密级 _____

中国科学院大学

University of Chinese Academy of Sciences

博士学位论文

语谱特性和噪声声学环境深度感知的语音增强方法研究

作者姓名 聂帅

指导教师 刘文举 研究员 中科院自动化研究所

学位类别 工学博士

学科专业 模式识别与智能系统

培养单位 中国科学院自动化研究所

2018 年 6 月

Deep Perception on Speech Spectral Characteristics and Noise Acoustic Environments for Speech Enhancement

A dissertation submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in Pattern Recognition and Intelligent Systems

By

Nie Shuai

Supervisor: Professor Liu Wenju

Institute of Automation, Chinese Academy of Sciences

University of Chinese Academy of Sciences

June, 2018

中国科学院大学

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。本人完全意识到本声明的法律结果由本人承担。

作者签名：
日　　期：2018.5.30

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大学有关保存和使用学位论文的规定，即中国科学院大学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：
日　　期：2018.5.30

导师签名：
日　　期：2018.5.30

摘　要

语音是人与机器最自然的交互方式之一，被普遍视为最有可能成为下一代信息和服务的入口。听觉信息处理是人工智能感知的重要组成部分，是目前最接近实用突破的研究方向，然而，在真实环境中，语音信号不可避免地受到噪声和混响的干扰，造成其可懂度和感知质量严重下降。语音增强旨在消除噪声和混响的同时尽可能保持语音质量不受影响，它对语音识别和语音通信等现实应用具有重要的价值，是语音信号处理领域最为关键的核心技术之一和重要研究课题。

由于语音的产生机制，语音信号本身具有明显的时序相关性、自回归性、时空结构和基本发音模式等频谱特性。此外，真实场景中的语音信号具有丰富的噪声声学环境信息。这些特性和信息为我们设计语音增强算法提供了许多有价值的声学线索，对提高语音增强性能具有重要作用。深度学习具有强大的感知能力，近年来，在语音和图像等领域取得了巨大的成功。本文在充分把握语音增强领域的基本理论和前沿方法的基础上，以深度学习为着眼点，瞄准语音固有的频谱特性和噪声声学环境的深度感知进行探索，深入研究了知识驱动(又称模式驱动)下的单声道语音增强算法。主要工作及创新点如下：

1. 通过判断时频单元被语音主导还是被噪声主导，可以实现语音和噪声的分离，自然地，单声道语音分离可以表达为一个二值分类问题。由于语音信号具有明显的时序信息，时间上相邻的时频单元，其被语音或噪声主导的概率具有很强的相关性，因此，前一时刻被语音主导的概率可以作为后一时刻的先验信息。深度层叠网络由若干个基础网络模块堆叠而成，前一个基础网络模块的输出作为先验信息输入到后一个基础网络模块，由于获得了更多的信息，通常后一个基础网络模块的性能会得到进一步提升。我们巧妙地利用深度层叠网络的独特网络结构，按照时间序列将混合语音帧依次输入到层叠的基础网络模块中，提出了带有时序的深度层叠网络（Deep Stack Network，DSN），实现对语音信号中的时序相关性进行有效建模，显著地提升了语音分离的性能。

2. 由于语音信号的短时连续性，语音信号可以表达成一个典型的自回归过程。通过N阶自回归模型，当前语音信号能够通过有限的历史语音信号进行预测。然而，在噪声环境中，语音信号不可避免地会受到噪声干扰，因此利用历史带噪的语音信号很难实现对当前纯净的语音信号的准确预测。但是历史分离的语音信号在一定程度上消除了绝大部分噪声的干扰，有效恢复了语音信号，因此，利用历史的分离语音信号可以实现对当前纯净语

音信号的预测。基于此，我们提出了基于循环网络结构的自回归语音分离网络，实现了语音自回归与分离的联合建模和优化，充分挖掘了语音信号的自回归特性，提高了语音分离的性能。

3. 理想时频掩蔽和目标语音频谱是语音分离中最常用的分离目标，一方面它们具有密切的联系，呈现了显著的相关性和很强的互补性，另一方面，由于语音的产生机制，语音信号具有明显的时频相关性，这导致无论是语音频谱还是时频掩蔽都具有明显的时空结构，而且由于语音的时频稀疏性，这些时空结构对声学环境能够保持相对的稳定性。为了挖掘这些特性以提高语音分离的性能，我们提出了两阶段多目标联合学习的语音分离方法，首先，利用自编码器通过自学习的无监督方式分别挖掘了语音听觉特征和分离目标的时空结构，然后通过线性映射将两个训练好的自编码器连接起来，构建了基于深度神经网络（Deep Neural Network，DNN）的多目标语音分离模型，最后利用多目标联合学习对构建的语音分离模型进行训练。所提出的语音分离方法一方面挖掘了语音输入特征和语音分离目标的时空结构，另一方面充分利用了理想时频掩蔽和目标语音频谱的相关性和互补性，提升了语音分离性能。

4. 语音是由一些基本发音模式产生的，因此，语音信号中隐含着一些基本的频谱结构模式。挖掘语音的频谱结构模式对于提高语音分离的性能具有重要意义。非负矩阵分解（Nonnegative Matrix Factorization，NMF）是著名的表示学习技术，能够有效挖掘语音信号中具有感知意义的基本时空模式，而DNN具有强大的建模能力，能够从混合信号中感知到其中的语音成分和噪声成分。基于DNN的语音分离方法通常直接学习一个从带噪特征到分离目标的映射函数，而忽略了语音的基本频谱结构模式。显然，DNN与NMF的有机联合可能是一个更好的策略。本文我们将NMF的重构生成方式融入到基于DNN的监督式语音分离中，提出了DNN和NMF联合协作的语音分离框架。NMF用来学习语音和噪声基本频谱模式，然后将学习到的基本频谱模式融入到基于DNN的监督式语音分离中直接重构目标语音和噪声的幅度谱。另外，为了进一步避免噪声残留和语音畸变，我们探索了一个带有稀疏约束和NMF重构约束的区分性训练目标。DNN和NMF有机联合的框架即充分利用了NMF对语音时空结构的优异表征能力，充分发挥了DNN超强的映射学习能力，同时避免了类似工作只学习NMF表征系数造成累计误差的缺陷。理论分析和实验结果都证实了该项工作显著地优于之前基于DNN的语音分离方法。

5. 在真实环境中，语音所处的噪声声学环境通常是复杂多变的，比如，噪声的平稳性随时间变化。基于传统信号处理的语音增强通常忽略了真实场景中噪声声学环境的不确定性，假定噪声声学环境是稳定的，采用确定性统

计信号模型解决语音增强问题。深度学习模型具有强大的感知能力，能够感知复杂环境中的语音和噪声声学环境。我们将深度学习强大的感知能力融入到基于信号处理的语音增强框架中，提出了融合信号处理和深度学习的语音增强方案，在这个方案中，深度学习用来感知混合信号中的语音存在概率和噪声声学环境的变化，而传统信号处理框架中的功率谱密度更新模块和维纳滤波模块用来增强最终的期望信号。深度学习和信号处理模块通过频谱近似的目标联合优化，系统的实验证明了所提出的语音增强方法在噪声匹配的条件和噪声不匹配的条件下都能取得较好的语音增强性能。

关键词： 语音增强，语音分离，深度学习，非负矩阵分解，语谱特性，声学环境

Abstract

Speech interaction is one of the most natural ways of human-computer interaction, and is widely regarded as the next major information and service portal. Auditory information processing is a crucial part of AI perception, which is one of the closest practical research directions at present. In real-world environments, the acquired speech signals are inevitably corrupted by various noises and reverberation, which causes the significant degradations of the speech intelligibility and quality. Speech enhancement aims to suppress the noise and reverberation components in the noisy speech while keeping the speech component undistorted. It is one of the key technologies and the most important research topics in the field of speech signal processing.

Due to speech production mechanisms, speech signals have some inherent spectral characteristics, such as temporal correlation, auto-regression, spectro-temporal structure and basic pronunciation pattern. In addition, the acquired speech signals in real-world environments contains rich noise acoustic environment information. These spectral characteristics and acoustic environment information provide a lot of valuable clues for speech enhancement. Deep learning has a powerful perceptual ability and has achieved great successes in the fields of speech and image processing. Based on the basic theories and the advanced research progress of speech enhancement, we focus on the deep learning-based perception on speech spectral characteristics and noise acoustic environments and devotes our efforts to the knowledge-driven single channel speech enhancement. We summarize our works and contributions as follows:

1. The separation of speech and noise can be implemented by judging whether the noisy time-frequency (T-F) unit is dominated by speech or by noise. Naturally, the single channel speech separation can be formulated as a binary classification problem. Due to speech production mechanisms, speech signal contains strong temporal correlation, in other words, there is a strong probability correlation on whether the neighboring T-F units are dominated by speech or by noise. Therefore, the probability that the previous T-F unit is dominated by the speech can be used as a priori probability of the next T-F unit. Deep stacking network (DSN) is stacked by several basic network modules. The output of the previous basic network are used as priori information to feed into the next. The performance of the next basic network

module will usually improves due to the obtained extra prior information. In order to exploit the temporal correlation of speech signals, we proposed a DSN with the time series (DSN-TS). It cleverly uses the unique structure of DSN to implement the effectively modeling of joint probabilities of the neighboring T-F units in time.

2. Speech signals can be described as an auto-regression process. Through an N-order autoregressive model (AR), the current frame of speech signals can be predicted by the limited historical frames of speech signals. Unfortunately, in noisy environments, speech signals are inevitably disturbed by various noises and its auto-regression is severely corrupted, which makes it very difficult to predict the current clean speech signals with the noisy historical speech signals. However, the separated speech largely avoids noise interference, and effectively preserves the harmonic structure of speech. Therefore, it is possible to use the historically separated speech signal to predict the current clean signal through an AR model. In this paper, we proposed a novel auto-regression speech separation network to jointly model and optimize the speech auto-regression and separation processes, which effectively exploits the auto-regression of speech for speech separation.

3. The ideal T-F masks and magnitude spectrums of target speech are the main targets of speech separation. On the one hand, they have a close relationship, and contain significant correlation and strong complementarity. On the other hand, the T-F masks and spectral features present prominent spectro-temporal structures due to speech production mechanisms. In addition, due to the sparsity of speech in the T-F domain, the spectro-temporal structures can keep relatively invariant to various auditory environments, which is very important to robust speech separation. Obviously, these characteristics are very worthy to be exploited for speech separation. In this paper, we propose a two-stage multi-target joint learning speech separation method. Firstly, we use two denoising autoencoders (DAE) to exploit the spectro-temporal structures of speech auditory features and speech separation targets by self-learning, respectively. Then the learned DAEs are combined by a linear transformation to build a multi-target DNN for speech separation. Finally, the multiple speech separation targets are jointly learning. Systematic experiments show that the proposed approach not only exploits the spatio-temporal structure of speech auditory features and speech separation targets but also make full use of the correlation and complementarity of ideal T-F

masks and speech magnitude spectrograms.

4. Deep neural network (DNN)-based speech separation usually uses a supervised algorithm to learn a mapping function from noisy features to separation targets. These separation targets, either ideal masks or magnitude spectrograms, have prominent spectro-temporal structures. Because speech is produced by some basic pronunciation patterns, these spectro-temporal structures contain some basic structure patterns. Nonnegative matrix factorization (NMF) is a well-known representation learning technique that is capable of capturing the basic spectro-temporal structures with physical or perceptual properties. While DNN has a powerful perceptual ability to speech and noise. Therefore, the combination of DNN and NMF as an organic whole is a smart strategy. In this paper, we propose a jointly combinatorial scheme for speech separation. NMF is used to learn the basis spectra that then are integrated into a DNN to directly reconstruct the magnitude spectrograms of speech and noise. Instead of predicting activation coefficients inferred by NMF, DNN directly optimizes an actual separation objective. Moreover, we explore a discriminative training objective with sparsity and reconstruction constraints to suppress noise and preserve more speech components further. The jointly combinatorial scheme of DNN and NMF concentrates the strengths of both DNN and NMF for speech separation. Systematic experiments show that the proposed models are competitive with the previous methods.

5. In real-world environments, the noise acoustic environment is usually complex and varied, for example, the smoothness of noise is varying with time. The signal processing-based speech enhancement usually assumes that the noise acoustic environment is stationary or slowly varying, and uses deterministic statistical signal model for speech enhancement while ignores the uncertainty of noise acoustic environment in real-world environments. Deep learning has a powerful perceptual ability to speech and noise. To address the limitations of conventional signal processing methods, in this paper, we propose a hybrid signal processing/deep learning scheme which incorporates the powerful perceptual capabilities of deep learning in the conventional speech enhancement framework. Deep learning is used to perceive the speech presence probability and the noise acoustic environment, while the signal processing-based power spectral density update module and Wiener filter are used to enhance the desired speech. The deep learning and sig-

nal processing modules are jointly optimized by a spectrum approximation objective. Systematic experiments demonstrate the proposed approach to noise suppression in noise-unmatched and noise-matched conditions.

Keywords: Speech enhancement, Speech separation, Deep learning, Nonnegative matrix factorization，Speech spectral characteristics, Acoustic environments

目　录

表　格

插　图

第一章　绪论

1.1　研究背景和意义

毫无疑问，语音是人与人以及人与机器之间最自然的交互方式之一。在物联网时代，语音被普遍视为最有可能成为下一代信息、内容以及服务的核心入口。目前，在语音通信，智能家居，车载以及智能机器人，可穿戴设备等智能设备上已得到广泛应用。特别是近几年以智能音箱为代表的语音交互产品在国内外的火爆，语音交互领域已汇集了包括谷歌、微软、苹果、阿里、百度、腾讯、京东等在内的几乎所有国内外IT巨头，呈现出百“箱”争鸣的景象[1]，极大地激发了语音交互技术的应用和发展，可以说语音交互的时代即将到来。

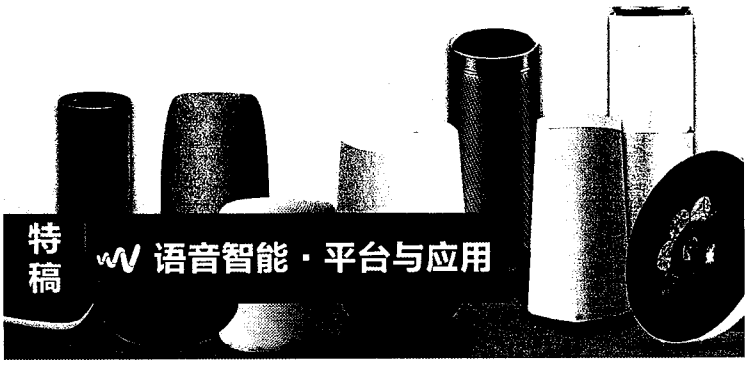


图 1.1: 百“箱”争鸣[1]。

在真实环境中，语音信号不可避免地受到噪声和混响的干扰，造成其可懂度和感知质量严重下降 [94,158]。特别在物联网时代，语音交互的应用场景从近场（拾音设备距离声源较近，通常在半米以内）过渡到远场（拾音设备距离声源较远，通常数米远）。由于声波在传播过程中其能量随传播距离呈指数衰减。在远场条件下，语音信号受到噪声和混响干扰更加严重，极大地影响了语音识别和语音通讯等语音交互应用的性能 [85,107]，真实场景下的远场语音交互依然面临着巨大的挑战。

人类的听觉系统具有显著的听觉信息处理和抗干扰能力，即便在鸡尾酒会这样复杂的噪声环境里，听力正常的人也能轻松地听清楚想要关注的声音，而忽略其他声音 [20,29,35,166,177]，显然人类的听觉系统对接收到的语音信号进行了增强处理，降低了干扰信号的影响。让机器具有类似的听觉信息处理能力是人工智能感知的重要组成部分，是目前最接近实用突破的研究方向，而语音

[1]图片来源:深圳湾，https://www.shenzhenware.com/articles/11698, 2017年7月23日

增强是其中最为关键的核心技术之一。语音增强旨在消除噪声和混响的同时尽可能避免语音畸变，以提高语音的可懂度和感知质量。从人耳的角度来讲，语音增强的目的是提高语音的感知质量，降低干扰对人耳听觉感受的影响，从机器的角度来讲，语音增强的目的是提高语音的可懂度，降低干扰对模型带来的数据不匹配。语音增强是影响语音通讯和语音识别等应用系统性能的关键要素，是语音信号处理领域里最为重要研究课题之一，多年来，受到国内外学者的广泛关注 [1,3-5,86,133,168,169]。

根据语音增强关注的侧重点，语音增强的含义分为以下三个方面 [113,115]：（1）语音降噪，消除或抑制背景噪声；（2）语音分离，从被干扰的混合语音信号中分离或提取目标语音信号；（3）语音降混响，抑制封闭空间里由语音信号反射而造成的混响。其中语音降噪和语音分离是本文的研究重点，为了避免混淆，我们在下文中对语音降噪和语音分离不做严格的区分，统称语音增强。根据语音增强所使用的麦克风个数，语音增强可以分为单通道语音增强和多通道语音增强。单声道语音增强只有信号本身的时域和频域信息可以利用，而多声道语音增强还可以利用空域信息，相对来说，由于可用的信息更少，单通道语音增强更具有挑战性。尽管多声道语音增强的性能一般要优于单通道语音增强，但是单声道语音增强对硬件的要求低，配置灵活，具有更大的市场价值，因此文本的研究重点放在单声道语音增强上。

1.2 语音增强的问题定义

在许多现实应用中，麦克风所处的声学环境中通常包含噪声，混响和回声等，导致麦克风所采集的声音信号除目标语音信号外还伴随着各种干扰信号。

噪声 由目标声源之外的声源发出的信号。根据噪声的平稳性可分为平稳噪声，短时平稳噪声和非平稳噪声，所谓平稳噪声是指噪声功率谱不随时间改变，比如，白噪声、空调噪声等，所谓短时平稳噪声是指噪声在一个较短的时间内相对平稳，比如人声干扰，所谓非平稳噪声是指噪声功率谱随时间快速变化，比如闹铃等瞬发声音。根据噪声的相干性又可分为相干噪声和非相干噪声。所谓相干噪声是指噪声由一个点源模型发出的，麦克风捕捉到的信号带有很强的相关性，所谓非相干噪声是指噪声由多个点源同时发出，不同麦克风因位置不同，所捕捉的信号相关性很低，在频谱上接近平稳，比如咖啡厅噪声。

混响 目标语音信号经过墙壁等障碍平面再次或者多次反射的信号。

回声 智能终端接收目标信号的同时，自己本身也发出声音信号（参考信号）所造成的干扰，比如智能音箱播放歌曲时产生的干扰被麦克风重新录取。

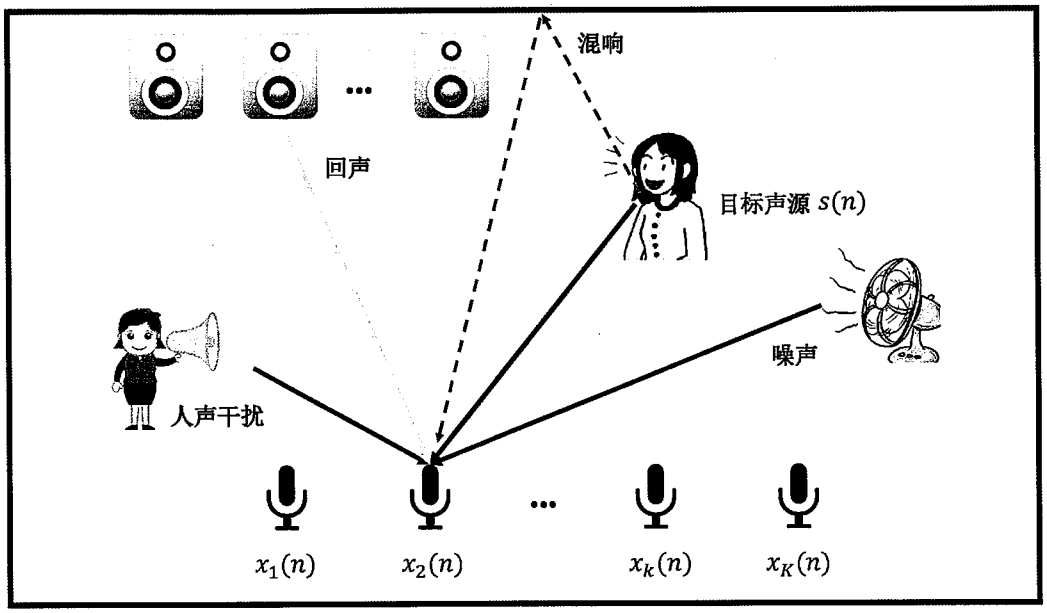


图 1.2: 典型的声学环境。

图1.2展示了一个典型的室内声学采集模型，由一个或多个麦克风组成的采集系统不仅接收到了直射的目标语音信号，还叠加了环境中的风扇噪声，人声干扰和回声信号，同时，由于墙壁等障碍平面的反射，目标语音信号经过多次反射相互叠加所形成的混响信号也不可避免地被麦克风采集到。我们定义$s(n)$和$v(n)$分别为时域目标语音信号和噪声信号，那么在一个封闭房间里，单个麦克风所采集到的观测信号可以用如下的数学形式表达：

$$x(n) = g(n) * s(n) + v(n) = y(n) + v(n), \tag{1.1}$$

其中，$*$定义卷积操作，$g(n)$是通道冲激响应。$y(n) = g(n) * s(n)$是混合信号中无噪的语音信号成分。如果不考虑混响，语音增强算法的任务就是从混合信号$x(n)$中尽可能精确地估计出我们想要的无噪语音信号$y(n)$，而消除噪声$v(n)$，我们定义$\hat{y}(n)$为估计的无噪语音信号。在信号处理领域，通常将时域信号转换到时频域进行处理。常用的时频分解方法包括短时傅里叶变换（Short-Time Fourier Transform，STFT） [55]和gammatone 听觉滤波 [135]。对时域信号进行短时傅里叶变换或gammatone 听觉滤波可以获得其时频表达，我们定义 $x(t,f)$，$y(t,f)$ 和 $v(t,f)$ 分别是时域信号 $x(n)$，$y(n)$ 和 $v(n)$的在第t时间帧、第f个频带的时频表达。假定$x(n)$，$y(n)$ 和 $v(n)$ 都是零均值的随机过程，那么满足以下公式

$$x(f,t) = y(f,t) + v(f,t). \tag{1.2}$$

不管在短时傅里叶变换域还是在gammatone听觉滤波域，都可以通过相应的逆变换方法从 $y(f,t)$ 恢复出 $y(n)$。因此，在时频域，语音增强问题可以转化为 $y(f,t)$ 的估计问题，设定 $\hat{y}(f,t)$ 为 $y(f,t)$ 的估计。由于在实际应用中，只有观

测信号 $x(t,f)$ 是已知的，语音增强问题可以进一步转化为一个滤波问题，即设计一个滤波器 $h(f,t)$ 作用到观测信号 $x(t,f)$ 上恢复语音信号 $y(f,t)$ 而消除噪声信号 $v(n)$。对于单通道的语音增强问题，滤波过程可以表达为如下公式：

$$\hat{y}(f,t) = x(t,f)h(f,t). \tag{1.3}$$

类似地，对于多通道的语音增强，滤波过程可以表达为：

$$\hat{\mathbf{y}}(f,t) = \mathbf{x}^H(f,t)\mathbf{h}(f,t), \tag{1.4}$$

其中 $(.)^H$ 表示共轭转置操作，$\mathbf{x}(f,t)$ 和 $\mathbf{h}(f,t)$ 分别定义为：

$$\mathbf{x}(f,t) = \begin{bmatrix} x_1(f,t) \\ ... \\ x_k(f,t) \end{bmatrix}, \mathbf{h}(f,t) = \begin{bmatrix} h_1(f,t) \\ ... \\ h_k(f,t) \end{bmatrix}, k = 1,2,...,K. \tag{1.5}$$

k为麦克风的序号，K为麦克风的个数。

1.3 语音增强的基本框架

图1.3给出了语音增强的一般框架，主要分为三个模块：1）时频分解，通过短时傅里叶变换或gammtone听觉滤波将一维的时域信号分解成二维的时频信号。2）信息抽取，从一个或多个麦克风信号中提取表征信号特点的信息，包括频谱特征，空间信息和声学环境。3）滤波器设计，根据从麦克风信号中提取的信息，依据一定的优化准则，设计滤波器，使其作用到观测信号上，能够消除干扰信号而恢复目标信号。

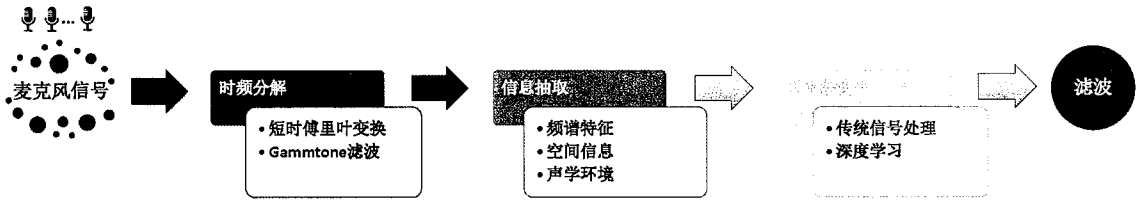


图 1.3: 语音增强的一般框架。

1.3.1 时频分解

通常的语音增强系统都要经过时频分解将时域信号转换成时频域信号，在时频域解决语音增强问题更加简单高效。常用的时频分解方法包括短时傅里叶变换和gammatone听觉滤波。

对于短时傅里叶变换，一维时域信号$y(n)$在时间帧为t、频带为f的短时傅里叶系数可以通过以下公式计算：

$$y(f,t) = \int_{-\infty}^{+\infty} y(n)w(n-t)\exp(-j2\pi fn)dn, \tag{1.6}$$

其中$w(n) = w(-n)$是一个对称的窗函数，j是虚数单位。短时傅里叶变换完备且稳定 [116]，通过短时傅里叶逆变换（Inverse Short-Time Fourier Transform，STFT）可以从时频域信号$y(f,t)$精确恢复时域$y(n)$。因此，只要我们能够从混合信号$x(t,f)$中估计出目标语音信号$\hat{y}(f,t)$，那么就可以实现语音增强。最终目标语音波形可以通过ISTFT计算得到：

$$\hat{y}(n) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \hat{y}(f,t) w(n-t) \exp(j2\pi f n) df dn, \tag{1.7}$$

如果不考虑相位的影响，单声道的语音增强可以进一步表达成目标语音幅度谱的估计问题，估计的目标语音幅度谱$|\hat{y}(f,t)|$ 联合混合信号的相位$\angle x(f,t)$，通过ISTFT即可得到目标语音的波形信号$\hat{y}(n)$。尽管相位信息对于提高语音增强性能具有一定的意义，但是自然语音的相位估计是非常困难的，许多研究表明，使用原始的混合语音相位依然能得到较好的语音增强性能 [171, 192]。

对于gammtone听觉滤波，使用一组听觉滤波器$g(n)$对时域信号进行滤波，得到一组滤波输出$g(n,f)$。滤波器组的冲击响应为:

$$g(f,t) = \begin{cases} t^{l-1} \exp(-2\pi bt) \cos(2\pi ft), & t \geq 0 \\ 0 & else \end{cases}, \tag{1.8}$$

其中，滤波器阶数 $l = 4$，f为滤波器的中心频率，中心频率沿对数频率轴等间隔的分布在[80 Hz, 5 kHz]，b为等效矩形带宽(Equivalent Rectangle Bandwidth，ERB)，等效矩形带宽与中心频率一般满足以下公式：

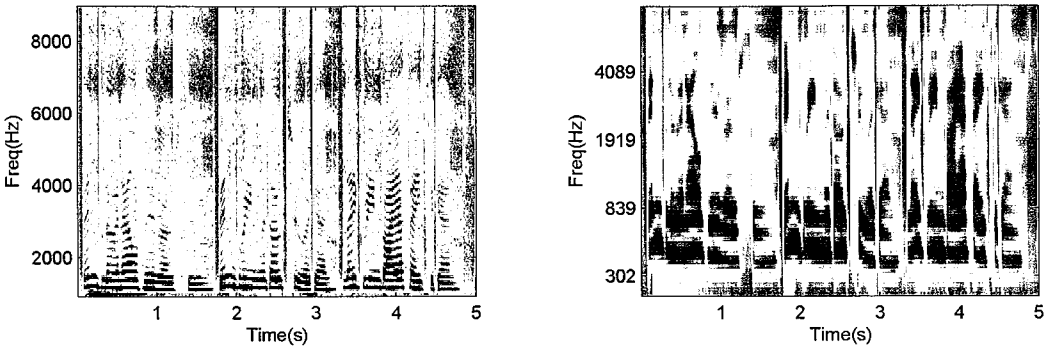
$$ERB(f) = 24.7(0.0043f + 1.0). \tag{1.9}$$

对于4阶的gammatone滤波器，Patterson等人给出了带宽的计算公式 [135].

$$b = 1.093 ERB(f). \tag{1.10}$$

最后对听觉滤波输出按照一定的帧长和帧移进行分帧加窗处理，即可得到输入信号的时频表达。通过计算每一个时频单元信号的听觉能量，就得到了语音信号听觉谱（cochleagram）。

不同于短时傅里叶变换，gammtone听觉滤波采用基于耳蜗基底膜（basilar membrane）模型的听觉滤波器组对时域信号进行时频分解，以模拟人耳对不同频率的感应程度，Gammtone听觉滤波在低频具有更高的分辨率在高频分辨率较低，更加符合人耳的感知机理，而短时傅里叶变换在高频低频采用相同的分辨率。图1.4给出了基于短时傅里叶变换的语谱图和基于gammtone听觉滤波的听觉谱。从图中可以看出，听觉谱在低频具有更高的分辨率。



(a) 语谱图 (b) 听觉谱

图 1.4: 语谱图和听觉谱的对比样例。

1.3.2 信息抽取

真实场景里的语音信号除本身具有能够表征自身特点的频谱特征外，同时还包含有丰富的空间信息，通过多个麦克风组成的麦克风阵列能感知信号的空间信息。此外，在真实的应用场景中，声源所处的声学环境是复杂多变的，声学环境可以为语音增强算法设计提供有价值的信息。

1）频谱特征：根据时频分解的方法不同，频谱特征可以分为短时傅里叶变换域特征和gammtone听觉滤波域特征。短时傅里叶变换域特征包括功率谱密度（Power Spectral Density，PSD）、短时傅里叶变换幅度谱、短时傅里叶对数幅度谱、梅尔倒谱系数（Mel-Frequency Cepstral Coeffient，MFCC）[38]、短时傅里叶滤波器组合（Filterbank，FBank）[146]、振幅调制谱（Amplitude Modulation Spectrogram，AMS）[93]、感知线性预测系数（Perceptual Linear Prediction，PLP）[69]、相对谱变换感知线性预测系数（Relative Spectral Transform PLP，RASTA-PLP）[70]。而gammtone听觉滤波域特征包括gammtone听觉谱（cochleagram），伽马通滤波器组频率倒谱系数（Gammatone Frequency Cepstral Coeffient，GFCC）[147,149]、多分辨率听觉谱（Multi-Resolution Cochleagram，MRCG）以及语音基音特征 [63,83,91,170]。

2）空间信息：包括到达时间延迟（Time Delay of Arrival，TDOA）[97]、广义互相关（Generalized Cross Correlation，GCC）[97]、空间协方差矩阵（Spatial Covariance Matrix，SCM）[188]以及双耳特征 [120,196]。

3）声学环境：由语音源、噪声源以及语音和噪声的传播环境构成了麦克风周围的声学环境，声学环境的先验知识对于语音增强具有重要意义。主要的声学环境包括噪声的统计特性和声场特性（如混响度）等。

1.3.3 滤波器设计

语音增强实际上是一个滤波过程，通过滤波滤除带噪语音信号中的噪声而恢复目标语音信号。滤波器的设计通常根据一定的优化准则，利用从观测

的混合信号中所提取的相关信息设计满足要求的最优滤波器。比如维纳滤波是信噪比意义下的最优滤波器 [26, 152]。除了基于传统信号处理的滤波器设计外 [133]，许多利用深度学习方法进行滤波器设计的方法最近被相续提出 [86, 142, 168, 169, 190, 193]。

1.4 语音增强的研究概况

语音增强旨在降低噪声的同时尽可能保持语音不失真，它对语音识别和语音通信等现实应用具有重要价值，是语音信号处理领域的一个重要研究课题。语音增强伴随着语音信号处理的研究至今已有几十年的历史，已有许多语音增强算法被提出。根据语音增强所使用的麦克风个数，语音增强可以分为单通道语音增强和基于麦克风阵列的多通道语音增强。单声道语音增强只有信号本身的时域和频域信息可以利用，理论上消除噪声的同时不可避免地会带来一定的语音畸变 [11, 12, 26]。相比于单声道语音增强，由多个麦克风组成的麦克风阵列能够提供额外的空间信息，产生空间滤波器，从所需要的语音源方向获取高品质的语音信号，同时抑制其他方向的干扰。大量的理论研究表明，多通道语音增强算法能够在消除噪声的同时有效控制语音畸变，其语音增强效果明显优于单声道语音增强方法 [11]。但是基于麦克风阵列的语音增强算法通常对麦克风阵列硬件要求比较苛刻，既要求组成阵列的每个麦克风具有较高的物理一致性，又要求较高精度的麦克风阵列几何构型，这会导致在实际应用中麦克风阵列的硬件成本急剧上升。另外，许多多通道语音增强算法，对麦克风阵列的几何构型和麦克风间距有一定要求，比如圆形或者线形，这对产品的外形结构提出了一定的要求并大大限制了麦克风在实际产品中的灵活摆放。尽管由于缺乏空间信息，单声道语音增强面临着更大的挑战，但是单声道语音增强对硬件的要求低，配置灵活，使用方便，应用范围广，具有更大的应用价值，因此文本侧重于单声道语音增强的算法研究，所以对语音增强研究概况的介绍会侧重于单声道语音增强，对多通道语音增强也会有所简介，如有遗漏和不足，敬请见谅。

1.4.1 多通道语音增强

目前主流的多通道语音增强算法大致分为自适应波束形成及后滤波法，广义旁辨消除法，多通道维纳滤波法以及基于深度学习的多通道语音增强方法。

1）自适应波束形成及后滤波法：波束形成可以依据目标语音的方位信息设计空域滤波器来增强目标声源方向而抑制来自其他方向的信号。从早期的“延时-加和”波束形成开始，许多固定波束形成或者自适应波束形成算法相续被提出 [11]。固定波束形成只和麦克风阵列的几何形状与目标方位有关，而自适应波束形成可以根据观测信号的统计特性，设计空域滤波器，从而得到更优的语音增强性能。1969年，Capone提出最小方差无失真响应（Minimum Varirance Distortionless Response，MVDR）滤波器 [22, 132]，在输出信号方差

最小的准则下，约束目标方位信号不失真，求解最优滤波器参数。随后Frost提出了线性约束最小方差（Linearly Constrained Minimum Variance，LCMV）滤波器 [13,52]，将MVDR从单约束条件推广到多约束条件 [48]，具有更好的自适应特性，逐渐成为最受关注的自适应波束形成器。虽然波束形成能够有效地提高语音质量，但在低频段，波束形成器的噪声消除效果较差。当混响较强或者处于散射噪声场时，简单地通过空间滤波得到的信号噪声残留较大。为了进一步消除噪声，可以采用单通道后滤波算法。例如，Allen将自适应波束形成方法和维纳滤波相结合 [7]，而Zeliinski提出最小均方（Least Mean-Square，LMS）后滤波方法 [195]。McCowan假设散射噪声场中麦克风的相关性是已知的，设计了相干性噪声情况下的有效后滤波算法 [121]。Lefkimmiatis将McCowan的方法进行了改进，提出了非线性准则的后滤波器设计方法，并可以适用于多种噪声场景 [105]。Cheng将听觉掩蔽效应引入后滤波算法，使后滤波器性能得到进一步的提升 [28]。

2）广义旁瓣消除法：1982年，Griffiths等人提出了广义旁瓣消除（Generalize Sidelobe Canceller，GSC）的空间滤波器框架 [62]，该框架将麦克风阵列语音增强分为三个子模块：固定波束形成、目标信号阻塞、自适应多通道噪声消除。经典的GSC算法所基于的重要假设是目标声源方位已知，所处环境无混响存在，且阵列几何特性已知，各个麦克风的特性相同。然而在实际情况下，由于混响的存在，声源方位的估计变得非常困难，且各个麦克风之间所接收到的信号不符合简单的时间延迟关系。在这种情况下，一方面，固定波束形成阶段不能有效抑制信号中的噪声，另一方面，目标信号阻塞环节不能完全消除目标语音。当目标语音在阻塞环节被泄露后，会引起多通道噪声消除的输出产生语音畸变，且噪声残留较大。因而，混响较强时，GSC的性能会出现显著下降。Nordholm和Bitzer等从理论上分析了不同噪声环境下GSC性能 [16,17,128]。为了解决这些问题，Hoshuyama、Herbordt和Greenberg等提出结合语音活性检测的算法 [61,68,79]，只在语音不存在的时间帧进行多通道自适应噪声消除滤波器的计算和更新，从而避免语音泄露所造成的负面影响。Cox [37]等人通过在自适应噪声消除环节增加鲁棒性约束条件，改善噪声消除的效果。Affes等提出了融合声源-麦克风传递函数的GSC 算法 [6]，Gannot等提出相对传递函数-广义旁瓣消除框架 [57,117]。在文献 [57]中，噪声的平稳性和语音的非平稳性被用于相对传递函数估计，相对传递函数被计算为符合连续多帧上功率谱方程组的最小二乘解。

3）多通道维纳滤波：无论是波束形成还是广义旁瓣消除均基于以下几个假设：a. 目标声源的空间位置已知；b. 麦克风阵列的几何形状已知；c. 麦克风阵列的各个麦克风不存在不匹配的情况，且信噪比一致，文献 [154]表明，麦克风不匹配会造成波束形成和GSC性能的严重下降。为了避免声源方位的估计和解决麦克风不匹配的问题，多通道维纳滤波器被提出。由于多通道维

纳滤波器不假设声源方位和麦克风阵列几何形状，在多噪声源以及散射噪声场中能够表现出比鲁棒性约束的GSC更好的性能 [154]。2002年，Doclo等人提出了基于广义奇异值分解的多通道最优滤波器 [40]，又被称为语音畸变加权多通道维纳滤波器（Speech Distortion Weighted - Multichannel Wiener Filter, SDW-MWF）。和GSC的自适应多通道噪声消除不同，SDW-MWF将语音畸变考虑到滤波器设计的框架内，将信号均方误差表达为噪声消除和语音畸变的折中，采用平衡因子控制噪声消除和语音增强的比例。文献 [41]中进一步提出了SDW-WMF的迭代算法，而 [139]使用矩阵QR分解实现对SDW-MWF的快速计算。随后，SDW-MWF的频域算法被进一步提出 [42]。2010年，Souden等人系统性地分析了SDW-WMF的频域算法 [152]。该工作分析了SDW-WMF的理论性能，并揭示了多通道维纳滤波、MVDR滤波器以及GSC的内在关系。和时域表示不同，在SDW-WMF的频域表示中，可以为每个时频单元选择单独的平衡因子，以更好地协调语音畸变和噪声消除。为了寻求更优的平衡因子，文献 [127] 提出以语音存在概率（Speech Presence Probability, SPP）为基础计算平衡因子，而文献 [156]采用信号直接路径和散射路径能量比（Direct-Diffuse-Ratio）作为平衡因子计算滤波器参数。尽管多通道维纳滤波不需要估计声源方位，且对麦克风阵列几何形状没有任何要求，但其性能严重依赖于噪声时间空间统计特性的估计，对于时间空间非平稳的噪声声学环境，噪声统计特性的估计非常困难，因此多通道维纳滤波在应对非平稳噪声声学环境时依然面临着巨大的挑战。另一方面，由于多通道维纳滤波将麦克风所接收到的语音信号作为参考，因此只能估计带有混响的纯净语音信号，难以处理混响，当混响较强时，会引起语音感知质量和可懂度的降低。

4）基于深度学习的多通道语音增强：传统的多通道语音增强技术通常需要事先估计噪声时间空间统计特性或者目标声源方位。在干扰或混响存在的情况下，目标声源方位的估计非常困难，而噪声统计特性的估计在非平稳声学环境下依然面临着巨大的挑战。除此之外，传统的多通道语音增强技术常常建立在对噪声特性、声场环境和麦阵系统理想的假设条件之上的，而这些假设条件在真实环境中通常又很难满足。基于数据驱动的多通道语音增强对噪声特性，声学环境和麦阵系统不做任何假设，直接从大量的多通道数据，智能感知声学环境，学习符合各种真实情况的多通道语音增强系统。目前基于深度学习的多通道语音增强方法大致可以概括为三类：a. 掩蔽预测的方法，利用深度学习模型估计混合信号被语音或者噪声主导的时频掩蔽，然后用估计的时频掩蔽计算多通道信号中语音或者噪声空间协方差矩阵，最后计算MVDR等多通道空间滤波器的滤波系数，比如 [49, 73, 74, 191]。这类方法一般使用单通道数据估计时频掩蔽，目前空间特征也开始应用到时频掩蔽的估计，如Zhang和May等人利用双耳特征估计时频掩蔽，然后利用估计的时频掩蔽来增强目标语音 [120, 196]；b. 空间滤波器预测法，利用深度学习模型从多通道数据中直接估计空间滤波器系数，

比如MVDR 滤波系数，一般使用从多通道信号中提取的空间特征作为深度学习模型的输入特征，比如广义互相关（GCC）或者空间协方差矩阵（SCM）等空间特征，典型的方法有 [43, 106, 188, 189]；c. 空间滤波器学习法，利用深度学习模型直接对原始的多通道数据进行建模，底部的短时卷积层被用来模拟“延时-加和”操作，实现利用特殊的深度学习的模型自动学习空间滤波的功能，比如 [78, 143, 144]。目前，通过学习系统，能够从多通道数据中自动挖掘空间信息，感知声学环境，并学习符合各种真实情况的多通道语音增强系统, 越来越多研究表明，深度学习在多通道语音增强问题显示了巨大的优势，日益成为多通道语音增强的一个新的研究趋势，正得到世界范围内的研究者的广泛关注。但同时由于过度地依赖于大量监督数据的学习，基于深度学习的多通道语音增强面临着较大的泛化问题，当面临不匹配的数据时，性能难以保证，另外，深度学习的方法通常需要较大的计算量，很难在本地达到实时性的计算要求，大大限制了基于深度学习的多通道语音增强的实际应用。

1.4.2 单通道语音增强

单通道语音增强由于只有一个麦克风，无法获取到声源的方位空间信息，因此只有信号本身的时域和频域信息可以利用。相对于多通道语音增强单声道语音增强面临着更大的挑战，但是由于其成本低、配置灵活，应用范围广等优势具有巨大的应用价值。自语音信号处理研究以来，单通道语音增强受到国内外研究者的广泛关注，单通道语音增强大致可以分为传统信号处理的方法，基于模型的方法，计算听觉场景分析的方法和监督式语音增强的方法。

1）传统信号处理的方法：传统信号处理的单通道语音增强算法可分为时域的方法和频域的方法 [2]。时域方法主要包括时域滤波的方法 [56]和信号子空间的方法 [47, 84]，由于时域的方法计算复杂度较高，语音增强问题通常会转化到频域进行研究，因此，本文对时域的语音增强方法不做过多介绍。谱减法 [19]和维纳滤波法 [26, 111]是语音增强领域最为经典的方法。这些方法通常依赖于噪声功率谱密度的估计，由于语音信号在时域和频域上是稀疏的，因此能够实现噪声功率谱密度的连续估计和追踪。短时滑动平均是最常用方法，在这个方法中，通常根据语音活性检测（VAD）来决定噪声功率谱密度的更新或者保持，利用时频单元的语音存在概率（Speech Presence Probability，SPP）对每一个频带的噪声功率谱密度进行更新是一种更加精细的方式 [45]，这种方式能够获得更加准确的噪声功率谱。然而对于非平稳噪声，语音存在概率估计非常困难，这会导致噪声功率谱密度的估计不准确，噪声功率谱密度的过估计会造成较大的语音失真，而欠估计虽然语音失真较小但同时也会造成较大的噪声残留。传统信号处理的方法通常假设噪声是平稳或者慢变的，难以应对非平稳噪声。为了应对非平稳噪声声学环境，Martin于1994年提出了最小统计量的语音增强方法 [118]，随后，进一步提出了更加平滑的最小统计量的噪声估计

方法 [119]，该类方法显著提高了单通道语音增强算法应对非平稳噪声的能力。Cohen进一步改进了此类方法，提出了最小控制迭代平均（Minima Controlled Recursive Averaging，MCRA）的噪声估计方法 [32,33]，基于MCRA的噪声估计方法具有估计误差小，对平稳噪声追踪比较快的特点。另一类重要的方法是基于最小均方误差的语音幅度谱估计的方法，在1984年被提出 [45]，随后进一步又提出了对数幅度谱的最小均方误差估计方法，该方法考虑了人耳听觉系统对声强的非线性感知 [46]。

2）基于模型的方法：利用纯净的语音和噪声数据分别构建语音模型和噪声模型，然后寻找对混合信号进行建模的最优模型组合，最后，利用选择的语音模型和噪声模型实现语音和噪声的分离。常见的语音和噪声的建模方法包括隐马尔科夫模型（Hidden Markov Model, HMM）和非负矩阵分解模型（Nonnegative Matrix Factorization, NMF）。Ephraim和Sameti基于最大似然准则利用HMM分别对语音数据和噪声数据分别进行建模，获得相应的语音和噪声模型，然后利用贝叶斯估计的方法计算观测信号的状态概率，进而实现语音和噪声的分离 [44,145]。随后，Roweis使用分解隐马尔可夫模型（Fctorial Hidden Markov Model，FHMM）对不同说话人信息进行建模，并通过计算时频掩蔽，通过滤波恢复目标语音信号 [140]。Hershey 等人在 [36]中提出了一种分层FHMM模型(1ayered FHMM)，利用了时间和语法的动态信息，进一步提高了语音增强的性能。NMF是另一类重要的语音和噪声建模方法。NMF的主要思路是将非负数据分解为一组非负基向量的非负线性组合，应用到纯净的语音信号或噪声信号时，能够获得语音或噪声的基本表示，这些基本表示可以认为是组成语音和噪声的基本元素。在测试的时候，通过求解组合混合信号的语音基向量和噪声基向量的权重系数，进而实现语音和噪声的分离。目前，NMF及其变体，比如稀疏的NMF（sparse NMF）、区分性的NMF（discriminative NMF）和卷积的NMF （convolutive NMF）已广泛应用到语音增强中 [101,129,151,182]，并取得较好的语音增强性能。然而非负矩阵分解是一个浅层的线性模型，很难挖掘语音数据中复杂的非线性结构，最近, Le Roux、Hershey和Hsu等将NMF扩展成深层结构并应用到语音分离中, 取得了较大的性能提升 [72,80,100]。基于模型的方法严重依赖于事先训练的语音和噪声模型，对于不匹配的语音和噪声，其性能通常会严重下降。

3）计算听觉场景分析的方法：计算听觉场景分析（Computational Auditory Scene Analysis，CASA）以听觉场景分析（Auditory Scene Analysis，ASA）为机理，试图通过计算机模拟人耳对声音的处理过程来解决语音分离问题 [166]。最早的CASA系统由Weintraub博士提出 [178]，实现了双说话人的语音分离。该系统基于频谱的周期性及时间连续性线索，利用双说话人的基音包络对时频单元进行组织，最后利用二值掩蔽从混合信号中重新合成目标语音时域信号。此后，英国谢菲尔德大学的Cooke利用频谱结构的连续特点以及

相邻时频单元在瞬时频率或者幅度调制率相似性的特点将独立的时频单元合并成同步流片段，然后利用基音作为线索对同步流片段进行组织，最后实现了语音和噪声的分离 [35]。紧随其后，Brown对Cooke系统进行了改进，将共同起止语音点引入到CASA系统的组织过程中 [20]。2004年，Hu和Wang进一步推动了CASA 技术的发展 [82]，提出了基于基音包络和幅度调制率的组织策略，并对低频确定性谐波和高频非确定谐波采用不同的处理策略，该系统显著提高了语音分离的性能，特别是高频段的语音分离。上述的方法能够实现具有谐波结构的浊音段的分离，很难处理没有明显谐波结构的清音部分，Shao等提出基于起止音的分割和基于模型的组织有效地实现了清音部分的分离 [148]。相对于其他的语音增强方法，计算听觉场景分析对噪声没有任何假设，具有更好的泛化性能。然而，计算听觉场景分析依赖于语音线索的检测，比如基频，而在噪声环境里，语音基频的检测面临着巨大的挑战，此外，由于缺乏谐波结构，计算听觉场景分析难以处理语音中的清音成分。

4）监督式单通道语音增强的方法：单声道语音增强旨在从混合信号中恢复出目标语音信号，可以很自然地表达成一个监督式学习问题。典型的监督式语音增强系统通过监督式学习算法学习一个从带噪特征到增强目标的映射函数 [1]。根据语音增强的目标不同，监督式语音增强进一步可以表达为一个二值分类问题或回归问题。Kim将语音分离表达成一个二值分类问题，使用高斯混合模型（Gaussian Mixture Model，GMM）对每一个频带中的语音主导时频单元和噪声主导时频单元进行建模 [94]，然后通过贝叶斯分类，对混合信号的每一个时频单元进行分类，从而实现语音和噪声的分离。随后，Han使用支持向量机（Support Vector Machine，SVM）来对时频单元进行分类 [64,65]，显著提升了语音分离的性能。Wang进一步又提出了深度神经网络-支持向量机（Deep Neural Network-Support Vector Machine，DNN-SVM）的组合模型来实现语音主导和噪声主导时频单元的分类 [175]，该方法使用DNN学习区分性的分类特征，SVM被用来最终的分类。基于分类的语音分离依赖于特征的选取，与此同时，Wang等人完善了基于分类的语音分离特征选择的研究 [170]。以上介绍的方法，都是对每一个频带单独建模，忽略了时频单元之间的时空相关性。目前绝大多数方法都是对整帧或者上下文多帧甚至是对句进行建模。Xu等人使用DNN直接对整帧进行建模，学习从带噪语音的对数幅度谱到目标语音的对数幅度谱的映射函数 [192,193]。由于语音的产生机制，语音信号具有明显的时序相关性，为此，Felix和Huang等人将循环神经网络（Recurrent Neural Network，RNN）或长短时记忆网络（Long Short-Term Memory，LSTM）应用到语音分离中，进一步提升了语音分离的性能 [89,179,181]。随后，能够抓住更多时空信息的卷积神经网络（Convolutional Neural Network，CNN）也被应用到语音分离中 [24,150]。对于监督式语音增强方法除了学习模型之外，输入特征和增强目标也是重要的组成部分。文献 [171]系统分析和比较了各种语音分离目标，

而 [27,168,170]总结和分析了各种语音分离特征。目前，监督式语音增强的方法一方面在朝所谓端到端的方向发展，比如Wang等人将傅里叶逆变换融入到深度学习模型中，直接估计目标语音波形 [174]，而Fu直接用原始波形作为输入特征 [53]。另一方面，监督式语音增强与其他应用目标，比如语音识别的目标，进行联合优化的方法也得到广泛关注 [98,176]。同时，一些新兴的生成式模型也开始应用于语音增强，比如生成式对抗网络（Generative Adversarial Network，GAN） [134]和深度的非负矩阵分解 [72,80,100]。对于监督式的多说话人语音分离目前研究得还不是很多，典型的方法有 [71,98,194]。

1.5 监督式的单声道语音增强

数据驱动的监督式语音增强是一个典型的监督式学习问题，通过大量数据的监督学习，学习模型可以自动发现数据的内在关联和区分模式从而实现语音和噪声的分离。语音和噪声，甚至是不同说话人的语音都存在各自独特的频谱特性，因此，对于单个麦克风采集的混合语音，实现语音和噪声的分离是可能的，人类的听觉系统在这方面表现出了显著的能力。事实上，人耳的语音分离能力也是通过不断学习提升的。这表明更多的学习可以提升语音分离的能力。图1.5给出了监督式的语音增强的一般结构框图，正如一个典型的监督式学习问题，它包括特征，目标和模型三个主要的模块，下面我们依次对这三个模块进行总结和介绍。

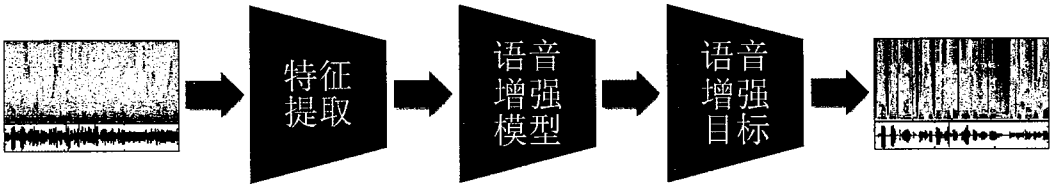


图 1.5: 监督式语音增强系统的结构框图。

1.5.1 特征

作为一个监督式学习问题，特征提取至关重要，是影响语音分离性能的重要因素。语音信号是宽带信号，将其分解到不同频率的子带往往是特征提取的第一步，短时傅里叶变换和gammtone听觉滤波是最常用的时频分解技术。根据特征提取的基本单位不同，语音分离特征可以分为时频单元级别的特征和帧级别的特征。时频单元级别的特征是从一个时频单元里的子带信号中提取的，这种特征粒度更细，更加关注微小的细节, 但缺乏对语音的全局性和整体性的描述，很难表征可感知的语音特性(例如, 音素) [1]。早期基于分类的监督式语音分离系统通常使用时频单元的特征 [64,65,94,175]。帧级别的特征是从一帧信号中提取的，这种特征粒度更大，关注语音的全局性和整体性，能够抓住语音的

时空结构，是目前监督式语音分离最常用的特征。正如章节1.3.2所介绍的，许多频谱特征被提出，并应用到监督式语音增强中，不同的特征既有互补性又有冗余性，针对各个特征之间的不同特性，Wang等人利用拉索（Lasso）方式系统分析各个特征，提出了最优的特征组合：梅尔倒谱系数、相对谱变换感知线性预测系数、振幅调制谱和语音基音特征 [170]，这个组合特征是早期监督式语音分离系统最常用的特征。后来Chen等人提出多分辨率的MRCG 特征，在低信噪比条件下取得了明显的性能优势 [27]，因此逐渐取代了先前的组合特征成为监督式语音分离最常用的特征之一。随着语音分离建模能力的提升，相对原始的短时傅里叶幅度谱及其对数能量谱越来越多地用到监督式语音分离中，并且取得良好的语音分离性能，逐渐成为更加主流的语音分离特征。为了利用信号中的短时变化特征，一阶和二阶差分信息常被计算并拼接到输入特征中。同时，为了抓住更多的上下文信息，输入特征通常会扩展上下文帧。Chen等还提出使用ARMA模型(Auto-Regressive and Moving Average Model) 对特征进行平滑处理，来进一步提高语音分离性能 [27]。

1.5.2 目标

语音增强的应用主要包括2个方面：一是以人耳作为目标受体，提高人耳对带噪语音的可懂度和感知质量；二是以机器为目标受体，提高机器对带噪语音的感知性能，比如提高噪声环境下语音识别的准确率。语音增强的这两个主要目标既存在联系又存在区别。以提高带噪语音的可懂度和感知质量为目标的语音增强系统通常可以作为噪声环境下语音识别的前端处理模块，能够显著地提高语音识别的性能 [125]，Felix指出语音分离系统的信号失真比(Signal-to-Distortion Ratio, SDR)和语音识别的字错误率(Word Error Rate, WER)有明显的相关性 [179]。但以提高语音的可懂度和感知质量为目标的语音分离系统侧重于去除混合语音中的噪声成分，往往会导致比较严重的语音畸变，而以提高语音识别准确率为目标的语音分离系统更多地关注语音成分，在语音分离过程中尽可能保留语音成分，避免语音畸变 [1]。针对语音增强的不同应用目的，许多具体的语音增强目标被提出，大致可以分为三类：时频掩蔽、频谱映射和隐式时频掩蔽。其中时频掩蔽和频谱映射的目标被证明能显著地抑制噪声，提高语音的可懂度和感知质量 [171]。而隐式时频掩蔽通常将时频掩蔽技术融入到具体的实际应用模型中，时频掩蔽作为中间处理过程来提高其他目标的性能，比如语音识别 [126,179]，目标语音波形的估计 [174]。

A 时频掩蔽

时频掩蔽是单声道语音增强最常用的目标，其中理想二值掩蔽（Ideal Binary Mask, IBM）和理想浮值掩蔽（Ideal Ratio Mask, IRM）是最基本的时频掩蔽目标，它们能显著地提高增强语音的可懂度和感知质量。理想二值掩蔽

和理想浮值掩蔽都是实数语音增强目标，并没有考虑相位对语音增强的影响，最近的一些研究显示相位信息对于提高语音的感知质量具有重要的作用 [131]，为此，考虑相位的时频掩蔽目标被提出，比如复数域的浮值掩蔽(Complex Ideal Ratio Mask, cIRM) [185]。另外，理想二值掩蔽和理想浮值掩蔽都假设语音和噪声相互独立，真实场景很难满足语音和噪声的独立性假设，为此，Liang等提出了信噪比意义下的最优浮值掩蔽（Optimal Ratio Time-Frequency Mask, ORM）[109]，下面对时频掩蔽目标进行简单的介绍。

（1）理想二值掩蔽（IBM）

IBM是计算听觉场景分析的主要计算目标 [165], 已经被证明能够极大地提高分离语音的可懂度 [67,108,124,167]。IBM是一个“0”和“1”的二值掩蔽矩阵，“1”表示被语音主导，“0”表示被噪声主导。可以根据时频单元的局部信噪比计算得到，如下所示：

$$IBM(f,t)=\begin{cases}1, & \text{if } SNR(f,t)>LC\\ 0, & \text{otherwise,}\end{cases}, \tag{1.11}$$

其中, $SNR(f,t)$定义了频带为f时间帧为t的时频单元的局部信噪比。LC是局部阈值(Local Criterion, LC)，它的选择对语音的可懂度具有重大的影响 [96], 一般为了保留足够多的语音信息，设置LC的值小于混合语音信噪比5dB。

（2）理想浮值掩蔽（IRM）

IRM定义如下:

$$IRM(f,t)=\left(\frac{|y(f,t)|^2}{|v(f,t)|^2+|v(f,t)|^2}\right)^{\beta}, \tag{1.12}$$

$|\cdot|$表示复数域或者实数域的绝对值操作，$|y(f,t)|^2$和$|v(f,t)|^2$ 分别表示纯净语音和噪声在频带为f时间帧为t的时频单元的能量。β是一个可调节的尺度因子。IRM在形式上和维纳滤波密切相关，大量的实验表明$\beta=0.5$是比较好的选择 [171]。

（3）最优浮值掩蔽(ORM)

IRM在语音和噪声独立的条件下, 能够取得最小均方误差意义下最大信噪比增益 [109,110]。真实环境里语音和噪声通常存在一定的相关性，针对这个问题，Liang等推导出一般意义下的最小均方误差的最优浮值掩蔽, 其定义如下:

$$ORM(f,t)=\frac{|y(f,t)|^2+\Re(y(f,t)v^*(f,t))}{|y(f,t)|^2+|v(f,t)|^2+2\Re(y(f,t)v^*(f,t))}, \tag{1.13}$$

其中$\Re(\cdot)$表示取复数的实部，*表示共轭操作。相对于IRM，ORM考虑了语音和噪声的相关性，但其变化范围更大，估计难度更大。

（4）复数域的理想浮值掩蔽(cIRM)

IRM没有考虑相位信息，定义在复数域cIRM考虑了相位信息。其目标是通过将cIRM作用到混合语音的STFT系数$x(f,t)$得到目标语音的STFT系数$y(f,t)$。具体如下：

$$y(f,t) = cIRM(f,t) * x^*(f,t), \tag{1.14}$$

其中‘$*$’定义复数乘法操作，通过数学推导我们能计算得到:

$$cIRM(f,t) = \frac{x_r(f,t)y_r(f,t) + x_i(f,t)y_i(f,t)}{x_r^2(f,t) + x_i^2(f,t)} + j\frac{x_r(f,t)y_i(f,t) - x_i(f,t)y_r(f,t)}{x_r^2(f,t) + x_i^2(f,t)}, \tag{1.15}$$

其中$x_r(f,t)$和$y_r(f,t)$分别是$x(f,t)$和$y(f,t)$的实部，$x_i(f,t)$和$y_i(f,t)$分别是$x(f,t)$和$y(f,t)$的虚部，j是虚数单位。

B 频谱映射

如果不考虑相位的影响，单声道语音增强可以转化为目标语音幅度谱估计的问题，因此，监督式语音增强就可以转化为频谱映射的学习问题。根据不同的时频分解技术，频谱映射可分为短时傅里叶变换域幅度谱映射和gammtone听觉滤波域的幅度谱映射。

（1）Gammtone幅度谱映射

时域信号经过gammtone听觉滤波，然后经过分帧加窗处理，可以计算出二维时频表示cochlegram。从混合信号的cochlegram中可以估计出目标语音的cochlegram。由于从cochlegram没有直接的逆变换方法得到目标语音的波形。我们可以通过估计的目标语音cochlegram和原始的混合语音cochlegram构造时频掩蔽，然后再使用gammtone逆变换得到目标语音的波形信号。

（2）傅里叶幅度谱映射

从混合信号中可以估计目标语音的傅里叶幅度谱，然后联合原始混合信号的相位，通过逆变换可以得到目标语音的时域信号。由于语音的能量主要集中在低频，高频段和低频段的傅里叶幅度差异较大，在实际应用中，一般通过对数操作来压制傅里叶幅度谱，从而减小高频段和低频段的幅度差异，因此，傅里叶对数幅度谱常被用来作为监督式语音增强的目标。

C 隐式时频掩蔽

时频掩蔽和频谱映射都是语音计算过程中的中间目标，并没有直接优化语音增强实际应用的目标，为此，隐式时频掩蔽被提取 [89,126,174]，在这些方法中，时频掩蔽并没有被用来作为计算目标，而是作为一个确定性的计算过程融入到具体应用模型中来直接优化实际应用的目标，比如语音识别的准确率或者目标语音波形。Huang等将时频掩蔽操作融入到目标语音的幅度谱估计中，在文献 [89] 中，深度神经网络作为语音分离模型，时频掩蔽函数作为额外的处

理层加入到网络的原始输出层，通过时频掩蔽，目标语音的幅度谱从混合语音的幅度谱中估计出来，然后用来计算误差，优化深度神经网络参数。Wang等人进一步将时频掩蔽融合到目标语音波形的估计中，在文献 [174]中，时频掩蔽作为神经网络的一部分，掩蔽函数从混合语音的STFT幅度谱估计目标语音的STFT幅度谱，然后通过ISTFT，利用混合语音的相位信息和估计的STFT幅度谱合成目标语音的时域波形，估计的时域波形与目标波形计算误差，最后通过反向传播更新网络权重。Narayanan等人提出将时频掩蔽融入到语音识别的声学模型中 [126]，时频掩蔽作为神经网络的中间处理层，从带噪的梅尔谱特征中掩蔽出目标语音的梅尔谱特征，然后输入到下层网络中进行声学状态概率估计。注意时频掩蔽仅仅是神经网络的中间处理层的输出，并不是以理想时频掩蔽作为目标学习而来的，确切地说是根据语音识别的误差学习而来的，实验结果显示，时频掩蔽输出具有明显的降噪效果，这从侧面显示了语音识别与语音分离之间存在密切联系 [126]。

1.5.3 模型

作为一个监督式学习问题，学习模型是影响监督式语音增强的关键因素。常用的学习模型可以概括为两类：浅层模型和深层模型。早期的监督式语音增强通常使用浅层模型对时频单元进行分类，比如GMM [94]和SVM [63]。NMF也是重要的浅层模型，也被广泛地应用到语音增强中 [123]。然而混合语音信号中蕴含着复杂的非线性关系，浅层模型表达能力有限，很难从复杂的数据中学习有用的特征表示，另外，为了挖掘语音的时频相关性，必须对更大尺度的整帧甚至多个上下文帧进行建模，浅层模型难以处理复杂的高维数据。深层模型是近几年来受到极大关注的学习模型，在语音和图像等领域都取得了巨大的成功。由于深层模型的层次化的非线性处理结构, 使得它能够自动抽取输入数据中对目标最有力的特征表示，相比于浅层模型，深层模型能够处理更原始的高维数据，对特征设计的知识要求相对较低，而且深层模型擅长于挖掘数据中的时空结构。由于语音的产生机制，语音分离的输入特征和输出目标都呈现了明显的时空结构，这些特性非常适合用深层模型来进行建模。近年来，许多深层模型被应用到监督式语音增强中, 包括DNN [192, 193]，CNN [24, 150], RNN [89, 181]和LSTM [179] 等。

1.6 论文研究内容

对于单声道语音增强，由于只有一个麦克风，可用的信息只有语音和噪声的固有特性，其中语音频谱特性和噪声统计特性是单声道语音增强的基本信息。一方面，不同声源，比如语音和噪声，必然存在一定的频谱差异，这些差异造成人耳对不同声音感受不同，得益于人类听觉系统强大的感知能力，人耳能轻易区分混合信号中的不同声源成分。同理，正是不同声源的频谱差异使得监督

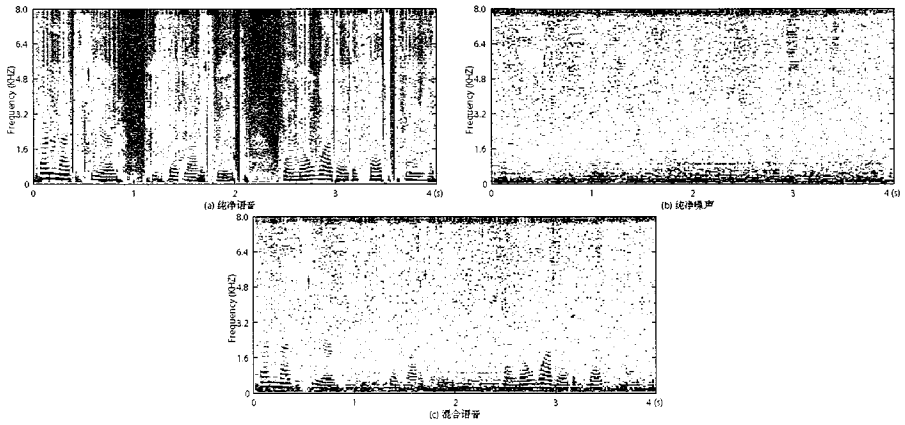


图 1.6: 纯净语音、噪声和混合语音的语谱图。

式语音分离成为可能。深度学习具有强大的感知能力，监督式语音增强通常利用深度学习模型学习一个从带噪特征到分离目标的映射函数，这种简单地将语音增强表达为一个监督式学习问题，过度地依赖于训练数据的粗暴训练，而忽略了语音本身的固有频谱特性，在面临不匹配的声学环境时，比如训练集中不存在的噪声类型和输入信噪比，性能会严重下降。事实上，由于语音的产生机制和语言学的限制，语音信号相对于其他噪声信号具有明显的频谱特性，图1.6展示了一段纯净语音和噪声的语谱图，我们可以发现，语音信号呈现了显著的时序相关性、自回归性、时空结构和基本发音模式等频谱特性，有效挖掘这些频谱特性不仅能够提高语音分离的性能还可以提高监督式语音增强模型的泛化能力。另一方面，相对于非语音的噪声信号，语音信号在时间和频率上具有明显的稀疏性，如图1.6所示，这种统计特性的差异性为传统语音增强算法的设计提供有力的依据。传统语音增强算法往往依赖于噪声统计特性的估计，假设噪声声学环境是平稳慢变的，利用传统信号处理的方法实现对噪声连续追踪和统计估计，而在真实环境中，语音所处的噪声声学环境通常是复杂多变的，基音信号处理的语音增强方法性能通常会严重下降，难以应付非平稳噪声声学环境。深度学习模型具有强大的感知能力，能够感知复杂环境中的语音和噪声声学环境的变化，深度学习和信号处理有机融合的框架既利用了深度学习对噪声声学环境和语音的强大感知能力，又充分发挥了信号处理的专家知识，能够显著提升语音增强的性能和实用性。本论文以深度学习为着眼点，瞄准语音固有频谱特性和噪声声学环境的深度感知进行探索，深入研究了知识驱动(又称模式驱动)下的单声道语音增强算法。主要内容包括：

（1）基于时序深度层叠网络的语音分离方法

语音信号具有明显的时序信息，时间上相邻的语音信号表现出很强的相关性。语音分离通常被表达成一个二值分类问题，即语音主导或噪声主导，时间上相邻的时频单元被语音或噪声主导的概率具有明显的相关性，前一时刻的时频单元被语音主导的概率可以作为后一时刻的先验信息。深度层叠网络是由若

干个简单基础网络堆叠而成的，前一个基础网络的输出作为先验信息输入到后一个基础网络，由于获得了更多的信息，通常后一个基础网络的性能会优于前一个基础网络。我们巧妙地利用深度层叠网络的独特结构，按照时间序列帧将混合语音依次输入到层叠的基础网络中，提出了带有时序的深度层叠网络，有效地对语音信号中的时序相关性进行建模，显著地提高了语音分离的质量和泛化性能。

（2）基于循环网络结构的语音分离与自回归联合优化的语音分离方法

由于语音的短时连续性，语音信号可以表达成一个典型的自回归过程，即通过N阶自回归模型，当前语音信号能够通过有限的历史语音信号进行预测。然而，在噪声环境中，语音信号不可避免地会受到噪声干扰，语音的自回归性受到严重破坏，很难利用带噪的历史语音信号预测当前的纯净语音信号。但历史分离的语音信号在一定程度上消除了绝大部分噪声的干扰，有效恢复了语音信号，因此，利用历史分离的语音信号可以实现对当前的纯净语音信号的预测。基于此，我们提出了一种基于循环网络结构的自回归语音分离网络，实现了语音分离和语音自回归的联合建模和优化。自回归语音分离网络由语音分离网络和语音自回归网络构成，语音分离网络的输出作为语音自回归网络的输入来预测下一帧的纯净语音信号，预测的下一帧的纯净语音联合混合语音输入到语音分离网络，来分离出当前的语音。语音分离网络和语音自回归网络相互协调，联合建模和优化，有效挖掘了语音的自回归特性，提高了语音分离的性能。

（3）两阶段多目标联合学习的语音分离方法

语音的分离目标，无论是理想时频掩蔽还是目标语音频谱，都具有明显的时空结构，此外，理想时频掩蔽和目标语音频谱具有密切的联系，呈现了显著的相关性和很强的互补性。为了有效挖掘语音分离目标的时空结构，充分利用理想时频掩蔽和目标语音频谱之间相关性和互补性，我们提出了两阶段多目标联合学习的语音分离方法。首先，我们利用自编码器通过自学习的无监督方式分别挖掘了语音输入特征和分离目标的时空结构，然后通过线性映射将两个训练好的自编码器连接起来，构建基于DNN的多目标语音分离模型，最后利用多目标联合学习对构建的语音分离模型进行训练。一方面通过自编码器挖掘了语音输入特征和语音分离目标的时空结构，另一方面通过多目标联合学习充分利用了理想时频掩蔽和目标语音频谱的相关性和互补性。

（4）基于DNN与NMF组合框架的语音分离方法

语音是由基本的发音模式产生的，因此，语音信号中隐含着一些固有的基本频谱模式。NMF是著名的表示学习技术，能够有效挖掘语音信号中的具有感知意义的基本频谱模式。我们将NMF的重构生成方式融入到基于深度神经网络（DNN）的监督式语音分离中，巧妙地将判别式和生成式过程融合在一起，提出了DNN和NMF联合协作的语音分离方法。该方法即充分利用了NMF对语音基本频谱模式的挖掘能力，又充分发挥了DNN强大的非线性映射学习能力，同时

避免了类似工作只学习NMF表征系数而造成累计误差的缺陷，理论分析和实验结果都证实了该项工作显著地优于之前基于DNN的语音分离方法。

（5）噪声声学环境深度感知的语音增强方法

在真实环境中，语音所处的噪声声学环境通常是复杂多变的，比如，噪声的平稳性是随时间变化的。基于传统信号处理的语音增强通常忽略了真实场景中噪声声学环境的不确定性，假定噪声声学环境是稳定的，采用确定性统计信号模型求解最优滤波器。而深度学习模型具有强大的感知能力，能够感知复杂环境中的语音和噪声。我们将深度学习模型的感知能力融入到基于信号处理的语音增强框架中，利用深度学习模型感知混合信号中的语音存在概率和噪声声学环境的变化，提出了融合信号处理知识和深度学习的深度噪声追踪网络，既利用了深度学习对噪声声学环境和语音的强大感知能力，又充分发挥了信号处理的专家知识，显著提升语音增强的性能和实用性。

1.7 论文结构安排

本论文结构安排如下：

第一章介绍论文的研究背景和意义、对语音增强的定义和基本框架进行了概述，并在此基础上介绍了语音增强的研究现状，最后从特征、目标和模型三个方面对本文的研究重点监督式的单声道语音增强进行了介绍，通过对监督式的单声道语音增强进行总结和分析，提出了本文的研究内容。本章是后续各章的基础，力图使读者对语音增强和本文的出发点有一个基本认识。

第二章首先介绍了基于时序深度层叠网络的语音分离方法，重点从网络结构、优化方法和实验三个方面对所提出方法进行阐述，并根据实验结果对该方法进行了分析和评估。

第三章介绍了基于循环网络结构的语音分离与自回归联合建模和优化的语音分离方法。首先解释了语音的自回归特性，然后提出了循环网络结构的自回归语音分离网络，最后给出了网络的联合优化目标。实验结果验证了该方法的有效性。

第四章介绍了两阶段多目标联合学习的语音分离方法，首先对语音分离目标和语音听觉特征的时空结构进行了解释，并指出不同语音分离目标之间的联系和互补性，然后提出了两阶段多目标联合学习的方法，并对两阶段的学习过程进行了介绍，最后在实验部分系统分析和评估了所提出的语音分离方法。

第五章介绍了基于DNN与NMF组合框架的语音分离方法，首先从信号模型的定义出发介绍了基于NMF的语音分离，然后系统地介绍四种不同的NMF模型，最后提出了DNN和NMF联合协作的语音分离框架，并从各个方面对模型进行了分析和评估。

第六章介绍了噪声声学环境深度感知的语音增强方法，从信号模型出发介绍了基于传统信号处理的语音增强方法的一般框架，然后从语音增强的一般框

架出发将噪声声学环境和语音存在的深度感知融入到传统语音增强的框架中，系统的实验显示深度学习和信号处理融合的方法提高了语音增强的性能。

第七章总结全文，对全文所做的工作进行全面的总结，并对未来的工作进行了分析和展望。

第二章　基于时序深度层叠网络的语音分离方法

2.1　引言

从混合语音信号中分离出我们感兴趣的目标语音信号，有许多重要的现实应用价值，比如助听器设计和鲁棒性语音识别。然而，在真实声学环境中，当前语音分离系统的性能远没达到令人满意程度，特别对于单声道语音分离，由于缺乏空间信息，语音分离更加困难。人耳听觉系统具有显著的语音分离能力，即便在鸡尾酒会这样复杂的噪声环境下，听力正常的人也能轻松地听清楚感兴趣的目标语音 [20, 29, 35, 166, 177]。受启发于人类的听觉感知，计算听觉场景分析（CASA）试图模拟人耳的听觉处理过程来解决语音分离问题。相比于传统的语音增强方法，比如，谱减法 [19]，维纳滤波法 [26] 和基于模型的方法 [130, 140]，CASA对干扰噪声没有任何假设，有更大潜力解决语音分离问题。然而，CASA严重依赖于语音基频和共同起始的可靠检测，在噪声环境下，语音基频和共同起始的检测非常困难，此外，由于缺乏谐波结构，CASA难以处理语音信号的清音部分。

基于人耳的听觉掩蔽效应，CASA将理想二值掩蔽（Ideal Binary Mask，IBM）作为自己的计算目标，许多研究证明IBM能够显著提高带噪语音的可懂度 [94, 167]。IBM是时频单元的二值矩阵，当时频单元的局部信噪比（Signal-to-Noise Ratio，SNR）大于某个指定的阈值时，该时频单元即被语音主导，对应的IBM元素为“1”，否则被噪声主导，对应的IBM元素为“0”。

从CASA的角度来看，语音分离可以转化为IBM的估计问题，即判断混合语音的时频单元是被语音主导还是被噪声主导。自然而然地，IBM的估计问题可以进一步转换成为一个二值分类问题，许多研究表明将语音分离表达为一个二值分类问题是一个有效的方案 [65, 94, 172, 175]。

由于语音的产生机制，语音信号具有时序相关信息，这些信息能够为语音分离提供丰富的线索。然而，许多基于分类的语音分离方法并没有专门地对语音信号的时序相关信息进行建模，仅仅将语音分离表达为二值分类问题。Kim等人利用高斯混合模型（Gaussian Mixture Model, GMM）分别对语音主导的时频单元和噪声主导的时频单元进行建模，然后利用贝叶斯分类器对混合语音时频单元进行分类 [94]。Han等人构建支持向量机（Support Vector Machine, SVM）分类器实现时频单元的二值分类 [65]。尽管这些方法通过差分或上下文特征考虑了相邻帧之间的相关信息，但这样的方法过于简单粗暴，不能有效挖掘语音信号中的时序相关信息。一方面拼接差分或上下文特征会增大输入向量的维度，从而造成建模难度的急剧增大，而且这种方式挖掘语音信号中的时序相关信息的能力非常有限，另一方面这些方法通常假定相邻时频单元是相互独

立的，对每个时频单元单独进行分类。

深度层叠网络（Deep Stacking Network, DSN）是由若干个简单基础网络模块堆叠而成的，前一个基础网络的输出作为先验信息输入到后一个基础网络，由于后一个基础网络获得了前一个基础网络提供的先验信息，通常相对于前一个基础网络性能会提高。我们利用深度层叠网络的独特结构，按照时间序列将混合语音特征按帧依次输入到层叠的基础网络，提出了带有时序的深度层叠网络（DSN with the time series，DSN-TS），实现了对语音信号中的时序相关性进行有效建模。为了进一步提升语音分离的性能，我们基于最大化命中率减误报率指标（Hit minus False Alarm Rates，HIT-FA）来动态调整分类阈值，实现最优的二值分类。实验结果表明所提出的语音分离方法显著优于基于DNN和DNN-SVM的语音分离方法。

针对语音信号的时序相关性，本章利用了深度层叠网络独特的网络结构，提出了带有时序的深度层叠网络，实现了对语音时序相关性的有效建模。我们首先介绍了深度层叠网络独特的网络结构，然后介绍了所提出的时序深度层叠网络，并将其应用到语音分离中。为了实现语音和噪声的有效分离，提出了基于最大化HIT-FA的分类阈值策略，同时介绍了语音分离所使用的听觉特征。最后对所提出的语音分离方法进行了系统的实验验证。

2.2 算法设计

2.2.1 时序深度层叠网络

所提出的DSN-TS是DSN的一个变体。DSN由若干个基础网络堆叠而成。基础网络由输入层，sigmoid隐含层和线性输出层组成，如图2.1-(a)所示。我们定义基础网络的上层连接权重矩阵为$\mathbf{U}$，下层连接权重矩阵为$\mathbf{W}$。给定权重$\mathbf{W}$，线性输出层使得上层连接权重$\mathbf{U}$可以通过一个闭式解计算得到。而下层连接权重矩阵$\mathbf{W}$可以通过随机梯度下降（Stochastic Gradient Descent，SGD）迭代更新得到。详细的优化算法可以参考文献 [39]。

DSN网络结构的独特之处在于堆叠的基础网络，其输入不仅包括原始的输入特征，还包括上一个基础网络的输出。这意味着高层基础网络的输出不仅依赖于原始的输入特征还依赖于低层基础网络的输出，也就是说，低层基础网络向高层基础网络提供了额外信息，能够帮助高层基础网络的分类。我们可以把DSN视为是循环神经网络（Recurrent Neural Network，RNN）的有限展开，如图2.2所示。这个循环网络能够对输入输出的联合概率分布$p(in, out)$建模。注意这个循环网络的输入始终是一样的，因此DSN并不能对时间上相邻的时频单元的概率相关性进行建模。为了克服这个缺点，我们通过引入时间序列扩展了DSN，得到了DSN的一个变体，称为带有时序的DSN（DSN-TS）。

假定时间序列$T = \cdots, (in_{k-1}, out_{k-1}), (in_k, out_k), (in_{k+1}, out_{k+1}), \cdots$，其中$(in_k, out_k)$对应于基础网络的输入和输出。对于语音分离，时间上相邻的两个

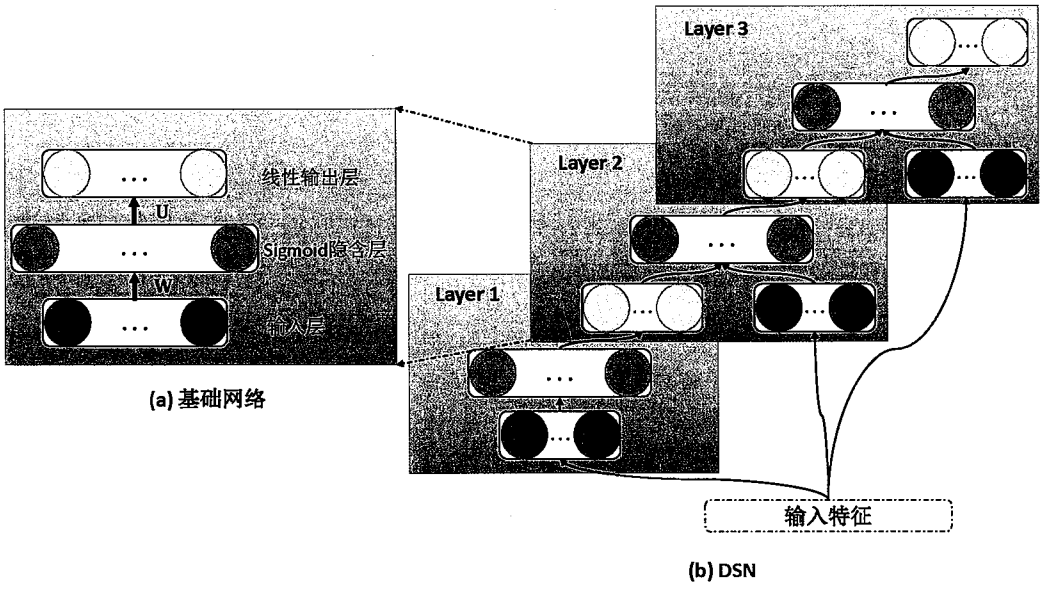


图 2.1: DSN的网络结构。

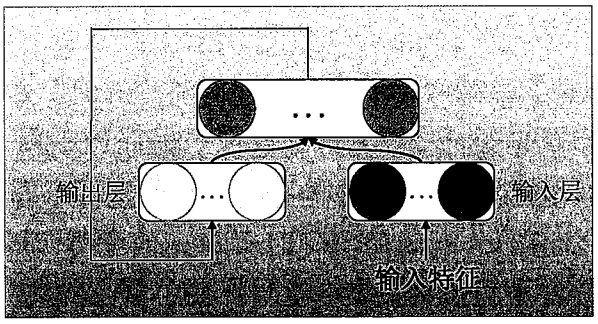


图 2.2: DSN作为RNN的有限展开示意图。

时频单元被语音或者被噪声主导的概率具有明显的相关性。类似马尔科夫模型，我们使用联合概率$p(out_{k-1}, in_k)$对时间序列T中相邻两项的相关性进行建模。类似地，DSN-TS能够像图2.2的循环网络结构一样对这个联合概率分布进行建模。相比于DSN，DSN-TS层叠的基础网络是按时间序列相互连接的，也就是说，基础网络输入的是当前帧的原始输入特征和上一时刻的概率输出，如图2.3所示。

本章我们使用DSN-TS作为我们的语音分离模型，DSN-TS对每一个频带的时频单元单独建模。对于语音分离，时间序列的元素(in_k, out_k)表示第k个时频单元的听觉特征和对应的IBM估计，图2.3说明了DSN-TS在语音分离任务中的应用。

对于深度神经网络，网络权重的初始化至关重要 [75]。受限的玻尔兹曼机（Restricted Boltzmann Machine，RBM）被广泛用来初始化DNN的网络权重。本章我们通过RBM预训练的方式来初始化基础网络的低层权重矩阵$\mathbf{W}$。

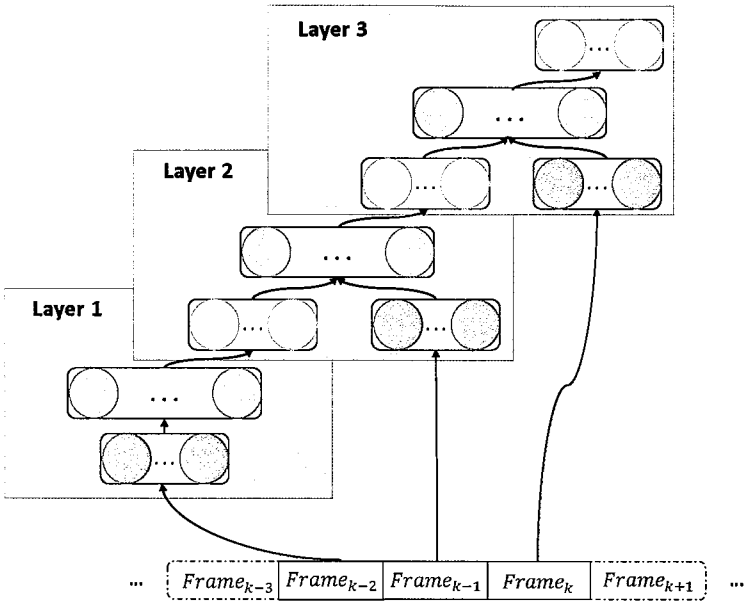


图 2.3: DSN-TS的网络结构。

2.2.2 基于HIT-FA的分类决策

因为DSN-TS的每个基础模块的输出是线性的，再加上每个频带被语音主导的时频单元和被噪声主导的时频单元可能是不平衡的，因此，DSN-TS输出的分布可能如图2.4所示，也就是说以0.5作为分类界限可能并不合理。因此我们需要基于某种准则寻找合适的分类阈值实现时频单元的分类，即根据DSN-TS的输出判断时频单元被语音主导还是被噪音主导。HIT-FA是语音分离性能的重要评价指标，既反映了时频单元的正确分类又反映了语音分离的错误分类，研究表明HIT-FA与语音的可懂度密切相关 [94]。本章我们通过最大化HIT-FA指标来选择最优的分类阈值。

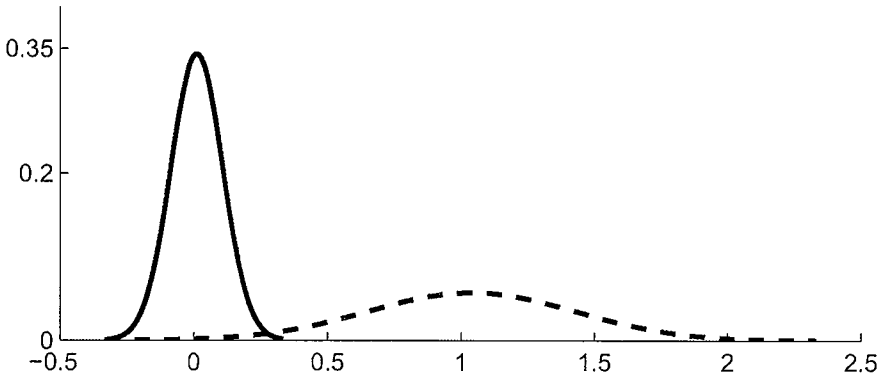


图 2.4: DSN-TS的输出分布示例。

首先我们通过一个修正的sigmoid函数将DSN-TS基础网络的原始输出转化为0到1的概率值，通过sigmoid函数转化的值可以解释为时频单元被语音或噪声

主导的概率 [136]。修正的sigmoid函数形式如下：

$$p(y=1|x)=\frac{1}{1+\exp(\alpha f(x)+\phi)}, \tag{2.1}$$

其中，$f(x)$是DSN-TS基础网络的输出，参数α和ϕ决定了sigmoid函数的形状。通过调整α和ϕ的值可以调整sigmoid函数的形状使得sigmoid输出关于0.5对称，也就是说，可以确定一个sigmoid函数使得HIT-FA在分类阈值取0.5时最大。我们在训练阶段通过最大化HIT-FA来确定最优的α和ϕ。

尽管通过sigmoid函数可以将DSN-TS的输出转化为0到1的概率值，并且以sigmoid函数输出的中心点0.5作为分类阈值可以在全局训练集上获得最大的HIT-FA。但是由于语音是时间变化的，不同时间段时频单元的IBM分布差异比较大，显然使用一个全局的分类阈值并不合适，本文使用一个基于固定长度片段的时变分类阈值来对时频单元的语音主导或噪声主导做最终判断。具体来讲，我们将DSN-TS的概率化输出，按照长度为256重叠为128分隔成若干个片段。在训练阶段，我们根据真实的IBM，基于最大化HIT-FA计算训练集数据中的每个片段的最优分类阈值，并且通过高斯核函数计算每个片段的DSN-TS概率化输出的直方图分布，利用这个直方图分布作为输入特征，最优分类阈值作为目标，使用单隐层DNN训练一个从直方图分布到最优分类阈值的映射函数。在测试阶段，我们同样地按照长度为256重叠为128对DSN-TS的概率化输出进行分段，然后通过高斯核函数计算每个片段的直方图分布，最后利用训练好的DNN估计片段的局部分类阈值，使用估计的分类阈值将该片段的时频单元分为语音主导和噪音主导。注意局部分类阈值的估计是在每一个频带每一个基础网络单独进行的。通过DSN-TS输出的概率化和局部阈值策略显著提升了DSN-TS的语音分离性能。

2.2.3 特征提取

对于分类问题，特征提取是至关重要的。我们首先使用64个gammtone听觉滤波器组对混合语音信号进行时频分解。这些gammtone听觉滤波器组的中心频率按照对数刻度等间隔地分布在50到8000Hz [166]。然后我们对每一个频带的gammtone滤波输出按照20ms窗长10ms窗移进行分帧处理。通过时频分解和分帧处理我们就可以得到时域信号的二维时频表示。对于每一个时频单元我们提取一系列互补的听觉特征，包括振幅调制谱（AMS），相对谱变换感知线性预测系数（RASTA-PLP），梅尔倒谱系数（MFCC）和语音基音特征。之前的研究表明这些互补特征是最好的语音分离特征组合 [170]。

2.3 实验及分析

2.3.1 数据集

我们使用“863计划”语音识别语料库对所提出的语音分离系统进行系统评

表 2.1: IBM命中与虚警的定义。

	真实的IBM	估计的IBM
Reject	0	0
Fa	0	1
Miss	1	0
Hit	1	1

估。“863计划”语音识别语料库分别包含了100个男性朗读者和100个女性朗读者，每人分别朗读500句。我们随机地从一个女性朗读者的语音中挑选50句纯净语音，然后分别和6种非人声噪声[1]以0dB信噪比进行混合，得到300句混合语音作为训练集。另外，我们从同一个朗读者的语音另外挑选20句与训练集不同的纯净语音，然后分别以-10dB，-5dB，0dB，5dB和10dB信噪比与14种非人声噪声进行混合，得到1400句混合语音作为测试集。其中有6种噪声和训练集噪声类型相同，剩下的8种噪音[2]和训练集噪声类型不同，以便用以测试语音分离模型对噪声类型的鲁棒性。这些噪声都是日常生成中的常见噪声，大部分都是非平稳噪噪声。

2.3.2 评价指标

我们使用时频单元的分类准确率，HIT-FA和SNR作为评价指标。根据时频单元真实的IBM和估计的IBM我们可以得到以下4种情况，其定义如表2.1所示，那么准确率（Accuracy）和HIT-FA可以通过公式（2.2）和（2.3）计算得到。

$$\text{Accuracy} = \frac{\#Hit + \#Reject}{\#Hit + \#Fa + \#Miss + \#Reject}. \tag{2.2}$$

$$\text{HIT-FA} = \frac{\#Hit}{\#Hit + \#Miss} - \frac{\#Fa}{\#Fa + \#Reject}. \tag{2.3}$$

其中$\#x$定义满足x条件的时频单元个数。更高的准确率和HIT-FA意味着更好的分离性能。

2.3.3 模型与设置

我们选择基于GMM和基于DNN-SVM的语音分离模型作为对比模型，分别定义为“GMM-based” [94]和“DNN-SVM-based” [175]。我们也比较基于DNN的语音分离模型，定义为“DNN-based”。为了更加系统而全面地评估所提出的语音分离系统，我们分别测试了噪声匹配和噪声不匹配的条件下的语音分离性能。

[1]babble, cocktail party, factory, siren, white and speech shaped noise

[2]bird chirp, crow noise, crowd, machine operation, engine start noise, alarm, traffic and wind noise.

除了最终的分类器不同之外，DNN-based和DNN-SVM-based具有完全相同的网络结构。不像DNN-based，DNN-SVM-based使用线性SVM作为最终分类器，线性SVM的输入不仅包括从混合信号中提取的听觉特征，还包括DNN最后一个隐层的输出。DNN的最后一个隐层的输出一般认为是原始输入特征的高度抽象，具有更好的线性区分性，但是由于经过多层非线性处理，最后的隐层输出通常会损失比较多的信息，和原始听觉特征级联的输入既避免了信息的损失又关注了特征的区分性，这种级联的特征作为SVM的输入能够得到更好的分类结果 [175]。DNN-based和DNN-SVM-based都使用双隐含层的深度神经网络，每个隐含层有200个神经元，隐含层的激活函数为sigmoid，输出层的激活函数为softmax。所有网络的连接权重通过RBM预训练的方式进行初始化 [75]，L-BFGS（Limited-Memory Broyden-Fletcher-Goldfarb-Shanno）优化器用来对DNN进行训练，最大迭代次数设置为500。

提出的DSN-TS由5个基础网络组成，每个基础网络有一个sigmoid隐含层，每个隐含层有200个神经元。同样地我们使用RBM预训练的方式对每个基础网络的连接权重进行初始化。由于DSN高效的优化方法，每个基础网络的训练仅迭代20次 [39]。另外，每个基础网络单独训练，并没有进行全局联合优化。

所有的语音分离系统使用相同输入特征。除了DSN-TS之外，其他语音分离模型使用相邻5帧的听觉特征作为输入。由于DSN-TS考虑了时序相关性，能够自动抓住上下文的相关性，因此DSN-TS的每个基础网络只输入单帧的听觉特征，并不包含上下文信息。

2.3.4　实验结果与分析

我们首先考察堆叠的基础网络模块个数是如何影响DSN-TS的语音分离性能的。图2.5报道了堆叠不同数量的基础网络模块时DSN-TS所取得的HIT-FA。我们可以观察到随着堆叠的基础网络模块数量的增加，DSN-TS取得的语音分离性能在不断地提高，但最终会达到一个稳定的值，实验中发现，当堆叠的基础网络模块增加到5个时，语音分离的性能就趋于稳定不再提升，因此在下面的实验中我们设置基础网络模块的个数为5。

表2.2报道了不同语音分离模型在噪声匹配和噪声不匹配的条件下所取得的语音分离性能，混合语音的输入SNR为0dB。从表中可以看出，DNN-based，DNN-SVM-based和DSN-TS的性能要明显优于GMM-based的性能，这表明神经网络模型对于语音和噪声的区分能力要优于GMM模型。我们也注意到DNN-SVM-based的性能要优于DNN-based。这主要得益于SVM的判别能力。另外我们也观察到在噪声匹配的条件下DSN-TS取得了最优的语音分离性能，要明显优于其他模型，而对于噪声不匹配的情况，DSN-TS在HIT-FA和Accuracy两个指标上明显优于其他模型。这表明时序相关信息对于语音分离具有重要的作用，所提出的DSN-TS模型有效地挖掘了语音信号的时序相关信息。

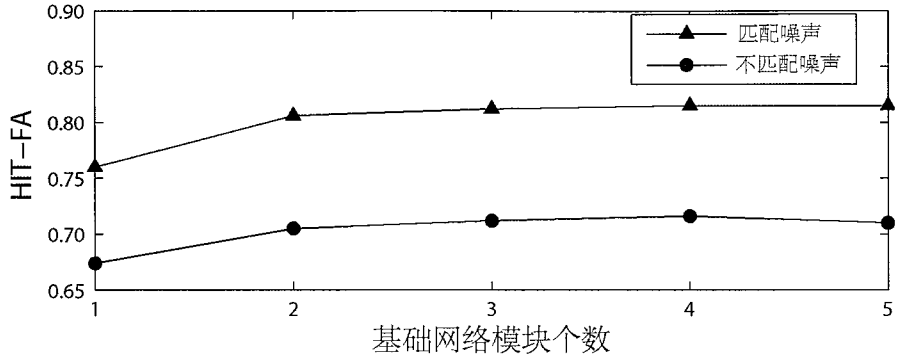


图 2.5: 堆叠不同数量的基础网络模块，DSN-TS取得的HIT-FA。

表 2.2: 不同语音分离模型在输入SNR为0dB时所取得的HIT(%)、FA(%)、HIT-FA(%)、Accuracy(%)和SNR(dB)。

	Models	HIT	FA	HIT-FA	Accuracy	SNR(dB)
Matched	GMM-based	79.99	34.92	45.07	69.15	5.00
	DNN-based	73.18	8.36	64.82	88.63	9.34
	DNN-SVM-based	77.06	6.03	71.02	91.19	9.10
	DSN-TS	**80.04**	**4.28**	**75.76**	**92.99**	**9.84**
Unmatched	GMM-based	80.40	36.58	43.82	66.92	5.11
	DNN-based	68.09	11.08	57.01	84.13	7.82
	DNN-SVM-based	69.89	8.05	61.84	87.14	7.30
	DSN-TS	75.40	8.79	**66.61**	**87.69**	7.44

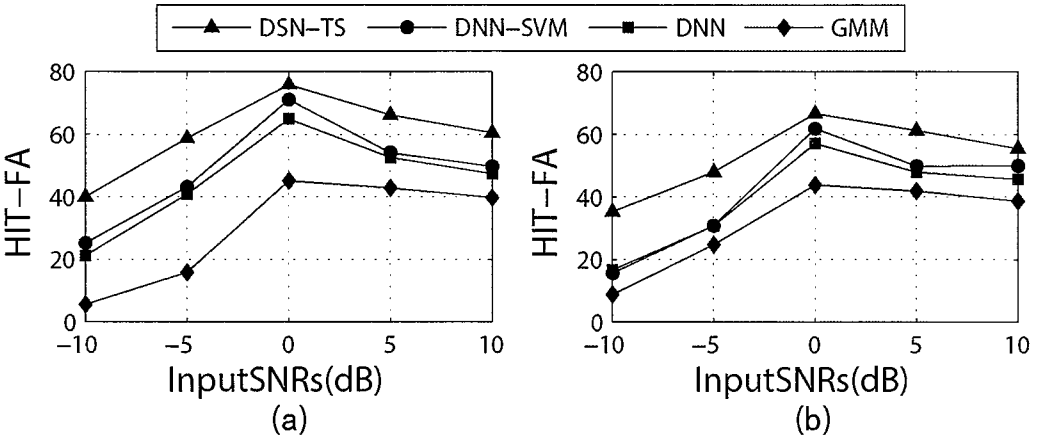


图 2.6: 语音分离模型在不同输入SNR条件下所取得的HIT-FA(%)。

最后，我们考察了语音分离模型对于输入SNR的鲁棒性。图2.6给出了语音分离模型在不同输入SNR条件下所取得的HIT-FA，其中图2.6(a)报道了噪声匹

配的情况，图2.6(b)报道了噪声不匹配的情况。我们可以看出无论在噪声匹配的条件下还是在噪声不匹配的条件下，所提出的DSN-TS在各种输入SNR条件下都取得了最优的性能。这主要得益于DSN-TS有效挖掘了语音的时序相关性，并且提出了基于HIT-FA的分类决策策略。

2.4　本章小结

针对语音的时序相关性，我们改进了DSN提出带有时序的DSN，实现了对语音时序相关性的有效建模。本章首先对DSN的网络结构进行了详细的介绍，然后仔细分析了语音的时序相关性，并将其表达为相邻时间帧的联合概率分布，通过改进DSN，提出了DSN-TS，分析了DSN-TS与DSN的区别，介绍了DSN-TS在语音分离中的应用流程。最后通过系统的实验验证了所提方法的有效性。

第三章　基于循环网络结构的声源分离与自回归联合优化的声源分离方法

3.1　引言

由于语音的产生机制，语音信号在频谱上具有明显的长时短时相关性，有效挖掘这些特性有助于提高语音分离的性能。最简单的方式是在输入向量中拼接更多的上下文帧或者计算输入特征的差分信息，这种方式一方面会造成输入维度急剧增加导致建模难度增大，另一方面挖掘长短时相关性的能力非常有限。本文第二章节利用深度层叠网络（DSN）的独特网络结构实现了对时间上相邻时频单元被语音或噪声主导的联合概率分布进行有效建模，提出了带有时序的深度层叠网络（DSN-TS），但DSN-TS的各个基础模块是分步优化的，而且仅能对相邻时频单元的相关性进行建模，此外DSN-TS是对每个频带单独建模，忽略了频带之间的相关性。循环神经网络（Recurrent Neural Network，RNN）被认为是非常有潜力的时序建模模型 [18, 122]。通过循环连接，RNN能够捕获语音数据中的长时短时相关性，但是梯度消失问题使得RNN的优化异常困难 [14]。

众所周知，由于语音的短时连续性，语音信号可以表达为一个典型的自回归过程 [137]，即通过N阶自回归模型，当前帧语音信号能够通过有限的历史帧语音信号预测出来 [137]。然而，在噪声环境中，语音信号不可避免地会受到噪声干扰，严重地破坏了语音的自回归性，因此利用带噪的历史语音信号很难预测当前的纯净语音信号。但历史分离的语音信号极大地消除了噪声的干扰，有效恢复了语音信号，因此，利用历史分离的语音信号可以预测当前的纯净的语音信号。反过来，通过历史分离的语音信号预测的当前纯净语音信号，又可以作为先验信息用来帮助当前语音信号的分离。基于此，我们提出了一种基于循环网络结构的自回归语音分离网络（Auto-Regression Separation Network，ARSN），ARSN由语音分离网络（Speech Separation Network，SSN）和语音自回归网络（Auto-Regression Network，ARN）构成。SSN的输出作为ARN的输入以自回归的方式预测下一时刻的纯净语音信号，而ARN预测的纯净语音信号反过来又作为先验信息联合对应的混合语音信号输入到SSN来实现该时刻语音信号的分离。SSN和ARN的相互协调，联合优化。此外，为了进一步减少噪声残留，我们探索了一个区分式的训练目标，有效抑制了噪声残留并尽可能避免了语音畸变。

针对语音的自回归性，本章提出了基于循环网络结构的自回归语音分离网络，实现了语音分离和语音自回归的联合建模和优化。在下面的章节中，首先我们介绍了所提出的自回归语音分离网络的结构，进一步介绍了所提出的区分

式训练目标，并对所提网络的优化过程进行了简单的介绍，最后进行了实验的验证和结果的分析。

3.2 自回归语音分离网络

3.2.1 网络结构

自回归分离网络由分离网络，N阶自回归网络和存储队列构成。为了方便介绍，我们以声乐分离为例来具体介绍所提出的自回归分离网络的结构以及流程。图3.1展示了一个应用于声乐分离的ARSN，它由一个声乐分离网络，一个歌唱自回归网络，一个背景音乐自回归网络和两个存储队列构成。声乐分离网络能够实现歌唱和背景音乐的分离，歌唱自回归网络能够通过历史分离的歌唱信号预测下一时刻的歌唱信号，而背景音乐自回归网络能够通过历史分离的背景音乐信号预测下一时刻的背景音乐信号。存储队列缓存了一段时间连续的分离歌唱和背景音乐，对于N阶自回归网络，那么存储队列的长度为N帧。因为存储队列的长度是有限的，因此分离歌唱或背景音乐按时间顺序进出存储队列。存储队列缓存的历史分离歌唱输入到歌唱自回归网络预测下一时刻的歌唱，同样地，存储队列缓存的历史分离音乐输入到背景音乐自回归网络预测下一时刻的背景音乐。反过来，自回归网络预测的歌唱和背景音乐作为先验信息联合对应的混合信号输入到声乐分离网络实现歌唱和背景音乐的分离。然后分离的歌唱和背景音乐再次进入存储队列进行缓存。这个过程按照时间序列依次重复进行，直到完成整个序列歌唱和背景音乐的分离。需要说明的是，这里ARSN各个网络的输入输出均是短时傅里叶变换幅度谱。

尽管所提出的ARSN和RNN有许多相似之处，但仍有许多不同，RNN是隐含层的自循环连接而ARSN是输入和输出层之间的循环连接。另外RNN 的循环连接是通过简单线性变换实现的，而ARSN的循环连接是通过一个由多个隐层构成的自回归网络实现的，并且自回归网络除了参与整个网络信息的循环传递外，也有自己明确的监督性目标，因此ARSN能够一定程度缓解RNN面临的梯度消失问题。

3.2.2 训练目标

语音分离的目标就是实现混合信号中语音和噪声的分离。最常用的方法是时频掩蔽技术 [171, 180]。将时频掩蔽$\mathbf{m}_t$应用到混合信号的STFT幅度谱$\mathbf{x}_t^{(m)}$即可获得分离语音的幅度谱$\tilde{\mathbf{y}}_t^{(s)}$。如果噪声和语音不相关，那么纯净语音$\mathbf{y}_t^{(s)}$和噪声$\mathbf{y}_t^{(n)}$幅度谱满足$\mathbf{x}_t^{(m)} \approx \mathbf{y}_t^{(s)} + \mathbf{y}_t^{(s)}$ [21]。在这个假设条件下分离噪声幅度谱可以通过应用$1-\mathbf{m}_t$到混合信号的幅度谱$\mathbf{x}_t^{(m)}$上计算得到。然而，在真实的声学环境中，噪声和语音并不是独立的，存在一定的相关性。本文我们没有假设噪声和语音相互独立，而是分别估计语音的时频掩蔽和噪声的时频掩蔽，并没有

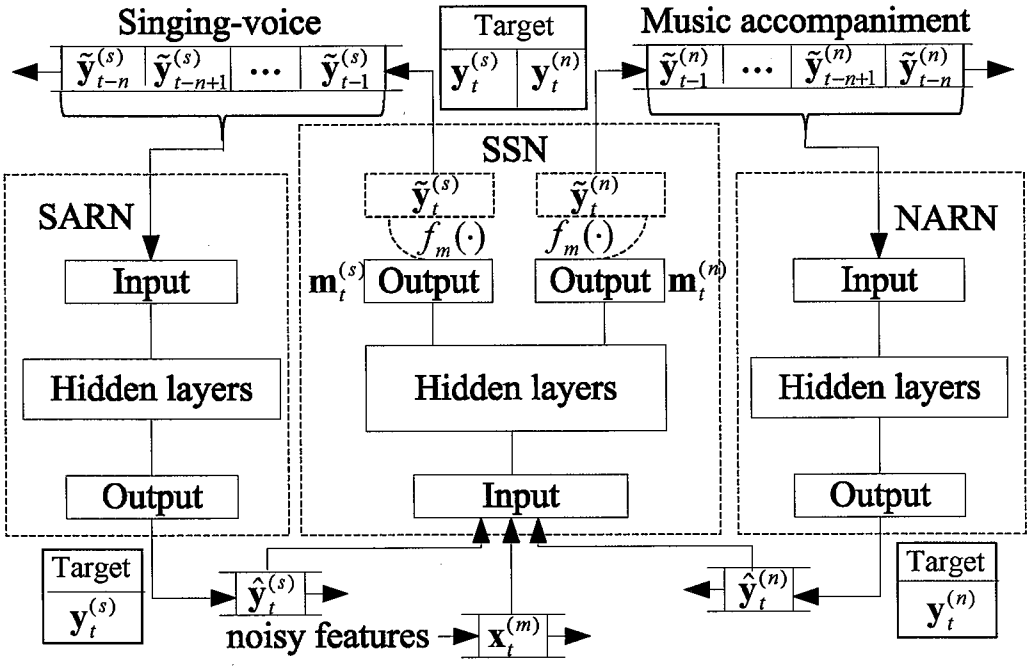


图 3.1: ARSN的网络结构。

限制两者相加之和为1。那么分离的语音和噪声幅度谱可以通过以下公式计算得到：

$$\tilde{\mathbf{y}}_t^{(s)} = f_m(\mathbf{m}_t^{(s)}) = \mathbf{m}_t^{(s)} \circ \mathbf{x}_t^{(m)}, \tag{3.1}$$

$$\tilde{\mathbf{y}}_t^{(n)} = f_m(\mathbf{m}_t^{(n)}) = \mathbf{m}_t^{(n)} \circ \mathbf{x}_t^{(m)}, \tag{3.2}$$

符号$\circ$定义了点乘运算，$\mathbf{m}_t^{(s)}$和$\mathbf{m}_t^{(n)}$是语音和噪声的理想时频掩蔽。

尽管理想浮值掩蔽可以直接作为语音分离的目标，但它仍然是一个中间计算目标，最终仍需估计目标语音或者噪声的幅度谱。直接估计目标语音和噪声的幅度谱可能更加直接，然而，目标幅度谱的取值范围通常比较大，难以控制数值边界，这可能导致训练比较困难 [171,192]。为此，我们提出基于掩蔽的幅度谱近似目标。在这个方案中，目标语音和噪声的幅度谱并不是直接预测得到，而是通过掩蔽技术从混合信号的幅度谱中计算得到的。具体来讲，我们将神经网络的输出作为时频掩蔽应用到混合信号的幅度谱中计算目标语音和噪声的幅度谱，最后利用分离的语音和噪声的幅度谱构建损失函数来优化神经网络。神经网络的输出层并没有明确的监督性目标，整个网络的优化来自最终的幅度谱近似目标。尽管神经网络的输出层并没有理想掩蔽作为监督目标，但在最终的频谱近似目标的监督指导下，神经网络的输出层自动学习到了和目标语音以及噪声相关的时频掩蔽。这种间接的基于时频掩蔽的频谱近似目标同时具备掩蔽近似和频谱近似的优势，能够获得到更加平滑的分离语音 [87]，同时也更容易优化 [171]。

混合信号中的不同声源通常具有一定的相关性和互补性，对多个声源同时

进行建模，可以提高声源分离的性能 [87]。因此本文我们对混合信号中语音和噪声同时进行建模和估计。即便如此，估计的目标语音中不可避免地存在一定的噪声残留，同时估计的噪声中也不可避免地存在一定的语音残留。为了进一步消除残留的语音和噪声，我们利用估计的目标语音和噪声构建一个区分式的损失函数，它不仅增大估计语音和期望语音以及估计噪声和期望噪音之间的相似性，同时也降低估计语音和期望噪声以及估计噪声与期望语音之间的差异性。另外，为了联合优化分离网络和自回归网络，我们提出了一个联合的优化目标，它不仅考虑了语音分离网络的损失还考虑了自回归网络的损失，如公式（3.3）所示。因此最终的损失函数包含三项，其中第一项和第二项是分离网络的分离误差和区分性损失，第三项是语音自回归和噪声自回归网络的估计误差。在一个模型里对它们联合建模和联合优化能够促进每个目标的学习。

$$\begin{aligned} J_t=&\frac{1}{2}(\left\|\tilde{\mathbf{y}}_t^{(s)}-\mathbf{y}_t^{(s)}\right\|_2^2+\left\|\tilde{\mathbf{y}}_t^{(n)}-\mathbf{y}_t^{(n)}\right\|_2^2)-\frac{\beta}{2}(\left\|\tilde{\mathbf{y}}_t^{(s)}-\mathbf{y}_t^{(n)}\right\|_2^2 \\ &+\left\|\tilde{\mathbf{y}}_t^{(n)}-\mathbf{y}_t^{(s)}\right\|_2^2)+\frac{\lambda}{2}(\left\|\hat{\mathbf{y}}_t^{(s)}-\mathbf{y}_t^{(s)}\right\|_2^2+\left\|\hat{\mathbf{y}}_t^{(n)}-\mathbf{y}_t^{(n)}\right\|_2^2), \end{aligned} \tag{3.3}$$

其中$\hat{\mathbf{y}}_t^{(s)}$和$\hat{\mathbf{y}}_t^{(n)}$分别是语音自回归网络和噪声自回归网络的输出。β和λ分别指明了区分性损失和自回归网络的预测损失相对整个优化目标的重要性，一般通过实验进行选择。

3.2.3 优化

我们使用沿时间反向传播（Backpropagation Through Time，BPTT）算法优化所提出的ARSN。BPTT的基本原则是沿时间展开，即将循环连接沿时间展开为深度前馈网络。图3.2显示了ARSN沿时间展开的结构。

3.2.3.1 前向传播

除了循环连接外，ARSN的前向传播过程和前馈网络类似。我们采用混合信号的幅度谱$\mathbf{x}_t^{(m)}$ $(t=1,2,\cdots,T)$ 作为输入序列，沿着时间序列依次计算ARSN的每层激活输出，如下所示：

$$\mathbf{z}_{t,l+1}=\mathbf{W}_l\times(\mathbf{a}_{t,l})^T;\ \ \mathbf{a}_{t,l+1}=f(\mathbf{z}_{t,l+1}), \tag{3.4}$$

其中$\mathbf{a}_{t,l+1}$是$(l+1)$层在时刻t的激活输出，$\mathbf{z}_{t,l+1}$是时刻t输入到$(l+1)$层的加权和，$\mathbf{W}_l$是l层和$(l+1)$层之间的连接权重。$f(\cdot)$是激活函数，符号$(\cdot)^T$定义了矩阵的转置操作。为了简化数学推导，我们定义$\mathbf{a}_{t,l+1}^{(m)}$、$\mathbf{z}_{t,l+1}^{(m)}$和$\mathbf{W}_l^{(m)}$是分离网络的相关符号，定义$\mathbf{a}_{t,l+1}^{(s)}$、$\mathbf{z}_{t,l+1}^{(s)}$和$\mathbf{W}_l^{(s)}$是语音自回归网络的相关符号，而$\mathbf{a}_{t,l+1}^{(n)}$、$\mathbf{z}_{t,l+1}^{(n)}$和$\mathbf{W}_l^{(n)}$是噪声自相关网络的相关符号。按照公式（3.4）的形式，它们能够依次被计算出来。

不同于前馈网络，分离网络和自回归网络的输入和输出是相互连接的。这里语音分离网络、语音自回归网络和噪声自回归网络的输入依次是$\mathbf{a}_{t,1}^{(m)}=$

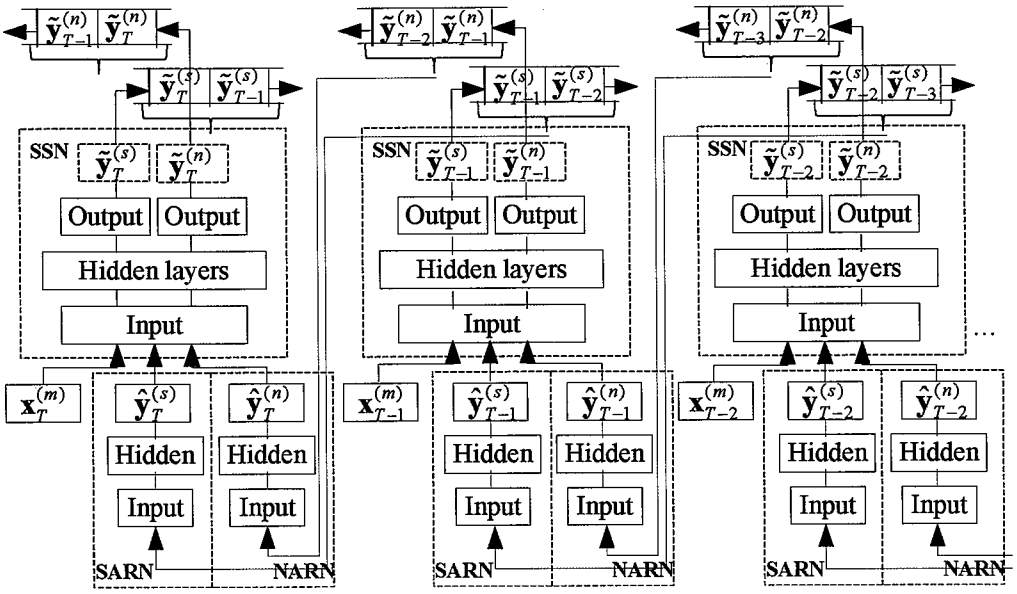


图 3.2: ARSN沿时间展开的网络结构。

$[1, \mathbf{x}_t^{(m)}, \hat{\mathbf{y}}_t^{(s)}, \hat{\mathbf{y}}_t^{(n)}]$、$\mathbf{a}_{t,1}^{(s)} = [1, \tilde{\mathbf{y}}_{t-2}^{(s)}, \tilde{\mathbf{y}}_{t-1}^{(s)}]$和$\mathbf{a}_{t,1}^{(n)} = [1, \tilde{\mathbf{y}}_{t-2}^{(n)}, \tilde{\mathbf{y}}_{t-1}^{(n)}]$。而它们的输出层依次是$[\mathbf{m}_t^{(n)}, \mathbf{m}_t^{(s)}] = \mathbf{a}_{t,n_l}^{(m)}$、$\hat{\mathbf{y}}_t^{(s)} = \mathbf{a}_{t,n_l}^{(s)}$和$\hat{\mathbf{y}}_t^{(n)} = \mathbf{a}_{t,n_l}^{(n)}$。$\tilde{\mathbf{y}}_t^{(s)}$和$\tilde{\mathbf{y}}_t^{(n)}$能够通过公式（3.1）和（3.2）计算得到。

3.2.3.2　反向传播

通过链式法则，我们可以从时刻T到1递归地计算损失函数关于各个连接权重的梯度，如下所示：

$$\nabla \mathbf{W}_{t,l} = \frac{\partial J_t}{\partial \mathbf{W}_l} = \frac{\partial J_t}{\partial \mathbf{z}_{t,l+1}} \frac{\partial \mathbf{z}_{t,l+1}}{\partial \mathbf{W}_l} = \frac{\partial J_t}{\partial \mathbf{z}_{t,l+1}} (\mathbf{a}_{t,l})^T. \tag{3.5}$$

为了简化符号，我们引入变量δ，定义$\delta_{t,l}^{(m)} = \frac{\partial J_t}{\partial \mathbf{z}_{t,l}^{(m)}}$，$\delta_{t,l}^{(s)} = \frac{\partial J_t}{\partial \mathbf{z}_{t,l}^{(s)}}$和$\delta_{t,l}^{(n)} = \frac{\partial J_t}{\partial \mathbf{z}_{t,l}^{(n)}}$。它们反映了ARSN各个网络层的神经元对损失函数的影响。

对于分离网络输出层（$l = n_l$）的δ项，能够通过以下公式计算得到：

$$\delta_{t,n_l}^{(m)} = [\mathbf{s}_t, \mathbf{n}_t] \circ [\mathbf{x}_t^{(m)}, \mathbf{x}_t^{(m)}] \circ f'(\mathbf{z}_{t,n_l}^{(m)}), \tag{3.6}$$

其中，

$$\mathbf{s}_t = -(\mathbf{y}_t^{(s)} - \tilde{\mathbf{y}}_t^{(s)}) + \beta(\mathbf{y}_t^{(n)} - \tilde{\mathbf{y}}_t^{(s)}) + \mathbf{I}_t(\delta_{t+1,1}^{(s)}) + \mathbf{I}_t(\delta_{t+2,1}^{(s)}), \tag{3.7}$$

$$\mathbf{n}_t = -(\mathbf{y}_t^{(n)} - \tilde{\mathbf{y}}_t^{(n)}) + \beta(\mathbf{y}_t^{(s)} - \tilde{\mathbf{y}}_t^{(n)}) + \mathbf{I}_t(\delta_{t+1,1}^{(n)}) + \mathbf{I}_t(\delta_{t+2,1}^{(n)}), \tag{3.8}$$

$\mathbf{I}_t(\cdot)$定义向量元素选择函数，$\mathbf{I}_t(\delta_{t+1,1}^{(s)})$和$\mathbf{I}_t(\delta_{t+2,1}^{(s)})$分别表示从$\delta_{t+1,1}^{(s)}$和$\delta_{t+2,1}^{(s)}$中选择和$\tilde{\mathbf{y}}_t^{(s)}$相关的$\delta$项。同样地，$\mathbf{I}_t(\delta_{t+1,1}^{(n)})$和$\mathbf{I}_t(\delta_{t+2,1}^{(n)})$从$\delta_{t+1,1}^{(n)}$和$\delta_{t+2,1}^{(n)}$中选择和$\tilde{\mathbf{y}}_t^{(n)}$相关的$\delta$项。

而$\delta_{t,n_l}^{(s)}$能够通过以下公式计算得到：

$$\delta_{t,n_l}^{(s)} = (-\lambda(\mathbf{y}_t^{(s)} - \hat{\mathbf{y}}_t^{(s)}) + \mathbf{I}_s(\delta_{t,1}^{(m)})) \circ f'(\mathbf{z}_{t,n_l}^{(s)}), \tag{3.9}$$

其中$\mathbf{I}_s(\delta_{t,1}^{(m)})$表示从$\delta_{t,1}^{(m)}$选择和$\hat{\mathbf{y}}_t^{(s)}$相关的$\delta$项。

对于第l层（$l = n_l - 1,\ n_l - 2,\ \cdots, 1$）的$\delta$项，$\delta_{t,l}^{(m)}$能够通过以下公式计算得到：

$$\delta_{t,l}^{(m)} = ((\mathbf{W}_l^{(m)})^T \times \delta_{t,l+1}^{(m)}) \circ f'(\mathbf{z}_{t,l}^{(m)}), \tag{3.10}$$

而$\delta_{t,l}^{(s)}$能够通过公式（3.11）计算得到：

$$\delta_{t,l}^{(s)} = ((\mathbf{W}_l^{(s)})^T \times \delta_{t,l+1}^{(s)}) \circ f'(\mathbf{z}_{t,l}^{(s)}). \tag{3.11}$$

类似于$\delta_{t,n_l}^{(s)}$和$\delta_{t,l}^{(s)}$，$\delta_{t,n_l}^{(n)}$和$\delta_{t,l}^{(n)}$能够通过类似公式（3.9）和（3.11）的方式计算得到。

计算时刻T到1的所有δ项之后，损失函数相对于所有连接权重的导数可以通过公式（3.12）计算得到。

$$\nabla \mathbf{W}_l = \sum_{t=1}^{T} \nabla \mathbf{W}_{t,l} = \sum_{t=1}^{T} (\delta_{t,l} \times (\mathbf{a}_{t,l-1})^T). \tag{3.12}$$

然后，我们使用L-BFGS算法 [112]优化更新网络的所有连接权重$\mathbf{W}_l$。

3.3 实验及其分析

3.3.1 数据集

我们将所提出的ARSN应用到声学分离任务上来检测所提出模型的有效性。需要说明的是声乐分离并非ARSN唯一的应用，可以很容易将其应用到其他的声源分离任务，比如语音分离。事实上，由于背景音乐的高度非平稳性，声乐分离相对于语音分离更具有挑战性。

我们使用MIR-1K数据集 [81]来评估所提出的ARSN在声乐分离任务上所取得的性能。MIR-1K数据集包含1000个音乐片段，总共时长133分钟。这些音乐片段是从11个男士和8个女士演唱的110首华语流行音乐中提取的，其中人声歌唱和背景音乐是单独录制并且存储在不同的通道。这就意味着对于每一个音乐片段，其人声歌唱和背景音乐事先是已知的，因此可以作为监督数据用来训练声乐分离模型。我们从中随机选择794个音乐片段作为训练集，训练集的音乐片段来自9个男士演唱者和6个女士演唱者。剩下的206个音乐片段用作测试集，其中176个音乐片段由另外的2个男士和2个女士演唱，它们用来测试声乐分离模型对演唱者的泛化性能，而测试集中的另外30个音乐片段和训练集的演唱者相同，但是他们的演唱内容并不相同。

3.3.2 评价指标

我们选择SIR（Source to Interference Ratio）、SAR（Source to Artifacts Ratio ）和SDR（Source to Distortion Ratio）作为评价指标 [160]，它们能够通过BSS Eval工具 [160]计算得到。更高的SIR、SAR和SDR意味着更好的分离性能。

3.3.3 模型与设置

我们选择Huang等人提出的基于深度RNN（定义为DRNN）的声乐分离方法 [88]作为对比模型，该方法被认为能够代表当前声乐分离技术的发展水平。在实验中，对于DRNN我们使用Huang提供的实现代码和实验配置[1]。为了保持和DRNN大致相同的参数量，我们设置ARSN的分离网络有3个隐含层，每个隐含层有1000个神经元，并且使用规整的线性函数（Rectified Linear Unit，ReLU） [59]作为隐含层的激活函数。另外设置ARSN的人声歌唱自回归网络和背景音乐自回归网络的隐含层层数为1，每个隐含层有250个神经元，隐含层的激活函数同样是ReLU。由于我们并没有假定语音和噪声相互独立，所以时频掩蔽$m_t^{(s)}$和$m_t^{(n)}$的值可能超过了1，为了避免过大的数值范围，我们通过一个有界的线性函数$f(x) = max(0, min(x, \theta))$ 将分离网络的输出限制到$[0, \theta]$。根据开发集[2] 的性能，我们设置θ等于2.5。

为了优化ARSN，我们设置BPTT的截断时间步长为100。为此，我们将训练集中的所有音乐片段拼接成一个长的序列，然后每100帧切出一个小片段作为ARSN的输入序列，相邻的两个小片段有50帧的重叠。使用L-BFGS优化器来训练ARSN，ARSN的网络权重是随机的并没有通过预训练的方式初始化。最大的迭代次数设置为400。另外根据开发集的性能，我们设置$\beta = 0.05$，$\lambda = 0.1$，设置自回归网络的阶数为5。另外为了避免过拟合问题，我们对自回归网络的输入$\hat{\mathbf{y}}_t^{(s)}$和$\hat{\mathbf{y}}_t^{(n)}$增加一个均值为0方差为0.2的高斯噪声，以便带来更多变化。

所有的实验使用混合信号的短时傅里叶变换幅度谱作为输入特征。我们也探索了MFCC特征和傅里叶变换对数幅度谱作为输入特征，初步的实验显示使用幅度谱作为输入特征能够取得更好的分离性能。时频分解使用1024个频点的STFT，分帧通过一个窗长为1024窗移为512的汉明窗加窗得到。另外我们发现上下文信息能够进一步提升ARSN的性能，因此我们使用一个3帧的上下文信息作为所有模型的输入。

3.3.4 实验结果与分析

首先，我们比较了不同的训练目标对声乐分离性能的影响。我们分别比较了IRM近似的目标 [171]、STFT幅度谱近似的目标 [171]、Huang提出

[1]https://sites.google.com/site/deeplearningsourceseparation

[2]Four clips, Ani_2_02, stool_4_01, bobon_1_01 and heycat_4_09.

表 3.1: 使用不同训练目标所取得声乐分离性能(dB)。

Objectives	Matched singer			Unmatched singer		
	SDR	SIR	SAR	SDR	SIR	SAR
Mixture	0.00	0.00	∞	0.00	0.00	∞
IRM	8.18	11.88	8.75	7.80	11.26	8.34
STFT-MAG	8.05	13.58	8.01	8.07	13.58	8.11
HuangObj	**9.48**	14.43	**9.80**	**9.04**	13.93	**9.36**
ProposedObj	9.43	**17.41**	9.62	8.88	**16.28**	9.07

的区分式幅度谱近似目标 [88]以及本文提出的区分式幅度谱近似目标，它们依次记为“IRM”、“STFT-MAG”、“HuangObj”和“ProposedObj”。其中，IRM近似的目标直接预测IRM，STFT幅度谱近似的目标直接预测目标人声歌唱的STFT幅度谱。HuangObj和ProposedObj通过掩蔽技术同时预测目标人声歌唱和背景音乐的幅度谱，并且利用预测的目标人声歌唱和背景音乐的幅度谱构建区分式的训练目标。HuangObj和ProposedObj的不同之处在于HuangObj假定人声歌唱和背景音乐是无关的，利用类似维纳滤波的掩蔽技术，而ProposedObj并没有假定人声歌唱和背景音乐不相关，而是通过两个掩蔽分别估计人声歌唱和背景音乐的幅度谱。为了简化实验，我们使用3个隐含层的DNN作为语音分离模型，每个隐含层有1000个神经元，隐含层的激活函数为ReLU。表3.1报告了不同训练目标所取得的声学分离性能。从表中可以看出，HuangObj和ProposedObj的性能要显著优于IRM和STFT-MAG，而HuangObj在SDR和SAR这两个评价指标上取得了最好的性能。但是相对于HuangObj，ProposedObj仅以很小的SDR和SAR代价便取得了显著的SIR性能提升，也就是说我们所提出的分离目标能够以很小的语音损失代价消除更多的噪声。这主要得益于ProposedObj放松了人声歌唱和背景音乐的独立性假设。

其次，我们系统分析和比较了不同声学分离模型在歌唱者匹配和歌唱者不匹配的条件下所取得的分离性能，需要说明的是这里DNN、DRNN和ARSN都使用本文所提出的基于掩蔽的区分式幅度谱近似目标。表3.2报道了DNN、DRNN和ARSN的声乐分离性能。从表中可以看出，DRNN和ARSN无论在歌唱者匹配和歌唱者不匹配的条件所取得的声乐分离性能都优于DNN。这主要得益于循环网络结构能够对时序数据中的长短时相关性进行有效建模，表明循环网络结构在声乐分离任务上要优于前馈网络结构。但是，由于面临着梯度消失问题，DRNN取得的性能提升比较有限。相比于DRNN，本文所提出的ARSN取得了显著的性能提升。这主要是因为一方面ARSN通过引入自回归网络有效挖掘了人声歌唱和背景音乐信号中的自回归性，另一方面自回归网络和分离网络相

表 3.2: 不同模型在歌唱者匹配和歌唱者不匹配条件下所取得的声乐分离性能(dB)。

Models	Matched singer			Unmatched singer		
	SDR	SIR	SAR	SDR	SIR	SAR
DNN	9.48	14.43	9.80	9.04	13.93	9.36
DRNN	9.96	15.53	10.22	9.47	15.03	9.69
ARSN	**10.24**	**19.92**	**10.48**	**9.50**	**18.26**	**9.70**

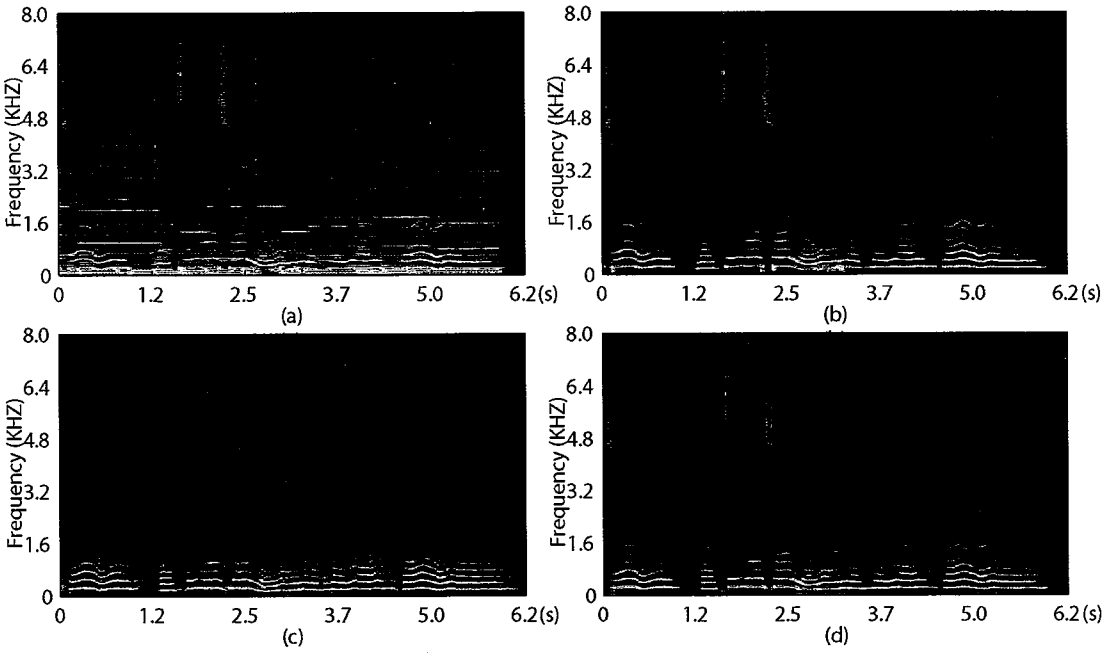


图 3.3: ARSN声乐分离样例。

互配合联合优化，充分发挥了自回归网络和分离网络各自的优势，另外由于循环网络具有明确的监督目标，缓解了循环网络结构面临的梯度消失问题。

最后，我们展示了一个声乐分离的例子来进一步阐述本文所提出的ARSN的各个模块的作用。图3.3(a)(b)(c)(d)依次是混合音乐、纯净人声歌唱、歌唱自回归网络预测的人声歌唱以及声乐分离网络分离的人声歌唱的语谱图。从图中可以看出，利用历史分离的歌唱，自回归网络成功预测了下一帧纯净歌唱的主要谐波结构，这表明歌唱自回归网络有效抓住了人声歌唱的自回归性。相比于自回归网络预测的纯净歌唱，分离网络分离的歌唱呈现了更加完整的结构和更加丰富的细节，特别在高频。这表明自回归网络预测的纯净歌唱能够为分离网络提供先验信息，并帮助帮助歌唱和背景音乐的分离。

3.4 本章小结

针对语音信号的自回归性，我们提出了自回归分离网络，实现了对语音自回归的有效建模。所提出的自回归分离网络由自回归网络和分离网络构成，自回归网络和分离网络通过循环连接的方式并在统一的框架下联合优化，既发挥了自回归网络对语音信号中长时短时相关信息的建模能力，又利用分离网络强大的分离能力，两者相互促进相互协调。另外，我们也探索了基于掩蔽的区分式频谱近似分离目标，显著降低了分离语音中的噪音残留和分离噪声中的语音残留。最后我们将所提出的分离方法应用到声乐分离任务上，在MIR-1K数据集中取得了显著的声乐分离性能。

第四章　两阶段多目标联合学习的语音分离方法

4.1　引言

语音信号除了在时间上具有很强的相关性之外，在频率上也呈现了显著的相关性，这导致语音信号在时频域上具有明显的时空结构。对于语音分离，时频域上的时空结构不仅表现在带噪语音的听觉特征，也表现在语音分离目标上，有效挖掘带噪语音听觉特征和语音分离目标的时空结构对提高语音分离性能具有重要价值。

语音分离旨在实现从混合信号中提取出我们感兴趣的语音信号或者实现混合信号中语音信号和噪声信号的分离，它能够很自然地表达为一个监督式的学习问题。典型的监督式语音分离通常利用监督式学习算法学习一个从带噪特征到分离目标的映射函数。理想时频掩蔽和目标语音频谱 [171] 是监督式语音分离的两类主要分离目标，理想掩蔽通常是利用纯净语音和纯净噪声计算得到的，比如理想二值掩蔽（IBM） [165]和理想浮值掩蔽（IRM） [171]。基于理想掩蔽目标的方法通常学习理想掩蔽的最优近似，而基于目标语音频谱的方法通常学习目标语音频谱的最优近似。不管是理想时频掩蔽还是目标语音频谱都可以实现语音和噪声的分离，并且能够极大地提高分离语音的可懂度和感知质量 [67, 108, 124, 167, 171]。根据所使用的时频分解技术的不同，理想时频掩蔽和目标语音频谱可分为短时傅立叶变换域和gammtone听觉滤波域下的两种表示方式。图4.1(a)(b)分别给出了一段混合语音在gammtone听觉滤波域下的目标语音听觉谱和IRM。如图所示，无论是听觉谱还是IRM都呈现出了明显的时空结构，而且由于语音的稀疏性这些时空结构对声学环境能够保持较强的稳定性，有效挖掘语音的时空结构能够提高语音分离的性能。另外直观来看，目标语音听觉谱和IRM呈现的时空结构具有明显的相似性。实际上，从理论上讲，IRM是由混合语音中的纯净语音和噪声听觉谱计算得到，如公式（4.1）所示，因此理想时频掩蔽和目标语音频谱具有密切的联系和相关性。

$$IRM(t,f) = \frac{S^2(t,f)}{S^2(t,f) + N^2(t,f)}, \tag{4.1}$$

其中$S^2(t,f)$和$N^2(t,f)$分别是纯净语音和噪声在时间帧为t频带为f的时频单元的听觉能量。由于语音在时间和频率上是稀疏的，语音听觉谱具有对声学环境相对稳定的谐波结构，但其取值范围通常比较大，估计难度大。而IRM的取值范围在$[0,1]$，相对于无界的听觉谱，更容易估计。综上，理想时频掩蔽和目标语音频谱既具有明显的相关性又具有很强的互补性，因此对它们联合建模有可能提高语音分离的性能。

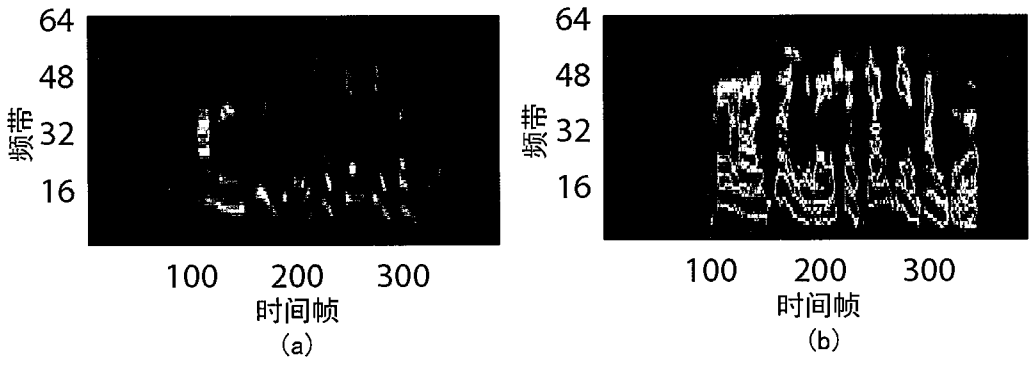


图 4.1: Gammtone听觉滤波域下目标语音听觉谱和IRM（0dB信噪比混合）。

本章我们提出了一种基于DNN的多目标联合建模的语音分离方法，使用DNN对目标语音频谱和IRM联合建模。另外为了更好地帮助DNN的学习和进一步提升分离语音的质量，我们提出了两阶段的学习方法。在第一阶段，为了挖掘语音听觉谱和IRM的时空结构，我们利用两个降噪自编码器（Denoising Autoencoders，DAE）通过自学习的方式来分别挖掘带噪输入听觉特征以及IRM和目标语音听觉谱联合的分离目标的时空结构。然后通过一个线性转换矩阵W_h将两个训练好的DAE连接起来，从而构建了基于DNN的多目标语音分离模型。最后通过监督式学习算法对构建的语音分离模型进行训练和微调。第一阶段预测的目标语音听觉谱和IRM具有较强的互补性，为了利用这些信息来进一步提升语音分离的性能，在第二阶段，我们利用DNN对第一阶段估计的目标语音听觉谱和IRM进行进一步的融合来得到最终的语音分离目标。实验证明两阶段的语音分离方法有效挖掘语音的时空结构，显著提升了语音分离的性能。

前两个章节主要针对语音信号在时间上的特性，分别挖掘了语音的时序相关性和自回归性。本章我们更多地从语音时频域上的时空结构出发，利用两阶段多目标联合学习的语音分离方法，不仅挖掘了理想时频掩蔽和目标语音频谱的时空结构，还有效地对它们的相关性和互补性进行了建模。在下面的章节里我们首先介绍了多目标联合学习的语音分离方法，接着介绍了基于DAE语音分离网络构建方法，并提出了带有偏好权重的梯度下降算法来考虑不同学习任务的差异，最后介绍了基于DNN的多目标融合方法并通过实验验证本章所提出的语音分离方法的有效性。

4.2 第一阶段：多目标联合学习

通常来讲，通过多任务联合学习能够促进共享相关信息的多个学习目标的学习，许多工作在理论和实验上都对此进行了验证 [9]。对于语音分离，理想时频掩蔽和目标语音频谱是两类主要的分离目标，它们具有密切的联系，正如上文所讨论的，IRM和目标语音听觉谱既具有明显的相关性又有很强的互补

性，因此通过多任务学习的方式对它们联合建模能够提高语音分离的性能。近年来，基于DNN的多任务学习受到国内外研究者的广泛关注，在许多应用领域取得巨大的成功 [34]。多目标的DNN输出层通常有多个相关的学习目标，这些学习目标共享隐含层。通过多任务联合学习，共享隐含层能够提取多个学习任务相关的共享特征表示。我们将基于DNN的多任务学习应用到语音分离中，利用DNN同时对两个常用的语音分离目标IRM和目标语音听觉谱进行联合建模。

4.2.1 基于自编码器的模型构建

在低性噪比条件下，语音信号通常会被噪声严重干扰，直接学习从带噪听觉特征到分离目标可能非常困难。然而，由于语音的产生机制，语音的听觉特征和分离目标都具有明显的时空结构，再加上语音信号通常在时频域是稀疏的，这些时空结构对于声学环境能够保持一定的稳定性 [173]。显然这些时空结构为语音分离提供了丰富的线索，能够帮助监督式语音分离的学习。为了挖掘语音听觉特征和分离目标的时空结构，我们使用两个DAE分别对带噪听觉特征以及IRM和目标语音听觉谱联合的分离目标进行无监督学习，通过自学习的方式，挖掘混合语音的带噪听觉特征以及语音分离目标的时空结构。具体来讲，其中一个DAE使用大量的带噪听觉特征进行训练，另一个DAE使用IRM和目标语音听觉谱联合的分离目标进行训练。另外，由于语音的短时连续性，我们使用上下文信息作为训练输入。训练好的DAE能够对新的输入进行特征抽取和编码。我们可以把DAE的编码层输出作为学习到的带噪听觉特征和分离目标的时空结构模式。然后利用线性变换的方式将带噪听觉特征的编码向量映射到语音分离目标的编码向量。线性变换的权重$\mathbf{W}_h$可以通过一对一的监督式学习方式得到，并且可以基于最小均方误差利用逆矩阵方式显示计算，即$\mathbf{W}_h = (\mathbf{H}_x{}^T\mathbf{H}_x)^{-1}\mathbf{H}_x{}^T\mathbf{H}_y$，其中$\mathbf{H}_x$是带噪听觉特征的编码输出，而$\mathbf{H}_y$是语音分离目标的编码输出。通过线性变换的方式我们将两个训练好的DAE的编码层和解码层连接起来，从而构建了一个前向传播网络，如图4.2所示，由于前向传播网络的权重已经通过DAE和线性变换的训练得到，如果在此基础上利用监督数据继续训练可以进一步提升语音分离系统的性能。事实上，DAE和线性变换的训练可以视为语音分离模型的预训练过程。令人欣喜的是，对于用以上方法构建的语音分离网络，即使不再进行训练，依然能够从带噪输入中分离出目标语音和噪声。这主要是由于DAE抓住了带噪听觉特征和语音分离目标的结构模式而线性变换学习到了从听觉特征模式到分离目标模式的映射。

4.2.2 基于偏好权重加权的梯度下降

对于多任务学习，不同任务的学习可以根据侧重点不同区别对待。事实上，深度神经网络的每一个输出神经元可以认为是一个单独的学习任务，对应地，神经网络的每个输出神经元的学习也可以区别对待。直观来讲，对于神经网络

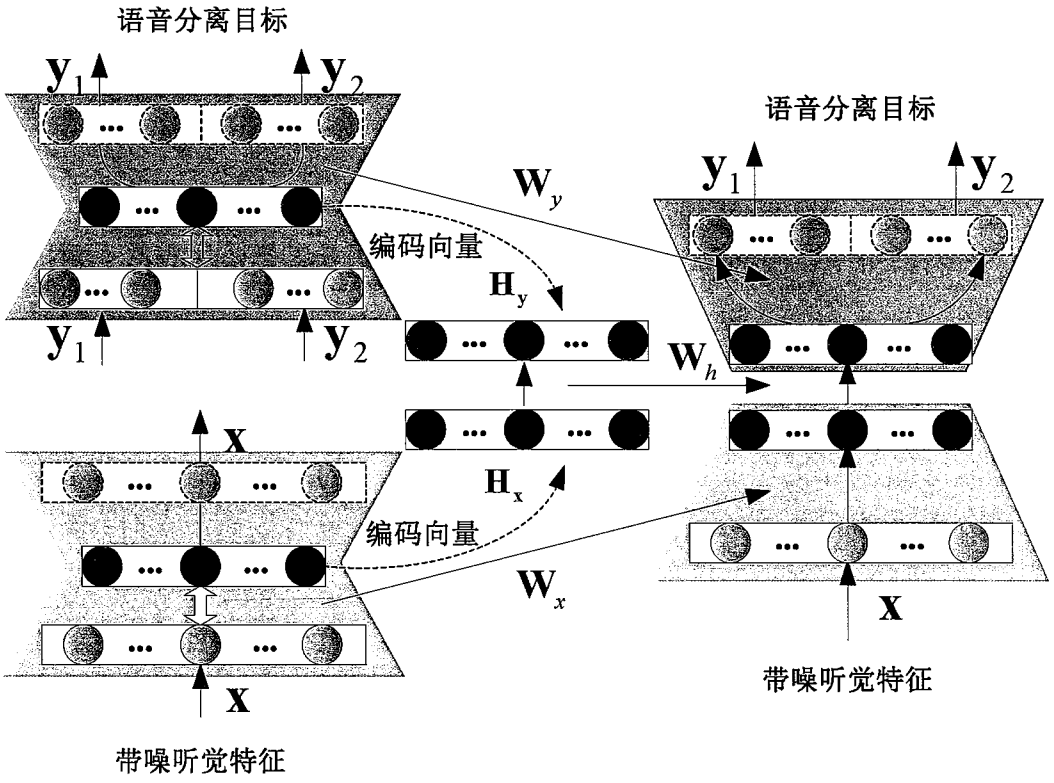


图 4.2: 基于DAE的语音分离模型构建方式。

输出层的每个神经元，预测误差较大的神经元没有得到充分学习，继续优化的潜力较大，应该加强其学习，而对于预测误差较小的神经元，应该减少对其学习的关注。基于反向传播（Backpropagation，BP）的优化算法均等地看待每一个目标的学习。对于IRM和目标语音听觉谱联合的分离目标，一方面IRM和语音听觉谱的数值范围差异较大，另一方面由于语音的稀疏性，语音听觉能量在不同频带上的分布差异较大，绝大部分集中在中低频，高频能量相对较小。因此对于多目标的语音分离模型其优化更需要差异化地对待不同学习目标。为此，我们提出了一种基于偏好权重加权的梯度下降算法，使用一个偏好权重对神经网络的输出误差进行加权。我们希望误差较大的输出神经元应该加强学习，加权权重较大，误差较小的输出神经元应该减弱学习，加权权重较小。偏好权重与预测误差密切相关，我们通过使用最小最大映射的归一化（Mapminmax Normalization）方式将预测误差归一化到$[0,1]$来计算偏好权重，如公式（4.2）所示。

$$\rho = \frac{|\mathbf{h}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) - \mathbf{y}| - \min(|\mathbf{h}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) - \mathbf{y}|)}{\max(|\mathbf{h}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) - \mathbf{y}|) - \min(|\mathbf{h}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) - \mathbf{y}|)}, \tag{4.2}$$

其中$|\cdot|$定义向量或矩阵元素的绝对值运算。$\mathbf{W}$和$\mathbf{b}$是神经网络的连接权重和偏置向量，而$\mathbf{x}$和$\mathbf{y}$神经网络对应的输入和目标。$\mathbf{h}_{\mathbf{w},\mathbf{b}}(\mathbf{x})$是神经网络的预测输出。$\rho$是神经网络输出误差的偏好权重，它是一个和神经网络输出神经元一一对应的

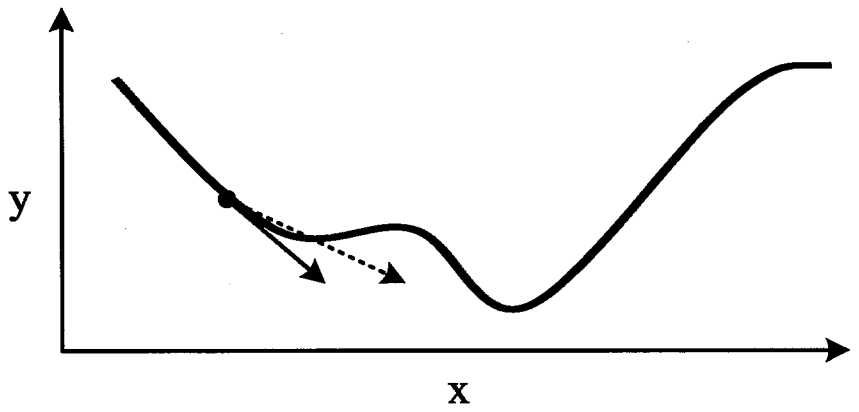


图 4.3: 最速梯度下降方向和偏好权重加权的梯度下降方向的示意图。

向量，用以对输出误差进行加权，如公式（4.3）所示。

$$J(\mathbf{W},\mathbf{b};\mathbf{x},\mathbf{y}) = \frac{1}{2}\|\sqrt{\rho} \bullet (\mathbf{h}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) - \mathbf{y})\|^2, \tag{4.3}$$

其中•定义点乘运算。从公式（4.2）和（4.3）可以看出。预测误差越大的输出神经元其加权权重越大，在最终的损失函数中所占的比重越高，对网络权重更新的影响越大，其学习会得到加强。

在基于梯度下降的优化算法中，最速梯度下降是最常用的优化算法 [141]。最速梯度下降算法总是寻找当前时刻梯度下降最快的方向，这种局部贪心的优化算法可能会导致神经网络陷入局部最小值从而需要更多的迭代数量才能收敛到相对最优值 [66]。本章节我们使用一个偏好权重对输出误差进行加权，从而使得最速梯度下降的方向偏向全局最优值的方向，如图4.3所示。尽管带有偏好权重的梯度下降方向并不是当前时刻最优的梯度下降方向，但它通过偏好权重使得优化方向侧重于误差更大的方向，从全局上看可能更容易收敛到全局最优值。

为了检验所提出的偏好权重加权的梯度下降算法的有效性，我们构建了一个基于DNN的语音分离实验，IRM作为DNN的训练目标，使用最小均方误差（MSE）作为目标函数，DNN的输出层使用sigmoid激活函数。我们首先使用不带偏好权重的梯度下降算法对其进行优化，记录训练集和验证集的MSE随迭代次数的变化情况，然后使用带有偏好权重的梯度下降算法对其进行优化，同样记录训练集和验证集的MSE随迭代次数的变化情况。图4.4展示两种优化算法的训练过程。从图中可以看出，无论在训练集还是在验证集，偏好权重加权的梯度下降算法能够取得更小的误差损失。这表明偏好权重加权的梯度下降算法能够避免一些局部最小值，有潜力取得更优的优化结果。

4.3　第二阶段：多目标融合

第一阶段预测的IRM和目标语音听觉谱，尽管都能用来分离语音和噪声，

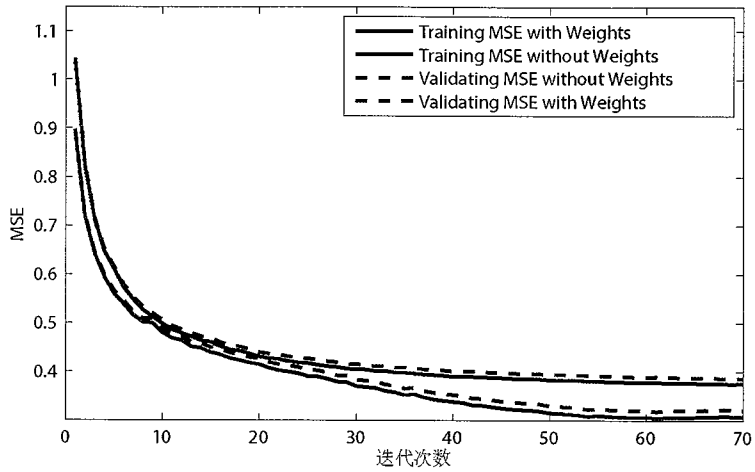


图 4.4: 最速梯度下降和偏好权重加权的梯度下降算法的训练过程。

但不同的分离目标侧重不同，各有优势，进一步将它们融合能够获得更好的分离性能。为此，我们使用DNN对第一阶段预测的IRM和目标语音听觉谱进行融合，从而得到最终的语音分离目标IRM。由于第一阶段预测的IRM和目标语音听觉谱损失了很多信息，因此第二阶段DNN的输入除了包括第一阶段预测的语音分离目标外还包括了原始的带噪听觉特征，它们共同作为DNN的输入来预测最终的IRM。利用最终估计的IRM我们能够获得更高感知质量的分离语音。

4.4 特征提取

从混合语音信号中提取合适的特征对于监督式的语音分离至关重要。我们首先使用64个gammtone听觉滤波器组将一维的时序信号分解为二维的时频信号。然后对每一个频带的gammtone滤波输出按照20ms窗长10ms窗移进行分帧处理，通过计算每一个时频单元的听觉能量我们能得到64维的gammtone听觉谱。除了gammtone听觉谱之外，我们还提取了振幅调制谱（AMS），相对谱变换感知线性预测系数（RASTA-PLP）和梅尔倒谱系数（MFCC） [170]。另外为了挖掘语音信号中的短时连续性，我们还计算了所有拼接特征的一阶差分，并且使用二阶自回归滑动平均模型（Auto-Regressive Moving Average，ARMA）沿着时间方向对所提取的特征进行平滑 [25]，最终形成一个拼接了gammtone听觉谱、AMS、RASTA-PLP和MFCC以及它们的差分的特征向量。

4.5 实验及其分析

4.5.1 数据集

我们使用“863计划”语音识别语料库对所提出的语音分离系统进行系统评估。“863计划”语音识别语料库分别包含了100个男性朗读者和100个女性朗读者朗读的100,000句纯净语音。作为一个先导性实验，我们在一个相对小的数据

集上进行实验验证。对于训练，我们从5个男性朗读者和5个女性朗读者的朗读语音中随机选择100句纯净语音，每人选择10句。然后以0dB和-5dB的信噪比随机和三种非语音噪声（babble, speech shaped noise 和 factory）进行混合。而对于测试，我们从同样的5个男性朗读者和5个女性朗读者中选择另外不同的50句朗读语音。然后分别以-10dB、-7dB、-5dB、-2dB和0dB的信噪比随机和6种非语音噪声（babble, speech shaped noise, factory, traffic, machine 和 cocktail [35]）进行混合。需要指出的是其中3种噪声是训练集里没有的，因此可以测试语音分离系统对不匹配噪音类型的泛化能力。

4.5.2 评价指标

我们使用短时客观可懂度（Short-Time Objective Intelligibility，STOI）[155]和信噪比（SNR）相对于原始带噪语音的绝对增益作为评价指标。另外我们也计算了语音质量的感知评价指标（Perceptual Evaluation of Speech Quality，PESQ）[138]来评估分离语音的感知质量。这些指标目前是语音分离领域的主流评价指标，认为这些指标的提升可以提高语音的可懂度和听觉质量。

4.5.3 模型与设置

在本章节中，我们比较了单目标的DNN、多目标的DNN [171]和DNN-NMF [173]等语音分离模型的性能，它们分别定义为“ST-DNN”、“MT-DNN”和“DNN-NMF”。对于所有语音分离模型，我们使用3隐层的DNN，每个隐含层的神经元个数依次为320、320和160，隐含层的激活函数为sigmoid。我们选择总共5帧的上下文特征作为DNN的输入。DNN的初始化通过DAE逐层预训练的方式 [162, 163]。然后利用小批量（mini-Batch）梯度下降算法对DNN进行训练，最大迭代次数设置为100，学习率从初始的0.1开始逐渐降低到0.001，在最开始的5代训练里使用0.5 的动量，之后动量保持为0.9。为了避免过拟合问题，我们对DNN的每个隐层使用dropout技术 [76]，dropout的比例设置为0.2。

ST-DNN的预测目标是全频带的IRM，MT-DNN的预测目标是全频带的IRM和目标语音听觉谱的联合分离目标，因此ST-DNN的输出层有64个神经元而MT-DNN的输出层有128个神经元。由于IRM的取值范围为$[0, 1]$，而语音听觉谱的取值非负但可能超过1，因此ST-DNN使用sigmoid作为输出层的激活函数，MT-DNN 使用sigmoid函数预测IRM，使用非负线性函数预测目标语音听觉谱。不同于ST-DNN和MT-DNN，DNN-NMF预测NMF事先推断的IRM的NMF激活系数，作为中间目标，通过非负线性组合，激活系数能够用来重构IRM，然后通过重构的IRM能够实现目标语音的分离。对于DNN-NMF我们使用文献 [173]的实验配置，设置NMF基的个数为128，使用IRM作为NMF的训练数据，同时拼接了上下文帧（左右各2帧），NMF的训练方式和文献 [173]完全相同。由于NMF基的激活系数是非负的而且并不一定在$[0, 1]$的范围内，因此DNN-NMF使用非负线性函数作为输出层的激活函数。

4.5.4 实验结果与分析

我们依次从以下几个方面验证了所提出的语音分离方法，包括IRM和目标语音听觉谱联合的多目标学习，偏好权重加权的梯度下降算法以及基于DAE的语音时空结构的学习。

首先，我们比较了单目标的ST-DNN和多目标的MT-DNN的语音分离性能。ST-DNN和MT-DNN的区别在于MT-DNN除了预测IRM之外还预测目标语音的听觉谱。需要说明的是尽管MT-DNN估计的IRM和目标语音听觉谱都能用来分离出最终的目标语音，但在这里呈现的语音分离结果是通过MT-DNN估计的IRM获得的。表4.1的第一行和第二行给出了ST-DNN和MT-DNN在输入信噪比为-5dB时所取得了语音分离性能。从表中可以看出，多目标的MT-DNN 无论在噪声匹配还是在噪声不匹配的条件下都明显优于单个目标的ST-DNN。这主要得益于多目标的DNN有效挖掘了IRM和目标语音听觉谱之间的相关性和互补性。

其次，我们评估了所提出偏好权重加权的梯度下降算法对多目标的MT-DNN的语音分离性能的影响。我们定义使用偏好权重加权的梯度下降算法优化的MT-DNN 为MT-DNN-PW-1。相比于MT-DNN-PW-1，MT-DNN使用最速梯度下降算法进行优化，并没有考虑不同目标的学习差异性。表4.1的第二行和第三行分别给出了MT-DNN和MT-DNN-PW-1在输入信噪比为-5dB时所取得了语音分离性能。相比于MT-DNN，MT-DNN-PW-1进一步提升语音分离的性能。这表明偏好权重加权的梯度下降算法考虑了不同学习目标的差异性，能够对多个目标进行更好地优化。

然后，我们评估了基于DAE的语音听觉特征和分离目标的时空结构学习是否能够提升语音分离的性能。本章小节4.2.1详细介绍了DAE的学习过程和基于DAE的语音分离网络构建过程。我们以MT-DNN-PW-1为例，定义基于DAE所构建的语音分离网络为MT-DNN-PW-2。MT-DNN-PW-1和MT-DNN-PW-2的网络结构、优化方法和训练配置完全相同。不同之处在于，MT-DNN-PW-1是通过DAE的逐层预训练初始化的，而MT-DNN-PW-2是通过连接两个训练好的DAE构建的，其中一个训练在输入的听觉特征，另一个训练在语音分离目标，两个DAE的连接是通过一个线性映射，如图4.2 所示。表4.1的第三行和第四行分别给出了MT-DNN-PW-1和MT-DNN-PW-2在输入信噪比为-5dB时所取得了语音分离性能。从表中可以看出，MT-DNN-PW-2在绝大数评价指标上要优于MT-DNN-PW-1。这一方面是由于DAE有效地学习到了语音听觉特征和语音分离目标的时空结构，另一方面相对于从输入特征开始的逐层预训练的方式，本文所提出的从输入特征和分离目标同步进行的预训练方式，既挖掘了输入特征的时空结构模式同时又挖掘了语音分离目标的时空结构模式。

表4.1的第五列给出了MT-DNN-PW-2预测的IRM和目标语音幅度谱进一步

表 4.1: 不同语音分离模型在输入SNR为-5dB时所取得语音分离性能。

System	Matched noise			Unmatched noise		
	gSTOI	gSNR(dB)	PESQ	gSTOI	gSNR(dB)	PESQ
ST-DNN	0.201	17.17	2.14	0.115	10.60	1.32
MT-DNN	0.211	17.67	2.22	0.124	10.75	1.36
MT-DNN-PW-1	0.216	18.13	2.24	0.124	10.62	1.35
MT-DNN-PW-2	0.221	18.69	2.21	**0.129**	**10.89**	**1.36**
TWO-STAGE	**0.231**	**19.41**	**2.39**	0.125	10.68	1.30
DNN-NMF	0.205	16.94	2.15	0.102	9.69	1.20

融合的结果，我们将其定义为“TWO-STAGE”。TWO-STAGE 以MT-DNN-PW-2的输出以及原始的带噪听觉特征作为输入，以DNN作为融合模型，输出目标为IRM。相对于MT-DNN-PW-2，TWO-STAGE在噪音匹配的条件下能够取得更好的语音分离性能，这表明对预测的多目标进一步融合能够提升语音分离的性能。但是在噪声不匹配的条件下，语音分离性能会微弱地下降，这可能是过拟合造成的。

最后，我们比较了对比系统DNN-NMF的语音分离性能。DNN-NMF试图通过NMF挖掘语音分离目标IRM的时空结构，从而实现提高语音分离性能的目的。相比于其他基于DNN的语音分离模型，DNN-NMF并没有直接预测语音分离目标IRM而是学习IRM的NMF激活系数，然后再通过DNN预测的激活系数利用NMF重构得到最终的IRM，从而实现语音和噪声的分离。表4.1的最后一行给出了DNN-NMF所取得的语音分离性能。DNN-NMF取得了比ST-DNN更优秀的性能，这表明挖掘语音分离目标的时空结构确实能够帮助语音分离，但相对于我们所提出的语音分离方法，DNN-NMF在所有评价指标上都处于劣势。一方面是由于我们考虑多个分离目标的互补性，另一方面我们同时挖掘了语音输入特征和语音分离目标的时空结构，而且由于DAE的深层次非线性结构相对于NMF浅层的线性结构具有更强的建模能力。同时我们也观察到DNN-NMF在噪声不匹配的条件下并没有取得很好的分离性能，这主要是由于DNN-NMF预测是中间目标，既存在DNN输出误差又存在NMF重构误差，导致最终IRM的重构对DNN的输出误差比较敏感。

测试数据的信噪比与训练数据的信噪比的不匹配可能会导致语音分离系统的性能下降，为此，我们比较了不同语音分离系统对输入信噪比的泛化性能。图4.5给出了不同语音分离系统在不同输入信噪比时取得了STOI增益，其中图4.5(a)是噪声匹配的条件下所取得语音分离性能，而图4.5(b)是噪声不匹配的条件下所取得语音分离性能。可以看出，我们所提出的方法在噪声匹配和噪声

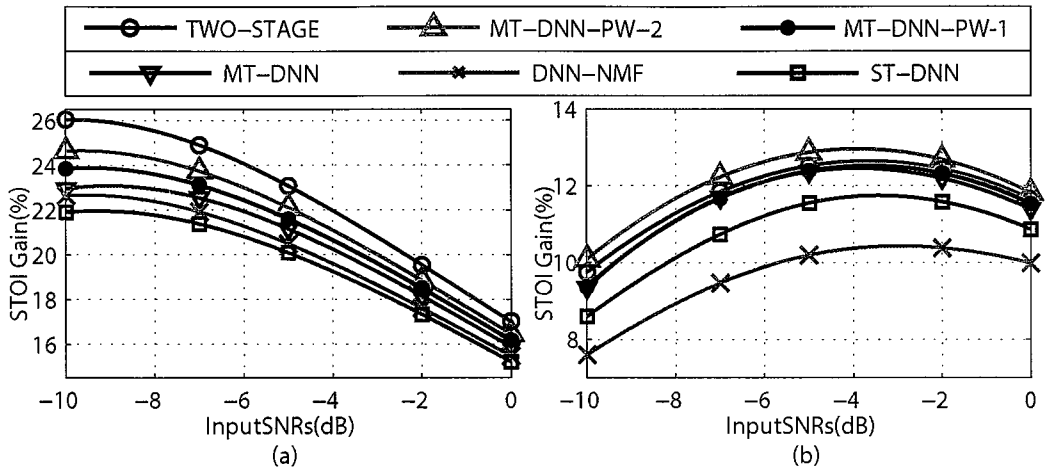


图 4.5: 不同语音分离模型在不同输入SNR条件下所取得的STOI绝对增益(%)。

不匹配的条件下对于各种输入信噪比都取得了最好的STOI增益。这主要是因为所提出的方法挖掘了输入听觉特征和输出分离目标的时空结构，而由于语音在时频域具有稀疏性，时空结构对信噪比能够保持一定的稳定性，因此能够对输入信噪比相对鲁棒。同时我们也考虑不同语音分离目标的互补性，避免了单个分离目标带来的评估误差风险。

4.6 本章小结

监督式语音分离通常学习一个从带噪听觉特征到理想掩蔽或目标语音频谱的映射函数。由于语音的产生机制和稀疏性，语音听觉特征和分离目标在时频域呈现了明显的时空结构，另外，语音分离的两类主要目标，理想时频掩蔽和目标语音频谱之间具有密切的联系和互补性。本章针对语音的时空结构和语音分离目标的互补性进行探索通过挖掘这些特性显著提升了语音分离的性能。

本章首先利用两个降噪自编码器以自学习的方式分别学习了带噪听觉特征和语音分离目标的时空结构，然后通过线性变换将两个训练好的降噪自编码器连接起来构建了语音分离网络，实验证明这种预训练方式能够显著提高语音分离的性能。其次我们也验证了IRM和目标语音听觉谱联合的分离目标要优于单个的IRM分离目标。考虑到多个目标的差异性，我们还提取出了偏好权重加权的梯度下降算法，实验证明该算法相对于最速梯度下降算法能够更好地优化多目标的深度神经网络。

第五章　基于DNN与NMF组合框架的语音分离方法

5.1　引言

由于语音的产生机制，语音信号在时间和频率上具有明显的相关性，导致语音信号在频谱上呈现了显著的时空结构，另外，又由于语言发音学的限制，所有的语音是由一些基本发音模式产生的，因此语音信号中隐含着一些基本的频谱结构模式，挖掘语音的频谱结构模式对于提高语音分离的性能具有重要意义。

语音分离旨在从带噪信号中提取出我们感兴趣的目标语音信号，它能很自然地表达为一个监督式的学习问题 [171]。深度学习是目前最令人兴奋的监督式学习算法，其层次化的非线性结构，使其具有强大的感知能力，能够学习极其复杂的映射关系。近年来，基于深度学习的语音分离得到国内外学者的广泛关注，取得了显著的性能提升，特别在非平稳噪声和较低信噪比条件下，相对于传统信号处理方法表现出了巨大的优势 [1,4,5,86,168,169]。

一个典型的监督式语音分离系统通常学习一个从带噪输入特征到分离目标的映射函数，最常用的语音分离目标包括理想时频掩蔽或目标语音频谱 [171]，许多研究表明它们能显著提升带噪语音的可懂度和感知质量 [8,89,94,167,193]。Wang等人利用分类模型或者回归模型从带噪听觉特征中学习理想二值或者浮值掩蔽 [63,171,175]。而Xu等人利用DNN直接学习从带噪频谱到目标语音频谱的映射函数 [193]。尽管这些方法相对于传统的语音增强方法 [113]取得了显著的性能提升，但它们简单地将语音分离表达成一个监督式学习问题，过度地依赖于大量监督数据的粗暴训练，忽略语音固有模式的挖掘，这也导致监督式语音分离面临着较大的泛化问题，即当遇到不匹配的声学环境时，比如训练数据中没有的噪声类型和输入信噪比，其性能可能会急剧下降。事实上，语音是由基本的频谱模式组成的，这些固有的基本频谱模式产生了各种内容的连续语音，因此语音可以看成是一个生成式的过程。此外，由于语音在时频域上是相对稀疏的，因此语音的基本频谱模式对各种声学环境保持了相对的稳定性。因此，利用生成式模型挖掘语音固有的基本频谱模式对于提高语音分离性能具有重要的价值。

非负矩阵分解（Nonnegative Matrix Factorization，NMF）是著名的表示学习技术，它能学习到非负数据中类似部件的局部表示，比如应用到人脸，NMF能学习到类似眼睛、鼻子、嘴等局部表示 [103,104]。在NMF框架中，非负数据被分解为基矩阵和激活矩阵的乘积。基矩阵由一系列非负基向量组成，基向量可以视为学习到的基本局部表示，利用这些基向量能够通过非负线性组合近似重构原始的非负数据。激活矩阵由对应基向量的激活系数组成，它们代表

原始非负数据是由哪些基向量构成的。如果将NMF应用到纯净语音幅度谱上，NMF能够学习到语音的基本频谱模式，这些频谱模式具有一定的物理感知意义，看起来像组成语音的发音音素 [129,151]。

NMF有许多重要的现实应用，已广泛应用到语音分离领域 [15,54,123,151,164]。如果将NMF应用到语音分离目标，比如理想时频掩蔽和目标语音频谱，NMF能学习语音分离目标的基本结构模式，这些结构模式能够重构语音分离目标。Williamson等人作为后处理将NMF应用到DNN分离的目标语音频谱上来进一步提高分离语音的质量 [183]。文献 [173]将NMF应用到开方的理想浮值掩蔽（IRM）上来发掘IRM的时空结构，DNN 用来学习IRM 的NMF激活系数，然后利用学习到的激活系数通过非负线性组合重构出IRM，从而实现目标语音和噪声的分离。而文献 [92] 将NMF应用到目标语音的幅度谱上，同样用DNN学习目标语音幅度谱的NMF激活系数，然后通过非负线性重构得到目标语音的幅度谱，从而实现目标语音和噪声的分离。尽管这些方法都利用了NMF挖掘了语音分离目标的时空结构模式，但基于DNN的激活系数估计和NMF重构是分步进行的，DNN优化的是NMF的重构激活系数并不是实际的语音分离目标，这可能导致激活系数估计的误差和NMF重构误差相互叠加造成语音分离性能下降。

综上，将DNN与NMF联合为一个有机的整体可能是更为可取的策略。本章我们提出了DNN与NMF联合协作的组合框架来实现语音分离。NMF用来从大量纯净的语音和噪声频谱中学习说话人无关的语音基向量和噪声类型无关的噪声基向量。然后学习到的语音基向量和噪声基向量作为NMF重构层融入到DNN中来直接重构目标语音和噪声的频谱。一方面，NMF挖掘了语音和噪声的基本频谱结构，另一方面，DNN直接实现了混合语音中语音和噪声的分离，优化的是一个最终的分离目标，而并没有通过预测NMF的激活系数来实现语音分离，因此相比于之前类似的方法，将激活系数估计和NMF重构分步进行，本章所提出的方法将NMF和DNN 有机地融合为一个整体，能够在一个模型里同时利用DNN和NMF的能力，并且避免了分步方法的估计误差多次累计的问题。除此之外，我们还探索了一个带有稀疏约束和重构约束的区分式的训练目标，该目标能够有效控制噪声残留和语音畸变。系统的实验证明本章所提出的方法要优于之前的语音分离方法。

语音信号在时间和频率上的相关性导致语音信号具有明显的时空结构，针对语音频谱的基本频谱模式，本章提出利用NMF挖掘了语音的基本频谱模式，同时将NMF 和DNN融合为一个有机的整体，充分发挥了DNN和NMF的优势，提升了语音分离的性能。在接下来的章节里，我们首先回顾了基于NMF语音分离，接着介绍了NMF 与DNN 组合的语音分离框架，然后针对NMF和DNN组合的语音分离框架提出了带有稀疏和重构约束的区分式语音分离目标，最后从多个方面对本文所提出的语音分离方法进行评估和讨论。

5.2 问题定义

我们定义$s(k)$和$n(k)$分别是纯净的语音信号和噪声信号，其中k是采样点坐标。如果我们仅考虑加性噪声的情况，那么混合信号可以通过直接将$s(k)$和$n(k)$相加得到，如公式（5.1）所示。

$$x(k) = s(k) + n(k). \tag{5.1}$$

在语音信号处理领域，短时傅里叶变换（STFT）是最常用的时频分解技术。当我们对时域信号进行STFT时，我们能获得它们的时频表示。我们定义$X(f,t)$、$Y_s(f,t)$和$Y_n(f,t)$分别是$x(k)$、$s(k)$和$n(k)$的复数STFT系数。其中f和t分别表示频率和时间帧坐标。在STFT域，公式（5.1）可以表达为如下公式：

$$X(f,t) = Y_s(f,t) + Y_n(f,t). \tag{5.2}$$

假设语音和噪声是相互独立的，由于语音的稀疏性，混合信号的STFT幅度谱可以通过以下公式近似得到 [109,110]:

$$|X(f,t)| \approx |Y_s(f,t)| + |Y_n(f,t)|, \tag{5.3}$$

其中$|\cdot|$是复数域的绝对值运算。为了简化符号表达，我们将公式（5.3）重写为矩阵的形式：

$$\mathbf{X} \approx \mathbf{Y}_s + \mathbf{Y}_n, \tag{5.4}$$

其中$\mathbf{X} \in \mathbb{R}_+^{F\times T}$，$\mathbf{Y}_s \in \mathbb{R}_+^{F\times T}$和$\mathbf{Y}_n \in \mathbb{R}_+^{F\times T}$分别定义了混合语音信号，纯净语音信号和噪声信号的STFT幅度谱，而F和T分别是总的频点数和时间帧数。

语音分离的任务就是要从混合语音信号$x(k)$中获得目标语音信号$s(k)$的估计$\hat{s}(k)$。如果不考虑相位的影响，语音分离任务可视为目标语音幅度谱的估计问题。尽管考虑相位信息能够获得更好的语音分离性能 [50]，但许多研究表明利用估计的语音幅度谱和原始混合信号的相位能够获得较高质量的分离语音波形 [192]。因此在我们的语音分离系统中，我们并没有考虑相位的估计而只预测了目标语音的幅度谱。一旦获得语音的幅度谱，我们利用混合信号的相位通过短时傅里叶逆变换（ISTFT）即可获得分离语音的波形信号。

5.3 基于NMF的语音分离

在NMF的框架下，混合信号的幅度谱能够近似为非负基矩阵和非负激活矩阵的乘积，如下所示：

$$\mathbf{X} \approx \mathbf{B}_x\mathbf{A}_x, \tag{5.5}$$

其中$\mathbf{B}_x \in \mathbb{R}_+^{F\times R}$是非负基矩阵，而$\mathbf{A}_x \in \mathbb{R}_+^{R\times T}$是非负激活系数矩阵。$\mathbf{B}_x$由$R$个非负基向量组成，每个基向量对应于矩阵的每一列。每一个基向量表示组成语

音或噪声的基本频谱。因为混合信号中既包含语音成分又包含噪声成分，因此，混合信号的基矩阵$\mathbf{B}_x$由两部分构成，一部分和语音相关另一部分和噪声相关。因此我们可以将$\mathbf{B}_x$表示成两个子矩阵的拼接，如下式所示：

$$\mathbf{B}_x = \begin{bmatrix} \mathbf{B}_s & \mathbf{B}_n \end{bmatrix}, \tag{5.6}$$

其中$\mathbf{B}_s \in \mathbb{R}_+^{F\times R_s}$是和语音相关的基矩阵，而$\mathbf{B}_n \in \mathbb{R}_+^{F\times R_n}$是和噪声相关的基矩阵。$R_s$和$R_n$分别是语音基向量和噪声基向量的个数，且满足$R_s + R_n = R$。对应地，激活矩阵也可以分解为两个子矩阵的拼接，如下所示：

$$\mathbf{A}_x = \begin{bmatrix} \mathbf{A}_s^{\mathrm{T}} & \mathbf{A}_n^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}, \tag{5.7}$$

其中$\mathbf{A}_s \in \mathbb{R}_+^{R_s\times T}$是对应于$\mathbf{B}_s$的激活矩阵，而$\mathbf{A}_n \in \mathbb{R}_+^{R_n\times T}$是对应于$\mathbf{B}_n$的激活矩阵。T定义了矩阵转置运算。将公式（5.6）和（5.7）带入公式（5.5）可得，

$$\begin{aligned} \mathbf{X} \approx \mathbf{B}_x\mathbf{A}_x &= \begin{bmatrix} \mathbf{B}_s & \mathbf{B}_n \end{bmatrix} \begin{bmatrix} \mathbf{A}_s^{\mathrm{T}} & \mathbf{A}_n^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \\ &= \mathbf{B}_s\mathbf{A}_s + \mathbf{B}_n\mathbf{A}_n = \hat{\mathbf{Y}}_s + \hat{\mathbf{Y}}_n, \end{aligned} \tag{5.8}$$

其中$\hat{\mathbf{Y}}_s \in \mathbb{R}_+^{F\times T}$和$\hat{\mathbf{Y}}_n \in \mathbb{R}_+^{F\times T}$分别是混合信号中的语音成分和噪声成分。它们可以通过$\mathbf{B}_s\mathbf{A}_s$和$\mathbf{B}_n\mathbf{A}_n$近似评估得到。为了进一步平滑分离结果，类维纳滤波技术通常被用来约束分离的语音$\hat{\mathbf{Y}}_s$和噪声$\hat{\mathbf{Y}}_n$之和等于混合信号 [87, 123, 182]，如公式（5.9）所示。

$$\begin{aligned} \tilde{\mathbf{Y}}_s &= \frac{\hat{\mathbf{Y}}_s}{\hat{\mathbf{Y}}_s + \hat{\mathbf{Y}}_n} \otimes \mathbf{X}, \\ \tilde{\mathbf{Y}}_n &= \frac{\hat{\mathbf{Y}}_n}{\hat{\mathbf{Y}}_s + \hat{\mathbf{Y}}_n} \otimes \mathbf{X}, \end{aligned} \tag{5.9}$$

其中$\tilde{\mathbf{Y}}_s \in \mathbb{R}_+^{F\times T}$和$\tilde{\mathbf{Y}}_n \in \mathbb{R}_+^{F\times T}$是目标语音幅度谱和噪声幅度谱的最终估计。这里的除法运算表示点除运算，而$\otimes$定义点乘运算。最后分离语音的波形$\hat{s}(k)$和分离噪声的波形$\hat{n}(k)$ 通过ISTFT得到。

在基于NMF的语音分离系统中，$\mathbf{B}_s$和$\mathbf{B}_n$通常事先从大量的纯净语音和噪声训练得到。而混合信号的基矩阵可以通过拼接$\mathbf{B}_s$和$\mathbf{B}_n$得到，如公式（5.6）所示。固定$\mathbf{B}_x$，激活矩阵$\mathbf{A}_x$可以通过最小化损失函数$D(\mathbf{X}|\mathbf{B}_x\mathbf{A}_x)$得到。损失函数$D(\mathbf{X}|\mathbf{B}_x\mathbf{A}_x)$测度了$\mathbf{X}$ 和$\mathbf{B}_x\mathbf{A}_x$的差异。最常用的损失函数包括欧拉距离（Euclidean Distance） [104]、KL散度（Kullback-Leibler Divergence，KL）[23, 104]和IS散度（Itakura-Saito Divergence，IS） [51]。对于语音分离，KL散度是最常用损失函数 [92]，它的表达形式如下：

$$D_{KL}(\mathbf{X}|\mathbf{B}_x\mathbf{A}_x) = \|\mathbf{X} \otimes \log(\mathbf{X}/\mathbf{B}_x\mathbf{A}_x) - \mathbf{X} + \mathbf{B}_x\mathbf{A}_x\|_F, \tag{5.10}$$

其中$\|\cdot\|_F$表示F范数（Frobenius Norm），符号/定义点除运算，对数运算作用到矩阵的每个元素。通常损失函数的最小化可以通过乘法更新规则迭代进行 [104]，

如下所示:

$$\mathbf{A}_x \leftarrow \mathbf{A}_x \otimes \frac{\mathbf{B}_x^{\mathrm{T}} \frac{\mathbf{X}}{\mathbf{B}_x\mathbf{A}_x}}{\mathbf{B}_x^{\mathrm{T}}\mathbf{1}}, \tag{5.11}$$

$$\mathbf{B}_x \leftarrow \mathbf{B}_x \otimes \frac{\frac{\mathbf{X}}{\mathbf{B}_x\mathbf{A}_x}\mathbf{A}_x^{\mathrm{T}}}{\mathbf{1}\mathbf{A}_x^{\mathrm{T}}}, \tag{5.12}$$

其中$\mathbf{1}$是元素全为1的$F \times T$矩阵。矩阵除法表示点除运算。给定一个测试数据，保持基矩阵$\mathbf{B}_x$不变，激活矩阵$\mathbf{A}_x$可以通过公式（5.11）计算得到，而基矩阵$\mathbf{B}_x$可以通过拼接$\mathbf{B}_s$和$\mathbf{B}_n$得到。对应语音的激活矩阵$\mathbf{A}_s$和对应噪声激活矩阵$\mathbf{A}_n$可以根据公式（5.7）划分$\mathbf{A}_x$得到。那么语音和噪声的分离可以通过公式（5.9）计算。

5.4　基于NMF的时空结构挖掘

NMF是著名的字典学习技术 [104]。通过非负约束它能够学习到非负数据中的潜在局部表示。在语音分离中，NMF通常用来学习语音和噪声的基本频谱模式$\mathbf{B}_s$和$\mathbf{B}_n$，它们通常通过大量的纯净语音和噪声训练得到。最常用的方法是将NMF应用到大量纯净语音和噪声的幅度谱上，从而获得语音和噪声的基本频谱模式，比如 [129, 151, 164, 181]。

我们定义$\mathbf{S} \in \mathbb{R}_+^{F\times T}$和$\mathbf{N} \in \mathbb{R}_+^{F\times T}$分别是纯净语音和噪声的幅度谱。它们通过拼接训练数据中所有的语音帧或噪声帧得到。在训练阶段，NMF模型$\mathbf{B}_s\mathbf{A}_s$和$\mathbf{B}_n\mathbf{A}_n$分别通过最小化$D(\mathbf{S}|\mathbf{B}_s\mathbf{A}_s)$和$D(\mathbf{N}|\mathbf{B}_n\mathbf{A}_n)$训练得到。乘法更新法则通常被用来优化目标函数。除了经典的NMF（typical NMF，TNMF）[103]之外，许多NMF的变体被相续提出，主要包括，稀疏NMF（Sparse NMF，SNMF）[101]、区分式NMF（Discriminative NMF，DNMF）[182]以及卷积NMF（Convolutive NMF，CNMF）[129, 151]，它们分别考虑了激活系数的稀疏性，语音的基向量和噪声的基向量之间的区分性以及非负数据的短时连续性，下面我们对它们依次做一个简单的介绍。

5.4.1　SNMF

得益于NMF的非负约束，NMF能够得到非负数据的稀疏表示 [103]。稀疏性对挖掘数据的局部表示至关重要 [103]。为了进一步提高表示学习的稀疏性，我们在学习基向量的时候对激活矩阵增加了一个稀疏约束，即要求重构原始数据的时候仅有少数的基向量被激活。不同于TNMF，SNMF的优化目标除了包括重构损失之外还包括激活系数的ℓ_1-norm的约束。因此，SNMF求解的是下面的优化问题：

$$\mathbf{B}_s, \mathbf{A}_s = \underset{\mathbf{B}_s,\mathbf{A}_s}{\arg\min}(D(\mathbf{S}|\mathbf{B}_s\mathbf{A}_s) + \mu\|\mathbf{A}_s\|_1), \tag{5.13}$$

其中$\|\cdot\|_1$定义了ℓ_1-norm运算。因为$\mathbf{A}_s$的ℓ_1-norm的稀疏约束对于$\mathbf{B}_s\mathbf{A}_s$并不是尺度不变的，也就是说可以无限最小化$\|\mathbf{A}_s\|_1$而不影响$\mathbf{B}_s\mathbf{A}_s$的值来最小化优化目

标，这是因为$\mathbf{B}_s\mathbf{A}_s = (\alpha\mathbf{B}_s)\left(\frac{1}{\alpha}\mathbf{A}_s\right)$，这里$\alpha > 1$。为了避免缩放的不确定性，在优化目标函数的时候，基向量需要固定在确定的尺度上。一个常用的方法是在优化目标中增加一个对基向量归一化的约束，因此最终的优化目标可以重写如下：

$$\begin{aligned}\mathbf{B},\mathbf{A} &= \arg\min_{\mathbf{B},\mathbf{A}}(D(\mathbf{S}|\widetilde{\mathbf{B}}\mathbf{A}) + \mu|\mathbf{A}|_1), with \\ \widetilde{\mathbf{B}} &= \left[\begin{array}{ccccc}\frac{\mathbf{b}_1}{\|\mathbf{b}_1\|} & \cdots & \frac{\mathbf{b}_i}{\|\mathbf{b}_i\|} & \cdots & \frac{\mathbf{b}_{R_s}}{\|\mathbf{b}_{R_s}\|}\end{array}\right],\end{aligned} \tag{5.14}$$

其中$\widetilde{\mathbf{B}}$是$\mathbf{B}$按照列向量归一化的版本。$\mathbf{b}_i \in \mathbb{R}_+^{F\times 1}$是第$i$个基向量，对应于$\mathbf{B}$的第$i$列。为了避免符号的混淆，我们去掉了下标$s$。对于一般化的KL散度优化目标，公式（5.14）可以通过乘法更新法则优化，具体优化算法可以查阅文献 [101]。

5.4.2 DNMF

在基于NMF的语音分离中，混合信号的基矩阵是通过拼接事先在纯净语音数据和噪声数据上训练的语音基矩阵和噪声基矩阵得到的。在真实声学环境中，由于语音和噪声存在一定的相关性，因此语音基向量和噪声基向量不可避免地存在一定的重叠，这会导致在测试阶段基于事先训练的基向量推断的激活系数将会不同于通过训练目标获得的激活系数。这主要是因为测试阶段激活系数的推断是基于事先训练的语音基向量和噪声基向量重构混合信号得到的，而训练阶段是基于语音的基向量或噪声的基向量单独重构语音和噪声得到的。显然，通过事先单独训练得到的语音和噪声的基向量拼接的混合信号基向量并不适合于测试阶段的混合信号。为了缓解这个问题，Weninger等人提取了一种区分式的方法来获得语音和噪声的基向量 [182]。这种方法在训练语音和噪声的基向量时考虑了测试阶段重构混合信号的优化目标，如下所示：

$$\begin{aligned}&\mathbf{B}_s,\mathbf{B}_n = \arg\min_{\mathbf{B}_s,\mathbf{B}_n}(\gamma_s D(\mathbf{S}|\mathbf{B}_s\widehat{\mathbf{A}}_s(\mathbf{X},\mathbf{B}_s)) + \gamma_n D(\mathbf{N}|\mathbf{B}_n\widehat{\mathbf{A}}_n(\mathbf{X},\mathbf{B}_n))), \\ &where \quad \mathrm{A}_x = \arg\min_{\mathbf{A}_x} D(\mathbf{X}|\widetilde{\mathbf{B}}_x\mathrm{A}_x), \\ &and \quad \mathrm{A}_x = \left[\widehat{\mathbf{A}}_s(\mathbf{X},\mathbf{B}_s); \widehat{\mathbf{A}}_n(\mathbf{X},\mathbf{B}_n)\right];\ \widetilde{\mathbf{B}}_x = \left[\begin{array}{cc}\widetilde{\mathbf{B}}_s & \widetilde{\mathbf{B}}_n\end{array}\right].\end{aligned} \tag{5.15}$$

这里$\widetilde{\mathbf{B}}_s$和$\widetilde{\mathbf{B}}_n$分别是$\mathbf{B}_s$和$\mathbf{B}_n$的归一化版本。γ_s和γ_n表示语音和噪声重构的相对重要性，且它们满足$\gamma_s + \gamma_n = 1$，根据不同的应用侧重的不同，比如有些应用关注语音的重构，那么γ_s就相对大一些。不同于单独学习语音基向量和噪声基向量，DNMF最小化一个整体的目标函数来联合学习语音和噪声的基向量，这样学习的语音和噪声的基向量具有一定的区分性，能够一定程度缓解语音基向量和噪声基向量重叠的问题。区分式的语音和噪声基向量既考虑语音和噪声的重构目标，也考虑它们拼接起来作为混合信号基向量的重构目标，因此公式（5.15）的优化非常困难。为此，Weninger提出一种简化的优化方法，具体可以查看文献 [182]。

5.4.3 CNMF

由于语音的产生机制，语音具有明显的短时连续性。然而传统的NMF忽略了语音的这些特性。为了对语音的短时相关性进行建模，Paris等人将传统的NMF推广到卷积的形式，提出了卷积的NMF（CNMF）。在传统的NMF中，非负矩阵$\mathbf{S}$是通过基矩阵$\mathbf{B}_s$和对应的激活矩阵$\mathbf{A}_s$的乘积得到的，即：$\mathbf{S} \approx \mathbf{B}_s\mathbf{A}_s$。而在CNMF中，非负矩阵是通过一系列连续的非负基向量和对应的非负激活系数卷积得到的。如下所示：

$$\mathbf{S} \approx \sum_{t=0}^{T_o-1} \mathbf{B}_s(t) \overset{t\rightarrow}{\mathbf{A}}_s, \tag{5.16}$$

其中，$\mathbf{B}_s(t) \in \mathbb{R}_+^{F\times R_s}$是一系列时间上连续的基向量。$T_o$是基向量的序列长度。$\overset{t\rightarrow}{(\cdot)}$定义了按列向右平移$t$步的操作，最左边用0填充。对应的$\overset{\leftarrow t}{(\cdot)}$定义向左平移，具体操作如下所示：

$$\begin{aligned}
\mathbf{A} &= \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix} & \overset{0\rightarrow}{\mathbf{A}} &= \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix} \\
\overset{1\rightarrow}{\mathbf{A}} &= \begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 5 & 6 & 7 \end{bmatrix} & \overset{2\rightarrow}{\mathbf{A}} &= \begin{bmatrix} 0 & 0 & 1 & 2 \\ 0 & 0 & 5 & 6 \end{bmatrix} \\
\overset{\leftarrow 0}{\mathbf{A}} &= \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix} & \overset{\leftarrow 1}{\mathbf{A}} &= \begin{bmatrix} 2 & 3 & 4 & 0 \\ 6 & 7 & 8 & 0 \end{bmatrix} \\
\overset{\leftarrow 2}{\mathbf{A}} &= \begin{bmatrix} 3 & 4 & 0 & 0 \\ 7 & 8 & 0 & 0 \end{bmatrix} & \overset{\leftarrow 3}{\mathbf{A}} &= \begin{bmatrix} 4 & 0 & 0 & 0 \\ 8 & 0 & 0 & 0 \end{bmatrix}.
\end{aligned} \tag{5.17}$$

为了评估基矩阵$\mathbf{B}_s(t)$和激活矩阵$\mathbf{A}_s$来近似$\mathbf{S}$，CNMF通常使用KL散度作为损失函数，对应的，针对CNMF的乘法更新算法在文献 [129, 151]中被提出。

5.5 DNN和NMF的组合框架

语音基向量和噪声基向量包含了构成语音和噪声的基本成分，通过非负线性组合，它们能够重构出纯净的语音和噪声。对于语音分离，语音基矩阵$\mathbf{B}_s$和噪声基矩阵$\mathbf{B}_n$可以事先通过大量的纯净语音和噪声训练得到。在语音基向量张成的空间里，可以训练DNN直接从带噪数据中重构目标语音，同样地，在噪声基向量张成的空间里，DNN能够重构出噪声。本章我们提出了DNN与NMF联合协作的组合框架来实现语音和噪声的分离，这个组合框架既利用了NMF局部表示学习的能力又充分发挥了DNN超强的非线性映射学习能力。我们定义所提出的DNN与NMF的组合模型为“Joint-DNN-NMF”，下面将对其进行详细的介绍。

5.5.1 网络结构

所提出的Joint-DNN-NMF是DNN的一个变体，图5.1给出了它的网络结构。和DNN类似，Joint-DNN-NMF由输入层，隐含层和输出层构成。但不同的

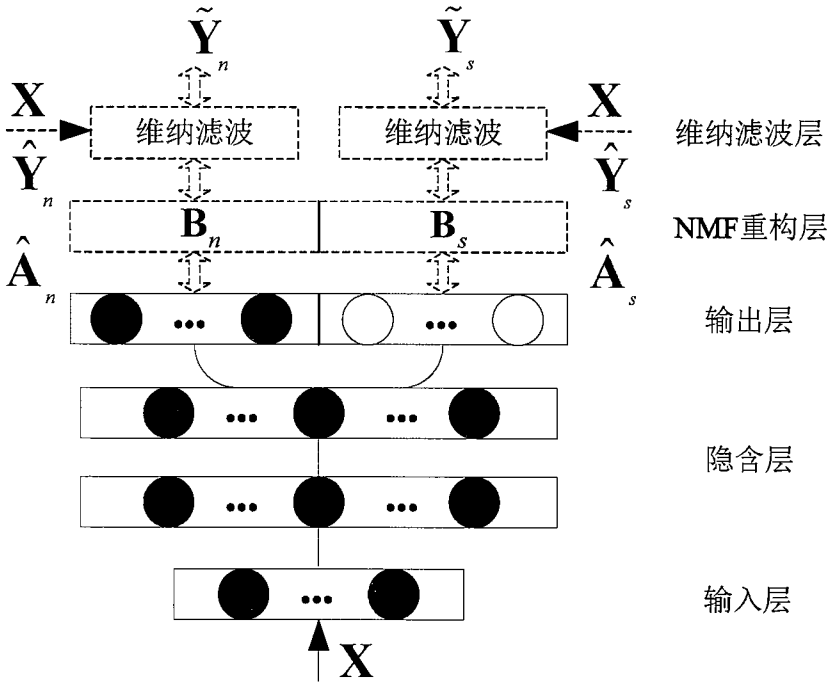


图 5.1: Joint-DNN-NMF的网络结构。

是Joint-DNN-NMF的输出层还紧跟一个NMF重构层，这个重构层的连接权重由事先训练的语音基向量$\mathbf{B}_s$和噪声基向量$\mathbf{B}_n$初始化得到。神经网络的原始输出层的输出作为激活系数，通过NMF重构层得到重构的目标语音和噪声的频谱$\hat{\mathbf{Y}}_s$和$\hat{\mathbf{Y}}_n$。另外，为了进一步平滑分离的语音和噪声，我们对NMF重构的语音和噪声进一步进行维纳滤波，从而限制重构的语音和噪声之和等于混合语音，维纳滤波可以视为Joint-DNN-NMF的维纳滤波层。需要强调的是NMF重构层和维纳滤波层尽管可以视为DNN的额外层，但它们是确定性的运算过程，并没有权重需要优化，它们产生的输出会用来计算误差从而来优化整个网络权重。

5.5.2 训练目标

语音分离的目的是实现混合信号中的语音和噪声的分离。从混合信号中分离语音和分离噪声具有很强的互补性，在一个模型中对它们同时建模可能会提高语音分离的性能 [87]。另外，利用分离的语音和噪声可以构建区分式的训练目标来进一步控制噪声残留和语音畸变。在本章所提出的语音分离系统中，我们同时预测语音和噪声的频谱，如图5.1。

给定分离语音和噪声幅度谱估计$\tilde{\mathbf{Y}}_s$和$\tilde{\mathbf{Y}}_n$，一方面我们希望尽可能增大预测频谱和对应的目标频谱的相似性，另一方面我们希望分离的语音中噪声残留尽可能小而分离噪声中语音残留尽可能小，因此我们希望训练目标能够增大预测频谱和其他声源频谱的差异性，如公式（5.18）所示。也就是说，通过增大分离语音和分离噪声之间的区分性来降低分离语音中的噪声残留和分离噪声中的语

音残留，从而实现控制噪声残留和语音畸变的目的。

$$J=\frac{1}{2}(\left\|\mathbf{Y}_s-\tilde{\mathbf{Y}}_s\right\|_2^2+\left\|\mathbf{Y}_n-\tilde{\mathbf{Y}}_n\right\|_2^2)-\frac{\lambda}{2}(\left\|\mathbf{Y}_s-\tilde{\mathbf{Y}}_n\right\|_2^2+\left\|\mathbf{Y}_n-\tilde{\mathbf{Y}}_s\right\|_2^2), \tag{5.18}$$

其中λ指定了目标函数中对应项的相对重要性，一般通过实验选择得到。

为了平滑分离结果，我们通常对重构的语音和噪声进行维纳滤波，从而得到最终的分离语音和噪声$\tilde{\mathbf{Y}}_s$和$\tilde{\mathbf{Y}}_n$，如公式（5.9）所示。当Joint-DNN-NMF的NMF重构层执行的是非负线性组合时，NMF的重构过程如下所示：

$$\hat{\mathbf{Y}}_s=\mathbf{B}_s\hat{\mathbf{A}}_s;\quad \hat{\mathbf{Y}}_n=\mathbf{B}_n\hat{\mathbf{A}}_n, \tag{5.19}$$

那么，

$$\begin{aligned}\tilde{\mathbf{Y}}_s&=\frac{\hat{\mathbf{Y}}_s}{\hat{\mathbf{Y}}_s+\hat{\mathbf{Y}}_n}\otimes\mathbf{X}=\frac{\mathbf{B}_s\hat{\mathbf{A}}_s}{\mathbf{B}_s\hat{\mathbf{A}}_s+\mathbf{B}_n\hat{\mathbf{A}}_n}\otimes\mathbf{X},\\ \tilde{\mathbf{Y}}_n&=\frac{\hat{\mathbf{Y}}_n}{\hat{\mathbf{Y}}_s+\hat{\mathbf{Y}}_n}\otimes\mathbf{X}=\frac{\mathbf{B}_n\hat{\mathbf{A}}_n}{\mathbf{B}_s\hat{\mathbf{A}}_s+\mathbf{B}_n\hat{\mathbf{A}}_n}\otimes\mathbf{X},\end{aligned} \tag{5.20}$$

其中$\hat{\mathbf{A}}_s\in\mathbb{R}_+^{R_s\times T}$和$\hat{\mathbf{A}}_n\in\mathbb{R}_+^{R_n\times T}$是DNN的输出，可以视作NMF重构语音频谱和噪声频谱的激活系数。

而当Joint-DNN-NMF的NMF重构层执行的是非负卷积运算时，NMF的重构如下所示

$$\hat{\mathbf{Y}}_s=\sum_{t=0}^{T_o-1}\mathbf{B}_s(t)\overset{t\rightarrow}{\hat{\mathbf{A}}}_s;\quad \hat{\mathbf{Y}}_n=\sum_{t=0}^{T_o-1}\mathbf{B}_n(t)\overset{t\rightarrow}{\hat{\mathbf{A}}}_n, \tag{5.21}$$

那么，

$$\begin{aligned}\tilde{\mathbf{Y}}_s&=\frac{\hat{\mathbf{Y}}_s}{\hat{\mathbf{Y}}_s+\hat{\mathbf{Y}}_n}\otimes\mathbf{X}=\frac{\sum_{t=0}^{T_o-1}\mathbf{B}_s(t)\overset{t\rightarrow}{\hat{\mathbf{A}}}_s}{\sum_{t=0}^{T_o-1}(\mathbf{B}_s(t)\overset{t\rightarrow}{\hat{\mathbf{A}}}_s+\mathbf{B}_n(t)\overset{t\rightarrow}{\hat{\mathbf{A}}}_n)}\otimes\mathbf{X},\\ \tilde{\mathbf{Y}}_n&=\frac{\hat{\mathbf{Y}}_n}{\hat{\mathbf{Y}}_s+\hat{\mathbf{Y}}_n}\otimes\mathbf{X}=\frac{\sum_{t=0}^{T_o-1}\mathbf{B}_n(t)\overset{t\rightarrow}{\hat{\mathbf{A}}}_n}{\sum_{t=0}^{T_o-1}(\mathbf{B}_s(t)\overset{t\rightarrow}{\hat{\mathbf{A}}}_s+\mathbf{B}_n(t)\overset{t\rightarrow}{\hat{\mathbf{A}}}_n)}\otimes\mathbf{X}.\end{aligned} \tag{5.22}$$

从公式（5.20）和（5.22）可以看出实现语音和噪声的分离是通过语音基向量和噪声基向量非负线性重构得到的。在重构阶段，一般来说希望被激活的基向量越少越好。为此我们对神经网络的输出$\hat{\mathbf{A}}_s$和$\hat{\mathbf{A}}_n$施加ℓ_1的稀疏约束，以使得重构语音和噪声时激活系数尽可能稀疏。然而因为ℓ_1的系数约束并不是尺度不变的，正如公式（5.23）所示，可以通过对$\hat{\mathbf{A}}_s$和$\hat{\mathbf{A}}_n$施加一个极小的尺度因子来最小化$\hat{\mathbf{A}}_s$和$\hat{\mathbf{A}}_n$而不影响维纳滤波输出。

$$\frac{\mathbf{B}_s\hat{\mathbf{A}}_s}{\mathbf{B}_s\hat{\mathbf{A}}_s+\mathbf{B}_n\hat{\mathbf{A}}_n}\otimes\mathbf{X}=\frac{\mathbf{B}_s(\frac{1}{\alpha}\hat{\mathbf{A}}_s)}{\mathbf{B}_s(\frac{1}{\alpha}\hat{\mathbf{A}}_s)+\mathbf{B}_n(\frac{1}{\alpha}\hat{\mathbf{A}}_n)}\otimes\mathbf{X}, \tag{5.23}$$

其中$\alpha > 1$。为了避免尺度不确定性问题，为此我们引入NMF重构约束，如下所示：

$$\begin{aligned} J=&\frac{1}{2}(\left\|\mathbf{Y}_s-\tilde{\mathbf{Y}}_s\right\|_2^2+\left\|\mathbf{Y}_n-\tilde{\mathbf{Y}}_n\right\|_2^2)-\frac{\lambda}{2}(\left\|\mathbf{Y}_s-\tilde{\mathbf{Y}}_n\right\|_2^2+\left\|\mathbf{Y}_n-\tilde{\mathbf{Y}}_s\right\|_2^2) \\ &+\frac{\gamma}{2}(\left\|\mathbf{Y}_s-\hat{\mathbf{Y}}_s\right\|_2^2+\left\|\mathbf{Y}_n-\hat{\mathbf{Y}}_n\right\|_2^2)+\mu(\left\|\hat{\mathbf{A}}_s\right\|_1+\left\|\hat{\mathbf{A}}_n\right\|_1), \end{aligned} \tag{5.24}$$

其中，λ，γ和μ定义了最终优化目标各项的相对重要性，它们可以通过实验选择得到。

5.5.3 相关工作

本章所提出的Joint-DNN-NMF不同于文献 [60, 92, 173, 184, 186]的相关工作。文献 [186]和 [60]将NMF作为后处理应用到基于DNN的语音分离输出上来进一步提高分离语音的质量。在文献 [173] 和 [184]中，Wang和Williamson等人使用NMF对开方的IRM进行建模来挖掘IRM的时空结构，他们使用DNN 学习IRM的NMF的激活系数，尽管DNN估计的激活系数能够用来重构IRM从而实现语音和噪声的分离，但是DNN优化的是NMF激活系数而不是实际的分离目标，整个语音分离过程是分步进行。使用类似的思路，Kang等人使用NMF对目标语音幅度谱进行建模，然后使用DNN学习目标语音幅度谱的NMF激活系数，最后通过NMF重构来合成最终的目标语音幅度谱。尽管Kang等人所提出的方法和本章所提出的方法都是对目标语音幅度谱进行建模，但是DNN估计和NMF重构依然是分开进行的，并没有统一到一个框架下。除此之外，我们还对DNN输出的激活系数施加了稀疏约束，确保NMF重构的时候只有少数基向量被激活，以提高语音分离的性能。

5.6 实验及其分析

在这一小节，我们系统评估了所提出DNN与NMF联合协作的组合语音分离框架。为了简化分析和比较，我们并没有探索NMF与其他神经网络结构的组合，但是这并不意味着所提出的语音分离框架只能是NMF与DNN的组合。事实上，很容易将所提出的组合框架应用到NMF与其他更先进的神经网络的组合，比如长短时记忆网络（Long Short-Term Memory，LSTM） [77]和卷积神经网络等（Convolutional Neural Network，CNN） [102]。

5.6.1 数据集

为了评估所提出的语音分离系统，我们使用TIMIT集 [58]和NOISEX-92集 [159]作为实验所用的语音和噪声数据集。TIMIT数据集包含630个说话人在8个对话场景中的美式英语朗读语音，每人朗读10句。NOISEX-92数据集由日常环境中15种常见的噪声组成，每种噪声时长大约4分钟。这些噪声覆盖了日常生活中的常见噪声，并且绝大部分是非平稳噪声。

对于TNMF、SNMF和CNMF的训练。我们从TIMIT的训练集中随机选择2000句纯净语音来训练语音基向量$\mathbf{B}_s$，从NOISEX-92中随机选择9种噪声数据[1]训练噪声基向量$\mathbf{B}_n$。对于DNMF 的训练，我们按照语音片段的时间长度从9种噪声数据[1]中随机切出2000个噪声片段，然后随机地和2000个纯净语音数据联合起来形成2000个语音噪声对儿。通过对语音噪声对儿联合训练语音和噪声的基向量。因为训练NMF的数据是随机选择的，因此语音基$\mathbf{B}_s$是说话人无关的，而噪声基$\mathbf{B}_n$是噪声类型无关的。

对于DNN的训练，我们从TIMIT的训练集中随机选择50个说话人的100句纯净语音，每个说话人选择2句，然后以-5dB到5dB的信噪比随机地和来自NOISEX-92 的9种噪声数据[1]进行混合产生2000句混合语音。

对于测试，我们从TIMIT的测试集中随机选择100个说话人的100句纯净语音，每个说话选择1句语音，然后以-10dB, -7dB, -5dB, -2dB, 0dB, 2dB, 5dB, 7dB 和10dB的信噪比随机地和NOISEX-92的所有噪声数据进行混合产生500句混合语音。其中有6种噪音[2]在训练集中没有出现过，用于测试语音分离模型对于不匹配噪声类型的泛化能力。以类似的方式，我们构建了500句混合语音作为验证集。需要指出的是测试集和验证集所选择的语音和训练集是完全不同的，没有共同的说话人。

因为NOISEX-92的每句噪声片段长达4分钟相对来说比较长，在构建混合语音的时候，为了确保噪声的不同部分参与混合，我们按照语音片段的长度随机地将NOISEX-92的噪声切分出不同的片段。

所有音频的采样率为16kHz。长度为512的汉明窗用于对时域音频数据进行加窗，产生窗长为32ms窗移为16ms的帧数据，然后利用频点数为512的STFT对时域帧数据进行时频分解得到其频域表示，最后对复数STFT系数取模并去掉对称部分得到257维的STFT幅度谱。在下面的实验中，我们使用混合语音的STFT幅度谱作为所有语音分离模型的输入特征，另外为了抓住短时信息，我们拼接上下文信息（左右各拼接2帧）作为最终的输入。

5.6.2 评价指标

我们使用SIR、SAR、SDR [160]、PESQ（$\in [-0.5, 4.5]$） [138]和STOI（$\in [0, 1]$） [155]作为评价指标。SIR、SAR和SDR可以通过BSS Eval计算得到 [160]。PESQ量化了语音客观感知质量而STOI量化了语音的客观可懂度。另外我们也计算了SIR, SDR, SNR, PESQ和STOI相对于未处理带噪语音的增益，计算公式

[1]babble, factory1, buccaneer1、destroyerengine, white, machinegun, pink, volvo, hfchannel.

[2]factory2, buccaneer2, destroyerops, f16, leopard, m109.

如下：

$$\begin{aligned} gSDR(\hat{s},s,x) &= SDR(\hat{s},s) - SDR(x,s), \\ gSIR(\hat{s},s,x) &= SIR(\hat{s},s) - SIR(x,s), \\ gSNR(\hat{s},s,x) &= SNR(\hat{s},s) - SNR(x,s), \\ gPESQ(\hat{s},s,x) &= PESQ(\hat{s},s) - PESQ(x,s), \\ gSTOI(\hat{s},s,x) &= STOI(\hat{s},s) - STOI(x,s), \end{aligned} \tag{5.25}$$

其中gSIR, gSDR, gSNR, gPESQ和gSTOI分别定义了SIR, SDR, SNR, PESQ和STOI的增益。s是纯净的语音信号，x是混合语音信号，$\hat{s}$是分离的语音信号。对于所有评价指标，我们根据测试数据的长度的计算加权平均值，更大的值代表更好的语音分离性能。

5.6.3 模型与设置

Joint-DNN-NMF：为了比较DNN与不同的NMF的组合，我们构建了4个Joint-DNN-NMF模型，它们分别联合了TNMF、SNMF、DNMF和CNMF，依次命名为Joint-DNN-TNMF、Joint-DNN-SNMF、Joint-DNN-DNMF和Joint-DNN-CNMF。这四个模型除了NMF重构层不同之外，其他结构完全一样。其中Joint-DNN-TNMF、Joint-DNN-SNMF和Joint-DNN-DNMF的NMF重构层是线性重构，如公式（5.19），而Joint-DNN-CNMF的重构层是卷积重构，如公式（5.21）。语音和噪声的基向量都是事先用纯净语音和噪声训练的。

对比模型：我们使用基于DNN和基于DNN-NMF的语音分离系统作为对比。基于DNN的语音分离模型，包括DNN-SPE-NOI-MAG [192]、DNN-SPE-MAG [192]和DNN-IRM [171]，它们使用DNN直接学习从带噪特征到分离目标的映射函数。DNN-SPE-MAG预测目标语音的幅度谱，而DNN-SPE-NOI-MAG同时预测目标语音和噪音的幅度谱，DNN-IRM从带噪特征中估计IRM。相对于直接估计语音分离目标，基于DNN-NMF的语音分离系统Kang-DNN-NMF使用DNN预测目标语音幅度谱的NMF激活系数。然后DNN预测的激活系数再通过事先学习的语音基向量或噪声基向量的线性组合用来重构出目标语音或噪声的幅度谱。

NMF配置：所有的NMF模型，包括TNMF、SNMF、DNMF和CNMF全部使用纯净语音和噪声的幅度谱训练，为了抓住频谱的短时信息，我们拼接相邻的5帧上下文（左右各2帧）作为TNMF、SNMF和DNMF的输入向量。相应地，我们设置CNMF建模的时序长度为5，即$T_o = 5$。根据文献 [101]、[129]和 [182]的实验结果，我们经验地设置语音和噪声基向量的个数为256。初步的实验显示更多数量的基向量个数（比如1000）似乎并没有取得明显的性能提升。另外，我们设置TNMF、SNMF、DNMF和CNMF的稀疏约束权重为0、5、5和0.3。对于TNMF、SNMF和CNMF，我们使用文献 [101]和 [129]提供的最好的参数配置和代码实现，而DNMF的算法实现和算法配置依据文献 [182]。

DNN配置：为了简化表述，除非明确提出，否则所有语音分离模型的DNN都有3隐含层，每个隐含层有1000个神经元，隐含层的激活函数为ReLU [59]。所有模型都从一个随机的初始值开始训练，优化器使用L-BFGS [112]，最大的迭代次数设置为400。为了防止过拟合，Dropout技术应用到每个隐含层，dropout的比例设置为0.15。输入特征使用混合语音的STFT幅度谱，并且使用训练集统计的均值方差将输入特征归一化到0均值和1方差。相邻5帧的上下文拼接成一个长向量作为所有模型的最终的输入。根据预测的目标，DNN-IRM的输出层有257个神经元，输出层的激活函数使用sigmoid。DNN-SPE-NOI-MAG的输出层有257×2个神经元，DNN-SPE-MAG的输出层有257个神经元，由于语音和噪声的STFT幅度谱取值虽然非负，但不一定在$[0, 1]$范围之内，因此我们选择softplus [59]作为它们输出层的激活函数。由于Kang-DNN-NMF使用的语音基向量和噪声基向量个数都为256，因此它的输出层有256×2个神经元。对于本章所提出的模型Joint-DNN-TNMF、Joint-DNN-SNMF、Joint-DNN-DNMF和Joint-DNN-CNMF，它们的原始输出将作为语音基向量和噪声基向量的激活系数输入到NMF重构层，根据语音基向量和噪音基向量的个数，它们输出层的神经元个数设置为256×2。又因为NMF的激活系数非负且取值可能超过$[0, 1]$，因此，对于所有基于DNN-NMF的模型，输出层使用softplus作为激活函数。

5.6.4　NMF模型的比较

本章小节我们将系统地分析和比较TNMF、SNMF、DNMF和CNMF四种NMF对语音和噪声基本频谱模式的学习能力以及在语音分离中所取得的性能。图5.2展示了不同NMF学习到的语音基和噪声基。它们呈现了明显的频谱时空结构，特别是对于语音，学习到的基向量看起来非常像构成语音的基本组成成分，比如音素。从图5.2(a)(b)(c)(d)可以看到，有些基向量关注了语音的高频清音成分，有些基向量表征了语音不同的谐波结构。尽管NMF是一个浅层的线性模型，但是由于非负约束，它确实学习到了语音频谱的局部结构表示，而且这些学习到的结构表示看似具有明显的物理感知意义。另外，我们也观察到对基向量的激活系数增加稀疏约束似乎能够增强基向量的表达能力。如图5.2(b)所示，SNMF 学习到的语音基向量相对于TNMF学习到的语音基向量细节更加丰富，结构更加完整，特别在高频部分，呈现出更清楚的结构。尽管DNMF学习到的基向量看起来非常像SNMF学习到的基向量，但是我们依然能够观察到一些明显的不同，比如图5.2(b)和(c)中用红色矩形框标注的基向量。似乎DNMF学习的基向量更加关注独特的局部。这主要是因为DNMF是联合学习语音的噪声的基向量的，试图使得学习到的语音基和噪声基具有差异性，尽可能避免频谱重叠的问题，而其他的NMF是分开地学习语音和噪声的基向量的，并没有考虑语音和噪声基之间的区分性。图5.2(d)呈现了CNMF学习到的语

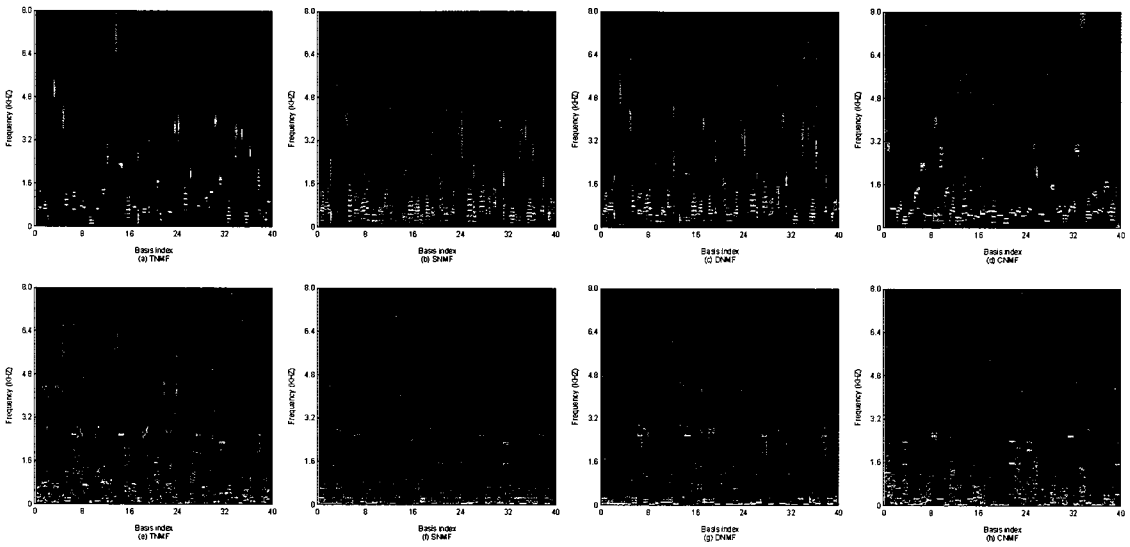


图 5.2: TNMF、SNMF、DNMF和CNMF学习到的语音和噪声基向量。

音基向量。我们观察到CNMF学习到的语音基向量具有更多的时变信息，这表明相对于其他NMF，CNMF更有能力抓住数据中的时序信息。

另外，我们将学习到的语音和噪声基向量应用到基于NMF的语音分离中进一步评估不同NMF的语音分离性能。为了约束分离语音和分离噪声之和等于混合语音，我们使用维纳滤波对NMF分离的语音和噪声进行后处理得到最终的分离语音和噪声。图5.3报道了不同NMF在同一个测试集上所取得的语音分离性能。DNMF在SDR、SIR和SNR三个指标上取得了最好的性能，特别是在SIR指标上，所取得的优势相当明显。但是相对于TNMF，在SAR上表现得并不突出。这表明DNMF能够去除更多的噪声，但同时也带来了更大的语音畸变。相对于其他NMF，DNMF取得的性能提升主要得益于DNMF考虑了语音和噪声频谱重叠的问题，使得学习到的语音基向量和噪声基向量具有一定的差异性。SNMF和CNMF在各个评价指标上都取得了相当的性能，但在SDR和SIR两个指标上明显优于TNMF。这主要是由于稀疏约束和卷积重构能够增强基向量的表达能力。值得指出的是我们所取得语音分离结果和文献 [101, 129, 181]报道的基本一致。

5.6.5 区分性权重λ的比较

本章小节旨在评估优化目标中的区分性权重λ对语音分离性能的影响。λ指明了区分性项对于优化目标的相对重要性。优化目标中的区分性项使得语音分离模型尽可能减少分离语音中的噪声残留以及分离噪声中的语音残留。λ的值越大意味着对残留噪声的惩罚力度越大，但是过度地消除噪声会带来较大的语音畸变。因此选择一个合适的λ对于语音分离至关重要。为了简化实验，我们根据Joint-DNN-TNMF在验证集上所取得语音分离性能来选择合适的λ，从

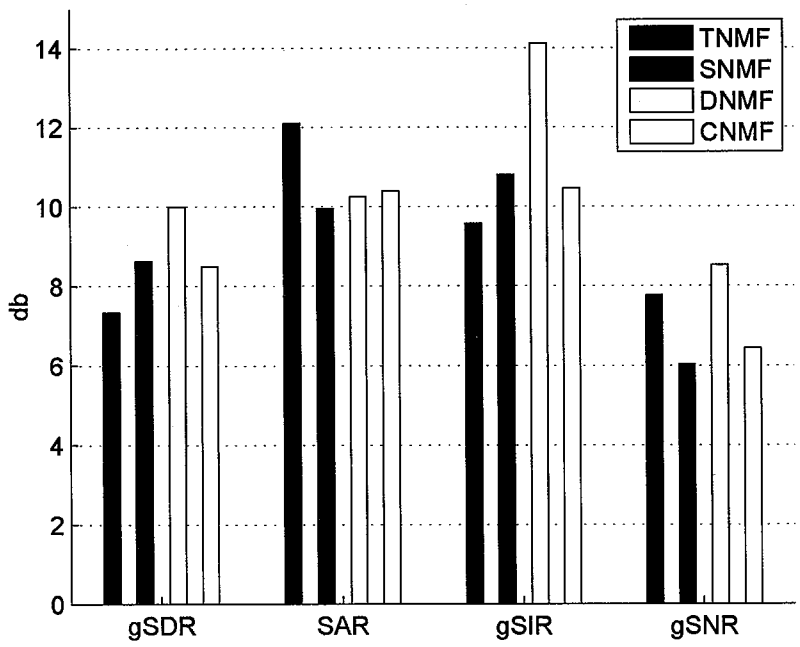


图 5.3: TNMF、SNMF、DNMF和CNMF在测试集上所取得了平均gSDR(dB)、SAR(dB)、gSIR(dB)和gSNR(dB)。

优化目标的角度来讲这个简化是合理的，因为λ并不关乎模型。在实验中，我们设置稀疏权重μ为0，从$\{0, 0.005, 0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 1.0\}$选择$\lambda$，测试Joint-DNN-TNMF在各种$\lambda$取值时所取得的语音分离性能。图5.4报道了Joint-DNN-TNMF在各种不同λ取值时所取得的gSDR、gSIR、SAR和gPESQ。从图中可以观察到随着区分性权重λ取值的变大所取得的gSIR越来越高，一直到$\lambda = 0.5$，这表明更大的区分性权重能够消除更多噪声减少噪声残留。另一方面我们也观察到，当$\lambda \leq 0.02$时，随着λ的取值变大，gSDR会有一个微弱的提升，而SAR和gPESQ会轻微的降低。这表明虽然更大的区分性权重能够消除更多噪声，但同时也带来了较大的语音畸变。相对较小的λ可以以比较小的语音畸变代价较大程度地消除噪声。因此，在下面的实验中，我们选择$\lambda = 0.02$。

5.6.6　重构约束权重γ的比较

本章小节旨在评估优化目标中的重构约束权重γ对语音分离性能的影响。γ指明了NMF重构约束对于优化目标的相对重要性。NMF重构约束使得NMF重构层重构的语音和噪声幅度谱尽可能接近目标语音和噪声的幅度谱。这个约束可以限制DNN的输出在一定尺度范围之内，从而避免维纳滤波层造成的激活系数尺度不确定的问题，如公式（5.23）。然而过大的重构约束会限制维纳滤波层对整个模型学习的影响。因此，需要选择之一个合适的γ使得语音分离模型达到最好的性能。稀疏约束要求激活系数的整体幅度尽可能小，而激活系数的幅度对维纳滤波输出并没有影响，故需要NMF重构约束来限制激活系数的幅度。为了凸显NMF重构约束γ对语音分离性能的影响，我们使用Joint-

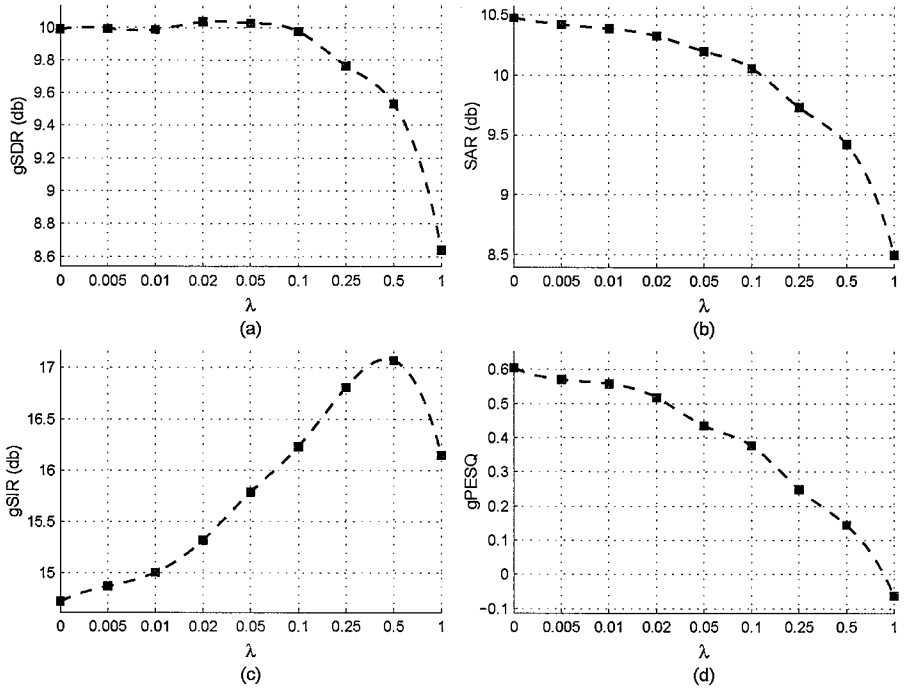


图 5.4: Joint-DNN-TNMF使用不同区分性权重λ时，在验证集上所取得到gSDR、SAR、gSIR和gPESQ。

DNN-SNMF来选择合适的γ，并且设置较大的稀疏约束($\mu = 1$) 来突现激活系数尺度不确定的问题。在实验中，我们分别设置γ等于$\{0, 0.01, 0.1, 0.5, 1\}$集合中的值。图5.5报告了Joint-DNN-SNMF使用不同重构权重γ所取得的gSDR、gSIR、SAR和gPESQ性能。实验结果表明Joint-DNN-SNMF在$\gamma = 0.1$附近取得最好的gSDR和gSIR。我们也观察到，随着γ变大，SAR和gPESQ的性能会持续提升。这表明更大的γ能够保持语音成分而不损害语音，但同时也限制了消除噪声的能力。因此在下面的实验中，我们设置$\gamma = 0.1$来兼顾消除噪声和避免语音畸变。

5.6.7 稀疏约束权重μ的比较

本章小节旨在评估优化目标中的稀疏约束权重μ对语音分离性能的影响。对$\hat{\mathbf{A}}_s$和$\hat{\mathbf{A}}_n$的ℓ_1稀疏约束能够使得DNN的输出尽可能是稀疏的，这样在NMF重构的时候就只有少数基向量被激活。μ是稀疏约束的权重系数，会影响语音分离的性能，因此需要选择合适的μ来使得语音分离的性能达到最优。我们测试Joint-DNN-TNMF、Joint-DNN-SNMF、Joint-DNN-DNMF 和 Joint-DNN-CNMF 在 μ 取集合$\{0, 0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$中的不同值时在验证集上分别所取得的语音分离性能来选择最优的μ。为了避免激活系数尺度不确定的问题，在实验中我们使用相对较大的重构约束权重。我们设置重构约束权重$\gamma = 1$。图5.6报道了Joint-DNN-TNMF、Joint-DNN-SNMF、Joint-DNN-DNMF 和 Joint-DNN-CNMF 使用不同稀疏权重所取得的语音分离性能。我们观

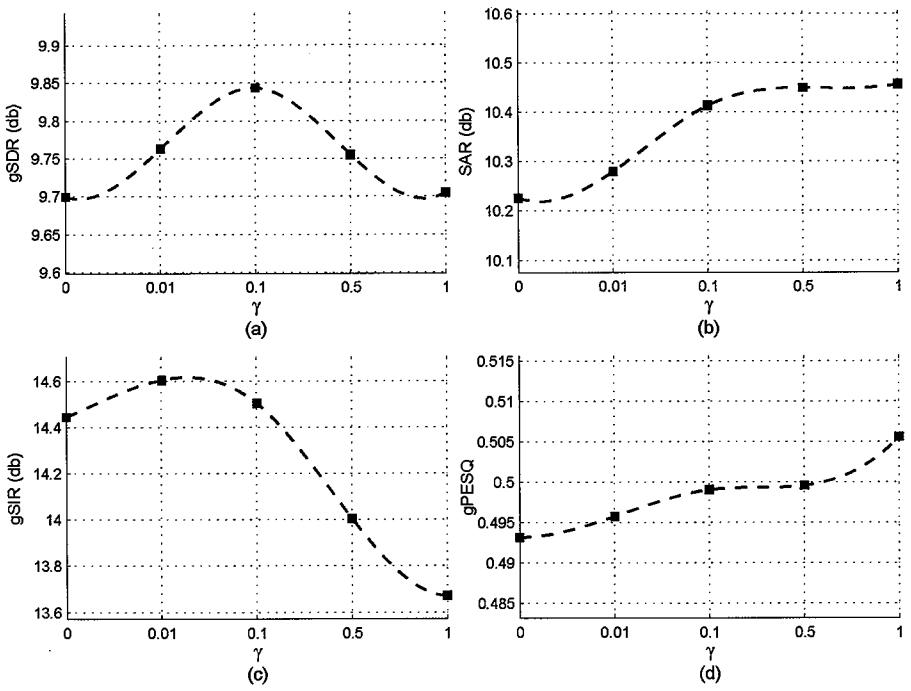


图 5.5: Joint-DNN-SNMF使用不同重构约束权重γ时，在验证集上所取得到gSDR、SAR、gSIR和gPESQ。

察到对于较小的稀疏权重($\mu < 0.1$)，随着稀疏权重的增大性能只有微弱的提升。随着μ继续增大，Joint-DNN-TNMF、Joint-DNN-SNMF 和 Joint-DNN-DNMF所取得的gSDR和gSIR性能逐渐提升。Joint-DNN-TNMF和Joint-DNN-SNMF分别在$\mu = 1$和$\mu = 5$时取得最好的gSIR，而Joint-DNN-DNMF取得的gSIR会继续持续提升。Joint-DNN-TNMF、Joint-DNN-SNMF 和 Joint-DNN-DNMF 的SAR和gPESQ对稀疏约束权重并不敏感，但是过大的稀疏约束($\mu > 1$)会降低SAR和gPESQ的性能。综合考虑，我们设置Joint-DNN-TNMF、Joint-DNN-SNMF 和 Joint-DNN-DNMF 的稀疏权重$\mu = 1$，而设置Joint-DNN-CNMF的稀疏权重$\mu = 0.1$。另外我们也观察到Joint-DNN-DNMF在性能指标gSDR、gSIR和gPESQ上要优于Joint-DNN-SNMF，这个结果和文献 [182]报道的结果一致。令人意外的是Joint-DNN-TNMF在各种评价指标上都取得了最好的性能，这表明使用TNMF学习的语音基向量和噪声基向量更加适合和DNN进行组合。

5.6.8 实验结果与分析

本章小节我们将系统评估本文所提出的模型和对比模型在测试集上所取得的语音分离性能。根据上面关于超参数λ、γ和μ的分析和讨论，我们设置Joint-DNN-TNMF、Joint-DNN-SNMF、Joint-DNN-DNMF和Joint-DNN-CNMF的区分性权重λ和重构约束权重γ分别为0.02和0.1，设置Joint-DNN-TNMF、Joint-DNN-SNMF和Joint-DNN-DNMF的稀疏约束权重μ为1，Joint-DNN-CNMF的稀疏权重为0.1。表5.1报道了不同语音分离模型在测试集上所取得总体语音分离

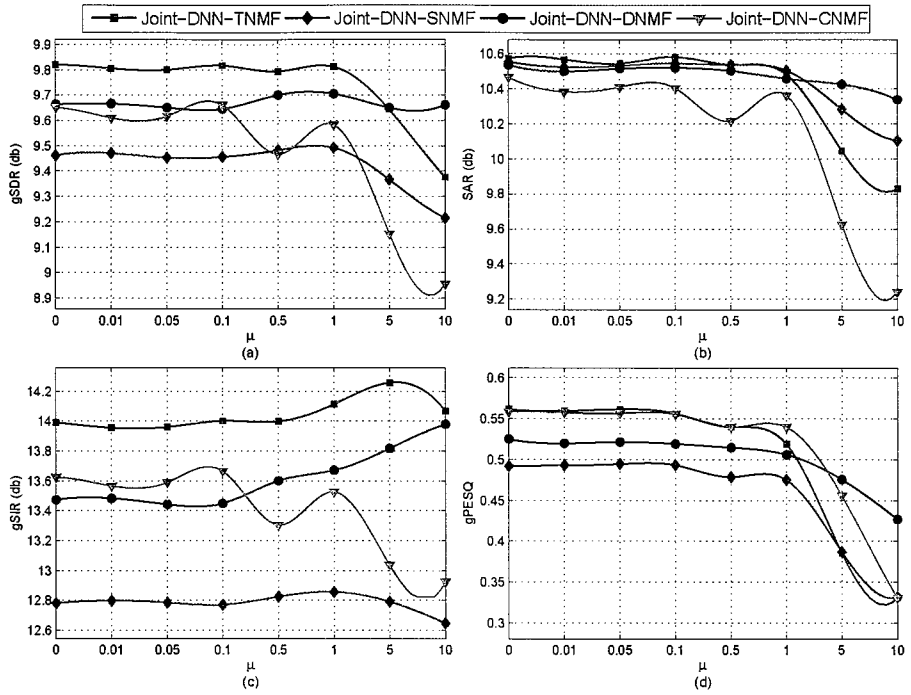


图 5.6: Joint-DNN-TNMF、Joint-DNN-SNMF、Joint-DNN-DNMF和Joint-DNN-CNMF使用不同稀疏约束权重μ，在验证集上所取得到的gSDR、SAR、gSIR 和gPESQ。

表 5.1: 不同语音分离模型所取得的全局语音分离性能。

Models	gSDR	SAR	gSIR	gPESQ
Joint-DNN-TNMF	**10.00**	10.51	**14.81**	0.54
Joint-DNN-SNMF	9.62	10.47	13.34	0.48
Joint-DNN-DNMF	9.87	10.39	14.50	0.52
Joint-DNN-CNMF	9.93	10.44	14.66	**0.55**
DNN-SPE-NOI-MAG [192]	9.72	**10.76**	13.46	0.50
DNN-SPE-MAG [192]	8.32	8.80	12.20	0.47
DNN-IRM [171]	8.36	10.67	10.79	0.43
Kang-DNN-NMF [92]	9.35	10.03	13.37	0.45

性能，另外，为了比较语音分离模型对不匹配噪声的泛化能力，我们还报道了所有语音分离模型在噪声匹配和噪声不匹配的条件下所取得语音分离性能，如表5.2所示。

首先，我们比较了不同语音分离目标对语音分离性能的影响。其中DNN-IRM使用理想掩蔽近似的语音分离目标，DNN-SPE-NOI-MAG和DNN-SPE-MAG都使用了幅度谱近似的语音分离目标，但是DNN-SPE-MAG只预测了目

表 5.2: 不同语音分离模型在噪声匹配的条件和噪声不匹配的条件下所取得的语音分离性能。

Models	Matched noise				Unmatched noise			
	gSDR	SAR	gSIR	gPESQ	gSDR	SAR	gSIR	gPESQ
Joint-DNN-TNMF	**11.30**	11.04	**16.78**	**0.65**	7.96	9.74	11.71	0.38
Joint-DNN-SNMF	10.68	10.77	14.83	0.55	7.93	10.05	11.00	0.36
Joint-DNN-DNMF	11.00	10.81	16.28	0.61	8.09	9.77	11.65	0.38
Joint-DNN-CNMF	11.05	10.83	16.39	0.63	**8.16**	9.89	**11.91**	0.42
DNN-SPE-NOI-MAG [192]	10.82	**11.07**	15.08	0.57	7.98	**10.31**	10.87	0.39
DNN-SPE-MAG [192]	9.12	9.06	13.25	0.52	7.03	8.44	10.47	0.38
DNN-IRM [171]	9.25	10.71	11.97	0.50	6.94	10.64	8.93	0.31
Kang-DNN-NMF [92]	10.46	10.39	14.89	0.51	7.61	9.53	10.97	0.35

标语音幅度谱，DNN-SPE-NOI-MAG同时预测了目标语音幅度谱和噪声幅度谱。Kang-DNN-NMF使用目标语音幅度谱和噪声幅度谱的NMF激活系数作为分离目标。Joint-DNN-TNMF、Joint-DNN-SNMF、Joint-DNN-DNMF和Joint-DNN-CNMF使用隐式掩蔽的幅度谱近似目标，它们通过NMF重构和维纳滤波同时预测目标语音幅度谱和噪声幅度谱。从表5.1可以看出，DNN-SPE-MAG和DNN-IRM在不同评价指标上各有优劣，DNN-SPE-MAG 在 gSIR 和 gPESQ 上取得更优的性能，而DNN-IRM在SAR上取得更好的性能，在gSDR取得相当的性能。相对于DNN-SPE-MAG和DNN-IRM，DNN-SPE-NOI-MAG在所有指标上都取得了显著的性能优势，这主要得益于分离语音和分离噪声密切相关存在一定的互补性，同时对它们建模能够提高语音分离的性能。

其次，我们比较了所提出的语音分离方法的性能。表5.1给出了全局的性能指标，表5.2给出了噪声匹配和噪声不匹配的条件下所取得语音分离性能。从全局性能上我们可以看出，除了SAR，本文提出的DNN与NMF协同组合的框架在绝大多数指标上都优于单独的DNN语音分离模型（DNN-IRM、DNN-SPE-NOI-MAG和DNN-SPE-MAG），特别是在SDR、SIR以及PESQ上取得了显著的性能提升。这主要得益于DNN和NMF的联合建模，一方面利用了NMF挖掘了语音和噪声频谱的时空结构，另一方面充分发挥了DNN强大的非线性映射学习的能力。尽管Kang-DNN-NMF也将DNN和NMF同时用于语音分离，但在Kang-DNN-NMF框架中，DNN和NMF是分步进行的，并没有联合到统一的模型中。这会导致DNN预测误差和NMF重构误差的双重累积。从表5.1和表5.2中可以看出Kang-DNN-NMF相对于DNN-SPE-NOI-MAG和DNN-SPE-MAG并没有取得显著的性能优势，特别是在噪声不匹配的条件下，性能会急剧下降，在很多性能指标上甚至比DNN-SPE-NOI-MAG和DNN-SPE-MAG都要差。

为了评估语音分离模型对带噪语音可懂度的提升情况，我们给出了不

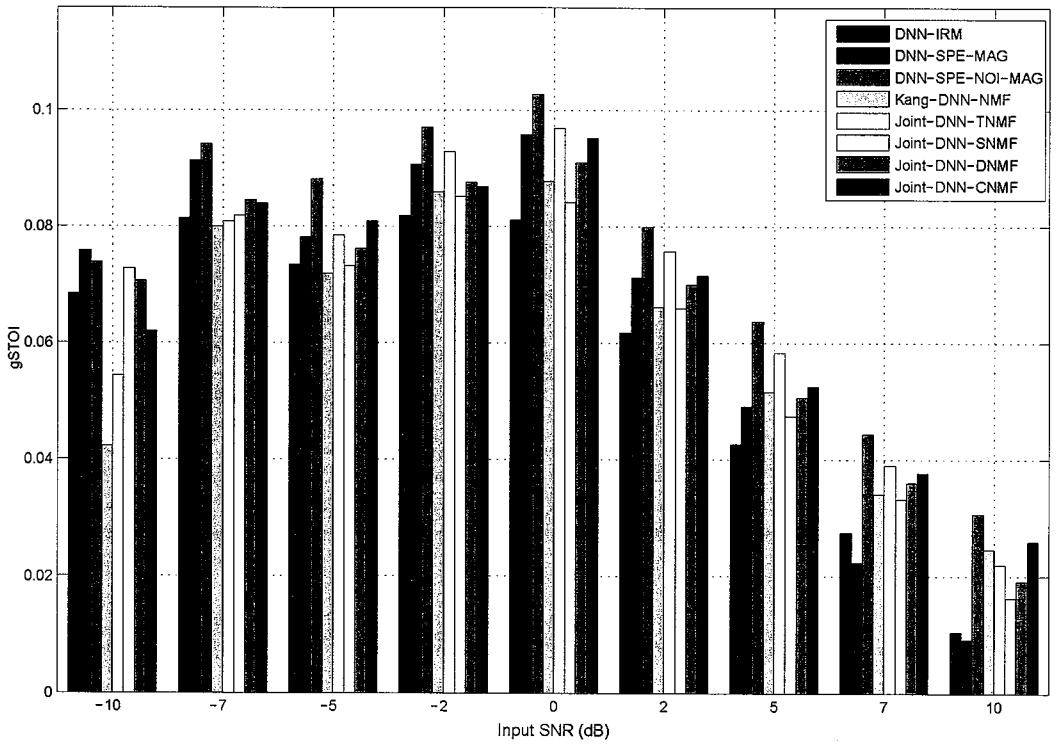


图 5.7: 不同语音分离模型在不同输入SNR条件下取得的STOI增益。

同语音分离模型在测试集上所取得的STOI增益。图5.7报道了语音分离模型在不同输入SNR条件下所取得的STOI增益值。从图中可以看出，相对于原始带噪语音，所有模型都显著地提升了语音的可懂度。DNN-SPE-NOI-MAG在绝大多数输入SNR条件下都能取得更高的STOI增益，这表明DNN-SPE-NOI-MAG在STOI指标上要优于其他模型。我们也观察到，本文所提出的DNN与NMF协同联合的模型在STOI指标上要优于Kang-DNN-NMF。这表明DNN与NMF的协同组合方式要优于分步建模的方式。

最后，我们展示了Joint-DNN-TNMF和DNN-SPE-NOI-MAG分离语音的对数能量谱来直观地理解不同语音分离模型的差异。图5.8(a)(b)(c)(d)分别是混合语音、纯净语音、DNN-SPE-NOI-MAG分离的语音和Joint-DNN-TNMF分离的语音的对数能量谱。从图中可以看到，相比于DNN-SPE-NOI-MAG，Joint-DNN-TNMF分离的语音噪音残留较小，呈现了更完整的频谱结构。这是因为Joint-DNN-TNMF 分离的语音是通过纯净语音的基向量重构得到的，而DNN-SPE-NOI-MAG是在一个没有任何约束的空间里直接将带噪语音映射到目标语音，相比较而言，Joint-DNN-TNMF利用了更多语音本身的先验知识，能够避免一定的噪声残留。

5.7 本章小结

语音的分离目标，无论是理想时频掩蔽还是目标语音频谱，都具有明显的时空结构。挖掘这些时空结构能够提升语音分离的性能。NMF具有强大

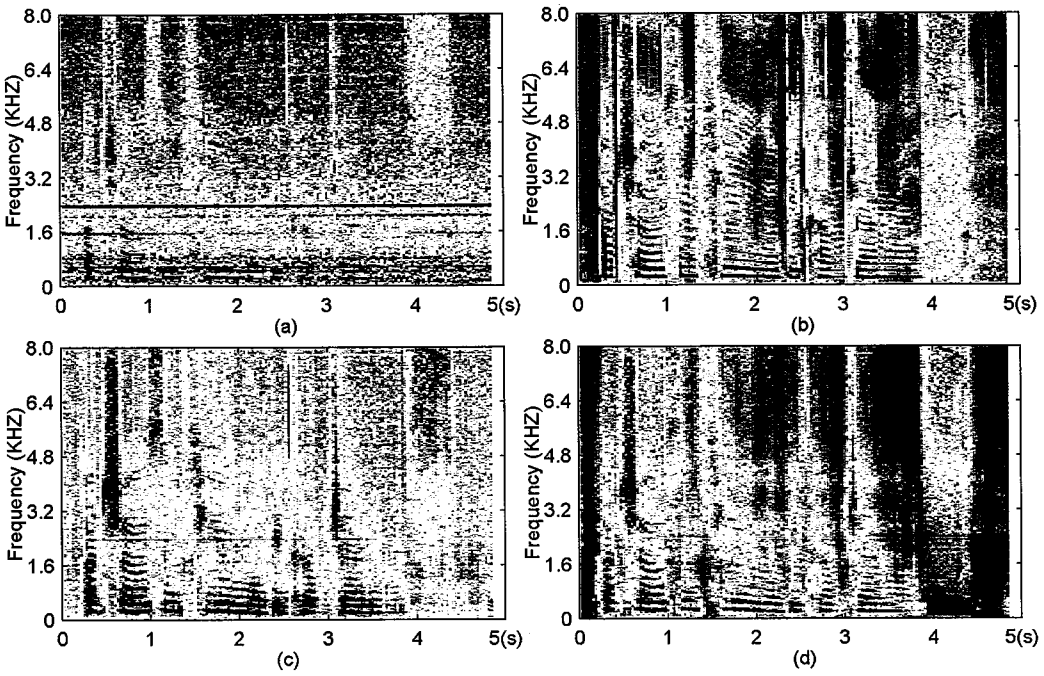


图 5.8: Joint-DNN-TNMF和DNN-SPE-NOI-MAG分离语音的频谱对比

的局部表示学习能力，能够挖掘非负数据的结构模式。本章我们提出了一种DNN和NMF联合协作的语音分离框架，既利用了NMF对语音分离目标的时空结构的挖掘能力，又充分发挥了DNN强大的映射学习能力。系统的实验证明，所提出的DNN与NMF的组合模型显著优于单独的DNN语音分离模型，也优于分步的DNN-NMF方法。本章节，我们首先综述了基于NMF的语音分离方法，然后回顾了几个典型的NMF模型。最后给出了本章所提出的语音分离方法，介绍DNN与NMF组合模型的网络结构，优化目标和相关工作。在实验部分，我们比较了不同NMF学习到的基向量，对比了它们的语音分离性能，同时也详细介绍了所提出的优化目标的各个参数对语音分离性能的影响，并从各个方面比较不同语音分离系统的性能差异。

第六章　噪声声学环境深度感知的语音增强方法

6.1　引言

真实环境中所采集的语音信号不可避免地会被各种噪声和混响的干扰，这些干扰显著地降低了语音的可懂度和感知质量 [94,158]，也会使自动语音识别的性能严重下降 [10,90,187]。为应对真实的声学环境，必须建立有效的语音增强技术，以消除噪声和混响的干扰，提高语音的可懂度和感知质量。目前，包括信号处理和深度学习等各种技术已广泛应用到语音增强中。

频谱特性是单声道语音增强算法的重要线索，监督式语音增强算法主要利用了语音和噪声的频谱特性使用监督学习算法实现了语音和噪声的分离，前面章节从语音频谱特性的各个方面进行了一系列的探索。得益于深度学习对语音和噪声具有强大的感知能力，基于深度学习的语音增强技术相对于传统的信号处理方法取得了显著的性能优势 [168,193]。然而基于深度学习的语音增强方法严重依赖于监督数据的训练，面对不匹配的声学环境时，由于语音或者噪声的错误评估造成其性能会严重下降。而在传统语音增强的框架下，通过计算噪声统计特性这些评估误差可以通过“平均”得到缓解。

对于单声道语音增强，除了语音频谱特性之外，噪声统计特性也是重要的线索。对于传统信号处理的语音增强方法，噪声统计通常是必不可少的，需要连续估计 [31,32,114,118,119]。由于语音信号在时间和频率上是稀疏的，因此可以实现对噪声统计特性的连续估计 [153]。短时滑动平均是估计噪声统计的常用方法，在这个方法中，通过使用语音活性检测（VAD）来决定更新或者保持噪声统计。而根据语音存在概率（Speech Presence Probability，SPP）来判断每一个时频单元是否包含语音活性成分是一种更加精细的方式，这种方式能够获得更加准确的噪声功率谱（Power Spectral Density，PSD）。这些方法通常假设噪声是平稳慢变的，而在真实环境中，噪声声学环境通常是复杂多变的，一方面传统信号处理的方法难以准确估计SPP，另一方面缺乏对噪声声学环境的智能感知，这导致基于信号处理的语音增强方法在真实的声学环境中并没有达到令人满意的性能。为了应对非平稳噪声声学环境，Martin提出了最小统计的方法来追踪带噪信号的最小频谱能量 [118]。Cohen 改进了最小统计方法，提出了一种所谓的极小控制递归平均（Minima Controlled Recursive Averaging，MCRA）的方法 [32,33]。尽管这些方法提升了传统语音增强方法应对非平稳噪声的能力，但对于真实声学环境依然是一个极具挑战的任务，其性能远没达到令人满意的程度。

综上，噪声统计特性和语音频谱特性是单声道语音增强的基本信息。基于信号处理的语音增强方法主要依赖于噪声统计特性的估计，对于平稳噪声的

处理相对较好，但仍难以应付非平稳噪声。而基于深度学习的语音增强方法主要依赖于对语音频谱特性的感知，由于深度学习强大的感知能力，基于深度学习的方法有能力处理非平稳噪声。本文前四章所提出的语音增强方法主要是利用深度学习模型学习一个从带噪特征到目标语音和理想掩蔽的映射函数，这种简单地将语音增强表达成一个监督式学习问题，过度依赖于训练数据而忽略了信号处理领域知识，通常在面临不匹配的声学环境时，比如训练集中不存在的噪声类型和输入SNR，性能会严重下降。显然，信号处理和深度学习深度融合的方案可能是一个更加明智的选择。本章我们提出了一个信号处理与深度学习融合的语音增强框架，将深度学习对语音和噪声强大的感知能力融入到传统的语音增强框架中。我们首先利用门控循环单元（Gated Recurrent Unit，GRU）[30]和前馈网络构建了一个新颖的深度噪音追踪网络（Deep Noise Tracking Network，DNTN）。GRU用于从带噪特征中估计语音存在概率，然后将GRU长时状态向量（GRU隐含层输出）与当前带噪特征拼接起来输入到前馈网络中，用以感知噪声声学环境估计噪声统计的更新因子。根据估计的语音存在概率（SPP）和噪声声学环境，利用成熟的时间移动平均技术对噪声信号和观测信号的PSD进行智能更新。最后利用估计的噪声信号和观测信号的PSD构建最优维纳滤波器，通过维纳滤波消除噪声提取目标语音。整个模型的所有模块通过频谱近似的目标进行联合优化。在数据集CHiME-4和NOISEX-92上的实验系统证明了信号处理和深度学习深度融合的方法能够显著提升语音增强的性能。

本章首先对信号模型和语音增强问题进行定义，然后介绍了信号处理和深度学习融合的语音增强框架。最后在CHiME-4和NOISEX-92数据集上，通过实验验证了所提出的方法的有效性。

6.2 信号模型和问题定义

这里我们考虑混响房间里单个麦克风的场景，假设房间里充满噪声并且只有一个点源目标语音。我们定义$s(k)$和$v(k)$分别是语音信号和加性噪声信号，其中k是离散时间坐标。观测信号能够通过以下信号模型表示：

$$x(k) = g(k) * s(k) + v(k) = y(k) + v(k), \tag{6.1}$$

其中，$*$定义卷积运算，$g(k)$是通道冲击响应。$y(k) = g(k) * s(k)$是无噪的语音成分。假定所有的信号是零均值的随机过程，在短时傅立叶变换（STFT）域，公式（6.1）可以重写为：

$$x(f,t) = y(f,t) + v(f,t), \tag{6.2}$$

其中，f和t分别定义频率和时间坐标。

严格来讲，语音增强不仅涉及到降噪还涉及到降混响。这里我们仅关注降噪问题。因此，我们的目标就是应用一个线性滤波器$h(f,t)$对观测信号$x(f,t)$ 进

行滤波来消除噪声信号$v(f,t)$恢复语音信号$y(f,t)$。维纳滤波被视为是许多噪声消除方法的基石，许多降噪算法和维纳滤波密切相关 [26]。假定语音信号和噪声信号是不相关的，维纳滤波器增益的一般形式如下：

$$h(f,t)=\frac{\phi_{yy}(f,t)}{\phi_{yy}(f,t)+\phi_{vv}(f,t)}, \tag{6.3}$$

其中$\phi_{vv}(f,t)$是噪声信号的PSD，可以通过以下公式计算得到：

$$\phi_{vv}(f,t)=E\left\{v(f,t)v^{*}(f,t)\right\}, \tag{6.4}$$

其中，*定义了复数的共轭运算。因为噪声信号和语音信号假设不相关，因此，目标信号的PSD可以通过观测信号和噪声信号的PSD计算得到，如下所示：

$$\phi_{yy}(f,t)=E\left\{y(f,t)y^{*}(f,t)\right\}=\phi_{xx}(f,t)-\phi_{vv}(f,t), \tag{6.5}$$

其中$\phi_{xx}(f,t)=E\left\{x(f,t)x^{*}(f,t)\right\}$是观测信号的PSD。$E\{\cdot\}$定义了数学期望运算。在实际应用中，为了满足实时性要求，短时递归平滑常被用来近似数学期望 [153]，具体来说就是，在当前时刻，噪声信号和观察信号的PSD可以通过滑动平均计算得到，如下所示：

$$\begin{aligned}\hat{\phi}_{vv}(f,t)&=\tilde{\alpha}_{v}(f,t)\hat{\phi}_{vv}(f,t-1)+\left(1-\tilde{\alpha}_{v}(f,t)\right)x(f,t)x^{*}(f,t)\\ \hat{\phi}_{xx}(f,t)&=\alpha_{x}(f,t)\hat{\phi}_{xx}(f,t-1)+\left(1-\alpha_{x}(f,t)\right)x(f,t)x^{*}(f,t),\end{aligned} \tag{6.6}$$

其中$0\leqslant\alpha_{x}(f,t)\leqslant 1$和$0\leqslant\tilde{\alpha}_{v}(f,t)\leqslant 1$分别是观测信号和噪声PSD的平滑因子。它们对正确更新观测信号和噪声信号的PSD至关重要。在实际应用中，$\alpha_{x}(f,t)$通常设置为一个合适的常数α_{x}，以在平滑噪声信号和跟踪语音信号之间达到一个好的平衡，而$\tilde{\alpha}_{v}(f,t)$是一个随时间和频率变化的滑动因子。当语音缺失的时候，它应该足够小，以迅速跟随噪声的变化，当语音出现的时候，它应该充分的大，以避免将语音成分带入噪声PSD中从而造成噪声PSD过估计。显然，$\tilde{\alpha}_{v}(f,t)$与语音存在或缺失的检测密切相关，因此，通常使用语音存在概率SPP$p(f,t)$来调节$\tilde{\alpha}_{v}(f,t)$ [153]，如下所示：

$$\tilde{\alpha}_{v}(f,t)\ =\ \alpha_{v}\ +(1-\alpha_{v})p(f,t), \tag{6.7}$$

其中，$\alpha_{v}\in[0,1)$决定了语音缺失的情况下噪声PSD的更新因子。在传统语音增强算法中，一般基于噪音声学环境平稳的假设，设置为一个常数，然而，在实际应用中，噪声声学环境平稳的假设通常很难满足。

6.3　信号处理和深度学习深度融合

图6.1展示了大多数传统语音增强算法的一般框架 [133]。在这个框架中，噪声PSD的估计是最重要的模块之一，因为它在很大程度上决定了维纳滤波

器输出的噪声残留 [153]。评估噪声PSD的常用方法是通过时变滑动平均技术根据当前的平滑因子不断更新噪声的PSD，平滑因子通常是通过SPP实时调节的 [118,119]。但当输入SNR比较低或者噪声不平稳时，传统的信号处理技术很难获得准确的SPP，这就限制了噪声追踪器在真实声学环境下的跟踪能力 [157]。

深度学习对语音和噪声具有很强的感知能力。为了解决传统信号处理方法的局限性，我们提出了信号处理和深度学习融合的方案，在这个方案里，深度学习技术代替传统语音增强框架里基于信号处理技术的SPP评估器，利用深度学习模型来估计语音存在概率，而保留传统语音增强框架中的基于信号处理技术的PSD更新模块和维纳滤波器。深度学习和信号处理模块融合为一个有机的整体，各个模块以端到端的方式联合协作，而不是以分步的方式进行。也就是说，深度学习模型是由最终维纳滤波输出的误差联合优化的，而不是由另外单独的目标进行优化的,比如理想SPP，而且没有必要事先单独训练深度学习模型。

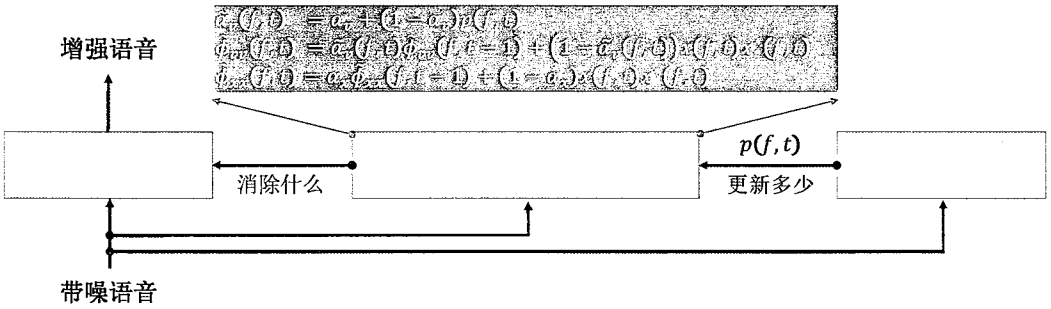


图 6.1: 传统语音增强系统的一般结构。

图6.2展示了所提出的信号处理和深度学习融合的语音增强框架的结构框图。整个方案紧跟如图6.1所示的传统语音增强系统的一般结构。我们构建了一个新颖的深度噪声追踪网络（DNTN）来代替图6.1中的SPP/VAD评估器来估计SPP或者VAD。DNTN由语音感知网络和噪声声学环境感知网络构成。语音感知网络使用双层GRU自动从带噪输入特征中感知SPP，输出层有F个sigmoid的神经元。噪声声学环境感知网络使用单层的前馈网络，用以感知当前噪声声学环境的平稳性，它的输入不仅包含当前时刻的带噪输入特征，还包括当前语音感知网络的长时状态（最后一个GRU的隐层输出），噪声声学环境感知网络的输出层仅有1个神经元，激活函数为sigmoid。语音感知网络输出一个$[0,1]$的$1 \times F$的向量，该向量被用作当前帧的语音存在概率$p(f,t)$来调节噪声PSD的平滑因子，噪声声学环境感知网络输出一个$[0,1]$系数α_v，反映了当前噪声声学环境的平稳性。在获得α_v和$p(f,t)$之后，通过公式（6.7）可以计算噪声PSD的平滑因子$\tilde{\alpha}_v(f,t)$，然后根据公式（6.6）更新噪声信号和观察信号的PSD，利用噪声信号和观察信号的PSD根据公式（6.3）能构建最优维纳滤波

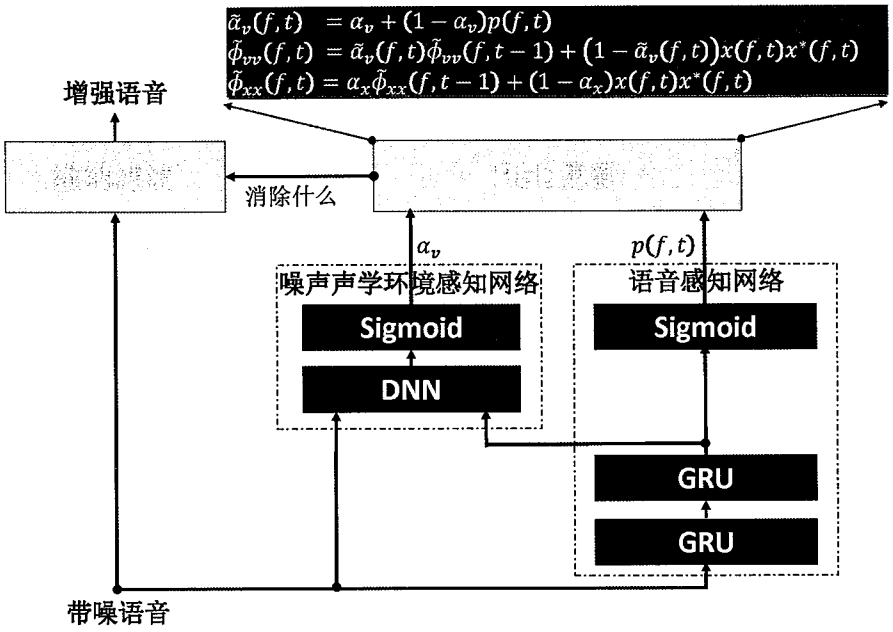


图 6.2: DNTN的结构。

器。最后利用构建的维纳滤波器对观测信号进行滤波我们可以得到去除噪声的期望信号，如下所示：

$$\tilde{y}(f,t) = \frac{\phi_{xx}(f,t) - \phi_{vv}(f,t)}{\phi_{xx}(f,t)} x(f,t). \tag{6.8}$$

需要说明的是语音感知网络和噪声声学环境感知网络并没有明确的目标，而是通过最终的维纳滤波输出进行优化的。

语音增强的目标旨在从噪声信号$v(f,t)$中恢复目标语音信号$y(f,t)$。因此，我们使用基于最小均方误差的频谱近似的优化目标来联合优化语音增强系统的各个模块，如下所示：

$$J = \frac{1}{T}\sum_{t=0}^{T}\sum_{f=0}^{F}|\tilde{y}(f,t) - y(f,t)|^2, \tag{6.9}$$

其中，$|\cdot|$定义了复数绝对值运算，F和T分别是频点数和时间帧数。

6.4 实验及其分析

6.4.1 数据集

我们使用CHiME-4 [161]和NOISEX-92 [159]数据集来评估所提出的DNTN的语音增强性能。CHiME-4数据库由真实数据和模拟数据组成，其中真实数据是在4个真实场景[1]中由实际说话人朗读录制的，而模拟数据是通过人为地将纯净语音与真实场景中采集的背景噪声混合而产生的，这意味着混合信号中的语音成

[1]bus、cafe、pedestrian area 和 street junction。

分和噪声成分事先都是已知的。因此, 模拟数据可以作为监督性数据用来训练所提出的DNTN。虽然CHiME-4数据集中的每个语句都包含6个通道, 但我们随机选择其中一个通道信号用于实验。NOISEX-92数据集包含15种日常生活中常见的噪声数据[2], 每一种噪声的录音长度大约为4分钟。需要强调的是这些噪声数据与CHiME-4数据集中的噪声有很大的不同。

CHiME-4的训练集由1,600句真实语音和7,138句模拟语音组成，我们选择其中的7,138句模拟数据作为本章实验的训练数据，另外使用CHiME-4的验证集中的模拟数据（410 (模拟) × 4 (环境)）作为本章实验的验证数据。对于测试数据，我们从WSJ0的验证集中随机选择1,000句纯净语音，然后将它们与NOISEX-92中的15种噪声数据以信噪比0dB到10dB随机地进行混合，产生1,000句混合语音。测试数据中的噪声数据不同于训练集，因此可以用来测试语音增强模型对不匹配噪声的泛化能力。在对语音和噪声进行混合时, 为了确保噪声的不同片段与纯净语音进行混合, 我们根据纯净语音的时间长度, 随机将NOISEX-92的每句噪声切分为不同的片段。

6.4.2 评价指标

我们使用SIR、SAR、SDR [160]以及PESQ（$\in [-0.5, 4.5]$） [138]作为评价指标。SIR、SAR和SDR可以通过BSS Eval计算得到 [160]，PESQ量化了语音客观感知质量。所有评价指标都是根据测试数据的时长加权平均计算得到，反映语音增强算法在不同数据上的平均性能。评价指标的值越高代表语音增强的性能越好。

6.4.3 模型与设置

基于深度学习的监督式语音增强通常学习一个从带噪输入特征到理想时频掩蔽或目标语音频谱的映射函数。本章我们采用掩蔽近似和频谱近似的监督式语音增强作为对比方法。掩蔽近似的方法使用两层GRU作为语音增强的模型，从带噪特征中预测IRM，该方法定义为GRU-IRM [171]，而频谱近似的方法使用两层GRU预测目标语音的频谱，我们将其定义为GRU-MAG。由于IRM的取值范围在0到1之间，因此GRU-IRM使用sigmoid作为输出层的激活函数。GRU-MAG并没有直接预测目标语音的频谱，而是将输出层的sigmoid的输出用作掩蔽运算应用到混合语音的频谱上来产生期望语音的频谱，这种方式可视为间接掩蔽 [89]。本章所提出的DNTN除了有相同的GRU网络外，还有一个单隐含层的前馈网络，前馈网络的隐含层的神经元个数为512，激活函数为ReLU [59]。前馈网络的输出层神经元个数为1，激活函数为sigmoid。根据STFT的频点个数，

[2]babble、factory1、buccaneer1、destroyerengine、white、machinegun、pink、volvo、hfchannel、factory2、buccaneer2、destroyerops、f16、leopard、m109。

GRU-IRM、GRU-MAG和DNTN的GRU网络的输出层有256个神经元。GRU-IRM、GRU-MAG和DNTN的每个GRU层的Cell维度为512。此外，我们将批量规范化应用于每个GRU层的输入之前，这可能会加快GRU训练的收敛速度 [99]。

在下面的实验中，我们使用256维的混合语音的对数功率谱作为输入特征，输入特征的每个维度利用训练集的均值和方差归一化到0均值1方差。所有网络都通过Adam优化器 [95]从随机初始值开始训练，初始学习率设置为0.001，最大的迭代次数设置为25代，批量化的尺寸设置为256。频谱特征通过将512点的STFT应用于32毫秒帧长的混合信号得到，分帧操作通过长度为512的汉明窗按照50%重叠加窗得到。256维的对数功率谱是通过对STFT功率谱求对数并去除对称部分得到的。

6.4.4　实验结果与分析

首先，我们系统地评估了参数α_x对语音增强性能的影响。α_x是观测信号PSD的平滑因子，决定了观察信号PSD的更新或保持速度，如公式（6.6）所示。过大的α_x限制了观测信号PSD的更新，不能及时跟踪语音信号和声学环境的变化，而过小的α_x不能平滑混合信号中的噪声信号。因此，它的选择对正确更新观察信号PSD的非常重要，应当设置为一个合适的值，以达到平滑噪声信号和跟踪语音信号的平衡。我们从集合 $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$ 中选择α_x的值，并测试语音增强模型在不同α_x取值时所取得的性能。图6.3报告了语音增强模型在噪声匹配（CHiME-4）和噪声不匹配（NOISEX-92）条件下使用不同平滑因子α_x时所取得的SDR的平均增益（gSDR）。gSDR的计算如下式：

$$gSDR(\tilde{y}, y, x) = SDR(\tilde{y}, y) - SDR(x, y), \tag{6.10}$$

其中$\tilde{y}$是增强的语音信号。gSDR反映了全局性能提升 [89]。从图中可以看出所提出的语音增强模型无论在噪声匹配的条件下还是在噪声不匹配的条件下都是在$\alpha_x = 0.8$附近取得最优的性能，这个经验值非常接近理论计算的值 [119]。因此，我们在下面的实验中设置$\alpha_x = 0.8$。

其次，我们系统评估了本章所提出的模型（DNTN）和对比模型（GRU-IRM和GRU-MAG）的语音增强性能。表6.1分别报告了不同模型在噪声匹配和噪声不匹配条件下所取得的语音增强性能。我们观察到，所有模型无论在噪声匹配或噪声不匹配的条件下都显著地消除了噪声增强了语音，但GRU-MAG和DNTN要显著优于GRU-IRM。这是因为GRU-MAG和DNTN直接优化的是实际语音增强目标，而GRU-IRM优化的是语音增强的中间目标IRM。这表明基于间接掩蔽的频谱近似目标要优于直接的掩蔽近似目标。我们还观察到，所提出的DNTN与GRU-MAG在噪声匹配条件下取得了大致相同的语音增强性能，DNTN在SDR、SAR、gSDR和PESQ等指标上稍微优于GRU-MAG。然而，

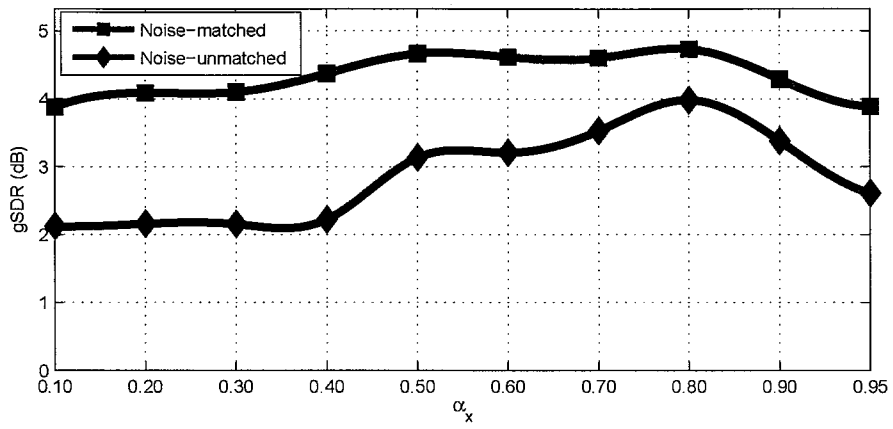


图 6.3: DNTN在噪声匹配和噪声不匹配条件下使用不同平滑因子α_x时所取得的gSDR(dB)。

表 6.1: 不同语音增强模型在噪声匹配和噪声不匹配的条件下所取得的性能。

	Models	SDR	SIR	SAR	gSDR	PESQ
Unmatched	Mixture	5.03	5.03	–	–	1.26
	GRU-IRM	6.29	7.53	8.97	1.26	1.37
	GRU-MAG	8.00	10.67	10.28	2.98	1.36
	DNTN	**9.00**	**12.08**	**10.65**	**3.97**	**1.41**
Matched	Mixture	3.86	3.86	–	–	1.17
	GRU-IRM	6.95	8.75	8.67	3.10	1.34
	GRU-MAG	8.54	**11.63**	10.62	4.69	1.34
	DNTN	**8.58**	11.19	**11.25**	**4.73**	**1.36**

在噪声不匹配条件下，DNTN在各种评价指标上达到最佳性能, 显著优于GRU-MAG。这表明DNTN对不匹配的噪声具有更好的泛化性能，这对于实际应用具有十分重要的意义。更好的泛化性能主要是由于对信号处理领域知识的有效利用。DNTN将深度学习对语音和噪声的强大感知能力融入到传统成熟的语音增强框架中，不仅利用了语音的频谱特性，而且还利用了噪声的统计特性，充分发挥了信号处理和深度学习各种的优势。事实上，信号处理领域知识的应用可以看成DNTN的正则化，避免模型过拟合。

最后，我们展示了DNTN的一些可视化结果来进一步理解DNTN的每一个模块发挥的作用。图6.4(a)(b)(c)(d)分别呈现了纯净语音、噪声、混合语音和DNTN增强的语音的对数能量谱。图6.4(e)(f) 分别是DNTN中的语音感知网络和噪声声学环境感知网络的输出。从图中可以看出，尽管并没有明确的监督性目标，但通过最终的维纳滤波输出误差作为指导，语音感知网络自动学会了

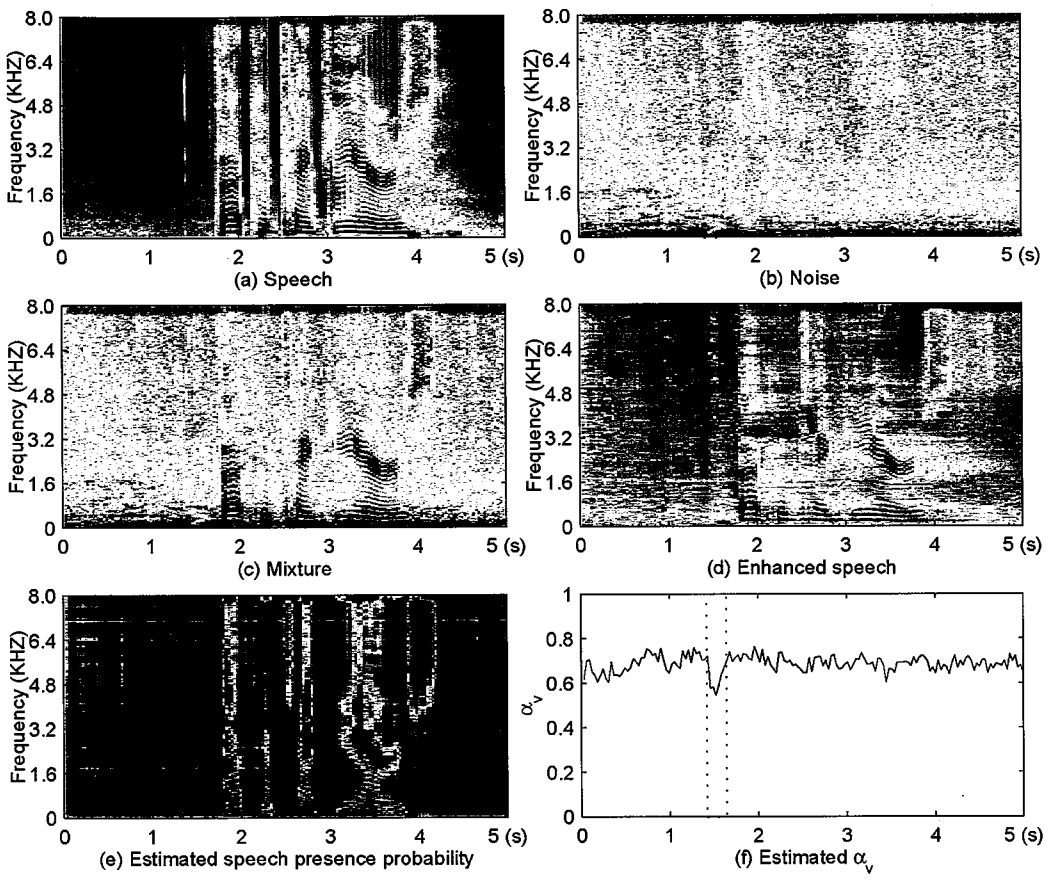


图 6.4: DNTN的可视化样例。

从带噪输入中感知语音存在概率，图6.4(e)通过颜色深浅标识了语音的存在概率。图6.4(f)是噪声声学环境感知网络的输出，从其波动曲线可以看出，它反映噪声声学环境的平稳性，如虚线所框住的区域，噪声频谱出现波动，能量降低，对应的噪声声学环境感知网络的输出也变低。这表明噪声声学环境感知网络能够感知到噪声声学环境的变化。

6.5　本章小结

本章在充分把握信号处理领域的基本理论和监督式语音增强前沿方法的基础上，对基于信号处理的语音增强方法和基于深度学习的语音增强方法进行深入分析和比较，针对信号处理的方法难以应对非平稳噪声而基于深度学习的方法对不匹配声学环境泛化能力差的问题，提出了信号处理和深度学习融合的语音增强框架，该框架将深度学习对语音和噪声声学环境的强大的感知能力融入到传统语音增强框架中，依照传统语音增强方法的一般结构，利用双隐层GRU和单隐层前馈网络构建了一个新颖的噪声追踪网络，整个网络分为语音感知网络和噪声声学环境感知网络，它们分别负责感知语音存在概率和噪声声学环境变化。模型的各个模块通过最终维纳滤波输出的频谱逼近目标进行联合优化。在CHiME-4和NOISEX-92上的系统实验证明所提方法显著优于单纯的深

度学习方法，对不匹配的声学环境具有很好的推广能力。本章所提方法对基于深度学习的单声道语音增强的实用化是一个积极的探索，未来可以很容易拓展到多通道的情况。

第七章　总结与展望

7.1　总结

对于单声道语音增强，由于只有一个麦克风，可用的信息只有语音的频谱特性和噪声的统计特性，它们是设计单通道语音增强算法的关键信息。传统信号处理的方法和近年来流行的监督式的语音增强的方法分别从这两个方面进行了一系列的探索。本文关注于单声道语音增强。一方面对传统信号处理方法，从信号模型、问题定义和一般框架等基本理论进行了简单的介绍。另一方面，重点介绍了监督式的语音增强方法，归纳总结了监督式语音增强的方法一般处理流程，基本框架和研究方法，针对监督式语音增强方法的特征、模型和目标三个方面进行了详细的介绍和比较。对目前单声道语音增强的各种前沿方法进行了分析和比较。传统基于信号处理的单声道语音增强方法主要依赖噪声统计特性的估计，通常假设噪声声学环境是平稳和慢变的，尽管对于平稳噪声表现出了很好的增强效果和推广性能，但仍难以应付非平稳噪声和复杂多变的声学环境。而监督式的单声道语音增强方法主要依赖于语音和噪声频谱特性的区分性来实现语音和噪声的分离，深度学习对语音和噪声具有强大的感知能力，是目前监督式语音增强的主流方法，监督式语音增强方法通常利用深度学习模型直接学习一个从带噪特征到分离目标的映射函数，在处理非平稳噪声方面相对于传统的语音增强方法具有明显的优势。但这些方法通常简单地将语音增强表达成一个监督式学习问题，过度地依赖于训练数据的粗暴训练，而忽略了语音学和信号处理领域专业知识，在面临不匹配的声学环境时，比如训练集中不存在的噪声类型和输入信噪比，性能会严重下降。针对这些问题，我们以深度学习为着眼点，瞄准语音固有的频谱特性和噪声声学环境的深度感知进行探索，深入研究了知识驱动下的单声道语音增强算法，提出了一系列新的方法和思路。

（1）由于语音的产生机制，语音信号本身具有明显的时序相关性、自回归性和时空结构等频谱特性。针对这些特性，我们提出了时序深度层叠网络、基于循环结构的自回归语音分离网络以及两阶段多目标的自编码网络等新颖网络结构和优化目标，巧妙地将人类对语音信号的先验认知融入到基于深度学习的语音分离中。这种知识驱动的方法有效克服了大多数数据驱动的监督式语音分离方法对数据的过度依赖而导致泛化性能较差的缺点，增强了语音分离系统的实用性。

（2）语音是由一些基本的发音模式产生的，因此，语音信号中隐含着一些固有的基本频谱模式。NMF能够有效挖掘语音信号中的具有感知意义的基本频谱模式。我们将NMF的重构生成方式融入到基于DNN的监督式语音分离中，

将判别式和生成式过程融合在一起，提出了DNN和NMF联合协作的语音分离框架。即充分利用了NMF对语音基本频谱模式的挖掘能力，又充分发挥了DNN强大的非线性映射学习能力，同时避免了类似工作只学习NMF表征系数而造成累计误差的缺陷，理论分析和实验结果都证实了该项工作显著地优于之前基于DNN的语音分离方法。

（3）在真实环境中，语音所处的噪声声学环境通常是复杂多变的，比如，噪声的平稳性随时间变化。基于传统信号处理的语音增强通常忽略了真实场景中噪声声学环境的不确定性，假定噪声声学环境是稳定的，采用确定性统计信号模型求解最优滤波器。而深度学习模型具有强大的感知能力，能够感知复杂环境中的语音和噪声。我们将深度学习强大的感知能力融入到基于信号处理的语音增强框架中，利用深度学习模型感知混合信号中的语音存在概率和噪声声学环境的变化，提出了融合信号处理知识和深度学习的深度噪声追踪网络，既利用了深度学习对噪声声学环境和语音的强大感知能力，又充分发挥了信号处理的专家知识，显著提升语音增强的性能和实用性。

7.2 展望

单声道语音增强发展至今已经有几十年了，基于传统信号处理的语音增强的基本理论和框架已基本成熟，积累了许多重要的经验和知识，但传统信号处理的方法通常假设噪声声学环境是平稳和慢变的，仍难以应付非平稳噪声和复杂多变的声学环境，大大限制了其实用性。最近十年里，监督式的语音增强方法得到极大的发展，特别是基于深度学习的语音增强方法，显著地提高了语音增强的性能，特别在处理非平稳噪声方面相对于传统的语音增强方法具有明显的优势，但由于过度地依赖于训练数据的粗暴训练，基于深度学习的语音增强方法面临着较大的泛化性能问题，同时由于深度学习方法所要求的计算量比较大，难以在实际终端上实时应用，这也限制了监督式语音增强方法的应用。另外，单声道语音增强不可避免地会带来语音畸变，由多个麦克风组成的阵列，能够有效控制语音畸变，实现语音不失真增强。综上，我们认为领域知识和深度学习融合的方案以及将深度学习融入到多通道语音增强框架是未来的研究趋势，未来可以从以下几个方面进行更深入的研究：

（1）泛化性：基于深度学习的监督式语音增强由于对数据的过于依赖存在较大的泛化性问题，一般的做法是扩大数据的覆盖范围，但现实情况很难做到覆盖大部分的声学环境，同时训练数据增大带来训练时间的急剧增大，不利于模型更新。我们认为通过有效融合先验知识可以显著增强模型的推广能力。融合先验知识可以从以下两个方面进行。a. 挖掘人耳的听觉心理学知识，将计算听觉场景分析的知识融入到监督式语音分离模型中。人耳听觉系统对声音的处理过程具有很强的鲁棒性，计算听觉场景分析试图模拟人耳对声音的处理过程，将语音分离分解为分段和组织两个过程，计算听觉场景分析对噪声没有任何假

设，甚至能够有效实现多说话人的分离。通过将深度学习和计算听觉场景分析有机融合在一起，可能会显著提高监督性语音分离的泛化性能。b. 传统信号处理和深度学习深度融合，传统信号处理的语音增强经过多年的发展，形成了一套成熟可靠的结构。将深度学习和信号处理进行深度融合既能发挥深度学习强大的感知能力又能充分利用信号处理领域的专业知识。本文章节六在这方面是一个有意义的尝试，但仅是基于维纳滤波的框架，解决了降噪的问题。还有更多信号处理的知识值得融合，特别是降混响的信号处理知识。

（2）实用性：单声道语音增强可利用的只有语音和噪声的固有特性，理论上消除噪声的同时不可避免地会带来一定的语音畸变。相比于单声道语音增强，由多个麦克风组成的麦克风阵列能够提供额外的空间信息，产生空间滤波器，从所需要的语音源方向获取高品质的语音信号，同时抑制其他方向的干扰。大量的理论研究表明，多通道语音增强算法能够在消除噪声的同时有效控制语音畸变，其语音增强效果明显优于单声道语音增强方法，因此在实际应用中，特别对于远场语音交互应用，基于麦克风阵列的前端处理是目前的主流。但基于麦克风阵列的语音增强算法通常对麦克风阵列硬件要求比较苛刻，既有要求组成阵列的每个麦克风具有较高的物理一致性，又要求较高精度的麦克风阵列几何构型，这会导致在实际应用中麦克风阵列的硬件成本急剧上升。另外，有些多通道语音增强算法，对麦克风阵列的几何构型和麦克风间距有一定要求，比如圆形或者线形，这对产品的外形结构提出了一定的要求，大大限制了麦克风在实际产品中的灵活性。由少数麦克风组成的微型麦克风阵列（通常为双麦）兼顾性能的同时成本低，配置灵活，具有极大的市场应用价值。微型麦克风阵列由于麦克风个数较少，对声源空间方位信息的估计比较困难，很难使用定向波束增强。微型麦克风阵列语音增强算法的设计一般基于多通道维纳滤波的框架，多通道维纳滤波无需声源方位的估计，但其性能强烈依赖于噪声的统计特性估计。基于传统信号处理的微型阵列语音增强通常假设噪声的统计特性是平稳的，而采用确定性统计信号模型求解最优滤波器。而在真实应用场景中，语音所处的声学环境通常是复杂多变的，基于传统信号处理的噪声统计估计面临着巨大的挑战。当噪声估计不准时，语音增强将产生较大语音畸变，而深度学习模型具有强大的感知能力，将深度学习模型对声学环境的感知能力融入到基于传统信号处理的微型麦克风阵列语音增强框架中，既利用了深度学习的强大感知能力，又充分发挥了信号处理领域的专家知识。

（3）优化目标：从人的角度来讲，语音增强的目的旨在提高带噪语音的高层感知单元的可懂度，而从机器的角度，语音增强的目的旨在减小真实测试数据和训练数据的分布差异。目前很多监督式语音增强算法的优化目标本质上都是在寻找局部信噪比意义下最优，具体来说就是在计算每个时频单元的优化误差，平等地看待每个时频单元，对人的感知而言没有考虑高层感知单元的完整性，对机器而言，没有考虑具体应用的最终目标。探索与最终目标更加直接的

或者更能反映语音本质的优化目标将具有很大的意义。