北京交通大学

博 士 学 位 论 文

面向情感语音合成的言语情感建模研究

The Modeling Research for Speech Emotion towards Expressive Speech Synthesis

作者：高莹莹

导师：朱维彬

北京交通大学

2019 年 7 月 ~~2015 年 12 月~~

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通大学可以将学位论文的全部或部分内容编入有关数据库进行检索，提供阅览服务，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。学校可以为存在馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：　　导师签名：

签字日期：　　年　月　日　　签字日期：　　年　月　日

学校代码：10004 密级：公开

北京交通大学

博士学位论文

面向情感语音合成的言语情感建模研究

The Modeling Research for Speech Emotion towards Expressive Speech Synthesis

作者姓名：高莹莹 学 号：10112060

导师姓名：朱维彬 职 称：副教授

学位类别：工学 学位级别：博士

学科专业：信号与信息处理 研究方向：情感语音合成

北京交通大学

2019 年 7 月 ~~2015 年 12 月~~

致谢

本论文的工作是在朱维彬老师的的悉心指导下完成的，从本科阶段开始跟朱老师做项目，每周一次的一对一辅导，朱老师从最基本的语音学概念开始，讲到时下最热门的算法技术，为我打开进入语音学世界的大门。一直以来，朱老师治学都很严谨，对学生的要求也很高，让我学习到了对待学术所该具备的一丝不苟和精益求精。朱老师敏锐的洞察力和富有前瞻性的指导，为我指明了研究的方向；在我迷茫彷徨之际，又像朋友亲人一样给予我鼓励，帮我重拾信心。在此，衷心感谢老师多年以来对我的辛勤栽培与付出。

本文写作期间，我的共同导师肖扬教授永远离开了我们，借此机会向他致以最诚挚的怀念和谢意。肖老师在重病期间，仍坚持指导学生论文和答辩工作，用切身实践诠释了教师这一职业的无私与伟大。

感谢徐金安老师在学术论文撰写期间对我的指导和帮助。感谢梁满贵教授、阮秋琦教授、裘正定教授、赵瑞珍教授等几位老师在开题答辩以及预答辩时对我的耐心指导。感谢中国传媒大学侯敏教授、何伟老师、邹煜老师和赵俐老师对我们数据采集与研究工作的全力配合和热情帮助。感谢中科院心理所杨玉芳研究员、杨晓虹助理研究员对我们工作提出的宝贵意见和建议。感谢社科院李爱军研究员、贾媛助理研究员对我们工作的帮助和支持。

感谢实验室的师姐、师妹和几位师弟在工作和学习上对我的帮助与支持。感谢同班张帅、安文娟、刘帅奇等几位同学对我提供的帮助和建议，特别感谢我的舍友李玲同学在生活和学习上的相互鼓励和支持。还要感谢那些曾经参与我们数据采集实验的学弟学妹们对我们工作的大力支持与配合。

感谢我的家人和朋友。感谢我的父母，在我漫长的求学生涯中始终给予我最无私的爱和支持。感谢我的姐姐，在我在外求学的十年里照顾父母，让我能够安心学业。感谢我的男友及其家人对我无微不至的关怀和照顾，让我在这个陌生的城市感受到家的温暖。感谢我的几位挚友，在我好与不好的时候，感谢有你们相伴。

最后，感谢母校十年来对我的培养。交大十年，母校给予我太多太多，离别之际，对母校的感情溢于言表，衷心祝愿母校越来越好！

摘要

语音作为人类重要的交际工具之一，除传递字面信息，还通过语气的变化传递情感。当前情感语音合成的研究，主要集中于某些特定情感状态与语音信号关联关系的探寻，虽然观察到一些情感与声学参数变化相关联的指向性线索，但由于情感表现的多样性和复杂性，导致了情感声学参数的数值分布多呈现较大的离散特性。构造一个堪用的情感语音合成系统，仍是一件未完成的任务。

论文~~我们希望~~从认知科学的角度出发，~~解析出与~~对语气变化相关的情感~~的~~发生及衍化过程进行分析，进而~~构建~~提出情感信息~~的自动~~预测模型~~，为研究情感与语音信号错综复杂的关联关系提供有针对性指引~~。言语情感建模主要涉及以下~~需要解决的~~问题~~主要有~~：1）相关的理论需要有所升华，尤其是要解决情感的准确刻画和动态衍化过程的描述；2）建模技术需要有所突破，考虑到影响情感因素及情感生成过程的复杂性，所需处理的特征参数可能会来自多个层面，模型应能支持多尺度特征处理及动态衍化过程刻画。

针对第一个问题，~~我们~~论文在心理学、朗读学、播音学与语音学等相关理论和实践指导下，采用心理语言学、感知语音学实验和数据分析相结合的方法，探索汉语朗读或播音等创作型有声语言活动中情感表达与言语特征间的关联关系，进而对言语情感生成及衍化机制进行归纳。以此为基础，提出多视角情感描述方案，分别从认知评价、心理感受、生理反应和发音描述四种视角描述言语情感的不同侧面，各视角互为补充，共同构成言语情感的分布式表达。各视角之间依据言语情感生成过程形成直接或间接的衍化关系，前面步骤的结果会影响后续步骤的反应，后续步骤的反应又会反馈回去进一步影响前面成分的变化。发音描述作为言语情感生成过程的最终输出结果，形成连接情感特征与声学特征的接口，有助于发现二者之间更为显性的映射关系。基于该描述方案，构建了一个新闻言语情感数据库，通过言语情感标注的实施以及后续预测模型的建立验证了言语情感生成过程及描述方案的合理性。

针对第二个问题，采用~~基于~~深度神经网络构建情感预测模型~~，~~一方面由于深度网络的多层非线性映射结构与多视角描述模型的多层分布式结构一致，另一方面便于模型实现对情感动态衍化过程以及多尺度特征关联关系的建模。具体来说，暂不考虑文本内容之外的影响因素，利用主题模型提取文本的语义空间向量表示，依次预测篇章级、段落级和句子级不同尺度的情感信息。各尺度内部，形成由认知到心理、生理再到发音的衍化关系，发音描述作为最终目标，其他成分作为其子目标，子目标依次作为后续预测目标的部分已知信息参与到后续模块的训练；不同尺度之间，构成由上至下的层级结构，大尺度单元的预测结果作为小尺度单元的部分

已知信息参与到小尺度单元的预测，为其提供更为全局的上下文参考。最后通过实验验证了所提方法的有效性，加入情感衍化关系以及多尺度特征间关联关系的影响，使模型最终预测结果的召回率、精准率和 F1 值分别相对提升了 31.8%、10.3%和 22.8%。

本文工作的主要创新点在于：

（1）基于言语情感生成过程的分析归纳，提出多视角情感描述模型：模型细致刻画了言语情感生成过程中各成分的变化及之间的衍化关系，并以发音描述作为连接情感与语音的接口，用于指导后续声学参数的调整；

（2）基于深度神经网络，构建文本-情感计算模型：模型综合考虑了言语情感生成过程中来自不同尺度特征的影响以及不同情感成分间的衍化关系，支持多尺度特征融合以及动态衍化关系刻画；

（3）将先验知识引入深度神经网络，实现网络中间结构的部分可见化：通过网络结构的直接显性设定，有效利用了言语情感生成的先验知识，降低了训练数据与网络规模的开销，预测性能亦有所提升。

关键词：情感语音合成；情感生成；情感描述；文本情感预测；深度神经网络；中间层可见化；

批注 [w1]: 请统一说法，正文中的、结论中的

设置了格式：突出显示

设置了格式：突出显示

ABSTRACT

Speech is one of the important human communication tools, in addition to passing the literal information, also expressing emotions through the changes of voice. The current research of emotion speech synthesis is mainly focused on the exploration of the relationship between some specific emotion state and the variations of speech signal. Some directive clues have been observed referring to the relevance between emotions and the changes of acoustic parameters. But because of the diversity and complexity of emotional expression, the numerical distribution of emotional acoustic parameters is more likely to be discrete. The construction of an available expressive speech system still has a lot of work to do.

Starting from the perspective of cognitive science, we hope to parse out the relevant mechanism about how the emotions generate and develop, further to build an automatic predicting model of emotional information, in order to provide targeted guidance for the study on the perplexing relationship between emotion and speech signal. There are two major problems to be solved: 1) related theories need to be sublimated, especially to depict the emotion accurately and describe the process of dynamic evolution; 2) modeling technology needs a breakthrough, which should be able to support multi-scale processing and dynamic evolution characterization, considering the complexity of the emotional factors during the emotion generating processes and the required parameters may be from multiple levels.

For the first proplem, under the guidance of the theory and practice of psychology, reading science, broadcasting science and phonetics, we use the method combining psycholinguistics, perceptual phonetics experiments and data analysis to explore the correlation between emotional expression and speech features during the producing process of acted speech like reading or broadcasting, further to summarize a theoretical mechanism about the generation and development of speech emotion and to guide the establishment of emotion describing model and predicting model. Based on this, an emotion describing model is proposed, explaining different aspects about speech emotion from different perspectives, including cognitive appraisal, psychological feeling, physical response and utterance manner. The four perspectives complement each other, forming a distributed representation for speech emotion. Each perspective has a direct or indirect effect with the others in line with the producing mechanism of speech emotion, which means that the outcomes of former steps will influence the reactions of subsequent steps,

and the reactions of subsequent steps will feed back to affect the changes of former steps. As the final output of the producing process of speech emotion, the utterance manner turns into an interface between emotional features and acoustic features, which is helpful to find a more obvious mapping relationship between them. On the basis of the describing scheme, a news speech emotion database is constructed and annotated manually, which verifies the rationality of the proposed theoretical system and describing model together with the following predicting experiments.

Towards the second problem, an emotion predicting model is built based on deep neural network, since the multi-layered nonlinear mapping structure of the network is consistent with the distribute structure of the multi-perspective describing model; on the other hand, it is easy to realize the modeling of the dynamic varying process of emotion and the correlations between features from different scales. In particular, excluding the effects beyond textual content, the topic model is adopted to extract a vector representation for texts in the semantic space. The emotional information on the document level, the paragraph level and the sentence level is predicted sequentially. Inside each level, a continuous process is formed from cognition to psychology then to physiology and utterance. The utterance is treated as the ultimate target and the other components are seen as sub targets. The sub targets participate in the training of their subsequent steps as part of known information. Among different levels, a top-down hierarchical structure is formed. The predicted results on higher levels take part in the prediction on lower levels as part of known information, providing a more global contextual reference. Finally the effectiveness of the predicting model is validated through experiments, which shows that the appending of the influence of sequential relationship inside an emotion epoch and the interrelationship among features from different scales improves the recall, precision and F1-value of the utterance manner prediction by 31.8%, 10.3% and 22.8% respectively.

The main innovation points of this paper are:

(1) Provide a multi-perspective emotion description scheme from the generation of emotion in speech, which gives detailed describing methods for the various factors affecting the generation of emotional speech, and supplies a comprehensive detailed reference for the study of the relationship between emotion and speech;

(2) Build a text-based emotion computing model based on deep neural networks, which considers the impacts of the factors from different aspects and in different levels during the generating process of speech emotion, and supports the modeling of dynamic derivative relationship and the processing of multi-scale features;

(3) Introduce priori knowledge to deep neural network to realize the supervision and guidance of the network structure, which improves the performance of the network, reduces the cost of training data and network scale, and has good scalability for the other types of features.

KEYWORDS： Expressive speech synthesis; emotion production; emotion description; text-based emotion prediction; deep neural network; visible intermediate layers

目录

1 引言

1.1 问题描述

随着人机“对话”的日趋频繁，人机交互接口技术已经从机械化时代跨入多媒体用户界面时代，人们不再满足于仅仅通过键盘、鼠标、触摸屏或手写笔等传统方式同机器交流，而是希望有一种类似于人与人之间的交流方式。语音是人类交际的重要工具，让计算机具备像人一样的“听”和“说”的能力，是人们长期追求的目标，而语音合成技术是实现这一目标的关键技术之一。当前，语音合成技术在自然语言处理、信号处理和随机过程处理等方法的推动下获得很大发展[1]，合成语音的清晰度、可懂度已基本达到要求，自然度也有较好表现，如何进一步提升合成语音的表现力，赋予其丰富的语气和感情色彩，成为语音合成领域新的亟待解决的问题，这也催生了语音合成领域一个重要分支——情感语音合成的发展。

在人与人的交流过程中，除了语言、脸像和行为所表达的直接的语义信息外，人类的情感也传递了重要的信息。对人类情感机理的研究与探索一直是科学研究的重要方向，人类的智能不仅表现为正常的理性思维和逻辑推理能力，也应表现为正常的情感能力。情感能力对于计算机与人的自然交往至关重要，在与计算机交互的过程中，人们希望机器能理解自己的需要和感受，并做出适当的反映，而如果缺乏这种情感理解和表达能力，就很难指望计算机具有类似人一样的智能，也很难期望人机交互做到真正的和谐与自然。情感计算（Affective Computting）就是研究如何赋予计算机类似于人一样的观察、理解和生成各种情感特征的能力，最终使计算机像人一样能进行自然、亲切和生动的交互[2]。

情感语音合成是情感计算与语音合成相结合的产物，借助情感计算的概念，从携带已知情感状态的语音信号中分析情感与语音的关联，并将这些情感特征运用于语音合成过程，从而获得具有丰富语气变化、能够模拟人类情感的自然、友好的合成语音。目前的语音合成技术，通常指文语转换（Text-To-Speech，TTS）技术，主要解决如何将文字信息转化为可听的声音信息。典型的语音合成系统，主要包含文本分析、韵律生成和合成语音三个模块（图 1-1 实线框）。文本分析模块实现对输入文本的语言学表达，包括分词、词性标注等操作；注音生成模块负责发音内容的符号化描述，韵律生成模块主要实现层级结构、重音、语调的韵律描述；合成语音模块则根据发音内容与韵律描述实现声学参数生成。情感语音合成系统在 TTS 基础上加入特定的情感信息，使其也参与到声学参数的控制，实现携带特定情感的

语音。通常情况下，情感信息来源于离线实验对某些特定情感状态与语音信号的关联关系的分析总结。

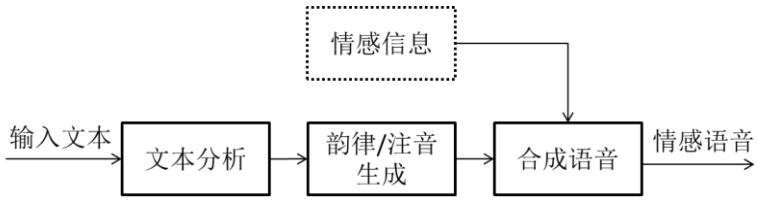


图 1-1 情感语音合成系统基本框图

Fig. 1-1 The basic framework of expressive speech synthesis system

目前，情感语音合成存在的问题主要有：当前该领域的研究主要集中在情感如何在语音中表达，而对于情感状态本身的描述问题则关注不够。在语音学的研究中，情感通常被简化为几个离散的基本类别或两到三维的超低维空间，这对于刻画具有时变特性的语音信号中的情感的发生与衍化显然是不够的，言语中的情感内容和感情方式是极为丰富的，各种情感之间还存在相互渗透、相互作用、相互转化的关联关系。而另一方面，当前的情感描述模型直接用于情感语音的研究时，还面临着情感类别与语音信号的变化之间映射关系不明显或者难以发现一一对应的映射关系，尤其对于处理某些复杂而微妙的情感，当前的情感描述方法则显得过于单一、刻画不够细腻。现有研究中对于言语中的情感的认识与刻画的不足，要求我们一方面借鉴心理学等相关领域成熟的研究成果，同时更为重要的是，要结合自身专业特色，探寻真正适合言语工程应用的情感的理论机制与描述模型。具体而言，新的情感刻画方案应考虑言语中情感的时变特性，而非某些离散的、静止的状态；还应考虑情感的多面性和复杂性，情感的产生受来自于社会、环境和过去经验等多方面因素的影响，同时会引起认知、心理、生理和行为等一系列的机体变化；除此之外，还应考虑不同方面的情感信息如何建模，如何量化且可计算，它们与语音学表达之间是否存在显性的关联关系等。综合这些问题，我们从人类产生情感语音的过程出发，结合认知科学的研究成果，分析归纳出言语中情感的生成及衍化过程；以此为基础，从认知、心理、生理和发音等不同角度给出情感的多视角描述模型，对情感进行全面、细腻、动态的刻画。

另一方面，自然、完备的人机交互接口技术，需要情感语音合成系统应能够结合文本内容与场景因素进行情感的自动预测。场景对情感特征的影响，属于社会学研究的范畴，一些社会学家和教育工作者进行了较早和较详细的分析，从感性的世界观理解不同场景，分析人的意境与场景的融合，并阐述了表达方式的不同[1]。我们将研究对象设定为朗读语音，相比于自然口语是更为规整的语言形式，情感的表

达也更稳定且可复制。在朗读语音背景下，场景的影响构成弱敏感环境因子[1]，背景音乐、环境以及发音人角色、面向的对象等都相对固定，因此，本文的研究中暂不考虑场景的影响，重点研究基于文本内容自动预测情感。从文本分析出情感并进行相应的语音学转换，能够使合成语音中的情感与文本内容一致，从而听感上更真实自然。文本的情感分析工作在自然语言处理领域有较多较成熟的经验和技术可供借鉴，但是在当前情感语音合成的应用背景下，对文本的情感分析与计算提出新的要求和挑战。首先，言语中的情感是一个动态衍化的连续过程，与传统的文本分类的任务不同，在情感预测过程中应考虑不同情感状态间的相互影响和衍化关系，这就要求模型应支持时序衍化关系的刻画。其次，情感语音合成是一个融合了多个不同尺度特征的任务，尤其对于篇章级的语音合成，除考虑单句的合成效果，还应考虑篇章中各句之间的关联关系，比如各句之间情感的延续性、篇章内部基调的一致性等，因此，模型还应支持多尺度特征之间关联关系的处理。除此之外，情感预测模型还应与情感描述模型相适应，例如：当采用情感的离散表示方式时，情感预测任务通常被建模为模式分类器；而采用维度表示方式时，情感预测任务则被建模为回归预测问题。当前情况下，我们的建模目标变为在言语情感生成及衍化过程的约束下实现多视角情感信息的预测，模型为一个具有多个层面、每个层面具有多个维度支撑的立体空间，与之对应，采用具有多层非线性映射结构的深度神经网络作为基本结构，在此基础上加入衍化关系与层级关系约束网络结构，使模型支持情感动态衍化过程的刻画以及多尺度特征的融合。

综上所述，本文主要解决两方面的问题：一、深化情感理论，完善情感描述体系，解决情感的准确、有效刻画和动态衍化过程的描述；二、建立融合不同影响因素和不同尺度特征的情感预测模型，解决情感的时序衍化特性和情感特征间相关性的建模。

本文研究的意义在于：1）理论层面，弥补当前情感语音研究中对于言语情感认识与刻画的不足，深化和发展认知科学中关于言语情感产生的相关理论；2）技术层面，为提高合成语音的情感表现力和情感识别的准确性提供理论依据和技术手段；3）社会学层面，加深对于智能本质的理解与研究，推动人机交互、人工智能等相关产业的发展。

1.2 研究现状

1.2.1 情感语音合成

情感语音合成属于语音合成技术的一个分支，伴随着语音合成技术的发展以

及人们对人机交互技术要求的不断提高，越来越多的国家和个人开始重视情感信息处理技术的研究。

情感语音合成的研究，主要涉及以下几个方面：

（1）情感生成机理与描述：情感生成机理的研究主要是在认知科学上探索大脑对信息的分析与处理的机理，解析大脑中情感概念的产生及衍化机制，并探索情感状态判定及与生理和行为之间的关系，为情感计算与情感语音合成的研究提供理论基础。人类情感的研究已经是一个非常古老的话题，心理学家、生理学家已经在这方面做了大量的工作，但目前关于情感的理论解释、情感如何表述等仍没有统一的定论。比如：在心理学上，就存在情感、情绪和表情等不同的概念表述，我们认为，它们均属于广义情感的研究范畴，描述了同一事物的不同方面，通常情况下，情绪是人对客观事物的态度和体验，情感则是就脑的活动而言，表情是指情感在有机体身上的外在表现[3]；在语言学上，则倾向于将情绪或情感与态度分开来对待，态度属于副语言学信息，而情感属于非语言学信息[4]。我们将态度作为认知评价的体现之一，将其归为广义情感概念的一部分。

（2）情感语音数据库构建：情感语音数据库是进行情感语音分析、情感语音合成与语音情感识别的基础，为其提供分析数据、合成语料或模型训练语料以及测试用语音。情感语音数据库的建立涉及到心理学、生理学、统计学和计算机科学等多门学科的交叉，且受性别因素、种族因素、语言因素等多方面因素的影响，因此到目前为止，在世界范围内仍没有一个构建情感语音数据库的标准。从语料来源来分，情感语音分为自然语音（Spontaneous Speech）、表演语音（Acted Speech）和诱导语音（Elicited Speech）三种，与之对应，情感语音数据的采集方式分为真实情感采集、模拟情感和诱导方式三种。据此，国内外很多研究机构建立了各自不同语言、不同类型、不同内容和不同规模的情感语音数据库。

（3）情感声学特征分析：情感声学特征分析是研究情感语音合成与语音情感辨识的前提，也是决定情感语音合成效果的关键因素与技术难点之一。具有丰富情感的语音与中性语音的差别，表现于音高、音长、音量和音色等各个方面。情感声学特征分析就是研究表征不同情感状态的声学特征参数，发现并建立情感与声学特征间的关联关系。研究主要分两类声学参数：一类是与韵律相关的超音段参数，如基频、时长、能量等，情感在语音上的表现主要体现在韵律上；除此之外，音质也对情感表达起一定作用，音质主要由反应发音时声门波形状变化的音段参数决定，如谱倾斜（Spectral Tilt）、声门波的 NAQ（Normalized Amplitude Quotient）等。

（4）情感语音合成方法：现有的合成情感语音的方法主要包含波形拼接、语音转换和统计参数合成三大类[5]。波形拼接方法通过录制大规模情感语料库，收集不同情感类型的语音，合成时从相应的情感语料库调取相应的语音片段，通过拼接

得到保留了原始录音语气的语音。语音转换方法通过分析不同情感类型的语音相对于中性语音在声学参数上的变化规律，对预先合成的中性语音进行调整转换得到新的情感语音。统计参数合成基于隐马尔科夫模型（Hidden Markov Models，HMM）等统计模型，对不同情感的语音进行参数化表征和声学建模，并以此为基础进行情感语音的声学预测，合成携带不同情感的语音。目前三种方法各有优劣，通过波形拼接方法得当的情感语音，自然度优于其他方法，但可合成情感类型受限于情感语料库现有的情感类型，扩展性差，且合成效果依赖于大规模语料库，建库成本高昂；语音转换方法依托于情感声学特征分析的研究，虽然观察到一些情感与声学参数变化相关联的指向性线索，但由于情感表现的多样性和复杂性，导致了情感声学参数的数值分布多呈现较大的离散特性；统计参数合成方法可以在短时间内自动构建一个新的合成系统，基本不需人工干预，对于数据的需求量也少于波形拼接的方法，合成情感类型也较前两种方法灵活，但是由于 HMM 产生的频谱和韵律模型过于平滑，使频谱和韵律模式上的细节丢失，影响其合成语音的自然度。

2000 年，ISCA 在北爱尔兰的贝尔法斯特召开了一个称为“语音与情感：研究的概念框架”的研讨会，第一次把致力于情感与语音研究的学者聚集在一起，大会丰富的成果指明之后情感语音的研究将沿两条线前进[13]：一、语音表达与情感对应关系的描述；二、情感状态本身的描述。这之后，一些新的研究方法均围绕着这两条路线展开。在合成方法上，日本 ATR 提出了一种基于语料库的情感语音合成方法，通过在具有不同情感的语料库间进行切换挑选合适的单元并采用波形拼接的方法合成情感语音[14]。东京工业大学的 Yamagishi 等人基于统计参数合成方法，采用模型插值的方法合成两种不同发音风格间的中间风格的语音，还通过模型适应的方法实现基于小规模语料库的情感合成[15]。近年来，随着深度学习的再度兴起，深层神经网络也被用于声学建模[16,17]、韵律预测[18,19]、语音转换[20,21]等情感语音合成的相关方面，由于深层神经网络强大的非线性学习能力，缓解了 HMM 模型对于参数细节刻画的缺失，进一步提升了合成语音的效果。在情感描述方面，Cowie 和 Cornelius 专门对语音中的情感特点进行分析，提出了情感的广义描述、结构化描述以及时序描述等建模方面的新要求[22]。在这一趋势的影响下，以 Scherer 的成分过程模型（Component Process Model，CPM）[23]为代表的认知模型得到越来越多的情感语音以及情感计算研究者的青睐，情感的描述从之前的静态、片面、离散的表示方式向动态、全面和连续的方式过渡。

同时，越来越多的语音合成研究者开始研究情感的自动预测或生成。Alm 等人为提升儿童童话故事合成器的表现力，在合成系统前端加入基于机器学习方法的文本倾向性检测模块，用以指导后续的声学信号调整[24]。吴志勇和他的同事借助 PAD（Pleasure-Arousal-Dominance）情感三维模型分别对文本语义不同尺度的情感

信息进行描述和预测，如：Pleasure 和 Arousal 两维基于词义描述较低尺度的韵律词的情感表达，Dominance 维基于对话行为描述句子级的整体发音，并基于此种设定分别对不同尺度的声学特征进行局部和整体的非线性转换，在一中文旅游信息口语对话系统上取得了不错的实验效果[25]。

目前，国际上研究情感语音比较活跃的单位有：美国 MIT 媒体实验室、MIT 人工智能实验室、CMU（Carnegie Mellon University）、英国爱丁堡大学、英国女王大学、瑞士日内瓦大学、日本 ATR 等。此外，微软、IBM、英国电信、索尼等公司也都相继成立了情感计算和智能交互的研究小组。国内，比较著名的单位有：中科院自动化所、中科院声学所、社科院语言所、清华大学、中国科技大学、哈尔滨工业大学以及东南大学等。中科院自动化所陶建华老师带领的团队，近期正在开展关于情感的多尺度时序建模的研究，重点研究时序上下文环境对于情感预测的影响以及情感特征之间的相互影响[26]。社科院李爱军老师与天津大学党建武老师合作，共同推动情感表达的跨文化多模态研究[27]。清华大学蔡莲红老师的团队与中科院心理所合作，将人机交互技术与心理学、认知科学相结合，在情感语音合成领域开辟了新的探索道路[25,28]。此外，东南大学的赵力老师[29]、清华大学的徐明星老师[30]、哈尔滨工业大学的韩继庆老师[31]和刘挺老师[32]等均在从事情感语音相关方向的研究。

1.2.2 情感描述

可用于言语情感研究的情感描述方案有很多，各有其合理性，但同时也反映了一个现实，学术界对“情感”概念本身的界定还未达成共识。究其原因，在于研究对象“情感”的复杂性以及关注角度的不同。

狭义来讲，情感通常被定义为心理上的主观体验，与心理感受等价，Cowie 和 Cornelius 将这部分成分定义为确切的情感状态（emotional states）；除此之外，他们认为，对于情感更广义的解释应该包含与情感产生相关的认知活动以及心理活动触发的生理唤起等情感相关状态（emotion-related states），这部分成分被认为更广泛的存在于日常对话中，且与语音的关联更紧密[22]。与 Cowie 和 Cornelius 的理论类似，Ortony 等人将情感状态分为生理/身体状态（physical/bodily states）和精神状态（mental conditions），前者对应于生理唤起水平，后者包含认知成分、心理成分和行为成分[33]。

关于情感生成理论的研究起源很早，也先后涌现出各种不同的派别，Cornelius 对心理学中的情感生成理论进行汇总，归纳出四种主要观点：达尔文主义（Darwinian）、詹姆斯主义（Jamesian）、认知观点和社会构成观点（Social

constructivist）[34]。Darwin 主义认为情感是与物种生存有关的进化现象，相同或相近的物种具有相同的情感表达方式；而以 William James 为代表的学派认为，情感由身体变化决定，这种说法成立的前提是将情感与主观感受等价，即主观感受是身体变化的反应；认知观点认为情感受对周围环境关乎自身好坏的评估影响，评价结果决定其他成分的反应模式；社会构成学说则认为情感的意义由社会文化所决定的行为和价值观构成。

目前，应用最广泛的情感描述方法是离散表示法和维度空间表示法。离散表示法基于达尔文主义构建，认为某些情感具有特定的触发条件和固定的生理及表情上的反应模式，称为基本情感或典型情感；其他情感由基本情感的组合或程度变化得到。最常用的基本情感是：高兴、悲伤、恐惧、厌恶、生气和惊讶[35]。维度表示法通过抽取情感空间的几个本质属性作为基本维度，不同情感间的相似性或差异性通过彼此在情感空间中的距离来表示。最常用的维度空间是一个由“愉悦度”（又称为“评价维”）和“激活度”两维构成的超低维空间[36]，为了进一步增强对于某些情感的区分能力，有时会增加一维“控制度”[37]。离散表示法可以很直观的为人们提供有关情感状态的整体印象，但是对于情感的语音学实现来说所提供的信息却过于抽象，不利于发现情感与声学参数间的转换关系。维度表示法一定程度上增强了情感与语音（激活维）的关联性，但是情感空间被认为是一个更复杂多变、甚至有可能存在层级性的多维空间[22]，将这样一个复杂空间投射到一个仅由两到三维的超低维空间，有可能引起信息不对等和信息丢失问题。

近年来，认知评估理论[38]逐渐被情感语音和情感计算的研究者广泛接受并采纳[39]。在该理论中，情感被定义为持续一段时间的连续过程，而非某些离散状态。情感由对过去经验和当前情境的评估触发，涉及认知、动机、感受、生理、表情等一系列变化，评估结果决定其他部分的反应，反应结果又会反馈回去影响评估的结果。Scherer 提出的成分过程模型 CPM[23]就是基于认知评估理论构建的情感描述体系，情感由评估结果和反应模式共同表示，不再拘泥于有限数目的离散状态，而是形成情感的分布式表示，内部构造及其发展衍化的过程也可得到体现。情感的多成分模式化表示为言语情感的细致刻画提供了新的可供借鉴的方法，但是由于该模型将更多的重心放在认知前端，而对其他部分的反应刻画得不够具体，不适合直接应用于言语情感的描述。本文将基于这些研究提出专门针对言语情感的描述体系。

1.2.3 文本情感计算

在本文中，文本情感计算是对文本中所负载情感进行判别或预测的问题，在此作为情感语音合成系统前端文本分析的重要环节，为合成系统提供与文本内容相

符的情感描述与刻画，使合成语音的情感表达听起来更自然、更恰当。

更为一般地，文本情感计算任务是自然语言处理领域中一项重要的基础性研究，在很多领域都有所应用。例如，政府机构通过舆情分析挖掘出人们对某个热点事件的观点倾向，从而引导某些决策的制定；电子商务网站通过分析用户对产品和服务的评论来调查市场对产品的反应，从而对产品质量和市场做出相应调整；用户也可根据公众对于某件商品的评价来指引自己的消费决策。

由于情感类型常被表示成离散类别，因此文本情感计算也通常被当作分类任务处理。根据类别划分粒度不同，文本情感预测可分为较粗颗粒度的主客观分类、倾向性判别（肯定、否定和中立）和较细颗粒度情感多分类问题（高兴、悲伤、恐惧、厌恶、生气和惊讶等）。根据文本分析单元的尺度不同，文本情感计算又可分为篇章级、段落级、句子级甚至更细尺度的情感分析。

文本情感计算的常用方法可以概况为以下三类[41,42]：

1) 基于关键词和语法规则的方法：该方法是判别文本情感最直观的方法，通过查询带情感信息标注的情感词典，结合相关的启发式语法规则对文本进行情感分类。例如，当基于离散表示模型时，情感词由基本情感类别标注，Calix 等人基于该种类型的情感词典提取词特征（互信息）实现文本情感分类[43]。Tokuhisa 等人则基于标注了词语正负倾向性的情感词典实现文本倾向性检测[44]。此外，Osherenko 等人尝试加入一些启发式语法规则，根据是否存在否定词和程度副词作为特征训练了一套情感预测模型[45]。基于关键词和语法规则的方法依赖于标注了情感信息的情感词典，对于不包含任何情感关键词的文本和出现具有歧义的关键词的情况难以判别其情感，语法规则的制定也强依赖于专家知识，这些问题都制约了该类方法的推广使用。

2) 基于语义的方法：针对上述方法中存在的一些问题，一些研究者试图从语义层面进行解决，提出了基于语义的文本情感分析方法。例如，Liu 等人通过引入知识库来扩展词语含义及其相互关系，避免了歧义情况的发生[46]。Bellegarda 基于潜在语义分析（Latent Semantic Analysis，LSA）的方法，综合考虑两类语义信息（领域相关和情感相关）的影响训练各情感类别的划分区域并计算各类别的区域中心点，通过计算文档与各情感类别中心点的相似性判别文本情感类别[47]。基于语义的分析方法提出了情感概念（emotion concept）的想法，将抽象的情感类别扩展为更为具体的情感概念，较之前的单纯依靠词典匹配的方法更具灵活性，但是情感概念的获取仍不能摆脱人工构建的知识库作为指引。

3) 机器学习的方法：机器学习的方法将文本情感分析问题视作统计分类问题，根据训练集中标注样本的比例可分为监督学习、半监督学习和无监督学习。基于监督学习的方法认为情感分类是一个标准的包含大量标注样本的统计分类问题，经

典的统计分类模型，如朴素贝叶斯、K 近邻、支持向量机、最大摘、条件随机场等都曾用于文本旳情感分类。Trilla 等人针对句子级别的文本倾向性判别问题，对比了不同分类器的分类效果以及不同特征组合对于分类结果的影响[48]。半监督学习通过少量标注样本和大量未标注样本进行学习，常用的技术手段包括聚类算法、迁移学习（Transductive Learning）和协同学习（Co-training）等。例如，Wan 针对中文情感标注语料较少而英文标注语料则相对丰富的情况，利用机器翻译和协同学习融合两种语言的信息来提高中文情感分类的效果[49]。无监督分类方法不使用任何有标注样本，但仍会借助少量的先验知识。如，Agrawal 和 An 构建了小规模的指示情感类别的种子词集，然后通过候选词语与种子词间的点互信息（Pointwise Mutual Information，PMI）来计算文本的情感类别[41]。基于机器学习的文本情感分类方法适用于较大尺度的文本分析单元，对于专家知识和情感模型的依赖性较低，因此本文选用该类方法基于新提出的情感描述体系进行文本情感预测。

在机器学习领域，深度神经网络技术近期得到了快速发展，DBN (Deep Belief Network)[50]、DSN (Deep Stacking Network)[51]等多种深度学习（Deep Leraning）模型涌现出来，并在图像识别[52-54]、语音识别[55-57]、自然语言处理[58,59]等应用领域获得了实实在在的成果。相对于传统的机器学习模型，深度神经网络具有多层非线性映射的深度结构，因而具有逼近更为复杂函数的能力；同时深度学习模型还可通过组合低层特征形成更加抽象的高层表示，通过分布式表示刻画所处理数据的复杂结构规则[60]。考虑到处理特征参数多层面、多尺度的特点，采用深度学习技术构建言语情感预测模型，显然是一个合理的选择。

已有研究中，文本情感计算常用的资源有：General Inquirer（GI）[61]英文词典，包括 1915 个褒义属性词和 2291 个贬义属性词；WordNet[62]英文结构化语义词典，收录 95600 个不同的词形和 70100 种词义，并按照词义将词分为不同的同义词集（Synsets）；SentiWordNet[63]是基于 WordNet 构造的情感词词典，描述了 WordNet 中的每个同义词集(Synset)褒、贬或中立的倾向及强度；WordNet-Affect[64]也依据类似方法构建，不过标注内容换为不同的情感类别，对于不同情感类别间的关系也给出了层级表示；近年来，一些研究者还结合认知评价理论构建了常识知识库 EmotiNet[65]；HowNet（知网）[66]是一个以中、英文词语所代表的概念为描述对象，以概念与概念之间、以及概念所具有的属性之间的关系为基本内容的语言知识库，归纳了 9193 个中文情感词语和 9142 个英文情感词语，是中文文本情感分析重要的基础资源。

1.3 研究目标与内容

1.3.1 研究目标

上文分析指出，目前的情感语音合成研究，在情感模型与情感预测方面，有两方面关键问题需要解决：一、相关的理论需要有所升华，尤其是要解决情感的准确刻画和动态衍化过程的描述；二、建模技术需要有所突破，考虑到影响情感因素及情感生成过程的复杂性，所需处理的特征参数可能会来自多个层面，模型应能支持多尺度特征处理及动态衍化过程刻画。针对这两个关键问题，本文的研究目标设定为：

（1）探索言语情感生成及其衍化过程，形成言语情感描述方案；

（2）以此为基础，面向情感语音合成的应用，构建言语情感预测模型。

1.3.2 研究内容

为实现上述研究目标，本文的研究工作将按照以下思路来进行：以相关领域研究为指导，采用心理语言学、感知语音学实验和数据分析相结合的方法，探索汉语朗读或播音等创作型有声语言活动中情感表达与言语特征间的关联关系，进而得出言语情感的生成及衍化过程，并基于此形成情感的多视角描述体系；以此为基础，利用机器学习等计算方法，构建基于文本分析的言语情感预测模型，考虑到特征参数的层级性，因此引入深度神经网络实施模型的计算，模型支持多尺度特征处理及动态衍化过程刻画。具体研究内容与技术思路为：

（1）言语情感生成及衍化过程

关于情感的生成与界定，至今没有统一的定论。心理学领域存在大量的情感理论与模型可供借鉴，基于对心理学情感理论与模型的分析，归纳汇总出言语中情感生成所涉及到的可能影响因素及各因素间相互关系；同时基于朗读学、播音学等实践经验指导，阐明各因素间的信息传递规律，归纳汇总出由文本到言语情感的生成过程及衍化机制。

（2）言语情感描述方案

通过对心理学已有的情感模型进行分析，发现这些模型分别基于不同的视角来解释或定义情感，各视角都有其存在的合理性和独特性，同时它们之间又有所交叠或关联。因此本研究中，以言语情感生成过程为指导，借鉴心理学、朗读学、播音学以及语音学相关成果，提出从不同视角刻画情感不同方面影响因素的多视角描述方案；然后采用心理语言学、感知语音学实验和数据分析相结合的方法，确认

结构模型中每个层面的主要因素，进而得到每个视角的具体表示方式。

（3）新闻言语情感数据库构建

情感数据库是研究情感计算及情感声学特征的基础，构建数据库是本文工作的重要组成部分。不同于自然口语（spontaneous speech）中情感的自发流露，新闻言语中的情感产生有着相对显性的过程，且在相当程度上可以重现，为言语情感研究带来些许便利。本文使用的新闻言语情感数据库采用表演语音的构建形式，构建过程包括：语料的收集与录制、基于多视角情感描述体系对该库进行人工标注、以及对标注结果进行统计分析等。文稿来源于来源于中国传媒大学有声媒体语言研究中心所积累的播音文稿数据库。录音工作由具有专业背景的发音人在专业录音棚完成。

（4）言语情感预测模型

本文所提出的言语情感描述方案从多个视角描述言语情感，影响情感的因素来自多个层面，由此归纳得到的言语情感生成模型是结构化的。因此，为了处理层次化、多尺度特征，采用结构化的深度神经网络来构建言语情感预测模型。具体来说，在言语情感生成及其衍化过程指导下，利用言语情感数据库的标注数据，以及基于语义分析得到的文本特征，训练深度学习模型，实现由朗读文本到言语情感的预测模型。以发音描述作为预测系统的最终结果输出，用于指导后续声学参数的调整。模型综合考虑来自篇章级、段落级以及句子级不同尺度单元的上下文环境的影响，以及情感衍化过程中各环节间的相互影响，形成多层嵌套的立体网络，可以处理结构化特征以及具有时序性和相关性特征的建模，且实现方式简单，具有良好的可扩展性。

1.4 论文结构安排

本文共分为六个章节，各部分安排如下：

第一章 引言：首先明确本文的研究问题与意义；然后对所涉及到的相关领域的国内外研究现状进行介绍；之后指明文本的研究目标和总体思路；最后对具体研究内容和技术路线进行阐述。

第二章 情感理论和描述体系：首先从心理学、朗读学和播音学等相关学科出发，介绍与研究言语情感生成过程及其描述体系的相关理论和模型参考；之后给出言语情感生成过程的具体阐述，并给出多视角情感描述体系的构建过程及具体表示形式。

第三章 新闻言语情感数据库构建：首先对情感数据库的构建方法及遵从原则进行总结；然后介绍本文所使用的新闻言语数据库的构建过程，包括语料的收集与录制、基于多视角情感描述体系对该库进行人工标注、以及对标注结果进行统计分析等。

第四章 基于深度神经网络的言语情感预测：首先介绍几种常见的深度神经网络的网络结构和存在的问题；然后介绍我们基于言语情感生成过程和深度堆叠网络搭建的预测模型的网络结构和建模策略；之后对模型的训练算法、优化措施和文本处理等内容进行说明；最后对模型的性能进行测试并对结果进行分析讨论。

第五章 多尺度情感预测建模：基于第四章提出的言语情感预测模型，搭建融合不同尺度特征影响的多尺度情感预测模型。本章首先对多尺度情感预测模型的网络结构和建模策略进行介绍；然后介绍多尺度文本特征提取的具体步骤；之后给出整个系统的框图；最后给出多尺度情感预测模型的实验测评结果和性能分析。

第六章 总结与展望：对全文的工作内容和取得的结果进行总结，并对下一步工作进行展望。

本文的组织结构如图 1-1 所示：

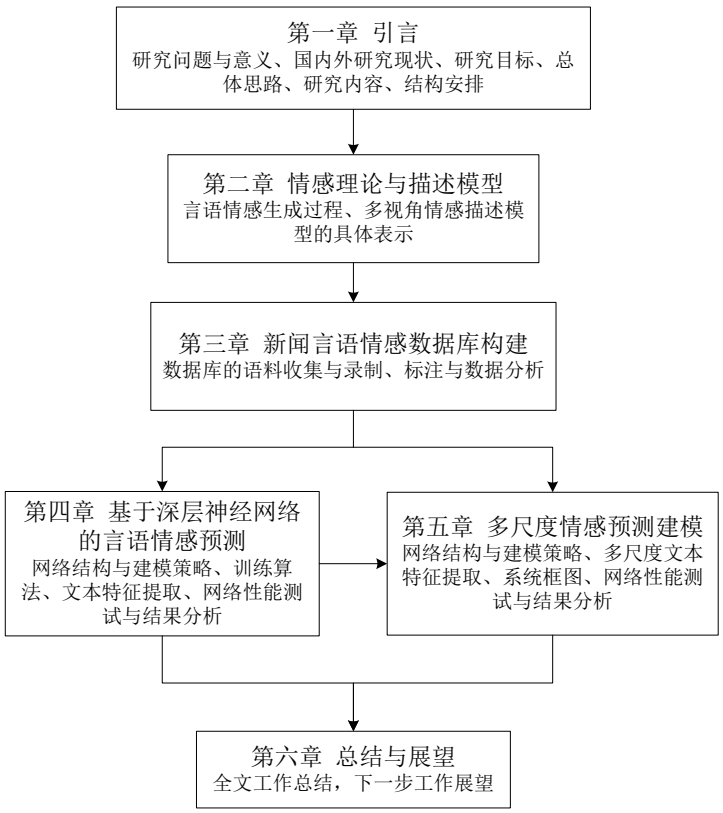


图 1-1 论文组织结构及各章主要内容

Fig. 1-1 The structure of the thesis and the main content of each chapter

2 情感理论与描述体系

目前，情感语音合成的研究主要集中在情感的语音学表达，对于情感自身的描述问题则关注不多。情感通常被简化为几个离散的基本类别或两到三维的超低维空间，这对于情感这一复杂多变的心理活动的刻画显然是不够的，同时当应用于情感语音合成时，也面临着情感类别与语音信号的映射关系不明显或生成情感类型过于单一等问题。

本文以文语转换系统中的情感生成为切入点，以心理学已有的大量情感理论和情感模型为基础，并将文语转换过程与朗读或播音等有声语言创作过程类比，参考朗读者或播音员在有声创作过程中对于情感的酝酿和把握，揭示言语情感的生成过程，并基于此构建情感的多视角描述体系，从认知、心理、生理和发音等不同视角更加全面、细致的刻画言语中的情感。

2.1 心理学关于情感的研究

心理学认为，情感或情绪（Emotion）一词源于词根“move”，即：使人“动”起来，是以生理唤起水平、面部表情、姿势和主观感觉的变化为特征的某种状态[67]。同时，情感和情绪又是人对客观事物与人的需要之间的关系的反映，产生于认知活动过程中，并影响认知活动的进行[68]。情感与情绪在日常生活中常被混用，但心理学认为二者有以下区别：

- 情绪是更多与生理需要满足与否相联系的心理活动，情感则是与社会性需要相联系的心理活动；
- 情绪发展先于情感体验；
- 情绪一般比较不稳定，情感则相对稳定；
- 情绪表现具有外显性，而情感表现多以内在感受形式存在。

某种意义上，情绪是情感的外部表现，情感则是情绪的本质内容，二者本质上存在一致性，从计算角度可以不做仔细分辨[68]。

关于情感的产生与界定的争论是一个古老的命题，且至今仍未停止。Scherer将这些争论汇总为“认知与情感的争论（Cognition-Emotion Debate）”，“意识和身体的争论（Mind-Body Debate）”，“生物与文化的争论（Biology-Culture Debate）”和“中央与周围的争论（Center-Periphery Debate）”[40]。其中，“认知与情感的争论”主要围绕情感与认知是相互独立的还是相互影响密不可分的；“意识和身体的争论”主要围绕意识活动与身体反应间的关系展开；“生物与文化的争论”则存在于生存环境的影响和社会文化的影响之间；“中央与周围的争论”围绕主观

感受和其他部分的关系。Cornelius 则进一步提炼出情感理论中的四种主要的观点：达尔文主义（Darwinian）、詹姆斯主义（Jamesian）、认知观点和社会生成（Social constructivist）观点[34]。本节将首先从情感成分的角度对这些情感理论予以阐述，然后介绍几种典型的描述情感的方法或模型。

2.1.1 情感理论

当代评估理论认为，情感不是孤立的离散状态，而是持续一段时间的连续过程，涉及机体一系列子系统的变化，这些子系统的变化被定义为情感的成分（components），包括：评价周围环境好坏与否的评估成分，指示行动倾向性的动机成分，描述身体反应的生理成分，表示表情与动作的行为成分，以及描述主观体验的感受成分[38]。Cowie 和 Cornelius 将这些情感成分分为确切的情感状态（emotional states）和情感相关状态（emotion-related states），前者指心理体验，后者包括对环境形势的评估、行动倾向、沟通行为和生理状态[22]。Ortony 等人则将它们分为生理/身体状态和精神状态，包含生理成分、认知成分、感受成分和行为成分[33]。我们认为，动机也可以作为评估的内容之一，对环境形势的评估成分和指示行动倾向性的动机成分可以合并为认知成分，因此我们认同 Ortony 的观点，将情感定义为包含认知、心理、生理和行为四种成分相互作用的连续过程。

- 认知成分

有关认知与情感的关系的讨论，可以追溯到 Plato 时期，他认为，灵魂可以分为认知、情感和动机三个方面，且这三方面是互相独立且对立的。而 Aristotle 却主张将情感与认知和动机分开是不可能的，不同级别的心理功能之间存在相互作用，这是认知理论的起源，但还没有明确认知在情感中的地位与作用[40]。19 世纪 60 年代，Schachter 提出，任何一种情感的产生，都是由外界环境刺激、机体的生理变化和对外界刺激的认识过程三者相互作用的结果，并通过实验证实，对外界刺激和身体变化的认知对于情感体验起决定性作用[69]。Arnold 将这种强调评估对于情感产生的主导作用的理论命名为评估理论（appraisal theories），评估对象是环境刺激对于个体的意义，如：是好还是坏，是威胁还是支持，与己有关还是与己无关，等等[70]。评估理论之后，社会心理学家又针对情感的差异性提出社会生成学说，认为情感是对社会文化进行评估的产物，社会文化决定评估的内容，并组织或约束人们的行为，因此才有了情感表达在文化与文化之间和人与人之间的差异。某种程度上，社会生成学说可以算作评估理论的一种补充，社会生成学说并没有否定认知评估对于其他成分的决定作用，而是扩展了需要评估的内容，即在原有的环境刺激的基础上增加社会文化差异的影响。基于认知的情感模

型就是依据认知成分对于情感的决定性作用而构建的，其中认知部分由若干评价维度表示。

- 心理成分

心理成分又被叫做主观体验（subjective experience）或情绪感受（affective feeling），指个体对不同情感状态的自我感受[71]。心理成分可以说是情感最本质的内容，认知过程、生理变化和行为表现都可被看作是与情感相关的状态，对于其是否属于情感的概念范畴都曾经或者仍然存在争议，只有心理成分无论何时都被确切的当作情感的必不可少的成分，在狭义的概念中甚至与情感等价，如：James 的早期观点就认为情感是身体变化的主观反应，情绪体验之前，必须先有身体上的表现发生[72]。这种观点指出了生理唤起和行为反馈等身体变化对情感体验的作用，但是忽略了认知的影响，仅仅将情感与心理反应等价，缩小了情感的概念范畴。心理感受属于身体内在的体验，不具有生理或行为上的外部特征，但是对生理或行为有重要的调节作用。因为心理成分常被用来描述最终的情感体验，因此通常采用一些日常语言标签来表示，如快乐、恐惧、愤怒等；又因为心理体验的交错复杂性，同一时间可能处于交织着不同性质的情感体验中，如悲喜交加、百感交集等，因此又有了基本情感和复合情感的区分和组合关系。

- 生理成分

生理成分指伴随特定情感状态出现的生理反应，特别是一切无意识反应[67]。常用的生理指标有心率、呼吸、血压、皮肤电活动、掌汗、皮质醇水平、相关事件电位、瞳孔直径和脑电波等[71]。这些反应由自主神经系统控制，其中交感神经分支使身体唤起并为紧急行动做准备，副交感神经分支导致身体的镇静和安静[67]。在情感的维度表示中，生理成分通常由激活度（激动-平静）、紧张度（紧张-松弛）等维度表示。

- 行为成分

行为成分又称为外在表现或表情，包含了面部表情、姿态表情和语调表情等身体各部分的外部表现。Darwin 在其 1872 年的著作《人与动物的情感表达》（The Expression of Emotion in Man and Animals）[73]一书中指出，情感是与物种生存有关的进化现象，相同或相近的物种具有相同的表情，如人类在愤怒时暴露牙齿的方式与猴子和狗等其他哺乳动物一样；表情之所以在进化过程中被保留下来，是因为向他人表达情感有助于个体生存。这种观点强调情感的生物适应性，是基本情感的理论基础，基本情感最主要的判别标准就是是否具有进化意义和跨种族、跨文化的一致性。与之相对，近些年兴起的社会生成学说则提出情感具有社会适应性，个体为了适应社会情境、文化规范和人际关系的需要，情感由先天预成性向随意性转化，不同文化背景下有不同的情感表达，如：亚洲文化强调群体和谐

性，在公共场合很少表现生气，因为生气会导致与他人的分离；北美文化则强调个人权利和需要的自由表达，生气也被认为是面对不公正待遇的自然反应。情感的社会适应性可以看作生物适应性在人类身上的延伸[68]。Darwin 的观点与社会生成学说都关注情感的行为表现，前者着眼于行为表现的普遍一致性，后者着眼于行为表现的文化差异性。与本文研究内容相关的语调表情通常通过韵律的变化（如音高、音强和语速等）表现，还可以通过音质的变化（如明暗、虚实等）体现。

对于以上四种成分的关系，普遍认可的是认知成分决定其他成分，其他成分是认知的反应。其他三者间也存在直接或间接的影响，但不一定是决定性的影响；三者发生的先后顺序也没有定论，James 认为先有生理反应和行为才产生情绪体验即心理上的主观感受，Cannon 通过实验手段观察到由丘脑同时发动肌体唤起、行为和情绪体验。情感过程是连续的且循环的，同一时间可能存在多种成分的共同作用，不同成分间还可能存在反馈，生理状态、心理感受和行为的反馈信息可能改变评价的结果，这一变化又进一步改变其他反应，而反应的改变会再次改变对事件的评价或认识，情感就是在这一过程中产生、发展与消亡。

2.1.2 情感描述

描述情感的模型有很多种，王国江、王志良等将情感模型分为基于任务的和基于设计的两种：前者侧重于自然情绪的决策、表现和行为的模拟，是为了实现具体的特定任务，并不过多注意自然情感的发生机理；后者则偏重于对触发情绪产生和变化的内部机理的研究，将情感看作认知机制中的一个组成部分，看重与动机、心境、认知评价等过程的结合，力图模拟情绪的自然产生过程[74]。更具体些，Scherer 根据关注重点、诱发机理和区分机制的不同将其分为离散情感模型、维度模型、意义导向（meaning oriented）模型和成分模型四种[40]，其中，前三种都没有或很少涉及对于发生机理的解释，因此可以看作是基于任务的模型，只有第四种属于基于设计的模型，基于认知理论阐释了情感的发生机理。接下来分别对这几种模型的理论基础、特点、形式及其相互关系作简要介绍。

- 离散情感模型

离散情感模型基于 Darwin 的生物进化理论，将情感看作是与物种生存相适应的进化现象，关注重点是情感在面部、肢体和声音等方面的动作表达或适应性行为模式，而对情感的发生机理没有明确的阐明，只提及由某些特定的情境决定。离散情感模型将情感分为相互独立的类别，这些相互离散的类别在外部表现、生理唤醒模式上都存在一定的差异。Izard[75]从生物进化和个体发展角度将情感分

为基本情感和派生情感（也被称为主要情感和次要情感）：基本情感是先天预成的，具有跨种族、跨文化的普遍一致性和各自独立的外显表情、内部体验生理神经机制以及不同的适应功能；派生情感随个体认知的成熟而逐渐发展，并随着文化的不同而变化，派生情感由基本情感的变化或者组合得到，Cowie 和 Cornelius 将这种生成理论命名为情感的调色板理论[22]。

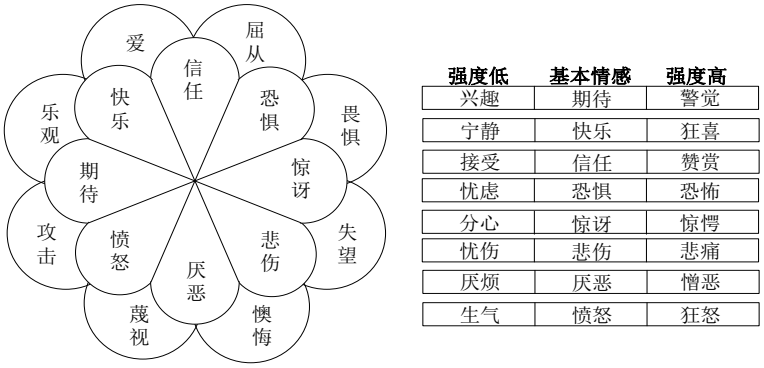
我国自古就有关于基本情感种类的各种说法。《中庸》将情感分为“喜、怒、哀、乐”四种。《左传》、《荀子》和《白虎道》均将情感分为“喜、怒、哀、乐、好/爱、恶”六类。《礼记》则提出七情说，曰“喜、怒、哀、惧、爱、恶、欲，七者弗学而能”。著名心理学家孟昭兰则从婴儿情绪的发生角度，认为人类婴儿有六种基本情绪：快乐、兴趣、厌恶、恐惧、痛苦和愤怒[76]。国外学者也对基本情感做了大量研究。Ekman 及其同事通过寻找面部表情的共性得到了六种基本情感：高兴（joy）、悲伤（sadness）、恐惧（fear）、厌恶（disgust）、愤怒（anger）和惊讶（surprise）[77]，Cornelius 根据其在心理学界和工程界的重要影响将其命名为“Big Six”[78]。Shaver 等人通过人工分类的办法将 135 个被认为是描述情感的词语按相似性分为六个基本类别，分别是：喜爱（love）、愉悦（joy）、惊讶（surprise）、愤怒（anger）、悲伤（sadness）和恐惧（fear）[79]，可以看出与 Ekman 的“Big Six”基本一致，仅将“厌恶”换成了“喜爱”而已。Plutchik 按照情感的相似性和两极性对情感进行划分，得出八个基本情感：恐惧（fear）、惊讶（surprise）、悲伤（sadness）、厌恶（disgust）、愤怒（anger）、快乐（joy）、期待（anticipation）和信任（trust）[80]。Ortony 和 Turner[81]、Laros 和 Steenkamp[82]都曾对心理学文献中所提及的基本情感类型进行了汇总，我们将所有这些被认为是基本情感的情感词根据被提及的频率排序，如表 2-1 所示，当词频大于等于 5 时，即有五位及五位以上学者都认同该情感词属于基本情感，可以得到 fear、anger、disgust、sadness、joy、love 和 surprise 七种认可度较高的基本情感，这也刚好是 Ekman 和 Shaver 两人结论的汇总。

派生情感由基本情感的不同组合或强度上的变化得到。如 Plutchik 在提出八种基本情感的同时还对基本情感的组合关系以及强度的变化做了探讨，图 2-1 画出了部分复合情感（即相邻基本情感间的混合，这种混合还可能发生在相距更远的情感之间[67]）和基本情感强度上的变化。有些派生情感可以命名，但大多数派生情感是难以命名的[68]。

表 2-1 心理学文献中所提及的基本情感词基于词频排序的结果

Table 2-1 The basic emotion words sorted by the frequencies in psychology literature

基本情感词	词频	基本情感词	词频
fear	11	contentment	1
anger	9	courage	1
disgust	7	dejection	1
sadness	7	despair	1
joy	6	elation	1
love	5	expectancy	1
surprise	5	grief	1
happiness	4	guilt	1
rage	4	hate	1
anxiety	3	hope	1
shame	3	hostility	1
contempt	2	liking	1
desire	2	panic	1
distress	2	pleasure	1
interest	2	pride	1
pain	2	sorrow	1
wonder	2	subjection	1
anticipation	1	tender	1
aversion	1	trust	1



强度低	基本情感	强度高
兴趣	期待	警觉
宁静	快乐	狂喜
接受	信任	赞赏
忧虑	恐惧	恐怖
分心	惊讶	惊愕
忧伤	悲伤	悲痛
厌烦	厌恶	憎恶
生气	愤怒	狂怒

图 2-1 Plutchik 情感模型中的基本情感和部分派生情感[67]

Fig. 2-1 The basic emotions and part of sencondary emotions in Plutchik's emotion model[67]

使用离散情感模型来描述情感符合人们的直觉和常识，有利于情感计算的成果在现实生活中的推广和应用。但是随着研究的深入，离散情感模型的问题也逐渐显现出来。首先，对于基本情感的划分与界定仍没有统一的意见，不同研究者基于自身需要关注不同的情感类型，限制了研究成果间的相互比较。其次，离散

情感模型关注重点是整个情感过程的后端，即身体和行为上的反应模式，对于情感的发生机制没有深入研究，而这对于研究情感的自动生成造成困难。

- 维度模型

维度模型认为所有情感分布在由几个情感维度构成的情感空间里，不同情感间的相似性或差异性通过彼此在情感空间中的距离来表示。所谓情感维度即情感的某些固有属性，情感在这些维度上是连续地、平滑地分布的。维度模型重点关注主观感受部分，对于情感的发生机理也没有明确解释。

最早提出情感维度观的是科学心理学的创始人 Wundt，他认为情感空间由愉快-不愉快、激动-平静和紧张-松弛三个维度构成，每种情感分布在三个维度的两极之间的不同位置上[83]。随后，Schlosberg 通过对面部表情的研究，提出愉快-不愉快、注意-拒绝和激活水平三个维度，如图 2-2 所示，椭圆长轴为快乐维，短轴为注意维，垂直于椭圆面的轴为激活水平维[84]。Izard 则在 Wundt 的三维模型的基础上提出情感的四维说，分别是：愉快度、紧张度、激活度和确信度，愉快度表示主观体验的享乐色调，紧张度形容个体对突发情境缺乏预料和缺少准备的程度，激动度表示兴奋的程度，确信度则用来形容个体胜任、承受感情的程度[85]。Russell 将这些维度简化为愉快度和强度两维，提出情感分类的环状模式，将多种情感排布在二维坐标的不同象限内[86]（图 2-3），如：愉快+高强度=高兴，愉快+中强度=轻松，不愉快+高强度=怕/愤怒，不愉快+中强度=厌烦。可以看出，Russell 的二维理论能区分大部分基本情感，但是对于某些情感却不能做有效区分，如愤怒和怕（恐惧），二者都属于不愉快且强度较高的情感，均处于图 2-3 第三象限的中间位置，距离较近，这对于两种基本情感来说类别差异不够显著。Mehrabian 在此基础上提出 PAD 三维模型[37]，P 代表愉悦度（Pleasure），表示个体情感的正负特性，有时也常被叫做评价维（Valence/Evaluation）；A 代表激活度（Arousal，有时也被写作 Activation），表示个体的神经生理激活水平，与强度、力度等是语义等价的[28]；D 代表优势度（Dominance/Control），有时也翻译为控制度，表示个体对情景和他人的控制状态，优势度的加入成功区分了前两个维度不能有效区分的一些情感，如愤怒属于优势度高的情感，恐惧则相反，优势度较低。

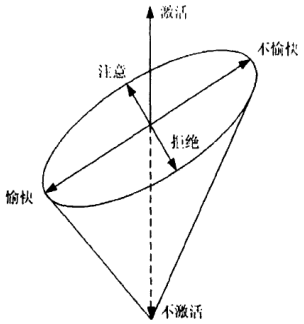


图 2-2 Schlosberg 的三维情感模型

Fig. 2-2 Schlosberg's 3-dimensional emotion model

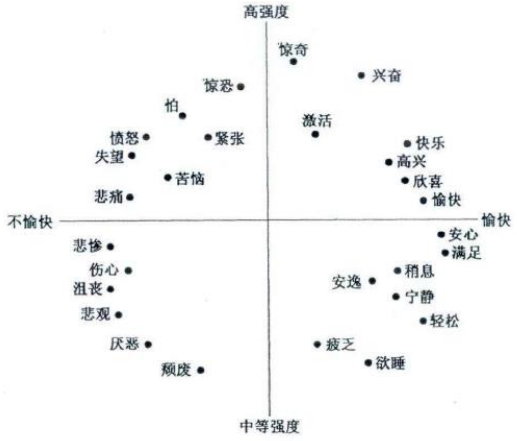


图 2-3 Russell 的环状情感模型

Fig. 2-3 Russell's circular emotion model

维度模型将主观的情感体验量化，相比离散模型具有更好的可操作性。但是，维度模型意图将错综复杂的可能存在层级结构的情感空间映射到仅具有两个或三个维度的超低维空间上，有可能会造成信息的不对等与严重丢失；同时，它只关注情感的心理感受部分在低维空间的投射，忽略了对于情感生成机制的研究。

- 意义导向模型

意义导向模型通过情感词的语义空间结构来实现对于情感的建模，关注重点也是情感的主观体验部分，即心理成分。情感过程涉及机体诸多子系统的变化，主观体验是其中最突出的部分，反映情感最本质的内容[79]。意义导向模型与离散模型有相似的地方，即二者都采用语言标签来标注情感，但离散模型更关注由生物进化决定的行为表现上的共性，所以有了基本情感的概念，意义导向模型则关注由社会文化决定的文化解析模式的相通性，所以用情感词的语义结构来表现心

理感受的潜在结构。

Ortony、Clore 和 Collins 的 OCC 模型[87]虽然很多情况下被认为是基于认知的模型，但同时也是结构化模型，定义了 24 种情感的阶层关系。Fox 提出的三级情感模型，按照情感的主动和被动程度将情感分为不同的等级[88]，如表 2-2 所示。Shaver 和他的同事选用 100 名被试对 135 个常用来表示情感的英语词汇根据相似性聚类，之后利用该人工聚类的结果进行层级聚类，得到了情感词的层级结构[79]。

表 2-2 Fox 的三级情感模型

Table 2-2 Fox's three-layer emotion model

第一级	approach			withdrawal		
第二级	joy	interest	anger	distress	disgust	fear
第三级	pride bliss	concern responsibility	hostility jealousy	misery agony	contempt resentment	horror anxiety

意义导向模型的主要实现形式是情感词的网格分类或树形结构，不仅可以得到离散模型中的基本情感类别，同时各类别间的关系及其内部成员也都显现出来，层级的分布结构既有对于情感的粗分，又有更加细致的划分，研究者可以根据需要自行选择情感的划分尺度，而不需再考虑情感究竟可以分为几类的问题。但是需要注意的是，意义导向模型只关注情感的心理感受部分，对于认知机制和生理、行为反应等都没有涉及，不能体现完整的情感过程。

- 成分模型

成分模型的理论基础是认知评估理论，即情感由对过去经验和当前情境的评估触发，其他成分如生理、表情、动机和感受等的反应模式都由评估结果决定，因此成分模型也被称为基于认知的情感模型。成分模型的关注重点是情感情境的评估和各种反应模式之间的联系，明确了情感的诱发机制是由文化和个体差异决定的对于一系列评价准则的认知，情感的区分机制是由此引发的适应性表情反应、生理反应和行为趋势的不同。

成分模型根据可描述情感的数目是否固定分为基于主题（theme）的模型和基于模式（modal）的模型两种。基于主题的模型情感数目是固定的，每个主题对应相应的情感类型，由于主题的有限导致情感类型的数目受限，情感间的关系也变得相对离散，如上文提到的 OCC 模型，通过对事件的结果、行动代理者、物体的外表等主题的单独评估定义了 24 种离散的情感。基于模式的模型摒弃了情感是相对独立的离散状态的概念，而是将其视作能持续一段时间的情感片段（emotion episode），情感片段的属性由评估结果的模式和由评估结果引发的其他成分的反应模式共同决定，评估准则的交错分布形成了错综复杂的评估结果，评估结果还会引起动机、生理、心理等各方面的反应，反应结果又会反馈给认知从

而改变认知的结果，这一过程的循环进行使情感变得不再离散，情感类型的数目也无从统计。Scherer[23]将这种由特定模式决定的情感定义为模式情感（modal emotion），并给出了成分模型的动态结构，如图 2-4 所示，对于事件的评估结果引起动机的改变，评估结果和动机的变化共同唤起自主神经系统（控制心率、呼吸等）和躯体神经系统（控制面部表情、声音和肢体等）的反应，认知结果、行动倾向、表情和生理反应集中整合到一起表示情感，其中一部分触发有意识的主观体验（即心理感受）可以划分为不同的情感类别并用情感词标注。Scherer 的成分模型的评价维度包含新颖性检查、内在愉悦度检查、目标重要性检查、应对能力检查和相容性检查五项。

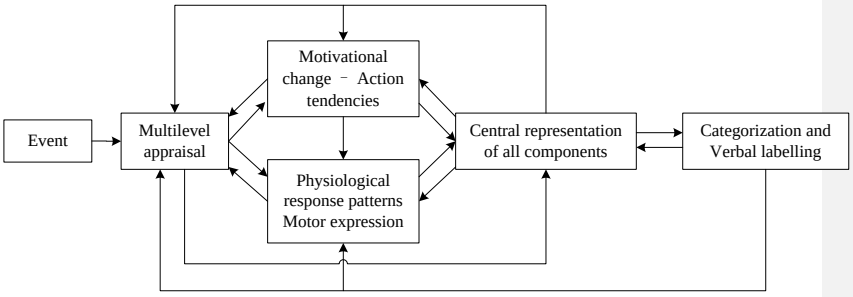


图 2-4 成分模型动态结构图[23]

Fig. 2-4 The dynamic architecture of the component process model[23]

成分模型定义了情感的诱发机制，并将认知评估和各反应模式联系到一起，通过各成分间持续地、双向地作用实现对于情感的分布式的、动态的刻画。模型局限性在于将更多的重心放在认知前端，而对其他部分的反应刻画得不够具体。

综合来看，以上四种模型彼此间存在着渗透，并不总是存在明显的界限，如 Shaver 的层级模型既属于意义导向模型，同时又体现了六种基本情感；OCC 模型既定义了情感的阶层关系，又定义了认知机制的决定作用；Scherer 的成分模型中使用离散情感类别来表示心理成分的主观体验。目前来看，四种模型各有利弊，也各自有自己的关注重点和存在意义，没有强有力证据证明某种模型可以取代其他模型。因此，各模型可以互为补充，相互融合，实现对于情感过程的完整刻画，例如：使用成分模型的情感诱发机制，保留认知对于其他成分的决定作用及各成分间的相互影响，认知内容通过一系列的认知维度表示，心理感受使用离散情感或意义导向的层级模型表示，生理反应则可以映射到诸如“激活度”、“控制度”这样的维度上。

2.2 朗读与播音中的情感研究

心理学关于情感理论和描述方法的研究为我们研究言语情感的产生和建模提供了重要的理论支持和方法指导。但是心理学研究的情感是广义的情感，言语情感作为其中的一个方面，有其一脉相承的共通点，也有由自身特点决定的特殊性。为了解析言语情感的生成过程，我们还借鉴了朗读学[89]和播音学[90]对于情感的触发与控制的研究。

朗读与播音，是把文字语言转化为有声语言的创作活动，其中，播音又分无稿播音和有稿播音两种，本文研究的是文语转换系统中的情感表达，因此这里仅研究有稿播音。播音是在朗读基础上发展起来的以新闻宣传为目的的创作形式，这里将朗读与播音产生的有声语言统称为朗读语音，意在与日常交流所用的自然口语区分。朗读语音与自然口语中的情感虽然都由声学特征体现，但朗读语音中的情感是朗读者或播音员接受语言符号的刺激而产生的能动反应，不同于自然口语中说话人直接接触客观事物的刺激而产生的无意识反应，因此朗读语音中的情感生成过程更接近于文语转换系统中的情感生成，即机器将文档转换成情感语音的过程可以和创作主体从接触文稿刺激到将其转换为携带恰当感情色彩的有声语言的过程进行类比。

从文字语言到有声语言的创作过程，是理解与感受的过程，理解在先、感受在后，但二者又是相辅相成、相互结合、相互渗透的。创作者首先接受稿件文字符号的刺激，理解字形、字音和字义，明确词语、语句和段落间的语法关系和逻辑关系。然后，随着对文字字形、字义的浅近认知，长期经验造成的字形与字音、字义的紧密联系使得创作主体产生了由此及彼、由表及里的某种感受，当稿件中的人、事、物、理被创作主体深切感受到之后，主体所面临的便是按自己的经验和社会价值的需要，做出对稿件内容的评估、判断，旗帜鲜明地肯定应该肯定的东西，否定应该否定的东西，赞颂真、善、美，揭露和抨击假、恶、丑。人对于客观事物和自身状况持有肯定或否定的态度这还不够，必须由这种态度引起以个体的某种心理感觉为特征的体验，并导致机体生理状态的一系列反应。最后，为了将作品引向情感，创作主体调整气息状态，使之与当前精神状态相符，变换高低、强弱、快慢、虚实等声音形式，形成负载着特定思想感情的语音。

2.3 言语情感生成及衍化过程

基于朗读学与播音学中创作主体从接触文稿刺激到将其转换为携带恰当感情色彩的有声语言的创作过程，我们将基于文本生成言语情感的过程概括为：文本分析->认知评价->心理感受->生理反应->发音调整，其中，文本分析为其他四步的生成提供原始的文本特征，后面四步构成言语情感的四种成分，分别对应着

言语情感生成过程中在认知、心理、生理和行为上发生的变化。认知评价是发音人根据自己的经验和社会价值的需要，做出的对稿件内容的评估判断，这种判断除了包含对稿件内容的正负倾向性的判断，还应根据发音人自身所代言的角色，受众的身份，以及所要达到的目的等来进一步把握言语过程中所应持有的态度和分寸；心理感受描述情感的主观体验部分，是个体中央神经系统有意识的心理反应；生理状态描述个体的生理唤起水平及机体对于这些反应的控制；发音调整属于情感的行为成分，发音人根据评估结果和心理、生理的反应调整发音器官，产生携带相应情感特征的语音信号。各成分内容可以看作是从不同视角审视情感，解析了情感的部分内容但又不完整，各方面信息互为补充、又不可替代，它们之间存在相互作用但是又没有一一映射的关系，各成分的模式特征共同组成了言语情感的分布式表达。

言语情感的产生过程是连续的、动态的，各步骤之间存在直接或间接的相互影响，前面步骤的结果会影响后续步骤的反应，后续步骤的反应又会反馈回去进一步影响前面成分的变化。根据认知理论，认知评价是其他步骤的先决条件。认知的结果不止受当前稿件内容的影响，还受发音人过去经验及文化背景的制约，这些因素可能超出了文本内容所涵盖信息的范围。但是在朗读语音合成的背景下，社会背景因素演变为发音人立场、代言角色、受众身份等决定朗读态度和分寸的因素，这些因素属于长期固定的特征，因此这些因素可以暂且不予考虑，目前仅关注由文本刺激引起的认知结果的变化。心理、生理和发音之间也存在直接或间接的相互影响，同时它们的反应也会反馈到认知系统而影响认知的结果。我们将反馈作用归结为对下一时刻情感变化的影响，为了简化计算的需要，暂不考虑反馈作用。图 2-5 给出在不考虑反馈的情况下，从输入文本生成言语情感各部分内容的过程及其内部关系。文本分析的结果作为原始输入依次输送给每个子模块，每个子模块分别对应言语情感一种成分的生成，前面模块的生成结果会对其后续模块产生影响，且影响会逐层累积向后传递。发音方式的描述作为系统最终的情感标注信息输出，可以视作从情感描述到语音声学特征变化的衔接。

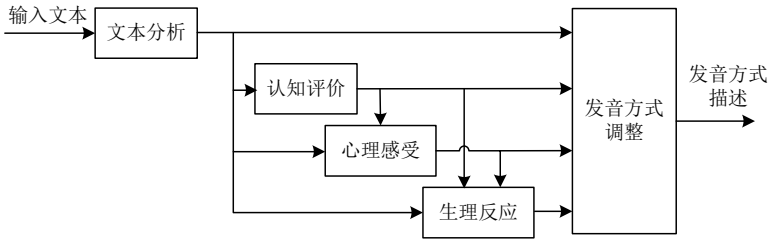


图 2-5 基于文本的言语情感产生过程

Fig. 2-5 The text-based producing process of speech emption

2.3 多视角情感描述体系

根据言语情感生成及衍化过程，言语情感由认知评价、心理感受、生理反应和发音描述四种成分构成，分别从不同视角刻画言语情感某一方面的变化，四种成分共同构成言语情感的多视角描述体系。模型不再拘泥于有限数目的基本情感或几个维度特征，而是形成情感的分布式层级表示，情感的内部构造及其发展衍化的过程也可得到体现。各视角内容的具体表示采用维度表示、离散类别表示和层级表示相结合的方式，每个视角形成一个超平面，而超平面的集合则支撑起一多层结构空间，可用于刻画言语情感内部的复杂关系。

以下分别介绍各视角内容的具体描述方案。

2.3.1 认知评价

已有的研究中，与认知评价关系最紧密的是关于态度的研究。朗读学将态度分为肯定和否定类、严肃和亲切类、祈求和命令类、客观和直露类以及坚定和犹豫类五类[89]。Fujisaki 和 Hirose 将态度分为指示（directive）、确定（confirmative）、疑问（interrogative）、劝告（exhortative）和犹豫（hesitative）五类[91]。顾文涛提出态度的五维说：友好-敌对（Friendly-Hostile）、礼貌-粗鲁（Polite-Rude）、严肃-戏谑（Serious-Joking）、褒扬-贬斥（Praising-Blaming）和确定-不确定（Confident-Uncertain）[92]。基于这些关于态度的描述方法，我们提出针对正负倾向性、正式程度、直露程度、话语温度和话语硬度的五个评价维度，各维度均具有两极性，两极分别为：

- 正负倾向性：否定-肯定；
- 正式程度：非正式-正式；
- 直露程度：委婉-直接；
- 话语温度：冷漠-热情；
- 话语硬度：柔和-刚硬。

出于尽可能全面刻画评价内容的考虑，以上各维度间可能存在交叠。为了探究这些维度间的相互关系，进行了以下两个心理学感知实验：

（1）相关性分析

本实验的目的是探究五个评价维度的相关性，以各维度两极上的词语为评价对象，分别对这些词语针对五个评价维度打分，每极分为“弱”、“中”、“强”三个强度，加上中性，每个维度分为七级刻度，分别用“-3”、“-2”、“-1”、“0”、

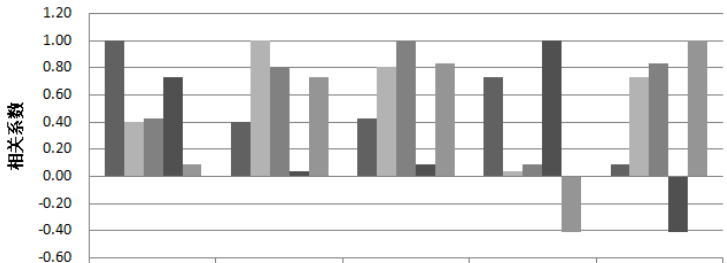
“1”、“2”、“3”表示，以“否定-肯定”维为例，-3~3 分别表示“强否定”、“否定”、“弱否定”、“既无否定也无肯定”、“弱肯定”、“肯定”和“强肯定”。除两极上的词语外，特加入一些与两极词语表意近似，但是程度存在差别的词语，以加深被试对于评价维度的理解。

最终的测试词集为：反对、非正式、否定、强硬、坚定、肯定、柔和、随和、随意、委婉、严肃、犹豫、冷漠、命令、批评、乞求、亲切、热情、赞扬、正式、支持、直接、庄重。

被试为经过播音学专业训练的中国传媒大学播音主持专业的高年级学生，规模为 49 人，男女比例是 19:30，以下实验均相同，不再赘述。通过对 49 份打分结果进行求和、取平均以及相关性分析，得到这几个维度间的相关性分析结果（如图 2-6 所示）。图中用红色椭圆标出相关系数大于 0.5 即相关性较强的值，可以看出：“否定-肯定”与“冷漠-热情”的相关性较高，“非正式-正式”、“委婉-直接”和“柔和-强硬”三者的相关性较高。通过分析，我们认为：

a) “否定-肯定”与“冷漠-热情”主要与话语内容的倾向性判断有关，其中“否定-肯定”又是倾向性的主要决定因素，因此设定“否定-肯定”为话语倾向性主导维度，“冷漠-热情”为辅助维度。

b) “非正式-正式”、“委婉-直接”和“柔和-强硬”主要与话语样式相关，播音学基于文体和节目形式的不同将话语样式分为三类：“宣读式”、“播报式”和“谈话式”，表达技巧主要体现于语言规整性、咬字力度、气息控制、语流速度、声音形式等特征的不同，其中语言规整性是区分这三种话语样式的主要特征，它与“非正式-正式”维所要表达的特征一致，因此设定“非正式-正式”为刻画话语样式的主导维度，“委婉-直接”和“柔和-强硬”为辅助维度，辅助刻画话语样式的更多细节以及态度分寸的把握。



	否定-肯定	非正式-正式	委婉-直接	冷漠-热情	柔和-强硬
否定-肯定	1.00	0.39	0.43	0.73	0.08
非正式-正式	0.39	1.00	0.80	0.03	0.73
委婉-直接	0.43	0.80	1.00	0.08	0.83
冷漠-热情	0.73	0.03	0.08	1.00	-0.42
柔和-强硬	0.08	0.73	0.83	-0.42	1.00

图 2-6 认知评价各维度相关性分析结果

Fig. 2-6 The correlation analysis results between the dimensions of cognitive appraisal

（2）刻画能力验证

以上实验中测试词语围绕评价维度有针对性选取，为了进一步验证评价维度的刻画能力，随机选取日常生活中常用于描述态度的词语作为评价对象，依旧对测试词语针对五个评价维度按七级刻度打分。

测试词语集为：尊敬、轻蔑、冷峻、坚决、犹豫、无礼、傲慢、恭敬、谦卑、和气、庄重、戏谑、夸奖、贬斥、指示、乞求、鼓励、压制、警告、恐吓。

由相关性分析得知，“非正式-正式”维和“否定-肯定”维为两个独立性较高的维度，将测试词语的打分结果投射到由这两维构成的平面上以观察其对不同的态度类型的区分能力，如图 2-7 所示，各词语分布在椭圆构成的区间内，椭圆圆心为被试打分的均值，长短轴半径为标准差。由图 2-7 可以看出，各词语的分布区间基本可以区分开，一些距离较近的词语也可以通过程度的差异进行细分，如犹豫和戏谑，均分布在否定、非正式象限，但是戏谑的否定程度更重；夸奖、鼓励、恭敬、尊敬都是肯定的评价，但是正式程度依次上升，恭敬和尊敬也存在肯定程度的差异；轻蔑、傲慢、压制、恐吓、警告都是否定的态度，但是正式程度上也有所差别。

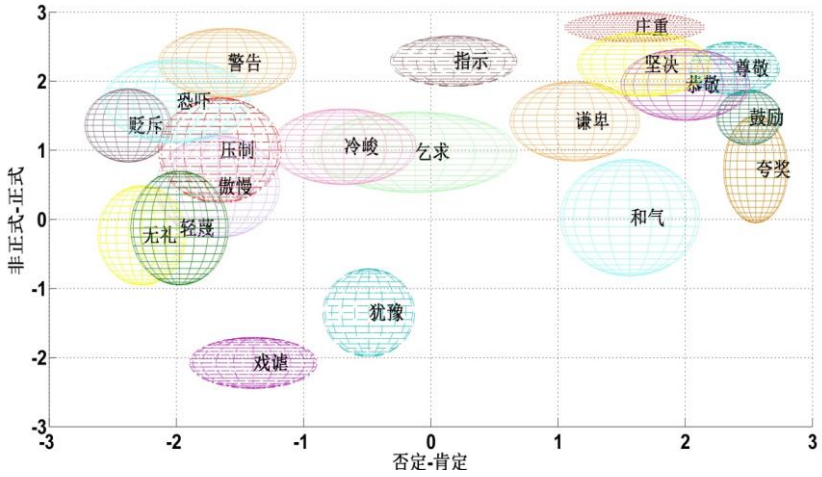


图 2-7 测试词语在“否定-肯定”、“非正式-正式”二维空间分布

Fig. 2-7 The distribution of test words in the “negative-positive” and “informal-formal” space

另外，由词语分布也可以看出维度间的相关性，图 2-7 中词语虽然多集中在一二象限（即“正式”的词语居多），但在“否定-肯定”维分布较平均，正式程度的差异也能辅助区分否定程度或肯定程度相同的词语，因此这两维的关系较为

独立，刻画了认知评价的两个不同的方面。图 2-8 给出测试词语在“否定-肯定”维和“冷漠-热情”维构成的二维空间的分布，可以看出测试词语大多集中在一、三象限，表明否定和冷漠、肯定和热情并存的情况较多，验证了相关性分析中得出的这两维有较强正相关关系的结论；且有些词语出现交叠几乎无法区分，显示了仅依靠这两维刻画能力有限，需要其他维度辅助。

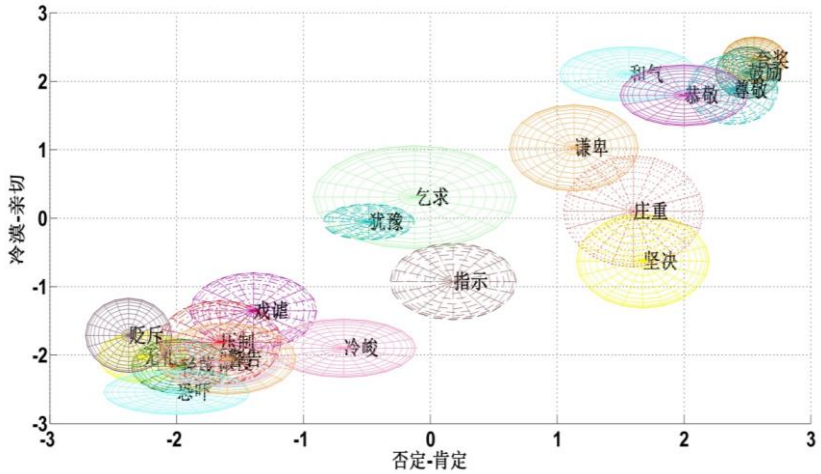


图 2-8 测试词语在“否定-肯定”、“冷漠-热情”二维空间分布

Fig. 2-8 The distribution of test words in the “negative-positive” and “Indifferent-Passionate” space

2.3.2 心理感受

根据对心理学情感模型的分析，心理感受通常用标注情感类别的情感词表示。基于情感词的模型有离散情感模型和意义导向模型，前者倾向于将情感词划分为几个离散类别，我们将其称为离散表示，后者倾向于挖掘情感词的语义空间结构，我们将其称为层级表示。

（1）离散表示

离散模型的依据是其他情感可以由基本情感的组合或者程度变化派生得到。2.1.2 节提到了对于情感的不同划分方法，根据认可度高低排序，得出七种认可度较高的基本情感类别：高兴、悲伤、愤怒、恐惧、惊讶、厌恶和喜欢，我们选用这七种情感作为基本情感。为了验证基本情感的刻画能力，我们挑选了一组常用来形容感受的词语作为测试词语，请被试选出测试词语所包含的基本情感类型，

并给出程度上（“0”-无、“1”-弱、“2”-中、“3”-强）的打分。

测试词语包含：满意、惊愕、寂寞、欢愉、羞愧、悲哀、自豪、妒忌、憎恶、愤恨、惊恐、懊悔、忧愁。

表 2-3 列出了打分的平均结果，为了方便观察每种情感的主要成分，程度相对较弱的成分我们暂且不予显示。

表 2-3 心理感受测试词的基本组成成分（程度均值±标准差）

Table 2-3 The basic components of psychogical feeling test words (mean±std)

	高兴	悲伤	恐惧	惊讶	愤怒	厌恶	喜爱
满意	2.7±0.5	--	--	--	--	--	2.6±0.5
惊愕	--	--	1.2±1	2.7±0.5	0.7±0.8	0.6±0.8	--
寂寞	--	2.2±0.7	0.6±0.7	--	--	1.2±1	--
欢愉	2.7±0.5	--	--	--	--	--	2.5±0.7
羞愧	--	1.1±0.7	--	--	0.7±0.8	1.5±0.9	--
悲哀	--	2.8±0.4	--	--	--	1.2±1	--
自豪	2.6±0.5	--	--	--	--	--	2.4±0.7
妒忌	--	0.7±0.7	--	--	1.3±0.9	2.3±0.7	--
憎恶	--	--	--	--	1.9±1	2.7±0.6	--
愤恨	--	0.9±1.1	--	--	2.7±0.5	2.6±0.6	--
惊恐	--	--	2.7±0.5	2.4±0.7	--	0.9±0.8	--
懊悔	--	2.2±0.7	--	--	--	1.3±1.1	--
忧愁	--	2.3±0.7	--	--	--	0.8±0.8	--

由表 2-3 的结果可以看出：愤恨、憎恶和妒忌的主要成分都是愤怒和厌恶，只是愤怒的程度不同，妒忌的愤怒程度最低，愤恨的愤怒程度最高；懊悔、忧愁、寂寞、羞愧、悲哀的主要成分都是悲伤和厌恶，其中悲哀的悲伤成分最重，忧愁的厌恶成分最轻，羞愧的悲伤成分最轻；满意、欢愉和自豪的主要成分都是高兴和喜爱，且程度相当；惊恐和惊愕的主要成分都是惊讶和恐惧，惊恐的恐惧成分更高。总体来说，使用七种基本情感的组合及程度变化能区分大部分的测试情感，但是同时还发现，一些情感经常同时出现，如高兴和喜爱、恐惧和惊讶、厌恶和愤怒等，因此使人们对于这种情感类别的划分方式的完备性产生质疑。

为了更好的描述情感，我们汇总了汉语中描述心理感受的情感词，并用聚类的办法重新对其类别划分进行研究，得到了下面的层级表示。

（2）层级表示

层级表示的方法相对离散表示更具灵活性和可操作性，一方面层级表示提供不同尺度的划分方式，研究者可根据需要自主选择停留在哪一层上进行研究；另一方面类别间的衍生关系清晰可见，为复合情感的生成提供便利。

关于汉语情感词的汇总及划分，已有一些成果可供借鉴，如吉大的李轶博士

将情感词根据心理感受分为愉快-中性-不愉快三大类共 32 个单一小类和 3 个复合类[93]；北师大的许小颖教授将现代汉语中基于心理感受的 390 个情感词划分为 23 个小类和一个其他类[94]，二人的研究成果在很大程度上存在共通性。这些小类既在一定程度上保证了感受类型的多样性和详尽性，又比直接应用数目庞大的情感词进行分类要省工省力的多，因此我们选择对二人的研究成果进行取并集操作（表 2-4 中斜体部分），汇总后共得到 43 类情感作为即将要聚类的情感词（其中“信”、“疑”两类情感被认为主要与认知评价相关，“激动”与生理激活关系更紧密，所以这里没有采用）。

表 2-4：汉语情感词汇总

Table 2-4 The summary of Chinese emotion words

李铁[65]											
宁静类	*舒适类*	*满足类*	*快乐类*	~~激动类~~	*喜爱类*	*希望类*	*荣耀类*	*尊敬类*	*骄傲类*	*平静类*	*清闲类*
怜悯类	*紧张类*	*惊奇类*	*寂寞类*	*倦怠类*	*沮丧类*	*颓废类*	*耻辱类*	*羞愧类*	*悲哀类*	*烦闷类*	*焦躁类*
怨恨类	*忧愁类*	*轻蔑类*	*懊悔类*	*嫉妒类*	*憎恶类*	*愤怒类*	*恐惧类*	~~复合类~~			
许晓颖、陶建华[66]											
喜 乐	爱	*愁 闷*	悲	*慌*	敬	激动	羞 疚	烦	*急*	傲	*吃惊*
怒	*失望*	*安心*	恨（恶）	嫉	*蔑视*	悔	*委屈*	*谅*	~~信~~	~~疑~~	其他
汇总											
宁静	舒适	体谅	失望	恐惧	愉悦	悲哀	委屈	喜爱	安心	着急	自豪
惊奇	快乐	怨恨	尊敬	吃惊	懊悔	耻辱	希望	嫉妒	羞愧	怜悯	清闲
愁闷	满足	紧张	愤怒	蔑视	倦怠	沮丧	骄傲	寂寞	忧愁	烦闷	憎恶
颓废	轻蔑	惊慌	平静	焦躁	愧疚	荣耀					

请被试人员对上述情感词表进行人工分类，分类规则基于被试自身对于这些词语所属心理感受类型的理解，为了避免分类过粗或者过细，对分类数目进行限定，最少不少于 4 类，最多不多于 10 类，分类完成后给每一类别进行命名以备我们后续参考。经过筛选最终使用的分类结果是 33 份，男女比例为 9:24，筛选原则主要基于被试对实验要求的理解是否准确，给出的划分结果是否遵循实验目的，类别规模是否存在严重不平衡等。

得到人工分类的结果后，首先构建一个可以表示数据集合中两两情感相似性的特征矩阵，由一个对称方阵表示，方阵每一行及每一列均对应情感词表中的一种情感，矩阵元素为该位置对应的两个情感词被归为一类的频数除以总人数（归一化）。利用该相似性矩阵，分别选用 K-means 和层次聚类两种算法进行聚类。

(a) K-means 算法

K-means 算法中，初始聚类中心的好坏会影响聚类效果，初始中心的随机设定也会使聚类结果具有随机性，由于本实验中样本集不大，所以这种影响会比较明显。为了更合理的设定初始中心，我们采用文献[95]提出的方法，即用前 k 个

最大特征值对应的特征向量乘以样本集作为初始中心。K-means 算法还需要预先设定聚类个数，人工聚类时聚类个数限定为 4~10 类，因此这里也将聚类个数分别设为 4~10 类。

观察结果发现分成 4 类的结果已具有一定合理性，如表 2-5 所示。随着聚类个数的增加，有些不希望再拆分的类别被越拆越细，而有些类别的混合度仍然较高。针对这个问题，我们采取了半人工干预的分层聚类的方法，第一层全部数据聚成四类，将平和类作为最终结果保留不参与之后的聚类；第二层其他三类类别内部再各自聚类，聚类个数分别试验了 2、3、4 类，最终得到的聚类结果为八类，即保留平和类，愁闷类拆分为 3 类，喜乐类和惊怒类各自拆为 2 类，如表 2-6 所示：

表 2-5 K-means 算法聚为 4 类结果

Table 2-5 The classification results via K-means algorithm (four classes)

愁闷类	失望 悲哀 委屈 沮丧 怜悯 倦怠 忧愁 烦闷 颓废 懊悔 羞愧 愁闷 愧疚 寂寞
喜乐类	愉悦 骄傲 荣耀 希望 自豪 快乐 尊敬 喜爱
平和类	**舒适 体谅 满足 清闲 平静 宁静 安心**
惊怒类	恐惧 紧张 愤怒 蔑视 轻蔑 惊慌 吃惊 惊奇 憎恶 耻辱 嫉妒 着急 怨恨 焦躁

表 2-6 K-means 算法分层聚类结果

Table 2-6 The layered classification results via K-means algorithm

第一层	第二层	第三层
平和类	平和类	舒适 体谅 满足 清闲 平静 宁静 安心
愁闷类	悔愧类	怜悯 懊悔 羞愧 愧疚
	愁苦类	倦怠 烦闷 颓废 愁闷 寂寞
	悲伤类	失望 悲哀 委屈 沮丧 忧愁
喜乐类	荣耀类	骄傲 荣耀 自豪 尊敬
	喜悦类	愉悦 希望 快乐 喜爱
惊怒类	惊恐类	恐惧 紧张 惊慌 吃惊 惊奇 着急 焦躁
	愤怒类	愤怒 蔑视 轻蔑 憎恶 耻辱 嫉妒 怨恨

(b) 层次聚类算法

层次聚类算法是无需人工干预的自动分层聚类方法，无需预先设定聚类数目。该算法采用二叉树聚类策略，可以得到聚类对象的树形结构，具体实施通过调用 Matlab 工具包中的 4 个聚类相关函数实现：

i. 调用向量距离函数 pdist() 计算两两变量之间的距离，令 Y=pdist(A,'distance')，A 为特征矩阵，'distance'是距离的类型，表 2-7 列出几种距离类型可供选择，Y 为返回的距离矩阵。距离类型的选择由第 iv 步的相似性分析结果决定。

ii. 调用连接函数 linkage()定义变量之间的连接关系，令 Z=linkage(Y)，Y 为距离矩阵，Z 为连接关系矩阵，Z 前两列是索引下标列，用以索引聚类节点，最后一列表示它们之间的距离。linkage 函数计算之后，二叉树式的聚类已经完成，只是 Z 数组可视化不强，还需下面的步骤生成可视化聚类树。

iii. 调用树图绘制函数 dendrogram()绘制可视化聚类树，键入 dendrogram(Z)即可生成二叉树的 figure 文件，Z 为上面得到的连接关系矩阵。

iv. 调用相似性分析函数 cophenetic()评价二叉聚类树与实际距离的相符程度，令 c=cophenetic(Z,Y)，c 为相关系数。表 2-7 是 i 中各类型的距离与聚类树的相关性评价对比，最终选用相关性最高的标准化欧式距离 seuclidean，即欧式距离用标准差归一化。

表 2-7 各类型距离与聚类树的相关性评价

Table 2-7 The correlation evaluation between different types of distances and clustering trees

距离	euclidean	***seuclidean***	cityblock	minkowski	cosine	correlation	hamming	jaccard	chebychev
相关系数	0.8685	***0.9292***	0.7911	0.8685	0.9027	0.9082	0.8658	0.8542	0.8997

图 2-9 即层次聚类算法生成的二叉聚类树，横轴为 43 个候选聚类词，纵轴是它们之间的标准化欧式距离。从图中可以看出，根据距离尺度由大到小的变化，聚类结果呈一种由粗到细的层级分布：最底层类别划分最细致，同一类别内成员间的融合最紧密；越往上类别数目越少，类别间区分性增加，类别内部成员间的融合度降低。

为了使层级结构更加明显，我们对二叉树做了微调，将整棵树调整为六层的层级结构，并给出每一层的类别名称，如图 2-10 所示。最顶层为心理感受；第二层分为积极感受和消极感受，二者叶子成员比例是 16:27，消极感受比例偏大；第三层分为 4 类，积极感受、消极感受各自分为两类，分别是舒畅、同情和哀怨、惊恐，该层的类别内部的复合度依然较高；第四层被进一步分为 7 类，舒畅被分为平和与喜乐，哀怨被分为愁苦、悔愧和厌恶，同情和惊恐保持不变，这一层大部分类别成分已经趋于单一化；第五层在此基础上进一步细分为 19 类，属于积极感受的 8 类，消极感受的 11 类；第六层即用来聚类的候选词，构成 43 个叶子节点。

通常情况下，第四层的划分结果已能保证工程应用的要求。这一结果与离散表示法中的基本情感有一致的地方，如喜乐、愁苦和厌恶可以视作分别对应着基本情感中的乐、哀、恶；此外，对于基本情感中相似性较高的情感进行了合并，如惊讶和恐惧这里合并为惊恐类，愤怒归并到厌恶类，喜爱归并到喜乐类；在此基础上，还对基本情感没有涉及到的情感类型进行了补充，如平和类和悔愧类，这在日常对话中也属于常见的情感类型。

至此，我们通过 K-means 算法和层次聚类算法分别得到了心理感受的分层聚类结果。二者结果有相通性，如：K-means 聚类结果的第二层的大部分类别（平和、悔愧、愁苦、惊恐）与层次聚类的第四层结果重合。由于层次聚类相对于 K-means 算法需要人为参与的部分更少，因此我们更倾向于使用层次聚类算法。

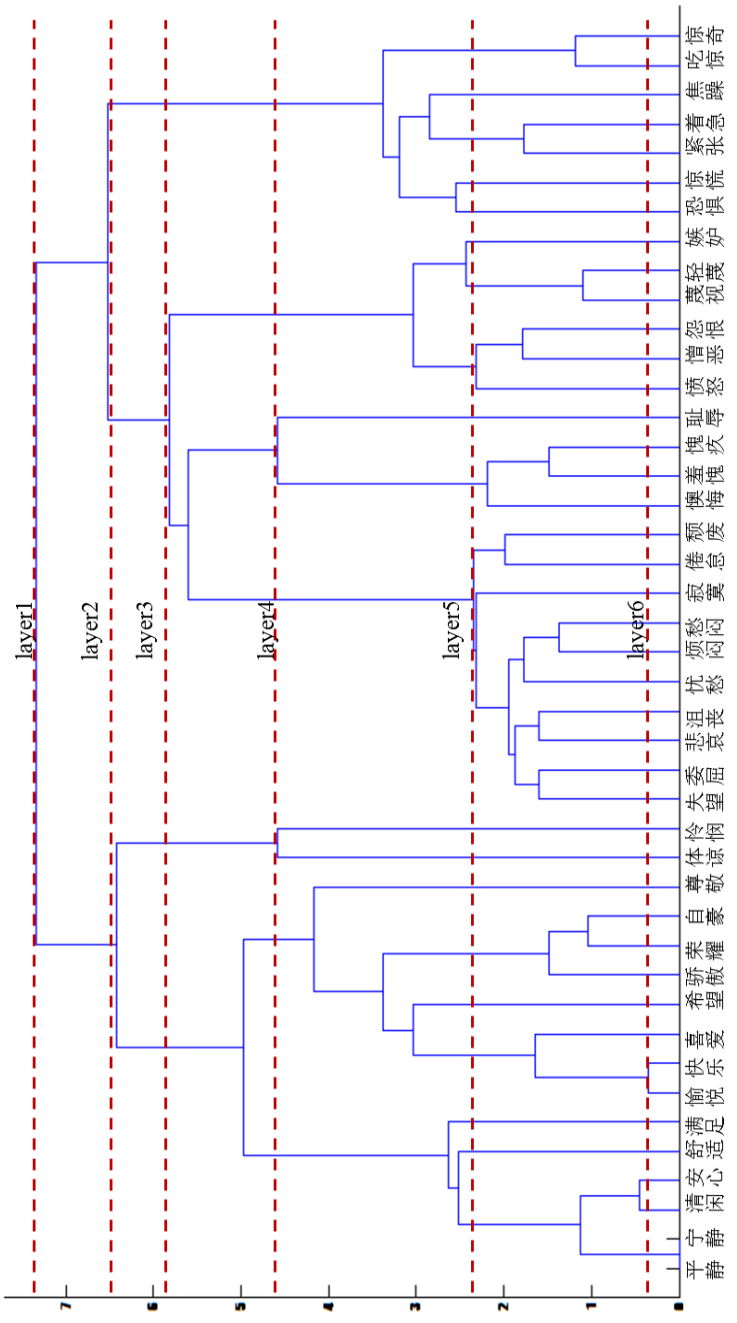


图 2-9 心理感受的层次聚类结果（二叉树）

Fig. 2-9 The hierarchical clustering result (binary tree) of psychological feelings

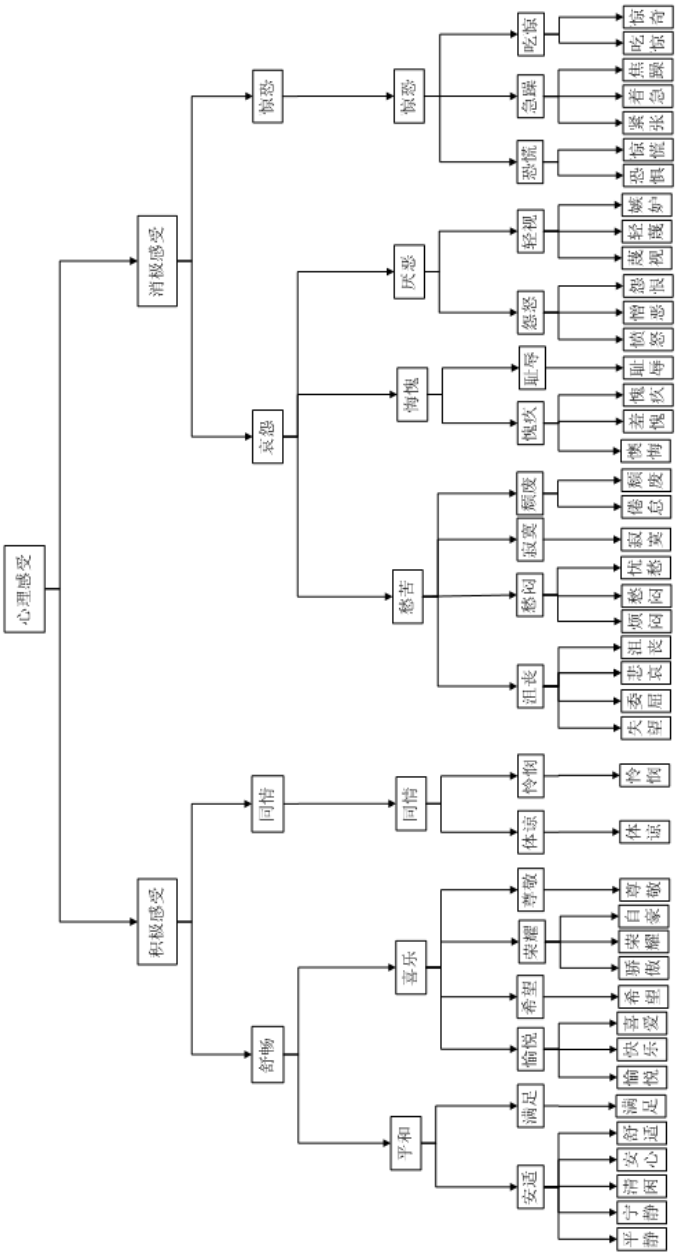


图 2-10 心理感受的层级结构

Fig. 2-10 The hierarchical structure of psychological feelings

2.3.3 生理状态

生理状态描述的是个体的生理唤起水平，因此通常用激活度表示。在此基础上，我们加入另一维度——控制度，注意这里的控制度有别于 PAD 模型中对情景和他人的控制，而是强调个体对于自身生理状态的控制水平。控制度的加入是为了进一步增强对情感的区分能力，如：“紧张”和“慌乱”是两种激活度都比较高的情感，但是“慌乱”要比“紧张”更失控，即控制度更低。

选用一组常用来形容情绪（强调生理上的变化）的词语作为测试词语对这两维的刻画能力进行验证，每个维度分为“0”、“1”、“2”、“3”四个刻度。

测试词语为：平和、激动、紧张、焦躁、烦闷、抑制、轻松、放纵、活跃、郁闷、压抑、失落、慌乱、惊慌失措、镇定。

图 2-11 显示测试词语在“激活度-控制度”构成的二维空间的分布，同样椭圆圆心表示被试打分的均值，长短轴为标准差。可以看出：激活度可以区分大部分情绪，但控制度的加入使得“平和”、“轻松”与“郁闷”、“压抑”等激活度均较低的情感区分开。具体来说：“平和”的激活度和控制度均较低，激活度越高越偏向“激动”，控制度越高越偏向“抑制”；激活度较高、控制度较低情况下是“放纵”、“活跃”，激活度较低、控制度较高情况下是“郁闷”、“压抑”，“失落”相比这两种情绪控制度较低；“烦闷-焦躁-慌乱-惊慌失措”是一组激活度逐渐递增，控制度逐渐递减的情绪，“紧张”与“慌乱”的激活度相当，控制度比“慌乱”高；“镇定”是一种激活度略高于“平和”，控制度又很高的情绪；“轻松”的激活度也略高于“平和”，但控制度与“平和”相当。

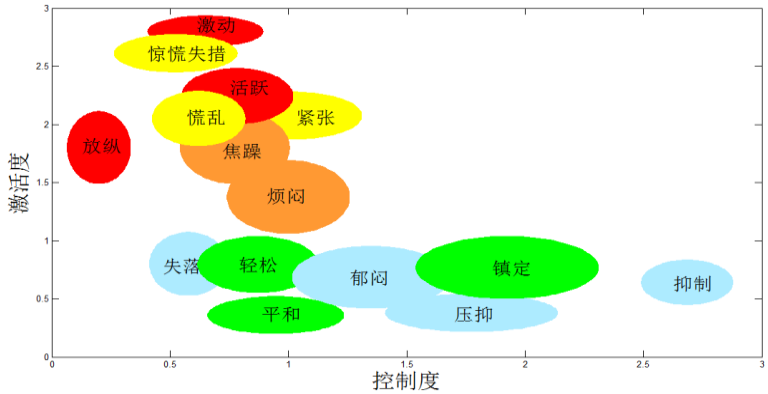


图 2-11 生理状态测试词在“激活度-控制度”二维空间的分布

Fig. 2-11 The distribution of physical states test words in “Activation-Dominance” space

综合来看，整个生理状态描述空间可以分为激活度高-控制度高、激活度高-控制度低、激活度低-控制度高和激活度低-控制度低四大类。目前测试的这些词语在激活度和控制度都很高的区域没有分布，这与人们对测试词语的理解和测试词语的选取都有关系。通过其他三大类以及类别内部程度的差别，可以对测试词语进行刻画和区分，从而验证了这个描述空间对常见情绪状态的刻画能力。

2.3.4 发音描述

发音方式的描述是连接前端抽象的情感信息和后端具体的声学特征的桥梁，相比声学参数可感知性更强，相比认知、心理和生理等特征与声学特征的对应关系更为明确、清晰。

关于不同情感发音方式的描述，在朗读学中，张颂先生将语气的感情色彩的描述为：爱的感情是“气徐声柔”的；憎的感情是“气足声硬”的；悲的感情是“气沉声缓”的；喜的感情是“气满声高”的；惧的感情是“气提声凝”的；欲的感情是“气多声放”的；急的感情是“气短声促”的；冷的感情是“气少声平”的；怒的感情是“气粗声重”的；疑的感情是“气细声黏”的；

在播音学中，有关于声音色彩的描述，声音色彩是人通过听觉所获得的对于一种声音的综合印象。声音色彩的发生和变化，与呼吸、共鸣、吐字等诸器官的运用技巧以及声带的闭合状况及其张力方面有关。刻画声音变化的对比形式包含：高与低、强与弱、实与虚、快与慢、松与紧、刚与柔、纵与收、厚与薄、明与暗。

这些关于发音的描述对于计算机建模来说是抽象的、难以理解的，为实现在合成语音系统中的应用，还需与声学参数对应起来。

Peter Roach[13]曾提出情感语音的语音学描述（phonetic description）体系，包含韵律（prosodic）特征和副语言学（paralinguistic）特征两大类特征：前者包括停顿、基频（高-低、宽-窄）、响度（响-静、渐强-减弱）和语速（快-慢、加快-减慢、清脆-吞吐（clipped -drawled））；后者包括假声（falsetto）、耳语（whisper）、气嗓（breathy）、粗嗓（rough）、挤喉嗓（creak）、鼻音（nasal）等描述音质类型的特征和清喉（clear-throat）、吸气（sniff）、吞咽（gulp）、打哈欠（yawn）、笑、哭、颤音（tremulous）等自然声。

陶建华等学者[1]将情感语音的声学特征分为韵律类、音质类和清晰度类三种：韵律类包含基频（均值、范围、高低线倾斜程度、抖动）、语速、重音突变、停顿连贯性、重音频度和音强；音质类包含呼吸声、明亮度和喉化度；清晰度分为正常、焦急、模糊和准确。

孟子厚[96]统计的音质的主观评价用语包括：清晰-模糊、丰满-干瘪、圆润-粗

糙、明亮-灰暗、柔和-坚硬、融合-散、清澈-浑浊、自然-呆板等，这些用语既有形容人声的，也有形容乐声的。

综合以上不同领域的研究，我们提取出刻画刻画发音方式的七个维度，每一维度与一个或多个声学参数参数相关：

- 音色

  1）暗-明：声音的明亮程度；

  2）瘪-满：发音是否到位，音节的饱满程度。

- 语调

  3）低-高：音调高低；

  4）平-曲：语调变化是否丰富；

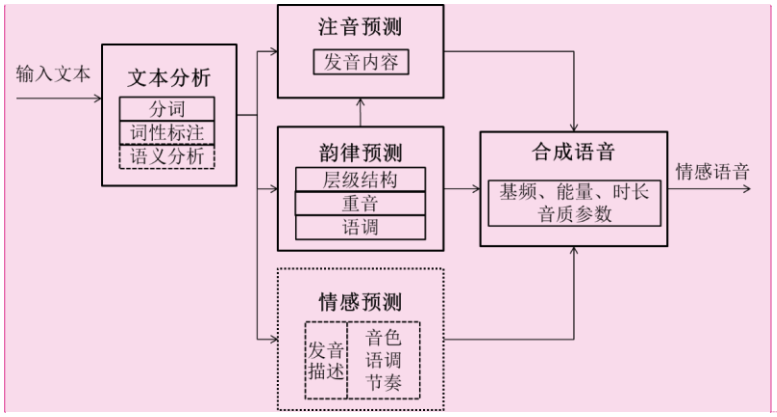
- 节奏

  5）慢-快：话语速度；

  6）散-粘：音节间相互关系，是否黏着；

  7）稳-变：节奏变化是否显著。

发音方式的描述是对情感的语音学表达的总体性描述，独立于个人生理特点和具体言语内容。发音描述将作为情感预测的最终输出内容，如图 2-12 所示，图中实线框为基于参数调整的中性语音合成框架，本文在此基础上，加入基于文本的情感预测（如图 2-12 虚线框所示），通过语义分析，触发认知、心理及生理等一系列子系统的情绪反应，最终体现于发音方式的变化上，发音方式又通过音色、语调和节奏等三个层面的语音学特征描述。关于韵律预测的研究，许多学者做过针对重音、语调合成的实验性研究[]，但当前实用语音合成系统中韵律特征仍主要由韵律层级结构刻画[]。本文通过加入发音方式的描述，与发音内容描述和韵律特征描述共同作用于合成语音模块，实现基频、能量、时长以及音质参数等声学参数的生成。



批注 [w2]: 还是改称 功能语调 更为妥贴

图 2-12 支持本文-情感预测的情感语音合成系统

Fig. 2-12 The expressive speech synthesis system supporting text-based emotion prediction

2.3.5 整体框架

最终确定的描述体系的整体框架如图 2-13 所示：

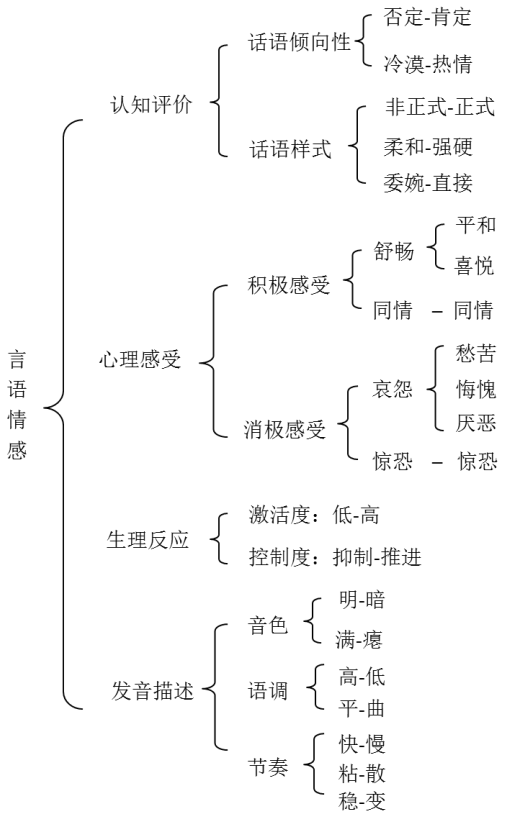


图 2-13 多视角情感描述体系框架

Fig. 2-13 The framework multi-perpective emotion description system

（1）认知评价：对稿件内容的整体倾向性评估，并对要采取的话语样式做出预判。采用五维表示法，其中，“否定-肯定”是话语内容倾向性的主导维度，“冷漠-热情”辅助刻画说话人的态度。话语样式以“非正式-正式”为主导维度，主要负责语言规整性等相关特征的描述，“委婉-直接”和“柔和-强硬”两维辅助刻画话语样式的更多细节，如吐字力度、气息控制等。

（2）心理感受：评估结果引发的心理反应和主观体验。采用离散表示与层级表示相结合的方法，汇总中文情感词中常用来形容心理感受的词语进行层次聚类，得到心理感受的层级表示（图 2-13 因为空间限制只列出部分结果），克服了离散类别表示法类别数目难以确定的问题，根据不同需要可以选择不同的划分粒度，不同类别间的派生关系也得以体现。

（3）生理反应：个体的生理唤起水平以及机体对于这些反应的控制。分别由

激活度和控制度两个维度刻画，注意这里的控制度与心理学模型中所指的对于周围场景的控制有所不同，这里强调个体对自身反应的控制。

（4）发音描述：对于发音方式的估计。针对音色、语调和节奏提取了七个特征维度。音色层包含声音的明亮度和饱满度，语调层包含音高以及语调变化，节奏层包含语速、音节黏着度和节奏变化。发音描述将作为情感预测的最终输出内容，与文本发音内容、韵律结构等信息共同控制声学参数的调整。

不同类别或维度构成了言语情感的向量表示，每个类别或维度又有“无”、“弱”、“中”、“强”不同程度的划分。其中，心理感受和生理状态用“0”、“1”、“2”、“3”四级刻度表示，认知评价和发音描述具有正负区分性，用“-3”、“-2”、“-1”、“0”、“1”、“2”、“3”七级刻度表示。

2.4 本章小结

本章工作主要围绕言语情感生成过程与情感描述问题展开：

（1）以心理学已有的大量情感理论和情感模型为参考，结合朗读学和播音学中关于情感的研究，将文语转换过程与朗读者或播音员将文稿转换成有声语言的过程类比，将基于文本的言语情感生成及衍化过程概括为文本分析、认知评价、心理感受、生理反应和发音调整等几步，各步之间存在直接或间接的相互影响，前面步骤的结果会影响后续步骤的反应，后续步骤的反应又会反馈回去进一步影响前面成分的变化；

（2）基于言语情感生成过程，提出了情感的多视角描述体系，包含认知评价、心理感受、生理反应和发音描述四种成分，分别从不同视角解读言语情感的不同方面，各视角互为补充、缺一不可，共同组成了言语情感的分布式表达；各视角具体表示采用维度表示、离散类别表示和层级表示相结合的方式，每个视角形成一个超平面，而超平面的集合则支撑起一多层结构空间，可用于刻画言语情感内部的复杂关系。

3 新闻言语情感数据库构建

基于第 2 章提出的多视角言语情感描述体系，我们构建了新闻言语情感数据库，一方面用于验证言语情感描述体系的合理性，一方面为下一章的情感预测模型的训练提供数据支持。

3.1 概述

情感语音按采集方式不同分为自然语音（spontaneous speech）、诱导语音（elicited speech）和表演语音（acted speech）。

自然语音情感的真实性、自然度都最好，但是采集起来也最困难。最理想的采集方式是在说话人不知情的情况下进行录音，此时说话人是完全放松的，情感表达最自然，但是这样做可能触及侵犯个人隐私等法律问题。另一方面，对于情感类别、发音内容、发音质量的控制等也不好掌控，因此要收集一定规模的情感覆盖广泛且均衡、语音质量又清晰可用的自然语音数据是非常困难的事。

诱导语音的采集通过让说话人在录音前接受一定形式的情感刺激而采集其应激情感，诱导方式包含文本诱发、音乐诱发、电影诱发、图片诱发和自我想象等。诱导语音的采集试图保留情感的真实性并实现对于情感类型的控制，但是由于个体接受刺激的程度及反应方式不同，情感的可控性仍然不强。

表演语音则是由专业的播音员或演员通过对某种情感的主观模仿获得。获取表演语音的途径有两种：一种是通过对电影、电视或广播的音频文件进行剪辑截取需要的情感类型，这种做法可以一定程度上保证情感的自然度，但是难以满足对于发音内容和语音质量的要求；另一种途径是确定好发音内容让专业播音人员朗读并在安静环境下进行录制，这种方法对情感类型、发音内容、发音人性别和录音环境等方面的要求都容易满足，是目前最常用的情感语音库构建方式。

本文使用的数据库是基于表演语音的形式构建。虽然一些学者质疑表演语音中的情感不是真实的情感，带有夸大成分，但是情感的真实性和可控性是一对难以调和的矛盾，我们试图从语料形式和发音人两方面来缓解这种矛盾。语料形式选择新闻言语，相对自然语言更规整，表达模式也相对固定，有规律可循；话语样式涵盖播报、宣讲、评论等各种风格，一定程度上满足情感丰富多样性的需要。发音人选择经过多年训练的专业播音员，情感的触发相比普通人更加稳定、可靠，具有可重复性；备稿过程等播音领域成熟的实践经验还可为我们的研究提供指导。

数据库的搭建流程包含以下几步：

(1) 语料准备阶段：包括播音稿件的收集与筛选；
(2) 语音数据采集：对筛选出的播音稿进行转录；
(3) 情感信息标注：基于本文提出的多视角情感描述体系对稿件进行情感标注；
(4) 标注数据处理：对多人标注结果进行整合以确定每篇稿件最终的情感标注类型。

接下来分别对每步进行具体介绍。

3.2 语料准备阶段

语料来自中央人民广播电台《新闻与报纸摘要》节目的播音稿，以篇章为单位，总共收录 29109 篇新闻播音稿。时间跨度从 2006 年到 2013 年，内容涵盖教育、医疗、科学、经济、文化、艺术、娱乐、军事以及政治等各领域。播音语体覆盖宣告、讣告、播报、通讯、评论五种：宣告指全文播发的中央会议公报、人员名单等，也包括讣告，因为讣告较为特殊所以单列为一类；播报指一般消息播报；通讯是对事件、人物的专题报道，较之于消息播报更加详细具体、形象生动；评论是对事物进行评论、发表看法，节目中一般指全文播发的评论文章、短评、解读等。

筛选工作分两步进行：

（1）首先筛选出长度适中的文档。一方面，因为后期应用中会涉及到篇章、段落、句子等不同尺度的情感预测，因此文档长度不能过短，否则篇章、段落和句子的差别不明显，无法观察篇章内部情感的衍化；另一方面，篇幅过长的稿件易造成播音员在录制过程中出现疲惫或松懈的状态，导致播音状态不稳定或者情感体验失真，因此文档长度也不宜过长。

最终保留段落数目大于 2 小于 6 的篇章，共筛选出 8600 篇长度适中的文档留作文本情感预测使用，包含 64568 句，平均每篇包含大约 8 句，每句平均包含 55 个汉字。

（2）录音材料筛选。因为此时还没有情感类型的标注信息，只能通过语料内容与播音语体对可能的情感类型及程度进行估计，尽可能做到情感类型的丰富且涉及各种程度（包括中性）。

最终从 8600 篇长度适中的文档中挑选出 600 篇作为准备录音的材料，各语体与内容的分布如表 3-1、3-2 所示。

表 3-1 600 篇录音语料的各语体篇章分布

Table 3-1 The distribution of the 600 chapters with different styles

语体	播报	讣告	评论	通讯	宣告
篇章数	519	9	21	48	3

表 3-2 600 篇录音语料的各种内容篇章分布

Table 3-2 The distribution of the 600 chapters with different contents

内容	计算机与因特网	教育	区域	人文与艺术	商业与经济	社会科学	社会与文化	新闻与媒体	医疗与健康	娱乐与休闲	政府与政治	自然科学
篇章数	7	33	26	5	67	5	316	8	29	12	80	12

3.3 语音数据采集

录音工作在中国科学院声学所消声室完成，发音人为中央人民广播电台有二十多年播音经验的资深播音员郑岚女士。录音过程分为备稿与正式录音两个阶段，其中备稿阶段是播音员理解稿件并触发感情的重要阶段，包括划分层次、概况主题、联系背景、明确目的、找出重点和确定基调等步骤。正式录音过程在实验人员监听下完成，主要监听播音员嗓音的状态，确保发音状态自始至终保持一致。

录音文件以 wav 格式保存，采样率为 44100Hz，精度为 16bit，单声道，总时长 5 小时 26 分 48 秒，共 1.61GB。

3.4 情感信息标注

从上述 600 篇经过录音的语料中再筛选出 150 篇有代表性的语篇进行人工标注。

标注人员由本实验室同一课题组的三名师生担任。

标注内容为第 2 章所提出的多视角情感描述体系的各成分的具体内容，每个维度或每个极性分为“无”、“弱”、“中”、“强”四种程度，心理感受和生理状态为单极维度，认知评价和发音描述的维度具有正负极性，因此由“-3”、“-2”、“-1”、“0”、“1”、“2”、“3”七级刻度表示。心理感受分为不同的层级，由于低层是高层的细分，高层由低层合并而来，所以从低层的标注一定程度上可以推出高层的标注，反之则不一定成立。因此为了减少标注负担，心理感受部分只标注最底层的维度，上层维度的数值通过对相应子层成员取最大值的操作获得。认知评价中的“委婉-直接”和发音描述中的“散-粘”两个维度在当前的新闻语料中变化不明显，因此也暂不予以标注。

标注单元分为篇章、段落和句子三个尺度，即每篇文档需要标注这三个尺度单

元下的情感内容，标注人员可以自行选择标注顺序，出于减少工作量的考虑，建议标注人员先标出篇章级的情感，在段落和句子级只需要标出与篇章情感不同的地方即可。

标注过程分为训练阶段和正式阶段两部分，训练阶段先选出 10 篇进行独立标注，之后几名标注人员比较结果并商讨标注细则，当观点基本一致时进行大规模正式标注。三份标注结果通过下一节的介绍方法汇总整合成每篇文档最终的标注结果。三份标注结果在不同尺度的一致率如表 3-3 所示。从表中可以看出，三个尺度的标注一致率近似，篇章级一致率略高于其他两级；三人之间两两的一致率好于三人共同的一致率，但三人共同的一致率也接近 80%，达到可用的水平。

表 3-3 标注结果一致率统计，“A”、“B”、“C”分别代表三位标注人员。

Table 3-3 The agreements analysis of the annotations. “A”, “B”, and “C” stand for the three annotators.

	A&B	A&C	B&C	A&B&C
篇章级	0.8264	0.843	0.8525	0.7821
段落级	0.8275	0.8392	0.8483	0.7798
句子级	0.8267	0.8378	0.8468	0.7788
全部	0.8269	0.8389	0.848	0.7795

3.5 标注数据处理

三份标注结果参考 Schuller 等人的提出方法[97]加权整合在一起。简单来说，首先计算每个维度三人标注结果的均值使其作为基准值；然后分别计算每人的标注结果与基准值的相关系数作为衡量其标注可信度的指标；最后将相关系数归一化作为个人标注的权重加权求和得到最终确定的标注结果。与之不同的是，本文中的标注数据含有很多标注值全为零的维度（主要出现在心理感受部分，表示语料中不含该维度所代表的情感成分，这与新闻语料属于领域相关的语料有关），标注值全为零意味着相关系数的分母会出现 0，因此此时无法计算相关系数。我们选择高斯权重来代替相关系数。高斯权重与相关系数一样也可以调整个人标注对于整合结果的贡献率，加强与均值接近的数值的贡献，抑制与均值偏差较大的数值的作用。相比直接采用算术平均值的做法，高斯加权可以削弱不可靠标注在整合结果中的比重，而算术平均值每个人所占比重没有差别。高斯权重及归一化加权的计算如公式（3-1）和公式（3-2）所示：

$$GW_k(i,j) = exp(-dist_{euc}\big(L_k(i,j), M(i,j)\big) / (2\sigma^2)) \tag{3-1}$$

$$L_{ad}(i,j) = (\sum_k GW_k(i,j) \times L_k(i,j)) / \sum_k GW_k(i,j) \tag{3-2}$$

其中， $L_k(i,j)$ 表示第 k 个标注人员对于第 i 个样本、第 j 维的标注结果，$M(i,j)$ 为三人标注结果的均值，$dist_{euc}$ 是单人标注结果与三人标注均值的欧式距离，σ 是尺度参数，$GW_k(i,j)$ 为相应的高斯权重值，$L_{ad}(i,j)$ 为个人标注结果归一化加权之后得到的调整之后的结果，作为文档最终的标注结果。σ 参数可以调节高斯权重与标注偏差的曲线分布，我们试验了 0.1 到 2 范围内的几个数值，如图 3-1 所示，最终将 σ 设定为 0.2，从图 3-1 可以看出 $\sigma = 0.2$ 时，与标注均值的偏差大于 1 的数值高斯权重已经变得很小，从而抑制了与偏差较大的标注对于整体的贡献。

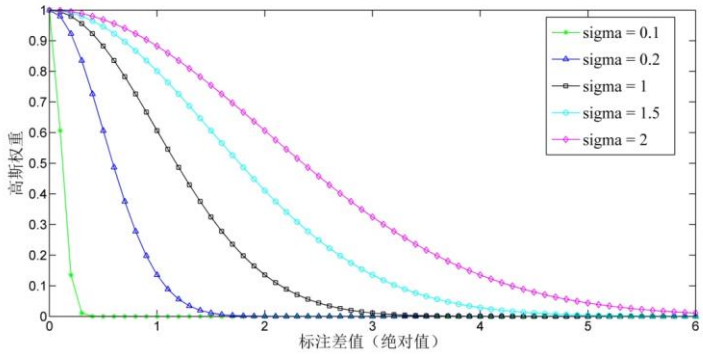


图 3-1 不同尺度参数的高斯权重曲线分布。其中，横轴为个人标注与均值的差值的绝对值，纵轴为高斯权值。

Fig. 3-1 The Gaussian weight curves varying with different scale parameters (sigma), where the horizontal axis represents the distance between the individual annotation and their mean value, and the vertical axis represents the Gaussian weight value.

3.6 标注结果分析

对整合后的标注数据进行统计学分析发现，标注结果存在两方面的分布特征：

（1）稀疏性：这一特征主要出现在心理感受部分，上面提到，心理感受标注最底层，共 43 种感受，每篇文档只包含这之中的几种感受，其他都为 0，因此这部分标注数据非零率很低（仅为 2.45%），即非常稀疏。分析原因可能由于底层分类过细，且有些感受在新闻语料中很少出现造成。稀疏性特征会影响前期数据的处理，比如无法计算相关系数，以及后期预测结果的测评指标的选择，比如若选择正确率（accuracy）作为测评指标，即使预测数值全为零，正确率还是会很高（97%以上），而这显然是一种很差的预测结果，因此我们根据数据特点选择更能体现系统性能的评价指标。

（2）非均衡性：心理感受部分大量全零维度（即此种感受没有样本）的存在一定程度上体现了语料在心理感受的类型间的分布不均衡。除了零与非零的区别，非零数值还有强度上的差别。图 3-2 画出了心理感受之外其他三种成分各维度不同强度值的样本分布（不包含“委婉-直接”和 “散-粘”两维）。可以看出：每个维度上不同强度间的样本不均衡分布是普遍存在的，有些维度集中在一到两种强度且一种强度占明显优势，如“激活度”、“控制度”、“瘪-满”；有些维度分布相对均匀但也不是所有强度都包含，如“否定-肯定”、“非正式-正式”和“柔和-强硬”。总体来看，认知评价部分的样本均衡性较好，生理反应和发音都易出现一种强度占绝大多数的情况，说明就新闻语料而言，认知评价的类型相对丰富，生理反应和发音方式相对比较单一，如控制度大多处于略微有所控制的状态；此外，各维度最强程度的极值都很少出现，强度为 1 的居多。分析出现这种现象的原因，除了与新闻言语特有的发音风格有关，还可能与标注人员的感知特点有关。心理学发现情感的感知与个体年龄、受教育程度和文化背景有关，随着年龄和文化程度的增长，人对于情绪的体验也趋于淡化，且东方文化崇尚中庸之道，凡事讲究谦和、内敛，因此三名标注人员给出的标注结果强度均趋于平和，程度极强的情况较少。此外，情感的类型也会影响其感知，比如有研究发现，出于生物进化的需要，人们对某些与生存相关的情感更为敏感[57]，因此可能会影响对某些情感类型的感知。

针对样本分布不均衡的问题，在之后的情感预测中会根据算法的特点进行均衡性调整，具体可参见第 4 章。

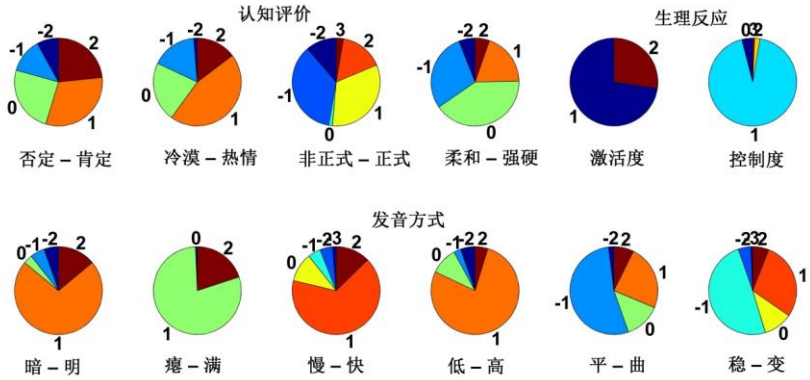


图 3-2 言语情感部分成分各维度不同强度值的样本分布

Fig. 3-2 The distribution of the samples of different intensities on each dimension in the speech emotion model

3.7 本章小结

本章主要工作是采用表演语音的形式构建新闻言语情感数据库。构建流程包括语料收集与筛选、语音数据采集、情感信息标注和标注数据处理几步，得到的结果有：

（1）总共收集了 29109 篇新闻播音稿；

（2）筛选出 8600 篇长度适中的语篇作为言语情感预测模型的训练语料；

（3）从 8600 篇中选出 600 篇聘请专业播音员在录音棚进行转录；

（4）从 600 篇中选出 150 篇进行多人人工情感标注，并对标注数据高斯加权整合为最终的标注结果；

（5）最后分析了标注结果的统计特性，发现其数据稀疏性和样本非均衡分布的特点，在之后的应用中会采取相应的措施进行调整。

4 基于深度神经网络的言语情感预测模型

在本文提出的言语情感描述方案中，言语情感的描述是从多视角进行的，影响情感的因素来自多个层面，由此归纳得到的言语情感预测模型自然而然地是结构化的。因此，为了处理层次化、多尺度特征，拟采用结构化的深度神经网络来构建言语情感预测模型。本章主要进行同一尺度下不同情感成分间的结构化建模，解决言语情感生成过程中动态衍化过程的建模，多尺度情感特征的融合将在下一章讨论。

4.1 问题分析

近年来，深度神经网络因其在计算机视觉、语音识别和自然语言处理等领域获得的成功而被广泛关注。深度神经网络具有多层非线性映射结构，通过低层特征到高层特征的逐层抽象可以得到数据的分布式特征，与我们提出的言语情感的多层分布式表示方式相契合。此外，受限波尔兹曼机（Restricted Boltzmann Machine，RBM）的引入使其在训练阶段可以利用大量的无标数据得到网络的初始参数，之后只需利用少量有标数据进行微调即可，从而实现网络的半监督训练。因此，我们选用深度神经网络作为言语情感预测的基础模型。

深度神经网络中间层的含义是未知的、不可见的，因此常被称为“隐含层”。本文提出的言语情感的多层表示对每层的含义及其相互关系都进行了明确设定，如何人工干预深度神经网络的内部结构，约束网络的学习方向，成为我们需要解决的问题。对于深度神经网络中间层的研究已有一定的工作基础，如：文献[98]在传统卷积神经网络的基础上，通过引入联合目标函数实现对包含输出层和中间层在内每一层的监督；文献[99]将无监督的聚类算法嵌入到深度网络有监督的判别任务中，实现对输出层和中间任一层的半监督学习；文献[100]提出一种新的深度神经网络——深度堆叠网络（Deep Stacking Network，DSN），网络由一系列具有相同或相似结构的单隐层神经网络模块堆叠构成，区别于传统神经网络的整体训练，DSN的每个模块均可单独有监督训练，并通过将前一模块输出累加到下一模块作为部分输入的方式实现对前面模块训练结果的继承与利用。对中间层监督力度的加强，有助于为目标任务提取到更具区分性的特征，也一定程度上缓解了多层神经网络训练过程中易陷入局部最优的难题。然而，这距中间层真正意义上的可见化还有一定距离，中间层的含义仍然是未知的，我们仍难以干预深度网络内部的学习过程并对学习结果给出解释。

基于深度堆叠网络，我们搭建了可以与言语情感多视角描述体系相整合的中间层部分可见的深度堆叠网络（Visible Deep Stacking Network，VDSN），利用多视角情感模型提供的先验知识对深度网络的内部结构进行部分人工干预或引导，赋予中间层具体的含义和显性的相互关系，使其变得可见化或部分可见，部分可见是为了保留对未知信息的提取能力和一定的容错能力。根据堆叠位置的不同，VDSN又分为输入层部分可见的深度堆叠网络（Input-layer Visible Deep Stacking Network，IVDSN）和隐含层部分可见的深度堆叠网络（Hidden-layer Visible Deep Stacking Network，HVDSN）。我们认为，中间层的可见化有助于深度网络性能的提升，尤其在标注数据有限的情况下，先验知识的引入可以有效地扩充可利用信息。本章我们会通过实验对网络性能进行测试，验证所提方法对于深度网络的优化效果，以及利用该模型进行言语情感预测的有效性和多视角情感描述体系的合理性。

接下来将介绍模型的网络结构和训练算法，以及文本特征的提取步骤和网络性能的优化措施，最后将给出利用所提网络进行言语情感预测的实验和结果分析。

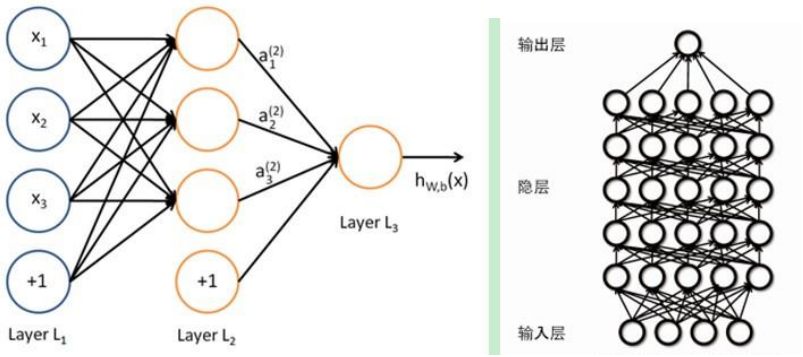
4.2 网络结构

4.2.1 深度神经网络 DNN

深度神经网络（Deep Neural Network，DNN），从字面理解就是深层次的人工神经网络，从人工神经网络发展而来。人工神经网络（Artificial Neural Network，ANN）是一种利用类似于大脑神经突触连接的结构进行信息处理的数学模型，在工程与学术界也常直接简称为神经网络或类神经网络。神经网络由大量的简单基本元件——“神经元”相互连接而成，每个神经元作为网络的一个节点，代表一种特定的输出函数，称为激活函数或激励函数（Activation Function）。每两个节点间的连接都代表一个对通过该连接信号的加权，相当于人工神经网络的记忆。按照生物神经元的特性，每个神经元有一个阈值，当该神经元所获得的输入信号的累积效果超过该阈值时，它就处于激发态；否则应处于抑制态。由此，最直接的激活函数是阈值函数，又称为阶跃函数。但为了使用反向传播算法进行有效学习，激活函数必须限制为可微函数。最常用的非线性连续可导的激活函数是 Sigmoid 函数，值域为[0,1]；当将 Sigmoid 函数稍作变换，还可以得到值域为[-1,1]的双曲正切函数。二者都能实现逻辑回归的映射关系，本文中使用 Sigmoid 函数作为激活函数。

多层感知机（Multilayer Perceptron，MLP） 是一种前向结构的人工神经网络，映射一组输入向量到一组输出向量。MLP 可以被看作是一个有向图，由多个节点层组成，每一层都全连接到下一层。除了输入节点，每个节点都是一个具有非线性

激活函数的神经元。图 4-1 给出多层感知机的示意图，包含输入层、输出层和隐含层三部分：输入层接受大量输入信息但不进行非线性计算；输出层汇总了对网络中输入层源节点产生的激励模式的全部响应；输入层与输出层之间众多神经单元和链接组成的各个层面由于不能在训练样本集中观测到它们的值而被称为“隐含层”。隐含层可以有一层或多层，当仅包含一层隐含层时（如图 4-1(a)所示），网络与传统隐马尔可夫模型（Hidden Markov Model，HMM）、条件随机场（Conditional Random Field，CRF）、最大熵模型（Maximum Entropy，MaxEnt）和支持向量机（Support Vector Machine，SVM）等一起统称为浅层学习结构，它们的共性是仅含单个将原始输入信号转换到特定问题空间的简单结构，在有限样本和计算单元情况下对复杂函数的表示能力有限，针对复杂分类问题其泛化能力受到一定制约。含多隐层的多层感知机构成一种深度学习结构（如图 4-1(b)所示）。



(a) 含单隐含层的浅层神经网络 (b) 含多隐含层的深度神经网络

(a) Shallow neural network with one hidden layer (b) Deep neural network with multiple hidden layers

图 4-1 多层感知机

Fig. 4-1 Multilayer Perceptron

深度学习结构可以很简洁地表示复杂函数；还可以通过组合低层特征形成更加抽象的高层表示（属性类别或特征），从而得到数据的分布式特征表示；此外，深度网络还展现了强大的从少数样本中学习数据集本质特征或潜在结构的能力。

常用的训练神经网络的算法是反向传播（Back-propagation，BP）算法，通过使输出层与目标值的误差最小反向逐层调整各层间的联系权重，但该误差函数是一个含多个极小值的高度非凸空间，因此在沿梯度下降的方向搜寻误差极小值时常使网络收敛于局部最小，尤其是从远离最优区域开始时（随机值初始化会导致这种情况的发生）；其次，在利用 BP 算法向其它层反向传播该梯度时，随着网络层数的增加，反向传播的梯度（从输出层到网络的最初几层）幅度值会急剧地减小，这使得最初几层的权重变化非常缓慢，以至于它们不能够从样本中进行有效

的学习，从而使得整个网络的性能与仅由最后几层组成的浅层网络性能相似，这种问题通常被称为“梯度的扩散（Gradient Diffusion）”；此外，BP 算法要求有监督训练，即需要标注语料，然而有标签的数据通常是稀缺的，因此对于许多问题，我们很难获得足够多的样本来拟合一个复杂模型的参数，而考虑到深度网络具有强大的表达能力，在不充足的数据上进行训练将会导致过拟合。这些原因在一段时期制约了深度神经网络的运用与发展，并使很多机器学习和信号处理的研究从深度网络转移到相对较容易训练的浅层学习结构。

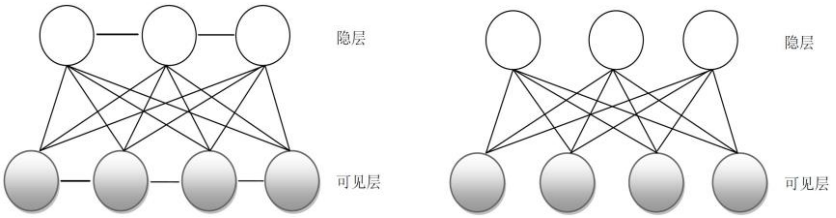
4.2.2 深度置信网络 DBN

2006 年，加拿大多伦多大学教授 Geoffrey Hinton 和他的学生 Ruslan Salakhutdinov 提出一种新的深度神经网络——深度置信网络（Deep Belief Network，DBN）[50]，为解决深度结构的优化难题带来希望。从结构上讲，DBN 与传统的多层感知机没有明显区别，并且在做有监督学习时也基于同样的算法（BP 算法）。不同的是，该网络在做有监督学习前要先做非监督预训练，以非监督学习到的网络参数作为初始值进行基于有标数据的微调。预训练的引入有效提高了网络的收敛速度，改善了网络易陷入局部最优的局面，从而使深度网络的广泛应用成为可能。

关于预训练有助于提升网络收敛速度的原因，最直接的解释是预训练为网络参数训练到一组合适的初始值，从这组初始值出发会令代价函数达到一个更低的值。另一种解释是，DBN 是一种生成型神经网络，通过引入无监督预训练可以得到观测数据的先验概率 P(x)，先验概率的获取有助于后验概率的学习，而传统的区分型神经网络只能通过有监督学习直接估计后验概率，相比之下，先验的学习可以利用大量的无标数据学习和发现数据中存在的模式，有助于避免因网络函数表达能力过强而出现过拟合情况。

DBN 由一系列受限波尔兹曼机（Restricted Boltzmann Machine，RBM）组成。RBM[101]是一种典型神经网络，源自玻尔兹曼机（Boltzmann Machine，BM）[102]。BM 是 Hinton 和 Sejnowski 于 1986 年提出的一种根植于统计力学的随机神经网络。该网络的神经元是随机神经元，神经元的输出只有两种状态（未激活、激活），一般用二进制的 0 和 1 表示，状态的取值根据概率统计法则决定。从功能上讲，BM 是由随机神经元全连接组成的反馈神经网络，且对称连接，无自反馈，包含一个可见层 **v** 和一个隐层 **h**，如图 4-2(a)所示。BM 具有强大的无监督学习能力，能够学习数据中复杂的规则，但是代价是训练时间长，且无法确切地计算 BM 所表示的分布，甚至得到服从 BM 所表示分布的随机样本也很困难。为此，Smolensky 提出了受限的玻尔兹曼机 RBM，与 BM 一样 RBM 也具有一个可见层一个隐含层，层

间对称连接，但是层内无连接，其结构如图 4-2(b)所示。RBM 具有很好的性质：在给定可见层的单元状态（输入数据）时，各隐单元的激活条件独立；反之，给定隐单元状态时，可见层单元的激活亦条件独立。同时又由于所有的 **v** 和 **h** 满足 Boltzmann 分布，因此，当输入 **v** 的时候，通过 $p(\mathbf{h}|\mathbf{v})$ 可以得到 **h**，而得到 **h** 之后，通过 $p(\mathbf{v}|\mathbf{h})$ 又能得到可视层。连接隐层单元与可视层单元的权重值即为待训练参数。在训练过程中，首先将可视向量值映射给隐单元；然后可视单元由隐层单元重建；这些新可视单元再次映射给隐单元，这样就获取新的隐单元。执行这种反复步骤叫做吉布斯采样（Gibbs Sampling），而隐层激活单元和可视层输入之间的相关性差别就作为权值更新的主要依据。具体训练过程将在下一节进行详细介绍。



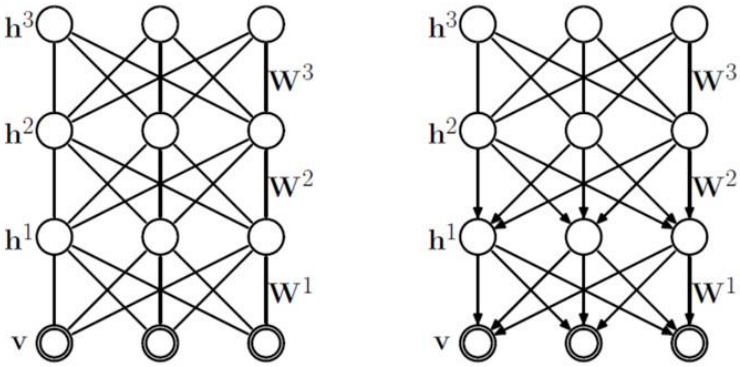
(a) BM 模型 (b) RBM 模型

(a) BM model (b) RBM model

图 4-2 BM 与 RBM 模型结构比较

Fig. 4-2 the comparison between BM and RBM

如果把隐含层层数增加，可以得到 Deep Boltzmann Machine（DBM）[103]（如图 4-3(a)所示）；如果在靠近可视层的部分使用贝叶斯信念网络（即有向图模型，当然这里依然限制层中节点之间没有链接），而在最远离可视层的部分使用 RBM，就可以得到 Deep Belief Network（DBN）（如图 4-3(b)所示）。DBM 中的权值连接均为无向的，或双向的；DBN 中除最顶层为权值连接均为单向的。因此 DBM 模型中 RBM 的训练需要考虑上下相邻 RBM 的影响；而 DBN 中的 RBM 可以独立训练，即仅需要考虑该 RBM 内部可视层与隐含层的相互影响。例如，在 DBM 中，生成第一个隐含层 **h**1 的条件概率为 $p(\mathbf{h1}|\mathbf{v},\mathbf{h2})$，即其相邻层 **v** 和 **h**2 均对其有影响，需要考虑双向的连接，而 DBN 中仅需考虑单向的生成权值连接即可，生成概率为 $p(\mathbf{h1}|\mathbf{h2})$。



(a) Deep Boltzmann Machine (DBM) (b) Deep Belief Network (DBN)

图 4-3 DBM 与 DBN 结构对比

Fig. 4-3 The comparison between DBM and DBN

DBN 通过贪婪逐层非监督方法预训练生成模型的权值，如图 4-4 所示，训练好一个 RBM 模型后，固定生成权值，然后在上面垒加一层新的隐层单元，原来 RBM 的隐层变成它的输入层，这样就构造了一个新的 RBM，然后用同样的方法学习它的权值。依此类推，可以堆叠多个 RBM。在最高两层，权值被连接到一起，这样更低层的输出将会提供一个参考的线索或者关联给顶层，这样顶层就会将其联系到它的记忆内容 y（这里指判别任务）。在微调阶段，DBN 通过带标签数据对判别性能做调整，确定网络的识别权值。

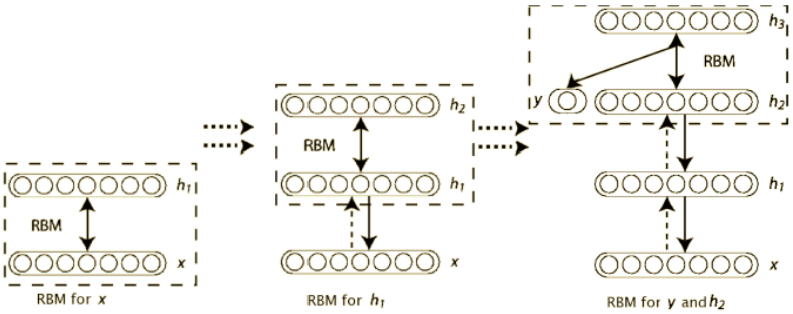


图 4-4 DBN 的形成过程——贪婪逐层无监督训练

Fig. 4-4 The forming process of DBN – greedy layer-wised unsupervised training

4.2.3 深度堆叠网络 DSN

DBN 通过引入贪婪逐层无监督预训练的方法实现可以利用大量无标数据和少

量有标数据的半监督学习，预训练的结果被认为是比随机初始值更优的初始值，可以有效缓解网络收敛于局部最小的问题。但是DBN的中间各层不能进行有效监督，随着层数的增加，梯度扩散现象仍可能发生。如果将无监督的贪婪逐层训练改为利用有标签数据的有监督贪婪逐层训练，每层训练的结果作为下一层的输入向量，如此堆叠成的网络即为深度堆叠网络（Deep Stacking Network，DSN）。DSN实现了深度网络各层的可监督训练，图4-5给出DSN的形成过程，每层的输入采用“替代”的方式，即新训练网络的隐含层替代原有输入成为新的输入。

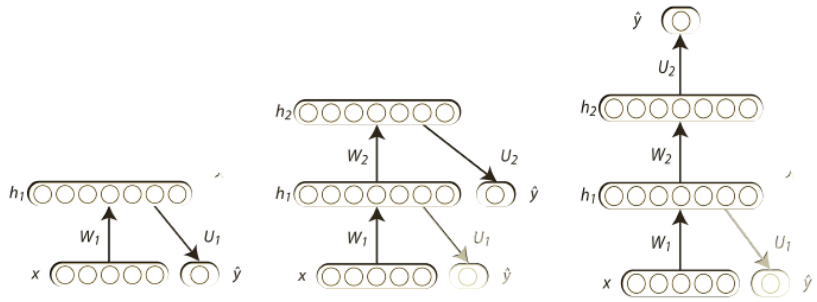


图4-5 DSN的形成过程——贪婪逐层有监督训练

Fig. 4-5 The forming process of DSN – greedy layer-wised supervised training

微软研究院的邓力博士和他的同事们采用输入层逐层堆叠的方式构建了更为完整的深度堆叠网络[104]，如图4-6所示。图中用不同的颜色表示不同的模块，网络由一系列模块堆叠而成，每个模块由一个线性输入层、一个非线性隐含层和一个线性输出层构成，连接输入层与隐含层间的权重W和连接隐含层与输出层的U为待训练参数。原始输入以及底层模块的输出逐层垒加到上层模块作为该层输入，相比替代型输入方式，输入层的逐层堆叠充分利用了各层的学习结果，并始终保留原始输入，避免了学习误差的逐层传递。各模块独立训练，均由目标任务进行监督，相对于传统神经网络训练复杂度更低，也避免了因层数过多引起的梯度扩散的现象；对中间层监督力度的加强，还有助于为目标任务提取到更具区分性的特征。然而，这距离中间层真正意义上的可见化还有一定距离，各层的训练均以最终目标为学习对象，仍然缺乏对于中间层真正含义的发掘，对于模型结构的设定也缺乏有针对性的指引。

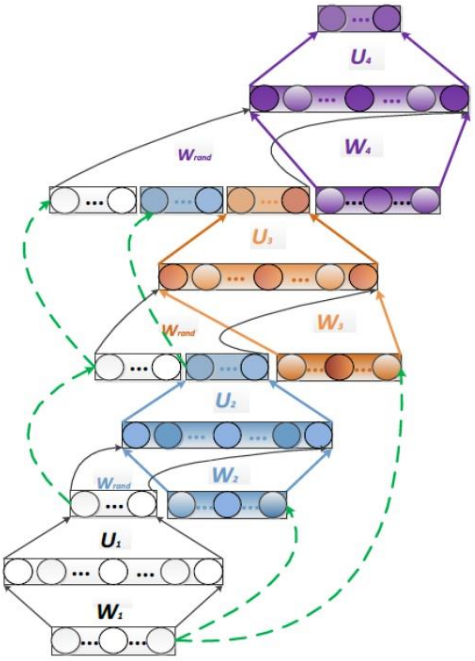


图 4-6 输入层逐层堆叠的 DSN[104]

Fig. 4-6 The DSN with input layer stacked[104]

4.2.4 中间层部分可见的深度堆叠网络 VDSN

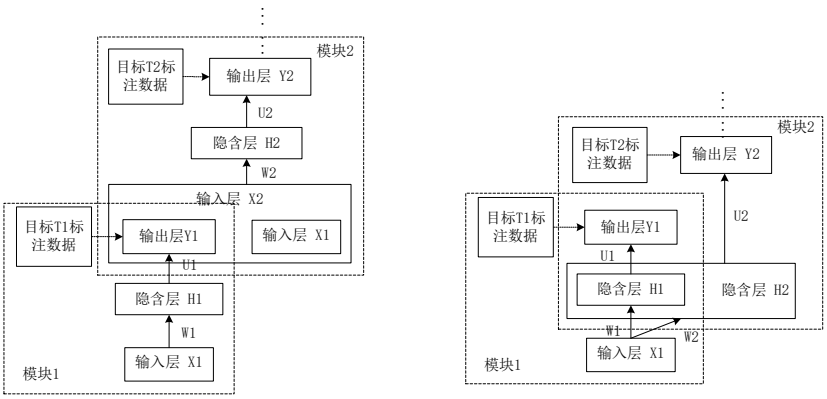
在 DSN 的基础上，我们加入先验知识的指引，使其参与到对网络结构的约束中，从而构建中间层部分可见的深度堆叠网络（Visible Deep Stacking Network, VDSN）。与 DSN 不同的是，VDSN 的每一模块都具有不同的子目标，各子目标按照预先设定的生成顺序依次训练并参与到后续目标及最终目标的学习中。在人工干预的同时，保留抽取新信息的能力和一定的容错能力，因此称为“部分可见”。有别于最终目标的子目标的设定，使网络的内部结构具有明确的含义和显性的关系，为网络结构的设计提供依据并指引学习的方向。至此，深度网络所描述的特征不再是静态的单一目标的特征，而是动态衍化的包含一系列子目标的连续过程，各个环节间可能存在的直接或间接的相互影响通过各模块的训练结果的堆叠进行传递。先验知识的加入可能只破解了网络结构的部分信息，尚有未解码信息存在的可能，各子目标间的相互关系也并非确定的一对一的映射关系，因此出于完备性考虑，前面所有模块的结果均会堆叠到后续模块以保证其对后续环节的潜在影响，原始

输入也始终作为输入传递给每个模块。

以 X 表示已有的输入特征，Y 表示最终的学习目标，对于 DSN 以及 DBN 等中间层不可见的深度网络，网络学习是已知 X 条件下对后验概率 P(Y|X)进行估计。DBN 通过引入无监督的预训练从而得到观测数据的先验概率 P(X)，相对于单纯的有监督学习而言能够利用大量的无标签数据学习和发现数据中存在的模式，有助于避免因网络函数表达能力过强而出现过拟合现象。DSN 将 DBN 的逐层无监督训练扩展到逐层有监督训练，避免了多层神经网络整体的反向误差调整；同时各模块均由最终目标监督训练，可以视为在逐层抽取更具区分性的特征。VDSN 延续了 DSN 基本模块的结构以及各模块独立训练的方式，同时 DBN 中无监督的预训练也可参与到网络参数的初始化；但是，VDSN 每个模块不再由最终目标监督，而是依次预测不同的子目标 Y_i，各子目标与最终目标间构成链式生成关系，各模块依次估计 $P(Y_1|X)$，$P(Y_2|X,Y_1)$，$P(Y_3|X,Y_1,Y_2)$，•••，$P(Y_M|X,Y_1,Y_2,\ldots,Y_{M-1})$，M 为子目标数。可见，VDSN 相比于中间层不可见的深度网络虽然也采用逐层训练的模式，但每层可利用的已知信息在原始输入的基础上得到扩展，在子目标设定合理的前提下，先验的引入有助于模型学习到更准确更具泛化能力的解。利用先验知识对已知信息进行扩展，对于数据有限尤其是有标数据难以获得的情况显得尤为重要，因为仅依赖小规模的数据可能很难靠网络本身自动发现数据中隐藏的结构模式，而借鉴相关领域的专业知识对网络结构进行部分的人工干预，可以减少对于训练数据的依赖，这可被视作从结构设定的角度对网络进行了监督或半监督引导。

VDSN 的每个模块也由输入层、单一隐含层和输出层构成，各模块的训练参数有两组：连接输入层 $\mathbf{X}$ 与隐含层 $\mathbf{H}$ 的权重矩阵 $\mathbf{W}$，和连接隐含层 $\mathbf{H}$ 与输出层 $\mathbf{Y}$ 的权重矩阵 $\mathbf{U}$。其中，隐含层为输入层的非线性映射 $\mathbf{H}=\sigma(\mathbf{b}+\mathbf{W}^{\mathrm{T}}\mathbf{X})$，$\sigma(x)=1/(1+\exp(-x))$，输出层为隐含层的线性组合 $\mathbf{Y}=\mathbf{U}^{\mathrm{T}}\mathbf{H}$。由此，各模块可向后面传递的结果有两种可能：一是将输出层传递给后续模块作为部分输入层，即两个权重矩阵 $\mathbf{W}$ 和 $\mathbf{U}$ 均向后传递并对其产生影响；二是将隐含层传递给后续模块作为部分隐含层，即仅向后传递连接输入层和隐含层的权重 $\mathbf{W}$。由于隐含层的值是非线性分布，而输入层和输出层的值均为线性分布，因此隐含层没有传递给后续模块作为部分输入层。第一种堆叠方式与输入层逐层堆叠的 DSN 的连接模式类似，但是各模块有不同的训练目标，每个模块的输出不再是最终目标的估计，而是有各自不同的子目标，由于训练结果的逐层堆叠发生在后续模块的输入层，采用这种方式搭建的 VDSN 称为输入层部分可见的深度堆叠网络（Input-layer Visible Deep Stacking Network, IVDSN）。第二种堆叠方式更直接地体现了中间层部分已知的情况，由于逐层堆叠体现在隐含层，将这种网络命名为隐含层部分可见的深度堆叠网络（Hidden-layer Visible Deep Stacking Network, HVDSN）。

图 4-7 给出了 IVDSN 和 HVDSN 两种网络中两个模块及其连接关系示意图。仅就这两个模块而言，目标**T**2 相当于最终目标，而**T**1 相当于其子目标，模块 1 的训练目标为使输出层**Y**1 与**T**1 的误差最小，而模块 2 的训练目标为使输出层**Y**2 与**T**2 的误差最小，各模块独立训练，分别基于各自目标进行本模块内部由输出层到输入层的反向误差调整。图 4-7(a)中，模块 1 的输出**Y**1 传递到模块 2 作为部分输入，同时为了防止信息丢失和误差累积，原始输入**X**1 也一并传递作为该模块的部分输入，因此模块 2 的输入层**X**2=[**X**1;**Y**1]。图 4-7(b)中，各模块输入均为**X**，但从模块 2 开始，隐含层的初始方式发生变化，即隐含层不再完全隐含，部分隐层节点通过模块 1 训练得到的隐含层进行初始化，同时扩张一部分未知节点（随机初始化或由 RBM 初始化），使模块 2 隐含层的初始状态变为**H**2=[**H**1;$\mathbf{H}_{unkonwn}$]，从而增加对更高层目标的学习能力。



(a) IVDSN (b) HVDSN

图 4-7 两种中间层部分可见的深度堆叠网络模块及其连接关系示意图

Fig. 4-7 Two modules of the two visible deep stacking networks (VDSN) and their connections

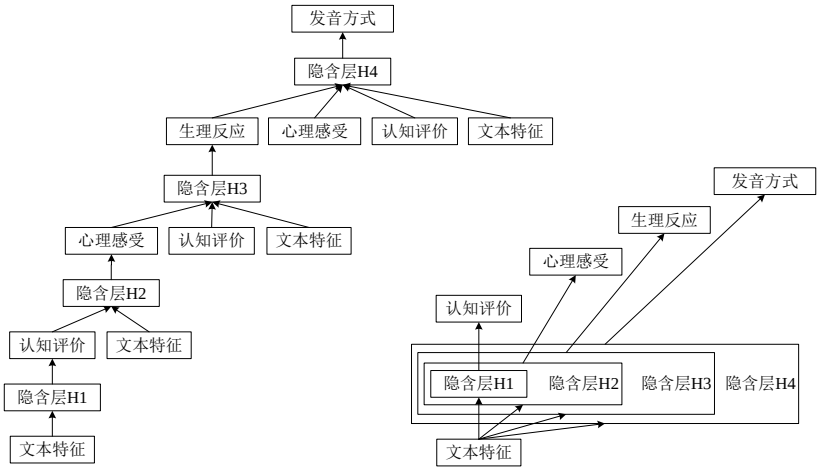
可见，HVDSN 直接延续了神经网络将隐含层作为中间层的概念，中间层部分可见、部分未知体现于隐含层的继承与扩张上；而 IVDSN 将低层模块视作高层模块的中间层，即各子目标作为实现最终目标的中间过程，中间层的部分可见化体现于中间过程的部分已知（各子目标的输出）和部分未知（原始输入内未被破解的部分）。网络结构上，IVDSN 的输入层随着训练目标的增加而逐层扩张；而 HVDSN 的输入层保持不变，隐含层的规模在逐层扩张。从计算角度分析，IVDSN 和 HVDSN 均延续了 DSN 各模块独立训练的训练方式，无需全局的反向误差调整，计算量上等同于几个单隐层的浅层神经网络；其差别仅在于，IVDSN 与 DSN 一样输入层维度需要逐层递增，因此比其他网络具有更多的输入层节点，而 HVDSN 隐含层维度

逐层递增，因此可能比其他网络具有更多的隐含层节点。

4.2.5 基于 VDSN 的言语情感预测模型

我们基于心理学、语音学、朗读学和播音学等领域的相关知识，提出言语情感的多视角描述体系，由认知评价、心理感受、生理反应和发音描述四部分构成，分别对应言语情感生成过程中认知、心理、生理和行为上的变化。言语情感的生成是一连续的、动态的过程，各步骤之间存在直接或间接的相互影响，还可能存在反馈作用（出于简化计算考虑，这里暂不考虑反馈）。在不考虑反馈的情况下，将文本分析得到的特征作为原始输入特征，依次生成认知、心理、生理和发音四种成分，每个环节对应于深度堆叠网络的一个模块，每个模块有各自不同的训练目标，输出层每个节点对应各自子目标的一个维度。当以发音描述作为最终目标，其他三个步骤则为中间过程，对应深度网络的中间层；类似地，每个环节都可称作其后续环节的中间过程。当直接从文本特征预测各环节内容而不考虑其中间环节的影响时，即为中间层“隐含”的情况。各模块间的影响关系通过将生成结果逐层累加到下一模块来实现。

当采用 IVDSN 时，前面模块的输出层依次堆叠到后面模块的输入层，如图 4-8(a)所示。每个模块为含单一隐含层的神经网络，由下至上依次对言语情感的四种成分进行预测。采用 HVDSN 时，前面模块的隐含层逐层堆叠作为后面模块部分隐含层的初始值，如图 4-8(b)所示。可以看出，IVDSN 形成明显的多层级联型网络，而 HVDSN 则构成嵌套型的网络，整体结构仍为单隐层神经网络。不同于传统神经网络由隐含层数目决定网络深度的方式，堆叠网络作为深度网络时网络深度由模块数决定，其中，IVDSN 和 HVDSN 的模块数与经由的中间环节的个数有关，二者最终都构成了深度为 4 的深度网络。两种网络对于未知信息和已知信息的获取及处理方式不同。IVDSN 中，已知信息与未知信息均位于输入层，已知信息来源于子目标的输出，未知信息仍隐含于文本特征中；而 HVDSN 的已知信息与未知信息均位于隐含层，二者均是对文本特征的非线性表示，差别仅在于已知信息由其他子目标进行了有监督微调，而未知节点仅做了无监督预训练（注意是这部分节点单独做无监督预训练，不包含已知节点）。



(a) 基于 IVDSN 的言语情感预测模型 (b) 基于 HVDSN 的言语情感预测模型

(a) The emotion prediction model based on IVDSN (b) The emotion prediction model based on IVD SN

图 4-8 基于 VDSN 的言语情感预测模型网络结构

Fig. 4-8 The emotion prediction model based on VDSN

4.3 训练过程

与 DBN 和 DSN 类似，VDSN 的训练也分为无监督的预训练（参数初始化）和需要少量标注数据的针对目标任务的微调两个步骤，训练以模块为单位进行，各模块的训练目标不同。IVDSN 输入层与隐含层间的权重全部由 RBM 进行初始化，HVDSN 隐含层部分节点由前面模块训练得到的隐含层初始化，剩余节点由 RBM 初始化。

4.3.1 基于 RBM 的参数预训练

RBM 即包含可视层和隐含层的双层对称网络，层内无连接，层间全连接。用向量 $\mathbf{v}$ 和 $\mathbf{h}$ 分别表示可见单元和隐单元的状态，给定一组状态 $(\mathbf{v},\mathbf{h})$，RBM 作为一个系统所具备的能量定义为：

$$E(\mathbf{v},\mathbf{h};\theta)=-\sum_{ij}W_{ij}v_i h_j-\sum_i a_i v_i-\sum_j b_j h_j \tag{4-1}$$

式中，$\theta=\{W,a,b\}$ 是 RBM 的参数，W_{ij} 是可见单元 v_i 与隐单元 h_j 之间的连接权重，a_i 表示可见单元 v_i 的偏置，b_j 表示隐单元 h_j 的偏置。基于该能量函数和 Boltzmann

分布公式，$(\mathbf{v},\mathbf{h})$的联合概率分布为：

$$P_\theta(\mathbf{v},\mathbf{h})=\frac{1}{Z(\theta)}\exp\left(-E(\mathbf{v},\mathbf{h};\theta)\right) \tag{4-2}$$

其中，$Z(\theta)=\sum_{\mathbf{v},\mathbf{h}}\exp\left(-E(\mathbf{v},\mathbf{h};\theta)\right)$ 为归一化因子或称为配分函数（partition function）。

对于一个实际问题，我们更关心的是由 RBM 所定义的关于观测数据 $\mathbf{v}$ 的分布 $P_\theta(\mathbf{v})$，即联合概率分布 $P_\theta(\mathbf{v},\mathbf{h})$ 的边际分布，也称为似然概率。

$$P_\theta(\mathbf{v})=\frac{1}{Z(\theta)}\sum_{\mathbf{h}}\exp\left(-E(\mathbf{v},\mathbf{h};\theta)\right) \tag{4-3}$$

该分布的确定需要计算归一化因子，因此很难准确计算。根据 RBM 的特殊结构以及一些近似手段，我们可以避开归一化因子的计算。

由 RBM 层内无连接，层间全连接的结构可知，当给定可见单元状态时，各隐单元的激活状态之间是条件独立的，此时，隐层第 j 个节点的激活概率为：

$$p_\theta\left(h_j=1|\mathbf{v}\right)=\sigma\left(b_j+\sum_i v_i W_{ij}\right) \tag{4-4}$$

其中，$\sigma(x)=1/\left(1+\exp(-x)\right)$为 Sigmoid 激活函数。由于 RBM 的结构是对称的，因此当给定隐单元状态时，各可见单元的激活状态也是条件独立的，第i个可见单元的激活概率为：

$$p_\theta\left(v_i=1|\mathbf{h}\right)=\sigma\left(a_i+\sum_j h_j W_{ij}\right) \tag{4-5}$$

以上为可见层为 0-1 分布的情况，若可见层为实数，则采用高斯分布，此时的能量函数变为：

$$E(\mathbf{v},\mathbf{h};\theta)=-\sum_{ij}W_{ij}v_i h_j-\frac{1}{2}\sum_i\left(v_i-a_i\right)^2-\sum_j b_j h_j \tag{4-6}$$

隐单元的激活概率公式不变，可见单元的激活概率变为：

$$p_\theta\left(v_i=1|\mathbf{h}\right)=N(a_i+\sum_j h_j W_{ij},1) \tag{4-7}$$

其中，N 为均值为$a_i+\sum_j h_j W_{ij}$，方差为 1 的高斯分布。

学习 RBM 的目标是求出参数θ的值，以拟合给定的训练数据。参数θ可以通过最大似然估计的方法学习得到，即选定该组参数时，模型所定义的关于观测数据

$\mathbf{v}$的似然概率最大：

$$\hat{\theta}=\arg\max L(\theta)=\arg\max\sum_{n=1}^{N}\log P_\theta\left(\mathbf{v}^{(n)}\right) \tag{4-8}$$

$\hat{\theta}$为最优参数，N 为样本数。似然概率的最大值可通过随机梯度上升法求得，其关于θ的梯度为：

$$\begin{aligned}
\frac{\partial L}{\partial\theta}&=\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\log P_\theta\left(\mathbf{v}^{(n)}\right)=\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\log\sum_{\mathbf{h}}P_\theta\left(\mathbf{v}^{(n)},\mathbf{h}\right)\\
&=\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\log\frac{\sum_{\mathbf{h}}\exp\left(-E\left(\mathbf{v}^{(n)},\mathbf{h};\theta\right)\right)}{\sum_{\mathbf{v}}\sum_{\mathbf{h}}\exp\left(-E\left(\mathbf{v},\mathbf{h};\theta\right)\right)}\\
&=\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\left(\log\sum_{\mathbf{h}}\exp\left(-E\left(\mathbf{v}^{(n)},\mathbf{h};\theta\right)\right)-\log\sum_{\mathbf{v}}\sum_{\mathbf{h}}\exp\left(-E\left(\mathbf{v},\mathbf{h};\theta\right)\right)\right)\\
&=\sum_{n=1}^{N}\left(\sum_{\mathbf{h}}\frac{\exp\left(-E\left(\mathbf{v}^{(n)},\mathbf{h};\theta\right)\right)}{\sum_{\mathbf{h}}\exp\left(-E\left(\mathbf{v}^{(n)},\mathbf{h};\theta\right)\right)}\times\frac{\partial\left(-E\left(\mathbf{v}^{(n)},\mathbf{h};\theta\right)\right)}{\partial\theta}\right.\\
&\left.-\sum_{\mathbf{v}}\sum_{\mathbf{h}}\frac{\exp\left(-E\left(\mathbf{v},\mathbf{h};\theta\right)\right)}{\sum_{\mathbf{v}}\sum_{\mathbf{h}}\exp\left(-E\left(\mathbf{v},\mathbf{h};\theta\right)\right)}\times\frac{\partial\left(-E\left(\mathbf{v},\mathbf{h};\theta\right)\right)}{\partial\theta}\right)\\
&=\sum_{n=1}^{N}\left(\left\langle\frac{\partial\left(-E\left(\mathbf{v}^{(n)},\mathbf{h};\theta\right)\right)}{\partial\theta}\right\rangle_{P_\theta\left(\mathbf{h}|\mathbf{v}^{(n)}\right)}-\left\langle\frac{\partial\left(-E\left(\mathbf{v},\mathbf{h};\theta\right)\right)}{\partial\theta}\right\rangle_{P_\theta(\mathbf{v},\mathbf{h})}\right)
\end{aligned} \tag{4-9}$$

其中，$\langle\cdot\rangle_P$ 表示关于分布 P 的数学期望。$P_\theta\left(\mathbf{h}\mid\mathbf{v}^{(n)}\right)$ 表示在可见单元限定为已知的训练样本$\mathbf{v}^{(n)}$时，隐层的概率分布，故式(4-9)中的前一项较容易计算。$P_\theta(\mathbf{v},\mathbf{h})$表示可见单元与隐单元的联合分布，由于归一化因子$Z(\theta)$的存在，该分布很难计算，导致式(4-9)中第二项很难获取，只能通过一些采样方法（如 Gibbs 采样）获取其近似值。假设只有一个训练样本，分别用 data 和 model 来简记$P_\theta\left(\mathbf{h}\mid\mathbf{v}^{(n)}\right)$和$P_\theta(\mathbf{v},\mathbf{h})$两个分布，则对数似然函数关于连接权重$W_{ij}$、可见单元偏置$a_i$和隐层单元偏置$b_j$的偏导数分别为：

$$\frac{\partial\log P_\theta(\mathbf{v})}{\partial W_{ij}}=\langle v_i h_j\rangle_{\text{data}}-\langle v_i h_j\rangle_{\text{model}} \tag{4-10}$$

$$\frac{\partial\log P_\theta(\mathbf{v})}{\partial a_i}=\langle v_i\rangle_{\text{data}}-\langle v_i\rangle_{\text{model}} \tag{4-11}$$

$$\frac{\partial\log P_\theta(\mathbf{v})}{\partial b_j}=\langle h_j\rangle_{\text{data}}-\langle h_j\rangle_{\text{model}} \tag{4-12}$$

基于 RBM 对称结构以及神经元的条件独立性，我们可以使用 Gibbs 采样方法得到式(4-10)~(4-12)中第二项的一个近似。Gibbs 采样是一种基于马尔可夫链蒙特卡罗(Markov Chain Monte Carlo，MCMC)策略的采样方法。对于一个 K 维随机向量 $\mathbf{X}=(X_1,X_2,\cdots,X_K)$，假设无法求得关于 $\mathbf{X}$ 的联合分布，但知道给定 $\mathbf{X}$ 中其他分量时，第 k 个分量 X_k 的条件分布 $P(X_k \mid X_{k^-})$，那么可以从 $\mathbf{X}$ 的任意状态（如 $[x_1(0),x_2(0),\cdots,x_K(0)]$）开始，利用上述条件分布迭代地对其分量依次采样。随着采样次数的增加，随机变量 $[x_1(t),x_2(t),\cdots,x_K(t)]$ 的概率分布将以 t 的几何级数的速度收敛于 $\mathbf{X}$ 的联合概率分布 $P(\mathbf{X})$。在 RBM 中进行 t 步 Gibbs 采样的具体步骤为：用一个训练样本初始化可见层的状态 $\mathbf{v}_0$，然后交替进行如下采样：

$$\begin{aligned}
&\mathbf{h}_0 \sim P(\mathbf{h}\mid\mathbf{v}_0),\ \mathbf{v}_1 \sim P(\mathbf{v}\mid\mathbf{h}_0),\\
&\mathbf{h}_1 \sim P(\mathbf{h}\mid\mathbf{v}_1),\ \mathbf{v}_2 \sim P(\mathbf{v}\mid\mathbf{h}_1),\\
&\qquad\cdots\cdots,\ \mathbf{v}_{t+1} \sim P(\mathbf{v}\mid\mathbf{h}_t).
\end{aligned}$$

在采样步数 t 足够大时，可以得到服从 RBM 所定义的分布的样本，但是该方法训练效率仍旧不高。2002 年，Hinton 提出了一种 RBM 的快速学习算法，即 CD-n（Contrastive Divergence，对比散度）算法，仅需 n（通常 n=1）步 Gibbs 采样便可得到足够好的近似[105,106]。在算法一开始，可见单元的状态被设置成一个训练样本，然后利用式(4-4)计算所有隐层单元的激活概率并抽取其二值状态。在所有隐层单元的状态确定之后根据式(4-5)或式(4-7)确定第 i 个可见单元 v_i 取值为 1 的概率，进而产生可见层的一个“重构（reconstruction）”。这样，在使用随机梯度上升法最大化对数似然概率时，各参数的更新规则为：

$$\Delta W_{ij}=\varepsilon\left(\langle v_i h_j\rangle_{\text{data}}-\langle v_i h_j\rangle_{\text{recon}}\right) \tag{4-13}$$

$$\Delta a_i=\varepsilon\left(\langle v_i\rangle_{\text{data}}-\langle v_i\rangle_{\text{recon}}\right) \tag{4-14}$$

$$\Delta b_j=\varepsilon\left(\langle h_j\rangle_{\text{data}}-\langle h_j\rangle_{\text{recon}}\right) \tag{4-15}$$

ε 是学习速率，$\langle\cdot\rangle_{\text{recon}}$ 表示一步重构后模型定义的分布期望。其中，可见单元和隐层单元的状态既可以用采样后的二值状态表示，也可以用值为 1 的概率表示，前者更接近 RBM 模型的数学定义，后者则引入较少的采样噪声。本文中隐层单元的状态使用采样后的二值状态，可见单元的状态使用式(4-7)计算的值为 1 的概率。

4.3.2 基于批量梯度下降的参数微调

BP 算法（反向传播算法）是常用来对神经网络参数进行微调的有监督学习算法，因为预测误差由输出层向输入层反向逐层传播而得名。该算法的优化准则是最小均方误差准则，执行该优化的方法是梯度下降法（Gradient Descent）。梯度下降法有两种迭代求解思路：随机梯度下降（Stochastic Gradient Descent）和批量梯度下降（Batch Gradient Descent）。随机梯度下降通过计算每条样本的损失函数并对参数求偏导得到对应的梯度，以此来迭代更新参数，即最小化每条样本的损失函数；批量梯度下降则是最小化所有训练样本的损失函数，每次迭代要用到所有训练集的数据。对比而言，随机梯度下降法的迭代速率要高于批量梯度下降法，因其可能不需要计算完所有样本的损失函数就达到收敛；但是后者最终求解的是全局最优解，而随机梯度下降法不是每次迭代得到的损失函数都朝着全局最优的方向，对于有多个极小值的非凸问题则可能收敛到局部最优。DSN 的训练算法是基于批量梯度下降的，同时，DSN 采用矩阵计算的形式，便于实现算法的并行计算[107]。

假设输入数据用矩阵 $\mathbf{X}$ 表示，$\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_i,\ldots,\mathbf{x}_N]$，每条样本的特征向量 $\mathbf{x}_i=[x_{1i},\ldots,x_{ji},\ldots,x_{Di}]^{\mathrm{T}}$，其中 D 为输入特征的维度，N 为样本数目。目标矩阵即标注数据 $\mathbf{T}=[\mathbf{t}_1,\ldots,\mathbf{t}_i,\ldots,\mathbf{t}_N]$，每条样本的目标向量 $\mathbf{t}_i=[t_{1i},\ldots,t_{ji},\ldots,t_{Ci}]^{\mathrm{T}}$，其中 C 为输出向量的维度。输出向量 $\mathbf{Y}=[\mathbf{y}_1,\ldots,\mathbf{y}_i,\ldots,\mathbf{y}_N]$。设隐含层单元数为 L，隐含层 $\mathbf{H}=[\mathbf{h}_1,\ldots,\mathbf{h}_i,\ldots,\mathbf{h}_N]$，$\mathbf{h}_i=[h_{1i},\ldots,h_{ji},\ldots,h_{Li}]^{\mathrm{T}}$。分别用 $\mathbf{W}$ 和 $\mathbf{U}$ 表示输入层和隐含层间、隐含层和输出层间的权重矩阵，$\mathbf{W}$ 大小为 $D\times L$，$\mathbf{U}$ 大小为 $L\times C$，二者构成网络的待训练参数。$\mathbf{W}$ 的初始值通过 RBM 的预训练得到。给定 $\mathbf{W}$，则可计算 $\mathbf{H}=\sigma\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}\right)$（此处省略了偏置）。隐含层与输出层间的映射为线性映射，$\mathbf{Y}=\mathbf{U}^{\mathrm{T}}\mathbf{H}$，相当于线性回归。

网络训练的目标是更新 $\mathbf{W}$ 和 $\mathbf{U}$ 使平方误差 E 最小。

$$E=\left\|\mathbf{Y}-\mathbf{T}\right\|^2=\mathrm{Tr}\left[\left(\mathbf{Y}-\mathbf{T}\right)\left(\mathbf{Y}-\mathbf{T}\right)^{\mathrm{T}}\right] \tag{4-16}$$

Tr 表示求矩阵的迹。

E 关于 $\mathbf{U}$ 的偏导数即梯度为：

$$\frac{\partial E}{\partial \mathbf{U}}=\frac{\partial \mathrm{Tr}\left[\left(\mathbf{U}^{\mathrm{T}}\mathbf{H}-\mathbf{T}\right)\left(\mathbf{U}^{\mathrm{T}}\mathbf{H}-\mathbf{T}\right)^{\mathrm{T}}\right]}{\partial \mathbf{U}}=2\mathbf{H}\left(\mathbf{U}^{\mathrm{T}}\mathbf{H}-\mathbf{T}\right)^{\mathrm{T}} \tag{4-17}$$

令该梯度为 0，因该函数是一凸优化问题，可以得到 $\mathbf{U}$ 的闭合形式的解：

$$\mathbf{U}=\left(\mathbf{H}\mathbf{H}^{\mathrm{T}}\right)^{-1}\mathbf{H}\mathbf{T}^{\mathrm{T}} \tag{4-18}$$

可以看出 $\mathbf{U}$ 的值由 $\mathbf{W}$ 决定，$\mathbf{W}$ 确定之后，$\mathbf{H}$ 则随之确定，进而可以得到 $\mathbf{U}$ 的值。因此，网络训练的重点是 $\mathbf{W}$ 值的更新。计算 E 关于 $\mathbf{W}$ 的梯度需要考虑 $\mathbf{W}$ 与 $\mathbf{U}$ 的关联关系，将式(4-18)代入 E 关于 $\mathbf{W}$ 的梯度计算公式可以得到：

$$\frac{\partial E}{\partial \mathbf{W}}=\frac{\partial \mathrm{Tr}\left[\left(\mathbf{U}^{\mathrm{T}}\mathbf{H}-\mathbf{T}\right)\left(\mathbf{U}^{\mathrm{T}}\mathbf{H}-\mathbf{T}\right)^{\mathrm{T}}\right]}{\partial \mathbf{W}}$$
$$=\frac{\partial \mathrm{Tr}\left[\left(\left[\left(\mathbf{H}\mathbf{H}^{\mathrm{T}}\right)^{-1}\mathbf{H}\mathbf{T}^{\mathrm{T}}\right]^{\mathrm{T}}\mathbf{H}-\mathbf{T}\right)\left(\left[\left(\mathbf{H}\mathbf{H}^{\mathrm{T}}\right)^{-1}\mathbf{H}\mathbf{T}^{\mathrm{T}}\right]^{\mathrm{T}}\mathbf{H}-\mathbf{T}\right)^{\mathrm{T}}\right]}{\partial \mathbf{W}}$$
$$=2\mathbf{X}\left[\mathbf{H}^{\mathrm{T}}\circ\left(1-\mathbf{H}\right)^{\mathrm{T}}\circ\left[\mathbf{H}^{+}\left(\mathbf{H}\mathbf{T}^{\mathrm{T}}\right)\left(\mathbf{T}\mathbf{H}^{+}\right)-\mathbf{T}^{\mathrm{T}}\left(\mathbf{T}\mathbf{H}^{+}\right)\right]\right] \tag{4-19}$$

其中，$\mathbf{H}^{+}=\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{H}^{\mathrm{T}})^{-1}$，$\circ$代表元素相乘的内积运算。

上式中，每个样本在每次迭代中所起的作用一样。为了加快算法的收敛速率，文献[107]引入一个权重矩阵$\mathbf{\Lambda}$，从而加大对预测误差大的样本的关注。$\mathbf{\Lambda}$为一对角阵，对角线上的元素$\lambda_{ii}=(N/E\|\mathbf{y}_i-\mathbf{t}_i\|^2+1)/2$，其中，$i$为样本索引，$N$为样本数量，样本预测误差越大其对应的$\lambda_{ii}$值越大，由此使得预测误差大的样本在全部样本的预测误差中拥有较大的权重，即在参数更新中起更主要的作用。如果该样本训练误差降低了，在下次迭代中该样本所占权重也会变小。通过这种方式使参数向误差最有效降低的方向更新，提升了算法的收敛速度，同时一定程度上降低了网络学习陷入局部最优的可能。引入$\mathbf{\Lambda}$后平方误差E变为$E=\mathrm{Tr}\left[\left(\mathbf{Y}-\mathbf{T}\right)\mathbf{\Lambda}\left(\mathbf{Y}-\mathbf{T}\right)^{\mathrm{T}}\right]$，$\mathbf{U}$的最优解变为：

$$\mathbf{U}=\left(\mathbf{H}\mathbf{\Lambda}\mathbf{H}^{\mathrm{T}}\right)^{-1}\mathbf{H}\mathbf{\Lambda}\mathbf{T}^{\mathrm{T}} \tag{4-20}$$

E关于$\mathbf{W}$的梯度公式变为：

$$\frac{\partial E}{\partial \mathbf{W}}=\frac{\partial \mathrm{Tr}\left[\left(\mathbf{U}^{\mathrm{T}}\mathbf{H}-\mathbf{T}\right)\mathbf{\Lambda}\left(\mathbf{U}^{\mathrm{T}}\mathbf{H}-\mathbf{T}\right)^{\mathrm{T}}\right]}{\partial \mathbf{W}}$$
$$=\frac{\partial \mathrm{Tr}\left[\left(\left[\left(\mathbf{H}\mathbf{\Lambda}\mathbf{H}^{\mathrm{T}}\right)^{-1}\mathbf{H}\mathbf{\Lambda}\mathbf{T}^{\mathrm{T}}\right]^{\mathrm{T}}\mathbf{H}-\mathbf{T}\right)\mathbf{\Lambda}\left(\left[\left(\mathbf{H}\mathbf{\Lambda}\mathbf{H}^{\mathrm{T}}\right)^{-1}\mathbf{H}\mathbf{\Lambda}\mathbf{T}^{\mathrm{T}}\right]^{\mathrm{T}}\mathbf{H}-\mathbf{T}\right)^{\mathrm{T}}\right]}{\partial \mathbf{W}}$$
$$=2\mathbf{X}\left[\mathbf{H}^{\mathrm{T}}\circ\left(1-\mathbf{H}\right)^{\mathrm{T}}\circ\left[\mathbf{H}^{+}\left(\mathbf{H}\mathbf{\Lambda}\mathbf{T}^{\mathrm{T}}\right)\left(\mathbf{T}\mathbf{H}^{+}\right)-\mathbf{\Lambda}\mathbf{T}^{\mathrm{T}}\left(\mathbf{T}\mathbf{H}^{+}\right)\right]\right] \tag{4-21}$$

其中，$\mathbf{H}^{+}=\mathbf{\Lambda}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{\Lambda}\mathbf{H}^{\mathrm{T}})^{-1}$。

4.3.3 优化措施

在网络训练过程中，存在两个分布不均衡的问题可能影响网络性能：特征维度不均衡问题和样本数目不均衡问题，我们分别对其进行了调整以优化网络性能。

（1）维度均衡调整

回顾 4.2 节介绍的言语情感预测模型，随着中间过程的依次生成，中间过程的

结果也会逐层累加到后续步骤对其产生影响。在隐含层部分可见的深度堆叠网络（HVDSN）中，各部分的影响差异可以通过隐层单元数的设定予以平衡。而在输入层部分可见的深度堆叠网络（IVDSN）中，每个情感成分的维度已经由情感描述体系予以设定，各情感成分之间以及它们相对于文本特征都存在严重的维度不平衡，如：认知评价的描述维度为 5 维；生理反应仅为 2 维；心理感受不同层级维度不一样，为了刻画最细致的感受类型，我们选用最底层类别最多的划分，即包含 43 维；而文本特征的维度由主题数决定，经后面实验测定维度在 100 维以上。当这些特征处于同一层作为输入特征，维度较少的情感成分的作用可能会淹没在维度众多的文本特征中而无法得以显现。因此，为了均衡不同方面的输入特征的影响，我们在各部分的输入特征前加入了与各自维度成反比的系数，维度越多，所乘系数越小。

（2）样本均衡调整

另一个可能影响训练结果的因素是不同类型的样本数目的不均衡，由第 3 章对标注样本的统计分析结果也可以看出当前训练数据集中各标注类型的样本数量存在严重的不均衡现象。上一节在介绍网络的训练算法时，为了加快算法的收敛速率，曾引入一个对角阵 $\mathbf{\Lambda}$，样本预测误差越大其对应的对角线上的元素 λ_{ii} 值越大，从而加大对预测误差大的样本的关注。在此基础上，我们增加了关于样本数量不均衡问题的调整，即在计算对角阵元素 λ_{ii} 时不仅考虑该样本预测误差的影响，同时考虑该类型样本数目的影响，样本数较少的数据在全局误差中赋予一个比大类别样本更大的权重，从而均衡不同类别样本在参数更新中所起的作用。调整后的 λ_{ii} 变为 $\lambda_{ii\_ad}=\left(N/E\sum_j\left(y_{ji}-t_{ji}\right)^2 v_{ji}+1\right)\Big/2$，$v_{ji}=1/\left(N_{jk}+1\right)$，同样 i 为样本索引，j 指示不同的特征维度，k 指示样本类别，即不同的强度值，N_{jk} 表示在第 j 维上标注为 k 的样本子集的大小。

4.4 文本特征提取

当前，计算机还不具备人脑的结构，无法理解自然语言，所以基于文本的情感预测首先需要将无结构的自然语言文本转化为计算机可计算的特征文本。本章的文本分析单元定为篇章级，不涉及更细尺度的分析，因此对输入文本的语言学处理主要包括分词、特征词提取和特征降维几步。

4.4.1 分词与特征词提取

我们采用 NLPIR 汉语分词系统（又名 ICTCLAS 2014）[108]进行分词处理，将连续文本分割成离散的词序列。分词之后提取与目标任务可能相关的词作为特征词，剔除无用词或关联不大的词从而缩小词典规模。通常情况下，文本情感分析采用常用来表达情绪感受的词作为特征词，称为“情感词”，如“高兴”、“难过”、“哭”、“生气”等。对于当前语料库，如果只采用情感词作为特征词，将发生特征过于稀疏，部分文本甚至不包含特征词的情况。除显性的情感词外，一些词语与语义表达相关，通过语义能传达出语句中的情感倾向，我们将这部分词语称为“内容词”，也将其作为与情感表达相关的特征词。内容词的获取通过将功能词等虚词过滤的办法，我们采用的是中文自然语言处理开放平台公布的中文停用词词表，包含标点符号、数词、语气助词、连词、副词、介词和叹词等，保留下来的实词全部被认为与语义内容相关。注意内容词过滤掉了一些与情感表达相关的词语，如副词和叹词等，因此也不能仅用内容词作特征词。情感词的提取利用 HowNet 提供的情感词词表进行过滤，包含“感受词”、“评价词”和“程度词”几类（见表 4-1）：“感受词”即常用来形容情绪、感受的动词、形容词等，“评价词”是常用来表示对某件事物或人的看法、态度或评价的词语；“程度词”则是用来形容情感级别的副词等。内容词和情感词都作为可能负载情感特征的载体，即“特征词”。

表 4-1 三种情感词示例

Table 4-1 Samples for the three types of emotional words

“感受词”示例	哀凄、哀切、哀伤、哀痛、懊悔、懊恼、抱疚、抱愧、抱歉、抱怨、暴跳、悲悯、悲戚、担心、担忧、爱怜、爱恋、爱慕、尊崇、尊敬、尊重、夸奖、夸赞、快活、快乐
“评价词”示例	蓬勃向上、品学兼优、飘洒、飘逸、漂亮、详明、详实、谐调、欣然、心灵手巧、心明眼亮、卑贱、笨手笨脚、病恹恹、波谲云诡、不对茬儿、索然无味、徒劳无益、顽固唯利是图、乌七八糟、虚弱
“程度词”示例	极其、最为、十分、很、很是、更加、更进一步、更为、越发、越加、多多少少、或多或少、略微、不甚、不怎么、过了头、过甚、过于

4.4.2 特征降维

向量空间模型（Vector Space Model, VSM）是最常用的文本特征表示方式，文档被表示成由特征词出现概率组成的多维向量，向量长度等于特征词个数。这种方法的好处是可以将文档转化成同一空间下的向量计算相似度，但是特征词较多时，以此为特征会增加网络空间和计算时间的开销，当标注数据有限时还易造成过拟

合，因此要对 VSM 表示的文本特征进行降维处理。此外，VSM 没有能力处理一词多义和一义多词问题，例如同义词也分别被表示成独立的一维，而某个词项有多个词义时却始终对应同一维度。

针对这一问题，一些学者开始寻找能挖掘文本潜在语义的表示模型，如较早期的潜在语义分析（Latent Semantic Analysis，LSA）模型[109]。LSA 打破了文档在词典空间进行表示的思维定式，在文档和词之间加入了一个语义维度，以此来探查词与词之间内在语义联系；除此之外，LSA 基于奇异值分解（Singular Value Decomposition，SVD）构造一个原始特征矩阵的低秩逼近矩阵，从而达到特征降维的目的。具体来说，SVD 将原始的词项文档矩阵 C 分解为 U、D、V 三个小矩阵，如图 4-9 所示，U 和 V 是 C 的奇异向量矩阵，D 为由奇异值组成的对角阵。每个奇异值对应的是每个“语义”维度的权重，在很多情况下，前 r 个最大的奇异值之和就占了全部奇异值之和的 99%以上（r 远远小于文档数和词项数）。为了压缩存储矩阵，只保留前 r 个对矩阵影响最大的奇异值，而较小的其他奇异值因为不重要而被删除，这样便得到 D 的低维近似矩阵 D_r，其所对应的奇异向量矩阵 U 变为 U_r（保留前 r 大奇异值对应的奇异向量），文档在潜在语义空间的映射通过 $C^TU_rD_r^{-1}$ 求得，该映射即为文本特征的降维表示。

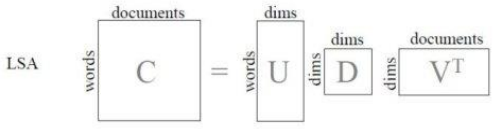


图 4-9 LSA 的降维实现——奇异值分解

Fig. 4-9 The dimension reduction via LSA –Singular Value Decomposition

尽管基于 SVD 的 LSA 取得了一定的成功，但是其缺乏严谨的数理统计基础，而且奇异值分解非常耗时。随着概率统计技术的发展，基于概率统计的分析模式逐渐取代了基于线性代数的分析模式。概率潜在语义分析（probabilistic Latent Semantic Analysis，pLSA）[110]就是 LSA 的概率拓展，它比 LSA 具有更坚实的数学基础。pLSA 继承了“潜在语义”的概念，通过“统一的潜在语义空间”（也就是 Blei 等人于 2003 年正式提出的“Topic”的概念[111]）来关联词与文档，如图 4-10 所示，通过在文档与词语之间增加一个语义空间来抽取文档的语义信息，该空间的每个维度称作一个主题，每个主题由词语的概率分布表示，每篇文档则表示成这些主题的概率分布，该“文档-主题”分布被用作文本特征的降维表示。在 pLSA 中，各个因素（文档、潜在语义空间、词）之间的概率分布求解通常采用 EM（Expectation–Maximization）算法。pLSA 模型中的参数随着文本集的增长而线性增长，容易出现过拟合情况，且模型中的文档概率值与特定的文档相关，没有提供文档的生成模

型，对于训练集外的文本无法分配概率。鉴于 pLSA 的缺点，Blei 等人于 2003 年进一步提出新的主题模型 LDA（Latent Dirichlet Allocation）[111]，它是一个层次贝叶斯模型，把模型的参数也看作随机变量，从而可以引入控制参数的参数，实现彻底的“概率化”。

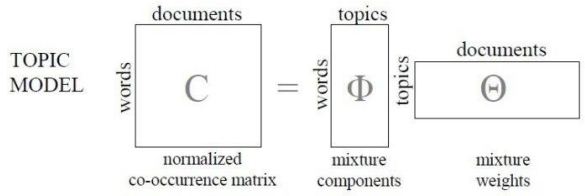


图 4-10 主题模型的降维实现

Fig. 4-10 The dimension reduction via topic model

本文采用 LDA 模型进行特征降维。与 pLSA 一样，LDA 也采用引入主题空间的方法，每个主题为词语的概率分布，文档为这些主题的随机组合。假设有T个主题，则所给文档中第i个词语w_i的概率分布可表示为：

$$P(w_i)=\sum_{j=1}^{T}P(w_i\mid z_i=j)P(z_i=j) \tag{4-21}$$

其中，z_i是潜在变量，表示第i个词语w_i取自该主题；$P(w_i\mid z_i=j)$表示给定主题j出现单词w_i的概率；$P(z_i=j)$表示当前文档出现给定主题j下单词的概率。

在 LDA 中，每个文档中词的主题分布服从多项式（Multinomial）分布，其先验服从 Dirichlet 分布；每个主题下词的分布服从多项式分布，其先验也同样服从 Dirichlet 分布。假定文档集共包含D篇文档，词典集共包含W个唯一词项。为方便表示，令$\phi_w^{(z=j)}=P(w|z=j)$表示主题j在W个词项上的多项式分布；令$\psi_{z=j}^{(d)}=P(z=j)$表示文档d在T个主题上的多项式分布，于是文本d中词项w的概率表示为：

$$P(w\mid d)=\sum_{j=1}^{T}\phi_w^{(z=j)}\psi_{z=j}^{(d)} \tag{4-22}$$

ϕ和ψ为模型的待估计参数，Blei 引入两个超参数（即参数的参数）α和β来控制这两个概率分布的分布，即：

$$z_i|\psi^{(d)}\sim Multinomial\left(\psi^{(d)}\right),\quad \psi^{(d)}\sim Dirichlet(\alpha)$$

$$w_i\mid z_i,\phi^{(z=j)}\sim Multinomial\left(\phi^{(z=j)}\right),\quad \phi^{(z=j)}\sim Dirichlet(\beta)$$

参数的估计属于概率图模型的推理问题，主要算法可以分为精确推理和近似推理两类。LDA 的精确推理很困难，一般采用近似推理的算法。Blei 提出的是使

用变分推理（Variational Inference）的方法[111]，Griffiths 提出了更为简便易懂、精度也更高的基于 Gibbs 采样的方法 [112]，本文使用该方法估计 LDA 的参数。上文已经提到，Gibbs 采样是 Markov Chain Monte Carlo 算法的一个特例。这个算法的运行方式是每次选取概率向量的一个维度，给定其他维度的变量值采样当前维度的值，不断迭代，直到收敛输出待估计的参数。具体步骤如下：

1) 随机给文本中的每个词语分配主题，此为 Markov 链的初始状态。

2) i 从 1 循环到 N（N 为文档中所有词语个数），分别计算 $P(z_i = j \mid z_{-i}, w_i)$，如式(4-23)所示，即排除当前词的主题分配，根据其他所有词的主题分配估计当前词分配各主题的概率。当得到当前词针对所有主题的概率分布后，根据该概率分布采样该词的新的主题，即为 Markov 链的下一个状态。

$$P(z_i = j \mid z_{-i}, w_i) = \frac{\frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}}{\sum_{j=1}^{T} \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}} \tag{4-23}$$

其中，z_{-i} 表示除第 i 个词语 w_i 外其他所有词的主题分配；$n_{-i,j}^{(w_i)}$ 表示主题 j 中除当前第 i 个词语 w_i 外其他与 w_i 相同的词语个数；$n_{-i,j}^{(\cdot)}$ 表示主题 j 中除当前第 i 个词语 w_i 外所有词语个数；$n_{-i,j}^{(d_i)}$ 表示当前词语所在文档 d_i 中分配给主题 j 的词语个数，不包括 $z_i = j$ 这次分配；$n_{-i,\cdot}^{(d_i)}$ 表示文档 d_i 中所有分配了主题的词语个数，同样不包括 $z_i = j$ 这次分配。

3) 迭代第 2)步足够次数以后，认为 Markov 链接近目标分布。用 w 表示 W 个词典集中的某个单一样本，ϕ 和 ψ 的值按下式估算得到：

$$\tilde{\phi}_w^{(z=j)} = \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + W\beta}, \quad \tilde{\psi}_{z=j}^{(d)} = \frac{n_j^{(d)} + \alpha}{n_{\cdot}^{(d)} + T\alpha} \tag{4-24}$$

其中，$n_j^{(w)}$ 表示词项 w 分配给主题 j 的频数；$n_j^{(\cdot)}$ 表示分配给主题 j 的所有词数；$n_j^{(d)}$ 表示文本 d 中分配给主题 j 的词数；$n_{\cdot}^{(d)}$ 表示文本 d 中所有词数。

ψ 为文档 d 在 T 个主题上的多项式分布，被用作降维之后的文本特征，作为情感预测模型的文本输入，降维之后文本特征的维度从原来的词项个数 W 变为主题数 T。关于主题数 T 和超参数 α、β 的设定将在下一节实验参数确定部分进行讨论。

表 4-2 给出部分主题模型提取出的主题维度，每一主题只列出概率排名前 10 的词语，可以看出，主题模型可以提取出媒体、灾情、医药、军事、农村、经济、网络、治安、电影、房地产等不同的语义维度。将文档映射到这些主题维度，一方面可以得到篇章级、段落级以及句子级等较长文本的固定维度的向量表示，这对于神经网络的训练是必须的；另一方面，还可以根据文档在这些主题维度的概率分布

得到文档的语义信息，进而根据不同的语义内容推断相应的情感变化。

表 4-2 主题模型训练结果示例

Table 4-2 Samples of the results of topic model

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10	Topic11	Topic12
中国	重建	药品	军队	农村	医疗	币	网	西藏	犯罪	电影	房产
声	震	生产	国防	农民	卫生	汇率	网络	藏	公安	漫	市场
中央广播电台	川	药	中央军委	农业	医	年	网站	治区	机关	影片	调控
广播	后	企业	军事	城镇	基	外汇	信息	高原	公安部	部	房价
直播	灾	国家	建设	城乡	病	美元	联网	拉萨	案件	票房	政策
媒体	恢复	剂	军	化	服务	货币	色情	青藏	警	中国	城市
节目	援建	中药	主席	新	医院	国家	手机	年	嫌疑人	动画	房
今天	灾区	限公司	全军	发展	机构	储备	播	生	民警	影视	卓
报道	汶川	使	新	土	公	中国	淫秽	农牧民	公安局	档	涨
电台	镇	食品	战	城市	改	升值	网民	族	击	市场	价

4.5 实验与讨论

本节将通过实验验证基于 VDSN 的情感预测模型的有效性以及言语情感描述体系中对于各情感成分间影响关系的设定的合理性。在这之前，将首先对实验采用的评价指标以及针对参数设置、数据规模的测试实验进行介绍。

4.5.1 评价指标

我们使用均方根误差（Root-Mean-Square-Error，RMSE）作为实验的评价指标，一方面由于本文所提出的言语情感描述体系采用多维空间表示，RMSE 能够刻画预测向量与目标向量在多维空间的距离；另一方面由于本文的预测任务不同于传统的情感分类任务，每个维度上有表示不同强度的刻度值，因此更倾向于回归预测，通过 RMSE 不仅可以看出预测结果的正确与否，还能看出偏离目标值的误差大小。RMSE 越小，则预测结果越好，计算公式为：

$$RMSE = \sqrt{\sum_{i}^{N} \left\| \mathbf{y}_i - \mathbf{t}_i \right\|^2 \Big/ N} \tag{4-25}$$

其中，i 为样本索引，N 为测试样本数，$\mathbf{y}_i$ 为预测模型的输出向量，$\mathbf{t}_i$ 为目标向量，即标注数据。

4.5.2 准备实验

准备实验的目的是为了确定训练主题模型和神经网络模型时的参数设置，以及训练数据规模对于预测结果的影响。采用第 3 章建立的情感语料库，其中 150 篇为有标签数据，8450 篇为无标签数据，无标签数据分两次加入到训练集与有标数据一起参与网络的预训练：第一次先加入一小部分（450 篇）无标数据测试加入无标数据能否提升预测效果；第二次再加入剩余全部无标数据（8000 篇）测试扩大无标数据的规模能否进一步改善预测结果。如此，训练数据分为三个不同规模的数据集：小规模（150 篇，全部有标）、中等规模（600 篇，150 篇有标+450 篇无标）和大规模（8600 篇，150 篇有标+8450 篇无标）。

（1）主题模型参数设置

LDA 主题模型用于文本特征的降维，有三个主要参数需要提前设定：主题数和两个超参数α、β。主题数即降维之后的文本特征维度，决定网络输入层的节点数，一定范围内增加主题数有助于提升主题模型对于数据集的刻画能力，但维度过多会增加网络预测模型的时间和空间开销，在样本有限的情况下还易造成过拟合。超参数α控制着文档在主题上的分布（“文档-主题”分布）的稀疏性，一个较小的α意味着模型倾向于用较少的主题描述文档；与之类似，β控制着主题在词项上的分布（“主题-词项”分布）的稀疏性，一个较小的β意味着模型倾向于分配给每个主题较少的词项。这三个参数存在交错复杂的相互影响，因此很难单独估计某一参数的取值。Griffiths[112]通过在不同数据集上的测试给出两个通常情况下表现良好的超参数的经验值，$\alpha=50/T$（T 为主题数），$\beta=0.01$。我们首先利用经验值固定超参数以进行主题数的估计，然后固定主题数和一个超参数，测试另一超参数的经验值是否可用。

通常情况下，主题模型对数据集的刻画能力通过模糊度（perplexity）[111]来评价。模糊度是一个数值上与单个词语平均似然度成反比的指标，值越低则表示模型对数据集的刻画越精准。这里采用模糊度作为评估参数好坏的指标，计算公式为：

$$perplexity=\exp\left(-\sum_{d}^{D}\log\left(p\left(w_d\right)\right)\Big/\sum_{d}^{D}N_d\right) \tag{4-26}$$

其中，D 为数据集的文档数，N_d 为每篇文档包含的词语数，$\log p(w_d)$ 表示文档中每个词语的对数似然概率。

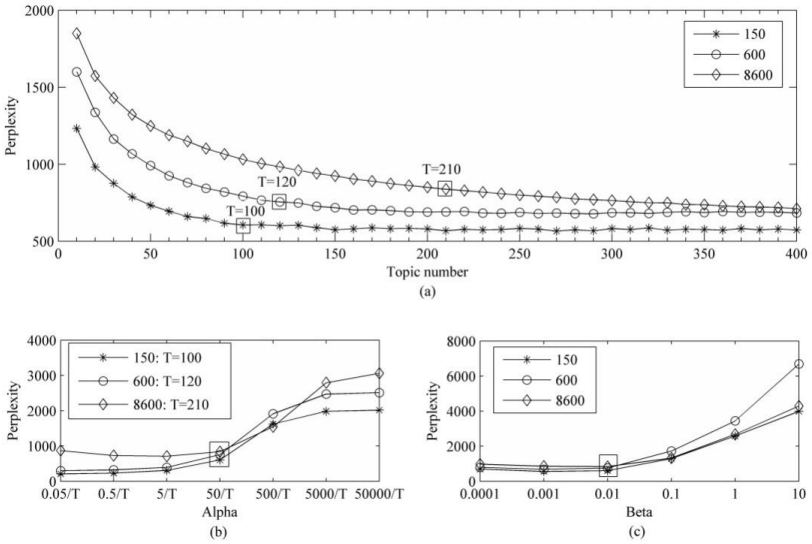


图 4-11 LDA 主题模型不同参数下模糊度分布曲线。(a) 不同主题数下模糊度的分布曲线；(b) 不同α下模糊度的分布曲线；(c) 不同β下模糊度的分布曲线。

Fig. 4-11 The perplexity curves varying with the three parameters in LDA model: (a) the perplexity curves varying with topic number; (b) the perplexity curves varying with α; (c) the perplexity curves varying with β.

图 4-11 给出在三种不同规模数据集下，LDA 主题模型随参数变化模糊度的分布曲线。其中，(a)为随主题数变化模糊度的变化，此时$\alpha=50/T$（T为主题数），$\beta=0.01$；(b)为超参数α的变化引起的模糊度的变化，主题数设定为由图(a)确定的主题数，$\beta=0.01$；(c)为超参数β的变化引起的模糊度的变化，同样主题数设定为由图(a)确定的主题数，$\alpha=50/T$。通过图(a)可以看出，随着主题数在[10,400]范围内变化，三个训练集的模糊度均会显著下降之后趋于平缓。我们将小、中、大三个训练集的主题数分别设定为 100、120 和 210，分别为三条曲线正切值开始小于 1 的点对应的主题数，意味着继续增加主题数，模糊度的下降变得不明显或者不再下降。固定了主题数和其中一个超参数，另一个超参数的变化引起的模糊度的分布曲线如图(b)和(c)所示，随着超参数的变大，模糊度值先是变化不明显之后显著上升。从图中可以看出，两个经验值均提供了较低的模糊度值，意味着它们是合理且可用的。

（2）神经网络参数设置

我们对三个神经网络训练有关的参数设置进行了测试，分别是：隐含层单元数、微调阶段的迭代次数和预训练阶段 RBM 的循环周期。测试采用的评价指标是发音

描述的预测误差 RMSE，因其将作为情感预测模型最终的输出结果。因为标注样本有限，采用五折交叉验证（5-fold Cross Validation）的方法进行网络的性能验证，即：将全部标注数据随机分成五等份，每次（one fold）取一份作为测试集，其余部分作为训练集（与无标注数据一起）。五折运算取平均作为用于评估的预测误差。此外，由于神经网络预测结果的随机性，运行 10 次五折交叉验证，取其平均结果作为最终预测误差。

图 4-12 给出隐含层单元数、微调迭代次数和 RBM 循环周期三个参数对于发音描述预测结果的影响。图(a)为随着隐层单元数的变化发音描述的 RMSE 的变化曲线，此时另两个参数设为 10。可以看出，随着隐层单元数的增加，三个训练集的预测误差均先下降后上升，尤其是最小规模数据库的误差增幅最大，分析原因可能由于隐层单元数过多而数据规模较小从而导致了过拟合现象。当采用 HVDSN 时，由于隐含层节点数是逐层扩张的，起始模块隐含层节点数不能过少而发生欠拟合，终止模块的隐含层节点数又不能过多而引起过拟合，因此将最底层的模块即认知模块的隐含层节点数设为 4，依次递增 2，心理模块、生理模块和发音模块的隐含层节点数分别设为 6、8、10。在 IVDSN 与 HVDSN 对比时，IVDSN 隐层节点设置与 HVDSN 相同。图(b)给出微调阶段的迭代次数对预测结果的影响，此时 RBM 周期设为 10，隐层节点数采用实验(a)确定的数目进行设置。可以看出，当迭代次数较小时（≤10），三个训练集的 RMSEs 均较小且变化不明显；随着迭代次数增加三个训练集的 RMSEs 均有所上升。因此最终将迭代次数设定为 10。图(c)为 RBM 的循环周期对于训练结果的影响，可以看出影响并不显著，其中最小规模的训练集的 RMSE 在振荡，其他两个训练集的 RMSEs 先略微下降之后趋于平稳。最终将 RBM 的循环周期数也设为 10。

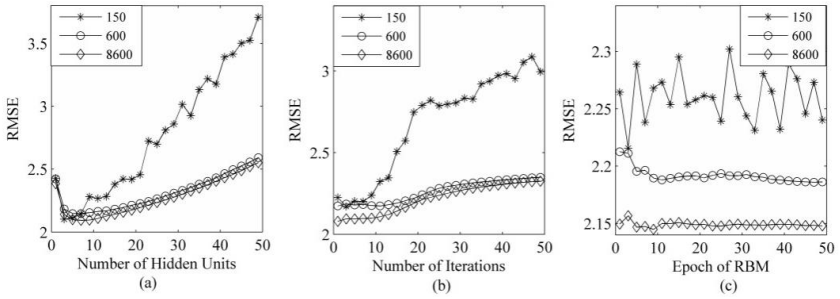


图 4-12 神经网络训练参数对发音描述预测误差（RMSE）的影响。(a) 不同隐含层单元数下 RMSE 分布曲线；(b) 不同微调迭代次数下 RMSE 分布曲线；(c) 不同 RBM 循环周期下 RMSE 分布曲线。

Fig. 4-12 The effects of the three parameters in neural network training for the prediction of utterance manner. (a) the RMSEs of the three data sets varying with the hidden unit number; (b) the

RMSEs of the three data sets varying with the number of iterations; (c) the RMSEs of the three data sets varying with the epoch of RBM.

（3）数据规模的影响

从图 4-12 中可以看出不同规模的训练集 RMSE 分布在不同的区间，在图 4-12(c)中这种差异尤为明显。为了进一步明确数据规模对于训练结果的影响，我们给出不同训练集下四种情感成分直接由文本特征预测而不考虑其他成分影响的 RMSE，如表 4-3 所示。从表中可以看出，最小的数据集四种情感成分的预测误差均最大，中等规模数据集的 RMSE 也处于居中水平，最大规模的数据集 RMSE 最小。这说明在有标数据基础上增加无标注数据参与网络的预训练有助于网络性能的提升；进一步扩大无标数据的规模可以进一步提升预测效果。之后的实验均采用最大规模的数据集（8600 篇）进行网络的训练与测试。

表 4-3 不同规模训练集各情感成分的预测结果（RMSE）

Table 4-3 The RMSEs of the four components based on different data sets

	150	600	8600
认知评价	2.29	2.17	2.10
心理感受	1.72	1.70	1.64
生理反应	0.63	0.62	0.61
发音描述	2.25	2.19	2.14

4.5.3 验证实验

确定了参数设置和数据规模之后，将通过实验验证基于 VDSN 搭建的言语情感预测模型的有效性以及言语情感描述体系的合理性，主要包括：（1）验证中间环节对于后续任务的影响，即中间层可见对预测任务的作用；（2）将 VDSN 与其他中间层不可见的深度网络对比，验证中间层部分可见化对于深度网络的优化作用；（3）维度均衡性调整和样本均衡性调整等措施对于网络的进一步优化作用。

（1）中间环节的作用

该实验以从文本特征直接生成言语情感各成分而不考虑其他环节影响的情况作为基准（表 4-4 中斜体行），分别与考虑了中间环节的各种情况作对比，当加入某些中间环节后预测误差变小则说明中间环节对于预测目标有促进作用，考虑该影响的情感描述体系具备合理性。考虑中间环节的情况分为包含全部中间环节和部分中间环节的多种情况，表 4-4 列出了不考虑反馈情况下各成分所有可能的预测方式和结果，分别采用 IVDSN 和 HVDSN 两种网络。

表 4-4 IVDSN 和 HVDSN 采用不同路径预测言语情感各成分的结果（RMSE）

Table 4-4 The predicted results (RMSE) of each component of speech emotion through different paths by IVDSN and HVDSN

预测路径	IVDSN	HVDSN
文本->认知	**2.07**	**2.07**
文本->心理	*1.59*	*1.58*
文本->认知->心理	**1.55**	**1.55***
文本->生理	*0.63*	*0.64*
文本->认知->生理	0.62	0.64
文本->心理->生理	0.62	0.62
文本->认知->心理->生理	**0.62**	**0.65**
文本->发音	*2.16*	*2.16*
文本->认知->发音	2.11	2.11
文本->心理->发音	2.13	2.11
文本->生理->发音	2.11*	2.11**
文本->认知->心理->发音	2.09**	2.08*
文本->认知->生理->发音	2.11***	2.09*
文本->心理->生理->发音	2.12**	2.09*
文本->认知->心理->生理->发音	**2.11****	**2.09**

*: p<0.05; **: p<0.01; ***: p<0.001

从表 4-4 可以看出，采用 IVDSN 和 HVDSN 两种网络的情况下，加入其他成分的影响均可一定程度降低预测误差。对各模块考虑其他成分影响的情况与未考虑其他成分影响的情况作差异显著性分析，可以得出：当采用 IVDSN 时，“文本->生理->发音”、“文本->认知->心理->发音”、“文本->认知->生理->发音”、“文本->心理->生理->发音”、以及“文本->认知->心理->生理->发音”的结果均与“文本->发音”的结果有显著差异，说明增加其他成分的影响有助于发音方式的预测；当采用 HVDSN 时，“文本->认知->心理”的预测效果要显著优于“文本->心理”，说明认知的加入有助于心理感受的预测，“文本->生理->发音”、“文本->认知->心理->发音”、“文本->认知->生理->发音”、“文本->心理->生理->发音”的预测效果也显著优于直接从文本预测发音。心理和认知对于生理的预测结果提升作用不明显，分析可能由于该成分预测误差本来就小，因此可提升空间不大。这些结果反映出中间环节对于后续任务的影响，考虑其他情感成分的影响可以提升网络的预测效果。最终确定的言语情感四种成分的预测方式如表 4-4 中黑体行所示，按生成顺序将结果逐步累积参与到后续目标的预测，与图 4-8 所示网络拓扑结构一致。

表 4-4 中还显示，四种成分的预测误差分布范围差异较大，这与 RMSE 的计算方式有关，RMSE 衡量的是多维表示的空间距离，与维度有关，如：发音描述计

算的是预测结果与标注值在 7 个维度上的距离平方和，而生理反应的描述空间只有 2 维。因此，当希望知道每一维度上的预测误差，需要根据公式 $RMSE_{\dim}=\sqrt{RMSE^2/dimensionality}$ 对 RMSE 进行转换。转换之后四种成分平均每个维度上的预测误差分别为 0.89，0.24，0.42 和 0.79，可见误差值均在 1 个刻度以内，认知和发音为 7 级刻度，误差较大于其他两种成分，但均可视作偏移程度可以接受的预测。

（2）与其他深度神经网络对比

本实验通过将两种中间层部分可见的网络与其他深度网络对比，以验证中间层的可见化对于网络性能的提升作用。以 DBN 作为未对中间层进行监督且未引入先验知识的神经网络，DSN 作为对中间层进行监督但未引入先验知识的网络，IVDSN 和 HVDSN 则为对中间层进行监督且引入先验知识的网络。通过 DSN 与 DBN 的对比可验证选用 DSN 作为基础网络，即增加对中间层监督的必要性；IVDSN、HVDSN 分别与 DSN 比较则可以验证引入先验知识将中间层部分可见化的效果；另外 IVDSN 和 HVDSN 也可以进行比较而测试两种堆叠方式的优劣。除此之外，本实验还考虑了网络深度对其性能的潜在影响。对于 DBN 来说，网络深度由隐含层数目决定，这与传统的深度神经网络一致；而对于 DSN、IVDSN 和 HVDSN 来说，每个训练模块均为单隐层的神经网络，因此将其作为深度网络考虑时网络深度则由模块数决定，其中，IVDSN 和 HVDSN 的模块数与经由的中间环节的个数有关。

表 4-5 列出了不同网络在不同深度条件下由文本特征预测发音方式的结果。其中，IVDSN 和 HVDSN 网络深度为 1 表示从文本直接预测发音的情况；网络深度为 2 表示从文本经认知或心理或生理某个中间环节预测发音的情况，表 4-5 中结果为三种情况的平均值；网络深度为 3 表示从文本经认知、心理、生理中某两个中间环节预测发音的情况，共有三种组合方式（如表 4-4 所示），表 4-5 中结果为三种情况的平均值；网络深度为 4 则表示从文本经认知、心理和生理三个中间环节预测发音的情况。

表 4-5 不同深度神经网络在不同深度条件下发音方式预测结果（RMSE）

Table 4-5 The predicted results (RMSE) of utterance manner by different deep networks and in different depths

网络深度	DBN	DSN	IVDSN	HVDSN
1	2.28	2.17	2.16	2.16
2	2.27	2.13	2.12	2.11
3	2.27	2.15	2.11	2.09
4	2.27	2.15	2.11	2.09

从表 4-5 可以看出，DSN 相对 DBN 预测效果有明显提升（*p*-value=0.0002），验证了对中间层进行监督对于深度神经网络性能的优化作用，以及选用 DSN 作为基础网络的优势所在。IVDSN 与 DSN 相比性能又有进一步提升（*p*-value=0.02），说明在网络结构相同的情况下，中间层的部分可见化可以进一步提升网络性能，同理，HVDSN 的性能也显著优于 DSN（*p*-value=0.03）。HVDSN 与 IVDSN 的对比中 HVDSN 的性能更优（*p*-value=0.04），一方面可能由于 IVDSN 输入层维度多于 HVDSN 的输入层，因此网络规模上 HVDSN 更精简，这在数据有限情况下会对训练结果产生一定影响；另一方面可能与已知信息和未知信息的相对作用有关，HVDSN 中，已知信息与未知信息的节点数相差不大，且已知节点占主导，而 IVDSN 中两部分特征维度相差较大且未知信息占主导，因此 IVDSN 中已知信息的作用可能被削弱，这也从侧面反映出信息可见化的重要；此外，数值分布情况也可能对结果造成影响，HVDSN 中已知信息与未知信息属于同样的数值分布（均为隐含层节点），而 IVDSN 中子目标输出值与文本特征的分布不同，可能会对两部分信息的融合产生影响。

纵观网络深度对各深层网络的影响，DBN 与 DSN 随网络深度的增加性能提升效果不如 IVDSN 与 HVDSN 明显，这意味着在数据规模一定的情况下，单纯增加网络深度并不总是能提升网络的性能，在此基础上增加一定的先验知识的指引可以进一步优化网络的性能。

（3）优化措施的作用

在 4.2.3 节，我们提出两种均衡调整措施用于进一步优化网络的性能：维度均衡调整和样本均衡调整，此实验用于验证两种优化措施的作用。由于维度均衡调整仅在使用 IVDSN 时有必要进行，因此该实验在 IVDSN 上运行，隐含层单元数均设为 10，其他参数同上。表 4-6 列出不加入任何均衡性调整措施以及依次加入两种调整措施之后的预测结果对比。从“Original”列可以看出，在各模块隐含层单元数均设为 10 的情况下也可以得到与表 4-4 中隐含层递增设置相接近的结果，即考虑其他成分的影响可以降低预测误差，且随着堆叠更多的中间环节的结果误差降低幅度有所增加。当加入维度均衡调整之后（“Adj1”列），这种误差随路径变长逐渐降低的趋势更为明显，这说明了调整维度影响之后可以更有效的体现情感成分间的相互作用，且对于整体的预测效果也有所提升。在此基础上，进一步加入样本均衡调整之后（“Adj1+Adj2”列），多数情感成分的预测误差均有所下降，即使直接从文本特征预测各情感成分预测误差也变得更小，充分说明考虑样本数量影响的必要以及对于样本均衡调整措施的有效性。用 t-检验分别分析各列间的差异，结果显示 *p*-value 均小于 0.05，说明这两种调整措施的优化效果显著。

表 4-6 通过不同预测路径及调整措施预测四种情感成分的结果（RMSE）。其中“Original”列表示未加入任何均衡性调整措施的预测结果；“Adj1”列表示加入维度均衡调整之后的预测结果；“Adj1+Adj2”列表示在此基础上进一步增加样本均衡调整之后的预测结果。

Table 4-6 The RMSEs of the four components via different stacking patterns and adjustments, where “Original” means the results without adjusting the imbalances, “Adj1” denotes the results after adjusting the dimensionality imbalance, “Adj1+Adj2” refers to the results further adding the adjustment on the sample imbalance.

	Original	Adj1	Adj1+Adj2
文本->认知	**2.08**	**2.08**	**2.07**
文本-> 心理	*1.66*	*1.66*	*1.58*
文本->认知->心理	**1.63**	**1.62**	**1.57**
文本-> 生理	*0.61*	*0.61*	*0.67*
文本->认知->生理	0.60	0.60	0.67
文本->心理->生理	0.61	0.61	0.64
文本->认知->心理->生理	**0.60**	**0.59**	**0.64**
文本-> 发音	*2.15*	*2.15*	*2.07*
文本->认知->发音	2.14	2.13	2.07
文本->心理->发音	2.11	2.06	2.03
文本->生理->发音	2.15	2.10	2.02
文本->认知->心理->发音	2.08	2.04	2.01
文本->认知->生理->发音	2.11	2.10	2.04
文本->心理->生理->发音	2.09	2.06	2.01
文本->认知->心理->生理->发音	**2.10**	**2.05**	**2.01**

4.5.4 讨论

我们通过对深度神经网络中间层的部分可见化实现了基于言语情感生成过程的情感预测建模。对于深度网络中间层可见化的已有研究[88-90]停留于基于相同目标（最终目标）的中间层可监督训练，仍缺乏对中间层真正含义的挖掘。我们在最终目标的基础上引入不同的子目标，相当于为网络训练引入了联合目标，训练结果既满足最终目标的误差最小，同时也满足各子目标的误差最小。联合目标的引入相当于在经验风险上增加了一个正则项，即在最小化经验误差同时约束网络的结构。从贝叶斯估计的角度来看，该约束对应于模型的先验分布。如果先验合适，则得到的学习结果更倾向于真解，这也解释了引入子目标之后最终目标的预测误差更小

的原因。

联合目标的训练可以先于最终目标而分步完成，也可与最终目标同步训练。文献[88]和文献[89]在对中间层进行监督或半监督学习时都采用联合目标与最终目标同步训练的做法。同步训练的方式沿用深度神经网络全局运用反向误差传递的训练模式，代价函数仍是一个含多个极小值的高度非凸空间，因此可能使网络最终收敛于局部最小，同时在反向传播的过程中还可能发生梯度消失的情况。我们采用分步训练的方式，每个模块为一浅层神经网络，代价函数由高度非凸变成凸函数或近似凸函数，降低了网络训练复杂度并提升了网络收敛于全局最优的可能。

本文中我们是先通过大量引入相关领域的研究成果汇总出可能与言语情感生成有关的多方面影响因素，并进一步推断出其可能的相互关系和组织结构，从而进行发音描述（最终目标）以及其他子目标的设定。除了本章中的应用，子目标还可有另一种解读，如在自然语言处理中常涉及的韵律层级结构，低层小尺度单元的训练任务可视作高层大尺度单元的子目标，从而由细到粗汇总出整个句子或篇章的训练结果（句子级情感分析常用的做法）；或者反之高层大尺度单元也可先于低层小尺度单元进行处理，从而为小尺度单元提供上下文参考（下一章的应用）。在图像处理领域，这种由细到粗或由粗到细的层级结构仍然存在，在人脸识别的研究中已经发现某些中间层学习到了组成人脸的结构特征（边缘或器官），验证了人脑视觉系统由具体到抽象逐层组合低层特征的分级处理模式。但是，基于深度网络自动发现文本或图像中结构特征的研究都离不开大规模数据的支持，组建的网络规模也很庞大，我们希望通过人为设定某些中间层的子目标从而使网络能快速高效地学习到期待的特征，减少数据和网络规模的开销。

4.6 本章小结

为表现多视角言语情感描述体系中各成分的相互关系及生成过程，我们采用具有多层非线性映射结构的深度神经网络搭建言语情感的预测模型。为将深度神经网络中间层可见化，基于深度堆叠网络（DSN）提出两种中间层部分可见的网络——输入层部分可见的深度堆叠网络（IVDSN）和隐含层部分可见的深度堆叠网络（HVDSN），从而实现按照既定的言语情感生成过程人工干预深度网络的结构设置和学习过程。中间层的部分可见意味着赋予其明确的含义和显性的相互关系，在可见同时保留网络对于未知信息的提取能力和一定的容错能力。验证实验结果表明：

（1）言语情感生成过程中各环节存在相互影响，考虑该影响可以提升发音描述的预测效果；

（2）中间层的部分可见化有助于提升深度网络的性能，尤其在数据规模有限的情况下，加入适当的先验知识可以使预测结果更趋近于真解，这也反向验证了言语情感生成及衍化过程的合理性；

（3）对于输入特征维度以及样本数量的均衡性调整措施可以进一步优化网络的性能；

（4）在 IVDSN 和 HVDSN 两种网络的对比中，对当前的预测任务来说，HVDSN 性能略优于 IVDSN。但是二者均可实现中间层的可见化建模，且均具有良好的可扩展性和简便易操作的训练过程。

言语情感生成过程以及多视角情感描述体系的提出为深度网络的建模提供了结构化指导，同时网络学习到的结果又为言语情感的预测提供了结构化特征。这里的结构化特征主要指同一尺度分析单元内言语情感各成分之间的级联型结构，下一章将在此基础上讨论不同尺度的情感信息间的层级型结构关系。

5 多尺度情感预测建模

情感的生成过程错综复杂，期间可能涉及来自不同方面的影响因素，上一章介绍的是同一文本单元内不同情感成分间的相互影响；同时，基于文本的情感预测还可能牵涉来自不同尺度的特征参数的相互交融。如何对这些特征的相互关系进行梳理，并将其与来自同一尺度的其他情感信息进行融合，构建更为健全的言语情感预测模型，是本章的主要研究内容。

5.1 问题分析

在情感语音合成的研究中，情感分析的尺度常定位于句子级，一方面由于句子级文本比短语和词包含更丰富的语义和情感信息，对于情感的判别更容易也更准确，另一方面由于句子级情感相比篇章级情感更细腻，能反映篇章或段落内部情感的衍化。本文中，输入文本是篇章级文档，我们将情感分析单元定为句子级，篇章和段落相对于句子级是更大尺度的分析单元。考虑到句子级可利用信息的有限性，对篇章级和段落级也进行情感预测，并将这些粗粒度下提取到的情感特征作为句子级情感预测的上下文环境，为句子级情感的衍化提供变化基调，同时扩展句子级情感预测的可用信息。

从篇章到段落再到句子这种由上至下的分层情感分析构成情感特征间的多尺度层级结构。上一章提到，多尺度层级结构也可视为另一种结构关系已知的情况（层级型结构关系），不同尺度信息的融合也可通过深度堆叠网络实现，不同尺度下的情感信息可以作为其他尺度下目标任务的子目标，因此多尺度情感建模也可以利用中间层部分可见的深度堆叠网络实现。除了不同尺度间的相互影响，上下文环境还有一种更为具体的存在形式，即同一尺度下不同文本单元间的影响。除此之外还有上一章所研究的同一文本单元内不同情感成分间的影响。综合这些影响，我们基于中间层部分可见的深度堆叠网络搭建了更为完善的支持多尺度情感分析的预测模型。

接下来将介绍多尺度情感预测模型的网络结构和多尺度下文本特征提取的主要步骤，然后给出整个系统的框图并对多尺度情感预测网络的性能进行测试。

5.2 多尺度情感预测模型

影响情感产生的因素有很多，本文关注的是由语言符号（文字）刺激产生的言

语情感。主要考虑两方面的影响因素：1）言语情感产生过程中各步骤之间的相互影响；2）不同的文本分析单元间的相互影响。第 2）点既包括不同尺度的文本单元间的影响，又包括同一尺度内不同分析单元间的影响；而第 1）点讨论的是同一分析单元内部情感的衍化过程。由此，就构成了由上至下、由粗到细的情感信息生成及衍化过程。图 5-1 给出了综合多尺度影响因素的文本-情感预测模型示意图，由上至下依次对篇章级、段落级和句子级的各种情感成分进行预测，在预测时融合同一分析单元内部各情感成分的影响和不同分析单元间的影响，大尺度特征为小尺度的预测提供全局上下文参考，小尺度单元间提供局部上下文参考。

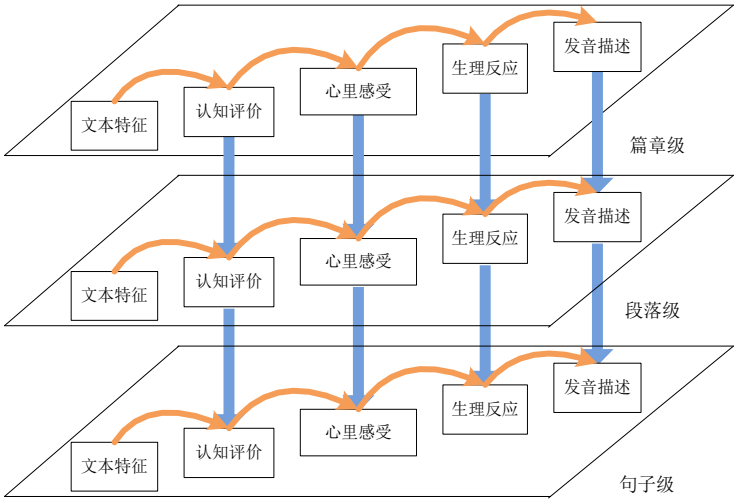


图 5-1 多尺度文本-情感预测模型网络结构

Fig. 5-1 The network of the multi-scale text-based emotion predicting model

图 5-1 中橘色弯箭头表示同一文本单元内部不同情感成分间的相互影响，即前面模块的输出依次堆叠到后面模块作为部分输入（本章使用的网络为 IVDSN）。图 5-1 出于可观性考虑仅给出方向指示性连接，部分连接未给出，实际实现过程中前面模块的所有输出都可输送给后面模块作为部分输入，这样做的好处是确保信息的完整利用，不足是会扩张网络规模，造成更多的网络参数，因此可根据所拥有训练数据的多少对堆叠信息进行取舍。

对于不同文本分析单元间的影响，在多尺度情感分析下可以有两种存在形式：一种是同一尺度不同文本分析单元间的影响，这是常用的构建当前所分析单元上下文环境的方法，类似于语言模型中的 n-gram；另一种是不同尺度间分析单元的

交互作用，该作用可以由粗粒度到细粒度，粗粒度信息为细粒度信息提供变化的基调，也可以由细粒度到粗粒度，通过整合细粒度信息推断出粗粒度信息，如基于关键词的句子级情感分析就是通过将当前句中所有情感关键词对应的情感属性进行求平均或取最大值等整合方式得到整句话的情感属性。此处我们关注的是句子级情感分析单元的上下文环境，因此不同尺度间的影响采用由粗到细的模式。在图 5-1 中用蓝色箭头表示大尺度信息对于小尺度单元的作用，篇章级可以作为段落级的上下文环境，段落级作为句子级的上下文环境，同样篇章级也可直接影响句子级，即当前句所在篇章和段落的情感信息均可传送给句子级单元作为基准参考，图 5-1 同样出于可观性考虑没有画出全部连接。

基于图 5-1，利用中间层部分可见的深度堆叠网络对多尺度情感预测进行建模。首先，提取篇章级情感信息，按第 4 章所述的生成顺序，采用 IVDSN 网络，各模块依次生成认知评价、心理感受、生理反应和发音描述四种成分，前面模块的输出逐层堆叠到后面模块的输入层作为部分已知信息，同时文本特征作为包含未解信息的原始输入传送给每个模块。之后，提取段落级情感信息，不同情感成分间的堆叠方式与篇章级相同，只是在每个模块堆叠本段落其他情感成分的同时，当前段所在篇章对应模块的输出也作为部分输入传送给段落级模块。为避免输入层节点扩张过多而造成网络过拟合，不同文本单元间的影响未采取四种情感成分全部传送的方式，只传送与当前预测目标相同的上级单元的预测结果，比如：段落级心理感受的预测，输入层包含本段落的文本特征、认知评价模块的输出结果以及篇章级心理感受模块的输出结果。最后，预测句子级情感信息，与上文叙述方式类似，不同情感成分的堆叠采用逐层累加的方式，段落级和篇章级情感信息可以均作为句子级的上下文信息进行传递，也可视情况选择其中一个，同样篇章或段落的情感成分只传递给句子级预测目标相同的模块。另一种上下文影响的实现形式与之类似，由于情感衍化存在一定连续性，我们选用当前分析句的前一句的情感预测结果作为其部分输入，前一句的预测结果也采用只传递给相同预测模块的方式，以避免输入层维度扩张过于严重。

至此，通过分层处理的方式搭建了多尺度文本-情感预测模型，考虑了来自不同尺度和不同方面影响因素的作用，各尺度内部遵从言语情感生成过程对于情感成分间相互关系的设定，不同尺度间采用由粗到细的层级结构，上级单元为下级单元提供变化的参考，下级单元的预测结果反映上级单元内部的衍化过程。在言语情感产生过程中，发音描述作为最终目标，其他情感成分作为其子目标，子目标依次作为后续预测目标中的已知信息，该生成关系构成级联型结构关系。不同尺度的依次生成与相互作用构成层级型结构关系，大尺度单元的预测目标作为小尺度单元的子目标，同样也为小尺度单元的预测提供已知参考，因此也可采用中间层部分可

见的深度堆叠网络进行建模。本章仅使用 IVDSN 进行实验以说明加入上下文影响的作用，同样也可采用 HVDSN 进行建模，即段落级或篇章级或前一句训练的隐含层作为句子级对应模块的部分隐含层的初始值。对于大尺度单元到小尺度的堆叠也许比较陌生，但同一尺度不同文本单元间训练结果的堆叠在其他应用中并不陌生，如语言模型中相邻词语的特征表示（词向量）的堆叠，以及语音识别中音素的特征参数的堆叠。训练特征的堆叠一定程度上可以扩展输入特征的时间跨度，这是深度学习在很多领域取得突破性进展的原因之一。

5.3 多尺度文本特征提取

基于篇章级文档的多尺度文本特征的提取主要经历分词、文本分割、特征词提取和特征降维四步，其中，分词仍采用 NLPIR 汉语分词系统，将连续文本分割成词序列；文本分割的目的是提取段落和句边界，将篇章级文档分割成段和句；特征词提取与第 4 章相同，将滤掉功能词等虚词之后的词称为内容词，HowNet 情感词典中的词作为情感词，内容词和情感词都作为可能负载情感特征的特征词；特征降维仍采用 LDA 主题模型，分别提取“文档-主题”分布、“段落-主题”分布和“句子-主题”分布作为三个尺度单元降维之后的文本特征。以下重点介绍第 4 章未涉及到的文本分割和多尺度的文本特征降维。

5.3.1 文本分割

本文中文本分割即将整篇文档分割成段落和句子。句边界可以通过标点符号（句号、问号、感叹号和省略号）判别。段落可被视为语义关联的文本块，而语义关联又可以被认为是具有相似的主题，因此，语义信息可以通过主题模型建模，而面向段落的文本分割可以采取基于主题模型的方法。TextTiling[113]是最早的非监督线性主题分割方法。该方法首先通过 LDA 提取文档集的“主题-词项”分布，分布中每个元素表示该位置对应词语属于该位置对应主题的概率，通过选取概率最大的主题为每个词语设定其对应的唯一主题 ID，然后用这些主题 ID 替代文档中的词语，将文档表示成主题序列。之后通过设定固定的窗长首先将文档切分成一系列的主题块，以主题块为基本单元分别计算每两个相邻主题块的余弦相似性，值越趋近于 1 表示这两个主题块语义相似性越高，趋近于 0 则表示这两个主题块语义相似性越低。接下来计算每个主题块间隔的“深度值”，用以描述余弦相似性的变化，深度值的计算公式如式(5-1)所示，其中 s_i 表示第 i 个主题块间隔对应的余弦相似性值，$hl(i)$ 表示该间隔左侧余弦相似性的峰值，$hr(i)$ 表示该间隔右侧余弦相似性的

峰值。若给定要分割的段落数n，则对深度值进行排序，选取前n个深度值对应的主题块间隔作为段落边界；否则设定一个阈值（通常采用$\mu-\sigma/2$，其中μ为所有深度值的期望，σ为深度值的标准差），若某个间隔对应的深度值大于阈值，该间隔则被认为是主题变化显著的位置，即段落边界。

$$d_i=\left(hl(i)+hr(i)-2s_i\right)/2 \tag{5-1}$$

TextTiling 的缺陷是可能使段边界落在一句话内部，TopicTiling[114]的方法在此基础上做了改进，以每句话作为基本单元，替代 TextTiling 中的主题块。我们的做法基于 TopicTiling 方法。由于句子长度不固定，而相似性的计算需要长度一致的特征向量，因此 TopicTiling 采用长度为主题数的向量表示每句话，每个元素为该位置所对应主题 ID 在当前句中出现的频数。然后利用这些向量计算每两个相邻句子间的余弦相似性，同样余弦相似性值越大代表两句话的语义相似性越高。之后搜寻这些余弦相似性的局部最小值作为候选位置，利用式(5-1)计算这些局部最小点的深度值，这是 TopicTilling 与 TextTiling 的另一个不同。但是这种做法对本文的应用并不适合，本文采用的数据集篇章长度有限，筛选出的局部最小点也非常有限，因此采用仅计算局部最小值的方式很容易造成文本的欠分割，所以我们直接计算了所有句间隔的深度值。另外，TopicTiling 也是通过设定边界数然后对深度值排序的方式或者设定阈值的方式决定最终的边界位置。我们采用设定阈值的方式，该方法可以自动决定边界数，但是不同于两种方法常用的$\mu-\sigma/2$（该阈值会造成我们所用文档集的过分割），我们将该阈值调大为$\mu+\sigma/2$，从而获得边界数目适当的分割结果。

5.3.2 多尺度特征降维

多尺度文本的特征降维同样通过 LDA 主题模型进行，模型具体训练过程已在第 4 章给出，这里介绍在文本单元具有多个尺度时的处理方式。首先通过 LDA 训练所有篇章级文档的主题模型，模型超参数和主题数根据上一章的测试实验设定。模型训练结果会得到两个分布：每篇文档在所有主题上的多项式分布（简称“文档-主题”分布）和每个主题在所有词项上的多项式分布（简称“主题-词项”分布）。其中，“文档-主题”分布直接作为篇章级文档降维之后的文本特征，特征维度从词典集大小降为主题的个数，每个元素对应当前文档属于各主题的概率。段落级和句子级的降维文本特征通过 LDA 的另一训练结果“主题-词项”分布得到。利用该分布，将文档中每个词替换成其所对应的最大可能的主题 ID（做法与文本分割中方法一致），然后通过计算每一段中每个主题出现的频数可以得到段落在所有主题数上的多项式分布（利用式(4-24)计算），简称“段落-主题”分布，被用于段落级文

本的降维特征。句子级的处理方式与段落类似，即计算每一句中每个主题出现的频数并通过式(4-24)计算“句子-主题”分布，用作句子级文本的降维表示。

5.4 系统框图

综合上述所有对基于文本的多尺度情感预测模型的介绍，图 5-2 给出模型整体框图及其在情感语音合成系统中的位置。模型主要分为两个模块：语言学分析（linguistic analysis）模块与情感生成（emotion generation）模块。这两部分的输出均会作为情感语音合成系统的输入传送给语音合成系统。其中，语言学分析的结果也为情感信息的生成提供文本特征。语言学分析包含分词处理、文本分割、特征词过滤和特征降维几步，其中，文本分割的结果是获取篇章级、段落级和句子级三个尺度的文本单元，特征词包含内容词和情感词两大类，均作为情感信息的载体。情感生成模块包含认知评价、心理感受、生理反应和发音描述四步，各步骤生成结果依次向后堆叠作为后续步骤的部分已知信息，以此方式刻画各步之间的相互影响，同时不同尺度以及不同分析单元间也存在交互作用，也采用类似方式进行刻画。接下来将对融合多尺度相互影响的预测模型进行测试。

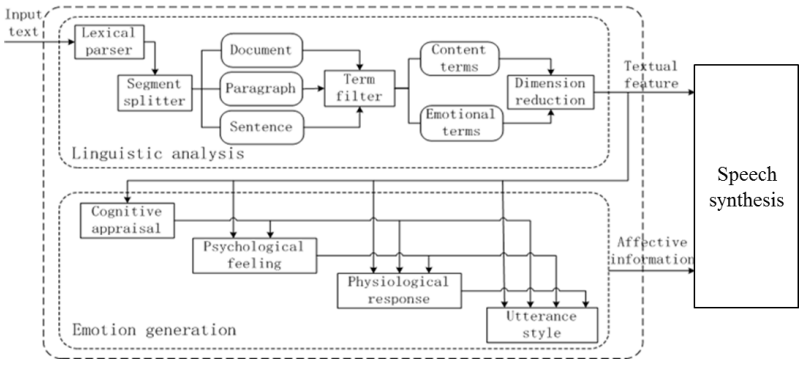


图 5-2 支持多尺度特征的情感语音合成系统

Fig. 5-2 The expressive speech synthesis system supporting multi-scale emotion prediction

5.5 实验与分析

本节将在第 4 章考虑成分间相互影响的基础上加入不同文本分析单元间的影响，即上下文环境的影响，包括不同尺度文本单元间的影响和同一尺度下不同分析

单元间的影响两种形式。本节将分别测试这两种形式的上下文信息对于句子级情感分析的作用，最后将给出融合所有不同形式的影响的综合结果。

评价指标仍以均方根误差 RMSE 为主，因为该预测任务仍是多维预测任务。除此之外，由于现有的情感分析工作多被当作分类任务，因此常使用的评价指标是可以综合衡量分类结果好坏的精准率（Precision）、召回率（Recall）和 F1 值，其中，精准率表示分给类别 C 中所有样本正确分类的比例，召回率表示分给类别 C 中正确分类的样本占应该分到类别 C 的比例，F1 值表示二者的调和均值，即 $F1=2\times Precision\times Recall/(Precision+Recall)$ 。为使我们的工作与他人具有可比性，我们在最后也将给出使用这三个指标计算的结果。由于我们使用的情感模型描述内容众多且具有多维多强度的划分（分布式表示），因此难以做到对每个标注类别（每个维度上的每种强度）计算精准率、召回率和 F1 值，我们以发音描述为最终结果，给出其所有维度和所有强度的平均 F1 值、平均精准率和平均召回率。

测试所用数据集为第 3 章建立的情感语料库，共 8600 篇新闻播音稿，可分割为 27431 段和 64568 句，其中 150 篇为有标数据，可以分割为 461 段和 661 句。

主题模型和深度网络训练参数按第 4 章测试的结果进行设置，主题数 $T=210$，超参数 $\alpha=50/T$，$\beta=0.01$。本章采用 IVDSN 网络，隐含层节点数均设置为 10，微调迭代次数和 RBM 循环周期也均设为 10。

网络仍通过首先利用全部样本（有标和无标）进行无监督预训练，之后利用有标数据进行有监督微调的方法进行训练。由于本章进行句子级情感预测，样本数相对篇章级有所增加，因此且采用十折交叉验证的方法进行网络性能验证；同时由于神经网络预测结果的随机性，运行 10 次十折交叉验证，取其平均结果作为最终预测结果。

5.5.1 不同尺度文本单元的影响

本实验用于测试加入大尺度文本单元的情感信息对于小尺度单元情感预测的影响，以及该方式作为提供上下文环境的方法的有效性。以不加入其他尺度情感特征仅考虑本尺度本单元特征的预测结果作为基准（表 5-1 中斜体行），分别预测了段落级和句子级的四种情感成分。本章实验在第 4 章提出的言语情感预测模型基础上进行，因此本尺度特征在文本特征基础上还要考虑其他情感成分的影响，即各其他成分按生成顺序依次堆叠作为后面成分的部分输入。跨尺度特征采用 5.2 节所叙述的大尺度单元相同情感成分间的堆叠。

表 5-1 加入不同尺度文本单元影响的段落级和句子级情感预测结果(RMSE)

Table 5-1 The RMSEs of the four components at paragraph level and sentence level with or without higher-level features stacked

	特征组合	认知评价	心理感受	生理反应	发音描述
段预测	*para1*	*2.40*	*1.64*	*0.55*	*2.25*
	para1+doc	**1.93**	**1.55**	**0.55**	**2.04**
句预测	*sen*	*2.38*	*1.60*	*0.58*	*2.21*
	sen + doc	1.93	1.48	0.61	2.06
	sen + para2	1.97	1.54	0.60	2.20
	sen + doc + para2	**1.91**	**1.48**	**0.57**	**2.05**

注：“sen”表示句子级本尺度特征（文本+其他情感成分）；“doc”表示篇章级跨尺度特征（相同情感成分）；“para1”表示段落级本尺度特征（文本+其他情感成分）；“para2”表示段落级跨尺度特征（相同情感成分）。

由表 5-1 可以看出，加入大尺度文本的情感特征，对于段落级和句子级情感预测的结果均有明显提升。其中，在段落级，除生理反应外（*p*-value=0.12），其他三种成分的 *p*-value 均小于 0.01，可见结果变化显著。在句子级，篇章级特征相对段落级特征对句子级情感预测的影响更明显，但是二者组合的方式误差降低幅度最大。同样针对基准结果做差异显著性分析，除只加入段落情感特征预测发音描述的结果变化不显著外（*p*-value=0.20），其他结果均发生显著变化（*p*-value<0.01）。生理反应的预测误差在分别加入篇章级特征和段落级特征时略微有所上升，但均处于较低的水平。总之，表 5-1 中结果表明加入其他尺度情感特征有助于本尺度文本单元的情感预测；同时，以大尺度特征为小尺度单元情感预测提供上下文参考的方式可行且有效。

5.5.2 同一尺度不同分析单元的影响

不同文本分析单元的影响除上面所说的不同尺度的影响外，还包括同一尺度内不同分析单元的影响，这里指当前所分析语句前一句情感分布的影响，篇章或段落级更大尺度的情感信息为句子级情感预测提供较为全局的参考，而前一句的情感信息为当前句提供更加细致的参考。由于仅加入篇章级特征和同时加入段落级和篇章级特征的预测结果相差不大（表 5-1 所示），为减少特征堆叠造成的网络扩张，这里仅采用篇章级情感特征作为大尺度特征。前一句情感特征与大尺度特征一样，均采用相同情感成分的堆叠。

表 5-2 加入同一尺度不同文本单元影响的句子级情感预测结果(RMSE)

Table 5-2 The RMSEs of the four components at sentence level with previous features at the same level stacked

特征组合	认知评价	心理感受	生理反应	发音描述
sen	*2.38*	*1.60*	*0.58*	*2.21*
sen + pre	1.98	1.55	0.59	2.10
sen + doc + pre	**1.88**	**1.45**	**0.58**	**2.01**

注：“pre”表示前一句情感特征（相同情感成分）。

表 5-2 给出加入前一句情感特征对当前句情感预测的影响，以及同时加入篇章级和前一句情感特征作为上下文参考的预测结果。同样以不加入其他单元情感特征仅考虑本尺度本单元特征的预测结果作为基准（表 5-2 中斜体行），可以看出，加入其他文本单元影响可以提升当前分析单元的预测效果，且提升效果显著（p-value<0.05）。同表 5-1 对比可以看出，前一句情感特征的影响不如篇章级特征明显，即采用篇章级特征时对预测误差的降低幅度最大，说明篇章级特征作为上下文参考相对其他两种特征（段落级特征和前一句特征）更有效。同时融合两种上下文特征，即同时加入篇章级情感特征和前一句情感特征对于预测结果的提升效果最明显（除生理反应外， p-value 均小于 0.01）。生理反应成分因为基准值已经处于较低的水平，因此提升效果不显著。总之，表 5-2 中结果表明加入前一文本单元情感特征有助于当前分析单元的情感预测，以该方式提供上下文参考也是一种有效的方式；此外，同时融合大尺度单元特征和前一文本单元特征的方式预测效果最佳。

5.5.3 综合结果比较

综合上述实验所得结果，我们给出以精准率、召回率和 F1 值为评价指标的融合所有不同形式的影响的最终结果。因为当前采用的情感描述体系的复杂性，仅给出发音描述标注类别（每个维度上的每种强度）的平均 F1 值、平均精准率和平均召回率。以既不考虑情感成分间相互影响也不考虑上下文影响的预测结果作为基准（表 5-3 中“Baseline”行）。本文所确定的最终的情感预测模型既包含了言语情感各成分间的相互影响，又融合了大尺度单元以及同一尺度其他文本单元的影响，由于这些影响采用堆叠的方式实现，表 5-3 中将这一结果标为“Stacking”。同一单元内不同情感成分按生成顺序依次堆叠作为后面成分的部分输入，大尺度单元的影响采用篇章级相同情感成分的堆叠，同样，同一尺度其他文本单元的影响采用前一句相同情感成分的堆叠。表 5-3 结果显示，本文所提出的情感预测的方法使发音描述的召回率、精准率和 F1 值均有明显提升，其中，精准率提升了 10.3%，F1 值

提高了 22.8%，召回率提高了 31.8%。

表 5-3 句子级发音描述的最终预测结果（平均召回率、平均精准率和平均 F1 值）。其中，“Baseline”表示既不考虑情感成分间相互影响也不考虑上下文影响的预测结果，“Stacking”表示同时考虑两种影响的结果，即本文所提方法。

Table 5-3 The mean recall, precision and F-value of the predictions of utterance manner at sentence level, in which “Baseline” represents the method without any other components or contextual features stacked, and “Stacking” refers to the proposed method with both other components and contextual features stacked.

	Recall	Precision	F1-value
Baseline	0.47	0.58	0.48
Stacking	**0.61**	**0.64**	**0.59**

此外，表 5-4 给出两篇新闻稿的人工标注情感与预测结果的对比（篇章级），以便观察分析所提模型的预测效果。从表中可以看出，模型预测结果基本与人工标注数值处于同一强度，表明预测模型可以提供堪用的情感预测信息。另外，两篇文稿虽然均属于人物悼念的类型，但发音方式截然不同，如果仅采用离散类别标注情感，则无法得到可供参考的更多细节，我们通过从情感生成的角度，逐一对该过程中涉及到的变化与反应进行评估，最终推得与文本内容所表达的情感相一致的发音方式预判，进而用以指导后续的语音信号的调整与变化。

5.6 本章小结

本章在第 4 章提出的融合不同情感成分间相互影响的言语情感预测模型基础上，增加了来自不同尺度和不同文本单元间特征的影响，提出了可以融合不同层级影响因素的多尺度情感预测模型。模型依然基于中间层部分可见的深度堆叠网络 VDSN 搭建，大尺度单元以及前一文本单元的情感特征作为当前分析单元的已知信息，为其提供不同颗粒度的上下文参考。实验表明：

（1）加入其他尺度情感特征有助于本尺度文本单元的情感预测，以大尺度特征为小尺度单元情感预测提供上下文参考的方式可行且有效；

（2）加入前一文本单元情感特征有助于当前分析单元的情感预测，以该方式提供上下文参考也是一种有效的方式；

（3）同时融合大尺度单元特征和前一文本单元特征的方式预测效果最佳；

（4）最终，本文所提出的情感预测模型（既包含言语情感各成分间的相互影响，又融合了大尺度单元以及同一尺度其他文本单元的影响）使预测目标的召回率、

精准率和 F1 值分别提升了 31.8%、10.3%和 22.8%。

表 5-4 人工情感标注与预测结果样例对比

Table 5-4 Comparison between artificial emotion labeling and predicting results

输入文本	人工标注	预测结果
我国著名电影艺术家，中国共产党党员孙道临的遗体昨天在上海龙华殡仪馆火化。孙道临因突发心脏病抢救无效，二零零七年十二月二十八号在上海逝世，享年八十六岁。孙道临逝世后，胡锦涛、江泽民、吴邦国、温家宝、贾庆林、曾庆红、李长春、习近平、李克强、贺国强、周永康等通过不同方式对孙道临的逝世表示沉痛哀悼，并向其家属表示深切慰问。昨天上午，近万人在贝多芬第七交响曲的旋律中送别孙道临。孙道临原名孙以亮，一九二一年生于北京，一九四三年加入中国旅行剧团，开始艺术生涯，六十多年来，他追求光明，追求进步，把个人命运、艺术追求和民族社会、国家的命运结合在一起，为繁荣中国电影事业贡献了毕生的智慧和力量。二零零五年，在纪念中国电影诞生一百周年大会上被授予“对国家有突出贡献电影艺术家”称号。	认知： 否定-肯定 1 冷漠-热情 0 非正式-正式 1 柔和-强硬 0 心理（主要成分）： 尊敬 1 悲哀 1 生理： 激活度 1 控制度 1 发音： 暗-明 -1 瘪-满 1 慢-快 -1 低-高 -1 平-曲（语调）-1 稳-变（节奏）-1	认知： 否定-肯定 0.69 冷漠-热情 0.72 非正式-正式 0.99 柔和-强硬 -0.38 心理（主要成分）： 尊敬 0.68 悲哀 1.89 生理： 激活度 1.26 控制度 1.01 发音： 暗-明 -0.89 瘪-满 1.33 慢-快 -1.27 低-高 -1.04 平-曲（语调） -0.62 稳-变（节奏）-1.40
大年初三一大早，大方县百那彝族乡拢住村农民党员武德顺接到挚友赵高勇的电话：“大方县东部海拔较高地区电力线路受损情况严重，请来帮帮忙。”“就来。”武德顺放下电话急匆匆地走出家门。二月十五日是吴德顺参加电力线路突击抢修的第七天。上午八点十分，武德顺带着两名抢险突击队员高建华、赵士远到达海拔两千米的山头。分配工作后，武德顺系上保险绳敏捷地爬上十米高杆，谁想意外瞬间发生，武德顺登上了五号杆，突然电杆齐根断倒。正在整理器具的高建华抬头，眼睁睁地看着电杆带着师傅朝他倒来。赵士远和高建华连奔带扑的跑到武德顺身边，武德顺已经不能说话。三年前的一个赶场天，武德顺和爱人赵施梅收养了一名弃婴，就是他们现在的宝贝女儿星月，守林下地、牧马串寨，武德顺总要把星月带在身边。得知武德顺牺牲的消息，赵施梅将脸贴在女儿脸上，任凭泪水留下。星月趴在妈妈	认知： 否定-肯定 1 冷漠-热情 1 非正式-正式 -2 柔和-强硬 -1 心理（主要成分）： 尊敬 1 怜悯 1 悲哀 1 生理： 激活度 1 控制度 1 发音： 暗-明 1 瘪-满 2 慢-快 1 低-高 1 平-曲（语调） 2 稳-变（节奏） 2	认知： 否定-肯定 1.04 冷漠-热情 1.16 非正式-正式 -2.48 柔和-强硬 -1.41 心理（主要成分）： 尊敬 1.59 怜悯 0.33 悲哀 0.51 生理： 激活度 1.44 控制度 1.13 发音： 暗-明 1.41 瘪-满 1.50 慢-快 0.92 低-高 0.97 平-曲（语调） 2.15 稳-变（节奏） 2.28

的怀里一边哭又一边地喊着爸爸、爸爸。		

6 总结与展望

6.1 全文工作总结

论文通过对~~本文模拟~~有声语言的创作过程的分析，从情感语音生成角度探究了与言语中情感的发生及衍化过程，形成了从多个视角刻画情感状态的描述模型，解决了细腻、复杂情感的精~~准~~确刻画问题，为研究情感语音~~信号~~的变化提供了更详尽的信息参照；在描述模型基础上，基于深层神经网络建立了从文本到情感的预测模型，解决了多尺度特征处理以及动态衍化过程的建模问题，实现了基于文本内容的情感状态的自动预测。完成的主要工作有：

（1）以心理学已有的大量情感理论和情感模型为参考，结合朗读学、播音学、以及语音学中关于情感的研究，将文语转换过程与朗读者或播音员将文稿转换成有声语言的过程类比，对基于文本的言语情感生成过程进行分析，将其归纳为文本分析、认知评价、心理感受、生理反应和发音调整等步骤，各步骤之间存在直接或间接的相互影响，前面步骤的结果会影响后续步骤的反应，后续步骤的反应又会反馈回去进一步影响前面成分的变化。

（2）基于言语情感生成过程分析，提出了情感的多视角描述体系，分别从认知评价、心理感受、生理反应和发音描述四种不同视角刻画言语情感的不同侧面，各视角互为补充，共同组成言语情感的分布式表达模型。各视角采用维度表示、离散类别表示和层级表示相结合的方式进行具体刻画，分别形成一个多维支撑或多层结构的超平面，各超平面根据影响关系构成体现言语情感内部复杂结构的多层特征空间。发音描述的引入形成连接情感特征与声学特征变化的接口。

（3）以汉语朗读语音为研究对象，基于新闻播音文稿，采用表演语音的形式构建了新闻言语情感数据库，为情感预测模型的训练以及将来的情感声学特征分析提供数据基础，同时通过言语情感标注的实施验证了情感描述体系的可操作性。数据库规模为：共包含 29109 篇新闻播音稿，其中 8600 篇长度适中的语篇作为言语情感预测模型的训练语料，600 篇既包含文本形式又进行了语音形式的转录，对其中 150 进行了基于多视角言语情感描述体系的情感标注。

（4）基于言语情感~~生成过程和~~多视角描述体系，利用深度神经网络建立了从文本到情感的预测模型。模型综合考虑言语情感生成过程中各成分之间的相互影响以及上下文环境的影响，支持动态衍化过程的刻画以及不同尺度特征的融合。具体而言，以深度堆叠网络为基础，将情感特征间的衍化关系以及多尺度特征的层级

关系应用于对深层网络结构的约束和引导。各尺度内部均遵从言语情感产生过程对于情感特征间衍化关系的设定，以发音描述作为最终目标，其他成分作为其子目标，子目标依次作为后续预测目标的已知信息，构成级联型生成关系；不同尺度间采用由粗到细的生成顺序，大尺度单元的预测结果为小尺度单元提供上下文参考，不同尺度的情感特征构成层级型结构关系。这些已知结构关系作为先验引入深度神经网络，形成中间结构可见或部分可见的深度神经网络 VDSN，根据可见位置的不同，又分为输入层部分可见（IVDSN）与隐含层部分可见（HVDSN）两种网络，二者均可实现诸如级联型或层级型结构的建模。

通过预测模型性能测试与验证实验，得到以下结论：

a）堆叠结构的网络模型支持动态衍化过程刻画以及多尺度特征融合，在模型中加入情感衍化过程中的相互影响以及不同尺度特征间相互影响，可以使模型的召回率、精准率和 F1 值相对提高 31.8%、10.3%和 22.8%；

b）深度神经网络中间层可见化的建模策略，可有效地将先验知识融入网络结构，进而提升网络的性能；特征维度均衡调整和样本均衡调整等优化措施，可以一定程度上解决网络学习过程中不同类型特征间的融合问题以及样本稀疏问题。

综上所述，本文的研究从认知科学角度归纳形成了言语情感描述体系，并从技术层面解决了情感的时序衍化特性和情感特征间相关性的建模问题。在一定程度上解决了当前情感语音研究中存在于言语情感认知与计算建模方面的问题，所利用的建模算法与策略可对相关研究起借鉴作用。

6.1 下一步工作展望

基于论文的研究积累，可以从以下几个方面对下一步工作进行展望：

（1）情感预测：目前论文的情感预测模型是基于文本内容建立，而对于可能影响情感生成及其表现的场景因素、发音人的社会文化背景因素等暂未考虑。论文的研究目标为朗读语音的合成，这些因素相对稳定或属于弱敏感因子。但对于自然对话系统，需要建立具备环境感知能力的情感预测机制，并分析不同场景、不同文化背景下情感在语音乃至其他模态特征的表现形式，形成更智能、友好、和谐的人机交互界面。

（2）情感理论研究：本文从情感语音的生成过程出发，给出言语情感生成过程和描述框架，深化了有关言语中情感的理解和表示。然而，对于情感这一复杂命题的研究与探索还远没有结束。人类情感是一个涉及到多个系统交互作用的连续过程，各种情感之间还存在着相互渗透、相互转化的关系。要实现对于情感的建模，

一方面需要有强有力的理论支撑，目前仍缺乏在心理学和生理学界都得到广泛认可的理论模型；另一方面在建模手段上还存在技术瓶颈，比如，情感既具有实时变化性又具有长期记忆性，对于特征信号检测与处理尺度的把握都造成困难。

（3）情感预测：在预测模型方面，我们采用具有多层非线性映射结构的深度神经网络作为基础网络，通过加入对衍化及层级关系的处理提升了预测效果，已知先验关系的引入相当于对网络结构加入了引导与约束，使网络能更高效地学习到期待的特征，在先验合理的前提下学习结果也更接近于真解。该网络结构可以应用于图像、语音等其他具有结构化特征的信号的处理，尤其对于数据与计算资源有限的情况，先验的引入可以降低训练数据与网络规模的开销，但是关于先验如何获取、先验是否合理等问题在不同的应用领域还需做进一步研究。

（4）情感语音合成：论文研究工作的出发点以及落脚点都是提升情感语音合成系统的合成效果，要做到这一点，还需要与韵律预测模块以及语音合成器模块的有效结合。论文中情感预测的输出是发音方式描述，欲将其转化为对合成语音声学特征参数的调整，进而通过合成器实现情感语音的合成输出，一方面涉及到更为精细有效的声学特征参数选取，另一方面涉及到整体性的发音描述如何投射到多尺度合成单元的实现策略，前者与语音合成器改进紧密相关，后者则有赖于对情感语音进行的更为深入的、针对性语音学探索。

批注 [w4]: 注意术语统一

设置了格式：突出显示

让机器像人一样具备感知并表达情感的能力是一项任重而道远的工作，情感语音合成的研究属于其中的一个方面，如何与其他学科有效结合，进而推动自身以及相关学科发展，是语音研究者亟需面临的挑战之一。

参考文献

[1] 陶建华, 许晓颖. 面向情感的语音合成系统[C]// 第一届中国情感计算及智能交互学术会议论文集. 2003.
[2] 蔡莲红. 情感计算[J]. 中国计算机学会通讯, 2010, (5):17-19.
[3] 李爱军. 面向言语工程的情感语音[N]. 中国社会科学院院报, 2006.
[4] Gu W, Zhang T, Fujisaki H. Prosodic analysis and perception of Mandarin utterances conveying attitudes[C]//Twelfth Annual Conference of the International Speech Communication Association. 2011.
[5] Govind D, Prasanna S R M. Expressive speech synthesis: a review[J]. International Journal of Speech Technology, 2013, 16(2):237-260.
[6] Cahn J E. Generating expression in synthesized speech[D]. Massachusetts Institute of Technology, Dept. of Architecture, 1989.
[7] Murray I R. Simulating emotion in synthetic speech[D]. University of Dundee, 1989.
[8] Arnfield S, Roach P, Setter J, et al. Emotional stress and speech tempo variation[C]//Speech under Stress. 1995.
[9] Gobl C, Chasaide A N. Acoustic characteristics of voice quality[J]. Speech Communication, 1992, 11(4): 481-490.
[10] Moriyama T, Saito H, Ozawa S. Evaluation of the relation between emotional concepts and emotional parameters in speech[J]. Systems & Computers in Japan, 2001, 32(3):56-64.
[11] Picard R W. Affective computing[J]. Igi Global, 1997, 1(1):71-73.
[12] 任蕊. 基于 Fujisaki 模型的情感语音信号分析与合成[D]. 北京交通大学, 2008.
[13] Roach P. Techniques for the phonetic description of emotional speech[C]// Proceedings of the ISCA Workshop on Speech and Emotion. 2000.
[14] Iida A, Campbell N, Higuchi F, et al. A corpus-based speech synthesis system with emotion[J]. Speech Communication, 2003, 40(1): 161-187.
[15] Yamagishi J, Onishi K, Masuko T, et al. Modeling of various speaking styles and emotions for HMM-based speech synthesis[C]//Proc. the 8th Eurpean Conference on Speech Communication and Technology. 2003 (III): 2461-2464.
[16] Kang S, Meng H. Statistical parametric speech synthesis using weighted multi-distribution deep belief network[C]//Proc. Interspeech. 2014: 1959-1963.
[17] Fan Y, Qian Y, Xie F, et al. TTS synthesis with bidirectional LSTM based recurrent neural networks[C]//Proc. Interspeech. 2014: 1964-1968.
[18] Fernandez R, Rendel A, Ramabhadran B, et al. Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks[C]//Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH). 2014.
[19] Yin X, Lei M, Qian Y, et al. Modeling DCT Parameterized F0 Trajectory at Intonation Phrase Level with DNN or Decision Tree[C]//Fifteenth Annual Conference of the International Speech Communication Association. 2014.
[20] Nakashika T, Takiguchi T, Ariki Y. High-order sequence modeling using speaker-dependent recurrent temporal restricted boltzmann machines for voice conversion[C]//Fifteenth Annual Conference of the International Speech Communication Association. 2014.

[21] Xie F L, Qian Y, Fan Y, et al. Sequence error (SE) minimization training of neural network for voice conversion[C]//Proc. Interspeech. 2014: 2283-2287.
[22] Cowie R, Cornelius R R. Describing the emotional states that are expressed in speech[J]. Speech communication, 2003, 40(1): 5-32.
[23] Scherer K R. The dynamic architecture of emotion: Evidence for the component process model[J]. Cognition and emotion, 2009, 23(7): 1307-1351.
[24] Alm C O, Roth D, Sproat R. Emotions from text: machine learning for text-based emotion prediction[C]//Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2005: 579-586.
[25] Wu Z, Meng H M, Yang H, et al. Modeling the expressivity of input text semantics for Chinese text-to-speech synthesis in a spoken dialog system[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2009, 17(8): 1567-1576.
[26] Chao L, Tao J, Yang M, et al. Bayesian Inference based Temporal Modeling for Naturalistic Affective Expression Classification[C]//Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on. IEEE, 2013: 173-178.
[27] Li A, Fang Q, Jia Y, et al. Emotional McGurk Effect? A Cross-Cultural Investigation on Emotion Expression under Vocal and Facial Conflict[M]//Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer Berlin Heidelberg, 2013: 214-226.
[28] 刘烨, 傅小兰, 陶霖密. 基于 PAD 三维空间的情感测量[J]. 中国计算机学会通讯, 2010, 6(5): 9-14.
[29] 赵力, 黄程韦. 实用语音情感识别中的若干关键技术[J]. 数据采集与处理, 2014, 29(2): 157-170.
[30] 张鼎天, 徐明星. 基于调制频谱特征的自动语音情感识别[C]//第十二届全国人机语音通讯学术会议 (NCMMSC'2013) 论文集. 2013.
[31] 邵艳秋, 穗志方, 韩纪庆, 等. 小规模情感数据和大规模中性数据相结合的情感韵律建模研究[J]. 计算机研究与发展, 2015, 44(9): 1624-1631.
[32] Zhao Y, Qin B, Che W, et al. Appraisal expression recognition with syntactic path for sentence sentiment classification[J]. International Journal of Computer Processing of Languages, 2011, 23(01): 21-37.
[33] Ortony A, Clore G L, Foss M A. The referential structure of the affective lexicon[J]. Cognitive science, 1987, 11(3): 341-364.
[34] Cornelius R R. Theoretical approaches to emotion[C]//ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion. 2000.
[35] Ekman P. An argument for basic emotions[J]. Cognition & emotion, 1992, 6(3-4): 169-200.
[36] Cowie R, Douglas-Cowie E, Savvidou* S, et al. 'FEELTRACE': An instrument for recording perceived emotion in real time[C]//ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion. 2000.
[37] Mehrabian A. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament[J]. Current Psychology, 1996, 14(4): 261-292.
[38] Moors A, Ellsworth P C, Scherer K R, et al. Appraisal theories of emotion: State of the art and future development[J]. Emotion Review, 2013, 5(2): 119-124.
[39] Calvo R A, D'Mello S. Affect detection: An interdisciplinary review of models, methods, and

their applications[J]. Affective Computing, IEEE Transactions on, 2010, 1(1): 18-37.
[40] Scherer K R. Psychological models of emotion[J]. The neuropsychology of emotion, 2000, 137(3): 137-162.
[41] Agrawal A, An A. Unsupervised emotion detection from text using semantic and syntactic relations[C]//Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on. IEEE, 2012, 1: 346-353.
[42] Calix R A, Javadpour L, Knapp G M. Detection of affective states from text and speech for real-time human–computer interaction[J]. Human Factors: The Journal of the Human Factors and Ergonomics Society, 2012, 54(4): 530-545.
[43] Calix R A, Mallepudi S A, Chen B, et al. Emotion recognition in text for 3-D facial expression rendering[J]. Multimedia, IEEE Transactions on, 2010, 12(6): 544-551.
[44] Tokuhisa R, Inui K, Matsumoto Y. Emotion classification using massive examples extracted from the web[C]//Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, 2008: 881-888.
[45] Osherenko A. Towards semantic affect sensing in sentences[C]//Proceedings of the AISB 2008 Symposium on Affective Language in Human and Machine. 2008: 41-44.
[46] Liu H, Lieberman H, Selker T. A model of textual affect sensing using real-world knowledge[C]//Proceedings of the 8th international conference on Intelligent user interfaces. ACM, 2003: 125-132.
[47] Bellegarda J R. Emotion analysis using latent affective folding and embedding[C]//Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text. Association for Computational Linguistics, 2010: 1-9.
[48] Trilla T, Alias F. Sentence-Based Sentiment Analysis for Expressive Text-to-Speech[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2013, 21(2): 223-233.
[49] Wan X. Co-training for cross-lingual sentiment classification[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1. Association for Computational Linguistics, 2009: 235-243.
[50] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.
[51] Deng L, Yu D, Platt J. Scalable stacking and learning for building deep architectures[C]//Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012: 2133-2136.
[52] Yoo H J. Deep Convolution Neural Networks in Computer Vision[J]. IEIE Transactions on Smart Processing & Computing, 2015, 4(1): 35-43.
[53] Oquab M, Bottou L, Laptev I, et al. Learning and transferring mid-level image representations using convolutional neural networks[C]//Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014: 1717-1724.
[54] Zhang C, Zhang Z. Improving multiview face detection with multi-task deep convolutional neural networks[C]//Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on. IEEE, 2014: 1036-1041.
[55] Sainath T N, Kingsbury B, Saon G, et al. Deep Convolutional Neural Networks for Large-scale Speech Tasks[J]. Neural Networks, 2014.
[56] Deng L, Hinton G, Kingsbury B. New types of deep neural network learning for speech

recognition and related applications: An overview[C]//Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 8599-8603.
[57] Bengio S, Heigold G. Word embeddings for speech recognition[C]//Proceedings of the 15th Conference of the International Speech Communication Association, Interspeech. 2014.
[58] Le Q, Mikilov T. Distributed Representations of Sentences and Documents[C]//The International Conference on Machine Learning, Beijing, China. 2014
[59] Srivastava N, Salakhutdinov R R, Hinton G E. Modeling Documents with a Deep Boltzmann Machine[J]. Uncertainty in Artificial Intelligence, 2013
[60] 孙志军, 薛磊, 许阳明, 等. 深度学习研究综述[J]. 计算机应用研究, 2012, 29(8): 2806-2810.
[61] Stone P J, Bales R F, Namenwirth J Z, et al. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information[J]. Behavioral Science, 1962, 7(4): 484-498.
[62] Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.
[63] Esuli A, Sebastiani F. Sentiwordnet: A publicly available lexical resource for opinion mining[C]//Proceedings of LREC. 2006, 6: 417-422.
[64] Strapparava C, Valitutti A. WordNet Affect: an Affective Extension of WordNet[C]//LREC. 2004, 4: 1083-1086.
[65] Balahur A, Hermida J M, Montoyo A, et al. Emotinet: A knowledge base for emotion detection in text built on the appraisal theories[M]//Natural Language Processing and Information Systems. Springer Berlin Heidelberg, 2011: 27-39.
[66] Dong Z, Dong Q. HowNet-a hybrid language and knowledge resource[C]//Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003 International Conference on. IEEE, 2003: 820-824.
[67] Coon D. 心理学导论: 思想与行为的认识之路[M]. 中国轻工业出版社, 2004.
[68] 陈俊杰. 图像情感语义分析技术[M]. 电子工业出版社, 2011.
[69] Schachter S. The interaction of cognitive and physiological determinants of emotional state[J]. Advances in experimental social psychology, 1964, 1: 49-80.
[70] Arnold M B. Emotion and personality[M]. New York: Columbia University Press, 1960.
[71] 蔡莲红. 情感计算[J]. 中国计算机学会通讯, 2010, (5):17-19.
[72] James W. What is an emotion?[J]. Mind, 1884 (34): 188-205.
[73] Darwin C. The expressions of the emotions in man and animals[M]. London: John Murray, 1872.
[74] 王国江, 王志良, 杨国亮, 等. 人工情感研究综述[J]. 计算机应用研究, 2006, 23(11): 7-11.
[75] Izard C E. Basic emotions, relations among emotions, and emotion-cognition relations[J]. 1992.
[76] 孟昭兰. 情绪心理学[M]. 北京大学出版社, 2005.
[77] 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): 712-717.
[78] Cornelius R R. The science of emotion: Research and tradition in the psychology of emotions[M]. Prentice-Hall, Inc, 1996.
[79] Shaver P, Schwartz J, Kirson D, et al. Emotion knowledge: further exploration of a prototype approach[J]. Journal of personality and social psychology, 1987, 52(6): 1061-1086.
[80] Plutchik R. Emotion: A psychoevolutionary synthesis[M]. Harpercollins College Division, 1980.

[81] Ortony A, Turner T J. What's basic about basic emotions?[J]. Psychological review, 1990, 97(3): 315-331.
[82] Laros F J M, Steenkamp J B E M. Emotions in consumer behavior: a hierarchical approach[J]. Journal of business Research, 2005, 58(10): 1437-1445.
[83] Wundt W. Outlines of psychology[M]. Springer US, 1980.
[84] Schlosberg H. Three dimensions of emotion[J]. Psychological review, 1954, 61(2): 81-88.
[85] Izard C E. The psychology of emotions[M]. Springer Science & Business Media, 1991.
[86] Russell J A. A circumplex model of affect[J]. Journal of personality and social psychology, 1980, 39(6): 1161-1178.
[87] Ortony A, Clore G L, Collins A. The cognitive structure of emotions[M]. Cambridge, UK: Cambridge University Press, 1988.
[88] Fox N A. If it's not left, it's right: Electroencephalograph asymmetry and the development of emotion[J]. American psychologist, 1991, 46(8): 863-872.
[89] 张颂. 朗读学. 中国传媒大学出版社, 2007
[90] 张颂. 中国播音学, 中国传媒大学出版社, 2003
[91] Fujisaki H, Hirose K. Analysis and perception of intonation expressing paralinguistic information in spoken Japanese[C]//ESCA Workshop on Prosody. 1993.
[92] Gu W, Zhang T, Fujisaki H. Prosodic analysis and perception of Mandarin utterances conveying attitudes[C]//Twelfth Annual Conference of the International Speech Communication Association. 2011.
[93] 李轶. 论情感类义位的范围及其分类[J]. 华夏文化论坛, 2010: 020.
[94] 许小颖, 陶建华. 汉语情感系统中情感划分的研究[C]//第一届中国情感计算及智能交互学术会议论文集 2003: 199-205.
[95] 钱线, 黄萱菁, 吴立德. 初始化 K-means 的谱方法[J]. 自动化学报, 2007, 33(4): 342-346.
[96] 孟子厚. 音质主观评价的实验心理学方法[M]. 北京:国防工业出版社, 2008.
[97] Schuller B, Weninger F, Dorfner J. Multi-Modal Non-Prototypical Music Mood Analysis in Continuous Space: Reliability and Performances[C]//ISMIR. 2011: 759-764.
[98] Lee C Y, Xie S, Gallagher P, et al. Deeply-supervised nets[J]. arXiv preprint arXiv:1409.5185, 2014.
[99] Weston J, Ratle F, Mobahi H, et al. Deep learning via semi-supervised embedding[M]//Neural Networks: Tricks of the Trade. Springer Berlin Heidelberg, 2012: 639-655.
[100] Deng L, Yu D, Platt J. Scalable stacking and learning for building deep architectures[C]//Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012: 2133-2136.
[101] Smolensky P. Information processing in dynamical systems: foundations of harmony theory[C]//Parallel distributed processing: explorations in the microstructure of cognition, vol. 1. MIT Press, 1986: 194-281.
[102] Hinton G E, Sejnowski T J. Learning and relearning in Boltzmann machines[J]. MIT Press, Cambridge, Mass, 1986, 1: 282-317.
[103] Salakhutdinov R, Hinton G E. Deep boltzmann machines[C]//International Conference on Artificial Intelligence and Statistics. 2009: 448-455.
[104] Tur G, Deng L, Hakkani-Tur D, et al. Towards deeper understanding: deep convex networks for semantic utterance classification[C]//Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012: 5045-5048.

[105] Hinton G, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527-1554.

[106] Hinton G E. A practical guide to training restricted boltzmann machines[M]//Neural Networks: Tricks of the Trade. Springer Berlin Heidelberg, 2012: 599-619.

[107] Yu D, Deng L. Accelerated Parallelizable Neural Network Learning Algorithm for Speech Recognition[C]//INTERSPEECH. 2011: 2281-2284.

[108] Zhang H. NIPIR/ICTCLAS(2014), Chinese word segmentation system on: http://ictclas.nlpir.org/newsdownloads?DocId=389

[109] Deerwester S C, Dumais S T, Landauer T K, et al. Indexing by latent semantic analysis[J]. JAsIs, 1990, 41(6): 391-407.

[110] Hofmann T. Probabilistic latent semantic indexing[C]//Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999: 50-57.

[111] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. the Journal of machine Learning research, 2003, 3: 993-1022.

[112] Griffiths T. Gibbs sampling in the generative model of Latent Dirichlet Allocation[J]. Standford University, 2002.

[113] Hearst M A. TextTiling: Segmenting text into multi-paragraph subtopic passages[J]. Computational linguistics, 1997, 23(1): 33-64.

[114] Riedl M, Biemann C. Text segmentation with topic models[J]. Journal for Language Technology and Computational Linguistics, 2012, 27(1): 47-69.

作者简历及攻读博士学位期间取得的研究成果

一、作者简历

高莹莹，女，汉族，1987 年 6 月出生，河北保定人。2005 年 9 月至 2009 年 7 月就读于北京交通大学计算机与信息技术学院生物医学工程专业，获工学学士学位。2010 年至今以硕博连读形式攻读北京交通大学信息科学研究所信号与信息处理专业博士学位，研究方向为情感语音合成与情感计算。

二、发表论文

[1] Gao Yingying, Zhu Weibin. Detecting affective states from text based on a multi-component emotion model[J]. Computer Speech and Language. 2016(36): 42-57 (SCI，IF:1.753，An3)

[2] 高莹莹，朱维彬. 深层神经网络中间层可见化建模[J]. 自动化学报，2015, 41(9): 1627-1637 (EI，An5)

[3] Gao Yingying, Zhu Wweibin. How to describe speech emotion more completely - An investigation on Chinese broadcast news speech[C]. 8th International Symposium on Chinese Spoken Language Processing (ISCSLP), IEEE, 2012:450-453. (EI)

[4] 高莹莹，朱维彬. 面向情感语音合成的言语情感描述与预测[C]. 第十三届全国人机语音通讯学术会议（NCMMSC2015），天津，2015

[5] 高莹莹，朱维彬. 言语情感描述体系的试验性研究[J]. 中国语音学报，第四辑，2013

[6] 高莹莹， 朱维彬. 汉语朗读语音中言语情感产生机制与计算模型研究[C]. 第十二届全国人机语音通讯学术会议（NCMMSC2013），贵阳，2013

[7] 高莹莹，朱维彬. 关于新闻语料语气标注的初步研究[C]. 第十届中国语音学学术会议论文集（PCC2012），上海，2012

[8] 高莹莹，朱维彬. NAQ 与韵律特征的关系初探[C]. 第十一届全国人机语音通讯学术会议论文集（NCMMSC2011），西安，2011

[9] 高莹莹，朱维彬. 基于新闻言语数据库的语气标注及其韵律特征分析[C]. 第九届中国语音学学术会议（PCC2010），天津，2010

三、参与科研项目

[1] 863 计划面上项目(2007AA01Z198) “面向汉语语音合成的言语语义计算模型研究”

[2] 863 计划重点项目子课题(2006AA010104) “语义分析研究与合成系统实现”

[3] 横向课题(K10L00030) “习惯传媒语音识别引擎系统软件开发”

[4] 国家语言资源监测与研究中心项目(YZYS10-10) “新闻播报中的基调置标”

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京交通大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：　　签字日期：　　年　　月　　日

学位论文数据集

表 1.1： 数据集页

关键词*	密级*	中图分类号	UDC	论文资助
情感语音合成；情感生成；情感描述；情感预测；深度神经网络；中间层可见化；	公开			
学位授予单位名称*		学位授予单位代码*	学位类别*	学位级别*
北京交通大学		10004	工学	博士
论文题名*		并列题名		论文语种*
面向情感语音合成的言语情感建模研究				汉语
作者姓名*	高莹莹		学号*	10112060
培养单位名称*		培养单位代码*	培养单位地址	邮编
北京交通大学		10004	北京市海淀区西直门外上园村 3 号	100044
学科专业*		研究方向*	学制*	学位授予年*
信号与信息处理		情感语音合成、情感计算	5	2016
论文提交日期*				
导师姓名*			职称*	
评阅人	答辩委员会主席*		答辩委员会成员	
电子版论文提交格式 文本（ ） 图像（ ） 视频（ ） 音频（ ） 多媒体（ ） 其他（ ） 推荐格式：application/msword；application/pdf				
电子版论文出版（发布）者		电子版论文出版（发布）地		权限声明
论文总页数*	115			
共 33 项，其中带*为必填数据，为 21 项。				