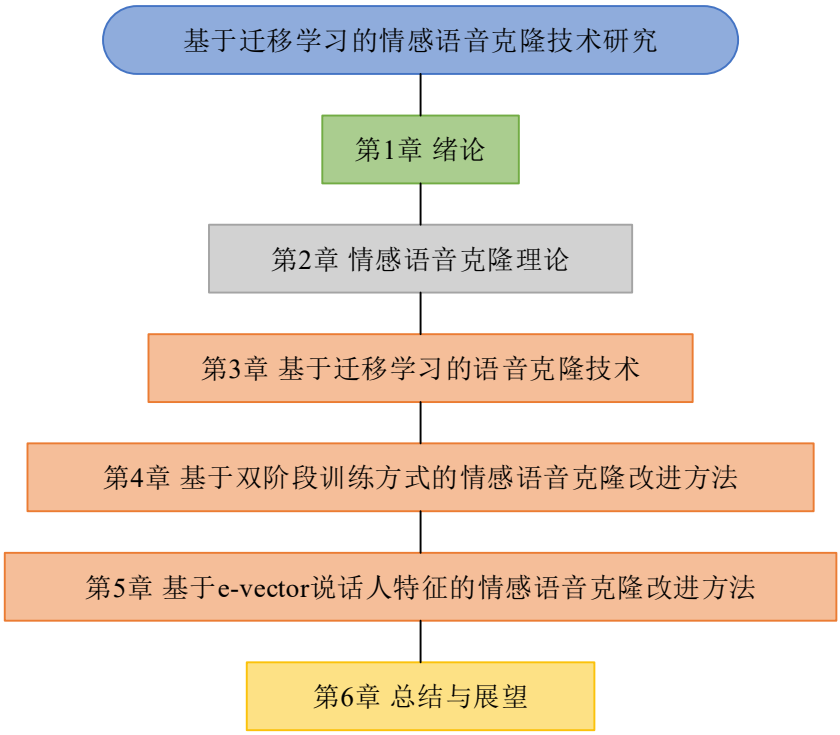


图 1-1 论文组织结构

Fig. 1-1 The structure of the thesis

第一章绪论部分，主要介绍了本研究相关的情感语音克隆技术的研究背景，接着分别对语音合成技术、说话人模型、语音克隆技术以及情感语音合成技术四个方面的国内外研究现状进行了分析介绍，并对本研究中主要的研究内容以及创新点进行了阐述。

第二章情感语音克隆相关理论部分，详细介绍了与情感语音克隆相关的理论和工具。具体来说，阐述了语音信号的产生原理，分析了语音特征参数：基音频率、时长、能量以及梅尔频率倒谱系数，最后介绍了语音编码技术。

第三章基于单阶段迁移学习的语音克隆模型部分，主要完成了语音克隆模型的搭建，并对此模型框架的组成部分进行了描述。此架构基于说话人识别技术与语音合成技术，利用迁移学习的方法，将语音识别框架中的说话人编码器与语音合成技术中的合成器技术、声码器技术相结合，对说话人编码器进行单阶段说话人训练，并对其运行原理、网络模型、数据处理过程以及训练方式进行了介绍。同时，介绍了实验数据的设置以及实验参数的选取，并从主客观角度对克隆语音的音色相似度以及情感相似

度进行了评价。

第四章基于双阶段迁移学习的情感语音克隆方法部分，首先分析了使用单阶段迁移学习的语音克隆模型中所生成的语音缺乏丰富表达能力的问题，接着提出了针对说话人编码器进行双阶段迁移学习的情感语音克隆方法。该方法包括两个阶段的训练过程，即针对说话人编码器进行说话人训练以及情感训练。同时详细介绍了训练数据选取、训练流程框架、具体训练过程以及使用到的损失函数。最后采用主客观评价方式，对该情感语音克隆方法的结果进行了分析。

第五章基于 e-vector 说话人特征的情感语音克隆改进方法部分，介绍了在对说话人编码器采用双阶段训练的情感语音克隆方法的实验结果进行分析所发现的问题，提出了在提取音频声纹特征的同时对情感韵律特征进行分析，然后对其进行多种特征融合的说话人编码方式，并将其命名为 e-vector。此外，介绍了利用 e-vector 对说话人编码器双阶段训练方式的改进。

第六章总结与展望部分，对本论文研究内容进行了总结，分析了目前实验所存在的局限性，并对下一步工作进行展望与设想。

1.5 本章小结

本章主要介绍了情感语音克隆技术相关领域的研究背景以及情感语音克隆的应用领域，对语音合成技术、说话人模型、语音克隆技术以及情感语音合成技术的国内外研究现状进行了分析总结，同时对本论文的研究内容和主要解决的问题进行了介绍，最后对本研究主要的行文框架进行了简要描述。

第 2 章 情感语音克隆相关理论

智能语音处理领域中，情感语音克隆技术是一个相当重要的分支，是生成具有丰富表达能力的个性化语音的关键技术。完整的情感语音克隆技术包括语音信号特征的提取、语音克隆模型的建立、说话人特征的处理、音频情感特征的处理以及语音生成等多个阶段。因此，研究情感语音克隆技术应该从最基本的理论知识出发，例如语音信号的产生及其所包含的相关特征所代表的意义。本章主要介绍了语音信号的产生原理、语音相关特征参数以及语音编码技术。

2.1 语音信号概述

语音信号的产生过程可以被描述为多个阶段的连续性处理。首先，当一个人想要表达某些信息时，这些信息会在大脑中生成，并通过音素序列、基频周期的变化、韵律特征和适宜的响度等语音特征进行编码。完成信息编码后，肌肉神经命令控制声带振动，在改变声道形状的同时发出包含目标信息的语音信号，并通过口唇、颚和舌头的发声运动，以及气流进入鼻腔软腭的时机来控制发音。在说话人完成整个发声过程后，语音信号将传播到听者的耳膜处。耳膜将信号转换为动态频谱分析，并通过神经传感器将所得到的频谱信号转换为对听觉神经的触动信号。最终，这些信号将在人脑更高层次的神经中枢中转化为语言编码，以使听者正确理解说话人所表达的信息[36]。

根据以上过程可以得出，人类用来产生语音的发音器官自下而上包括肺部、气管、喉、咽、鼻腔、口腔和唇[37]，这些器官形成了一个连续的管道。在该管道中，喉部被称为声门，而喉部以上的区域称为声道，其形状会随着所需表达的语音信息的不同而发生变化。在人体呼吸系统中，肺起着关键作用，由一团有弹性的海绵状物质组成，位于胸腔内。在呼吸期间，空气进入肺部存储。当需要发声时，人体腹部肌肉的收缩使得横膈膜向上移动，逼出肺部存储的空气，形成气流，从而成为语音产生的原动力。该气流通过气管到达咽喉，在气流的冲击下，紧闭的声带被迫振动，快速地开合，产生一系列气流，这些气流在被截断后变成准周期的脉冲。接下来，这些脉冲经过咽喉、口腔或鼻腔的动作进行进一步调制，形成不同的语音。最后，由于嘴唇开合的变化，这些语音通过口腔被散发出去，并传达给听者[36]。

2.2 语音特征参数

语音相关特征在情感语音合成、语音识别、说话人识别等任务中得到了广泛应用。

实验证明，此类语音特征在解决各种语音任务时都是有效的。以下将介绍在情感语音合成以及说话人识别任务中常用的语音特征参数。

2.2.1 基音频率

在语音信号的生成过程中，声带会因为气流冲击而不断地收缩和扩张，从而产生振动。其中，声带完成一次完整的开合动作所需的时间即称为基音周期，其倒数代表基音频率。声带的大小、厚度、松紧度等特征以及声门上下间的气压差也会对基音频率的值产生影响。通常情况下，声带越紧、越薄、越长，基音频率也会越高。同时，声门的形状也会随之变得更加纤细[36]。

一般而言，基音频率的大小范围会根据说话人的性别、年龄以及其他具体条件而产生变化，最低可到 80Hz，最高可达到 500Hz。当男性年龄增长时，说话的声音比年轻时更低沉，基音频率下降，而青年女性和孩童则具有较高的基音频率。此外，根据以往的研究可以得出，发音者说话时的情感状态也会对基音频率产生影响[38]。此研究分析了人在不同情感状态下的语音韵律特征参数，根据结果可以得出，尽管发音人的性别可能会对语音的相关情感特征产生影响，但也表现出了共同的规律。例如，在愤怒和喜悦等情感下，人发出的语音中基音频率均值较高；而在悲伤等情感下，基音频率均值则较低。

2.2.2 时长

说话人表达一句话所用的时间长度被称为音频时长，这可以反映出说话人的语速快慢。语速是语音韵律特征之一，对于语言节奏的表达具有重要作用。在不同的环境或情感状态下，说话人表达的急切程度也会随之改变。因此，语速的变化成为了表达情感的重要方式之一。在相关韵律参数的研究中，对不同情感下语音的时长、平均发音速率进行了提取，结果如表 2-1 所列。

表 2-1 不同情感语音的时长及平均发音速率[39]

Table 2-1 Duration and average articulation rate of different emotional speech[39]

情感类型	时长（s）	平均发音速率（音节/s）
中立	4.78	0.209
愤怒	2.59	0.385
厌恶	5.15	0.194
困倦	4.93	0.203
开心	3.93	0.252

由表中数据可以看出，在文本内容和说话人保持不变的前提条件下，不同的情感状态对语音的语速产生了明显的影响。其中，当说话人处于愤怒状态时，其语速最快，

其次是开心。在这两种情感状态下，说话人的激活度较高。而当人处于厌恶状态下时，语速最慢。

2.2.3 能量

音频的能量特征反映了说话人在发音时的强度。根据研究结果显示，音频所表达情感的变化与声音的强度之间有着较强的相关性，由于这种性质与人耳的听觉系统相关，因此并非简单的线性关系。尽管声音的强度对情感表达的影响十分复杂，但通过情感语料库对其进行统计分析，仍可以得到具有重要意义的结论。

在计算时，声音的强度通常采用短时能量来表示，第 n 帧语音信号 $x(n)$ 的短时能量定义为：

$$E_n = \sum_{m=n-(N-1)}^{n} [x(m)w(n-m)]^2 \qquad \text{（公式 2-1）}$$

公式中，$w(n-m)$ 是窗口移动函数，n 是窗口的时间位置，它可以是窗的起点、中点或者终点，N 是窗的有效宽度。本研究通过实验提取了情感语料库中所有音频的短时能量后，根据不同的情感标签，利用平均值来比较分析语音的短时能量均值、能量最大值以及最小值，计算结果如表 2-2 所列。

表 2-2 不同情感语音的能量

Table 2-2 Energy of different emotional speech

情感类型	能量均值/dB	能量最大值/dB	能量最小值/dB
中性	62.89	69.56	51.55
喜悦	66.90	86.32	50.71
愤怒	71.34	89.40	54.26
悲伤	64.12	70.08	42.49
惊讶	67.05	88.54	51.09

由实验结果可以得出，中性情感的能量均值（62.89）最低，其次为悲伤情感（64.12），喜悦情感（66.90）与惊讶情感（67.05）的能量均值基本处于同一水平，能量均值最强的为愤怒情感（71.34），总体符合人们在日常生活中利用语音表达情感的规律。当人们经历愤怒、惊讶或喜悦等情感时，他们通常会使用大声说话的方式来表达自己；而当人们感到悲伤时，他们则倾向于以低沉的语调来发声。

2.2.4 梅尔频率倒谱系数

梅尔频率倒谱系数（Mel Frequency Cepstrum Coefficient，简称 MFCC）结合了语音的生成机制与人耳听觉感知过程中的相关特性，具有表征说话人身份的能力[40]，因此是语音识别领域以及说话人识别技术中常用的语音特征，在大多数语音识别系统中得

到了广泛使用。

迄今为止，在研究人耳听觉特性方面，不仅局限于语音声学领域，心理声学领域也有大量相关研究。心理学相关研究表示，由于人类听力系统的特性，如个体差异、听觉疲劳等，对声音的感知会产生影响。因此有些声音尽管客观存在，但人们听到的声音与实际的声波并不完全一致[41]。这一现象充分说明了人耳具有复杂的功能，它不但能灵敏地捕获到各种声音，还能对接收到的声音进行主观选择，起到分析器的作用。耳蜗是人耳的关键组成部分，它使得人即使处在十分嘈杂的环境中或者具有复杂变化的情况下，仍能分辨出各种语音。由于耳蜗各个部位的内部形状、外毛细胞以及绒毛情况有所差异，使得其对频率选择有所不同，因此耳蜗的整体工作过程相当于一个滤波器组，将不同频率的声音分离出来并进行处理。不同于传统的滤波器，耳蜗的滤波作用是基于对数频率尺度的，这是由耳蜗结构的特殊性质所决定的。根据这一原则，研究者得出了梅尔滤波器组。Hz 频率 f 与 Mel 频率 f_{mel} 之间的转换可以用如下公式进行表示：

$$f_{mel} = 2595 \times \lg(1 + f / 700) \qquad \text{（公式 2-2）}$$

得到了梅尔频谱后，梅尔频率倒谱系数的具体计算过程如下：

（1）将原始语音信号切分为若干帧，然后对每一帧应用预加重和汉明窗处理，接着使用短时傅里叶变换来获取该帧语音信号的频谱信息；

（2）对频谱信号进行平方处理，并利用 A 个梅尔带通滤波器进行滤波操作。此外，还需要进行叠加操作；

（3）求出每个频带的对数功率谱后进行反离散余弦变换，便得到了 L 个梅尔频率倒谱系数，L 通常取值为 12~16，具体计算如下式所示：

$$C_n = \sum_{a=1}^{A} \log x'(a) \cos\left[\pi(a - 0.5)n / A\right], n = 1, 2, \cdots, L \qquad \text{（公式 2-3）}$$

公式中，$x'(a)$ 表示第 a 个滤波器的输出功率谱，A 表示第 A 个梅尔带通滤波器，n 表示第 n 个梅尔频率倒谱系数；

（4）将得到的梅尔频率倒谱系数进行一阶与二阶差分，便得到了对应的动态特征。

2.3 语音编码

语音信号是一种模拟信号，通过脉冲编码调制（Pulse Code Modulation，简称 PCM）等技术可以将其转换为数字信号，实现语音信号的数字传输，这是通信发展的重要推动因素之一。模拟通信是指利用模拟信号和相关技术来传输信息，并在接收端还原原始模拟信号。与之相比，数字通信在语音质量、抗干扰性和传输效率等方面具有更优

越的性能。将语音信号数字化的最简单方法是直接进行模/数转换。但是，为了实现高质量的数字语音，需要采用足够高的采样率。然而，高采样率也意味着需要更多的数据存储空间来存储数字语音数据。为了解决这个问题，可以采用压缩编码的方式对数据进行预处理，以减少传输码率。传输码率或比特率表示每秒钟传输音频数据所需的比特数。

语音编码的目标是在保持音频质量和语义信息的前提下，最大程度地减少传输数据量[36]。语音编码的方式根据其实现过程可以分为三种：波形编码、参数编码以及混合编码。波形编码是直接利用数字编码的形式对原始波形进行表示，使得波形尽量保持不变。因此，该种编码方式能较好地保证音频的质量，同时传输音频也具有良好的抗噪性。但是由于它直接对原始波形进行数字编码，对数码率的要求较高，通常为 16 千比特每秒至 64 千比特每秒。参数编码与波形编码处理方式不同的是，它首先对原始波形进行处理提取特征参数，然后对特征进行编码传输，在接收端利用解码后的参数重构语音。因此，参数编码并不保证波形相同，主要注重于从听觉的角度对语音的重现，即使得接收端重构后的语音与发送端的语音听起来是相同的。因此，相比较于波形编码，参数编码对数码率的要求较低。混合编码即是上述两种方法的结合，同时从波形以及参数两方面构造语音编码，在提升语音自然度的同时，对数码率进行有效的降低[36]。

综上所述，不同的编码方式有不同的优缺点和适用场景，需要根据实际情况选择恰当的编码方式。在当今较流行的语音合成框架技术中大多采用的都是参数编码方式，在本研究中沿用了此方法。

2.4 本章小结

本章主要介绍了本研究中所使用的一些基本理论和工具。首先对语音信号的产生机制进行介绍，并结合本文的研究目的对语音信号的相关特征参数进行分析。最后，对语音合成技术中所使用到的语音编码原理进行了介绍。

第 3 章 基于单阶段迁移学习的语音克隆模型

随着智能语音技术的不断发展并广泛应用于生活中，生成个性化语音已成为研究人员关注的一个方向。因此，本研究构建了基于单阶段迁移学习的语音克隆模型，以生成特定说话人的语音。本章主要介绍了语音克隆模型的构建，并通过实验对所生成的语音的质量进行了评估。

3.1 基于迁移学习的说话人编码器

尽管传统的语音合成方法可以根据给定的文本生成自然的语音，但这些合成语音并不能适用于任意目标说话人，其主要原因在于生成语音的过程中，缺乏目标说话人的有效信息。本研究在此基础上，构建了基于单阶段迁移学习的语音克隆模型，有效解决了生成任意目标说话人语音的问题。

本研究所采用的具体方法为使用迁移学习的方式，利用说话人识别技术中的说话人编码器，提取目标音频中的说话人特征，此特征能对说话人的身份进行表示。随后，将该特征作为语音合成网络的输入，对语音的生成过程进行调节，从而生成具有目标说话人音色的克隆语音。

3.1.1 迁移学习

在认知、学习以及探索新知识的过程中，人类会借助于已有的知识和经验，以便更快速地进行研究[42]。因此，当人们试图学习新知识或面对新的任务时，他们会运用已有的知识体系加深对新知识的理解，并利用已有的经验来解决相应的问题。由此可以得出，迁移学习是人类普遍表现出的心理特征之一，指的是新知识与已有知识的联系越紧密，人们越容易掌握它。心理学研究证实了迁移学习在人类认知过程中的普遍表现和重要性。而在机器学习领域，传统的算法仅为解决特定问题而设计，这忽略了算法及相关知识之间的联系。相比之下，迁移学习方法尝试通过利用已经学习到的相关知识，来提升目标任务的学习质量，从而使机器学习更加高效[43]。目前，迁移学习已被广泛应用于传统机器学习难以处理的问题，例如图像处理、语音识别和自然语言处理[44]。

在迁移学习中，常将已经完成的任务称为源任务，而对于待完成的新任务则称之为目标任务。迁移学习旨在通过利用在源任务中所学到的知识和经验来提高目标任务的学习效果。具体而言，该方法利用在某个源任务上训练好的模型，将其应用于目标任务中，以帮助解决新问题。这种迁移只有当源任务与目标任务之间存在相关性时才

能发挥作用。因此，相较于从头开始训练，利用已有的参数初始化模型可以避免大量时间和资源的浪费，从而提高模型性能。值得注意的是，从头开始训练模型通常还需要非常大规模的数据集，这对于数据的采集、存储以及处理都提出了较高的要求。与此不同的是，利用迁移学习的方法，可以在已有模型的基础之上，使用相对较少的数据集来微调其参数，以达到目标任务的效果。这种方法一方面减轻了数据采集和处理的负担，另一方面也避免了从头训练的过程中可能存在的过拟合等问题。

关于迁移学习的研究最早出现于心理学，后又广泛应用于教育学[45]。随着深度学习的发展，出现了一系列有关于迁移学习的方法，按照目的可以将其分类为：（1）迁移预训练模型：利用预训练模型提取数据的基本特征，在此基础上进行新模型的训练。（2）利用预训练模型提取特征：将预训练模型提取的特征作为新任务的输入，对模型的生成过程进行调节以改变结果的某些特征。（3）微调现有的预训练模型：此方式与前两种方式的区别在于需要对已有的模型进行操作，根据源任务与目标任务的关系分析结果，对已有的模型进行再次训练，使其适应新任务的目的。

说话人识别是一种根据音频中所包含的特征信息识别说话人的身份的技术[46]。在传统的语音合成模型中，生成过程仅针对输入的文本信息，导致合成的语音缺乏目标说话人的身份信息。因此，本研究利用说话人识别框架对说话人编码器进行预训练，以提取目标音频中的说话人特征。然后通过迁移学习中的第二种方法，将此特征作为语音合成模型的输入，对语音的生成过程进行调节以生成具有目标说话人身份特征的语音。

3.1.2 说话人编码器

在传统的语音合成模型中，为了提高生成语音的自然度，通常需要对训练数据集进行高要求：包含大量高质量且带有相应文本的语音数据。而对于支持多个说话人的语音合成模型，则需要利用每个说话人数十分钟的语料进行训练[6]。同时，这种方法存在一个问题，即生成的语音不能推广到未参与训练过程的说话人。为了使该模型适应语音克隆的目标，本研究通过利用迁移学习的方法，将说话人识别技术中的说话人编码器与语音合成模型相结合，有效解决了传统语音合成模型中缺乏说话人身份信息的问题。

具体而言，本研究采用了一种基于复合长短时记忆网络的说话人编码器。该网络由三个长短时记忆网络层（Long Short-Term Memory，简称 LSTM）[47]和一个 256 个单元的全连接层构成。长短时记忆网络是一种改进后的递归神经网络，具有处理不同长度时序信息的能力。该网络通过使用记忆单元和门机制实现选择性记忆和噪声信息的过滤，避免了循环神经网络中的梯度消失问题[48]，使得网络能够更好地处理具有时间序列信息的数据。由于语音信号具有短时平稳性和明显的时序性信息，使用长短时记忆网络可以获取帧特征之间的时序关系，从而相较于传统模型，能够体现更好的性能。

单个长短时记忆网络结构如图 3-1 所示。

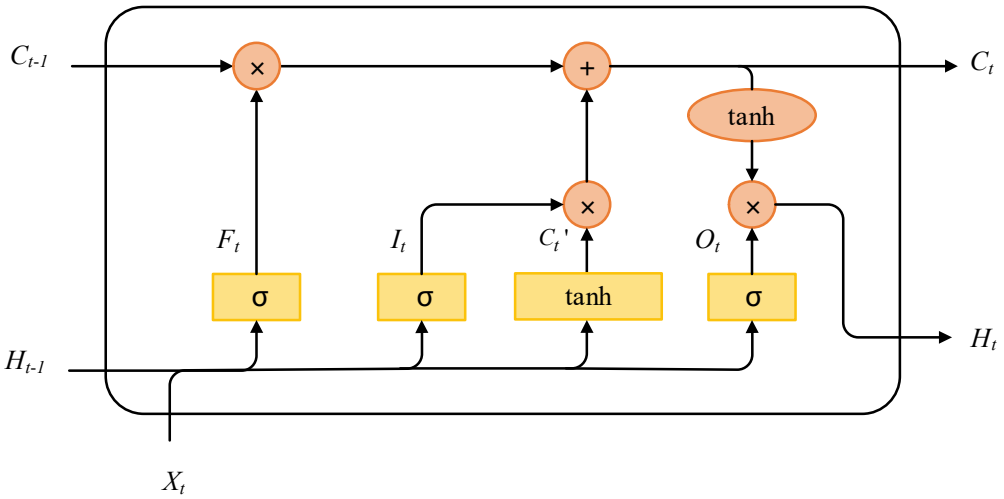


图 3-1 单个长短时记忆网络结构

Fig. 3-1 Structure of the single long short-term memory

长短时记忆网络通常由多个基本时间单元组成，每个单元具有相应的输入门、遗忘门、输出门以及记忆单元。各个门的状态按公式 3-1 至公式 3-5 进行更新。

$$F_t = \text{sigmoid}(X_t W_{xf} + H_{t-1} W_{hf} + b_f) \quad \text{（公式 3-1）}$$

$$O_t = \text{sigmoid}(X_t W_{xo} + H_{t-1} W_{ho} + b_o) \quad \text{（公式 3-2）}$$

$$C_t' = \tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c) \quad \text{（公式 3-3）}$$

$$C_t = F_t \odot C_{t-1} + I_t \odot C_t' \quad \text{（公式 3-4）}$$

$$H_t = O_t \odot \tanh(C_t) \quad \text{（公式 3-5）}$$

公式中，F_t 表示遗忘门，X_t 为当前时刻的输入，H_{t-1} 为前一时刻的隐藏状态，W 和 b 分别为模型的权重和偏置值，O_t 表示输出门，C_t' 为候选记忆单元，C_t 表示为记忆单元，C_{t-1} 为前一时刻的记忆单元，I_t 表示输入门，H_t 表示隐藏状态，W 和 b 分别为模型的权重和偏置值。由网络结构图可知，三个门状态都是由 H_{t-1} 与 X_t 的输入决定的。其中，遗忘门通过控制记忆细胞中的信息来实现对历史状态的控制。它能够决定需要保留的信息，并结合当前信息产生新的记忆单元。输入门用于对输入的信息进行处理。最后，输出门结合记忆单元的状态对输入信息进行处理，从而产生当前时刻的状态。公式中，sigmoid 代表 Sigmoid 激活函数，将结果映射到 0~1 之间，如公式 3-6 所示；tanh 表示 Tanh 激活函数，如公式 3-7 所示。

$$\text{Sigmoid}(x) = \frac{1}{1+e^{-x}} \quad \text{（公式 3-6）}$$

$$\mathrm{Tanh}(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad \text{（公式 3-7）}$$

在说话人识别技术中，使用说话人编码器进行目标说话人的身份验证。该编码器经过训练后能够将目标说话人的音频波形图转化为对应的说话人特征[49]。本研究将该说话人特征作为合成器的输入，对梅尔频谱图的生成过程进行调节，使其具有目标说话人的身份特征。为了实现这一目标，本研究通过在文本无关的说话人验证（Text-independent Speaker Verification，简称 TI-SV）任务下对说话人编码器进行训练。在训练过程中，模型仅通过分析说话人的语音特征，例如声音频率、音调、语速等，来识别说话人的身份，而不需要考虑具体内容。经过训练后，说话人编码器能够准确提取表征说话人身份信息的特征，该特征直接决定了生成的克隆语音与原始语音之间的相似度，训练流程如图 3-2 所示。

图 3-2 说话人编码器训练流程图

Fig. 3-2 Training process of speaker encoder

具体而言，本研究采用说话人验证框架对说话人编码器进行训练。在预处理阶段，首先对任意长度的训练音频进行处理，生成 40 通道的梅尔频谱图。接着，提取能够代表说话人身份的音频帧特征，并利用三层长短时记忆网络对其时序信息进行处理。每个长短时记忆网络层由 768 个单元组成，将最后一个隐藏层的输出作为该段训练音频的特征向量，便得到了具有时序信息的说话人特征。最后，使用全连接层将该特征映射到固定 256 维的向量中。通过这种方法，可以训练出高效的说话人编码器，从而提高克隆语音的相似度。

3.2 基于单阶段迁移学习的语音克隆模型

在具备了能够根据目标音频提取说话人身份特征信息的前提下，本研究构建了一个基于单阶段迁移学习的语音克隆模型。该方法的流程如图 3-3 所示。具体而言，在该语音克隆方法中共包含三个模块：（1）说话人编码器，其基于目标说话人的输入音频提取符合该说话人身份特征的信息；（2）合成器，则以文本内容和说话人特征向量为输入，旨在生成具有目标说话人音色并能够陈述文本内容的语音信号所对应的梅尔频谱图；（3）声码器，用于将梅尔频谱图转换为对应的语音波形，从而得到最终的克隆语音。

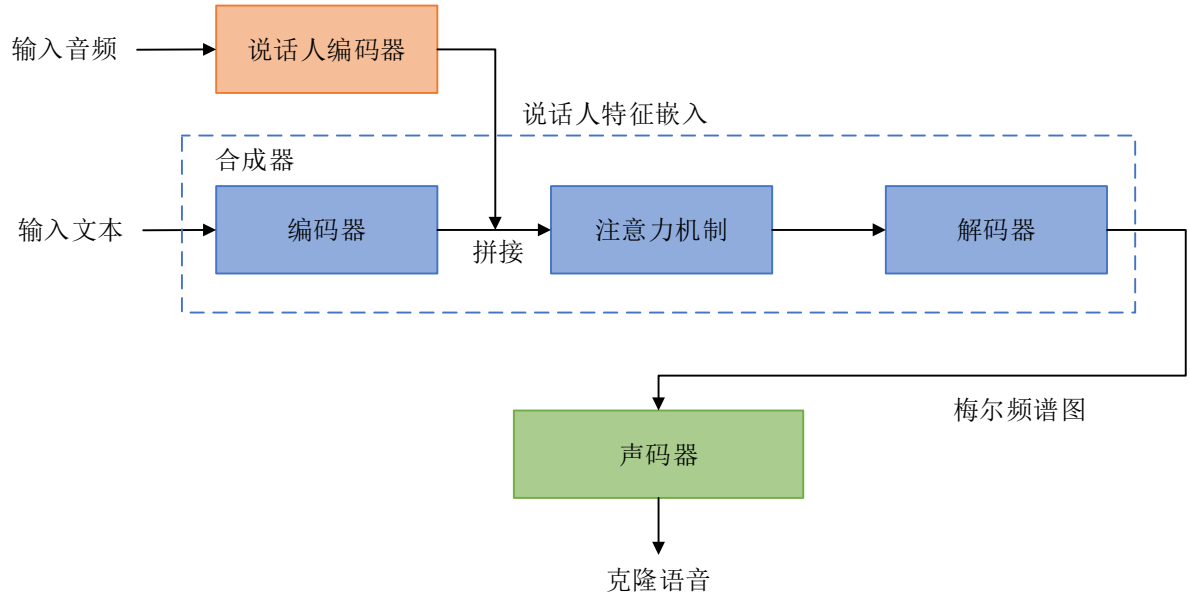


图 3-3 语音克隆方法流程图

Fig. 3-3 Flowchart of voice cloning method

本研究通过实验发现，如果采用端到端的训练方式，使得生成的克隆语音与输入的目标语音之间的损失最小化，那么说话人编码器、合成器和声码器需要在同一数据集上进行训练。这就需要训练数据集满足各个模块不同的需求：必须有大量的说话人、附带相应文本以及低噪声等。同时，如果使用数据量较小的数据集进行训练，可能会导致泛化能力较低的问题。基于上述问题，本研究采用分模块独立训练的方式来训练语音克隆方法。具体而言，整个训练流程包括以下几个步骤：（1）在说话人验证语料库上训练说话人编码器；（2）将训练完成的说话人编码器加入合成器，利用语音合成语料库对其进行训练；（3）在语音合成语料库上训练声码器。经过实验验证，此训练方式取得了较好的效果。

3.2.1 合成器

在传统的语音合成模型中，由于未使用目标说话人的身份信息，因此往往不能生成具有任意说话人音色的语音。本研究在此基础上，将符合说话人身份的特征嵌入作为语音合成模型的输入，对梅尔频谱图的生成过程进行调节，有效地解决了多说话人的问题。

具体而言，本研究采用去除了 WaveNet 的 Tacotron2 架构[50]作为合成器。该合成器架构如图 3-4 所示。Tacotron2 是一种采用循环序列到序列模型的语音合成架构，其主要目的是根据文本生成对应的梅尔频谱图。该架构由编码器和解码器组成，并且包含了注意力机制来减少解码过程中可能出现的子序列缺失或重复问题。编码器由字符嵌入层、三层卷积神经网络和双向长短时记忆网络组成，用于将输入文本转化为具有语义信息的特征向量。解码器则是一个自回归循环神经网络，由预处理网络、两层单向长

短时记忆网络、线性全连接层和五层卷积后处理网络组成。在生成预测结果时，预处理网络对编码器的输出进行处理，并与注意力上下文向量拼接作为单向长短时记忆网络的输入，然后通过线性投影层生成目标语音的频谱帧以及必要的停顿信息。最终，后处理网络对预测结果进行改善，输出生成的梅尔频谱图。

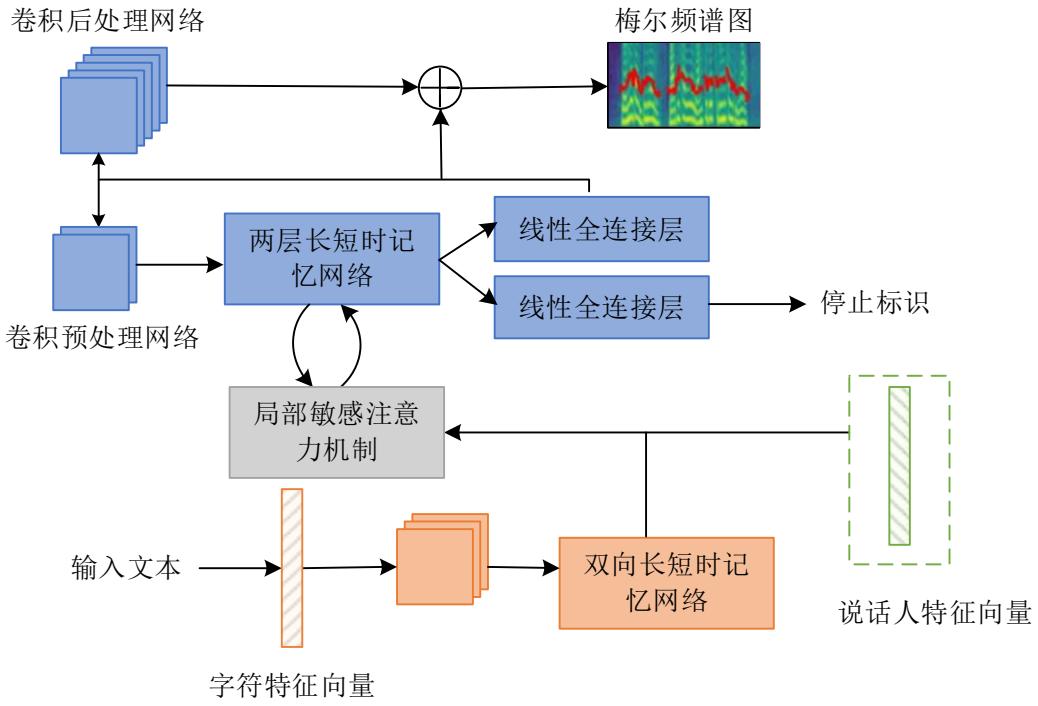


图 3-4 合成器框架图

Fig. 3-4 The framework of the synthesizer

由此过程可得，传统语音合成方法在生成梅尔频谱图的过程中，缺乏目标说话人的身份信息，因此本研究在此基础之上对 Tacotron2 架构做出如下更改：将说话人编码器根据输入音频提取的说话人特征嵌入作为合成器的输入，与合成器中的编码器所生成的字符特征向量进行拼接，对梅尔频谱图的生成过程进行调节。解码器首先根据编码器所输入的序列中的字符特征预测出相应的梅尔频谱图，再利用序列中的说话人特征信息对相关特征进行调节，使得其能表达目标信息，最终便能得到具有说话人特征的梅尔频谱图。

3.2.2 声码器

在本研究中，使用改进的 WaveRNN[51]作为声码器。该声码器的架构如图 3-5 所示。由图可知，网络主要由类残差网络、上采样网络、密集层以及门控循环单元（Gated Recurrent Unit，简称 GRU）组成，可以实现根据梅尔频谱图到音频波形的转换。其中，上采样网络对输入的梅尔频谱图进行处理，使其长度与输入的目标波形一致。类残差

网络提取其频谱段特征生成信息向量，利用此信息向量对音频波形的生成过程进行调节，使得生成波形具有目标信息。

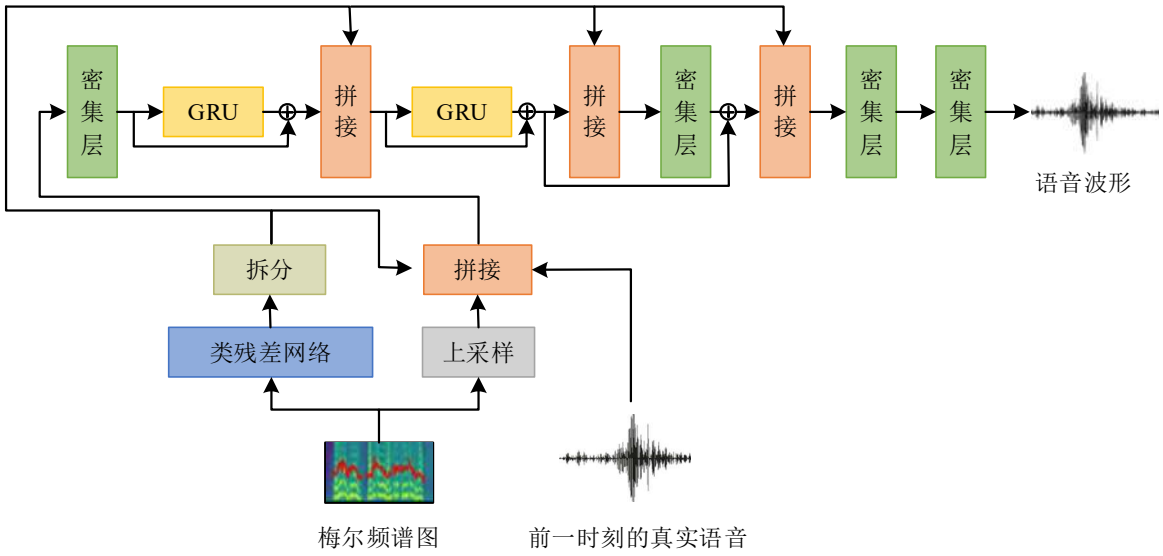


图 3-5 声码器框架图

Fig. 3-5 The framework of the vocoder

门控循环单元是一种循环神经网络，与长短时记忆网络类似，它可以有效解决长时记忆问题。与之不同的是，门控循环单元通过简化复杂的结构，使得网络计算量更小，从而提高了训练和推理效率。因此，在 WaveRNN 中使用门控循环单元对波形的序列信息进行建模，同时提升了序列生成速度，其结构如图 3-6 所示。

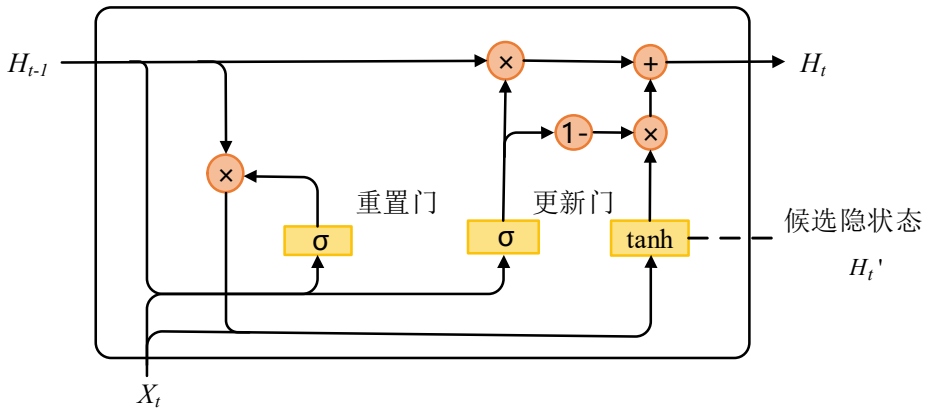


图 3-6 门控循环单元结构图

Fig. 3-6 The Structure of the gated recurrent unit

由图可知，门控循环单元由两个门以及一个隐状态组成，分别为重置门、更新门以及候选隐状态。重置门可根据学习到的经验忽略不重要的数据，而更新门可对重要的信息进行关注并继续向后传递。因此门控循环单元可以有选择地对过去的经验进行

学习，同时避免了循环神经网络中的梯度爆炸等问题。各个门的状态按公式 3-8 至公式 3-11 进行更新。

$$R_t = \text{sigmoid}(X_t W_{xr} + H_{t-1} W_{hr} + b_r) \quad \text{（公式 3-8）}$$

$$Z_t = \text{sigmoid}(X_t W_{xz} + H_{t-1} W_{hz} + b_z) \quad \text{（公式 3-9）}$$

$$H_t' = \tanh(X_t W_{xh} + (R_t \odot H_{t-1}) W_{hh} + b_h) \quad \text{（公式 3-10）}$$

$$H_t = Z_t \odot H_{t-1} + (1 - Z_t) \odot H_t' \quad \text{（公式 3-11）}$$

公式中，R_t 代表重置门，Z_t 代表更新门，H_t' 代表候选隐状态，X_t 为当前时刻的输入，H_{t-1} 为上一时刻的隐含状态，H_t 为当前时刻的隐含状态。其中 sigmoid 代表 Sigmoid 激活函数，tanh 代表 Tanh 激活函数，W 和 b 分别为对应的权重矩阵以及偏置值。由公式 3-8 及公式 3-9 可知，R_t 的范围以及 Z_t 的范围在 0~1 之间，其值可以控制当前信息 X_t 与过去信息 H_{t-1} 对输出状态的影响。例如，当 Z_t 为 0，R_t 为 1 时，此时情况与循环神经网络相同；当 R_t 以及 Z_t 都为 0 时，只考虑当前信息；当 Z_t 为 1 时只考虑过去信息，可以根据在数据中学习到的经验对此参数进行调整。

本研究在对声码器进行训练时，首先对音频数据集进行预处理，提取相应的梅尔频谱图，并将其与对应的音频波形切分成等长的段进行训练。在训练过程中，该模型以频谱段 T 和其前一段波形 W-1 作为输入，生成与 W-1 等长的波形段 W。为保证 W 和 W-1 长度相同，使用上采样网络对输入的频谱段进行调整。同时，利用类残差神经网络提取频谱段特征，并将其重复拼接以与波形段长度保持一致。然后，根据通道数将该特征向量进行拆分。拆分的第一部分与经过上采样的频谱段 T 和波形段 W-1 进行拼接，生成的向量再次通过密集层和 GRU 进行处理，在每层处理之后与特征向量的其他部分进行拼接。最后，通过两个密集层产生波形段 W。

3.3 实验与分析

为了评估基于单阶段迁移学习的语音克隆模型所生成的语音的质量，本研究采用主观分析方法对其与目标说话人的相似度进行验证。

3.3.1 数据集及预处理

本研究实验是在 Ubuntu18.04 操作系统、256GB 内存、4 张 Nvidia Tesla V100S 32G 显卡的硬件配置条件下进行。训练中使用了通用数据集 LibriSpeech 语音识别语料库[52]。LibriSpeech 是一个基于 LibriVox 的公共领域有声读物数据集，旨在辅助自动语音识别系统的训练和测试。该数据集包含 1000 小时的音频数据，采样率为 16kHz，并对每个

录音进行了文本标注和说话人标注。作者将语料库根据音频质量划分为9个部分。其中包含两个口音较纯正的训练集 train-clean-100 以及 train-clean-360，音频时长为 463 个小时，说话人为 1172 个，男女比例 1:1，每个人的语气和风格都有较大差异，这为训练提供了便利条件。因此，本研究采用这两个部分进行实验。

在训练前，本研究首先对音频数据进行了预处理。针对训练集中 train-clean-100 和 train-clean-360 的数据存在较为明显的环境背景噪音以及静音片段的问题，本研究采用谱减法[53]对其进行去噪处理。同时，为了避免对语音进行切分后得到含有过长静音的片段，进而影响训练效果，本研究还对训练音频进行了语音活性检测（Voice Activity Detection，简称 VAD）。语音活性检测方法通常包括特征提取和语音判决两部分，在处理某段音频时，它将产生一个二进制标志。当该标志为 1 时表示非静音，为 0 时表示静音。为了保证语音的自然韵律，需要对该标志数组进行调整。当静音时长不足 0.2 秒[27]时，将其标志修改为 1。根据得到的标志数组，对原始音频进行调整，并对静音部分进行处理。最后，为了调整说话人的音量，对语音进行归一化处理。

本研究在对语音克隆模型各个部分进行训练时所使用到的数据集如表 3-1 所列。train-clean-360 数据集中包含 439 名女性和 482 名男性说话人，这些说话人提供了丰富的语音样本。因此，在说话人编码器网络的训练过程中，本研究使用了该数据集。此外，train-clean-360 数据集中每位说话人录制了 25 分钟的语音，总计达到 363.6 小时，数据量十分庞大。因此，在合成器的训练中也采用了此数据集。声码器网络的训练采用了 train-clean-100 数据集。

表 3-1 训练数据集

Table 3-1 Datasets used for training

网络	子数据集	数据时长/h	说话人个数
说话人编码器 合成器	train-clean-360	363.6	921
声码器	train-clean-100	100.6	251

本研究在训练说话人编码器时，尽管模型能够处理不同长度的音频数据，但是当数据长度一致时，其能够体现出更高的运算效率[27]。因此，为了得到更好的训练效果，本研究使用滑动窗口的方法对音频进行切分处理，如图 3-7 所示。滑动窗口的大小设置为 1.6 秒[8]，同时为了保证获取特征的时序性，片段间保留了 50%的重叠，因此滑动窗口的步长设置为 0.8 秒，最后将网络输出所有片段的特征嵌入的平均值作为说话人特征。随机梯度下降（Stochastic Gradient Descent，简称 SGD）是一种常见的优化算法，用于训练神经网络模型。由于说话人编码器模型较简单，因此本研究采用 SGD 进行训练，初始学习率为 0.01，批处理数量为 32。

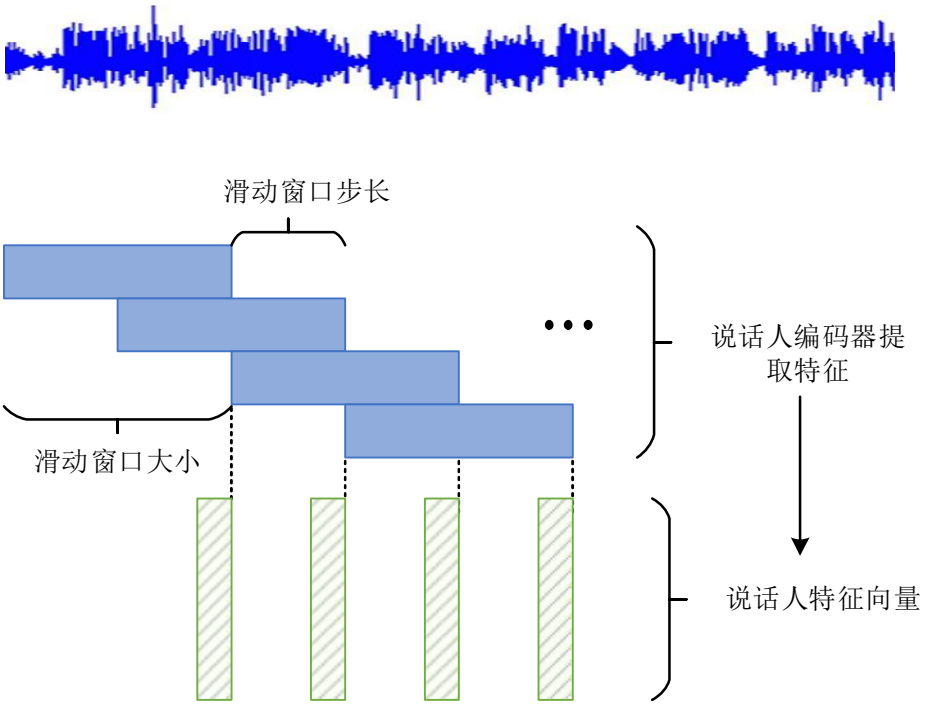


图 3-7 训练音频的滑动窗口

Fig. 3-7 Sliding window for training speech

在训练合成器的过程中，本研究使用说话人编码器从目标说话人的语音数据中提取说话人特征嵌入，并将其作为合成器的输入。但是，在训练过程中不更新训练好的说话人编码器的参数，只对合成器的参数进行更新。为了提高模型的收敛速度和性能，使用自适应矩估计（Adaptive Moment Estimation，简称 ADAM）优化算法，初始学习率设置为 0.001，批处理数量为 32。

在训练声码器时，使用合成器所输出的梅尔频谱图作为输入，并利用目标语音波形对声码器的输出进行判断。使用 ADAM 作为优化器，初始学习率为 0.0001，批处理数量为 8。

3.3.2 性能评估

本研究旨在通过实验评价克隆语音的质量，具体包括生成语音的自然度以及该语音与目标语音的音色相似度。在此背景下，平均意见得分是常用的主观评价指标，被广泛应用于语音合成技术的质量评估中。通常情况下，研究人员采用 95%置信区间下的平均意见得分对生成语音的质量进行评估。置信区间是指由样本统计量所构成的总体参数的估计区间，而 95%置信区间指的是某个总体参数的真实值有 95%的概率会落在测量结果的区间内[54]。在主观评测中，评测者需要根据评分表对听到的语音做出相应的评价。本研究纳入了 100 名评测者，在年龄、性别、职业等方面具有多样性，以确保实验结果的可靠性。具体而言，生成的语音的自然度打分标准如表 3-2 所列，与目标

语音的相似度打分标准如表 3-3 所列。

表 3-2 语音自然度打分标准

Table 3-2 Scoring criteria for speech naturalness

分值	打分标准
1	非常差，听觉无法忍受
2	很差，严重的失真，不自然
3	中等，听不太清，明显失真
4	良好，失真不明显
5	流畅自然，无瑕疵

表 3-3 语音相似度打分标准

Table 3-3 Scoring criteria for speech similarity

分值	打分标准
1	非常差，不是人声
2	很差，与目标说话人相差较大
3	中等，与目标说话人接近
4	良好，与目标说话人相似
5	优秀，与目标说话人一致

3.3.3 实验结果与分析

为了验证基于单阶段迁移学习的语音克隆模型所生成的克隆语音的质量，本研究使用了 20 位目标说话人的语音数据进行测试，其中 10 位说话人在训练数据中随机选取（ID 分别为：30，55，100，119，154，166，176，192，204，274），10 位说话人未出现在训练数据中（ID 分别为：61，121，237，260，367，672，908，1089，1188，1221），每组数据中包含男性和女性各 5 名说话人，每位说话人选取了 10 句测试音频，每句测试音频的时长为 5 秒。评测者依据表 3-2 及表 3-3 的标准对所提供的克隆语音的自然度和相似度进行打分，同时采用真实语音作为对比。具体得分情况如表 3-4 及表 3-5 所列。

表 3-4 语音自然度得分情况

Table 3-4 Scores for naturalness of speech

说话人 ID	性别	是否参与训练	克隆语音	真实语音
30	女	是	3.59	4.89
55	男	是	3.74	4.87
61	男	否	3.39	4.79
100	女	是	3.84	4.94
119	男	是	3.71	4.95
121	女	否	3.56	4.92
154	男	是	3.75	4.89
166	女	是	3.68	4.91
176	男	是	3.80	4.92
192	女	是	3.69	4.98
204	男	是	3.89	4.88
237	女	否	3.49	4.85
260	男	否	3.57	4.86
274	女	是	3.84	4.99
367	女	否	3.38	4.76
672	男	否	3.64	4.85
908	男	否	3.42	4.94
1089	女	否	3.65	4.89
1188	男	否	3.48	4.82
1221	女	否	3.52	4.87

表 3-5 语音相似度得分情况

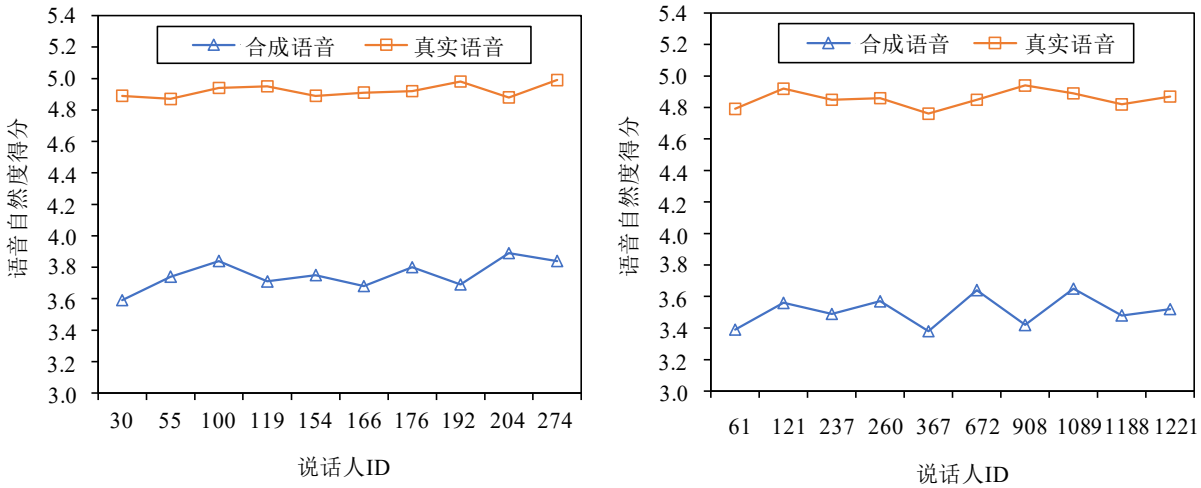
Table 3-5 Scores of speech similarity

说话人 ID	性别	是否参与训练	克隆语音	真实语音
30	女	是	4.18	4.74
55	男	是	4.11	4.6
61	男	否	3.82	4.72
100	女	是	4.01	4.67
119	男	是	4.20	4.68
121	女	否	3.59	4.8
154	男	是	4.04	4.73
166	女	是	3.97	4.65
176	男	是	4.14	4.72

表 3-5 续表

说话人 ID	性别	是否参与训练	克隆语音	真实语音
192	女	是	4.06	4.63
204	男	是	3.98	4.71
237	女	否	3.89	4.73
260	男	否	3.94	4.7
274	女	是	4.09	4.65
367	女	否	3.78	4.79
672	男	否	3.88	4.71
908	男	否	3.62	4.68
1089	女	否	3.70	4.67
1188	男	否	3.92	4.70
1221	女	否	3.68	4.68

同时，为了对语音的自然度以及相似度得分情况做进一步分析以便得出结论，本研究对参与训练与未参与训练的目标音频所得到的结果进行了比较。具体来说，本研究采用 95%置信区间下的平均意见得分对其进行了统计分析，得到的实验结果如表 3-6 所列。同时，为了获得更加直观的实验结果，使用折线图进行对比，结果如图 3-8 至图 3-9 所示。



（a）参与训练　　（b）未参与训练

图 3-8 语音自然度得分对比

Fig. 3-8 Comparison of scores for naturalness of speech

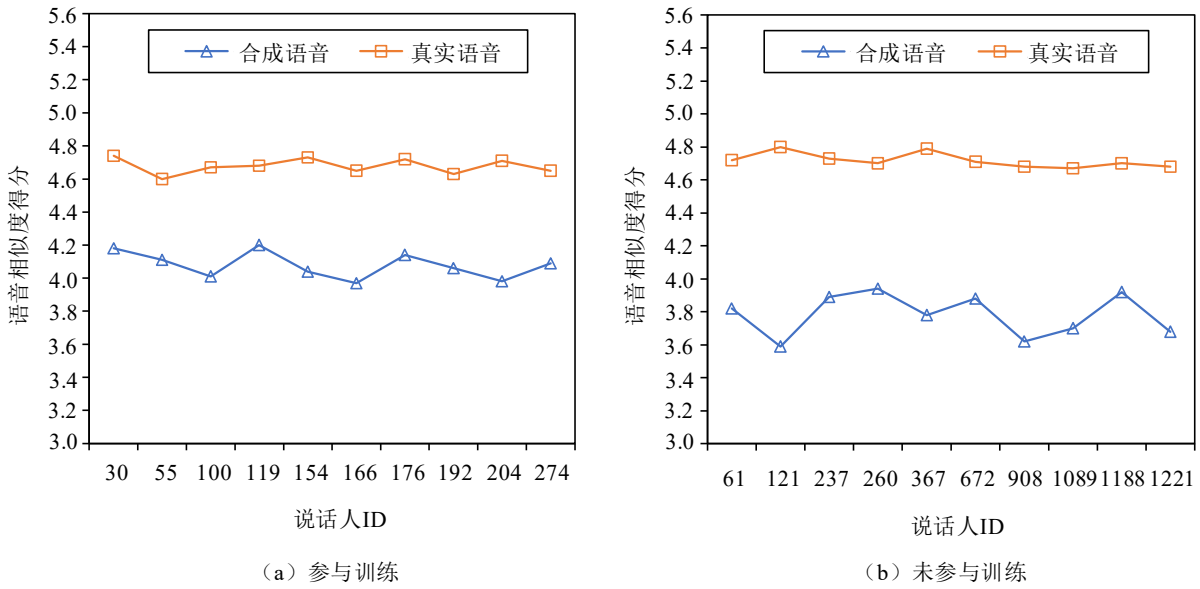


图 3-9 语音相似度得分对比

Fig. 3-9 Comparison of scores for similarity of speech

表 3-6 95%置信区间下的语音自然度和相似性得分

Table 3-6 Naturalness and similarity score evaluations with 95% confidence interval

音频	语音自然度	语音相似度
参与训练	3.753 ± 0.17	4.078 ± 0.15
未参与训练	3.510 ± 0.18	3.782 ± 0.25
真实语音	4.889 ± 0.11	4.698 ± 0.09

从实验结果可以看到，参与训练的说话人所生成的克隆语音的相似度整体高于未参与训练的说话人的结果。由表 3-5 可以得出在音色相似度的实验结果中，其之间的差距最大为0.61，最小为0.03。对实验过程进行分析可以得出，这是由于在训练中说话人编码器的泛化差异所影响的。同时由图 3-9 可以看出，无论说话人是否参与训练，音色相似度的整体结果与真实语音的结果都较为相近，其中参与训练的说话人与真实语音的音色相似度最大差距为0.73，最小差距为0.48，未参与训练的说话人最大差距为 1.21，最小差距为 0.76。

同时由实验结果可以得出，克隆语音的自然度仍有待提升，具体表现为参与训练的说话人与真实语音的自然度相差最大为 1.3，最小为 0.99，而未参与训练的说话人最大为1.52，最小为1.21。通过对模型分析可以得出其主要原因是在说话人编码器的训练中，仅考虑到了音频中与说话人身份相关的特征，未考虑到与情感相关的韵律特征，而情感的表达对语音的质量有很大的影响。因此，为了提高克隆语音的质量，本研究

着重对说话人编码器进行了改进。

3.4 本章小结

本章主要介绍了基于单阶段迁移学习的语音克隆方法的整体模型架构，并对构建好的模型进行了分析实验，主要包括：对迁移学习方法的原理、分类以及应用进行了介绍，分析了说话人识别技术中说话人编码器的结构及原理，然后将该说话人编码器与语音合成技术相结合，构建了基于单阶段迁移学习的语音克隆模型，并对合成器及声码器的模型原理及训练方式进行了介绍。同时，对基于单阶段迁移学习的语音克隆模型进行了实验，介绍了实验相关设置以及数据集预处理过程，分析了实验中所使用到的主要模型参数，阐述了实验结果的评价标准，并基于标准对生成语音的质量进行了评价，最后对实验结果进行了分析。

第 4 章 基于双阶段迁移学习的情感语音克隆方法

在第三章中完成了基于单阶段迁移学习的语音克隆模型的搭建，解决了生成特定目标说话人语音的问题。但是由实验结果可知，生成的克隆语音的自然度较低，缺乏一定的表达能力。为了解决此问题，本章提出了基于双阶段训练方式的情感语音克隆方法。具体来说，首先分析了传统语音克隆方法的不足，然后提出了一种基于双阶段迁移学习的情感语音克隆方法，同时对说话人编码器的音色克隆阶段以及情感克隆阶段进行了详细说明，并通过主客观分析对实验结果进行评测。

4.1 基于双阶段迁移学习的情感语音克隆方法

在前一章中，提到了基于单阶段迁移学习的语音克隆模型生成的语音自然度较低，缺乏表达能力。本研究对该模型进行分析后发现，在传统的说话人识别网络中，其任务仅涉及音频与说话人所属关系的判别。因此，在训练说话人编码器时，仅使用说话人识别语料库，并利用说话人标签进行声纹特征提取的训练。根据这一过程可以得出结论：这种训练方法没有考虑到音频中的韵律特征，而只关注说话人的音色特征。由于缺乏表达情感韵律特征的能力，这种训练方式无法提高克隆语音的自然度，而这种能力对于提升合成语音的自然度非常重要。

情感是一种复杂的人类体验，而语音是主要的情感表达手段之一。在日常交流以及特定场景中（如有声读物、电影配音和新闻播报等），恰当地表达情感对于听众的感知至关重要，因为不合适的情感表达可能导致内容表述不清，甚至出现歧义。为了解决这些问题，近年来兴起了情感语音合成研究领域，该方法相较于传统语音合成方式，考虑到了说话人情感的表达，从而可以生成具有丰富表达能力的语音。在有声读物、智能语音助手、人机交互等应用场景下，合成语音能调整情感以适应特定场景，这将大大提升用户的使用体验。据此可得出结论，传统语音克隆模型中的说话人编码器无法捕捉音频中的情感特征，导致所合成的克隆语音缺乏自然度和丰富的表达能力。为了解决这一问题，在情感语音克隆方法中需要对传统方法中的说话人编码器进行改进，以便从目标说话人语音中提取出表征特定说话人的情感特征。此特征需要同时考虑声纹特征和情感特征，才能在相似度和自然度方面发挥最佳效果。

基于单阶段迁移学习的语音克隆模型中，采用说话人识别框架对说话人编码器进行说话人训练，因此无法表达情感特征。本研究在此基础上提出了一种基于双阶段迁移学习的情感语音克隆方法，能有效解决克隆语音表达能力不足的问题。具体而言，该方法利用音色克隆阶段和情感克隆阶段对说话人编码器进行训练。在音色克隆阶段，

使用说话人识别语料库训练编码器，学习提取说话人身份特征；在情感克隆阶段，使用情感语料库对其进行训练，学习提取情感特征。相比于传统的单一训练方式，本方法显著提升了克隆语音的自然度。图 4-1 展示了该方法的框架。

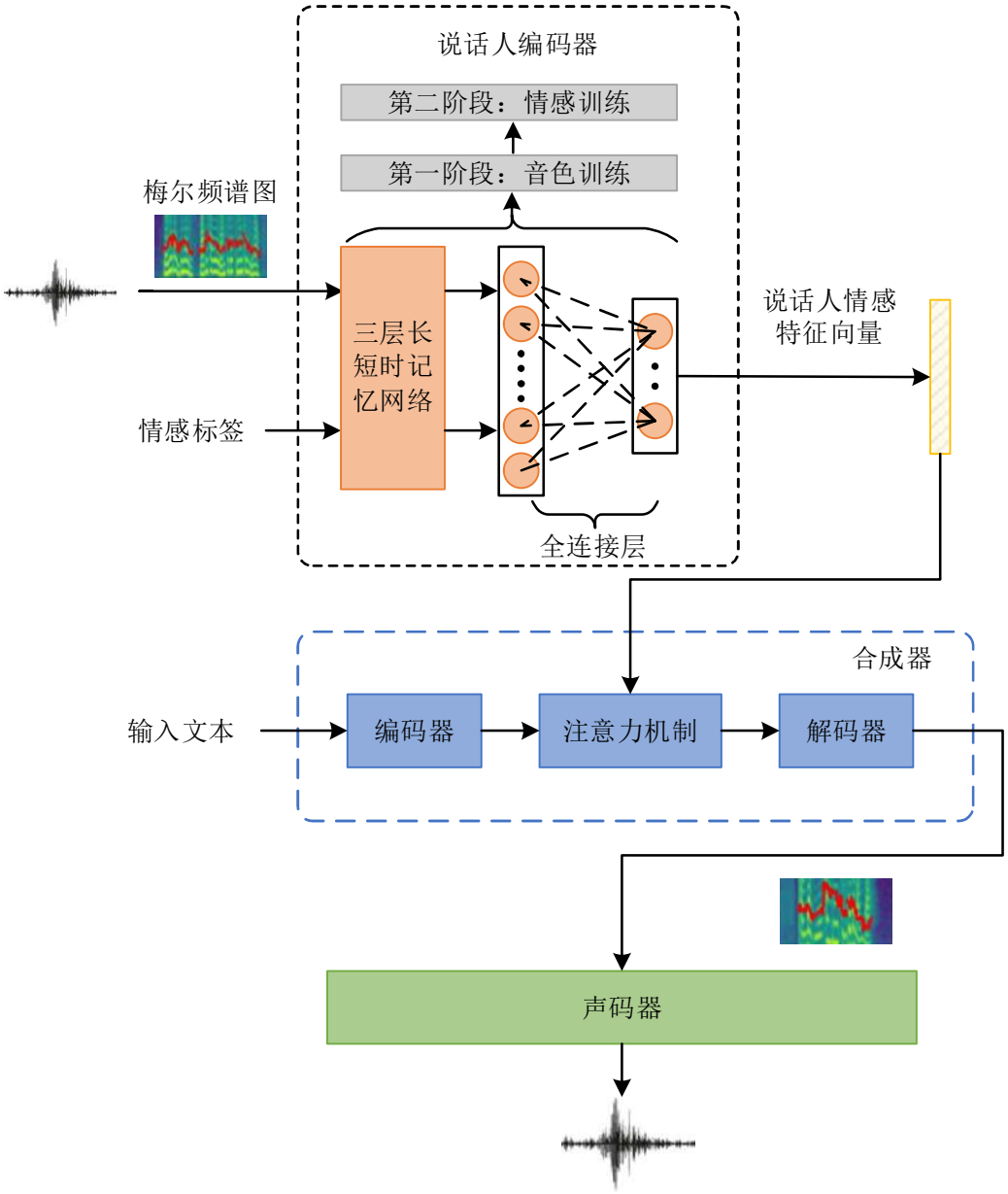


图 4-1 情感语音克隆方法框架图

Fig. 4-1 Framework of emotional voice cloning method

4.2 损失函数的定义

在训练中，每个批次包含 $N \times M$ 句音频。这些音频来自 N 个不同的标签，每个标签分别对应 M 句音频。在音色克隆阶段，选择不同的说话人作为标签；在情感克隆阶段中，则将不同的情感作为标签。特征向量 $X_{ij}(1 \le i \le N, 1 \le j \le M)$ 代表根据第 i 个标签

的第 j 句话所提取的特征。为了获取具有时序信息的语音特征，使用长短时记忆网络对每句话所提取的特征向量 X_{ij} 进行处理。在最后一层长短时记忆网络中，将网络最后一帧的输出送入线性全连接层进行线性处理。使用 $f(X_{ij};w)$ 代表整个网络的输出，w 代表整个神经网络的参数，其中包括长短时记忆网络的层数以及线性全连接层的层数等。将网络的输出进行 L2 正规化后作为说话人的嵌入向量，如公式 4-1 所示。

$$e_{ij}=\frac{f(X_{ij};w)}{\left\|f(X_{ij};w)\right\|_2} \tag{公式 4-1}$$

公式中，e_{ij} 代表第 i 个标签中第 j 句话的嵌入向量，将每个标签具有的所有嵌入向量的中心定义为 c_k，则 c_k 的计算方式为：

$$c_k=\frac{1}{M}\sum_{m=1}^{M}e_{km} \tag{公式 4-2}$$

接着，利用余弦相似度计算每个嵌入向量 e_{ij} 与所有 $c_k(1\le i\le N,1\le j\le M,1\le k\le N)$ 之间的相似矩阵 $S_{ij,k}$，计算方式为：

$$S_{ij,k}=w\cdot\cos(e_{ij},c_k)+b \tag{公式 4-3}$$

公式中，w 和 b 为待学习的参数，同时为了使得当向量间的余弦相似度越大的同时相似矩阵的值也越大，规定 w 的值必须大于 0。

在训练过程中，通过相似度来衡量输入数据与中心值之间的距离。因此，要使得同一个标签中所有句子的特征嵌入都尽可能接近它的中心值，同时与其他标签的中心距离较远。所以，本研究使用 Softmax 函数对相似矩阵 $S_{ij,k}$ 的结果进行判别，当 $k=i$ 时为同一标签，得到的相似矩阵即为 $S_{ij,i}$，此时相似度结果应接近于 1，反之则应接近于 0。e_{ij} 的损失函数定义如下：

$$L(e_{ij})=-S_{ij,i}+\log\sum_{k=1}^{N}\exp(S_{ij,k}) \tag{公式 4-4}$$

即要使得每个 e_{ij} 距离其所属标签的向量中心 $c_k(k=i)$ 更近，距离其它 $c_k(k\ne i)$ 较远。

为了使得训练效果更加稳定，在计算 e_{ij} 与 c_k 的相似性时，若 $k=i$，则 e_{ij} 不参与 c_k 的计算，使用 $c_k(-j)$ 代替，计算方法如公式 4-5。相应地，$S_{ij,k}$ 的计算更新如公式 4-6 所示。

$$c_k(-j)=\frac{1}{M-1}\sum_{m=1,m\ne j}^{M}e_{km} \tag{公式 4-5}$$

$$S_{ij,k}=\begin{cases}w\cdot\cos\left(e_{ij},c_k(-j)\right)+b & if\ \text{k=i};\\ w\cdot\cos\left(e_{ij},c_k\right)+b & otherwise.\end{cases} \tag{公式 4-6}$$

最终网络训练的损失函数为所有 e_{ij} 的损失值之和：

$$L(S)=\sum_{i,j}L(e_{ij}) \quad \text{（公式 4-7）}$$

公式中，$L(e_{ij})$ 表示嵌入向量 e_{ij} 的损失值，$L(S)$表示相似矩阵 $S_{ij,k}$ 的损失值。

4.3 说话人编码器的双阶段训练

传统语音克隆方法所生成的语音中包含文本信息以及目标说话人身份信息，但是缺乏一定的情感信息，导致结果自然度较低。本研究在上一章所述的语音克隆框架基础上，提出了改进说话人编码器的训练方式。具体算法如下：为了使得说话人编码器能够根据目标音频和情感标签提取符合特定情感的说话人特征嵌入，利用上一节所述的损失函数，对说话人编码器进行说话人训练以及情感训练。

4.3.1 音色克隆阶段：说话人训练

在对说话人编码器音色克隆阶段的训练中，为了使得该编码器可以根据目标音频对说话人身份信息进行描述，本研究针对语音的梅尔频率倒谱系数进行处理。梅尔频率倒谱系数结合了语音生成机制与人耳听觉感知过程的特点，因此能有效地表征说话人身份。具体而言，本研究在训练过程中，首先使用工具提取输入音频的梅尔频率倒谱系数作为静态特征。由于音频是一段连续时间内的平稳信号，采用静态特征无法很好地处理其时间内的联系，因此对静态特征进行一阶与二阶差分处理，将原始的特征与处理后的特征进行拼接得到动态特征。

在相关研究中，通常将梅尔频率倒谱系数的维度设置在 12 到 16 之间。表 4-1 展示了采用 13 维梅尔频率倒谱系数及其不同阶数的动态特征对说话人识别系统性能影响的研究结果[36]。该研究以 13 维的线性预测倒谱系数（Linear Predictive Cepstral Coefficient，简称 LPCC）作为基线系统，结果表明，梅尔频率倒谱系数可以有效地提升说话人识别系统的性能。同时，根据实验结果显示，增加梅尔频率倒谱系数的维数对系统性能没有产生积极的影响。当采用一阶与二阶动态特征后，误识率降低了 20%。类似地，进一步提升动态阶数也未能改善系统性能。因此，在本研究中，采用目标音频的 13 维的梅尔频率倒谱系数及其一阶与二阶动态特征来描述说话人声纹特征。

表 4-1 动态特征对系统识别性能的影响[36]

Table 4-1 Impact of dynamic features on system recognition performance[36]

特征集合	相对误识率的降低	特征集合	相对误识率的降低
13 维的 LPCC 特征	基线系统	1 阶和 2 阶动态特征	+20%
13 维的 MFCC 特征	+10%	3 阶动态特征	+0%
16 维的 MFCC 特征	+0%	-	-

为了使得生成的语音具有目标说话人的音色，本研究将可以代表说话人身份的特征向量加入语音合成过程中进行处理。根据已有的研究可知，说话人识别技术的任务是对输入音频所属的说话人身份进行确认。因此，在音色克隆阶段训练中，为了使得说话人编码器可以根据任意输入的目标音频提取动态梅尔频率倒谱系数特征，本研究通过迁移学习的方法，使用说话人识别任务的训练框架对此编码器进行训练。具体而言，本研究选用已有的说话人识别框架[55]，使用公开的语音识别语料库[52]，以不同的说话人标识作为训练标签对说话人编码器进行音色克隆阶段的训练。

语音信号的分析方法有多种，其中包括时域特征分析、频域特征分析、倒谱域特征分析等方法[56]。由于语音信号是一种时域信号，因此最早采用的是时域分析方法。时域分析具有直观简单、易懂明了等优点，但是无法对信号的感知特性进行深入的分析。相比之下，频域分析可以反映信号在不同频率上的分布情况。研究表明，人耳对不同频率的语音信号的敏感度并不相同，对低频信号更为敏感。因此，在频域上两个距离相等的频率对于人耳来说其距离并不一定相等。为了解决这个问题，研究者引入了梅尔刻度，使用这种刻度可以使得任意距离相等的两个频率对于人耳来说距离也相等。通过将频谱图进行梅尔刻度变换，可以得到梅尔频谱图，从而更好地分析语音信号在频域上的特征。

因此，在对说话人编码器进行音色克隆阶段的训练时，本研究首先对训练音频进行预处理，生成通道数为 40 的梅尔频谱图。在此基础上，提取了能够代表说话人身份的梅尔频率倒谱系数动态特征向量。为了创造具有时序信息的特征，本研究采用长短时记忆网络对帧级别的特征进行处理。具体来说，采用了三层长短时记忆网络和一个全连接层作为说话人编码器，提取多段目标说话人音频的特征嵌入，并计算其平均值作为最终的说话人特征嵌入。同时，将每个音频的特征嵌入与求得的所有说话人特征嵌入的相似矩阵作为损失函数，以此来体现当所属说话人一致时应呈现出较高的相似度，反之则相似度应较低。

4.3.2 情感克隆阶段：情感训练

经过音色克隆阶段的训练后，说话人编码器可以接受任意长度的目标音频作为输入，通过特征提取和规范化处理将其映射到维度固定为256的特征嵌入，用于表征说话人的身份信息。然而，在进行音色克隆阶段的训练时，该编码器以说话人作为标签进行训练，因此该特征向量虽然能够表征目标说话人的身份，但却缺乏丰富的表达能力，这种表达能力是必要的。

为了使得说话人编码器可以提取目标音频中带有情感因素的说话人特征向量，本研究进行了情感克隆阶段的训练。具体而言，以音色克隆阶段学习到的参数初始化模型，然后使用情感语料库，以情感作为标签对说话人编码器进行情感克隆阶段的训练，对该模型的参数进行微调。经过此双阶段训练后，说话人编码器可以根据目标音频以

及情感标签提取出带有特定情感的目标说话人特征嵌入。训练流程如图 4-2 所示。

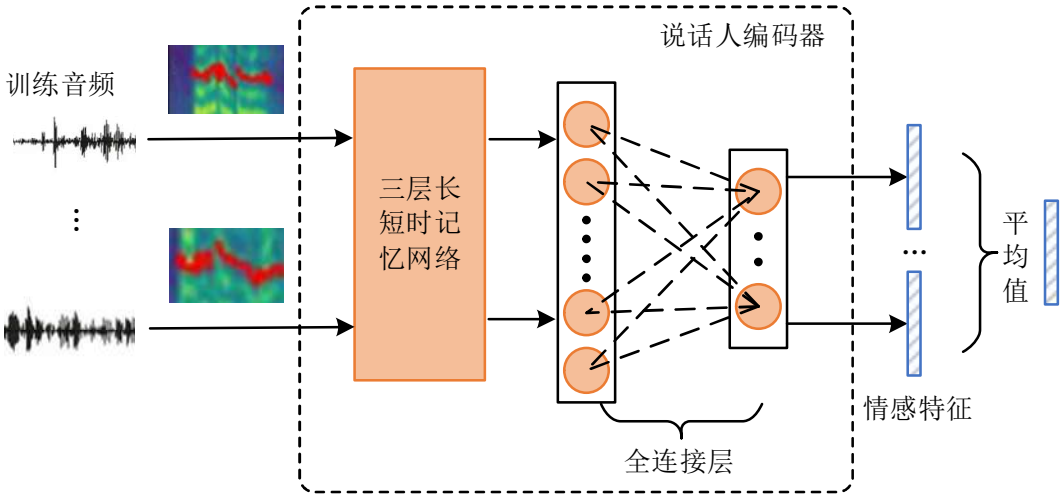


图 4-2 训练流程图

Fig. 4-2 Flow of training

在情感克隆阶段的训练中，本研究选择了四种常见情感标签：喜悦、愤怒、悲伤和中性，并对每个情感标签选择了 M 个音频进行训练。据相关研究表明，音频的基音频率、时长和能量等特征与语音中情感变化密切相关[38]。在多个情感音频中，这些特征的变化呈现出一致性的规律。因此，在情感阶段的训练中，本研究对音频的上述情感特征进行处理分析。具体的操作过程如下：

第一步：与音色克隆阶段一致，首先对训练音频进行预处理，得到梅尔频谱图；

第二步：利用相关工具提取音频中的基音频率、时长和能量等特征，将此特征使用向量拼接的方式进行特征融合；

第三步：由于在上一步中提取的特征向量为帧级特征，因此本研究采用 LSTM 网络对多个帧特征进行处理，以分析前后帧之间的相关信息，并建立具有长时信息的情感特征嵌入；

第四步：为了得到每个音频的情感特征嵌入，本研究对每个音频中各个分段的特征嵌入进行 L2 正则化处理，并求其平均值作为该音频对应的情感特征嵌入。之后，对于每个情感标签，本研究将输入的 M 句音频的特征嵌入的平均值作为该情感的最终特征嵌入表示；

第五步：采用余弦相似性对所有输入音频的特征嵌入与每个情感标签的特征嵌入之间的相似度进行判断。当情感标签一致时，应呈现出较高的相似度。

4.4 实验与分析

音色克隆阶段训练参数的设置与 3.3.1 节中所述内容相同。在情感克隆阶段的训练

中采用了情感语音数据集（Emotional Speech Dataset，简称 ESD）[57]。ESD 中包含 3 万条语音数据，总时长为 29 小时，采样率为 16kHz，信噪比高于 20dB，说话人由 10 位以英语为母语和 10 位以中文为母语的发声者组成，男女比例相等。每个文本均被发声者分别用 5 种情感类别（中性、喜悦、愤怒、悲伤和惊讶）进行表达，并且每条录音都配有相应的文本标注。在训练中，本研究选取了中性、喜悦、悲伤以及愤怒四种情感，使用 SGD 优化器，初始学习速率为 0.001，批处理数量设置为 32。

由于在合成语音的波形频谱图中难以直观地观察到音色与情感的克隆效果，所以本研究从客观、主观两个维度对合成语音的效果进行评价，评价分为音色相似性与情感相似性两个方面。前者旨在衡量合成语音与目标语音的音色相似度，而后者则强调评估合成语音与目标情感之间的匹配程度。本章主要介绍了评价中使用到的主观评价指标与客观评价指标，并对实验结果进行了描述，最后在此基础上做出总结。

4.4.1 客观评价指标

在客观评测中需对目标语音与合成语音的基频、时长、说话人特征向量的相关指标进行计算，并根据结果对相似度进行评价。具体来说，在音色相似度的测评中，对目标语音与合成语音提取的说话人特征向量间的余弦相似度进行评测；在情感相似度的测评中，采用均方根误差（Root Mean Square Error，简称 RMSE）以及梅尔倒谱失真（Mel-Cepstral Distortion，简称 MCD）来对其进行评测。

（1）余弦相似度

余弦相似度是一种常用的特征向量相似度的评价方法，其计算方式为通过计算两个向量间夹角的余弦值来表征它们之间的相似程度。在本研究中，通过计算目标语音的说话人特征向量 R 与合成语音中说话人特征向量 R' 的余弦相似度对其音色的相似度进行评价，当夹角越小时，余弦相似度的值越接近于 1，表明向量间的相似度越高。具体计算公式如下：

$$\cos(\theta)=\frac{\sum_{i=1}^{n}R_i\times R'_i}{\sqrt{\sum_{i=1}^{n}\left(R_i^2\right)}\times\sqrt{\sum_{i=1}^{n}\left(R'^2_i\right)}} \quad \text{（公式 4-8）}$$

（2）均方根误差

均方根误差是一种描述数据差异程度的指标，计算方法为预测值与真实值偏差平方和除以观测次数后开平方。在语音合成技术中，通常通过计算真实语音与合成语音在基频和时长上的均方根误差来衡量其之间的偏差，计算公式如下：

$$\mathrm{RMSE}=\sqrt{\frac{1}{V}\sum_{i=1}^{V}E_i^2} \quad \text{（公式 4-9）}$$

公式中，评价语音对比数为 V，误差参数 E_i 反映的是合成语音 c 和目标语音 d 之间

的差异。其中，基频误差定义为 $F0_i^c - F0_i^d$；时长误差定义为 $T_i^c - T_i^d$。为了确保评价数据的准确，利用 Praat 语音分析软件获取语音的有效时长。

（3）梅尔倒谱失真

梅尔倒谱失真是一种用来评测两个梅尔倒谱序列间差异的方式，计算方式为：

$$\mathrm{MCD[dB]} = \frac{10}{\ln 10}\sum_{i=1}^{K}\sqrt{2\sum_{i=1}^{U}\left(m_{k,i}^{d} - m_{k,i}^{c}\right)^2} \qquad \text{（公式 4-10）}$$

公式中，$m_{k,i}^{d}$ 和 $m_{k,i}^{c}$ 分别表示目标语音 d 和合成语音 c 中第 k 帧中的第 i 个梅尔倒谱系数（Mel-Cepstral Coefficients，简称 MCEP），U 为某一帧中所选取的特征个数。根据前人的研究成果，梅尔倒谱系数向量在其维度达到 24 时，可更为准确地表征语音的情感特征[58]，在本研究中遵循了该结论。由式可知，梅尔倒谱失真越小，原始语音与合成语音之间的差异越小，效果越好。

4.4.2 主观评价指标

主观评价需要评测者根据听到的合成语音与目标语音按照给定的评价标准进行评分。在音色相似性的评价中，评测方式与 3.3.2 中所述一致；在情感相似度的评价中，采用情感平均意见评分（Emotional Mean Opinion Score，简称 EMOS）以及偏好测试来对其进行评测。评测者与 3.3.2 中所述一致。

（1）情感平均意见评分

EMOS 主要用来评测合成语音中的情感表达质量，评测者在听到目标情感语音与合成语音之后，对两者情感的相似度进行打分，评分标准如表 4-2 所列。

表 4-2 EMOS 打分标准

Table 4-2 Scoring criteria for emotional mean opinion score

分值	打分标准
1	非常差，不是人声
2	很差，与目标情感相差较大
3	中等，与目标情感接近
4	良好，与目标情感相似
5	优秀，与目标情感一致

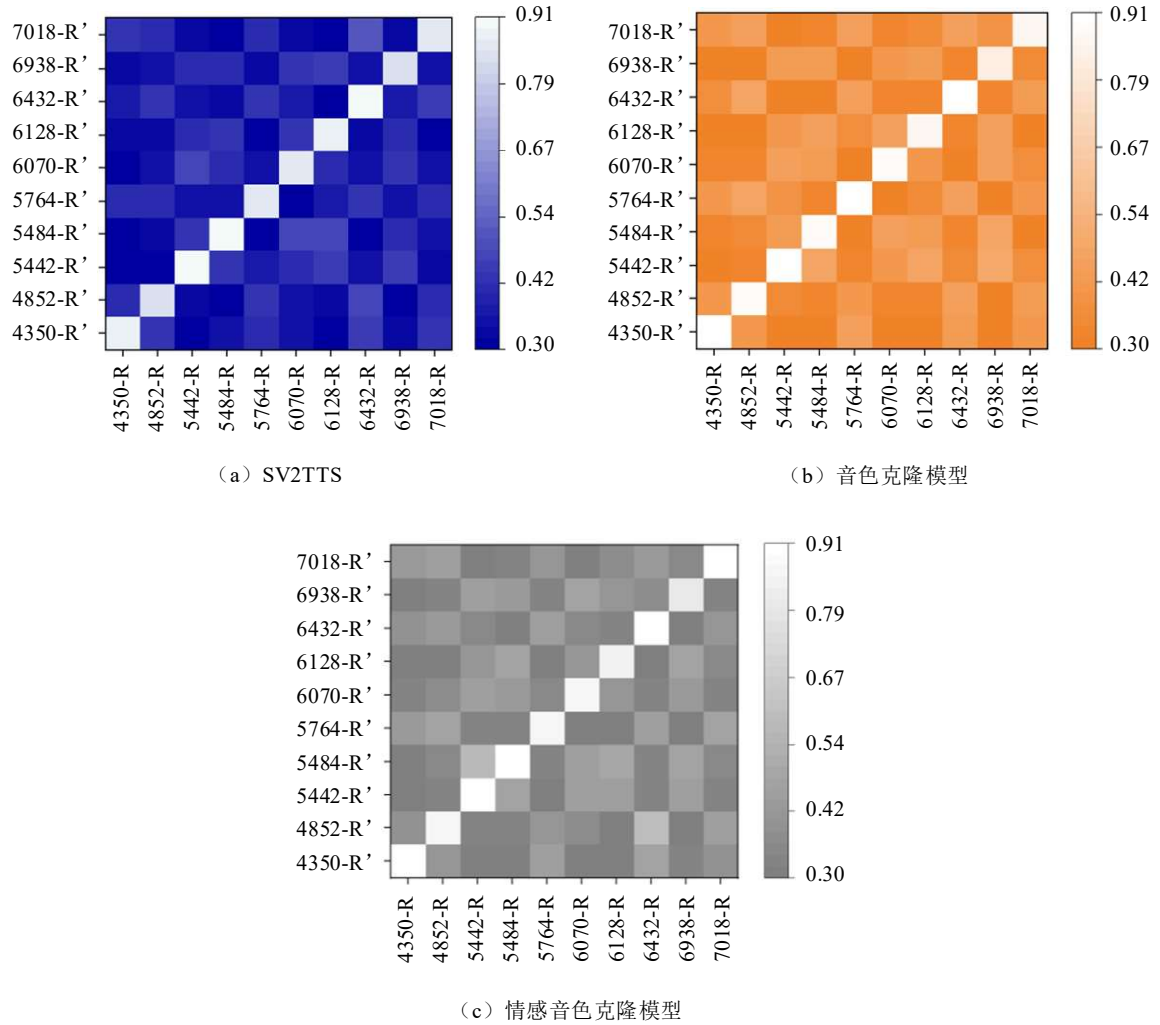
（2）偏好测试

为了验证所提出的方法与传统语音克隆模型[8, 10]以及音色克隆模型之间的差异，同时与情感语音合成模型[59]对比，本研究进行了偏好测试以获得更加客观的实验结果。具体来说，在给出对比模型的合成语音和情感标签后，评测者需听取所有评测语音并选择最符合目标情感的语音，或选择无偏好。

4.4.3 实验结果与分析

本章综合分析了使用双阶段迁移学习训练说话人编码器的情感语音克隆方法所生成的情感克隆语音的质量，具体分为音色相似性与情感相似性两个方面，同时结合音色克隆模型以及传统语音克隆方法 SV2TTS[8]的实验结果进行对比分析。

为了验证情感语音克隆方法的音色克隆效果，本研究在 LibriSpeech 中的 test-other 测试集中随机选取 10 位说话人，第一次随机选取 5 位男性（ID 分别为：4350、4852、5764、6432、7018），第二次随机选取 5 位女性（ID 分别为：5442、5484、6070、6128、6938）。采用经过双阶段训练的说话人编码器，提取目标输入音频中的说话人情感特征向量。接着，利用公式 4-8 计算这些特征向量之间的余弦相似性，以分析其音色相似性。实验结果见图 4-3。



（a）SV2TTS　　（b）音色克隆模型

（c）情感音色克隆模型

图 4-3 特征向量间相似度的实验结果

Fig. 4-3 Experimental results of similarity between feature vectors

图 4-3 中，颜色越深表示说话人特征向量间的相似度越低，反之则表示相似度越高。

在 ID 标记定义中，*R* 表示根据目标真实语音提取的说话人特征向量，*R'*表示根据合成语音提取的说话人特征向量。本研究对参与测评的 10 位说话人的真实语音的说话人特征向量与合成语音的说话人特征向量进行两两分析，即进行 10×10 次对比分析。结果表明，所有子图当中均可以看出，同一说话人的特征向量间相似度普遍较高，不同说话人间相似度较低。同时，对三个模型同一说话人特征向量间的相似度进行比较，结果如图 4-4 所示。

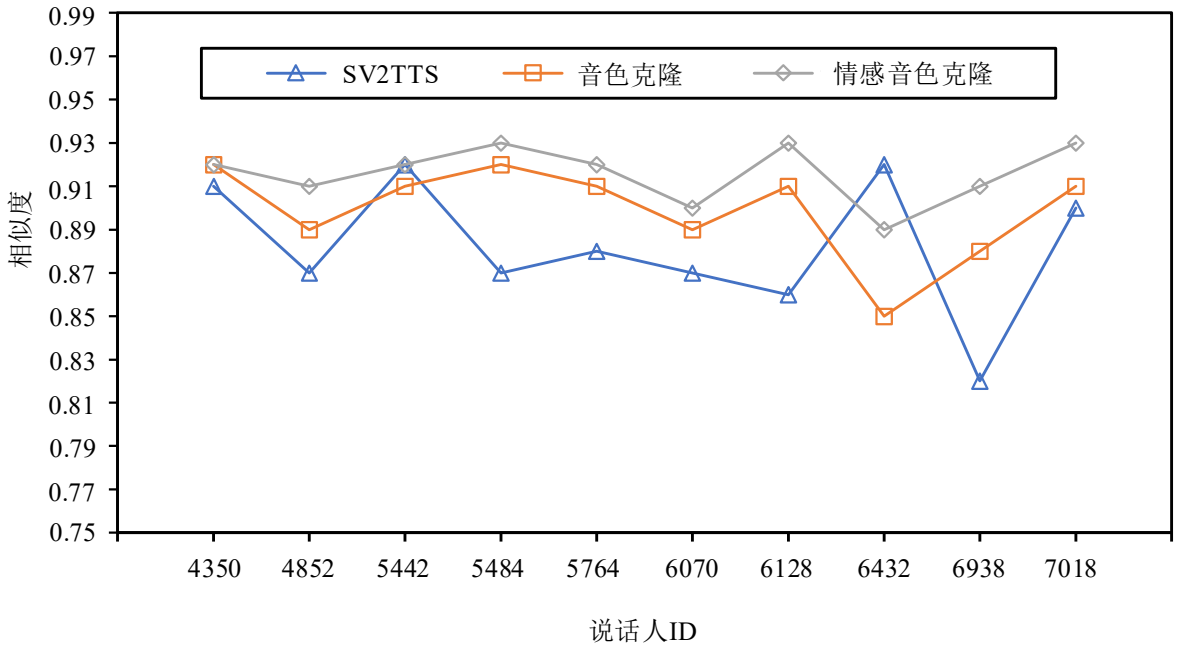


图 4-4 同一说话人特征向量间相似度比较

Fig. 4-4 Comparison of similarity between feature vectors of the same speaker

由图 4-4 可知，情感音色克隆模型所得的说话人相似度整体高于传统模型 SV2TTS，其中有 8 位说话人表现出了更好的语音相似性，并且体现出了稳定的性能。具体而言，在传统 SV2TTS 语音克隆模型中，同一说话人特征向量间的相似度方差为 8.76E-4。然而，在音色克隆模型中，这个方差值降低到了 4.29E-4。在此基础上，情感音色克隆模型进一步进行了改进，将相似度方差降低到了 1.64E-4，表明该模型在保持说话人身份特征的同时，提供了更加稳定的语音合成效果。

同时，为了验证合成语音的音色相似性，本研究进行了主观评价实验。具体而言，在 LibriSpeech 和 ESD 评价数据集中随机选择参与训练和未参与训练的各 10 位说话人进行实验。将目标说话人的真实音频与合成音频一一配对，并采用说话人相似性平均意见得分来评估合成语音的音色相似性。评测者需按照表 3-3 中所列的标准，对它们的相似度进行打分。根据评分标准可得，评分越高表明两者之间的音色相似度越高。同时，本研究还将传统模型的实验结果与所提出的方法进行了对比分析。得分情况如表 4-3 及表 4-4 所列，并将统计结果汇总于表 4-5。

表 4-3 语音相似度得分情况 - LibriSpeech

Table 4-3 Similarity score of speech - LibriSpeech

说话人	性别	是否参加训练	真实语音	合成语音
4586	男	是	4.32	4.18
4592	女	是	4.65	4.42
4680	女	是	4.31	4.23
4731	女	是	4.55	4.37
4859	男	是	4.28	4.02
4973	男	是	4.39	4.10
5012	女	是	4.41	4.32
5049	男	是	4.50	4.37
5115	女	是	4.45	4.16
5152	男	是	4.34	4.19
4350	男	否	4.36	3.91
4852	男	否	4.28	3.82
5442	女	否	4.37	3.61
5484	女	否	4.56	3.52
5764	男	否	4.60	3.88
6070	女	否	4.29	3.73
6128	女	否	4.54	3.89
6432	男	否	4.47	3.54
6938	女	否	4.38	3.68
7018	男	否	4.35	3.79

表 4-4 语音相似度得分情况 - ESD

Table 4-4 Similarity score of speech - ESD

说话人	性别	是否参加训练	真实语音	合成语音
0002	女	是	4.29	4.18
0004	男	是	4.35	3.92
0005	男	是	4.41	3.87
0007	女	是	4.35	4.02
0008	男	是	4.27	3.86
0009	女	是	4.38	3.69
0011	男	是	4.50	3.75
0012	男	是	4.42	4.10
0015	女	是	4.27	3.95

表 4-4 续表

说话人	性别	是否参加训练	真实语音	合成语音
0019	女	是	4.46	3.88
0001	女	否	4.39	3.46
0003	女	否	4.52	3.51
0006	男	否	4.47	3.42
0010	男	否	4.54	3.21
0013	男	否	4.38	3.42
0014	男	否	4.39	3.28
0016	女	否	4.38	3.13
0017	女	否	4.40	3.39
0018	女	否	4.36	3.44
0020	男	否	4.29	3.18

表 4-5 95%置信区间下的语音相似度平均意见得分

Table 4-5 Speaker similarity mean opinion score with 95% confidence interval

方法	研究	评价数据集	得分
传统方法	CLMSTTS[26]	VCTK + ST	真实语音：4.788 ± 0.254[26] 参与训练：3.418 ± 0.449[26] 未参与训练：3.453 ± 0.390[26]
	SV2TTS[8]	VCTK	真实语音：4.670 ± 0.040[8] 参与训练：4.220 ± 0.060[8] 未参与训练：3.280 ± 0.070[8]
		LibriSpeech	真实语音：4.330 ± 0.080[8] 参与训练：3.280 ± 0.080[8] 未参与训练：3.030 ± 0.090[8]
	NVC[30]	VCTK	真实语音：3.910 ± 0.030[30] 未参与训练：2.590 ± 0.120[30]
本文方法	双阶段训练说话人编码器的方法	LibriSpeech	真实语音：4.420 ± 0.110 参与训练：4.236 ± 0.120 未参与训练：3.737 ± 0.140
		ESD	真实语音：4.391 ± 0.080 参与训练：3.922 ± 0.150 未参与训练：3.344 ± 0.130

由表中数据可以得到，在基于双阶段迁移学习的情感语音克隆方法中，不论是针对训练集中已经出现的语音或者未曾出现的语音，相较于传统的语音克隆模型，其性能都有不同程度的提升。进一步综合以上音色相似性的评测结果可以得出，采用双阶段训练说话人编码器的方式，对生成的克隆语音音色的相似度影响较小，可以基本保持音色的相似度。

此外，本研究通过客观和主观的评估方法，对生成语音的情感相似性进行了评估。为验证双阶段训练说话人编码器方法在情感表达方面的优越性，本研究将传统模型 SV2TTS、ENVC 与新方法进行了横向对比。此外，考虑到结果的有效性，本研究还以情感语音合成模型的结果作为参照对象进行了实验。

在客观测评中，本研究采用均方根误差来衡量目标真实语音与合成语音在时长和基频方面的偏差，并采用梅尔倒谱失真来评估其梅尔频谱之间的差异。均方根误差实验结果如图 4-5 至图 4-6 所示。图 4-5 为原始语音与合成语音基频的均方根误差测评结果，图 4-6 为时长的均方根误差测评结果。由图可知，传统语音克隆模型 SV2TTS 以及单阶段音色克隆模型所生成的语音与目标情感的偏差较大。ENVC 通过语音风格的迁移提升了合成语音的表达能力，但与目标情感还有一定偏差。本研究所提出的双阶段迁移学习情感音色克隆方法与情感语音合成模型的结果最为接近，证明其与目标情感最为相似。

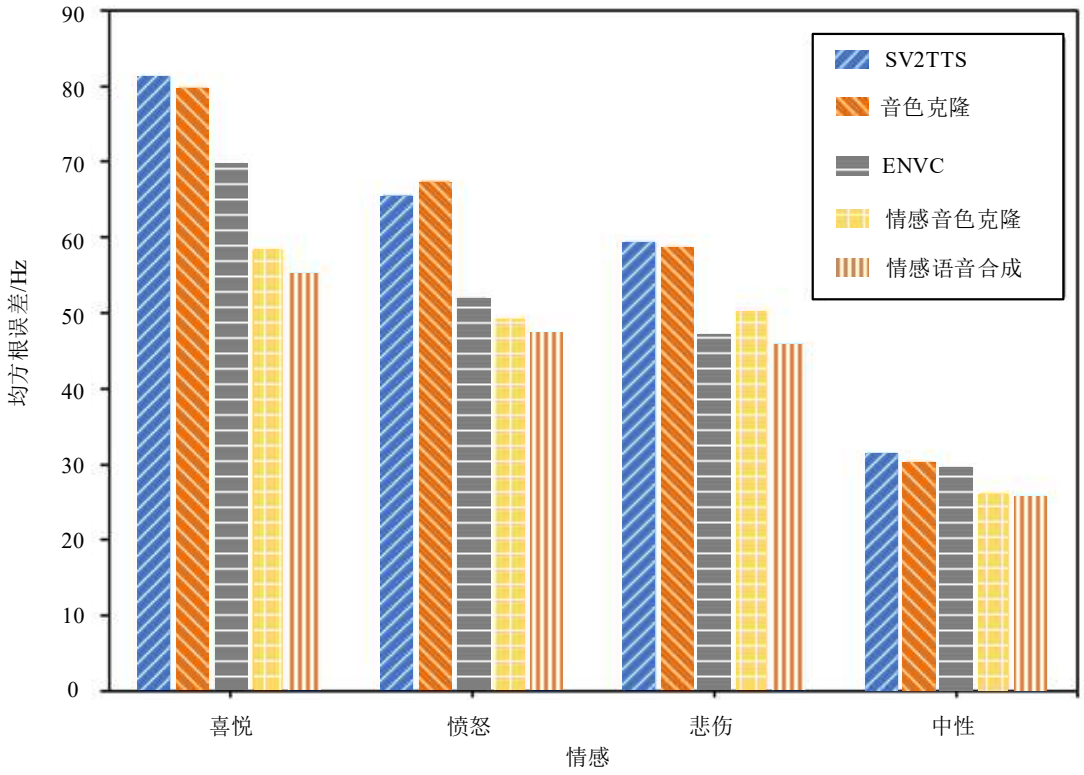


图 4-5 RMSE 实验结果 - 基频

Fig. 4-5 Experimental results of RMSE - pitch frequency

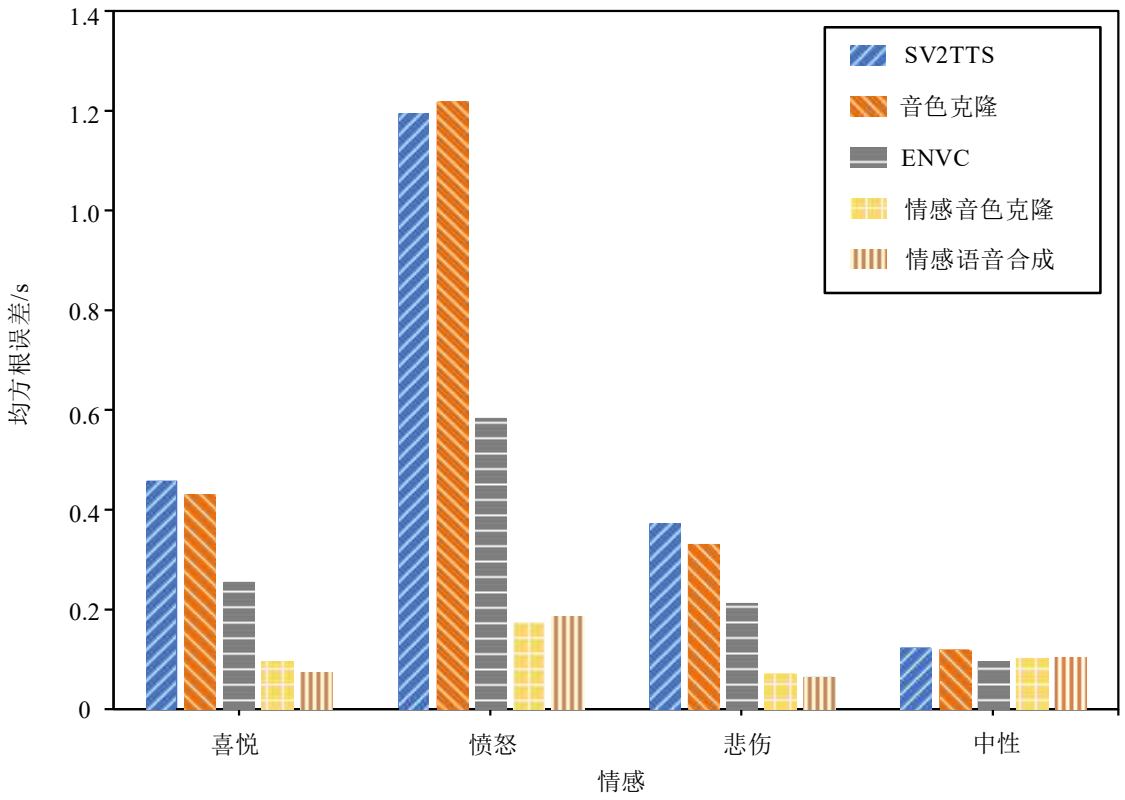


图 4-6 RMSE 实验结果 - 时长

Fig. 4-6 Experimental results of RMSE - duration

MCD 测评结果如表 4-6 所列。前人研究表明，当梅尔倒谱失真小于 8 时，语音识别系统可以正确识别语音[60]。从表中可以看出，传统模型 SV2TTS 与单阶段迁移学习音色克隆模型的结果除中性情感外，均大于 8，而 ENVC 也未能对情感进行较好的表达。与之不同的是，情感音色克隆与情感语音合成模型的实验结果均小于 8。

表 4-6 客观评测结果 - MCD

Table 4-6 Objective evaluation results - MCD

研究	MCD/dB			
	喜悦	愤怒	悲伤	中性
SV2TTS	9.27	10.96	11.28	5.24
音色克隆	9.15	9.64	13.47	7.15
ENVC	6.49	7.52	8.06	6.23
情感音色克隆	5.72	7.04	6.98	5.81
情感语音合成	5.36	6.47	7.62	5.29

在主观评测中，本研究从模型生成的四种情感中各选取 10 句语音，共计 160 个样本参与评价。100 位评测者接受偏好测试和 EMOS 测试来对合成语音进行评估。图 4-7

所示为偏好测评实验结果。由图可知，除中性情感外，传统模型占比最大为 7%，ENVC 最大占比为 15%，本研究所提出的情感音色克隆模型与情感语音合成模型在偏好测试中占比较大，较音色克隆模型分别提升了 24%、19%、14%。

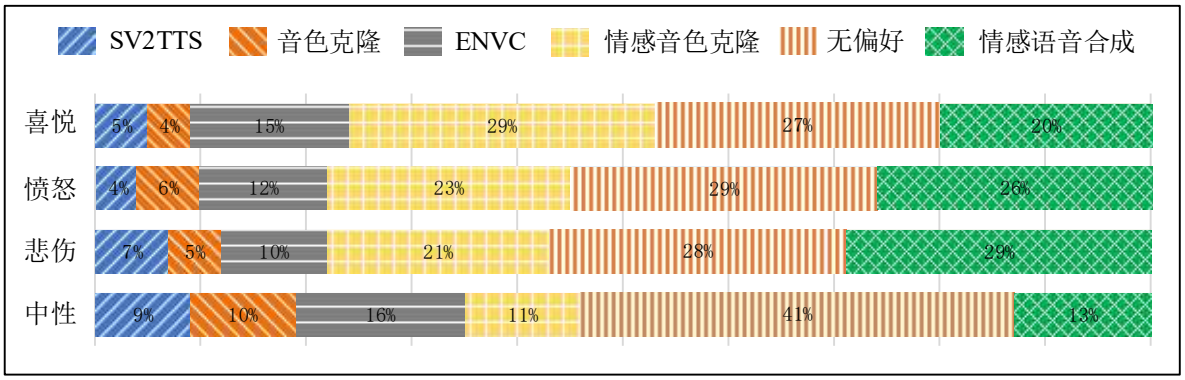


图 4-7 偏好测试实验结果

Fig. 4-7 Experimental results of the preference test

EMOS 测试结果如表 4-7 所列。由结果可知，相比较传统模型 SV2TTS，本研究所提出的情感音色克隆模型在喜悦、愤怒、悲伤分别提升了 1.56、3.14、2.46。在中性情感的测试中，参与评测模型的实验结果没有较大差异。由以上实验结果可以得出，本研究所提出的对说话人编码器进行双阶段迁移学习的情感语音克隆方法，能有效提高克隆语音的情感相似度。

表 4-7 95%置信区间下的情感相似度平均意见得分

Table 4-7 Emotional mean opinion score with 95% confidence interval

研究	EMOS			
	喜悦	愤怒	悲伤	中性
SV2TTS	1.91 ± 0.15	0.68 ± 0.02	1.22 ± 0.14	3.02 ± 0.10
音色克隆	1.25 ± 0.20	0.97 ± 0.13	1.83 ± 0.09	3.11 ± 0.17
ENVC	2.97 ± 0.12	3.03 ± 0.09	2.59 ± 0.21	3.08 ± 0.04
情感音色克隆	3.47 ± 0.05	3.82 ± 0.11	3.68 ± 0.02	3.15 ± 0.18
情感语音合成	3.36 ± 0.10	4.01 ± 0.11	3.84 ± 0.06	3.26 ± 0.05
真实语音	4.76 ± 0.02	4.89 ± 0.06	4.92 ± 0.01	3.50 ± 0.08

4.5 本章小结

本章主要介绍了基于双阶段迁移学习的情感语音克隆方法的过程，并利用此方法进行了相关分析实验，主要包括：对训练中使用到的损失函数进行了介绍，并利用此损失函数对说话人编码器进行音色克隆阶段以及情感克隆阶段的双阶段训练，对训练

过程中所使用到的数据集以及相关参数进行了说明，分析了说话人训练以及情感训练中所使用到的不同的音频特征，同时利用此说话人编码器搭建了情感语音克隆方法，并根据相关的客观评价指标与主观评价指标对合成的情感克隆语音的音色相似度以及情感相似度进行了测评，最后对实验结果进行了分析说明。

第 5 章 基于 e-vector 说话人特征的情感语音克隆改进方法

在第四章中实现了一种基于双阶段迁移学习的情感语音克隆方法，该方法能够有效解决克隆语音自然度较低的问题。然而，通过实验发现情感克隆语音与目标情感之间仍存在一定的差距，并且情感克隆阶段的训练对音色相似度产生了一定的影响。针对上述问题，本研究提出了一种基于 e-vector 说话人特征的情感语音克隆改进方法。本章重点分析了传统说话人编码器进行双阶段训练时存在的问题，并提出了一种基于多特征融合的说话人编码方式，学习提取音频中的说话人情感特征。本章详细地介绍了该方法的原理和实现过程，并对实验结果进行了对比分析。

5.1 e-vector 说话人编码

传统的说话人编码方式[20, 21]主要应用于说话人识别领域，因此其重点关注的是根据目标音频提取出符合特定说话人身份的特征向量，以完成对音频与说话人身份之间所属关系的判断。在语音克隆技术的研究中，旨在刻画目标说话人的音色特征，因此获取能够代表说话人身份的音频特征至关重要。因此，本研究通过迁移学习的方式，将语音识别网络中的说话人编码器作为语音合成框架的说话人嵌入层，从而实现了合成任意目标说话人的语音。

而在情感语音克隆的研究中，不仅要完成对目标音频中音色的克隆，同时还要使得合成语音具有丰富的表达能力，这就意味着说话人编码器不仅要完成说话人身份特征的提取，同时要使得该特征具有丰富的表达能力，具有相应的情感特征。因此，传统说话人识别网络中的说话人编码器无法完全适应这一任务。为了解决这一问题，在本研究提出的基于双阶段迁移学习的情感语音克隆方法中，对说话人编码器分别进行了音色克隆阶段与情感克隆阶段的双阶段训练。

根据主客观实验结果显示，使用本研究所提出的双阶段训练说话人编码器的方法能够提升克隆语音的情感相似度，但是与目标情感仍然存在一定差距。同时，生成语音的音色相似度并未得到显著改善。本研究对上述训练过程进行了分析，发现传统的说话人编码器只针对音频单方面的特征进行学习，因此对情感特征的学习产生了一定影响。此外，在对说话人编码器进行双阶段训练时，情感阶段的训练仅关注情感特征的变化，而未考虑到说话人身份特征，这也会对生成的克隆语音中音色相似性造成一定影响。

在此研究基础上，本章提出了一种新的 e-vector 说话人编码方式，以解决前述方法中未考虑到音频中情感特征的问题。为了适应情感语音克隆的目标，该编码器在根据

目标音频进行说话人编码时，不再是学习单一的说话人身份特征，而是结合情感韵律特征，同时对音频多个方面的特征进行提取。具体来说，在训练过程中，e-vector 说话人编码器除了提取音频中能够代表说话人身份的动态梅尔频率倒谱特征以外，还会提取与情感相关的特征，例如基音频率、时长以及能量特征，从而提取具有情感的说话人特征嵌入。此外，由于这些特征都属于声学特征，包含的信息存在很大的相似性[61]，为了提升不同特征之间的优势互补效果并改善模型的学习效率，本研究利用深度受限玻尔兹曼机（Deep-restricted Boltzmann Machine，简称 DBM）对其进行融合和降维。同时，由于语音信号具有长时随机性和短时平稳性，本研究使用时延神经网络[62]对融合的帧级特征进行时序性建模，并将最终输出的特征嵌入称为 e-vector。

5.2 网络构建

本章的关键在于构建一种新的说话人编码器。前人的研究发现，受限玻尔兹曼机（Restricted Boltzmann Machine，简称 RBM）由于其网络结构的优势，能够学习到数据的高层特征[63]。而深度受限玻尔兹曼机则是深度信念网络（Deep Belief Network，简称 DBN）的基本单元，由多个 RBM 堆叠组合而成，因此具有多层的非线性变换结构，从而能够对复杂的线性函数进行模拟。同时，有相关研究表明，时延神经网络由于其独特的层连方式，能够对节点间的时序关系进行描述，并且能够有效地对长时信息进行建模。因此，本研究构建了一种基于 DBM-TDNN 网络结构的说话人编码器，并将该编码器根据目标音频所提取的说话人特征定义为 e-vector。该说话人编码器架构如图 5-1 所示。

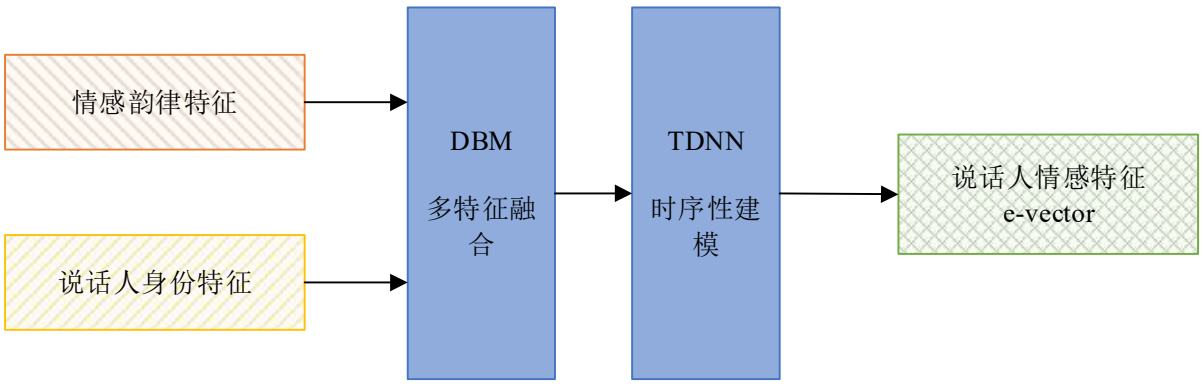


图 5-1 基于 DBM-TDNN 的说话人编码器

Fig. 5-1 Speaker encoder based on DBM-TDNN

5.2.1 深度受限玻尔兹曼机

受限玻尔兹曼机为两层网络结构，如图 5-2 所示，分别为可见层（Visual Layer，简称 VL）与隐藏层（Hidden Layer，简称 HL），其受限体现于将玻尔兹曼机的完全图变

为了可见层与隐藏层间的二分图[64]。

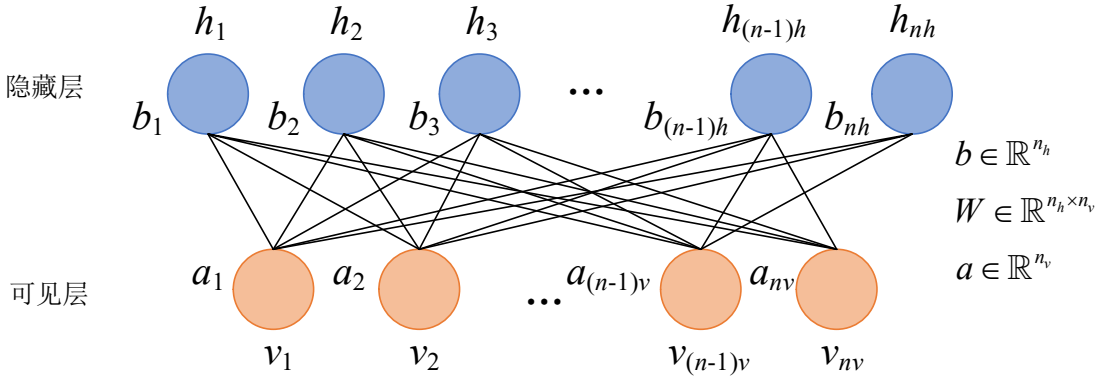


图 5-2 受限玻尔兹曼机结构图

Fig. 5-2 Structure of the restricted Boltzmann machine

受限玻尔兹曼机通过能量函数作为驱动，运用对比散度快速学习方法（Contrastive Divergence，简称 CD）对输入的多个特征进行重构，并生成可以描述输入特征相关性的新特征。其网络能量可以表示为：

$$E(v,h\,|\,\theta)=-\sum_{i=1}^{V}a_i v_i-\sum_{j=1}^{H}b_j h_j-\sum_{i=1}^{V}\sum_{j=1}^{H}v_i W_{ij} h_j \qquad \text{（公式 5-1）}$$

公式中，a_i 和 b_j 分别表示可见层的第 i 个和隐藏层的第 j 个神经元的偏置值，v_i 和 h_j 分别表示其期望值，W_{ij} 表示其之间的连接权重。设可见层与隐藏层的神经元个数分别为 V 和 H。在已知可见层（或隐藏层）的所有神经元状态时，便可以根据条件概率求取隐藏层（或可见层）的某个神经元被激活的概率，求解方法为[65]：

$$p(h_k=1|v)=\text{sigmoid}(b_k+\sum_{i=1}^{V}w_{i,k}v_i) \qquad \text{（公式 5-2）}$$

$$p(v_k=1|h)=\text{sigmoid}(a_k+\sum_{j=1}^{H}w_{j,k}h_j) \qquad \text{（公式 5-3）}$$

由于传统的受限玻尔兹曼机网络结构仅适用于服从二值分布（0-1 分布）的数据，因此该网络在模拟非二值分布的语音信号等数据时，存在一定困难。为了解决这一问题，本研究采用高斯-伯努利分布的受限玻尔兹曼机[66]，以更好地适应非二值分布的数据。同时，DBM 由多层该 RBM 网络堆叠组成，将下层输出作为上层输入，利用能量函数以及神经元的概率求解（参见公式 5-4 至公式 5-6），从而实现了对非二值分布数据的有效建模。

$$E(v,h|\theta)=\sum_{i=1}^{V}\frac{(v_i-a_i)^2}{2\sigma_i^2}-\sum_{j=1}^{H}b_j h_j-\sum_{i=1}^{V}\sum_{j=1}^{H}\frac{v_i}{\sigma_i}W_{ij}h_j \qquad \text{（公式 5-4）}$$

$$P(v_k=1|h)\sim N(a_k+\sigma_i\sum_{j=1}^{H}w_{j,k}h_j,\sigma_i^2) \quad \text{（公式 5-5）}$$

$$P(h_k=1|v)=sigmoid(b_k+\sum_{i=1}^{V}w_{i,k}\frac{v_i}{\sigma_i}) \quad \text{（公式 5-6）}$$

公式中，σ 表示高斯函数中的标准方差，θ 表示由 W、a、b 以及所构成的参数集合，即 $\theta=(W,a,b,\sigma)$，其余字符含义均与公式 5-1 中所述一致。图 5-3 展示了 DBM 结构的示意图。

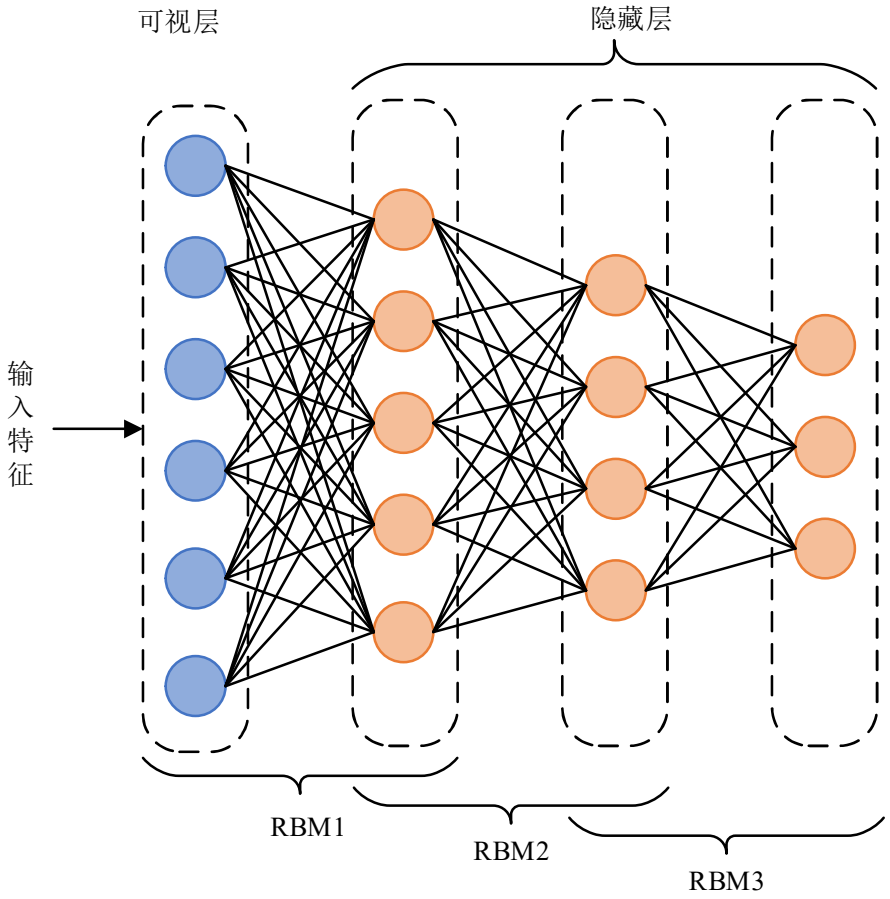


图 5-3 DBM 结构

Fig. 5-3 Structure of deep-restricted Boltzmann machine

5.2.2 时延神经网络

时延神经网络是一种多层的前馈神经网络[67]，其最初被提出来用于解决语音识别任务。由于语音信号具有时序性质，所包含的信息会随着时刻的推移而产生变化，因此不同帧之间的信息具有一定的时序相关性。某一时刻的静态声学特征无法良好表达这种相关性，因此对这种长时间的时序动态信息相关性进行建模至关重要。为了解决这个问题，TDNN 通过拼接前后若干时刻的特征来描述每个隐藏层的特征。与传统神经网络相比，TDNN 的输出不仅描述了语音信号当前时刻的特征，还考虑到了前后若

干时刻与当前时刻的关系，因此可以有效地对语音信号的上下文信息进行建模。TDNN 的网络结构如图 5-4 所示。

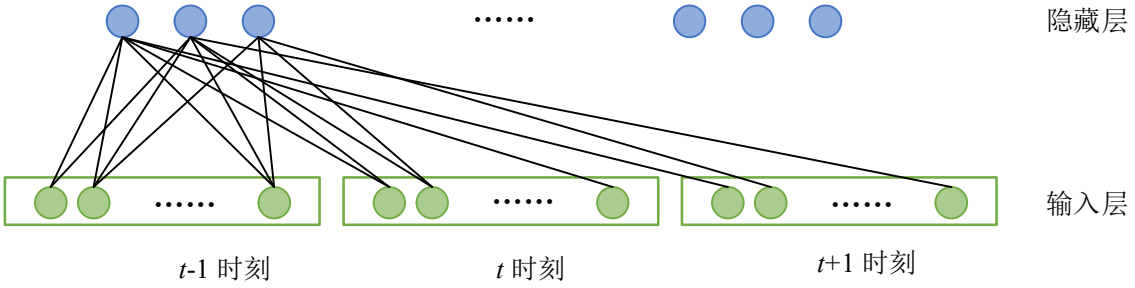


图 5-4 单层时延神经网络结构

Fig. 5-4 Structure of single-layer time-delay neural network

因此，在使用时延神经网络处理语音片段时，需考虑时延问题。具体而言，当前时刻的输入不仅涵盖当前时刻的信息，还需包括前一时刻和后一时刻的信息。图中，t 表示当前时刻，那么前一帧表示为$t-1$，后一帧表示为$t+1$。将 t、$t-1$ 以及$t+1$三个时刻的输入信息进行拼接，作为隐藏层的输入，这样就实现了对语音序列的时延操作[68]，从而更加准确的处理语音。在实际应用中，可以通过增加时延网络层数以及时延范围，使网络捕获到更大时域范围内的特征。

时延神经网络有两个重要优点：移位不变性和权重共享。移位不变性指的是对于一个输入信号，在不同时间上的输入信息被等同对待，即对于同一段语音信号不同时刻的信息，模型对其的处理是相同的。这种方法可以有效地捕捉到语音信号中的时间相关性，并使模型具有时间上的不变性。权重共享是指在模型的卷积层中，每个神经元在所有时间步骤上共享相同的权重。这种方法可以显著减少模型的参数数量，同时也能降低需要训练的参数数量，从而加速训练过程。若把图 5-4 所示的时延神经网络结构看成是一个窗口，那么当使用时延神经网络处理语音片段时，该窗口会沿着时间轴进行滑动，处于相同时间位置的输入将共享同一组权重矩阵。该网络的工作原理如图 5-5 所示。

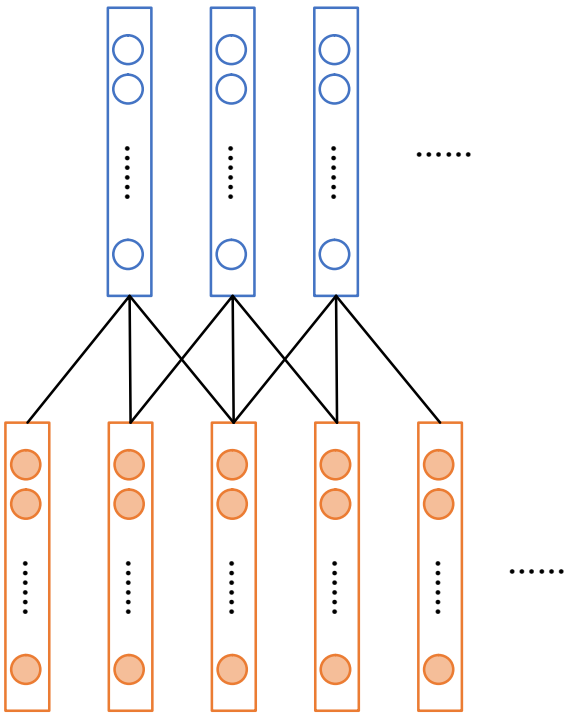


图 5-5 时延神经网络工作原理

Fig. 5-5 How time-delay neural networks work

5.3 特征提取

传统的说话人编码器利用说话人识别框架进行训练，以学习提取能够表征说话人身份的音频特征向量。然而，此向量缺乏对情感韵律特征的表征，不适用于情感语音克隆的目标。为了解决这一问题，本研究提出了一种基于 DBM-TDNN 的说话人编码方法，有效地提取了音频中的说话人情感特征。具体而言，本研究首先对训练音频进行预处理，提取能够反映情感变化和说话人身份的相关特征，并将这些特征作为 DBM 层的输入。通过 DBM 的网络结构进行拼接和降维操作，对输入特征进行重构，多次重复该操作后，基音频率、音频时长、能量以及 MFCCs 的低层次特征被统计映射为更适合情感语音克隆目标的说话人情感特征。由于语音是一种短时平稳的时序性信号，本研究利用 TDNN 网络对 DBM 的输出特征进行进一步处理，以充分利用帧前后的关系及信息，并建立语音的时序性特征。

5.3.1 DBM 特征融合

为了从目标音频中提取具有情感的说话人特征向量，本研究首先构建了 DBM 网络，对音频中的情感韵律特征和说话人身份特征进行多特征融合。DBM 网络结构如图 5-6 所示。

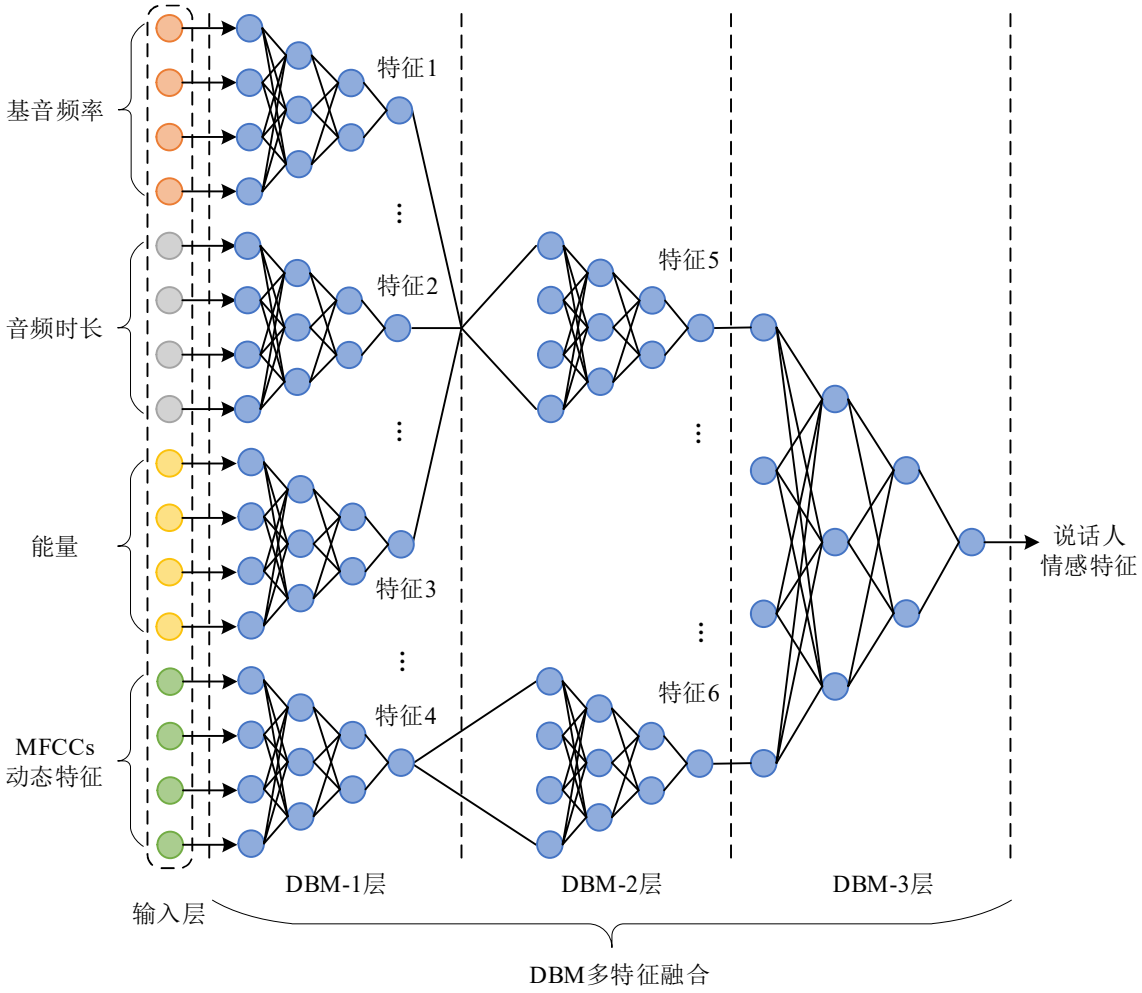


图 5-6 DBM 特征融合结构图

Fig.5-6 Structure for feature fusion using DBM

该 DBM 网络由多层 RBM 组合而成，在本研究中，每个 RBM 子块都选取了四层网络结构，各个子块的具体神经元数目情况如表 5-1 中所列。

表 5-1 DBM 网络结构神经元数目

Table 5-1 Number of neurons in the DBM

DBM-1 层	DBM-2 层	DBM-3 层
(8, 55, 45, 6)		
(8, 55, 45, 6)	(18, 45, 35, 15)	(45, 50, 40, 30)
(8, 55, 45, 6)		
(42, 55, 45, 40)	(40, 45, 35, 30)	

在本研究的训练过程中，首先对音频进行了预处理，包括消除数据中的噪声和静音部分，随后提取了基音频率、音频时长、能量以及 MFCCs 动态特征。最后，采用

DBM 网络对这些特征进一步处理。具体输入网络的特征情况如表 5-2 所列。

表 5-2 DBM 输入特征

Table 5-2 Input features of DBM

特征类型	特征名称	特征维数
情感韵律特征	基音频率	1 ~ 8
	音频时长	9 ~ 16
	能量	17 ~ 24
说话人身份特征	MFCCs 动态特征	25 ~ 66

本研究利用 DBM 网络处理输入特征的流程如表 5-3 所列，特征处理以及融合操作的具体过程为：

第一步：提取训练音频的情感韵律特征与说话人身份特征，输入 DBM-1 层进行降维操作，得到隐藏层输出的特征 1、特征 2、特征 3 以及特征 4；

第二步：根据音频特征的相关属性，基音频率、音频时长、能量都是能有效地反映语音中所蕴含情感的特征，因此本研究将特征 1、特征 2 以及特征 3 进行线性拼接，输入 DBM-2 层进行特征融合以及降维操作，得到特征 5，同时对特征 4 进行降维处理得到特征 6；

第三步：为了从目标音频中提取说话人情感特征向量，本研究将特征 5 与特征 6 进行线性拼接，将其作为 DBM-3 层的输入进行特征融合与降维，并将最后一个隐藏层的输出作为最终结果。

表 5-3 DBM 特征处理过程

Table 5-3 The process of processing features by DBM

DBM-1 层		DBM-2 层		DBM-3 层	
输入特征	输出特征	输入特征	输出特征	输入特征	输出特征
基音频率	特征 1	特征 1+特征 2+特征 3	特征 5	特征 5+特征 6	说话人情感特征
音频时长	特征 2				
能量	特征 3				
MFCCs 动态特征	特征 4	特征 4	特征 6		

5.3.2 TDNN 时序建模

为了获取帧级别的说话人情感特征，本研究采用了基于 TDNN 网络的结构对 DBM 网络输出的特征进行进一步处理，如图 5-7 所示。由于语音是一种短时平稳的时序性信号，因此，使用 TDNN 网络可以有效地捕捉到每个说话人的发音特征中蕴含的时序结构信息。在本研究中，多层 TDNN 网络的学习可以使系统充分利用帧前后的关系及信

息，进而提高说话人特征的有效性。最后，输出层对最后一层 TDNN 隐藏层的结果进行处理，所得到的即为音频帧级别的说话人情感特征。

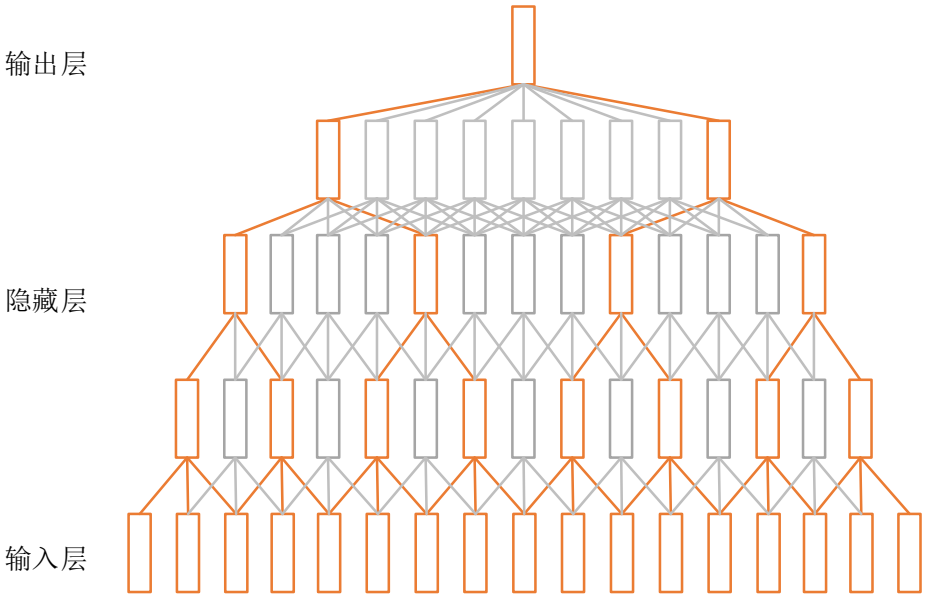


图 5-7 TDNN 时序建模结构图

Fig. 5-7 Structure for temporal modeling using TDNN

本研究提出的 DBM-TDNN 模型包含一个输入层、三个隐藏层和一个输出层，其中每个隐藏层均为 TDNN 网络。使用 DBM 所提取的音频融合特征向量作为输入。在第一个隐藏层中，将帧特征 $\{t-1,t,t+1\}$ 进行时序拼接后作为输入。然而，由于帧特征在处理过程中会被重复计算，导致结果间存在信息重叠。因此，本研究采用了子采样[69]的方法来解决这个问题。在第二层隐藏层处理时，本研究采取了跨度为 1 的方式对帧特征进行子采样，将 $\{t-1,t+1\}$ 的帧级特征进行拼接作为输入。类似地，最后一个隐藏层对输入特征进行处理时，采取了跨度为 3 的方式对其进行子采样。最终，输出层将最后一个隐藏层的结果进行拼接，生成包含所有输入信息的帧级别特征向量，作为训练语音的最终说话人情感特征。因此，在时序建模的过程中，本研究通过调整 TDNN 网络每一层的参数来改变输出特征拼接的时间步长以及特征之间的依赖关系，从而提高说话人特征的有效性。

在本研究中，参考了 x-vector 说话人识别系统[21]的思想，在语音帧级别的说话人特征基础上进行了改进。为了更好地利用音频的整体信息，本研究采用了统计层和全连接层来构建音频句子级别的特征。该方法考虑到了音频的整段信息，因此能够更好地利用数据的整体信息。具体而言，本研究首先利用统计层计算特征的均值和标准差，以聚合帧特征。然后，将得到的特征进行拼接，并输入到两层全连接层中进行再次处理。最终，将全连接层的输出定义为说话人情感特征 e-vector。该方法的结构如图 5-8 所示。

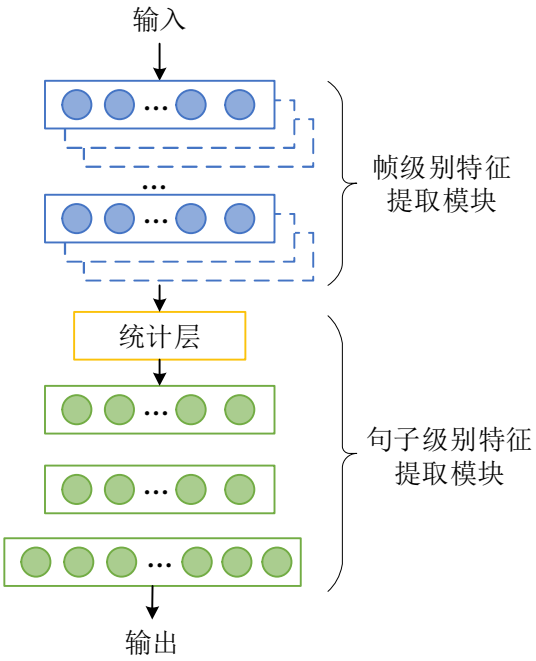


图 5-8 x-vector 说话人识别系统框架[70]

Fig. 5-8 Framework for x-vector speaker recognition system[70]

5.4 说话人编码器双阶段训练方式的改进

在训练基于 DBM-TDNN 的说话人编码器时，目的是从输入音频中提取说话人特征向量 e-vector，这需要同时学习音频中的身份特征与情感特征。若采用传统的一阶段训练方法，则对训练语料库提出了较高的要求。具体而言，为了学习说话人的身份特征，训练数据集中需要包含丰富的说话人标签。而需要完成音频中情感特征的学习，这就意味着训练音频需同时具有相应的情感标签。因此，获取符合这些标准的数据集具有挑战性。除此之外，此训练方式需同时达到对说话人与情感标签判断的高准确性，这可能导致训练时间较长、难以收敛、泛化能力较低等问题。为了解决这个问题，本章采用了前一章中描述的说话人训练以及情感训练的双阶段训练方式，并对训练过程做出了相应改进。

本研究采用了基于语音识别语料库的预学习方法，对声纹特征的表示进行学习。接着，利用情感语料库进行情感训练，以完成对说话人特征和情感特征的良好表征。具体而言，在音色克隆阶段的训练中，情感特征不参与训练，因此在提取训练音频中的相关特征时，情感特征被设置为全 0 占位。当说话人标签一致时，应表现出较高的相似度。在情感克隆阶段的训练中，e-vector 作为说话人情感特征，将音频中的说话人信息和情感信息进行融合，因此，训练中使用的标签也从情感转变为具体说话人的情感。当情感标签和说话人标签同时一致时，应体现出较高的相似性；当两个标签都不一致

时，应体现出较低的相似性；而当其中部分标签一致时，应体现出适度的相似度。训练中使用的损失函数与 4.2 节中所述的一致。

5.5 实验与分析

为了评估说话人模型 e-vector 的性能，本研究利用该说话人编码方式对情感语音克隆模型进行改进，并利用主观分析方法对模型生成的语音的音色相似度以及情感相似度进行验证。

5.5.1 实验设置

在训练说话人编码器时，音色克隆阶段使用的是通用语音数据集 LibriSpeech 中的 train-clean-360，该部分中说话人较丰富，共有 921 位说话人，因此适合用于学习说话人身份特征。在情感克隆阶段的训练中，使用了 ESD 情感语音数据集。ESD 数据集中每条记录不仅有相应的文本及情感标注，还有相应的说话人信息，因此适应情感阶段的训练需求。由于基于 DBM-TDNN 网络结构的说话人编码器模型较为复杂，本研究使用 ADAM 作为优化器来训练模型。在训练过程中，初始学习速率设置为 0.001，每次训练的批处理数量为 32。

为了验证说话人编码器对音频中说话人身份特征以及情感特征的提取效率，本研究采用了两部分数据集进行实验。第一部分为 LibriSpeech-test，其中包含 40 位说话人，男女的比例为 1:1，每位说话人具有 8 分钟的音频数据。第二部分为情感语音数据集 IEMOCAP[71]，该数据集中包含了来自 10 名演员的大约 12 小时的多模态数据，其中包括语音、视频、面部表情、身体语言等。这些数据是在模拟的情境中进行收集，演员需要按照指定的情感状态对内容进行表演。情感状态包括愤怒、厌恶、喜悦、无聊、悲伤和中性。在此数据集中，作者将对话处理为 3 秒至 15 秒的音频，并对其进行了情感标注。

5.5.2 实验结果与分析

本章综合分析了使用 e-vector 说话人特征的情感语音克隆改进方法所生成的情感克隆语音的质量，并结合第四章中所述基于传统说话人模型所构建的情感语音克隆模型的性能进行对比分析。为了验证情感语音克隆改进方法的音色克隆效果，在 LibriSpeech 中的 test-other 测试集中随机选取 10 位说话人，第一次随机抽取 5 位男性（ID 分别为：3005，4198，5764，7018，8188），第二次随机抽取 5 位女性（ID 分别为：3331，5442，6128，6938，8280），将真实音频与合成音频进行配对，使用说话人相似性平均意见得分来评估合成语音的音色相似性，由评测者按照表 3-3 标准对其进行打分，得分情况如表 5-4 所列。

表 5-4 说话人相似度得分情况

Table 5-4 Speaker similarity scores

说话人 ID	MOS		
	传统说话人模型	e-vector	真实语音
3005	3.66	3.86	4.41
3331	3.75	4.01	4.76
4198	3.52	3.76	4. 50
5442	3.88	3.80	4.54
5764	3.54	3.94	4.65
6128	3.51	3.66	4.58
6938	3.62	3.68	4.55
7018	3.81	3.85	4.46
8188	3.52	3.69	4.12
8280	3.50	3.90	4.37

本研究为了比较传统说话人编码器和 e-vector 的得分情况，分析它们的趋势以及比较不同模型之间得分的差异和相似之处，采用了折线图进行可视化分析。研究结果如图 5-9 所示。

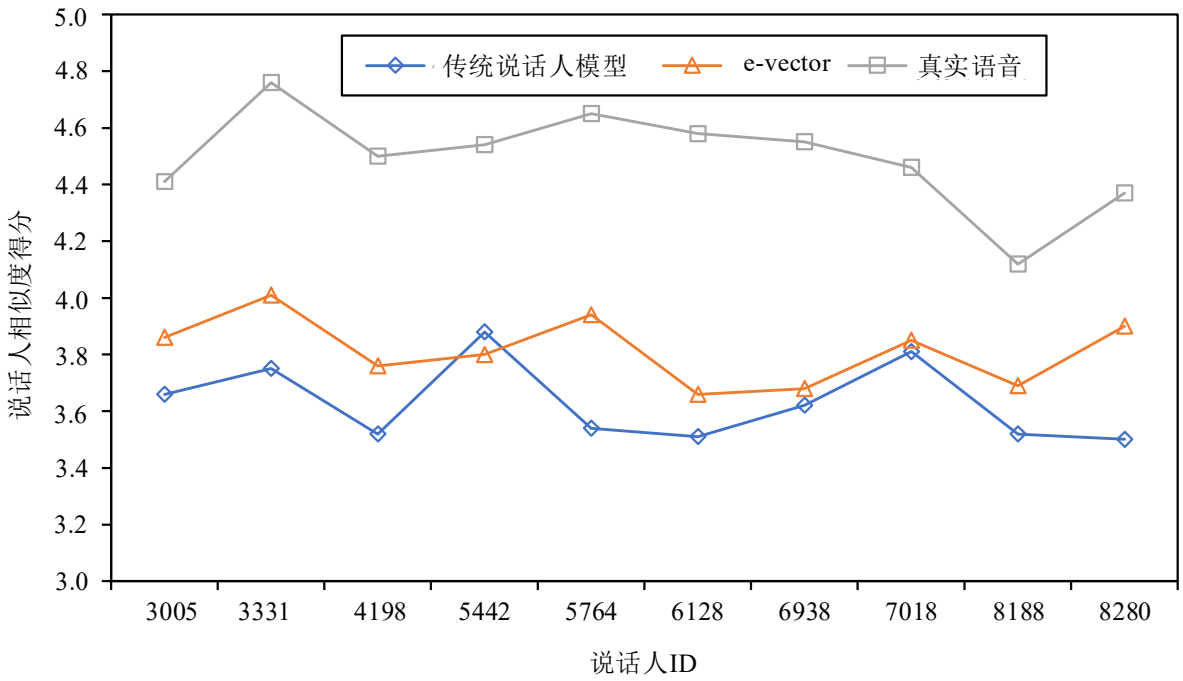


图 5-9 说话人相似度得分统计结果

Fig. 5-9 Statistical results of speaker similarity scores

从实验结果可以得出，相对于传统的说话人模型，采用基于 e-vector 说话人编码方

式的情感语音克隆方法，在生成的情感克隆语音中所表现出的音色相似度整体得到了显著提升，并且模型表现出了更加稳定的性能。具体而言，采用 e-vector 方法所得到的说话人相似度结果的总体方差为 1.25E-3，而传统说话人模型所得结果的总体方差为 1.74E-3。

为了验证基于 e-vector 说话人特征的情感语音克隆改进方法的情感克隆效果，本研究使用了未参与训练的情感语音数据集 IEMOCAP 进行测试。具体方法是从数据集中随机选取了 10 位演员的中性、喜悦、愤怒和悲伤情感各 10 句音频，并进行了 EMOS 测评。测试结果列于表 5-5 中。

表 5-5 95%置信区间下的情感相似度平均意见得分

Table 5-5 Emotional mean opinion score with 95% confidence interval

情感	模型	
	传统说话人模型	e-vector
中性	3.15 ± 0.12	3.21 ± 0.20
喜悦	3.54 ± 0.06	3.79 ± 0.13
愤怒	3.71 ± 0.05	3.92 ± 0.19
悲伤	3.60 ± 0.18	3.79 ± 0.11

根据实验结果，本研究提出的基于 e-vector 说话人编码方式的情感语音克隆改进方法在提高克隆语音的情感表达能力方面表现出了显著的效果。具体而言，该方法在愤怒情感方面表现最佳，而在中立情感方面得分相对较低。此外，在喜悦和悲伤情感方面，该方法也取得了较好的结果。因此，该方法对于改善传统说话人编码器合成的克隆语音的情感相似度具有积极作用。

5.6 本章小结

本章主要介绍了基于 e-vector 说话人特征的情感语音克隆改进方法的过程，并对此说话人编码器构建的情感语音克隆模型进行相关实验分析，主要包括：对传统的说话人编码器进行分析并提出了符合本研究目标的 e-vector 说话人编码器，利用深度受限玻尔兹曼机与时延神经网络构建模型，并对其原理及优势进行了说明。之后，利用上述构建的说话人编码器提取说话人情感特征嵌入。其中，深度受限玻尔兹曼机用于融合说话人特征与情感特征，再通过时延神经网络对特征进行时序性建模，同时对说话人编码器的双阶段训练方式进行了相应改进。最后介绍了实验相关设置参数，并对实验结果进行了分析。

第 6 章 总结与展望

6.1 本文工作总结

语音克隆技术是语音合成技术的一个重要分支，致力于生成具有特定说话人特征的语音，以实现个性化语音的生成。目前主要的语音克隆方法包括说话人自适应和说话人编码。尽管这些方法可以生成与目标说话人相似的语音，但由于在训练说话人编码器时，未考虑音频中的情感韵律特征，因此生成的克隆语音缺乏表达能力和较好的自然度。然而，表达能力对于提升克隆语音的质量至关重要。具有丰富表达能力的克隆语音可以改善智能驾驶的交互体验，为影视作品提供逼真的配音，并有利于创造智能情境下的虚拟主播等。因此，进一步提升克隆语音的表达能力是当前研究的重要方向。针对以上问题，本研究首先构建了基于单阶段迁移学习的语音克隆模型，并提出了基于双阶段迁移学习的情感克隆语音方法，对说话人编码器进行说话人训练以及情感训练。此外，本研究还提出了基于 DBM-TDNN 的多特征融合说话人编码方式 e-vector 以进一步提升情感语音克隆方法的性能。

具体而言，本研究在说话人识别技术基础之上，通过迁移学习的方法，利用说话人编码器提取目标音频中的说话人身份特征，并将此特征作为语音合成模型的输入，对梅尔频谱图的生成过程进行调节。然后，利用声码器对合成器所输出的梅尔频谱图进行处理，生成最终的克隆语音。在本研究中，采用去除了 WaveNet 的 Tacotron2 架构作为合成器，并对其做出了相应改进。声码器采用的是改进后的 WaveRNN 架构。同时，通过主观实验对克隆语音的音色相似度以及自然度进行评价。根据实验结果可知，参与训练的输入音频所得到的音色相似度与自然度与真实语音的平均差距分别为 0.600 和 1.169，未参与训练的语音的平均差距分别为 0.936 和 1.345，由此结果可以得出，克隆语音的自然度还有待提升。为了解决此问题，本研究提出了对说话人编码器进行音色克隆阶段以及情感克隆阶段的双阶段迁移学习的情感语音克隆方法。具体而言，在音色克隆阶段中，利用说话人识别语料库，使用说话人标签进行训练，学习音频中目标说话人身份特征的提取；在情感克隆阶段中，以音色克隆阶段的参数初始化模型，利用情感语料库，使用情感标签进行训练，学习提取音频中的相关情感韵律特征。为了验证克隆语音的质量，本研究进行了主客观实验对克隆语音的音色相似度以及情感相似度进行评价。所有实验结果均表明，克隆语音的情感相似度得到了提升，同时音色相似度也有一定的改善。但是，由 EMOS 评测结果可以得出，生成的语音与目标情感还有一定偏差，除中性情感外，其中差距最大的为喜悦（1.36），最接近的为愤怒（1.24）。对此问题进行研究发现，传统说话人编码器利用说话人识别框架进行训练，

目的是根据目标音频提取能代表说话人身份的特征向量，进而对语音和说话人的所属关系进行判别。根据此过程可以得出，传统的说话人编码器在训练时，只针对音频单个方面的特征进行学习，因此对情感特征的学习产生了一定影响。同时对双阶段训练过程进行分析可以得出，即使以音色克隆阶段的训练结果初始化说话人编码器的参数，在情感阶段的训练过程中，仅关注情感特征的变化，未考虑到说话人身份特征，也会对音色相似度的提升产生影响。为了解决此问题，本研究提出了基于 DBM-TDNN 的说话人编码方式 e-vector。该说话人模型在训练时，利用多特征融合的方式，对能代表说话人身份的MFCCs动态特征与情感韵律特征进行同时学习。实验结果表明，e-vector较传统说话人模型，音色相似度最高提升了 0.26，情感相似度最高提升了 0.48（悲伤）。

6.2 未来工作展望

近年来，语音克隆作为生成个性化语音的重要技术得到了更进一步的发展，但是克隆语音的表达能力是此类研究中重要但经常被忽视的方面。在此背景下，本研究首先构建了基于单阶段迁移学习的语音克隆模型以生成任意目标说话人的语音，并提出了基于双阶段迁移学习的情感语音克隆方法以提升克隆语音的自然度，最后提出新的说话人模型 e-vector 对传统说话人编码方式进行再次改进。研究结果表明，通过对 e-vector 说话人编码器进行双阶段训练，并利用其在目标音频中提取的说话人情感特征构建情感语音克隆方法，所生成的语音具有较好的音色相似性与情感相似性，能更好的对说话人身份特征以及情感特征进行表达。尽管目前的研究在实验中取得了较好的结果，但是仍然存在局限性。

首先，根据实验结果显示，未参与训练的说话人在音色相似性方面未能获得明显的改善，与参与训练的说话人之间存在较大的差距。对实验过程分析得出，这是由于说话人编码器受到参与训练和未参与训练的说话人之间泛化差异的影响。说话人自适应的语音克隆方法虽可避免此问题，但是该方法需要对模型进行上千步的微调才能达到理想的效果[31]。因此未来可考虑在利用说话人编码方式的同时，结合说话人自适应的方法，利用目标语音对模型进行少量微调，以进一步提升未参与训练的说话人的音色相似性。

其次，本研究虽然利用双阶段训练说话人编码器的方法改善了传统语音克隆模型中缺乏情感表达的问题，但是由实验结果可以得出与目标情感还有一定的差距。分析说话人编码器训练过程可以发现，情感克隆阶段的训练中，选取的情感特征为基音频率、能量以及音频时长三个基础特征。然而，由于语音信号的复杂性，还有一些复杂特征能体现出情感的变化。例如与基音频率相关的特征还有均方差等，短时能量变化率、短时平均振幅与能量相关，语速、短时平均过零率等与时长相关。因此，在未来

的研究中可以丰富情感阶段中所训练的特征，以进一步提升克隆语音的情感相似度。

然后，在实验过程中发现，模型生成克隆语音的耗时较长，这可能是由于模型的复杂度过高导致的。为了平衡克隆语音质量与模型复杂度之间的关系，可以采取优化网络结构，减少模型中神经元的个数或者层数，或者采用更合适的激活函数来提高网络的运行效率。

最后，本研究在实验中选取了四个有代表性的情感：喜悦、愤怒、悲伤、中性参与训练。研究者们提出了多种关于人类情感分类的理论，其中美国心里学家 Ekman 提出的喜悦、伤心、惊讶、憎恶、气愤和恐惧 6 项基本情感理论被广泛接受[72]。但是由于可获取到的情感语料库的限制，本文未对其进行充分研究。因此在今后的研究中，可以利用更加丰富的情感语料库对实验结果进行多重验证。

参考文献

[1] 张祖红.语音信号的处理技术及其应用分析[J]. 电子技术, 2022, 51(12): 151-153.

[2] 杨帅, 乔凯, 陈健, 等. 语音合成及伪造、鉴伪技术综述[J]. 计算机系统应用, 2022, 31(07): 12-22.

[3] KAUR N, SINGH P. Conventional and contemporary approaches used in text to speech synthesis: A review[J]. Artificial Intelligence Review, 2022: 1-44.

[4] CHEN B, DU C, YU K. Neural fusion for voice cloning[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 1993-2001.

[5] 张雅欣, 张连海. 一种基于 x-vector 说话人特征的语音克隆方法[J]. 信息工程大学学报, 2020, 21(06): 664-669.

[6] GIBIANSKY A, ARIK S, DIAMOS G, et al. Deep voice 2: Multi-speaker neural text-to-speech[J]. Advances in Neural Information Processing Systems, 2017, 30: 852-943.

[7] KONS Z, SHECHTMAN S, SORIN A, et al. High quality, lightweight and adaptable TTS using LPCNet[C]//2020 IEEE Spoken Language Technology Workshop (SLT). Piscataway: IEEE, 2021: 317-324.

[8] JIA Y, ZHANG Y, WEISS R, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis[J]. Advances in Neural Information Processing Systems, 2018, 31: 825-836.

[9] 仝玉杰, 王骁, 毛悟宇, 等. 智能数字替身技术在抑郁症诊疗中的研究进展[J]. 国际精神病学杂志, 2022, 49(04): 598-601.

[10] NEEKHARA P, HUSSAIN S, DUBNOV S, et al. Expressive neural voice cloning[C]//Asian Conference on Machine Learning. New York: PMLR, 2021: 252-267.

[11] SHEN J, PANG R, WEISS R J, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Canada: IEEE, 2018: 4779-4783.

[12] GRAVES A. Generating sequences with recurrent neural networks[J]. arXiv preprint arXiv:1308.0850, 2013.

[13] VASWANI A, SHAZEER N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[14] LI N, LIU S, LIU Y, et al. Neural speech synthesis with transformer network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii: AAAI, 2019, 33(01): 6706-6713.

[15] REN Y, RUAN Y, TAN X, et al. Fastspeech: Fast, robust and controllable text to speech[J]. Advances in Neural Information Processing Systems, 2019, 32: 3171–3180.

[16] PARK J, ZHAO K, PENG K, et al. Multi-speaker end-to-end speech synthesis[C]//AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 10046-10054.

[17] KENNY P, DUMOUCHEL P. Experiments in speaker verification using factor analysis likelihood ratios[C]//ODYSSEY04-The Speaker and Language Recognition Workshop. Toledo: IEEE, 2004: 219-226.

[18] ROUF R J, ARIFIANTO D. Speaker forensic identification using joint factor analysis and i-vector[C]// Journal of Physics: Conference Series. Temple Circus: IOP Publishing, 2021, 1896(1): 012026.

[19] DEHAK N, KENNY P J, DEHAK R, et al. Front-end factor analysis for speaker verification [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 19(4): 788-798.

[20] VARIANI E, LEI X, MCDERMOTT E, et al. Deep neural networks for small footprint text-dependent speaker verification[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processi-

ng(ICASSP). Florence: IEEE, 2014: 4052-4056.
[21] SNYDER D, GARCIA-ROMERO D, SELL G, et al. X-vectors: Robust dnn embeddings for speaker recognition[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Jersey: IEEE, 2018: 5329-5333.
[22] BAI Z, ZHANG X L. Speaker recognition based on deep learning: An overview[J]. Neural Networks, 2021, 140: 65-99.
[23] VALIN J M, SKOGLUND J. LPCNet: Improving neural speech synthesis through linear prediction [C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Ottawa: IEEE, 2019: 5891-5895.
[24] OORD A, DIELEMAN S, ZEN H, et al. Wavenet: A generative model for raw audio[C]// Neural Information Processing Systems (NIPS). California: Curran Associates, 2016: 6199-6207.
[25] CHEN Y, ASSAEL Y, SHILLINGFORD B, et al. Sample efficient adaptive text-to-speech[J]. arXiv preprint arXiv:180910460, 2018.
[26] CHEN M, CHEN M, LIANG S, et al. Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding[C]//Interspeech. Graz: ISCA, 2019: 2105-2109.
[27] JEMINE C. Real-time-voice-cloning[D]. Liége: University of Liége, 2019.
[28] CAI W, CHEN J, LI M. Exploring the encoding layer and loss function in end-to-end speaker and language recognition system[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(12):2287-2302.
[29] 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]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York: IEEE, 2020: 6184-6188.
[30] ARIK S, CHEN J, PENG K, et al. Neural voice cloning with a few samples[J]. Advances in Neural Information Processing Systems, 2018, 31: 6817-6827.
[31] HUANG S F, LIN C J, LIU D R, et al. Meta-TTS: Meta-learning for few-shot speaker adaptive text-to-speech[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 1558-1571.
[32] LEE Y, RABIEE A, LEE S-Y. Emotional end-to-end neural speech synthesizer[C]// Interspeech. Graz: ISCA,2019: 1142-1146.
[33] HU T Y, SHRIVASTAVA A, TUEL O, et al. Unsupervised style and content separation by minimizing mutual information for speech synthesis[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York: IEEE, 2020: 3267-3271.
[34] LEI Y, YANG S, WANG X, et al. Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 853-864.
[35] PAMISETTY G, SRI RAMA MURTY K. Prosody-TTS: An end-to-end speech synthesis system with prosody control[J]. Circuits, Systems, and Signal Processing, 2023, 42(1): 361-384.
[36] 韩纪庆, 张磊, 郑铁然. 语音信号处理.第 2 版[M]. 北京: 清华大学出版社, 2013.
[37] 王洪雅, 冯培勋, 程明亮, 等. 人体发音器官数字化三维模型构建及可视化研究[J]. 解剖学研究, 2021, 43(02): 188-190.
[38] 张昕, 胡航烨, 曹欣怡, 等. 基于 Tacotron 模型和韵律修正的情感语音合成方法[J]. 数据采集与处理, 2022, 37(04): 909-916.
[39] 曹欣怡. 基于韵律参数优化的情感语音合成[D]. 南京: 南京师范大学, 2020.
[40] 张兴明, 杨凯. 深度学习说话人识别中语音特征参数提取研究[J]. 现代计算机, 2021, (08): 3-7+13.

[41] 姜路遥,李兵兵. 汉语听觉阈下启动效应：来自听觉掩蔽启动范式的证据[J/OL]. 心理学报:1-13[2023-03-20].http://kns.cnki.net/kcms/detail/11.1911.B.20221229.1820.008.html.

[42] SIMONSMEIER B A, FLAIG M, DEIGLMAYR A, et al. Domain-specific prior knowledge and learning: A meta-analysis[J]. Educational Psychologist, 2022, 57(1): 31-54.

[43] CHAN J Y L, BEA K T, LEOW S M H, et al. State of the art: a review of sentiment analysis based on sequential transfer learning[J]. Artificial Intelligence Review, 2023, 56(1): 749-780.

[44] NIU S, LIU Y, WANG J, et al. A decade survey of transfer learning (2010–2020)[J]. IEEE Transactions on Artificial Intelligence, 2020, 1(2): 151-166.

[45] CODY T, BELING P A. A systems theory of transfer learning[J]. IEEE Systems Journal, 2023, 17(1): 26-37.

[46] FARSIANI S, IZADKHAH H, LOTFI S. An optimum end-to-end text-independent speaker identifycation system using convolutional neural network[J]. Computers and Electrical Engineering, 2022, 100: 107882.

[47] XIA W, LU H, WANG Q, et al. Turn-to-diarize: Online speaker diarization constrained by transformer transducer speaker turn detection[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver: IEEE, 2022: 8077-8081.

[48] TANG J, SU Q, NIU L, et al. Emotion analysis of Chinese reviews based on fusion of multi-layer CNN and LSTM[C]//2022 the 5th International Conference on Image and Graphics Processing (ICIGP). New York: Association for Computing Machinery, 2022: 351-356.

[49] CHEN Z, CHEN S, WU Y, et al. Large-scale self-supervised speech representation learning for automatic speaker verification[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Vancouver: IEEE, 2022: 6147-6151.

[50] WANG Y, SKERRY-RYAN R J, STANTON D, et al. Tacotron: Towards end-to-end speech synthesis [C]//Interspeech. Stockholm: ISCA, 2017: 4006-4010.

[51] KALCHBRENNER N, ELSEN E, SIMONYAN K, et al. Efficient neural audio synthesis[C] //International Conference on Machine Learning. New York: PMLR, 2018: 2410-2419.

[52] PANAYOTOV V, CHEN G, POVEY D, et al. Librispeech: an asr corpus based on public domain audio books[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York: IEEE, 2015: 5206-5210.

[53] LI A, LIU W, ZHENG C, et al. Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1829-1843.

[54] 盛骤. 概率论与数理统计:第三版[M]. 北京: 高等教育出版社, 2001.

[55] LIU T, DAS R K, LEE K A, et al. MFA: TDNN with multi-scale frequency-channel attention for text-independent speaker verification with short utterances[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver: IEEE, 2022: 7517-7521.

[56] 罗海涛. 语音信号频谱的获取[J]. 电脑与电信, 2022, (04): 555-557.

[57] ZHOU K, SISMAN B, LIU R, et al. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto: IEEE, 2021: 920-924.

[58] SISMAN B, YAMAGISHI J, KING S, et al. An overview of voice conversion and its challenges: From statistical modeling to deep learning[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 29: 132-157.

[59] ZHOU K, SISMAN B, LI H Z, et al. Limited data emotional voice conversion leveraging text-to-speech: Two-stage sequence-to-sequence training[C]// Interspeech. Brno: ISCA, 2021: 846-851.

[60] Chandolikar N, Joshi C, Roy P, et al. Voice recognition: A comprehensive survey[C]//2022 International Mobile and Embedded Technology Conference (MECON). New York: IEEE, 2022: 45-51.

[61] 王怡, 王黎明, 柴玉梅. 融合多特征的语音情感识别方法[J]. 小型微型计算机系统, 2022, 43(06): 1232-1239.

[62]金浩, 朱文博, 段志奎, 等. 基于注意力机制的 TDNN-LSTM 模型及应用[J]. 声学技术, 2021, 40 (04): 508-514.

[63] PATEL S, CANOZA P, SALAHUDDIN S. Logically synthesized and hardware-accelerated restricted Boltzmann machines for combinatorial optimization and integer factorization[J]. Nature Electronics, 2022, 5(2): 92-101.

[64] DE OLIVEIRA A C N, FIGUEIREDO D R. Network connectivity and learning performance on restricted Boltzmann machines[C]//2022 International Joint Conference on Neural Networks (IJCNN). Padua: IEEE, 2022: 1-8.

[65] DABELOW L, UEDA M. Three learning stages and accuracy–efficiency tradeoff of restricted Boltzmann machines[J]. Nature Communications, 2022, 13(1): 5474.

[66] GU L, YANG L, ZHOU F. Approximation properties of Gaussian-binary restricted Boltzmann machines and Gaussian-binary deep belief networks[J]. Neural Networks, 2022, 153: 49-63.

[67] KUMAR A, AGGARWAL R K. Hindi speech recognition using time delay neural network acoustic modeling with i-vector adaptation[J]. International Journal of Speech Technology, 2022, 25(1): 67-78.

[68] 蔡国都. 基于 x-vector 的说话人识别研究[D]. 北京: 北京交通大学, 2019.

[69] PEDDINTI V, POVEY D, KHUDANPUR S. A time delay neural network architecture for efficient modeling of long temporal contexts[C]//Sixteenth Annual Conference of the International Speech Communication Association. Dresden: ISCA, 2015.

[70] SNYDER D, GARCIA-ROMERO D, POVEY D, et al. Deep neural network embeddings for text-independent speaker verification[C]// Interspeech. Stockholm: ISCA, 2017: 999-1003.

[71] BUSSO C, BULUT M, LEE C C, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. Language Resources and Evaluation, 2008, 42: 335-359.

[72] EKMAN P, FRIESEN W V, O'SULLIVAN M, et al. Universals and cultural differences in the judgments of facial expressions of emotion[J]. Journal of Personality and Social Psychology, 1987, 53(4): 712.