中国科学院大学

University of Chinese Academy of Sciences

硕士学位论文

基于深度循环神经网络的 LDV 测声系统语音降噪

作者姓名：白涛

指导教师：吴谨　研究员

中国科学院电子学研究所

学科类别：工学硕士

学科专业：物理电子学

培养单位：中国科学院电子学研究所

2018 年 6 月

Deep Recurrent Neural Network Based Speech Denoising in LDV Remote Voice Acquisition System

A Thesis Submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Master of Science in Engineering

in Physical Electronics

By

Bai Tao

Supervisor: Professor Wu Jin

Institute of Electronics, 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：白涛

日　　期：2018.6.1

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：白涛　　导师签名：

日　　期：2018.6.1　　日　　期：2018.6.1

摘 要

激光多普勒测振仪(Laser Doppler Vibrometer，LDV)是利用激光多普勒效应测量物体振动的一种仪器，具有测量精度高、非接触和能够远距离探测的优点，所以被广泛的应用在工业生产和精密测量等领域。LDV 测声系统是利用 LDV 检测由声音经空气耦合引起物体的振动，从而还原声音信息的设备，俗称“激光窃听器”。激光测声设备具有非接触、隐蔽性好、使用方便等特点，在反恐、军事侦察、安全监听等领域得到重要应用。

然而，由于环境噪声、振动目标运动、大气湍流、相干散斑效应、电路噪声等多种因素的影响，高灵敏的 LDV 测声系统输出语音噪声大，可懂度低。目标距离越远，这种现象更加严重。如何降低噪声，提高输出语音的质量，成为 LDV 测声系统需要解决的关键问题。

本文应用神经网络语音降噪技术，对 LDV 测声系统的语音降噪问题开展了研究。

首先利用 1550nm 激光建立了基于同差探测的 LDV 测声系统的实验装置，基于 LabVIEW 软件编制了 LDV 测声系统控制程序，用小波分解等手段，实现了数据实时采集和分析。

其次，采用神经网络语音降噪的方法，基于深度循环神经网络，编写了具有多层网络结构的语音降噪程序。采用高斯加性噪声模型构造含噪语音模拟数据集，完成了深度循环神经网络的训练和优化。模拟结果表明，所构建的深度循环神经网络语音降噪程序，在信噪比-6dB~6dB 范围内，实现有效的语音降噪。

最后，开展了基于深度循环神经网络语音降噪的 LDV 测声系统实验。通过提取 LDV 测声系统的语音信号特征和噪声信号特征，送入构建好的神经网络中进行训练，根据多种语音信号质量评价指标，对处理后的语音信号的质量的量化评价，形成了用于 LDV 测声系统语音降噪的最优网络结构。利用扬声器播放语音信号，在 10m 和 75m 两个距离上，实验了深度循环神经网络语音降噪的性能。实验数据表明，深度循环神经网络语音降噪方法，与传统的语音降噪方法在 SDR、SIR、SAR 三项指标上相比，有 5dB 到 15dB 的提升，在 PESQ 指标上有 0.2dB 到 0.5dB 的提升，在 STOI 指标上有 0.05 到 0.1 的提升。在使用短时傅里叶变换

提取语音特征，使用的网络层数为3，目标函数为二阶方差，激活函数为RELU时，所构建的网络的性能最好。

结果表明，深度循环神经网络可以有效的去除从激光多普勒测声系统采集回来的语音信号中的噪声，并且与传统的语音降噪方式相比，处理后得到的语音信号的质量有明显的提升。

关键词：激光技术；激光多普勒测振仪；语音降噪；深度循环神经网络

Abstract

Laser Doppler vibrometer is an instrument for measuring object vibration based on laser Doppler effect. It has the advantages of high accuracy, non-contact and long distance detection. So it is widely applied in the fields of industrial production and precision measurement. LDV remote voice acquisition system is a device for reducing sound information, LDV detection sound through the air coupling caused by the vibration of the objection, commonly known as "laser eavesdropper". LDV remote voice acquisition system has the characteristics of non-contact, good concealment and convenient c[illegible]ation, has important applications in military reconnaissance, anti-terrorism, security monitoring and other fields.

However, due to environmental noise, vibration target motion, atmospheric turbulence, coherent speckle effect, circuit noise and other factors, the highly sensitive LDV system has large noise and low intelligibility. The farther the target is, the more serious this phenomenon is.

In this paper, we use the application of deep recurrent neural network noise reduction technology, the speech de-noising problem of LDV remote voice acquisition system is studied.

First, by using 1550nm laser, based on the homodyne detection LDV remote voice acquisition system is built. The control program of LDV remote voice acquisition system is compiled based on LabVIEW software, wavelet analysis and other means are used to achieve real-time data acquisition and analysis.

Then, by using the neural network method of speech denoising, based on deep recurrent neural network, a speech denoising program with multi-layer network structure is written. Using Gauss additive noise model to construct noisy speech simulation dataset, complete the training and optimization of deep recurrent neural network. Simulation results show that the speech noise reduction program based on deep recurrent neural network achieve effective noise reduction of speech in -6dB~6dB SNR.

Finally, an experiment of LDV remote voice acquisition system based on deep recurrent neural network speech noise reduction is carried out. The characteristics of speech signal and noise signal of LDV remote voice acquisition system are extracted, then fed into a constructed deep recurrent neural network for training. According to

the quality evaluation index of multiple speech signals, the quality of processed speech signals is quantitatively evaluated, the optimal network structure for speech noise reduction in LDV remote voice acquisition system has been formed. Using loudspeakers to play voice signals, on the two distances between 10m and 75m, experiments on the performance of deep recurrent neural network speech noise reduction. Experimental data show that, the deep recurrent neural network speech denoising method is compared with the traditional speech denoising method, compared to the three indexes of SDR, SIR and SAR, have promotion of 5dB to 15dB. In the index of PESQ, have promotion of 0.2dB to 0.5dB. In the index of STOI, have promotion of 0.05 to 0.1.

The results show that deep recurrent neural network can effectively remove the noise from the voice signals collected from the LDV remote voice acquisition, and compared with the traditional speech signal denoising method, the quality of speech signal after processing is obviously improved. This is of practical significance for improving the performance of LDV remote voice acquisition system.

Key Words: Laser Technology; Laser Doppler Vibrometer; Speech Signal Denoising; Deep Recurrent Neural Network

目 录

图目录

第 1 章 绪论

1.1 课题的背景和意义

激光多普勒测振仪(Laser Doppler vibrometer，LDV)是利用光学多普勒效应检测物体振动的设备，它可以实现对振动目标非接触、高灵敏度和作用距离远的测量[1]。由于声音通过空气传播，通过空气耦合，信号源附近物体，会随之振动，所以，理论上，利用 LDV 探测说话人附近物体的振动情况，可以还原说话人的声音。相应的设备便是 LDV 测声系统，俗称“激光窃听器”。

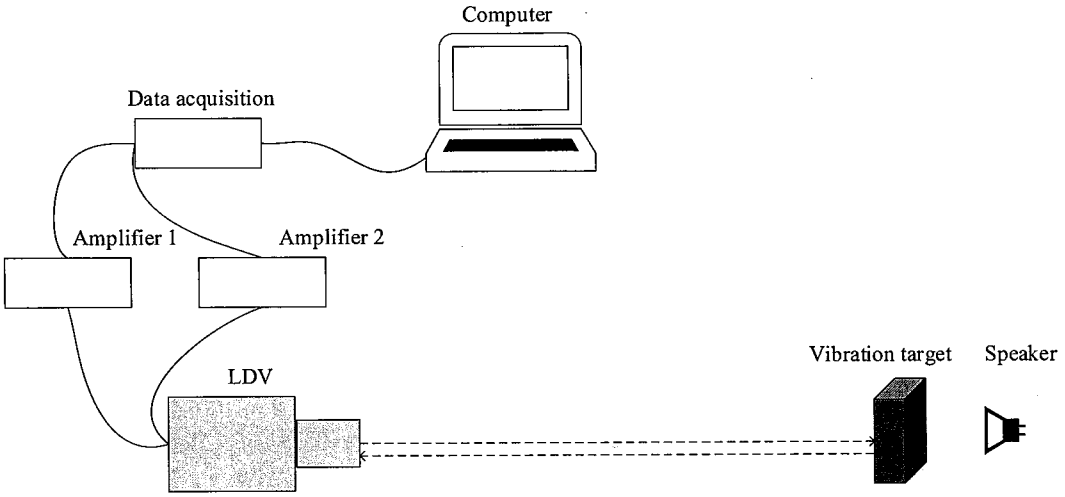


图 1.1 LDV 测声系统

Figure1.1 LDV voice acquisition system

LDV 测声系统获取的声音过程如图 1.1 所示。目标附近的声音经过空气耦合，引起目标的振动。LDV 发射激光照射目标，同时接收来自目标的散射回波。通过光学外差探测，将散射回波光信号转换为电信号。再经过放大和数字化，由电脑处理获得目标振动所包含的声音信息。

LDV 测声系统具有如下特点：

（1） 无需前置，安全、方便；

不需要在窃听现场先期设置任何设备，对窃听目标及其所处环境没有任何干扰和危害，被窃听目标也无法事先探测。

（2） 可以采用人眼安全不可见的红外激光，隐蔽性好；

探测激光可采用人眼安全的红外激光，工作时，人眼看不见，被窃听目标没

有任何其他感知。

（3）作用距离远；

有效作用距离可以达到100m以上，对合作目标可达300m，隐蔽性、安全性好[2]。

（4）可以透过障碍物

可以穿透部分玻璃、纱窗等障碍物，可探测车内、室内语音。探测光束很小，很小透光面积就可以满足遥感探测的需要。

（5） 语音恢复性好，可懂度高；

许多材料都有良好的声频振动响应，获取的语音信息可懂度高。

（6）有较大的激光入射角适应范围；

遥感能力不依赖于特定激光入射角度，有较大的激光入射角适应范围。

（7）功耗低，电源适应广；

设备功耗不大，适用多种电源（市电、电池、汽车电源）。

（8）布设、操作简单。

设备布设、操作简单，易学易用，开机即用。

事实上，LDV测声系统已在反恐、安全等领域得到了成功应用。

据美国探索频道有关节目报道，当年为了捕获本拉登，在其巴基斯坦住所远处，中央情报局（FBI）通过激光窃听器秘密监听了几个月，获得了本拉登的声频信息，从而确认了本拉登在寓所的时间。最终成功击毙本拉登，使基地组织的走向灭亡。

然而，LDV测声系统具备高灵敏等一系列优点的同时，也具有一大缺点，即所获得的语音信号是单声道高噪声的。目标距离越远，噪声越严重，所获的语音信息质量差，可懂度低。

LDV测声系统的主要噪音来源如下：

（1） 环境噪声

一般，被测目标不是处在声音暗室，其周遭会有复杂的噪声存在，如马路上汽车奔驰声、电脑的风扇声、周围工地的打桩声等等。所有任何声音都会通过空气耦合至目标，参与目标及测声设备的振动。另外，如果目标或测声设备位于建筑物内，建筑物本身的振动也会耦合进来。因此，仅目标本身的振动，就是非常复杂的，感兴趣的语音信息，仅是目标振动信息的极小部分[3]。

（2） 目标运动引入的噪声

被检测目标对象一般也不是静止的，会存在不同程度的无规则运动。目标的运动一方面引起回波强度的变化，同时也引入额外的多普勒频率，形成噪声。

（3） 大气湍流噪声

大气湍流引起激光传输路径上的大气折射率变化，引起激光强度的衰减、波阵面的破坏、光强分布的变化等，破坏了激光的相干性，也引入了额外的相位噪声。目标距离越远、大气湍流越大，在 LDV 测声系统中引入的噪声越大，获得的语音质量越差。

（4） 相干散斑噪声

相对于激光的短波长，几乎所有的被测表面都可以看作是粗糙的。对粗糙表进行相干探测时，散斑效应是不可避免的现象。当目标距离远，目标上的光斑尺寸大时，散斑效应更加明显。散斑效应服从高斯随机分布，也是相干测量的一种噪声来源[4]。

（5） 电路噪声

LDV 测声系统通过光学外差探测获得目标的振动信息。光学外差信号一般都需要经过滤波、放大和数字化等电子学处理。这些电路处理也会引入噪声，影响 LDV 测声系统最终输出效果。

总之，LDV 测声系统应用于复杂的自然环境，其应用条件的特殊性（远距离、隐蔽性、不可控制的周遭环境、不合作的目标对象等）使这种设备获取的信息一般都是噪声严重的，需要依赖强大的后期处理（主要是语音降噪与增强）工作。

另一方面，具有优异语音降噪与增强性能的技术，可以使 LDV 测声系统具有更加强健的性能，如更远的作用距离、更好的环境适应性等，从而使这种设备更能为用户创造价值，应用更加广泛。

目前，LDV 测声系统常采用单一激光束遥感目标信息，其降噪问题，适用于单通道语音信号降噪算法。目前，单通道语音的降噪处理已有大量的研究成果，但是，将这些方法应用于 LDV 测声系统的语音信号处理，却并不十分有效。LDV 测声系统的语音降噪问题，值得进一步探索。

LDV 测声系统的单通道语音信号降噪算法，应当是一个能够随着探测环境的改变，自适应调整降噪策略的、具有‘学习’当前 LDV 测声系统噪声模式和语

音信号结构的算法。为此，本文尝试在LDV测声系统中应用基于神经网络的语音降噪算法。

与传统的降噪手段相比较，基于神经网络的语音降噪最大的特点是其强大的特征提取能力。利用神经网络强大的非线性拟合能力，通过大数据的训练，能够不需要对系统有过多的研究分析，即可用神经网络来表征系统内部结构。

1.2 LDV测声系统研究

利用LDV感知声音的工作已经有很长时间的研究历史，国内外均有成型的设备应用于实务工作中。但是，随着激光光源、探测器技术、电子技术、数字信号处理技术等的发展和进步，LDV 测声技术一直在不断发展。下面罗列了进入新世纪以来，国内外科技文献中关于LDV测声技术的一些研究报道。

2000年，德国Peter Lutzmann 等人使用10.6um和1.54um的激光远程测量物体的振动特性，由于多普勒分辨率足够大，所以可以对隐藏和伪装的目标进行识别[5]。

2003年，罗海俊利用5mW功率 790nm-820nm波长的激光器探测探测由声音信号引起的附近的玻璃的振动，来还原声音信号。同时研究了发射装置和反射面之间的入射角度对接收到的信号的影响[6]。

2005年，美国朱志刚等人利用红外摄像头和POLYTEC公司的LDV相结合的方式，构建了一种多模态监视系统。该系统为了实现LDV系统的自动对焦和选择合适的振动目标，将红外线（IR），云台变焦彩色相机（PTZ）和LDV相结合。并利用简单的高斯滤波器和语音自适应增益算法对语音信号做降噪处理。该系统实现了没有反光条带下100米的语音信号测量，和在有反光带条件下300米的语音信号的测量[7]。

2007年叶嘉熊等利用LDV探测从玻璃目标返回的振动信号，并且对探测器和反射光之间的夹角对声音探测的影响做了深刻的研究。实验结果表明，如果振动目标为音室玻璃，探测距离在 6-63m 内，夹角在±4° 内，可以接收到带有较大噪声的声音；同时指出，通过降低系统内部的电流噪声和采用具有较好性能的光学滤波片是提高性能的有效手段[8]。

2009 年尚建华等人基于外差探测的相关理论，构建了全光纤外差技术的激光多普勒测振系统，并利用该系统透过玻璃探测振动信息，结果表明该系统能够

透过玻璃获取 25 米外的振动信息，并且讨论了 LDV 到振动目标的距离和玻璃的材质，对探测信号的信噪比的影响[9]。

2009 年，Weihong Li 等人研究了对 LDV 测声系统的噪声形成的原因进行了研究，并表明检测到的语音信号被多种的噪声源所污染，如激光的散斑噪声，目标的运动，以及背景噪声等。并利用高斯带通语音增强算法和维纳滤波器对获取到的语音信号做降噪处理。并指出选择合适的振动目标和提出相应的语音降噪手段是提升 LDV 测声系统性能的必要条件[10]。

2010 年，刘彬彬等人提出了偏振激光干涉仪的校正方法。用来对非线性误差信号进行实时的校正。该方法利用的原理是椭圆匹配，对由参数误差、位置误差和外界因素产生的非线性误差校正，信号的信噪比在经过非线性误差校正之后可以有 30dB 的提高，测量分辨率优于 $10\,pm/Hz^{1/2}$ [11]。

2011 年，Li Rui 等人研究了振动目标结构特性以及表面材料与其振动特性之间的关系，并建立了振动振幅模型。然后通过仿真和实际的传感器实验，研究了不同材料和结构的典型表面的振动特性。并根据人类声音频率范围内的频率响应，将目标分为三类。并对这三类进行 LDV 测声实验。实验结果显示玻璃板和纸盒有更好的性能，并且玻璃板对高频分量有更好的响应，意味着语音信号有更高的清晰度。此外，这些目标的频率响应曲线，不仅可以帮助选择更好的振动目标，也可以利用这些目标振动曲线进行 LDV 测声系统语音信号增强算法的研究[12]。

2012 年，李斐斐等人研究了利用外差探测方式来搭建全光纤 LDV，并对通过空气耦合的振动目标的振动特性进行测量和分析。实验探究了 6 种振动目标在 200-2700Hz 的振动响应曲线。实验结果表明对振动目标的振动响应影响最大的是目标材料，其他的因素只会对局部的细节有影响[1]。

2014 年，梁娜等人研究了基于零差探测技术的 LDV，讨论了激光多普勒信号的解调方法，并且采用光学偏振元件，设计了光学六端口混频器，用于校正相位不平衡。最后用设计好的六端口混频器构建了基于零差探测的 LDV 系统，并给出相应的测振实例，将语音信号实现了成功的还原[13]。

1.3 单通道语音降噪研究

语音降噪旨在通过各种算法来降低语音信号中的各种噪声，从而尽可能的恢

复原始的语音信号。有关语音降噪的研究已近有悠久的历史，对于单通道的语音降噪，从降噪手段上分为无监督和有监督两种。

无监督的语音降噪算法的典型代表有谱减法和维纳滤波法等。

谱减法的原理简单，但是，应用谱减法时，如果对噪声的方差估计过高，就会引起语音信号的失真，而如果对噪声方差欠估计，就会产生音乐噪声，同时能够使用的信噪比的范围较窄[14]。

维纳滤波法在降噪的过程中虽然不会引入音乐噪声，但是维纳滤波法对噪声信号的要求较为苛刻，要求噪声信号是平稳信号，同时与语音信号不相关，所以维纳滤波法对非平稳的噪声信号处理能力非常有限[15]。

有监督的语音降噪算法主要有基于非负矩阵分解的方法、基于隐马尔可夫模型的方法和神经网络的方法。

基于非负矩阵分解(Nonnegative Matrix Factorization, NMF)的方法，将干净的语音信号和噪声信号的平方谱或者幅度谱，近似的分为两个非负的，系数矩阵和基矩阵的乘积。因此，非负矩阵分解可以从语音信号和噪声的混合信号中有效的分辨出每个信号的频谱，从而将语音信号和噪声分离[16]。

对于隐马尔可夫模型，它对非平稳噪声有较好的抑制能力，它利用最大似然准则，通过状态的高斯模型，拟合语音信号和噪声信号，对于每一句带噪的语音信号，通过遍历语音和信号的状态池来寻找最佳组合，以此来对语音信号进行降噪处理[17]。

上述的几种方式为传统的单通道语音去噪方式。传统的单声道语音信号去噪算法要求噪声比较平稳，以便在非语言段对噪声进行估计，再依据估计出来的噪声对带噪语音段进行处理。同时，算法模型的构建往往基于高斯假设，以及对语音信号和噪声之间的关系做了独立性假设。但在实际情况中，噪声具有随机性和突变性，使得对噪声的跟踪和估计变得困难，而且传统的语音增强方法易引入非线性失真。

第三种有监督的语音信号处理算法基于神经网络，是近年来有监督的语音降噪算法中发展最为迅速，而且效果最为显著的一种方法。

近年来，神经网络（Neural Network, NN）在语音信号处理中有着很多成功的应用，并且适应性好，限制条件少。基于大数据的训练，神经网络可以充分学习噪音和干净语音之间的复杂的非线性关系，它能记住一些噪声模式，因而可以

很好的抑制一些非平稳噪声。

将神经网络应用于语音信号降噪领域的研究很早[19]。开始的研究工作使用一个单层隐含层，只有 160 个神经元的浅层神经网络（SNNs），由于该网络的规模太小，不能充分学习噪声特征和语音信号之间的关系。在一系列的因素的限制下，利用神经网络对语音信号做降噪处理并没有取得交到的处理效果。

神经网络的重大的突破始于 Hinton 提出的一种贪婪的无监督的网络优化处理算法[20]。

在该因素的促使下，在语音信号处理领域，神经网络在最初成功应用于语音识别[21]，并且处理的效果与传统的基于统计的语音识别方式有很大的提升。

在此之后，利用神经网络的方式对语音降噪，和与传统的方式相比，有了突破性的进展，2012 年，Andrew L. Mass 等人，提出了用于语音识别的基于循环神经网络的语音信号降噪算法[22]，该文章将深度循环自动编码器应用于语音信号降噪以提高语音识别率，这个网络模型没有对噪声信号的模型做任何的假设，相反，只要给予足够的数据，可以通过网络模型去自动的提取噪声信号的特征。

2013 年，Philemon Brakel 等人提出了基于双向截断神经网络网络结构用于语音降噪[23]。Xu-Gang LU 提出了利用深度降噪自编码器来增强语音信号的质量[24]。

2015 年，徐勇等人提出了一种基于深度神经网络的语音降噪算法[25]，并利用一系列的方式去提高网络的范化性能。该算法和传统的 MMSE 算法相比，取得了显著的改进，并且在处理过程中不会引入音乐噪声。

2015 年，在 Andrew L. Mass 等人的研究基础上，Po-Sen Huang 文章提出了掩蔽和深度循环神经网络联合优化的神经网络结构，用于单通道的信号分离和语音信号的去噪，该算法的语音降噪效果不止优于传统的非负矩阵分解算法，同时也要优于深度神经网络结构[26]。

2017 年，在 Po-Sen Huang 研究的基础上，Zhe-Cheng Fan 提出了使用 VPNN 网络的多声源信号分离网络[27]。作者首先将输入的信号映射到三维空间中，然后送入三维向量积神经网络（输入、输出、权重），最后将网络的输出映射到实数。该网络结构很好的解决了时间-频率单元信息利用不足的问题。

纵观语音信号处理最近的发展，可以看出利用神经网络对语音信号处理已逐渐成为主流的方式，而深度循环神经网络又是其中降噪效果最为显著的网络结

构。处理的效果与传统的方式相比也有质的飞跃。所以，针对LDV测声系统的系统特性，利用深度循环神经网络来对从LDV所获取的语音信号进行处理，势必会给LDV测声系统性能的提升带来更显著的效果。

1.4 本文主要内容

在神经网络成熟应用之前，大家对系统的研究分析，都是从分析物理结构、构建数据模型、实验分析的流程来对一个复杂的系统进行研究。然而，通过神经网络，可以利用系统的输入和输出数据，让计算机自己去构建系统的模型，极大的简化了系统拟合的流程。本文尝试利用神经网络进行LDV测声系统的语音降噪，从而提升其测声性能。

本文的结构安排如下：

第一章，介绍了课题的研究背景和研究现状。

第二章，建立了基于零差探测的LDV测声系统，编制了基于LabVIEW的信号采集与分析软件，实现了对振动信息的获取。

第三章，介绍了语音降噪的基本原理和神经网络用于语音降噪的原理，包括多种不同参数结构的深度循环神经网络。

第四章，介绍了实验设备、实验流程以及实验数据的采集，构建用于训练神经网络的数据集，并利用该数据集和已构建好的网络模型，对从LDV测声系统所获取的语音信号做降噪处理。对不同参数结构的网络所获取的实验结果进行分析，用语音信号评价指标量化的评价语音信号的质量，来帮助寻找最优的用于LDV测声系统降噪的网络结构。

第五章，简要总结。

第 2 章　建立 LDV 测声系统实验装置

2.1　LDV 测声系统基本原理

在这里假设入射光波的振幅和参考光的振幅分别是：

$$U_s(t)=a_s\sin(\omega_s t+\varphi_s) \qquad \ldots(2.1)$$

$$U_o(t)=a_o\sin(\omega_o t+\varphi_o) \qquad \ldots(2.2)$$

其中，$\omega_s=2\pi v_s$ 和 $\omega_o=2\pi v_o$ 分别是入射光和参考光的角频率，a_s 和 a_o 分别是入射光和参考光的振幅。将混合光束输入到混频器上后，输出的光强为：

$$I_{hs}=K\left|U_s(t)+U_o(t)\right|^2=K\left\{U_s^2(t)+U_o^2(t)+2U_oU_s(t)\right\}$$

$$=\frac{K}{2}\left\{a_s^2+a_o^2-a_s^2\cos(2\omega_s t+2\varphi_s)-a_o^2\cos(2\omega_o t+2\varphi_o)\right\}$$

$$-2a_oa_s\cos\left[(\omega_s+\omega_o)t+(\varphi_s+\varphi_o)\right]+2a_oa_s\cos\left[(\omega_s-\omega_o)t+(\varphi_s-\varphi_o)\right] \qquad \ldots(2.3)$$

其中，K 表示光电探测器的光电灵敏度。从式（2.3）中可以看出，信号在经过混频后包含直流分量、二倍信号光、二倍参考光和信号光和参考光的差频和和频。通过低通滤波后，输出为频率为 $\Delta v=v_s-v_o$ 的差频信号：

$$I_{hs}=Ka_sa_o\cos(2\pi\Delta vt+\Delta\varphi),\quad \Delta\varphi=\varphi_s-\varphi_o \qquad \ldots(2.4)$$

而当本振光和信号光的频率相同，上式变为

$$I_{hs}=Ka_sa_o\cos(\Delta\varphi) \qquad \ldots(2.5)$$

式（2.5）为零差探测的表达式，也是构建基于零差探测的 LDV 测声系统的主要方式。

假设入射光的相位为 $\omega_s=2\pi f_s(t-\Delta t)$，参考光的相位为 $\omega_o=2\pi f_o(t-t_o)$，其中 Δt 表示振动目标和分光镜之间的传输延迟。

用 $L(t)$ 表示分光棱镜和目标之间的距离，则对于做简单的余弦运动的振动目标，

$$L(t)=R_o+R(t)=R_o+a\cos(\omega t+\phi) \qquad \ldots(2.6)$$

其中 R_o 表示振动目标和探测器之间的距离，$R(t)$ 为振动目标的振动形式，a

为振动目标的振幅，ω为振动目标的振动频率，ϕ为振动目标相位。

那么，

$$\Delta t = \frac{2L(t)}{c} \qquad \ldots (2.7)$$

在零差探测的条件下，

$$\begin{aligned}\Delta\varphi &= \omega_s - \omega_o = 2\pi f_o(\Delta t - t_o) \\ &= 2\pi f_o \frac{2\left(R_o + a\cos(\omega t + \phi)\right)}{c} - 2\pi f_o t_o \end{aligned} \qquad \ldots (2.8)$$

将式（2.8）代入式（2.5），得

$$I_{hs} = Ka_s a_o \cos\left(2\pi f_o \frac{2\left(R_o + a\cos(\omega t + \phi)\right)}{c} - 2\pi f_o t_o\right) \qquad \ldots (2.9)$$

可以看出，所需要的振动信息包含在I_{hs}的相位之中，通过I_{hs}构建两路正交信号，解相位即可获得所测量的振动信息。

2.2 LDV测声系统构建

激光多普勒测振技术分为零差激光多普勒测振技术和外差激光多普勒测振技术。零差激光多普勒测振技术和外差激光多普勒测振技术相比，有其独特的优势，由于它的构建不需要对参考光做移频处理，系统不需要声光移频模块，具有相对简单的系统结构，方便实验实现。

图2.1为零差LDV测声系统的结构框图。包括激光器（Laser），实验中，选用波长为1550nm的单频窄线宽激光器。激光器发出的激光束经过光纤耦合器（OC）分为两束，一束是探测光，一束是本征光。探测光经过环路器（Circulator）后通过光束聚焦镜（BF）聚焦在振动目标表面，声源通过空气使得振动目标表面产生振动，从而使得目标的散射光产生多普勒频移。目标散射回波原路返回，由光束系统聚焦收集，通过环路器作为信号光，然后输入六端口混频器（Six-Port）。本振光直接通过光纤耦合输入六端口混频器。六端口混频器输出四路混频光分别进入平衡外差探测器（Balanced Detector）产生I\Q两路信号。两路信号经过滤波和放大器放大，然后通过采集卡同时采集[13]。

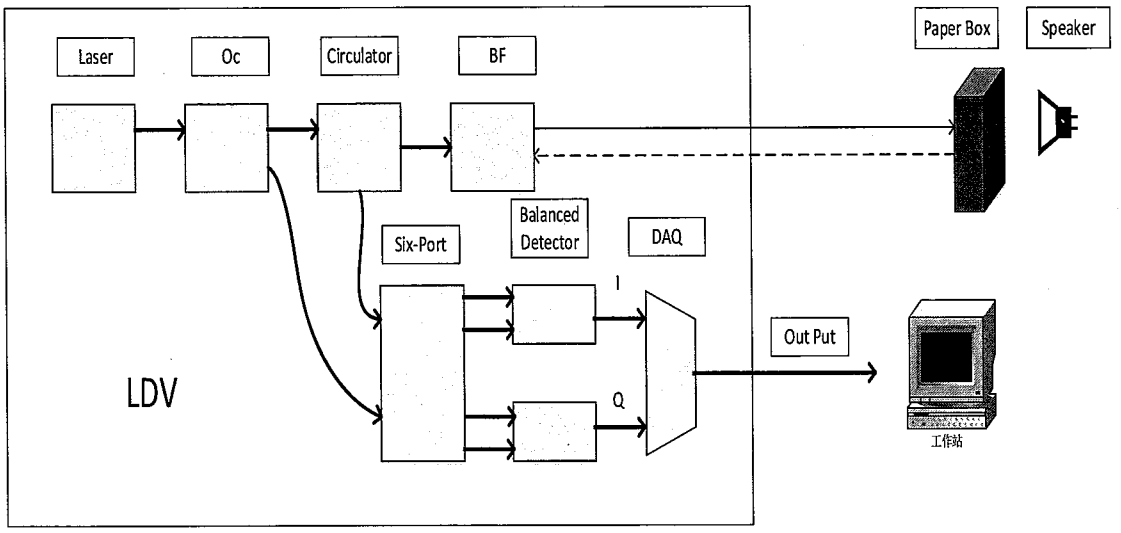


图 2.1 LDV 测声系统

Figure2.1 LDV remote voice acquisition

假设采集到的两路 I\Q 信号分别为 $m_I(t)$ 和 $m_Q(t)$，则可以构建如下复函数：

$$m(t)=m_I(t)+j\cdot m_Q(t) \qquad \ldots (2.10)$$

振动信息 $s_v(t)$ 包含在 $m(t)$ 的相位中，通过对相位解缠绕处理，就能获得了随时间变化的振动函数，即

$$s_v(t)=\text{unwrap}\left(\text{angle}\left[m(t)\right]\right) \qquad \ldots (2.11)$$

继续对振动信息 $s_v(t)$ 进行滤波等处理，就获得了语音信号。不过，这样获得的语音信号一般噪声大、语音弱，需要采取专门的语音降噪与增强措施，才能使其中的语音信息被人听懂和理解，成为有价值的信息。

2.3　基于 LabVIEW 的数据采集与分析处理

LabVIEW(laboratory virtual instrument engineering workbench)是一种由美国国家仪器（NI）研发的程序开发环境，用图形化的编辑语音 G 来编写程序，生成框图形式的程序。通过改变软件，能够实现不同仪表仪器的功能，已经在科学实验和工业领域得到广泛应用。

一个 LabVIEW 程序由两部分组成：

（1）前面板，包含了交互式的程序接口，类似于真实的仪器的面板，上面包含旋钮、开关、图标等，用户通过鼠标和键盘可以控制仪器的输出和结果的显

示，数据的存储等。

（2）数据流程框图，数据流程框图是程度的代码部分，是解决问题的编程化方法。这里使用的是图形化的编程语言。

所构建的 LabVIEW 程序的前面板如图 2.2 所示：

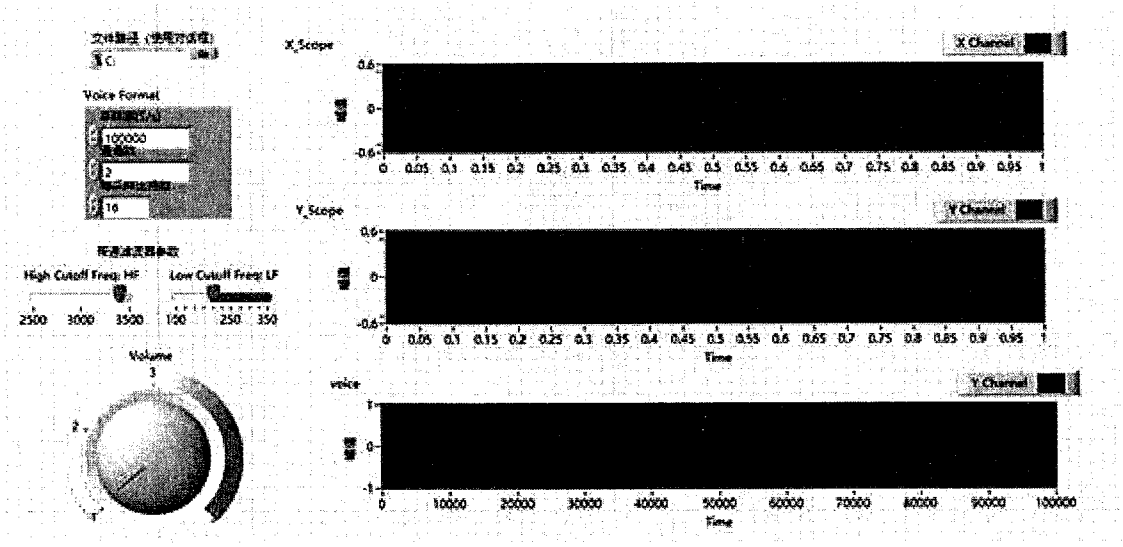


图 2.2 LabVIEW 数据采集分析系统前面板

Figure2.2 LabVIEW front panel of data acquisition and analysis system

该前面板包含采集过程中的参数设置部分：带通滤波器参数设置、采样率设置、语音信号强度设置、文件保存设置。以及显示部分：I\Q 两路正交信号的实时显示、从 LDV 测声系统所获取的语音信号的波形的实时显示。除此之外，也可以将语音信号的频谱图实时显示。

所构建的 LabVIEW 程序框图如图 2.3 所示。设置采集卡的采样率为 100 kHz，对 I\Q 两路正交信号采集后，将这两路数据实时的显示在前面板上，同时将数据保存，以便于后续的研究处理工作。在这之后，解调信号中的相位来还原振动信息，接着用带通滤波器滤波，然后就可以播放和存储所获取的语音信号。

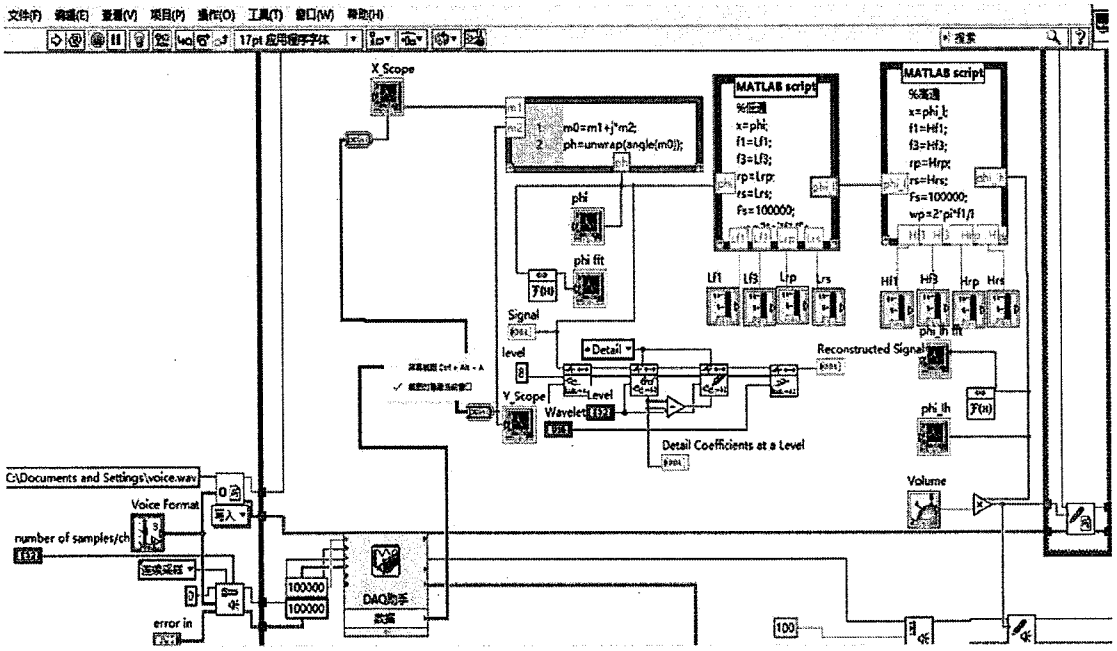


图 2.3 LabVIEW 数据采集与分析系统程序框图

Figure2.3 LabVIEW block diagram of data acquisition and analysis system

另外，还构建了 LabVIEW 小波分解模块，用于实时的显示从 LDV 测声系统采集回来的系统噪声或语音信号，通过观察各种小波基分解后的波形，有利于了解 LDV 测声系统的运转特性。小波分解模块的 LabVIEW 前面板如图 2.4 所示：

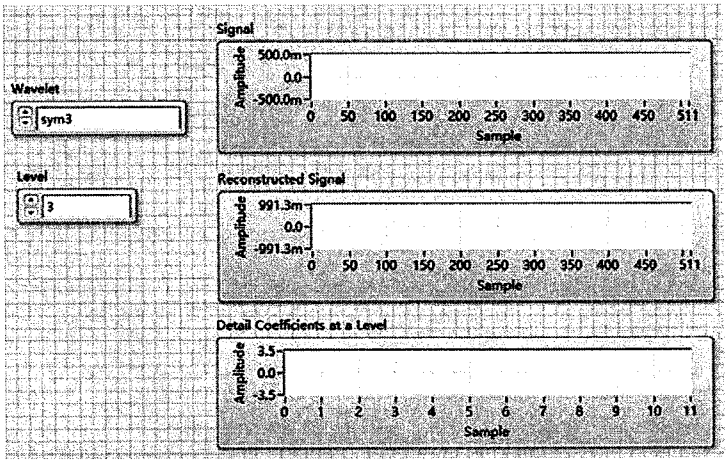


图 2.4 LabVIEW 小波分解系统前面板

Figure2.4 Front panel of wavelet decomposition system

进一步，还引入了 LabVIEW 小波语音信号降噪模块，可以选择多种小波基，多种小波分解层数，以及降噪策略，如“软阈值降噪”、“硬阈值降噪”等，对从 LDV 测声系统采集回来的语音信号进行实时的降噪处理。该模块的结构如图 2.5 所示：

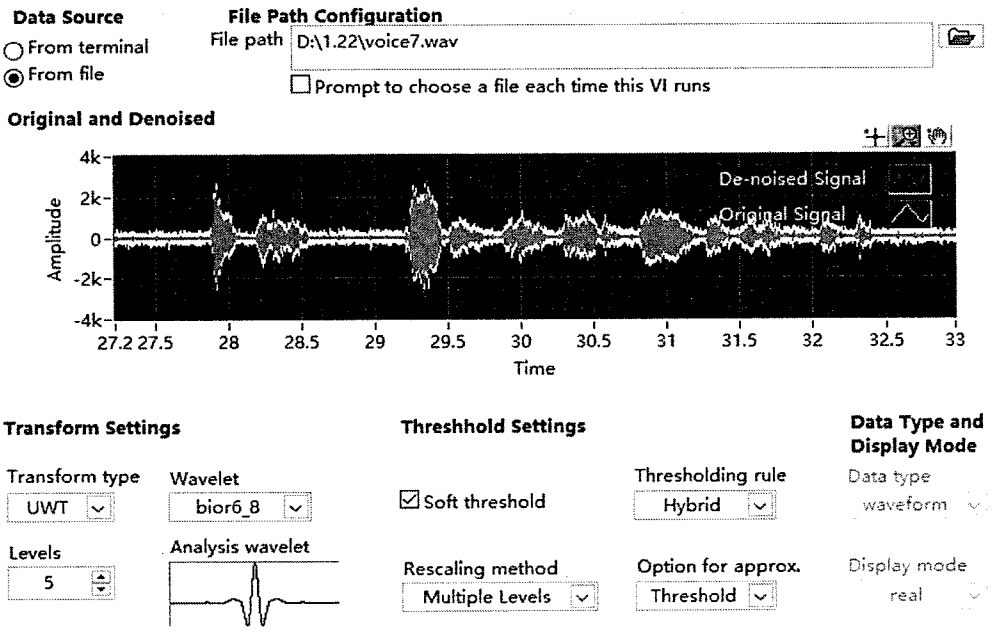


图 2.5 基于 LabVIEW 的小波语音降噪前面板

Figure2.5 Front panel of wavelet speech signal denoising based on LabVIEW

图 2.5 中白色的曲线为去噪前的结果，红色的曲线为去噪后的结果。从图中可以看出，选择合适的小波基、合适的小波分解层数以及阈值选择策略，可以去除一部分 LDV 测声系统的噪声。但是，实验发现，该方式在去除的同时，对语音信号的质量也有一定程度的损害。而这种对带噪语音信号中的信息造成损害的处理方式，是利用 LDV 测声系统探测语音信号的过程中，需要克服的。所以，寻求能够在去除 LDV 测声系统噪声的同时，对原有的语音信号的信息产生较小影响的处理方式，才适合目前的场景所需的。

2.4 LDV 测声系统语音信号特性

为了对从 LDV 测声系统所采集回来的语音信号的特性有全面的细致的了解，使用多种语音信号分析手段对从 LDV 测声系统采集回来的语音信号做研究分析。

通过振动目标附近的扬声器播放如图 2.6 所示的干净语音信号，图 2.6 上方图为该语音信号的时域波形，图 2.6 下方的图为该语音信号的语谱图。

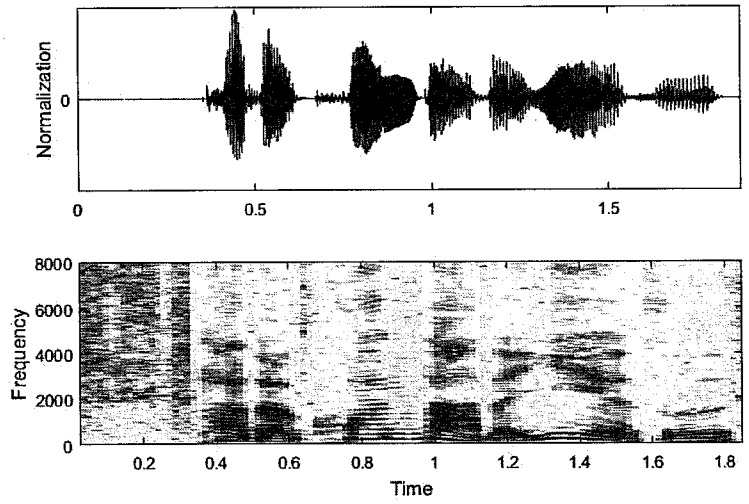


图 2.6 干净的语音信号及其语谱图

Figure2.6 Clean voice signals and their spectrogram

通过控制扬声器播放强度，从 LDV 测声系统采集得到-4dB 到 4dB 的语音信号。同时，对采集回来的语音信号做统一的滤波和归一化处理。图 2.7 上方所示为从 LDV 测声系统获取的不同信噪比的语音信号的时域波形，下方所示为语音信号相应的语谱图。

与图 2.6 对照可以发现，语音信号通过空气耦合驱动振动目标振动，然后通过激光多普勒测声系统来还原语音信号的过程，给语音信号中增加了很多噪声。并且，语音信号的强度越弱，相应的由激光多普勒测声系统所引入的噪声越强。同时，语音信号的语谱结构也有很大的破坏，尤其是在高频段，从语谱图中基本看不出任何语音信号的频谱。而且，在语音信号损失的同时，语音信号当中也引入了一定的非平稳噪声，这将给对语音信号的处理带来了更大的挑战。

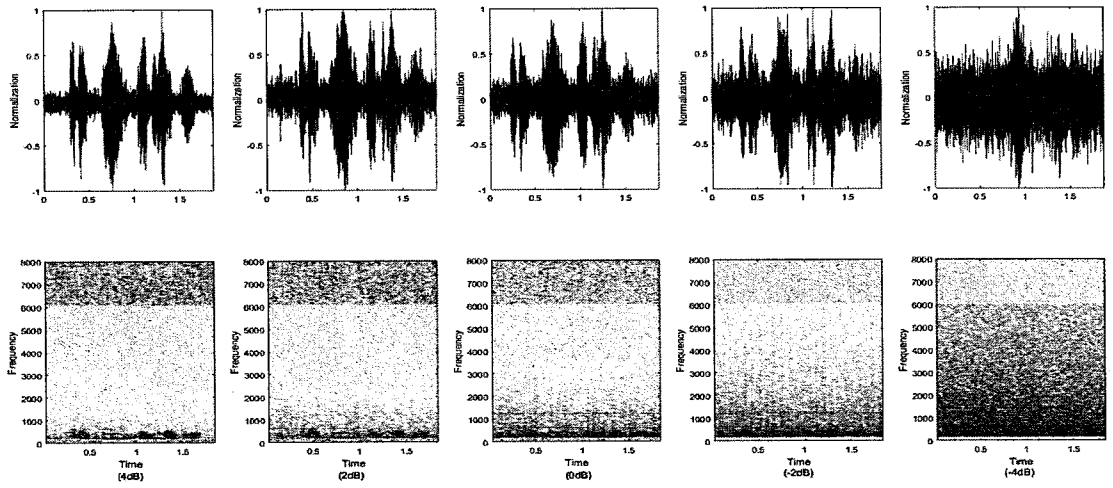


图 2.7 不同信噪比的语音信号及其语谱图

Figure2.7 Speech signal and spectrogram of different SNR

本文利用前文所构建的基于 LabVIEW 小波的语音降噪系统，对从 LDV 采集来的数据进行降噪处理。

首先将图 2.7 所示的语音信号通过 LabVIEW 小波的语音降噪系统，当使用 bior 小波基，分解层数为 5 层，使用极大极小值算法(Minimax)为降噪阈值选择算法时，得到了如图 2.8 所示的降噪结果。

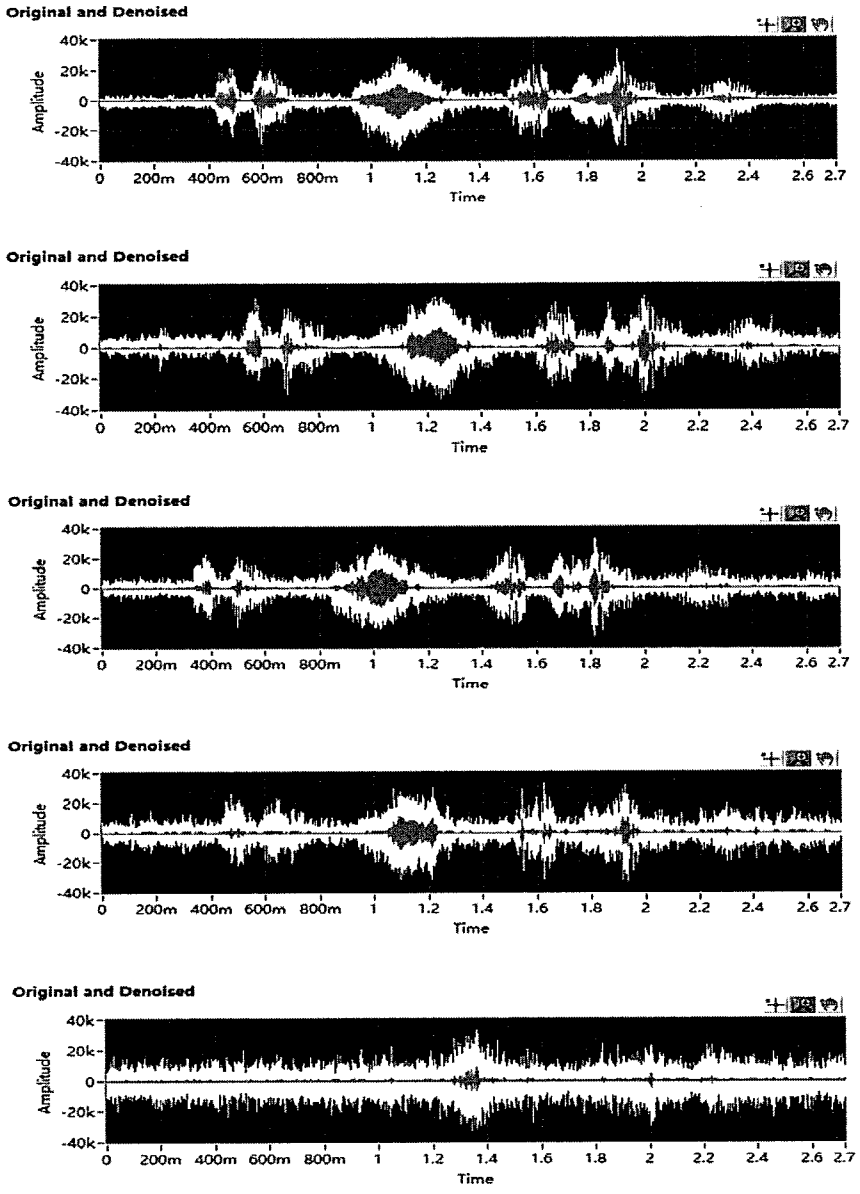


图 2.8 基于 LabVIEW 小波语音降噪系统处理结果

Figure2.8 Result of wavelet speech signal denoising based on LabVIEW

从图 2.8 中可以看出，基于 LabVIEW 小波语音降噪系统虽然可以很好的降低 LDV 测声系统中的噪声。但是，在降噪的同时，对语音信号的质量也有很大的损害。而对于 LDV 测声系统说，在降噪的同时保留语音信号中的信息，是选择降噪算法时所必须要考虑的条件。从图 2.8 所示的结果可以看出，小波语音降噪的手段显然不满足该使用场景所需的条件。

所以，我们必须寻求一种在降低语音中噪声的同时，能够保留语音中原有信息的降噪方式。通过本文第一章的论述，本文选择使用神经网络语音降噪的方式完成该工作。

第 3 章　语音信号特征提取与质量评价

3.1　语音信号特征提取

对语音信号来说，频域的特性较为稳定，不会因为时域少数的奇异值发生大的变化。另外，对语音信号的感知，主要也是对语音信号的频谱的特性进行辨识。所以，本文在将信号输入神经网络之前，首先要做的便是对其进行特征提取。在的工作中，尝试使用四种不同的特征提取方式对语音信号特征进行提取。分别是信号的短时傅里叶变换、对数功率谱和梅尔倒谱、对数梅尔倒谱。本文用如下的方式来构建语音信号特征提取后的结果，以及利用神经网络处理后结果的关系模型。

语音信号和 LDV 之间的作用关系非常复杂，但加性噪声是影响听感的主要因素[25]，在这里可以将模型简化为加性噪声：

$$y(t)=x(t)+n(t) \qquad \ldots（3.1）$$

其中 $y(t)$ 表示 t 时刻从 LDV 测声系统采集的带噪的语音信号，$x(t)$ 表示 t 时刻播放的语音信号，$n(t)$ 表示由 LDV 系统在 t 时刻添加的语音信号噪声。

信号的短时傅里叶变换、对数功率谱和梅尔倒谱、对数梅尔倒谱分别用如下的方式来定义。

3.1.1　短时傅里叶变换

对等式（3.1）两边做离散傅里叶变换可以表示为：

$$Y=X+N, \qquad \ldots（3.2）$$

其中 X 和 N 表示干净的语音信号、和LDV测声系统噪声的离散傅里叶变换。Y 表示他们的和。

先对信号进行分帧处理，然后计算每帧信号的 DFT 系数：

$$Y(w)=\sum_{t}^{T-1}y(t)h(t)e^{-j2\pi dt/T},\ \mathrm{w}=0,1,\ldots,T-1, \qquad \ldots（3.3）$$

其中 $h(t)$ 表示窗函数，w 表示信号的频率[28]。对时域的语音信号进行分帧加窗，然后对每帧的信号做离散傅里叶变换，即可得原时域信号的短时傅里叶变换。

3.1.2 对数功率谱

假设 $Y(w)$ 的定义为等式（3.3）中所定义的，可以用如下公式定义对数功率谱：

$$Y(w)=\log\left|Y(w)\right|^2,\ w=0,1,...,W-1, \qquad \ldots（3.4）$$

其中，$W=T/2+1$[29]。取语音信号的对数功率谱 $X(w)$ 和从 LDV 获取的噪声的对数功率谱 $N(w)$，将它们的和 $Y(w)$ 作为神经网络的输入，假设从神经网络预测出的语音信号的对数功率谱为 $\hat{X}(w)$，由于人耳对相位的微小变化不敏感，所以可以通过带噪声的语音信号的相位重构去噪后的语音信号 $\hat{x}(t)$。

3.1.3 梅尔倒谱

梅尔倒谱（Mel-Frequency Cepstral Coefficient，MFCC）广泛的应用在语音识别和语音信号降噪领域，是最常用到的语音特征[30]。由于人耳对不同频率的语音信号的敏感度不同，使得响度高的频率成分会掩蔽响度低的频率成分，使得高频的成分不容易被人耳所听到。所以，将从低频到高频的信号作用到一组由密到疏的滤波器，将滤波器输出的信号的能量作为信号的特征，然后对特性进一步处理后就可以作为语音信号的特征。这种特征在利用了声学模型的同时，不对原始信号做任何的假设，符合人耳的听觉特性，所以取得了广泛的应用。

梅尔倒谱的处理过程如图 3.1 所示：

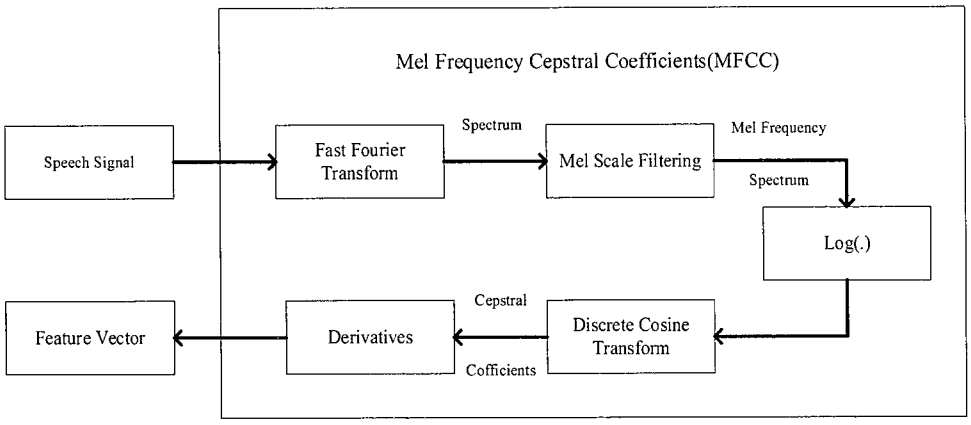


图 3.1 梅尔倒谱处理流程

Figure3.1 MFCC process flow

梅尔倒谱的计算分为以下步骤：

首先将信号分成 25ms 的帧，帧步长为 10ms，允许帧中有重叠。然后对每一

帧的信号做离散傅里叶变换，然后求每帧信号的功率谱。在得到功率谱后，将其通过梅尔滤波器组。梅尔滤波器组是 20-40 个的三角滤波器。在得到通过梅尔滤波器所获得的语音信号的频谱后，计算每个滤波器中的能量。在得到能量后，取每个能量的离散余弦变换（Discrete Cosine Transform，DCT），保持离散余弦变换的一部分系数，最后这部分特征就构成语音信号的梅尔倒谱特征向量。如果在计算滤波器的能量后取对数，所最后获取的特征就变成对数梅尔倒谱（Log Mel-Frequency Cepstral Coefficient，LOGMFCC）。

3.2　语音信号质量评价指标

为了对去噪后的语音信号有一个量化的衡量标准，本文引入了 5 个在语音信号处理领域常用的评价指标，分别是语音感知质量评价（Perceptual Evaluation of Speech Quality，PESQ）[31]，短时客观可懂度测量（Short-time Objective Intelligibility Measure，STOI）[32]。以及信号-失真-比率（Source to Distortion，SDR）、信号-人工-比率（Source to Artifacts Ratio，SAR）、信号-干扰-比率（Signal to interference ratio，SIR）,这三种评价指标可以用来衡量带噪声的语音信号经算法处理后，语音信号和噪声分离的性能[33]。

3.2.1　语音感知质量评价

因为语音信号的感知主体是人，所以主观评价能够反映人对语音信号质量的真实评价。但是主观评价也有着非常显著的缺点，就是在评价过程中非常的费时费力，而且测试人员的主观因素对语音质量的影响很大[31]。为了解决这个问题，常用一个映射函数来拟合主观评价和客观评价之间的关系：

$$\hat{S}=T\left(O_i\right) \qquad \ldots(3.5)$$

其中 $\hat{S}$ 为客观评价转换为主观评价的值，$T(\cdot)$ 为转换函数，O_i 为客观评价所得到的分值。

它是语音质量主观评价，和客观评价相关度最好的一个评价算法，是衡量语音信号质量的一个重要的指标。

3.2.2　短时客观可懂度测量

短时客观可懂度测量是近几年为了衡量语音信号的可懂度而提出的。语音信

号在高信噪比的时候，要听懂语音信号的内容并不困难，但是，在低信噪比的情况下，提升语音信号的可懂度更为重要，所以短时客观可懂度评价指标在低信噪比的情况下更有意义。

3.2.3 语音信号分离测量

为了比较几种算法对语音信号和噪声分离的性能，衡量最佳的算法，文献提出了三种评价指标来衡量信号的失真。此种评价方式将信号看作是由（1）需要提取的目标信号，称作真实信号，记作 s_{target}。（2）干扰目标提取的信号，称其为干扰信号，记作 e_{interf}。（3）传感器等产生的噪声，称其为噪声信号，记作 e_{noise}。（4）提取真实信号的过程中,由算法引入的信号，称其为人工信号，记作 e_{artif}，这几部分组成。

在此基础上，定义了三种评价指标，

信号失真比率（Source to Distortion Ratio，SDR）：

$$\text{SDR} := 10\log_{10}\frac{\left\|s_{\text{target}}\right\|^2}{\left\|e_{\text{interf}}+e_{\text{noise}}+e_{\text{artif}}\right\|^2} \quad \ldots(3.5)$$

信号干扰比率（Source to Interferences Ratio，SIR）：

$$\text{SIR} := 10\log_{10}\frac{\left\|s_{\text{target}}\right\|^2}{\left\|e_{\text{interf}}\right\|^2} \quad \ldots(3.6)$$

信号人工比率（Sources to Artifacts Ratio，SAR）：

$$\text{SAR} := 10\log_{10}\frac{\left\|s_{\text{target}}+e_{\text{interf}}+e_{\text{noise}}\right\|^2}{\left\|e_{\text{artif}}\right\|^2} \quad \ldots(3.7)$$

第 4 章　基于深度循环神经网络的语音降噪

4.1　深度循环神经网络语音降噪

4.1.1　神经网络简介

深度神经网络（Deep Neural Networks）是人工神经网络(Artificial Neural Network)的改进，主要用于解决人工神经网络在训练过程中所引入的过拟合问题和局部最优问题[22]。图 4. 1 所示是深度神经网络最为典型的一种结构，

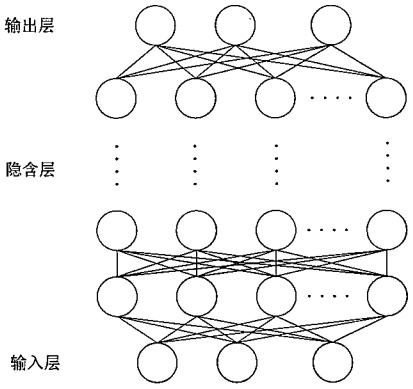


图 4.1 深度神经网络

Figure4.1 Deep Neural Network

它由输入层，输出层和隐含层组成，每层网络的输出可由式 4.1 表示：

$$h^k = \text{logistic}\left(b^k + W^k h^{k-1}\right) \quad \ldots(4.1)$$

其中，h^k 表示第 k 层的网络的输出，b^k 表示第 k 层的网络的偏移量，W^k 表示第 $k-1$ 层的网络和第 k 层的网络之间的连接矩阵，激活函数为：

$$\text{logistic}\left(x\right) = \frac{1}{1+e^{-x}} \quad \ldots(4.2)$$

激活函数也可以有多种不同的形式，如 tanh 函数，

$$\tanh\left(x\right) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad \ldots(4.3)$$

和 Relu，

$$\text{Relu}\left(x\right) = \max\left(0, x\right) \quad \ldots(4.4)$$

使用 sigmoid 激活函数，假设网络的层数为 l，则第 l 层的网络的输出为：

$$h_i^l = \frac{e^{b_i^l} + W_i^l h^{l-1}}{\sum_j e^{b_j^l} + W_j^l h^{l-1}} \quad \ldots(4.5)$$

其中 h_i^l 表示第 l 层网络的第 i 个神经元的输出结果，W_i^l 表示第 l 层，第 i 行的神经元的权重矩阵。

4.1.2 深度循环神经网络结构

前面所述的神经网络，在网络的结构确定下来之后，能够输入的数据的维度也就随之确定，而循环神经网络（Recurrent Neural Network，RNN）则通过时序性的网络结构，能够处理可变维度的输入数据，这使得 RNN 网络结构在序列问题时有更大的优势[34]。RNN 网络相当于是神经网络结构的扩展，但网络的形态和神经网络相比，已经有了较大的不同。4.1.1 节所述的深度神经网络只是单纯接受输入到网络中的数据，而 RNN 可以通过时序性的网络结构找出序列信号前后之间的相关性。

RNN 之所以能够提取时序特性，关键在于状态（state）的引入。在一个样本输入到网络中后，RNN 网络可以先得到一个状态，然后通过这个状态再去获取输出，同样，这个状态还会和下一刻的输入一起影响下一个输出，甚至整个时序中的状态，通过这样的网络结构，能够将输入的时序的前后关联发掘出来。可以把状态看作是一个个的记忆结构，它能帮助记住样本的时序结构信息，从而帮助给新的样本做决策。在这个具体的任务中，相当于去判决一个带噪的频谱信号中有多大的比例是语音信号，多大的比例是系统噪声。

朴素的 RNN 网络结构可以用图 4.2 表示,

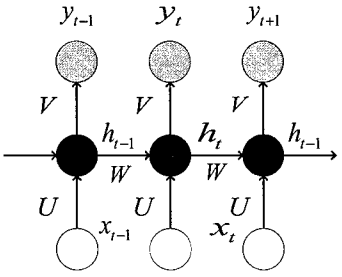


图 4.2 RNN 网络参数结构

Figure4.2 RNN parameter architecture

假设循环神经网络的输入为 x_t，输出为 y_t，隐含层状态为 h_t，t 代表时间。

则系统可以定义为，

$$h_t = f_h\left(x_t, h_{t-1}\right) \quad \ldots(4.6)$$

$$y_t = f_o\left(h_t\right) \quad \ldots(4.7)$$

给定 N 个训练序列，

$$D = \left\{\left(x_1^{(n)}, y_1^{(n)}\right), \ldots, \left(x_{T_n}^{(n)}, y_{T_n}^{(n)}\right)\right\}_{n=1}^{N} \quad \ldots(4.8)$$

网络中每个输出的计算方式如下，

$$y_1 = Vh_1 = V\left(Ux_1\right) \quad \ldots(4.9)$$

$$y_2 = Vh_2 = V\left(Wh_1 + Ux_2\right) \quad \ldots(4.10)$$

$$y_t = Vh_t = V\left(Ws_{t-1} + Ux_t\right) \quad \ldots(4.11)$$

从中可以看出，对每一个时刻而言，RNN 都可以看作是普通的神经网络，例如，$x_i \rightarrow h_i \rightarrow o_i$，$\left(i = \ldots, t-1, t, t+1, \ldots\right)$ 的传递结构。

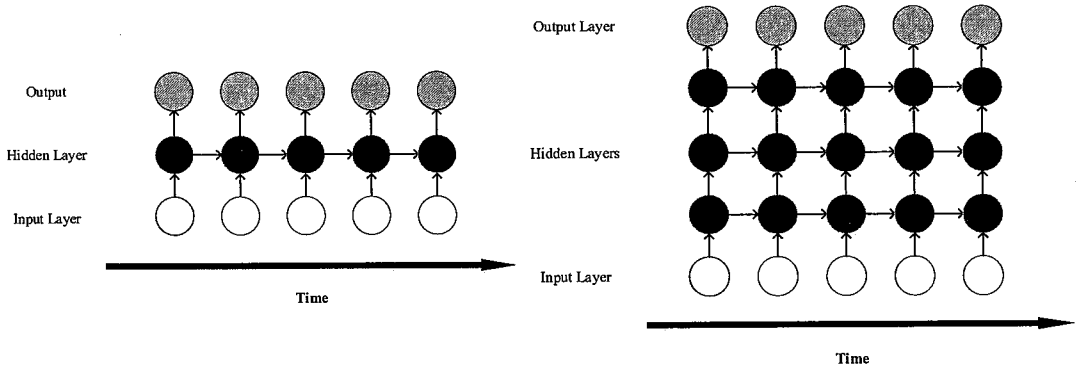


图 4.3 RNN 和 DRNN 网络结构对比

Figure4.3 RNN and DRNN architecture contrast

而对于深度循环神经网络（Deep Recurrent Neural Network，DRNN）[35]，它的结构如图 4. 3 右侧所示。深度循环神经网络的隐含层可以被定义为，

$$h_t^{(l)} = f_h^{(l)}\left(h_t^{(l-1)}, h_{t-1}^{(l)}\right) = \phi_h\left(W_l^T h_{t-1}^{(l)} + U_l^T h_t^{(l-1)}\right) \quad \ldots(4.12)$$

其中 $h_t^{(l)}$ 表示在时间 t 第 l 层的隐含层的状态，ϕ_h 表示非线性激活函数。W_l 表示第 l 层的权重矩阵，U_l 表示第 l 层输入的权重矩阵。而当最后一层的隐含层计算完成，可通过，

$$y_t = f_o\left(h_t, x_t\right) = \phi_o\left(V^T h_t\right) \quad \ldots(4.13)$$

计算出输出结果。V^T 为输出层的权重矩阵。

4.1.3 循环神经网络结构优化

本文利用时序反向传播算法(back propagation through time，BPTT)来优化前述的循环神经网络[36]。时序反向传播算法是一种用于优化循环神经网络的算法模型。时序反向传播算法是反向传播算法（back propagation，BP）的基础上构建的[37]。反向传播算法的流程可以用如下的过程来描述，

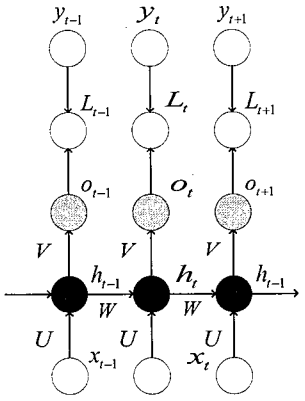


图 4.4 RNN 参数结构

Figure4.4 RNN parameter structure

RNN 网络的参数结构如图 4.4 所示。RNN 在训练时，首先要进行前向传导的过程，从网络结构中，可以得到如下等式：

$$o_t = \varphi(Vh_t) = \varphi\left(V\phi\left(Ux_t + Wh_{t-1}\right)\right) \quad \ldots(4.14)$$

其中，

$$h_0 = 0 = (0,0,\ldots,0)^T, \quad \ldots(4.15)$$

假设，

$$o_t^* = Vh_t, \quad h_t^* = Uh_t + Wh_{t-1}, \quad \ldots(4.16)$$

则有，

$$o_t^* = Vh_t, \quad h_t^* = Uh_t + Wh_{t-1}, \quad o_t = \varphi\left(o_t^*\right), \quad s_t = \varphi\left(s_t^*\right), \quad \ldots(4.17)$$

从而，

$$\frac{\partial L_t}{\partial o_t^*} = \frac{\partial L_t}{\partial o_t}\frac{\partial o_t}{\partial_t^*} = \frac{\partial L_t}{\partial o_t} * \varphi'\left(o_t^*\right) \quad \ldots(4.18)$$

$$\frac{\partial L_t}{\partial V}=\frac{\partial L_t}{\partial Vh_t}\times\frac{\partial Vh_t}{\partial V}=\left(\frac{\partial L_t}{\partial o_t}\times\varphi'\left(o_t^*\right)\right)\times h_t^T \quad \ldots(4.19)$$

由于 $L=\sum_{t=1}^{n}L_t$，从而可得，

$$\frac{\partial L_t}{\partial V}=\sum_{t=1}^{n}\left(\frac{\partial L_t}{\partial o_t}\times\varphi'\left(o_t^*\right)\right)\times h_t^T \quad \ldots(4.20)$$

时序反向传播算法和反向传播算法的不同是，反向传播算法只需要按照空间结构进行传播 $o_t\to h_t\to x_t$，而时序反向传播算法还需要沿时间通道传播 $h_t\to h_{t-1}\to\ldots\to h_1$。假设时间 t 由 n 开始循环到 1，则时间通道上的梯度可以表示为：

$$\frac{\partial L}{\partial h_t^*}=\frac{\partial h_t}{\partial h_t^*}*\left(\frac{\partial h_t^T V^T}{\partial h_t}\times\frac{\partial L_t}{\partial Vh_t}\right)=\phi'\left(h_t^*\right)\left[V^T\times\left(\frac{\partial L_t}{\partial o_t}\times\varphi'\left(o_t^*\right)\right)\right] \quad \ldots(4.21)$$

$$\frac{\partial L_t}{\partial h_{k-1}^*}=\frac{\partial h_k^*}{\partial h_{k-1}^*}\times\frac{\partial L_t}{\partial h_k^*}=\phi'\left(h_{k-1}^*\right)*\left(W^T\times\frac{\partial L_t}{\partial h_k^*}\right),\left(k=1\ldots,t\right) \quad \ldots(4.22)$$

利用时间通道上的梯度可以计算 U 和 W 的梯度：

$$\frac{\partial L_t}{\partial U}=\sum_{k=1}^{t}\frac{\partial L_t}{\partial h_k^*}\times\frac{\partial h_k^*}{\partial U}=\sum_{k=1}^{t}\frac{\partial L_t}{\partial h_k^*}\times x_k^T \quad \ldots(4.23)$$

$$\frac{\partial L_t}{\partial W}=\sum_{k=1}^{t}\frac{\partial L_t}{\partial h_k^*}\times\frac{\partial h_k^*}{\partial W}=\sum_{k=1}^{t}\frac{\partial L_t}{\partial h_k^*}\times h_{k-1}^T \quad \ldots(4.24)$$

假设 RNN 网络在时间 t，第 l 层网络的输出为 $h_t^{(l)}$：

$$h_t^{(l)}=\tanh\left(Ux_t+Wh_{t-1}\right)。 \quad \ldots(4.25)$$

至此，就可以利用训练好的网络结构参数，以及测试集的数据，得到带噪语音信号降噪后的结果。

4.2　LDV 测声系统语音降噪流程

语音信号是一种时变的、非平稳的随机信号。语音信号是由发声的器官的运动生成的，声道的不同尺寸和形状使得语音信号有不同的频谱特性[38]。

通过 LDV 测声系统获取的语音信号会被各种各样的噪声所污染。所面对的语音信号降噪问题是单通道条件下的语音信号降噪问题。

LDV 测声系统的噪声构成非常复杂，为简化问题，从加性噪声出发开展研

究。对于加性噪声，信号模型表示为：

$$y(t)=x(t)+n(t) \quad \ldots(4.26)$$

其中，$y(t)$表示t时刻从LDV采集的带噪的语音信号，$x(t)$、$n(t)$分别表示t时刻无噪声的语音信号和噪声。

将带噪声的语音信号用窗函数分割成一个个重叠的帧，然后将其转换到频域。由于短时傅里叶变换是线性的，所以可以得到如下的变换形式：

$$Y(k,t)=|Y(k,t)|e^{j\angle Y}=X(k,t)+N(k,t)=|X(k,t)|e^{j\angle X}+|N(k,t)|e^{j\angle N} \ldots(4.27)$$

其中k表示频率，t表示时间。$X(k,t)$、$N(k,t)$、$Y(k,t)$分别表示干净的语音信号、噪声、带噪的语音信号在时频点(k,t)的频谱。$\angle X$、$\angle N$、$\angle Y$表示对应的相位信息。由于人耳对相位不敏感，所以可以将模型近似为：

$$|Y(k,t)|=|X(k,t)|+|N(k,t)| \quad \ldots(4.28)$$

式（4.28）所表示的信号的时频谱可以用如下的矩阵形式来表示：

$$\boldsymbol{Y}=\boldsymbol{X}+\boldsymbol{N} \quad \ldots(4.29)$$

其中$\boldsymbol{Y}\in R^{K\times T}$、$\boldsymbol{X}\in R^{K\times T}$、$\boldsymbol{N}\in R^{K\times T}$分别表示带噪的语音信号、干净的语音信号和LDV系统噪声的频谱。其中K表示频点的数目，T表示时间帧。

如图4.5所示为时刻t利用神经网络去噪的示意图。输入网络的信号x_t是含噪声的语音信号，然后取其特征。神经网络的输出为两部分$\hat{y}_{1t}$和$\hat{y}_{2t}$，分别表示对语音信号的预测结果和对噪声的预测结果。其中隐含层的层数为l，假设原信号为y_{1t}和y_{2t}。

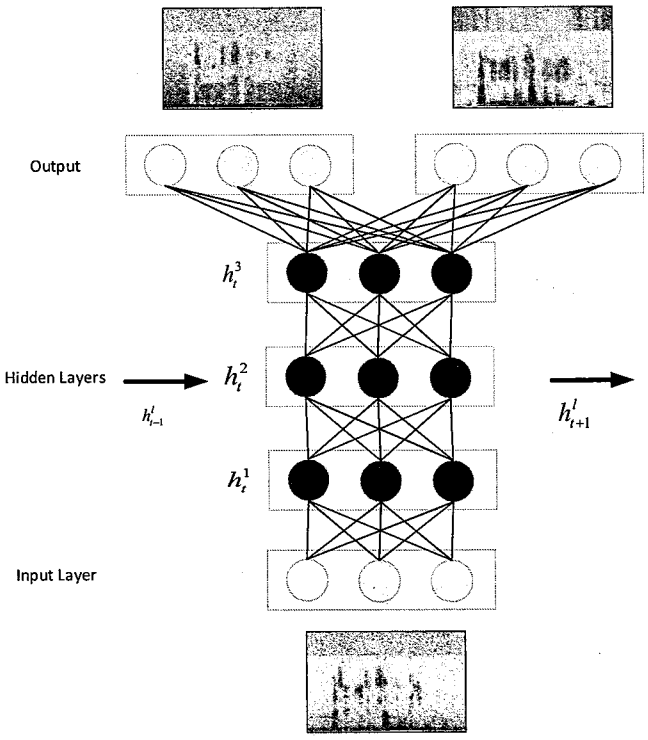


图 4.5 去噪网络结构

Figure4.5 Denoised network architecture

在频域完成语音信号的降噪处理之后，再利用带噪声的语音信号的相位，作为对干净的语音信号的相位的估计，再经过逆傅里叶变换即可得到信号的时域波形[26]。

取语音信号的频谱 $X(w)$ 和从 LDV 获取的噪声的频谱 $N(w)$，将它们的和 $Y(w)$ 作为神经网络的输入，假设从神经网络预测出的语音信号的频谱为 $\hat{X}(w)$，由于人耳对相位的微小变化不敏感[39]，所以可以通过带噪声的语音信号的相位重构去噪后的语音信号 $\hat{x}(t)$：

$$\hat{X}(w)=\exp\left\{\hat{X}(w)/2\right\}\exp\left\{j\angle Y(d)\right\} \quad \ldots(4.30)$$

通过逆傅里叶变换可以得到时域信号 $\hat{x}(t)$：

$$\hat{x}(t)=\frac{1}{T}\sum_{k=0}^{L-1}\hat{X}(k)e^{j2\pi kt/T} \quad \ldots(4.31)$$

同理可以得到由神经网络模型估计得到的系统噪声 $\hat{n}(t)$。

从上面的网络结构中可以看出，用于语音信号降噪的深度循环神经网络处理方式，主要由两个部分组成，一个部分是特征提取部分，另一个部分是网络结构

部分。

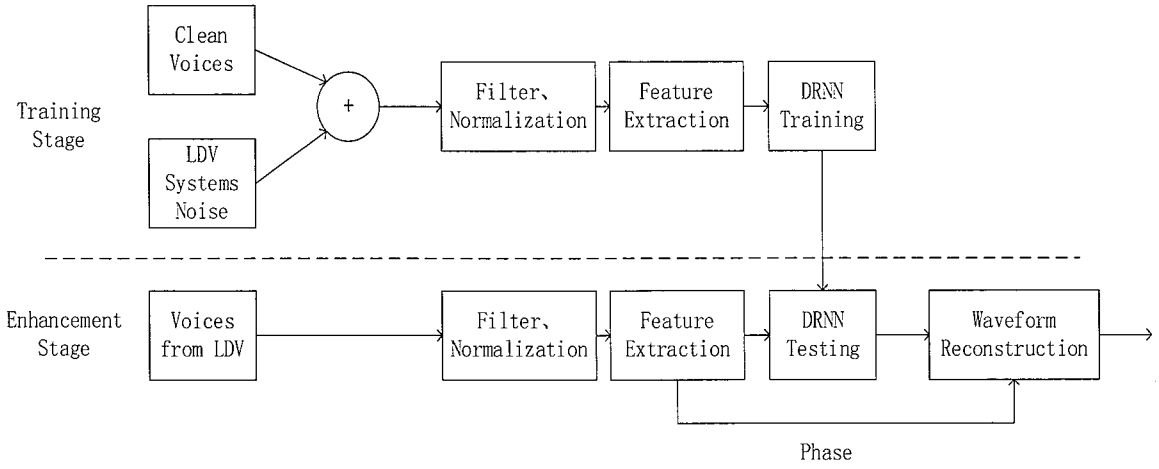


图 4.6 数据处理流程

Figure4.6 Data processing flow

如图 4.6 所示，数据处理的流程主要分为两个环节。训练环节和测试环节，在训练的阶段，先对语音信号和 LDV 系统噪声中的特征进行提取，然后对语音信号和 LDV 混合信号的特征进行提取，然后将上述数据输入到网络结构中对神经网络进行训练。将每个帧长设置为 256，帧移是 128，然后计算重叠帧的特征。

为了测试训练好的网络的性能，首先用干净的语音信号与从 LDV 采集回来的系统噪声做直接相加，然后做滤波、归一化处理，然后进行特征提取，接着送入网络进行测试。为了测试训练好的网络在实际应用中的性能，将从 LDV 获取的带噪语音信号在提取特征后送入网络中进行测试。

4.3 深度循环神经网络语音降噪算法仿真

为了验证前面所构建的网络模型对语音降噪的有效性，利用多种不同强度的噪声，和语音信号，对基于神经网络的语音降噪算法的有效性进行验证。

使用高斯噪声，以及干净的语音信号构建用于测试所构建的网络有效的数据集。将高斯噪声和干净的语音信号放置于同一个音频文件的两个声道，然后对该文件进行统一的滤波和归一化处理。所使用的音频文件的采样率为 16000Hz，每一个音频文件的时长为 10s。

在训练过程中，首先将每个训练文件的两个声道的信息进行分帧，然后对每一帧的信号提取其特征，将它们的和作为神经网络的输入。首先对网络进行随机

初始化，然后通过初始化参数求得网络的输出。利用网络的输出，以及每一帧信号的特征，求其目标函数。然后用前文所述的网络优化方式对网络进行优化，在这里用了 1000 个迭代周期。

在网络训练完成后，利用文件结构和训练集一样，但是语音信号内容不一样的文件构建训练集。构建了含高斯噪声的不同信噪比的语音信号，送入网络中进行测试。

如图 4.7 所示，为-6dB 到 6dB 的含高斯噪声的语音信号的处理结果。从图中可以看出，在 SDR、SIR、SAR、STOI 四项指标上，随着信噪比的提升，语音信号的质量也有明显的提升。

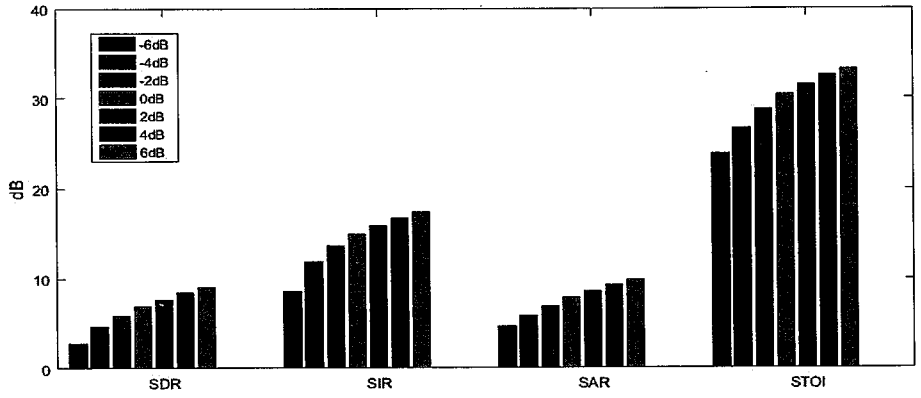


图 4.7 DRNN 处理带高斯白噪声语音信号结果

Figure4.7 DRNN processing results of speech signals with Gauss white noise

如图 4.8 所示为使用在本章节中构建的神经网络，处理含有高斯噪声的不同信噪比的语音信号，处理前语音信号的波形如图 4.8 第一行所示，处理后的结果如第二行所示，每一列对应处理前后的结果。从图中可以看出，所构建的神经网络，在处理平稳的高斯噪声时，有较好的性能。在滤除绝大部分噪声的同时，语音信号也得到了很好的保留。

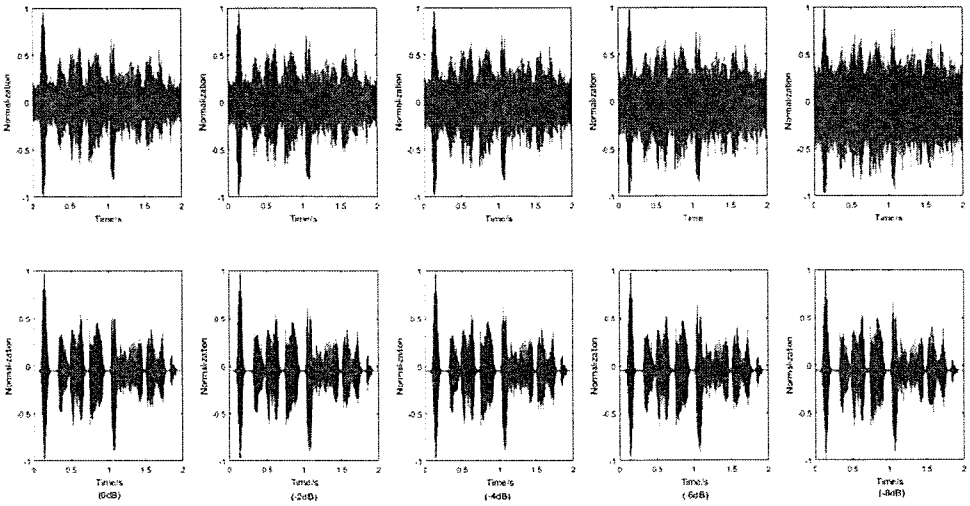


图 4.8 利用 DRNN 处理不同信噪比的含高斯噪声的语音信号

Figure4.8 Using DRNN to process different SNR speech signals with Gauss noise

如图 4.9 所示为图 4.8 所示语音信号的波形所对应的语谱图。从图 4.9 的第一行中可以看出，虽然语音信号中加入了非常强的高斯白噪声，但是语音信号的语谱结构仍然相对完整。在使用神经网络对该语音信号进行处理时，先将语音信号进行特征提取，神经网络可以将语音信号和高斯噪声的特征很好的辨识出来，然后对其做相应的分离处理。从图 4.9 第二行的结果中也可以看出，在去除高斯噪声后，语音信号的语谱结构仍然得到了很好的保留。

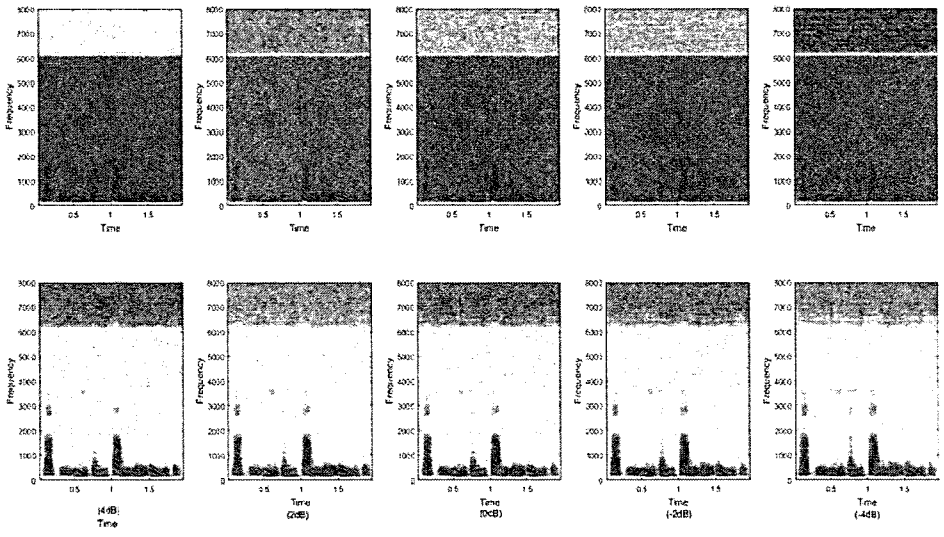


图 4.9 处理前后语音信号的语谱图

Figure4.9 Spectrogram of speech signal before and after processing

第 5 章　LDV 测声系统语音降噪实验

5.1　数据集的构建

用于训练神经网络的数据分为两部分，语音信号和噪声信号。干净的语音信号取自标准男声和女声朗读，总共 2 小时的声音素材。噪声信号来自于 LDV 测声系统。当在目标附近没有声源信号时，LDV 测声系统的输出即作为噪声信号。干净的语音信号与噪声信号可通过（5.1）式构建大量的平行语料，用于神经网络的训练。

$$Y = X + \alpha N(k),\quad k = \beta, \beta+1, \ldots, T, \ldots, \beta-1 \qquad \ldots(5.1)$$

上式中，α 为噪声信号的能量因子，β 是一个随机因子。调整 α 的值，可以控制输入神经网络的信号的信噪比。

通过设置不同的 α 值，可构建信噪比为-6dB，-4dB，-2dB，0dB，2dB，4dB，6dB 的含噪语音信号。进一步，将含噪语音信号降采样到 16kHz，分成时长为 10s 的小段文件，做统一的滤波和归一化处理，形成数据集。然后将数据集按 8:2 的比例分为训练集和测试集两部分。

为了测试网络结构在实际应用中的降噪性能，用扬声器播放标准男声和女声朗读，驱动振动目标振动，再通过 LDV 采集振动信息，通过 LDV 还原振动目标的振动信息后，即可获得带噪的语音信号，将它作为网络的另一部分测试样本。

5.2　深度循环神经网络参数对降噪结果的影响

为了提高网络的性能，测试网络结构参数对降噪效果的影响，构造多种参数结构的 DRNN 降噪网络，同时，用添加了-6dB, -4dB, -2dB, 0dB, 2dB, 4dB, 6dB 从 LDV 获取的噪声信号去测试网络，用 SDR、SAR、SIR、PESQ、STOI 五个评价指标对网络的降噪效果做量化评估。

控制以下几个方面的参数来进行实验，分别是 DRNN 网络的层数、每层的神经元的个数、不同的特征提取方式、激活函数的形式、目标函数的形式。

5.2.1　网络的层数

首先是 DRNN 网络的层数。为了测试不同的网络层数对降噪结果的影响，构建不同参数结构的 DRNN 网络。

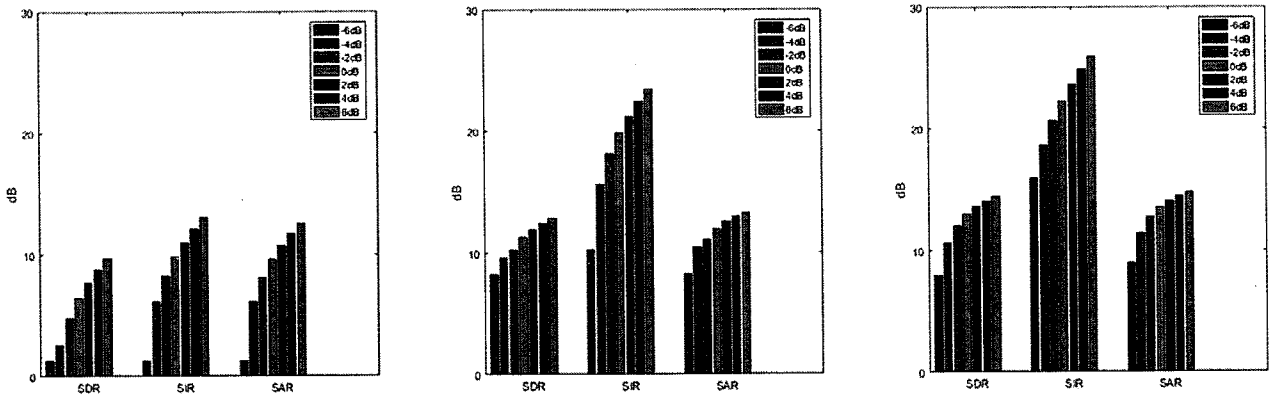
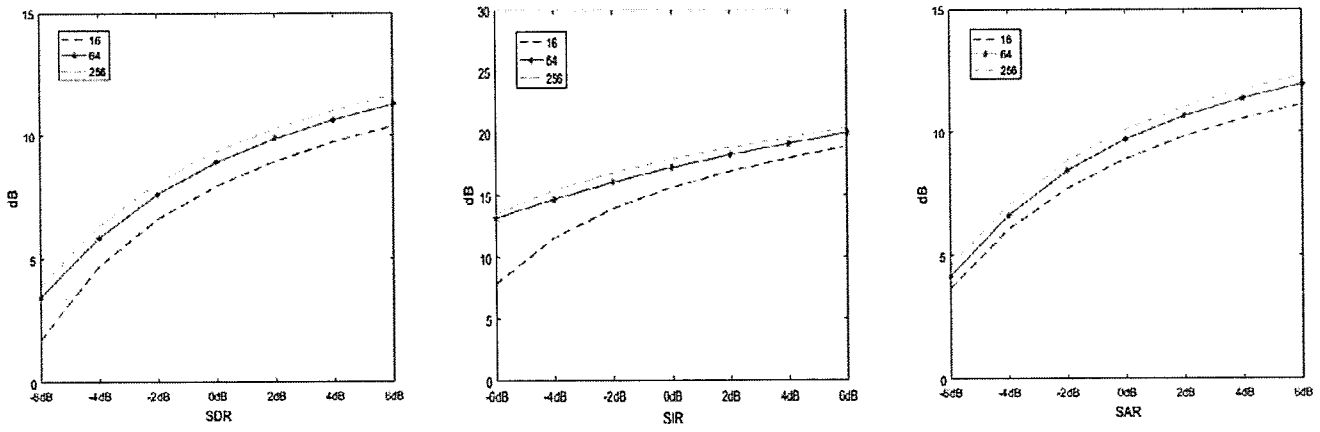


图 5.1 网络层数对降噪结果影响

Figure5.1 Influence of network layers on denoised results

如图 5.1 所示为使用隐含层层数为 1、2、3，隐含层神经元个数为 256 的网络结构参数下的测试集的平均测试结果。用不同颜色的柱状图表示对不同信噪比的信号的处理结果。对比图 5.1，可以看出网络层数的增加可以对明显地提升网络的去噪性能。由此可以看出，网络层数的增加，可以使得网络拟合 LDV 系统噪声和语音信号的能力更强，所以比单一层数的 RNN 网络有着更好的降噪效果。

5.2.2 每层神经元的个数



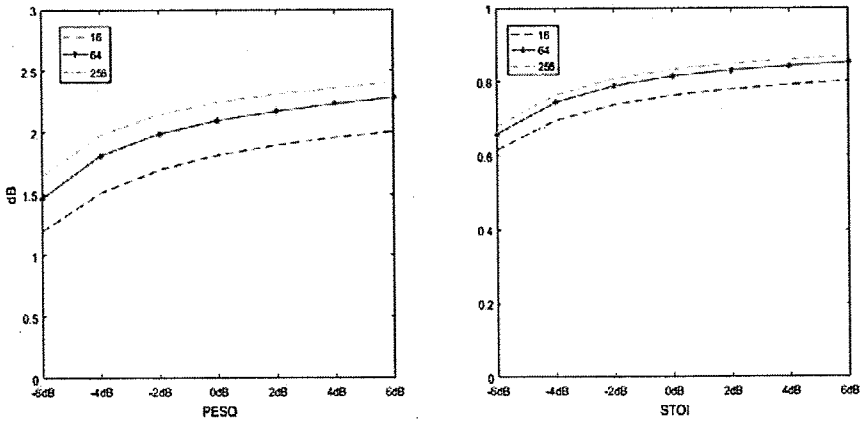


图 5.2 每层神经元个数对降噪结果的影响

Figure5.2 Influence of the number of neurons in each layer on denoised results

图 5.2 所示为 RNN 网络每层使用神经元的个数为 16、64、256，在不同信噪比时，网络的平均降噪结果。不同形状的折线表示对不同网络的处理结果。其中，5.2 上方三个图表示 SDR、SAR、SIR 三个语音信号评价指标的结果。从图中可以看出，当每层的神经元的个数增加时，降噪效果也随之变好，在三个指标上都有相同的结果。网络每层的神经元的个数，对 SIR 评价指标的影响最大，说明网络每层神经元的数量的增加，能够明显的提升网络提取 LDV 系统噪声特征和语音信号特征的能力，可以明显的降低噪声对语音信号的影响。

图 5.2 下方两个图为使用 PESQ 和 STOI 两个评价指标的结果。从图中可以看出，随着每层网络的神经元个数的增加，网络的降噪效果也有明显的改进。但是，降噪的效果并不会随着神经元个数的增加而成比例的增加。网络每层的神经元数量的增加对 PESQ 指标的影响最为显著，所以，适当的增加每层网络的神经元的数量，可以有效的提升网络的降噪效果，以及降噪后语音信号的可懂度。

5.2.3　不同的特征提取方式

图 5.3 所示为使用 STFT、LOGPOW、MFCC、LOGMFCC 四种特征提取方式，在不同信噪比时网络的平均降噪结果。不同类型的折线表示不同的特征提取方式。对比 SDR、SIR、SAR 三个衡量指标，从图中可以看出 STFT 的效果最好，而且在不同信噪比下的降噪效果也基本超过另外三种特征提取方式。MFCC 特征提取方式的效果最差，这种特征提取方式虽然在 SDR、SIR、SAR 上取得的效果很差，但是在 PESQ 上取得与前三个指标类似的结果，这是由于 MFCC 本来就是针对人对语音

信号的感知设计的，所以在主观语音质量评估中能取得较好的结果。同时，也注意到 MFCC 在低信噪比的时候能更有效的提取语音信号特征。

但是 LOGMFCC 可以取得与 STFT、LOGPOW 两种特征提取方式类似的结果。在 PESQ 和 STOI 两个评价指标上也有类似的结果。

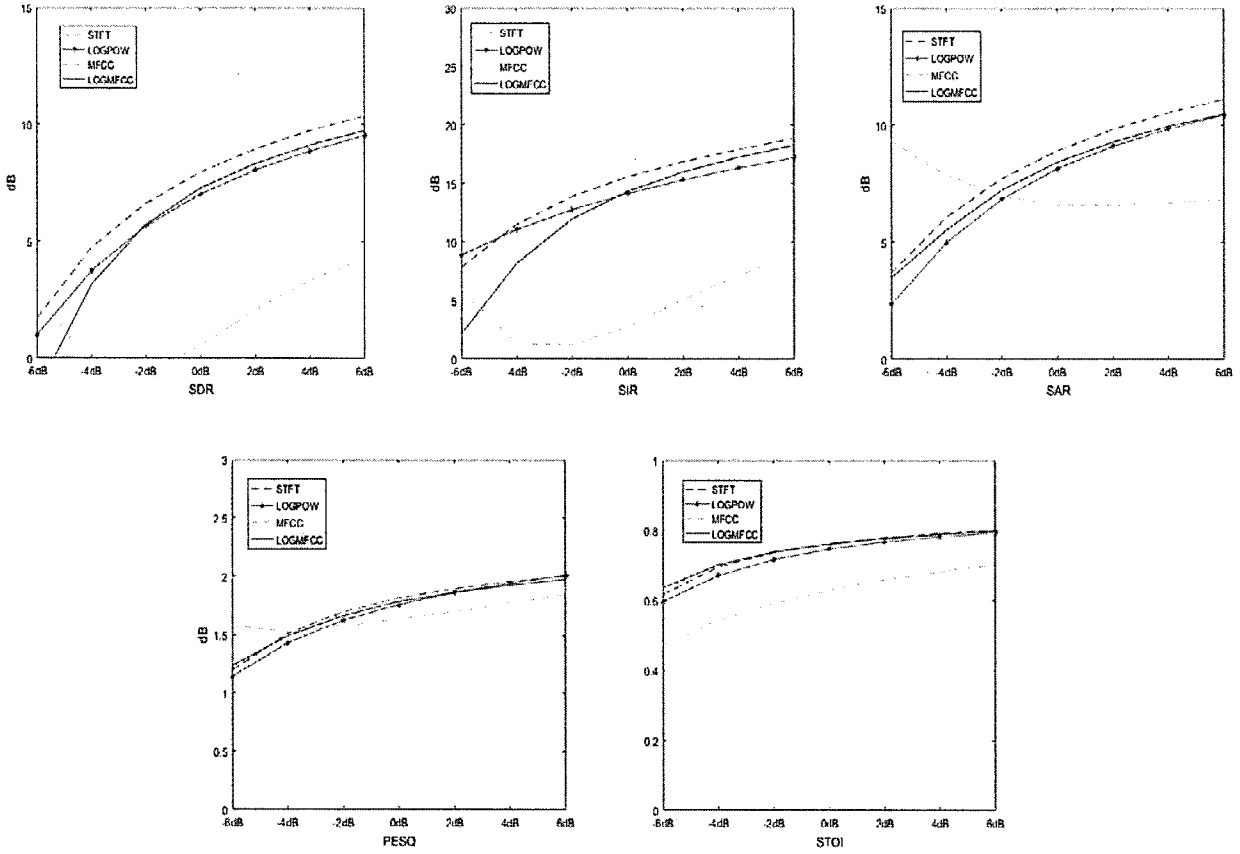


图 5.3 特征提取方式对降噪结果的影响

Figure5.3 Influence of feature extraction methods on denoised results

5.2.4 激活函数形式

激活函数是神经网络的一个非常重要的元素，它可以为网络提供非线性建模的能力。如果的网络没有激活函数所提供的非线性结构，那么神经网络不管有多少层，也只能提供对线性系统的拟合，只有在给网络加入非线性的结构之后，网络才具备有非线性映射学习能力。

激活函数应当具有可微性，用来保证网络可用反向传播算法求解；应当具有单调性，用来保证单层的网络是凸函数；应当具有有限的输出范围，用来保证网络在优化的时候的稳定性。

在网络的构建过程中，使用了如图 5.4 所示的三种不同结构的激活函数。分

别是 tanh、Relu（Rectified Linear Units）、和 Logistic。

tanh 激活函数的形式如下：

$$f(x)=\frac{1-e^{-2x}}{1+e^{-2x}} \qquad \ldots(5.2)$$

Relu 激活函数的形式如下：

$$f(x)=\max(0,x) \qquad \ldots(5.3)$$

Logistic 激活函数的形式如下：

$$f(x)=\frac{1}{1+e^{-x}} \qquad \ldots(5.4)$$

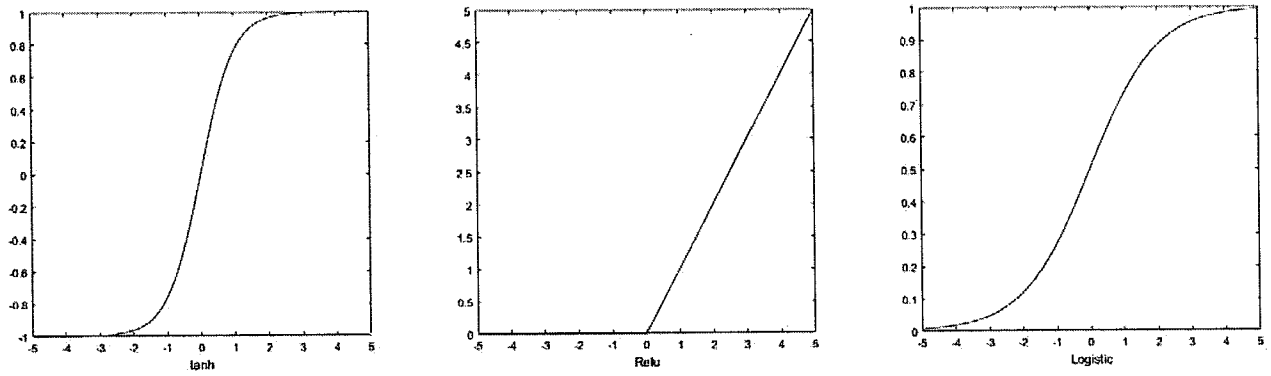
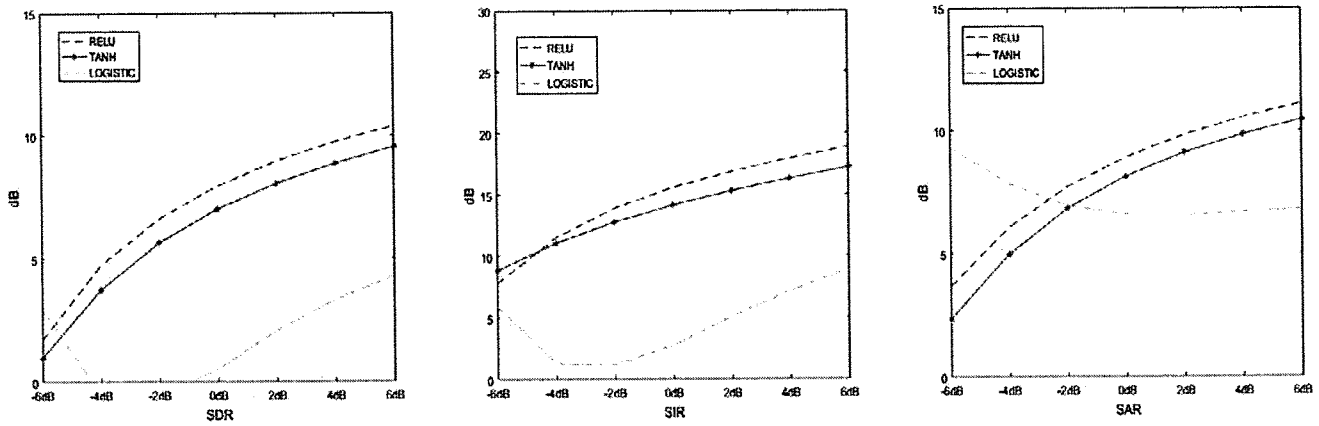


图 5.4 不同的激活函数曲线

Figure5.4 Different activation function curves

图 5.5 为分别利用上述三种函数作为激活函数的处理结果，从图中可以看出，Relu 函数作为神经网络的激活函数，的网络能取得最好降噪结果，使用 Logistic 函数作为神经网络的激活函数，取得的降噪效果最差。但是在信号的信噪低的时候，Logistic 激活函数能够取得较好的降噪结果。



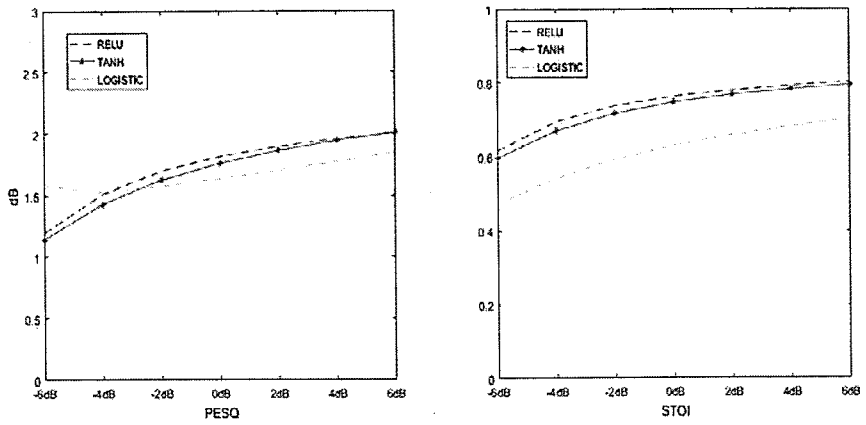


图 5.5 激活函数对降噪结果的影响

Figure5.5 Influence of activation function on denoised results

5.2.5 目标函数形式

为了测试不同的目标函数对网络降噪效果的影响，构建了 3 种不同形式的目标函数。

第一种目标函数使用交叉熵：

$$E(y_t,\hat{y}_t)=-y_{1t}\log\hat{y}_{1t}-y_{2t}\log\hat{y}_{2t} \quad \ldots(5.5)$$

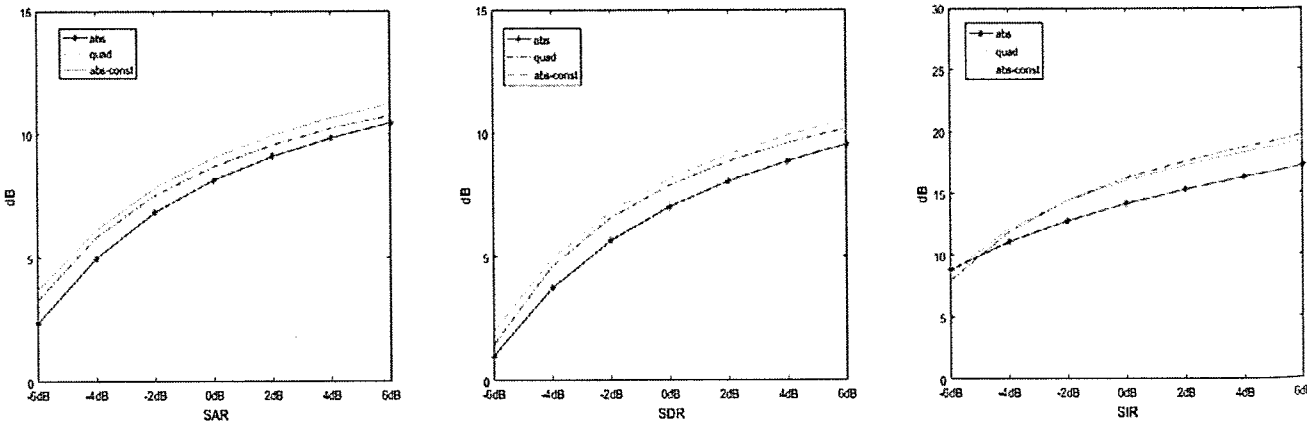
$$E(y,\hat{y})=\sum_t E_t(y_t,\hat{y}_t)=-\sum_t y_t\log\hat{y}_t \quad \ldots(5.6)$$

代价函数为二阶矩

$$J_{MSE}=\sum_{t=1}^{T}\left(\|\hat{y}_{1t}-y_{1t}\|_2^2+\|\hat{y}_{2t}-y_{2t}\|_2^2\right) \quad \ldots(5.7)$$

同时可以引入正则化项提高网络的范化能力，将其设为的第三种目标函数：

$$J_{MSE}=\sum_{t=1}^{T}\left(\|\hat{y}_{1t}-y_{1t}\|_2^2+\|\hat{y}_{2t}-y_{2t}\|_2^2-\gamma\|\hat{y}_{1t}-y_{2t}\|_2^2-\gamma\|\hat{y}_{2t}-y_{1t}\|_2^2\right) \quad \ldots(5.8)$$



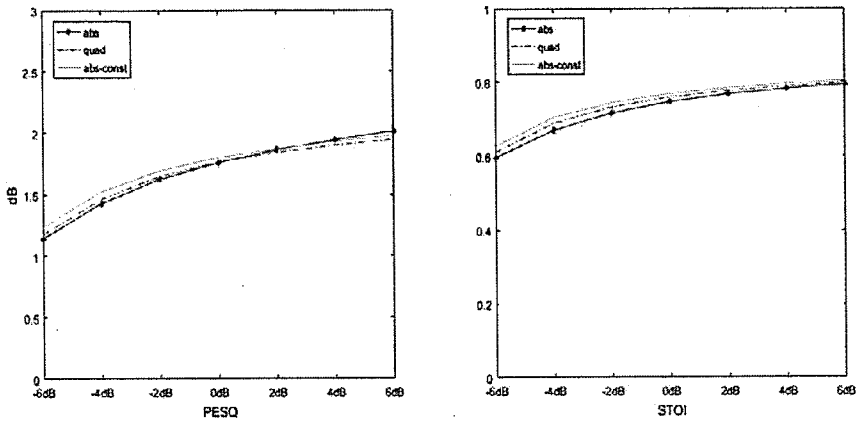


图 5.6 目标函数对降噪结果的影响

Figure5.6 Influence of objective function on denoised results

从图 5.6 上方三个图中可以看出，使用交叉熵的网络在 SDR、SIR、SAR 三个指标上的降噪效果与另外两个相比最差。在引入正则化系数后，网络的性能有一定的提升，但是并不明显。这说明在引入正则化系数后，能够在一定程度上提高网络的降噪能力。

从图 5.6 下方两个图，PESQ 和 STOI 两个指标中可以看出，目标函数的形式，对提高降噪后的语音信号的可懂度影响并不显著。

5.2.6　网络参数配置总结

结合 5.2 节内容的分析，可以得到如下的结论：在使用 DRNN 网络对 LDV 测声系统的噪声进行处理时，使用的每层网络的神经元个数为 256，使用 RELU 激活函数、使用 3 个隐含层的网络、使用 STFT 为特征提取函数、使用二阶方差加正则化函数时，可以得到最好的降噪效果。

使用如上的网络参数训练神经网络，然后利用训练好的网络对从 LDV 测声系统所获取到的语音信号做降噪处理。为了验证 DRNN 网络对不同信噪比的从 LDV 获取的带噪语音信号的处理结果，利用扬声器控制播放的音量，以获取不同信噪比的带噪语音信号。

我们使用的 LDV 测声系统的实物图如图 5.7 所示，

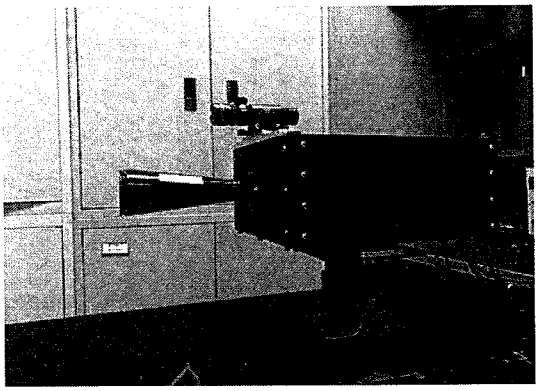


图 5.7 LDV 测声系统实物图

Figure5.7 LDV remote voice acquisition system target photo

我们使用的振动目标以及扬声器如图 5.8 所示：

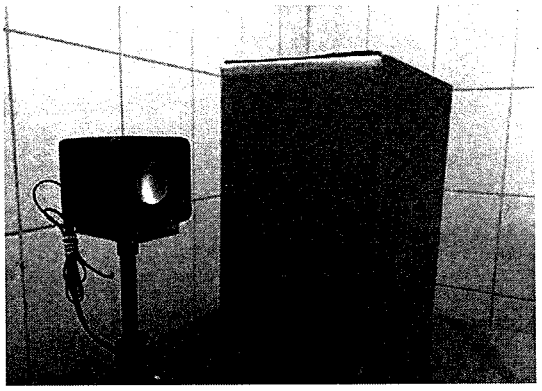


图 5.8 振动目标和扬声器

Figure5.8 Vibration target and speaker

从 LDV 获取带噪的语音信号和处理后的结果如图 5.9 所示：

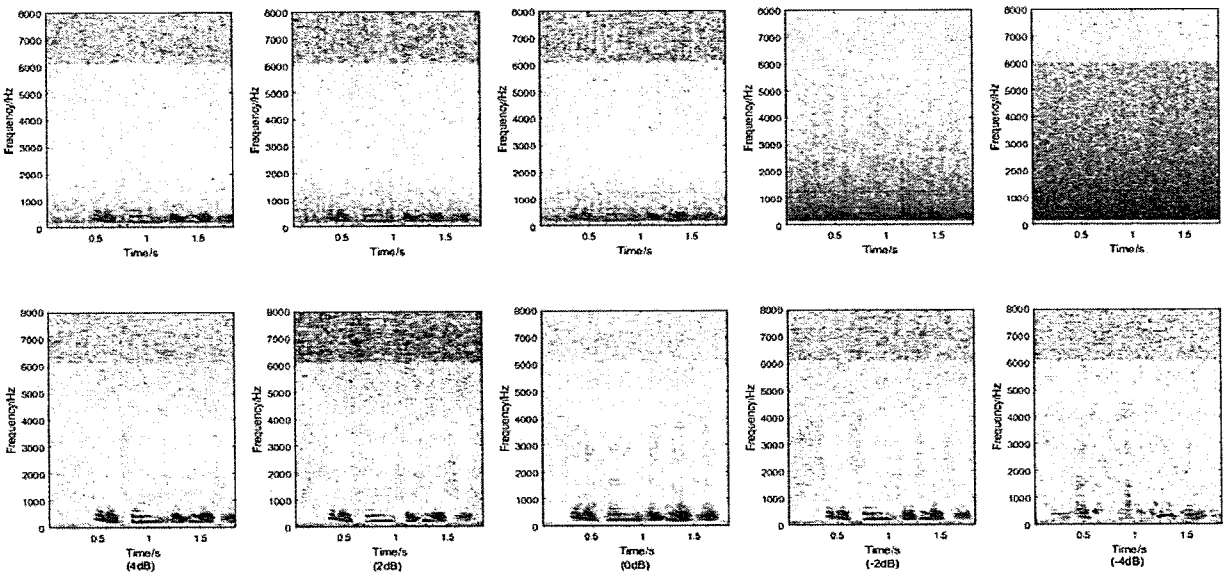


图 5.9 从 LDV 获取的语音信号和降噪后的结果

Figure5.9 Voice be obtained from LDV and denoised results

从图 5.9 中可以看出，所构建的 DRNN 网络在处理 4dB 到-2dB 的从 LDV 获取的语音信号时，可以得到很好的处理结果，语谱图中语音信号的结构也很清晰，但当信号的信噪比降到-4dB，虽然背景噪声去除了很多，但是语音信号有明显的失真。

5.3　神经网络与传统的方法对比

LDV 测声系统的语音降噪问题，可以归结为单通道语音信号的降噪问题。虽然主要以神经网络为提升 LDV 测声系统性能的主要手段，但是，为了对单通道语音降噪有全面系统的了解，本文从传统的单通道语音降噪算法入手，去寻找最适于 LDV 测声系统的语音降噪手段。

同时，传统的单声道语音降噪算法，也对基于神经网络的降噪算法有借鉴意义。比如在将信号送入网络中进行处理时，首先要做的就是语音信号的特征提取，这个步骤在传统的语音信号处理中也常常用到，而且有很多种的方式供参考利用，比如用短时傅里叶变换、对数功率谱、梅尔倒谱、对数梅尔倒谱等。

另外，为了对比本文所构建的用于 LDV 测声系统语音降噪网络的性能，势必要与多种算法做比对分析，这样就需要构建传统的语音降噪算法，来对从 LDV 测声系统所获取的语音信号做降噪处理。

5.3.1　神经网络与谱减法处理结果对比

谱减法（Spectral Subtraction）是语音降噪中较为成熟、而且最为常用的一种方法[40]。该算法假设噪声是统计平稳的，而且语音信号和噪声不相关。该算法用无语音的间隙获得噪声的频谱估计，然后与含噪的语音频谱相减，以此来获得语音信号的频谱估计。谱减法运算量小，易于实现，广泛的应用于语音信号降噪处理。但在使用谱减法的过程中，会不可避免的引入“音乐噪声”，“音乐噪声”的产生是由于对噪声谱的估计不准确，而且对谱减算法中负数部分的处理为非线性的。

谱减法的流程如图 5.10 所示：

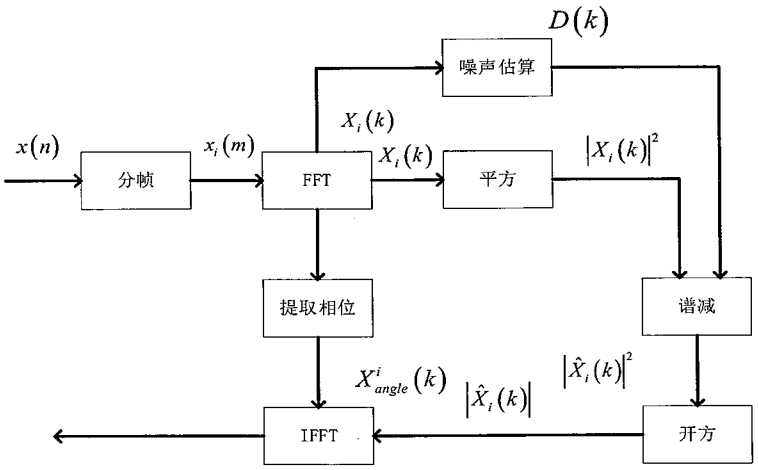


图 5.10 谱减法原理图

Figure5.10 Principle of spectral subtraction

首先，对信号进行分帧处理，对信号噪声的功率谱进行估计。接着通过对带噪声的语音信号做傅里叶变换来得到带噪语音信号幅度谱，然后，将得到的幅度谱减掉前面估计的噪声的幅度谱，得到对原始的语音信号幅度谱的估计。由于的人耳对相位信息不敏感，所以利用前面估计出来的幅度谱和带噪语音信号的相位来重构信号。

假设语音信号为 $x(n)$，则对语音信号分帧后做 DFT，

$$X_i(w)=\sum_{m=0}^{N-1}x_i(m)\exp\left(j\frac{2\pi mw}{N}\right),\quad w=0,1,2,\ldots,N-1 \qquad \ldots(5.9)$$

于此同时求信号的相位：

$$X_i^{angle}(w)=\arctan\left[\frac{\mathrm{Im}(X_i(w))}{\mathrm{Re}(X_i(w))}\right] \qquad \ldots(5.10)$$

对噪声段的信号计算能量的平均值：

$$E(w)=\frac{1}{N}\sum_{i=1}^{N}|X_i(w)|^2 \qquad \ldots(5.11)$$

然后用原始的信号减去这个噪声分量：

$$X_i(w)=\begin{cases}|X_i(w)|^2-a\times E(w), & |X_i(w)|^2\geq a\times E(w)\\ b\times E(w), & |X_i(w)|^2<b\times E(w)\end{cases} \qquad \ldots(5.12)$$

其中 a 是一个过减因子常量，b 是一个增益补偿因子常量。通过此方法，可以在频域得到干净的语音信号。对信号做傅里叶逆变换，再由相位信息，即可得

到降噪后的语音信号。

用谱减法降噪的方法和 DRNN 网络的方法做对照实验。图 5.9 给出了在相同的数据集下的测试结果。从图 5.11 中可以看出，在低信噪比时，利用 DRNN 网络降噪所得的结果，在 STOI 评价指标的表现上，明显要比谱减法降噪所得的结果要好。而在高信噪比时，利用 DRNN 网络降噪所得的结果，在 PESQ 准则的表现上，明显要比谱减法降噪所得的结果要好。这说明 DRNN 网络在处理低信噪比的信号时，所得到的结果比利用谱减法所得到的结果的可懂度更高。在这两个评价指标上，DRNN 网络降噪的性能均比利用谱减法所得到的结果要好。

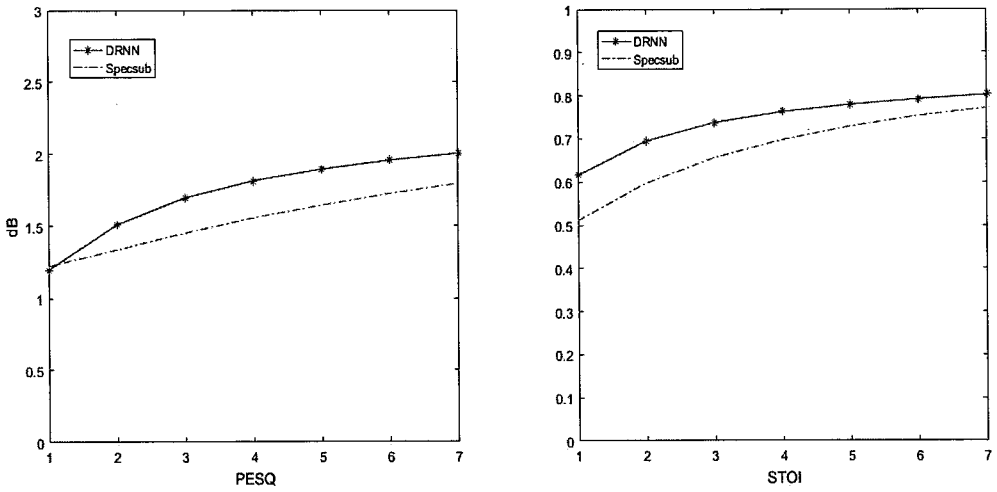


图 5.11 DRNN 和谱减法对比结果

Figure5.11 Comparison between DRNN and spectral subtraction

5.3.2　DRNN 与非负矩阵分解处理结果对比

谱减法假设噪声是固定的，而且对话音检测激活技术有很大的依赖，如果噪声是非平稳的，而且信噪比比较低，那么处理的结果会非常差[41]。而子空间法利用空间分解，将整个空间划分为几个独立的噪声子空间和信号子空间，然后通过对这两个子空间进行处理来实现语音增强。

利用非负矩阵分解（Nonnegative Matrix Factorization, NMF），可以将语音信号的平方谱，或者幅度谱，分解为系数矩阵和两个非负的基矩阵的乘积。而非负矩阵的基矩阵能够表征语音信号的频谱特征。通过干净的语音信号和噪声来训练数据，然后利用非负矩阵分解可以将混合语音中的不同的元素分离开来。

使用非负矩阵分解算法对语音信号进行重建，在低信噪比和噪声环境有较大变换时仍然适用。非负矩阵分解模型对语音信号增强分为训练阶段和测试阶段[43]，系统的处理流程如图 5.12 所示：

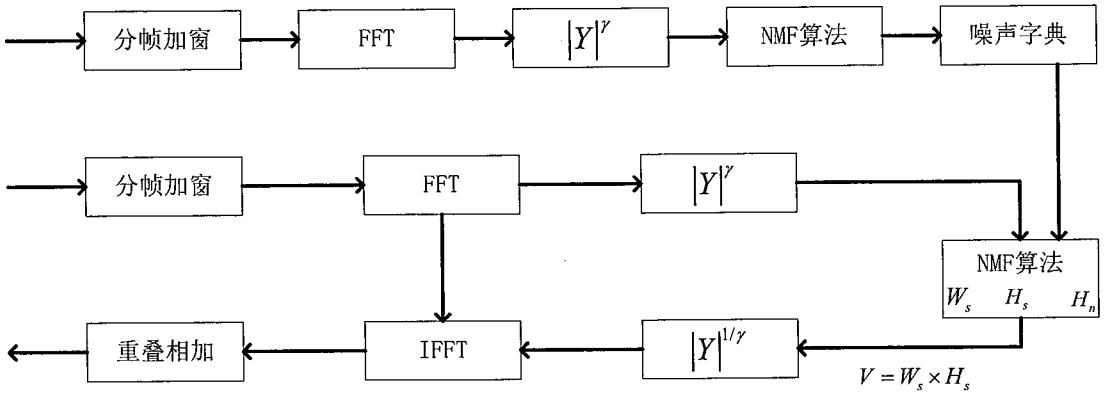


图 5.12 NMF 算法流程

Figure5.12 NMF algorithm flow

该算法在训练阶段首先将噪声转换到频域，然后取二阶范数，然后通过非负矩阵分解算法对噪声信号的频谱进行分解，用来得到噪声的字典矩阵。在增强阶段，首先将带噪的语音转换到频域，然后取二阶范数，通过非负矩阵分解算法分解得到带噪的语音的编码矩阵和字典矩阵，然后通过先验信息，通过相应的迭代计算得到语音的编码矩阵和字典矩阵，来构建相应的语音信号[42]。

用非负矩阵分解降噪的方法和本文所用的方法做对照实验，如图 5.13 所示，左侧为使用非负矩阵分解的方法做不同信噪比 LDV 测声系统语音降噪的实验所得的结果，而右侧为使用本文所使用的最优参数的 DRNN 网络对 LDV 测声系统语音降噪的结果。可以看出本文所用的方法，在 SDR、SIR、SAR 三项评价指标中，较非负矩阵分解降噪的方法有较大的提升。充分的体现了利用深度循环神经网络的方式对从 LDV 测声系统所获得的语音信号做降噪处理，和传统的方式相比的优越性。

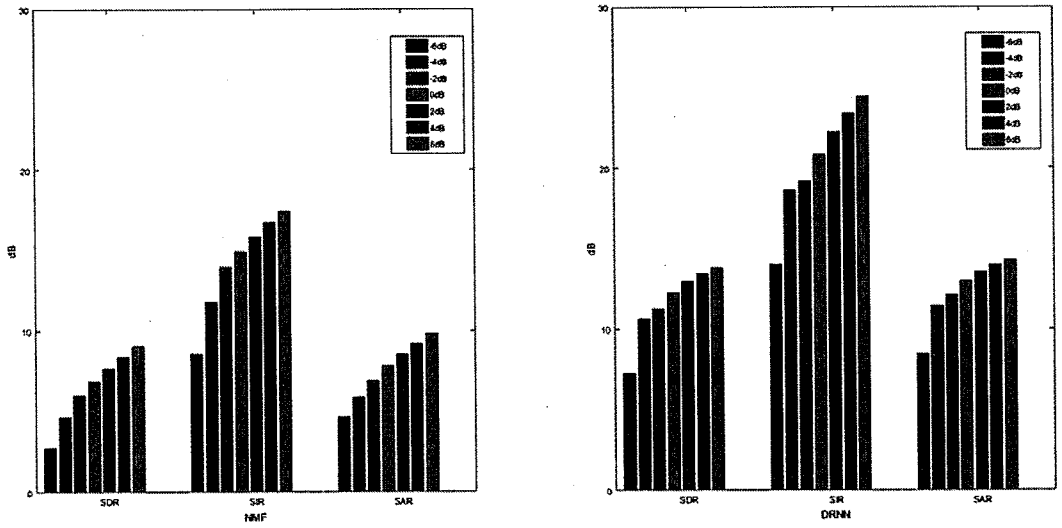


图 5.13 NMF 和 DRNN 对比结果

Figure5.13 NMF and DRNN results

5.3.3　DRNN 网络与维纳滤波器处理结果对比

维纳滤波法（Wiener Filtering）也是语音降噪中的一种常用的手段[43]，它分为时域和频域两种形式，但是时域的方法由于需要求解协方差矩阵的逆，在数据比较长的时候会有很大的计算量，所以在这里采用频域法[44]。

假定观测到的时间序列为，

$$z(n)=s(n)+v(n),\quad n=n_0,n_0+1,\ldots,n_f \qquad \ldots(5.13)$$

其中 $v(n)$ 为噪声，$s(n)$ 为原始信号，n_0 和 n_f 分别为观测起始时刻和观测结束时刻。把观测序列通过离散时间线性系统，

$$\hat{s}\left(\frac{n}{n_f}\right)=\sum_{k=n_0}^{n_f}h(n,k)z(k) \qquad \ldots(5.14)$$

然后做原始信号的线性最小均方估计，通过最小均方估计的正交原理来求解，

$$E\left\{\left[s(n)-\sum_{k=n_0}^{n}h(n,k)z(k)\right]z(i)\right\}=0,\quad (i=n_0,n_0+1,\ldots,n) \qquad \ldots(5.15)$$

即，

$$R_{sz}(n,i)=\sum_{k=n_0}^{n}h(n,k)R_z(k,i),\quad (i=n_0,n_0+1,\ldots,n) \qquad \ldots(5.16)$$

上式即为 Winer-Hoff 方程，由于把信号及其观测过程看作是平稳的随机序列，而且是因果连续时不变系统，所以可以取 $n_0=-\infty$，此时上式变为，

$$R_{sz}(n)=\sum_{l=n_0}^{+\infty}h(l)R_z(n-l)=h(n)*R_z(n),\quad n\geq 0 \qquad \ldots(5.17)$$

对式子两边做 Z 变换，即可得，

$$G_{sz}(z)=H(z)G_z(z) \qquad \ldots(5.18)$$

$$H(z)=\frac{G_{sz}(z)}{G_z(z)} \qquad \ldots(5.19)$$

其中，$H(z)$即所要求的维纳滤波器。将带噪语音信号通过此维纳滤波器来获得降噪后的语音信号。

用维纳滤波器去噪的方法和本文所用的方法做对照实验，结果如图 5.14 所示，

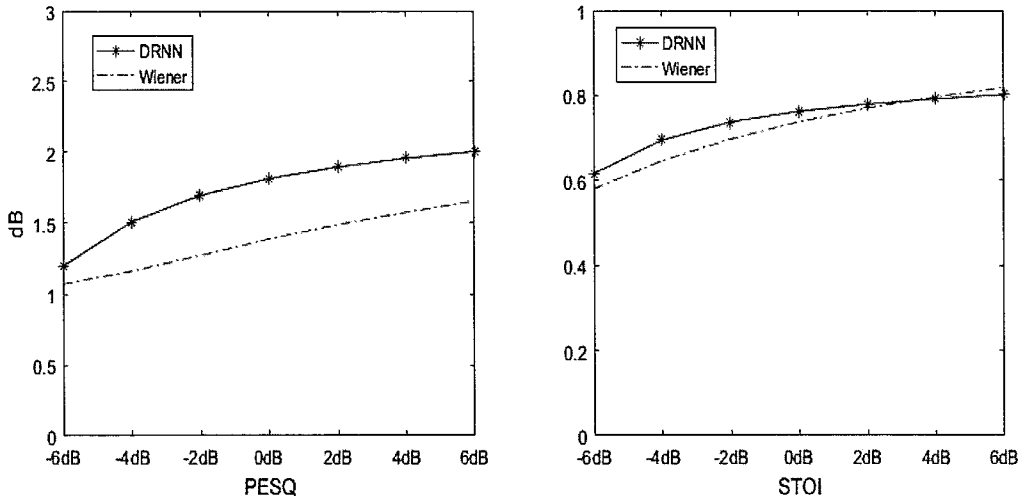


图 5.14 DRNN 和 Wiener 滤波器对比结果

Figure5.14 Contrast results of DRNN and Wiener filters

从图 5.14 可知，以 PESQ 和 STOI 两项语音信号质量评价的标准来看，使用深度循环神经网络，处理不同信噪比的语音信号，所获得的语音的质量普遍比使用维纳滤波器所获得的效果要好。但在 STOI 这项指标中，维纳滤波器也取得了较好的效果。

第 6 章　总结和展望

6.1　论文总结

本文主要针对从 LDV 测声系统所获取的语音信号，通过 DRNN 网络的方法做降噪处理的研究。实验探究的过程中，首先对 LDV 测声系统的基本原理及其构造进行了学习研究。并且，对 LDV 测声系统国内外的发展现状进行了系统性的了解。于此同时，为了使得采集实验数据和分析实验结果更加简便，构建了基于 LabVIEW 的 LDV 测声系统语音信号实时采集和分析装置，给实验带来了极大的便利。

为了量化的印证 DRNN 网络对 LDV 测声系统的系统性能的提升，以及寻找最优的用于 LDV 测声系统的网络结构参数，使用多种类型的网络结构，多种类型的特征提取方式，对从 LDV 测声系统采集回来的信号进行处理分析。于此同时，本文构造了多种传统的语音信号处理手段，对从 LDV 测声系统采集回来的信号处理分析，并与利用 DRNN 网络降噪的方式做对比分析。实验结果证明：利用 DRNN 网络降噪的方式，在多个不同的语音信号质量评价指标下，均比传统的手段得到的处理结果要好。这给 LDV 测声系统又提供了可能的发展方向。

6.2　本文的主要贡献及其创新点

随着神经网络技术的发展和成熟，在语音信号处理领域也取得了诸多成功的应用。因此，也给 LDV 测声系统性能的进一步提升带来了新的契机。本文所做的主要贡献如下：

1.将 DRNN 网络语音降噪的方式应用于 LDV 测声系统：

本文提出利用 DRNN 网络降噪的方式，提升从 LDV 测声系统获取的语音信号的质量。并构建多种结构和参数的网络，寻找最适用于 LDV 测声系统语音降噪的网络结构。

2.综合利用多种降噪方式对 LDV 测声系统所获得的信号进行处理：

本文对从不同条件下 LDV 采集会来的信号做对比分析，将多种语音降噪的手段应用于 LDV 测声系统语音降噪。通过利用多种降噪手段的过程，研究 LDV 测声系统的系统噪声的特性，寻找最适用于 LDV 测声系统语音降噪的方式。

6.3 后续研究工作

虽然本文针对 LDV 测声系统所获取的语音信号中的噪声太大的问题，利用 LDV 测声系统所固有的系统特性，提出了基于 DRNN 网络的语音降噪方式，寻找最优的网络参数，最优的网络结构，并利用该方式所得到的处理结果与传统的语音信号做了对比试验。但是，在的研究过程中，依然有一些问题并未很好的解决，在以下的内容中，对研究过程中所遇到的一些问题，进行总结分析，并为下一步工作的可能的方向提供一些思路。

6.3.1 低信噪比语音信号信息丢失

利用 DRNN 网络去噪的方法，对从 LDV 测声系统采集回来的低信噪比的信号处理时，由于低信噪比的语音信号中的语谱结构不完整，DRNN 网络在处理过程中，很难以将该帧信号中的语音信号识别出来，以至于无法对该帧带噪的语音信号做降噪处理。在更严重的情况下，LDV 测声系统的噪声信号会将原来的语音信号的语谱结构所污染，这样就使得 DRNN 网络会误将一帧噪声信号当作语音信号，进行增强处理。这样会使得原本信噪比很低的语音信号变得更为嘈杂。

6.3.2 振动目标的非空气振动引起的位移

在利用 LDV 测声系统侦测语音信号的过程中，振动目标会随着空气的振动而振动。但在现实的场景中，能引起振动目标振动的因素并不仅仅是单纯的由空气的振动引起的，还有很多其它的因素。比如说，如果振动目标放置的平台发生抖动，势必会引起振动目标的抖动，而且会叠加在由空气振动引起的抖动上。如果非语音信号引起的振动目标抖动的频率，正好是语音信号的频段内，那么将会给测声的结果带来很大的影响。

6.3.3 与传统的语音降噪手段相结合

虽然经过本文的论述，DRNN 网络在 LDV 测声系统语音降噪的场景下，与各种传统的算法相比，都有着明显的优势。但是，传统的语音降噪算法，特别是“学习型”的算法，例如非负矩阵分解，也对从 LDV 测声系统所获取的带噪语音信号有一定的效果，而且算法结构简单，在实际使用过程中易于计算。

所以，设想，如果能将传统的语音降噪算法，和基于 DRNN 网络的语音降噪的算法，针对目标的 LDV 测声系统语音降噪的使用场景，构建一体化的语音

降噪模型，充分结合它们的优势。势必会给 LDV 测声系统语音降噪的研究工作，带来新的突破。