基于神经网络的语音情感特征表示学习

（申请清华大学工程硕士专业学位论文）

培 养 单 位：计算机科学与技术系

工 程 领 域：计算机技术

申 请 人：代 东 洋

指 导 教 师：贾 珈 副教授

联合指导教师：吴 志 勇 副研究员

二〇二〇年五月

基于神经网络的语音情感特征表示学习

代东洋

Speech Emotion Representation Learning Based on Neural Network

Thesis Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the professional degree of

Master of Engineering

by

Dai Dongyang

(Computer Technology)

Thesis Supervisor : Associate Professor Jia Jia

Associate Supervisor : Associate Professor Wu Zhiyong

May, 2020

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容。

本人保证遵守上述规定。

（保密的论文在解密后应遵守此规定）

作者签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘 要

情感是交流中的一大要素，它可以帮助人们更好地理解意图、表达需求。随着语音技术的发展，语音交互正在成为一种主流人机交互方式。从语音中学习有效的情感特征表示，并将其用于语音情感识别和情感语音合成中，可以帮助机器更好地理解说话人的意图，提高机器生成语音的自然度，有利于建立更智能的人机语音交互系统。本课题主要研究如何利用神经网络从语音中提取有效的情感特征表示，并将其用于提高语音情感识别任务的性能，实现情感可控的语音合成。主要工作和贡献如下：

一、针对神经网络从语音中所提取的特征不具有足够的情感区分度的问题，提出了一个基于中心损失函数的端到端语音情感识别模型，提升了语音情感识别的性能。由于情感的主观性，神经网络很难从语音中提取具有情感区分度的特征。本文的模型同时使用交叉熵和中心损失函数，交叉熵保证模型所提取的特征可分离，中心损失函数保证所提取特征具有更小的类内距离。在中心损失函数和交叉熵的联合作用下，模型提取出更具有情感区分度的特征，语音情感识别的性能得到提高。

二、针对测试数据和训练数据分布存在差异时，语音情感识别模型性能不稳定的问题，提出了一个基于最大平均差异损失函数的端到端跨域语音情感识别模型。语音情感识别在实际应用场景下，常常会遇到测试数据和训练数据分布存在差异的情况。本文在端到端语音情感识别模型中引入最大平均差异损失函数，保证模型所提取的情感特征能够减小这种分布差异，最终保证模型在测试数据上有稳定的情感识别效果。

三、提出了一种基于全局风格令牌的情感合成方案，解决了在开源无情感标签信息的合成语料库上进行情感可控语音合成的问题。该方案使用全局风格令牌提取局部文本信息无关的情感特征作为合成网络的条件输入，使用跨域语音情感识别模型预测合成数据的情感软标签，并根据情感软标签建立情感标签到风格令牌权重的映射，实现了情感可控的语音合成。另外，为了实现根据效价度和激活度相互独立地控制生成语音的情感，在模型中引入情感预测损失函数，在风格令牌权重的不同部分引入不同的监督信息。最终，本文的方案实现了根据情感类别标签以及根据情感空间表示两种方法控制生成语音的情感。

关键词：语音情感识别；情感语音合成；中心损失函数；最大平均差异；风格令牌

Abstract

Emotion is an important element in communication, it helps people understand and express better. With the development of voice technology, voice interaction is becoming a mainstream human-computer interaction method. Learning effective emotion feature representations from speech and using them in speech emotion recognition and emotional speech synthesis can help machines better understand the user's intentions, improve the naturalness of machine-generated speech, and help build more intelligent Human-machine voice interaction system. This topic mainly studies how to use neural networks to extract effective emotion feature representations from speech and use them to improve the performance of speech emotion recognition tasks and achieve emotion-controllable speech synthesis. The key work and contributions are as follows:

1. An end-to-end speech emotion recognition model based on the center loss function is proposed, which aims to solve the problem that the features extracted by the neural network from speech are not discriminative enough, and the performance of speech emotion recognition is improved. Due to the subjectivity of emotions, it is hard for neural networks to extract features with emotional discrimination from speech. Our model uses both cross-entropy and center loss. Cross-entropy ensures that the features extracted by the model can be separated, and the center loss function guarantees that the extracted features have a smaller intra-class distance. Under the joint effect of the center loss and cross-entropy, the model extracts features with more emotional discrimination, and the performance of speech emotion recognition is improved.

2. An end-to-end cross-domain speech emotion recognition model based on the maximum average difference loss function is proposed to solve the problem of unstable model performance when the distribution of test data and training data is different. In actual application scenarios, speech emotion recognition often encounters a difference in the distribution of test data and training data. We introduce the maximum mean discrepancy loss function in the end-to-end speech emotion recognition model to ensure that the emotion features extracted by the model can reduce the distribution difference, and finally ensure that the model has a stable emotion recognition effect on the test data.

3. An emotional speech synthesis solution based on Global Style Tokens(GST)

model is proposed for emotion-controllable speech synthesis on an open-source synthesis corpus without emotion label information. In this solution, we use a series of global style tokens to extract emotion features not related to local text information as the conditional input of the basic speech synthesis network, uses a cross-domain speech emotion recognition model to predict emotional soft labels of synthesized data, and establishing a mapping of emotion labels and style token weights to achieve emotion controllable speech synthesis. Besides, in order to realize that the emotions of the generated speech are affected independently of each other according to valence and arousal, an emotion prediction loss function is introduced into the model to give different supervision information to different parts of the weight of the style token. Finally, our solution implements two methods to control the emotion of generating speech based on the emotion category label and the emotion space representation.

Key Words: Speech Emotion Recognition; Emotional Speech Synthesis; Center Loss; Maximum Mean Discrepancy; Global Style Tokens

目 录

主要符号对照表

SER	语音情感识别（Speech Emotion Recognition）
LPC	线性预测编码（Linear predictive coding）
MFCC	梅尔倒谱系数（Mel Frequency Cepstral Coefficents）
SVM	支持向量机（Support Vector Machine）
KNN	K-近邻模型（K-Nearest Neighbors）
GMM	高斯混合模型（Gaussian Mixture Model）
DNN	深度神经网络（Deep Neural Network）
ELM	极限学习机（Extreme Learning Machine）
LSTM	长短期记忆网络（Long Short-Term Memory）
BLSTM	双向长短期记忆网络（bi-directional Long Short-Term Memory）
CNN	卷积神经网络（Convolutional Neural Network）
RNN	循环神经网络（Recurrent Neural Network）
HMM	隐马尔可夫模型（Hidden Markov Model）
TTS	文语转换（Text To Speech）
GRU	门循环单元（Gated Recurrent Unit）
CBHG	多时间步长卷积网络组、高速网络和门循环单元（1-D convolution bank + highway network + bidirectional GRU）
GST	全局风格令牌（Global Style Tokens）
MMD	最大平均差异（Maximum Mean Discrepancy）
STFT	短时傅里叶变换（Short-time Fourier transform）
ReLU	修正线性单元（Rectified Linear Units）
UA	非加权正确率（Unweighted Accuracy）
WA	加权正确率（Weighted Accuracy）
PCA	主成分分析（Principal Component Analysis）
RKHS	再生核希尔伯特空间（Reproducing Kernel Hilbert Space）
t-SNE	t-分布邻域嵌入算法（t-Distributed Stochastic Neighbor Embedding）
MAE	平均绝对误差（Mean Absolute Error）

第 1 章　引言

1.1　研究背景与意义

在现代社会，计算机技术已经应用到了社会生产和生活的各个方面，它与人们日常的工作、学习和生活息息相关。如何让计算机更自然地与人进行交互，一直是计算机科学的一个重要研究课题。

近几十年来，随着计算机相关技术的发展，人与计算机的交互方式也在不断地更新演化。从最初的命令行界面到图形用户界面，计算机的使用门槛大大降低，从而让计算机在大众中得以普及。智能手机时代，更自然的触屏交互方式，成为了主要的人机交互方式。随着语音技术的发展，以及智能硬件的普及，语音交互方式正在获得人们越来越多的关注，逐渐成为一种主流的人机交互方式。

情感是人等高等动物特有的一种能力，由于各种情感的存在，我们的社会、生活变得丰富多彩。情感在人类信息沟通中意义重大，在人与人之间的交流中，为了理解对方，不仅要分辨出对方语音中的文字内容，还要识别对方的情感，而能相互理解情感的人往往可以成为知心朋友。

在语音交互中，情感要素也发挥着重要的作用。一方面，同样的内容，以不同情感的声音说出往往表达不同的意图。另一方面，对于某些特定内容的语音，需要配以合适的情感才会听起来更自然。如何从语音中提取有效的情感信息，并加以利用，已经成为近年来语音领域的热门问题。目前语音中情感信息提取的相关技术，已经在抑郁症检测、电子书朗读、语音助手、虚拟人等领域得到了应用。充分考虑情感要素的语音交互方式，正在一点点融入我们的生活，改变我们的社会。

本课题主要研究如何从语音中提取有效的情感特征表示，并将其用于识别说话人的情感以及合成具有情感的个性化声音，为建立更智能更自然的人机语音交互系统提供基础。提取语音中的情感特征主要用于语音情感识别和情感语音合成两个方面，其中语音情感识别可以帮助机器更好地理解说话人的状态和意图，而情感语音合成可以使机器产生更自然的声音，让人机语音交互更智能。

1.2　研究现状

1.2.1　情感的表示

情感是心理状态的描述，它在潜意识中产生，是对某些特定外部或者内部事件的自主生理反应[1]。情感是人类精神世界的重要组成部分，是人们生活中不可

或缺的因素。因此关于情感的研究，一直是心理学、神经科学、医学、历史学、社会学和计算机科学等学科的重要部分。

很多心理学家认为，情感是由高兴、悲伤、生气、惊喜等一些离散的基本情感（图1.1）组成[2]。复杂的情感由简单的情感组合而成。因此，在情感识别任务中，可以直接根据基本情感来判断情感的状态，这种情形下，情感识别任务就是一个分类任务。

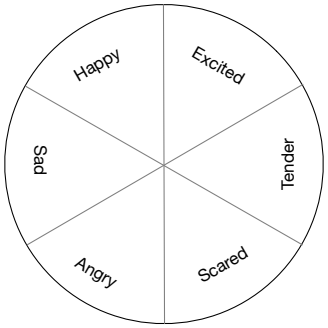


图 1.1　基本情感

心理学家们也尝试使用因子分析的方法把与情感相关的反应映射到数量有限的维度上，以捕捉到不同情绪之间的异同[3]。因子分析所揭示的前两个维度是效价度（Valence, 体验的消极或积极程度）和激励度（Arousal, 体验的激励或活力程度）。可以在 2-D 坐标图上描绘这两个维度（如图1.2），即情感的二维坐标表示。

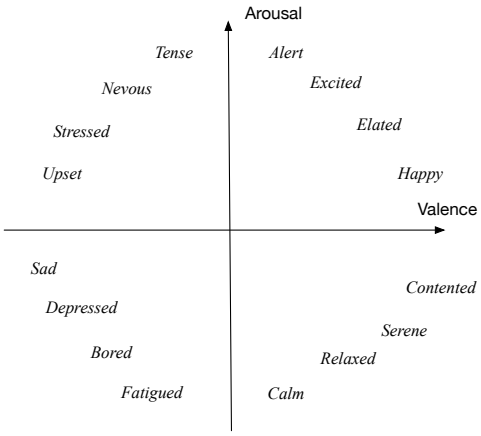


图 1.2　情感的二维表示

本课题主要研究从语音中提取有效的特征表示，以建立语音和情感之间的映射关系。其中语音情感识别为从语音到情感的映射，情感语音合成为情感到语音的映射。对于语音情感识别，如果预测基本情感或者效价度、激励度的正负值，则语音情感识别即为一个分类任务；如果预测效价度、激励度的具体值，语音情感识别则变成了一个回归任务，本课题中主要将语音情感识别当作分类问题来处理。

对于情感语音合成，本课题分别尝试了通过指定基本情感类别和情感的二维坐标表示来控制生成语音的情感。

1.2.2　语音情感识别研究现状

语音情感识别（SER，Speech Emotion Recognition）的流程如图1.3所示，首先从语音中提取出情感相关的特征，然后再使用分类器预测对应的情感类别。早期语音情感识别工作，会结合语音学背景知识，手工提取一些特征用于情感识别。其中包含情感信息的特征主要包括以下三类[4]：

a. 韵律学特征，包括时长（Duration）、基频（Pitch）、能量（Energy）等；
b. 基于谱的相关特征，包括线性预测编码（LPC, Linear predictive coding），梅尔倒谱系数（MFCC，Mel Frequency Cepstral Coefficents）等；
c. 声音质量相关特征，包括共振峰频率及其带宽、声门参数等。

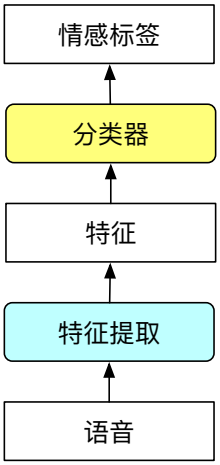


图 1.3　语音情感识别的流程

然后根据以上手工提取的特征，使用支持向量机（SVM，Support Vector Machine)[5]、K-近邻模型（KNN, K-Nearest Neighbors)[6]、贝叶斯模型[7]、高斯混合模型（GMM, Gaussian Mixture Model)[8] 或者决策树模型[9]，预测语音所对应的情感。

由于神经网络[10] 能够从原始数据中自动提取高层次的特征表示。在神经网络逐渐普及以后，越来越多的研究人员开始尝试利用神经网络从原始数据中提取情感相关的特征表示。其中，Han 等人使用深度神经网络（DNN, Deep Neural Network）和极限学习机（ELM, Extreme Learning Machine[11]）从低层次手工特征中直接提取高层次的情感特征[12]。而 Lee 等人使用双向长短期记忆网络（BLSTM，bi-directional Long Short-Term Memory[13]）提取情感特征的高层次表示[14]。Trigeorgis 等人使用神经网络直接从原始语音波形中提取情感特征[15]。Satt 等人使用卷积神经网络

（CNN，Convolutional Neural Network）和循环神经网络（RNN，Recurrent Neural Network）直接从语谱图片段中提取情感特征表示，用于语音情感识别任务[16]。

然而，情感本身是主观的、容易混淆的，不同的情感类别往往是难以区分的[17]。因此如何从语音中提取具有情感区分度的有效特征是一个有挑战性的工作。一种可行的策略是引入度量学习（Metric Learning）[18] 的思想，设计合适的损失函数，指示神经网络提取具有更小的类内距离以及更大的类间距离的特征。为提取更有效的情感特征，Lian 等人提出了一种基于余弦相似度损失（cosine similarity loss）的模型[19]。Huang 等人所提出的神经网络模型使用三元组损失函数（Triplet Loss）[20] 从语音中提取具有情感区分度的特征[21]。

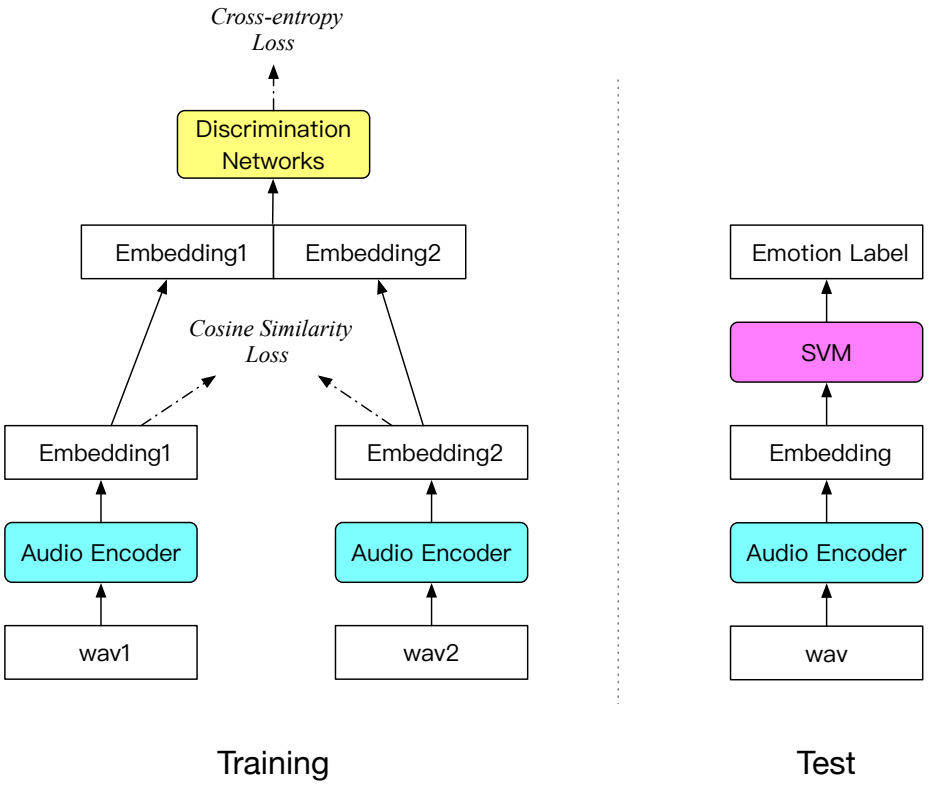


图 1.4　使用 Cosine Similarity Loss 提取有效的特征

Lian 等人的工作[19] 如图1.4所示，在孪生神经网络（Siamese network）[22] 中引入余弦相似性损失（Cosine Similarity Loss），以提取更有效的情感相关的特征。在训练时，模型接受一个音频对（wav1 和 wav2）作为输入，wav1 和 wav2 分别通过一个共用参数的 Audio Encoder，得到 embedding1 和 embedding2。如果 wav1 和 wav2 属于同一种情感，Cosine Similarity Loss 则令 embedding1 和 embedding2 之间的差异尽可能地小，如果 wav1 和 wav2 属于不同的情感，Cosine Similarity Loss 则令 embedding1 和 embedding2 之间的差异尽可能大。通过 Audio Encoder 提取的特

征，再使用 SVM 进行分类，得到最终的情感标签。

Huang 等人的工作[21] 在语音情感识别任务中引入三元组损失函数（triplet loss），使属于同一情感的特征之间距离更小，属于不同情感的特征间距离更大（如图 1.5）。模型接受一个音频三元组（wav1、wav2 和 wav3）作为输入，其中 wav1 和 wav2 属于同一种情感，wav1 和 wav3 属于不同的情感。对 wav1、wav2、wav3 所提取的特征分别为 embedding1、embedding2、embedding3。Triplet Loss 的作用便是让 embedding1 和 embedding2 之间的距离尽量小，让 embedding1 和 embedding3 之间的距离尽量大。最后，使用 SVM 对 embedding 进行分类，得到其情感的标签。

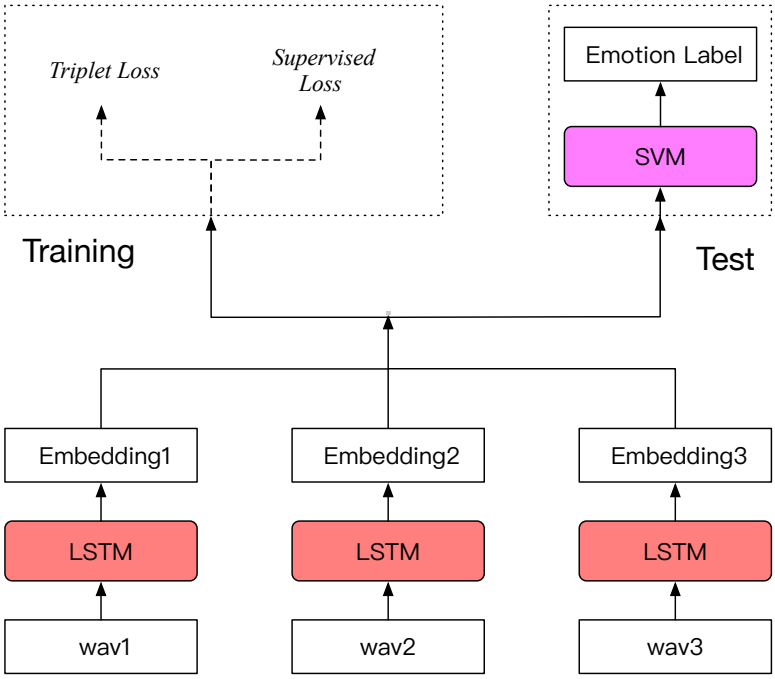


图 1.5　使用 Triplet Loss 提取特征

以上工作使用 Consine Similarity Loss 或 Triplet Loss 来提取更具有情感区分度的特征，以提升语音情感识别任务的性能。然而，它们采用”两步”策略进行情感识别，第一步首先提取情感相关的特征，第二步再使用支持向量机（SVM）对其进行分类。因为这两个步骤的目标可能不完全一致，这样的“两步”策略可能会造成误差的积累，从而降低语音情感识别的性能。另一方面，Consile Similarity Loss 的工作需要构造音频对作为输入，Triplet Loss 的工作中需要构造音频三元组作为输入，这些方法的性能很大程度上依赖于音频对或音频三元组的构造策略。为了取得比较好的语音情感识别效果，需要在构造音频对或音频三元组上花费较大的精力。

此外，在语音情感识别的任务中，由于数据收集、文化习惯等的不同，模型的训练数据集和测试数据集之间可能有较大的差异。一种情形就是跨语言语音情感

识别任务。某些语言的情感数据资源非常丰富（有大量的带情感标签的语音）。而某些语言的情感数据资源相对比较匮乏（语料库中有很少情感标签或没有情感标签）。我们称情感数据资源丰富的语言为源语言，情感数据资源匮乏的语言为目标语言。如何根据源语言丰富的情感数据资源训练一个有效的语音情感模型，保证在目标语言的数据上也能有很好的性能，一直是一个值得研究的问题。Neumann 等人尝试使用卷积神经网络和注意力机制提取情感相关特征，在英语和法语之间实现跨语言语音情感识别[23]。但该工作在训练模型时，没有充分利用目标语言的数据，因此模型的性能还有一定的提升空间。

1.2.3　情感语音合成研究现状

语音合成又叫文语转换（TTS，Text to Speech），指使用计算机将文本转换为语音的技术。最早的语音合成技术包括拼接式的语音合成[24] 和基于隐马尔可夫模型（HMM，Hidden Markov Model）的参数化语音合成[25]。其中拼接式合成方法生成的语音自然度、清晰度更好，而基于 HMM 的参数合成方法虽然语音质量相对较差，但其语速、音调等可以调节。由于神经网络在大量的数据下能够产生比传统方法更好的效果，引入了神经网络模型后，参数合成方法在语音自然度、清晰度等方面逐渐达到甚至超过了拼接式语音合成方法。基于神经网络的参数合成方法逐渐成为了主流。目前主流的网络结构主要分为两种: 级联模型结构[26-30] 和端到端结构[31-35]。

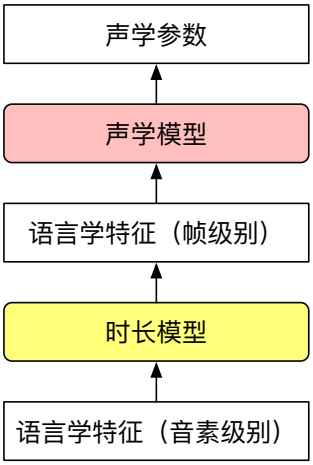


图 1.6　级联模型结构

级联模型结构（如图1.6），主要由两部分组成: 时长模型（Duration Model）和声学模型（Acoustic Parameter Model）。其中，时长模型接受音素级别的语言学特征（Linguistic features）作为输入，得到每个音素的时长，进而得到帧级别的语言学特征。然后帧级别的语言学特征再通过声学模型得到声学参数（Acoustic features）。

声学参数可以直接使用声码器（Vocoder）生成语音。常见的声码器有 WORLD[36]、STRAIGHT[37]、WaveNet[38]、LPCNet[39] 等。

在级联结构模型的基础上，百度提出 EMPHASIS 模型[40] 生成带情感的语音（如图1.7）。相对于普通的级联结构模型，EMPHASIS 模型增加了包含情感信息的韵律特征（Emotional & prosodic features）作为模型额外的条件输入。在 EMPHASIS 模型中，时长模型和声学模型的实现包括多时间步长卷积网络组、高速网络和门循环单元（CBHG，1-D convolution bank + highway network + bidirectional GRU）。另外，由于相对于音素相关的特征，韵律特征是弱势特征。为了防止模型对音素级别的特征过拟合，网络先用不同的卷积分别处理音素相关的特征和韵律特征。

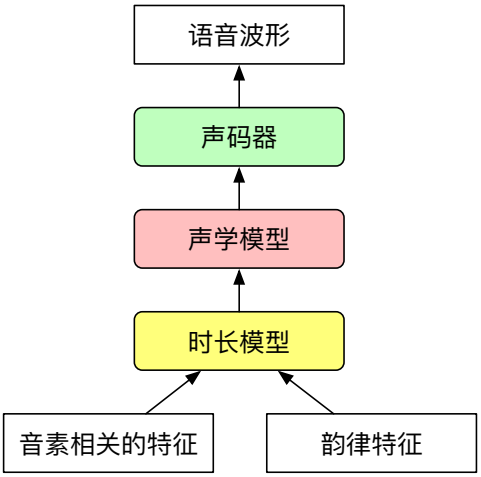


图 1.7　EMPHASIS 模型结构

由于级联模型需要有语音的时长信息标注，而标注每个音素的时长信息需要比较高成本的人力物力。因此，很难使用很大的语料库训练级联结构模型。而端到端的模型所需要的标注信息比级联模型简单很多，使用端到端模型在数据量很大的语料库上进行训练成为可能。

端到端模型的输入是文本的序列（音素的序列），输出是可以直接生成声音的声学参数序列。它使用基于注意力机制的序列到序列（seq2seq）模型实现（如图1.8）。它包含编码器（Encoder）和解码器（Decoder）。其中 Encoder 接受文本的序列作为输入，并对其上下文以及时序关系进行建模，得到一个包含上下文信息的中间特征序列。Decoder 是一个自回归结构，它使用自身上一个时间步的输出作为下一个时间步的输入，此外，它还通过注意力机制，对 Encoder 输出的特征序列进行加权平均，作为 Decoder 每个时间步额外的条件输入。常见的端到端语音合成模型有 Char2Wav[31]、Tacotron[32]、DeepVoice3[33]、Tacotron2[34]、Transformer TTS[35] 等。

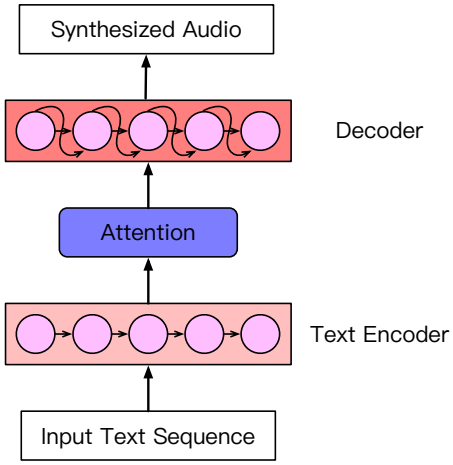


图 1.8　端到端模型结构

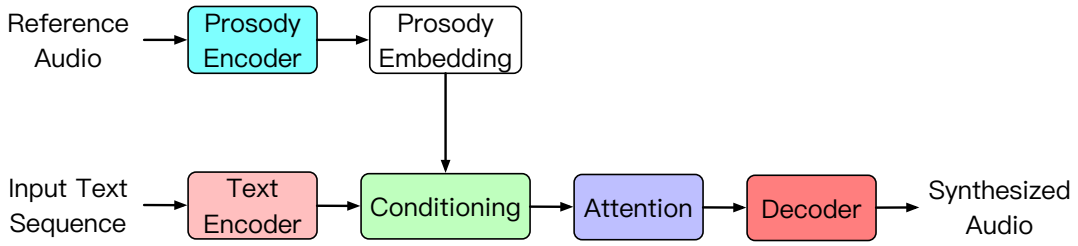


图 1.9　基于 Tacotron 的韵律迁移

谷歌在 Tacotron 模型的基础上，提出了韵律迁移[41] 的方法（如图1.9）。该模型接受一个韵律特征（Prosody Embedding）作为注意力（Attention ）层前额外的条件输入。而韵律嵌入向量（Prosody Embedding ）通过韵律编码器（Prosody Encoder ）从参考语音（Reference Audio ）中提取。通过该模型，可以生成与参考语音具有相同的韵律特点语音（Synthesized Audio）。

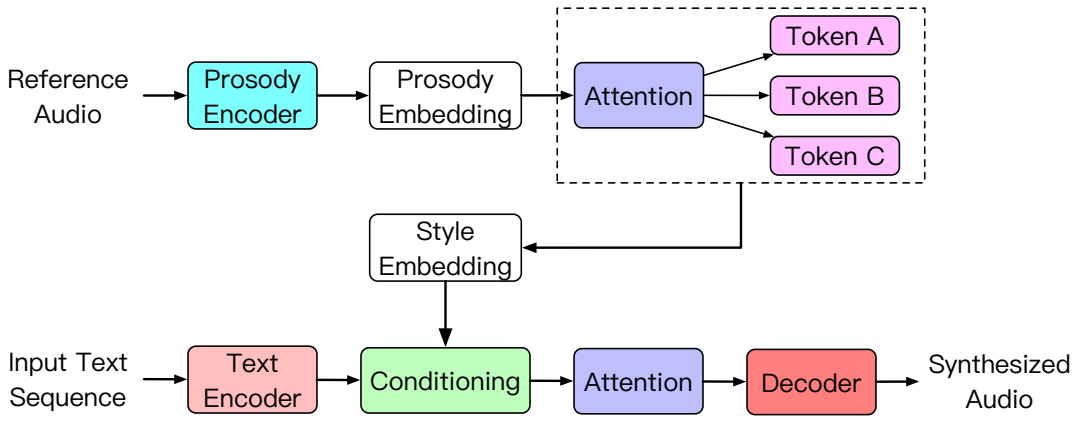


图 1.10　Global Style Tokens 控制生成语音的风格

然而，韵律迁移[41] 的方法，即使在生成语音时，也需要一个额外的参考语音

（Reference Audio ）作为输入。为了在预测语音时不需要再使用额外的参考语音，也为了更灵活地控制生成语音的风格。谷歌的研究人员进一步提出了全局风格令牌（GST，Global Style Tokens）的概念和相关模型[42]（如图1.10）。该方法并没有直接把韵律嵌入向量作为合成网络（Tacotron）的条件输入，而是比较韵律嵌入向量和一系列风格令牌（Style Token ）的相似度，并得到相应的权重。然后这些令牌的加权平均——风格嵌入向量（Style Embedding ）作为合成网络的条件输入。在训练的时候模型接受参考语音作为输入，在预测时，可以直接显式地指定各令牌的权重。与韵律迁移[41] 的方法相比，该方法可以更加灵活地控制生成语音的风格。然而，每个令牌代表什么风格并不明确。

如上所述的 EMPHASIS 模型[40]、韵律迁移[41] 方法和 GST[42]，都是在一个语音合成基础网络的基础上，添加一个包含表现力信息（情感、说话风格等）的特征作为条件输入，以实现带情感语音的合成（如图1.11）。这些方法能够合成带有情感的具有表现力的声音，但还不能显式地控制生成语音的情感。

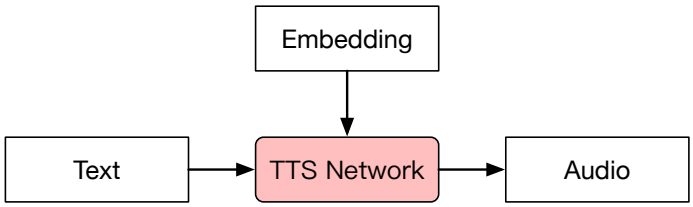


图 1.11　带情感语音合成方法

为实现情感可控的语音合成，Inoue 等人直接将情感类别的独热（one-hot）编码作为级联模型的条件输入[43]，Lee 等人将情感类别的 one-hot 编码作为端到端模型的条件输入[44]，这些方法实现较为简单，但是却需要合成语料库中每句话都有对应的情感标签，因此使用这些方法进行情感语音合成时，需要专门录制相应的语料库，成本较高，而且这些语料库往往只能实现单个说话人的情感语音合成，很难实现情感控制在说话人之间的迁移。

Hodari 等人在语音情感语料库上训练了一个情感识别模型，然后识别合成语料库中每一条语音的情感作为合成模型的条件输入以实现情感语音合成[45]。首先在情感识别语料库上训练一个情感识别模型，然后使用该模型预测合成语料的情感标签，这其实也是一个跨域语音情感识别的问题。Hodari 等人的方法[45] 没有考虑合成语料库和情感识别语料库之间的差异性，在训练跨域情感识别模型时也没有充分使用合成语料库的数据，在情感识别语料库上训练得到的模型，并不能在合成语料库上取得很稳定的效果，这也影响了最终模型生成的语音所表达的情感的准确性。

1.3　本文主要研究内容和贡献

1.3.1　研究内容和各章简介

情感是交流中的重要因素，从语音中提取有效的情感特征表示有利于建立更自然更智能的人机语音交互系统。本文主要研究如何从语音中提取有效的情感特征表示，并将包含情感的特征表示用于语音情感识别和情感语音合成。本文的工作主要包含以下三个方面：（1）在语音情感识别任务中使用神经网络提取更具有情感区分度的情感特征表示。（2）在应用语音情感识别模型时，考虑测试数据和训练数据之间的差异，解决跨域语音情感识别的问题。（3）从语音中提取有效的情感特征表示，并将其用于情感可控的语音合成。

第一章为引言，介绍了语音情感识别和情感语音合成工作的背景、意义、发展历史和研究现状。

第二章介绍了基于中心损失函数（Center Loss）[46] 的端到端的语音情感识别模型。该模型直接接受变长的语谱图作为输入，节省了一系列繁琐的特征工程工作。模型采用卷积神经网络从语谱图中提取局部特征，并在时间维度上进行降维；采用循环神经网络把卷积神经网络的变长输出序列变为定长的输出；采用由全连接网络组成的分类器输出每个情感的后验概率。该模型可以直接根据输入的变长语谱图预测情感，同时模型中引入了中心损失函数，以保证神经网络提取的特征更具有情感区分度。在 IEMOCAP 数据集[47] 上的实验证明，引入了中心损失函数后，语音情感识别任务的性能得到明显的提高。

第三章介绍了基于最大平均差异（MMD，Maximum Mean Discrepancy）[48] 损失函数的端到端跨域情感识别模型。该模型同样接受变长的语谱图作为输入，直接预测对应的情感。为了解决目标域数据（测试数据）和源域数据（训练数据）之间分布存在差异的问题，训练模型时同样使用了没有情感标签的目标域数据以保证在目标域情感识别模型也有很好的性能。模型使用交叉熵训练情感识别模型的同时，引入了一个辅助训练目标：最小化源域特征和目标域特征之间的最大平均差异，以保证模型能够提取域无关的情感特征。跨域语音情感识别可以应用于跨语言情感识别以及识别合成数据情感任务中。我们在 IEMOCAP 数据集和 RECOLA 数据集[49] 上进行跨语言情感识别的实验，结果表明，相对于基准模型，基于最大平均差异损失函数的模型在目标语言上能够更好地预测语音的情感。

第四章介绍了基于全局风格令牌（GST）的情感语音合成方案。该方案可以直接使用没有情感标签的开源合成数据进行情感可控的语音合成。该方案首先使用跨域语音情感识别模型识别合成数据的情感标签。然后使用全局风格令牌提取局部信息无关的情感特征，作为端到端模型的条件输入，以影响所合成语音的情感。

通过建立情感标签到情感特征的映射，实现了情感可控的语音合成。同时，该方案提出了根据情感类别控制合成语音情感以及根据情感的二维坐标表示控制合成语音情感的两种模型，在 Blizzard Challenge 2013 数据[50] 上均取得了很好的合成效果。

第五章对从语音中提取有效的情感特征表示，并将其用在语音情感识别以及情感语音合成中的相关研究成果进行了总结。同时，关于如何从无监督语音数据中提取情感特征以及如何把情感要素结合语音中的其他副语言要素相结合，对未来的研究方向进行了展望。

1.3.2　本文主要贡献

本文主要研究如何利用神经网络从语音中提取有效的情感特征表示，并将所提取的更有效的情感特征用于提高语音情感识别正确率、提高跨域语音情感识别的正确率以及情感可控的语音合成。本文研究内容框架如图1.12所示，主要有以下三个贡献点：

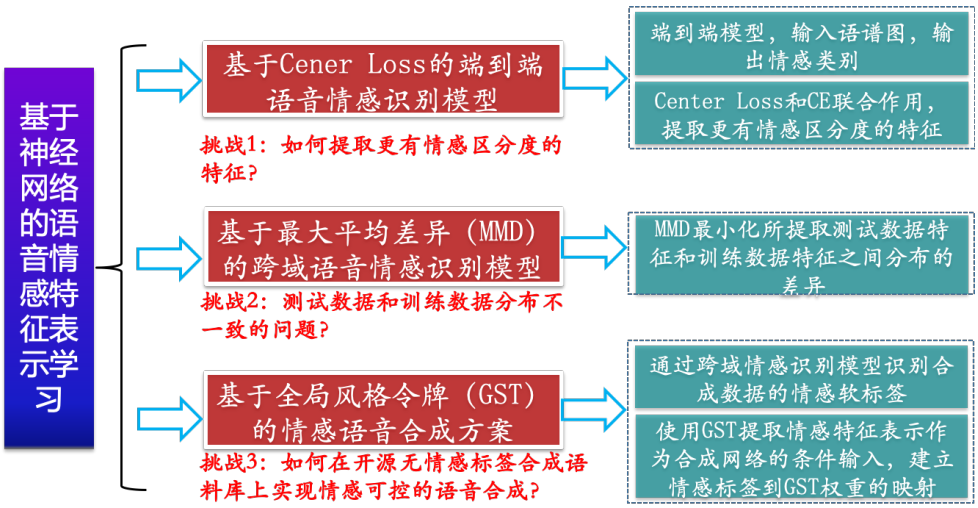


图 1.12　本文研究内容框架

一、提出了基于中心损失函数（Center Loss）的端到端语音情感识别模型。该模型直接根据变长的语谱图预测情感类别，并且在 Center Loss 的作用下，所提取的特征具有更好的情感区分度，语音情感识别任务也有更高的正确率。由于神经网络能够在大量数据下自动进行特征学习，而且往往比利用专家知识进行特征工程所提取的特征具有更好的效果，利用神经网络建立端到端的模型逐渐成为一种主流，该模型直接根据变长的语谱图预测情感，省去了一系列繁琐的特征工程工作。现有一些工作已经尝试使用 Triplet Loss 等提取更具有情感区分度的特征，以

提升语音情感识别任务的性能，然而现有的方法都是先利用神经网络提取具有情感区分度的特征，然后再利用支持向量机预测对应的情感类别，这种“两步”的策略会造成误差的积累。本模型直接在端到端的语音情感模型中引入中心损失函数，保证所提取的特征具有情感区分度的同时，直接根据输入的变长语谱图预测情感。而且实验效果表明，引入 Center Loss 确实有助于提取更具情感区分度的特征，也能带来语音情感识别正确率的提升。

二、提出了基于最大均值差异（MMD）的端到端跨域语音情感识别模型。该模型充分考虑了源域数据和目标域数据分布之间的差异性，从变长语谱图中直接提取域无关的情感特征，因此即使测试数据和训练数据之间分布不一致，情感识别模型也能在测试数据上取得稳定的性能。在应用语音情感识别模型时，往往会遇到测试数据和训练数据分布不一致的情况。我们称这种情况下的语音情感识别任务为跨域语音情感识别，我们称训练数据为源域数据，测试数据为目标域数据。一种跨域语音情感识别场景就是跨语言语音情感识别：源语言中有大量的带情感标签的语音，目标语言带情感标签的语音资源很匮乏，如何利用源语言的数据训练一个稳定的语音情感识别模型识别目标语言语音的情感。在源语言上训练一个语音情感识别模型直接应用于目标语言上往往不能取得稳定的效果。本模型训练语音情感识别模型时同时利用了无情感标签的目标域数据，在训练神经网络时引入最小化 MMD 的子任务，以从原始数据中提取域无关的情感特征，保证在目标域数据上语音情感识别模型也有稳定的效果。在跨语言语音情感识别任务中，本文的模型相对于基准方法有了明显的提升。

三、提出了一种基于全局风格令牌（GST）的情感语音合成方案。该方案使用跨域语音情感识别模型预测合成数据的情感监督信息，并建立从标签信息到风格令牌权重的映射，最终实现了在开源的无情感标签合成数据，也能根据情感类别或者情感的二维坐标表示显式地控制合成语音的情感。目前主流的带情感语音的合成模型都是接受一个包含情感信息的特征表示作为条件输入，而该特征表示从语音中提取，然而由于合成数据不具有情感标签，因此这些方法只能合成带有情感的具有表现力的语音，却不能显式地控制合成语音的情感。本方案采用跨域的语音情感识别模型预测合成数据的情感标签，充分考虑了合成数据和情感数据分布之间的差异，进一步建立情感类别和令牌权重的关系，实现了根据情感类别标签控制合成语音的情感。此外，该方案中使用情感识别得到的软标签作为监督信息，在基于 GST 的情感语音合成模型中引入情感预测损失函数，进一步实现了根据效价度和激活度相互独立地影响生成语音的情感。

第 2 章　基于中心损失函数的端到端语音情感识别

2.1　本章引论

根据语音学等专业领域知识，语音的情感和能量、基频、韵律学特征以及声音质量的相关特征有关[51]。然而特征工程相对比较繁琐，而且需要一定的领域知识。由于深度神经网络可以从原始数据中学习高层次特征表示[52], 而且深度学习在很多领域都带来了突破[53]。因此也有越来越多的研究人员尝试从原始数据中直接提取情感相关的特征用于语音情感识别，其中 Satt 等人使用神经网络直接从定长的语谱图片段中提取特征识别情感，相对于传统的方法，只需要求对语音波形进行短时傅里叶变换（STFT，Short-time Fourier transform），节省了特征工程的工作[16]。但是语谱图输入模型前，需要先截取成定长的语谱图片段，这可能导致情感相关信息的损失。

由于情感本身是比较主观含糊的，因此从语音中提取有效的情感特征是困难的。为了从语音中提取出更有效的情感特征表示，研究人员尝试设计合适的损失函数以保证所提取的特征更具有情感区分度：类内距离更小，类间距离更大。Lian 等人使用 Consine Similarity Loss 来学习输入的音频对之间的相似性和区别[19]，在 Cosine Similarity Loss 的作用下，如果输入的音频对属于同一种情感，则它们所对应的情感特征尽可能接近，反之它们所对应的情感特征差异尽可能大。Huang 等人从音频三元组中提取情感特征，使用 Triplet Loss 保证属于同一类情感的特征之间的距离尽可能小，属于不同情感的特征之间距离尽可能大[21]。以上语音情感识别方案的性能，很大程度上依赖于音频对或音频三元组的构造方式。而且它们都只使用神经网络提取特征，而不是直接预测情感，从神经网络中提取具有情感区分度的特征后，再通过支持向量机预测对应的情感，这种非端到端的模型也会造成误差的累积。

本章提出了基于中心损失函数（Center Loss）[46] 的端到端语音情感识别模型。在端到端的语音情感识别模型中引入中心损失函数，以提取更有情感区分度的特征表示。该模型由卷积神经网络，循环神经网络和全连接网络组成，其中卷积神经网络用于提取局部特征以及在时间方向上降低维度，循环神经网络把变长的特征序列变为定长向量，全连接网络把特征向量映射到目标特征空间以及输出每个情感的后验概率。该模型接受变长的语谱图作为输入，直接输出各个情感类别的后验概率，同时该模型在中心损失函数的作用下，保证了所提取的情感特征在类间可分离的同时，类内距离更小。

2.2　基于中心损失函数的端到端语音情感识别模型

2.2.1　模型概述

基于中心损失函数的端到端语音情感识别模型框架如图 2.1所示。其中包括了由若干二维卷积层组成的卷积网络模块（CNN Layers），一个双向循环神经网络组成的循环网络模块，以及两个全连接神经网络层（FC1 和 FC2）。模型同时使用了交叉熵损失函数（Softmax Cross-entropy Loss）和中心损失函数（Center Loss）进行多任务训练。

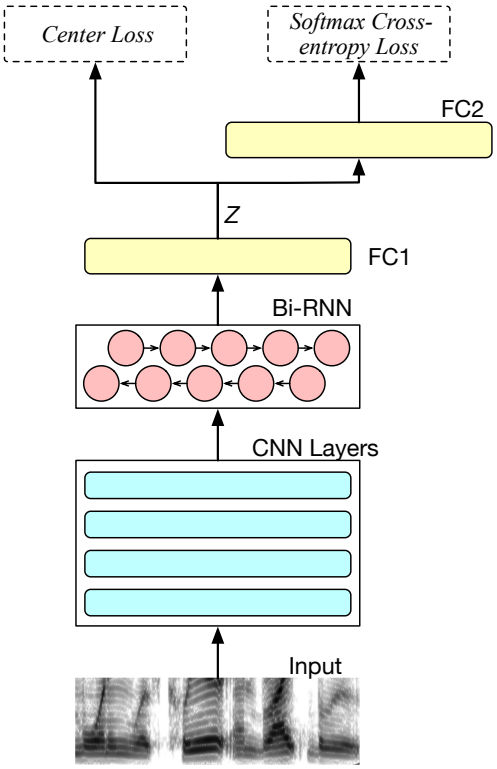


图 2.1　模型基本框架

直接从原始语音波形得到的语谱图即可作为模型的输入，因为原始语音的长度是不确定的，因此得到的语谱图是变长的。卷积网络模块从变长的语谱图中提取局部信息，并且对数据从时间维度上进行降维。由双向循环网络组成的循环网络模块把卷积网络模型输出的变长特征序列转化为定长的高维向量。第一层全连接神经网络（FC1）把循环神经网络模块输出的定长高维向量映射到目标维度空间，第二层全连接神经网络（FC2）以 Softmax 作为激活函数，它的输出代表各个情感类别的后验概率。模型同时受交叉熵损失函数（Softmax Cross-entropy Loss）和中心损失函数（Center Loss）的作用。其中交叉熵损失函数促进网络提取情感可分离的特征，而中心损失函数保证所提取特征的类内间距更小。在交叉熵和中心损失

函数的联合作用下，模型提取更有情感区分度的特征。

2.2.2　模型细节

模型接受变长的语谱图作为输入，它可以是线性谱，也可以是梅尔谱（Mel-spectrogram）。线性谱可以直接从原始语音波形中，通过短时傅里叶变换（STFT）得到。线性谱通过一组梅尔滤波器（Mel-scale Filter Banks），即可得到梅尔谱。假设输入的语谱图大小为 $L_T \times L_F$，其中 L_T 代表时间维度的长度，它依赖于原始音频的时长，L_F 代表频域相关的维度。

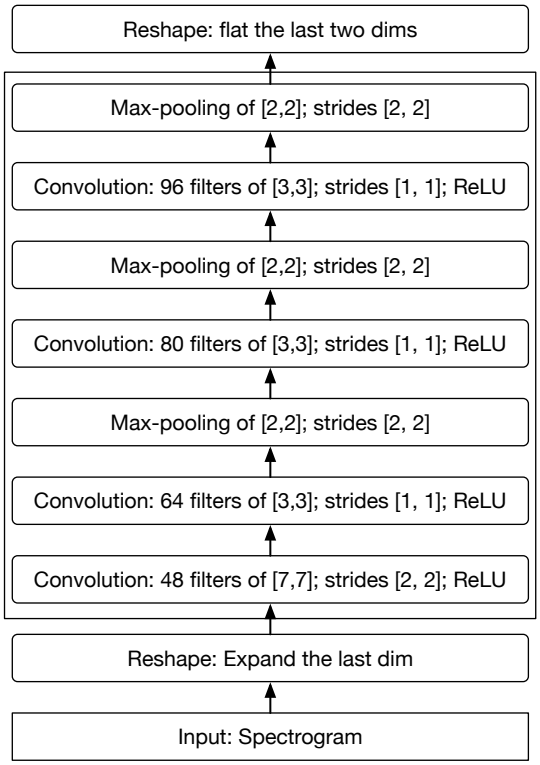


图 2.2　卷积网络模块细节

卷积网络模块主要用于从语谱图中提取局部特征，此外卷积神经网络还对输入数据的时间维度进行降维，以提高模型的运算效率。根据计算机视觉的相关经验，多层卷积网络中，第一个卷积层使用比较大的卷积核，其他卷积层使用比较小的卷积核往往性能比较好[54-55]。同时，通过几十次实验后，最终决定卷积网络模块的细节如图 2.2所示。

卷积网络模块的输入为 $L_T \times L_F$，由于卷积网络模块由二维卷积层组成，因此进行卷积操作前，需要先把输入数据转换为 $L_T \times L_F \times 1$ 的格式，增加一个代表通道的维度。卷积网络模块由四层卷积神经网络组成。其中第一个卷积层卷积核的大小是 7×7，步长大小为 3×3，其后没有池化层。第二、三、四层卷积层的

卷积核的大小都是 3×3，步长大小都为 1×1，每次卷积操作后都会跟一个步长为 2×2 的 2×2 最大池化层。第一、二、三、四层卷积的激活函数都是修正线性单元（ReLU，Rectified Linear Units），卷积核的个数分别为 48、64、80 和 96 个。为了后续的循环神经网络处理，把卷积层的 3-D 格式输出的后两个维度展开，将数据转化为转换为 2-D 格式。

卷积神经网络可以直接处理变长的数据，输出数据的长度随着输入数据的长度而变化。然而，实际训练神经网络时，都是批（batch）处理数据。每次的输入是由一个 batch 的数据组成的张量，数据中时间最长的样本作为张量的时间维度，其他数据会在时间维度上后续补零至张量的时间维度。为了防止这种补零的操作在训练神经网络时对前向传播以及梯度反向传播造成影响，在每次卷积操作后，输出数据都会乘以一个掩码，该掩码根据每个数据的有效输出的时间维度计算得到，其中有效的时间维度部分值为 1，无效的时间维度部分值为 0。每次卷积操作后输出的有效时间维度按照公式（2-1）计算，其中 l_{in} 代表输入的有效时间维度的长度，l_{out} 代表输出的有效时间维度的长度，k 代表卷积核时间维度的大小，s 代表步长时间维度的值。

$$l_{out} = 1 + \left\lfloor \frac{l_{in} - k}{s} \right\rfloor \tag{2-1}$$

循环神经网络模块由一层双向的循环神经网络组成，它接受由卷积神经网络输出的变长特征序列作为输入，通过拼接前向循环网络和后向循环网络最后一个时间步的输出，实现了把变长的特征序列压缩为定长的特征向量（如图 2.3）。由于门循环单元（GRU，Gated Recurrent Unit）[56] 相对于普通的循环网络增加了门函数（Gated Function），能够更好地解决时间序列的长时依赖的问题，同时 GRU 相对于 LSTM 结构更简单、网络参数更少。综上这里采用 128 维的 GRU 来实现循环神经网络模块，拼接前向循环网络和后向循环网络最后一个时间步的输出作为循环网络模块的输出，因此循环网络模块的输出为 256 维。

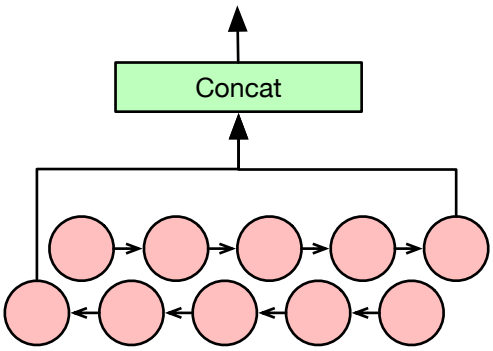


图 2.3　循环网络模块压缩变长特征序列为固定长度向量

第一层全连接网络把循环网络模块输出的定长向量映射到目标维度 d（在我们的实验中 $d=64$）。第一层全连接网络激活后的输出值我们设为 z（$z \in \mathbb{R}^d$），我们把 z 视为模型学到的深度情感特征，并根据 z 来计算中心损失函数。由于 ReLU 激活函数会损失所有负值的信息，这不利于我们利用中心损失函数提取具有情感区分度的特征，因此第一层全连接网络采用 PReLU（Parametric Rectified Linear Unit）[57] 作为激活函数。PReLU 的计算公式如（2-2）所示，其中 a 是一个可训练的参数，它随着神经网络参数的更新而更新。

$$f(y)=\begin{cases} y & y \geqslant 0 \\ ay & y<0 \end{cases} \tag{2-2}$$

第二层全连接网络的输出维度等于情感类别的个数，采用 Softmax 作为激活函数，第二层全连接网络的输出为各个情感的后验概率，我们根据第二层全连接的输出预测语音的情感。

2.2.3　交叉熵损失函数

交叉熵可以衡量模型预测的后验概率和实际分布之间的差异，因此交叉熵损失函数被广泛地应用于分类任务中。在交叉熵损失函数的作用下，语音情感识别网络可以提取情感可分离的特征。在多分类任务下，交叉熵的计算方式如公式（2-3）所示。

$$L_s^0=-\frac{1}{m}\sum_{i=1}^{m} log(\frac{e^{W_{y_i}^{\mathrm{T}} z_i+b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^{\mathrm{T}} z_i+b_j}}) \tag{2-3}$$

其中 m 代表每个 batch 中样本的个数，n 代表情感类别的个数。$z_i \in \mathbb{R}^d$ 代表对应于第 i 个样本的情感特征，它属于第 y_i 个情感类别，y_i 是第 i 个输入样本的情感标签，而且 $y_i \in \{1,2,3,...,n\}$。$W \in \mathbb{R}^{d\times n}$ 是第二层全连接网络的权重参数，$b \in \mathbb{R}^n$ 是第二层全连接网络的偏置参数。$W_j \in \mathbb{R}^d$ 代表 W 中的第 j 列，b_j 代表向量 b 的第 j 个元素。

2.2.4　中心损失函数

为了减少所提取情感特征的类内间距，本章在端到端的语音情感模型中引入了中心损失函数（Center Loss）[46]。中心损失函数最早在人脸识别任务中提出，以保证同一个人的人脸所提取的特征差异尽可能小。模型为每一个情感类别维护一个全局的类中心点，在神经网络训练的过程中，会保证所提取的情感特征到其对

应的类中心点的距离尽可能小。中心损失函数其实就是情感特征到类中心点的平方距离，其公式如（2-4）。

$$L_c^0 = \frac{1}{m}\sum_{i=1}^{m}||z_i - c_{y_i}||^2 \tag{2-4}$$

其中 c_j（$j \in \{1,2,...,n\}$）代表第 j 个情感标签所对应的全局类中心点。通过最小化中心损失函数，属于同一情感类别的情感特征之间的距离会缩小。全局类中心点 c 的初始值设为 0，在训练过程中每训练一步更新一次。全局类中心点的更新依赖于局部类中心点 $\dot{c}$，其中 $\dot{c}_j$ 代表第 j 个情感所对应的局部类中心点，$\dot{c}_j$ 的值根据公式（2-5）计算。$\delta(\cdot)$ 函数的计算方式见公式（2-6）。

$$\dot{c}_j = \frac{\sum_{i=1}^{m}\delta(y_i = j)z_i}{\sum_{i=1}^{m}\delta(y_i = j)} \tag{2-5}$$

$$\delta(condition) = \begin{cases} 1 & condition = True \\ 0 & condition = False \end{cases} \tag{2-6}$$

全局类中心点 c_j 按照公式（2-7）更新。其中 α 是控制 c_j 更新速度的超参数，当新一个 batch 的数据中包含第 j 个情感的样本时，α 的值越大，c_j 的值受新一个 batch 中的数据的影响越大。当新一个 batch 的数据中不包含第 j 个情感的样本时，c_j 维持原来的值不变。c_j^t 和 $\dot{c}_j^t$ 分别代表训练过程中第 t 次迭代时 c_j 和 $\dot{c}_j$ 的值，全局类中心点每训练一步更新一次。

$$c_j^{t+1} = \begin{cases} (1-\alpha)c_j^t + \alpha\dot{c}_j^t & \sum_{i=1}^{m}\delta(y_i = j) > 0 \\ c_j^t & \sum_{i=1}^{m}\delta(y_i = j) = 0 \end{cases} \tag{2-7}$$

2.2.5　模型的损失函数

由于训练数据存在不同情感之间数据不均衡的问题，因此在训练模型时，不直接使用 L_s^0 和 L_c^0，而是使用加权的交叉熵损失函数和中心损失函数。如公式（2-8）和公式（2-9）所示。

$$L_s = -\frac{1}{\sum_{i=1}^{m}\omega_{y_i}}\sum_{i=1}^{m}\omega_{y_i}log(\frac{e^{W_{y_i}^{\mathrm{T}}z_i + b_{y_i}}}{\sum_{j=1}^{n}e^{W_j^{\mathrm{T}}z_i + b_j}}) \tag{2-8}$$

$$L_c = \frac{1}{\sum_{i=1}^{m}\omega_{y_i}}\sum_{i=1}^{m}\omega_{y_i}||z_i - c_{y_i}||^2 \tag{2-9}$$

其中权重 w_j（$j \in \{1,2,...,n\}$）代表第 j 种情感对应的权重，它与训练集中该情感样本的个数成反比。

本章的模型在训练时使用的是由交叉熵损失函数和中心损失函数组成的联合损失函数，其计算方法如公式（2-10），其中 λ 是权衡中心损失函数相对于交叉熵损失函数重要性的超参数，λ 越大，中心损失函数所起的作用越大。当 $\lambda = 0$ 时，联合中心损失函数不发挥作用，模型仅仅受交叉熵损失函数的影响，这种情况下是我们的基准模型。

$$L = L_s + \lambda L_c \tag{2-10}$$

2.2.6　模型的学习算法

模型同时使用交叉熵损失函数和中心损失函数来提取具有情感区分度的更有效的特征，其学习算法总结见表格 2.1。

表 2.1　模型的学习算法

Algorithm 1 基于 Center Loss 的语音情感识别模型学习算法

Input: 训练数据 $\{(x_i, y_i)\}$，初始化好的神经网络参数 θ，初始化好的全局类中心点 $\{c_1, c_2, ..., c_n\}$，超参数 α, λ 的值，学习率 μ 的值。
Output: 训练好的神经网络模型参数 θ。
1. **while** not converge **do**
2.　　计算情感特征 z_i 的值，其中 $i \in \{1,2,...,m\}$
3.　　根据公式（2-7）更新 $\{c_j | j = 1,2,...,n\}$
4.　　计算模型的损失函数 L，其中 $L = L_s + \lambda L_c$
5.　　根据公式 $\theta \leftarrow \theta - \mu \frac{\partial L}{\partial \theta}$ 更新 θ
6. **while**

2.3　实验与分析

2.3.1　数据集介绍

本章实验在 IEMOCAP（The Interactive Emotional Dyadic Motion Capture）[47] 数据集上进行。IEMOCAP 数据集是一个使用的多模态多说话人数据集，包含视频、语音等数据。它由两部分组成，参与者在其中即兴表演或者表演剧本的数据。IEMOCAP 数据集由多个标注人员为数据标注情感标签，情感标签包括情感类别的方式以及情感维度表示的方式。

本章实验中我们仅使用 IEMOCAP 中的音频数据，根据情感类别标签训练语

音情感识别模型。我们筛选出数据标注为中性（neural）、生气（angry）、高兴（happy）、悲伤（sad）和兴奋（excited）的数据，舍弃掉标注为其他情感类别的数据量相对比较少的数据。同时由于标注为高兴的数据与标注为兴奋的数据在情感维度空间分布上差异不大，因此合并两种情感标签，统一用高兴（happy）表示。各种情感标签下数据的分布如表 2.2。

表 2.2　IEMOCAP 中不同情感标签音频数据的样本数

情感类别	中性	生气	高兴	悲伤
样本数量	1708	1103	1636	1084

2.3.2　实验设置

我们同时在对数刻度的梅尔谱和对数刻度的线性谱上进行实验。为了得到线性谱，一系列汉明窗（Hamming Window）应用在语音信号上。其中窗长 40 毫秒，窗移 10 毫秒。语音信号的采用率为 16kHz，进行傅立叶变换时频域维度为 1024，当计算梅尔谱时，梅尔滤波器的个数为 128。我们认为 14s 长的语音已经足够包含情感相关的信息，为了提升计算效率，对于时长大于 14s 的语音（占所有数据总数的 2.07%），我们仅提取中间 14 秒的片段来计算频谱图。此外，为保证模型性能稳定，数据集中的语谱图以帧为基本单位进行标准化（Normalization）。

由于各类别之间的数据不平衡-中性（占总数据的 30.9%），生气（占 19.9%），快乐（占 29.6%）和悲伤（占 19.6%），我们同时采用了非加权正确率（UA，Unweighted Accuracy, 即各个类别召回率的平均值）和加权正确率（WA，Weighted Accuracy, 预测正确的样本数除以测试集样本总数）作为评估语音情感识别系统性能的指标。在保证情感分布一致的前提下，我们把数据集随机分为 5 个子集，其中 4 个子集用于训练，一个子集的一半用作验证集，一半用作测试集。我们将重复交叉验证 5 次得到的平均结果作为最终结果。

在训练阶段，我们使用 Adam[58] 优化器，将学习率设置为 0.0003。每个 batch 包含的样本个数为 32，我们将训练过程中使得验证集上 UA 最高的神经网络参数作为模型的最终参数。

2.3.3　语音情感识别结果

模型中超参数 α 控制全局类中心点的更新速度，λ 代表中心损失函数的权重。我们首先进行实验，在梅尔谱作为输入时，研究超参数 α 和 λ 不同的值对语音情感识别性能的影响。实验共分为两组，第一组固定 $\lambda = 0.3$，设置 α 的值分别为 0.1、

0.3、0.5、0.7、0.9。第二组固定 $\alpha = 0.5$，分别设置 λ 的值为 0、0.003、0.03、0.3 和 3。

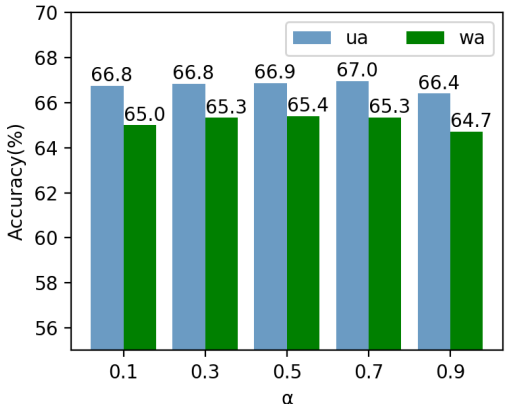


图 2.4　在梅尔谱输入下，固定 $\lambda = 0.3$，语音情感识别性能随 α 的变化规律

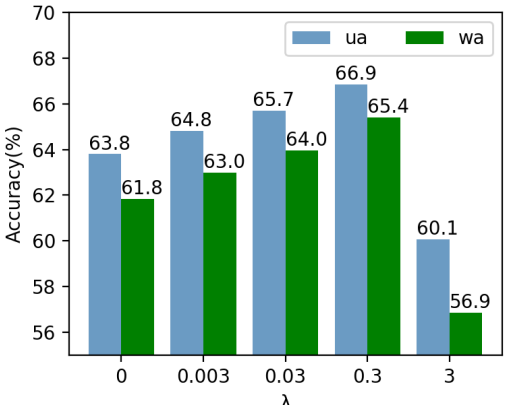


图 2.5　在梅尔谱输入下，固定 $\alpha = 0.5$，语音情感识别性能随 λ 的变化规律

第一组的实验结果如图 2.4所示，可以发现，当 α 从 0.1 变化到 0.9 时，非加权正确率和加权正确率的波动都不大，α 的值对语音情感识别性能的影响不大。第二组的实验如图 2.5所示，根据第二组实验结果可知，当固定 $\alpha = 0.5$ 时，λ 的值从 0.003 变化到 3 时，情感识别的正确率先增加后减小，λ 的值对语音情感识别的性能影响较大。

由于当 $\lambda = 0$ 时，端到端语音情感识别模型仅仅受交叉熵损失函数的影响，这是我们的基准模型。根据第二组实验结果，可以发现，当梅尔谱作为模型输入时，$\lambda = 0.3$、$\alpha = 0.5$（实验设置-1）时，模型的 UA 为 66.86%，WA 为 65.40%；而 $\lambda = 0$（实验设置-2）时，模型的 UA 和 WA 分别是 63.80% 和 61.83%。由此可见，在模型中引入中心损失函数，且把超参数 λ 设置为合适的值后，语音情感识别模型的性能得到显著提高。在梅尔谱作为输入的实验下，通过引入中心损失函数，并设置了合适的超参数值后，模型的非加权正确率和加权正确率均有了超过 3% 的提高。

同时我们也得出了实验设置-1 和实验设置-2 下的混淆矩阵（如表 2.3），由于我们执行了 5 次 5 折交叉验证，所以最终混淆矩阵是 5 次 5 折交叉验证的平均值。通过表 2.3可知，在梅尔谱作为输入的情况下，引入中心损失函数后，四种情感识别的正确率都得到了提升。

表 2.3　不同设置下情感识别混淆矩阵，（a）实验设置-1，（b）实验设置-2

	中性	生气	高兴	悲伤
中性	0.575	0.095	0.164	0.166
生气	0.119	0.691	0.155	0.035
高兴	0.211	0.162	0.511	0.115
悲伤	0.138	0.026	0.060	0.776

(a)

	中性	生气	高兴	悲伤
中性	0.637	0.067	0.167	0.127
生气	0.108	0.705	0.167	0.020
高兴	0.219	0.131	0.556	0.094
悲伤	0.128	0.025	0.070	0.777

(b)

更进一步地，我们在线性谱上进行实验，探究当模型使用线性谱作为输入时，中心损失函数的作用与影响。在线性谱输入下，我们分别比较了 $\lambda = 0.3$、$\alpha = 0.5$（实验设置-3）以及 $\lambda = 0$（实验设置-4）两种实验设置下，模型的性能。

实验结果表明，当 $\lambda = 0.3$、$\alpha = 0.5$ 时，模型的 UA 和 WA 分别为 65.13% 和 62.96%；当 $\lambda = 0$ 时，模型的 UA 和 WA 分别为 60.98% 和 58.93%。由此可见，引入中心损失函数，对于线性谱的输入依然能够提升语音情感识别的性能。

实验设置-3 和实验设置-4 下的混淆矩阵如表 2.4，通过比较表 2.4（a）和表 2.4（b）可以发现，在中心损失函数的作用下，模型提取更有效的更具有情感区分度的特征，这对于每个情感类别的识别正确率都有提升。

表 2.4　不同设置下情感识别混淆矩阵，（a）实验设置-3，（b）实验设置-4

	中性	生气	高兴	悲伤
中性	0.544	0.093	0.185	0.177
生气	0.127	0.681	0.167	0.025
高兴	0.216	0.186	0.476	0.122
悲伤	0.161	0.039	0.062	0.737

(a)

	中性	生气	高兴	悲伤
中性	0.573	0.073	0.196	0.157
生气	0.103	0.720	0.153	0.022
高兴	0.205	0.161	0.518	0.114
悲伤	0.125	0.028	0.053	0.793

(b)

我们按照表2.5在四种不同的设置下分别进行交叉验证实验，以比较不同的实验设置（表 2.5）对语音情感识别任务的影响。不同设置下模型的性能如图 2.6所示。通过比较实验设置-1 和实验设置-2，或者比较实验设置-3 和实验设置-4 可以发现，模型中引入中心损失函数可以提高语音情感识别任务的性能。通过比较实

验设置-1 和实验设置-3，或者比较实验设置-2 和实验设置-4 可以发现，由于梅尔谱按照人类的听觉特点对线性谱进行压缩，使用梅尔谱作为输入能取得更好的性能。

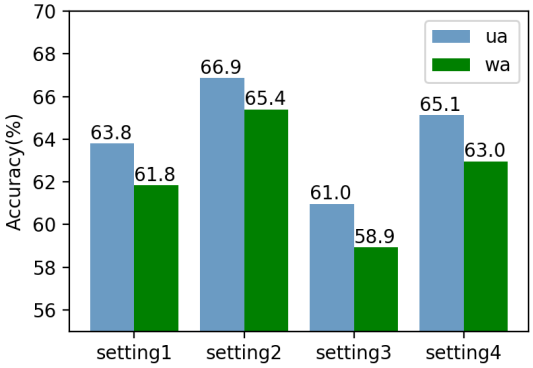


图 2.6　不同的实验设置下语音情感识别的正确率

表 2.5　各种实验设置下的超参数和输入

	超参数	输入
实验设置-1 （Setting1）	λ=0	梅尔谱
实验设置-2 （Setting2）	λ=0.3, α=0.5	梅尔谱
实验设置-3 （Setting3）	λ=0	线性谱
实验设置-4 （Setting4）	λ=0.3, α=0.5	线性谱

2.3.4　情感特征可视化结果

在中心损失函数的作用下，模型能够提取更具有情感区分度的特征。我们使用主成分分析（PCA，Principal Component Analysis）[59] 嵌入神经网络所提取的情感特征，以可视化中心损失函数的作用。我们在实验设置-1 和实验设置-2 中的 5 次交叉验证实验中随机选取一次，对其训练集和测试集产生的特征 z，分别进行 PCA 嵌入，产生的结果如图 2.7。

其中，图 2.7（b）和图 2.7（d）是在使用了中心损失函数的模型中所提取的特征。将图 2.7（b）和图 2.7（a）进行比较，或者将 2.7（d）和图 2.7（c）进行比较，可以发现，使用中心损失函数时，属于同一个情感类别的特征更加紧凑，具有更好的聚类效果。这也解释了为什么使用中心损失函数能够带来语音情感识别性能的损失，因为中心损失函数能够促进情感特征更接近其类中心点，而交叉熵损失函数则保证所提取的情感特征可分离，在两个损失函数的共同作用下，模型能够提取更具有情感区分度的特征，这最终带来了语音情感识别性能的提升。

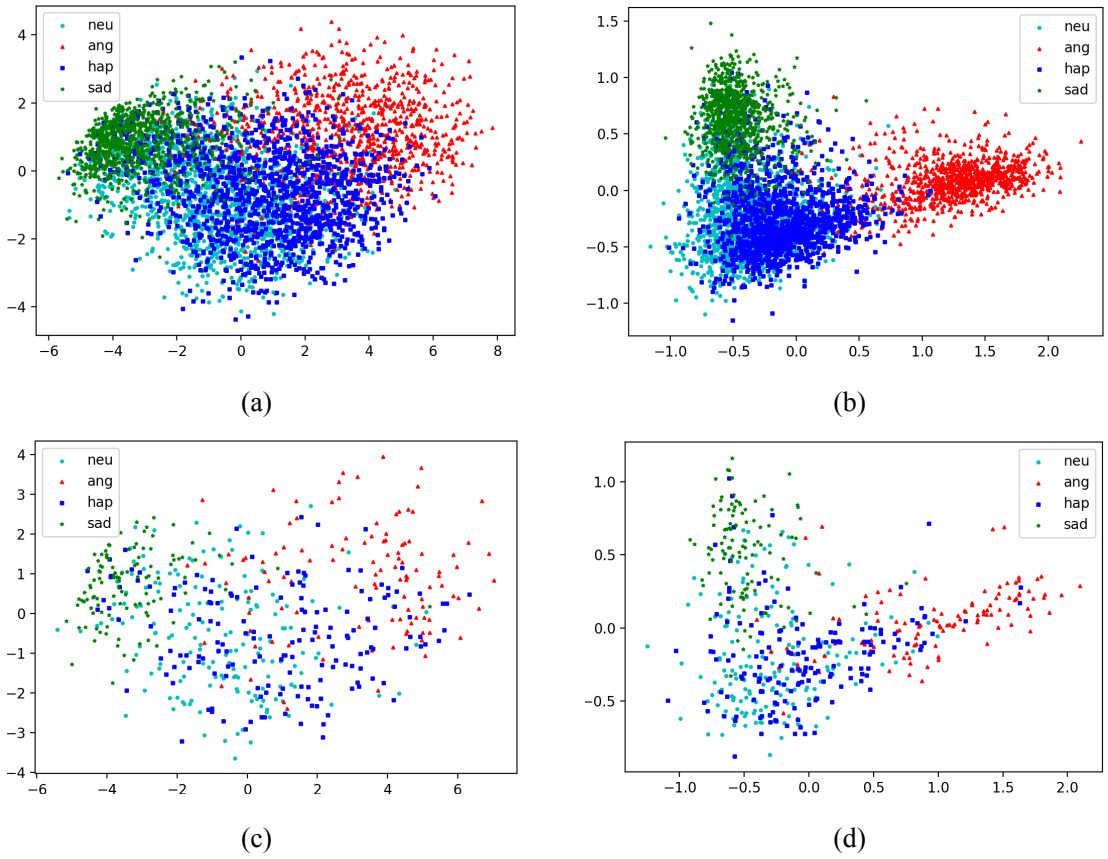


图 2.7　情感特征的 PCA 表示，（a）实验设置-1 下的训练集，（b）实验设置-2 下的训练集，（c）实验设置-1 下的测试集，（d）实验设置-2 下的测试集

2.4　本章小结

从语音中识别语音的状态对于建立更自然的语音交互系统至关重要。由于情感本身是主观含糊的，提取具有情感区分度的特征比较困难。在本章，我们提出了一种基于中心损失函数的端到端语音情感识别模型，它接受变长的语谱图作为输入，直接预测对应的情感。该模型同时使用中心损失函数和交叉熵损失函数，中心损失函数促进属于同一类的情感特征有更小的类内距离，交叉熵损失函数保证了所提取的情感特征是可分离的。在中心损失函数和交叉熵损失函数的共同作用下，神经网络能够更有效地识别语音的情感。

在 IEMOCAP 数据集上的实验结果表明，引入中心损失函数后，语音情感识别模型在梅尔谱的输入下，UA 和 WA 都提高了 3% 以上，当线性谱作为输入时，UA 和 WA 都提高了 4% 左右。通过可视化情感特征的 PCA 嵌入特征，我们可以发现，引入了中心损失函数的模型情感特征具有更好的聚类效果，在中心损失函数的作用下，模型提取的特征更具有情感区分度。

第 3 章　基于最大平均差异损失函数的跨域语音情感识别

3.1　本章引论

随着计算机相关技术尤其是神经网络的发展，语音识别准确率不断上升，语音合成产生的语音越来越自然、逼真。越来越多的语音产品出现在人们的日常生活中，为提高语音产品的智能化程度，语音情感识别相关技术也开始逐步应用到各个产品中。然而在实际场景中应用语音情感识别模型时，往往遇到实际应用场景的数据分布和模型的训练数据分布不一致的情形，这往往会导致语音情感识别模型在实际应用场景下不稳定。

语音情感识别任务中，当测试数据与模型的训练数据分布不一致时，称之为跨域语言情感识别问题。其中，训练数据被称为源域数据，测试数据被称为目标域数据。跨语言情感识别是一种典型的跨域情感识别问题：某些语言的情感语料库相对比较容易获取，比如英语、法语等；而某些语言的情感语料库相对比较缺乏，而且数据收集处理成本较高，比如越南语、马来语等小语种。为识别小语种下的语音的情感，根据英语等语言的情感语料库训练一个语音情感识别模型，应用在越南语等小语种上不失为一种可行的方式。另外，目前绝大多数语音合成数据是没有情感标签的，为了实现情感可控的语音合成，需要使用在情感语料库上训练好的语音情感识别模型预测合成数据的情感标签。然而不同于一些情感语料库包含情感非常强烈表演性的语音数据，合成语料库中的数据情感相对没有那么明确、强烈。合成语料库和情感语料库之间数据分布存在差异，因此识别合成数据的情感也是一个跨域语音情感识别任务。

Neumann 等人使用一个基于注意力机制的模型实现了英语和法语之间的跨语言情感识别[23]，Hodari 等人使用训练好的语音情感识别模型识别合成数据的情感标签以作为合成网络的条件输入[45]。但是以上工作均没有考虑源域和目标域数据分布之间的差异。在本章中，我们提出了一个基于最大平均差异损失的端到端跨域语音情感识别模型，该模型直接接受变长语谱图作为输入，直接输出情感标签的预测值，在和训练数据分布存在差异的测试数据上也能有稳定的性能。该模型同时使用交叉熵和最大平均差异（MMD）作为损失函数，其中交叉熵用于训练情感分类任务，最大平均差异损失函数用于降低源域特征和目标域特征之间分布的差异性。在最大平均差异损失函数的作用下，模型提取域无关的情感特征，这保证了模型在目标域数据上也能有稳定的性能。

3.2　基于最大平均差异损失函数的端到端跨域语音情感识别模型

3.2.1　模型概述

模型框架如图 3.1所示，模型由孪生的特征提取器（网络参数相同）和分类器组成。

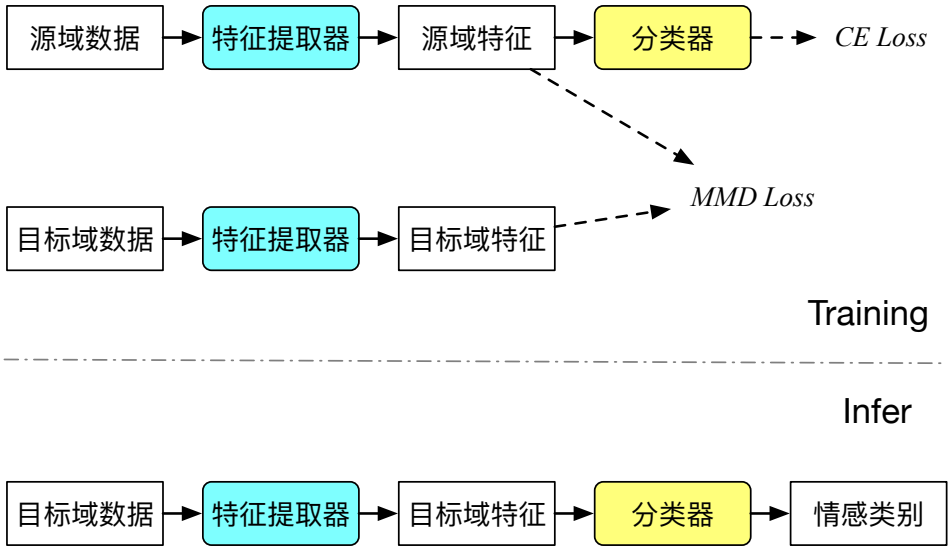


图 3.1　模型基本框架

模型在训练时同时接受源域数据和目标域数据作为输入，其中源域数据包含情感标签，目标域数据不包含情感标签信息。源域数据和目标域数据经过孪生的特征提取器后分别得到源域特征和目标域特征。源域特征再通过分类器，得到情感的标签的预测值。模型在训练过程中同时最小化交叉熵损失函数（CE Loss）和最大平均差异损失函数（MMD Loss）。交叉熵损失函数根据分类器的预测输出和情感标签的差异计算得到，保证模型特征提取器提取的是情感相关的特征。最大平均差异损失函数用于衡量源域特征和目标域特征之间分布的差异，促进模型提取域无关的特征。

模型在测试时接受测试数据（目标域数据）作为输入，经过特征提取器得到测试数据的特征，然后再经过分类器得到测试数据的预测标签。

3.2.2　模型细节

模型直接接受变长语谱图作为输入，语谱图直接由语音信号数据得到，没有过多的特征工程工作。语谱图可以是线性谱或者梅尔谱，根据章节2.3.3中的实验结果，梅尔谱作为输入能取得更好的语音情感识别结果，因此本章实验中都是使

用梅尔谱作为输入。

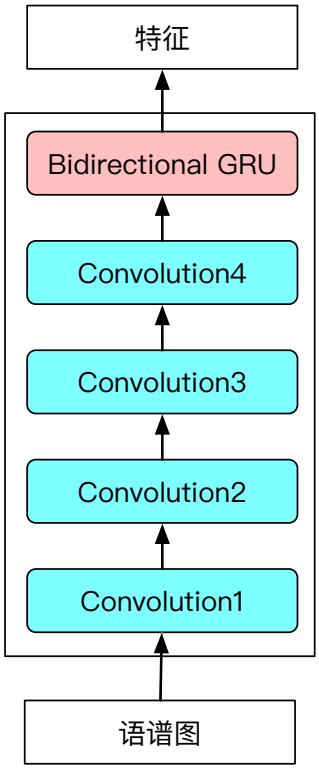


图 3.2　特征提取器的网络结构

特征提取器接受变长的语谱图作为输入，输出定长的特征表示。它采用 CNN+RNN 的模型架构，其中卷积神经网络（CNN）用于提取局部特征以及在时间维度降维，循环神经网络提取时序特征并将其最后一个时间步的输出作为特征提取器模块的输出。特征提取器网络结构模型如图 3.2所示，它包括四个卷积层和一个双向循环网络层。每一层具体参数和操作见表 3.1。

表 3.1　特征提取器每一层的具体的参数设置

	参数设置
Bidirectional GRU	前向 GRU：128 维；后向 GRU：128 维
Convolution4	卷积核个数:96；卷积核大小 3×3； 卷积步长 2×2；2×2 最大池化
Convolution3	卷积核个数:80；卷积核大小 3×3； 卷积步长 2×2；2×2 最大池化
Convolution2	卷积核个数:64；卷积核大小 3×3； 卷积步长 2×2；2×2 最大池化
Convolution1	卷积核个数:48；卷积核大小 7×7；卷积步长 2×2

分类器的网络结构如图 3.3所示，它由两层全连接网络组成。由于特征提取器的输出层为 128 维的双向 GRU 网络，因此分类器的输入特征的维度为 256。第一层全连接网络（Dense1）的输出维度为 64，即把 256 维的特征映射为 64 维；第二层全连接网络（Dense2）的激活函数为 Softmax，其输出代表对各个情感类别后验概率的预测值。

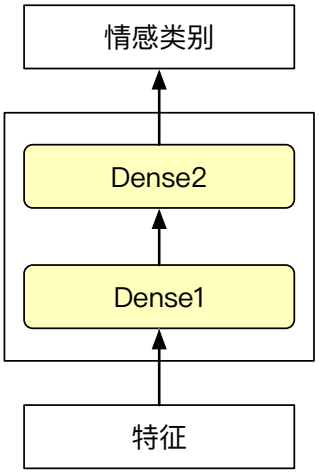


图 3.3　分类器的网络结构

3.2.3　最大平均差异损失函数

假设源域特征为 s，服从分布 $p(s)$，目标域特征为 t，服从分布 $q(t)$。最大平均差异（MMD）[48] 描述的是分布 $p(s)$ 和分布 $q(t)$ 之间的差异，如公式（3-1）所示，它表示特征 s 和特征 t 经过一系列函数 f 映射后，期望值 $E_p(f(s))$ 和 $E_q(f(t))$ 差值的上确界，其中 $\mathcal{F}$ 是所有函数 f 可能的值组成的集合。

$$\mathrm{MMD}(\mathcal{F},p,q)=\sup_{f\in\mathcal{F}}(E_p(f(s))-E_q(f(t))) \tag{3-1}$$

对于 $\mathcal{F}$ 是再生核希尔伯特空间（RKHS，Reproducing Kernel Hilbert Space）的单位球时，根据里斯表示定理（Riesz's Representation theorem），可以得到公式（3-2）。

$$f(x)=<f,\phi(x)>_{\mathcal{H}} \tag{3-2}$$

因此：

$$\begin{aligned}E_p(f(s))&=<f,E_p(\phi(s))>_{\mathcal{H}}=<f,\mu_p>_{\mathcal{H}}\\E_q(f(t))&=<f,E_q(\phi(t))>_{\mathcal{H}}=<f,\mu_q>_{\mathcal{H}}\end{aligned} \tag{3-3}$$

当 $\mathcal{F}$ 是再生核希尔伯特空间的单位球时，即 $\|f\|_{\mathcal{H}}\leqslant 1$ 时，计算 MMD 的平

方，得到公式（3-4）。

$$\begin{aligned}\mathrm{MMD}^2(\mathcal{F},p,q)=&[\sup_{\|f\|_{\mathcal{H}}\leqslant 1}(E_p(f(s))-E_q(f(t)))]^2\\=&[\sup_{\|f\|_{\mathcal{H}}\leqslant 1}(<f,\mu_p>_{\mathcal{H}}-<f,\mu_q>_{\mathcal{H}})]^2\\=&[\sup_{\|f\|_{\mathcal{H}}\leqslant 1}(<\mu_p-\mu_q,f>_{\mathcal{H}})]^2\end{aligned}\tag{3-4}$$

我们希望找到最好的 f 使得 $<\mu_p-\mu_q,f>_{\mathcal{H}}$ 最大，当两个定长向量同向时，其乘积最大，所以：

$$\begin{aligned}\mathrm{MMD}^2(\mathcal{F},p,q)=&[\sup_{\|f\|_{\mathcal{H}}\leqslant 1}(<\mu_p-\mu_q,f>_{\mathcal{H}})]^2\\=&\left\|\mu_p-\mu_q\right\|_{\mathcal{H}}^2\end{aligned}\tag{3-5}$$

而 $\mu_p=E_p(\phi(s))$，$\mu_q=E_q(\phi(t))$，因此：

$$\begin{aligned}\mathrm{MMD}^2(\mathcal{F},p,q)=&\left\|\mu_p-\mu_q\right\|_{\mathcal{H}}^2\\=&<\mu_p-\mu_q,\mu_p-\mu_q>_{\mathcal{H}}\\=&<\mu_p,\mu_p>_{\mathcal{H}}+<\mu_q,\mu_q>_{\mathcal{H}}-2<\mu_p,\mu_q>_{\mathcal{H}}\\=&E_p<\phi(s),\phi(s')>_{\mathcal{H}}+E_q<\phi(t),\phi(t')>_{\mathcal{H}}\\&-2E_{p,q}<\phi(s),\phi(t)>_{\mathcal{H}}\end{aligned}\tag{3-6}$$

我们定义核函数 $k(x,x')=<\phi(x),\phi(x')>$，则在计算 $\mathrm{MMD}^2(\mathcal{F},p,q)$ 直接根据核函数计算即可，不需要关注 $\phi(x)$ 具体的形式。

实际场景中，我们按照公式（3-7）计算最大平均差异损失函数。

$$\begin{aligned}L_{\mathrm{MMD}}=\mathrm{MMD}^2(\mathcal{F},p,q)=&\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}k(s_i,s_j)+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(t_i,t_j)\\&-\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}k(s_i,t_j)\end{aligned}\tag{3-7}$$

其中 m 代表源域特征的数量，n 代表目标域特征的数量，s_i 代表第 i 个源域特征，t_i 代表第 i 个目标域特征。由于再生核希尔伯特空间往往是高维甚至是无限维的，对应的核函数一般选用表示无穷维的高斯核，如公式（3-8）。

$$Gaussion(x,x')=e^{-\frac{\left\|x-x'\right\|^2}{2\sigma^2}}\tag{3-8}$$

其中 σ 为超参数，模型的性能受 σ 值的选取的影响。为了避免去选择一个最优的 σ，参考 Long 等人的方法[60]，我们使用多核核函数，核函数的计算方式如公式（3-9）。

$$k(x,x') = \frac{1}{\sum_i^K \beta_i} \sum_i^K \beta_i G_i(x,x') \tag{3-9}$$

其中 K 是子核函数的个数，β_i 代表第 i 个子核函数的权重，第 i 个子核函数采用高斯核，见公式（3-10）。

$$G_i(x,x') = e^{-\frac{\left\|x-x'\right\|^2}{2\sigma_i^2}} \tag{3-10}$$

本章实验中，K 的取值为 19，$\beta_1 = \beta_2 = ... = \beta_{19} = 1$。$\{\sigma_i | i = 1,2,..,19\}$ 的取值依次为 $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 5, 10, 15, 20, 25, 30, 35, 10^2, 10^3, 10^4, 10^5, 10^6\}$。

最大平均差异损失函数用于衡量源域特征的分布 p 和目标域特征的分布 q 之间的差异，在最大平均差异损失函数的作用下，特征提取器所输出的源域特征和目标域特征之间分布的差异得以减小。

3.2.4　模型的训练与应用

为了保证特征提取器所提取特征是情感相关的，除了最大平均差异损失外，模型在训练时还使用交叉熵作为损失函数。交叉熵计算方法见公式（3-11）。

$$L_{\mathrm{CE}}^0 = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} y_j^i log(\hat{y}_j^i) \tag{3-11}$$

其中，d 代表情感类别的个数，即第二层全连接层的输出维度。n 代表一个 batch 中样本的个数。$y \in \mathbb{R}^d$ 是情感标签的独热（One Hot）编码，$\hat{y} \in \mathbb{R}^d$ 是第二层全连接成经过 Softmax 激活函数后输出的情感类别的后验概率。y_j^i 代表第 i 个样本的标签独热编码中的第 j 个单元，$\hat{y}_j^i$ 表示模型对第 i 个样本属于第 j 个情感的预测概率。

由于不同类情感数据不均衡的原因，我们采用带权重的交叉熵损失函数训练模型，其中属于不同类情感的样本分别有不同的权重，如公式（3-12）。

$$L_{\mathrm{CE}} = -\frac{1}{\sum_{i=1}^{n} \omega_{argmax_j(y_j^i))}} \sum_{i=1}^{n} (\omega_{argmax_j(y_j^i))} \sum_{j=1}^{d} y_j^i log(\hat{y}_j^i)) \tag{3-12}$$

其中 ω_k 表示第 k 类情感对应的权重，它与训练集中第 k 类情感样本数目成反比。

模型在训练时，孪生的特征提取器分别接受源域和目标域的变长语谱图作为输入，得到源域特征 s 和目标域特征 t，模型根据 s 和 t 按照公式（3-7）计算最大平均差异损失函数，然后源域特征 s 经过分类器，得到情感的预测概率 $\hat{y}$，模型根据 $\hat{y}$ 和情感标签 y 按照公式（3-12）计算交叉熵损失函数。模型在交叉熵损失函数和最大平均差异损失函数的联合作用下，提取情感相关域无关的特征，以保证在跨域语音情感识别任务中有稳定的性能。模型总的损失函数见公式（3-13），其中超参数 λ 调节最大平均差异损失函数相对于交叉熵损失函数的权重，这里 λ 取 0.5。

$$L = L_{\mathrm{CE}} + \lambda L_{\mathrm{MMD}} \tag{3-13}$$

模型在目标域数据上进行推断时，目标域的语谱图通过特征提取器得到目标域特征，然后分类器接受目标域特征作为输入，预测其对应的情感后验概率。

3.3　实验与分析

3.3.1　数据集介绍

本章在 RECOLA[49] 和 IEMOCAP[47] 两个情感语料库上进行跨语言语音情感识别模型实验，以测试模型的性能。

RECOLA 多模态数据集由瑞士弗里堡大学的研究人员提供，该数据旨在研究人们在交流时的情感反馈。该数据集包含 9.5 个小时的音频、视觉和生理（心电图和皮肤电活动）记录，这些数据记录了多位讲法语的参与者之间的在线二元互动。有 6 位标注人员在帧级别上标注了情感信息，标注信息包括效价度（Arousal）和激活度（Valence），其取值范围都是 $[-1, 1]$。由于我们关注的是一整句话的情感，因此把一句话中所有帧的标注值进行平均，得到整句话的标签。本章实验中我们使用了来自 23 个说话人的共计 1308 句音频数据。

本章实验中我们分别对效价度（正向/负向）和激活度（高/低）进行分类。对于 RECOLA 数据，我们把效价度或激活度值属于 $[-1, 0]$ 的记为 0，值属于 $(0, 1]$ 的记为 1。由于 IEMOCAP 中，效价度和激活度值的为 $[1, 5]$。对于 IEMOCAP，我们把效价度或激活度值属于 $[1, 2.5]$ 的标注为 0，值属于 $(2.5, 5]$ 的标注为 1。每个数据集下不同标签情感的计数见表 3.2。

3.3.2　实验设置

根据第2章中的实验结果，相对于线性谱，模型接受梅尔谱作为输入有更好的语音情感识别结果，本章实验中我们只在梅尔谱作为输入的情形下进行实验，提取梅尔谱时的参数设置如表 3.3。

表 3.2　各种数据集下不同标签情感的计数

数据集	标签类型	标注为 0 的数量	标注为 1 的数量
IEMOCAP	Arousal	3121	6918
IEMOCAP	Valence	5356	4683
RECOLA	Arousal	520	788
RECOLA	Valence	246	1062

表 3.3　梅尔谱计算的参数配置

参数名称	参数值
语音信号采样率	16kHz
窗函数	汉明窗 (Hamming windows)
窗长	40ms
窗移	10ms
DFT 频率维度	1024
梅尔滤波器个数	128

我们在英文语料库（IMOCAP）和法语语料库（RECOLA），分别针对激活度和效价度进行实验。我们把 $\lambda = 0$ 时的模型作为基准模型，当 $\lambda = 0$ 时，最大平均差异损失不发挥作用，模型等价于在源数据上训练（不使用目标域数据的信息）后，直接应用到目标域数据上。

按照源语言目标语言的不同，以及标签类型的不同，是否使用最大平均差异损失，共有八组不同的实验设置，见表 3.4。当使用 IEMOCAP 作为源数据时，随机取其中 500 句音频作为验证集，当使用 RECOLA 作为源数据时，随机取其中 200 句音频作为验证集。为避免随机误差，每组实验重复 5 次，取其平均值作为模型最终结果。我们采用非加权正确率（UA）和加权正确率（WA）作为模型性能评判的指标。

表 3.4　不同设置下的八组实验

	λ 取值	源数据	目标数据	标签类型
实验设置-1	0	IEMOCAP	RECOLA	Arousal
实验设置-2	0	IEMOCAP	RECOLA	Valence
实验设置-3	0	RECOLA	IEMOCAP	Arousal
实验设置-4	0	RECOLA	IEMOCAP	Valence
实验设置-5	0.5	IEMOCAP	RECOLA	Arousal
实验设置-6	0.5	IEMOCAP	RECOLA	Valence
实验设置-7	0.5	RECOLA	IEMOCAP	Arousal
实验设置-8	0.5	RECOLA	IEMOCAP	Valence

为保证训练过程的稳定，对于训练集、测试集和验证集的数据，我们都按照公式（3-14）以帧为单位对输入的梅尔谱进行标准化。

$$x'_{frame} = \frac{x_{frame} - \mu}{\sigma} \tag{3-14}$$

其中 x_{frame} 指正则化前某帧的值，x'_{frame} 代表正则化后该帧的值，μ 代表训练集数据以帧为单位所求得的均值，σ 代表训练集数据以帧为单位所求得的标准差。

在训练过程中，使用 Adam[58] 优化器来最小化交叉熵函数和最大平均差异函数。初始学习率（Warm Up Learning Rate）设为 0.00003，初始学习率训练 100 步后，学习率变为 0.0003 一直训练，一个 batch 样本的个数为 32。我们取验证集下取得最好结果（UA 最大）的模型参数作为模型的最终参数。

3.3.3　跨语言语音情感识别结果与分析

我们通过实验验证最大平均差异损失函数（MMD Loss）在不同的实验设置下的作用，得到实验结果如表 3.5 和表 3.6所示。表格中“基准模型正确率”指 $\lambda = 0$ 时（即不使用 MMD Loss 时）模型的正确率，“MMD 模型正确率”指 $\lambda = 0.5$ 时（即使用 MMD Loss 时）模型的正确率。

表 3.5　不同实验设置下的模型的非加权正确率（UA）

源数据	目标数据	标签类型	基准模型正确率	MMD 模型正确率
IEMOCAP	RECOLA	Arousal	0.529	0.550
IEMOCAP	RECOLA	Valence	0.470	0.489
RECOLA	IEMOCAP	Arousal	0.500	0.543
RECOLA	IEMOCAP	Valence	0.505	0.518

表 3.6　不同实验设置下的模型的加权正确率（WA）

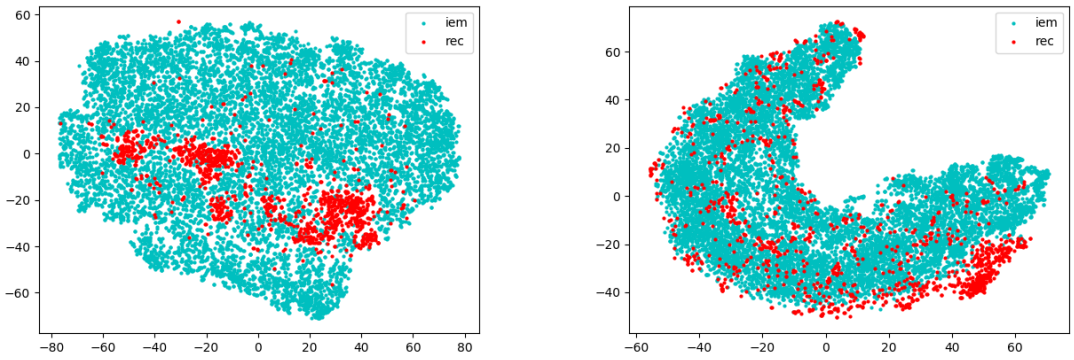
源数据	目标数据	标签类型	基准模型正确率	MMD 模型正确率
IEMOCAP	RECOLA	Arousal	0.534	0.538
IEMOCAP	RECOLA	Valence	0.632	0.720
RECOLA	IEMOCAP	Arousal	0.617	0.642
RECOLA	IEMOCAP	Valence	0.488	0.500

表 3.5表示不同的跨语言设置、不同的标签类型下，引入最大平均差异损失函数前后，模型的非加权准确率变化。通过表 3.5可知，引入最大平均差异损失函数后，非加权准确率提升了 1.3% ~4.3%。表 3.6表示不同的实验设置下，是否使用最大平均差异损失函数，模型的加权准确率的对比结果。通过表 3.6可知，在使用最

大平均差异损失函数的情况下，模型的加权准确率在四种实验设置下平均提升了约 3.2%。实验结果表明，在跨语言语音情感识别任务中，引入平均差异损失函数来提取域无关特征的方法，对提高跨语言语音情感识别模型的性能是有效的。

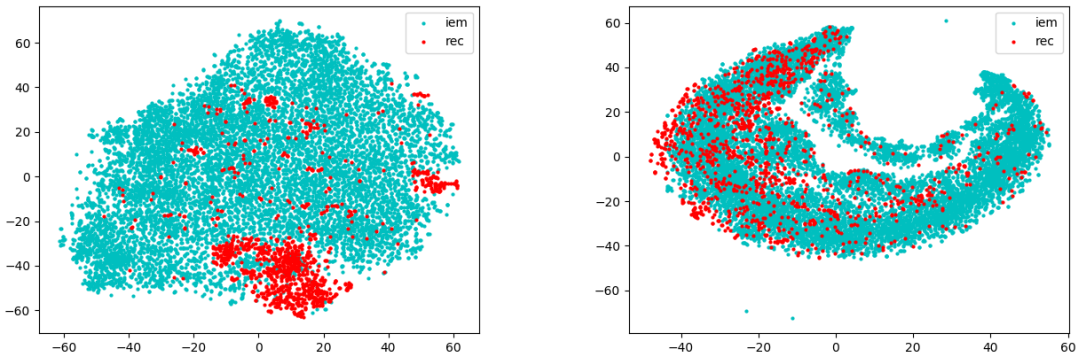
3.3.4　不同域情感特征可视化表示

最大平均差异用于衡量源域特征和目标域特征的分布的差异，本小节使用 t-分布邻域嵌入算法（t-SNE，t-Distributed Stochastic Neighbor Embedding）[61] 将通过训练好的模型得到的两数据集中所有的情感特征（特征提取器的输出）降到 2 维，以可视化最大平均差异损失函数对两数据集情感特征分布的影响。



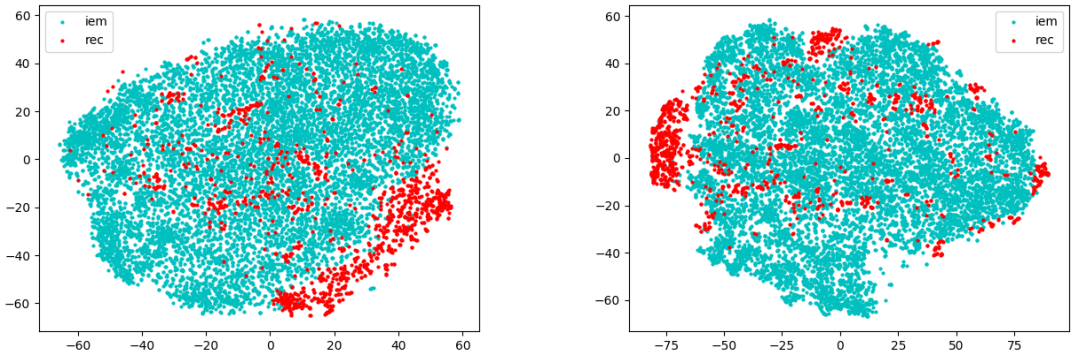
(a)　　(b)

图 3.4　源数据为 IEMOCAP，目标数据为 RECOLA，标签类型为 Arousal 时两个数据集情感特征的分布，（a）未使用 MMD Loss（b）使用 MMD Loss



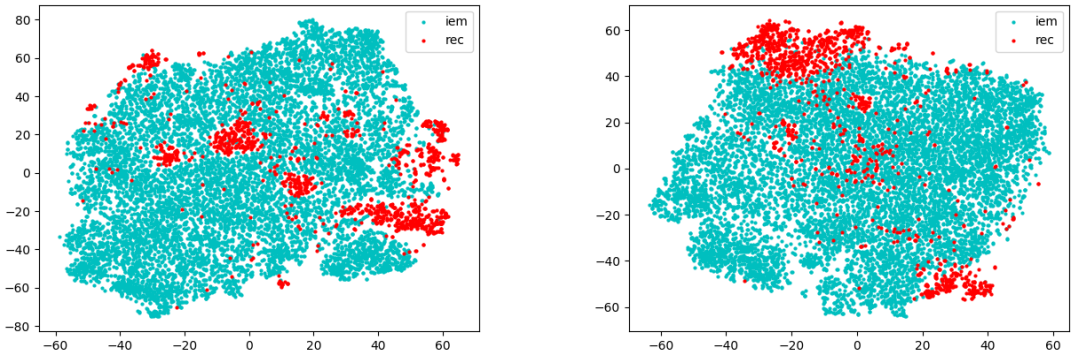
(a)　　(b)

图 3.5　源数据为 IEMOCAP，目标数据为 RECOLA，标签类型为 Valence 时两个数据集情感特征的分布，（a）未使用 MMD Loss（b）使用 MMD Loss



(a)　　(b)

图 3.6　源数据为 RECOLA，目标数据为 IEMOCAP，标签类型为 Arousal 时两个数据集情感特征的分布，（a）未使用 MMD Loss（b）使用 MMD Loss



(a)　　(b)

图 3.7　源数据为 RECOLA，目标数据为 IEMOCAP，标签类型为 Valence 时两个数据集情感特征的分布，（a）未使用 MMD Loss（b）使用 MMD Loss

图3.4 ~ 图3.7 分别表示四种不同的实验设置下，引入最大平均差异损失函数，对两数据集（IEMOCAP 和 RECOLA）情感特征分布的影响。其中子图-(a) 表示基准模型（未使用最大平均差异损失函数）下，两数据集情感特征的分布；而子图-(b) 表示模型使用最大平均差异损失函数的情况。比较子图-(a) 和子图-(b) 可以发现，使用最大平均差异损失函数后，RECOLA 数据集的情感特征分布 IEMOCAP 数据集的情感特征分布更为接近，这验证了最大平均差异损失函数可以从分布存在差异的数据中提取分布一致的特征，因此引入最大平均差异损失函数后跨语言语音情感识别任务性能得以提升。

3.4　本章小结

在语音情感识别的实际应用场景中，常常存在测试数据和训练数据分布不一致的情况。本章在端到端语音情感识别模型中引入最大平均差异损失函数，以实现跨域语音情感识别。最大平均差异衡量的是源域特征和目标域特征通过一系列映射后其分布差值的上确界，通过里斯表示定理以及核技巧，把最大平均差异转化为可微分的、更方便计算的形式（公式（3-7））。在最大平均差异损失函数的作用下，模型从分布不一致的源域数据和目标域数据中提取出分布一致的情感特征，在目标域上实现稳定的语音情感性能。

本章通过跨语言语音情感识别任务验证了最大平均差异损失函数，引入最大平均差异损失函数后，跨语言语音情感识别任务的准确率得以明显提升，进一步又通过 t-SNE 图可视化了最大平均差异损失函数对情感特征分布的影响。

第 4 章　基于全局风格令牌的情感语音合成

4.1　本章引论

语音合成，又叫文语转换（TTS），指把文字信息转化为声音信息的技术。随着神经网络的发展，语音合成生成语音的质量得到极大的提升，尤其是端到端语音合成框架[31-35] 和基于神经网络的声码器[38-39] 出现后，目前使用计算机合成的语音在清晰度、自然度等方面的表现接近甚至超过了人声，达到了可商用的标准。随着越来越多语音合成相关产品的出现，对合成语音的控制性有了更多的需求，比如如何合成控制语速、音色、情感等信息。控制生成语音的情感，实现在特定的场景下生成带有合适情感的语音，能够使语音相关产品更智能，使人们的生活更便利。

关于带情感语音的合成，目前的研究主要关注于从语音中提取有效的特征表示，作为合成网络的条件输入，以控制合成语音的情感风格[41-42,62]。以上工作通过改变作为条件输入的情感特征影响生成语音的情感风格，但还无法实现情感显式可控的语音合成，即输入情感标签信息合成对应情感的声音。企业中为了实现可控的语音合成，需要向数据公司订购特定的情感合成语料库，一般使用情感标签的独热（One hot）编码作为合成模型的条件输入。然而公司采购的情感合成语料库以表演型语料居多，每句话的情感都是很明确且强烈的，不是情感 A 就是情感 B，很少有情感 A 和情感 B 共存的情况。另外，相对于从语音中提取情感特征表示，使用独热编码得到的情感特征不能很好地建模情感之间的相互关系。以上两点决定了目前企业中采用的方案不能对合成语音的情感进行更精细地控制，比如控制某种情感的剧烈的程度，尝试合成同时具有情感 A 和情感 B 的声音。

本章提出了一种基于全局风格令牌的情感语音合成方案，该方案使用从语音中提取的特征作为端到端合成网络的条件输入，在没有情感标签的开源合成数据上也实现了情感可控地语音合成。本方案具有以下亮点：

a. 本方案使用全局风格令牌提取情感特征，加强语音中的副语言信息。同时在全局风格令牌中引入情感预测损失函数，以实现在情感二维空间表示下，效价度和激活度对生成语音情感相互独立的控制；
b. 本方案使用全局风格令牌的权重作为情感信息的表征，这是一种相对低维的特征，方便推断时建立情感到特征的映射；
c. 本方案首先先使用跨域的语音情感识别模型预测合成数据的标签。为模型训练时引入情感预测损失函数、模型推断时建立情感到特征的映射等工作提供

基础。

4.2　基于情感损失函数的情感语音合成模型

4.2.1　模型概述

本章中，我们使用一个跨域的语音情感识别模型预测出了合成数据的情感标签信息。在训练情感合成模型时，为了充分利用这些情感标签的监督信息，以提取更有效的情感特征，我们提出了一个基于全局风格令牌的情感语音合成模型。

如图 4.1所示，模型包含四个子模块，分别是文本编码器（Text Encoder）模块、解码器（Decoder）模块、韵律编码器（Prosody Encoder）模块和全局风格令牌（GST）模块。其中 Text Encoder 模块和 Decoder 模块构成了端到端语音合成的基础模型。Prosody Encoder 从参考语音（Reference Audio）中提取韵律嵌入向量（Prosody Embedding），在训练阶段参考语音就是训练合成网络的目标语音，即 Reference Audio = Ground Truth Audio。Prosody Embedding 通过 GST 模块后其文本信息得到过滤，情感相关信息得到加强，最终输出情感嵌入向量（Emotion Embedding），Emotion embedding 即为基础端到端合成网络的条件输入，最终影响合成语音的情感。

GST 模块中包含一组风格令牌（Style Token），它们是一组定长的向量，随着模型的训练而更新。通过比较各个 Style Token 和 Prosody Embedding 之间的相似度，得到每个 Token 的权重，这些 Token 的加权平均即为 Emotion Embedding。由于这些 Token 是在整个训练集上训练得到的，所以称之为全局风格令牌（GST，Global Style Token）。

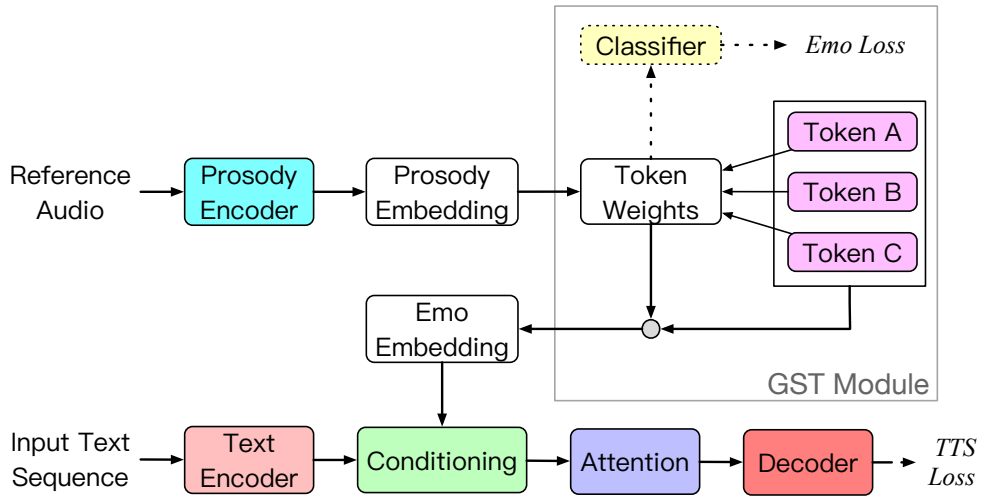


图 4.1　基于全局风格令牌的情感语音合成模型结构

在模型推断阶段（如图 4.2），我们根据输入的情感标签信息得到其对应 Token 权重（Weights），然后各个 Token 根据其权重进行加权平均，最终得到 Emotion Embedding，作为基础合成网络的条件输入，实现情感可控的语音合成。

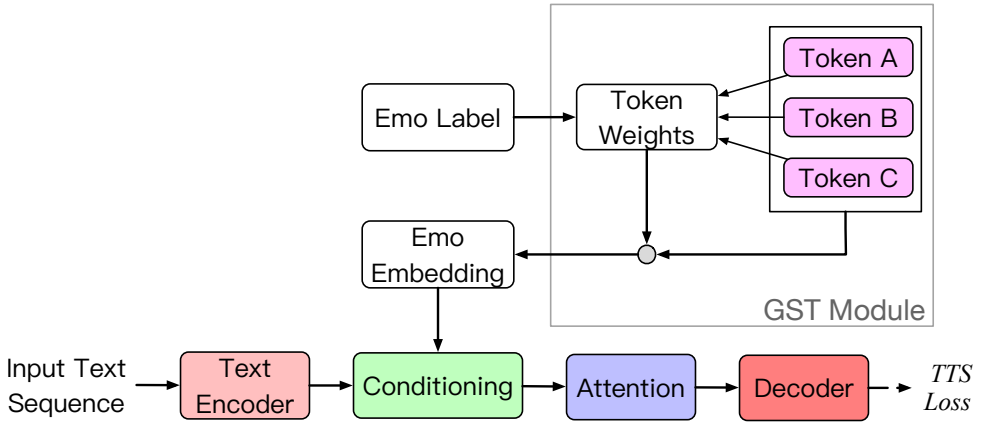


图 4.2　模型推断时的结构及流程

4.2.2　端到端语音合成模型基础

本章所提出的方法是在一个基础端到端语音合成模型的基础上，增加一个带情感信息的特征表示作为条件输入，以实现情感语音合成。我们使用的基础端到端语音合成模型是 Tacotron[32]，这里就 Tacotron 模型进行简明介绍。

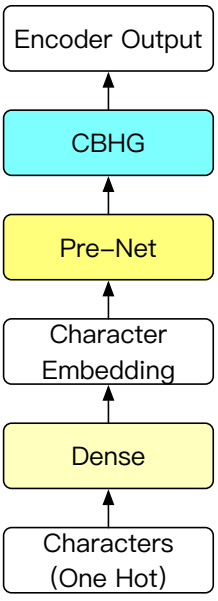


图 4.3　Tocotron Encoder 的结构

Tacotron 的 Encoder 直接接受字符的独热编码组成的序列作为输入，然后对输入文本的上下文关系进行建模，最终输出一个和输入序列等长的特征序列（Encoder

Output）。

Encoder 的结构如图 4.3所示，输入文本字符独热编码组成的序列经过一层全连接层[①]后，得到字符编码序列（Character Embedding）。字符编码序列经过前处理（Pre-Net）网络子结构，以及 CBHG 网络子结构，最终得到 Encoder 的输出序列（Encoder Output）。其中 Pre-Net 网络子结构由多层全连接网络组成。

CBHG 网络子结构如图 4.4所示，它包含一个多时间步长卷积网络组、多个普通卷积网络层、高速网络组和由 GRU 单元组成的双向循环网络。其中多时间步长卷积网络组指卷积网络的有多组大小的卷积核，以同时建模不同距离大小下的上下文信息。高速网络组（Highway Network）[63] 由多层全连接层实现，与普通全连接层不同的是，高速网络组中每一层后分别有一个门函数以防止梯度消失或梯度爆炸。最后由 GRU 单元实现的双向循环网络进一步建模输入文本序列的上下文信息和时序信息。

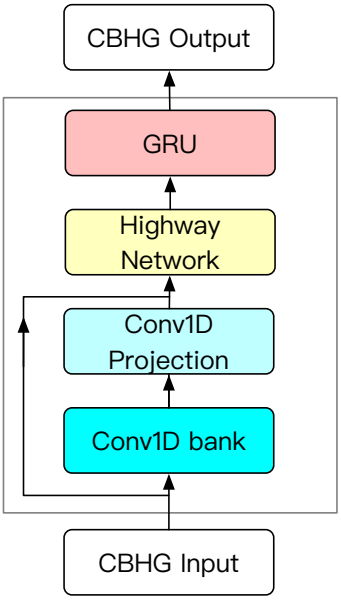


图 4.4　CBHG 的结构

值得注意的是，Encoder 的输出是一个特征序列，而我们用来做条件输入的情感嵌入向量（Emotion Embedding）是一个定长的向量，因此我们需要将 Emotion Embedding 复制 t_{enc} 份，分别与 Encoder Output 的每个单元拼接起来，其中 t_{enc} 指 Encoder 的时间步个数。

Decoder 通过一个自回归的结构逐步预测语音的频谱。其结构如图 4.5所示，包括 Pre-Net 网络子结构、Attention-RNN 网络子结构、Decoder-RNN 网络子结构和 CBHG 网络子结构。其中 Pre-Net 网络子结构、CBHG 网络子结构的模型结构同

① 本小节中所述的全连接层都是以字符为级别或者以帧级别（以时间步为级别）的。

Encoder 中的相关模块。Attention-RNN 网络子结构包含多个循环网络层，其输出用作 Bahdanau Attention[64] 的查询向量（Query）。Decoder-RNN 网络子结构包含多个循环网络层和一个全连接层，它同时接受 Attention-RNN 的输出以及由 Attention 提供的上下文信息特征作为输入，输出梅尔谱的预测值。CBHG 网络子结构把预测梅尔谱转换为预测线性谱。最后我们使用 Griffin-Lim 算法[65]，将线性谱转换为语音信号。

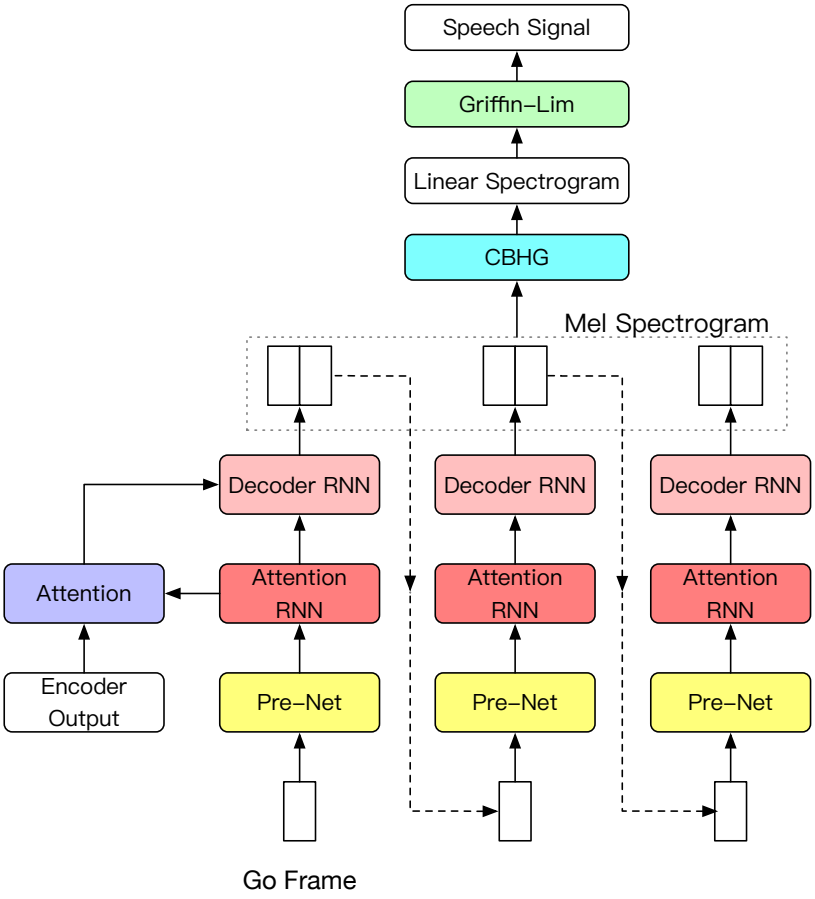


图 4.5　Tocotron Decoder 的结构

合成网络的损失函数 L_{TTS} 用于衡量预测频谱（梅尔谱、线性谱）和实际频谱之间的差异。我们在开源数据上训练，数据质量相对较差，而平均绝对误差（MAE，Mean Absolute Error）对异常数据点相对不敏感，因此我们采用 MAE 计算合成网络的损失函数。

频谱由原始语音提取得到，由于原始语音不是等长的，因此在模型训练时一个 batch 中所有样本，其频谱长度往往是不同的。为了并行处理训练数据，加快计算速度，实际训练时我们把一个 batch 中最长音频的频谱长度作为所有频谱数据频谱长度，有效频谱长度不足的数据，后续补 0。

计算 MAE 时是否考虑频谱补 0 的部分，也会对合成效果产生较大的影响。如果计算 MAE 时忽略频谱补 0 的部分，训练出来的模型合成语音时往往不能在文本内容结束后及时停止，而如果计算 MAE 时直接使用频谱补 0 的部分，又往往会使合成语音过早停止。综合以上实验反馈结果，最终我们计算 MAE 时，分别对频谱有效部分和频谱补 0 部分设置不同的权重（频谱有效部分权重为 1，频谱补 0 部分权重为 0.1），以实现较好的合成效果。

4.2.3　情感离散表示下的特征提取-1（使用全局风格令牌）

在 Tacotron 模型的基础上，我们添加额外的情感特征作为基础合成模型的条件输入来实现情感语音合成。这里我们采用 Wang 等人提出的 GST 模型[42] 来提取语音情感特征表示。其中，情感特征是由一组在整个训练集上得到的风格令牌（Global Token）组合得到的，因此所提取的情感特征更好地过滤掉了参考语音中的文本信息，更好地保留了副语言信息。

如图 4.6，相比于 Tacotron 模型，GST 模型增加了 Prosody Encoder 模块和 GST 模块。其中 Prosody Encoder 模块从参考语音的梅尔谱中提取 Prosody Embedding，Prosody Embedding 通过 GST 模块后得到过滤掉文本信息的 Emotion Embedding，Emotion Embedding 作为 Tacotron 的条件输入，最终影响合成语音的情感风格。

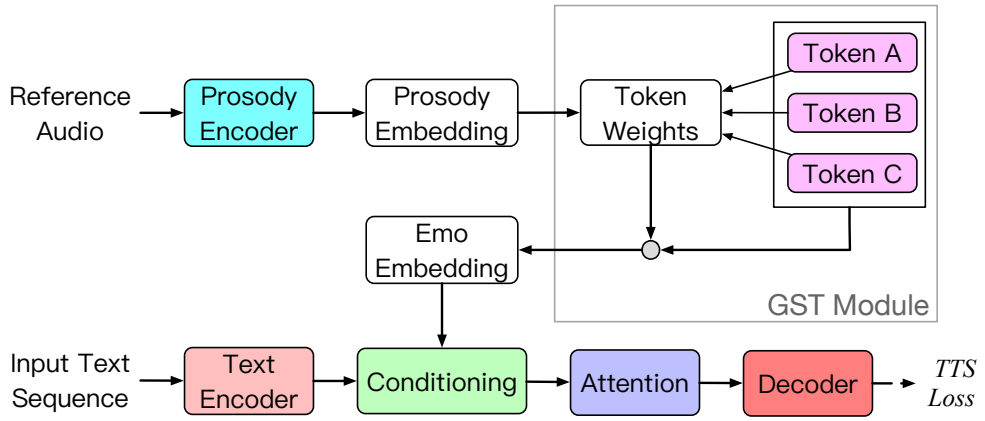


图 4.6　使用 GST 提取情感特征，并用于语音合成

Prosody Encoder 模块的网络结构包括四层卷积网络和一层循环网络，具体网络细节可以参考原始论文，这里不再赘述。

GST 模块包括一组可训练的 Style Token，通过计算 Token_i 与 Prosody Embedding 之间的相似度，得到其对应的权重 w_i，相似度的计算方法见公式（4-1）。

$$w_i = v \cdot tanh(q + k_i + b) \tag{4-1}$$

其中 v、b 都是可训练的变量，k_i 指 Token_i，q 指 Prosody Embedding 或由 Prosody Embedding 转换得到的向量。

最终 Emotion Embedding 的值 e 按照公式（4-2 ）计算得到。

$$\begin{aligned} \alpha_i &= \frac{e^{w_i}}{\sum_j^k e^{w_j}} \\ e &= \sum_i^k \alpha_i k_i \end{aligned} \tag{4-2}$$

其中 k 表示 Token 的个数。因为 Emotion Embedding 是 Token 的加权平均，而 Token 是在整个训练集上训练得到的，过滤掉了数据中的文本信息而只保留了副语言信息，因此 Emotion Embedding 相对于 Prosody Embedding 是更有效的副语言表征向量。

这里，我们仿照 Vaswani 等人的做法[66]，把 Prosody Embedding 映射到 h 个子空间，得到 h 个子向量再分别和各个 Token 之间进行相似度，因此 w_i 有 $h \times k$ 个不同的取值情况。为了方便描述，我们把 w_i 所有可能的取值组成的向量称为 w。

本章实验中，Prosody Embedding 为 256 维的向量，Token 维 64 维定长的向量，h 的值设置为 4，Token 的个数 k 设置为 10。因此 w 为一个 40 维的向量。模型推断时，直接建立情感标签信息到 w 的映射（具体方法见4.3.2节），即可实现指定特定的情感合成对应情感的声音。

4.2.4　情感离散表示下的特征提取-2（引入情感预测损失函数）

通过跨域语音情感识别模型，我们识别出了合成数据的情感标签信息。为了在情感语音合成中充分利用情感标签信息，我们在 GST 模型中引入情感损失函数（Emotion Loss），以促进网络能够提取和情感关联度更大的特征表示。

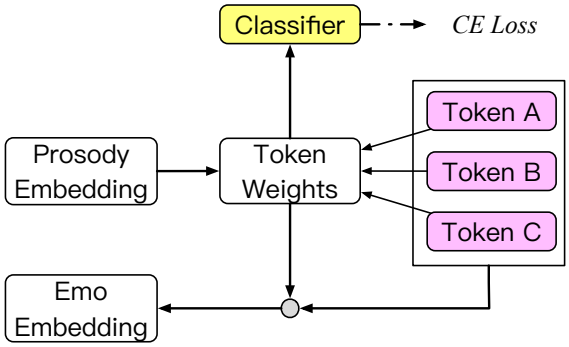


图 4.7　GST 模块中引入情感损失函数

引入情感损失函数的情感语音合成模型中，Text Encoder 模块、Decoder 模块

以及 Prosody Encoder 模块与 GST 模型相同，这里不再累述。在 GST 模块中，如图 4.7所示，我们增加了一个分类器（Classifier），接受 Token 的权重（Token Weights）作为输入，输出参考语音（Reference Audio）的对各类情感的后验概率，然后我们根据 Classifier 的输出和情感信息标签分布的差异，计算出情感损失函数。

我们用交叉熵衡量 Classifier 的输出和情感信息标签分布的差异，情感损失函数的计算公式见（4-3）。

$$L_{\text{emo}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} y_j^i log(\hat{y}_j^i) \tag{4-3}$$

其中，d 代表情感类别的个数。n 代表一个 batch 中样本的个数。$y \in \mathbb{R}^d$ 是情感的软标签，即跨域语音情感识别模型预测的参考语音对各个情感的后验概率，$\hat{y} \in \mathbb{R}^d$ 是本模型中分类器的输出值。y_j^i 代表第 i 个样本的情感软标签的第 j 个元素，$\hat{y}_j^i$ 代表第 i 个样本对应的分类器输出的第 j 个元素。

模型在训练时同时最小化 L_{TTS} 和 L_{emo}，总的损失函数为：

$$L = L_{\text{TTS}} + L_{\text{emo}} \tag{4-4}$$

4.2.5　情感二维空间表示下的特征提取

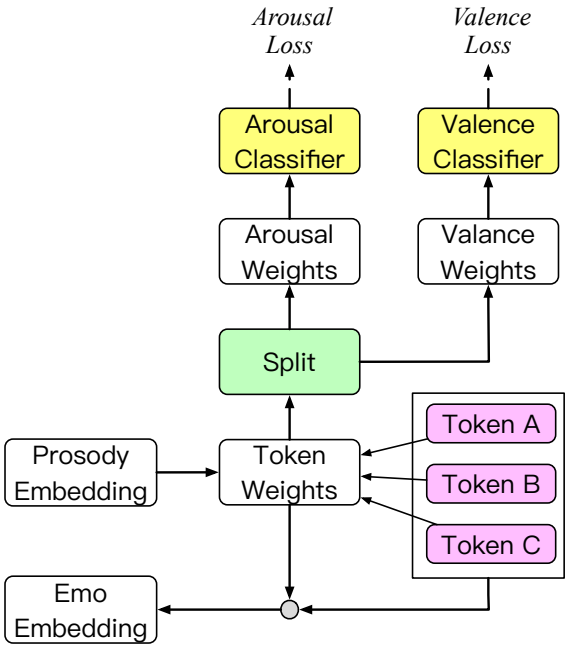


图 4.8　情感二维空间表示下的 GST 模块

本章中我们还提供了一种根据情感二维空间表示控制生成语音情感的方法。为了保证激活度和效价度相互独立地影响生成语音的情感，我们按照图 4.8设计模

型中的 GST 模块。

Token 的权重 w 是一个 $h \times k$ 的向量，我们将其分为两部分，其高 $\frac{h \times k}{2}$ 维称为激活度相关权重（Arousal Weights），低 $\frac{h \times k}{2}$ 维称为效价度相关权重（Valence Weights）。然后 Arousal Weight 通过激活度分类器（Arousal Classifier），会输出激活度为高或低的后验概率，根据对应的激活度标签信息，计算得到激活度损失函数 L_{arousal}。同理，Valence Weight 通过效价度分类器（Valence Classifier）后，可以得到效价度激活函数 L_{valence}。在 L_{arousal} 和 L_{valence} 的作用下，Arousal Weights 控制的是激活度相关的信息，Valence Weights 控制的是效价度相关的信息，最终实现激活度和效价度相互独立地控制生成语音的情感。在情感二维空间表示的情况下，模型总的损失函数为：

$$L = L_{\text{TTS}} + L_{\text{arousal}} + L_{\text{valence}} \tag{4-5}$$

4.3　情感合成方案

4.3.1　预测合成数据的情感标签

我们使用跨域的语音情感识别模型预测合成数据的情感标签信息。其模型如图4.9所示。模型由一个孪生的特征提取器和分类器组成。模型在训练时同时使用交叉熵损失函数（CE Loss）和最大平均差异损失函数（MMD Loss），保证特征提取器所提取的特征具有情感可区分度的同时，SER 特征和 TTS 特征之间的分布也能保持一致。这保证了 TTS 特征作为分类器的输入时，也能有稳定的情感识别性能。

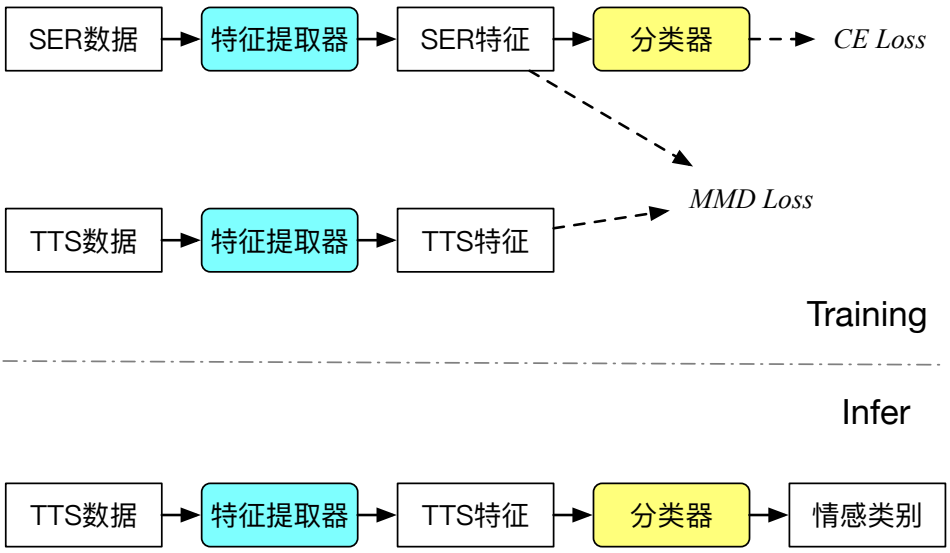


图 4.9　跨域语音情感识别模型（SER Corpus → TTS Corpus）

4.3.2　情感到特征的映射

模型在推断阶段，需要根据情感标签信息找到其对应的 Token 权重，以实现情感语音合成（如图 4.2）。因此，我们需要一种方法，建立情感标签到 Token 权重的映射。

当情感标签采用情感类别表示时，我们先建立基本情感到其对应权重的关系。这里阐述得到第 i 类情感对应的权重 w_i 的方法。跨域语音情感识别模型可以输出合成数据对第 i 个情感的后验概率 $\hat{y}_i$，找出整个合成数据集 $\hat{y}_i$ 最大的 K 条音频数据，而这 K 条音频对应的 Token 权重的均值，作为 w_i 的值。更进一步地，我们用 $\alpha w_i + (1-\alpha)w_j$ 表示同时具有第 i 类情感和第 j 类情感的特征（Token 权重）。该方法同样可以用于基准模型下建立从情感标签到特征的映射。本章实验中，情感标签采用情感类别表示的情况下，K 取 50。

对于情感标签采用 Arousal-Valence 的形式表示的情况。我们用 w^h 表示 w 的高 $\frac{h\times k}{2}$ 维特征，用 w^l 表示 w 的低 $\frac{h\times k}{2}$ 维特征，假设 Arousal, Valence 的取值范围都是 $[0,1]$。由于 w^h 控制语音的 Arousal 信息，我们令 w_0^h 代表 Arousal 值取 0 时的权重，w_1^h 代表 Arousal 值取 1 时的权重。同理，w_0^l 代表 Valence 值取 0 时的权重，w_1^l 代表 Arsoual 值取 1 时的权重。对于 w_0^h 的值，跨语言模型可以预测合成数据 Arousal 值取 0 或 1 的概率，找到整个合成数据集预测合成数据 Arosual 的值为 0 的后验概率最大的 K 个音频，求其 w^h 的平均值，作为 w_0^h 的值。同理可得 w_1^h、w_0^l、w_1^l 的值。因此当指定情感标签 Arousal=β_1、Valence=β_2 时，其对应的 $w^h = (1-\beta_1)w_0^h + \beta_1 w_1^h$、$w^l = (1-\beta_2)w_0^l + \beta_2 w_1^l$。然后 w^h 和 w^l 拼接起来，作为最终的 Token 权重，控制生成语音的情感。本章实验中，对于情感标签采用情感 2 维空间表示的情况，K 取 100。

4.4　实验与分析

4.4.1　数据集介绍

情感语料库使用 IEMOCAP 数据集[47]，关于 IEMOCAP 的介绍，以及在情感标签在情感类别表示下和在情感二维空间表示下数据的分布和处理，详见第2.3.1节和第3.3.1节，这里不再赘述。

合成语料库使用 Blizzard Challenge 2013 数据集[50]。该数据集是语音识别比赛 2013 届的数据，目前已经公开。它由一个说话人读电子书的数据组成，我们过滤掉文本长度小于 5 大于 90，以及语音时长小于 0.5s 大于 14s 的数据后，共有 95640 条音频数据。

4.4.2　实验设置

本章实验需要训练跨域语音情感识别模型和语音合成模型。对于跨域语音情感模型，数据处理、模型设置等完全参考第3章，这里不再赘述。

训练合成模型时需要提取线性谱和梅尔谱，提取频谱时相关参数如表 4.1所示，此外，在进行短时傅里叶变换变换之前，需要对每个音频的音量进行标准化以及按照系数 0.97 进行预加重。GST 模块中，情感分类器（Classifier）采用单层全连接实现，网络其余参数设置，完全参考 Tacoton 模型[32] 以及 GST 模型[42] 相关论文。

表 4.1　频谱计算的参数配置

参数名称	参数值
语音信号采样率	16kHz
窗函数	汉宁窗 (Hann windows)
窗长	50ms
窗移	12.5ms
DFT 频率维度	2048
梅尔滤波器个数	80

4.4.3　离散情感表示下情感语音合成主观评测结果

在离散情感表示下，我们分别使用第4.2.3节和第4.2.4节中的情感特征提取方法进行情感语音合成。两种方案下，分别合成 10 组语音，每组语音 5 个音频，其中四个音频对应于四种情感，另一个音频为两个基本情感的情感特征的插值作为条件输入，作为混淆项。我们请 15 位人员根据其主观感受判断所合成的语音的情感。

在第4.2.3节的方案下（情感离散表示下的特征提取-1），每个情感类别下的合成语音其主观标注下的情感分布如图 4.10所示。其中对于指定合成为中性的语音,64.67% 被标注合成为中性；对于指定合成为愤怒的声音，52% 被标注为愤怒；对于指定为高兴的声音，38.67% 被标注为高兴；对于指定为悲伤的声音，54% 被标注为悲伤。标注情感和指定情感平均一致率为 (64.67% + 52% + 38.67% + 54%) / 4 = 52.34%。

对于第4.2.4节中的方案（情感离散表示下的特征提取-2），其主观评测结果如图4.10所示。该方案中，指定合成为中性的语音中 46.67% 被标注合成为中性；指定合成为愤怒的声音中 51.33% 被标注为愤怒；对于指定为高兴的声音中 50.67% 被标注为高兴；对于指定为悲伤的声音中 44.67% 被标注为悲伤。标注情感和指定

情感平均一致率为 48.34%。

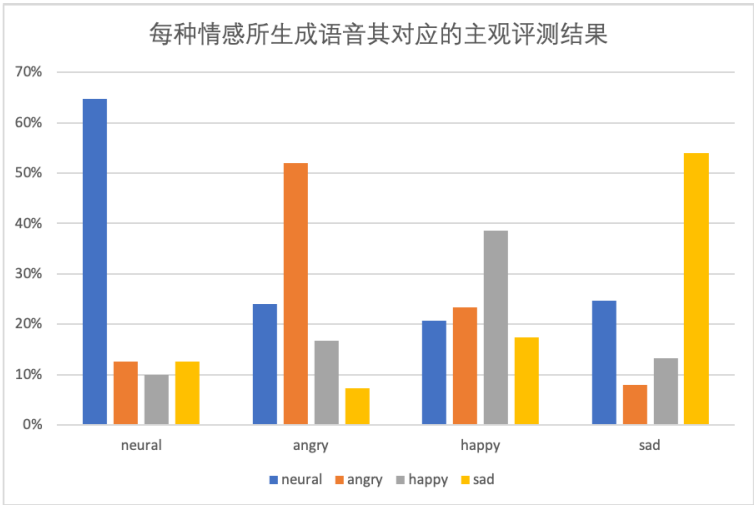


图 4.10　离散情感表示下基于 GST 的情感语音合成主观评测结果（不使用 emo loss）

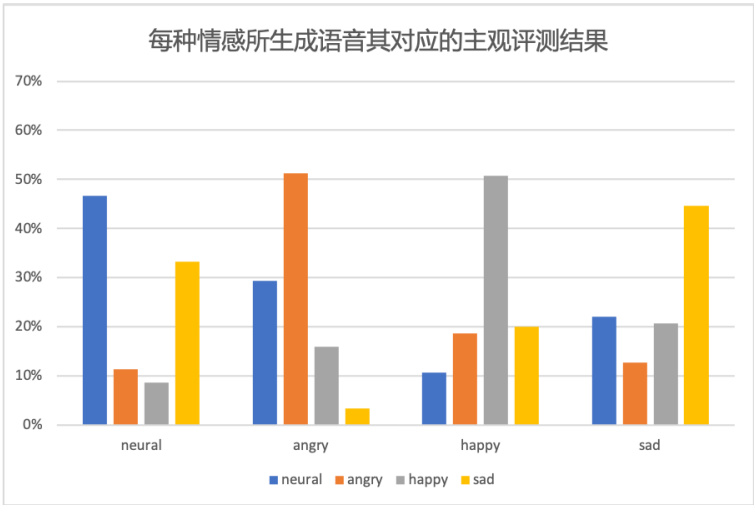


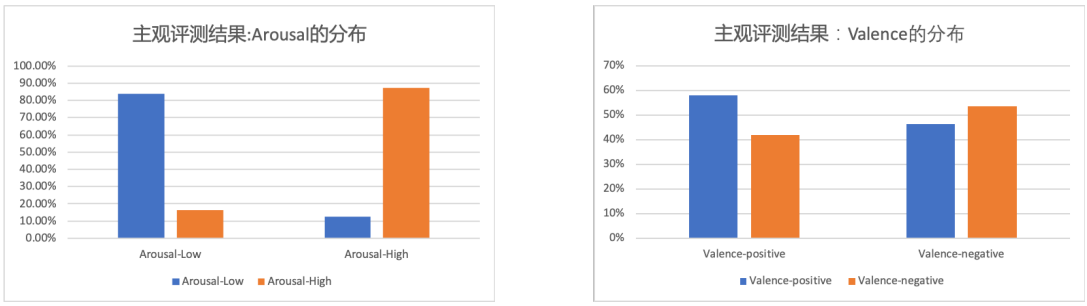
图 4.11　离散情感表示下基于 GST 的情感语音合成主观评测结果（使用 emo loss）

比较图4.10和4.11的主观评测结果可以发现，在 GST 模块中引入情感预测损失函数（emo loss）并没有带来情感语音合成效果的提升，全局风格令牌的机制已经能够保证模型从参考语音中学习到比较好的情感特征表示。此外，四种情感中，“高兴”被预测准确的概率最低，这受 IEMOCAP 数据集的影响，和第2章中的实验结果相一致。

4.4.4　情感空间表示下情感语音合成主观评测结果

按照第4.2.5节方案，我们实现了同时指定效价度和激活度信息实现情感语音合成。我们生成了 10 组音频，每组包含四个音频，其效价度和激活度分别是（0.2,

0.2)、(0.2, 0.8)、(0.8, 0.2) 和 (0.8, 0.8) 四种不同的组合，请 15 位人士对所生成的语音的效价度和激活度进行评价，得到结果如图4.12所示。



(a)　　(b)

图 4.12　情感空间表示下基于 GST 的情感语音合成主观评测结果，(a) 激活度，(b) 效价度

由图4.12可知，我们的评测结果中，对于指定激活度为 0.2 的音频，有 83.67% 被认为激活度是低的，对于指定激活度为 0.8 的音频，有 87.33% 被认为激活度是高的；指定效价度为 0.2 的音频中有 58% 被认为它是消极的，指定效价度为 0.8 的音频中有 53.67% 被认为它是积极的。根据主观评测结果可知，对于生成语音的情感的二维控制，激活度比效价度更为明显，这也和第3章中的实验结果相一致。

4.5　本章小结

随着语音技术的发展，语音合成的自然度已经达到了可以商用的标准。在此基础上，人们希望能够进一步控制生成语音的情感、音色等信息。本章提出了一种基于 GST 的情感语音合成方案，该方案利用跨域的语音情感识别模型识别合成数据的情感标签，进一步建立情感标签信息到风格令牌权重的映射以实现情感可控的语音合成。此外模型中引入情感预测损失函数，把风格令牌分成两组，分别对影响模型的效价度和激活度。实验结果表明，在情感类别表示下，本章的方案对合成语音的情感实现了较好的控制效果，同时也实现了根据效价度和激活度控制生成语音的情感。

第 5 章　总结

5.1　研究工作总结

情感是语音中的一个重要的因素，从语音中提取有效的情感特征表示有利于提高语音交互产品的智能度。本文研究课题为语音中的情感特征表示学习，研究内容包括从语音中提取有效的情感特征表示以用于语音情感识别任务和情感语音合成任务。研究工作主要包括以下三个方面：

一、提出了基于中心损失函数的端到端语音情感识别模型，提取更具有情感区分度的特征以提升模型情感识别的性能。由于情感本身是比较主观的感受，利用神经网络很难从语音中提取具有情感区分度的特征。我们在端到端的语音情感识别模型中引入中心损失函数，端到端语音情感识别模型使用交叉熵损失函数，保证提取的特征是可分离的，而中心损失函数提取的情感特征到其类中心点距离更小。在交叉熵损失函数和中心损失函数的联合作用下，模型所提取的特征具有更好的情感区分度，语音情感识别任务的性能得到提升。在 IEMOCAP 数据集[47] 上的实验表明，引入中心损失函数后，模型的非加权正确率和加权正确率均得到了明显的提升。

二、提出了基于最大平均差异损失函数的端到端跨域语音情感识别模型，保证提取的源域特征和目标域特征分布一致以实现稳定的跨域语音情感识别性能。语音情感识别模型在实际应用场景中，往往遇到测试数据和训练数据分布不一致的情况，我们称训练数据为源域数据，测试数据为目标域数据。为了保证模型在目标域数据上有稳定的语音情感识别性能，我们在端到端语音情感识别模型中引入最大平均差异损失函数，用于最小化源域特征和目标域特征分布之间的差异。在交叉熵和最大平均差异损失函数的作用下，模型提取出情感相关域无关的情感特征，进而保证模型在目标域数据下也能取得稳定的情感识别效果。使用本文的模型在 IEMCAP 数据集[47] 和 RECOLA 数据集[49] 上进行跨语言语音情感识别实验，结果表明引入最大平均差异损失函数后，跨语言语音情感识别任务的正确率得到显著提升。

三、提出了基于全局风格令牌的情感语音合成方案，实现了在无情感监督信息的开源合成数据下，训练出情感可控制的语音合成模型。随着语音技术的发展，文语转换技术产生的声音已经能够和真人相媲美。为了让语音合成产品更智能，人们希望可以对合成语音的情感进行控制。为了解决情感语音合成的需求，本文提出了一个基于全局风格令牌的情感语音合成方案，该方案中我们先使用跨域语音

情感识别模型预测合成数据的情感软标签，然后建立了情感标签到风格令牌权重的映射，最终实现了对合成语音情感的控制。我们以 IEMCAP 数据集[47] 作为源域数据，Blizzard Challenge 2013 数据集[50] 作为目标域数据训练跨域语音情感识别模型；并在 Blizzard Challenge 2013 数据集上训练情感合成模型，最终实现了情感可控制的语音合成。

5.2　未来工作展望

本文主要研究如何利用神经网络从语音中提取有效的特征以提升语音情感识别的性能，并将语音情感识别所提取的特征用于语音合成中，实现了情感语音合成。在现有工作的基础上，仍然有一些问题可以继续研究。

神经网络能够从大量数据中自动提取特征，以实现特定的任务。如果能充分利用容易获取的海量无监督的数据提取特征，更有利于发挥神经网络的优势。谷歌于 2018 年所提出的 BERT 模型[67]，正是基于该思想，BERT 模型从海量无监督的文本数据中提取有效的上下文语义特征，并将所提取特征用于特定的自然语言处理任务中，刷新了多个自然语言处理的任务。类似地，如何巧妙地设计模型从容易获取的无监督语音数据中提取有效的特征表示，使语音情感识别的性能不再受限于数据量有限的情感语料库，是一个非常值得研究的问题。

另外，随着语音合成技术的发展，相关产品越来越多，人们希望对合成的语音能有更多的控制能力。除了对情感的控制外，也有对生成语音音色进行控制的需求。其中，说话人转换[68]、声音克隆[69] 是语音合成领域中很热门也很有趣的研究方向。如何将语音中的音色、情感等副语言信息进行解耦，以实现对合成语音进行在音色、情感等多个维度进行相互独立的控制也是一个值得研究的问题。

参考文献

[1] Schacter D, Gilbert D, Wegner D, et al. Psychology: European edition[M]. London, United Kingdom: Macmillan International Higher Education, 2011.

[2] Handel S. Classification of emotions[J]. http://www.theemotionmachine.com/classi-lication-of-emotions, 2012.

[3] Osgood C E, Suci G J, Tannenbaum P H. The measurement of meaning: number 47[M]. Champaign, IL, USA: University of Illinois press, 1957.

[4] 韩文静, 李海峰, 阮华斌, 等. 语音情感识别研究进展综述[J]. 软件学报, 2014, 25(1):37-50.

[5] Lin Y L, Wei G. Speech emotion recognition based on hmm and svm[C]//International Conference on Machine Learning and Cybernetics. Guangzhou, China: IEEE, 2005: 4898-4901.

[6] Pao T L, Liao W Y, Chen Y T. A weighted discrete knn method for mandarin speech and emotion recognition[J]. Speech Recognition, 2008:411.

[7] Ververidis D, Kotropoulos C. Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition[J]. Signal processing, 2008, 88(12): 2956-2970.

[8] Mishra H K, Sekhar C C. Variational gaussian mixture models for speech emotion recognition[C]//Seventh International Conference on Advances in Pattern Recognition. Kolkata, India: IEEE, 2009: 183-186.

[9] Mao Q, Wang X, Zhan Y. Speech emotion recognition method based on improved decision tree and layered feature selection[J]. International Journal of Humanoid Robotics, 2010, 7(02):245-261.

[10] Goodfellow I, Bengio Y, Courville A. Deep learning[M]. Cambridge, MA, USA: MIT press, 2016.

[11] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: theory and applications[J]. Neurocomputing, 2006, 70(1-3):489-501.

[12] Han K, Yu D, Tashev I. Speech emotion recognition using deep neural network and extreme learning machine[C]//Fifteenth annual conference of the international speech communication association. Singapore: ISCA, 2014: 223-227.

[13] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[14] Lee J, Tashev I. High-level feature representation using recurrent neural network for speech emotion recognition[C]//Sixteenth annual conference of the international speech communication association. Dresden, Germany: ISCA, 2015: 1537-1540.

[15] Trigeorgis G, Ringeval F, Brueckner R, et al. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network[C]//IEEE international conference on acoustics, speech and signal processing. Shanghai, China: IEEE, 2016: 5200-5204.

[16] Satt A, Rozenberg S, Hoory R. Efficient emotion recognition from speech using deep learning on spectrograms.[C]//Eighteenth annual conference of the international speech communication association. Stockholm, Sweden: ISCA, 2017: 1089-1093.

[17] Mower E, Metallinou A, Lee C C, et al. Interpreting ambiguous emotional expressions[C]//Third International Conference on Affective Computing and Intelligent Interaction and Workshops. Amsterdam, The Netherlands: IEEE, 2009: 1-8.

[18] Bellet A, Habrard A, Sebban M. Metric learning[J]. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2015, 9(1):1-151.

[19] Lian Z, Li Y, Tao J, et al. A pairwise discriminative task for speech emotion recognition[J]. arXiv preprint arXiv:1801.01237, 2018.

[20] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Boston, MA, USA: IEEE, 2015: 815-823.

[21] Huang J, Li Y, Tao J, et al. Speech emotion recognition from variable-length inputs with triplet loss function.[C]//Nineteenth annual conference of the international speech communication association. Hyderabad, India: ISCA, 2018: 3673-3677.

[22] Bromley J, Guyon I, LeCun Y, et al. Signature verification using a siamese time delay neural network[C]//Advances in neural information processing systems. Denver, Colorado, USA: Curran Associates, 1994: 737-744.

[23] Neumann M, et al. Cross-lingual and multilingual speech emotion recognition on english and french[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 5769-5773.

[24] Hunt A J, Black A W. Unit selection in a concatenative speech synthesis system using a large speech database[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. Atlanta, Georgia,USA: IEEE, 1996: 373-376.

[25] Tokuda K, Yoshimura T, Masuko T, et al. Speech parameter generation algorithms for hmm-based speech synthesis[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing. Istanbul, Turkey: IEEE, 2000: 1315-1318.

[26] Ze H, Senior A, Schuster M. Statistical parametric speech synthesis using deep neural networks [C]//IEEE International Conference on Acoustics, Speech, and Signal Processing. Vancouver, BC, Canada: IEEE, 2013: 7962-7966.

[27] Lu H, King S, Watts O. Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis[C]//Eighth ISCA Workshop on Speech Synthesis. Barcelona, Spain: ISCA, 2013: 261-265.

[28] Qian Y, Fan Y, Hu W, et al. On the training aspects of deep neural network (dnn) for parametric tts synthesis[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: IEEE, 2014: 3829-3833.

[29] Wu Z, Valentini-Botinhao C, Watts O, et al. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. South Brisbane, Queensland, Australia: IEEE, 2015: 4460-4464.

[30] Hashimoto K, Oura K, Nankaku Y, et al. The effect of neural networks in statistical parametric speech synthesis[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. South Brisbane, Queensland, Australia: IEEE, 2015: 4455-4459.

[31] Sotelo J, Mehri S, Kumar K, et al. Char2wav: End-to-end speech synthesis[C]//International Conference on Learning Representations. Toulon, France: OpenReview.net, 2017.

[32] Wang Y, Skerry-Ryan R, Stanton D, et al. Tacotron: Towards end-to-end speech synthesis[C]// Annual Conference of the International Speech Communication Association. Stockholm, Sweden: ISCA, 2017.

[33] Ping W, Peng K, Gibiansky A, et al. Deep voice 3: Scaling text-to-speech with convolutional sequence learning[C]//International Conference on Learning Representations. Vancouver, BC, Canada: OpenReview.net, 2018.

[34] Shen J, Pang R, Weiss R J, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, AB, Canada: IEEE, 2018: 4779-4783.

[35] Li N, Liu S, Liu Y, et al. Neural speech synthesis with transformer network[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 33. Honolulu, Hawaii, USA: AAAI, 2019: 6706-6713.

[36] Morise M, Yokomori F, Ozawa K. World: a vocoder-based high-quality speech synthesis system for real-time applications[J]. IEICE TRANSACTIONS on Information and Systems, 2016, 99 (7):1877-1884.

[37] Kawahara H. Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds[J]. Acoustical science and technology, 2006, 27(6):349-353.

[38] Oord A v d, Dieleman S, Zen H, et al. Wavenet: A generative model for raw audio[J]. arXiv preprint arXiv:1609.03499, 2016.

[39] Valin J M, Skoglund J. Lpcnet: Improving neural speech synthesis through linear prediction[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, United Kingdom: IEEE, 2019: 5891-5895.

[40] Li H, Kang Y, Wang Z. Emphasis: An emotional phoneme-based acoustic model for speech synthesis system[J]. arXiv preprint arXiv:1806.09276, 2018.

[41] Skerry-Ryan R, Battenberg E, Xiao Y, et al. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron[J]. arXiv preprint arXiv:1803.09047, 2018.

[42] Wang Y, Stanton D, Zhang Y, et al. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis[J]. arXiv preprint arXiv:1803.09017, 2018.

[43] Inoue K, Hara S, Abe M, et al. An investigation to transplant emotional expressions in dnn-based tts synthesis[C]//Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Kuala Lumpur, Malaysia: IEEE, 2017: 1253-1258.

[44] Lee Y, Rabiee A, Lee S Y. Emotional end-to-end neural speech synthesizer[J]. arXiv preprint arXiv:1711.05447, 2017.

[45] Hodari Z, Watts O, Ronanki S, et al. Learning interpretable control dimensions for speech synthesis by using external data.[C]//Nineteenth Annual Conference of the International Speech Communication Association. Hyderabad, India: ISCA, 2018: 32-36.

[46] Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition [C]//European conference on computer vision. Amsterdam, The Netherlands,: Springer, 2016: 499-515.

[47] Busso C, Bulut M, Lee C C, et al. Iemocap: Interactive emotional dyadic motion capture database [J]. Language resources and evaluation, 2008, 42(4):335.

[48] Borgwardt K M, Gretton A, Rasch M J, et al. Integrating structured biological data by kernel maximum mean discrepancy[J]. Bioinformatics, 2006, 22(14):e49-e57.

[49] Ringeval F, Sonderegger A, Sauer J, et al. Introducing the recola multimodal corpus of remote collaborative and affective interactions[C]//IEEE international conference and workshops on automatic face and gesture recognition. Shanghai, China: IEEE, 2013: 1-8.

[50] King S, Karaiskos V. The blizzard challenge 2013[J]. 2013.

[51] El Ayadi M, Kamel M S, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases[J]. Pattern Recognition, 2011, 44(3):572-587.

[52] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8):1798-1828.

[53] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553):436-444.

[54] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv:1409.1556, 2014.

[55] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: IEEE, 2016: 770-778.

[56] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[57] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]//IEEE international conference on computer vision. Santiago, Chile: IEEE, 2015: 1026-1034.

[58] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[59] Wold S, Esbensen K, Geladi P. Principal component analysis[J]. Chemometrics and intelligent laboratory systems, 1987, 2(1-3):37-52.

[60] Long M, Cao Y, Wang J, et al. Learning transferable features with deep adaptation networks[J]. arXiv preprint arXiv:1502.02791, 2015.

[61] Maaten L v d, Hinton G. Visualizing data using t-sne[J]. Journal of machine learning research, 2008, 9(Nov):2579-2605.

[62] Zhang Y J, Pan S, He L, et al. Learning latent representations for style control and transfer in end-to-end speech synthesis[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom: IEEE, 2019: 6945-6949.

[63] Srivastava R K, Greff K, Schmidhuber J. Highway networks[J]. arXiv preprint arXiv:1505.00387, 2015.

[64] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[65] Griffin D, Lim J. Signal estimation from modified short-time fourier transform[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1984, 32(2):236-243.

[66] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. Long Beach, CA, USA: Curran Associates, 2017: 5998-6008.

[67] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[68] Sun L, Li K, Wang H, et al. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training[C]//IEEE International Conference on Multimedia and Expo (ICME). Seattle, WA, USA: IEEE, 2016: 1-6.

[69] Arik S, Chen J, Peng K, et al. Neural voice cloning with a few samples[C]//Advances in Neural Information Processing Systems. Montréal, Canada: Curran Associates, 2018: 10040-10050.

致　谢

衷心感谢导师吴志勇老师和贾珈老师对我的悉心指导，也感谢腾讯 AI Lab 康世胤博士为我提供的帮助。

感谢实验室的同学们对我入学以来提供的帮助，尤其是与钟括同学在度量学习、迁移学习方面的交流给我的工作带来了很大的启发。

感谢清华大学深圳研究生院为我提供了舒适的学习和生活环境，让我度过了难忘的三年时光。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1994 年 5 月 17 日出生于河南省太康县。

2017 年 7 月同济大学本科毕业并获得软件工程专业学士学位。

2017 年 9 月进入清华大学计算机系攻读工程硕士学位至今。

发表的学术论文

[1] **Dai D**, Wu Z, Li R, Wu X, Jia J, and Meng H. Learning discriminative features from spectrograms using center loss for speech emotion recognition[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, United Kingdom: IEEE, 2019: 7405-7409.(Regular Paper, EI:20193007228731,CCF-B)

[2] **Dai D**, Wu Z, Kang S, Wu X, Jia J, Su D, Yu D and Meng H. Disambiguation of chinese polyphones in an end-to-end framework with semantic features extracted by pre-trained[C]//Twentieth Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA, 2019: 2090-2094.(Regular Paper, EI:20194607674520,CCF-C)

[3] Wu X, Liu S, Cao Y, Li X, Yu J, **Dai D**, Ma X, Hu S, Wu Z, Liu X and Meng H. Speech emotion recognition using capsule networks[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, United Kingdom: IEEE, 2019: 6695-6699. (Regular Paper, EI:20192907201454,CCF-B)

[4] Lu H, Wu Z, **Dai D**, Li R, Kang S, Jia J and Meng H. One-shot voice conversion with global speaker embeddings[C]//Twentieth Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA, 2019: 669-673. (Regular Paper, EI:20194607674295,CCF-C)

专利

[1] 吴志勇, **代东洋**, 康世胤, 苏丹, 俞栋. 确定多音字发音的方法及装置 (申请中, CN110277085A).

[2] 吴志勇, **代东洋**. 一种基于对抗学习的端到端的跨语言语音情感识别方法 (申请中, CN110364186A).

参加的科研项目

[1] 国家自然科学基金香港政府研究资助局（NSFC-RGC）合作项目：面向互联网口语对话的交互属性挖掘与特色语音生成的研究（资助号：61531166002、N_CUHK404/15）

[2] 国家自然科学基金重点项目：互联网话语理解的心理机制与计算建模（资助号：61433018）

[3] 国家社会科学基金重大项目：社会情感的语音生成与认知的跨语言跨文化研究（资助号：13&ZD189）

[4] 腾讯 AI Lab 犀牛鸟联合研究项目：普通话端到端语音合成技术研究（资助号：JR201803）