中国科学院大学

University of Chinese Academy of Sciences

博士学位论文

基于深度神经网络的远场语音识别声学建模研究

作者姓名：张宇

指导教师：颜永红 研究员

中国科学院声学研究所

学位类别：工学博士

学科专业：信号与信息处理

培养单位：中国科学院声学研究所

2018 年 6 月

Research on Acoustic Modeling Based on Deep Neural Network for Distant Speech Recognition

A dissertation submitted to the

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in Signal and Information Processing

By

Zhang Yu

Supervisor: Professor Yan Yonghong

Institute of Acoustics, Chinese Academy of Sciences

June, 2018

中国科学院大学
研究生学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的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，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对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或致谢。

作者签名：张宇

日　　期：2018.5.25

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：张宇　　导师签名：

日　　期：2018.5.25　　日　　期：2018.5.25

摘　要

近年来，随着计算机技术和深度学习理论的发展，基于深度神经网络（DNN）的声学建模方法获得广泛应用，相较于传统的高斯混合模型-隐马尔科夫模型（GMM-HMM），其显著提升了语音识别系统性能，识别系统在说话人距离麦克风距离较近的近场场景下已具有较高的识别准确率。与此同时，语音识别的研究热点转向了更加实际也更具挑战性的远场语音识别（DSR）。在说话人距离麦克风较远的远场环境中，语音信号受到噪声、混响以及非目标人声干扰等因素影响，导致识别准确率大幅度降低。本文对基于深度神经网络的远场语音识别声学建模技术展开研究。为提高远场语音识别性能，本文分别从声学模型的网络结构、输入特征以及训练目标值三个方面进行研究。本文的主要研究工作和创新点包括：

1. 提出一种基于注意力长短时记忆（LSTM）神经网络和多任务学习的声学建模方法。基于深度神经网络的声学模型一般将上下文多帧特征简单地拼接作为输入，缺点是忽略了对每帧特征本身所包含的时间信息的利用，因为不同时刻的特征对于当前时刻状态预测的贡献不一定是相同的。为此，本文针对基于LSTM的混合声学建模框架提出一种注意力机制，自动学习调整对上下文扩展输入特征的关注度。同时，在模型训练阶段采用联合预测声学状态和干净特征的多任务网络结构，来进一步提升声学模型在远场场景下的鲁棒性。

2. 提出一种基于空间特征补偿的多通道声学建模方法。远场语音识别系统通常采用多个麦克风录制语音信号。与采用单个麦克风相比，其优势在于可以提供目标说话人的空间信息。传统的多通道语音识别系统一般采用两个独立的系统模块，即前端的多通道语音增强和后端的语音识别器。然而，当最终目标为提高语音识别准确率时，将增强模型独立于声学模型单独优化并非最优解决方案。为解决此问题，一些研究者提出前后端联合优化的方法。但是，这些方法一般需要引入若干层神经网络来估计波束形成的滤波系数，导致最终用于识别的模型参数量较大。本文中，我们提出将编码声源位置信息的信道间相位变换广义互相关（GCC-PHAT）

做为空间特征补偿，与多通道声学特征拼接输入深度神经网络声学模型。该方法通过利用神经网络输入特征的灵活性，有效地提高了深度神经网络对多通道语音信号的声学建模能力。与此同时，将此方法与之前提出的注意力机制相结合，系统性能得到进一步提升。

3. 将教师学生迁移学习框架应用于远场语音识别声学建模。远场语音信号受噪声和混响等因素的干扰，不同声学单元之间的区别性变得模糊。若仍然采用强制对齐得到的0-1分布做为目标值，基于深度神经网络的声学模型很难学习。因此，含有更丰富信息的软判决标注更适合远场语音声学模型的训练。为得到可靠的非0-1分布目标值，本文挖掘与远场语音同步录制的近场语音信号中蕴含的信息，利用教师学生迁移学习框架训练远场语音识别声学模型。实验结果显示，与0-1分布目标值训练的声学模型相比，该模型在远场单通道和多通道语音识别任务上均取得了性能提升。与上述提出的两种方法结合后，识别系统可获得进一步的性能改善。

关键词： 远场语音识别，声学模型，深度神经网络

Abstract

With the development of computer technology and deep learning theory, deep neural network (DNN) based acoustic models have been widely used. Compared to the traditional Gaussian mixture model-hidden Markov model (GMM-HMM), it greatly improved automatic speech recognition accuracy. And state-of-the-art speech recognition systems perform well in close-talking conditions. At the same time, the research focus of speech recognition switched to distant speech recognition (DSR), which is a more practical and challenging task. In the distant talking scenarios, the speech is captured by one or more microphones located far away from the speaker, which makes it susceptible to distortion from noise, reverberation, and overlapped speech. And the performance of distant speech recognition degrades significantly even with DNN based acoustic models. This thesis concentrates on DNN-HMM based hybrid acoustic modeling for distant speech recognition. In order to improve the performance of distant speech recognition, acoustics models are optimized from three aspects: network structure, input features, and target labels of training data. The main contributions of this thesis include:

1. An attention-based LSTM with multitask learning is proposed for acoustic modeling in distant speech recognition. The input of a conventional neural network acoustic model is formed from a context window of acoustic features, which results in that the temporal information within the input layer is ignored. However, the contribution from each frame at the input layer to the state prediction may be not the same. Therefore, an attention mechanism is proposed to be embedded within an LSTM based acoustic model to automatically tune its attention to the input frames at each time step. At the same time, multitask learning architecture is incorporated to improve robustness, in which the network is trained to perform both a primary senone classification task and a secondary feature enhancement task.

2. A spatial feature compensation method is proposed for DNN based acoustic models in multichannel speech recognition. Acoustic signals from microphone arrays are used to improve performance in distant speech recognition due to the availability of spatial information. Multichannel speech recognition systems often adopt a two-part architecture, in which a beamforming algorithm is applied to enhance the speech, followed by conventional acoustic modeling approaches. Since the speech enhancement part is usually separate from the speech recognition part, the system fails to optimize towards the final objective, i.e. speech recognition accuracy, which leads to a suboptimal solution. To obtain an optimal performance, joint training of speech enhancement and acoustic model was proposed to improve speech recognition accuracy. However, several layers of neural network were often added in those studies to estimate the beamforming filter coefficients, resulting in a larger amount of model parameters. In this thesis, we propose to improve multichannel speech recognition by supplying the generalized cross correlation with phase transform (GCC-PHAT) between microphones, which encodes the location of the speaker, as input features to DNN based acoustic model in parallel with the multichannel acoustic features. The proposed model improves the ability of DNN to model the acoustic signals from the microphone array by utilizing the flexibility in input features of neural network. In addition, further improvements can be achieved by combining this method with the attention mechanism.

3. The teacher-student transfer learning framework is applied to acoustic modeling in distant speech recognition. The distinction between different acoustic units becomes more ambiguous as speech in distant scenarios is corrupted by noise and reverberation. If the hard label obtained by forced alignment is still used, DNN based acoustic models may be hard to learn. Therefore, the soft label which provides much more information than hard label is more suitable to train acoustic models for distant speech recognition. In order to obtain a reliable soft label, the information contained in the simultaneously recorded close-talking data is investigated, and the

teacher-student transfer learning framework is used to train acoustic models. The experimental results show that the model trained with the soft targets instead of the hard labels achieves improvements on single channel and multichannel speech recognition tasks. And it obtains further performance improvements through combining with the above two proposed methods.

Key Words: Distant Speech Recognition, Acoustic Model, Deep Neural Network

目　录

表　格

插　图

第一章 绪论

1.1 研究背景及意义

随着计算机技术和互联网的高速发展，人类迈入了信息化时代。近年来，移动互联网的兴起更是对人们的生活和沟通方式产生了重要的影响。手机、平板电脑、智能手表以及智能音箱等智能终端设备的迅速发展和普及，使得人机交互领域成为新的研究热点。语音是人类沟通和信息传递的最主要方式。因此，利用语音进行人机交互是一个非常重要的研究课题。自动语音识别（Automatic Speech Recognition, ASR）技术作为语音交互方式的核心技术，被应用于越来越广泛的实际应用场景中。

语音识别是将语音信号通过计算机转化为相应文本或命令的一项技术，其使得计算机具有类似人类的听觉功能，能够听懂人类的语音，并理解人类的意图。语音识别是一门交叉学科，它涉及的领域包括生理学、声学、语言学、信号处理、信息理论、模式识别等。语音识别技术最早可追溯到二十世纪五十年代，AT&T Bell实验室搭建了针对10个英文数字的语音识别系统。然而，由于早期的识别技术发展不成熟，系统的识别准确率过低，语音识别一直没有在真正意义上走进人们的日常生活。近年来，语音识别系统的性能得到显著提升，其原因可以主要归结为以下三点：

(1) 使用了极具表达能力的声学模型，如深度神经网络（Deep Neural Network, DNN）。声学模型作为语音识别技术的重要模块，对整个识别系统的性能起到至关重要的作用。传统的语音识别系统采用隐马尔科夫模型（Hidden Markov Model, HMM）表达语音信号的时变特性，高斯混合模型（Gaussian Mixture Model, GMM）建模语音信号的发音多样性 [1]。最近几年，深度学习理论在机器学习领域兴起，其对语音识别技术同样产生了深远的影响。微软语音研究人员推动了深度神经网络在语音识别声学建模中的应用。文献 [2]中的实验结果表明，基于深度神经网络-隐马尔科夫模型的混合声学建模方法显著提升了语音识别系统性能，其成功取代传统的GMM-HMM声学建模技术，成为主流语音识别系统的标配。此后，比简单前馈网络更为复杂的神经网络模型被提出，如卷积神经网

络（Convolutional Neural Network, CNN） [3]和长短时记忆递归神经网络(Long Short-Term Memory Recurrent Neural Network, LSTM RNN) [4]，进一步提升了语音识别系统的准确率。

(2) 拥有海量的训练数据。伴随着互联网和云计算持续发展，我们获得越来越多的数据资源。这些数据会尽可能地覆盖语音中可能出现的变化，如噪声环境的变化，说话人语速的变化等。通过利用从真实场景收集的大量数据训练模型，可以提升模型的鲁棒性，使模型在不同的应用场景下均能表现出较好的性能。

(3) 通用计算图形处理器（General Purpose Graphical Processing Unit, GPGPU）的使用使得在海量数据上训练复杂神经网络模型成为可能。它强大的并行处理能力很大程度上加速了神经网络的计算，从而减少了神经网络声学模型的训练时间。另外，基于CPU/GPU 集群的任务调度方法进一步加速了模型的训练。

目前的语音识别技术在近场场景下已经取得了良好的性能，达到了实际应用的门槛。常见的一些实际应用包括互联网语音搜索、手机语音助手以及广播电视新闻自动语音标注等。因此，语音识别技术的研究热点也从近场语音识别转向更具挑战性的远场语音识别（Distant Speech Recognition，DSR)。远场与近场的区别在于，在远场条件下收集语音信号的麦克风距离说话人的位置更远。当人的声音传到麦克风时，声音衰减比较严重。因而在近场场景中很难见到的一些问题，在远场环境下就变得很明显。背景噪声、混响以及人声干扰等因素大幅度降低了远场语音识别的准确率 [5]。用户在许多远场语音识别的实际应用场合，如家居场景和会议场景等，会觉得识别效果欠佳。为了提升远场语音识别的准确率，远场场景中语音信号的录制从单麦克风向多通道的麦克风阵列转变，例如，智能音箱Google Home使用双麦克风，Amazon Echo 采用六加一的环形麦克风阵列。与单通道相比，使用麦克风阵列的优势在于其可以提供空间上的区分度，因此被广泛用来提升远场语音识别系统的性能。

信息化时代的人们对于远场语音交互系统的需求日益增长，如人机对话、远程音视频会议、智能交互电视等实际应用。声学建模技术作为整个语音识别系统中的关键技术，对识别性能有着重要的影响。因此，针对基于深度神经网络的远场语音识别声学建模技术展开研究，对于提升远场语音识别的性能，并

将其扩展到实际应用需求中具有重要的现实意义。

1.2 国内外研究现状

目前已有的关于远场语音识别的研究大致可以分为三类：基于前端的方法，基于后端的方法以及前后端联合训练的方法。基于前端的方法对语音信号或特征增强，试图最大限度的去除噪声和混响，增强后的信号或特征做为识别器的输入。基于后端的技术则是从模型层面解决问题，更新变换声学模型结构或参数，使声学模型能够更有效的建模远场语音信号。前端增强算法的优化准则一般是信号级别的准则，后端声学模型的优化标准是识别准确率。由于前端的语音增强与后端的识别模块分开优化，整个系统未能针对最终目标（语音识别准确率）进行优化，导致了次优解问题 [6]。因此，许多前后端联合优化的算法相继被提出。以下将对三类方法分别做总结介绍。

1.2.1 前端算法

语音增强方法可以分为传统经典框架下的语音信号增强算法和深度学习框架下的语音信号增强算法。比较而言，经典框架下的方法侧重于信号处理与统计量估计，不需要或仅需要少量的先验数据。深度学习框架下的方法侧重于数据分析和模型的设计训练，需要大量的先验语音数据。

1.2.1.1 经典框架下的语音增强方法

在单通道的场景下，由于只有单一的观测信号，算法往往依赖很强的假设，如假设噪声服从高斯分布或者是准平稳的。1979 年，Boll提出了谱减法，通过在频域减去观测信号中的噪声成分来增强语音信号 [7]。另一个经典的解决方案是维纳滤波，它提供了统计意义下的最小均方误差(Minimum Mean Square Error, MMSE)解。在频域中，Ephraim和Malah提出了高斯模型下的短时幅度谱MMSE最优增益估计 [8]和对数幅度谱的MMSE最优增益估计 [9]。2001 年D. D. Lee和H. S. Seung [10]提出了非负矩阵分解（Non-negative Matrix Factorization, NMF）的完整理论方法，很多研究将其应用于解决语音增强问题。NMF 将信号频谱分解为字典矩阵和权重矩阵，并通过优化二次代价函数或KL 散度，进行迭代式求解。由于可以利用少量先验语音数据来获得语音或

噪声的字典化表示，NMF 在处理结构化噪声时具有很大优势，但迭代式求解带来的问题是算法的计算复杂度较高。

在多通道的场景下，波束形成类算法集中体现了对多通道空间信息的利用，它主要包括固定波束形成和自适应波束形成两类算法。固定波束形成的滤波器系数需要在使用前进行设定，且不随时间和输入信号的变化而变化。典型的固定波束形成算法有延迟相加 [11]、超指向性 [12] 等。自适应波束形成的滤波器系数随输入数据的变化而发生改变，从而能适应时变的声学环境，得到更好的结果。最小方差无失真响应波束形成器(Minimum Variance Distortionless Response, MVDR)是一种被广泛研究的自适应波束形成算法，其在目标语言无失真的条件下最小化输出能量 [13]。MVDR的一种等价实现方式是广义旁瓣抵消（Generalized Sidelobe Canceller, GSC） [14]，它由固定波束形成器、阻塞矩阵和自适应滤波器构成。GEV(Generalized Engine Value)波束形成 [15]设计目的是最大化输出信噪比。此算法可能引入不确定的语音失真，但可以通过系数补偿来减少这种失真。

1.2.1.2 深度学习框架下的语音增强方法

深度学习的代表性框架是各种结构化的深度神经网络。在单通道场景下，它最直接的应用是将带噪信号映射为干净信号，在文献 [16]中Xu等人成功实现了幅度谱的映射并详细分析了该方法中数据的设定以及网络结构的设计。频谱映射的方法虽然取得了比传统信号估计方法更好的性能，但处理后的语音存在机械式的失真。与频谱回归映射类似，时频掩蔽分类同样采用基于深度神经网络的方法，训练目标变为掩蔽值，代价函数可以为最小均方误差或交叉熵，它通过设定合理的语音特征集和目标掩蔽，即可达到语音增强和分离效果。Wang等分析了语音特征集的选取以及理想二值掩蔽（Ideal Binary Mask, IBM)和理想率掩蔽（Ideal Ratio Mask, IRM）的性能特点 [17]，提出将逆傅里叶变换引入DNN中，并获得了更好的语音重构性能 [18]。在文献 [19] 中，Erdogan等分析了频谱损失代价和掩蔽损失代价的差异，并提出了相位敏感的掩蔽（Phase-Sensitive Mask, PSM)。

在多通道场景下，目前的大部分研究主要集中在将深度神经网络与经典波束形成方法进行融合 [20, 21]。该类方法利用DNN 估计目标源的掩蔽或能量，进而得到语音和噪声的二阶统计特性，并用于波束形成滤波器的计算。

1.2.2 后端算法

后端算法也称为基于模型的算法。提高声学模型对远场语音的建模能力是提升远场语音识别率的另一重要途径。此类算法一般将远场场景下录制的带噪语音不经过任何处理直接输入到后端的识别系统，让神经网络声学模型自动发现带噪语音与目标音素之间的关系。拥有长时信息建模能力的神经网络对噪声和混响的建模能力更强。因此，与DNN 相比，基于LSTM神经网络的声学模型 [22]进一步改善了远场语音识别系统的性能。基于后端的算法中，一个被广泛采用的方法就是多条件训练（multi-condition training） [23–25]。此算法在训练数据中提供由各种不同噪音引起的声学变化，从而减少测试数据与训练数据之间的声学分布不匹配问题。另一方面，Setzer等人将与噪声或混响相关的信息参数化，与声学特征拼接作为声学模型的输入 [26, 27]。此类方法试图在模型训练时提供更多关于噪声或混响的信息，使模型能够对远场语音更有效的建模。在实际应用中，更准确的状态级对齐标注可以带来明显的性能改善 [28]。因此，可以采用与远场麦克风平行录制的近场麦克风的数据做强制对齐得到状态级标注。

为研发智能音箱Google Home，Google的语音研究人员针对多通道语音识别的声学建模做了大量的研究。Sainath等人提出直接使用原始时域波形信号训练多通道声学模型。文献 [29]采用一层CNN同时做空间和频域滤波，滤波后的特征做为LSTM神经网络声学模型的输入。为进一步改善性能，文献 [30] 将空间滤波和频域滤波分解为网络中的两个单独层。然而，这些方法只能估计固定的滤波系数，潜在地限制了模型处理训练数据未出现的环境的能力。因此，文献 [31]提出神经网络自适应波束形成技术来解决此问题，此算法将波束形成神经网络的输出做为滤波系数。为降低模型的复杂度，文献 [32] 将上述算法在频域中实现。另一个研究方向是将端到端语音识别系统 [33–35]应用到多通道语音声学建模。Ochiai等人提出一种多通道端到端语音识别系统 [36]。Kim等人提出利用注意力机制动态调整模型对多通道信号的注意力，并将其应用于基于HMM的混合声学建模框架 [37]。

1.2.3 前后端联合训练算法

前后端联合训练算法一般在声学模型前加入一个包含神经网络结构的前端对特征做增强，后面连接神经网络声学模型。前后端所有模型参数采用声学模

型训练准则做联合优化。Gao等人提出将用于特征映射和声学建模的DNN单独训练，然后再将二者连接使用交叉熵准则联合更新微调参数 [38]。Xiao等人将声学处理的各个阶段包括多通道信号波束形成、特征提取和声学建模，作为一个统一的计算型网络 [39]。首先，频域波束形成器的参数由神经网络依据信道间的相关性特征估计。然后，将得到的滤波参数对多通道信号滤波求和得到增强后的单通道信号。最后，将增强信号提取的声学特征输入到深度神经网络进行声学建模。网络中的所有模型参数通过使用交叉熵目标函数联合训练。Zhong等人则是将多通道信号的短时傅里叶变换通过LSTM 网络预测波束形成系数 [40]。Heymann等人使用神经网络估计统计意义上最优波束形成器的掩蔽值。为更新模型参数，将它与声学建模的网络联合训练 [41]。该方法有效的避免了前端增强模块和后端声学模型之间的不匹配问题。

1.3 研究内容与结构安排

本文研究的主题是基于深度神经网络的远场语音声学建模技术，研究目标是在远场场景下的大词汇连续语音识别（Large Vocabulary Continuous Speech Recognition，LVCSR）任务上获得性能的改善和提升。本文分别从模型的网络结构，输入特征以及训练目标值三方面对远场语音识别声学模型进行优化，并通过实验对提出的方法做分析验证。本文的各章节主要内容如下：

第二章，自动语音识别系统。本章首先对语音识别的基本原理和框架进行阐述，然后描述了语音识别中常用的神经网络结构及误差反向回传更新训练的算法，最后介绍隐马尔科夫模型理论及其在声学建模中的应用。

第三章，基于注意力LSTM和多任务学习的远场语音识别声学建模。基于深度神经网络的声学模型一般简单地将上下文多帧特征拼接做为输入，因此忽略了对每帧声学特征所包含时间信息的利用。对于此问题，本章针对基于LSTM的混合声学建模框架提出一种注意力机制，使它可以对输入层中不同时刻特征向量加权，并利用多任务学习框架提升模型鲁棒性，对远场语音的识别做进一步优化。多任务学习框架中，将与远场语音同步录制的近场麦克风数据的声学特征做为网络的另一个输出，训练过程中声学状态预测任务与干净特征映射任务联合优化。该模型可以有效提升语音识别系统对远场语音的识别准确率。

第四章，基于空间特征补偿的的多通道声学建模。与单通道相比，麦克风

阵列的优势在于其可以提供空间区分度。因此，远场场景下语音识别系统一般采用麦克风阵列收集语音信号。传统的语音识别系统一般先采用增强算法得到增强后的语音信号，再利用增强信号进行声学建模。然而，这种分开优化方式引入了次优解问题。为解决此问题，一些联合优化方法被提出。但是这些方法大多数需要为增强算法引入若干层神经网络，导致声学模型参数量较大。本章提出一种基于空间特征补偿的多通道声学模型。此方法充分利用神经网络输入特征的灵活性，将空间信息的编码表示相位变换广义互相关（GCC-PHAT）做为辅助输入特征，帮助多通道声学模型的训练。该方法能够改善深度神经网络对多通道语音信号的声学建模能力。

第五章，基于教师学生迁移学习的远场语音识别声学建模。当语音信号在嘈杂的环境中录制时，不同声学单元之间的区分性更模糊，采用区分性过强的硬判决标注会增加远场声学模型训练的难度。因此，本章选用含有更丰富信息的软判决做为目标值进行模型的训练。我们利用与远场语音同步录制的近场语音数据获取可靠的软标签，通过教师学习迁移学习框架训练远场语音识别声学模型，优化后的目标值使远场语音识别声学模型的训练变得更轻松更容易。

第六章，总结与展望。对全文的内容进行总结，并对未来的研究方向进行展望。

第二章　自动语音识别系统

2.1　引言

自动语音识别技术经过半个多世纪的长足发展才取得如今的成绩。一直以来，语音识别技术都被认为是人与机器的沟通桥梁。然而，在过去的几十年里，语音并没有成为人机交互的重要形式，其主要原因是当时技术的落后使得语音识别技术在大多数真实场景中都不可用。20 世纪70年代，研究人员提出使用隐马尔科夫模型替代原本的模板匹配方法建模语音的时变性。此后，基于HMM 的混合声学建模框架逐步发展成熟，并在此基础上提出采用高斯混合模型描述HMM状态的特征分布。21世纪初期，基于GMM-HMM 的声学建模框架被广泛采用。最近几年，深度学习理论在机器学习领域兴起，其对语音识别同样产生了深远的影响。2010年，微软研究人员将深度神经网络取代高斯混合模型建模语音的发音多样性，并取得了巨大的成功。至此，语音识别技术才走向实际应用，真正意义上实现了人与机器语音沟通的愿望。

本章对基于隐马尔科夫模型声学建模的语音识别系统做介绍：首先描述语音识别的基本原理，对系统各个模块做简单介绍；然后对声学模型中用于计算声学后验概率的深度神经网络结构及其训练算法进行阐述；最后对隐马尔科夫模型及其基本算法做介绍。

2.2　语音识别的基本原理

自动语音识别的目标是将语音信号通过自动化系统转化为语言文本形式的信息。用公式表示为如下形式，即给定语音信号O，寻找一个产生概率最大的文本序列W，使得后验概率$P(W|O)$最大：

$$W = \arg\max_{W} P(W|O) \tag{2.1}$$

经过前人的探索总结，自动语音识别已经被划分为几个相对简单的模块逐级解决，如图2.1 所示。

在实际工程中，声学模型和语言模型作为机器学习中的分类系统，常常输出的是一个概率分布，所以整体上需要一个解码器在声学模型、语言模型及发

图 2.1 语音识别系统基本组成

Figure 2.1 Basic components of automatic speech recognition

音词典构建的解码空间中寻找当前语音信号最可能对应的一条语句文本。因此，包含解码器的典型自动语音识别系统如图2.2 所示。

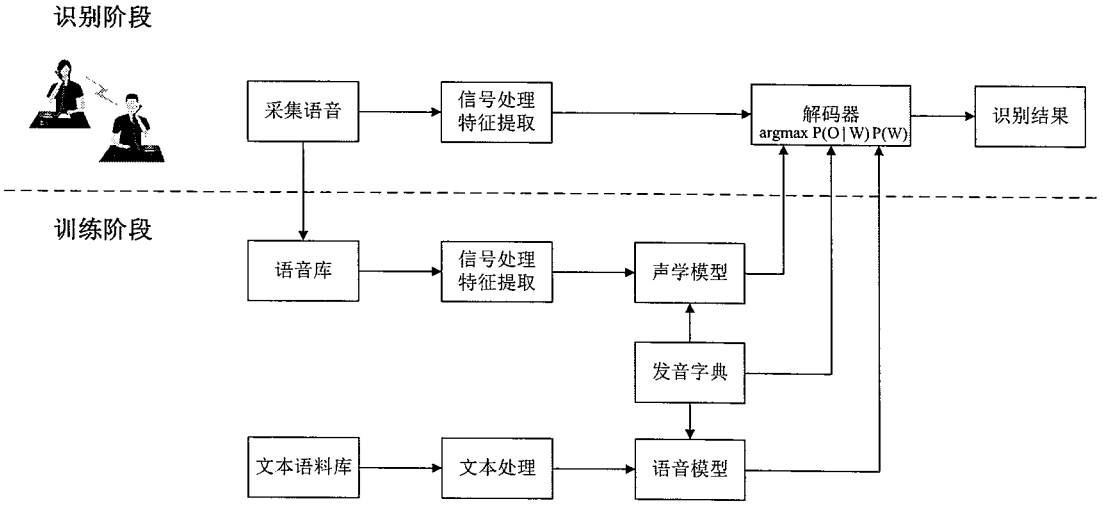


图 2.2 语音识别系统处理流程

Figure 2.2 Block diagram of a typical speech recognition system

可见，一个典型的ASR系统包含如下几个模块：

- 信号处理与特征提取：从语音信号中提取声学建模使用的特征。
- 声学模型：将声学特征对应到声学发音建模单元，并给出对应的观察概率。
- 发音词典：将发音建模单元与语言基本组成单元（如汉字、单词）建立联系。
- 语言模型：将字词的线性组合映射到语句，给出这一组合产生的概率。

- 解码器：将声学模型、语言模型的概率信息综合，在音素和字词组成的解码空间中寻找概率最高的语句输出。

若发音词典中文字序列W对应的声学建模单元序列为A，通过贝叶斯公式，公式2.1 可以转化为

$$\begin{aligned} W &= \arg\max_{W} P(W|O) \\ &= \arg\max_{W} \frac{P(O|W)P(W)}{P(O)} \\ &= \arg\max_{W} P(O|W)P(W) \\ &= \arg\max_{W} \sum_{A} P(O|A)P(A|W)P(W) \end{aligned} \tag{2.2}$$

对于确定的语音信号O，概率$P(O)$为固定值，因此公式2.2 中的第三行将此项省略。上式中，$P(O|A)$表征声学特征与声学建模单元的相似度，可由声学模型计算得到；发音词典产生语言基本组成单元字词与声学建模单元的对应关系$P(A|W)$；$P(W)$为文字序列W出现的概率，其由语言模型计算得出。下面将对语音识别系统中的各个模块展开介绍。

2.2.1　信号预处理及特征提取

语音识别系统首先需要将语音信号经过预处理和特征提取步骤得到适合声学建模的特征，声学特征应具有如下特点：（1）尽可能保留对建模目标有益的特征，如对声学发音建模单元具有鉴别性的共振频率；（2）尽可能去除对建模目标无益或有害的特征，如各种噪声；（3）考虑语音长时非平稳和短时平稳特性，只在短时使用适合平稳信号的时频分析方法，同时长时上下文相关信息也尽可能保留（如韵律）；（4）为了适合处理并且避免高维灾难，声学特征应该保持较低维度。

目前多种符合上述条件的声学特征已被提出，在实际中常用的有梅尔频率倒谱系数（Mel-Frequency Cepstral Coefficients，MFCC）[42]和感知线性预测系数（Perceptual Linear Prediction，PLP）[43]等。这里以常见的MFCC特征提取过程为例（如图2.3所示），详述语音信号预处理及特征提取步骤。语音信号的预处理通常包含预加重、分帧和加窗。特征提取包括快速傅里叶变换（Fast Fourier Transform, FFT）、梅尔滤波、取对数、离散余弦变换

（Discrete Cosine Transform，DCT）、高阶差分等步骤。为增强特征鲁棒性，通常会对提取的声学特征使用一些特征规整技术，如均值方差规整技术（Mean and Variance Normalization，MVN）、声道长度规整技术（Vocal Tract Length Normalization，VTLN）以及去相关和降维技术。以下将对上述步骤分别进行介绍。

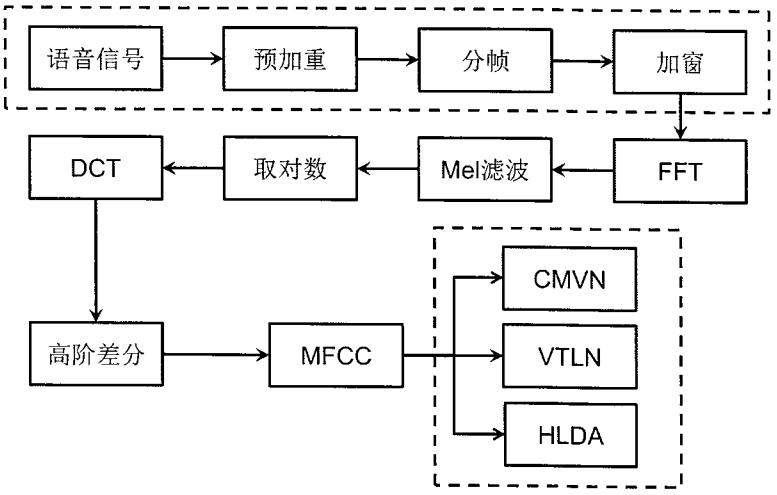


图 2.3 MFCC特征提取流程图

Figure 2.3 Block diagram of MFCC feature extraction

- 预加重：对语音的高频部分进行加重，使频谱整体更为平坦，以便进行频谱分析。语音信号由声带产生，经过喉、咽、口腔，最后由嘴唇发出。为了减少口唇对高频部分的抑制并增加语音的高频分辨率，常对高频部分进行加重。预加重的方式是通过一个一阶高通滤波器，其用差分方程实现的公式为：

$$y(n) = x(x) - ax(n-1) \tag{2.3}$$

上式中，预加重前后的时域采样信号分别为$x(n)$和$y(n)$，a为预加重系数。

- 分帧：语音信号本身是非平稳信号，在一个音素发音周期内又具有短时平稳信号的特性。因此，为方便分析以及得到平稳的信号估计特征，常用25ms的时域窗口截取语音采样后的数字信号做为一帧，以进行后续处理。在分帧操作中，帧与帧之间有部分信号重叠，使其保留每帧信号之间的上下文相关性。每次移动窗口的步长一般为10ms。

- 加窗：分帧后的截断信号在进行傅里叶变换时会产生吉布斯效应，因此需要在时域加窗对其进行抑制。特征处理中常用的窗函数有汉明窗等。
- 快速傅里叶变换：将时域信号进行快速傅里叶变换，得到频域频谱信号。
- 梅尔滤波：将频谱信号从频率尺度变换到贴近人耳感知变化的梅尔尺度。频域赫兹到梅尔刻度的变化公式：

$$\mathbf{Mel} = 1127 \cdot \ln(1 + \mathbf{Hz}/700) \tag{2.4}$$

在梅尔尺度上，将当前采样频率对应的奈奎斯特频率做为上限，约60hz做为下限，均匀划分出N+1 个频带，每两个频带分配一个三角滤波器，滤波器间重叠一个频带，得到如图2.4所示的滤波器组。 将频域信号经过

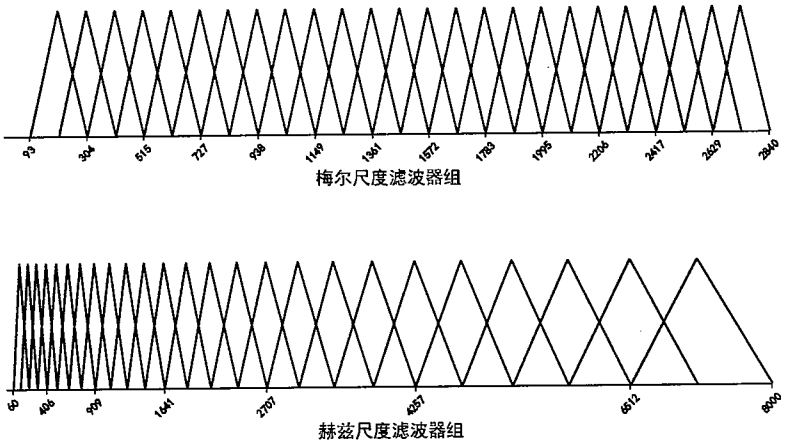


图 2.4 梅尔滤波器组

Figure 2.4 Schematic diagram of Mel filter bank

上述梅尔滤波器组处理，该滤波器组输出即为梅尔谱特征。

- 取对数：为压缩特征动态范围，并模拟听觉的非线性效应，对得到的梅尔谱取log对数。至此，我们已经得到一个深度学习常用的语音声学特征，即梅尔对数滤波器组特征（Mel-Scaled Log Filter-Bank，MS-LFB）。
- 离散余弦变换：语音频谱信号中，声道信息变化较慢，基因信息变化较快，利用快速傅里叶变换可以有效分离两者。因此，对MS-LFB 特征

经过DCT变换到倒谱域。经过DCT之后的倒谱特征即为标准的MFCC特征。

- 高阶差分：上一步得到的MFCC特征提取于当前窗内的短时语音信号。然而，语音信号的前后帧存在一定相关性。为弥补这个问题，一般在标准的MFCC特征后拼接一阶和二阶差分特征，其计算公式如下 [44]：

$$d(t) = \frac{\sum_{\theta=1}^{\Theta}(O(t-\theta) + O(t+\theta))}{2\sum_{\theta=1}^{\Theta}\theta^2} \tag{2.5}$$

其中，d(t)为差分后特征，Θ为差分窗长，二阶差分可在一阶结果上进行。

- 均值方差规整技术：信号在采集过程中经过信道传输，通常认为信道参数是稳定的，可在倒谱域减去均值来排除信道所带来的影响，此方法又称为倒谱均值减。均值规整会影响方差的估计，所以方差也需重新计算。实际工程应用中，一般对声学特征做均值方差规整，使其分布贴近标准正太分布。

- 声道长度规整技术：在“声源激励-声道传输”的模型中，声道差异是不同说话人语音差异的重要来源，例如声道长度的差异直接影响共振峰频率。VTLN的作用是消除不同说话人的声道差异对声学特征的影响，其通过对不同说话人数据做频谱变换，将共振峰位置调整到近似统一的位置。

- 去相关和降维技术：在采用高斯混合模型对隐马尔可夫状态的发生概率建模时，常出于计算复杂度考虑，采用对角协方差矩阵，所以特征各维之间需要满足线性无关。因此声学特征一般需要采用线性判别分析（Linear Discriminative Analysis，LDA） [45] 和异方差线性判别分析（Heteroscedastic Linear Discriminative Analysis，HLDA） [46]等技术进行去相关和降维处理。

2.2.2 声学模型

语音识别中声学模型目标是将声学特征对应到声学建模单元。声学模型主要解决两个问题，分别是语音信号的可变长特性和发音多样性。可变长特性可

采用隐马尔科夫模型建模。发音多样性由很多因素引起，如说话人的特性（性别、健康情况等）、说话风格、语速、方言差异等。一个良好的声学模型需要能够应付上述复杂的声学变化。因此，声学建模是影响语音识别系统性能的关键技术。声学建模过程可分为以下步骤：

- 选择声学建模单元：由不同模型的建模能力决定。对于基于HMM的声学模型来说，常使用音素或上下文相关的三音素（Triphone）做为声学建模单元。包含上下文信息的三音素充分考虑了协同发音的问题，因此它对发音的建模也更精确。然而，使用三音素会带来训练数据稀疏的问题。为平衡声学建模的准确性和鲁棒性，一般会对三音素的声学状态通过决策树进行聚类，使具有相似发音的建模单元可以共享训练数据。另一方面，对最近受到广泛关注的基于端到端语音识别系统 [33–35]而言，音素、音节、字、词都可用来做为建模单元。由于本文主要围绕基于HMM的声学建模技术进行研究，这里对其不做阐述。

- 确定声学输入特征：声学特征的选择要与声学模型相对应。为减少模型的参数量，GMM通常要求声学特征的每个维度统计独立。因此，MFCC与PLP为GMM-HMM声学建模中常用的特征。神经网络对输入特征没有此类限制，可选用更原始的MS-LFB特征做为DNN 的输入。

- 选择及构建声学模型：目前语音识别系统中常用的仍是基于HMM框架的声学模型。由于GMM具有拟合任意复杂分布的能力，在过去相当长的一段时间里，语音识别系统使用GMM建模语音信号的发音多样性。随着深度学习理论的兴起，基于DNN-HMM 的声学建模方法显著提升了语音识别系统性能，其取代了传统GMM-HMM框架成为主流声学建模方法。

- 训练声学模型：首先，使用最大似然估计准则通过期望最大化算法（Expectation Maximization，EM）和前后向算法实现GMM-HMM声学模型的参数训练。然后，利用语音的文本标注和已训练完成的GMM-HMM模型通过维特比算法强制对齐得到声学特征的状态级标注。最后，使用声学特征及其状态级标注利用误差反向回传算法更新训练神经网络参数，即可得到用于语音识别的声学模型。

由于声学建模是本文的主要研究内容，在本章随后的小节中对DNN和HMM及其在语音识别应用中涉及的算法做详细介绍。

2.2.3 发音词典

发音词典构建了语音信息向语言信息的转变，其建立了声学建模单元与语言基本组成单元（如汉字、单词）之间的联系，并通过音形联系将声学信息传递到语言学。从整个语音识别系统的角度来看，发音字典将声学模型和语言模型有效的连接起来，其完成了声学建模单元与语言模型建模单元之间的映射关系。这种映射关系可以是简单的一一对应，也可以是一对多（如同音字）或者多对一（如多音字）。

发音词典在声学层面规定了发音方式，因此，由说话人差异、轻度方言引起发音变异现象需要在发音词典上做相应处理。一个解决方案是建模其所有发音并对每种发音赋予相应的发音概率，即构建一个发音模型。另外，在涉及双语种多语种混合识别的应用中，发音词典也需要配合声学建模单元的选择，构建一套适合多语种混合识别的发音词典，如使用中英混合音素集构建中英混合ASR 系统。

以下为中文发音字典的一个实例：

挨aa ai1

挨aa ai2

挨打aa ai2 d a3

挨着aa ai1 zh e4

2.2.4 语言模型

语言模型的作用是将字词的线性组合映射到语句，并给出这一组合的发生概率。目前常用的是基于N元文法（N-Gram）的统计语言模型。N元文法是一种基于n-1阶马尔可夫链的概率序列建模模型，通过n个词语联合出现的概率来评估序列的发生概率。因此，整个词语序列W的概率为

$$\begin{aligned} P(W) =& P(w_1, w_2 \dots w_l \dots w_L) \\ =& \prod_{l=1}^{L} P(w_l | w_{l-N+1}, w_{l-N+2} \dots w_{l-1}) \end{aligned} \tag{2.6}$$

上式中的条件概率值是从大量的训练文本数据中统计得到的，实际中常用$N = 2, 3, 4$的N元文法语言模型。

由于训练语料的稀疏性，N元文法语言模型不可能覆盖所有的词语组合，导致概率值$P(w_l|w_{l-N+1}, w_{l-N+2}\dots w_{l-1})$不一定存在。此时，需要使用回退概率 [47]计算语句的概率

$$P(w_l|w_{l-N+1}, w_{l-N+2}\dots w_{l-1}) = \alpha_{w_{l-N+1}, w_{l-N+2}\dots w_{l-1}} P(w_l|w_{l-N+2}\dots w_{l-1}) \tag{2.7}$$

上式中，$\alpha_{w_{l-N+1}, w_{l-N+2}\dots w_{l-1}}$为回退系数。另一方面，N元文法语言模型还需要考虑集外词（out-of-vocabulary，OOV）的影响。为了解决此问题，语言模型中引入了平滑估计算法，如Kneser-Ney平滑算法 [48] 等。

随着深度学习的发展，基于深度神经网络的语言模型 [49]受到了广泛的关注。但在实际工程中，N-gram 模型得益于计算的高实时性，在线解码的ASR系统中仍然大多使用N-gram语言模型，而在后处理阶段即二遍解码时，采用神经网络语言模型进行重估。

2.2.5　解码器

解码器将声学模型和语言模型的概率信息综合，在音素和字词组成的解码空间中，通过动态规划算法（维特比算法）寻找一条路径，使该条路径上的观察概率最高。在解码过程中，声学状态发生概率需要根据当前特征实时进行计算，N元文法的语言模型可以部分或完全嵌入到解码空间中。解码器根据语言模型概率信息在解码构图中的嵌入情况，可以分为：

- 动态解码器：动态的查询当前路径对应的语言模型概率。解码速度相对较慢，其构成的解码网络占用空间小。
- 静态解码器：一般使用加权有限网络转换机（Weighted Finite State Transducers，WFST）将语言模型中的概率信息完全嵌入到解码空间中。解码速度相对较快，其构成的解码网络占用空间大。

为衡量语音识别系统的性能，解码器输出文字序列的解码结果后需对其准确率进行评估，主要评估指标有字错误率（Character Error Rate，CER）和词错误率（Word Error Rate，WER）。通过动态规划对齐识别结果与参考答案，可以得到以字或词为单元的替代错误、删除错误和插入错误的统计结果。错误率的计算公式即为

$$\text{错误率} = \frac{\text{替代错误数} + \text{删除错误数} + \text{插入错误数}}{\text{正确答案文字数}} \tag{2.8}$$

本文的实验均是在英语数据集上进行的，因此实验中采用WER做为评价模型性能的标准。

2.3 深度神经网络

在人类对自然的探索过程中，其间的发明创造常常来源于对大自然的观察与模仿。在上世纪50 年代，一种与人脑神经元类似具有二元线性分类功能的“感知器”被提出 [50]。它做为最早的人工神经网络（Artificial Neural Network，ANN），包含了沿用至今的神经元模型。

2.3.1 神经元模型

神经元模型模仿人类神经元机制，其接收来自神经突触的电流脉冲，并在处理后输出新的电流脉冲。图2.5为神经元结构示意图。

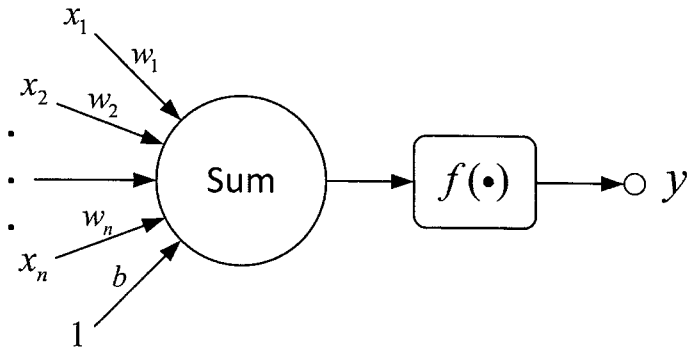


图 2.5 神经元模型

Figure 2.5 Diagram of artificial neuronal model

如图所示，输入向量为$x = [x_1, x_2 \cdots x_n, 1]$，每条输入路径对应一个权重参数$w_i$，其构成了权重向量$w = [w_1, w_2 \cdots w_n, b]$，偏置$b$ 可理解为神经元的常态活跃度。两者内积后再经过激活函数的变换就得到一个标量输出y，用公式表示上述过程为：

$$y = f(\sum_{i=1}^{n} wx^{\mathrm{T}}) = f(a) \tag{2.9}$$

其中，$f(\cdot)$为符号函数：

$$f(x)=\begin{cases}1, & \text{if } x>0\\ 0, & \text{otherwise}\end{cases} \tag{2.10}$$

单个神经元具备线性二元分类能力，其构造了一个超平面$w\cdot x^{\mathrm{T}}=0$，将来自n 维向量空间的向量划分为平面上和平面下两类。由于感知器分类能力局限在线性可分，人们逐渐对其失去信心。随着时间推移，人们意识到可以将神经元模型并联和串联组成多层感知网络（Multilayer Perceptron，MLP），图2.6为包含两个隐含层MLP结构示意图，并成功突破了工程优化的算法难题，找到了优化多层感知网络的优化算法–误差反向传播算法（Error Back-Propagation, EBP），在80 年代又开启了神经网络的新篇章。

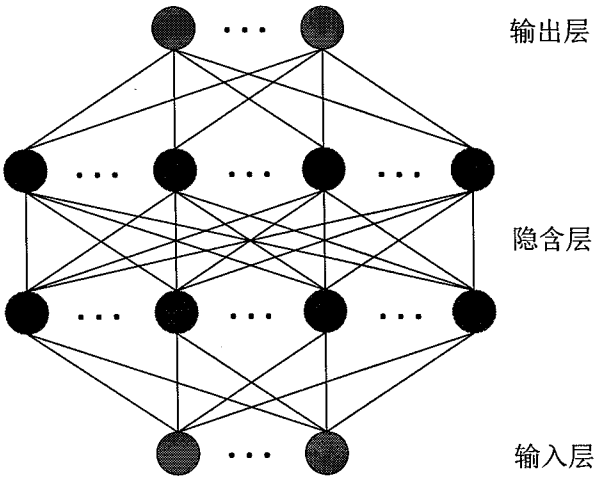


图 2.6 多层感知器结构示意图

Figure 2.6 Schematic diagram of MLP

2.3.2 常见的神经网络结构

本小节对语音识别中常用的神经网络结构做简单介绍。

- 前馈神经网络（Feed-Forward Neural Network，FFNN）：前馈神经网络的结构与图2.6中所示的多层感知器相同。网络中每个隐层的输出做为下一隐层的输入，其用公式表示为：

$$\boldsymbol{h}^{l}=f(\boldsymbol{W}^{l}\boldsymbol{h}^{l-1}+\boldsymbol{b}^{l}) \tag{2.11}$$

其中，$\boldsymbol{h}^l$ 为第l个隐层的输出，$\boldsymbol{W}^l$ 和$\boldsymbol{b}^l$ 分别为第l个隐层的权重矩阵和偏置向量，$f(\cdot)$ 为隐层的非线性激活函数。常用的非线性激活函数包含：

1）Sigmoid函数：

$$f(x) = \frac{1}{1+e^{-x}} \tag{2.12}$$

2）Tanh函数

$$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2.13}$$

3）ReLU函数

$$f(x) = \max(0, x) \tag{2.14}$$

当神经网络用于分类任务时，输出层的每个节点代表一个类别$i \in \{1, 2, \cdots, C\}$，其中$C$为类别个数。网络的输出对应输入数据属于每一类的概率。若输出向量为$\boldsymbol{v}$，则其需满足$v_i \geq 0$ 和$\sum_{i=1}^{C} v_i = 1$。因此，输出层通常采用softmax函数进行归一化，其计算公式为

$$\text{softmax}(v_i) = \frac{e^{v_i}}{\sum_{j=1}^{C} e^{v_j}} \tag{2.15}$$

- 时延神经网络（Time-Delay Neural Network，TDNN）：与最简单的前馈神经网络相比，时延神经网络能够看到更长的上下文信息。图2.7为TDNN的网络结构示意图，图中t 时刻每个隐层的输入为上一隐层在$t-1$，t 和$t+1$ 时刻输出的拼接向量，经过三个隐层的扩展之后，网络的输出层在t时刻的输出依赖于$t-4$到$t+4$时刻的输入。因此，TDNN 具有一定的长时建模能力。图2.7为举例说明的TDNN结构。实际中，各个层的时间扩展可完全不同，如Peddinti等人提出采样扩展的TDNN结构，其在声学建模中取得了优于简单前馈神经网络的性能 [51]。

- 卷积神经网络（CNN）：CNN在计算机视觉领域取得了前所未有的成功，因此，有些研究将其应用于语音识别 [3]。与普通神经网络的结构不同，CNN由卷积层和池化层构成。图2.8为网络结构示意图。网络的输入是二维矩阵。针对语音识别任务，一维表示时域，另一维表示频域。例如，对于上下文共扩展11帧的40维MS-LFB特征，CNN的输入为11*40的矩阵。卷积层利用矩阵卷积核在输入矩阵上分别沿两个维度平移，进而对输入

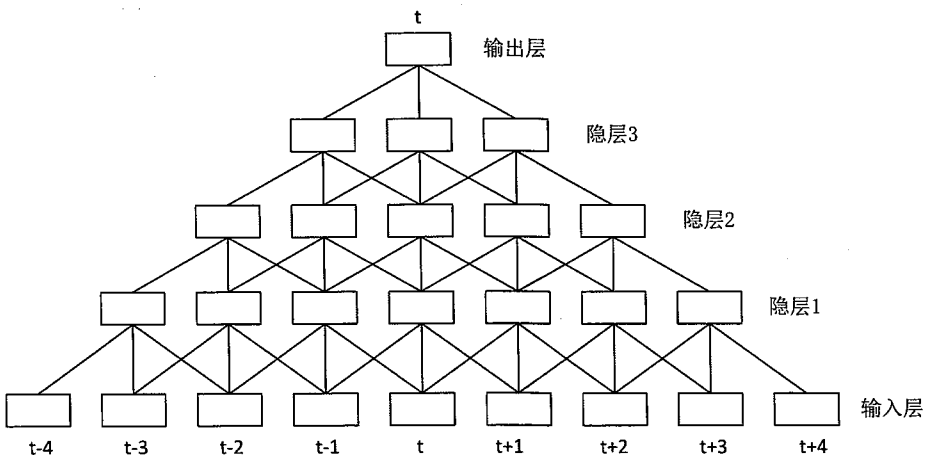


图 2.7 时延神经网络结构示意图

Figure 2.7 Schematic diagram of time-delay neural network

特征做卷积运算。随后，对卷积后的值经过非线性函数即得到一个特征平面。一个特征平面由一个卷积核计算得到，一个卷积层通常包括若干个特征平面。为压缩数据和减少参数量，在两个连续的卷积层间存在一个池化层，其对卷积层的输出进行降采样。依据池化的方法不同，可分为最大池化和平均池化。

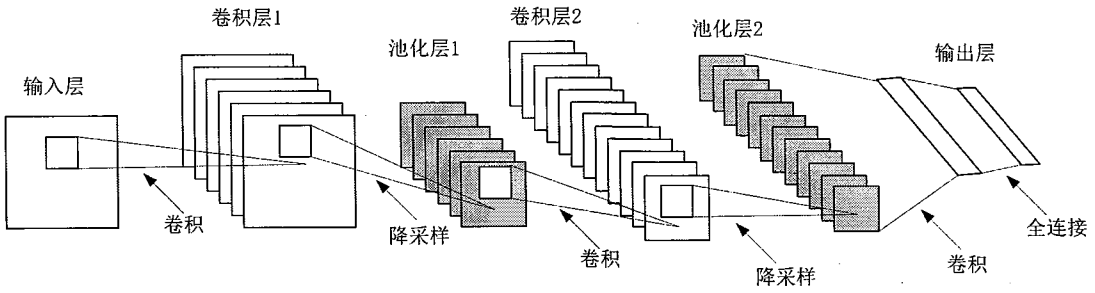


图 2.8 卷积神经网络结构示意图

Figure 2.8 Schematic diagram of convolutional neural network

- 递归神经网络（RNN）：与简单的前馈神经网络相比，递归神经网络在隐层之间加入反馈连接，使其可以考虑历史信息。然而，实际使用中RNN有严重的梯度消失问题，距离当前时刻越远的信息对当前的输出影响越小，因此不能对长时依赖性有效地建模。为了解决此问

题，许多基于RNN 的变种结构被提出，如GRU（Gated Recurrent Unit）和LSTM。由于本文实验的基线模型采用LSTM 结构，LSTM相关内容在第三章3.2节做具体介绍。

2.3.3 误差反向回传算法

误差反向传播算法的面世颇为曲折，经过多人的贡献 [52–55]，最终被广泛应用于寻找多层神经网络最优解中。反向传播算法则是寻找系统损失函数的极小值，其核心是梯度下降法，即每次朝误差曲面的梯度反方向前进一步（用此梯度更新权重参数值），经过多轮迭代后达到一个极小值。

误差反向传播算法主要由两部分组成，即前向计算过程和误差的反向传播过程。为方便对算法进行描述，我们定义如下：

- 训练数据和标注为$S = \{(x_n, t_n) \mid n = 1, 2, \cdots, N\}$。向量$x_n$ 为神经网络的第n个训练样本，t_n为x_n对应的标注向量；
- 神经网络参数$\lambda = (W, b, L)$。$W = (W^0, W^1, \cdots, W^{L-1})$ 和$b = (b^0, b^1, \cdots, b^{L-1})$为$L$ 层神经网络的连接权重和偏置；
- f_l为第l层非线性激活函数，x_n^l为第n个训练样本在第l层的激活值，y_n^l为网络第l 层的非线性函数输出值；
- J为神经网络训练准则，网络参数更新的学习率为η，θ为收敛阈值。

则误差反向传播算法可描述为算法1。

依据误差反向传播算法每次迭代中用于计算梯度和更新权重使用的样本情况，可得到以下三种学习模式：（1）批量梯度下降法：每次迭代使用训练全集计算更新误差，模型收敛较慢。（2）随机梯度下降法（Stochastic Gradient Descent，SGD）：每次迭代随机选择一个样本计算更新误差，计算量较大。（3）Mini-batch 梯度下降法(Mini-batch Gradient Descent)：每次迭代从全集中随机选择若干样本，使用当前mini-batch内的数据计算误差。此方法为上述两种更新方式的平衡。在工程实现上，mini-batch 的大小适合于当前硬件加速的矩阵运算（如GPU）。但是，由于引入mini-batch 超参数，工程上常需要结合经验对其调节。

Algorithm 1 误差反向传播算法（Error Back-Propagation，EBP）

1. 随机初始化：W 和 b

While $(n \leq N$ and $||J(\lambda)|| \leq \theta)$ **do**

$$y_n^0 = x_n;$$

2. 前向计算

for $l = 1$ to L **do**

$$\begin{aligned} x_n^l &= W_{n-1}^{l-1} y_n^{l-1} + b_{n-1}^l \\ y_n^l &= f_l(x_n^l) \end{aligned}$$

end for

$dy_n^L = \partial J(y_n^L, t_n|\lambda)/\partial y_n^L$

3. 误差反向回传

for $l = L$ down to 1 **do**

$$\begin{aligned} dx_n^l &= dy_n^l f_l'(x_n^l) \\ dW_n^{l-1} &= dx_n^l (y_n^l - 1)^T, \quad db_n^{l-1} = dx_n^l \\ \triangle W_n^{l-1} &= -\eta dW_n^{l-1}, \quad \triangle b_n^{l-1} = -\eta db_n^{l-1} \\ W_n^{l-1} &= W_{n-1}^{l-1} + \triangle W_n^{l-1}, \quad b_n^{l-1} = b_{n-1}^{l-1} + \triangle b_n^{l-1} \end{aligned}$$

end for

$n = n + 1$

end While

2.4　基于HMM的声学模型

2.4.1　隐马尔科夫模型

在一个随机过程中，若当前时刻状态已知，其下一时刻的状态仅与当前状态有关，与过去时刻的状态无关，这种特性被称为马尔科夫性，此随机过程被称为马尔科夫过程。其用数学公式表示为如下：假设$\{X(t),\ t \in T\}$为一随机过程，其状态空间为Ω。若对于任意的$t_1 < t_2 < \cdots t_n < t$ 和$\{x_1, x_2, \cdots, x_n, x\} \in$

Ω，在已知变量$X(t_1)=x_1,\cdots,X(t_n)=x_n$下的随意变量$X(t)$的条件概率满足

$$P\{X(t)=x|X(t_n)=x_n,\cdots,X(t_1)=x_1\}=P\{X(t)=x|X(t_n)=x_n\} \quad (2.16)$$

若随机变量的状态空间Ω是有限的离散状态集合，这种离散的马尔科夫过程被称为马尔科夫链。一个马尔科夫链可由转移概率和初始状态分布概率表示：

转移概率：

$$a_{ij}=P(q_t=s_j|q_{t-1}=s_i),\ \ 0\le a_{ij}\le 1;\ \ \sum_{j=1}^{N}a_{ij}=1 \quad (2.17)$$

初始状态分布概率：

$$\pi_i=P(q_1=s_i),\ \ 0\le \pi_i\le 1;\ \ \sum_{i=1}^{N}\pi_i=1 \quad (2.18)$$

上式中N为状态空间Ω中包含的状态数目。

上述马尔科夫链中每个状态的取值是唯一的，其不能对语音信号的发音多样性进行建模。为解决此问题，需要在马尔科夫链模型基础上引入每个状态产生观察向量的概率分布，这样的随机过程被称为隐马尔科夫过程。

隐马尔科夫模型可用如下五个元素描述 [56]：

1. 有限状态集合：$\Omega=\{s_i\}$，$i=1,2,3,\cdots,N$;

2. 状态间转移概率矩阵：$A=[a_{ij}]$，$i,j=1,2,3,\cdots,N$，其满足

$$a_{ij}=P(q_t=s_j|q_{t-1}=s_i),\ \ 0\le a_{ij}\le 1,\ \ \sum_{j=1}^{N}a_{ij}=1 \quad (2.19)$$

3. 观察序列：$X=\{x_t\}$;

4. 各状态上观察概率：$B=\{b_i(x)\}$，其满足

$$b_i(x)\ge 0,\ \ \int b_i(x)dx=1 \quad (2.20)$$

5. 各状态初始概率：$\Pi=\{\pi_i\}$，其满足

$$\pi_i=P(q_1=s_i),\ \ 0\le \pi_i\le 1,\ \ \sum_{i=1}^{N}\pi_i=1 \quad (2.21)$$

其中特定状态s_i的观察概率$B = \{b_i(x)\}$来自另一模型，如高斯混合模型（GMM），这样就组成了GMM-HMM模型，这一模型成为过去20年ASR系统的主流声学模型。

实际应用中，HMM的建模需要解决三个基本问题：

1. 评估问题：已知HMM的模型参数$\lambda = (\pi, A, B)$ 和观察序列O，快速计算观察序列的概率得分$P(O|\lambda)$；
2. 解码（对齐）问题：已知HMM的模型参数$\lambda = (\pi, A, B)$和观察序列O，寻找使观察概率最大的状态序列S。
3. 训练问题：给定观察序列O，如何调整HMM参数使观察概率$P(O|\lambda)$最大。

HMM的理论发展过程中以上三个问题已经得到很好的解决，下面对这三种基本算法进行介绍。

2.4.2　隐马尔科夫模型的基本算法

2.4.2.1　评估问题–前后向算法

前后向（Forward-Backward）算法属于动态规划算法，其利用了HMM中马尔可夫性质，即下一时刻的状态只与当前状态有关，高效地解决HMM评估问题。

首先定义两个统计量，其中$O = \{o_1, o_2, \cdots, o_T\}$为0到$T$时刻的观察序列：

- 前向概率$\alpha_i(t)$：表示系统在t时刻处于状态i时，并且t时刻及之前时刻的观察序列为$(o_1, o_2, \ldots, o_t)$ 的概率；

$$\alpha_i(t) = P(o_1, o_2, \ldots, o_t, s_t = i) \tag{2.22}$$

- 后向概率$\beta_i(t)$：表示系统在t时刻处于状态i，并且t时刻之后的观察序列为$(o_{t+1}, o_{t+2}, \ldots, o_T)$ 的概率。

$$\beta_i(t) = P(o_{t+1}, o_{t+2}, \ldots, o_T | s_t = i) \tag{2.23}$$

因此，模型产生观察序列O的概率为，在终点时刻所有状态上的前向概率之和，或在起点时刻所有状态上的后向概率之和，即：

$$P(O)=\sum_{i=1}^{N}\alpha_i(T)=\sum_{i=1}^{N}\beta_i(0) \tag{2.24}$$

前后向算法利用动态规划保存中间序列的概率值，以此高效解决评估问题。具体实施过程如算法2所示。

Algorithm 2 前后向算法

前向计算：

初始化：$\alpha_i(1)=\pi_i b_i(o_1), 1\leq i\leq N$

递归计算：$\alpha_i(t)=\{\sum_{j=1}^{N}a_{ji}a_j(t-1)\}b_i(o_t), 1\leq i\leq N, 2\leq t\leq T$

后向计算：

初始化：$\beta_i(T)=1, 1\leq i\leq N$

递归计算：$\beta_i(t)=\{\sum_{j=1}^{N}a_{ji}b_j(o_{t+1})\}\beta_i(t+1), 1\leq i\leq N, 1\leq t\leq T-1$

2.4.2.2 解码问题–维特比算法

上述的评估问题计算观察序列在所有状态序列上产生的概率和，使用维特比（Viterbi）算法 [57]可以进一步得到所有状态序列中对应观察概率最大的状态序列，此问题被称为解码问题，用数学公式描述如下：已知观察序列$O=(o_1,\ o_2,\ o_3,\ \cdots,\ o_T)$，寻找最优状态序列$\hat{S}=(s_1,\ s_2,\ s_3,\ \ldots,\ s_T)$，使得

$$\hat{S}=\arg\max_{S}P(O|S) \tag{2.25}$$

如同经典的有向无环图中寻找最短路径的算法，Viterbi算法也利用了动态规划思想，其能够高效实时地解决HMM解码问题。

如果从可能的N^T个状态序列中逐个计算寻找使$P(O|S)$最大的状态序列S，这样做的计算复杂度太高。Viterbi算法通过一个近似假设来减小搜索空间，即t时刻搜索的最优路径只与$t-1$时刻的最优路径有关，与未来路径无关。此

假设将全局最优化搜索问题转换为局部最优化搜索问题。基于上述假设，全局最优状态序列的搜索问题可利用动态规划记录每个时刻的最优状态序列来近似解决。

首先定义两个变量，即当前时刻最优路径$V_t(i)$和用于回溯全局最优路径的变量$\Psi_t(i)$：

$$V_t(i) = \max_{j \in S}\{[V_{t-1}(j)a_{ji}]b_i(x_t)\} \tag{2.26}$$

$$\Psi_t(i) = \arg\max_{j \in S}\{V_{t-1}(j)a_{ji}\} \tag{2.27}$$

Viterbi算法如算法3所示。

Algorithm 3 维特比算法

1：**初始化**

$V_1(i) = \pi_i b_i(o_1), \Psi_1(i) = \phi \qquad 1 \leq i \leq N$

2：**递归计算**

$V_t(i) = \max_{j \in S}\{[V_{t-1}(j)a_{ji}]b_i(o_t)\}$

$\Psi_t(i) = \arg\max_{j \in S}\{V_{t-1}(j)a_{ji}\} \qquad 1 \leq i \leq N, 2 \leq t \leq T$

3:：**计算最优状态序列概率**

$p(O, \hat{S}|\lambda) = \max_{i \in S} V_T(i)$

4：**回溯最优状态序列**

$\hat{s} = \Psi_{(t+1)}(\hat{s}_{t+1}) \qquad 1 \leq t \leq T$

$\hat{S} = (\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_T)$

2.4.2.3　训练问题–期望最大化算法

由于HMM只能观察到特征序列，其对应的状态序列并不可见，导致HMM的参数估计尚无一个闭式解。期望最大化（Expectation Maximization，EM）算法 [58]可以用来解决此类不完全数据的最大似然参数估计。将EM算法应用在HMM中，隐含的微观状态序列就是不完全数据，其通过求解不完全数据的概率期望，并将这一期望固定为对不完全数据的猜测，从而将不完全数据

问题转为完全数据问题。将EM算法应用于HMM模型参数估计的算法又被称为Baum-Welch算法 [59]。

Jenson不等式给出了积分的凸函数值和凸函数的积分值间的关系，反映到离散概率理论就是，对于凸函数φ，当$a_i > 0, \sum_i a_i = 1$，存在下列不等式：

$$\varphi(\sum_i a_i x_i) \leq \sum_i a_i \varphi(x_i) \tag{2.28}$$

该不等式的左边为一个积分的凸函数值，右边为凸函数的积分。由此公式可推导出，一个凸函数，其变量期望值的函数值，总是小于该函数值的期望；反之，一个凹函数，其变量期望值的函数值，总是大于该函数值的期望。

在HMM的训练问题中，包含两个变量空间，一个是所有隐含的状态序列S，一个是需要求解的模型参数λ，(S, λ) 组成的变量空间过于巨大，然而在固定当下λ参数下，通过求目标函数在所有隐含状态序列S上的期望，可以得到对目标函数当前最大值的猜测，该猜测基于Jenson 不等式推导出，基于这一猜测我们便能迭代更新模型参数。因此，基于EM算法的HMM模型训练流程，就是基于最大似然估计准则（Maximum Likelihood Estimation，MLE），使用当前模型参数λ 计算所有隐含状态序列上的期望观察概率，使该期望概率函数最大的参数作为新模型参数$\hat{\lambda}$。为方便计算，此处引入log域，得到如下公式：

$$\begin{aligned}\hat{\lambda} &= \arg\max_{\lambda} \log P(O|\lambda) \\ &= \arg\max_{\lambda}(\log \frac{P(O,S|\lambda)}{P(S|O,\lambda)}) \\ &= \arg\max_{\lambda}(\log P(O,S|\lambda) - \log(P(S|O,\lambda)))\end{aligned} \tag{2.29}$$

根据EM算法的思想，首先固定O和λ，对公式2.29两边在S上求期望，得到：

$$\mathrm{E}_S(\log P(O|\lambda)) = \sum_S P(S|O,\lambda) \log P(O|\lambda) = \log P(O|\lambda) \tag{2.30}$$

$$\begin{aligned}&\mathrm{E}_S(\log P(O,S|\lambda) - \log(P(S|O,\lambda)) \\ &= \sum_S P(S|O,\lambda) \log P(O,S|\lambda) - \sum_S P(S|O,\lambda) \log(P(S|O,\lambda))\end{aligned} \tag{2.31}$$

式2.31中后一项根据jenson不等式，令φ为凹函数log，则可得出其为参数λ单调减函数：

$$\begin{aligned}
&\sum_S P(S|O,\lambda)\log(P(S|O,\hat{\lambda})) - \sum_S P(S|O,\lambda)\log(P(S|O,\lambda)) \\
&= \sum_S P(S|O,\lambda)\log(\frac{P(S|O,\hat{\lambda})}{P(S|O,\lambda)}) \\
&\leq \log\sum_S P(S|O,\lambda) = 0
\end{aligned} \tag{2.32}$$

因此，为了最大化目标函数，目标函数公式2.29可变为：

$$\begin{aligned}
\hat{\lambda} &= \arg\max_{\lambda} \log P(O|\lambda) \\
&= \arg\max_{\lambda} \sum_S P(S|O,\lambda)\log P(O,S|\lambda)
\end{aligned} \tag{2.33}$$

声学建模中，HMM状态的特征概率分布通常使用GMM建模。因此，EM算法期望最大化步骤中，带入GMM概率密度公式，令其导数为0，即可推导出HMM-GMM声学模型参数的训练更新公式，其具体数学推导过程，请参见文献 [60]。

2.4.3　HMM在语音识别声学建模中的应用

上世纪70年代，研究人员提出使用隐马尔科夫模型替代原本的模板匹配方法建模语音信号的时变性。基于HMM的声学建模是在以下三点假设的基础上实现的：

1. 短时平稳性假设：语音信号本身为非平稳信号，一般在声学建模中将其分帧成时间较短的语音片段，在该段时间内的信号可以被看作平稳信号，这样可以使用高斯混合模型等静态模型来描述其概率分布；
2. 马尔科夫假设：语音信号下一帧跳转的状态仅与当前时刻的状态有关；
3. 条件独立假设：当前时刻声学特征出现的概率仅与当前属于的HMM声学状态有关，与上下文的声学特征和声学状态无关。

在语音识别系统声学建模中，通常使用HMM描述声学建模单元的状态跳转，这种跳转是由于声道运动引起的。声学模型一般采用音素做为建模单

元，每个音素的建模采用包含三个状态的HMM模型，其分别表示音素发音的起始、稳定和消退三个过程。图2.9为以音素为建模单元的HMM拓扑结构图，两个HMM通过分别建模“n”和“i3”的发音组成了对音节“n i3”的建模，HMM的每个状态均有自身跳转和跳转到相邻状态两种状态转移方式。

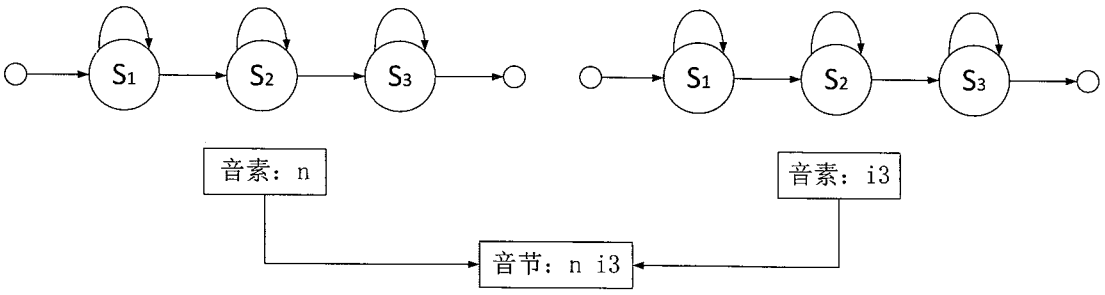


图 2.9 音素的HMM拓扑结构

Figure 2.9 Typical HMM topology for phoneme

语音信号中一般存在许多静音和短时停顿，为更好地对上述两种情况进行建模，通常会对其进行特殊处理。如图2.10 所示，静音采用中间状态可跨越的HMM拓扑结构，短时停顿采用可跨越的单状态HMM进行建模。

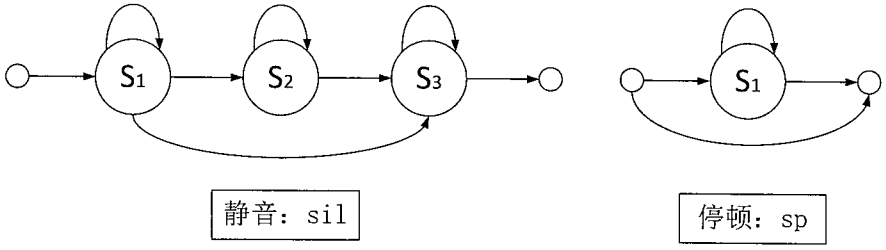


图 2.10 静音和短时停顿的HMM拓扑结构

Figure 2.10 Typical HMM topology for silence and shot pause

由于语音信号当前的发音会受到其前后文临近音的影响，为提升建模准确率，常使用包含上下文信息的三音素建模，其形式如“a-b+c”。然而，这种建模方式会导致建模单元数目过多。另一方面，很多三音素在训练数据中可能不存在，会带来训练数据稀疏的问题。为平衡声学建模的准确性和鲁棒性，一般会对三音素的声学状态通过决策树进行聚类，使具有相似发音的建模单元可以共享训练数据。

2.4.3.1　GMM-HMM声学模型

由于高斯混合模型具有模拟任意复杂分布的能力，传统的语音识别系统采用隐马尔科夫模型表达语音信号的时变特性，高斯混合模型建模语音信号的发音多样性。因此，GMM用于对2.4.1节中介绍的各状态上的观察概率$B=\{b_i(x)\}$进行建模，即

$$b_i(x)=\sum_{m=1}^{M}\frac{c_{im}}{\sqrt{(2\pi)^D|\sum_{im}|}}\exp\{-\frac{1}{2}(x-\mu_{im})^T\sum_{im}^{-1}(x-\mu_{im})\} \tag{2.34}$$

上式中，D为声学特征的维度，M为表征声学状态分布的高斯混合模型中的高斯数目，μ_{im}和$\sum_{im}$为建模声学状态i的第m个混合高斯的均值和协方差矩阵，c_{im}为各个高斯的混合权重，其满足

$$\sum_{m=1}^{M}c_{im}=1 \tag{2.35}$$

为减少模型参数量，将协方差矩阵$\sum_{im}$转化成对角矩阵，GMM-HMM声学模型通常要求声学特征的每个维度统计独立，因此MFCC与PLP为GMM-HMM声学建模常用特征。

在HMM-GMM声学模型训练中，有如下几个常用的训练准则：

- 最大似然准则：寻找使观察概率最大的模型参数。
- 基于最小贝叶斯分类错误的准则：
  - 最小句错误率准则 [61]：由最大互信息准则（Maximum Mutual Information，MMI）推导而来，可用于整句错误率优化。
  - 最小字错误率准则（Minimum Charecter Error，MCE）：用于字错误率优化，使用情况较少。
  - 最小音素错误率准则（Minimum Phone Error，MPE） [62]：代表性的鉴别性训练准则，与最大似然准则相比，提升了模型的鲁棒性。

在语音识别系统中，声学模型为基于数据驱动的机器学习模型，因此存在训练和测试数据分布不一致问题。为提升模型在实际应用场景的性能，通常使用与应用场景数据相近的自适应数据调整模型参数，从而对声学模型进行自适

应学习。一般可用的自适应数据较少，自适应技术仅学习少量新参数来调整旧的模型，当数据更多时，可以考虑重新训练声学模型。GMM-HMM的主要建模参数为GMM的均值方差，常见的有如下两种自适应方法：

- 最大似然线性回归（Maximum Likelihood Linear Regression，MLLR）[63, 64]：使用少量自适应数据，结合人工标注或解码器产生标注，对GMM的均值或方差学习一个线性变换矩阵，训练准则为最大化自适应数据上的观察概率。

- 最大后验概率（Maximum A Posteriori，MAP）[65]：将先验知识和自适应数据中得到的知识以线性插值的方式结合，并将结果作为自适应后的参数值。此种方法在自适应数据量较大时更有效。

由于上述自适应技术是针对GMM提出的，在神经网络声学建模中性能提升幅度较小，当前已经很少用到上述自适应技术，这里就不做过多阐述。下一节对目前主流的声学建模方式DNN-HMM进行介绍。

2.4.3.2 DNN-HMM声学模型

尽管GMM具有拟合任意复杂分布的能力，但它也有一个严重的缺陷，即对非线性数据建模的效率低下。因此，相关研究人员提出采用人工神经网络代替GMM，建模HMM状态后验概率。但是由于当时计算能力有限，很难训练两层以上的神经网络模型，所以其带来的性能改善非常微弱。21世纪以来，机器学习算法和计算机硬件的发展使得训练多隐层的神经网络成为可能。实践已经表明，DNN 在各种大型数据集上都取得了远超过GMM的识别性能。因此，DNN-HMM替代GMM-HMM成为目前主流的声学建模框架。

如图2.11为上下文相关（Context-Dependent）的DNN-HMM声学模型框架图。HMM建模语音信号的时变性，DNN建模声学特征后验概率$P(S|O)$。DNN的输入为语音信号提取的声学特征，输出层的每个节点分别代表经过决策树聚类的声学状态，因此输出层的维度与声学状态数相同。由公示2.2可知，解码中使用声学似然概率计算得分，因此DNN的输出需要通过如下贝叶斯公式转换再用于后续解码

$$P(O|S) = \frac{P(S|O)}{P(S)} \cdot P(O) \tag{2.36}$$

由于声学特征已知，$P(O)$为固定值，把此项省略不会影响识别结果，$P(S)$为声学状态的分布概率，通常从训练数据中统计得到。

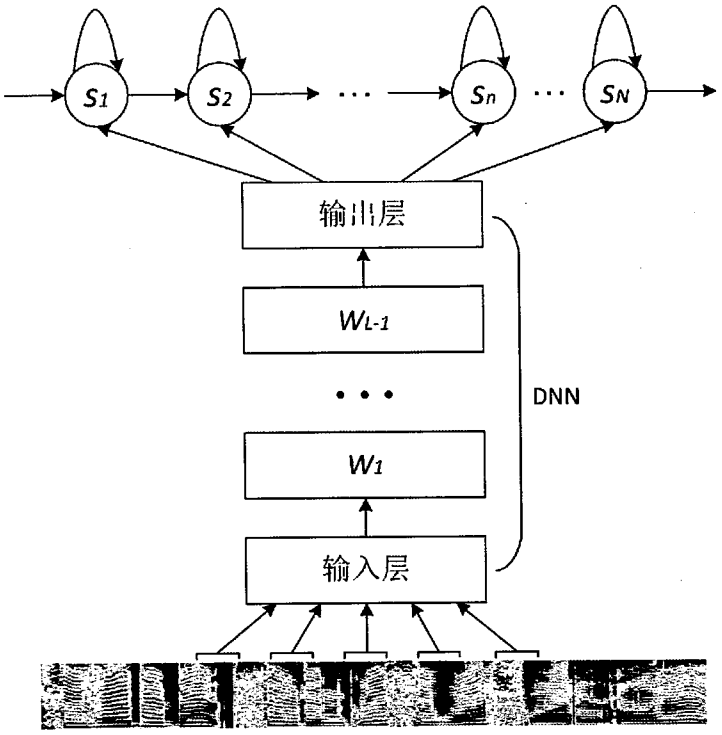


图 2.11 CD-DNN-HMM声学模型框架图

Figure 2.11 Diagram of CD-DNN-HMM based acoustic model

DNN的训练需要训练数据及其对应的标注。为得到每帧声学特征的状态标注，需要利用语音的文本标注和已训练完成的GMM-HMM模型通过维特比算法做强制对齐。然后使用声学特征及其对应的状态级标注利用2.3.3节中介绍的误差反向回传算法更新训练神经网络的模型参数。

训练DNN前首先需要确定目标函数，神经网络建模历史中出现过很多目标函数，在声学模型训练中，常见的最小均方误差（MSE）效果不如交叉熵（CE）损失函数，因此一般选用交叉熵做为神经网络声学模型训练的准则。交叉熵是来自信息论的概念，设建模数据的真实分布为p，神经网络计算得到的分布为q，则二者的交叉熵公式为：

$$\begin{aligned} H(p,q) &= E_p(-\log q) = -\sum p\log q = -\sum p\log p - \sum p\frac{\log q}{\log p} \\ &= H(p) + D_{KL}(p||q) \end{aligned} \tag{2.37}$$

上式中$H(p)$为真实数据的熵，$D_{KL}(p||q)$为衡量两个分布之间的距离的KL散度，所以声学建模最小化交叉熵等同于最小化KL 散度。声学模型训练中将语音特征的状态级标注做为上式中的p。p是各个维度之和为1的向量，通常在标注状态对应位置的值为1，其它维度的值为0。因此，式2.37可以简化为

$$H(p,q) = E_p(-\log q) = -\sum p \log q = -\sum_{label:1} \log q \tag{2.38}$$

即最小化交叉熵此时等价于最大似然估计准则。

交叉熵目标函数仅考虑了当前帧数据的分布。然而，语音作为一个序列建模问题，最优化的解一定要考虑序列信息。在GMM-HMM框架下，人们已经尝试过将序列信息引入GMM参数调整，并取得了成功，如MPE使用包含剪枝过的隐含状态序列集即词图，调整GMM参数使其增大正确和错误HMM状态观察概率的差距。因此，在DNN-HMM框架下，人们也尝试在DNN损失函数中引入词图上正确和错误状态的鉴别性信息，其被称为鉴别性训练（Discrimination Training）。鉴别性训练一般使用交叉熵模型对训练数据解码生成词图，然后以交叉熵模型做为初始种子模型更新权重参数，使其增大解码词图中正确路径的概率并降低错误路径的概率。神经网络声学模型常用的鉴别性训练准则有最大互信息准则（MMI） [66, 67]和状态级别最小贝叶斯风险准则（state-level Minimum Bayes Risk，sMBR） [68, 69]。

虽然相比于GMM模型，基于DNN的语音识别系统性能有显著的提升，但是测试数据与训练数据的不匹配仍然会对识别系统性能有较大的影响。因此，许多关于DNN声学模型的自适应技术被相继提出。最直接的自适应策略是重训练方法，即直接利用自适应数据依照误差反向传播算法调节神经网络权重，但这种方法的缺点是容易产生过拟合。一般采用相近数据训练得到的较为稳健的模型作为重训练的种子模型，以此来缓解过拟合问题。另一种相对稳健的方法是在神经网络中加入线性变换层 [70, 71]。此种算法在自适应过程中只更新线性层权重，其它层的权重保持不变。同样基于线性变换的方法还有fDLR算法 [72]和oDLR 算法 [73]。此外，基于正则化的自适应方法也有很多，如L2 正则化 [74]和相对熵正则化 [75]。正则化算法通过在目标函数中加入正则项达到缓解过拟合的目的。在说话人自适应方面，i-vector [76]和speaker code [77]的引入提升了系统对特定说话人的识别性能。

2.5　本章小结

本章首先简要介绍了语音识别系统的基本原理和框架，其主要包含信号处理与特征提取、声学模型、发音字典、语言模型和解码器。然后针对声学模型进行了详细描述，介绍了神经网络及隐马尔科夫模型的原理以及三个基本问题，随后对GMM-HMM和DNN-HMM声学模型及其自适应技术进行了介绍。本文的研究重点为基于深度神经网络的声学建模，因此有关GMM的内容未进行深入的分析和讨论。

第三章　基于注意力LSTM和多任务学习的远场语音识别声学建模

3.1　引言

自20世纪80年代起，GMM-HMM一直是语音识别声学建模的主流框架。HMM用于建模语音信号的时变特性，GMM表征HMM状态的声学特征分布。EM 算法的出现推动了GMM 在真实世界语音识别任务的应用。在随后的40年里，许多针对此框架优化的算法相继被提出。文献 [78]提出使用上下文相关的音素作为建模单元。为解决数据稀疏问题，文献 [79] 将决策树和数据驱动结合对状态进行聚类。21世纪初，鉴别性训练准则的出现，如最大互信息准则（MMI） [80]、最小分类错误准则（MCE） [81]等，进一步降低了识别错误率。

尽管GMM具有拟合任意复杂分布的能力，但它也有一个严重的缺陷，即对非线性数据建模的效率低下。例如，若对一系列球面的点集建模，只要选择合适的模型，就可以使用很少的参数。但是对于GMM 来说，却需要许多对角高斯分布或全协方差高斯分布才能拟合。因此，相关研究人员提出采用人工神经网络代替GMM，建模HMM状态后验概率 [82]。因为当时计算能力有限，很难训练两层以上的神经网络模型，所以其带来的性能改善非常微弱。

21世纪以来，机器学习算法和计算机硬件的发展使得训练多隐层的神经网络成为可能。实践已经表明，DNN 在各种大型数据集上都取得了远超过GMM的识别性能 [83]。随着深度学习理论的发展，更多结构化的神经网络被提出。与简单的前馈神经网络相比，递归神经网络在隐层之间加入反馈连接，使其可以考虑历史信息。标准的递归神经网络结构如图3.1 所示，图左是包含循环的递归神经网络结构，图右是循环结构沿时间展开的递归神经网络。但在实际应用中，此网络会有严重的梯度消失问题。因此，它对长时信息的建模能力仍然有限 [84]。长短时记忆模型（Long Short-Term Memory，LSTM）的提出有效的解决了此问题 [85]，并在随后被应用到各种深度学习的相关任务中。远场场景下麦克风录制的语音容易受到噪声和混响等因素的干扰，因此导致识别率大幅度下降。在实际应用中，拥有长时建模能力的模型对噪声和混响处理能力更好。因此，本文将基于LSTM网络的声学模型作为基线，对HMM状

态的后验概率进行建模。

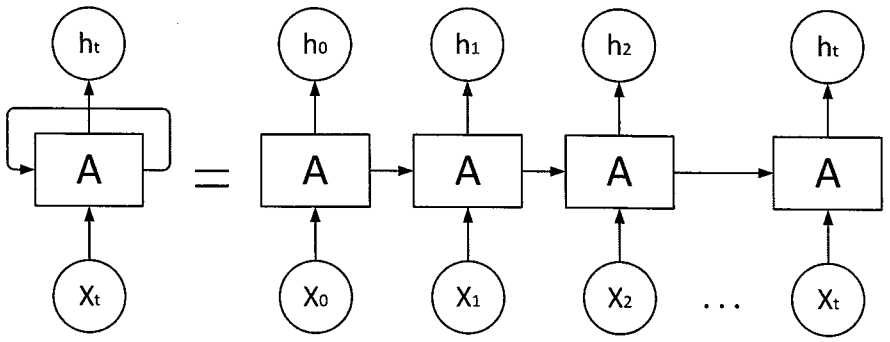


图 3.1 递归神经网络

Figure 3.1 Diagram of recurrent neural network

最近，一种基于注意力的递归神经网络模型在端到端语音识别系统中成功应用 [33–35]。此系统将注意力机制嵌入到模型中，利用其学习输入特征与文本标注之间的对齐。另一方面，文献 [3]中提出将位置相关的注意力信息应用于深度卷积神经网络声学模型，得到的系统性能优于基于双向LSTM声学模型的识别系统。Kim等人将类似的想法应用到多通道远场语音识别中，将声学特征的时间和空间信息结合，通过神经网络来预测声学状态 [37]。其结果表明，注意力机制可以自动将模型的注意力集中于更可靠的输入源上，从而提升模型的识别准确率。

传统的基于神经网络的声学模型一般简单地将上下文多帧特征拼接做为输入，忽略了输入层每帧的时间信息。受上述工作启发，本章针对基于LSTM 的混合声学建模框架提出一种注意力机制，使其对输入层中不同时刻特征向量加权。同时，采用多任务学习（Multi-Task Learning，MTL）框架提升模型在远场场景下的鲁棒性。

本章首先介绍LSTM神经网络的结构，然后对基于LSTM的注意力机制和多任务学习框架分别进行介绍，最后对公开数据集上的实验结果进行比较和分析。

3.2 LSTM神经网络

传统的简单前馈神经网络的输出仅与当前时刻的输入有关。因此，它最主要的缺点是不能充分地对具有上下文相关性的序列进行建模，语音正是这样一

种上下文具有强相关性的时间序列。为加入历史信息，相关研究人员提出在网络中加入递归循环。标准的递归神经网络前向传播的公式为：

$$\boldsymbol{h}_t = \phi(\boldsymbol{W}_{hx}\boldsymbol{x}_t + \boldsymbol{W}_{hh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_h) \tag{3.1}$$

上式中ϕ表示非线性激活函数。然而，实际使用中RNN会有梯度消失的问题，导致距离当前时刻越远的信息对当前的输出影响越小，因此不能对长时依赖性有效地建模。RNN的梯度消失问题如图3.2所示。为了解决此问题，许多基于RNN 的变种结构被提出，如GRU，LSTM等。本节针对本文所使用的LSTM网络做简单描述。

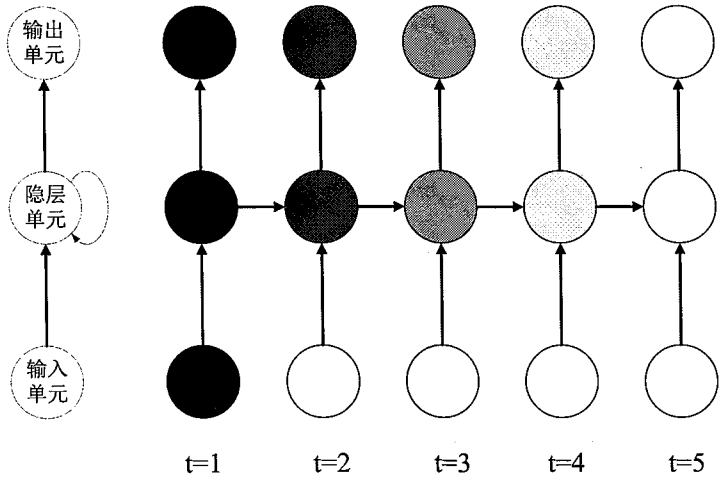


图 3.2 递归神经网络梯度消失示意图

Figure 3.2 The vanishing gradient problem for RNN

用于声学建模的LSTM网络包含输入层，输出层以及它们之间的若干递归隐层。递归隐层由若干记忆模块构成。每个模块包含一个或多个自连接的记忆单元以及控制信息流动的三个门：输入门、输出门和忘记门。LSTM记忆模块的结构如图3.3所示。 已知输入序列表示为$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T)$，递归隐层按照时刻$t = 1$到$T$依次计算三个门和记忆单元的激活值。$t$时刻的计算公式可以表示为如下：

输入门：

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{ix}\boldsymbol{x}_t + \boldsymbol{W}_{ih}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{ic}\boldsymbol{c}_{t-1} + \boldsymbol{b}_i) \tag{3.2}$$

忘记门：

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{fx}\boldsymbol{x}_t + \boldsymbol{W}_{fh}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{fc}\boldsymbol{c}_{t-1} + \boldsymbol{b}_f) \tag{3.3}$$

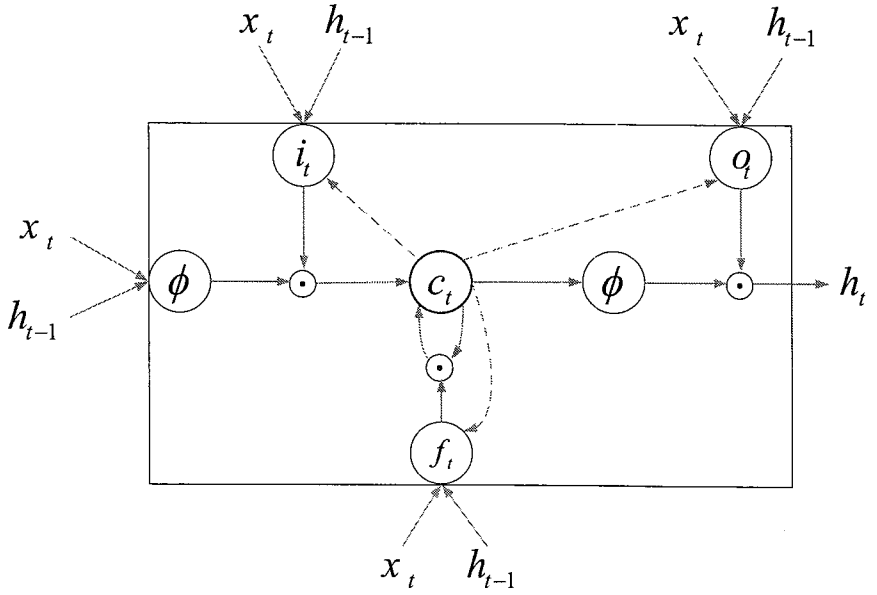


图 3.3 LSTM结构示意图

Figure 3.3 Diagram of an LSTM block

记忆单元：

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \phi(\boldsymbol{W}_{cx}\boldsymbol{x}_t + \boldsymbol{W}_{ch}\boldsymbol{h}_{t-1} + \boldsymbol{b}_c) \tag{3.4}$$

输出门：

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{ox}\boldsymbol{x}_t + \boldsymbol{W}_{oh}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{oc}\boldsymbol{c}_{t-1} + \boldsymbol{b}_o) \tag{3.5}$$

隐层输出：

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \phi(\boldsymbol{c}_t) \tag{3.6}$$

上式中，$\boldsymbol{i}_t$, $\boldsymbol{o}_t$, $\boldsymbol{f}_t$, $\boldsymbol{c}_t$分别是输入门，输出门，忘记门和记忆单元的输出。$\boldsymbol{W}_{.x}$权重矩阵为来自上一隐层的输入$\boldsymbol{x}_t$与记忆模块之间的连接矩阵，$\boldsymbol{W}_{.h}$权重矩阵为当前隐层上一时刻的输出$\boldsymbol{h}_{t-1}$ 与记忆模块之间的连接矩阵，$\boldsymbol{W}_{.c}$为记忆模块内部连接三个门与记忆单元的对角矩阵，$\boldsymbol{b}$ 为偏置向量。σ为sigmoid 非线性函数，ϕ为双曲正切非线性函数，$\odot$ 为向量间的逐个元素相乘符号。当前时刻隐层的输出$\boldsymbol{h}_t$将做为下一隐层的输入。网络输出层包含矩阵线性变换和softmax归一化函数，归一化函数的输出即为声学状态的后验概率。网络中使用输入门、忘记门和输出门分别控制当前、历史以及输出的信息流。

3.3　基于LSTM的注意力机制

随着深度学习和人工智能的发展，神经网络中的注意力机制引起了越来越多研究人员的兴趣。在神经科学和计算神经科学中，对基于注意力的神经过程已经存在广泛的研究 [86,87]。一个典型例子是视觉注意力，许多动物仅专注于其视觉输入的特定部分就可以计算适当的响应。受这个原理启发，在神经元计算时，我们只需选择最相关的信息就可高效地计算出相应的响应，而不是所有可用的信息，因为很大一部分与计算响应无关的输入是可以忽略的。这个类似的想法已经被应用到很多深度学习的应用领域，例如机器翻译 [88]，计算机视觉 [89]，语音识别 [33–35]等。针对语音识别任务，Chorowski等人抛弃了传统的基于HMM的混合声学建模框架，提出了基于注意力机制的端到端语音识别系统 [33]，该系统利用注意力机制对齐输入的声学特征和标注的字符序列。类似听觉注意力原理，模型在对字符解码预测时自动选择与该字符发音有关的声学特征进行关注。

注意力机制严格意义上讲是一种思想，而不是某种模型的实现，因而它的实现以及应用方式可以完全不同。一般情况下，注意力的具体表现形式是一个简单的向量，通常是softmax函数的输出。使用注意力向量对输入或隐层的表达做加权，使模型的关注点集中于与当前预测结果更相关的输入。

传统神经网络声学模型在t时刻的输入$\boldsymbol{x}_t$为L帧上下文特征扩展而成，特征向量经过若干隐层和输出层做前向计算，最终得到网络输出的后验概率。这种做法的缺点是忽略了输入L帧特征内部的时间信息，因为每帧特征对于当前时刻的状态预测贡献不一定是相同的。因此，本文在输入层引入注意力机制，使其对每帧特征给予不同的注意力。受已有关于注意力机制研究工作的启发 [33–35]，本文提出的基于注意力机制的LSTM模型如图3.4所示。三层LSTM网络作为基线声学模型，用来预测HMM状态后验概率。虚线方框部分为多任务学习框架引入的结构。关于多任务学习的内容将在下一节进行介绍。

注意力机制通过神经网络计算输入特征$\boldsymbol{x}_t$ 的注意力权值$\boldsymbol{\alpha}_t$来对L帧特征加权，加权后的特征$\hat{\boldsymbol{x}}_t$替代原始输入$\boldsymbol{x}_t$作为LSTM声学模型的输入。它的具体实现方式可描述为如下：

$$\boldsymbol{e}_t = Attend\left(\boldsymbol{x}_t, \boldsymbol{s}_{t-1}, \boldsymbol{\alpha}_{t-1}\right) \tag{3.7}$$

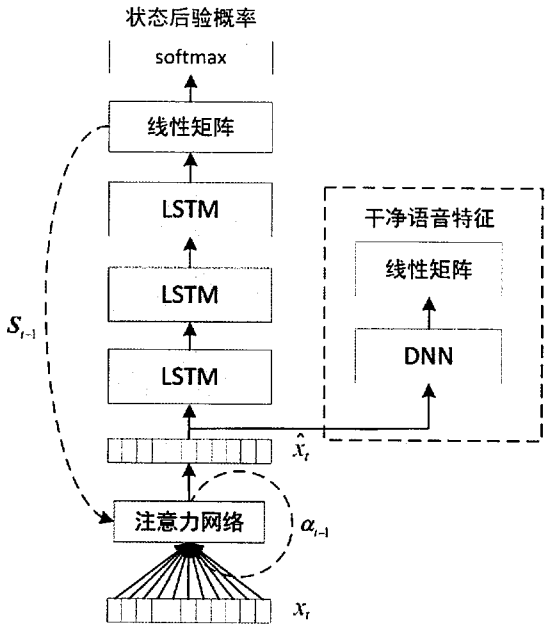


图 3.4 基于注意力机制和多任务学习框架的LSTM声学模型

Figure 3.4 Attention-based LSTM with multitask learning

$$\alpha_{tl} = \frac{\exp(e_{tl})}{\sum_{l=1}^{L} \exp(e_{tl})} \tag{3.8}$$

$$\hat{\boldsymbol{x}}_{tl} = \alpha_{tl}\boldsymbol{x}_{tl} \tag{3.9}$$

$$p(\boldsymbol{s}|\boldsymbol{x}_t) = LSTM(\hat{\boldsymbol{x}}_t) \tag{3.10}$$

上式中，$Attend(.)$表示计算注意力得分$\boldsymbol{e}_t$的神经网络，$LSTM(.)$代表预测声学状态的LSTM网络。如公式3.7 所述，注意力得分$\boldsymbol{e}_t$取决于当前时刻输入$\boldsymbol{x}_t$，上一时刻的预测$\boldsymbol{s}_{t-1}$以及上一时刻的注意力权值$\boldsymbol{\alpha}_{t-1}$。公式3.8表示将注意力得分$\boldsymbol{e}_t$通过softmax函数规整为0到1之间的注意力权值$\boldsymbol{\alpha}_t$。公式3.9利用获得的注意力权值$\boldsymbol{\alpha}_{tl}$对输入L 帧特征中的特征向量$\boldsymbol{x}_{tl}$ 加权，最终得到加权后的特征表示$\hat{\boldsymbol{x}}_t$。模型利用$\boldsymbol{\alpha}_t$ 对来自不同时刻的帧特征给予不同的关注度。将$\hat{\boldsymbol{x}}_t$通过上一节内容所描述的LSTM网络的声学模型即可得到预测的声学状态后验概率。

3.4　多任务学习

3.4.1　多任务学习基本框架

多任务学习已经成功应用于许多机器学习相关的领域中，如自然语言处理、语音识别和计算机视觉等。一般来说，只要优化的损失函数多于一个，就可以被称作是多任务学习。多任务学习的目标是利用相关任务的训练数据中所包含的特定领域信息来提升模型的泛化能力。因此，多任务学习其实是一种知识的归纳和迁移，它充分利用了隐含在各个任务中的特定领域信息，通过辅助任务的额外信息调整模型参数，来达到提升模型泛化能力的目的。

图3.5为深度神经网络中常用的多任务学习框架，其通过在所有任务之间共享隐层来实现。误差反向回传时，来自多个任务的梯度相加对模型参数更新训练。辅助任务可以是一个或者多个，任务的选择是提升模型性能的关键因素。一般情况下，与主要任务具有较强相关性的任务都可以做为辅助任务。在模型训练完成后，与主要任务无关的网络参数会被丢弃，因此模型的参数量不会因辅助任务的引入而增加。

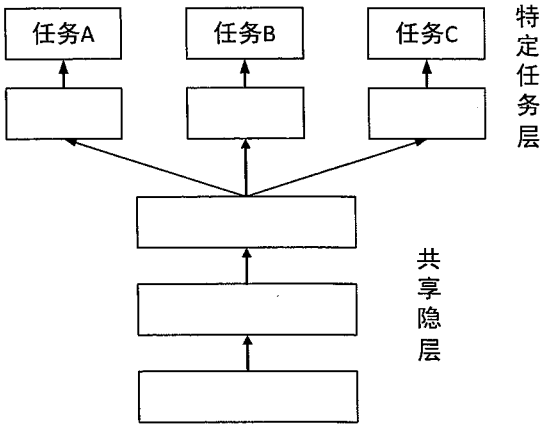


图 3.5 多任务神经网络架构示意图

Figure 3.5 Schematic diagram of multitask learning in DNN

示例图3.5中有三个任务，所有的任务共享输入层和部分隐层，每个任务有自己的特定任务层，若干隐层和输出层构成了特定任务层。网络更新的目标函数为：

$$E(\theta) = \alpha E_A(\theta) + \beta E_B(\theta) + (1-\alpha-\beta)E_C(\theta) \tag{3.11}$$

其中，$0 < \alpha,\ \beta < 1$。E_A，E_B，E_C分别为A,B,C任务的目标函数，θ为待更新的模型参数。插值系数α 和β用来对各个任务的重要性做出平衡。一般情况下，主要任务会被赋予较大的权重值。整个框架的模型参数通过优化插值目标函数联合更新学习。

3.4.2 多任务学习在声学建模中的应用

目前已有很多将多任务学习应用于语音识别声学建模的研究。多个学习任务中主要任务的输出是HMM 状态的后验估计，用于最终的识别任务。文献 [90] 中提出使用多任务学习框架联合优化上下文相关和上下文无关音素的状态建模，并证明此训练方法有效地提升了模型的鲁棒性。共享隐层的多语种神经网络声学模型 [91] 是多任务学习框架的另一个应用实例，其完成了模型级别的跨语种神经网络声学建模。文献 [92] 中将多任务学习应用于小资源任务的语音识别，文献 [27] 使用多任务学习改善混响环境中的语音识别。

提升声学模型的噪声鲁棒性一直是语音识别系统性能改善的关键。远场麦克风录制的语音信号易受噪声的干扰。因此，本文中将特征增强选为远场语音识别声学建模的辅助任务，利用与远场语音同步录制的近场语音特征做为特征增强辅助任务的标注。深度神经网络的每一层都可以看做一个特征提取器，加入的特征增强辅助任务使特征提取器的输出更具声学鉴别区分性。因此，可以在一定程度上提升模型的鲁棒性。

图3.4中虚线方框部分为多任务学习框架引入的结构，在注意力加权的声学特征$\hat{\boldsymbol{x}}_t$ 后连接一层神经网络和一个线性矩阵，线性矩阵的输出为映射后的干净语音特征。模型有两个输出，一个是预测三音子状态的后验概率输出，另一个是经过神经网络映射的干净语音特征。模型的一个分支对声学特征做状态分类，另一个分支对远场语音特征做去噪增强。在模型训练阶段，状态分类任务与特征增强任务分别使用交叉熵准则与最小均方误差准则优化相关模型参数。联合优化的目标函数表示为

$$E(\theta) = \alpha \sum_t p(\boldsymbol{s}_t|\boldsymbol{x}_t) + (1-\alpha) \sum_t (\hat{\boldsymbol{y}}_t - \boldsymbol{y}_t)^2 \tag{3.12}$$

上式中，$\boldsymbol{x}_t$为远场麦克风录制的语音在t时刻的声学特征，$\boldsymbol{s}_t$为t时刻声学特征$\boldsymbol{x}_t$的三音子状态标注，$\hat{\boldsymbol{y}}_t$ 是神经网络模型在特征增强辅助任务上的输出，$\boldsymbol{y}_t$是近场麦克风录制的语音在t时刻的声学特征，α是平衡两任务重要性的目标

函数插值系数。模型训练完成后，丢弃虚线方框部分中仅与辅助任务相关的模型参数。

3.5　实验与分析

本文的所有实验均在公开数据集AMI [93]和ICSI [94]上进行。本节首先对实验中使用的两个数据集和实验配置做介绍，然后分析了输入帧数对注意力机制的影响，最后对单通道和多通道数据上的实验结果进行分析。

3.5.1　实验数据集

- AMI数据集：AMI数据集包含100小时英文会议语音数据，数据类型为完全自然的口语对话。每个会议通常有四个参与者，参与者中大部分人的母语非英语，会议时长为一个半小时左右。语音的录制采用多个麦克风，其中包含头戴式麦克风，翻领麦克风以及多个麦克风阵列。8个麦克风组成的10厘米半径均匀圆形阵列作为主麦克风阵列，次麦克风阵列的几何形状随录制地点不同而变化。AMI 语料官方网址对数据的划分方式为80小时训练数据，10小时开发集以及10 小时测试集。

- ICSI数据集：ICSI数据集是由位于美国加利福尼亚州伯克利市的国际计算机科学研究所（International Computer Science Institute，ICSI）收集的英文会议语音数据，通过数据联盟（Linguistic Data Consortium，LDC）于2004年发布。语料中包含75 个会议的录音数据，它们通常是ICSI每周举行的工作组会会议。因此，语音数据类型同样为自然的口语对话。会议的时长在17到103分钟之间，但大部分会议的时长不超过一小时。每个会议最多有十人参加。语音的录制采用头戴式麦克风和4 个桌面式麦克风组成的1m间距线性阵列。语料的总时长为72小时。

3.5.2　实验配置

本章实验在AMI和ICSI两个数据集上进行。远场单通道（Single Distant Microphone, SDM）实验中，AMI数据集采用主麦克风阵列的第一通道数据，ICSI数据集采用线性阵列的通道E的数据。在远场多通道（Multiple Distant Microphones, MDM）实验中，AMI数据集使用主麦克风阵列录制的八通道数

据，ICSI 数据集使用线性阵列录制的四通道数据。头戴式麦克风同步录制的近场语音特征做为多任务框架中增强辅助任务的标注。AMI实验采用官方数据划分方式，使用80小时数据训练声学模型。本文的所有实验中，开发集数据未参与任何声学模型参数的训练和调节。因此，开发集和测试集都用来测试对比模型的性能。ICSI实验将六个完整的会议语音做为测试集（Bed008, Bmr005, Bmr020, Bmr026, Bro015 和Bro016），测试集时长为5小时，其余67小时数据做为训练数据。

系统采用由训练集标注和Fisher语料训练的三元文法语言模型，发音字典使用CMU字典。基线声学模型采用三层LSTM 网络，每层包含1024 个记忆单元。为减小模型参数量，每层递归隐层后衔接512 维的映射层。声学特征使用40维梅尔对数滤波器组特征（MS-LFB），将其做均值方差规整后输入到LSTM声学模型。我们采用截断的沿时反向传播算法(back propagation through time, BPTT)训练更新模型参数，截断长度为20 帧。为平衡两个训练目标，多任务目标函数中的插值系数α设为经验最优值0.9。多任务学习引入的结构（即图3.4 中虚线框部分）为一个1024 维的DNN 隐层（即线性变换层与非线性函数）和一个线性变换层。神经网络模型的两个输出分别为聚类后的三音子状态和映射的40 维MS-LFB干净语音特征。

3.5.3 输入层特征帧数选取

首先，在不引入多任务学习框架的情况下，探究输入层扩展帧数对注意力机制的影响。将输入层的上下文扩展帧数做为变量，寻找注意力机制下的最优配置。输入窗长实验在AMI的远场单通道和多通道数据集上进行。在多通道实验中，LSTM基线模型的输入是多通道特征的拼接。引入注意力机制时，将多通道数据中来自同一个时刻的40维特征拼接做为单帧输入。表3.1为变化输入帧数时，基线LSTM 声学模型与嵌入注意力机制的LSTM 声学模型（以下简称为ALSTM）在AMI开发集(dev)和测试集(eval)上的识别词错误率（WER）对比结果。表格中第二列表示扩展的上下文，例如，由$t-3$ 到$t+3$的输入扩展表示为[-3, 3]。本小节实验遍历了三组参数：[-3, 3]，[-5, 5]和[-7, 7]。

由表3.1可以看出，调整输入层的上下文扩展帧数对基线系统性能的影响相对比较小，基于注意力机制的LSTM 模型对输入层扩展帧数的变化更敏感，其在[-5, 5]扩展处取得了最优性能。单通道实验中，ALSTM模型在开发集和测试

表 3.1 输入帧数变化时的识别性能对比结果

Table 3.1 Performance comparison at different number of input frames

数 据	输入上下文	LSTM		ALSTM	
		dev	eval	dev	eval
单通道	[-3, 3]	43.0	47.5	43.0	47.6
	[-5, 5]	42.8	47.2	41.7	46.2
	[-7, 7]	43.1	47.3	42.1	46.7
多通道	[-3, 3]	37.5	42.4	36.7	41.7
	[-5, 5]	37.8	42.7	36.0	41.4
	[-7, 7]	38.0	43.3	36.4	41.5

集上分别有1.1% 和1% 的词错误率下降。多通道实验中，开发集和测试集上的识别词错误率下降1.8%和1.3%。表3.1 的结果表明11帧输入特征对于注意力机制来说是足够的。因此，接下来的多任务学习实验也采用上下文各扩展5帧的配置。

3.5.4 单通道实验

会议语音数据的一个重要特征是语音重叠，即某一时刻有多个人同时说话。我们在模型训练阶段没有去掉同时包含多说话人的语音片段，以下的实验表格中给出了模型在测试全集和去掉重叠语音片段的测试子集上的结果。带星号的测试集为去掉重叠语音后的测试子集。表3.2和表3.3为引入注意力机制和多任务学习的模型在AMI和ICSI单通道数据上的识别词错误率对比结果。

表 3.2 AMI单通道数据集上的识别性能对比结果

Table 3.2 Performance comparison in the SDM case of AMI

声学模型	dev	dev*	eval	eval*
LSTM	42.8	34.3	47.2	38.3
ALSTM	41.7	33.6	46.2	37.6
ALSTM-MTL	41.3	33.1	45.8	37.2

多任务学习框架加入对远场语音特征的增强任务，将映射干净语音的40维

表 3.3 ICSI单通道数据集上的识别性能对比结果

Table 3.3 Performance comparison in the SDM case of ICSI

声学模型	eval	eval*
LSTM	39.5	34.2
ALSTM	38.7	33.5
ALSTM-MTL	38.4	33.2

声学特征做为第二目标。表格中使用ALSTM-MTL 表示基于注意力和多任务学习的LSTM声学模型。由表格中的结果可以看出，语音重叠导致系统识别性能大幅度下降。当只考虑不含重叠语音的子集时，两个数据集上的识别词错误率降低5%以上。在AMI单通道实验中，基于注意力的LSTM模型在全集和子集分别取得了平均2.3%和2%的相对提升。类似的提升在ICSI数据集上得到了验证。引入多任务学习框架后，模型在两个数据集上进一步取得了绝对0.3~0.5%的词错误率下降。虽然多任务学习带来的性能改善幅度较小，但在所有的测试集上均有稳定的性能提升。此结果验证了多任务学习框架提升模型鲁棒性的结论。总的来说，与基线LSTM模型相比，提出的模型结构在单通道数据集上平均取得了相对3.1%的性能改善。

3.5.5 多通道实验

为进一步验证提出模型的有效性，本小节在多通道数据集上进行实验。针对多通道语音识别任务，文献 [95,96]直接将多通道声学特征拼接起来训练神经网络声学模型，并取得了优于单通道模型的识别结果。本章提出的模型结构只是对输入层的时间信息进行利用，不涉及对空间信息的利用。因此，对比的基线模型仍然采用简单的声学特征拼接。对于注意力机制来说，同一时刻的多通道声学特征拼接形成的多维特征做为公式3.9中的$\boldsymbol{x}_{tl}$。模型在多通道数据上的实验结果见表3.4和表3.5。

表中第一行为单通道数据上LSTM基线模型的识别性能，第二行是多通道数据上LSTM基线模型的结果。与单通道模型相比，简单的多通道特征拼接训练LSTM模型就可以取得显著的性能改善。此对比结果表明，使用麦克风阵列采集远场语音信号是提升远场语音识别性能的重要手段。对比第二行和第三行的结果可知，LSTM模型加入注意力机制后，多通道声学模型在两个数据集的

表 3.4 AMI多通道数据集上的识别性能对比结果

Table 3.4 Performance comparison in the MDM case of AMI

数 据	声学模型	dev	dev*	eval	eval*
单通道	LSTM	42.8	34.3	47.2	38.3
多通道	LSTM	37.8	30.7	42.7	34.5
	ALSTM	36.0	29.7	41.4	33.6
	ALSTM-MTL	35.5	29.1	41.0	33.2

表 3.5 ICSI多通道数据集上的识别性能对比结果

Table 3.5 Performance comparison in the MDM case of ICSI

数 据	声学模型	eval	eval*
单通道	LSTM	39.5	34.2
多通道	LSTM	29.8	25.2
	ALSTM	29.0	24.5
	ALSTM-MTL	28.7	24.2

测试全集上平均获得相对3.5%的性能提升。多任务学习的引入使模型的性能进一步提升。与多通道LSTM 基线模型相比，最终的ALSTM-MTL 模型性能相对提升4.5%。

3.6 本章小结

本章简单介绍了实验中使用的LSTM基线模型，并针对远场语音识别任务提出一种基于注意力机制和多任务学习框架的LSTM 声学模型。模型中嵌入的注意力机制使其自动学习调整对输入层扩展上下文特征输入的关注度。为提升模型对噪声的鲁棒性，训练阶段引入多任务学习框架，使其联合预测声学状态和干净特征。本章的最后一部分在AMI和ICSI数据集上验证了提出模型的有效性。同时，单通道与多通道模型的性能对比结果表明，麦克风阵列在远场语音识别场景中的应用是一个必然趋势。因此，如何有效地利用多通道数据对远场语音进行声学建模是本文的下一步研究内容。

第四章　基于空间特征补偿的多通道声学建模

4.1　引言

受益于计算机技术与深度学习理论的发展，语音识别系统在近场场景下已具有良好的识别性能。然而，人们对远场语音识别技术有着更广泛的应用需求。现实生活中大部分的应用场景是用户距离麦克风较远的远场环境，例如家居场景，会议场景等。在这种远场环境中，语音信号受到噪声和混响的严重干扰，系统的识别率大幅度下降。因此，识别系统在远场场景中的表现仍然差强人意。为提高识别性能，语音识别系统通常采用多个麦克风的信号来增强语音信号，减少混响和噪声的影响。与单通道相比，使用麦克风阵列的优势在于多通道语音信号可以提供空间上的区分度。

大部分的多通道语音识别系统采用两个独立的系统模块。首先，使用麦克风阵列语音增强模块对多通道信号做语音增强，其通常包含定位、波束形成和后滤波阶段。然后，将增强后的单通道语音信号通过语音识别模块做特征提取和声学建模。波束形成将每个通道的麦克风信号通过不同的滤波器后相加求和，以增强来自目标方向的信号并衰减来自其他方向的噪声。常用的波束形成算法有延迟相加，最小方差无失真估计等。然而，当最终目标为提高语音识别准确率时，将增强模型独立于声学模型单独优化可能不是最佳的解决方案。为解决此问题，Seitzer等人在早期研究 [6]中提出了一种似然最大化波束形成（likelihood-maximizing beamforming，LIMABEAM），其将波束形成的参数与GMM-HMM声学模型联合优化。文中的实验结果表明，此算法优于传统的延迟相加波束形成。与大部分的语音增强算法一样，LIMABEAM是一种基于模型的解决方案，其需要在声学模型推断和增强模型优化之间交替迭代估计模型参数。

目前的语音识别系统一般使用基于深度神经网络的声学模型，神经网络模型的参数使用梯度学习算法进行优化。为了使前端语音增强与后端神经网络声学模型可以联合优化，许多研究将前端增强模块引入神经网络。Heymann等人使用神经网络估计用于波束形成滤波系数计算的掩蔽值 [41]，并将此网络与声学模型联合优化。Xiao等人将信道间的广义互相关（Generalized Cross

Correlation，GCC）通过神经网络预测各个通道的频域滤波系数，随后对增强后的信号提特征输入神经网络声学模型 [39]。Zhong等人将多通道信号的短时傅里叶变换通过LSTM网络预测频域滤波系数 [40]。然而，上述的工作一般要引入若干层神经网络预测语音信号的掩蔽值或波束形成的滤波系数，导致最终用于识别的模型参数量较大。

从上一章的单通道和多通道声学模型的识别词错误率对比结果可以看到，与单通道相比，基于多通道语音信号的声学建模显著改善了远场语音系统的识别性能。但是在上一章实验中，多通道信号提供的空间信息没有被充分利用，因此，需对其做进一步的改进。本章在上一章的基础上，针对多通道语音识别声学建模，提出了一种基于空间特征补偿的方法。其在引入较少模型参数量的前提下，有效地提升了多通道语音识别系统的准确率。

本章的主要安排如下：首先，4.2节简单回顾了传统框架下的多通道语音识别系统；其次，在4.3节详细介绍了信道间的相位变换广义互相关（Generalized Cross Correlation with Phase Transform，GCC-PHAT）；然后，在4.4 节总结了特征补偿在声学建模中的应用，并提出基于GCC-PHAT 空间特征补偿的多通道声学模型；实验与分析在4.5 进行；最后，对本章内容做出总结。

4.2 多通道语音增强算法

图4.1为传统的多通道语音识别系统框图。首先，将多通道语音信号经过语音增强生成单通道语音信号。然后，对增强的单通道信号做特征提取得到声学特征。最后，将声学特征经过识别器解码得到识别的文字序列。

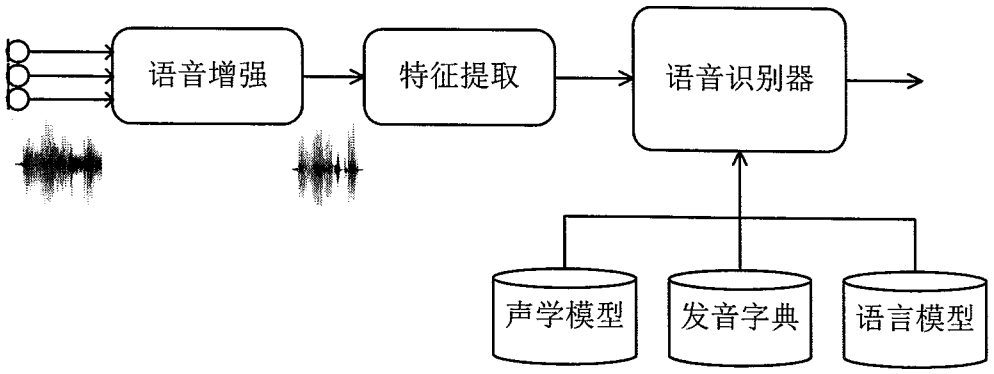


图 4.1 传统的多通道语音识别系统框图

Figure 4.1 Block diagram of a conventional multichannel speech recognition system

若收集语音信号的麦克风阵列由J个麦克风构成，图4.2为多通道语音增强算法示意图。目标语音信号为$x(t)$，第j个麦克风接收到的信号为$y_j(t)$。考虑

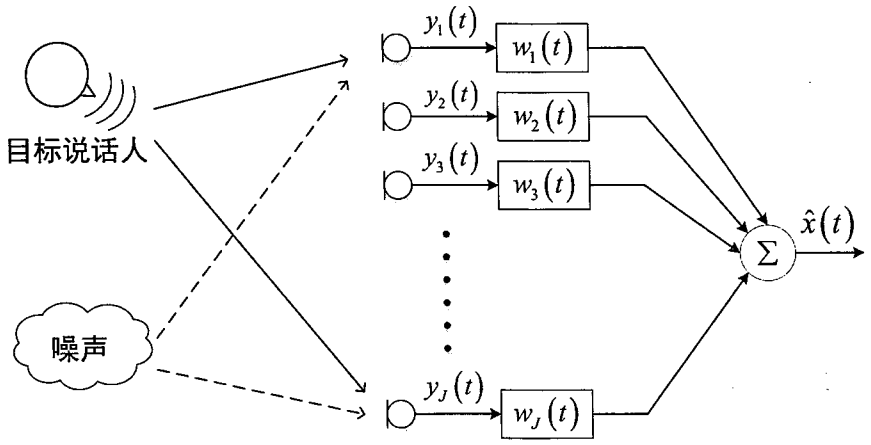


图 4.2 多通道语音增强算法示意图

Figure 4.2 Schematic diagram of a multichannel speech enhancement system

到混响和加性噪声因素，时域接收信号$y_j(t)$ 可以表示为：

$$y_j(t) = h_j(t) * x(t) + n_j(t), \quad j = 1, 2, \cdots, J \tag{4.1}$$

其中，$h_j(t)$是目标语音信号到第j个麦克风的声学冲激响应，$n_j(t)$为加性噪声信号，$*$为卷积运算符。语音增强算法一般在频域实现。对第j个麦克风接收到的信号$y_j(t)$分帧加窗，然后做短时傅里叶变换（Short-Time Fourier Transform，STFT），其频域表示为

$$Y_j(k,l) = \sum_{\ell=0}^{L-1} y_j(lR+\ell) m_j(\ell) e^{-\frac{j2\pi k\ell}{L}}, \quad j = 1, 2, \cdots, J \tag{4.2}$$

上式中，L为帧长，R为帧移，l为短时分帧加窗的帧序号，k为频率子带序号，m_j为窗函数。

对式4.1做短时傅里叶变换，得到J个麦克风接收到的信号为

$$\boldsymbol{Y}(k,l) = \boldsymbol{H}(k)X(k,l) + \boldsymbol{N}(k,l) \tag{4.3}$$

其中，

$$\boldsymbol{Y}(k,l) = [Y_1(k,l), Y_2(k,l)), \cdots, Y_J(k,l)]^{\mathrm{T}} \tag{4.4}$$

$$\boldsymbol{H}(k) = [H_1(k), H_2(k), \cdots, H_J(k)]^{\mathrm{T}} \tag{4.5}$$

$$\boldsymbol{N}(k,l) = [N_1(k,l), N_2(k,l), \cdots, N_J(k,l)]^{\mathrm{T}} \tag{4.6}$$

上式中，T为转置操作，$Y_j(k,l)$，$X(k,l)$以及$N_j(k,l)$分别为$y_j(t)$，$x(t)$和$n_j(t)$的短时傅里叶变换。假定短时分析时，声学传递函数是时不变的。$H_j(k)$是$h_j(t)$的短时傅里叶域表示，即目标说话人到第j个麦克风的声学传递函数。描述目标声源到麦克风阵列传递函数的向量$\boldsymbol{H}(k)$通常被称为导向矢量。

如图4.2所示，每个麦克风接收到的信号$Y_j(k,l)$经过滤波器$W_j(k,l)$后相加生成增强后的单通道信号$\hat{x}(t)$。将滤波器的增益函数表示为

$$\boldsymbol{W}^{\mathrm{H}}(k,l) = [W_1(k,l), W_2(k,l), \cdots, W_J(k,l)] \tag{4.7}$$

式中，H表示共轭装置。因此，多通道语音增强算法的输出为

$$\hat{X}(k,l) = \boldsymbol{W}^{\mathrm{H}}(k,l)\boldsymbol{Y}(k,l) \tag{4.8}$$

波束形成是一种被广泛采用的多通道语音增强技术。它通过空间滤波操作，将麦克风阵列的输出转换为单通道信号，使其关注目标说话人方向的语音，并衰减来自其他方向的噪声。基于波束形成的多通道语音增强算法使用不同优化准则得到不同的滤波参数。波束形成算法一般先对多通道信号做空间信息提取，如通道信号间的到达延时（Time Delay Of Arrival，TDOA），导向矢量或信号的空间相关矩阵等。然后依据相应的滤波准则，并利用提取的空间信息计算得到最终的滤波参数。波束形成是阵列信号处理中的一个长期研究课题。由于本文主要研究内容为语音识别声学建模，以下只针对两种常见的波束形成算法做简单介绍，即延迟相加（Delay and Sum）波束形成和最小方差无失真响应(MVDR) 波束形成。

4.2.1 延迟相加波束形成

在最简单的延迟相加波束形成中，由于目标说话人距离不同麦克风的距离不同，导致语音信号到达麦克风的时间不同。因此，不同的麦克风收集到的语音信号会有时间差。延迟相加波束形成算法通过对不同麦克风收集的信号做时间补偿，来同步各个麦克风收集到的语音信号。将同步后的各个通道信号相加

求和得到增强后的单通道语音信号。由此可得，延迟相加波束形成算法的滤波器系数为

$$\boldsymbol{W}^{\mathrm{H}} = [a_1 \exp(j2\pi f\tau_1),\ a_2 \exp(j2\pi f\tau_2),\ \cdots,\ a_J \exp(j2\pi f\tau_J)] \tag{4.9}$$

其中，$\boldsymbol{a} = [a_1,\ a_2,\ \cdots,\ a_J]^{\mathrm{T}}$ 是J个通道信号的加权，$\boldsymbol{\tau} = [\tau_1,\ \tau_2,\ \cdots,\ \tau_J]^{\mathrm{T}}$是J个通道信号的延时。因此，延迟相加波束形成算法的时域表示为

$$\hat{x}(t) = \sum_{j=1}^{J} a_j y_j(t + \tau_j) \tag{4.10}$$

为估计通道信号间的到达延时，一般通过对混响更具鲁棒性的GCC-PHAT计算得到。关于GCC-PHAT的内容在下一节做详细介绍。

4.2.2　最小方差无失真响应波束形成

将式4.3与式4.8相结合，得到增强后单通道信号的频域表示为

$$\hat{X}(k,l) = \boldsymbol{W}^{\mathrm{H}}(k,l)\boldsymbol{H}(k)X(k,l) + \boldsymbol{W}^{\mathrm{H}}(k,l)\boldsymbol{N}(k,l) \tag{4.11}$$

为使目标说话人的语音信号保持无失真，式4.11需满足

$$\boldsymbol{W}^{\mathrm{H}}(k,l)\boldsymbol{H}(k) = 1 \tag{4.12}$$

此时，式4.11可以等价于

$$\hat{X}(k,l) = X(k,l) + \boldsymbol{W}^{\mathrm{H}}(k,l)\boldsymbol{N}(k,l) \tag{4.13}$$

与此同时，为得到最佳滤波系数，使波束形成器输出$\hat{X}(k,l)$中的噪声最小化。因此，最小方差无失真响应波束形成器的滤波参数表示为

$$\begin{aligned} \boldsymbol{W}^{MVDR} = &\arg\min_{\boldsymbol{W}} E\{|\boldsymbol{W}^{\mathrm{H}}(k,l)\boldsymbol{N}(k,l)|^2\} \\ &\text{subject to } \boldsymbol{W}^{\mathrm{H}}(k,l)\boldsymbol{H}(k) = 1 \end{aligned} \tag{4.14}$$

方程4.14的解为

$$\boldsymbol{W}^{MVDR} = \frac{\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{H}(k)}{\boldsymbol{H}^{H}(k)\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{H}(k)} \tag{4.15}$$

其中，$\mathbf{\Phi}_{NN}$为噪声功率谱密度矩阵，测量不同麦克风接收的噪声信号之间的相关性：

$$\mathbf{\Phi}_{NN} = \begin{bmatrix} \phi_{N_1N_1} & \phi_{N_1N_2} & \cdots & \phi_{N_1N_J} \\ \phi_{N_2N_1} & \phi_{N_2N_2} & \cdots & \phi_{N_2N_J} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{N_JN_1} & \phi_{N_JN_2} & \cdots & \phi_{N_JN_J} \end{bmatrix} \tag{4.16}$$

因此，将空间信息提取步骤估计的导向矢量和噪声空间相关矩阵代入，就可得到求得滤波器系数。

4.3 相位变换广义互相关

正如上一节4.2.1所介绍，GCC-PHAT已成功用于确定两个空间分离的麦克风接收声波的到达时间延迟。因此，当麦克风阵列的几何形状固定时，多个麦克风对间的到达时延可以用来参数化声源位置 [97, 98]。对于语音信号来说，GCC-PHAT 是对说话人的位置信息进行编码，其提供了单通道语音信号无法获取的空间信息。

已知两个信道接收到的语音信号为$x_i(t)$和$x_j(t)$，GCC-PHAT计算公式表示为如下：

$$gcc_{ij}(t) = IFFT\left(\frac{X_i(f)X_j^*(f)}{|X_i(f)X_j^*(f)|}\right) \tag{4.17}$$

上式中，$X_i(f)$和$X_j(f)$分别为信号$x_i(t)$和$x_j(t)$的傅里叶变换，$*$表示复数共轭操作。

由于说话人到每个麦克风的距离不同，语音信号到达不同麦克风时会有时延差，因此，$X_i(f)$ 与$X_j(f)$之间存在相位差。举例来说，在不考虑混响和噪声的情况下，两个麦克风接收的信号$x_i(t)$和$x_j(t)$的关系为

$$x_j(t) = x_i(t - \tau_{ij}) \tag{4.18}$$

其中，τ_{ij}为两麦克风接收信号的时间差。上式的是频域表示为

$$X_j(f) = X_i(f)e^{-j2\pi f\tau_{ij}} \tag{4.19}$$

将上式代入公式4.17，我们得到

$$gcc_{ij}(t) = IFFT(e^{j2\pi f\tau_{ij}}) = \delta(t + \tau_{ij}) \tag{4.20}$$

经过傅里叶逆变换得到的GCC-PHAT在变量t为两信道间的延时处取得最大值，因此，两个麦克风之间的TDOA估计为

$$\hat{d}(i,j) = \underset{t}{\text{argmax}}\ gcc_{ij}(t) \tag{4.21}$$

理想情况下，$gcc_{ij}(t)$会在一段有限范围内呈现一个峰值，出现峰值的位置对应着麦克风i和j之间接收信号的到达时间差。麦克风间距确定了信号到达时延出现的范围。由于每个麦克风的位置不同，信号从声源位置到各个麦克风之间的声学路径长度不同。信号到两个麦克风间的最大路径差为麦克风的间距。当麦克风间距更远时，信号到达的时间差范围就会更大。因此，时延对应的峰值出现的有效范围由麦克风间距除以声速来决定。

由于本文的实验在AMI和ICSI两个数据集进行，此处使用AMI数据集举例说明GCC-PHAT的时延有效范围。AMI数据集采用8个麦克风10厘米半径的均匀分布圆形阵列，任意两个麦克风间的最大间距为20厘米，麦克风间的最大时间延迟为$\tau = 0.2m/340m/s = 0.588ms$。在16khz 的语音信号采样率时，它对应着10个采样点的延迟。因此，AMI数据集中麦克风间TDOA 范围是-10到10，其对应着GCC-PHAT中心21个相关系数。阵列中麦克风对的数目为$C_8^2 = 28$，所以，$21 \times 28 = 588$ 维相关系数足够用来编码说话人的位置。相应地，对于ICSI 数据集来说，麦克风最大间距为3m。因此每对麦克风间采用281个相关系数，整个麦克风阵列的相关系数维度是$C_4^2 \times 281 = 1686$。

4.4　特征补偿在声学建模中的应用

在GMM-HMM声学建模框架中，设计良好的声学特征是提升语音识别性能的重要手段之一。许多研究人员通过改善特征的鲁棒性来达到提升识别率的目的 [99, 100]。为减少模型的参数量，GMM通常要求输入特征的每个维度统计独立。梅尔倒谱系数（MFCC）和感知线性预测系数（PLP）是GMM-HMM声学建模框架下最常用的特征。它们的计算均源自于梅尔对数滤波器组特征（MS-LFB）。MS-LFB特征的各个维度之间存在相关性，因此不能直接做为GMM-HMM声学模型的特征。然而，MS-LFB到MFCC 或PLP之间的特征转换过程可能会丢失对声学状态分类有用的信息。与GMM 相比，DNN表现更出色的主要原因是它具有联合学习复杂特征表示和分类器的能力。它不需要人为设计的高层次特征做为输入。DNN 可以看作一个所有隐层组成的特征学习模

型和一个用softmax层表示的对数线性分类器。每个隐层的输出都是对原始声学特征的一种表达形式。越高层的特征越抽象，并且对输入的变化越不敏感。因此，文献 [101] 中直接使用MS-LFB特征做为神经网络声学模型的输入。

语音信号中通常存在两种变化类型：说话人的变化和环境的变化。声学模型对以上两种变化的鲁棒性是评价模型好坏的关键标准。为解决说话人变化的问题，Andreou等人针对GMM-HMM系统提出声道长度归一化（VTLN）算法。Gales提出特征空间最大似然线性回归（fMLLR），其对特征向量仿射变换，以使得变换后的特征更好地匹配模型 [64]。文献 [72] 对比了以上技术在GMM和DNN 上的效果。实验发现，VTLN和fMLLR可以为GMM系统带来显著的错误率降低，但在DNN系统上性能的提升相对较少，针对GMM系统提出的说话人自适应算法对DNN模型不是非常有效。因此，文献 [76]中提出将带有说话人特性的向量（i-vector）做为输入特征，与声学特征并行输入神经网络，使DNN声学模型适应目标说话人。图4.3为加入i-vector的声学模型结构示意图，不同的说话人之间i-vector 是变化的，在训练和测试阶段，将说话人向量i-vector连接到该说话人的声学特征。在300 小时Switchboard 数据集上的实验结果表明，与仅在声学特征上训练的模型相比，该方法将模型的性能提升了相对10%。另一方面，为了解决环境变化的问题，基于GMM-HMM声学模型的许多自适应技术被提出，如向量泰勒级数（Vector Taylor Series，VTS）自适应 [102, 103]，最大似然线性回归（MLLR） [104]等。针对DNN声学模型，文献 [26] 提出噪声感知训练(Noise-aware Training，NaT)，该算法在每一个观察输入之后增加信号噪声的估计值。文献 [27]则利用语音信号提取出与房间混响特性有关的向量，并将此向量作为训练DNN声学模型的补充特征。

综上所述，为应对说话人和环境的变化，GMM-HMM声学模型一般对特征做变换或对模型参数进行更新。相对来说，基于DNN-HMM 声学模型的处理方式更简单直观，其通过灵活地使用输入特征来适应语音信号中环境和说话人的变化。上述基于DNN 的研究工作说明，尽管DNN 具有强大的特征学习能力，使用信号或者环境中提取的辅助先验信息作为输入仍然可以进一步提升DNN的声学建模能力。辅助信息通常被表示为一个固定长度的向量，在训练和解码阶段与声学特征同步输入。

语音识别的声学特征提取一般保留幅度信息忽略相位信息。文献 [95]与[96]直接将多通道声学特征拼接训练DNN声学模型，并且取得了优于单通道模

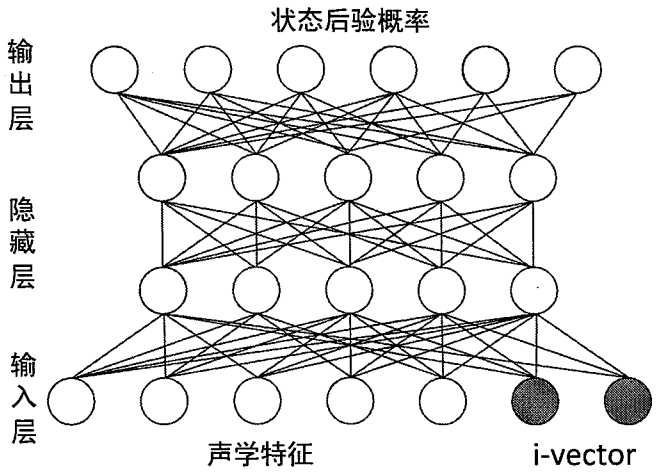


图 4.3 输入增加i-vector的神经网络声学模型示意图

Figure 4.3 Diagram of a neural network with inputs augmented with i-vectors

型的性能。然而，与单通道信号相比，多通道信号提供了额外的空间信息，即声源的位置信息，此信息存储于相位信息中。因此，简单的利用幅度信息会丢失多通道数据引入的关键信息。上一节4.3已经介绍过，GCC-PHAT 一般用来估计信道间的TDOA，各个通道间的到达时间延迟可以用来参数化声源位置。因此，GCC-PHAT包含麦克风信道间接收信号的时延信息，是一种对说话人位置信息编码的表达。受到上述研究工作的启发，本章提出将编码说话人位置信息的GCC-PHAT做为空间特征补偿，与多通道声学特征同步输入到神经网络声学模型。

4.5 实验与分析

本节实验在AMI和ICSI数据集上进行。实验主要包含三部分内容：第一部分分析GCC-PHAT提取窗长对识别性能的影响，并根据实验结果选取最佳窗长；第二部分在AMI和ICSI两个实验数据集上验证基于GCC-PHAT空间特征补偿方法的有效性；第三部分将第二章提出的注意力机制与空间特征补偿方法相结合，并对两个数据集上的实验结果进行对比分析。

4.5.1 实验配置

本章实验的基线模型与上一章相同。为叙述方便，这里对训练数据

和LSTM相关模型参数进行简单复述。AMI数据集包含80小时训练数据，10小时开发集和10小时测试集。ICSI数据集包含67小时训练集以及5小时测试集。基线模型为三层LSTM，每层含1024个记忆单元，递归隐层后衔接512维的映射层用来降低模型参数。神经网络声学模型的输入层为均值方差规整后的11帧40维MS-LFB特征的拼接，输出层维度对应HMM的聚类状态数目。如4.3节所述，AMI和ICSI数据集采用的GCC-PHAT维度分别为588和1686。为做性能对比，实验结果表格中列出远场单通道和传统的多通道波束形成的结果。单通道实验使用麦克风阵列第一个麦克风的数据。由于BeamformIt工具包 [105]利用GCC-PHAT计算信道间到达时延以实现加权延迟相加波束形成，所以采用它做为波束形成的基线。

4.5.2 GCC-PHAT分析窗长选取

LSTM声学模型的输入为多通道信号提取的声学特征与多通道信道间GCC-PHAT特征的拼接。为了能够及时对说话人的位置信息变化做出响应，分析窗口沿着每对麦克风语音信号移动计算GCC-PHAT。若分析窗口较大，会导致说话人位置变化的分辨率降低。另一方面，如果分析窗口较小，用来计算信道间相关性的声学信号较短，会降低信道间相关性估计的鲁棒性。因此，窗长的选取需要在分辨率和鲁棒性之间权衡。我们首先分析当计算GCC-PHAT的窗长变化时，识别性能的变化。为与声学特征帧数一致，麦克风间相关性的计算同样采用10ms的窗移。

表4.1为AMI数据集的词错误率结果。由表格中的实验结果可知，声学模型在105ms的计算窗长获得最优性能，达到鲁棒性和分辨率的最佳平衡。因此，接下来的实验均采用105ms窗长计算GCC-PHAT。

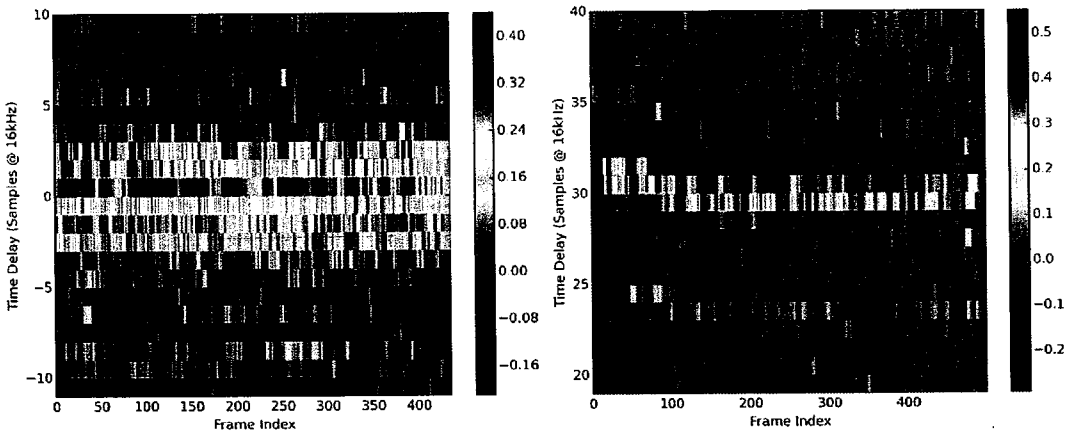
表 4.1 不同GCC-PHAT计算窗长的识别词错误率结果

Table 4.1 WER for different window sizes of GCC-PHAT

GCC-PHAT计算窗长(ms)	25	55	75	105	155
dev	36.6	36.4	35.9	35.8	36.5
eval	41.7	41.5	41.0	40.8	41.5

以AMI和ICSI数据集中的一句话为例，图4.4描述了麦克风阵列中前两个麦克风间的GCC-PHAT。图像的纵轴为时间延迟，横轴为每句话的帧索引。图

像的颜色代表GCC-PHAT的幅度。AMI数据集的时延范围在-10到10之间，图片的纵轴有21个取值。为方便与AMI比较，ICSI数据集只画出包含峰值在内的21个时间延迟。由图4.4可以看到，纵轴上对应GCC-PHAT最大值点即为估计的TDOA。



(a) AMI　　(b) ICSI

图 4.4 AMI和ICSI数据集上两麦克风间计算的GCC-PHAT

Figure 4.4 GCC-PHAT features between first two microphones on AMI and ICSI

4.5.3 基于GCC-PHAT空间特征补偿的多通道声学建模

本小节实验使用的声学模型如图4.5所示。为进一步提升模型鲁棒性，模型在训练阶段仍采用多任务学习框架。实验中包含三个基线声学模型：（1）在麦克风阵列第一通道数据上训练的单通道LSTM 声学模型；（2）将多通道信号经过波束形成算法增强为单通道数据，利用增强后的单通道数据训练的LSTM声学模型；（3）直接将多通道声学特征拼接训练的LSTM 声学模型。结果表格中带星号的测试集表示去掉重叠语音后的测试子集。以下对AMI和ICSI两个数据集上的识别结果进行分析。

表4.2为GCC-PHAT特征补偿实验在AMI数据集上识别词错误率结果（%）。对比单通道与多通道的实验结果，基于多通道数据训练的声学模型显著改善了远场语音识别性能。在不包含重叠语音的测试子集上，波束形成信号训练的基线模型表现得比多通道特征简单拼接训练的基线模型稍好，但在包含重叠语音的全测试集上，特征拼接模型的词错误率更低。此对比结果表明，在存在干扰

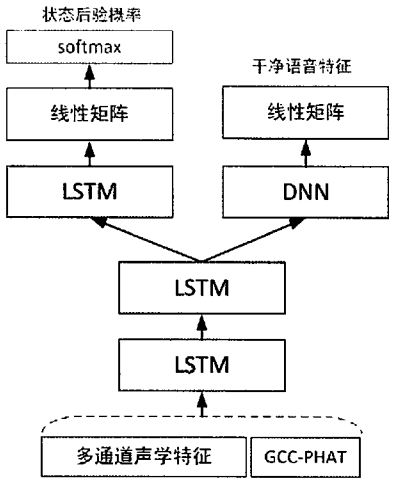


图 4.5 基于GCC-PHAT空间特征补偿的LSTM多通道声学模型

Figure 4.5 Diagram of an LSTM acoustic model with augmented GCC-PHAT inputs

声源的情况下，使用原始多通道特征拼接替代波束形成处理后的单通道特征可以使神经网络声学模型学习到更好的特征表示。

表格第四行为将多通道特征与588维GCC-PHAT特征拼接训练模型的识别结果。与上面两行多通道基线模型相比，基于GCC-PHAT空间特征补偿的模型在测试全集和不存在语音重叠的子集上都取得了显著的性能提升。多任务学习框架使模型又获得了绝对0.3~0.6%的WER降低。与多通道特征拼接训练的模型相比，最终的模型在测试全集和去除语音重叠片段的子集上平均取得相对5.7%的词错误率降低。与波束形成信号训练的模型相比，它在测试全集和去除语音重叠片段的子集上平均获得相对8.3%和4.1%的性能改善。

表4.3为基于GCC-PHAT特征补偿实验在ICSI数据集上识别词错误率结果(%)。对比第二和第三行两个多通道基线模型的识别结果发现，与波束形成的单通道语音信号训练的模型相比，简单的多通道特征拼接就可以获得大幅度性能提升。与多通道特征拼接训练的基线模型相比，加入空间特征补偿和多任务学习后，模型在测试全集和去除语音重叠片段的子集上获得相对3.0% 和4.3%的词错误率降低。

在训练阶段，多通道特征拼接训练的基线模型与图4.5所示的声学模型在训练集和验证集上的帧正确率变化如图4.6所示。图中横轴为迭代次数，纵轴为帧正确率(%)。由图可知，在两个数据集的训练集和验证集上，帧正确率都有明显的提升。总的来说，将编码说话人信息的GCC-PHAT做为辅助特征，与声学

表 4.2 基于GCC-PHAT空间特征补偿的识别结果—AMI数据集

Table 4.2 Performance of GCC-PHAT based spatial feature compensation on AMI

数　据	模　　型	dev	dev*	eval	eval*
单通道	-	42.8	34.3	47.2	38.3
多通道	波束形成	39.5	30.1	43.3	34.0
	多通道特征拼接	37.8	30.7	42.7	34.5
	+ GCC-PHAT	35.8	29.5	40.8	32.9
	+ GCC-PHAT + MTL	35.5	29.1	40.4	32.3

表 4.3 基于GCC-PHAT空间特征补偿的实验结果—ICSI数据集

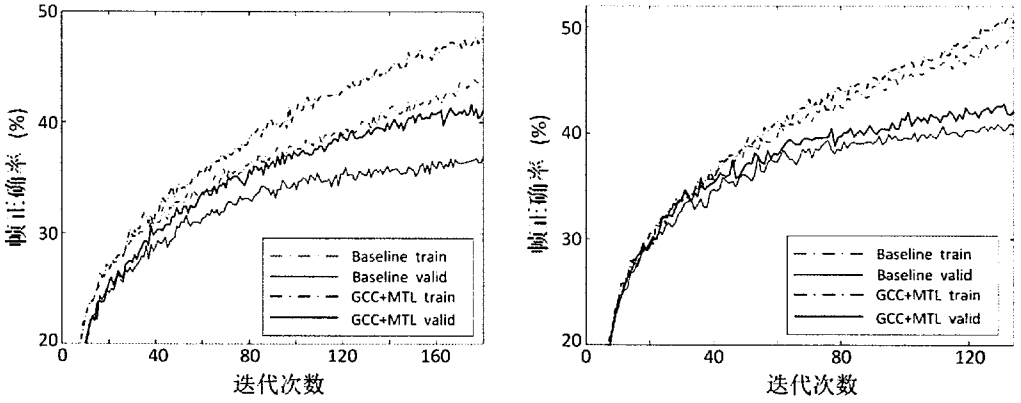
Table 4.3 Performance of GCC-PHAT based spatial feature compensation on ICSI

数　据	模　　型	eval	eval*
单通道	-	39.5	34.2
多通道	波束形成	33.8	27.4
	多通道特征拼接	29.8	25.2
	+ GCC-PHAT	29.2	24.5
	+ GCC-PHAT + MTL	28.9	24.1

特征同时输入神经网络声学模型，有效地提升了神经网络对多通道语音信号的建模能力。

4.5.4　基于时空信息的多通道声学建模

上一章提出基于注意力的LSTM模型，它通过注意力机制调整对输入层扩展上下文的关注度，并在多通道数据集上取得了一定程度的性能提升。然而，基于注意力的LSTM模型没有对多通道语音信号提供的空间信息做进一步的利用。本小节将上一章提出的注意力机制与基于GCC-PHAT的空间特征补偿相结合，图4.7为模型框图。将注意力权重加权后的特征向量与提取的GCC-PHAT特征拼接做为LSTM网络的输入。注意力机制利用输入层不同帧的时间



(a) AMI (b) ICSI

图 4.6 空间特征补偿模型在训练集和验证集上的帧正确率对比图

Figure 4.6 Frame accuracy on the validation set and training set during training

信息，GCC-PHAT辅助特征利用多通道语音信号提供的空间信息，两种方法分别通过时间信息和空间信息提升神经网络对多通道语音信号的建模能力。

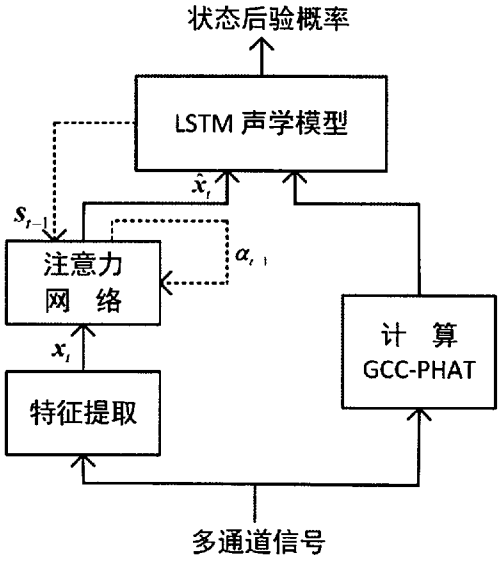


图 4.7 基于时空信息的LSTM多通道声学模型

Figure 4.7 Structure of an LSTM acoustic model for multichannel speech recognition

表格4.4和4.5为两个数据集上的识别词错误率(%)。表格第四行为上一章中多通道注意力模型的性能，第五行为4.5.3节中基于GCC-PHAT特征补偿模型

表 4.4 基于时空信息的多通道声学模型性能—AMI数据集

Table 4.4 Performance of the model using the spatio-temporal information on AMI

数　据	模　　型	dev	dev*	eval	eval*
单通道	-	42.8	34.3	47.2	38.3
多通道	波束形成	39.5	30.1	43.3	34.0
	多通道特征拼接	37.8	30.7	42.7	34.5
	+ 注意力机制	36.0	29.7	41.4	33.6
	+ GCC-PHAT空间特征	35.8	29.5	40.8	32.9
	+ 注意力机制 + GCC-PHAT空间特征	34.4	28.7	39.5	31.5

表 4.5 基于时空信息的多通道声学模型性能—ICSI数据集

Table 4.5 Performance of the model using the spatio-temporal information on ICSI

数　据	模　　型	eval	eval*
单通道	-	39.5	34.2
多通道	波束形成	33.8	27.4
	多通道特征拼接	29.8	25.2
	+ 注意力机制	29.0	24.5
	+ GCC-PHAT空间特征	29.2	24.5
	+ 注意力机制 + GCC-PHAT空间特征	28.6	23.9

的性能，最后一行是将以上两种方法结合后模型的识别结果。两种方法分别从不同的角度对模型做改进，即时间和空间。因此将两种方法结合的模型获得了最优性能。与多通道特征拼接训练的基线模型相比，基于时空信息的多通道声学模型在AMI和ICSI数据集上分别取得相对8.2%和4.0%的性能改善。

4.6　本章小结

本章提出了一种基于GCC-PHAT空间特征补偿的多通道深度神经网络声学模型，其在不需要多通道波束形成增强算法的前提下，有效地提升了深度神经网络对多通道语音信号的建模能力。本章首先简单回顾了传统框架的多通道语音识别框架。其次，对编码声源位置信息的相位变换广义互相关函数GCC-PHAT进行介绍。接着，总结了深度神经网络声学模型通过灵活使用输入特征

提高识别性能的相关工作。受上述研究工作启发，提出一种基于GCC-PHAT空间特征补偿的多通道声学建模方法，它将多通道语音信号提供的空间信息以辅助特征的形式加入到深度神经网络声学模型。将此方法与上一章的注意力机制结合后，基于时空信息的多通道声学模型在AMI和ICSI数据集上分别取得了相对8.2%和4.0% 的词错误率降低。

第五章　基于教师学生迁移学习的远场语音识别声学建模

5.1　引言

近年来，基于DNN-HMM的声学模型在自动语音识别领域取得了巨大的进步。在语音识别系统中，DNN可以被看作是一个分类器，它将输入层的声学特征映射到输出层的各个类中，即HMM的三音子状态，最终得到该声学特征属于每个状态的概率。DNN模型的有监督训练通常需要大量的标注数据。因此，获取正确的标注对于模型训练来说非常重要。在LVCSR系统中，为得到训练DNN的标注，声学特征的状态级标签一般利用语音的文本标注和已训练完成的GMM-HMM模型通过维特比算法对齐得到。由于标注质量很大程度上影响DNN 模型的性能，因此用于状态对齐的GMM-HMM模型需具备良好的性能。为了得到更准确的状态标注，文献 [106]中就利用训练好的DNN-HMM模型对训练数据重新做强制对齐。

状态级别对齐会得到每帧特征的三音子状态标注，当用于神经网络训练时，将状态标注转换为硬标签，即对应状态的目标值被设为1，并且所有其它状态的目标值为0。神经网络训练过程中不同的训练样本同时更新，因此可以认为每个类别的例子是其它类别的反例。然而，不同发音单元在音素层面可能会存在相似性，例如，/iy/ 和/ei/ 两个音素的结尾发音具有较强的相似性。为解决此问题，文献 [107]中依据状态的发音相似性来生成非0-1分布的目标向量。实验结果表明，此方法在孤立词识别任务上获得了相对20%的词错误率降低。另一方面，文献 [69]提出使用序列鉴别性准则对DNN 声学模型进行训练，目标向量为解码词图上计算得到的后验概率值。与0-1目标值交叉熵准则训练的DNN模型相比，此模型取得相对7~13% 的性能改善。

在远场场景中，语音信号受到噪声和混响的严重干扰，不同声学单元之间的区分性变得更加模糊。这种情况下，特征标注仅限于单个声学状态并非最佳解决方案。针对远场语音识别的声学建模，目标值位于0和1之间的软判决应该更适合远场语音声学模型的训练。因此，本章使用软判决标签代替0-1分布的硬判决标签训练远场语音声学模型。在第三章中，与噪声语音平行的干净语音数据被用来做为特征增强辅助任务的输出。本章探索如何利用平行数据对更有

效地建模远场语音，通过教师学生迁移学习框架得到每一帧声学特征的软判决标注，然后对远场语音识别的声学模型进行训练。

本章首先介绍迁移学习在声学建模中的相关研究应用，然后阐述本章实验所采用的基于教师学生迁移学习框架的远场语音识别声学建模方法，接着给出在AMI和ICSI单通道和多通道远场语音数据上的性能对比结果，最后一节对本章的工作内容做出总结。

5.2 迁移学习在声学建模中的应用

迁移学习 [108,109]是将知识从一个模型转移到另一个模型的机器学习方法。由于大部分的数据或任务存在相关性，迁移学习通常把已经训练好的模型学到的知识通过某种方式迁移到新模型来帮助加快或优化模型的训练。本节内容介绍迁移学习在语音识别领域的应用，选择迁移学习取得成功的两个突出领域进行介绍，即跨语言和多语言知识迁移与跨模型知识迁移。

5.2.1 跨语言和多语言知识迁移

早期关于跨语言和多语言的研究一般是基于语言学的映射，例如，通用的音素集或成对的音素映射。随着深度学习的普及，基于DNN的多语言声学建模受到研究者的广泛关注，它的基本思想是DNN模型可以看作是一个特征提取器，每个隐层输出都是对输入的一种特征表示，在低层学习到的特征一般是语言无关的，高层学习到的特征与语言更相关。因此，可以使用多语言数据训练多语言DNN声学模型，其中低层隐层参数在所有语言之间共享，高层隐层是特定语言相关的。Swietojanski等人在跨语种识别这方面做出了早期探索 [110]，其利用多语言数据初始化目标语言的DNN模型。在多语言识别任务中，DNN结构的隐藏层是跨语言共享的，每种语言都拥有自己的输出层 [91,111]。图5.1 为模型结构示意图，其属于第三章3.4.1节中多任务学习的一个应用实例。通过共享隐层结构，DNN可以利用多语言数据学习更好的隐层特征表示。上述的多任务学习可以扩展到更一般的框架。例如，文献 [112]中将音素识别和字形识别做为两个不同的任务训练DNN声学模型。

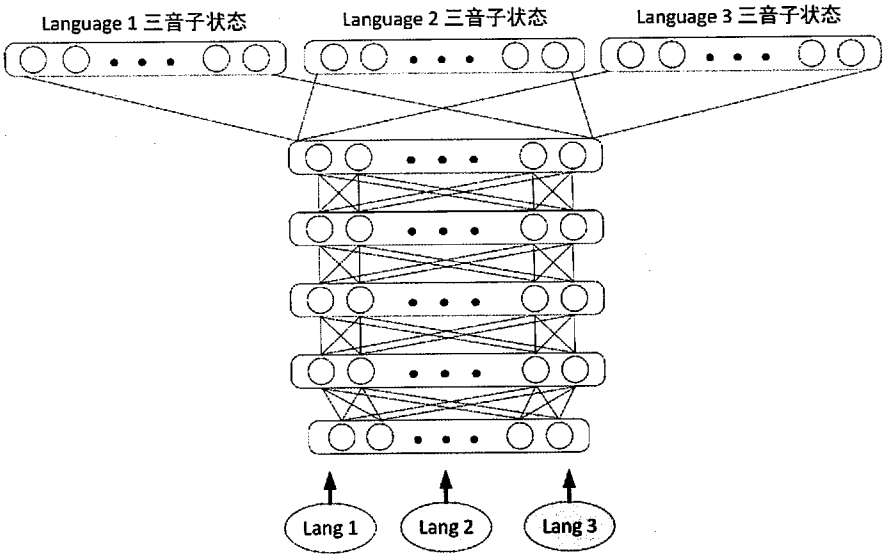


图 5.1 共享隐层的多语言深度神经网络模型示意图

Figure 5.1 Architecture of the shared-hidden-layer multilingual DNN

5.2.2 跨模型知识迁移

跨模型迁移学习框架一般是利用现有的复杂模型（教师模型）训练一个新的简单模型（学生模型）。传统的教师学生训练旨在训练一个复杂度低的模型，并使其性能接近高复杂度的教师模型的性能。学生模型试图使用通过教师模型获得的软判决标签来模拟教师的表现。因此，它一般用来做模型的压缩。图5.2为传统的教师学生训练框架示意图。教师模型为已训练好的复杂神经网络模型。将t时刻声学特征表示为$\boldsymbol{o}_t$，教师模型和学生模型的输出后验概率分别为$p_T(s|\boldsymbol{o}_t)$和$p_S(s|\boldsymbol{o}_t)$。学生模型的输出后验应尽可能的逼近老师模型的后验分布。用KL散度（Kullback-Leibler divergence）表示两个模型后验概率差异。因此，模型的训练目标为最小化下列函数：

$$D_{KL}(p_T||p_S) = \sum_t \sum_s p_T(s|\boldsymbol{o}_t) \log(\frac{p_T(s|\boldsymbol{o}_t)}{p_S(s|\boldsymbol{o}_t)}) \tag{5.1}$$

由于只更新学生模型参数，最小化公式5.1等价为最小化

$$E(\theta) = -\sum_t \sum_s p_T(s|\boldsymbol{o}_t) \log(p_S(s|\boldsymbol{o}_t)) \tag{5.2}$$

学生模型通过上述目标函数做误差反向回传更新。

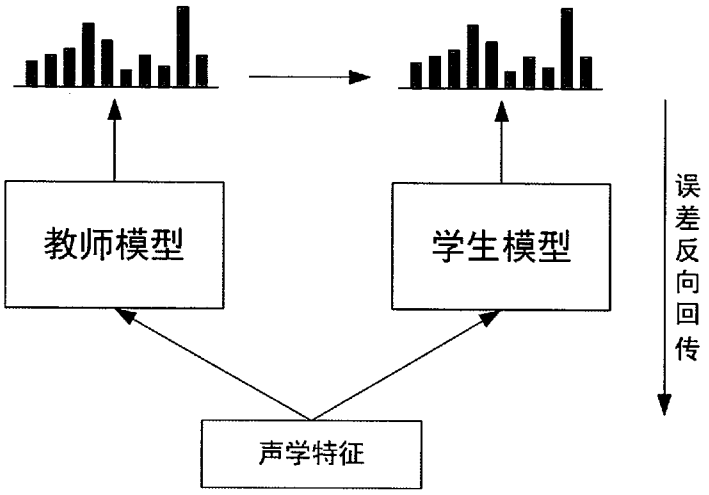


图 5.2 教师学生训练框架图

Figure 5.2 Framework of conventional teacher-student learning

Hinton等人在文献 [113]中提出知识提取的概念，并指出了使用软标签替代硬标签进行DNN模型知识提取的重要性。Chan等人将循环神经网络做为教师模型，利用其指导一个简单的前馈网络进行声学模型训练 [114]。文献 [115]中在大量数据未标注的情况下利用大规模的教师DNN训练小规模的学生DNN。与压缩模型的目的不同，另外一些研究使用教师学生训练模式来提升模型的鲁棒性。文献 [116]中首先使用数据仿真获得成对的干净语音和带噪语音，然后利用干净语音训练教师模型得到每帧特征的软判决标签，基于噪声语音训练的学生模型的目标函数为强制对齐获得的标注与软判决标签的结合。实验结果证明此方法有效提升了模型的鲁棒性。文献 [117]中为增强模型的抗噪能力，使用噪声数据和相应的增强数据做为学生和教师模型的输入。另一些研究则用教师学生训练框架做声学模型的自适应。文献 [118]在拥有少量自适应语料的条件下，使用教师学生训练框架将声学模型自适应到新的方言。Li等人利用此训练框架做声学模型的领域自适应 [119]。

5.3 基于教师学生学习框架的远场语音识别声学模型

远场语音信号受噪声和混响等因素的干扰，不同声学单元之间的区别性变得很模糊。在这种情况下，如果采用硬判决做为标注，对于声学模型训练来说是很有难度的。因此，含有更丰富信息的软判决标注更适合远场语音声学模型

的训练。为得到可靠的软判决标注，本章挖掘利用与噪声语音同步录制的近场语音信号中蕴含的信息。受文献 [119]中关于跨模型知识迁移研究的启发，本章采用相同的教师学生训练框架对远场语音声学模型做训练。图5.3为模型的训练流程图。

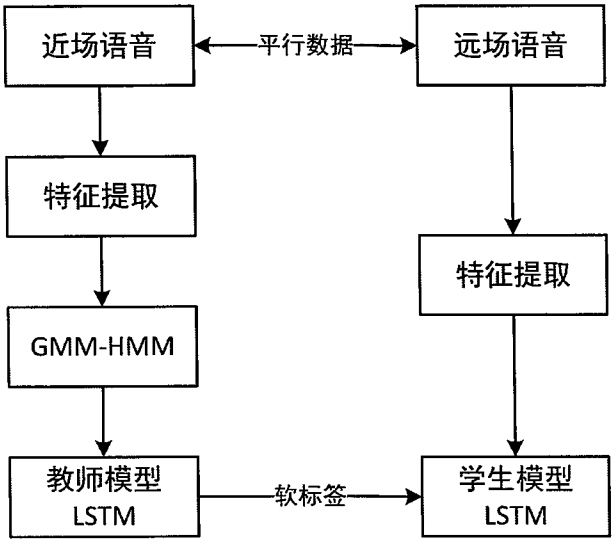


图 5.3 基于教师学生训练框架的远场语音识别声学建模流程图

Figure 5.3 Block diagram of teacher-student learning based acoustic modeling for DSR

这里以本节实验使用的LSTM声学模型为例，整个训练流程可以描述为以下三个步骤：

1. 训练教师模型：(a)提取近场语音的声学特征并训练GMM-HMM声学模型；（b）利用GMM-HMM模型对近场语音声学特征做强制对齐；（c）通过强制对齐得到状态级标注，并利用交叉熵准则训练LSTM声学模型。

2. 获得软判决标签：将近场声学特征通过步骤1训练的LSTM声学模型，得到每帧特征对应的后验概率分布。

3. 训练学生模型：提取远场语音的声学特征，使用步骤2计算的与其对应近场语音的软判决标签做为标注，通过误差反向回传训练与步骤1相同结构的LSTM声学模型。

若远场语音的声学特征和其同步录制的近场语音的声学特征分别表示

为$\boldsymbol{o}_{t,noisy}$和$\boldsymbol{o}_{t,clean}$参照公式5.2，学生模型训练的目标函数为

$$E(\theta) = -\sum_{t}\sum_{s} p_T(s|\boldsymbol{o}_{t,clean}) \log(p_S(s|\boldsymbol{o}_{t,noisy})) \tag{5.3}$$

上述步骤得到的学生模型即为远场语音识别系统中使用的声学模型。

教师模型是基于干净的近场语音数据训练的，将干净特征通过教师模型得到的软标签包含着比0-1分布的硬标签更丰富的信息，它代表着干净模型是如何看待样本数据的。例如，如果目标音素是/ah/，教师模型通常对于音素/ah/的声学状态节点有很高的概率。与此同时，由于发音相似性，它可能在其它元音音素上也具有合理的概率值，如/ax/和/ae/等，但是对于诸如/v/和/b/等辅音音素的概率非常低。当语音信号在嘈杂的环境中录制时，不同声学单元之间的区分性更模糊。对于DNN声学模型而言，某个状态概率为1其他状态概率为0的硬标注是很难训练的。因此，上述的软标签信息对于模型的训练会非常有帮助。图5.3中教师学生训练框架将干净模型学到的知识分享给噪声模型，干净模型提供的合理软标注让远场语音声学模型的训练变得更轻松更容易。

5.4 实验与分析

本章实验在AMI和ICSI两个数据集上进行。实验分为单通道（SDM）实验和多通道（MDM）实验两部分。针对单通道实验，5.4.2 节在LSTM网络上验证图5.3中学生模型对远场语音识别性能的改进，并将此方法与第三章提出的基于注意力机制和多任务学习的LSTM声学模型结合做进一步改进。5.4.3节中以上一章中提出的多通道声学模型为基线，同样利用教师学生训练框架对声学模型的目标值做软判决优化。

5.4.1 实验配置

实验采用的数据集与本文之前的实验一致，具体介绍可参见前一章4.5.1节的实验配置。本小节对实验中使用的神经网络参数做简单介绍。声学特征仍然采用40维的MS-LFB特征，均值方差规整后的特征做为神经网络输入。远场单通道实验采用麦克风阵列第一通道的数据训练模型。单通道和多通道中LSTM模型采用相同的配置：输入上下文扩展11帧；3层递归隐层，每层

含1024个记忆单元，递归隐层后衔接512 维映射层；输出层各节点对应HMM的聚类状态。实验结果表格中带星号的测试集为去掉重叠语音后的测试子集。

5.4.2　单通道实验

本小节在远场单通道数据上验证方法的有效性。表5.1和表5.2分别为声学模型在AMI和ICSI数据集上的词错误率（%）结果。表格中的前两行为11帧特征拼接训练的LSTM声学模型，后两行为第三章3.5.4节中基于注意力机制和多任务学习的LSTM声学模型的结果。

对比前两行结果，使用可靠的软判决标签后替代0-1分布的硬标签后，模型在两个数据集的近场和远场语音测试集上平均取得相对5.2%和2.0%的词错误率下降。对比后两行实验结果，在ALSTM-MTL声学模型上也可以看到类似的性能提升。实验表明，在语音信号受到噪声和混响干扰的远场场景中，合理的软判决标签更适合做为声学模型训练的标注。由于远场语音信号的声学区分性更模糊，软判决标注让神经网络声学模型的训练变得更容易，使模型在近场和远场场景下的鲁棒性均有提升。对比表格中第一行和第四行的识别结果，使用优化的软判决标签对ALSTM-MTL声学模型训练后，我们在远场单通道语音识别任务上获得了相对4.9%的性能改善。

表 5.1 基于教师学生训练框架的单通道实验结果—AMI数据集

Table 5.1 Performance on teacher-student learning in the SDM case of AMI

模　型	训练目标值	近场测试集		远场测试集			
		dev	eval	dev	dev*	eval	eval*
LSTM	硬标签	34.5	40.6	42.8	34.3	47.2	38.3
LSTM	软标签	32.7	38.3	41.8	33.4	46.1	37.3
ALSTM-MTL	硬标签	33.8	39.7	41.3	33.1	45.8	37.2
ALSTM-MTL	软标签	32.5	38.1	40.5	32.2	44.8	36.1

5.4.3　多通道实验

本小节在多通道数据上对比使用硬标注和软标注对模型性能的影响。表5.3和表5.4分别为AMI和ICSI 数据集上的词错误率（%）结果。表格中前两

表 5.2 基于教师学生训练框架的单通道实验结果—ICSI数据集

Table 5.2 Performance on teacher-student learning in the SDM case of ICSI

模 型	训练目标值	近场测试集	远场测试集	
		eval	eval	eval*
LSTM	硬标签	27.0	39.5	34.2
LSTM	软标签	25.7	38.7	33.4
ALSTM-MTL	硬标签	26.5	38.4	33.2
ALSTM-MTL	软标签	25.4	37.7	32.3

行为多通道特征拼接基线模型的识别结果，后两行为第四章4.5.4节中时间注意力机制与空间特征补偿结合的声学模型。

由两表格中的实验结果可知，软判决标注同样可以为远场多通道声学模型带来性能改善。首先对比前两行结果，当输入采用简单的多通道特征拼接时，使用软标签替代0-1分布硬标注训练声学模型，其在测试全集上平均获得相对3.6%的词错误率降低。将此方法与第四章提出的基于时空信息的多通道声学模型相结合，声学模型从网络结构，输入特征和目标值三方面进行优化，其取得了显著的性能提升。对比第一行和第四行识别结果，与0-1分布的硬判决训练的多通道特征拼接基线模型相比，最终的模型在两个数据集上平均取得相对9.8% 的性能改善。

表 5.3 基于教师学生训练框架的多通道实验结果—AMI数据集

Table 5.3 Performance on teacher-student learning in the MDM case of AMI

模 型	训练目标值	dev	dev*	eval	eval*
多通道特征拼接	硬标签	37.8	30.7	42.7	34.5
多通道特征拼接	软标签	36.3	29.3	41.1	32.7
+ 注意力机制 + GCC-PHAT空间特征	硬标签	34.4	28.7	39.5	31.5
+ 注意力机制 + GCC-PHAT空间特征	软标签	33.2	27.7	38.1	30.2

5.5 本章小结

本章针对远场语音声学模型训练的目标值展开研究，并将教师学生迁移学

表 5.4 基于教师学生训练框架的多通道实验结果—ICSI数据集

Table 5.4 Performance on teacher-student learning in the MDM case of ICSI

模　型	训练目标值	eval	eval*
多通道特征拼接	硬标签	29.8	25.2
多通道特征拼接	软标签	28.8	24.0
+ 注意力机制 + GCC-PHAT空间特征	硬标签	28.6	23.9
+ 注意力机制 + GCC-PHAT空间特征	软标签	27.8	23.0

习框架应用于远场语音识别声学建模，其通过从同步录制的近场语音中提取可靠的软标注目标值来提升远场语音声学建模能力。本章首先简单总结了迁移学习在语音识别声学建模中的两个主要应用，然后介绍了本章中所采用的基于教师学生迁移学习框架的远场语音识别声学建模方法。最后对此方法的有效性在两个数据集上做验证。其在远场单通道和多通道语音识别任务上均带来性能改善。此外，与前文提出的注意力机制和空间特征补偿结合后，模型取得了进一步性能提升。

第六章　总结与展望

6.1　论文总结

近年来，自动语音识别系统的性能显著提升，其原因可以归结为：（1）使用极具表达能力的深度神经网络建模声学后验概率；（2）拥有海量的训练数据；（3）使用通用计算图形处理器加速神经网络的计算。目前的语音识别技术在近场场景下已经取得了良好的性能。信息化时代的人们对于远场语音交互系统的需求日益增长，如人机对话、远程音视频会议、智能交互电视等实际应用。然而，在麦克风距离说话人较远的远场语音识别任务中，背景噪声、混响以及人声干扰等问题大幅度降低了识别准确率。声学模型作为语音识别系统的关键模块，对识别性能有重要影响。因此，本文针对基于深度神经网络的远场语音识别声学建模技术展开学习和研究工作。本文的组织结构为：

第二章介绍了自动语音识别系统的基本原理和框架，并着重阐述了声学模型相关知识，即神经网络和隐马尔科夫模型的基本算法及其在声学建模中的应用。

第三章介绍了实验中使用的LSTM网络结构，并针对远场语音识别任务提出一种基于注意力机制和多任务学习的LSTM声学模型。基于深度神经网络的声学模型一般简单地将上下文多帧特征拼接做为输入。然而，每帧声学特征对于当前时刻状态预测的贡献不一定是相同的。为解决此问题，本章提出的注意力机制使模型对不同时刻的输入特征调整关注度。为提升模型的噪声鲁棒性，训练阶段使用多任务学习框架，使其联合预测声学状态和干净特征。在AMI和ICSI两个数据集上的实验表明，该模型在远场单通道和多通道实验中分别取得了相对3.1%和4.5%的识别性能提升。

第四章研究了多通道语音识别的声学建模方法，并提出一种基于GCC-PHAT空间特征补偿的神经网络声学模型。传统的多通道语音识别系统一般将前端增强模型与后端声学模型分开训练，此方法并非最优解决方案。因此一些研究人员提出前后端联合优化的方法。然而，这些方法通常在前端增强模块中引入神经网络模型，导致用于识别的模型参数量较大。为此，本章提出一种基于GCC-PHAT空间特征补偿的多通道声学建模方法，其将多通道信号提供的空

间信息以输入特征的形式加入到神经网络声学模型。实验结果表明，该方法有效提升了神经网络对多通道信号的声学建模能力。将此方法与上一章的注意力机制结合后，与多通道特征拼接训练的基线模型相比，模型在AMI和ICSI数据集上分别取得了相对8.2%和4.0%的词错误率下降。

第五章介绍了迁移学习在语音识别声学建模中的两个重要应用，并将其中的教师学生迁移学习框架应用于远场语音识别的声学建模。由于远场语音信号不同声学单元之间的区分性更模糊，因此含有丰富信息的软判决标签更适合做为模型训练的目标值。为得到可靠的训练目标值，与远场语音同步录制的近场语音被用来训练教师模型，并将声学特征通过教师模型来获得软判决标签。实验结果表明，该方法有效提升了模型的鲁棒性。将此方法与上述提出方法结合后，系统识别性能获得显著提升。

6.2 主要工作和创新

本文从声学模型的网络结构，输入特征以及训练目标值三方面改进远场语音识别声学模型，主要工作和创新包含：

(1) 提出一种基于注意力LSTM和多任务学习的远场语音识别声学模型。注意力机制使模型自动调整对输入层上下文特征的关注度。与此同时，为提升声学模型在远场场景下的鲁棒性，模型在训练阶段使用联合预测声学状态和干净特征的多任务网络结构。

(2) 针对远场多通道语音识别任务，提出一种基于GCC-PHAT空间特征补偿的声学模型。此方法利用神经网络输入特征的灵活性，将编码声源位置信息的信道间广义互相关做为输入特征补偿。此外，该方法与注意力机制结合可进一步提升识别准确率。

(3) 将教师学生迁移学习应用于远场语音识别声学建模。该方法通过挖掘近场语音特征中蕴含的信息，获得合理的软判决标注，使远场声学模型的训练更容易，有效提升了声学模型的鲁棒性。另外，将此方法与上述提出方法结合后，模型在远场单通道和多通道语音识别任务上有显著性能提升。

6.3 未来工作展望

远场语音识别任务一直充满着挑战，尽管本文对于远场语音识别声学建模

提出一些改进方法，但还有许多工作值得进一步深入探索：

首先，在第三章基于注意力LSTM和多任务学习的远场语音识别声学建模中，文中提出将注意力机制加在LSTM模型的输入层。目前已有研究表明TDNN-LSTM声学模型在多个数据集上取得了比LSTM更好的声学建模能力 [120]。因此可以考虑使用TDNN-LSTM做为基线模型，在TDNN隐层上下文扩展中引入注意力机制，以获得更好的识别性能。

其次，在第四章基于空间特征补偿的多通道声学建模中，GCC-PHAT对声源位置估计的准确性直接影响着声学模型的性能。为了平衡GCC-PHAT对声源位置估计的鲁棒性和分辨率，本文通过实验遍历寻找GCC-PHAT的最佳计算窗长。然而，空间特征的提取算法仍有很大的改进空间。空间特征提取的过程中如何在提升分辨率的同时保持鲁棒性是下一步需要细化研究的问题。

最后，在第五章基于教师学生迁移学习的远场语音识别声学建模中，文中利用同步录制的近场语音数据和由其训练的教师模型来得到可靠的软判决标注。但在实际应用中，近场语音数据通常是不可获得的。因此，在同步录制的干净语音数据不可用的情况下，如何生成合理的软标注是下一步需要考虑的问题，此研究更具有实际工程应用的意义。