

西安电子科技大学

硕士学位论文



基于深度学习神经网络的语音增强方法研究

作者姓名 _____ 刘 浩 _____

指导教师姓名、职称 _____ 马鸿飞 教授 _____

申请学位类别 _____ 工学硕士 _____

学校代码 10701
分 类 号 TN91

学 号 1401120152
密 级 公开

西安电子科技大学

硕士学位论文

基于深度学习神经网络的语音增强方法研究

作者姓名：刘 浩

一级学科：信息与通信工程

二级学科：通信与信息系统

学位类别：工学硕士

指导教师姓名、职称：马鸿飞 教授

学 院：通信工程学院

提交日期：2017 年 6 月

Research on Speech Enhancement Method Based on Deep Learning Neural Networks

A thesis submitted to
XIDIAN UNIVERSITY
in partial fulfillment of the requirements
for the degree of Master
in Communications and Information Systems

By
Liu Hao
Supervisor: Ma Hongfei Professor
June 2017

西安电子科技大学

学位论文独创性（或创新性）声明

秉承学校严谨的学风和优良的科学道德，本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得西安电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同事对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文若有不实之处，本人承担一切法律责任。


本人签名： 刘浩 日期： 2017.6.12

西安电子科技大学

关于论文使用授权的说明

本人完全了解西安电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权属于西安电子科技大学。学校有权保留送交论文的复印件，允许查阅、借阅论文；学校可以公布论文的全部或部分内容，允许采用影印、缩印或其它复制手段保存论文。同时本人保证，结合学位论文研究成果完成的论文、发明专利等成果，署名为西安电子科技大学。

保密的学位论文在____年解密后适用本授权书。

本人签名： 刘浩 导师签名： 
日期： 2017.6.12 日期： 2017.6.12

摘要

语音是人们相互交流所使用的最基本的手段,然而在现实环境中,语音总是受到各种噪声的干扰。噪声的存在不但会降低语音的质量,还影响语音的可懂度。不仅如此,噪声干扰还会导致语音处理系统性能的急剧恶化。语音增强技术就是要抑制噪声,从带噪语音中提取出尽可能纯净的语音信号。传统的比较成熟的单声道语音增强方法有谱减法、维纳滤波法、基于统计模型的方法等,近十几年人们对语音增强又进行了许多新的探索,例如小波变换法,听觉掩蔽法等。

本文系统地研究了传统的单声道语音增强算法,研究过程中发现在低信噪比条件下,传统算法大都存在性能严重下降的缺点。为了提高增强的效果,本文对 BP 神经网络以及深度学习中的栈自动编码器和深度信念网络这两种主流模型进行了深入的研究,神经网络能够模拟人脑的工作原理,具有自学习能力和强大的非线性映射能力。在此基础上,本文提出了基于深度信念网络的噪声幅度谱估计语音增强方法,此方法中, BP 算法被用来对网络进行微调。此外,本文对子空间语音增强算法进行了改进。

在本文所提出的语音增强方法中,通过对深度信念网络进行预训练和微调,使之可以实现从带噪语音幅度谱到噪声幅度谱的非线性映射。在得到对带噪语音中噪声的幅度谱估计之后,使用带噪语音幅度谱减去网络估计的噪声幅度谱得到对纯净语音幅度谱的估计,最后利用人耳对相位信息不敏感的特性,使用带噪语音的相位信息采用重叠相加法恢复增强语音的时域波形,这种方法在低信噪比下具有较好的性能。

为了进一步提高增强语音的质量,本文训练了多个适应于不同类型噪声的深度信念网络,并在进行噪声幅度谱估计之前,加入了噪声分类模块。

本文分析了子空间语音增强算法中存在的不足之处:子空间方法中使用 VAD 检测对噪声进行估计,无法在语音帧及时更新噪声;在低信噪比条件下, VAD 检测的性能迅速降低,导致子空间语音增强算法性能也迅速下降。针对这两点不足之处,本文结合所提出的语音增强方法分别给出了改进方案。一种方案是在 VAD 检测的语音帧使用网络估计的噪声进行噪声的更新,第二种方案是不进行 VAD 检测,在每一帧都使用网络估计的噪声进行噪声的更新。

最后,本文在 MATLAB 平台上对子空间语音增强方法,基于深度信念网络的噪声幅度谱估计语音增强方法,两种改进的子空间语音增强方法进行了仿真并对比其性能。性能指标使用 PESQ 得分。仿真结果显示,基于深度信念网络的噪声幅度谱估计语音增强方法在低信噪比条件下,性能超过子空间法。两种改进的子空间方法的性能无论在高信噪比还是低信噪比条件下,性能均优于子空间法,并且,信噪比越低,较子空间法越优。

关 键 词： 语音增强， BP 神经网络， 深度学习， 深度信念网络， 栈自动编码器

ABSTRACT

Speech is the most basic means used by people to communicate with each other, but in reality, the speech is always disturbed by various kinds of noise. Noise would not only reduce the speech quality but also affect the intelligibility of speech. Besides, it can also lead to the performance of speech processing system becomes worse quickly. The purpose of speech enhancement technology is to suppress the noise, withdrawing the pure primitive speech from noisy speech as far as possible. At present, speech enhancement technology is widely used in speech coding, speech recognition, military, medical and many other fields. The traditional monaural speech enhancement methods include spectral subtraction, wiener filtering, enhancement methods based on statistical models and so on. In the past decades, several new researches have been made on speech enhancement, such as the methods based on wavelet transform and auditory masking.

Several traditional monaural speech enhancement algorithms have been studied in the thesis. In the process of algorithm analysis, it was found that in the low signal-to-noise ratio, the traditional algorithms usually performed terribly. In order to improve the effectiveness of the enhancement, the back propagation (BP) neural networks and two mainstream models of deep learning named stack automatic encoding (SAE) and deep belief networks (DBN) have been studied deeply as well. The neural networks are able to simulate the principle of the human brain which has the ability of non-linear mapping. Based on these, the speech enhancement method with the help of DBN was proposed, in this method, the BP algorithm was used for fine tuning. And at last the subspace speech enhancement algorithm was improved in the thesis.

The speech enhancement method based on DBN was carried out by pre-training and fine-tuning so that it can well learn the relationship between the noisy speech and the noise. After obtaining the estimation of the noise amplitude spectrum in the noisy speech, the estimation of the pure speech amplitude spectrum can be obtained by using the noisy speech amplitude spectrum minus the noise amplitude spectrum. Finally, because of the characteristic that the human ear is insensitive to the phase information, the time-domain waveforms of the enhanced speech were reconstructed by overlapping addition method with the phase spectrum of the noisy speech.

In order to further improve the quality of the speech, a number of deep belief networks that adapted to different types of noise were trained, and the noise classification module was added before noise amplitude spectrum estimation module.

It was found that there were two shortcomings in traditional subspace speech enhancement algorithm: Voice activity detection (VAD) was used in the method to estimate the noise, but it was unable to update the noise in time; the performance of VAD has reduced rapidly in low signal-to-noise ratio, which leads to the performance of subspace speech enhancement algorithm decline rapidly too, so two improvement programs were proposed respectively to solve the problems: One solution was to updated the noise with the estimated noise of the networks in speech frame detected by VAD, the second solution was updated noise in each frame with the estimated noise of the networks.

Finally, the subspace method, the proposed method and two improved subspace methods were simulated on the MATLAB platform. By comparing the performance of the methods, it was found that the proposed speech enhancement method was better than subspace method in low SNR, the performances of the two improved subspace methods were superior to the traditional subspace method in both high SNR and low SNR, and the lower the SNR was, the better the improved subspace methods performed.

Keywords: Speech Enhancement, BP Neural Networks, Deep Learning, Stacked Auto Encoder, Deep Belief Networks

插图索引

图 3.1	神经元的数学模型	17
图 3.2	BP 神经网络示意图	18
图 3.3	自动编码器基本结构	23
图 3.4	栈自动编码器的结构	24
图 3.5	受限玻尔兹曼机的连接方式	24
图 3.6	一叠 RBM 和对应的深度信念网络	27
图 4.1	基于 DBN 的噪声幅度谱估计语音增强方法框图	30
图 4.2	带噪语音幅度谱到噪声幅度谱之间的映射	30
图 4.3	特征提取框图	31
图 4.4	波形重构框图	31
图 4.5	深度信念网络的预训练和微调	33
图 4.6	高通滤波器的幅频特性及相频特性	35
图 4.7	浊音信号预加重处理前后的时域波形	35
图 4.8	浊音信号预加重处理前后的频谱	36
图 4.9	帧长帧移示意图	36
图 4.10	语音信号的分解和合成	37
图 4.11	各窗函数的合成误差	39
图 4.12	噪声分类方法系统框架	40
图 4.13	训练样本迭代次数对分类正确率的影响	41
图 4.14	训练样本数据量的大小对正确率的影响	43
图 4.15	SAE,BPNN 正确率随迭代次数变化图	44
图 5.1	PESQ 测度计算框图	49
图 5.2	不同数据量训练网络的语音增强效果	51
图 5.3	带噪语音处理前后时域波形及相应的语谱图（白噪声 SNR = 0dB）	53
图 5.4	四种语音增强方法的性能比较	54

表格索引

表 4.1	不同 BP 网络模型的正确率	42
表 5.1	MOS 评分表	48
表 5.2	SNN、DBN 语音增强方法平均 PESQ 得分	50

符号对照表

符号	符号名称
dB	分贝
Hz	赫兹
ms	毫秒

缩略语对照表

缩略语	英文全称	中文对照
AAI	AI-Articulation Index	清晰度指数
LPC	Linear Prediction Coding	线性预测编码
HMM	Hidden Markov Models	隐式马尔科夫模型
MMSE	Minimum Mean Square Error	最小均方误差估计
SVD	Singular Value Decomposition	奇异值分解
EVD	Eigen Value Decomposition	特征值分解
TDC	Time Domain Constrained estimator	时域约束估计器
SDC	Frequency Domain Constraint estimator	频域约束估计器
ANN	Artificial Neural Networks	人工神经网络
NN	Neural Networks	神经网络
DL	Deep Learning	深度学习
SAE	Stacked Auto Encoder	栈自动编码器
RBM	Restricted Boltzmann Machine	受限玻尔兹曼机
DBN	Deep Belief Networks	深度信念网络
BP	Error Back Propagation Algorithm	误差反向传播算法
SNN	Shallow Neural Networks	浅层神经网络
CD	Contrastive Divergence	对比散度
DFT	Discrete Fourier Transformation	离散傅里叶变换
MOS	Mean Opinion Score	平均意见得分
DAM	Diagnostic Acceptability Measure	满意度测量
PESQ	Perceptual Evaluation Of Speech Quality	语音质量感知评估
PSQM	Perceptual Speech Quality Measure	感知语音质量测量
SNR	Signal Noise Ratio	信噪比
VAD	Voice Activity Detection	语音活性检测

目录

摘要	I
ABSTRACT	III
插图索引	V
表格索引	VII
符号对照表	IX
缩略语对照表	XI
第一章 绪论	1
1.1 语音增强的意义	1
1.2 语音增强的研究背景及现状	2
1.3 神经网络的研究背景及现状	3
1.4 本论文的工作和章节安排	4
第二章 传统单声道语音增强算法研究	7
2.1 语音的特性	7
2.2 噪声的特性	8
2.3 单声道语音增强算法概述	9
2.4 子空间语音增强算法的原理	10
2.4.1 语音信号的线性模型	11
2.4.2 信号与噪声子空间	11
2.5 基于线性估计的子空间算法	13
2.6 本章小结	15
第三章 深度神经网络基础理论	17
3.1 神经元数学模型	17
3.2 误差反向传播算法研究	18
3.3 深度学习简介	21
3.3.1 浅层结构和深层结构	21
3.3.2 深度学习的训练机制	22
3.3.3 深度学习的主流模型	22
3.4 栈自动编码器	23
3.5 深度信念神经网络	24
3.6 本章小结	28
第四章 基于深度学习的语音增强方法	29

4.1	基于 DBN 的噪声幅度谱估计语音增强方法	29
4.2	窗函数的选择	34
4.3	噪声分类模块设计	39
4.3.1	BP 神经网络分类	41
4.3.2	栈自动编码器分类	43
4.4	子空间语音增强方法的改进	44
4.4.1	传统子空间语音增强方法的不足之处	44
4.4.2	改进的子空间语音增强方法	45
4.5	本章小结	46
第五章	实验结果及分析	47
5.1	语音质量评价标准介绍	47
5.1.1	主观评价	47
5.1.2	客观评价	48
5.2	基于 DBN 的噪声幅度谱估计语音增强方法性能分析	49
5.3	改进的子空间语音增强方法性能分析	52
5.4	本章小结	55
第六章	总结与展望	57
6.1	本文工作总结	57
6.2	展望	57
参考文献	59
致谢	63
作者简介	65

第一章 绪论

1.1 语音增强的意义

语言是人们首要的不可或缺的用于交流信息的工具。语音是语言的载体,是人们彼此之间互相传递信息的最基本,最有效,也是最便利的方法。然而,我们却生活在一个极其复杂的声学环境中。无论身处何处,我们都被噪声所环绕。例如,在街道上汽车的喇叭声,商场所播放的音乐声,街道上施工声,办公室里的电脑风扇声,空调的噪声,别人的讲话的声音等等。语音也通常会被这无所不在的噪声所干扰。噪声不仅会导致语音的质量下降,使人们听起来感觉刺耳、粗糙,而且会导致语音处理系统的处理能力迅速降低^[1]。例如,噪声会对语音处理系统中的预测参数产生影响,在使用线性预测编码(Linear Predictive Coding, LPC)参数的方法进行低误码率语音通信时,汽车发动机的噪声,街道上的施工噪声等都会对其造成影响,从而使得重构的语音可懂度大幅下降,以至于很难被人理解。所以,在复杂噪声环境下,为了使语音处理系统仍能够高效率地运行,必须对受污染语音采用增强算法处理。

语音增强技术是指当纯净的语音被噪声污染之后,尽量从受污染的语音中恢复出原始的语音信号。并使处理后的语音具有高度舒适感,可懂度和清晰度。从而使接收者乐于接受。语音增强的目的就分为两个方面,一是主观方面,二是客观方面。主观方面是指抑制语音中的噪声,提升语音的质量,使接收者容易接受。客观方面是指提高语音的可懂度。然而通常情况下,它们是互相冲突的。因此在应用中,通常会依据实际的情况选择不同的语音增强算法,使增强处理后的语音在主观或者客观度量标准上有所侧重。例如,在语音识别方面的应用上,就侧重于客观度量标准,着重提高其可懂度。在通话方面的应用上,就侧重于主观度量标准,着重提高人耳听觉的舒适感。

语音增强技术被广泛地应用于许多领域^[2],如:

(1) 语音编码方面的应用。语音编码技术在压缩语音信号传输带宽和增加信道传输速率方面担当重要角色。近年来,语音通信领域对高效的编码技术的需要日渐迫切,然而,这要求语音信号是尽量纯净的。所以,在进行编码之前,有必要对带噪语音进行增强。

(2) 语音识别方面的应用。使用人类的语言与计算机交流,让计算机听懂人类语言的含义,这是人们长久以来不断追求的事情。然而,即使是目前最优秀的识别工具,噪声的存在也会导致其工作时稳定性的明显降低,因此,在进行语音识别前,也有必要对带噪语音进行增强。

(3) 医疗方面的应用。医生在为病人诊断时使用的诊断仪器里存在放大电路,助

听器在放大病人心跳声的同时，也会放大环境噪声，为了使医生能够听清楚病人的心跳声从而更加准确地进行诊断，在仪器里就需要加入语音增强技术。

(4) 军事方面的应用。比如战斗机飞行员在和地面指挥站进行通话时，由于战斗机引擎噪声的影响，飞行员的话会被污染，甚至会完全听不清楚，这时就需要进行语音增强。

总之，语音增强技术在很多领域都有广泛的应用，意义重大。

1.2 语音增强的研究背景及现状

二十世纪初，随着电话的使用逐渐普遍，人们开始研究语音的可懂度^[3]。Bell 实验室首先提出了清音和浊音的概念。这一时期，语音可懂度的清晰度指数 (AI-Articulation Index, AAI)及其计算方法也被提出。

二十世纪五十年代，声压计和语谱图等工具被人们大量应用在建筑声学，语言学，噪声控制等方面的研究中。大量的实验研究表明，语音可懂度与单词的出现频率和单词的长度有关^[4]：出现频率高的，单词长的容易被听懂。

二十世纪六十年代，人们在语音增强理论中提出了维纳滤波理论^[5]和卡尔曼滤波理论^[6]。

二十世纪七十年代，得益于电子计算机超快的运行速度，快速傅里叶变换被提出并得到实际应用，从而一些新的语音增强方法被提出。例如梳状滤波器语音增强方法。其中，1979 年 Boll 提出的谱减法^[7]最具影响力。Berouti 对谱减法进行了改进，增加了过减因子和谱下限参数，这两个参数使谱减法具备了极大的灵活性^[8]，可以大幅度地减少音乐噪声，显著地提高了谱减法的性能。谱减法具有结构简单且实用的优点，是抑制宽带背景噪声使用最多的一种语音增强技术。

二十世纪八九十年代，基于统计模型的方法逐渐成为语音增强研究中的重点。如贝叶斯估计器、最大后验估计器等。隐马尔科夫模型方法 (Hidden Markov Model, HMM)^[9]是这些方法的典型代表，它不仅可以应用于语音识别任务中，而且可以将语音增强任务加入其框架中。

进入二十一世纪之后，移动通信得到飞速的发展，人们对语音增强技术的研究又做出了许多探索。一方面，别的领域中的许多好的算法被借鉴到语音增强的研究中，例如基于小波变换的增强方法^[10]，基于人耳听觉掩蔽效应的增强方法^[11]，基于信号子空间的增强方法^[12]，基于神经网络的增强方法^[13-15]等。另一方面，得益于计算机运算速度的提高，许多以往因计算复杂度高而无法实用的算法得以应用于实际的语音处理系统。

基于小波变换的语音增强方法的原理是：在不同的频段中对信号进行分解，选择

适当的阈值，将阈值之下的系数去掉，最后用剩余的系数对语音信号重构。基于听觉掩蔽的语音增强方法的原理是：利用人耳的无法很好的分辨出语音信号中强度较低的信号的听觉特性，人们在语音增强研究中提出了听觉掩蔽来提高算法的性能。基于信号子空间的语音增强算法的主要思想是：带噪语音信号空间可以分解为一个噪声信号占主导地位的子空间和一个语音信号占主导地位的子空间，可以采取置零落在噪声信号子空间中的带噪语音分量来估计纯净语音信号。基于神经网络的语音增强方法的原理是：用带噪语音和纯净语音数据对训练网络，利用深层神经网络强大的非线性映射能力，将带噪语音信号的功率谱映射为纯净语音信号的功率谱，这就使网络具备了去除噪声的能力。本文的一部分工作就是在此研究基础上，结合深度学习的方法对噪声的幅度谱进行估计从而进行语音增强。

1.3 神经网络的研究背景及现状

人工神经网络的发展历史，按照其发展的历程，可以大致分为下面的三个阶段：

第一阶段—启蒙时期

1943 年，第一个人工神经元的模型被生物学家 McCulloch 和数学家 Pitts 共同提出^[16]，这拉开了人们研究神经网络的序幕。1949 年，生物学家 Hebb 提出了 Hebb 法则^[17]，认为神经元之间的连接强度是可以改变的，而且这种可塑性被认为是人脑能够进行学习基础。1952 年，著名的非线性动力学微分方程被生物学家 Hodgkin 和 Huxley 提出。这一方程解决了神经元的许多问题，具有巨大的理论价值，如解决自激振荡，多重稳定问题等。1960 年，ADALINE 神经网络模型被科学家 Widrow 和 Hoff 联名提出^[18]。这两位科学家还研究了线性可分的问题。

第二阶段—低潮时期

1969 年，Minsky 和 Papert 在《Perceptrons》一书中指出单层的感知机不能很好解决线性不可分的问题。这个结论的提出使人们研究神经网络的热情遭到极大的冲击，并导致了神经网络发展历史上很长一段时期的低潮期。造成神经网络研究进入低潮期的另一个原因是：二十世纪七十年代后，集成电路迅速发展起来，微电子技术也进入全盛时期，人工智能在逻辑符号处理方法上取得了巨大成功，这就掩盖了人们研究新型神经网络的迫切性，许多的科学家将其研究转入了别的领域。当然，仍然有一部分学者在继续着神经网络的研究，并有另一部分新的学者投身到这一研究中，这为神经网络研究的低潮期添加了一笔亮色。

第三阶段—复兴时期

这一时期是神经网络研究发展的黄金时期。1982 年，美国国家科学院在其期刊上刊登了 Hopfield 模型的理论^[19]。该模型对神经网络的数据存储和输出等功能做了数

学上的综合概括,首次提出了学习算法的概念。1983年,Kirkpatrick在NP完全组合优化问题的求解中引入了模拟退火算法^[20]。1986年,Hinton提出了可大规模并行学习的Boltzmann机,在Boltzmann机中,Hinton首次提出并引入了隐单元的概念^[21]。1988年,由Rumelhart和McClelland提出了目前备受人们关注,应用范围最广的经典神经网络学习算法:误差反向传播算法^[22](Error Back Propagation, BP)。

然而随着神经网络层数加深,网络的性能并没有继续变好。科学家们研究发现,随着网络层数的增加,会产生梯度弥散的现象,这导致网络的参数更新非常缓慢。2006年,Hinton在《科学》发表了一篇文章^[23],文章中指出,人工神经网络具备强大的学习能力,但是如果网络深度过深,会造成学习上的困难,文章同时提出了解决这一问题的方法:引入“贪心逐层预训练”(Greedy Layer-wise Pre-training)的机制来解决这一困难。至此,人们在对深度神经网络的研究上取得了里程碑式的进展。随后,深度学习领域出现了几种主流模型,如深度信念网络(Deep Belief Networks, DBN),栈自动编码器(Stacked AutoEncoder, SAE)等。神经网络已经成为人们解决许多棘手问题的有利工具,但是要知道,目前,人类对大脑的工作原理的认识仍然很肤浅,相信随着科技的发展,神经网络的研究能够实现更加重大的进展。

1.4 本论文的工作和章节安排

本文的工作是在充分研究传统的经典语音增强算法和神经网络及深度学习的理论基础上,提出了使用栈自动编码器对噪声进行分类的方案,并在噪声分类的基础上提出了自己的基于深度信念网络的噪声幅度谱估计的语音增强方法,最后对子空间语音增强方法中采用话音活动检测来进行噪声估计的两个不足之处提出了自己的改进方案。

第一章主要论述了语音增强的研究意义,并分别介绍了语音增强算法和神经网络的发展历史及它们的研究现状。

第二章主要首先介绍了研究语音增强所需要了解的语音和噪声的特性,然后概述了传统的单声道语音增强算法,并重点研究了子空间语音增强算法的原理,语音的信号估计的基于线性估计器的方法。

第三章首先介绍了神经网络中的基础理论,如神经元的数学模型,神经网络的结构,然后深入研究了神经网络中应用最为广泛的BP算法,最后介绍了深度学习中的主流模型栈自动编码器和深度信念网络。

第四章首先介绍了本文所提基于DBN的噪声幅度谱估计语音增强方法,然后对语音分解合成进行实验,解释了窗函数选择为三角窗的原因,然后整个算法中的分类模块进行了详细介绍。最后针对子空间语音增强算法中存在的不足之处提出了相应的

改进方案。

第五章首先对几种常见的语音增强性能评测标准进行了分析,然后根据标准对本文提出的基于深度信念神经网络的噪声幅度谱估计增强方法进行性能评测。最后对本文所提出的两种子空间语音增强算法的改进方案进行仿真并分析其性能。

第六章对全文进行总结,指出本文所提的噪声幅度谱估计的语音增强算法的优缺点。以及本论文需要进一步研究的方向。

第二章 传统单声道语音增强算法研究

迄今为止,人们已经对语音增强研究了很长一段时间了,但是却并没有找到一种可以应用于多种场合的有效的增强办法。这不但是因为语音信号的处理涉及到各个学科领域的研究,例如声学处理领域,语音学,语言学,甚至医学上的人耳的构造以及感知特性,人的心理活动等方面。不同种类的噪声,甚至是同一种噪声在其不同强度下,都会对语音处理系统产生影响。所以,在进行语音增强的研究之前,我们首先要对语音和噪声的特性有一个充足的了解。

2.1 语音的特性

人类发出的语音是一种连续的波,因此,语音具备一般声波所具备的某些特性。但同时,它也具备着自己特有的特性^[24],例如不同人所发出的声音具有不同的音色、强度等。

(1) 清音和浊音信号

人类的声道可以被看作是一个滤波器,它可以对来自声带的气流进行调整从而产生不同的音色。声带可以为声道提供激励。由于声带所处的状态的不同,声带所提供的激励可以是周期的,也可以使非周期的。当这个激励是周期的,产生的语音信号就称为浊音信号,当这个激励是非周期的,产生的语音信号就称为清音信号。浊音信号的周期性并不是始终不变的,随着声带开启和关闭的快慢,浊音的频率也发生变化。清音和浊音在发声的方式上存在着很大的区别,这也导致了它们的特性存在很大差异。

(2) 语音信号的短时平稳性

语音信号是一种随时间不停变化的非平稳随机信号。但是,由于发声的过程中声带声道的运动变化有一定的限度,通常可以认为在较短的时间内具有一定的稳定性。因此,在处理语音信号的时候,不应当在较长的时间段内对其进行分析。而应该将其分割为较短的语音段,这样做,就可以对语音信号采用分析平稳随机过程的方法处理。

语音信号的另一特性是频率范围有限且相对集中。在对语音信号进行频谱分析时可以发现其频率主要分布在 300-3400Hz 范围内。

(3) 语音的听觉感知特性

人耳对语音的感知特性在研究中发挥着重大的作用,这是因为语音增强的最终目的是为了^[25]提高人的听觉感受。研究过程中人们发现人耳对环境中的噪声有着极强的自适应抑制能力。了解其中的原理可以大大帮助人们对语音增强的研究。尽管现阶段

在这方面的研究还不够透彻，但仍有一些可以应用在语音增强方法中的结论：

人耳主要对语音信号的幅度谱变化敏感，对相位谱的变化不敏感。

人耳对语音频率高低的感受与频率值的对数大致成正比。

人耳具有听觉掩蔽效应，强度大的声音会遮盖强度小的声音，使人耳感觉不到这个强度小的声音。并且研究发现掩蔽的程度是声强与频率的二元函数。

浊音所具有的共振峰结构对语音信号的可懂度非常重要，尤其是第二共振峰。

人耳有很强的自动识别能力，在两人或两人以上的多人会话中，人耳能够准确地识别出需要听取的声音。

2.2 噪声的特性

噪声无所不在，来源非常广泛。物体无规则震动多产生的声音是噪声，不希望听到、导致人烦闷的声音是噪声，声音过大甚至会引起人听觉受损的声音是噪声，打扰别人工作，影响别人休息的声音是噪声等等。噪声的特性复杂多样，可以是加性的或者是非加性的。本文主要讨论加性噪声。加性噪声大致上分为周期噪声、脉冲噪声、宽带噪声、同声道语音等^[26]。

(1) 周期噪声

周期噪声具有周期性。例如电风扇，电动机，汽车引擎等所产生的噪声就是周期噪声，我们所使用的交流电也会产生周期噪声。周期噪声在频谱上通常表现为离散的窄带，可以使用梳状滤波器来进行滤除。

(2) 脉冲噪声

脉冲噪声是不连续的，在时域波形中突然升起的不规则的幅值。主要来源比如爆炸的冲击，电磁攻击，突然的大叫声，打雷声等。对此类噪声的消除，可以计算一段时间内带噪语音在时域振幅的平均值，并把这个平均值作为一个阈值，当某一点处的振幅值大于这个阈值时，就认为该处存在脉冲噪声，将这一点处的振幅值除去。

(3) 宽带噪声

宽带噪声很常见也很难消除的一种噪声。由于它只能在无语音段独立存在，所以要去除这类噪声，就只能在无语音段时对宽带噪声的均值进行估计。

(4) 同声道语音

同声道语音是指信道在传输我们所希望听到的语音的同时，也同时传输着我们不希望听到的语音。人耳的鸡尾酒会效应，可以从混杂的声音中提取出他感兴趣的声音，但鸡尾酒会效应在单声道中同时存在多个声音的情况下，效果将会明显下降。可以利用希望声音和干扰声音在基音周期上的差别对它们进行分辨。

2.3 单声道语音增强算法概述

语音增强算法的研究已经经历了几十年发展,在人们不断的探索中取得了丰富的研究成果,有诸多不同的增强算法被提出。按要处理的语音的通道数来划分,语音增强分为单声道语音增强和多声道语音增强。两者相比,多声道语音增强可以更加轻松地实现增强任务,但单声道语音增强对硬件成本要求低。本文主要研究的是单声道语音增强。

单声道语音增强算法也分为很多类,但是现实环境中噪声种类和特性的多种多样,各种增强算法应用于不同类型的噪声时,其效果也各不相同。根本无法找到一种通用的算法来处理各类不同的噪声。总体而言,目前应用最多,比较流行的单声道语音增强算法有以下几种:

(1) 谱减法

谱减法(Spectral-subtractive)是迄今为止最容易理解和实现的增强算法,该算法基本原理是:假设带噪语音信号是短时平稳的,利用带噪语音中前面几帧无语音段来估计噪声的功率谱,在之后对带噪语音每一帧的处理中,都进行带噪语音功率谱减去噪声功率谱的操作,这样得到的就是增强语音的功率谱。由于人耳无法很好感知相位谱的变化,所以利用带噪语音的相位谱结合估计的增强语音的功率谱进行时域波形的重构。谱减法的实现简单,算法难度低,做为一个经典的增强算法而被广泛使用。但是谱减法会产生具有多频音音质的音乐噪声。目前,谱减法有着多种的改进,如引入过减因子和谱下限参数的谱减法,非线性谱减法,多带谱减法^[27]等。

(2) 维纳滤波法

谱减法利用了加性噪声的特点,认为可以通过从带噪语音的信号谱中减去噪声谱来得到纯净语音的信号谱。而维纳滤波法则是通过数学上易于处理的最优化均方误差准则来计算得到增强信号^[28]。维纳滤波器是这样一个系统:使得其输出信号 $\hat{d}(n)$ 在均方误差意义上尽量逼近期望信号 $d(n)$,这可以通过计算估计误差 $\hat{e}(n)$ 并使其最小化来实现。将这个系统的输入信号替换为带噪语音,期望信号替换为纯净语音,就可以将这个系统用于语音增强的目的。维纳滤波器也有着很多的变体,如平方根维纳滤波器,参变维纳滤波器等。和谱减法相比,维纳滤波法不会残留很大的背景噪声。

(3) 基于统计模型的方法

基于统计模型的方法可以充分利用语音信号和噪声信号的统计特性^[29],它以谱估计为理论基础,采用均方误差估计准则来得到增益函数。该方法在建立模型之后需要经过一个训练的过程来获得最初的参数,并且在模型的使用过程中还要依据实际数据来对模型的参数进行实时的更新,以便模型能够更加适应实际的情况。基于统计模型

的方法里主要的代表有：最小均方误差估计器(Minimum Mean Square Error, MMSE), 对数谱估计最小均方误差估计器(Minimum Mean Square Error Log-Spectral Amplitude, MMSE-LSA)等。另外，在统计模型方法中，可以集成人耳听觉掩蔽效应，语音存在概率等方法，将这些方法和统计估计器集成之后，可以明显地减少残留噪声。

(4) 子空间法

子空间法主要是基于线性代数上的数学理论，具体而言，该类算法是基于纯净语音信号的空间可以被视为带噪语音信号欧式空间的一个子空间的理论。即可以将带噪语音信号向量分解为纯净语音占主要地位的子空间和噪声占主要地位的子空间。采用将噪声子空间变为零的办法，就可以得到对纯净语音信号的一个估计。而要将带噪语音信号空间分解为两个互相正交的信号子空间和噪声子空间，可以采用奇异值分解(Singular Value Decomposition, SVD)和特征值分解(Eigen Value Decomposition, EVD)。

以上所介绍的增强方法中，比较成熟的是谱减法和统计模型的方法。其中谱减法及其变种形式因出现的时间早，算法简单易实现的特点，是目前使用最广泛的方法。在加入了过减因子和谱下限参数后谱减法可以调节语音失真大小并控制残留音乐噪声的多少。但是，因为对信号谱估计始终存在着误差，所以不可能完全避免音乐噪声的残留。基于统计模型的方法引入了最优化的均方误差准则，可以有效地减少增强后语音中的音乐噪声，但是它缺少控制语音失真和残留噪声多少的有效机制，因而其应用也受到了限制。

将子空间语音增强方法和其他的增强方法进行对比可以发现，子空间的方法有着残留的噪声少，引起的语音失真小，并且音乐噪声不明显等优点。子空间的增强方法还具有统计模型方法所不具有的控制残余噪声和语音失真关系的机制。基于子空间算法所具有的种种优势，目前，其已成为国内外语音增强研究的热点问题。因此，本文在研究了多种传统的语音增强方法的基础上，重点对子空间增强方法进行了研究。

2.4 子空间语音增强算法的原理

上文介绍了子空间语音增强算法的基本原理，即带噪语音信号空间可以被分解为两个互相正交的子空间：纯净语音占主导地位的子空间和噪声占主导地位的子空间。进行空间分解的常用的方法为 SVD 和 EVD。由于使用 EVD 分解得到的纯净语音信号的最优估计和使用 SVD 分解得到的纯净语音信号的最优估计可以相互转换。然而，对比这两种分解方法，EVD 的方法具有清楚的物理含义，方便理解。现阶段大多数的子空间增强方法都是基于特征值分解的，因此，本文主要研究基于 EVD 的子空间语音增强方法^[30]。

子空间的增强算法中,需要进行以下假设:噪声与语音信号是相互正交的零均值的随机过程,且具有各态历经性,语音信号具有短时平稳的特性。

2.4.1 语音信号的线性模型

假设纯净语音信号为 s , 将其通过一个无失真的信道传输, 在传输的过程中受到了加性噪声 n 的污染, 得到了带噪语音 y , 则 y 可以表示为:

$$y = s + n \quad (2-1)$$

上式中 $s = [s_1, s_2, \dots, s_k]^T$, $n = [n_1, n_2, \dots, n_k]^T$, $y = [y_1, y_2, \dots, y_k]^T$, 可以将矢量 s , n , y 看作是空间 C^K 的一部分。

用线性模型来表示纯净语音信号:

$$s = Vz = \sum_{i=0}^M z_i v_i \quad M \leq K \quad (2-2)$$

$z = [z_1, z_2, \dots, z_M]^T$ 是 M 个均值为零的随机变量构成的矢量, $\{v_1, v_2, \dots, v_M\}$ 是 M 维的相互之间线性独立的基向量, 即矩阵 V 的秩等于 M 。

2.4.2 信号与噪声子空间

带噪语音信号矢量可以表示为:

$$y = Vs + n \quad (2-3)$$

求 y 的协方差矩阵:

$$R_y = E\{yy^T\} = VR_s V^T + R_n \quad (2-4)$$

式中, R_y 、 R_s 和 R_n 分别表示带噪语音、纯净语音和噪声的协方差矩阵, 当噪声种类为白噪声时, 可求得 $R_n = \sigma_n^2 I$ 。对式(2-4)中的矩阵进行 EVD 分解, 可得它们的特征值具有如下关系:

$$\lambda_y(k) = \begin{cases} \lambda_s(k) + \sigma_n^2, & k = 1, 2, \dots, M \\ \sigma_n^2, & k = M + 1, \dots, K \end{cases} \quad (2-5)$$

矩阵 R_y 的 EVD 分解为:

$$R_y = U \Lambda_y U^T \quad (2-6)$$

$$\Lambda_y = \text{diag}[\Lambda_{y,1}, \sigma_n^2 I] \quad (2-7)$$

$$\Lambda_{y,1} = \text{diag}(\lambda_y(1), \lambda_y(2), \dots, \lambda_y(M)) \quad (2-8)$$

矩阵 R_s 的 EVD 分解为:

$$R_s = U \Lambda_s U^T \quad (2-9)$$

$$\Lambda_s = \text{diag}[\Lambda_{s,1}, 0I] \quad (2-10)$$

$$\Lambda_{s,1} = \text{diag}(\lambda_s(1), \lambda_s(2), \dots, \lambda_s(M)) = \Lambda_{y,1} - \sigma_n^2 I \quad (2-11)$$

称对角阵 $\Lambda_{y,1}$ 中包含的特征值为 R_y 的主特征值。将 U 写为 $U = [U_1, U_2]$ 的形式, 则 U_1 为 $K \times M$ 阶的矩阵, 它是由矩阵 R_y 的主特征向量构成的。即:

$$U_1 = \{u_k : \lambda_y(k) > \sigma_n^2\} \quad (2-12)$$

矩阵 U 是正交的, 则有:

$$I = U_1 U_1^T + U_2 U_2^T \quad (2-13)$$

不难看出矩阵 $U_1 U_1^T$ 是一个正交的投影矩阵。称 U_1 的列向量所张成的子空间为“信号”子空间, U_2 的列向量所张成的子空间为“噪声”子空间。这两个子空间互补。

概括起来, 使用 EVD 对带噪语音向量 y 进行分解:

$$\begin{aligned} y &= U_1 U_1^T y + U_2 U_2^T y \\ &= y_1 + y_2 \end{aligned} \quad (2-14)$$

由于假设了语音和噪声都是零均值的, 则:

$$E\{U^T y\} = 0 \quad (2-15)$$

$$\text{cov}\{U^T y\} = \text{diag}(\Lambda_{y,1} + \sigma_n^2 I, \sigma_n^2 I) \quad (2-16)$$

$$\text{cov}\{U_2^T y\} = \sigma_n^2 I \quad (2-17)$$

这表明, 我们可以认为 y 投影的分量 y_2 中不包含语音信号。因此, 在使用子空间法对纯净语音进行估计时, 可以简单地直接去除掉投影到噪声子空间中的分量, 这样所得到的增强信号 \hat{S} 为:

$$\hat{S} = U_1 U_1^T y = Hy \quad (2-18)$$

式中 H 是一个 $K \times K$ 的估计矩阵。

2.5 基于线性估计的子空间算法

对于噪声的抑制来说, 仅仅将带噪语音信号无修改地投影到信号子空间中是不够的。这种做法虽然不会产生语音信号失真, 却会引入大量的残留噪声。能够减小残留噪声的一种方法是在重构语音信号前, 推导在某种意义上最优的线性估计器。子空间语音增强方法有时域约束估计器(Time Domain Constraint estimator, TDC)和频域约束估计器(Frequency Domain Constraint estimator, SDC)两种, 由于 TDC 的性能要优于 SDC, 所以本文主要研究 TDC 估计器。

假设噪声为加性噪声, 且噪声和语音不相关, 则带噪语音可表示为:

$$Y(k) = S(k) + N(k) \quad (2-19)$$

令 $\hat{S} = HY$ 表示对纯净语音信号的一个估计。那么该估计器所得到的误差 ε 可以计算如下:

$$\begin{aligned} \varepsilon &= \hat{S} - S \\ &= (H - I) \cdot S + H \cdot N \\ &= \varepsilon_S + \varepsilon_N \end{aligned} \quad (2-20)$$

式中, ε_S 代表语音失真, ε_N 代表残留的噪声。那么, 语音信号失真的能量为:

$$\begin{aligned} \overline{\varepsilon_S^2} &= E[\varepsilon_S^T \varepsilon_S] = \text{tr}(E[\varepsilon_S^T \varepsilon_S]) \\ &= \text{tr}(HR_S H^T - HR_S - RH^T + R_S) \end{aligned} \quad (2-21)$$

以及残留噪声的能量为:

$$\begin{aligned}
 \overline{\varepsilon_N^2} &= E[\varepsilon_N^T \varepsilon_N] = \text{tr}(E[\varepsilon_N^T \varepsilon_N]) \\
 &= \text{tr}\{(H \cdot N) R_N (H \cdot N)^T\} \\
 &= \text{tr}(H R_N H^T) \\
 &= \sigma_N^2 \text{tr}(H H^T)
 \end{aligned} \tag{2-22}$$

然后通过对下面的优化问题求解来得到最优的线性估计器：

$$\min_H \overline{\varepsilon_S^2} \quad (\text{在 } \frac{1}{k} \overline{\varepsilon_N^2} \leq \alpha \sigma_N^2 \text{ 条件下}) \tag{2-23}$$

这里 α 的取值范围为 0 到 1 之间。该约束条件的意义是在噪声的残差能量在约束范围内时，使语音的失真最小。该最小化问题可以通过使用 Kuhn-Tucker 必要条件计算，即如果满足 Lagrange 函数：

$$L(H, \mu) = \overline{\varepsilon_S^2} + \mu (\overline{\varepsilon_N^2} - \alpha K \sigma_N^2) \tag{2-24}$$

并且满足：

$$\mu (\overline{\varepsilon_N^2} - \alpha K \sigma_N^2) = 0, \mu > 0 \tag{2-25}$$

式中 μ 为 Lagrange 乘子。通过令 $\nabla_H L(H, \mu) = 0$ ，求得 H ，可以得到最优估计器为：

$$H_{opt} = R_S (R_S + \mu R_N)^{-1} \tag{2-26}$$

噪声为白噪声时，有 $R_N = \sigma_N^2 I$ ，则最优估计器可以写为：

$$H_{opt} = R_S (R_S + \mu \sigma_N^2 I)^{-1} \tag{2-27}$$

对矩阵 R_S 进行 EVD 分解，即 $R_S = U \Lambda_S U^T$ ，上面的估计器就可以重写为：

$$H_{opt} = U \Lambda_S (\Lambda_S + \mu \sigma_N^2 I)^{-1} U^{-T} \tag{2-28}$$

则增强信号 \hat{S} 为:

$$\hat{S} = H_{opt} \cdot Y \quad (2-29)$$

2.6 本章小结

本章首先介绍了研究语音增强算法所必须要了解的语音特性和噪声特性,然后对已有的单声道语音增强算法进行了简要概述,包括谱减法,维纳滤波法,基于统计模型的增强方法,子空间法等。在这些算法中,子空间法有着残留噪声少,引起的语音失真小,音乐噪声不明显,且具有统计模型方法所不具有的控制残余噪声和语音失真关系的机制等优点,因此在本章接下来的内容中,重点介绍了子空间语音增强算法。在子空间语音增强算法的介绍中,由于使用 EVD 的方法与使用 SVD 的方法得到的最优估计之间可以彼此转换,但 EVD 方法具有清晰的物理意义。基于线性估计器的子空间语音方法可以分为 TDC 和 SDC 两种,因 TDC 具有更好的性能,所以本章最后对 TDC 进行了重点介绍。

第三章 神经网络基础理论

人工神经网络^[31](Artificial Neural Networks, ANN), 简称为神经网络(Neural Networks, NN)是二十世纪四十年代开始出现, 并在二十世纪末至今得到迅速发展的一门技术。神经网络是交叉性学科, 涉及神经学, 特别是动物脑部科学, 数学, 和工程学等多个领域, 它有能力模拟人脑内的大量神经元之间的连接, 有着非常出色的非线性映射能力、自组织学习能力和并行计算的能力。

神经网络通过神经元之间的层级连接, 一层一层的传递并处理数据, 采用由底至顶的自学习方法, 借助于其强大的映射能力, 可以对实际应用中复杂的难于用工程学方法建立的非线性系统建模。神经网络是对人类大脑内生物神经网络的模拟, 具备一定程度的人脑功能, 可以解决许多复杂的模式识别问题, 比如手写数字的识别, 车牌号的识别, 图片内物体的识别等。

3.1 神经元数学模型

神经元, 又称为节点, 是神经网络的基本单元。它将神经元的工作原理高度简化, 它模仿真实的人脑内神经元的结构和功能。它是在医学, 尤其是神经学的研究成果上提出的对人脑神经元的抽象, 远非对真实神经元的真实描述, 只是在一定程度上反映了真实神经元的功能。作为 NN 的基本构成单元, 人们对真实的生物神经元的主要特性进行了高度的抽象, 使得其在数学上表现为一个元器件, 该元器件具有多个输入接口, 一个输出接口。神经元的数学模型如图 3.1 所示:

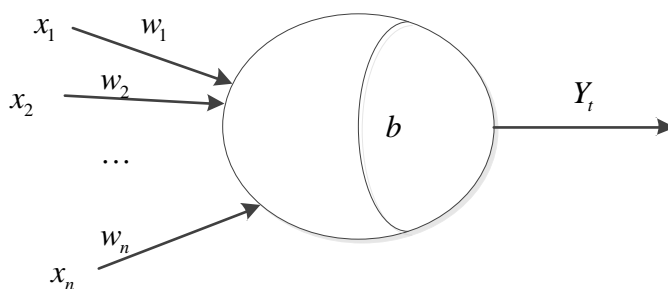


图 3.1 神经元的数学模型

图 3.1 中 x_i 表示一个神经元输入信号, w_i 和 b 表示其权值和偏置, Y_t 表示其输出:

$$Y_t = f\left(\sum x_i w_i - b\right) \quad (3-1)$$

式中 $f(\cdot)$ 为神经元的传输函数，它有多种函数可以选择，常见的有三种：硬极限传输函数，线性传输函数，对数 S 型传输函数。其数学表达式如下：

$$f(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (3-2)$$

$$f(x) = x \quad (3-3)$$

$$f(x) = \frac{1}{1 + e^{-x}} \quad (3-4)$$

3.2 误差反向传播算法研究

误差反向传播(Back Propagation, BP)神经网络(简称 BPNN)是实际应用中最为广泛的一种。BPNN 是有监督的多层前向网络，基于 BP 算法。BPNN 是为了解决此前多层网络中大都存在的算法不收敛问题而提出的，其基本思想是构造一个非线性系统，使用均方误差作为其性能指数，这一点来看 BP 算法可以看作是对最速下降法的近似，它们的区别在于对倒数的计算方式上。

典型的 BPNN 的结构如图 3.2 所示，通常为三层：输入层，隐藏层和输出层。

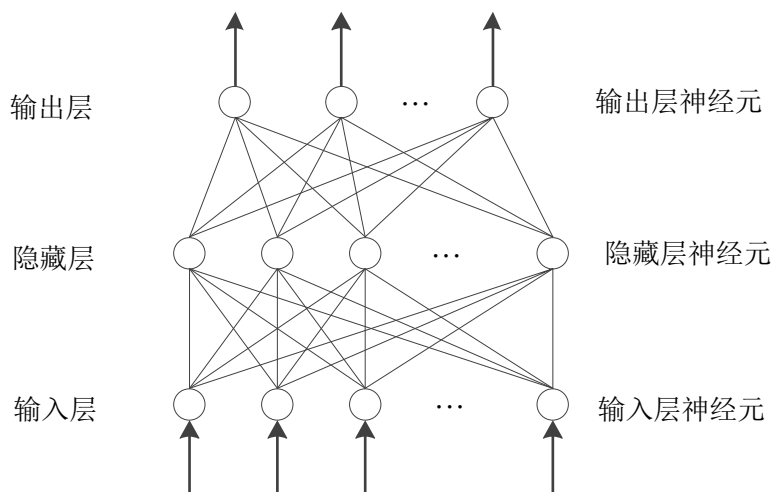


图 3.2 BP 神经网络示意图

假设在包含一个隐藏层的 BP 神经网络中，其输入向量为 $X = (x_1, x_2, \dots, x_i, \dots, x_n)^T$ ，隐藏层输出向量为 $Y = (y_1, y_2, \dots, y_j, \dots, y_m)^T$ ，输出层输出向量为 $O = (o_1, o_2, \dots, o_k, \dots, o_l)^T$ ，期望输出向量为 $d = (d_1, d_2, \dots, d_k, \dots, d_l)^T$ ，输入层和隐藏层之间的权值矩阵为 $V = (V_1, V_2, \dots, V_j, \dots, V_m)$ ，隐藏层和输出层之间的权值矩阵为 $W = (W_1, W_2, \dots, W_k, \dots, W_l)$ 。

误差反向传播算法的基本原理是：学习过程可以看做是由两部分组成，即输入信

号的前向传播过程和误差信号的反向传播过程。当给定了一个输入信号，数据从输入层传递到隐藏层再到输出层，由输出层的神经元处理后产生一个输出信号。当这个输出信号与期望的监督信号不符时，得到一个误差信号。反向传播过程是指误差信号由输出层到隐藏层，再到输入层，各层的神经元在得到误差之后，将其作为调节自己参数的依据。这两个过程循环重复进行，各神经元的参数也得到不断地调整，这就是BPNN的学习过程。

对于输出层：

$$o_k = f(net_k) \quad k = 1, 2, \dots, l \quad (3-5)$$

$$net_k = \sum_{j=0}^m w_{jk} y_j \quad k = 1, 2, \dots, l \quad (3-6)$$

对于隐藏层：

$$y_j = f(net_j) \quad j = 1, 2, \dots, m \quad (3-7)$$

$$net_j = \sum_{i=0}^n v_{ij} x_i \quad j = 1, 2, \dots, m \quad (3-8)$$

输出误差 E 的定义为：

$$E = \frac{1}{2} (d - o)^2 = \frac{1}{2} \sum_{k=1}^l (d_k - o_k)^2 \quad (3-9)$$

输出误差传递到隐藏层变为：

$$E = \frac{1}{2} \sum_{k=1}^l \left[d_k - f \left(\sum_{j=0}^m w_{jk} y_j \right) \right]^2 \quad (3-10)$$

进一步传递到输入层变为：

$$E = \frac{1}{2} \sum_{k=1}^l \left\{ d_k - f \left[\sum_{j=0}^m w_{jk} f \left(\sum_{i=0}^n v_{ij} x_i \right) \right] \right\}^2 \quad (3-11)$$

则权值的调整量为：

$$\Delta w_{jk} = -\eta \frac{\partial E}{\partial w_{jk}} \quad j = 0, 1, 2, \dots, m; k = 1, 2, \dots, l \quad (3-12)$$

$$\Delta v_{ij} = -\eta \frac{\partial E}{\partial v_{ij}} \quad i = 0, 1, 2, \dots, n; j = 1, 2, \dots, m \quad (3-13)$$

式中负号表示梯度下降， $\eta \in (0, 1)$ 表示调节系数，即网络的学习率。对于输出层神经元，上面的公式可写为：

$$\Delta w_{jk} = -\eta \frac{\partial E}{\partial w_{jk}} = -\eta \frac{\partial E}{\partial net_k} \frac{\partial net_k}{\partial w_{jk}} \quad (3-14)$$

对隐藏层神经元，可写为：

$$\Delta v_{ij} = -\eta \frac{\partial E}{\partial v_{ij}} = -\eta \frac{\partial E}{\partial net_j} \frac{\partial net_j}{\partial v_{ij}} \quad (3-15)$$

分别定义以下输出层和隐藏层的误差信号，即令：

$$\delta_k^o = -\frac{\partial E}{\partial net_k} \quad (3-16)$$

$$\delta_j^y = -\frac{\partial E}{\partial net_j} \quad (3-17)$$

则输出层的权值调整式改写为：

$$\Delta w_{jk} = \eta \delta_k^o y_j \quad (3-18)$$

隐藏层的权值调整式改写为：

$$\Delta v_{ij} = \eta \delta_j^y x_i \quad (3-19)$$

计算出式(3-18)和(3-19)中的误差信号，就可以得到权值调整式。对于输出层， δ^o 可展开为：

$$\delta_k^o = -\frac{\partial E}{\partial net_k} = -\frac{\partial E}{\partial o_k} \frac{\partial o_k}{\partial net_k} = -\frac{\partial E}{\partial o_k} f'(net_k) \quad (3-20)$$

对于隐藏层, δ^y 可展开为:

$$\delta_j^y = -\frac{\partial E}{\partial net_j} = -\frac{\partial E}{\partial y_j} \frac{\partial y_j}{\partial net_j} = -\frac{\partial E}{\partial y_j} f'(net_j) \quad (3-21)$$

最后计算出误差对各层输出的导数, 并将结果代入式(3-18), (3-19), 可得权值的调整量为:

$$\Delta w_{jk} = \eta \delta_k^o y_j = \eta (d_k - o_k) o_k (1 - o_k) y_j \quad (3-22)$$

$$\Delta v_{ij} = \eta \delta_j^y x_i = \eta \left(\sum_{k=1}^l w_{jk} \right) y_j (1 - y_j) x_i \quad (3-23)$$

3.3 深度学习简介

3.3.1 浅层结构和深层结构

无论是经典的 BP 神经网络还是 2006 年之后提出的深度学习, 人们最初的动机都是希望通过模拟人脑神经元的架构, 并加入特定的算法来让机器模仿人脑的工作模式。BPNN 可以基于误差反向传播算法从大量的样本数据中找到规律, 从而建立起输入样本和期望输出之间复杂的映射关系, 这样就可以对未知的输入数据产生相应的输出。通常人们认为, 如果神经网络的隐藏层数越多, 即网络的结构越深, 就具有越强大的学习能力和非线性映射能力, 也就会具有更好的分类能力^[32]。人们就希望通过加深神经网络的结构来增强其分类能力。

然而, 随着 BPNN 隐藏层数的增加, 它所潜在的缺陷就体现出来了, 主要是以下几方面:

(1) 会陷入局部最优

使用 BP 算法的 BPNN, 可以看作是对非凸问题的求解优化。然而, 这类非凸问题可能存在许多的极值点, 加之 BPNN 的初始参数是随机生成的较小的值, 训练过程中依靠梯度下降法来寻找最优值, 很难确保 BPNN 最终收敛在全局的最优值。

(2) 梯度弥散现象

随着隐藏层数的增加, 梯度值会迅速变小, 导致代价函数对前面几层神经元权值的偏导也急剧减小, 以至于网络不能有效的学习样本^[33]。

(3) 标签获取困难

现实生活中,人们获取数据是比较容易的,但是要为每一个数据打上相应的标签却很困难^[34],工作量非常巨大。当增加 BPNN 的隐藏层数时,还可能导致其出现过拟合。

3.3.2 深度学习的训练机制

接下来的十几年中,人们始终无法有效解决 BPNN 中出现的问题。直至 2006 年,Hinton 教授提出了“贪心逐层预训练”的机制。

深度神经网络与传统的神经网络从结构上而言区别并不大。Hinton 教授所提出的“贪心逐层预训练”的机制是在训练过程中分层进行,让每一层都贪心地达到自己的最优权值,每一层的训练都是无监督的。当所有的层都训练完成后,用各层所得到的参数对整个网络进行初始化,最后再基于有监督的训练方式对网络微调^[35-37]。

使用“贪心逐层预训练”方式得到的各层的参数对网络整体参数进行初始化,和 BP 神经网络的使用较小随机值初始化参数的方式相比,可以使网络的初始参数位于较优的区间,以这些较优的初始参数为基础对网络进行微调,有利于克服网络陷入局部极值的缺陷。另外,贪心逐层训练的方法让每一层网络都达到自己的最优,便于在微调时进行全局的优化。最后,预训练采取的是无监督训练机制,这就解决了有监督训练机制中设置标签困难的问题。

深度神经网络是对传统人工神经网络的继承和发展,因此也被称为“新一代神经网络”^[38]。

3.3.3 深度学习的主流模型

深度学习理论中有许多实用的模型,栈自动编码机和深度信念网络是其中的主流模型,人们也早已在许多的领域中应用它们。

SAE 是由多个 AE 堆叠起来构成的^[39]。单个 AE 的结构可以看作是一个三层的 NN,它可以在输出层对输入层的数据进行重构。将多个自动编码器堆叠构成 SAE 之后,采用逐层预训练的机制,训练完一个 AE 之后,将所有的原始数据输入,在隐藏层保存所有输入数据的特征,将这些特征作为输入数据继续对下一个 AE 训练。

深度信念网络是由受限玻尔兹曼机(Restricted Boltzmann Machine,RBM)堆叠而构成的,这一点与栈自动编码器相似。RBM 是一种基于统计力学的随机网络,它具有两层结构,分别称为显层和隐层。RBM 中引入了能量函数作为损失函数来进行优化^[40],它可以根据隐层数据分布的概率模型,对输入的数据进行无监督分布建模。深度信念网络的训练是通过 Hinton 教授的对比散度(Contrastive Divergence, CD)算法逐层对每一个 RBM 无监督训练实现的。所有的 RBM 都训练完成之后,在其顶端再添加

一层有监督的网络，采用 BP 算法对整体网络进行优化。这种机制可以简化训练过程，避免了直接训练多层网络的复杂性。

3.4 栈自动编码器

SAE 作为深度学习中的一种主流模型，得到了广泛的应用。其基础是自动编码器。AE 是一种无监督的学习算法，其结构如图 3.3 所示。

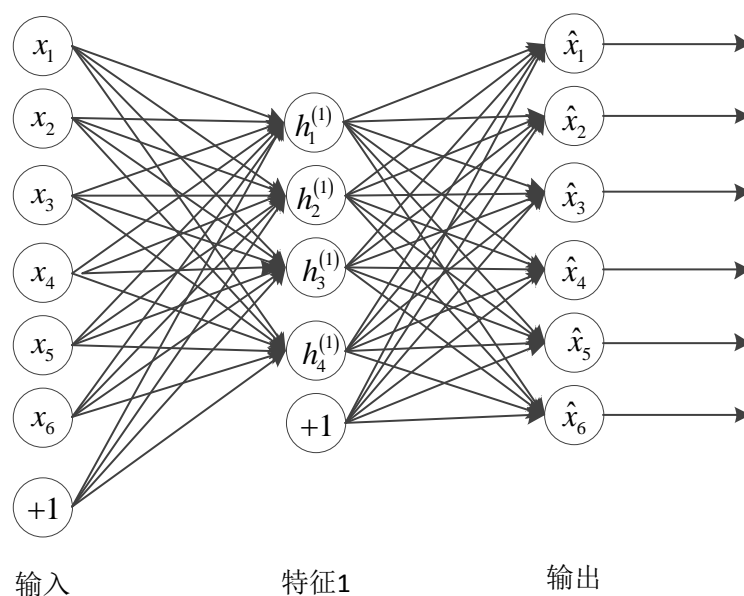


图 3.3 自动编码器基本结构

单个的 AE 可以被看作是具有三层结构的网络。区别是 AE 的输出等于其输入 X 而不是神经网络中所期望的标签 Y ，即： $y(i) = x(i)$ 。这表示，自动编码器的训练目的是为了使其输出逼近输入^[41]。

当给 AE 加上一些限制条件或者按照某些要求确定了其结构之后，比如当隐藏层的宽度小于输入层时，它就可以学到输入数据的一种压缩表达，通过这种压缩表达可以恢复出原始的输入数据^[42]。即对于输入矢量 X 来说， $h^{(1)}$ 可以看作是它的一种特征描述，但 $h^{(1)}$ 更为简洁，它既包含 X 的所有信息，又具有更小的数据量。因此，可以将 $h^{(1)}$ 看作 X 的一阶特征。

对于单个自动编码器而言，完成了上面的预训练之后，可以利用其参数初始化新构建一个神经网络。利用 BP 算法继续调节所构建网络的权值，将 X 的一阶特征映射为期望的输出标签。这个过程通常被称为微调^[43]。

栈自动编码器中的每一个 AE 的原理都是一样的。其训练过程可以看作将 SAE 拆分成多个 AE 逐个训练。第一个 AE 训练完成后隐藏层的矢量 $h^{(1)}$ 可看作是对输入矢量 X 的另一种表达，将其作为第二个 AE 的输入矢量。同理，训练第二个 AE 时，

可以得到一个 $h^{(1)}$ 到 $h^{(1)}$ 的三层神经网络, 得到第二个 AE 的隐藏层矢量 $h^{(2)}$, 则 $h^{(2)}$ 可以看作是 $h^{(1)}$ 的另一种表达, 称 $h^{(2)}$ 是原始输入矢量 X 的二阶特征矢量。二阶特征矢量 $h^{(2)}$ 也可以用来形成到期望的输出标签 Y 的映射。以此类推, 可以堆叠出更深层的栈自动编码器。

逐个训练每一个 AE 结束之后, 将它们的权值保存下来。新建一个具有相应结构的新的神经网络, 用保存下来的权值初始化这个新的网络的各参数, 这样就得到了输入层、多个隐藏层和输出层级联的栈自动编码器。如图 3.4 所示, SAE 可以提取原始输入矢量 X 的高阶特征, 并以其为根据对 X 分类。

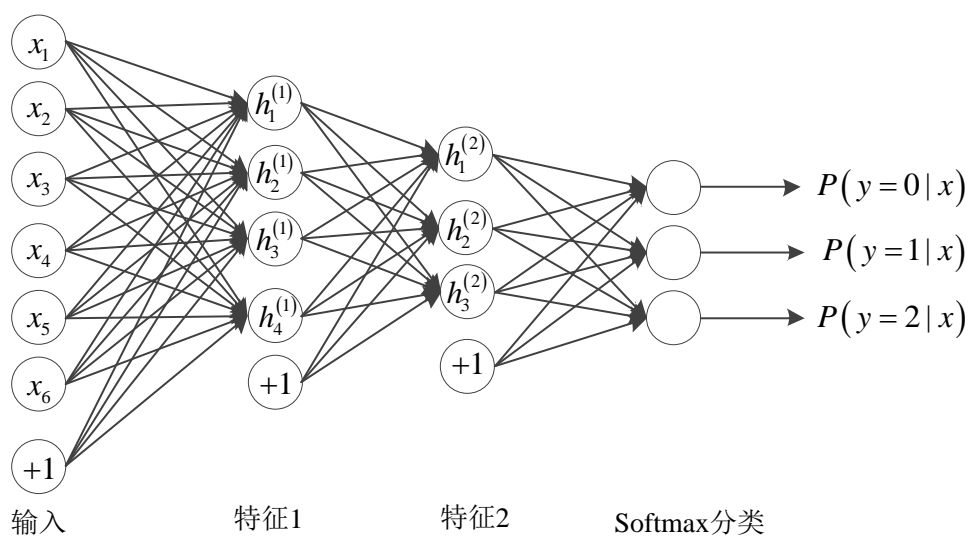


图 3.4 栈自动编码器的结构

3.5 深度信念神经网络

Hinton 教授在 1986 年提出了 RBM, 它是显层和隐层构成的一种随机神经网络^[21]。RBM 的显层和隐层之间存在连接, 同层的单元相互之间没有连接, 如图 3.5 所示。

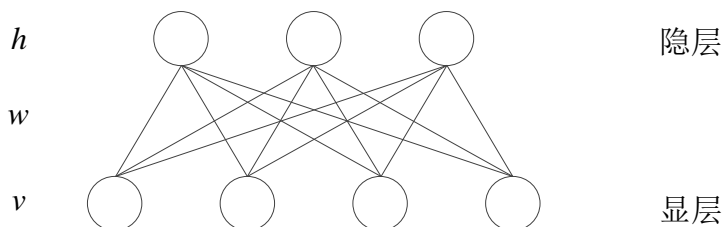


图 3.5 受限玻尔兹曼机的连接方式

图 3.5 所示的 RBM 显层有 4 个单元, 构成矢量 v , 隐层有 3 个单元, 构成矢量 h , W 为一个 4 行 3 列的矩阵, 表示显层与隐层之间的连接权重。

RBM 是一种基于能量的模型，任意一个变量都可以被定义成一种能量。通过能量函数来定义模型中的概率分布。RBM 模型的显层矢量和隐层矢量的能量联合配置为：

$$E(v, h; \theta) = -\sum_{ij} W_{ij} v_i h_j - \sum_i b_i v_i - \sum_j a_j h_j \quad (3-24)$$

式中， $\theta = \{W, a, b\}$ 是 RBM 模型的参数， W_{ij} 是其显层与隐层之间的权值矩阵， b_i 和 a_j 是显层和隐层的偏置。在获得显层矢量 v 和隐层矢量 h 的联合能量之后，能够推导出它们的联合概率为：

$$P(v, h; \theta) = \frac{1}{z(\theta)} \exp(-E(v, h; \theta)) \quad (3-25)$$

式中， $z(\theta)$ 为归一化因子， $P(v, h; \theta)$ 为 Boltzmann 分布函数。可以将式(3-25)写为：

$$P(v, h; \theta) = \frac{1}{z(\theta)} \exp\left(\sum_{i=1}^D \sum_{j=1}^F W_{ij} v_i h_j + \sum_{i=1}^D v_i b_i + \sum_{j=1}^F h_j a_j\right) \quad (3-26)$$

基于 RBM 独特的结构，其隐层单元与显层单元得以清楚地区分开来，可以求得条件概率：

$$P(v | h) = \prod_{i=1}^n P(v_i | h) \quad (3-27)$$

$$P(h | v) = \prod_{j=1}^m P(h_j | v) \quad (3-28)$$

那么，使似然函数 $P(v; \theta)$ 最大化可得：

$$P(v; \theta) = \frac{1}{Z(\theta)} \exp(b^T v) \prod_{j=1}^F \left(1 + \exp\left(a_j + \sum_{i=1}^D W_{ij} v_i\right)\right) \quad (3-29)$$

令 $L(\theta)$ 记为：

$$L(\theta) = \frac{1}{N} \sum_{n=1}^N \log P(v^{(n)}; \theta) \quad (3-30)$$

最大化 $L(\theta)$ ，需要求出 $L(\theta)$ 对 W 的偏导，得到：

$$\frac{\partial L(\theta)}{\partial W_{ij}} = \frac{1}{N} \sum_{n=1}^N \frac{\partial}{\partial W_{ij}} \log \left(\sum_h \exp \left[v^{(n)T} W h + a^T h + b^T v^{(n)} \right] \right) - \frac{\partial \log Z(\theta)}{\partial W_{ij}} \quad (3-31)$$

化简可得：

$$\frac{\partial L(\theta)}{\partial W_{ij}} = E_{P_{data}} [v_i h_j] - \sum_{v,h} v_i h_j P_{\theta}(v, h) \quad (3-32)$$

式(3-32)中，在所有的数据集上求 $v_i h_j$ 的平均可求得其前半部分。然而，后半部分与 v 和 h 所有组合下的情况均有关，难以计算。为此，Hinton 等研究者共同提出了一种对比散度学习算法。该算法的基本思想为：首先利用 RBM 显层矢量 v 来得到隐层矢量 h ，再利用隐层矢量 h 来对显层矢量 v 重构得到 v_1 ，在利用 v_1 对隐层矢量 h 重新生成，得到 h_1 。

$$P(h_j = 1 | v) = \frac{1}{1 + \exp(-\sum_i W_{ij} v_i - a_j)} \quad (3-33)$$

同理可得：

$$P(v_i = 1 | h) = \frac{1}{1 + \exp(-\sum_j W_{ij} h_j - b_i)} \quad (3-34)$$

可以将重构的显层矢量 v_1 和隐层矢量 h_1 当做是对 $P(v, h)$ 的一次抽样的结果，进行了多次抽样之后会得到一个样本的集合，将这个集合当做是对 $P(v, h)$ 的近似，这样就可以对式(3-32)的后半部分进行计算。

深度信念网络(Deep Belief Networks, DBN)是由多个 RBM 组成的^[44]。其结构如图 3.6 所示。DBN 是一个概率生成模型，它能够建立输入数据和它的标签之间的联合分布。DBN 既需要无监督的预训练，又需要有监督的调优。

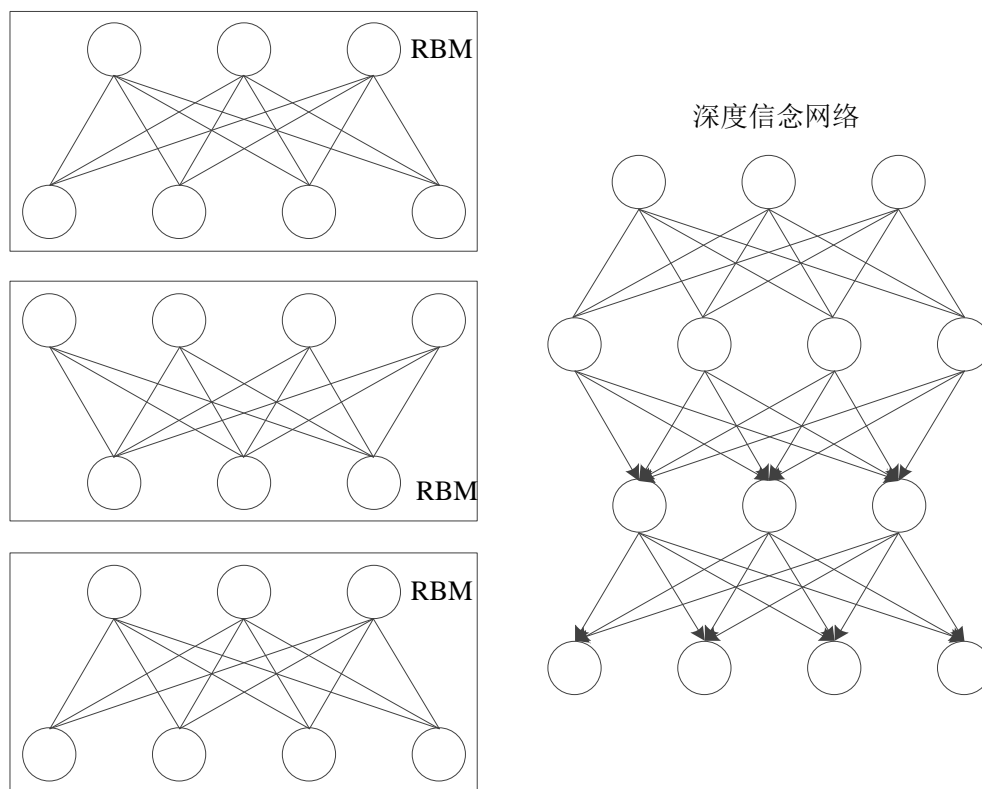


图 3.6 一叠 RBM 和对应的深度信念网络

由于 DBN 是由 RBM 堆叠而成的，所以它也是层与层之间有连接，层间单元无连接的结构。DBN 采用逐层训练的方式。一个 RBM 训练结束之后，将这个 RBM 的隐层当做下一个 RBM 的显层来对下一个 RBM 进行训练。一个只含单个隐层的深度信念模型，它的隐层矢量与显层矢量之间用概率公式可表示关系如下所示：

$$P(x, h^2, \dots, h^l) = \left(\prod_{k=2}^{l-1} P(h^k | h^{k+2}) \right) P(h^{l-2}, h^l) \quad (3-35)$$

在机器学习的领域中，DBN 的训练过程被看作是生成对象模型的过程。当 DBN 模型是一个含有多个隐藏层深度结构的时候，直接进行全局的优化是异常困难的。因此需要使用贪心模式逐层地预训练。这样，就把整个 DBN 模型参数的学习过程就分为预训练和微调两个部分。

初始化 DBN 就是将其拆分为多个 RBM 的过程。通过逐个地训练这一系列的 RBM，来完成初始化 DBN 网络的目的。过程如下：首先对第一个 RBM 输入样本进行训练，训练结束后，固定第一个 RBM 的权值，将所有的训练样本再次输入第一个 RBM 得到其相应的隐层矢量，这时将所得到的第一个 RBM 的隐层矢量作为第二个 RBM 的输入数据来对第二个 RBM 进行训练。

这种递归的模式对每一个 RBM 进行训练，得到每一个 RBM 的权值矩阵。所有的 RBM 训练完成后，用这些 RBM 的权值矩阵初始化 DBN，最后采用 BP 算法对 DBN

进行微调, 这样可以使 DBN 网络的表达能力得到大幅的提高。

逐层贪心的方法使得网络在学习过程中无论是时间复杂度还是空间复杂度都是线性的, 当要使用网络对大数据量进行学习的时候, 这种预训练加微调的模式是非常有利的。

3.6 本章小结

本章介绍了神经元模型, 推导了 BPNN 参数更新的公式。然而神经网络由浅层结构加深为深层结构时, 出现的梯度弥散, 局部最优等问题, 并由此引入了深度学习的介绍, 包括深度学习的训练机制和主流模型。其中重点介绍了在下一章中要使用的栈自动编码器和深度信念网络。

第四章 基于深度学习的语音增强方法

第二章主要研究了传统单声道语音增强算法中的子空间法,第三章中重点介绍了神经网络中的 BP 算法,和深度学习中 SAE 和 DBN 这两个主流模型。在本章中,将基于前面介绍的深度信念网络提出噪声幅度谱估计的语音增强方法,在该方法中首先对带噪语音信号中的噪声类型进行分类,分类使用的是栈自动编码器。然后根据分类结果将带噪语音交给适应于该噪声类型的深度信念网络,使用深度信念网络对带噪语音中的噪声幅度谱进行估计。最后对于子空间语音增强算法中的不足之处,提出了相应的改进方案。

4.1 基于 DBN 的噪声幅度谱估计语音增强方法

在本节,将介绍文本所提出的基于噪声幅度谱估计的语音增强方法。图 4.1 给出了本文所提方法的框图。整个过程可以分为两个阶段:训练阶段和增强阶段。

在训练阶段,使用纯净语音和 NOISEX_92 噪声数据库中的八种噪声,然后通过公式 $y = s + \alpha n$, α 为噪声系数,通过直接相加的方式生成大量的不同信噪比下的带噪语音,同时保留其对应的噪声。这些带噪语音和噪声的数据对为训练样本。然后进行特征提取,特征提取过程中,首先对训练样本进行预处理和分帧加窗操作,这里使用的特征是幅度谱,即进行 DFT 变换后取模。提取了大量的训练样本的幅度谱之后,对其归一化,即将全部的特征归一化到 0 和 1 之间,这有利于深度信念网络学习到带噪语音幅度谱和对应的噪声幅度谱之间的非线性映射关系。DBN 的学习分为两步:基于 RBM 的无监督预训练和基于 BP 算法的有监督微调。在训练阶段,我们用八种不同的带噪语音/噪声数据对分别训练了与之对应的 DBN 网络。

在增强阶段,拿到一句带噪语音之后,首先使用带噪语音的第一帧对其中的噪声进行分类,然后进行特征提取,特征使用的是幅度谱。根据对噪声的分类结果,将所提取的带噪语音的幅度谱进行归一化后输入到与该噪声相适应的已经训练好的 DBN 模型中进行计算,得到模型的输出后,先进行反归一化,得到的是 DBN 模型对带噪语音中噪声的幅度谱的估计,再用带噪语音的幅度谱减去 DBN 模型对其中噪声幅度谱的估计,就得到了对纯净语音的幅度谱估计,最后结合带噪语音的相位谱利用重叠相加法恢复出增强语音的时域波形。在这里,恢复增强语音的时域波形时使用带噪语音的相位信息是利用了人耳对相位变化不敏感的特性。尽管如此,为了合成更高质量的增强语音,仍然需要对纯净语音的相位信息进行准确估计。这一课题将作为本文的后续工作进行研究。基于 DBN 的噪声幅度谱估计语音增强方法如图 4.1 所示。

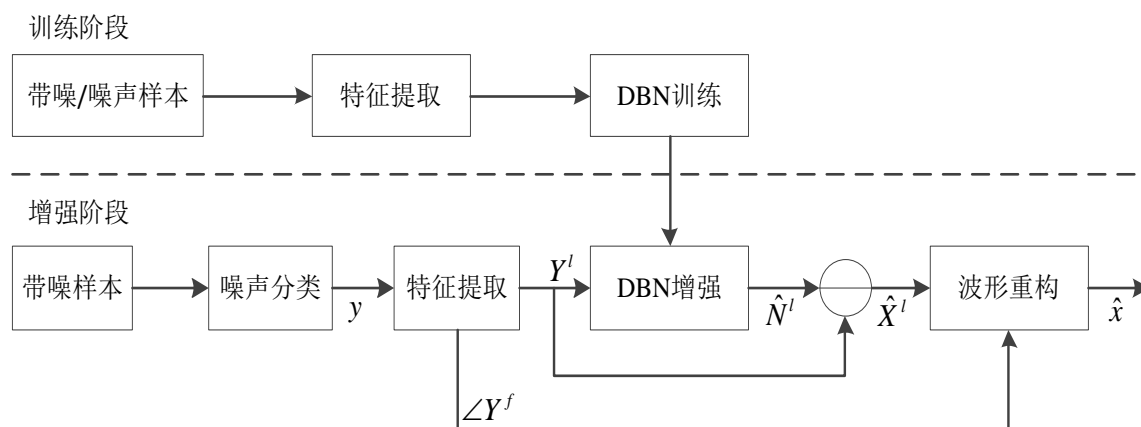


图 4.1 基于 DBN 的噪声幅度谱估计语音增强方法框图

图 4.1 中， y 为带噪语音的时域信号， Y^l 为带噪语音的幅度谱， \hat{N}^l 为深度信念网络估计的噪声幅度谱， \hat{X}^l 为 Y^l 减去 \hat{N}^l 得到的增强语音的幅度谱， $\angle Y^f$ 为带噪语音的相位谱， \hat{x} 为恢复出的增强语音的时域信号。

在传统的单声道语音增强算法中，通常有纯净语音和噪声不相关且具有零均值的假设，而在基于深度信念网络的语音增强中，使用 DBN 网络作为学习带噪语音和噪声幅度谱之间关系的模型。如图 4.2 所示，DBN 网络具有的强大的非线性映射能力，可以对带噪语音和噪声幅度谱之间复杂的非线性映射关系进行充分学习。并且，在该语音增强方法中，没有做任何其他的假设。



图 4.2 带噪语音幅度谱到噪声幅度谱之间的映射

(1) 构造训练数据

因为基于深度信念网络的噪声幅度谱估计语音增强方法需要让 DBN 网络学习带噪语音和噪声幅度谱之间的关系，学习过程分为预训练和微调，因此它是一个有监督的模型。它需要很多的训练样本对。根据噪声的加性模型，可以使用公式 $y = x + \alpha n$ 合成需要的训练数据对，这里 α 是用来控制噪声大小的参数，作用是调节向纯净语音中叠加噪声的幅度，以控制合成的带噪语音的信噪比。按照这样的方式，可以合成大量的训练 DBN 模型所需要的包含各种信噪比，各种噪声种类的训练数据对。这里需要说明的是，通过这种加噪方式获得的训练数据对并不能真正反映实际中的声学场景，这里的噪声仅是加性噪声，而实际场景中，还有存在非加性噪声。由于非加性噪声可以通过转换变为加性噪声，且实际中人们的听感主要是受到加性噪声的影响，所以本文仅讨论加性噪声的情况。

(2) 幅度谱特征提取

幅度谱特征的提取过程如图 4.3 所示。

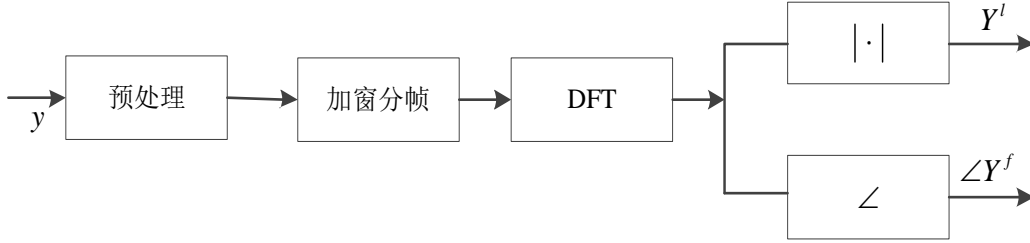


图 4.3 特征提取框图

图 4.3 中, 在进行特征提取时首先对信号进行预处理和加窗分帧, 然后计算 DFT 变换的系数。为了使恢复的信号比较平滑, 人耳的听感相对舒服一些, 各帧之间是相互重叠的, 通常相邻两帧之间的重叠率是帧长的二分之一。对于预处理和分帧加窗的操作, 在下一节进行介绍。对分帧之后的信号应用 DFT 变换, 变换公式如下所示:

$$Y(d) = \sum_l^{L-1} y(l)w(l)e^{-j2\pi dl/L} \quad d = 0, 1, \dots, L-1 \quad (4-1)$$

式中, d 是频率维度, $w(l)$ 是窗函数。本文使用的窗函数是三角窗。这里需要说明的是, 进行 DFT 变换的点数若能够增加, 则变换后的输入的特征将具有更高的频率分辨率, 这将更有助于深度信念网络的学习。该方法中所使用的特征是信号的幅度谱。幅度谱的定义如下:

$$Y^l = |Y(d)| \quad d = 0, 1, \dots, D-1 \quad (4-2)$$

这里 $D = L/2 + 1$, 而对于 $d = D, \dots, L-1$, $Y(d)$ 的时候的幅度值可通过对称准则来得到, 即 $Y(d) = Y(L-d)$ 。

(3) 重叠相加法波形重构

波形重构过程如图 4.4 所示。

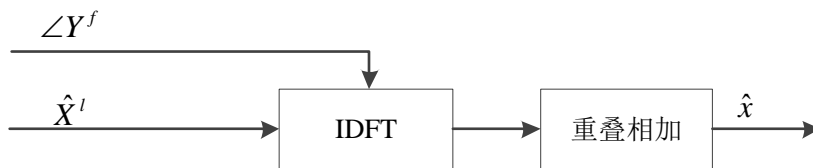


图 4.4 波形重构框图

图 4.4 为使用重叠相加法重构增强语音时域波形的框图。在利用已经训练好的深度信念网络得到其估计出的带噪语音中噪声的幅度谱特征 \hat{N}^l 之后，用带噪语音的幅度谱减去 DBN 网络估计出的噪声的幅度谱，即 $\hat{X}^l = Y^l - \hat{N}^l$ ，就得到了增强语音的幅度谱，之后就需要使用重叠相加法对波形进行重构，以恢复增强语音的时域波形。如下公式所示：

$$\hat{X}(d) = \hat{X}^l \cdot \exp\{j\angle Y^f\} \quad (4-3)$$

式中 $\hat{X}(d)$ 表示增强语音的 DFT 变换，这里的相位信息 $\angle Y^f$ 使用的是带噪语音 y 的相位信息。

每一帧增强语音的时域数据 $\hat{x}(l)$ 通过对 $\hat{X}(d)$ 做 IDFT 变换得到：

$$\hat{x}(l) = \frac{1}{L} \sum_{k=0}^{L-1} \hat{X}(k) e^{j2\pi kl/L} \quad (4-4)$$

最后，整个增强语音的时域波形是通过重叠相加法来进行合成的。

(4) 无监督预训练

深层神经网络的初始权值如果使用随机的数来初始化，最终完成训练的网络很可能会收敛于局部极值，而无法得到全局的最优结果，尤其是在网络具有深层结构，而用于训练网络的数据又较少的情况下。因此，采用深度信念网络作为估计噪声幅度谱的模型。DBN 可以使用逐层地无监督预训练机制来获得比较好的初始网络参数，在此参数的基础上再进行有监督的调优，这样可以使具有深层结构的 DBN 避免陷入局部极值点。深度信念网络的预训练和微调过程如图 4.5 所示。图 4.5 中，左图为使用带噪语音幅度谱训练 DBN 拆分成的一系列 RBM 的示意图。这些 RBM 是采用对比散度算法无监督逐层贪心的方式进行训练的。

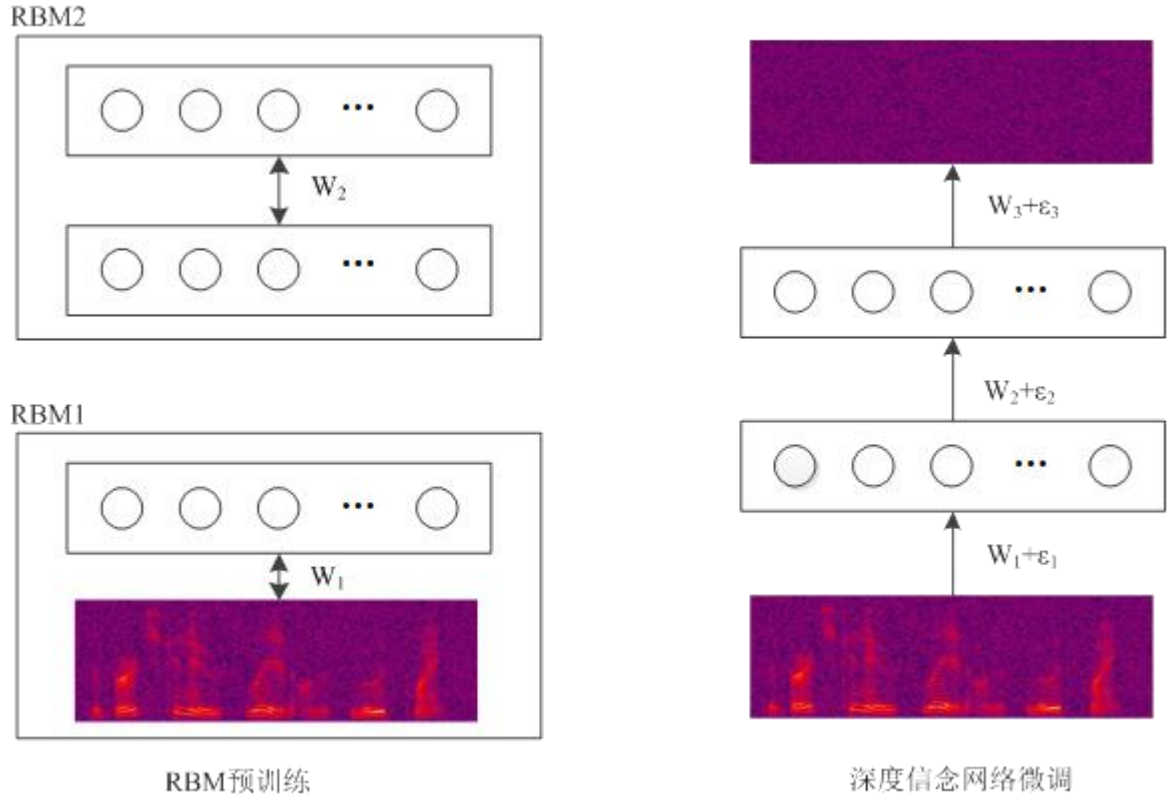


图 4.5 深度信念网络的预训练和微调

(5) 有监督微调

预训练完成之后, 就进行有监督的微调。微调采用的是 BP 算法, 误差信号为 DBN 网络估计的噪声幅度谱和期望的噪声幅度谱之间的误差。图 4.5 右图为使用期望的噪声幅度谱进行有监督微调的过程。微调过程中使用最小批模式以加快深度信念网络的学习速度。公式如下所示:

$$E = \frac{1}{N} \sum_{n=1}^N \sum_{d=1}^D \left(\hat{X}_n^d(W^\ell, b^\ell) - X_n^d \right)^2 \quad (4-5)$$

这里 E 表示 DBN 网络在最小批上的均方误差的均值, $\hat{X}_n^d(W^\ell, b^\ell)$ 和 X_n^d 分别为在第 n 帧, 第 d 个频率处的 DBN 网络估计的噪声幅度谱和目标噪声幅度谱。 N 表示的是最小批的大小, 也就是每一组内有多少个训练样本。 D 为幅度谱矢量的长度, (W^ℓ, b^ℓ) 为在 ℓ 层的有待更新的神经元的参数。这里需要说明的是, 深度信念网络的输入输出数据都是进行了归一化之后的数据, 归一化的公式为:

$$y = (y_{\max} - y_{\min}) * (x - x_{\min}) / (x_{\max} - x_{\min}) + y_{\min} \quad (4-6)$$

公式中, x_{\max} 和 x_{\min} 分别为进行归一化之前向量的最大值和最小值, 而 y_{\max}

和 y_{min} 则分别为归一化之后的向量的最大值和最小值, x 表示某一维度的原始向量的值, y 就是 x 归一化之后对应的值。在这里, 设置 y_{min} 为 0, y_{max} 为 1, 即将幅度谱特征归一化到 0 和 1 之间。

在网络的训练阶段, 没有对语音和噪声做任何的假设, 所以深度信念网络能够很好地学到带噪语音幅度谱和噪声幅度谱之间的复杂的映射关系。

4.2 窗函数的选择

上面所介绍的噪声幅度谱语音增强方法中, 在提取特征参数时, 先对带噪语音进行了预处理和分帧加窗的操作, 在将增强语音恢复为时域波形时, 采用的是重叠相加法, 以往的研究经验表明, 在这些操作过程中, 窗函数的选择对合成的增强语音的质量具有很大的影响, 本节将对预处理和分帧加窗进行介绍。

预处理的作用是对语音的高频部分进行加重, 提高它的高频分辨率, 可以通过一阶 FIR 高通数字滤波器来实现。公式如下:

$$H(z) = 1 - \alpha z^{-1} \quad (4-7)$$

式中, α 是预加重系数, 一般取 0.98, 对语音进行预加重的结果为:

$$y(n) = x(n) - \alpha x(n-1) \quad (4-8)$$

图 4.6 描述了高通滤波器 FIR 的幅频特性及其相频特性。图 4.7 和图 4.8 表示的是对一段浊音预加重前后的时域波形和频谱。对比发现, 预加重后, 信号的高频部分得以提升。

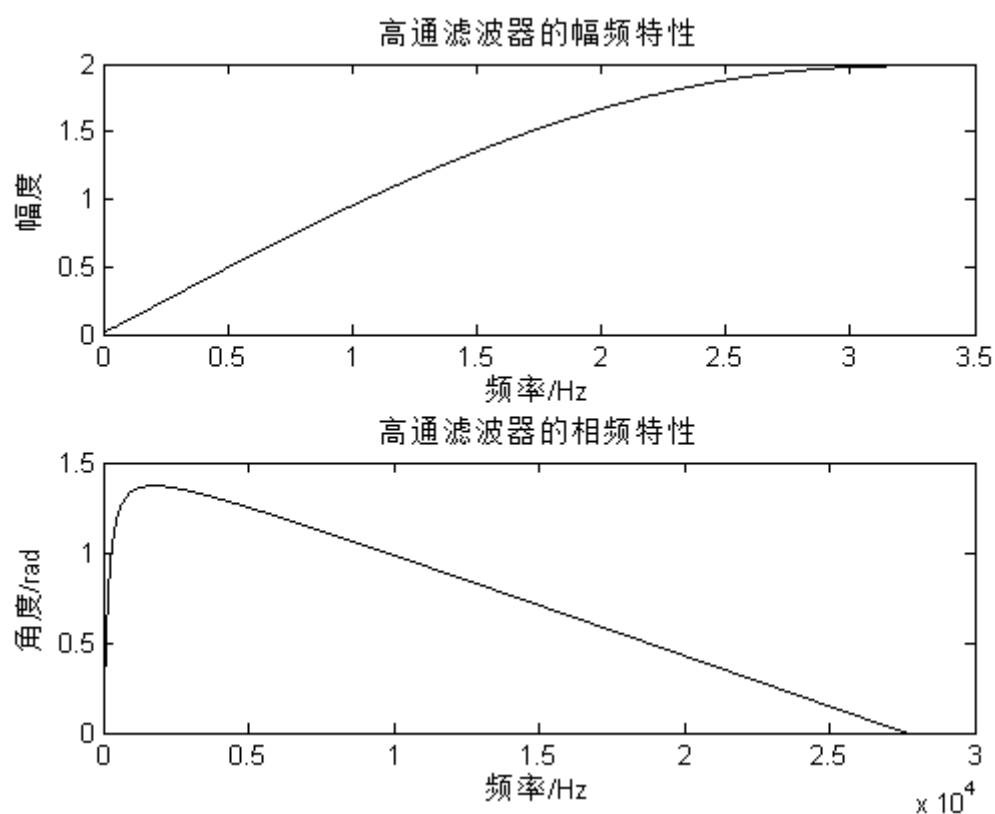


图 4.6 高通滤波器的幅频特性及相频特性

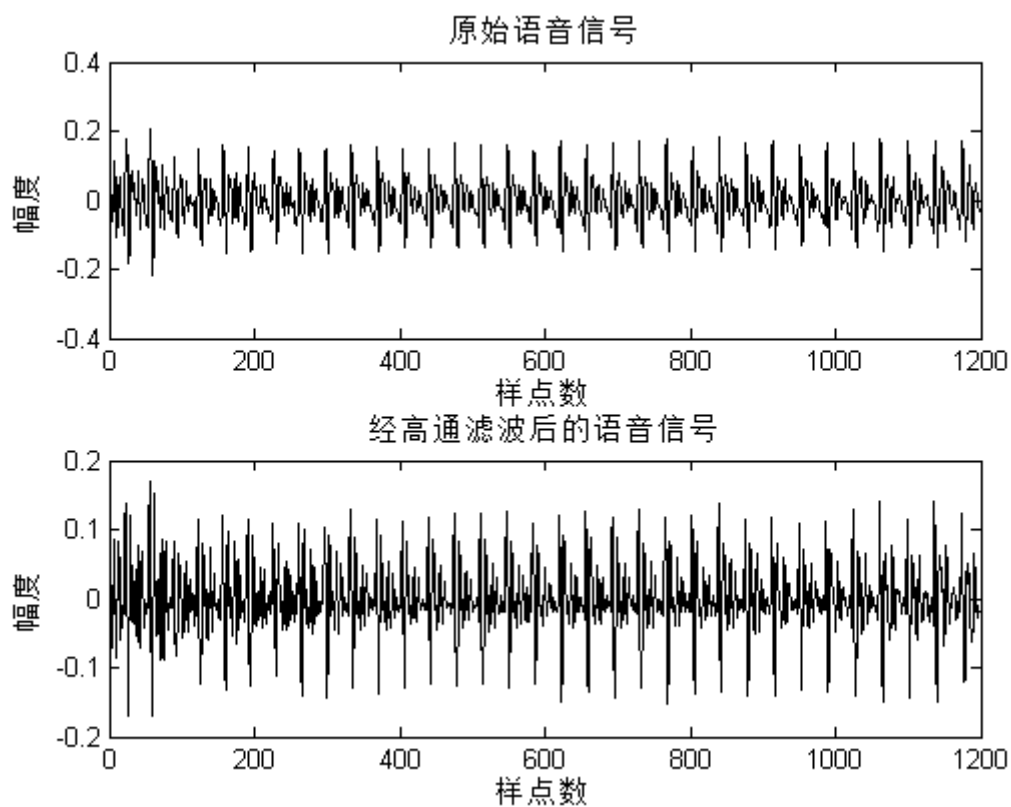


图 4.7 浊音信号预加重处理前后的时域波形

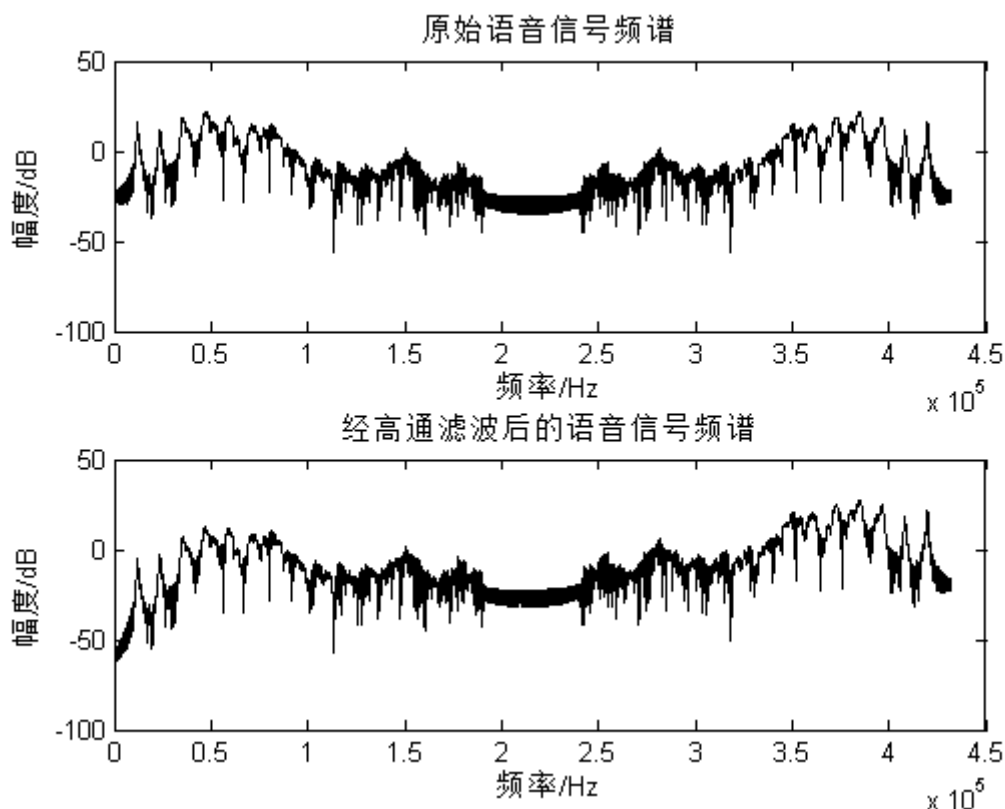


图 4.8 语音信号预加重处理前后的频谱

在预加重之后，把语音分帧加窗。通常认为语音信号在短时间内(10ms~30ms)是相对稳定的。利用其短时平稳性，将长的语音分割为若干帧分别处理。为了保证各帧之间的平滑，分帧通常采取互相重叠的方式。通常重叠部分设置为帧长的二分之一。图 4.9 是帧长帧移示意图。

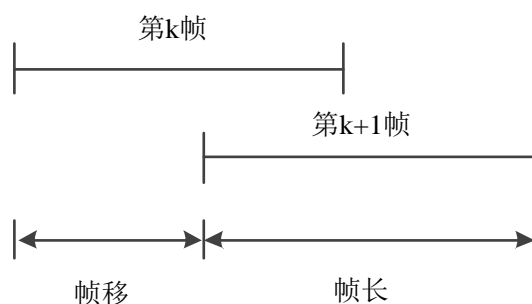


图 4.9 帧长帧移示意图

实际应用中的语音窗有矩形窗、汉明窗、三角窗等。
矩形窗函数如下：

$$w(n) = \begin{cases} 1 & 0 \leq n \leq N-1 \\ 0 & \text{其它} \end{cases} \quad (4-9)$$

汉明窗函数如下：

$$w(n) = \begin{cases} 0.54 - 0.46 \cos[2\pi n / (N-1)] & 0 \leq n \leq N \\ 0 & \text{其它} \end{cases} \quad (4-10)$$

三角窗函数如下：

L 为奇数时：

$$w(n) = \begin{cases} \frac{2n}{L+1} & 1 \leq n \leq \frac{L+1}{2} \\ \frac{2(L-n+1)}{L+1} & \frac{L+1}{2} < n < L \end{cases} \quad (4-11)$$

L 为偶数时：

$$w(n) = \begin{cases} \frac{2n}{L} & 1 \leq n \leq \frac{L}{2} \\ \frac{2(L-n+1)}{L} & \frac{L}{2} + 1 \leq n \leq L \end{cases} \quad (4-12)$$

在本文所提的噪声幅度谱估计语音增强方法中，需要用到重叠相加法来合成时域的语音波形，而窗函数的选择对合成语音的误差有着较大的影响，为了提高语音的合成质量，需要选择适当的窗函数。因此本节进行了以下实验，如图 4.10 所示：

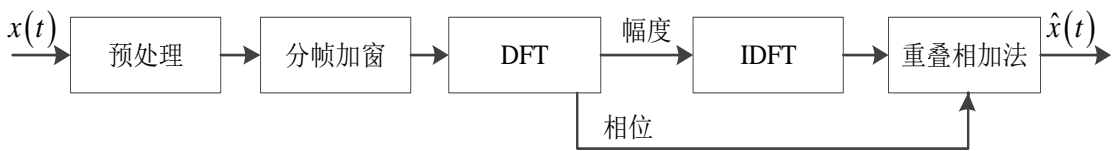
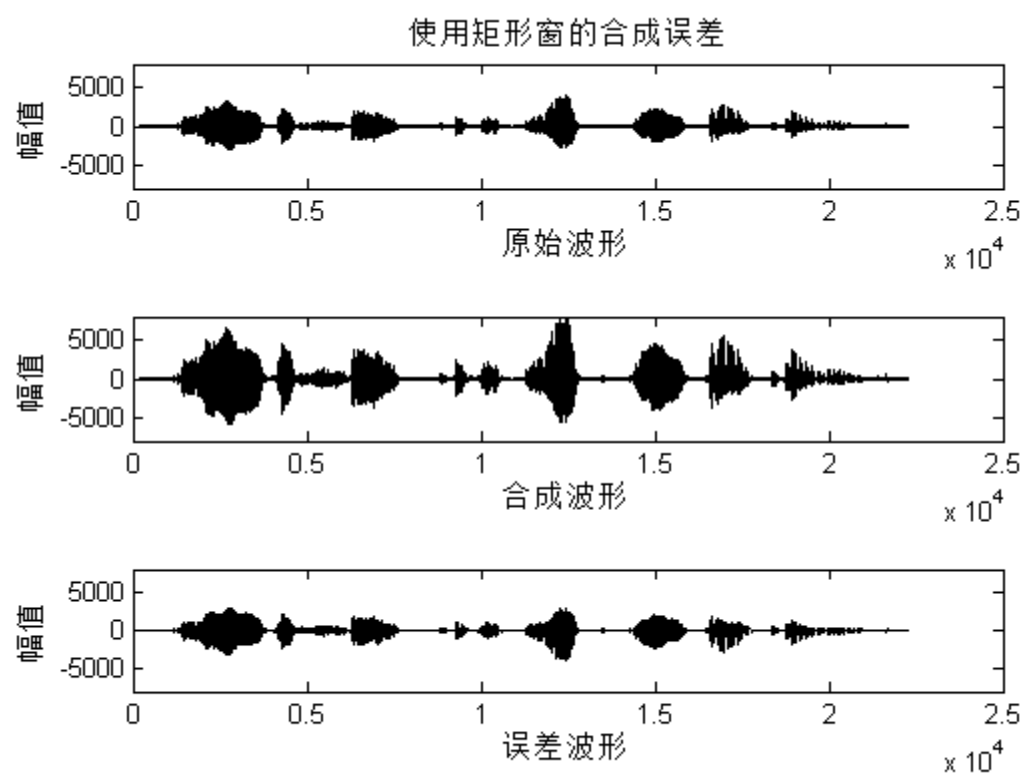
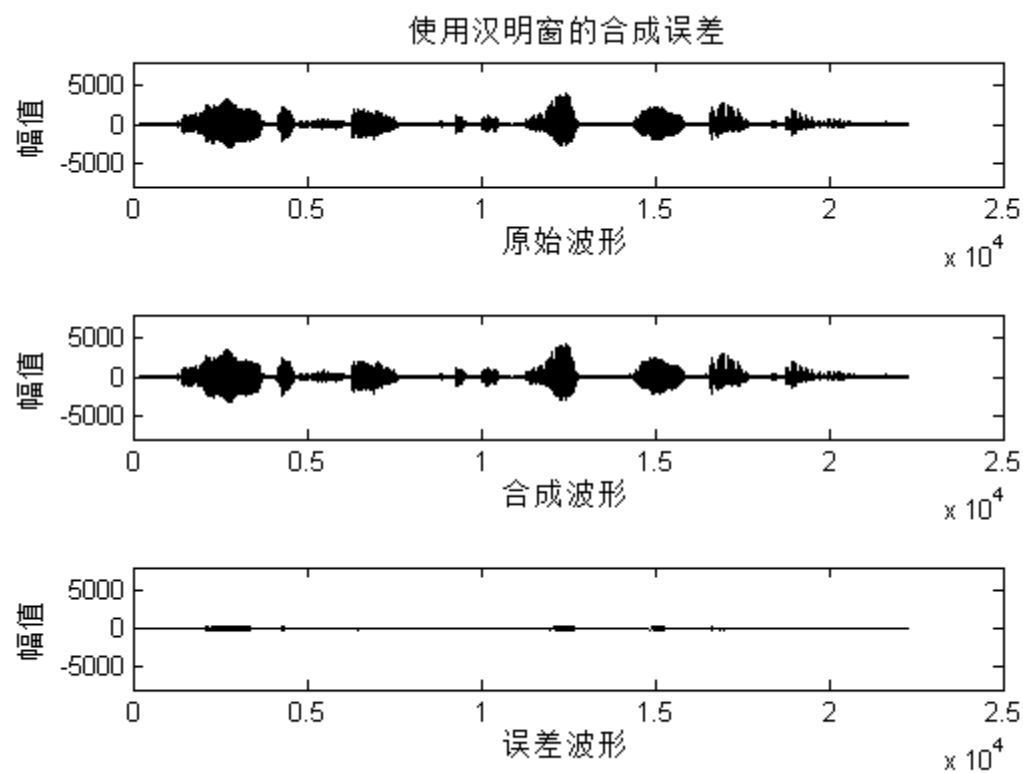


图 4.10 语音信号的分解和合成

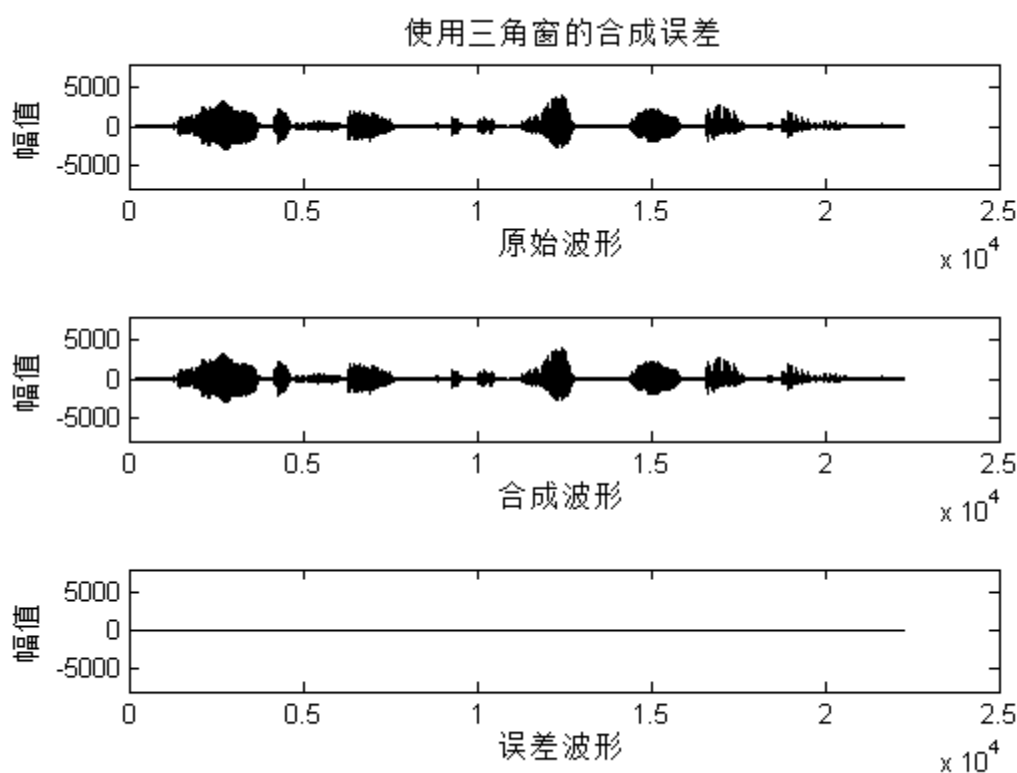
语音信号的分解和合成实验中，使用的语音采样率为 8kHz，在分帧加窗阶段，使用的帧长为 256 个采样点，帧移为 128 个采样点。分别采用了矩形窗，汉明窗和三角窗，并得到了各自的采用重叠相加法得到的时域波形 $\hat{x}(t)$ ，将其与原始波形 $x(t)$ 进行比较，观察误差的大小，结果如图 4.11 所示。



(a)



(b)



(c)

图 4.11 各窗函数的合成误差

图 4.11 中(a)为使用矩形窗的合成误差, (b)为使用汉明窗的合成误差, (c)为使用三角窗的合成误差。各图中的误差波形是使用原始波形减去合成波形得到的。

从图 4.11 中可以看出, 当使用矩形窗时, 合成误差最大, 汉明窗的合成误差次之, 使用的窗函数为三角窗时, 合成波形和原始波形的误差最小。所以, 本文以后的实验中, 所使用的窗函数都为三角窗。

4.3 噪声分类模块设计

在噪声幅度谱估计的语音增强方法的增强阶段, 对于带噪语音, 首先使用其第一帧对带噪语音中的噪声类型进行了分类, 然后提取带噪语音的幅度谱特征, 根据噪声分类的结果, 使用与之对应的深度信念网络来进行处理。这样, 对带噪语音中噪声的分类正确与否将会对最终的增强语音的质量产生巨大影响。本节将对噪声分类模块进行详细介绍。

本节分别使用 BP 神经网络和栈自动编码器对噪声数据库 NOISEX-92 中的八种噪声进行分类, 这八种噪声分别为飞机引擎噪声, 餐厅内噪声, 车辆噪声, 机枪噪声, 白噪声, 粉红噪声等。图 4.12 为噪声分类模块的系统框架。

基于 BPNN 或 SAE 的噪声分类实验整体的步骤为：

(1) 对噪声进行分割，将 NOISEX-92 中的时域噪声数据分割为两部分，分别为训练样本和测试样本，其中训练样本占全部数据的五分之四，测试样本占全部数据的五分之一。

(2) 对噪声进行预处理，分帧加窗操作，这里使用的窗函数也为三角窗。

(3) 提取特征参数，与基于深度信念网络的噪声幅度谱语音增强方法一致，这里使用的特征为噪声的幅度谱。

(4) 为各类噪声的特征参数设置标签，例如使用向量 $(1,0,0,0,0,0,0)^T$ 作为餐厅内噪声的标签，使用向量 $(0,1,0,0,0,0,0)^T$ 作为白噪声的标签，别的噪声类型的标签可依次类推。

(5) 使用训练样本对模型进行预训练和微调。

(6) 使用测试样本对模型的分类效果进行测试，并统计对各类噪声分类的正确率和整体的正确率。

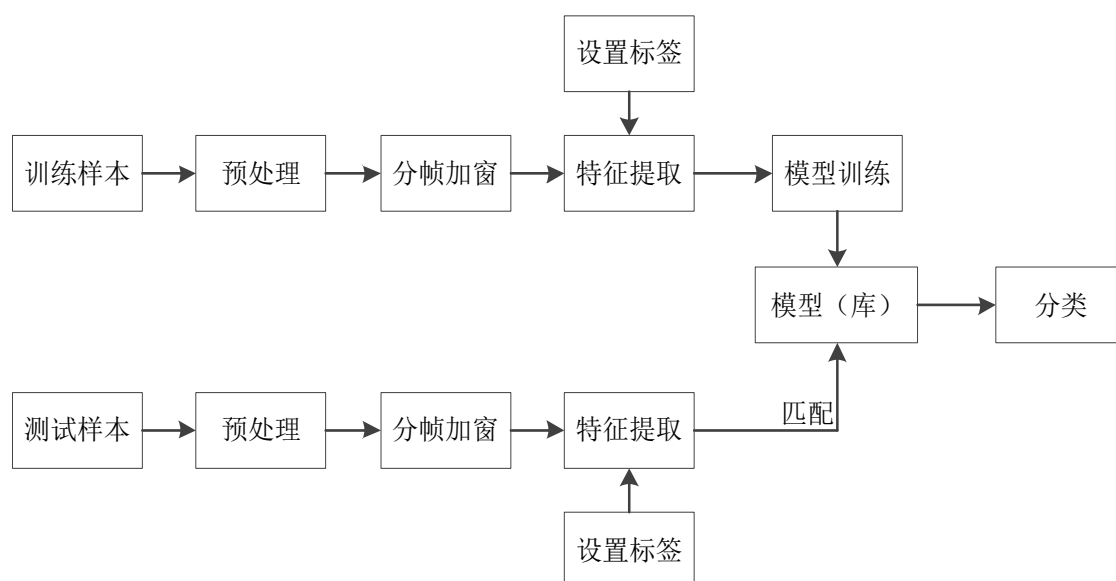


图 4.12 噪声分类方法系统框架

图 4.12 所示的噪声分类方法框架同时适用于 BP 神经网络和栈自动编码器，只需要将所使用的模型改为 BP 神经网络或者栈自动编码器就可以了。

NOISEX-92 噪声数据库中的噪声采样频率均为 8kHz，时长为 3 分 55 秒。

实验中，采用帧长为 32ms，即 256 个采样点，帧移为 16ms，即 128 个采样点。因为帧长为 256 个采样点，进行 DFT 变换之后得到频谱长度为 256，进行取模后得到噪声的幅度谱。由于幅度谱对称，所以幅度谱向量的前 129 个数据即可包含该帧的所有幅值信息。因此本实验所设计的网络模型的输入层为 129 维，即输入层有 129 个神经元。另外，因为所要分类的噪声种类共有 8 种，所以模型的输出应该为一个 8

维矢量，即输出层有 8 个神经元。

4.3.1 BP 神经网络分类

BPNN 可以很好的构建数据从输入到输出的非线性转换。本实验采用加入动量项的 BP 算法。这样可以既可以缩短训练时间，又能保持训练过程中误差的平稳下降。

下面本文将分别讨论训练数据迭代次数和神经网络结构以及训练数据量的大小对正确率的影响。

首先研究训练数据的迭代次数对分类正确率的影响。进行这项研究时的神经网络的结构设置为 129-100-100-8。正确率的变化如图 4.13 所示：

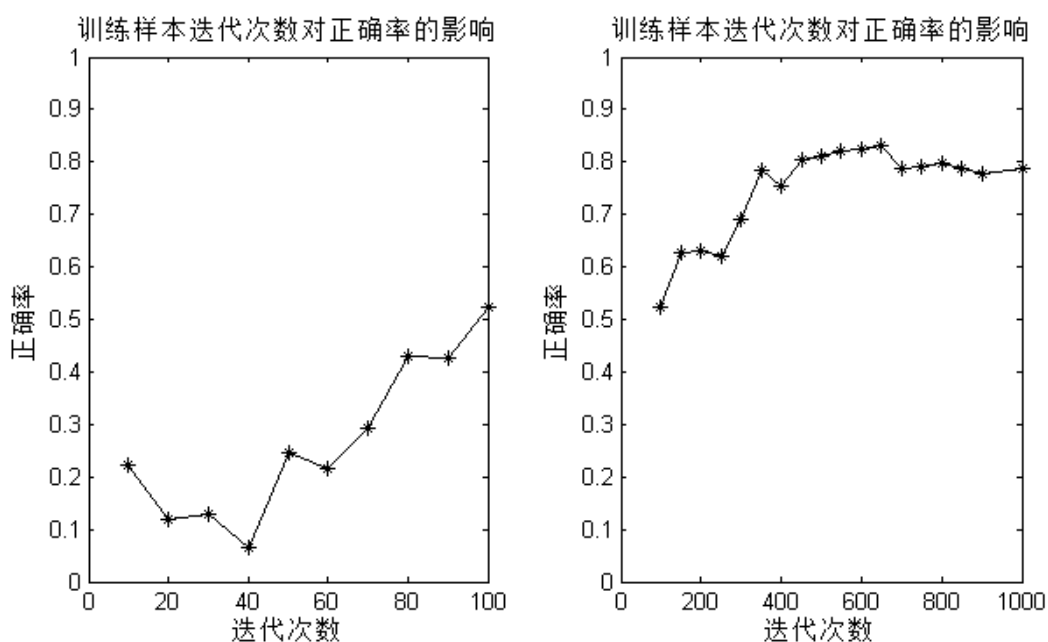


图 4.13 训练样本迭代次数对分类正确率的影响

图 4.13 中绘制了两张图，其中左图是迭代次数为 10 到 100 次时的正确率，右图是迭代次数为 100 到 1000 次时的正确率。需要说明的是，10 到 100 次之间的间隔是 10, 100 到 1000 次之间的间隔是 50。

从图中不难看出，对于隐藏层为 100-100 的 BP 网络，迭代次数较少时，即迭代次数为 10 到 60 次时，正确率的走势并未体现出规律性。这是由于迭代次数少时网络未能充分的学到输入数据和监督数据之间的非线性关系，加之网络的初始权值和偏置值是随机产生的，会对最终训练完成的网络产生一定的影响，这些因素可能会造成网络的分类性能不稳定。迭代次数为 60-650 次时，训练过程中误差越来越接近于最小值，使得网络的性能呈逐步提高的趋势，正确率仍有波动也是因为 BPNN 的初始网络参数是随机产生的原因。当迭代的次数非常大，超过 700 次时，出现了正确率不再明显提升，反而略有下降的情况，这是由于随着迭代的次数逐渐增加，误差逐步减小，

造成网络可能会出现过拟合的现象，这表现为网络对其学习过的样本能很好的分类，而对新的样本分类能力差。然而实验中，测试样本与训练样本是不同的，它们之间的分布存在一定差异，当网络过拟合时，就会导致正确率下降。另外一个可能的原因是训练过程中网络收敛到了一个局部最优值。因此，对于给定结构的 BPNN 的训练需要选择适当的迭代次数。

在本文的实验中，保存了分类正确率最高时，网络的权值和偏置值，即保存了性能最好的网络。采取的措施是，当新一轮的网络训练完成时，测试其正确率，当新的网络的性能更好时，则替换原先保存的网络。在图 4.13 中，当迭代次数为 650 次时，达到最高的正确率 83.1%。

下面我们讨论 BP 神经网络的层数以及节点数对其分类正确率的影响。由于网络的层数过多时，训练会变得十分困难，并且容易陷入局部最优值，因此，只讨论隐藏层分别为 1、2 层时，隐藏层神经元个数为 20、50、100 时的几种情况。对于每一种结构的网络，我们选取的迭代次数为 200。表 4.1 给出了不同模型的正确率。

表 4.1 不同 BP 网络模型的正确率

隐藏层数	节点结构	正确率	节点结构	正确率	节点结构	正确率
一层	20	0.4390	50	0.5302	100	0.6421
二层	20-20	0.4542	50-20	0.6858	100-20	0.5533
	20-50	0.5615	50-50	0.5450	100-50	0.6969
	20-100	0.5191	50-100	0.5262	100-100	0.7176

分析表 4.1 可以得出以下结论：

纵向来看，含两个隐藏层的 BP 神经网络的分类正确率要倾向于高于只含一个隐藏层的网络，原因在于随着网络隐藏层数的增加，其特征表达的能力越来越强，非线性映射能力也越强，因此网络进行分类的性能也就上升。

横向来看，并非隐藏层神经元的个数越多正确率就越高，例如 50-20 结构的正确率要高于 100-20 结构的网络，原因在于过多的神经元个数会引起对训练样本的过拟合，从而导致对测试样本不能很好的识别。

最后，讨论训练样本数据量的多少对于 BPNN 分类性能的影响。设置网络的结构为 129-100-100-8，分别使用训练数据的 1/4，2/4，3/4 和全部对网络进行训练，迭代次数设定为 200 次。图 4.14 给出了训练数据量的大小对网络分类性能的影响。

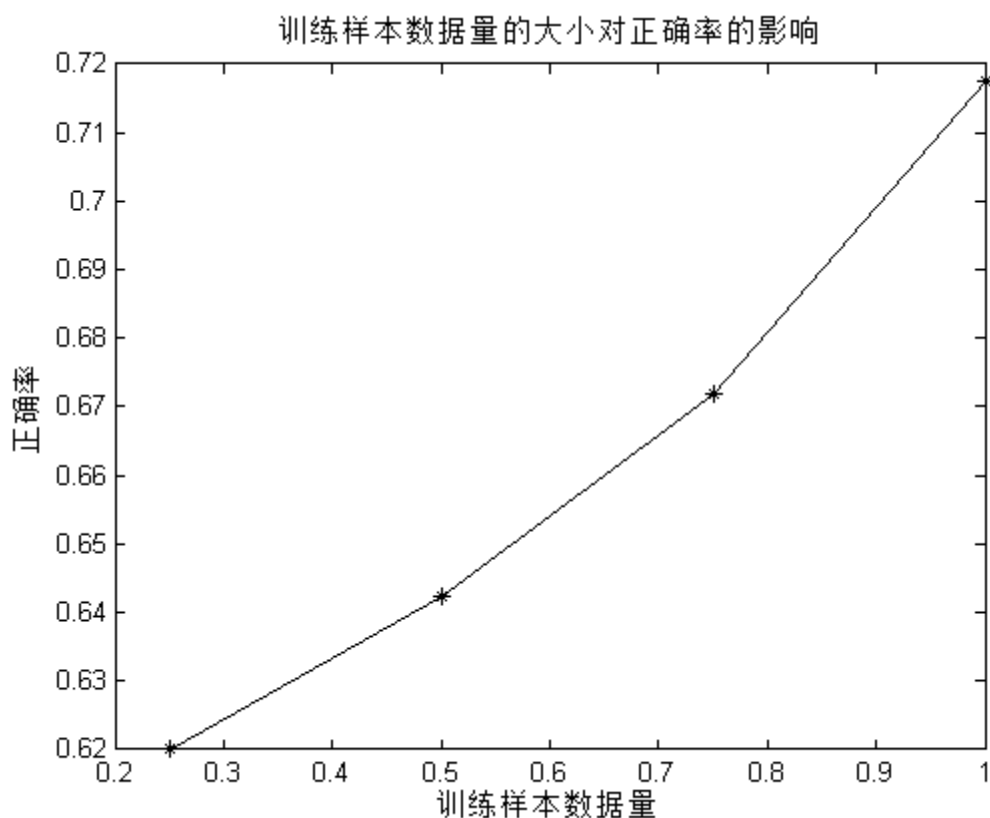


图 4.14 训练样本数据量的大小对正确率的影响

从图 4.14 中可以看出，随着训练数据量的增加，正确率也逐步提高。这是因为更多的训练数据使得网络的特征表达能力更接近于全体数据的分布。

4.3.2 栈自动编码器分类

栈自动编码器对噪声进行分类需要无监督预训练和有监督微调相结合。与对 BP 神经网络对噪声的分类类似，下面讨论训练数据迭代次数对正确率的影响，并与 BP 神经网络的分类效果进行对比。

设置 BP 神经网络和栈自动编码器的网络结构均为 129-100-100-8，令栈自动编码器的预训练次数设置为 10 次，将微调的迭代次数和 BPNN 的迭代次数设置为 100, 200, ..., 1000。将 SAE, BPNN 对噪声的分类正确率绘制于图 4.15 中：

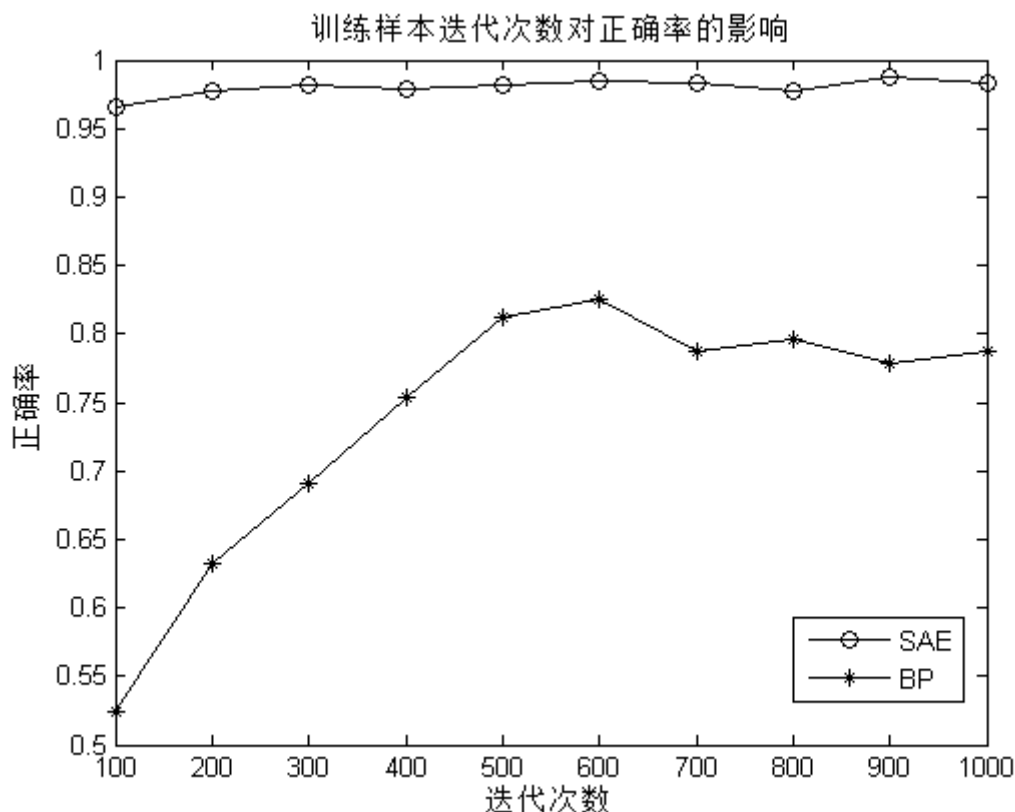


图 4.15 SAE,BPNN 正确率随迭代次数变化图

从图 4.15 中可以看出,使用栈自动编码器对噪声进行分类的正确率要远高于使用 BPNN 的正确率。并且,仔细观察不难看出,BPNN 在达到较高的迭代次数之后,正确率出现震荡,不但幅度较大而且正确率还有下降的趋势,而经过预训练的栈自动编码器的正确率比较稳定。这表明预训练可以使网络的初始参数的分布更接近全局的最优值,因此有效地避免了局部震荡。在图 4.15 中,使用 SAE 的分类的正确率最高为 98.7%。

此外还讨论了栈自动编码器的结构和训练数据量的大小对于正确率影响,这与 BP 神经网络有着相似的结论,即网络具有较深的结构时,正确率通常比较浅结构下的正确率要高,用来训练网络的数据量越多,网络也更倾向于具有更高的正确率。

4.4 子空间语音增强方法的改进

4.4.1 传统子空间语音增强方法的不足之处

在传统的子空间语音增强方法中,需要估计噪声的协方差矩阵 R_N 并进行特征值分解来估计噪声特征值 Λ_N ,通常的做法是对带噪语音进行语音活动检测(Voice Activity Detection,VAD),将其明确的区分为有声段或者无声段,从而仅在无声段估计

噪声的协方差矩阵 R_N ，对其做 EVD 分解得到噪声的特征值估计 Λ_N 。并且，用这个估计来作为后续有声段中的噪声特征值估计。这就存在两个问题，一个问题是当后续的 VAD 检测出的有声段中噪声的特性发生了变化时，无法得到对应的噪声来及时地更新对噪声特征值的估计，第二个问题是在信噪比较低或者是噪声类型为非平稳噪声的情况下，通过语音活动检测来区分有声段和无声段也将变得十分困难，随着语音活动检测的误判率增加，对噪声特征值的估计更加得不到及时的更新，这就会导致子空间语音增强算法性能的下降。

4.4.2 改进的子空间语音增强方法

上一小节中，指出了传统子空间法在估计噪声特征值 Λ_N 时存在的两个问题，针对子空间语音增强算法中该模块存在的这两个问题，我们将分别提出两种对应的方案。

首先是第一个问题，即采用 VAD 检测的方法，在有声段时，噪声的特征值无法及时更新的问题。解决方法是：在 VAD 检测为有声段时，利用 4.1 节中提出的深度信念网络噪声幅度谱估计的语音增强方法（该深度信念网络可以将带噪语音的幅度谱映射到对应的噪声的幅度谱，再用幅度谱相减的方式获得对增强语音的幅度谱估计，最后结合带噪语音的相位谱采用重叠相加法恢复出时域的波形）得到有声段的增强语音之后，用带噪语音的时域波形减去增强语音的时域波形，得到对有声段的噪声的估计，那么此时就可以使用该噪声估计来更新噪声的特征值 Λ_N 。这就有效解决了在有声段无法更新噪声的问题。该方案对子空间增强方法的改动部分步骤如下：

(1) 对一帧带噪语音进行 VAD 检测，将该帧分为语音帧或是噪声帧。

(2) 如果第一步中，VAD 检测的结果是有语音帧，则对该帧进行提取特征的操作（这里提取的是幅度谱特征），将所提取的幅度谱向量输入已经训练好的深度信念网络，得到网络的输出向量，即对该帧中噪声的幅度谱估计。使用带噪语音的幅度谱减去 DBN 估计的噪声幅度谱，取得对增强语音的幅度谱估计。并利用带噪语音的相位谱，进行 IDFT 变换合成该帧增强语音的时域波形。最后用该帧带噪语音的时域波形减去增强语音的时域波形，得到噪声的时域波形，即对该有语音帧的噪声的估计。

(3) 得到对有语音帧中的噪声的估计后，使用这个估计出噪声计算协方差矩阵并做 EVD 分解来估计噪声的特征值。

(4) 如果第一步中，VAD 检测的结果是噪声帧，则直接计算该帧噪声的协方差矩阵并估计特征值。

(5) 使用第三步或者是第四步中得到的噪声特征值更新整体的噪声特征值。

然后是第二个问题，即在信噪比较低或者噪声类型为非平稳噪声的情况下，VAD 检测的误判率会上升从而导致子空间法性能下降的问题。对于这个问题，解决方案依

旧是利用 4.1 节中提出的深度信念网络噪声幅度谱估计的语音增强方法，但是不再进行 VAD 检测。具体思路为：对带噪语音中的噪声进行估计，并把估计出的噪声当做是原始带噪语音中的噪声。那么就可以使用这个噪声在子空间法中每一帧都对噪声特征值进行更新，这就可以避免 VAD 检测误判对增强效果的影响。该方案对子空间增强方法的改动部分步骤如下：

(1) 使用基于深度信念网络的噪声幅度谱估计语音增强方法对带噪语音进行处理，得到一个增强语音。

(2) 在时域用带噪语音减去增强语音，得到一个噪声，这个噪声即为对原始带噪语音中噪声的估计。

(3) 使用子空间法对原始带噪语音进行处理，在每一帧都使用第二步中得到的噪声来更新整体的噪声特征值。

4.5 本章小结

本章首先介绍了本文所提出的基于深度信念网络的噪声幅度谱估计的语音增强方法，然后介绍了语音增强方法的预加重和加窗分帧部分，并对为何选择三角窗作为窗函数进行了实验分析。接着对噪声分类模块的设计进行了详细介绍，并分别使用 BP 神经网络和栈自动编码器用于噪声分类，对分类模块的仿真表明，栈自动编码器的分类性能要明显好于传统的 BP 神经网络。接下来详细分析了传统子空间语音增强方法中存在的不足之处：使用 VAD 检测对噪声进行估计，无法及时更新噪声的特征；低信噪比时，VAD 检测性能的下降也会导致子空间方法性能的下降。最后结合本文所提出的基于深度信念网络的噪声幅度谱估计的语音增强方法分别对传统子空间方法的不足之处给出了对应的解决方案。

第五章 实验结果及分析

5.1 语音质量评价标准介绍

语音质量的评价标准主要包括两个方面，分别是它的清晰度和可懂度。评价可以通过主观听音测试或者客观音质测量。主观评价主要包括让一组测试人员试听原始的纯净语音和处理后产生失真的语音并进行比较，根据预设好的标准对音质进行等级划分。客观评价是对原始纯净语音和进行处理后的语音在数学上进行对比，利用它们之间的“距离”的大小来量化音质。因此，为了确认客观评价的可信程度，其结果需要与主观听音测试的结果有较好地关联。

5.1.1 主观评价

主观评价是指根据测试人员的主观听觉感受来对受测试的语音进行评价。一个好的增强算法，不但要使增强后的语音可懂度提高，还要能够使语音质量提高。这是因为语音有可能可懂度很高，但是音质很差。还有可能两个不同算法处理后的语音，它们具有相同或相近的可懂度，但听起来一个流畅自然，一个粗糙难以接受。由于质量评估是非常主观的，所以其可靠性是一个问题。常用的主观评价方法有：平均意见得分(Mean Opinion Score, MOS)，判断满意度测量(Diagnostic Acceptability Measure, DAM)等。下面分别对它们进行介绍。

(1) 平均意见得分

MOS 是使用最广泛的直接主观质量评估方法^[45]。它根据测试人员对语音的主观听觉感受来对受测试语音的质量进行打分，打分标准如表 5.1 所示。MOS 测试分为两个阶段：训练和评估。在训练阶段，测试人员收听一组分别代表了非常好、中等和很差的质量等级的参考信号。这个阶段也成为“锚定阶段”，是非常重要的，因为它用来均衡测试人员对语音评级的主观范围的。在评估阶段，测试人员收听测试语音，并按照表 5.1 中的 5 个等级对受测语音的质量打分。

表 5.1 MOS 评分表

MOS 评分	质量等级	失真度级别
5	优	不易察觉
4	良	稍有察觉
3	中	有察觉, 稍微可厌
2	差	察觉明显, 可以忍受
1	坏	无法忍受

(2) 判断满意度测量

判断满意度测量是由 Voiers 提出的多维音质评价方法^[46]。DAM 测试从三个不同的方面评价语音的质量：参变、元变和同变。测试人员需要在 6 个维度上对语音失真进行评分，在 4 个维度上对背景失真评分。DAM 测试总共产生 16 个分数，其分数的模式取决于语音失真的类型。DAM 中的参变方法可以给出准确可靠的语音质量评分。与 MOS 测试相比，DAM 也有着它的缺点，其测试耗时很长，因为它需要事先对测试人员进行认真的训练。此外，在挑选测试人员时，还要根据他能否一直给出稳定的评分。尽管 DAM 比较麻烦和费时，但对语音质量的评价通过多维而不是单维的概念能得到更好的描述。

5.1.2 客观评价

主观测试方法对语音质量的评价非常可靠，但是因为它事先需要对测试人员进行培训，所以主观测试的方法非常耗时。因此，人们研究了客观的语音质量测度。常见的客观评价方法有：分段信噪比测度，对数谱失真，感知语音质量评价等。

(1) 分段信噪比测度

$$SNR_{seg} = \frac{1}{L} \sum_{i=0}^{N-1} \gamma \left\{ 10 \lg \frac{\sum_{n=0}^{N-1} x^2(n + \frac{LN}{2})}{\sum_{n=0}^{N-1} [x(n + \frac{LN}{2}) - \hat{x}(n + \frac{LN}{2})]^2} \right\} \quad (5-1)$$

式中， L 表示总帧数。 $\gamma = \min[\max(x, -10), 35]$ 的作用是去除过高或者过低的信噪比。因为过高或者过低的数值对整个语音的感知意义不大。 SNR_{seg} 的计算是对所有的语音帧的信噪比求几何平均。

(2) 对数谱失真

$$LSD = \frac{1}{L} \sum_{l=0}^{L-1} \left\{ \frac{1}{N/2+1} \sum_{k=0}^{N/2} [10\lg TX(k,l) - 10\lg T \hat{X}(k,l)]^2 \right\}^{1/2} \quad (5-2)$$

对数谱失真是在频域上定义的,用来衡量语音的失真大小的测度。对数谱失真的动态范围在 50dB 以内,是纯净语音谱和增强语音谱的比值。**LSD** 的值越大表示受测试语音的失真程度越大,**LSD** 的值越小,表示受测试语音失真越小,与原始语音越接近。

(3) 语音质量感知评估

语音质量感知评估是一种在各种编解码器和网络条件下均有着可靠性能的新的客观测度。2000 年, PESQ 测度在 ITU-T 第 12 研究组组织的评比中,取代了旧的基于感知语音质量测量(Perceptual Speech Quality Measure, PSQM),被选为 ITU-T P.862 建议^[47]。PESQ 测度的计算结构如图 5.1 所示,纯净语音和增强语音被调整到相同听觉强度水平后通过一个滤波器滤波。信号按时间对齐以校正延时的影响,然后通过听觉变换处理以获得响度谱。在计算响度谱之间的差分后,在时间和频率上求平均以得到 PESQ 评分。

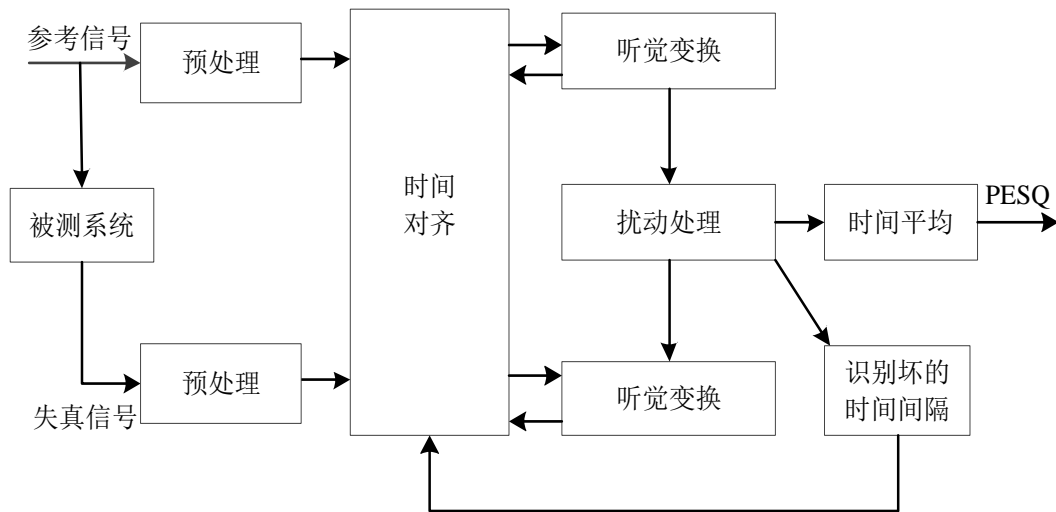


图 5.1 PESQ 测度计算框图

PESQ 得分的范围是-0.5 到 4.5 之间,并且在大多数情况下,得分与 MOS 得分近似。

5.2 基于 DBN 的噪声幅度谱估计语音增强方法性能分析

本节将对 4.1 节中所提出的基于深度信念网络的噪声幅度谱估计的语音增强方法

进行实验仿真。实验是基于 IEEE-Harvard 语料库^[48]中的纯净语音构建的。噪声数据来自于 NOISEX-92 噪声数据库。采用纯净语音和噪声相加的方法，可以合成大量的包含各种信噪比，各种类型噪声的训练数据对。本实验中，选取了 IEEE-Harvard 语料库中的 720 句纯净语句，这些语句由男性说话人发音，原始的以 25kHz 采样的宽带录音首先通过一个 300~3200Hz 的带通滤波器，然后下采样至 8kHz。纯净语音和八种噪声以不同的噪声系数叠加在一起，合成大量的信噪比分别为 20dB，15dB，10dB，5dB，-5dB 的带噪语音，这样就得到了充足的带噪语音和噪声数据对用来训练基于深度信念网络的增强模型。针对每一种噪声，分别对带噪语音和噪声数据对进行幅度谱特征提取的操作，并且构造不同的深度信念网络进行训练，即对每一种类型的噪声，都训练出了一个与之对应的适合估计其幅度谱特征的 DBN 网络。

实验中，纯净语音和噪声的采样率均为 8kHz，利用它们合成的带噪语音的采样率也为 8kHz。在对训练数据对进行分帧加窗时，采用的帧长为 32ms，即 256 个采样点，帧移为 16ms，即 128 个采样点，窗函数为三角窗。使用 DFT 变换得到每帧的傅氏变换系数，取模得到幅度谱特征。然后，129 维的幅度谱特征被用来训练深度信念网络，每个 RBM 的预训练迭代次数设置为 20 次，微调的迭代次数设置为 10 次。训练深度信念网络前，将所有的幅度谱特征归一化到 0 和 1 之间。

首先讨论 DBN 网络的结构深浅对语音增强性能的影响。为了方便讨论，以白噪声为例进行讨论。表 5.2 为使用不同的网络进行语音增强的平均 PESQ 得分。

表 5.2 SNN、DBN 语音增强方法平均 PESQ 得分

	Noisy	SNN1	SNN2	DBN1	DBN2	DBN3
SNR20	2.828	3.114	3.151	3.095	3.123	3.158
SNR15	2.466	2.846	2.907	2.830	2.885	2.892
SNR10	2.171	2.596	2.624	2.580	2.652	2.642
SNR5	1.861	2.289	2.311	2.288	2.361	2.325
SNR0	1.634	2.020	2.035	2.029	2.097	2.057
SNR-5	1.448	1.697	1.717	1.705	1.813	1.728
Ave	2.068	2.427	2.458	2.421	2.488	2.467

在表 5.2 中，比较了 SNN 和 DBN 用于语音增强的平均 PESQ 结果。SNN 表示浅层神经网络(SNN, Shallow Neural Networks)。SNN1 有一个隐藏层，隐藏层节点是 512，SNN2 有一个隐藏层，隐藏层节点数 2048(=1024*2)，他们使用的训练数据为 90 个语句（包含 -5dB, 0dB, ..., 20dB 等不同的信噪比）。可以发现，SNN2 能取得比 SNN1 更好的性能（前者的平均 PESQ 值是 2.458，后者是 2.427），这就表示更多的隐藏层神

神经元节点数可以使得带噪语音幅度谱到噪声幅度谱的映射更容易。DBN1、DBN2 和 DBN3 中的数字表示的是隐藏层的层数，每个隐藏层包含 1024 个神经元。DBN 使用的训练数据与 SNN 相同，为 90 个不同信噪比的语句。DBN2 取得了最好的性能，与 SNN2 相比，虽然二者的神经元个数是一样的，但是具有更深的结构的 DBN2 更有利于学到带噪语音幅度谱到噪声幅度谱的映射，所以 DBN2 的性能比 SNN2 要好。

接下来讨论训练数据量的多少对 DBN 网络增强性能的影响。固定网络的结构为两个隐藏层，每个隐藏层包含 1024 个神经元。使用训练数据量分别为 30 个语句，90 个语句，150 个语句，300 个语句来训练 DBN，结果如图 5.2 所示。

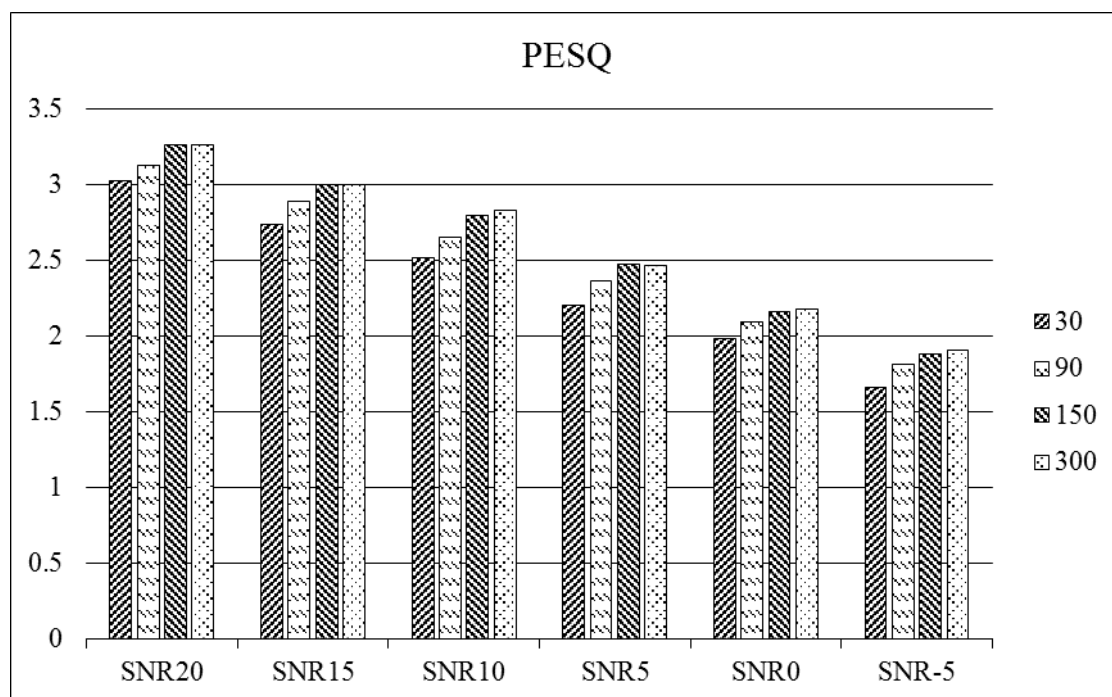


图 5.2 不同数据量训练网络的语音增强效果

图 5.2 比较了不同的训练样本数据量在白噪声条件下得到的 DBN 网络的增强性能。训练数据量如果只有 30 条语句的时候，性能是比较差的，这就说明训练数据量的大小对模型的性能有非常重要的影响。另外，在用于训练的数据量只有 30 条语句的时候，考虑到有 6 种不同的信噪比，实际上分配到每种信噪比的语句是非常少的，只有五条语句，这就做不到完全覆盖噪声的所有特征，因此导致了比较差的性能。随着训练数据量的增加，系统的性能一直处于单调提升的状态，一直到 150 条语句。其中当把训练数据量从 30 条语句增加到 90 条语句，和从 90 条语句增加到 150 条语句的时候系统的性能都有着比较大的提升。但是从 150 条语句的训练数据量增加到 300 条语句的过程中，性能并没有显著变化，这就说明在基于 DBN 噪声幅度谱估计的语音增强系统中，总的训练数据量并不需要太多，但是不同的信噪比带噪语音和噪声数据对要有一些训练数据来覆盖。在训练数据量为 150 条语句的时候，和训练数据量为

300 条语句的时候相比，语音增强的性能并没有太大的变化。这是因为噪声数据库中实际有效的白噪声只有 3 分 55 秒，所以当把训练数据从 150 条语句，增加到 300 条语句时候，300 条语句的训练数据中包含了很多冗余信息，有效信息和 150 条语句时候的训练数据可能差不多。

5.3 改进的子空间语音增强方法性能分析

本节将对 4.4 节中所提出的针对传统子空间语音增强算法中的两点不足之处的两个改进方案进行试验仿真，并和传统的子空间法与上一节分析的噪声幅度谱估计的语音增强方法一起进行比较。为方便起见，在这里，将上一节中的噪声幅度谱估计的语音增强方法称为深度信念网络法，将子空间方法的两种改进方案分别称为改进的子空间法 1 和改进的子空间法 2。其中深度信念网络法的网络配置是两个隐藏层，每个隐藏层 1024 个神经元，使用 150 条语句进行训练。

图 5.3 所示为噪声类型为白噪声，带噪语音信噪比为 0dB 时，分别采用四种增强方法进行处理的仿真结果。

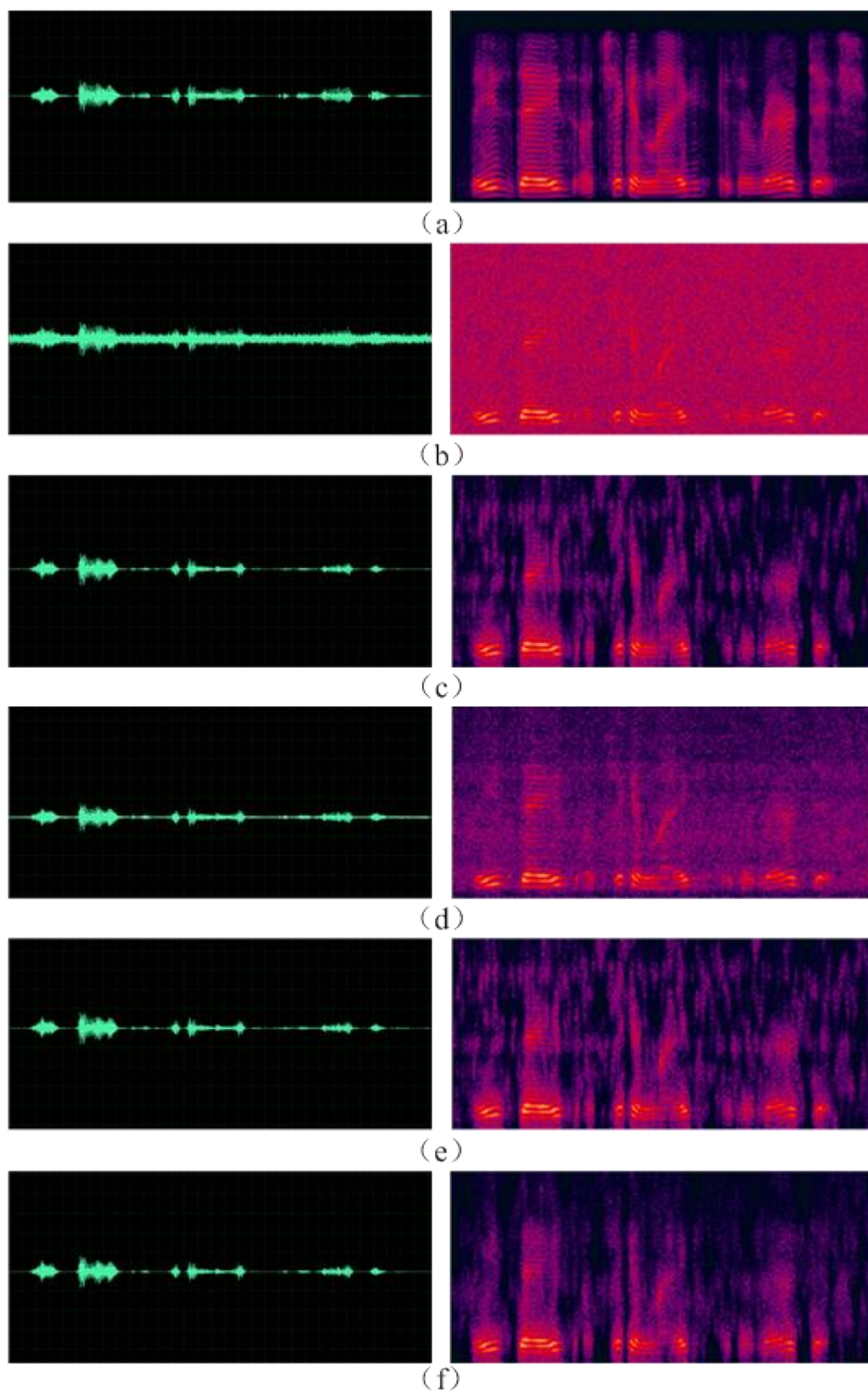


图 5.3 带噪语音处理前后时域波形及相应的语谱图（白噪声 $\text{SNR} = 0\text{dB}$ ）

图 5.3 中由上至下分别为：(a)纯净语音的时域波形及相应的语谱图、(b)带噪语音的时域波形及相应的语谱图、(c)传统子空间方法得到的增强语音的时域波形及相应的语谱图、(d)深度信念网络法得到的增强语音的时域波形及相应的语谱图、(e)改进的子空间法 1 得到的增强语音的时域波形及相应的语谱图、(f)改进的子空间法 2 得到的增强语音的时域波形及相应的语谱图。

对比图(c)和(e)中的时域波形和语谱图，发现二者的区别并不明显，这是因为改进的子空间法 1 只是在 VAD 检测为语音段时，才使用深度信念网络法进行改进，而在信噪比较低的情况下，VAD 检测的误判率会大大提高，这也就导致了改进的子空间法 1 的改进程度受到了制约。对比图(c)和图(f)中的时域波形和语谱图可以发现，在静音段改进的子空间方法 2 所得到的增强语音比子空间法所得到的增强语音对噪声的去除更干净一些。这一点可以从二者的语谱图中更明显的看出。

为了更清晰地比较四种方法的效果，图 5.4 所示为四种语音增强方法在不同信噪比下的平均 PESQ 得分。

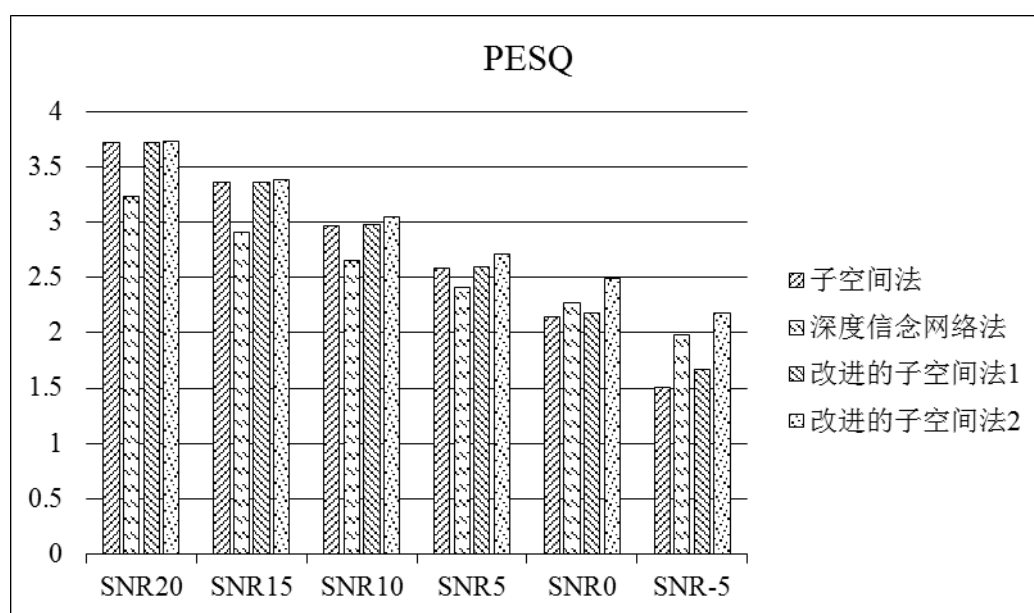


图 5.4 四种语音增强方法的性能比较

如图 5.4 中所示，比较子空间法与深度信念网络法，在信噪比较高时，子空间法的性能要好于深度信念网络法，随着信噪比的下降，二者的 PESQ 得分差距逐渐缩小，并且在信噪比为 0dB 和 -5dB 时，深度信念网络法的性能超过了子空间法。这是因为随着信噪比的降低，子空间法中 VAD 检测的误判率上升，算法无法及时更新噪声的协方差矩阵并进而更新噪声特征值，这就导致了在信噪比低时，子空间法的性能下降严重。而深度信念网络法中经过训练的 DBN 网络可以比较稳定的将带噪语音的幅度谱特征映射为其所包含的噪声的幅度谱特征，因此深度信念网络法在低信噪比降低时

表现得更为稳定一些，甚至在 0dB 和-5dB 的信噪比时性能超过了子空间法。

比较子空间法和改进的子空间法 1，发现子空间法 1 的 PESQ 得分在信噪比较高时略高于子空间法，在信噪比较低时，二者的得分差距逐渐变大。这是因为在语音帧时，子空间法无法及时更新噪声协方差矩阵，而改进的子空间法 1 却可以在语音帧时利用神经网络法得到的噪声更新噪声的协方差矩阵，这就使得信噪比高时，改进的子空间法 1 的性能略高于子空间法。随着信噪比的降低，VAD 检测的误判率上升，子空间法对噪声协方差矩阵的更新更加不及时，这就导致了改进的子空间法 1 与子空间法的性能差距逐渐变大。

比较子空间法和改进的子空间法 2，在信噪比较高时，改进的子空间法 2 有着与改进的子空间法 1 类似的性能表现，即性能略高于子空间法。但是随着信噪比的降低，改进的子空间法 2 的优势就逐渐显现出来。如图 5.4 中所示，在信噪比为 0dB 和-5dB 时，改进的子空间法 2 的性能要远高于子空间法。这是因为改进的子空间法 2 完全舍弃了 VAD 检测，取而代之的是在每一帧都使用深度信念网络法得到噪声来更新噪声的协方差矩阵。随着信噪比的下降，在子空间法因为 VAD 检测的误判率的迅速提高而导致性能严重下降时，改进的子空间法 2 完全不受此影响。但是有一点需要注意的是，改进的子空间法 2 的性能受相应的深度信念网络性能的影响，当训练得到的网络能够很好的学到带噪语音幅度谱到噪声幅度谱的非线性关系时，也就可以很好的估计出带噪语音中的噪声，这就会对改进的子空间法 2 提供很大的帮助。当训练所得到的深度信念网络的性能不好时，也会使得子空间方法 2 的性能下降。改进的子空间法 1 中也使用了深度信念网络，基于同样的原因，它也会受到网络性能的影响。

5.4 本章小结

增强算法的性能可以从其处理后的语音的可懂度和质量等方面进行评估。本章首先从主观方面和客观方面对语音质量评价的常用标准进行了介绍。主观方面主要有平均意见得分 MOS，判断满意度测量 DAM，客观方面主要有分段信噪比，对数谱失真，语音感知质量测度等。这些评价标准各有其优缺点，在使用时，可以根据被测试系统的具体目标有选择的使用。

然后对本文所提出的基于深度信念网络的噪声幅度谱估计语音增强方法进行了仿真分析，将其性能分别从 DBN 网络的层数和训练数据量的大小进行了比较。最后对子空间算法的两种改进方案进行了仿真，并和传统子空间算法，深度信念网络法进行比较，仿真结果表明，深度信念网络法在低信噪比下性能比子空间法好，改进的子空间方法 1 的性能在各个信噪比条件下都要好于子空间法，改进的子空间方法 2 的性能在各信噪比下较子空间法有明显的提高，尤其是在信噪比低的条件下。

第六章 总结与展望

6.1 本文工作总结

语音作为人们之间相互沟通最基本、最有效、最方便的方式，却常常被无所不在的噪声所干扰，语音增强技术可以抑制噪声，提取出纯净语音，在语音编码，语音识别，医疗，军事，通信网络等方面有着广泛的应用。语音增强技术经历了几十年的研究发展，取得了丰硕的成果，本文在充分研究了传统单声道语音增强算法的基础上，结合深度学习技术，提出了语音增强的新方案，并且针对子空间语音增强算法中使用 VAD 检测方法进行噪声估计的不足之处提出了改进方案。本文的具体工作如下：

(1) 对传统单声道语音增强算法中的谱减法，维纳滤波法，基于统计模型的语音增强法等进行了系统的研究，对其中的子空间语音增强方法进行了深入的研究分析。

(2) 对神经网络中最为经典的误差反向传播算法和深度学习中的主流模型栈自动编码机和深度信念网络进行了深入研究。

(3) 使用 BP 神经网络和栈自动编码器对 NOISEX_92 噪声数据库中的噪声进行分类，使用的特征为噪声信号的幅度谱，仿真结果表明，神经网络的神经元个数和网络的层数以及训练网络的数据量对分类性能有着很大的影响。栈自动编码器较 BPNN 有着更好的分类性能。

(4) 提出了基于深度信念网络的噪声幅度谱估计语音增强方法，与对噪声的分类类似，从深度信念网络的层数和训练数据量的多少对网络的语音增强性能进行分析。

(5) 子空间语音增强算法中使用 VAD 检测方法对噪声进行估计，这就无法在语音帧时及时更新噪声，而且在低信噪比条件下，VAD 检测的性能迅速降低，导致子空间语音增强算法也性能迅速下降，针对这两点不足，本文分别提出了改进的方案，并且将改进的方案和原始的子空间算法进行了仿真对比，结果表明，在低信噪比条件下，本文提出的改进方案可以明显提高子空间算法的性能。

6.2 展望

本文的研究工作仍有许多需要改进的地方，在以后的研究工作中，可以从以下几个方面进行考虑：

(1) 本文在研究 BP 神经网络和栈自动编码器对噪声的分类情况时，使用的是 NOISEX_92 噪声数据库。而实际环境中，噪声的种类异常丰富，该噪声数据库中的噪声种类确实有限的，所以，需要使用更多种类的噪声来对网络的噪声分类能力进行检验。

(2) 基于深度信念网络的噪声幅度谱估计语音增强方法中, 利用人耳对相位信息的不敏感性, 使用了带噪语音的相位来恢复增强语音的时域信号。然而, 带噪语音的相位毕竟和纯净语音的相位存在偏差, 且信噪比越低, 偏差越大。相位的失配将可能使人感觉到音质“粗糙”。因此, 在后续的工作中, 可以考虑如何对纯净语音的相位进行估计。

(3) 在使用深度信念网络对噪声的幅度谱进行估计前, 先使用 SAE 对带噪语音中的噪声进行了分类, 然后根据分类结果将带噪语音交给相应的深度信念网络进行处理。这里在进行分类时, 仅使用了带噪语音的第一帧, 这有可能是不可靠的, 毕竟 SAE 的分类正确率没有达到百分之百。因此, 可以对带噪语音的前几帧都进行分类, 对每一帧的分类结果进行统计, 将结果数最多的噪声最为最终的结果, 这样做可以提高分类的可靠性。

参考文献

- [1] 杨行峻,迟惠生等.语音信号数字处理.北京:电子工业出版社,1995
- [2] 王晶, 傅丰林, 张运伟. 语音增强算法综述[J]. 声学与电子工程, 2005(1):22-26.
- [3] M.Hawley,editor,Speech Intelligibility and Recognition, Hutchinson&Ross Inc,1977
- [4] French N R, Steinberg J C. Factors Governing the Intelligibility of Speech Sounds[J]. Journal of the Acoustical Society of America, 1945, 19(1):90-119.
- [5] Lim J S, Oppenheim A V. All-pole Modeling of Degraded Speech[J]. IEEE Transactions on Acoustics Speech & Signal Processing, 1978, 26(3):197-210.
- [6] Paliwal K, Basu A. A Speech Enhancement Method Based on Kalman Filtering[C]// Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP. IEEE, 1987:177-180.
- [7] Boll S. Suppression of Acoustic Noise in Speech Using Spectral Subtraction[J]. IEEE Transactions on Acoustics Speech & Signal Processing, 1979, 27(2):113-120.
- [8] Berouti M, Schwartz R, Makhoul J. Enhancement of Speech Corrupted by Acoustic Noise[C]// Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP. IEEE, 1979:208-211.
- [9] Ephraim Y. Statistical Model Based Speech Enhancement Systems[J]. Proceedings of the IEEE, 1992, 80(10):1526-1555.
- [10] Na W, Dezhong Z, Shuang X, et al. A New Algorithm for Speech Enhancement Using Wavelet Packet Transform Based on Auditory Model.[C]// International Conference on Computer Science and Software Engineering. IEEE, 2008:1000-1003.
- [11] Ma N, Bonchard M, Goubran R A. Frequency and Time Domain Auditory Masking Threshold Constrained Kalman Filter for Speech Enhancement[C]// International Conference on Signal Processing, 2004. Proceedings. Icsp. IEEE, 2004:2659-2662 vol.3.
- [12] You C H, Koh S N, Rahardja S. Signal Subspace Speech Enhancement for Audible Noise Reduction[C]// IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. IEEE Xplore, 2005:145-148.
- [13] Xu Y, Du J, Dai L R, et al. Global Variance Equalization for Improving Deep Neural Network Based Speech Enhancement[C]// IEEE China Summit & International Conference on Signal and Information Processing. IEEE, 2014:71-75.
- [14] Xu Y, Du J, Dai L R, et al. An Experimental Study on Speech Enhancement Based on Deep Neural Networks[J]. IEEE Signal Processing Letters, 2014, 21(1):65-68.
- [15] Han K, Wang Y, Wang D L, et al. Learning Spectral Mapping for Speech Dereverberation and

- Denoising[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2015, 23(6):982-992.
- [16] McCulloch, Warren S, Pitts, et al. A Logical Calculus of The Ideas Immanent in Nervous Activity[J]. Bulletin of Mathematical Biology, 1990, 52(1):99-115.
- [17] HEBB Donald O. The Organization of Behavior[J]. 1949, 9(3):213-218.
- [18] Widrow B, Hoff M E. Adaptive Switching Circuits[M]// Neurocomputing: foundations of research. MIT Press, 1988.
- [19] Hopfield J J. Neural Networks and Physical Systems with Emergent Collective Computational Abilities[J]. Proceedings of the National Academy of Sciences of the United States of America, 1982, 79(8):2554.
- [20] Kirkpatrick S, Gelatt C D, Vecchi M P. Optimization by Simulated Annealing[J]. Science, 1983, 220(4598):671.
- [21] Hinton G E, Sejnowski T J, Ackley D H. Boltzmann Machines: Constrained Satisfaction Networks that Learn[J]. 1986.
- [22] McClelland J L, Rumelhart D E. Explorations in Parallel Distributed Processing:, A Handbook of Models, Programs, and Exercises.[M]// Explorations in parallel distributed processing: a handbook of models, programs, and exercises. MIT Press, 1988:435.
- [23] Hinton G E, Salakhutdinov R R. Reducing the Dimensionality of Data with Neural Networks.[J]. Science, 2006, 313(5786):504.
- [24] Kent R D, Read C. The Acoustic Analysis of Speech[M]// The acoustic analysis of speech /. Singular/Thomson Learning, 2002:360-380.
- [25] Levitt H, Rabiner L R. Binaural Release from Masking for Speech and Gain in Intelligibility.[J]. Journal of the Acoustical Society of America, 1967, 42(3):601.
- [26] Ramabadran T V, Ashley J P, McLaughlin M J. Background Noise Suppression for Speech Enhancement and Coding[C]// Speech Coding For Telecommunications Proceeding, 1997, 1997 IEEE Workshop on. IEEE Xplore, 1997:43-44.
- [27] Singh L, Sridharan S. Speech Enhancement Using Critical Band Spectral Subtraction[C]// The, International Conference on Spoken Language Processing, Incorporating the, Australian International Speech Science and Technology Conference, Sydney Convention Centre, Sydney, Australia, November -, December. DBLP, 1998.
- [28] Lim J, Oppenheim A. All-pole Modeling of Degraded Speech[J]. IEEE Transactions on Acoustics Speech & Signal Processing, 1978, 26(3):197-210.
- [29] Breithaupt C, Martin R. MMSE Estimation of Magnitude-squared DFT Coefficients with SuperGaussian priors[C]// IEEE International Conference on Acoustics, Speech, and Signal

- Processing, 2003. Proceedings. IEEE, 2003:I-896-I-899 vol.1.
- [30] Ephraim Y, Trees H L V. A Signal Subspace Approach for Speech Enhancement[J]. IEEE Transactions on Speech & Audio Processing, 1993, 3(4):251-266.
- [31] Hagan M T, Demuth H B, Beale M. Neural Network Design[M]. China Machine Press, 2002.
- [32] Huang W, Hong H, Song G, et al. Deep Process Neural Network for Temporal Deep Learning[C]// International Joint Conference on Neural Networks. 2014:465-472.
- [33] Larochelle H, Bengio Y, Louradour J, et al. Exploring Strategies for Training Deep Neural Networks.[J]. Journal of Machine Learning Research, 2009, 1(10):1-40.
- [34] Najafabadi M M, Villanustre F, Khoshgoftaar T M, et al. Deep Learning Applications and Challenges in Big Data Analytics[J]. Journal of Big Data, 2015, 2(1):1.
- [35] Bengio Y, Guyon G, Dror V, et al. Deep Learning of Representations for Unsupervised and Transfer Learning[J]. Workshop on Unsupervised & Transfer Learning, 2011, 7.
- [36] Hinton G E. Learning Multiple Layers of Representation.[J]. Trends in Cognitive Sciences, 2007, 11(10):428.
- [37] Bengio Y, Lamblin P, Popovici D, et al. Greedy Layer-wise Training of Deep Networks[C]// International Conference on Neural Information Processing Systems. MIT Press, 2006:153-160.
- [38] Schmidhuber J. Deep Learning in Neural Networks: An overview.[J]. Neural Netw, 2015, 61:85-117.
- [39] Liu H, Taniguchi T, Takano T, et al. Visualization of Driving Behavior Using Deep Sparse Autoencoder[C]// Intelligent Vehicles Symposium Proceedings. IEEE, 2014:1427-1434.
- [40] Sawada Y, Kozuka K. Transfer Learning Method Using Multi-prediction Deep Boltzmann Machines for A Small Scale Dataset[C]// Iapr International Conference on Machine Vision Applications. IEEE, 2015.
- [41] Akinin M V, Akinina N V, Taganov A I, et al. Autoencoder: Approach to the Reduction of the Dimension of the Vector Space with Controlled Loss of Information[C]// Embedded Computing. IEEE, 2015:171-173.
- [42] Lander S, Shang Y. EvoAE -- A New Evolutionary Method for Training Autoencoders for Deep Learning Networks[C]// IEEE, Computer Software and Applications Conference. IEEE Computer Society, 2014:790-795.
- [43] Deng J, Zhang Z, Marchi E, et al. Sparse Autoencoder-Based Feature Transfer Learning for Speech Emotion Recognition[C]// Affective Computing and Intelligent Interaction. IEEE, 2013:511-516.
- [44] Roux N L, Bengio Y. Representational Power of Restricted Boltzmann Machines and Deep Belief Networks.[J]. Neural Computation, 2014, 20(6):1631-1649.
- [45] Rothauser E H, Chapman W D, Guttman N, et al. IEEE Recommended Practice for Speech Quality

- Measurements[C]// IEEE No. IEEE, 2016:1-24.
- [46] Voiers W. Diagnostic Acceptability Measure for Speech Communication Systems[C]// Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP. IEEE Xplore, 1977:204-207.
- [47] Rix A W, Hollier M P, Hekstra A P, et al. Perceptual Evaluation of Speech Quality (PESQ) The New ITU Standard for End-to-End Speech Quality Assessment Part I--Time-Delay Compensation[J]. Journal of the Audio Engineering Society, 2002, 50(10):755-764.
- [48] Rothausen E H, Chapman W D, Guttman N, et al. 297-1969 - IEEE Recommended Practice for Speech Quality Measurements[J]. 1969, 17:225-246.

致谢

在论文截稿之际在此我想感谢在整个论文撰写过程中给予过我帮助的老师 and 同学们。首先我要感谢我的导师马鸿飞教授，在研究生生涯中给予了我很大的指导和帮助。在我的论文撰写过程中，无论是论文开题，还是难点突破都是马老师的谆谆教导让我得以顺利进行。马老师牺牲了自己宝贵的时间和精力才使得本文写作得以顺利完成。还有马老师专业的知识储备和丰富的科研经验使得我在整个论文的撰写过程中少走了很多弯路。在此感谢马老师在我整个研究生生涯的帮助和指导。

感谢实验室的师兄师姐，在刚刚步入研究生生活时，是师兄师姐们的热心帮助让我快速适应了新的环境。感谢实验室的同窗们，他们的陪伴给了我很大的动力，并且给予了我很多无私的帮助和生活上的关怀。通过与他们的讨论使得自己在科研的道路上灵光闪现，使得难题得到解决，同时他们也为我的论文撰写提供了宝贵的意见。非常感谢他们在过去三年时间里的陪伴与帮助。

总之，在此论文截稿之际感谢整个过程中给予过我帮助和关心的老师和同学们！

作者简介

1. 基本情况

刘浩，男，河南济源人，1990年8月出生，西安电子科技大学通信工程学院通信与信息系统专业2014级研究生。

2. 教育背景

2009.09～2013.06 华北水利水电大学，本科，专业：电子信息工程

2014.09～ 西安电子科技大学，硕士研究生，专业：通信与信息系统

3. 攻读硕士学位期间的研究成果

[1] 马鸿飞，赵月娇，刘珂，刘浩等 一种采用栈自动编码机的语音分类算法[J]. 西安电子科技大学学报，2017,(05):13-18.



西安电子科技大学
XIDIAN UNIVERSITY

地址：西安市太白南路2号

邮编：710071

网址：www.xidian.edu.cn