语音情感识别中特征空间选择的研究

（申请清华大学工程硕士学位论文）

培 养 单 位: 计 算 机 科 学 与 技 术 系
学　　科: 计 算 机 技 术
研　究　生: 马　习
指 导 教 师: 吴　志　勇　　副　研　究　员

二〇一八年三月

Feature Space Selection in Speech Emotion Recognition

Thesis Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the professional degree of

Master of Engineering

by

Ma Xi

(Computer Technology)

Thesis Supervisor : Associate Professor Wu Zhiyong

March, 2018

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

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

本人保证遵守上述规定。

（保密的论文在解密后应遵守此规定）

作者签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘 要

在人机交互系统逐渐变得普遍的今天，只是理解语音中的语言学信息已经不足以满足所有需求，何提取语音中的情感信息在许多的应用场景中也变得越来越重要。传统的语音情感识别可以分为情感相的关声学特征的提取和情感分类模型的构建两部分，原始语音通常会先被映射到情感信息相关的声学特征，然后采用各种分类模型将将特征向量映射到对应的情感类别。近年来，随着深度学习方法的发展与普及，深度神经网络开始越来越多被应用到这一领域，并且取得了不错的效果。此外，特征提取和情感分类两个部分也开始被整合到一起，通过深度神经网络将可以构建从原始语音直接到情感类别的端到端的识别系统。但是如何引入情感相关的心理学信息以及处理长度变化的语音，并没有被现有的研究广泛的关注。

针对这两个问题，本文将分别从传统的语音情感识别方法和端到端的深度学习方法入手，设计对应的方法来提升系统的识别率。主要的研究工作和和贡献如下：

一、提出一种基于情感对的语音情感识别框架，并在最后的决策融合过程中引入心理学的情感空间模型，从而提升了系统的识别率。传统的语音情感识别系统通常为所有的情感选取相同的声学特征来完成最后的情感分类，但实验证明不同的情感和不同的声学特征的相关性并不同。针对这一问题，我们将分别为不同的情感对选取不同的特征集合，将原先的多分类问题转变为多个二分类问题，并在最后的决策融合过程中通过贝叶斯分类器引入情感空间的信息。在公开的情感语音数据集 IEMOCAP 上，我们方法取得了比传统的识别框架更好的准确率。

二、设计了一种能够处理变长语音段的神经网络结构，消除了语音分段带来的中性语音和情感语音的混淆，从而提升了系统的识别率。在使用深度神经网络实现端到端的语音情感识别系统时，由于卷积神经网络和循环神经网络很难处理变长的输入，通常会把变长的语音句子切分成多段等长的语音段，然后将所有语音段都标记为对应句子的情感标签, 但这样会导致部分中性语音段被标记为有情感。针对于这一问题，我们采用补齐和掩码的方式来处理神经网络中变长的输入序列，避免了错误标注带来了的效果变差。相对于切分等长语音段的方法，我们直接输入整个变长语音的方法可以在相同的数据集上取得更好的准确率。

关键词：语音情感识别；情感对；情感空间模型；变长语音段；深度神经网络

Abstract

An abstract of a dissertation is a summary and extraction of research work and contributions. Included in an abstract should be description of research topic and research objective, brief introduction to methodology and research process, and summarization of conclusion and contributions of the research. An abstract should be characterized by independence and clarity and carry identical information with the dissertation. It should be such that the general idea and major contributions of the dissertation are conveyed without reading the dissertation.

An abstract should be concise and to the point. It is a misunderstanding to make an abstract an outline of the dissertation and words “the first chapter”, “the second chapter” and the like should be avoided in the abstract.

Key words are terms used in a dissertation for indexing, reflecting core information of the dissertation. An abstract may contain a maximum of 5 key words, with semi-colons used in between to separate one another.

Key words: TEX; LATEX; CJK; template; thesis

目 录

第 1 章　绪论

1.1　研究背景和意义

语音作为人与人之间交流最为自然的一种媒介，在我们的日常生活中起着至关重要的作用，这也使得语音一直被许多研究者认为是最有效的人机交互方式。在最近的二十年间，语音识别技术已经取得飞速的发展，它的目的是将人们说出的语音转换成对应的词序列。尽管语音识别技术已经被广泛的应用，但我们和真正的自然人机交互仍然有很大的距离，因为机器仍然无法理解说话人的情感状态，而语音通常会包含说话人当时的情感状态信息。这使得语音情感识别技术开始受到重视，它的目的就是从语音中抽取说话人的情感状态。

语音情感识别在许多需要自然人机交互的系统特别有用，例如网页影评和计算机辅助教育这些需要用户反馈的应用场景。在车载系统中，情感识别可以通过语音检测驾驶员当前的精神状态，并在合适的时候做出提醒从而保证车辆安全。语音情感识别还可以作为心理咨询师提供诊断的辅助工具。在同声传译系统中，语音情感识别也可以通过检测说话人情感状来调整翻译内容。对于语音识别系统来说，加入情感信息同样可以提高识别准确率。还有在电话客服系统中，客户由于长时间的等待或者紧迫的需求，声音会变的烦躁和愤怒。在这种情况下，情感识别系统就可以通过检测语音中愤怒的程度来为这些客户安排服务优先级。此外，在聊天机器人中，语音情感识别系统可以通过检测对方的情感状态来做出不同的应答。这种功能同样也可以运用在玩具中，当孩子不高心的时候，玩具就可以通过一些方式安抚他们。在电子游戏中，尤其是在一些有语音交互的游戏中，情感识别可以通过模拟更加真实的交互场景来提升游戏体验。因此，语音情感识别在近几年来收到越来越多的关注。无论是在前沿模型算法的研究，还是在相关产品的落地，很多研究机构和公司都在这一领域投入大量的精力。

语音情感识别也存在许多挑战性的问题。首先，很难确定哪些声学特征与情感信息最相关。一些在语音识别领域公认的特征，例如梅尔频率倒谱系数（Mel-Frequency Cepstral Coefficients, MFCC），在语音情感识别上并没有取得好的效果，其次，一段语音中有时会包含多种情感，如何找到每种情感对应的语音边界是个很有挑战性的问题。此外，由于说话人的文化背景不同，也会导致表达情感的方式不太相同。例如有的人愤怒时会说话很大声，而有的人却声音很低沉。语言也是影响情感表达的一个主要问题。还有就是情感并没有公认的规范定义，只能通过人们的主观感觉来区分，但这会导致有许多的情感类别。许多的研究者都倾向

于一种“调色板”理论，就是只定义几种基本的情感，然后其他的情感都是由这些基本情感以不同的比例调和而成，这些基本情感被称为原型情感。情感语音的数据相对来说也很难获取，这会导致许多复杂的模型无法得到充分训练。

尽管目前还没有非常成熟的语音情感识别的产品，但许多公司已经在极力推动这一领域的产品落地。Facebook 早在 2012 年就开始进行对用户的情绪检测实验，期望能够通过情感信息来优化他们的推荐系统。微软小冰也嵌入了语音情感识别模块，希望能够和人们更自然的聊天。IBM 与软银合作推出了具有情绪感知能力的机器人 Pepper。苹果也开始在自己的产品中加入语音情感识别的功能。国内的公司在近几年也开始这一领域的产品化，科大讯飞公司已经开始推出语音情感识别相关的技术支持，百度视频也推出基于情感识别的内容推荐系统。此外，也有一些创业公司开始关注语音情感识别，例如竹简智能。

1.2　研究现状

语音情感识别根据处理流程不同大致可以分为两种类型，一种是传统从语音信号中抽取声学特征，然后通过分类器区分情感类别，另一种是直接将原始语音信号映射到情感类别的端到端系统，下面将分别介绍这这两类方法。

1.2.1　传统的语音情感识别

传统的语音情感识别主要可以分为两个部分，第一部分是抽取和情感信息相关的声学特征，第二部分是使用分类器将特征向量映射到对应的情感类别，图 1 是传统语音情感识别的流程图。

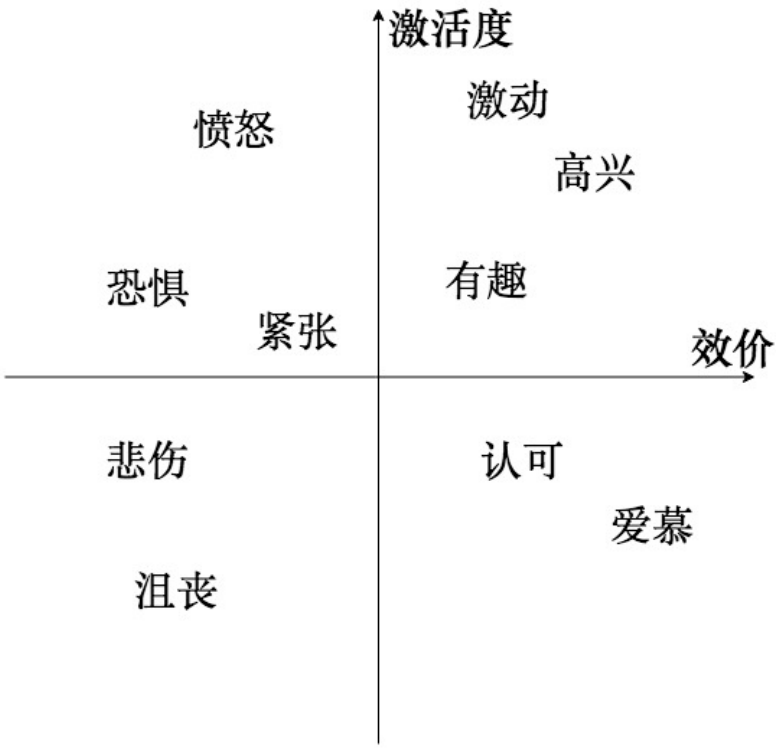

图 1.1　激活度-效价情感空间模型

特征选择是一个在所有分类问题中都存在的问题，目的是抽取与分类目标最相关的特征子集。语音情感识别作为一个分类任务，在这方面已经有很多的研究工作。首先，韵律学相关的特征已经被广泛的应用在语音情感识别，包括基频相关的特征，能量相关的特征和时长相关的特征。其次，谱相关的特征也在情感识别中起到了重要的作用，例如线性预测系数（Linear Prediction Coefficients, LPC），线性预测倒谱系数（Linear Prediction Cepstral Coefficients, LPCC）以及梅尔频率倒谱系数（Mel-frequency Cepstral Coefficients, MFCC）。此外，声音质量相关的特征也被证明同情感识别任务相关。

从语音信号中抽取声学特征后，接下来就变成了一个基本的分类问题，有许多的分类模型已经被运用在情感识别任务中。隐马尔可夫模型（Hidden Markov Model, HMM）是被广泛使用在语音识别中的一种模型，它的运作原理和语音产生的机制十分相似，因此这种模型同样也在语音情感识别中被使用。高斯混合模型（Gaussian Mixture Models, GMM）是一种概率分布模型，它使用多变量的高斯分布来预测当前语音句子属于不同情感类别的概率，在一些公开数据集上取得了不错的效果。支持向量机（Support Vector Machine, SVM）是一种被广泛使用在许多模式识别任务中的分类模型，它通过核函数（Kernel Functions）将低维空间无法区分的特征向量映射到更高维的空间，然后用线性分类器将其区分，这种模型在许多

语音情感识别的研究中也被使用。随着计算能力和存储容量的提升，深度学习模型在近几年变得逐渐流行，在许多的任务上都超过了传统的机器学习算法。在语音情感识别领域，深度学习的方法同样也被广泛使用。除了上面这些单独的分类模型，还有一些工作尝试将多种模型混合使用，然后共同决策最后的情感类别，期望能够提高系统的鲁棒性。

1.2.2　端到端的语音情感识别

在最近几年，深度学习的方法和工具已经被运用到语音处理领域，包括用于特征抽取，分类和回归任务，或者两者兼而有之。一些试验结果显示在原始语音信号上使用深度神经网络提取特征可以比采用人工定义的声学特征得到更好的效果。这导致许多的研究者开始采用端到端的系统，即省略声学特征提取的过程，直接建立从原始语音信号到任务目标的映射，图 2 是端到端的语音情感识别流程图。

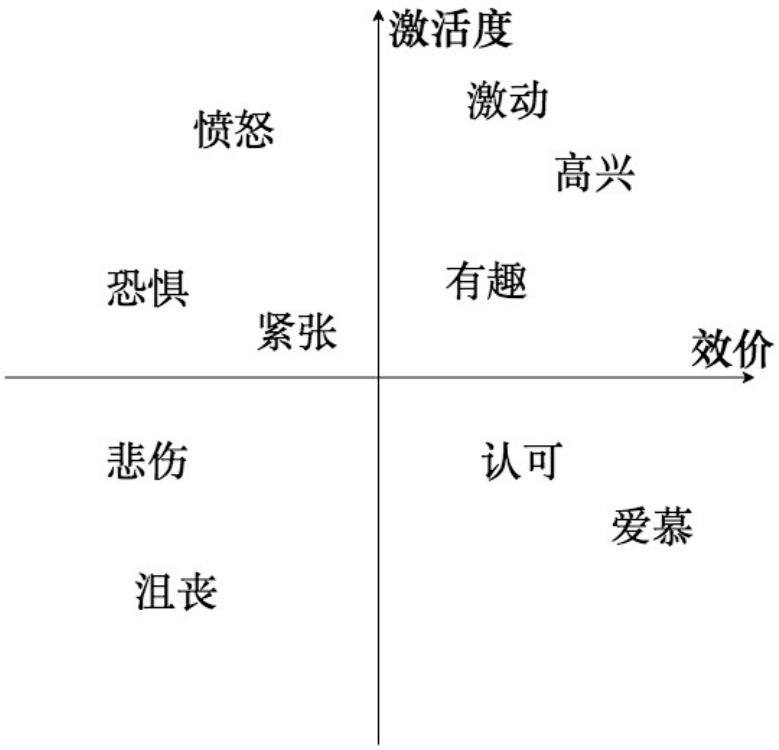

图 1.2　激活度-效价情感空间模型

这种端到端的系统最早出现在语音识别领域，最早的工作是 Jaitly 等人通过受限玻尔兹曼机（Restricted Boltzmann Machine, RBM）从原始语音信号上得到一种有利于语音识别的中间表示。Bhargava 等人则是通过堆叠的全连接神经网络从原始语音信号得到瓶颈特征（Bottleneck Feature），并且取得了和使用梅尔频率倒谱系数（Mel-frequency Cepstral Coefficients, MFCC）相近的效果。Sainath 等人将

CNN 和 LSTM-RNN 用在大词表语音识别上。Hannun 和 Amodei 在线性语谱图（Linearly-Spaced Spectrogram）上采用深度神经网络，并搭建出了当时最好的语音识别系统。此外，梅尔语谱图（Mel-Scale Spectrogram）上使用深度学习方法，在说话人识别上也取得了很不错的效果

在语音情感识别领域也开始有一些端到端的方法被提出。George 等人提出一种使用 CNN 从原始语音信号提取特征，然后通过 LSTM-RNN 捕获输入序列的时序信息并最终输出不同情感的后验概率，并且他们发现 LSTM-RNN 不同节点的输出和一些声学特征有很强的相关性。Satt 等人也采用了相似的神经网络结构，但不同的是他们从语谱图（Spectrogram）上抽取特征而非原始语音信号。他们认为在语谱图（Spectrogram）上可以更方便的进行去噪的操作，并且他们在公开情感语音数据集 IEMOCAP 上取得了超过之前最好结果（The State of Art）的准确率。

1.3　本文的主要研究内容和贡献

1.3.1　研究内容和各章介绍

语音情感识别效果与声学特征的选取是密切相关的，抽取什么特征以及如何抽取特征最终会影响到识别率。本文会从传统的语音情感识别和端到端的语音情感识别两类方法分别研究特征抽取的方法，意使最终的情感识别准确率得到提高。

第2章主要介绍语音情感识别相关的基础知识，包括情感的定义，情感语音数据库，情感相关的声学特征以及情感分类模型这四个方面的内容。首先，我们介绍了两种不同的情感定义方式，并分析了各自的优缺点。接下来我们介绍了不同类型的情感语音数据库，包括他们的采集方式和标注方式。然后我们会对情感相关的声学特征做一个全面的介绍，包括特征类型，特征选择算法和深度神经网络抽取特征。最后我们会对不同的语音情感分类模型做出介绍，包括传统的机器学习模型和深度学习模型。

第3章介绍了基于情感对的语音情感识别框架。通过将任意两种不同的情感组成情感对，并为不同的情感对分别选择不同的特征子集构建二分类器，从而得到更精确的二分类结果。进一步，我们发现在维度情感空间中，不同情感之前的距离并不相同，距离越近表示情感之间更相似，越远表示越不相似。依据这种信息，我们构建贝叶斯分类器来对所有情感对的输出结果做决策融合，从而得到最终的情感类别。在实验结果上我们超过了为所有情感的分类选取相同特征集的方法，同时我们的效果也优于基于决策树的分层识别框架，这种框架的设计思想和我们很相似，但我们的方法效果更好，而且当情感类别变化时更方便构建。

第4章介绍了基于变长语音段的端到端的情感识别方法。在这一章我们

将不再采用传统的声学特征提取方法，而是转而使用深度神经网络将语谱图（Spectrogram）映射到对应的情感类别。我们会将语音信号首先转换为语谱图（Spectrogram），然后采用 CNN 直接在语谱图（Spectrogram）上抽取特征。由于语音信号属于时间序列信号，因此在 CNN 抽取特征后我们将采用 RNN 建模输入序列的时序信息，并将最后一个时间步的输出传给后面的全连接网络，最后通过 softmax 将输入映射到每种情感的后验概率。此外，之前的语音情感识别模型在抽取特征时，要么是对一个句子的所有语音帧的特征计算统计量，要么是将句子切分成多个更短的等长语音段，并给将每个语音段都标记为句子所属的情感类别，因为大多数网络结构更容易去处理定长的输入序列。但是计算所有语音帧的特征统计量就会完全丢失不同帧之间的时序关系，而将句子切分成更短的等长语音段又无法保证每个语音段都包含有情感信息，这会导致网络在训练时对中性语音和情感语音产生混淆。因此，我们设计了一种能够处理变长语音段的神经网络结构，通过掩码可以保证在训练和测试的过程中都不需要对语音段进行切分，这样就可以避免上面的问题。实验结果显示，我们的方法得到的结果在公开的情感语音数据库 IEMOCAP 上超过了之前报道的最好的结果。

第5章对本文中关于语音情感识别中声学特征的相关研究成果进行了总结，同时，还对基于情感对的识别框架和变长神经网络的使用前景进行了展望。

1.3.2　本文主要贡献

本文的主要贡献点有以下几个方面：**一、提出一种基于情感对的语音情感识别框架，为不同的情感对选取不同的声学特征，并在最后的决策融合过程中引入心理学的情感空间模型，将所有情感对的二分类结果通过贝叶斯分类器后得到最终的情感类别，从而提升了系统的识别率。**

传统的语音情感识别系统通常为所有的情感选取相同的声学特征来完成最后的情感分类，但实验证明不同的情感和不同的声学特征的相关性并不同。针对这一问题，我们将分别为不同的情感对选取不同的特征集合，将原先的多分类问题转变为多个二分类问题，从而得到更精确的二分类结果。此外，由于不同的情感在维度情感空间中的分布位置不同，相似的情感会相对较近。依据这种信息，我们构建贝叶斯分类器来对所有情感对的输出结果做决策融合，从而得到最终的情感类别。在公开的情感语音数据集 IEMOCAP 上，我们方法取得了比传统的识别框架更好的准确率。

二、设计了一种能够处理变长语音段的神经网络结构，消除了语音分段带来的中性语音和情感语音的混淆，从而提升了系统的识别率。

在使用深度神经网络实现端到端的语音情感识别系统时，由于卷积神经网络（Convolution Neural Network, CNN）和循环神经网络（Recurrent Neural Network, RNN）很难处理变长的输入，通常会把变长的语音句子切分成多段等长的语音段，然后将所有语音段都标记为对应句子的情感标签, 但这样会导致部分中性语音段被标记为有情感，从而造成中性语音和情感语音的混淆。针对于这一问题，我们采用补齐和掩码的方式来处理神经网络中变长的输入序列，避免了错误标注带来了的效果变差。相对于切分成更小的等长语音段的方法，我们直接输入整个变长的语音句子的方法可以在相同的数据集上取得更好的准确率。

第 2 章　语音情感识别的相关工作

2.1　本章引论

目前，语音情感识别方面的研究工作越来越多。由于情感感知的人为主观性比较强，这导致当前的研究不只是模型算法方面的研究，还包括心理学方面的研究和人文社会学方面的研究。此外，情感语音数据库的采集和标注也是一个很大的挑战。本章我们首先对语音情感识别的相关工作进行简单的介绍，其中包括了情感的定义、情感语音数据库、声学特征的抽取以及情感分类模型的构建四个方面。

2.2　情感的定义

情感在心理学上的定义为：“人对客观现实的一种特殊反映形式，是人对于客观事物是否符合人的需要而产生的态度的体验”，但这种定义太过宽泛，在实际的语音情感识别任务中无法运用，只有将情感通过数学量化表示后才能够被模型处理。目前对于情感的主流定义方式大致分为两种，分别是离散的情感标签定义和连续的情感维度空间定义，下面将分别对这两种定义进行介绍。

2.2.1　离散的情感类别标签

在我们的日常生活中，我们在描述自己的主观感受时，通常会用一些特定的词汇，例如高兴，愤怒，悲伤等等。在情感识别中通常会将这些词汇作为情感的类别，进而将任务转化为多分类问题。关于具体应该将情感分为那些类别，不同的学者有着不同的定义和划分，下面的表格2.1列举了不同的定义方式。

这种定义情感的方式的优点是简单、易懂，而且可以有比较明显的应用场景，这也使得当前关于情感识别的研究主要都是基于这种定义进行的，本文也主要是基于这种情感定义展开的。但缺点是对情感的描述能力有限，因为情感标签的描述太过模糊，对于一些复杂的情感无法准确的定义。例如，愤怒可以分为冷愤怒和热愤怒；又比如，高兴又可以分为不同的等级，从喜上眉梢，到眉飞色舞，再到手舞足蹈。

表 2.1　不同学者对情感的定义

学者	情感类别
Arnold	愤怒，厌恶，无畏，忧郁，渴望，绝望，珍视，憎恨，希冀，爱慕，悲伤
Ekman, Friesen, Ellsworth	愤怒，厌恶，恐惧，高兴，悲伤，惊讶
Fridja,Gray	希冀，高兴，有趣，惊讶，渴望，悲伤
Izard	愤怒，轻蔑，厌恶，悲伤，恐惧，内疚，有趣，高兴，羞愧，惊讶
James	恐惧，悲伤，爱慕，愤怒
McDougall	恐惧，厌恶，高兴，顺从，柔和的情感，渴望
Oatley, Johnson-Laird, Panksepp	愤怒，厌恶，焦虑，高兴，悲伤
Plutchik	认可，愤怒，希冀，厌恶，高兴，恐惧，悲伤，惊讶
Tomkins	愤怒，有趣，轻蔑，厌恶，悲伤，恐惧，高兴，羞愧，惊讶
Watson	恐惧，爱慕，愤怒
Weiner, Graham	高兴，悲伤

2.2.2　连续的情感维度空间

另外一种情感的定义方式是连续的情感维度空间定义，这种定义将情感映射到一个笛卡尔空间坐标系中，不同的坐标轴分别代表不同的心理学属性，每一种情感都可以被视为坐标系中的一个点。常用的情感空间模型有二维情感空间（激活度-效价）和三维情感空间（激活度-效价-支配力），下面的图 2 和图 3 分别是两种情感维度空间的示意图。

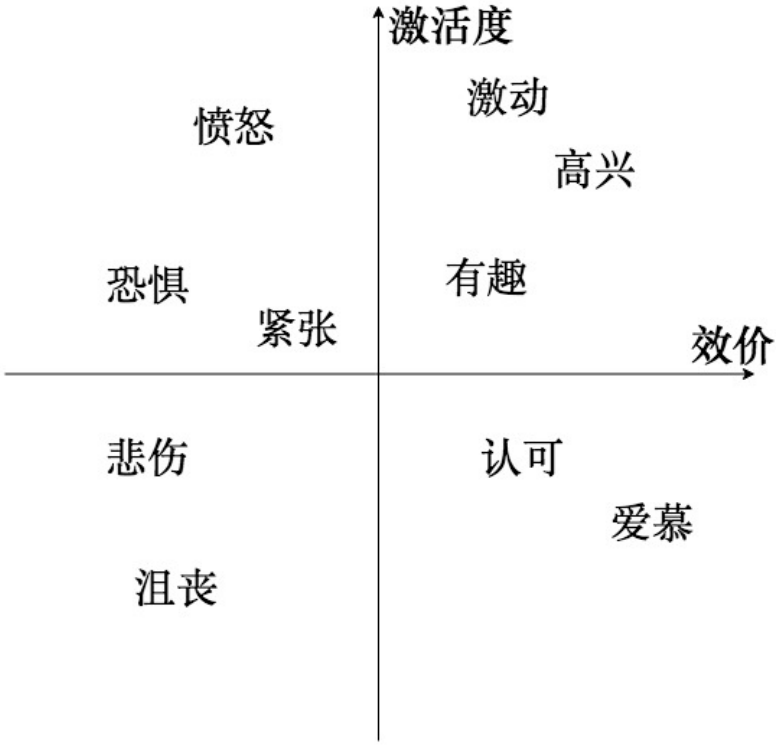

图 2.1　激活度-效价情感空间模型

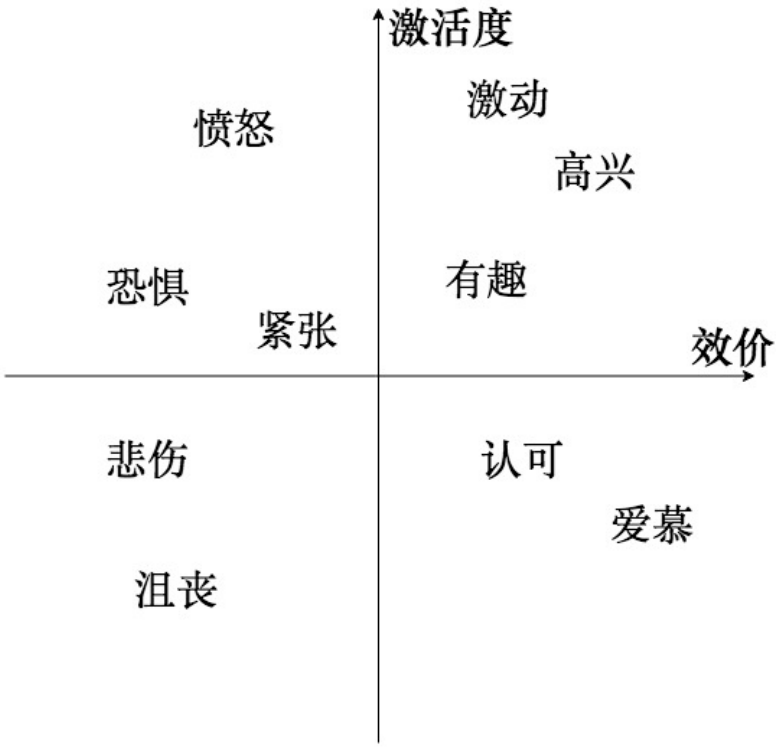

图 2.2　激活度-效价情感空间模型

这种情感空间模型理论上可以描述任何情感，其中激活度代表情感的强烈程

度，例如愤怒就是激活度非常高的情感。效价代表情感的积极性，例如悲伤就是一种积极性很低的情感，所以它的效价很低。支配力代表情感对别人的影响程度，例如高兴就是对别人影响比较大的一种情感，所以支配力比较大。情感空间模型将原先的标签分类问题转换为的对心理学属性值的回归问题，从而能够描述更为复杂的情感。但这种模型的的缺点是标记数据的成本太高，因为将主观情感量化为客观数值是一个繁重且难以保证质量的过程。

2.3　情感语音数据库

在大多数的监督性机器学习问题中，训练数据一直是一个很大的问题，尤其是如何得到高质量的训练数据，对于语音情感识别更是如此。由于语音中情感信息的标注主要是要依靠人来完成的，但不同的人对情感的感知程度是不同的，所以同一句话有可能被不同的人标注为不同的情感。下面的两小节将主要介绍情感语音数据库的设计准则和一些常用的情感语音数据库。

2.3.1　设计准则

如何判断情感语音数据库能够模拟真实的应用场景，这需要相应设计准则来指导，下面将介绍几种主要的设计准则。

自然语音还是表演式语音？通常来说，最符合实际的语音数据应该是从日常生活的对话中收集，例如广播电台，电话客服系统等，这样的录音包含有最自然的情感表达。但不幸的是，由于一些法律和道德的原因，这样的数据被禁止用作研究目的。所以现在大多数情感语音数据库都采用了另一种替代方式，聘请一些专业的演员在录音室中去演绎预定的情感语音。尽管有一些学者认为这样的得到的语音情感表现过于夸张，和实际的自然语音不一致，但是这并不影响用这种数据库来探索声学表现和情感之间的相关性。

录制时如何唤醒情感？在录制情感语音数据库时，首先需要做的就是唤醒说话人的情感，通常有三种唤醒方式。第一种就是让说话人根据规定情感进行表演，但这种方式得到情感表现过于夸张，和自然语音中的表现不一致。第二种就是将说话人置于某些特定的环境下来激起对特定的情感反应，例如通过一些诱发式的交谈，或者一些交互式的游戏。第三种就是让标注者从生活中录制的自然语音去标注出有情感的句子，这种语音最为真实，但是标注成本太高，需要大量的人力劳动。

不同情感的数据量是否平衡？由于在日常生活中，不同的情感触发的几率并不相同，所以会导致包含不同情感的语音数量也不同，例如中性语音是日常生活

中出现最多的。这种分布不平衡的数据库会导致在训练分类器时出现偏置，使得分类器更趋向于预测为数据量多的那种情感。有一些数据库为了保持分布平衡会保证不同情感的句子数量基本一致。但大多数研究者认为这种分布正体现了实际应用场景的情感出现概率，所以应该通过调整模型来包含这种信息。

情感语音是否应该保证说话人以及说话内容无关？由于不同的人表达情感的方式不相同，如果数据库中只包含个别人的语音就会导致模型不够强健，无法识别其他人的语音。应该保证尽可能多的说话人。还有就是语音中的语言学信息通常和情感都是强相关的，在录制数据库时是否应该排除掉语言学信息的影响。现在大多数研究者的观点是对于提前准备台词的表演型数据库，由于情感触发和文本是相关联的，所以并不适合用于语音情感识别。

2.3.2　常用的情感语音数据库

由于大多数的情感语音数据库都不是公开的，所以只有很少的基准数据库可以被研究者们共享。但由于情感语音数据的录制没有标准的规范，所以导致不同数据库的录制方式各不相同，下面的表格列举了一些常用的情感语音数据库。

这里主要介绍下 IEMODB 这个情感语音数据库，因为本文的研究工作主要是以这个数据库作为实验基础的。IEMODB 主要被设计用于多模态情感表现研究的，它包括肢体动作，音频和视频，一共有 5 个部分，每个部分包括 10 个主题，总共有接近 12 个小时的数据。每一个部分包含一个不同的对话场景，会有一个男演员和一个女演员分别表演规定好的剧本，以及在一个对话中诱发情感。至少三个标记员对同一句话标记情感类别，包括高兴，悲伤，中性，愤怒，惊讶，激动，沮丧，厌恶，恐惧这些情感标签。这个数据库被许多的研究工作采用，因此可以用来与别人的实验结果作对比。

2.4　声学特征的抽取

语音情感识别中一个重要的问题就是抽取与情感相关的声学特征，因为这些特征作为模型的输入会直接影响到最终的分类效果。下面将会对特征的选择和抽取方式做出介绍。

2.4.1　人工选择情感相关的声学特征

在大多数研究中，声学特征都是通过以往的一些经验选择或者设计出来的。用于语音情感识别的声学特征大致可以分为三类，分别为韵律学相关的特征，谱相关的特征以及声音质量相关的特征。

表 2.2　常用的情感语音数据库

数据库名	语言	大小	来源
LDC	英语	7 人 ×15 种情感 ×10 个句子	专业演员
柏林情感语音数据库	德语	10 人 ×7 种情感 ×10 个句子	专业演员
丹麦情感语音数据库	丹麦语	4 人 ×5 种情感	非专业演员
Natural	普通话	11 人 ×2 种情感	呼叫中心
ESMBS	普通话	12 人 ×6 种情感	非专业演员
INTERFACE	英语，斯洛文尼亚语，西班牙语，法语	635 个句子	专业演员
KISMET	美式英语	3 人 ×5 种情感	非专业演员
BabyEars	英语	12 人 ×3 种情感	父亲和母亲
SUSAS	英语	16000 个句子	压力下的模仿
MPEG-4	英语	2440 个句子	美国电影
北航情感语音数据库	普通话	7 人 ×5 种情感 ×20 个句子	非专业演员
FERMUS III	德话，英语	13 人 ×7 种情感	诱发环境
KES	韩语	5400 个句子	非专业演员
CLDC	汉语	1200 个句子	非专业演员
Pereira	英话	2 人 ×5 种情感 ×8 个句子	非专业演员
IEMODB	英话	10 人 ×9 种情感	专业演员

韵律是指人说话时的节奏，轻重，快慢和音高等方面的的变化，它与语音中携带的语言学信息并没有太大的关联，但却决定着一句话给听众的感觉，因此又被称为“超音段特征”或“辅助语言学特征”。这种韵律学相关的特征被广泛的应用在语音情感识别领域，主要包括基频，时长，能量，共振峰等。根据 Williams 和 Stevens 的研究，语音情感的激活度会显著的影响频谱上的能量分布，基频的大小以及停顿的时长，其他一些研究也证明了这一结论。此外，有研究证明这些特征也和基本情感类别有着很强的关联，例如 Murray 和 Arnott 的研究证明快的说话速

率与愤怒是相关联的。但也有些研究表明部分情感的韵律学比较相似，例如愤怒，恐惧，高兴和惊讶都有相似的基频。

谱相关的特征被认为与声道对语音信号的调制相关联，这类特征之前一直被语音识别广泛的应用，但现在一些研究证明这类特征在情感识别中也发挥很大的作用，例如线性预测系数（Linear Prediction Coefficients, LPC），线性预测倒谱系数（Linear Prediction Cepstral Coefficients, LPCC）以及梅尔频率倒谱系数（Mel-frequency Cepstral Coefficients, MFCC）。Nwe 等人发现语音信号不同频段的能量分布和情感类别有着相关性，例如高兴地语音通常在高频段有着较高的能量，而悲伤的语音在高频段的能量却相对较低。

声音质量特征

2.4.2　选择算法筛选情感相关的声学特征

2.4.3　深度神经网络抽取情感相关的声学特征

2.5　情感分类模型的构建

2.5.1　基于传统机器学习的情感分类模型

2.5.2　基于深度学习的情感分类模型

第 3 章　基于情感对的语音情感识别框架

第 4 章　基于变长语音段的端到端的情感识别

第 5 章　总结与展望

5.1　本文工作总结

5.2　未来工作展望

插图索引

表格索引

公式索引

参考文献

薛瑞尼. 2017. ThuThesis: 清华大学学位论文模板[EB/OL]. https://github.com/xueruini/thuthesis.

致　谢

衷心感谢导师 xxx 教授和物理系 xxx 副教授对本人的精心指导。他们的言传身教将使我终生受益。

在美国麻省理工学院化学系进行九个月的合作研究期间，承蒙 xxx 教授热心指导与帮助，不胜感激。感谢 xx 实验室主任 xx 教授，以及实验室全体老师和同学们的热情帮助和支持！本课题承蒙国家自然科学基金资助，特此致谢。

感谢 LaTeX 和 ThuThesis薛瑞尼 (2017)，帮我节省了不少时间。

声　明

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

签　名：_____ 日　期：_____

附录 A　外文资料原文

The title of the English paper

Abstract: As one of the most widely used techniques in operations research, *mathematical programming* is defined as a means of maximizing a quantity known as *bjective function*, subject to a set of constraints represented by equations and inequalities. Some known subtopics of mathematical programming are linear programming, nonlinear programming, multiobjective programming, goal programming, dynamic programming, and multilevel programming[1].

It is impossible to cover in a single chapter every concept of mathematical programming. This chapter introduces only the basic concepts and techniques of mathematical programming such that readers gain an understanding of them throughout the book[2,3].

A.1　Single-Objective Programming

The general form of single-objective programming (SOP) is written as follows,

$$\begin{cases} \max \ \ f(x) \\ \text{subject to:} \\ \qquad g_j(x) \le 0, \quad j = 1, 2, \cdots, p \end{cases} \tag{123}$$

which maximizes a real-valued function f of $x = (x_1, x_2, \cdots, x_n)$ subject to a set of constraints.

Definition A.1： In SOP, we call x a decision vector, and $x_1, x_2, \cdots, x_n$ decision variables. The function f is called the objective function. The set

$$S = \left\{ x \in \Re^n \;\middle|\; g_j(x) \le 0,\ j = 1, 2, \cdots, p \right\} \tag{456}$$

is called the feasible set. An element x in S is called a feasible solution.

Definition A.2：　A feasible solution x^* is called the optimal solution of SOP if and only if

$$f(x^*) \geq f(x) \tag{A-1}$$

for any feasible solution x.

One of the outstanding contributions to mathematical programming was known as the Kuhn-Tucker conditionsA-2. In order to introduce them, let us give some definitions. An inequality constraint $g_j(x) \leq 0$ is said to be active at a point x^* if $g_j(x^*) = 0$. A point x^* satisfying $g_j(x^*) \leq 0$ is said to be regular if the gradient vectors $\nabla g_j(x)$ of all active constraints are linearly independent.

Let x^* be a regular point of the constraints of SOP and assume that all the functions $f(x)$ and $g_j(x), j = 1, 2, \cdots, p$ are differentiable. If x^* is a local optimal solution, then there exist Lagrange multipliers $\lambda_j, j = 1, 2, \cdots, p$ such that the following Kuhn-Tucker conditions hold,

$$\begin{cases} \nabla f(x^*) - \sum\limits_{j=1}^{p} \lambda_j \nabla g_j(x^*) = 0 \\ \lambda_j g_j(x^*) = 0, \quad j = 1, 2, \cdots, p \\ \lambda_j \geq 0, \quad j = 1, 2, \cdots, p. \end{cases} \tag{A-2}$$

If all the functions $f(x)$ and $g_j(x), j = 1, 2, \cdots, p$ are convex and differentiable, and the point x^* satisfies the Kuhn-Tucker conditions (A-2), then it has been proved that the point x^* is a global optimal solution of SOP.

A.1.1　Linear Programming

If the functions $f(x), g_j(x), j = 1, 2, \cdots, p$ are all linear, then SOP is called a *linear programming*.

The feasible set of linear is always convex. A point x is called an extreme point of convex set S if $x \in S$ and x cannot be expressed as a convex combination of two points in S. It has been shown that the optimal solution to linear programming corresponds to an extreme point of its feasible set provided that the feasible set S is bounded. This fact is the basis of the *simplex algorithm* which was developed by Dantzig as a very efficient method for solving linear programming.

Table 1　This is an example for manually numbered table, which would not appear in the list of tables

Network Topology		# of nodes	# of clients			Server
GT-ITM	Waxman Transit-Stub	600	2%	10%	50%	Max. Connectivity
Inet-2.1		6000				
Xue	Rui	Ni	ThuThesis			
	ABCDEF					

Roughly speaking, the simplex algorithm examines only the extreme points of the feasible set, rather than all feasible points. At first, the simplex algorithm selects an extreme point as the initial point. The successive extreme point is selected so as to improve the objective function value. The procedure is repeated until no improvement in objective function value can be made. The last extreme point is the optimal solution.

A.1.2　Nonlinear Programming

If at least one of the functions $f(x), g_j(x), j = 1, 2, \cdots, p$ is nonlinear, then SOP is called a *nonlinear programming*.

A large number of classical optimization methods have been developed to treat special-structural nonlinear programming based on the mathematical theory concerned with analyzing the structure of problems.

Figure 1　This is an example for manually numbered figure, which would not appear in the list of figures

Now we consider a nonlinear programming which is confronted solely with maximizing a real-valued function with domain $\Re^n$. Whether derivatives are available or not, the usual strategy is first to select a point in $\Re^n$ which is thought to be the most likely place where the maximum exists. If there is no information available on which to base such a selection, a point is chosen at random. From this first point an attempt is made to construct a sequence of points, each of which yields an improved objective function value over its predecessor. The next point to be added to the sequence is chosen by analyzing the behavior of the function at the previous points. This construction continues

until some termination criterion is met. Methods based upon this strategy are called *ascent methods*, which can be classified as *direct methods*, *gradient methods*, and *Hessian methods* according to the information about the behavior of objective function f. Direct methods require only that the function can be evaluated at each point. Gradient methods require the evaluation of first derivatives of f. Hessian methods require the evaluation of second derivatives. In fact, there is no superior method for all problems. The efficiency of a method is very much dependent upon the objective function.

A.1.3 Integer Programming

Integer programming is a special mathematical programming in which all of the variables are assumed to be only integer values. When there are not only integer variables but also conventional continuous variables, we call it *mixed integer programming*. If all the variables are assumed either 0 or 1, then the problem is termed a *zero-one programming*. Although integer programming can be solved by an *exhaustive enumeration* theoretically, it is impractical to solve realistically sized integer programming problems. The most successful algorithm so far found to solve integer programming is called the *branch-and-bound enumeration* developed by Balas (1965) and Dakin (1965). The other technique to integer programming is the *cutting plane method* developed by Gomory (1959).

Uncertain Programming　(*BaoDing Liu,* 2006.2)

References

NOTE: These references are only for demonstration. They are not real citations in the original text.

[1] Donald E. Knuth. The TEXbook. Addison-Wesley, 1984. ISBN: 0-201-13448-9

[2] Paul W. Abrahams, Karl Berry and Kathryn A. Hargreaves. TEX for the Impatient. Addison-Wesley, 1990. ISBN: 0-201-51375-7

[3] David Salomon. The advanced TEXbook. New York : Springer, 1995. ISBN:0-387-94556-3

附录 B　外文资料的调研阅读报告或书面翻译

英文资料的中文标题

摘要：本章为外文资料翻译内容。如果有摘要可以直接写上来，这部分好像没有明确的规定。

B.1　单目标规划

北冥有鱼，其名为鲲。鲲之大，不知其几千里也。化而为鸟，其名为鹏。鹏之背，不知其几千里也。怒而飞，其翼若垂天之云。是鸟也，海运则将徙于南冥。南冥者，天池也。

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}, y)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})} \tag{123}$$

吾生也有涯，而知也无涯。以有涯随无涯，殆已！已而为知者，殆而已矣！为善无近名，为恶无近刑，缘督以为经，可以保身，可以全生，可以养亲，可以尽年。

B.1.1　线性规划

庖丁为文惠君解牛，手之所触，肩之所倚，足之所履，膝之所倚，砉然响然，奏刀騞然，莫不中音，合于桑林之舞，乃中经首之会。

表 1　这是手动编号但不出现在索引中的一个表格例子

Network Topology		# of nodes	# of clients			Server
GT-ITM	Waxman Transit-Stub	600	2%	10%	50%	Max. Connectivity
Inet-2.1		6000				
Xue	Rui	Ni	ThuThesis			
	ABCDEF					

文惠君曰：“嘻，善哉！技盖至此乎？”庖丁释刀对曰：“臣之所好者道也，进乎技矣。始臣之解牛之时，所见无非全牛者；三年之后，未尝见全牛也；方今之时，臣以神遇而不以目视，官知止而神欲行。依乎天理，批大郤，导大窾，因其固然。技经肯綮之未尝，而况大軱乎！良庖岁更刀，割也；族庖月更刀，折也；今臣之刀十九年矣，所解数千牛矣，而刀刃若新发于硎。彼节者有间而刀刃者无厚，以

无厚入有间，恢恢乎其于游刃必有余地矣。是以十九年而刀刃若新发于硎。虽然，每至于族，吾见其难为，怵然为戒，视为止，行为迟，动刀甚微，謋然已解，如土委地。提刀而立，为之而四顾，为之踌躇满志，善刀而藏之。”

文惠君曰：“善哉！吾闻庖丁之言，得养生焉。”

B.1.2　非线性规划

孔子与柳下季为友，柳下季之弟名曰盗跖。盗跖从卒九千人，横行天下，侵暴诸侯。穴室枢户，驱人牛马，取人妇女。贪得忘亲，不顾父母兄弟，不祭先祖。所过之邑，大国守城，小国入保，万民苦之。孔子谓柳下季曰：“夫为人父者，必能诏其子；为人兄者，必能教其弟。若父不能诏其子，兄不能教其弟，则无贵父子兄弟之亲矣。今先生，世之才士也，弟为盗跖，为天下害，而弗能教也，丘窃为先生羞之。丘请为先生往说之。”

图 1　这是手动编号但不出现索引中的图片的例子

柳下季曰：“先生言为人父者必能诏其子，为人兄者必能教其弟，若子不听父之诏，弟不受兄之教，虽今先生之辩，将奈之何哉？且跖之为人也，心如涌泉，意如飘风，强足以距敌，辩足以饰非。顺其心则喜，逆其心则怒，易辱人以言。先生必无往。”

孔子不听，颜回为驭，子贡为右，往见盗跖。

B.1.3　整数规划

盗跖乃方休卒徒大山之阳，脍人肝而餔之。孔子下车而前，见谒者曰：“鲁人孔丘，闻将军高义，敬再拜谒者。”谒者入通。盗跖闻之大怒，目如明星，发上指冠，曰：“此夫鲁国之巧伪人孔丘非邪？为我告之：尔作言造语，妄称文、武，冠枝木之冠，带死牛之胁，多辞缪说，不耕而食，不织而衣，摇唇鼓舌，擅生是非，以迷天下之主，使天下学士不反其本，妄作孝弟，而侥幸于封侯富贵者也。子之罪大极重，疾走归！不然，我将以子肝益昼餔之膳。”

附录 C　其它附录

前面两个附录主要是给本科生做例子。其它附录的内容可以放到这里，当然如果你愿意，可以把这部分也放到独立的文件中，然后将其 `\input` 到主文件中。

个人简历、在学期间发表的学术论文与研究成果

个人简历

xxxx 年 xx 月 xx 日出生于 xx 省 xx 县。

xxxx 年 9 月考入 xx 大学 xx 系 xx 专业，xxxx 年 7 月本科毕业并获得 xx 学士学位。

xxxx 年 9 月免试进入 xx 大学 xx 系攻读 xx 学位至今。

发表的学术论文

[1] Yang Y, Ren T L, Zhang L T, et al. Miniature microphone with silicon- based ferroelectric thin films. Integrated Ferroelectrics, 2003, 52:229-235. (SCI 收录, 检索号:758FZ.)

[2] 杨轶, 张宁欣, 任天令, 等. 硅基铁电微声学器件中薄膜残余应力的研究. 中国机械工程, 2005, 16(14):1289-1291. (EI 收录, 检索号:0534931 2907.)

[3] 杨轶, 张宁欣, 任天令, 等. 集成铁电器件中的关键工艺研究. 仪器仪表学报, 2003, 24(S4):192-193. (EI 源刊.)

[4] Yang Y, Ren T L, Zhu Y P, et al. PMUTs for handwriting recognition. In press. (已被 Integrated Ferroelectrics 录用. SCI 源刊.)

[5] Wu X M, Yang Y, Cai J, et al. Measurements of ferroelectric MEMS microphones. Integrated Ferroelectrics, 2005, 69:417-429. (SCI 收录, 检索号:896KM)

[6] 贾泽, 杨轶, 陈兢, 等. 用于压电和电容微麦克风的体硅腐蚀相关研究. 压电与声光, 2006, 28(1):117-119. (EI 收录, 检索号:06129773469)

[7] 伍晓明, 杨轶, 张宁欣, 等. 基于 MEMS 技术的集成铁电硅微麦克风. 中国集成电路, 2003, 53:59-61.

研究成果

[1] 任天令, 杨轶, 朱一平, 等. 硅基铁电微声学传感器畴极化区域控制和电极连接的方法: 中国, CN1602118A. (中国专利公开号)

[2] Ren T L, Yang Y, Zhu Y P, et al. Piezoelectric micro acoustic sensor based on ferroelectric materials: USA, No.11/215, 102. (美国发明专利申请号)

综合论文训练记录表

学生姓名		学号		班级	
论文题目					
主要内容以及进度安排	指导教师签字： 考核组组长签字： 年　　月　　日				
中期考核意见	考核组组长签字： 年　　月　　日				

指导教师评语	指导教师签字：_____ 年　　月　　日
评阅教师评语	评阅教师签字：_____ 年　　月　　日
答辩小组评语	答辩小组组长签字：_____ 年　　月　　日

总成绩：_____

教学负责人签字：_____

年　　月　　日