密级 _____

中国科学院大学
University of Chinese Academy of Sciences

博士学位论文

基于数据训练的单通道语音增强算法研究

作者姓名 李煦

指导教师 潘接林 研究员　李军锋 研究员

中国科学院声学研究所

学位类别 工学博士

学科专业 信号与信息处理

培养单位 中国科学院声学研究所

2017 年 5 月

Research on single-channel speech enhancement based on data training

By

Xu Li

A Dissertation Submitted to
University of Chinese Academy of Sciences
In partial fulfillment of the requirement
For the degree of
Ph.D. of Signal and Information Processing

Institute of Acoustics
Chinese Academy of Sciences

May, 2017

中国科学院声学研究所
学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：本论文的所有工作，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：李照

日期：2017年5月27日

学位论文使用授权说明

本人完全了解中国科学院大学关于收集、保存、使用学位论文的规定，即：

- 按照中国科学院大学要求提交学位论文的印刷本和电子版本；
- 中国科学院大学与中国科学院声学研究所有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务；
- 中国科学院大学与中国科学院声学研究所可以采用影印、缩印、数字化或其它复制手段保存论文；

（保密论文在解密后遵守此规定）

论文作者签名：李照 导师签名：

日期：2017年5月27日

摘　要

随着通信技术的发展，电话会议系统、车载免持电话和VoIP等通信系统不断涌现，并且随着智能可穿戴设备、智能家居和智能车载系统等领域的兴起，越来越多的设备如智能音箱，智能手表等具有语音交互功能。这些设备需要满足用户在各种实际声学环境下进行语音通信或交互的需求，如在嘈杂的户外和在混响较强的室内等。在这些实际环境中，目标语音容易受到环境噪声、非目标说话人干扰和房间混响等信号的影响。为了提高语音质量，实际中需要设计算法对传声器采集到的信号进行增强处理，使得语音增强技术研究成为一个重要课题。而且很多设备由于尺寸和成本的限制，通常只有一个传声器拾取信号，使得单通道语音增强技术成为一个研究热点。传统的单通道语音增强算法一般只能处理较平稳的噪声，对非平稳噪声抑制效果不佳，限制了这类算法在实际场景中的应用。本文主要研究了复杂环境下基于数据训练的单通道语音增强算法，利用训练数据本身带来的先验信息来提高算法对非平稳噪声的抑制能力。主要研究工作及创新点包括：

1. 对语音信号的时间连续性进行研究。标准的基于非负矩阵分解的单通道语音增强算法不需要假设噪声是平稳信号，而且能够利用语音和噪声训练数据中的先验信息，因此能够对非平稳噪声有较好的抑制效果。然而该算法假设语音相邻帧是相互独立的，没有考虑语音信号的时间连续性。本文提出了一种基于非负矩阵分解和k均值聚类的语音建模方法，能同时对语音信号的频谱结构信息和时间连续性信息建模。并且将该语音建模方法和因子条件随机场结合对混合信号的时间动态特性建模，用于分离两个说话人的语音信号以及分离语音和噪声信号。实验结果表明，该算法相比其它一些算法在主客观评价指标上都有较大的提升。

2. 对基于非负矩阵分解的无监督语音增强算法进行研究。在基于非负矩阵分解的无监督语音增强算法中，首先通过大量语音训练数据得到全局语音模型，然后通过组稀疏惩罚项从全局语音模型中选择少量的说话人字典描述测试信号中未见过的说话人语音信号，同时从测试信号中估计噪声字典，实现无监督语音增强。本文针对算法中的组稀疏惩罚项进行研究，分别提出了基于自适应组稀疏惩罚项和动态组稀疏惩罚项的无监督增强算法，能够更好地从全局语

音模型中选择匹配的语音字典。结果表明，所提组稀疏惩罚项能够在保留语音成分的同时抑制更多的噪声信号，提高增强效果。除此之外，我们提出了一种与说话人无关的语音模型，相比全局语音模型能够更好地对语音频谱进行建模，然后将所提语音模型用于无监督在线语音增强，对测试信号逐帧进行增强处理，具有重要的实践价值。

3. 对基于非负矩阵分解和深度神经网络的语音增强算法进行研究。近年来，非负矩阵分解和深度神经网络已经被结合用于单通道语音增强，其中非负矩阵分解用于描述语音信号的频谱结构，深度神经网络用于估计非负矩阵分解的权重矩阵。在本文中，我们将性别信息引入到基于非负矩阵分解和深度神经网络的语音增强算法中，通过引入新的先验信息，即性别信息来进一步提高语音增强效果。在训练阶段，针对男性和女性说话人分别训练深度神经网络-非负矩阵分解模型；在增强阶段，提出了一种性别鉴定算法用于判断每段测试信号中的说话人性别，然后选用对应性别的模型进行语音增强。实验结果表明增加说话人性别这一先验信息能够有效提高增强效果。

4. 对混响环境下的语音增强算法进行研究。在本文中，我们提出了一种混响环境下的理想浮值掩蔽定义，将目标语音的直达声和早期混响部分作为期望信号，晚期混响部分和噪声作为残余信号。然后采用深度神经网络模型估计新定义的理想浮值掩蔽，最后将估计得到的浮值掩蔽用于原始信号进行语音增强。同时我们也进行了一系列实验来对噪声和混响对于增强结果的影响进行了研究。实验结果表明，在很多测试环境下，本文所提算法相比原始信号都能够有效提高语音质量和可懂度。

关键词： 单通道语音增强，噪声抑制，非负矩阵分解，深度神经网络

Abstract

With the development of communication technology, communication systems such as teleconferencing systems, hands-free car telephone and VoIP systems continue to emerge. Moreover, the areas of smart wearable devices, intelligent vehicles and smart home have been light in competition over the past years, thus more and more devices such as smart loudspeaker and smart watch are designed with voice interaction features. These new systems are expected to meet the need of voice communication and interaction in different realistic acoustic scenarios, such as in noisy outdoor and strong reverberant indoor environments. In these realistic scenarios, the target signal is usually corrupted by background noise, non-target speaker's speech and reverberation, which affect the experience of voice communication and interaction. In order to improve the quality of speech signal, algorithms should be designed to enhance the mixture signal picked up by the microphone, making the research on speech enhancement technology an important issue. And due to the limitation of the size of equipment and the cost, sometimes only one microphone is applied to pick up the mixture signal, making the single-channel speech enhancement a hot research point. Traditional single-channel speech enhancement algorithms can only suppress stationary noise, which restrict the application of traditional speech enhancement algorithms in real world. In this thesis, we study the single-channel speech enhancement algorithms based on data training in complex environments. A prior information included in the training data is employed to improve the performance of single-channel speech enhancement algorithms in suppressing non-stationary noise. The main contents and contributions of thesis are summarized as follows:

1. Research on the temporal continuity of speech signal. Standard non-negative matrix factorization (NMF) based methods can employ the prior information included in speech and noise training data and do not assume the stationary of noise signal, which can suppress non-stationary noise well. However, one problem of standard NMF-based methods is that they ignore the tempo-

ral continuity of speech signal. A speech modeling method based on NMF and k-means clustering is proposed to simultaneously model the spectral structure and temporal continuity of speech signal. Then, the proposed speech modeling method and factorial conditional random field (FCRF) model are combined to jointly model the temporal dynamics of mixture signal for speech separation and speech enhancement. Experiments show that the proposed algorithm provides significant improvements in subject and objective evaluation metrics over some NMF-based methods.

2. Research on non-negative matrix factorization based unsupervised speech enhancement. In NMF-based unsupervised speech enhancement algorithms, a universal speech model (USM) is first learned from large amount of speech training data, and then in the enhancement stage only a very small number of speakers' dictionaries (groups) in the USM that best fit the unseen speaker's signal in the test data are active using a group sparsity penalty. In this thesis we propose adaptive group sparsity penalty and dynamic group sparsity penalty for unsupervised speech enhancement, which can select the matched groups from the USM more accurately. Results show that the proposed penalties are able to suppress more noise without introducing more artifacts. What's more, we propose a novel speaker-independent speech model to describe different local spectral structure of speech signal and use the proposed speech model for online speech enhancement, which is appropriate to use in real applications.

3. Research on speech enhancement based on non-negative matrix factorization and deep neural networks (DNN). Recently, NMF and DNN has been combined for single-channel speech enhancement, in which NMF is used to describe the spectral structure of speech signal and DNN is used to estimate the activations of NMF. We introduce the gender information into the DNN-NMF based speech enhancement algorithm. Specifically, in the training stage, two gender-correlated DNN-NMF models are trained using the gender-specific training data. In the test stage, an algorithm based on NMF and group sparsity penalty is proposed to identify the gender of the speaker in the test signal. Then the corresponding DNN-NMF model is selected for speech enhancement. Experimental results show that introducing the gender information can provide

substantial improvements over some NMF-based and DNN-based methods.

4. Research on speech enhancement in noisy reverberant conditions. We propose a new ideal ratio mask (IRM) definition in reverberation conditions where the direct sound and early reflections of target speech are regarded as the desired signal. DNN is employed to estimate the IRM in the noisy reverberant conditions. The estimated IRM is then applied to the noisy reverberant mixture for speech enhancement. Experiments are conducted to examine the effect of noise and reverberation on the enhancement results. Results show that the estimated IRM provides substantial improvements in speech intelligibility and speech quality over the unprocessed mixture signals in all test conditions.

Keywords: single-channel speech enhancement, noise suppression, non-negative matrix factorization, deep neural networks

目　录

表　格

插　图

第一章　绪论

1.1　研究背景和意义

近年来，随着信息技术的发展，电话会议系统、车载免持电话、视频通信和VoIP等通信系统不断涌现，语音通信技术受到越来越广泛的应用。然而，人们在使用通信系统时，不可避免地会受到各种噪声的干扰。例如，电话会议系统中的回声干扰、车载免持电话中的汽车噪声和公路噪声以及视频通信中的干扰语音等，都会影响语音通信的质量。为了抑制噪声，得到纯净的语音信号，需要设计算法对传声器采集到的信号进行处理，使得语音增强算法成为当前语音通信领域的研究热点。

另一方面，随着2011年苹果公司在手机端推出Siri智能语音助手，语音交互应用开始进入普通消费者的视野，并且随着智能可穿戴设备、智能车载和智能家居等领域的兴起，越来越多的电子设备具备语音交互功能，如亚马逊推出的Echo智能音箱，苹果公司推出的Apple Watch等智能电子设备。为了满足用户在各种实际声学环境下实现语音交互的需求，需要语音识别技术有较高的准确率。然而，实际环境中的各种噪声信号尤其是非平稳噪声会严重影响语音识别系统的性能，导致很多语音交互设备很难在实际环境中应用。因此，为了提高语音识别的准确率，一种常见的做法是在语音识别的前端加入语音增强模块，首先通过增强算法抑制噪声得到纯净的语音信号，再送入识别器进行识别。然而，很多语音增强算法在抑制噪声的同时，也造成了语音失真，导致并不能带来识别率的提高。因此，研究适用于语音识别的语音增强技术也越来越受到人们的关注。

此外，语音增强技术在助听器领域也得到了广泛应用。随着人口老龄化，听力障碍不仅是一个公共卫生问题，也成为一个社会问题。全球65岁以上人口中，每三位就有一位患听力障碍。我国第二次残疾人抽样调查显示，全国60岁以上老年人患听力残疾的比例达11%，也就是说老年人听力残疾人总数超过2000万。为了解决听力受损问题，需要借助助听器设备。为了提高信噪比，很多助听器系统都把语音增强作为预处理模块，用于抑制噪声信号。然而，研究表明很多语音增强算法虽然能够抑制噪声和提高信噪比，但是并不能

带来语音可懂度的提高，只是提高了听者的舒适度。因此，研究适用于助听器设备的语音增强技术具有重要的实用价值。

1.2 研究历史和现状

早在20世纪60年代，贝尔实验室的先驱者Schroeder就开始研究语音增强技术，至今已有五十多年的历史，语音增强技术的研究也得到了快速发展。目前国内外的诸多大学和研究机构以及国内外很多知名企业都在致力于语音增强的研究并取得了很多成果。在国内，一些高校如清华大学、北京大学、中国科学院声学研究所以及自动化所等单位都开展了语音增强的研究工作。国外的一些研究机构如University of Texas at Dallas, BarIlan University, KU Leuven以及Ohio State University也都提出了很多语音增强算法，大大推动了语音增强技术的发展。国内外的很多科技企业如Microsoft、NTT以及华为等也都在进行语音增强技术的研究，致力于语音增强技术的实用化。

语音增强算法根据通道数目来分类，可以分为单通道语音增强算法和多通道语音增强算法。下面就语音增强算法的国内外研究现状做一个简单的介绍。

1.2.1 单通道语音增强算法

单通道语音增强算法主要是利用语音和噪声信号在时频域分布的不同特点，从而区分语音和噪声信号。传统的单通道语音增强算法主要有谱减法、维纳滤波方法、统计模型方法和子空间方法等。近年来基于数据训练的方法也获得了广泛关注，主要有基于非负矩阵分解的方法和基于深度神经网络的方法等。

1.2.1.1 谱减法

1979年，Boll提出了基于短时幅度谱估计的单通道语音增强算法，也称为谱减法 [1]。谱减法由于原理简单、性能稳定和计算量小而得到了广泛的关注和研究。具体而言，假设噪声是加性的，然后通过从带噪信号的幅度谱中减去估计得到的噪声谱即可得到增强后的纯净语音谱。然而，谱减法存在严重的“音乐噪声”（Musical Noise）问题 [2] 以及会带来语音失真。为了解决这些问题，很多改进型的谱减法被提出。Berouti提出了一种采用过减因子的谱减法，通过从带噪信号幅度谱中减去过估计的噪声信号，同时限制得到的语音谱不小

于一个最小值 [2]。Lockwood考虑到噪声对语音不同频段的影响是不同的，提出了一种非线性谱减法，对不同的频率选取不同的过减因子 [3]。沿着这一思路，有学者提出了多频带谱减法，将信号分成不同的频带，对每个频带分别进行谱减法，减少了语音失真 [4]。Sim等人提出了一种基于最小均方误差的参数选取方法来最优地选择谱减法的参数，相比依照经验设置参数的算法能有效提高增强效果 [5]。Sovka将维纳滤波和谱减法结合起来，能够随时更新噪声谱信息，提高了算法的鲁棒性 [6]。Gustafsson等人提出了自适应平滑增益函数的谱减法来解决音乐噪声问题 [7]。与此同时，Virag提出了一种基于听觉感知的谱减法，将人类听觉系统特性考虑进来提高语音可懂度 [8]。

1.2.1.2 维纳滤波算法

另一种单通道语音增强算法是基于维纳滤波的方法。不同于谱减法主要是基于直觉和经验的准则，维纳滤波算法通过解决一个最优化问题来增强语音信号。维纳滤波主要是通过最优化算法得到一组滤波器系数，使得原始信号经过滤波器的输出和参考信号的均方误差（Minumum Mean-Square Error, MMSE）最小 [9–11]。接着有学者提出了参数化的维纳滤波方法，能够根据选取不同的参数来控制语音失真和噪声抑制量的相对大小，实现两者之间的折中 [12]。Lim和Oppenheim提出了一种迭代的维纳滤波方法，并对语音信号采用全极点AR模型建模，然后反复迭代得到增强后的语音信号 [13]。这种方法也存在一些问题，例如无法确定何时终止迭代以及算法会产生虚假的谱峰。为了解决迭代维纳滤波算法存在的问题，一些改进算法被提出来。有些学者提出了通过限制相邻帧间谱变化或者对应AR参数变化的方法来去除虚假的谱峰。另一些学者提出通过限制相邻迭代次数之间自相关系数的变化来去除虚假谱峰，相比无限制的维纳滤波算法和谱减法在性能上有明显提升 [14–16]。另外有学者提出了带约束的维纳滤波算法，通过解决一个带约束的最小值问题来同时控制语音失真和噪声抑制，然而这种算法很难选择合适的拉格朗日参数使得残余噪声足够小 [17, 18]。有学者提出了基于人耳听觉系统的带约束维纳滤波算法来解决这个问题，并取得了不错的增强效果 [19–21]。Scalart等通过决策导向算法估计一个低方差的先验信噪比，然后利用估计得到的先验信噪比来计算维纳增益函数，从而有效去除了音乐噪声 [22]。维纳滤波算法是一个线性滤波器，然而在很多情况下非线性滤波器能得到更好的结果。

1.2.1.3 基于统计模型的语音增强算法

不同于维纳滤波方法，基于统计模型的方法是一种非线性估计方法，通过不同的统计模型和最优化方法得到不同的非线性估计器进行语音增强 [23]。例如，给定观测数据，也就是带噪数据，然后通过最大似然估计或者贝叶斯估计得到增强后的语音信号。McAulay和Malpass最先将最大似然估计用于语音增强，假设噪声是高斯分布，然后最大化似然函数估计得到增强后的语音谱，相比谱减法和维纳滤波算法该算法的衰减较小 [24] 。Ephraim和Malah提出了基于高斯模型假设的短时谱幅度最小均方误差估计子（MMSE-STSA） [25] 和短时对数谱幅度估计子（MMSE-LSA） [26] ，有研究表明这两种方法能有效抑制音乐噪声。在MMSE-STSA和MMSE-LSA算法中，需要估计先验和后验信噪比，而先验信噪比相比后验信噪比更难估计。很多算法被提出来估计先验信噪比，如基于最大似然估计的方法 [25] 、基于决策-导向（decision-directed）的方法 [25] 以及各种改进的决策- 导向方法 [27–31]。考虑到高斯分布并不能很好地描述语音信号傅里叶变换系数的分布，很多学者提出了基于非高斯模型的MMSE估计算法，如基于拉普拉斯概率分布的模型和基于伽马分布的概率模型 [32, 33]。实验结果表明基于非高斯模型的算法相比基于高斯模型算法在客观评价指标上有微弱提高。后续又有学者提出了基于最大后验概率的谱估计方法，相比基于MMSE的方法更加容易计算 [34, 35]。接着又有学者将语音存在概率和基于统计模型的方法结合起来，能够有效去除残余噪声 [36–38]。

1.2.1.4 基于子空间的增强算法

子空间方法是一种主要基于线性代数理论的语音增强算法。具体而言，首先将混合信号的向量空间分解为由语音信号组成的子空间以及由噪声信号组成的子空间，然后去除混合信号中在噪声子空间的向量成分，保留语音子空间中的向量成分，从而进行语音增强。具体的子空间分解方法有基于奇异值分解（SVD）的算法 [39–41]、基于特征值分解的算法 [42]以及基于Karhunen-Loeve变换（KLT）的算法 [43, 44]。研究表明，基于子空间的方法能够有效去除噪声，同时对语音有较好的保留，并且能够有效抑制音乐噪声。然而，由于需要对每一帧混合信号都进行SVD或者KLT分解，导致算法计算量大，影响了实际使用。后续又有学者提出了子空间跟踪的方法，通过迭代更新的算法估计特征值和特征向量来降低算法计算复杂度 [45, 46]。

在上述几种方法中，一个很关键的问题是需要估计噪声信号。传统的噪声估计算法是基于平稳噪声的假设，然后根据某种判决机制跟踪估计噪声。Martin提出了基于最小统计量（minimum-statistics, MS）的噪声估计方法 [47–49]，通过在一个窗长内跟踪带噪信号幅度谱最小值作为噪声幅度谱。算法由于跟踪最小值导致噪声估计偏小，而且不能及时跟踪噪声的变化。Doblinger提出了一种基于最小跟踪（minimum tracking）的噪声估计算法，相比MS算法运算复杂度低而且能够连续更新噪声估计，但是在语音存在段有可能导致噪声过估计 [50]。另外一类比较常用的噪声估计方法是基于时间递归平滑的方法。Lin提出了基于后验信噪比的递归平滑算法，将平滑因子设置为关于后验信噪比的sigmold函数 [51]。Hirsch提出了基于加权谱平滑的噪声估计方法，并且提出了一种方法用于控制噪声的更新与否 [52]。Cohen等人提出了基于最小控制递归平滑的噪声估计方法 [53, 54]，相比MS算法能够获得较小的估计误差，在实际中也应用更多。然而，这些噪声估计方法大多对平稳噪声有较好的估计效果，而对非平稳噪声不能准确估计，导致了对非平稳噪声抑制效果不佳。

1.2.1.5 基于NMF的语音增强算法

近年来，基于数据训练的单通道语音增强算法由于对非平稳噪声有较好的抑制效果而得到了广泛关注和研究，如基于非负矩阵分解（non-negative matrix factorization, NMF）的语音增强算法和基于深度神经网络（deep neural networks，DNN）的语音增强算法。Lee和Seung首先提出了NMF的概念，主要利用NMF学习图像中的局部信息，后来学者们将其用于语音增强 [55, 56]。基于NMF的语音增强算法能够利用语音和噪声训练数据的先验信息，并且没有了平稳噪声的假设，相比传统的单通道语音增强算法能够对非平稳噪声有更好的抑制效果。这类算法一般包含两个步骤：步骤一为训练阶段，通过对语音和噪声的训练数据进行非负矩阵分解得到对应的字典矩阵，这些字典用于描述语音和噪声的频谱结构；步骤二是增强阶段，对测试信号幅度谱进行非负矩阵分解，固定语音和噪声的字典矩阵，迭代得到权重矩阵。最后将语音字典和其对应的权重矩阵相乘来重构增强后的语音信号幅度谱 [57]。针对上述的语音增强框架，学者们进行了大量研究并提出了很多改进算法。首先针对分解时目标函数的选择，学者们提出了不同的目标函数，主要有均方误差、广义KL散度准则（generalized Kullback-Leibler divergence） [58, 59] 和IS散度（Itakura-Saito

divergence） [60]。广义KL散度准则和IS散度准则由于能够对语音信号的高频和低频成分误差都具有较好的描述能力而得到了广泛应用 [61]。另一个研究工作是关于矩阵分解算法的选取。一种比较常用的分解算法是由Lee和Seung提出的乘法迭代准则（multiplicative updates rules） [56]。该算法实现简单，而且能保证误差函数值一直下降，并收敛到局部最优点。此外还有其它一些分解算法被提出，如基于二阶矩的方法 [62]、基于映射梯度的算法 [63] 以及基于激活集牛顿算法 [64,65]等。基于NMF的语音增强算法另一个很关键的问题是字典的获取。好的字典主要有以下一些性质：第一是字典能够准确描述该音源的频谱结构，并且针对该音源有较好的泛化能力；第二是字典要尽量小，从而降低计算复杂度；第三是字典要有足够的鉴别能力，一个音源的字典不能用于描述另一个音源。目前有比较常用的几种获取字典的方法，如通过矩阵分解算法得到字典 [66,67]、通过对训练数据幅度谱进行k均值聚类的方法以及对训练数据随机采样的方法 [65,68,69]。除此之外，为了提高语音增强的效果，经常需要对算法加入一些约束来尽可能满足语音信号的一些性质。例如，由于语音信号本身的稀疏特性，需要对权重矩阵加入稀疏惩罚项使得只有少量的语音字典向量用来描述语音信号 [58,70–72]；另外由于语音信号具有很强的短时连续性，需要对其连续性建模 [58,73–80]。而且为了方便算法能够实际应用，很多学者提出了基于NMF的在线增强算法，能够实时处理混合信号 [81–83]。

1.2.1.6 基于DNN的语音增强算法

在基于深度神经网络的语音增强算法中，语音增强被当作是一个有监督的学习任务，通过监督学习算法训练一个模型来预测目标语音 [84]。根据训练目标的不同，上述算法可以分为基于掩蔽值的增强算法和基于谱映射的增强算法。在进行模型训练时，模型的输入一般是混合信号的特征，如幅度谱，MFCC特征等。基于掩蔽值的语音增强算法通过学习一个映射函数描述混合信号特征到时频掩蔽值的非线性关系，然后利用估计得到的掩蔽值结合混合信号得到增强后的语音信号。时频掩蔽值根据计算方法的不同主要分为理想二值掩蔽（ideal binary mask, IBM）和理想浮值掩蔽（ideal ratio mask, IRM）。对于IBM [85,86]，如果该时频点信噪比大于某个阈值，则该时频点IBM设置为1，表明语音占主导地位，否则IBM设置为0。对于IRM [87–89]，其IRM值为目标语音能量和混合信号能量的某种比值，是一个连续值。Wang第一次将DNN引入到语音增强中用于估计IBM，实验表明这类基于DNN的增强算法明显优于

传统的语音增强算法 [90]。接下来，Wang等人针对DNN估计目标的选取进行了大量实验，结果表明IRM相比IBM更能提高增强后的语音质量 [91]。Healy等人研究表明基于DNN的IBM和IRM估计算法对于听力受损人群能够有效提高语音信号的可懂度 [92, 93]。Huang将IRM融合到DNN框架中，最小化语音谱均方误差，相比基于NMF的算法能有效提高增强效果 [94, 95]。Weninger将信号近似（signal approximation, SA）作为目标函数，并将LSTM(long short-term memory)引入进来，相比基于DNN和NMF的算法能进一步提高了增强效果 [96]。Williamson提出了一种复数域的IRM用于语音增强，能够同时估计语音信号的幅度和相位信息，显著提高了增强效果 [97]。

基于谱映射的增强算法主要是学习从混合信号到目标语音的回归函数。Xu等人训练了一个DNN模型作为回归模型来估计目标语音的对数能量幅度谱，结果表明增强效果明显优于传统的增强算法 [98, 99]。Han等人训练了一个DNN模型用于学习从混响语音到纯净语音信号的回归模型，能够有效提高信噪比和语音可懂度 [100, 101]。Tu等人训练了DNN模型来同时估计目标语音和干扰信号，结果表明同时估计目标和干扰能进一步提高增强效果 [102]。此外，由于NMF能够描述音源的频谱结构信息，DNN能够学习大量输入数据之间的非线性关系，DNN和NMF已经被结合用于单通道语音增强，相比基于DNN的算法和基于NMF的算法能有效提高增强效果 [103–105]。

1.2.2 多通道语音增强算法

与单通道增强算法相比，多通道语音增强算法除了能够利用时、频域信息外，还可以利用空间上的区分度进行噪声抑制。主要的多通道语音增强算法有波束形成、维纳滤波和盲源分离等方法。

波束形成算法又被称为空域滤波算法。这类算法大致可以分为两类，第一类为固定波束形成（Fixed Beamforming），通过使用一组固定的滤波器系数来增强某一特定区域的声源，而尽可能抑制来自于其它方向的声源。经典的固定波束形成算法有延迟相加和超指向波束形成算法等 [106]。固定波束形成算法由于滤波器系数固定，不随时间或者输入信号的变化而变化，因此当声学环境复杂多变时，噪声抑制能力有限。第二类波束形成算法为自适应波束形成算法（Adaptive Beamforming），其滤波器系数根据输入的变化而变化，能够适应时变的声学环境，因此能得到更好的结果。线性约束最小方差（Linearly

constrained minimum variance, LCMV）算法是最早被提出的自适应波束形成算法之一 [107]。此后，广义旁瓣抵消（Generalized sidelobe canceller, GSC）算法将LCMV中的带约束条件的优化问题转化为无约束优化问题，实现简单，得到了广泛研究 [108]。然而，在实际环境下，如在扩散场噪声和混响环境下，自适应波束形成算法的性能下降显著。近年来，很多改进的自适应波束形成算法被提出来提高算法的鲁棒性 [109–111]。另一种改善波束形成算法的思路是在波束形成后级联一个后置滤波器。Zelinski首先提出了波束形成器的后滤波方法，但是该方法只对白噪声抑制能力较好 [112]。McCowan等人对Zelinski的算法进行了扩展，引入了扩散噪声模型。Yousefian对复相干函数进行了研究，在混响较小情况下对相干干扰有不错的抑制能力 [113]。

与单通道维纳滤波方法类似，多通道维纳滤波（Multi-channel Wiener Filter, MWF）也是通过解决一个最优化问题来增强目标语音信号。在计算增益函数时，MWF需要估计噪声的协方差矩阵，然而在目标和干扰信号同时存在的情况下，很难准确估计方向性干扰的协方差矩阵。Simon将语音失真引入到最优化准则中，提出了一种语音失真加权的多通道维纳滤波算法（Speech distortion weighted multi-channel Wiener filter, SDW-MWF），能够在语音失真和噪声抑制二者之间进行折中 [114]。Spriet等人提出了基于空间预处理的SDW-MWF算法，使用延迟相加的结果作为参考信号，使得算法具有了一定的空间分辨能力，并且指出了加权因子对于语音失真和噪声抑制的影响 [115]。在这之后，基于语音存在概率、人耳听觉特性和残余噪声控制的加权因子被陆续提出，使得MWF算法在语音失真和噪声抑制方面能够得到更好的折中 [116–118]。

盲源分离算法（Blind source separation, BSS）是一种只依赖当前的输入信号，而不需要事先获取阵列拓扑结构，目标声音位置等先验信息的多通道增强算法。算法假设需要分离的信号源是相互独立的，通过解决一个最优化问题，如最大化非高斯特性，最小化互信息等问题对各个信号源信号进行分离 [119]。这类算法不依赖阵列拓扑结构，因此对模型不匹配较为鲁棒，然而也存在一些问题。首先，算法假设各个音源信号是相互独立，而在实际环境中很难满足；其次，算法复杂度较高，计算量大，并且存在排列奇异性等问题，导致算法很难在实际环境中应用。

1.3 本论文主要研究内容

本论文主要对复杂环境下基于数据训练的单通道语音增强算法进行研究，通过利用大量训练数据带来的先验信息提高算法在复杂噪声环境下的性能。各章节内容简述如下：

第一章绪论，介绍了论文工作的研究背景和意义，综述了单通道和阵列语音增强相关算法的研究历史和现状，并说明了本文的研究内容。

第二章介绍了单通道语音增强的基础知识。一方面介绍了传统的基于信号处理的语音增强算法，并说明了这些经典算法所面临的一些问题。另一方面介绍了最近几年提出的基于数据训练的单通道语音增强算法。

第三章对语音信号的时间连续性进行研究。提出了一种基于非负矩阵分解和k均值聚类的语音建模方法，通过语音训练数据得到若干语音字典和一个状态序列来同时对语音信号的频谱结构信息和时间连续性信息建模。并且将该语音建模方法和因子条件随机场结合对混合信号的时间动态特性建模，用于分离两个说话人的语音信号以及分离语音和噪声信号。该算法通过采用多个小字典描述语音信号的频谱，提高了字典的鉴别能力，并且采用因子条件随机场对时间特性建模，进一步提高了算法的增强效果。

第四章研究了基于非负矩阵分解的无监督语音增强算法。针对全局语音模型，首先提出了基于自适应组稀疏惩罚项的无监督增强算法，能够自适应确定稀疏参数来选择对应的语音字典。随后又提出了基于动态组稀疏惩罚项的无监督增强算法，根据不同帧信号选取不同的语音字典，这样能够动态选择最合适的语音字典，从而提高了算法的增强效果。另外，我们提出了一种与说话人无关的语音模型，相比全局语音模型能够更好地对语音频谱进行建模，然后将所提语音模型用于无监督在线语音增强，对混合信号逐帧进行处理，具有重要的实践价值。

第五章研究了基于非负矩阵分解和深度神经网络的语音增强算法。将性别信息引入到基于非负矩阵分解和深度神经网络的语音增强算法中，通过引入新的先验信息来进一步提高语音增强效果。在训练阶段，针对男性和女性说话人分别训练深度神经网络-非负矩阵分解模型；在增强阶段，提出了一种性别鉴定算法用于判断每段测试信号中的说话人性别，然后选用对应性别的模型进行语音增强。实验结果表明增加说话人性别这一先验信息能够有效提高增强效果。

第六章对混响环境下的语音增强算法进行研究。提出了一种混响环境下的理想浮值掩蔽定义，将目标语音的直达声和早期混响部分作为期望信号。然后采用深度神经网络模型估计新定义的理想浮值掩蔽，最后将估计得到的掩蔽值用于混响带噪的原始信号进行语音增强。同时我们也进行了一系列实验来对噪声和混响对于增强结果的影响进行了研究。

第七章对全文进行总结，指出现阶段研究存在的不足，并指出下一步的研究目标。

第二章　单通道语音增强基础理论

本章首先介绍一些与语音增强相关的基础知识；然后介绍单通道语音增强算法中一些经典的算法，同时也对最近研究较多的基于数据训练的增强算法进行介绍。

2.1　语音增强基础知识

在进行语音增强时，对语音和噪声信号的特性了解是十分有必要的。本节主要介绍一些关于语音增强的基础知识，包括语音和噪声信号的主要特性、人耳感知特性和声学环境等。

2.1.1　语音信号主要特性

经过研究表明，语音信号具有以下一些特性：

（1）语音信号是高度非平稳的随机信号，功率谱随时间而变化。然而，在一段比较短的时间内（10-30ms）语音信号的谱结构相对平稳。因此，语音信号的短时谱具有相对稳定性 [120]。在语音增强技术中，很多算法都利用了语音信号的这种短时平稳特性 [25,26]；

（2）通过对语音信号产生的机制进行研究得知，语音可以分为浊音和清音两大类。浊音在时域和频域具有很强的规律性，如在时域具有明显的周期性，而在频域具有清晰的共振峰结构。而且浊音的能量较强，且大部分能量都集中在低频段。清音则没有明显的规律性，具有很强的随机性，在频域也没有共振峰结构，而且能量较弱，且大部分能量集中在中高频段；

（3）语音信号在时频域分布具有稀疏性。语音在时间轴上并不是连续出现，在语音段之间会存在静音段；其次，语音信号也不是在每个频带上都有能量，在频谱上具有明显的稀疏性。因此很多语音增强算法都是利用了语音稀疏特性，假设语音在频域比噪声信号更加稀疏，然后再进行语音增强 [47,53,54]。

2.1.2 噪声信号主要特性

（1）噪声信号在实际场景中无处不在，例如街道上汽车行驶的噪声、车内发动机的声音以及餐厅里人们交谈的背景声音等。其次，噪声信号频谱特性千变万化，既可能是平稳信号，如空调发出的噪声，也可能是非平稳信号，如在餐厅里多人聊天的背景噪声等。非平稳噪声由于随时间而变化，因此更加难以抑制；

（2）噪声信号的频谱形状，特别是能量在不同频段的分布也是不一样的。与语音信号浊音能量主要分布在低频段，而清音能量主要分布在中高频段不同，噪声信号能量分布随噪声类型不同而不同。例如，风噪能量主要集中在低频段，特别是500Hz以下，而车噪能量则分布在较宽的频带范围内 [120]；

（3）噪声信号能量在不同实际环境中也是不同的。在一些室内场所，如医院、教室和办公室等，噪声能量相对较低，大约在50-55dB声压级，信噪比较高，而在列车和飞机等环境下，噪声能量相对较高，大约在70-75dB声压级，导致信噪比较低 [120]。因此，在设计语音增强算法时，需要针对不同信噪比情况在语音失真和噪声抑制方面进行折中。

2.1.3 人耳听觉特性

在实际的听觉场景中，即使是在非常嘈杂的环境中，人耳也能够有效抑制背景噪声和混响。因此，了解人耳听觉特性有助于语音增强技术的发展。人耳听觉分析是一个很复杂的问题，涉及到生理学、心理学和声学等诸多领域。目前，已有的一些研究成果主要包括：

（1）人耳对频率高低的感受近似与该频率的对数成正比 [121]；

（2）人耳在进行定位时，对低频信号主要通过双耳时间差进行定位，而对高频信号主要通过双耳响度差进行定位 [122]；

（3）人耳具有掩蔽效应，主要包括频域掩蔽和时域掩蔽等。频域掩蔽是指在某频段能量高的声源对另一个在该频段附近能量低的声源具有掩蔽作用。时间掩蔽是指在时间上先后进入听觉系统的两个声音之间会造成掩蔽作用，具体又分为前向掩蔽和后向掩蔽效应 [121]。

（4）人耳在嘈杂的环境中会有意识地选择自己所关注的目标声音，忽略或抑制其它干扰噪声信号。这个过程被称作听觉场景分析 [123]。有学者对该

过程进行分析建模，提出了计算听觉场景分析（Computational Auditory Scene Analysis, CASA）模型 [86] 以及基于该模型的语音增强算法 [85,90,92]。

2.1.4　声学环境

在实际场景中进行语音通信和语音交互时，会受到声学环境中很多不利因素的影响，主要包括环境噪声，封闭空间带来的混响等。

对于环境噪声，根据噪声源是否具有明确的声源到达方位角，可以将背景噪声分为具有明确方位的方向性干扰源和来自四面八方的扩散场噪声。例如在室内环境中，干扰说话人的声音具有明确的声源到达角，属于方向性干扰源；而房间内距离传声器较远的空调等电器发出来的背景噪声经过房间墙壁的反射则被认为是扩散场噪声。在实际环境中最难处理的是干扰说话人的语音，因为它和目标语音具有相似的时频域特性，非常难以去除。

除了上述加性噪声之外，封闭声学环境中的混响也会影响语音通信质量和语音交互系统中识别器的性能。声音信号经过封闭空间的反射到达人耳或传声器时可以分为三个部分：直达声，早期反射声和晚期反射声 [124]。直达声是通过直达路径到达人耳或传声器的声音；早期反射声是在直达声之后50到100ms之内的反射声，有研究表明早期反射声能够有效提高语音可懂度 [125]；晚期反射声包含大量来自四面八方的反射，难以区分，一般可以用扩散声场来表征。

2.2　单通道语音增强算法

语音通信和交互系统在实际应用中，不可避免地受到各种噪声的影响，导致了通信质量和交互体验的下降。单通道语音增强算法由于只需要一个传声器、成本较低、使用方便、没有传声器一致性的要求、对声源位置没有要求以及对混响不太敏感的优势而得到了广泛关注和研究。单通道语音增强算法由于只有一路混合信号，只能利用混合信号的时、频域信息提取目标信号。传统的单通道语音增强算法一般假设噪声相比语音更加平稳，首先通过噪声估计算法得到噪声谱，然后再计算增益函数提取增强后的语音谱，只对较为平稳的噪声有较好的抑制能力。而在实际场景中，噪声和干扰更多的是非平稳信号，特别是干扰说话人的语音信号，由于和目标语音有相似的时频特征，更加难以抑制。近年来，基于数据训练的单通道语音增强算法由于能够利用大量训练数据

的先验信息，对非平稳噪声有较好的抑制效果，为未来的单通道增强算法提供了一个新的解决思路。

2.2.1 信号模型

在单通道语音增强算法中，一般假设噪声是加性的。传声器接收到的信号$x(t)$主要包括纯净语音信号$s(t)$和噪声信号$n(t)$。因此，时域接收信号可以写成如下形式：

$$x(t) = s(t) + n(t) \tag{2.1}$$

语音增强算法通常在频域进行处理，下面给出上述时域信号模型的频域表示。带噪信号$x(t)$经过分帧加窗和短时傅里叶变换（short-time Fourier transform, STFT）后得到：

$$X(k,l) = \sum_{m=0}^{M-1} x(lR+m)h(m)e^{-\frac{j2\pi km}{M}} \tag{2.2}$$

其中，$h(m)$为时间窗函数，$\sum_{m=0}^{M-1} h(m) = 1$，$M$为帧长，$R$是帧移，$k$为频率子带序列，$l$为帧序号。

对公式(2.1)做STFT后可得传声器接收信号在STFT域的表示如下：

$$X(k,l) = S(k,l) + N(k,l) \tag{2.3}$$

其中，$X(k,l)$，$S(k,l)$和$N(k,l)$分别为$x(t)$，$s(t)$和$n(t)$的短时傅里叶域表示。

单通道语音增强算法的关键在于从混合信号$X(k,l)$中估计得到$S(k,l)$。按照不同的增强原理，我们主要介绍以下几种单通道语音增强算法：谱减法、维纳滤波算法、基于统计模型的语音增强算法和基于数据训练的语音增强算法。

2.2.2 谱减法

谱减法主要是基于直观和经验的考虑，假设噪声是加性的，首先估计得到噪声谱，然后直接从带噪语音谱中减去噪声谱，即可得到增强后的语音谱 [1,2]。重写公式(2.3)，省略帧序号，得到：

$$X(k) = S(k) + N(k) \tag{2.4}$$

然后将式(2.4)写成极坐标形式：

$$|X(k)|e^{j\phi_x(k)} = |S(k)|e^{j\phi_s(k)} + |N(k)|e^{j\phi_n(k)} \tag{2.5}$$

其中，$|X(k)|$，$|S(k)|$和$|N(k)|$分别是$X(k)$，$S(k)$和$N(k)$的幅度谱，$\phi_x(k)$，$\phi_s(k)$和$\phi_n(k)$ 是其对应的相位信息。在进行语音增强时，噪声幅度谱$|N(k)|$可以通过噪声估计算法得到，噪声相位$\phi_n(k)$ 一般用带噪信号相位$\phi_x(k)$ 代替。通过替代后，可以得到估计后的纯净语音谱：

$$\hat{S}(k) = [|X(k)| - |\hat{N}(k)|]e^{j\phi_x(k)} \tag{2.6}$$

其中$|\hat{N}(k)|$是估计得到的噪声幅度谱，“$\hat{}$”表示该参数是估计值。增强后的语音时域信号可以通过对$\hat{S}(k)$进行逆短时傅里叶变化（inverse Short-time Fourier Transform, iSTFT）得到。

然而，在实际情况下由于噪声估计的不准确，导致增强后的语音信号幅度谱$|\hat{S}(k)| = (|X(k)| - |\hat{N}(k)|)$ 可能为负值。因此，为了保证谱减后得到的幅度谱为非负值，可以通过半波整流来加以限制：

$$|\hat{S}(k)| = \begin{cases} |X(k)| - |\hat{N}(k)| & \text{if } |X(k)| > |\hat{N}(k)| \\ 0 & \text{else} \end{cases} \tag{2.7}$$

谱减法更一般的形式可以通过下面公式来描述：

$$|\hat{S}(k)|^p = |X(k)|^p - |\hat{N}(k)|^p \tag{2.8}$$

其中p是幂指数，当$p = 1$时，得到最初的基于幅度的谱减法；当$p = 2$时，得到基于功率的谱减法。

由于半波整流的存在，导致增强后的语音幅度谱会存在很多孤立的峰，这样就会产生比较严重的“音乐噪声” [2]。在某些情况下，音乐噪声相比干扰噪声更加令人难以容忍。很多改进型的谱减法被提出来解决音乐噪声问题。

Berouti等人提出了采用过减因子的谱减法 [2]，通过减去过估计的噪声幅度谱，同时限制谱减后的语音幅度谱不小于一个下限值，具体做法如下：

$$|\hat{S}(k)|^2 = \begin{cases} |X(k)|^2 - \alpha|\hat{N}(k)|^2 & \text{if } |X(k)|^2 > (\alpha+\beta)|\hat{N}(k)|^2 \\ \beta|\hat{N}(k)|^2 & \text{else} \end{cases} \tag{2.9}$$

其中，α是过减因子（$\alpha \geq 1$），β是谱平滑参数（$0 < \beta \ll 1$）。α可以通过以下公式逐帧进行调整：

$$\alpha = \alpha_0 - \frac{3}{20}\text{SNR}, \quad -5\text{dB} \leq \text{SNR} \leq 20\text{dB} \tag{2.10}$$

其中，α_0是α在0dB信噪比时的取值，SNR是每一帧估计得到的信噪比。

通过减去过估计的噪声谱，能够削弱谱减后的残余噪声，而且通过对小于门限的幅度谱进行平滑处理，填补孤立谱峰之间的“凹处”，可以有效降低音乐噪声。同时，将过减因子α和信噪比SNR结合起来，当信噪比较高时，减去较少的噪声，反之减去较多的噪声，自适应地进行噪声抑制。

2.2.3 维纳滤波算法

维纳滤波是一种基于MMSE准则的线性滤波器，以最小化输出和参考信号之间的均方误差为目标求解滤波器系数 [9,11]，其目标函数为：

$$\boldsymbol{h} = \arg \ \min \ E\{|d(n) - \hat{d}(n)|^2\} \tag{2.11}$$

其中，$d(n)$是期望信号，即目标语音信号，$\hat{d}(n)$是估计的语音信号。假设滤波器是有限冲击响应（finite impulse response, FIR），$\hat{d}(n)$可表示为：

$$\begin{aligned} \hat{d}(n) &= \boldsymbol{h}^H \boldsymbol{x} \\ &= \sum_{k=0}^{M-1} h_k x(n-k) \ \ n = 0, 1, 2, ... \end{aligned} \tag{2.12}$$

其中，$\boldsymbol{h}^T = [h_0, h_1, h_2, ..., h_{M-1}]$为所求的滤波器系数，$\boldsymbol{x}^T = [x(n), x(n-1), x(n-2), ..., x(n-M+1)]$为包含过去$M$个采样点的观测信号向量，在这里指带噪信号。

上述最优化问题可以通过对式(2.11) 求梯度并令其等于零，得到：

$$-2\boldsymbol{r}_{xd} + 2\boldsymbol{R}_{xx}\boldsymbol{h} = 0 \tag{2.13}$$

其中，$\boldsymbol{r}_{xd}$为观测信号和期望信号的互相关，$\boldsymbol{R}_{xx}$为观测信号的自相关。将式(2.13) 展开得：

$$\sum_{k=0}^{M-1} h_k r_{xx}(m-k) = r_{xd}(-m), \ \ m = 0, 1, ..., M-1 \tag{2.14}$$

因此可以通过解M个方程组来得到滤波器系数$\{h_k\}$。

通过求解式(2.13)，得到维纳滤波器的解，也称作维纳霍夫解 [10]：

$$\boldsymbol{h} = \boldsymbol{R}_{xx}^{-1}\boldsymbol{r}_{xd} \tag{2.15}$$

下面推导维纳滤波算法的频域表示，假设滤波器为无限冲激响应（infinite impulse response, IIR），则估计得到的语音信号为：

$$\hat{d}(n) = \sum_{k=-\infty}^{\infty} h_k x(n-k), \ -\infty < n < \infty \tag{2.16}$$

将式(2.16)写成卷积的形式：

$$\hat{d}(n) = h(n) * x(n) \tag{2.17}$$

其中$*$表示卷积符号。将式(2.17)转换成频域：

$$\hat{D}(k) = H(k)X(k) \tag{2.18}$$

其中，$H(k)$和$X(k)$是滤波器系数$h(n)$和观测数据$x(n)$的离散傅里叶变换系数。

维纳滤波算法在频域的目标函数为：

$$H(k) = \arg \ \min \ E\{|D(k) - \hat{D}(k)|^2\} \tag{2.19}$$

其中，$D(k)$是期望信号$d(n)$的离散傅里叶变换系数。将式(2.18)代入目标函数求梯度并令其等于零，得到：

$$H(k)P_{xx}(k) - P_{dx}(k) = 0 \tag{2.20}$$

其中，$P_{xx}(k) = E|X(k)|^2$是观测信号的功率谱，$P_{dx}(k) = E[D(k)X^*(k)]$是观测信号和期望信号的互功率谱。通过求解式(2.20)，得到维纳滤波器在频域的表示：

$$H(k) = \frac{P_{dx}(k)}{P_{xx}(k)} \tag{2.21}$$

在式(2.21)中，需要计算$P_{dx}(k)$和$P_{xx}(k)$。在降噪中，$D(k) = S(k)$，$X(k) = S(k) + N(k)$，并且假设语音和噪声信号是相互独立的，则：

$$\begin{aligned} P_{dx}(k) &= E[S(k)\{S(k) + N(k)\}^*] \\ &= E[S(k)S^*(k)] + E[S(k)N^*(k)] \\ &= P_{ss}(k) \end{aligned} \tag{2.22}$$

类似的，

$$\begin{aligned} P_{xx}(k) &= E[\{S(k)+N(k)\}\{S(k)+N(k)\}^*] \\ &= E[S(k)S^*(k)]+E[S(k)N^*(k)]+E[N(k)S^*(k)]+E[N(k)N^*(k)] \\ &= P_{ss}(k)+P_{nn}(k) \end{aligned} \tag{2.23}$$

最后，将式(2.22)和(2.23)代入式(2.21)，得到：

$$H(k)=\frac{P_{ss}(k)}{P_{ss}(k)+P_{nn}(k)} \tag{2.24}$$

定义在频率k的先验信噪比ξ_k为：

$$\xi_k=\frac{P_{ss}(k)}{P_{nn}(k)} \tag{2.25}$$

式(2.24)可进一步改写为：

$$H(k)=\frac{\xi_k}{\xi_k+1} \tag{2.26}$$

维纳滤波算法最重要的部分就是噪声信号和先验信噪比的估计，很多算法被提出用于估计噪声和先验信噪比 [22,25,27,47,50,53]。

2.2.4 基于统计模型的算法

2.2.4.1 MMSE-STSA估计子

MMSE-STSA是基于MMSE准则的非线性滤波器，从观测信号中估计得到目标语音信号的幅度谱 [25]。滤波器系数主要通过最小化估计值和真实幅度谱之间的均方误差而得到，其目标函数为：

$$e=E\{(\hat{S}_k-S_k)^2\} \tag{2.27}$$

其中，$\hat{S}_k$是滤波器估计得到的语音幅度谱，$S_k=|S(k)|$是纯净语音的真实幅度谱。

通过贝叶斯理论，最小化式(2.27)得到MMSE估计器：

$$\begin{aligned} \hat{S}_k &= E[S_k|X(k)] \\ &= \int_0^\infty s_k p(s_k|X(k))ds_k \\ &= \frac{\int_0^\infty s_k p(X(k)|s_k)p(s_k)ds_k}{\int_0^\infty p(X(k)|s_k)p(s_k)ds_k} \end{aligned} \tag{2.28}$$

假设语音和噪声傅里叶变换系数服从高斯分布，然后通过推导式(2.28)，得到MMSE幅度估计子：

$$\hat{S}_k = \sqrt{\lambda_k}\Gamma(1.5)\Phi(-0.5, 1; -\nu_k) \tag{2.29}$$

其中，$\Gamma(\cdot)$为伽马函数，$\Phi(a, b : c)$为合流超线几何函数，λ_k定义为：

$$\begin{aligned} \lambda_k &= \frac{\lambda_s(k)\lambda_n(k)}{\lambda_s(k) + \lambda_n(k)} \\ &= \frac{\lambda_s(k)}{1 + \xi_k} \end{aligned} \tag{2.30}$$

其中，$\lambda_s(k) = E\{|S(k)|^2\}$为纯净语音信号第$k$个频率成分的方差，$\lambda_n(k) = E\{|N(k)|^2\}$为噪声信号第$k$个频率成分的方差。

ν_k的定义为：

$$\nu_k = \frac{\xi_k}{1 + \xi_k}\gamma_k \tag{2.31}$$

其中，ξ_k和γ_k是先验和后验信噪比，定义为：

$$\xi_k = \frac{\lambda_s(k)}{\lambda_n(k)} \tag{2.32}$$

$$\gamma_k = \frac{X_k^2}{\lambda_n(k)} \tag{2.33}$$

其中，$X_k = |X(k)|$是输入观测信号的幅度谱。

将式(2.31)-(2.33)代入式(2.30)得到：

$$\sqrt{\lambda_k} = \frac{\sqrt{\nu_k}}{\gamma_k}X_k \tag{2.34}$$

最终的MMSE幅度估计子为：

$$\hat{S}_k = \frac{\sqrt{\pi}}{2}\frac{\sqrt{\nu_k}}{\gamma_k}\exp(-\frac{\nu_k}{2})[(1 + \nu_k)I_0(\frac{\nu_k}{2}) + \nu_k I_1(\frac{\nu_k}{2})]X_k \tag{2.35}$$

其中，$I_0(\cdot)$和$I_1(\cdot)$分别表示零阶和一阶贝塞尔函数。在实际计算时，MMSE-STSA算法需要估计先验信噪比ξ_k和后验信噪比γ_k。

2.2.4.2 MMSE-LSA估计子

MMSE-LSA估计器也是基于MMSE准则的非线性滤波器，是通过最小化对数幅度谱的均方误差而得到 [26]，其目标函数为：

$$e = E\{(\log\hat{S}_k - \log S_k)^2\} \tag{2.36}$$

其中，$\log\hat{S}_k$是滤波器估计得到的语音对数幅度谱，$\log S_k$是纯净语音的真实对数幅度谱。

最小化式(2.36)得到最优log-MMSE估计器：

$$\log\hat{S}_k = E\{\log S_k|X(k)\} \tag{2.37}$$

$$\hat{S}_k = \exp(E\{\log S_k|X(k)\}) \tag{2.38}$$

假设语音和噪声傅里叶变换系数服从高斯分布，根据贝叶斯理论推导式(2.38)，得到MMSE对数谱幅度估计子：

$$E\{\log S_k|X(k)\} = \frac{1}{2}\log\lambda_k + \frac{1}{2}\log\nu_k + \frac{1}{2}\int_{\nu_k}^{\infty}\frac{e^{-t}}{t}dt \tag{2.39}$$

$$\hat{S}_k = \frac{\xi_k}{\xi_k+1}\exp\{\frac{1}{2}\int_{\nu_k}^{\infty}\frac{e^{-t}}{t}dt\}X_k \tag{2.40}$$

其中，ν_k如式(2.31)所示，ξ_k是先验信噪比，如式(2.32)所示。

研究表明，log-MMSE估计子相比线性MMSE估计子在保证语音失真一致的前提下，能获得更多的噪声抑制量 [26]。

2.2.5 基于数据训练的算法

2.2.5.1 基于非负矩阵分解的语音增强算法

非负矩阵分解将一个非负矩阵（如音频信号的幅度谱）$\boldsymbol{X} \in \mathbb{R}_+^{M\times N}$分解为两个非负矩阵$\boldsymbol{W} \in \mathbb{R}_+^{M\times K}$和$\boldsymbol{H} \in \mathbb{R}_+^{K\times N}$的乘积 [55]：

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H} \tag{2.41}$$

其中$\boldsymbol{W}$为字典矩阵，$\boldsymbol{H}$为权重矩阵，K为字典矩阵列向量的数目。$\boldsymbol{W}$的列向量是描述$\boldsymbol{X}$频谱结构的基向量，$\boldsymbol{H}$中的值是$\boldsymbol{W}$中基向量对应的激活值。式(2.41)可以通过解决下述目标函数进行分解：

$$\boldsymbol{W}, \boldsymbol{H} = \arg\ \min\ d_{\mathrm{KL}}(\boldsymbol{X}|\boldsymbol{W}\boldsymbol{H}) \tag{2.42}$$

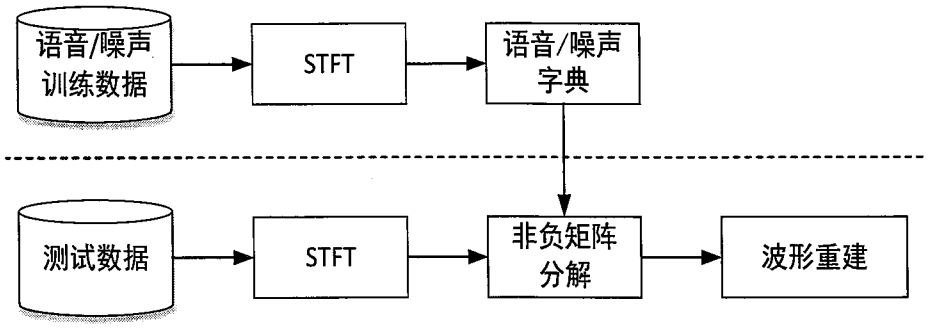

图 2.1: 基于NMF的单通道语音增强算法系统框图。虚线上方为训练阶段，虚线下方为增强阶段

其中，d_{KL}为广义KL散度 [58]，定义如下：

$$d_{\mathrm{KL}}(\boldsymbol{A}|\boldsymbol{B}) = \sum_{m,n} A_{m,n}\log\frac{A_{m,n}}{B_{m,n}} - A_{m,n} + B_{m,n} \tag{2.43}$$

式(2.42)可以通过以下乘法迭代准则 [56]得到$\boldsymbol{W}$和$\boldsymbol{H}$：

$$\begin{aligned}\boldsymbol{W} &\leftarrow \boldsymbol{W} \otimes \frac{\frac{\boldsymbol{X}}{\boldsymbol{W}\boldsymbol{H}}\boldsymbol{H}^T}{\mathbf{1}\boldsymbol{H}^T} \\ \boldsymbol{H} &\leftarrow \boldsymbol{H} \otimes \frac{\boldsymbol{W}^T\frac{\boldsymbol{X}}{\boldsymbol{W}\boldsymbol{H}}}{\boldsymbol{W}^T\mathbf{1}}\end{aligned} \tag{2.44}$$

其中，$\mathbf{1}$是值全为1的矩阵，维度和$\boldsymbol{X}$相同。$\otimes$和$\frac{a}{b}$表示矩阵的逐点相乘和相除运算。首先将$\boldsymbol{W}$和$\boldsymbol{H}$初始化为非负值，然后通过式(2.44)反复迭代更新$\boldsymbol{W}$和$\boldsymbol{H}$即可得到最终分解的$\boldsymbol{W}$和$\boldsymbol{H}$。

基于非负矩阵分解的单通道语音增强算法框图如图 2.1 所示。在训练阶段，首先收集特定说话人和特定噪声类型的训练数据，通过NMF得到语音和噪声字典矩阵；在增强阶段，对测试信号幅度谱进行非负矩阵分解，固定字典矩阵，估计得到权重矩阵，然后将语音字典和其对应的权重矩阵相乘重构出增强后的语音信号。算法主要由以下几个步骤组成 [57]：

（1）计算训练数据中语音和噪声信号的幅度谱$\boldsymbol{X}_S$和$\boldsymbol{X}_N$，以及测试信号的幅度谱$\boldsymbol{X}$。在这里，我们将很多帧的信号幅度谱结合起来写成矩阵的形式，矩阵中的每一列为一帧信号幅度谱。

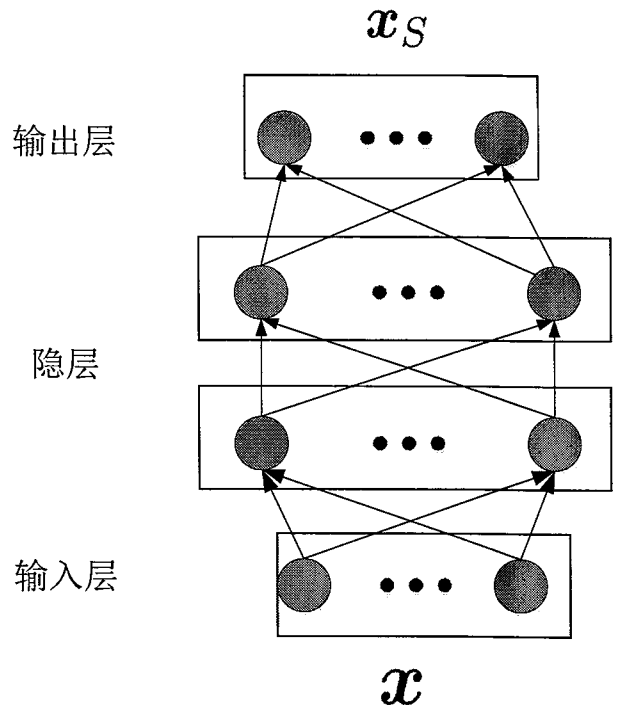

图 2.2: DNN模型结构图

（2）分别对$\boldsymbol{X}_S$和$\boldsymbol{X}_N$通过式(2.41)进行非负矩阵分解，得到对应的字典矩阵$\boldsymbol{W}_S$和$\boldsymbol{W}_N$，并令$\boldsymbol{W}=[\boldsymbol{W}_S \ \ \boldsymbol{W}_N]$。

（3）对$\boldsymbol{X}$进行非负矩阵分解，即$\boldsymbol{X}\approx\boldsymbol{W}\boldsymbol{H}$，固定$\boldsymbol{W}$，通过式(2.44)第二个公式迭代得到权重矩阵$\boldsymbol{H}$。

（4）将权重矩阵$\boldsymbol{H}$拆分为$\boldsymbol{H}=[\boldsymbol{H}_S^T \ \ \boldsymbol{H}_N^T]^T$，$\boldsymbol{H}_S$为语音字典对应的权重矩阵，$\boldsymbol{H}_N$为噪声字典对应的权重矩阵，然后重构分离后的语音幅度谱$\hat{\boldsymbol{X}}_S=\boldsymbol{W}_S\boldsymbol{H}_S$和噪声幅度谱$\hat{\boldsymbol{X}}_N=\boldsymbol{W}_N\boldsymbol{H}_N$，最后通过维纳滤波的形式得到最终增强后的语音幅度谱：

$$\tilde{\boldsymbol{X}}_S=\boldsymbol{X}\otimes\frac{\hat{\boldsymbol{X}}_S}{\hat{\boldsymbol{X}}_S+\hat{\boldsymbol{X}}_N} \tag{2.45}$$

（5）将$\tilde{\boldsymbol{X}}_S$和测试信号的相位结合通过iSTFT以及重叠相加恢复出增强后的语音时域信号。

上述算法由于同时需要语音和噪声的训练数据，因此被称作全监督增强算法。如果只需要语音或者噪声的训练数据，则称作半监督语音增强算法。在半

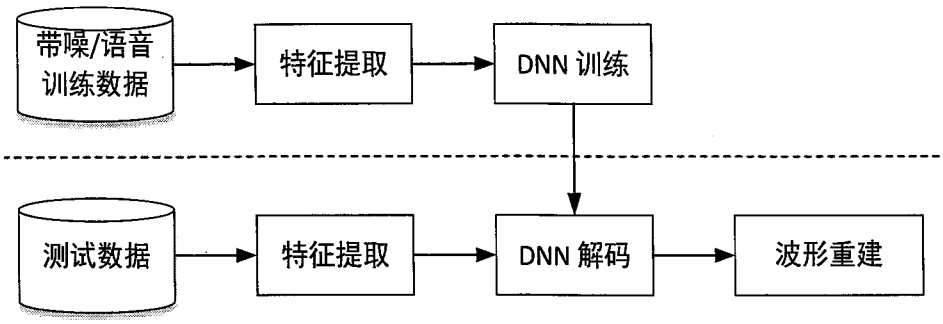

图 2.3: 基于DNN的语音增强算法框图。虚线上方为训练阶段，下方为增强阶段

监督增强算法中，如果无法事先获取噪声的训练数据，则在步骤（3）中通过测试信号估计得到噪声字典$\boldsymbol{W}_N$。

2.2.5.2　基于深度神经网络的语音增强算法

在基于深度神经网络的语音增强算法中，语音增强被当作是一个有监督的学习任务，通过大量的训练数据训练一个判别式模型，如DNN模型来估计目标信号 [84]。

图 2.2 为一种常见的用于语音增强的DNN模型结构图，共包含一个输入层，一个输出层和两个隐层（隐层数目可以根据训练数据量设定），层与层之间采用全连接结构。$\boldsymbol{x}$为模型的输入特征如混合信号的幅度谱等，中间两层隐层用于学习输入特征与输出之间的非线性关系。$\boldsymbol{x}_S$为$\boldsymbol{x}$对应的模型输出，根据训练目标的不同主要分为两类：第一类输出目标为掩蔽值，如IBM，IRM等 [85–89]；第二类目标为纯净语音频谱 [98, 99]。在DNN模型中，第l层隐层的输出由以下公式得到：

$$h^l(\boldsymbol{x}) = f(\boldsymbol{U}^l h^{l-1}(\boldsymbol{x}) + \boldsymbol{b}^l) \tag{2.46}$$

其中，$\boldsymbol{U}^l$是DNN第$l-1$层和第l层之间的权重矩阵，维度和这两层的节点数相关，$\boldsymbol{h}^{l-1}(\boldsymbol{x})$是第$l-1$层的输出，$\boldsymbol{b}^l$是第$l$层的偏置向量，第0层为输入层，$h^0(\boldsymbol{x}) = \boldsymbol{x}$。函数$f$为非线性函数，一般采用ReLU（Rectified Linear Units）激活函数$f(x) = \max(0, x)$ [126]或者Sigmoid函数$f(x) = 1/(1 + e^{-x})$。输出层

函数要根据训练目标而定，如果训练目标为掩蔽值，则输出层函数一般采用sigmoid函数；如果训练目标为目标语音谱，则输出层函数一般采用线性函数。

图 2.3 是基于DNN的语音增强算法框图。在训练阶段，首先确定DNN模型的参数，如层数、每层节点数、隐层非线性函数、训练目标和输出层函数，然后从混合训练信号中提取特征作为DNN的输入，并且计算参考输出目标，如理想条件下的掩蔽值或参考语音谱。确定输入输出之后采用后向传播算法估计DNN模型参数；在增强阶段，首先提取测试信号的特征，然后送入已训练好的DNN模型中估计目标值。如果估计目标为掩蔽值，则将掩蔽值结合混合信号频谱得到增强后的语音谱；如果估计目标为幅度谱，则直接将输出作为增强后的语音幅度谱。最后结合测试信号相位信息恢复出增强后的语音时域波形。

第三章　基于语音连续性建模的单通道语音增强算法研究

3.1　引言

在单通道语音增强算法中，由于没有空间信息可以利用，只能从一路混合信号中得到增强后的语音信号，因此需要引入一些先验信息进行处理。传统的单通道语音增强算法引入了噪声相比语音更加平稳的先验假设信息，首先采用诸如跟踪平滑的方法估计噪声信号，然后再进行语音增强。然而这种先验假设和很多实际情况并不相符，如很多噪声比语音信号更具有非平稳特性，或者很多干扰噪声本身也是语音信号，导致无法准确估计噪声，从而降噪效果不佳。因此，为了能够对非平稳噪声有较好的抑制效果，需要引入更加准确的先验信息。近年来，基于数据训练的方法被广泛提出来用于语音增强，从语音和噪声的训练数据中学习得到一些先验信息，然后利用这些先验信息进行语音增强。例如，通过语音和噪声的训练数据分别对语音和噪声进行建模，然后利用建立的模型区分测试信号中的语音和噪声，从而达到语音增强的目的。在这些算法中，基于NMF的单通道语音增强算法由于效果显著而受到了广泛关注和研究。

在基于NMF的增强算法中，首先通过语音和噪声训练数据分别得到语音和噪声的字典矩阵，这些字典矩阵分别表征了语音和噪声信号的频谱结构信息，然后在增强阶段将测试信号幅度谱投影到语音和噪声字典列向量张成的子空间里，从而进行语音和噪声的分离。在进行增强时，为了能够有效区分语音和噪声信号，需要语音和噪声字典具有较强的鉴别性，即一个音源的字典不能用来描述另一个音源的信号；同时，为了保证一个音源的字典能够描述该音源训练数据之外的信号，需要保证字典具有一定的泛化能力。因此，字典的选取对于增强算法的性能有至关重要的影响。

由于语音固有的发声机理，语音信号具有很强的时间连续性，相邻时刻的语音信号具有相似的频谱特性。然而，标准的基于NMF的增强算法假设语音相邻帧是相互独立的，并没有考虑到语音信号的时间连续性。因此，很多基于NMF的改进算法被提出来对语音信号的时间连续性建模，提高了语音增强的效果。很多算法通过对权重矩阵加入正则化约束，防止权重矩阵相邻两帧的值变化过大来对语音时间连续性建模 [58, 73–75]；还有一些算法通过卡尔曼滤波

来描述语音信号的时间连续性 [76–78]；Mysore等人将NMF和隐马尔科夫模型（hidden Markov model, HMM）结合来描述语音信号的频谱动态变化特性 [79]，又将NMF和因子隐马尔科夫模型（factorial hidden Markov model, FHMM）结合来描述混合信号的时间动态特性，用于语音增强 [80]。

本章我们继续对基于NMF的增强算法进行研究，首先结合NMF和k均值聚类算法提出了一种新的语音信号建模方法，能够同时描述语音信号的频谱结构信息和时间连续性信息。其次，我们将该语音建模方法和因子条件随机场（factorial conditional random field, FCRF）结合对混合信号的时间动态特性建模，然后用于分离两个说话人的语音信号以及分离语音和噪声信号。

3.2 基于NMF和k均值聚类的语音信号建模

标准的基于NMF的语音增强算法对语音信号采用一个字典建模，使用一个语音字典来描述语音信号所有帧的谱结构信息。为了能够涵盖所有语音帧的信息，一般采用较大的语音字典，然而这样会导致语音字典有可能用来描述噪声信号，从而不能够有效区分语音和噪声信号。在本节中，我们提出了一种语音建模方法同时对语音信号的频谱结构信息和时间连续性信息建模。具体而言，针对不同的语音帧采用不同的语音小字典进行建模，这样就增强了语音字典的建模能力，同时为了描述语音信号的时间连续性，在不同的语音小字典之间加入了状态跳转，使得相邻帧的语音信号尽可能采用同一个语音小字典建模。

语音建模方法的具体实现框图如图 3.1 所示。具体而言，首先对纯净语音训练数据进行STFT得到语音频谱，并取绝对值得到幅度谱$\boldsymbol{X}$。然后通过k均值聚类算法 [65] 得到一个较完备的语音字典$\boldsymbol{W}$，聚类的目标函数为KL散度 [58]。在聚类中，$\boldsymbol{X}$被聚成K个类别，聚类完成之后的每一个聚类中心为语音字典$\boldsymbol{W}$中的列向量，因此在$\boldsymbol{W}$中含有K个列向量。随后，再对幅度谱$\boldsymbol{X}$进行一次k均值聚类，聚类类别的数目M是语音信号状态的数目。第二次聚类的目的是为了得到$\boldsymbol{X}$中频谱结构类似的语音帧以及得到语音相邻帧之间的状态序列。

在第二次聚类完成之后，$\boldsymbol{X}$中的每一帧语音都有一个类别标记来表示该帧语音信号属于哪一个聚类类别，那么所有的语音帧就对应了一个标记序列，这个标记序列就是从训练数据得到的状态序列，用来描述语音信号的时间连续性。图 3.2 给出了语音相邻帧对应的状态转移概率的示意图，从图中可以看出

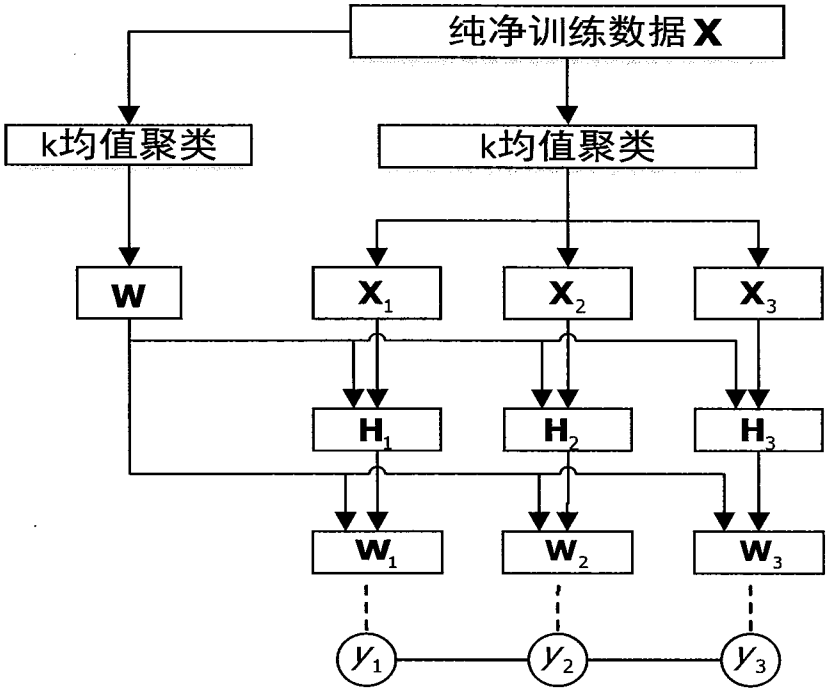

图 3.1: 本节提出的语音建模的算法框图。为了简化，这里我们假设只有三个状态

语音相邻帧属于同一类别的概率要远大于属于不同类别的概率，说明了语音信号具有很强的时间连续性。然后将$\boldsymbol{X}$中属于同一类别的语音帧集合在一起，对其进行非负矩阵分解得到每一个状态（类别）对应的语音字典。具体而言，令$\boldsymbol{X}_j$表示属于第j类的语音帧组成的幅度谱矩阵，然后对$\boldsymbol{X}_j$进行如下非负矩阵分解：

$$\boldsymbol{X}_j = \boldsymbol{W}\boldsymbol{H}_j, \quad j = 1, ..., M \tag{3.1}$$

其中，M为状态数目，$\boldsymbol{H}_j$为分解后得到的权重矩阵，$\boldsymbol{H}_j$中的值表示$\boldsymbol{W}$中对应列向量对于$\boldsymbol{X}_j$的重要程度。在得到$\boldsymbol{H}_j$之后，通过$\boldsymbol{H}_j$中的值来从$\boldsymbol{W}$中挑选一些列向量组成状态j对应的字典$\boldsymbol{W}_j$，具体按照如下步骤得到：将$\boldsymbol{W}$中每个列向量$\boldsymbol{w}_n$在$\boldsymbol{H}_j$中对应的值进行相加：

$$a_j^n = \sum_{t=1}^{T_j} H_j^{(nt)} \tag{3.2}$$

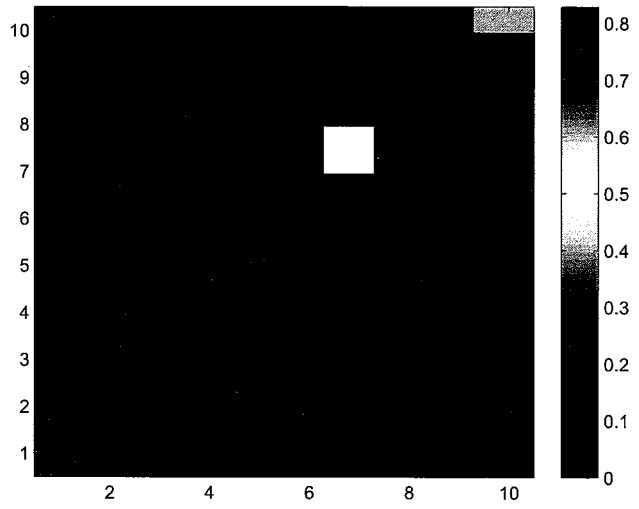

图 3.2: 相邻帧对应的状态转移概率示意图，状态类别数为10

其中，T_j是$\boldsymbol{X}_j$中的语音帧数，$H_j^{(nt)}$是$\boldsymbol{H}_j$的第n行第t列个元素。如果

$$a_j^{(n)} \geq \beta \times \max_n\{a_j^{(n)}\} \tag{3.3}$$

说明$\boldsymbol{w}_n$对于描述$\boldsymbol{X}_j$比较重要，因此将$\boldsymbol{w}_n$从$\boldsymbol{W}$挑选出来组成第j个类别对应的语音字典$\boldsymbol{W}_j$，这样就可以从$\boldsymbol{W}$中挑选合适的列向量组成$\boldsymbol{W}_j$。其中，β为稀疏参数，用于控制$\boldsymbol{W}_j$中的列向量数目。根据上述方法，就可以通过语音训练数据得到对应的状态序列和每个状态对应的语音字典。

3.3 基于语音连续性建模的语音分离算法

在本节中，我们将3.2节提出的语音建模方法和FCRF模型结合用于分离两个说话人的语音信号。算法的系统框图 3.3 所示，主要包括两个部分。在训练阶段，首先对每个说话人对应的训练数据采用3.2节提出的语音建模方法得到一个状态序列和若干语音字典同时对语音信号的时间连续性和频谱结构建模。其中，每个字典对应状态序列中的一个状态。然后将两个说话人的训练语音进行混合，得到混合训练语音。对混合训练语音进行特征提取，并结合两个说话人的状态序列，训练FCRF模型用于描述两个说话人混合语音的时间动态特性。在分离阶段，对测试信号进行特征提取，然后将提取的特征送入已训练好

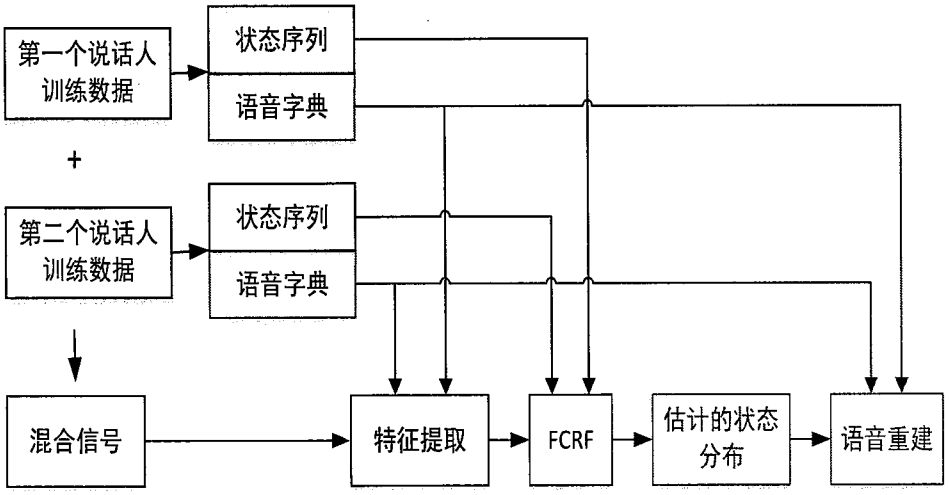

图 3.3: 本节提出的单通道语音分离算法系统框图。红色部分为训练阶段，黑色部分为训练阶段和分离阶段共用，绿色部分为分离阶段

的FCRF模型中进行解码，得到每帧混合语音对应两个说话人状态标记的联合后验概率，最后再结合训练阶段得到的语音字典重构每个说话人的语音信号。

3.3.1　语音信号建模

在算法的训练阶段，首先针对每个说话人的训练数据采用3.2节提出的语音建模算法得到一个状态序列和若干语音字典，同时对语音信号的时间连续性和频谱结构建模。具体的算法描述见3.2节。

3.3.2　特征提取

FCRF相比FHMM去除了相邻状态对应的观测之间条件独立的限制，而且FCRF模型能够容纳多种类型的观测数据，因此能够更好地对语音信号的时间动态特性建模 [127]。在训练FCRF模型时，我们使用了两种类型的观测数据：第一种是混合语音信号的幅度谱；第二种观测数据是对混合信号幅度谱进行非负矩阵分解之后得到的权重矩阵。在分解中，每个说话人的字典是该说话人在训练阶段得到的所有小字典的集合。

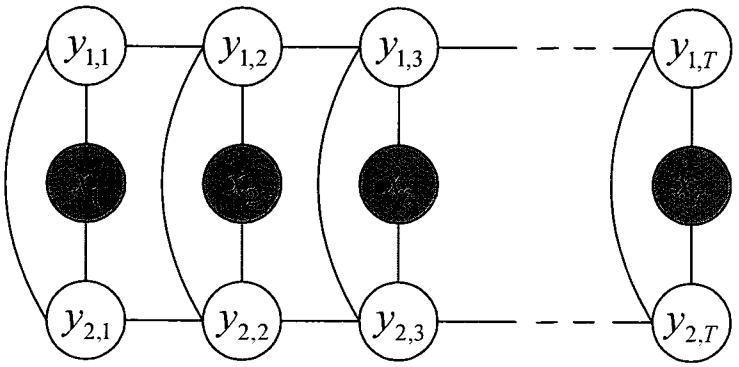

图 3.4: FCRF的图模型

3.3.3 FCRF

FCRF模型首先在 [128]中被提出，并且有研究结果表明FCRF在自然语言处理中相比CRF模型具有更优的性能表现。FCRF的图模型结构如图3.4 所示。其中，$\boldsymbol{x} = [x_1, x_2, \cdots, x_T]$是观测数据，$\boldsymbol{y}_1 = [y_{1,1}, y_{1,2}, \cdots, y_{1,T}]$和$\boldsymbol{y}_2 = [y_{2,1}, y_{2,2}, \cdots, y_{2,T}]$ 是两条线性标记链。每个观测x_k对应两个状态标记$y_{1,k}$和$y_{2,k}$。在给定观测$\boldsymbol{x}$后，FCRF模型在对应状态序列为$\{\boldsymbol{y}_1, \boldsymbol{y}_2\}$时的条件概率为：

$$p(\{\boldsymbol{y}_1, \boldsymbol{y}_2\}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \left(\prod_{t=1}^{T-1} \prod_{l=1}^{2} \Phi_l(y_{l,t}, y_{l,t+1}, \boldsymbol{x}, t) \right) \left(\prod_{t=1}^{T} \prod_{l=1}^{1} \Psi_l(y_{l,t}, y_{l+1,t}, \boldsymbol{x}, t) \right) \tag{3.4}$$

其中，l是线性链的标记序号，$Z(\boldsymbol{x})$是配分函数，用于归一化概率分布。$\{\Phi_l\}$是每条线性链内的势能函数，$\{\Psi_l\}$是两条线性链之间的势能函数。势能函数可以进行以下分解：

$$\Phi_l(y_{l,t}, y_{l,t+1}, \boldsymbol{x}, t) = \exp\left\{ \sum_k \lambda_k f_k(y_{l,t}, y_{l,t+1}, \boldsymbol{x}, t) \right\} \tag{3.5}$$

$$\Psi_l(y_{l,t}, y_{l+1,t}, \boldsymbol{x}, t) = \exp\left\{ \sum_k \lambda_k f_k(y_{l,t}, y_{l+1,t}, \boldsymbol{x}, t) \right\} \tag{3.6}$$

其中，$\{f_k\}$是特征函数，$\{\lambda_k\}$是权重参数。

在训练FCRF模型时，首先提取混合训练语音的特征作为FCRF模型的输入，通过两个说话人训练数据得到的状态序列作为两条线性标记链，然后估

计参数λ_k。参数λ_k通过limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) 算法 [129] 估计得到。在分离阶段，首先提取测试信号的特征作为已训练好的FCRF模型的输入，针对每帧测试信号估计得到两个说话人对应状态标记的联合后验概率$p(y_{1,t}, y_{2,t}|\boldsymbol{x})$。具体的估计算法采用loopy belief propagation (LBP) [128]。后验概率被用来重构每个说话人的语音幅度谱。

3.3.4 语音重建

在通过FCRF模型得到$p(y_{1,t}, y_{2,t}|\boldsymbol{x})$之后，结合在训练阶段得到的每个说话人的语音字典，通过以下方式重构每个说话人的语音幅度谱：

$$\hat{\boldsymbol{x}}_{1t} = \sum_{m=1}^{M}\sum_{n=1}^{M} p(y_{1,t}=m, y_{2,t}=n|\boldsymbol{x})\boldsymbol{W}_{1,m}\boldsymbol{h}_{1mn,t} \tag{3.7}$$

$$\hat{\boldsymbol{x}}_{2t} = \sum_{m=1}^{M}\sum_{n=1}^{M} p(y_{1,t}=m, y_{2,t}=n|\boldsymbol{x})\boldsymbol{W}_{2,n}\boldsymbol{h}_{2mn,t} \tag{3.8}$$

其中，$\hat{\boldsymbol{x}}_{1t}$和$\hat{\boldsymbol{x}}_{2t}$是分离得到的两个说话人对应第$t$帧的语音幅度谱。$p(y_{1,t}=m, y_{2,t}=n|\boldsymbol{x})$是在第$t$帧时第一个说话人对应状态是$m$和第二个说话人对应状态是$n$的联合后验概率，$\boldsymbol{W}_{1,m}$是第一个说话人第$m$个状态对应的语音字典，$\boldsymbol{W}_{2,n}$是第二个说话人第$n$个状态对应的语音字典。$\boldsymbol{h}_{1mn,t}$和$\boldsymbol{h}_{2mn,t}$是将第$t$帧测试信号幅度谱$\boldsymbol{x}_t$进行非负矩阵分解得到的权重向量：

$$\boldsymbol{x}_t = \begin{bmatrix}\boldsymbol{W}_{1,m} & \boldsymbol{W}_{2,n}\end{bmatrix}\begin{bmatrix}\boldsymbol{h}_{1mn,t}\\ \boldsymbol{h}_{2mn,t}\end{bmatrix}, \quad m,n=1,\cdots,M \tag{3.9}$$

式(3.7)-(3.8)可以看做是加权的基于NMF的语音分离算法，每个说话人有多个字典进行建模，两个说话人每对字典组合对应的权重通过FCRF估计得到，用于描述这对语音字典对于该帧测试语音的重要程度。

在得到$\hat{\boldsymbol{x}}_{1t}$和$\hat{\boldsymbol{x}}_{2t}$之后，再通过维纳滤波的形式得到两个说话人最终分离后的语音幅度谱：

$$\tilde{\boldsymbol{x}}_{1t} = \boldsymbol{x}_t \otimes \frac{\hat{\boldsymbol{x}}_{1t}}{\hat{\boldsymbol{x}}_{1t}+\hat{\boldsymbol{x}}_{2t}} \tag{3.10}$$

$$\tilde{\boldsymbol{x}}_{2t} = \boldsymbol{x}_t \otimes \frac{\hat{\boldsymbol{x}}_{2t}}{\hat{\boldsymbol{x}}_{1t}+\hat{\boldsymbol{x}}_{2t}} \tag{3.11}$$

其中，$\otimes$和$\frac{a}{b}$表示向量对应元素的相乘和相除。

最后，将$\tilde{\boldsymbol{x}}_{1t}$和$\tilde{\boldsymbol{x}}_{2t}$分别结合测试信号的相位信息，并通过iSTFT以及重叠相加算法得到分离后两个说话人的时域语音信号。

3.3.5 实验评价和讨论

3.3.5.1 实验设置

在实验部分，我们通过从混合信号中分离两个说话人的语音信号来对所提算法进行评价。具体而言，在实验中我们分别使用了仿真数据和实录数据。对于仿真数据，我们使用了Grid数据集 [130]中的语音数据，该数据集共包含34个说话人的语音数据。本实验中的测试集共包含100条混合语音，每条混合语音由两个说话人的语音信号混合而成，其中每个人的语音信号是从该说话人的语料中随机选择10条语音连接而成。所有测试信号大约长16s，而且按照不同的信噪比进行混合（-6，-3，0，3和6dB），在混合时将第二个说话人的语音信号当做“噪声”，主要是为了描述两个说话人音量不同的情况。对于测试集中每个说话人而言，从其Grid语料库中随机选择100条语句作为其训练数据，每个说话人的测试语音和训练语音不重合。

对于实录数据，我们使用了一个线性两通道传声器阵列实录了20条混合语音数据，然后取出第一个通道的信号作为实录测试信号。在录音时，不同性别的说话人分别站在阵列前方一米处，而且两人音量大致相同。所有的实录信号大约长10s。

所有的语音信号首先通过重采样到16kHz，然后通过STFT得到其频谱，在变换时采用汉宁窗，帧长32ms，帧移16ms。

3.3.5.2 评价指标

为了评价算法的分离效果，使用了三种客观评价指标：

1. source-to-interferences ratio (SIR) [131]。这项指标是目标语音和干扰信号的功率比，用于衡量算法对于干扰的抑制能力。

$$\mathrm{SIR} = 10\log_{10}\frac{||\mathrm{s_{target}}||^2}{||\mathrm{e_{interf}}||^2} \tag{3.12}$$

2. source-to-artifact ratio (SAR) [131]。这项指标是原始信号和算法引入失真的功率比，用于衡量算法自身引入失真的大小。

$$\mathrm{SAR} = 10\log_{10}\frac{||\mathrm{s_{target}} + \mathrm{e_{interf}}||^2}{||\mathrm{e_{artif}}||^2} \tag{3.13}$$

3. source-to-distortion ratio (SDR) [131]。这项指标是目标语音信号和其它非目标信号的功率比，用于综合衡量算法的分离性能。

$$\mathrm{SDR} = 10\log_{10}\frac{||\mathrm{s_{target}}||^2}{||\mathrm{e_{interf}} + \mathrm{e_{artif}}||^2} \tag{3.14}$$

其中，$\mathrm{s_{target}}$表示目标信号，$\mathrm{e_{interf}}$表示干扰信号，$\mathrm{e_{artif}}$表示算法引入的失真部分。对应这三项评价指标而言，分数越大说明分离效果越好。

除了客观评价指标之外，我们也采用了主观测听实验（Mean Opinoin Score test, MOS）。在主观测听实验中，十位听力正常的测试人员（五男五女）对分离后的语音信号进行了主观打分，其中5分表示分离效果最好，而1分表示分离效果最差。将所有人的打分结果进行平均得到最后的MOS结果。

3.3.5.3　比较算法

为了比较，我们考察了以下几种算法的结果：

1. 标准的基于NMF的语音分离算法 [57]。

该算法采用KL散度作为目标函数以及乘法迭代准作为分解算法。每个说话人的语音字典包含50个基向量。

2. 稀疏NMF算法（sparse NMF, SNMF） [65]。

该算法语音字典通过k均值聚类得到，分解算法通过激活集牛顿算法（Active-set Newton Algorithm, ASNA） [65] 得到。稀疏参数γ设置为1。

3. NFHMM算法 [79]。

在NFHMM算法中，每个说话人类别数为20，每个类别对应的字典包含10个基向量。

3.3.5.4　结果分析

首先，我们探讨状态数M和稀疏参数β对于所提算法分离结果的影响。M主要用于描述语音信号频谱的变化程度，M值越大说明语音信号的频谱变化越大，反之亦然。β主要用于控制每个状态对应小字典中基向量的个数，β值越大说明每个小字典中包含的基向量越少，反之亦然。表3.1是当$K = 50$，输入SNR为0dB时，不同M和不同β对应的SDR均值。从结果来看，当$M = 10$和$\beta = 0.05$时能取得最好的分离效果。因此在后续的实验中我们就采用上述参数设置。

表 3.1: 在K为50，输入信噪比为0dB时，不同M和不同β组合下的SDR均值

SDR(dB)		β			
		0.01	0.025	0.05	0.1
M	5	5.32	5.59	5.78	5.50
	10	5.60	5.84	**6.16**	6.08
	15	5.81	5.97	6.02	5.89
	20	5.68	5.82	5.97	5.90

图3.5-图3.8给出了各种算法在仿真测试集上针对不同输入信噪比时的SDR、SIR、SAR和MOS结果。从结果来看，对于所有输入信噪比而言，本节提出的算法相比其它算法能够有效提高SDR和MOS指标，说明所提算法能够得到更好的分离结果和更优的主观听觉体验。具体而言，所提算法在SIR指标上有较大提升而在SAR上有一定提升，说明该算法在不引入更多失真的前提下能够有效抑制干扰信号。而且，与输入信噪比为0dB时的结果相比，所有算法在输入信噪比较高或较低时（6，-6dB）分离结果都有所下降。这主要是因为在信噪比较高或较低时，能量强的语音信号会掩盖能量弱的语音信号，导致更难以分离两个说话人的语音信号。

图3.9给出了各种算法在分离后得到的语音频谱图，左上角图是0dB混合时的混合信号语谱图，右上角图为混合前第一个说话人对应的纯净语音频谱图，中间左图是本节所提算法分离后得到的语谱图，中间右图是SNMF算法分离后得到的语音谱，左下角图是NFHMM算法分离后得到的语谱图，右下角图是标准NMF算法分离后得到的语谱图。从语谱图结果可以看出，本节所提算法能够获得更好的分离结果。具体而言，该算法能够得到更清晰的频谱结构以及更少的频谱重叠（例如在2.2s和4.5s处）。这也是和客观评价指标对应的。

除此之外，我们比较了不同算法在以下两种混合情况下的分离结果：同性说话人混合（same gender, SG）和异性说话人混合（opposite gender, OG）。具体的SDR结果如表3.2所示。从结果来看，在同性说话人混合时，本节所提算法相比其它算法能够获得更大的SDR提升，这主要是因为该算法采用了不同的字典描述不同的语音帧，增强了语音字典的鉴别性，降低了一个说话人字典描述另一个说话人信号的可能性。此外，对于所有算法而言，同性说话人混合时的分离结果相比异性混合时SDR都有所下降，这也是正常的，同性说话人由于

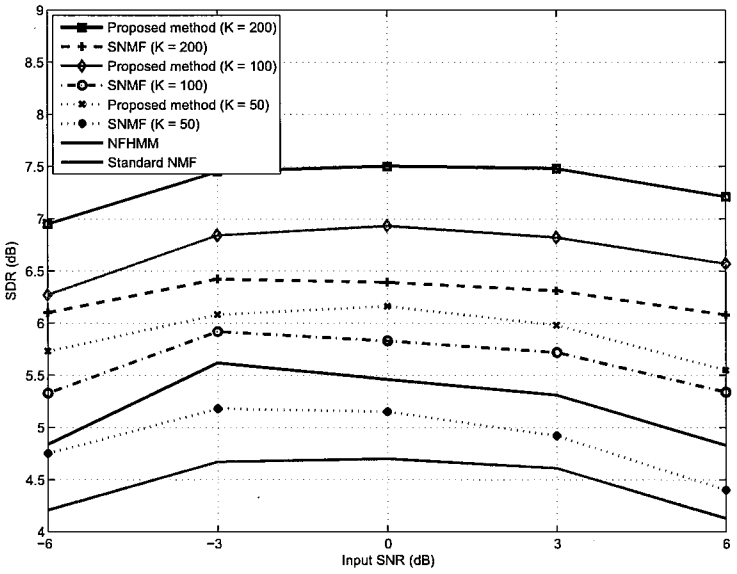

图 3.5: 各种算法在不同输入信噪比时的SDR均值

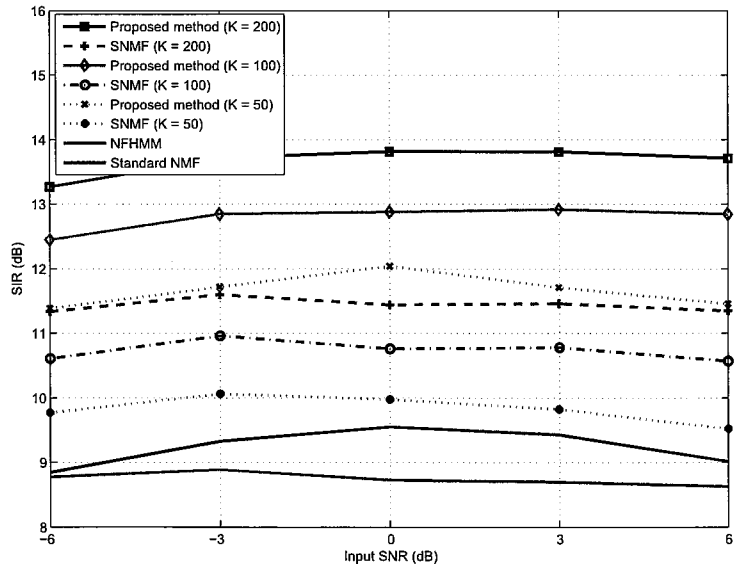

图 3.6: 各种算法在不同输入信噪比时的SIR均值

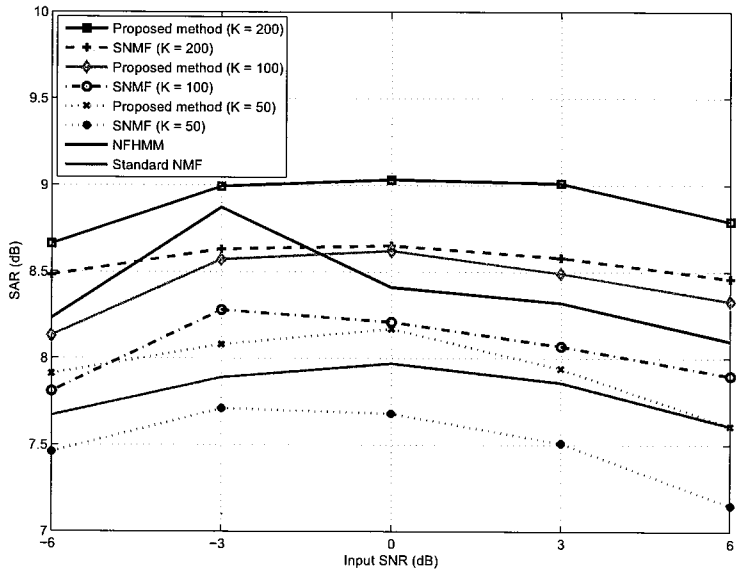

图 3.7: 各种算法在不同输入信噪比时的SAR均值

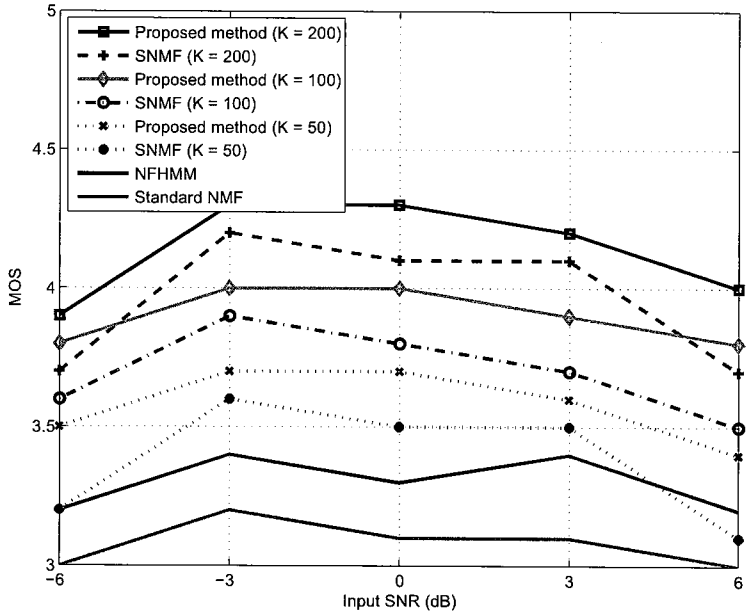

图 3.8: 各种算法在不同输入信噪比时的MOS均值

表 3.2: 各种算法在以下两种混合情况下的SDR结果。SG：同性说话人混合，OG：异性说话人混合

SDR(dB)	SG	OG
Proposed method(K=200)	**5.79**	**9.08**
SNMF(K=200)	4.37	8.25
Proposed method(K=100)	5.11	8.61
SNMF(K=100)	3.76	7.74
Proposed method(K=50)	4.22	7.95
SNMF(K=50)	3.04	7.07
NFHMM	3.67	7.11
Standard NMF	2.94	6.33

语音信号频谱更加相似而更难以分离。

图3.10是不同算法在实录测试集上的MOS结果。对于实录测试集而言，由于每个说话人没有对应的训练数据，因此我们利用仿真数据训练得到的语音模型来对实录测试集中的说话人信号建模。从结果来看，由于训练数据和测试数据不匹配，语音模型无法对测试集中的说话人信号较好的建模，导致所有算法的MOS结果下降较大。

3.3.6 小结

在本节中，我们将3.2节提出的语音建模方法结合FCRF模型用于分离两个说话人的语音信号。具体而言，首先通过语音建模方法得到每个说话人对应的状态序列和若干语音字典，然后训练FCRF模型用于描述混合语音信号的时间动态特性进行语音分离。本节所提算法由于对不同的语音帧采用了不同的语音字典进行建模，增强了字典的鉴别性，提高了对于非目标语音信号的抑制能力。同时采用FCRF对混合信号的时间动态特性进行建模，能够容纳更多类型的观测信息，进一步提高了算法的分离性能。实验结果表明，无论是客观评价指标还是主观测听实验，本节提出的算法相比其它一些算法能够获得更好的分离效果。

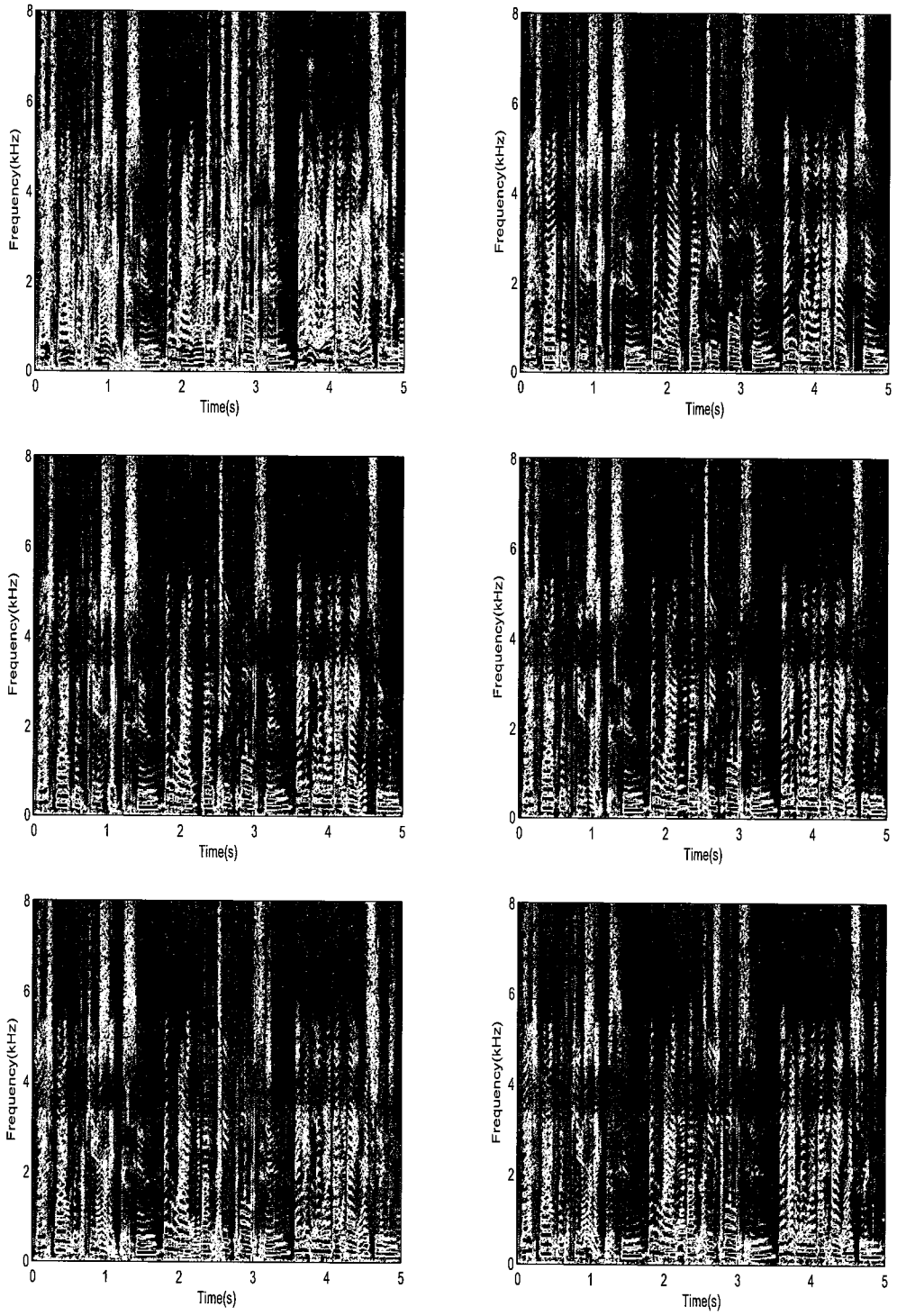

图 3.9: 各种算法分离后的语音频谱图

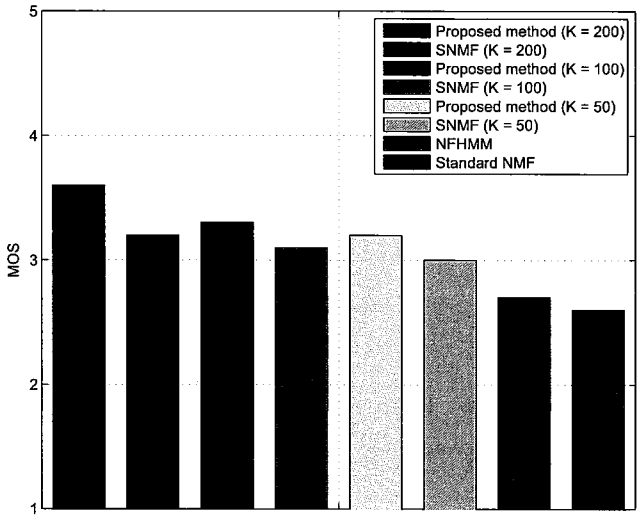

图 3.10: 各种算法在实录测试集上的MOS均值

3.4　基于语音连续性建模的语音增强算法

在3.3节中，我们将3.2节提出的语音建模方法和FCRF模型结合用于分离两个说话人的语音信号。实验结果表明，该算法相比其它一些算法无论在主观测听实验还是客观评价指标上都能够有效提高分离效果。在本节中，我们将3.2节提出的语音建模方法和FCRF模型结合用于语音增强，分离语音和噪声信号。算法的系统框图如图 3.11 所示，包括两个部分。在训练阶段，首先针对语音训练数据通过3.2节的算法得到一个状态序列和若干语音字典用于描述语音信号的时间连续性和频谱结构，同时对噪声训练数据通过k均值聚类得到噪声字典。然后训练FCRF模型用于描述混合信号的时间动态特性。在增强阶段，提取测试信号的特征送入已训练好的FCRF模型进行解码，得到每帧测试信号中语音对应状态的后验概率，再结合训练阶段得到的字典重构增强后的语音信号。

3.4.1　语音和噪声信号建模

在训练阶段，首先针对语音训练数据采用3.2节提出的语音建模算法得到一个状态序列和若干语音小字典，来描述语音信号的时间连续性和频谱结构特性。其次，对噪声训练数据通过k均值聚类得到噪声字典，来描述噪声信号的

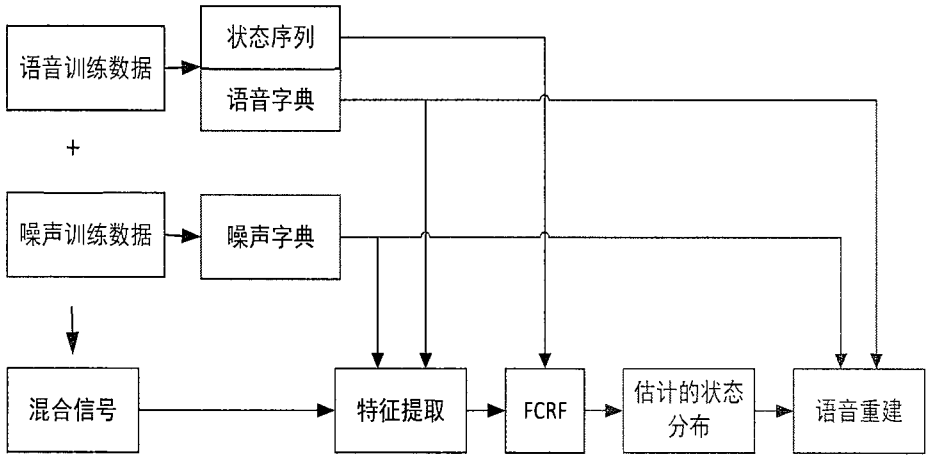

图 3.11: 本节提出的单通道语音增强算法系统框图。红色部分为训练阶段，黑色部分为训练阶段和增强阶段共用，绿色部分为增强阶段

谱结构信息。在这里我们假设语音信号具有时间连续性，因此采用一个状态序列对其连续性建模，而不对噪声信号做时间连续性假设，因此只使用一个噪声字典对噪声信号建模。

3.4.2 特征提取

在训练FCRF模型时，我们采用了和3.3节语音分离算法同样的观测数据类型：第一种为语音和噪声混合信号的幅度谱；第二种观测数据是对混合信号幅度谱进行非负矩阵分解之后得到的权重矩阵。在进行非负矩阵分解时，语音字典是训练阶段得到的所有语音小字典的集合，噪声字典是通过对噪声训练数据聚类而得到。

3.4.3 FCRF

在本节提出的语音增强算法中，我们采用和3.3节语音分离算法同样的FCRF模型。FCRF的图模型如图 3.4 所示。需要指出的是，在本节语音增强算法所使用的FCRF模型中，第一条线性链$\boldsymbol{y}_1$用来描述语音信号的时间连续性，包含多个状态，每个状态对应一个语音字典矩阵，而第二条线性链$\boldsymbol{y}_2$用来描述噪声信号，由于不对噪声信号做时间连续性假设，因此该线性链只有一个状

态，而且只对应一个噪声字典矩阵。

在训练FCRF模型时，首先提取混合训练数据的特征作为FCRF模型的输入，通过语音训练数据得到的状态序列作为线性标记链$\boldsymbol{y}_1$，而另一个线性标记链$\boldsymbol{y}_2$代表噪声信号，只有一个状态。在训练时，通过L-BFGS算法 [128] 估计模型参数λ_k。在增强阶段，提取测试信号的特征作为已训练好的FCRF模型的输入，然后针对每帧测试信号得到语音对应的状态后验概率分布$p(y_{1,t}|\boldsymbol{x})$。该后验概率被用来重构增强后的语音信号幅度谱。

3.4.4　语音重建

在通过FCRF模型解码得到$p(y_{1,t}|\boldsymbol{x})$之后，结合训练阶段得到的语音和噪声字典，通过以下方式重构语音和噪声的幅度谱：

$$\hat{\boldsymbol{x}}_{S,t} = \sum_{m=1}^{M} p(y_{1,t} = m|\boldsymbol{x}) \boldsymbol{W}_{S,m} \boldsymbol{h}_{S,m,t} \tag{3.15}$$

$$\hat{\boldsymbol{x}}_{N,t} = \sum_{m=1}^{M} p(y_{1,t} = m|\boldsymbol{x}) \boldsymbol{W}_{N} \boldsymbol{h}_{N,m,t} \tag{3.16}$$

其中，$\hat{\boldsymbol{x}}_{S,t}$和$\hat{\boldsymbol{x}}_{N,t}$是增强后得到的语音和噪声信号对应第$t$帧的幅度谱。$p(y_{1,t} = m|\boldsymbol{x})$是在第$t$帧时语音信号对应状态是$m$的后验概率。$\boldsymbol{W}_{S,m}$是语音信号第$m$个状态对应的字典矩阵，$\boldsymbol{W}_N$是噪声信号对应的字典矩阵。$\boldsymbol{h}_{S,m,t}$和$\boldsymbol{h}_{N,m,t}$是将第$t$帧测试信号幅度谱$\boldsymbol{x}_t$进行非负矩阵分解得到的权重向量：

$$\boldsymbol{x}_t = \begin{bmatrix} \boldsymbol{W}_{S,m} & \boldsymbol{W}_N \end{bmatrix} \begin{bmatrix} \boldsymbol{h}_{S,m,t} \\ \boldsymbol{h}_{N,m,t} \end{bmatrix}, \quad m = 1, \cdots, M \tag{3.17}$$

式(3.15)-(3.16)可以看做是加权的基于NMF的语音增强算法，语音信号采用多个小字典建模，每个语音字典对应的权重通过FCRF估计得到，该权重用于描述对应语音字典对于该帧语音信号的重要程度。

在得到$\hat{\boldsymbol{x}}_{S,t}$和$\hat{\boldsymbol{x}}_{N,t}$之后，再通过维纳滤波的形式得到最终增强后的语音信号幅度谱：

$$\tilde{\boldsymbol{x}}_{S,t} = \boldsymbol{x}_t \otimes \frac{\hat{\boldsymbol{x}}_{S,t}}{\hat{\boldsymbol{x}}_{S,t} + \hat{\boldsymbol{x}}_{N,t}} \tag{3.18}$$

其中，$\otimes$和$\frac{a}{b}$表示向量对应元素的相乘和相除运算。

最后，将$\tilde{\boldsymbol{x}}_{S,t}$结合测试信号的相位信息，并通过iSTFT以及重叠相加算法得到增强后的语音时域信号。

3.4.5 实验评价和讨论

3.4.5.1 实验设置

在实验中语音数据来自于Grid数据集 [130]，该数据集共包含34个说话人的语音数据。噪声数据我们使用了三种噪声类型：airport噪声、babble噪声和restaurant噪声。从Grid数据集中随机选择20个说话人语音信号（10位男性，10位女性说话人）和3种噪声混合生成测试集数据。在每条测试数据中，语音信号是从对应说话人的语料中随机选择10句语音连接而成，噪声信号从每种噪声类型的前一半数据中得到。每条测试语音大约长16s，而且按照不同的信噪比进行混合（-6，-3，0，3和6dB）。对于测试集中每个说话人而言，从其Grid数据库中随机选择100条语句作为其训练数据，训练语音和测试语音不重合。每种噪声信号的前一半数据用于用于生成测试数据，后一半数据用于训练噪声字典。

由于噪声信号为8kHz采样，因此我们首先将语音信号降采样到8kHz，然后再进行处理。在进行STFT时，采用汉宁窗，帧长32ms，帧移16ms。

3.4.5.2 评价指标

为了评价算法的增强结果，我们使用了如下客观评价指标：

1. SDR、SIR和SAR [131]。这三种评价指标主要是用来衡量算法对于非目标信号的抑制能力，以及算法自身引入失真的大小。

2. Perceptual Evaluation of Speech Quality (PESQ) [132]。该评价指标主要是用来衡量算法增强后的语音质量。

这几种评价指标都是分数越高，表明增强结果越好。

3.4.5.3 算法描述

在实验中，我们比较了以下几种算法的增强结果：

1. 标准的基于NMF的语音增强算法 [57] 。

该算法采用KL散度作为目标函数以及乘法迭代准则作为分解算法。语音字典和噪声字典各包含50个基向量。

2. 稀疏NMF算法（sparse NMF, SNMF） [65]。

该算法字典通过k均值聚类得到，语音字典包括200个基向量，噪声字典包括50个基向量。矩阵分解算法通过激活集牛顿算法（Active-set Newton Algorithm, ASNA） [65]得到。稀疏参数γ设置为1。

3. NFHMM算法 [80]。

在NFHMM算法中，语音信号类别数为20，每个类别对应字典大小为10个基向量。噪声信号只有一个字典，包含50个基向量。

4. 本节提出的算法。

在本节提出的算法中，语音信号类别数$M = 10$，权重参数$\beta = 0.05$，语音字典大小$K = 200$，噪声字典大小$K = 50$。

3.4.5.4　结果分析

图3.12-图3.15给出了各种算法在不同输入信噪比时的SDR、SIR、SAR和PESQ的结果。从结果来看，对于所有输入信噪比情况而言，本节提出的算法相比其它算法而言能够有效提高SDR和PESQ指标，说明该算法能够得到更好的增强结果和语音质量。具体而言，本节所提算法在SIR指标上有较大提升而在SAR指标上有所下降，这主要是因为所提算法采用了较小的语音字典来描述不同的语音帧，因此能够对噪声信号有更好的抑制效果，与此同时由于语音字典较小，导致其对语音的描述能力变弱，因此导致SAR指标下降。但是从综合评价指标SDR来看，本节所提算法还是能够获得最好的增强效果。

3.4.6　小结

本节中，我们将3.2节提出的语音建模方法结合FCRF模型用于语音增强。在所提算法中，我们假设语音信号具有时间连续性，采用状态序列和多个小字典进行建模，而噪声信号没有时间连续性，只采用一个噪声字典建模，并且训练FCRF模型描述混合信号的时间动态特性进行语音增强。该算法由于对不同的语音帧采用了不同的语音字典进行建模，提高了对于噪声信号的抑制能力。同时采用FCRF对混合信号的时间动态特性进行建模，能够容纳更多的观测信息，进一步提高了增强结果。实验结果表明，本节提出的算法相比其它一些算法能够获得更好的增强效果，并且能够有效提高语音质量。

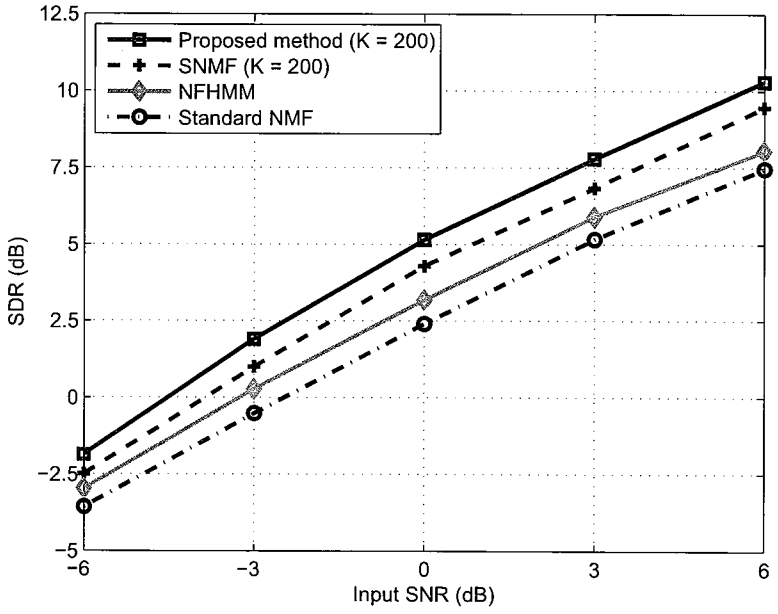

图 3.12: 各种算法在不同输入信噪比时的SDR均值

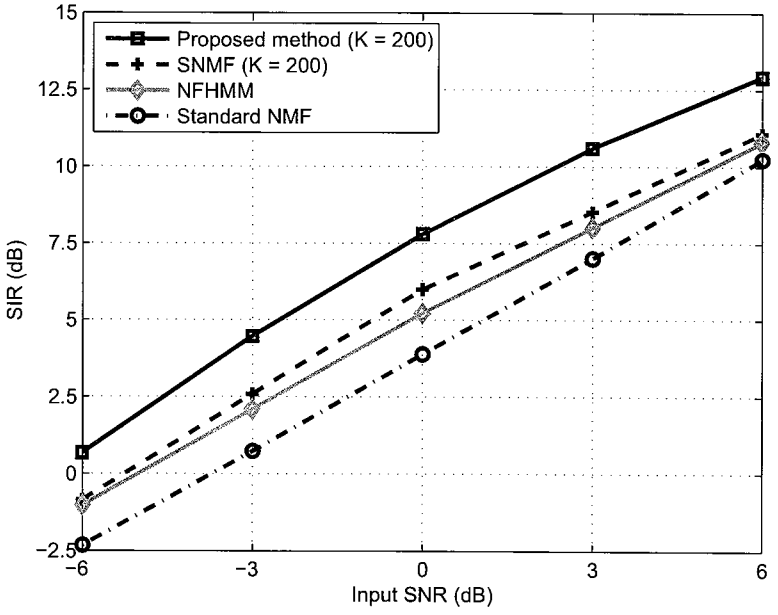

图 3.13: 各种算法在不同输入信噪比时的SIR均值

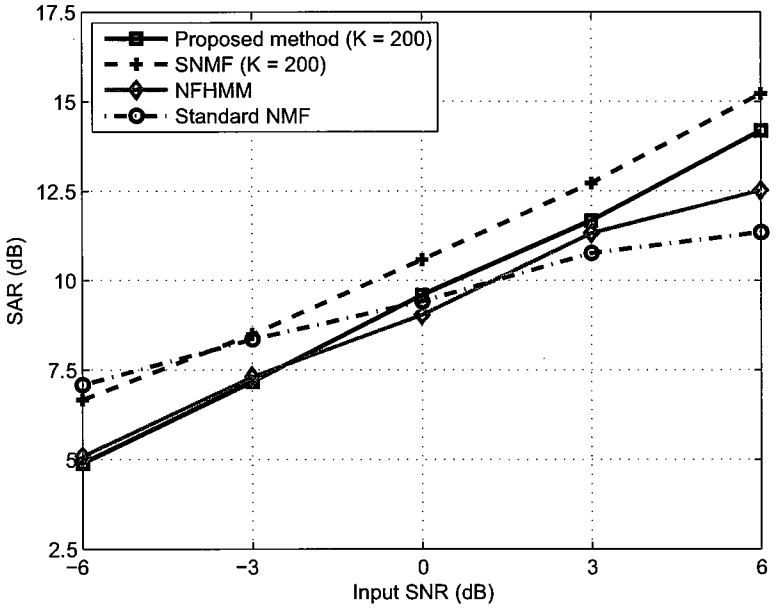

图 3.14: 各种算法在不同输入信噪比时的SAR均值

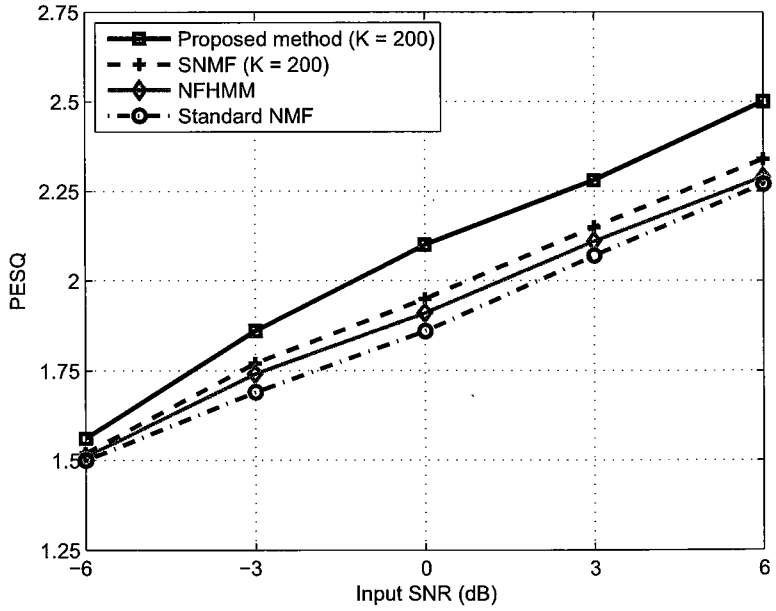

图 3.15: 各种算法在不同输入信噪比时的PESQ均值

第四章　基于非负矩阵分解的无监督语音增强算法研究

4.1　引言

在基于NMF的语音增强算法中，需要事先得到对应说话人或者对应噪声类型的训练数据。然而，在很多实际场景中很难获得匹配的训练数据，限制了这类算法在实际中的应用。很多基于NMF的改进算法被提出来解决上述问题。有学者将语音活动检测（voice activity detection，VAD）和NMF结合，首先通过VAD算法得到纯噪声段，然后通过得到的噪声段数据训练噪声字典，最后从混合信号中估计得到语音字典，进行无监督语音增强，这样就不需要语音和噪声的训练数据。然而这类算法非常依赖VAD算法的准确性，当非平稳噪声存在时，VAD准确性下降，导致增强效果不佳。后续有学者针对NMF算法提出了基于全局语音模型（Universal speech model，USM）的无监督语音增强算法 [133]，首先通过大量说话人的训练数据得到每个说话人对应的语音字典，将这些语音字典组成一个全局语音模型，然后在增强阶段，通过组稀疏惩罚项来从全局语音模型中选择少量的说话人语音字典描述测试信号中未见过的说话人语音信号，同时从混合信号中估计噪声字典，去除了需要特定说话人和特定噪声类型训练数据的限制。实验结果表明，该算法在某些情况下的增强结果甚至优于基于说话人依赖的半监督语音增强算法 [133]。

除此之外，很多基于NMF的语音增强算法都是对一段带噪语音进行降噪处理。在很多应用场景中，需要算法能够对带噪信号进行实时处理。因此，很多基于NMF的在线语音增强算法被提出来，如Duan和Joder等人分别提出了基于NMF的在线语音增强算法 [81, 82]，首先针对特定说话人的训练数据得到语音字典，然后逐帧进行非负矩阵分解，估计噪声字典和权重向量进行语音增强。该算法能够实现在线增强处理，但是需要测试信号中说话人的匹配训练数据，限制了这类算法在实际中的应用。Germain等人提出了基于USM的在线语音增强算法 [83]，通过语音训练数据得到全局语音模型，然后逐帧进行无监督非负矩阵分解，估计噪声字典和权重向量进行语音增强。该算法去除了需要特定说话人训练数据的限制，而且能够在线处理，因此受到了广泛关注。

本章我们继续对基于USM的语音增强算法进行研究，并对算法中的组稀疏

惩罚项进行改进，分别提出了基于自适应组稀疏和动态组稀疏惩罚项的增强算法，能够更好地从全局语音模型中选择匹配的语音字典，在保留语音成分的同时尽可能抑制噪声成分，提高增强效果。而且我们提出了一种与说话人无关的语音模型，并将其用于在线语音增强算法，非常利于实际使用。

4.2 基于USM的无监督语音增强算法回顾

图 4.1 是基于USM的无监督语音增强算法 [133] 框图。首先在训练阶段，采集大量说话人的纯净语音数据，训练全局语音模型；然后在增强阶段，对测试信号幅度谱进行无监督非负矩阵分解，从测试信号中估计噪声字典和权重矩阵，并且在分解时对全局语音模型进行组稀疏惩罚，选取最匹配的语音字典描述当前测试信号中的说话人语音信号。

4.2.1 全局语音模型

首先采集很多说话人的纯净语音数据，然后通过STFT得到每个说话人的语音幅度谱$\boldsymbol{X}_S^{(i)}$，对$\boldsymbol{X}_S^{(i)}$进行非负矩阵分解：

$$\boldsymbol{X}_S^{(i)} = \boldsymbol{W}_S^{(i)}\boldsymbol{H}_S^{(i)}, \ \ i = 1, ..., G \tag{4.1}$$

其中，$\boldsymbol{W}_S^{(i)}$是第i个说话人对应的语音字典，G是所有说话人的个数。式(4.1)可以通过乘法迭代准则 [56]进行分解。

将所有说话人的语音字典结合，组成一个较大的字典矩阵：

$$\boldsymbol{W}_S = [\boldsymbol{W}_S^{(1)}, \boldsymbol{W}_S^{(2)}, ..., \boldsymbol{W}_S^{(G)}] \tag{4.2}$$

$\boldsymbol{W}_S$即是通过大量说话人语音数据得到的全局语音模型，该模型包含很多说话人的语音字典。

4.2.2 无监督非负矩阵分解

在增强阶段，对测试信号幅度谱$\boldsymbol{X}$进行无监督非负矩阵分解，最小化如下目标函数：

$$\boldsymbol{W}_N, \boldsymbol{H} = \arg \ \min \ d_{\mathrm{KL}}(\boldsymbol{X}|\boldsymbol{W}\boldsymbol{H}) + \lambda\Omega(\boldsymbol{H}_S) \tag{4.3}$$

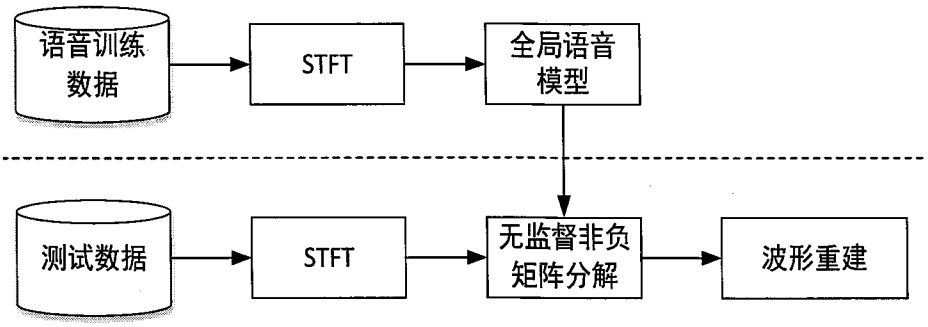

图 4.1: 基于USM的无监督语音增强算法框图。虚线上方为训练阶段，下方为增强阶段

其中，$\boldsymbol{W} = [\boldsymbol{W}_S \quad \boldsymbol{W}_N]$，$\boldsymbol{W}_N$是噪声字典，$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_S \\ \boldsymbol{H}_N \end{bmatrix}$是字典对应的权重矩阵，$\boldsymbol{H}_S = [\boldsymbol{H}_S^{(1)^T}, \cdots, \boldsymbol{H}_S^{(G)^T}]^T$是$\boldsymbol{W}_S$对应的权重矩阵，$\boldsymbol{H}_S^{(g)}$是$\boldsymbol{W}_S$中语音字典$\boldsymbol{W}_S^{(g)}$对应的权重矩阵。式(4.3)中目标函数第一项为KL散度距离，用来进行标准的非负矩阵分解。目标函数的第二项为$\log/l_1$组稀疏惩罚项 [133]：

$$\Omega(\boldsymbol{H}_S) = \sum_{g=1}^{G} \log(\epsilon + ||\boldsymbol{H}_S^{(g)}||_1) \tag{4.4}$$

其中，ϵ是一个极小正值，防止log函数自变量为负值，$||\ ||_1$为矩阵第一范式。在式(4.4)中，每个说话人字典对应的权重矩阵为一组。组稀疏惩罚项的目的在于对不同说话人语音字典对应的权重矩阵进行惩罚，从而抑制大多数说话人的语音字典，选择少数几个说话人的语音字典来描述测试信号中未见过说话人的语音信号。λ是稀疏参数，用于控制组稀疏惩罚项的相对大小。

在进行分解时，$\boldsymbol{W}_S$是全局语音模型，在分解时保持不变，从测试信号中估计噪声字典$\boldsymbol{W}_N$以及权重矩阵$\boldsymbol{H}$。这样就不需要事先获得噪声训练数据，同时由于全局语音模型和测试信号中的说话人无关，也不需要测试信号中对应说话人的训练数据，因此能进行无监督语音增强，扩大了算法的适用范围。而且，算法假设当全局语音模型足够大时，总是能从该模型中找到若干语音字典来描述测试信号中的未见过说话人的语音信号。为了防止全局语音模型描述测试信号中的噪声信号，需要对全局语音模型进行组稀疏性惩罚，抑制大多数说话人的语音字典，选择少数最匹配的语音字典。式(4.3) 可以通过以下算法 1 分

解得到:

Algorithm 1 基于全局语音模型的无监督非负矩阵分解算法

1) 输入：$\boldsymbol{X} \in \mathbb{R}_{+}^{M \times N}$，$\{\boldsymbol{W}_S^{(g)} \in \mathbb{R}_{+}^{M \times R_S} | 1 \leq g \leq G\}$，$R_N$
2) 输出：$\boldsymbol{W}_N$，$\boldsymbol{H}_S$，$\mathbf{H}_N$
3) 使用非负值随机初始化$\boldsymbol{W}_N$，$\boldsymbol{H}_S$和$\boldsymbol{H}_N$
4) **重复**

$\boldsymbol{W} = [\boldsymbol{W}_S \quad \boldsymbol{W}_N]$

$\boldsymbol{H} = [\boldsymbol{H}_S^T \quad \boldsymbol{H}_N^T]^T$

$\boldsymbol{R} \leftarrow \boldsymbol{X}./(\boldsymbol{W}\boldsymbol{H})$

$\boldsymbol{H} \leftarrow \boldsymbol{H}.*(\boldsymbol{W}^T\boldsymbol{R})$

for $g = 1, \cdots, G$

$\boldsymbol{H}_S^{(g)} \leftarrow \boldsymbol{H}_S^{(g)}/\{1 + \frac{\lambda}{\epsilon + ||\boldsymbol{H}_S^{(g)}||_1}\}$

end for

$\mathbf{W}_N \leftarrow \mathbf{W}_N.*(\boldsymbol{R}\boldsymbol{H}_N^T)$

$\mathbf{W}_N \leftarrow \mathbf{W}_N./(\mathbf{1}\mathbf{1}^T\boldsymbol{W}_N)$

5) **直至收敛**

.*和./表示矩阵对应元素之间的点乘和点除运算。

4.2.3 波形重建

在通过算法 1完成无监督分解之后，通过以下方式重构处理后的语音和噪声幅度谱：

$$\hat{\boldsymbol{X}}_S = \boldsymbol{W}_S\boldsymbol{H}_S \tag{4.5}$$

$$\hat{\boldsymbol{X}}_N = \boldsymbol{W}_N\boldsymbol{H}_N \tag{4.6}$$

在得到语音幅度谱$\hat{\boldsymbol{X}}_S$和噪声幅度谱$\hat{\boldsymbol{X}}_N$之后，再通过维纳滤波的形式得到最终增强后的语音幅度谱：

$$\tilde{\boldsymbol{X}}_S = \boldsymbol{X} \otimes \frac{\hat{\boldsymbol{X}}_S}{\hat{\boldsymbol{X}}_S + \hat{\boldsymbol{X}}_N} \tag{4.7}$$

其中，$\otimes$和$\frac{a}{b}$表示矩阵对应元素的相乘和相除运算。

最后，将$\tilde{\boldsymbol{X}}_S$结合测试信号的相位信息，并通过iSTFT以及重叠相加算法得到增强后的语音时域信号。

4.3　基于自适应组稀疏惩罚项的无监督语音增强算法

在4.2节基于USM的无监督语音增强算法中，目标函数中的组稀疏惩罚项对于算法的性能有至关重要的影响。组稀疏惩罚项的目的在于从很多说话人字典中选取少量字典来描述测试信号中的语音信号，在保留语音成分的同时抑制噪声成分。然而，4.2节算法中的组稀疏惩罚对全局语音模型中所有说话人字典对应的权重矩阵都采用了同样的稀疏参数，并没有考虑不同说话人语音字典的相对重要程度。因此，在本节中我们提出了自适应选择稀疏参数的方法来根据字典的重要程度不同选择不同的稀疏参数。当语音字典对于测试信号中的语音比较重要时，则对该语音字典采用较小的稀疏参数，反之则采用较大的稀疏参数。算法的具体描述如下所示。

4.3.1　全局语音模型

在训练阶段，同样采集大量说话人的语音训练数据，然后对每个说话人训练数据进行分解，得到其语音字典$\boldsymbol{W}_S^{(g)}$，再将所有说话人语音字典组合得到全局语音模型$\boldsymbol{W}_S$：

$$\boldsymbol{W}_S = [\boldsymbol{W}_S^{(1)}, \boldsymbol{W}_S^{(2)}, ..., \boldsymbol{W}_S^{(G)}] \tag{4.8}$$

其中，G是训练数据中所有说话人的个数。

4.3.2　自适应组稀疏无监督非负矩阵分解

在增强阶段，对测试信号幅度谱$\boldsymbol{X}$进行无监督矩阵分解，最小化如下目标函数：

$$\boldsymbol{W}_N, \boldsymbol{H} = \arg\ \min\ d_{\mathrm{KL}}(\boldsymbol{X}|\boldsymbol{W}\boldsymbol{H}) + \sum_{g=1}^{G} \lambda^{(g)} \log(\epsilon + ||\boldsymbol{H}_S^{(g)}||_1) \tag{4.9}$$

其中，$\boldsymbol{W} = [\boldsymbol{W}_S \quad \boldsymbol{W}_N]$，$\boldsymbol{W}_N$是噪声字典，$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_S \\ \boldsymbol{H}_N \end{bmatrix}$是字典对应的权重矩阵，$\boldsymbol{H}_S = [\boldsymbol{H}_S^{(1)^T}, \cdots, \boldsymbol{H}_S^{(G)^T}]^T$是$\boldsymbol{W}_S$对应的权重矩阵，$\boldsymbol{H}_S^{(g)}$是$\boldsymbol{W}_S$中语音字典$\boldsymbol{W}_S^{(g)}$对应的权重矩阵。式(4.9)中目标函数第一项为KL散度距离，用来进行标准的非负矩阵分解。目标函数的第二项为自适应组稀疏惩罚项，目的在于对不同说话人语音字典对应的权重矩阵进行稀疏惩罚，使得大多数字典对应的权重矩阵值足够小，只选择少量的语音字典用于重构测试信号中的语音。

需要注意的时，与式(4.3)中的组稀疏惩罚项不同，式(4.9)中的组稀疏惩罚项对不同语音字典对应的权重矩阵采用了不同的稀疏参数$\lambda^{(g)}$，而且$\lambda^{(g)}$的值在算法中自适应变化。如果一个语音字典能够较好地描述测试信号中的语音信息，那么其对应的稀疏参数就应该选择较小的值来保留该语音字典。反之如果该语音字典不能很好地描述测试信号中语音信息，则应该尽可能抑制该语音字典，使其字典对应的权重值尽可能小，从而需要选择较大的稀疏参数。我们按照以下规则自适应更新稀疏参数：

$$\lambda^{(g)} = \lambda_0 \frac{\max\limits_{g}\{H_{S,sum}^{(g)}\}}{H_{S,sum}^{(g)}} \tag{4.10}$$

其中，$H_{S,sum}^{(g)} = ||\boldsymbol{H}_S^{(g)}||_1$，$\lambda_0$是初始化的稀疏参数值。从式(4.10)可以看出，每个语音字典对应的稀疏参数$\lambda^{(g)}$和其对应的权重矩阵的和$H_{S,sum}^{(g)}$是负相关关系。$H_{S,sum}^{(g)}$值越大，说明其对应的语音字典$\boldsymbol{W}_S^{(g)}$越重要，则选择较小的稀疏参数，反之亦然。因此，通过式(4.10)可以保留重要的语音字典，抑制不太重要的语音字典，能够在保留语音成分的同时抑制噪声成分。式(4.9)可以通过算法 2进行分解得到。

4.3.3 波形重建

在通过算法 2完成无监督分解之后，通过以下方式重构处理后的语音和噪声幅度谱：

$$\hat{\boldsymbol{X}}_S = \boldsymbol{W}_S \boldsymbol{H}_S \tag{4.11}$$

$$\hat{\boldsymbol{X}}_N = \boldsymbol{W}_N \boldsymbol{H}_N \tag{4.12}$$

在得到语音幅度谱$\hat{\boldsymbol{X}}_S$和噪声幅度谱$\hat{\boldsymbol{X}}_N$之后，再通过维纳滤波的形式得到最终增强后的语音幅度谱：

$$\tilde{\boldsymbol{X}}_S = \boldsymbol{X} \otimes \frac{\hat{\boldsymbol{X}}_S}{\hat{\boldsymbol{X}}_S + \hat{\boldsymbol{X}}_N} \tag{4.13}$$

其中，$\otimes$和$\frac{a}{b}$表示矩阵对应元素的相乘和相除运算。

最后，将$\tilde{\boldsymbol{X}}_S$结合测试信号的相位信息，并通过iSTFT以及重叠相加算法得到最终增强后的语音时域信号。

Algorithm 2 基于自适应组稀疏惩罚的无监督非负矩阵分解算法

1) 输入：$\boldsymbol{X} \in \mathbb{R}_{+}^{M \times N}$，$\{\boldsymbol{W}_S^{(g)} \in \mathbb{R}_{+}^{M \times R_S} | 1 \leq g \leq G\}$，$R_N$
2) 输出：$\boldsymbol{W}_N$，$\boldsymbol{H}_S$，$\boldsymbol{H}_N$
3) 采用基于USM算法的结果初始化$\boldsymbol{H}_S$和$\boldsymbol{H}_N$，随机初始化$\boldsymbol{W}_N$
4) **重复**

$$\boldsymbol{W} = [\boldsymbol{W}_S \quad \boldsymbol{W}_N]$$
$$\boldsymbol{H} = [\boldsymbol{H}_S^T \quad \boldsymbol{H}_N^T]^T$$
$$\boldsymbol{R} \leftarrow \boldsymbol{X}./(\boldsymbol{W}\boldsymbol{H})$$
$$\boldsymbol{H} \leftarrow \boldsymbol{H}.*(\boldsymbol{W}^T\boldsymbol{R})$$
for $g = 1, \cdots, G$
$$H_{S,sum}^{(g)} = ||\boldsymbol{H}_S^{(g)}||_1$$
end for
for $g = 1, \cdots, G$
$$\lambda^{(g)} = \lambda_0 \frac{\max_g\{H_{S,sum}^{(g)}\}}{H_{S,sum}^{(g)}}$$
$$\boldsymbol{H}_S^{(g)} \leftarrow \boldsymbol{H}_S^{(g)}/\{1 + \frac{\lambda^{(g)}}{\epsilon + ||\boldsymbol{H}_S^{(g)}||_1}\}$$
end for
$$\mathbf{W}_N \leftarrow \mathbf{W}_N.*(\boldsymbol{R}\boldsymbol{H}_N^T)$$
$$\mathbf{W}_N \leftarrow \mathbf{W}_N./(\mathbf{1}\mathbf{1}^T\boldsymbol{W}_N)$$

6) **直至收敛**

4.3.4 实验评价和讨论

4.3.4.1 实验设置

在实验部分，我们通过从带噪信号中估计得到语音信号来对所提算法进行评价。具体而言，在实验中语音数据来自于TIMIT数据集 [134]，噪声数据来自于 [81]，共包含10种噪声类型，主要是非平稳噪声，如键盘敲击声和鸟鸣声等。从TIMIT训练集中随机选择20个说话人（每个说话人10条语音）作为语音训练数据得到全局语音模型。从TIMIT测试集中随机选择5个说话人（每个说话人1条语音）和 [81]中10种噪声信号混合得到50条测试信号。测试集信号按照不同的信噪比进行混合（-10，-5，0和5dB）。

所有的信号首先通过重采样到16kHz，然后通过STFT得到其频谱，在变换

时采用汉宁窗，帧长64ms，帧移16ms。

4.3.4.2 评价指标

为了评价算法的增强效果，使用了三种客观评价指标：SDR、SIR和SAR [131]。这三种评价指标主要是用来衡量算法对于非目标信号的抑制能力，以及算法自身引入失真的大小。

4.3.4.3 比较算法

为了评价算法性能，我们比较了以下几种算法的结果：

1. 标准的基于NMF的半监督语音增强算法 [57]。

该算法采用KL散度作为目标函数以及乘法迭代准则作为分解算法。对测试信号中的每个说话人，采用TIMIT测试集中该说话人剩余的9条语音训练与该说话人匹配的语音字典，字典包含20个基向量。噪声字典大小根据噪声类型的不同选择最优的参数 [81]。

2. 基于USM的无监督语音增强算法 [133]。

该算法在训练全局语音模型时，每个说话人语音字典包含$R_S = 10$个基向量，共有$G = 20$个说话人字典。迭代次数和稀疏参数λ的选取通过调整来获得最大的SDR指标。噪声字典大小R_N根据噪声类型的不同选择最优的参数 [81]。

3. 本节提出的算法。

在本节提出的算法中，语音字典采用全局语音模型，$R_S = 10$，$G = 20$，噪声字典大小R_N根据噪声类型的不同选择最优的参数 [81]。在算法运行时，首先迭代基于USM的算法10次，得到权重矩阵作为本节算法的初始值。λ_0的选择通过调整来获得最大的SDR指标。

4.3.4.4 结果分析

首先我们比较了几种算法在无监督情况下的结果。在无监督情况下，噪声字典通过测试信号迭代得到。图4.2-图4.4给出了各种算法在不同输入信噪比时的SDR、SIR 和SAR的结果。从结果来看，对于所有输入信噪比情况而言，本节提出的算法相比基于USM的算法能够有效提高SDR指标，说明所提算法能够得到更好的增强效果。具体而言，所提算法在保持SAR 的情况下在SIR上有较大提升，说明该算法在不引入更多失真的前提下能够去除更多的噪声信号。这

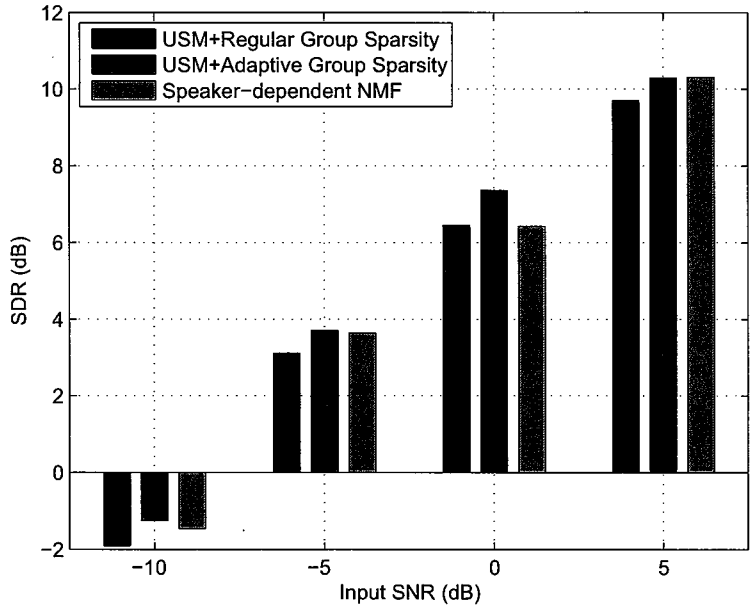

图 4.2: 各种算法在不同输入信噪比时的SDR均值

主要是因为算法针对不同的说话人字典选择了不同的稀疏参数，这样能够保留最重要的若干语音字典，同时去除不重要的字典以防止其用来描述噪声信号，从而去除更多的噪声。

其次，我们进一步比较了几种算法在半监督情况下输入信噪比为0dB时的结果。在半监督情况下，针对每种噪声类型我们事先得到了其训练数据，用于训练匹配的噪声字典。表4.1给出了几种算法在半监督情况下SDR、SIR和SAR的结果。相比其它算法而言，本节所提算法并没有在指标上有明显提升。一个可能的解释是当语音和噪声字典都固定时，标准的组稀疏惩罚项就能够找到最合适的语音字典，而当噪声字典无法事先获取时，则难以找到最匹配的语音字典，自适应组稀疏惩罚则更有优势。

4.3.5　小结

在本节中，我们基于全局语音模型，提出了基于自适应组稀疏惩罚项的算法用于无监督语音增强。算法能够针对不同的语音字典自适应地选择稀疏参数，从而能够保留最匹配的语音字典，抑制无关的语音字典，进而能够保留更多的语音成分同时抑制噪声信号。实验结果表明，本节提出的算法相比标准

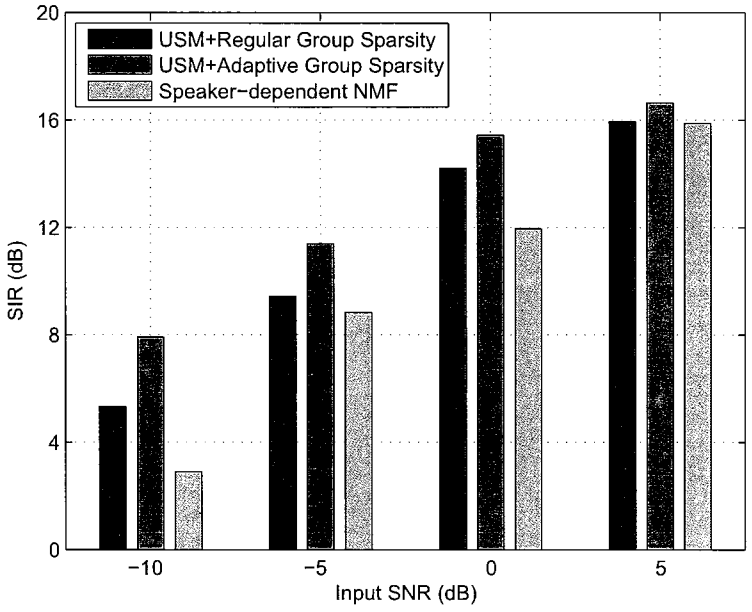

图 4.3: 各种算法在不同输入信噪比时的SIR均值

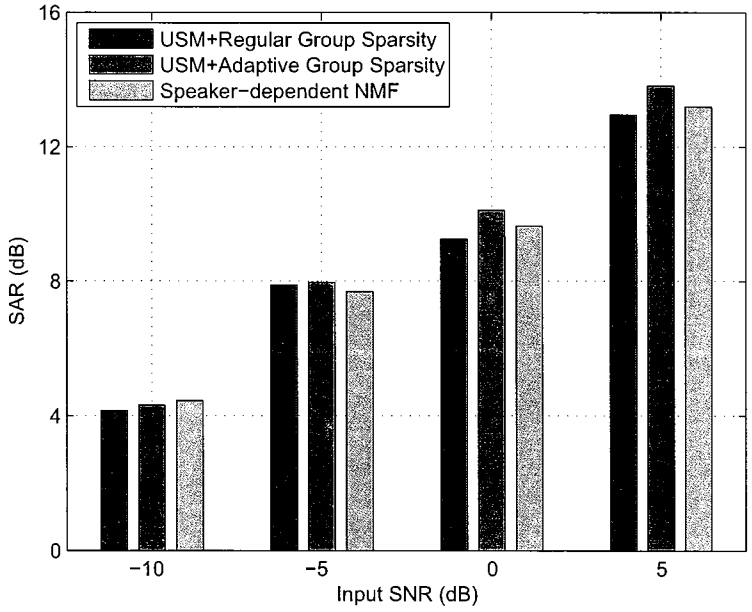

图 4.4: 各种算法在不同输入信噪比时的SAR均值

表 4.1: 各种算法在半监督情况下SDR、SIR和SAR的结果

	SDR(dB)	SIR(dB)	SAR(dB)
USM+Regular Group Sparsity	9.98	19.15	11.36
USM+Adaptive Group Sparsity	9.87	20.12	11.02
Speaker-dependent NMF	9.54	15.97	11.57

的NMF算法，以及固定参数的组稀疏惩罚项算法能够有效提高语音增强效果。

4.4 基于动态组稀疏惩罚项的无监督语音增强算法

在4.3节中我们提出了一种自适应组稀疏惩罚项的方法用于改进4.2节中的算法性能。在本节中我们提出了一种动态组稀疏惩罚项的方法从另一个方面改进4.2节中的算法性能。在4.2节的增强算法中，目标函数中的稀疏惩罚项对所有说话人的语音字典采用一种全局惩罚的方式，即对所有测试信号语音帧选择相同的语音字典建模，并没有考虑到语音谱的动态变化。因此，在本节中我们提出了动态组稀疏惩罚函数针对不同的语音帧采用不同的语音字典进行建模，并对相邻的语音帧采用同样的字典建模，这样能够同时描述语音信号的频谱动态变化和时间连续特性。算法的具体描述如下所示。

4.4.1 全局语音模型

在训练阶段，同样采集很多说话人的训练数据，然后对每个说话人训练数据进行分解，得到其语音字典$\boldsymbol{W}_S^{(g)}$，再将所有说话人语音字典组合得到全局语音模型$\boldsymbol{W}_S$：

$$\boldsymbol{W}_S = [\boldsymbol{W}_S^{(1)}, \boldsymbol{W}_S^{(2)}, ..., \boldsymbol{W}_S^{(G)}] \tag{4.14}$$

其中，G是训练数据中所有说话人的个数，每个说话人字典对应组稀疏中的一个组。

4.4.2 动态组稀疏无监督非负矩阵分解

在增强阶段，对测试信号幅度谱$\boldsymbol{X}$进行无监督矩阵分解，最小化如下目标函数：

$$\boldsymbol{W}_N, \boldsymbol{H} = \arg\ \min\ d_{\mathrm{KL}}(\boldsymbol{X}|\boldsymbol{W}\boldsymbol{H}) + \lambda \sum_t \Omega(\boldsymbol{h}_{S,t,L}) \tag{4.15}$$

其中，$\boldsymbol{W}=[\boldsymbol{W}_S \ \ \boldsymbol{W}_N]$，$\boldsymbol{W}_N$是噪声字典，$\boldsymbol{H}=\begin{bmatrix}\boldsymbol{H}_S\\ \boldsymbol{H}_N\end{bmatrix}$是字典对应的权重矩阵，$\boldsymbol{H}_S=[\boldsymbol{h}_{S,1},\cdots,\boldsymbol{h}_{S,t},\cdots,\boldsymbol{h}_{S,T}]$是$\boldsymbol{W}_S$对应的权重矩阵，$\boldsymbol{h}_{S,t}$是第$t$帧对应的权重向量，$\boldsymbol{h}_{S,t}=[\boldsymbol{h}_{S,t}^{(1)^T},\cdots,\boldsymbol{h}_{S,t}^{(G)^T}]^T$，$\boldsymbol{h}_{S,t}^{(g)}$是$\boldsymbol{W}_S$中语音字典$\boldsymbol{W}_S^{(g)}$对应第$t$帧的权重向量。式(4.15)中目标函数第一项为KL散度距离，目标函数的第二项为动态组稀疏惩罚项：

$$\Omega(\boldsymbol{h}_{S,t,L})=\sum_{g=1}^{G}\log(\epsilon+||\boldsymbol{h}_{S,t-L:t+L}^{(g)}||_1) \tag{4.16}$$

其中，$\boldsymbol{h}_{S,t-L:t+L}^{(g)}=[\boldsymbol{h}_{S,t-L}^{(g)},\cdots,\boldsymbol{h}_{S,t}^{(g)},\cdots,\boldsymbol{h}_{S,t+L}^{(g)}]$。该稀疏惩罚项针对不同短时窗内的语音帧分别进行稀疏惩罚，在同一窗内的语音帧选择同样的语音字典进行描述，而不在同一窗内的语音帧则选择的不同的语音字典，这样就增强了算法对不同语音帧的建模能力，能够描述语音幅度谱的动态变化特性。L是短时窗的长度，可以自适应调整，如果$L=1$，则说明对不同的帧都选择不同的语音字典建模。考虑到语音信号的时候连续性，一般选择$L>1$，使得相邻的语音帧采用的同样的字典建模。式(4.15)可以通过算法 3进行分解得到。

4.4.3 波形重建

在通过算法 3完成无监督分解之后，通过以下方式重构处理后的语音和噪声幅度谱：

$$\hat{\boldsymbol{X}}_S=\boldsymbol{W}_S\boldsymbol{H}_S \tag{4.17}$$

$$\hat{\boldsymbol{X}}_N=\boldsymbol{W}_N\boldsymbol{H}_N \tag{4.18}$$

在得到语音幅度谱$\hat{\boldsymbol{X}}_S$和噪声幅度谱$\hat{\boldsymbol{X}}_N$之后，再通过维纳滤波的形式得到最终增强后的语音幅度谱：

$$\tilde{\boldsymbol{X}}_S=\boldsymbol{X}\otimes\frac{\hat{\boldsymbol{X}}_S}{\hat{\boldsymbol{X}}_S+\hat{\boldsymbol{X}}_N} \tag{4.19}$$

其中，$\otimes$和$\frac{a}{b}$表示矩阵对应元素的相乘和相除运算。

最后，将$\tilde{\boldsymbol{X}}_S$结合测试信号的相位信息，并通过iSTFT以及重叠相加算法得到最终增强后的语音时域信号。

Algorithm 3 基于动态组稀疏惩罚的无监督非负矩阵分解算法

1) 输入：$\boldsymbol{X} \in \mathbb{R}_{+}^{F \times T}$，$\{\boldsymbol{W}_S^{(g)} \in \mathbb{R}_{+}^{F \times R_S} | 1 \leq g \leq G\}$，$R_N$，$L$
2) 输出：$\mathbf{W}_N$，$\boldsymbol{H}_S$和$\boldsymbol{H}_N$
3) 使用非负值随机初始化$\boldsymbol{W}_N$，$\boldsymbol{H}_S$和$\boldsymbol{H}_N$
4) **重复**

$$\boldsymbol{W} = [\boldsymbol{W}_S \quad \boldsymbol{W}_N]$$
$$\boldsymbol{H} = [\boldsymbol{H}_S^T \quad \boldsymbol{H}_N^T]^T$$
$$\boldsymbol{R} \leftarrow \boldsymbol{X}./(\boldsymbol{W}\boldsymbol{H})$$
$$\boldsymbol{H} \leftarrow \boldsymbol{H}.*(\boldsymbol{W}^T\boldsymbol{R})$$
for $g = 1, \cdots, G, t = L+1, \cdots, T-L$
$$\boldsymbol{h}_{S,t}^{(g)} \leftarrow \boldsymbol{h}_{S,t}^{(g)} / \{1 + \frac{\lambda}{\epsilon + ||\boldsymbol{h}_{S,t-L:t+L}^{(g)}||_1}\}$$
end for
for $g = 1, \cdots, G, t = 1:L, T-L+1:T$
$$\boldsymbol{h}_{S,t}^{(g)} \leftarrow \boldsymbol{h}_{S,t}^{(g)} / \{1 + \frac{\lambda}{\epsilon + ||\boldsymbol{h}_{S,t}^{(g)}||_1}\}$$
end for
$$\mathbf{W}_N \leftarrow \mathbf{W}_N.*(\boldsymbol{R}\boldsymbol{H}_N^T)$$
$$\mathbf{W}_N \leftarrow \mathbf{W}_N./(\mathbf{1}\mathbf{1}^T\boldsymbol{W}_N)$$

5) **直至收敛**

4.4.4 实验评价和讨论

4.4.4.1 实验设置

在实验部分，语音数据来自于TIMIT数据集 [134]，噪声数据来自于 [81]，共包含10种噪声类型，主要是非平稳噪声，如键盘敲击声和鸟鸣声等。从TIMIT训练集中随机选择20个说话人（每个说话人10条语音）作为训练数据得到全局语音模型。从TIMIT测试集中随机选择5个说话人（每个说话人1条语音）和 [81] 中10种噪声信号混合得到50条测试信号。测试集信号按照不同的信噪比进行混合（-10，-5，0和5dB）。

所有的信号首先通过重采样到16kHz，然后通过STFT得到其频谱，在变换时采用汉宁窗，帧长64ms，帧移16ms。

4.4.4.2 评价指标

为了评价算法的增强效果，使用了三种客观评价指标：SDR、SIR和SAR [131]。这三种评价指标主要是用来衡量算法对于非目标信号的抑制能力，以及算法自身引入失真的大小。

4.4.4.3 比较算法

为了评价算法性能，我们比较了以下几种算法的结果：

1. 标准的基于NMF的半监督语音增强算法 [57]。

该算法采用KL散度作为目标函数以及乘法迭代准则作为分解算法。对测试信号中的每个说话人，采用TIMIT测试集中该说话人剩余的9条语音训练与该说话人匹配的语音字典，字典包含20个基向量。噪声字典大小根据噪声类型的不同选择最优的参数 [81]。

2. 基于USM的无监督语音增强算法 [133]。

该算法在训练全局语音模型时，每个说话人语音字典包含$R_S = 10$个基向量，共有$G = 20$个说话人字典。迭代次数和稀疏参数λ的选择通过调整来获得最大的SDR指标。噪声字典大小R_N根据噪声类型的不同选择最优的参数 [81]。

3. 本节提出的算法。

在本节提出的算法中，语音字典采用全局语音模型，$R_S = 10$，$G = 20$，噪声字典大小R_N根据噪声类型的不同选择最优的参数 [81]。λ和L的选择通过调整来获得最大的SDR指标。

4.4.4.4 结果分析

在实验中我们首先比较了参数L的取值对所提算法增强结果的影响。表4.2给出了在无监督条件下，不同输入信噪比和不同L取值时的SDR结果。其中，$L = 0$对应了算法对每一帧语音信号都选择不同的语音字典进行描述，而$L > 0$则对应算法对语音相邻的L帧信号选择同样的语音字典，考虑了语音信号的时间连续性，而当L等于输入信号的帧数时，算法则退化为原始的基于USM的算法。从结果来看，在相同输入信噪比时，$L > 0$时的SDR要高于$L = 0$时的SDR结果，说明考虑语音信号的时间连续性能够有效提高增强效果。其次，最优的L取值一般是$L = 3$或者$L = 4$，当L取值更大时，算法性能下

降。这主要是因为语音信号是短时平稳的，如果间隔的语音帧过长，则需要采用不同的语音字典建模。

表 4.2: 所提算法在无监督情况下不同L取值时的SDR结果

SDR(dB)		L						
		0	1	2	3	4	5	6
SNR	-10	-1.4	-1.5	-1.4	-1.3	−1.3	**-1.1**	-1.3
	-5	3.59	3.71	3.79	3.79	**3.82**	3.81	3.71
	0	6.67	6.96	7.13	**7.18**	7.12	7.01	6.85
	5	9.8	9.88	9.98	**10.09**	9.98	9.89	9.90

其次我们比较了几种算法在无监督情况下的结果。在无监督情况下，噪声字典通过测试信号迭代得到。图4.5-图4.7给出了各种算法在不同输入信噪比时的SDR、SIR 和SAR的结果。从结果来看，对于所有输入信噪比情况而言，本节提出的算法相比基于USM算法而言能够有效提高SDR指标。具体而言，所提算法在保持SAR的情况下在SIR上有较大提升，说明该算法在不引入更多语音失真的前提下能够有效去除噪声信号。这主要是因为算法针对不同的语音帧选择了不同的语音字典进行建模，而且对于相邻的语音帧采用同样的语音字典，能够同时对语音谱结构和时间连续性建模。除此之外，在很多测试条件下，本节所提算法相比基于说话人依赖的NMF算法也能够提高增强效果。这主要是因为尽管无法事先获取测试信号中对应说话人的训练数据，采用大量说话人的语音字典并且对其进行动态选择能够对未见过的说话人语音信号有很好的建模能力。

最后，我们进一步比较了几种算法在半监督情况下输入信噪比为0dB时的结果。表4.3给出了几种算法在半监督情况下SDR、SIR 和SAR的结果。相比其它算法而言，本节所提算法在指标上有微弱提升。一个可能的解释是当语音和噪声字典都固定时，标准的组稀疏惩罚项就能够对语音和噪声信号进行较好的区分，导致本节所提的动态组稀疏惩罚算法性能提升不明显。

4.4.5 小结

在本节中，我们基于全局语音模型，提出了动态组稀疏惩罚项的算法用于无监督语音增强。算法能够针对不同的语音帧选择不同的字典进行描述，而且

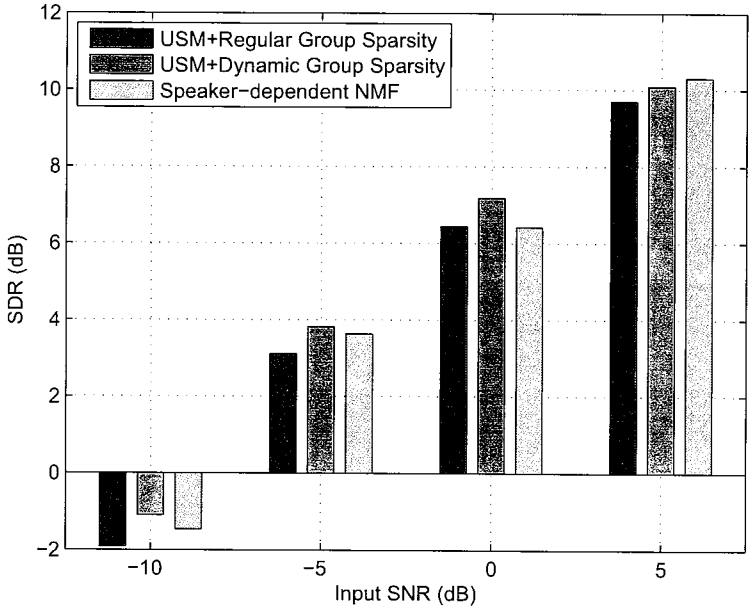

图 4.5: 各种算法在不同输入信噪比时的SDR均值

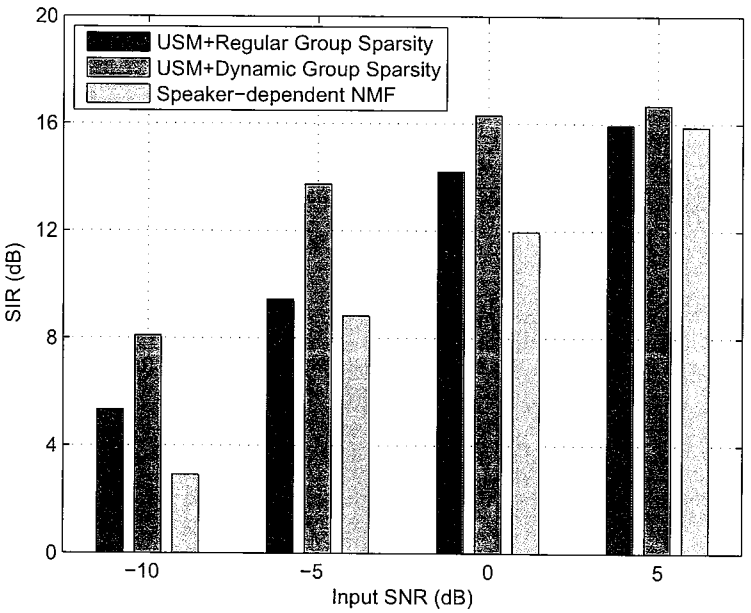

图 4.6: 各种算法在不同输入信噪比时的SIR均值

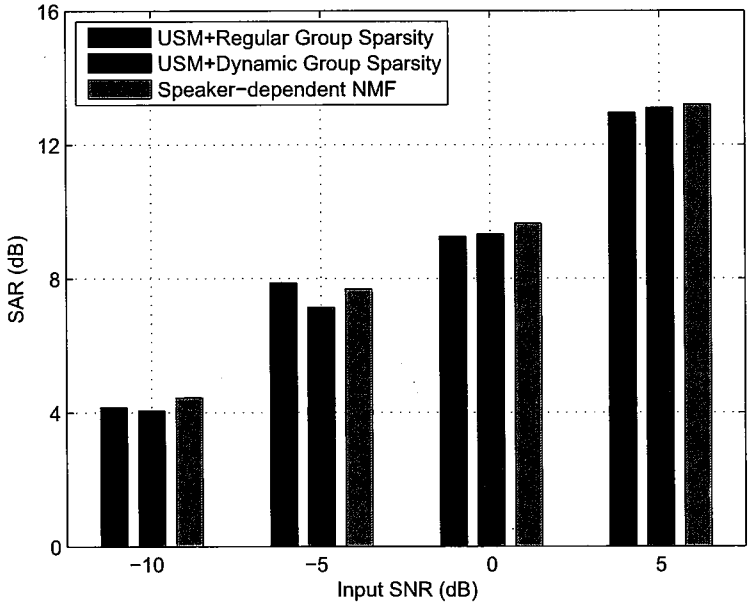

图 4.7: 各种算法在不同输入信噪比时的SAR均值

表 4.3: 各种算法在半监督情况下SDR、SIR和SAR的结果。

	SDR(dB)	SIR(dB)	SAR(dB)
USM+Regular Group Sparsity	9.98	19.15	11.36
USM+Dynamic Group Sparsity	10.08	20.27	11.06
Speaker-dependent NMF	9.54	15.97	11.57

对于相邻的语音帧选择同样的字典建模，能够同时描述语音信号谱结构变化和时间连续性，提高了算法对语音信号的建模能力。实验结果表明，本节提出的算法相比标准的组稀疏惩罚算法，甚至是基于说话人依赖的NMF算法都能够有效提高语音增强效果。

4.5 无监督在线语音增强算法

在前面几节描述的无监督语音增强算法中，不需要事先获得特定说话人和特定噪声类型的训练数据，扩大了算法的适用范围。而且算法不需要平稳噪声的假设，能够对非平稳噪声有较好的抑制能力。然而上述算法都是对一段测试

语音进行处理，在很多实际应用中，需要算法能够对带噪数据实时处理。在本节中，我们提出了一种无监督在线语音增强算法。具体而言，在训练阶段首先通过k均值聚类和NMF算法从大量说话人的训练数据中得到与说话人无关的语音模型，然后在增强阶段对测试信号逐帧进行无监督非负矩阵分解，估计噪声字典和权重向量，进行在线语音增强。算法的具体描述如下所示。

4.5.1 与说话人无关的语音模型训练

在训练阶段，首先采集大量说话人的训练数据，通过STFT得到幅度谱，然后对所有说话人的幅度谱$\boldsymbol{X}_S$进行k均值聚类，在聚类时采用KL散度作为目标函数。聚类完成之后，将属于同一类别的语音帧集合在一起，得到每一类别对应的语音幅度谱$\boldsymbol{X}_S^{(g)}$，再对其进行非负矩阵分解：

$$\boldsymbol{X}_S^{(g)} \approx \boldsymbol{W}_S^{(g)} \boldsymbol{H}_S^{(g)}, \quad g = 1, \cdots, G \tag{4.20}$$

其中，$\boldsymbol{W}_S^{(g)} \in \mathbb{R}_+^{F \times K_S}$是第$g$个类别对应的语音字典，用于描述语音不同的频谱结构信息，G是聚类的类别数。最后将所有描述不同谱结构信息的语音字典结合，组成一个与说话人无关的语音模型：

$$\boldsymbol{W}_S = [\boldsymbol{W}_S^{(1)}, \cdots, \boldsymbol{W}_S^{(G)}] \tag{4.21}$$

需要注意的是，不同于前面几节所采用的全局语音模型，本节所提的语音模型中每个字典包含很多说话人的信息，用于描述不同的频谱结构信息。

4.5.2 无监督在线非负矩阵分解

在增强阶段，当前时刻测试信号到达之后，对其进行STFT得到其幅度谱$\boldsymbol{x}_t$，然后进行无监督在线非负矩阵分解。具体而言，将前L帧已经处理过的测试信号幅度谱$\boldsymbol{X}_L = [\boldsymbol{x}_{t-L}, \cdots, \boldsymbol{x}_{t-1}]$和当前帧测试信号幅度谱$\boldsymbol{x}_t$结合作为矩阵分解算法的输入$\boldsymbol{X} = [\boldsymbol{x}_{t-L}, \cdots, \boldsymbol{x}_{t-1}, \boldsymbol{x}_t]$，然后解决如下最优化问题：

$$\boldsymbol{W}_N, \boldsymbol{h}_t = \arg \ \min \ d_{\mathrm{KL}}(\boldsymbol{X} || \boldsymbol{W}\boldsymbol{H}) + \lambda \sum_{g=1}^{G} \log(\epsilon + ||\boldsymbol{h}_{S,t}^{(g)}||_1) \tag{4.22}$$

其中，$\boldsymbol{W} = [\boldsymbol{W}_S \ \ \boldsymbol{W}_N]$，$\boldsymbol{W}_N \in \mathbb{R}_+^{F \times K_N}$ 是噪声字典，$\boldsymbol{H} = [\boldsymbol{H}_S^T \ \ \boldsymbol{H}_N^T]^T$是权重矩阵，$\boldsymbol{H}_S = [\boldsymbol{h}_{S,t-L}, \cdots, \boldsymbol{h}_{S,t-1}, \boldsymbol{h}_{S,t}]$ 是语音模型对应的权重矩阵，$\boldsymbol{h}_{S,t} = [\boldsymbol{h}_{S,t}^{(1)^T}, \cdots, \boldsymbol{h}_{S,t}^{(G)^T}]^T$是当前帧信号对应的语音权重向量。

式(4.22)中第一项为KL散度距离，第二项为log/l_1组稀疏惩罚项，通过对当前帧信号对应的语音权重向量进行组稀疏惩罚来选择若干最匹配的语音字典来描述当前帧混合信号中的语音信息。由于语音模型$\boldsymbol{W}_S$中包含很多语音字典描述语音信号不同的频谱结构信息，因此针对某一帧语音信号只需选择少量的语音字典即可描述。参数λ用来控制组稀疏惩罚项的权重大小。在分解时，前L帧测试信号对应的权重向量$\boldsymbol{h}_{t-L},\cdots,\boldsymbol{h}_{t-1}$在之前的分解中已经估计得到，在当前帧分解中保持固定，只对当前帧信号对应的权重向量$\boldsymbol{h}_t$和噪声字典$\boldsymbol{W}_N$进行估计。具体的分解算法如算法4所示：

Algorithm 4 在线矩阵分解估计噪声矩阵和权重向量

1) 输入：$\boldsymbol{x}_t$, $[\boldsymbol{x}_{t-L},\cdots,\boldsymbol{x}_{t-1}]$, $\boldsymbol{W}_S$, $[\boldsymbol{h}_{t-L},\cdots,\boldsymbol{h}_{t-1}]$
2) 输出：$\boldsymbol{W}_N$, $\boldsymbol{h}_t$
3) 使用非负值随机初始化$\boldsymbol{h}_t$，使用上一帧估计得到的噪声字典初始化$\boldsymbol{W}_N$
4) 令$\boldsymbol{X}=[\boldsymbol{x}_{t-L},\cdots,\boldsymbol{x}_{t-1},\boldsymbol{x}_t]$, $\boldsymbol{H}=[\boldsymbol{h}_{t-L},\cdots,\boldsymbol{h}_{t-1},\boldsymbol{h}_t]$
5) 令$\boldsymbol{W}=[\boldsymbol{W}_S \quad \boldsymbol{W}_N]$ (assume $\mathbf{1}^T\boldsymbol{W}=\mathbf{1}$)
6) **重复**
   $\boldsymbol{V}\leftarrow\boldsymbol{X}./\boldsymbol{W}\boldsymbol{H}$
   $\boldsymbol{h}_t\leftarrow\boldsymbol{h}_t.*(\boldsymbol{W}^T\boldsymbol{v}_t)$
   for $g=1,\cdots,G$
   $\boldsymbol{h}_{S,t}^{(g)}\leftarrow\boldsymbol{h}_{S,t}^{(g)}/\{1+\frac{\lambda}{\epsilon+||\boldsymbol{h}_{S,t}^{(g)}||_1}\}$
   end for
   $\boldsymbol{W}_N\leftarrow\boldsymbol{W}_N.*(\boldsymbol{V}\boldsymbol{H}_N^T)$
   $\boldsymbol{W}_N\leftarrow\boldsymbol{W}_N./(\mathbf{1}\mathbf{1}^T\boldsymbol{W}_N)$ (归一化$\boldsymbol{W}_N$)
7) **直至收敛**

.* and ./ 表示对应元素的相乘和相除运算。

4.5.3　无监督在线语音增强算法描述

本节所提出的无监督在线语音增强算法主要包括以下几个步骤：

1）首先采集大量说话人的语音训练数据，通过k均值聚类和NMF算法得到与说话人无关的语音模型$\boldsymbol{W}_S$。

2）对输入信号$\boldsymbol{X}=[\boldsymbol{x}_{t-L},\cdots,\boldsymbol{x}_{t-1},\boldsymbol{x}_t]$按照算法4进行分解，得到权重向量$\boldsymbol{h}_t$和噪声字典$\boldsymbol{W}_N$。

3）将权重向量$\boldsymbol{h}_t$进行分解$\boldsymbol{h}_t=[\boldsymbol{h}_{S,t}{}^T\ \ \boldsymbol{h}_{N,t}{}^T]^T$，重构估计得到的语音幅度谱$\hat{\boldsymbol{x}}_{S,t}=\boldsymbol{W}_S\boldsymbol{h}_{S,t}$和噪声幅度谱$\hat{\boldsymbol{x}}_{N,t}=\boldsymbol{W}_N\boldsymbol{h}_{N,t}$。

4）通过维纳滤波的形式得到当前帧最终增强后的语音幅度谱：

$$\tilde{\boldsymbol{x}}_{S,t}=\boldsymbol{x}_t\otimes\frac{\hat{\boldsymbol{x}}_{S,t}}{\hat{\boldsymbol{x}}_{S,t}+\hat{\boldsymbol{x}}_{N,t}} \tag{4.23}$$

5) 将$\tilde{\boldsymbol{x}}_{S,t}$结合测试信号的相位信息，并通过iSTFT以及重叠相加算法得到增强后的语音时域信号。

6）更新前L帧处理后的幅度谱（加入最新的当前帧幅度谱$\boldsymbol{x}_t$，去除最早的一帧幅度谱），保留估计得到的噪声字典$\boldsymbol{W}_N$用于进行下一帧增强处理。

4.5.4 实验评价和讨论

4.5.4.1 实验设置

在实验部分，语音数据来自于TIMIT数据集 [134]，噪声数据来自于两个数据集，分别是NOSISEX-92数据集 [135]和 [81]中使用的噪声数据集，共包含24种噪声类型，既有平稳噪声，也有非平稳噪声。从TIMIT训练集中随机选择20个说话人（每个说话人10条语音）作为训练数据得到与说话人无关的语音模型。从TIMIT测试集中随机选择5个说话人（每个说话人1条语音）和24种噪声信号混合得到120条测试信号。测试集信号按照不同的信噪比进行混合（-10，-5，0，5和10dB）。

所有的信号首先通过重采样到16kHz，然后通过STFT得到其频谱，在变换时采用汉宁窗，帧长64ms，帧移16ms。

4.5.4.2 评价指标

为了评价算法的性能，主要采用了如下几种评价指标：

1. SDR、SIR和SAR [131]。这三种评价指标主要是用来衡量算法对于噪声的抑制能力，以及算法自身引入失真的大小。

2. PESQ [132]。该评价指标主要是用来衡量算法增强后的语音质量。

表 4.4: 各种算法的参数取值

Proposed algorithm	$G = 20, K_s = 10, K_N = 20, L = 60$ $\lambda = 1$, MM iter./frame = 10
Online USM	$G = 20, K_s = 10, K_N = 10, L = 60$ $\lambda = 32$, MM iter./frame = 20
Online Speaker-dependent	$K_s = 10, K_N = 5, L = 60$ MM iter./frame = 5

4.5.4.3 算法描述

为了评价算法性能，我们比较了以下几种算法的增强结果：

1. 谱减法（spectral substraction, SS） [4]。

该算法是一种传统的语音增强算法，首先进行噪声估计，然后从混合信号中减去噪声信号进行语音增强。

2. 维纳滤波算法（Wiener filtering, WF） [22]。

该算法是一种基于最优化准则的增强算法，通过估计线性滤波器来进行语音增强。

3. RESET-SIM算法 [136]。

该算法是一种在线增强算法，计算简单，但是算法假设噪声信号要有重复性的谱结构特性。

4. 在线USM算法（Online USM） [83]。

该算法首先通过语音训练数据得到全局语音模型，然后进行在线增强。

5. 说话人依赖的NMF在线算法（Online Speaker-dependent） [81]。

该算法针对测试集中每个说话人训练与其匹配的语音字典，然后进行在线增强处理。

6. 本节提出的算法。

在本节所提算法中，首先通过语音训练数据得到与说话人无关的语音模型，然后进行在线语音增强。

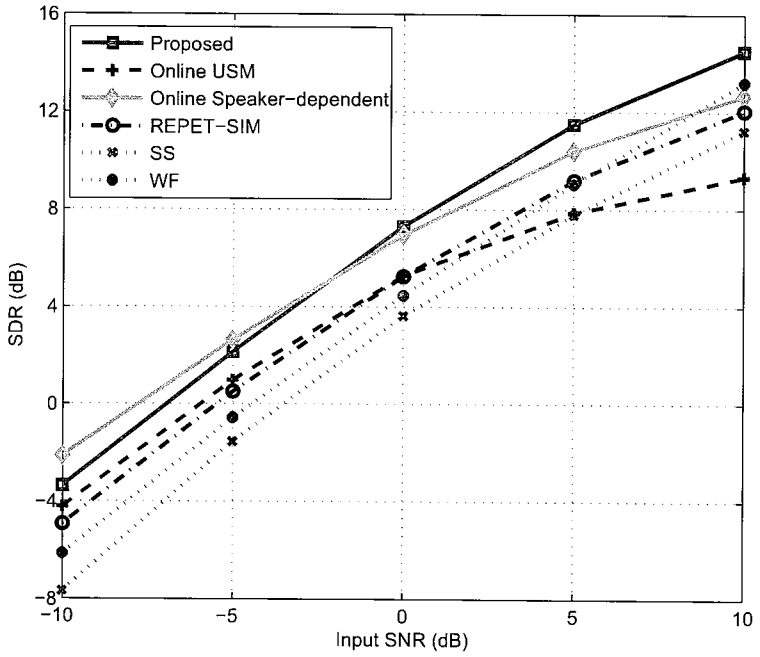

图 4.8: 各种算法在不同输入信噪比时的SDR均值

4.5.4.4 结果分离

表4.4首先给出了不同算法的参数取值，通过选取不同的参数来达到最优的增强结果。图4.8-图4.11给出了各种算法在不同输入信噪比时的SDR、SIR、SAR和PESQ的结果。从结果来看，本节所提算法要明显优于基于USM的在线增强算法，这主要是因为我们采用了不同的语音建模方法。基于USM的算法对于每帧语音信号采用少数几个人的语音字典建模，而本节提出的方法也是采用少量几个语音字典建模，但是每个字典都包含了所有说话人的信息，所以能够对语音信号更好地建模。此外，从结果来看我们的算法在某些情况下甚至优于基于说话人依赖的NMF算法。一个可能的原因是基于说话人依赖的NMF算法虽然能够事先得到该说话人的训练数据，但是由于数据量过少，从而导致无法很好地描述语音信号，而我们的算法采用大量说话人的训练数据，反而能够更好地描述未见过的说话人语音信号。

本节提出的算法也明显由于传统的语音增强算法，如谱减法和维纳滤波算法。这主要是因为这两种算法都假设噪声是相对平稳的，但是在测试集中有很

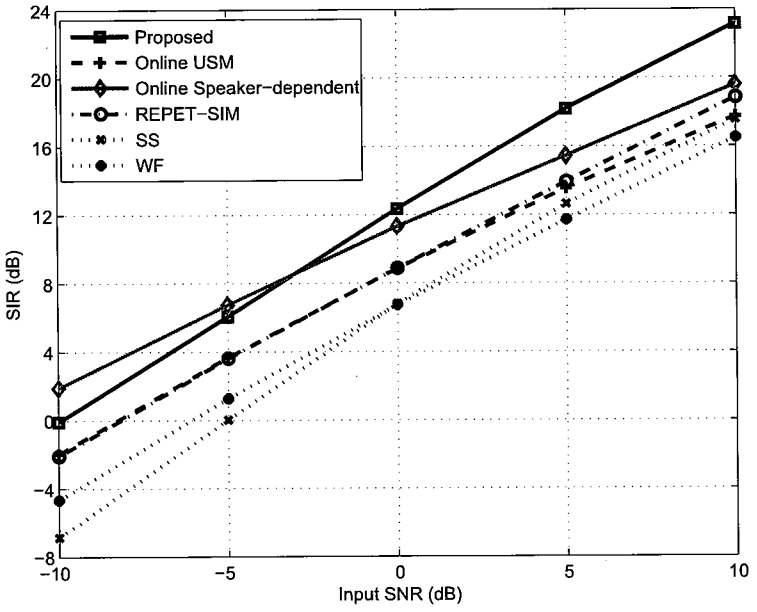

图 4.9: 各种算法在不同输入信噪比时的SIR均值

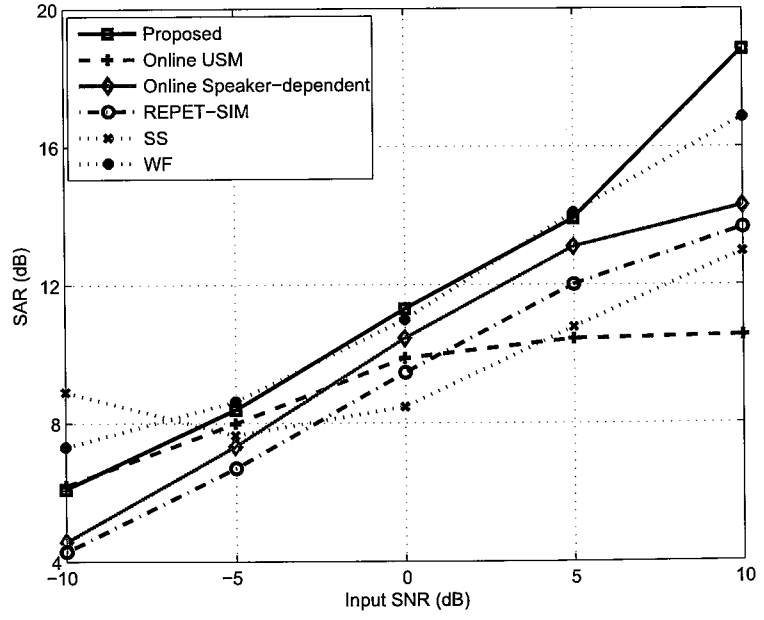

图 4.10: 各种算法在不同输入信噪比时的SAR均值

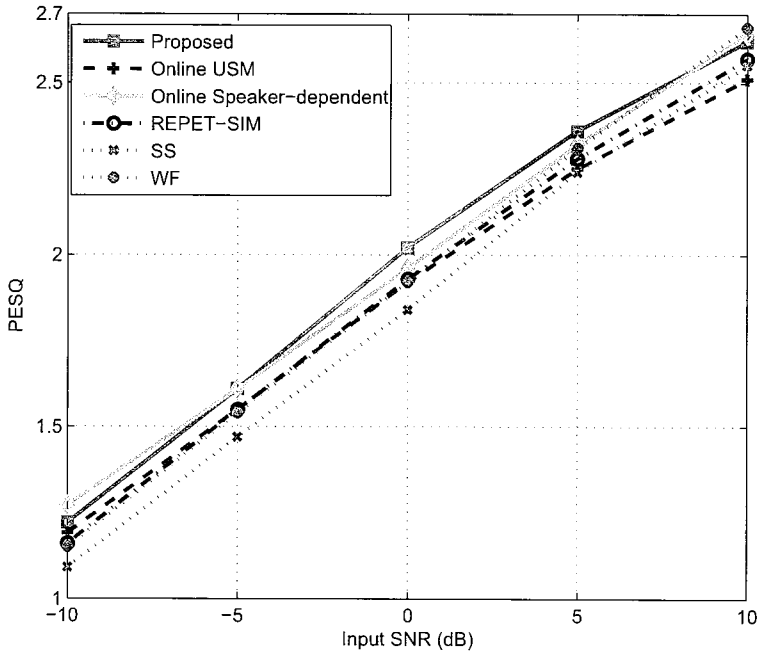

图 4.11: 各种算法在不同输入信噪比时的PESQ均值

多噪声是非平稳的，导致增强效果不佳。而所提算法无需平稳噪声的假设，对非平稳噪声也有较好的抑制能力。

4.5.5 小结

在本节中，我们提出了一种无监督在线语音增强算法。首先通过大量说话人的训练数据采用NMF和k均值聚类算法得到与说话人无关的语音模型，然后通过在线无监督矩阵分解算法逐帧估计噪声字典和权重向量，实现在线增强。实验结果表明，本节提出的算法相比基于USM的在线语音增强算法和一些传统的语音增强能够有效提高语音增强效果。

第五章　基于非负矩阵分解和深度神经网络的语音增强算法研究

5.1　引言

在前面的几章中，我们对基于非负矩阵分解的单通道语音增强算法进行了研究。基于NMF的语音增强算法能够利用语音和噪声训练数据中的先验信息，通过矩阵分解的方式学习语音和噪声信号的局部信息，并且没有了平稳噪声的假设，因此相比传统的单通道语音增强算法能获得更好的增强效果。然而，非负矩阵分解是一个浅层的线性模型，难以学习语音和噪声数据中复杂的非线性结构信息。此外，在基于NMF的增强算法中，一个最基本的假设是语音字典和噪声字典张成的子空间不重叠，这样才能有效去除噪声信号。然而在很多实际场景中，语音和噪声子空间经常重叠，导致很难准确估计权重矩阵及有效增强语音信号。

近年来，单通道语音增强被当做一个有监督的学习问题，通过一个监督学习算法训练鉴别模型来估计目标语音的幅度谱或掩蔽值。Wang等人第一次将DNN用于二值分类来进行语音分离，而且分离结果明显优于之前提出的算法 [84, 90]。Healy等人提出了基于DNN的IBM估计算法，并且表明该算法能够有效提高带噪信号的可懂度，特别是对于听力受损的人群 [92]。由于人类发声的机理，语音频谱具有很强的时频结构特性，如浊音的谐波结构等。如果能够在算法中有效利用语音信号的时频结构特性，将很有可能提高增强效果，然而很多基于DNN的增强算法忽略了这种语音频谱结构特性，直接估计目标语音的幅度谱或掩蔽值 [91, 94, 98]。近年来，DNN和NMF已经被结合用于单通道语音增强 [103–105]，其中NMF用于描述语音信号的频谱结构，DNN用于估计NMF中的权重矩阵。实验结果表明，这种结合DNN和NMF的增强算法相比基于DNN的算法和基于NMF的算法能够有效提高语音增强效果。

在本章节中我们将性别信息引入到基于DNN-NMF的语音增强算法中，通过引入新一维的先验信息，即性别信息来进一步提高语音增强效果。首先，在训练阶段，针对男性和女性说话人分别训练一个DNN-NMF模型；在增强阶段，提出了一个基于NMF和组稀疏的性别鉴定算法来判断每段测试信号说话

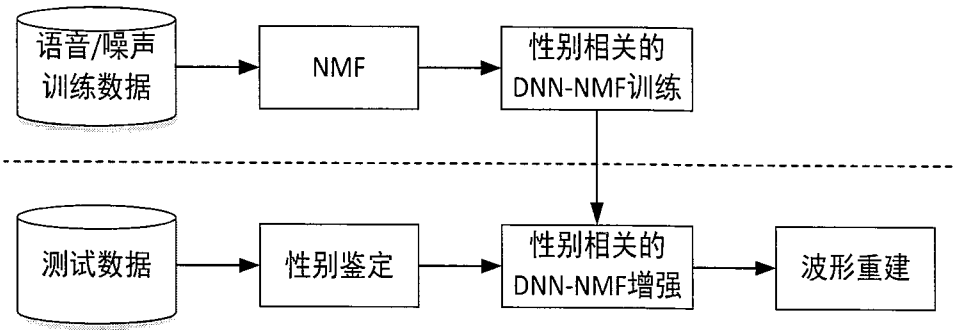

图 5.1: 本章节提出的语音增强算法框图。虚线上方为训练阶段，下方为增强阶段

人的性别，然后选择对应的DNN-NMF模型进行语音增强。提出训练与性别相关的DNN-NMF模型的出发点在于，如果针对不同性别分别训练一个DNN-NMF模型，那么每个模型只需要学习较少的输入变化，因此能够提高模型的学习能力，并且不需要太多参数。除此之外，相比采用同一个字典来描述所有说话人语音信号，针对不同性别分别训练一个语音字典能够更好地刻画对应性别的语音频谱结构，而且字典具有更强的鉴别性。

5.2 融合性别信息的基于DNN-NMF的语音增强算法

本节提出的语音增强系统框图如图 5.1 所示。算法主要分为两个部分：在训练阶段，通过对训练语音中的男性和女性说话人数据分别进行非负矩阵分解，得到对应的语音字典$\boldsymbol{W}_S^{(m)}$和$\boldsymbol{W}_S^{(f)}$，同时对噪声训练数据进行分解得到噪声字典$\boldsymbol{W}_N$。然后对不同性别的语音训练数据和噪声训练数据分别训练与性别相关的DNN-NMF模型；在增强阶段，我们提出了一种基于NMF和组稀疏惩罚项的性别鉴定算法，首先利用该算法判断测试信号中的说话人性别，然后选取对应性别的DNN-NMF模型进行语音增强。

5.2.1 DNN-NMF模型结构

本节采用文献 [105]提出的深度学习框架作为所提算法的DNN-NMF模型结构。该模型将NMF和DNN融合到一个学习框架中，降低了训练误差，并且能够训练鉴别性较强的网络。DNN-NMF的模型结构如图 5.2所示，包括一个输入

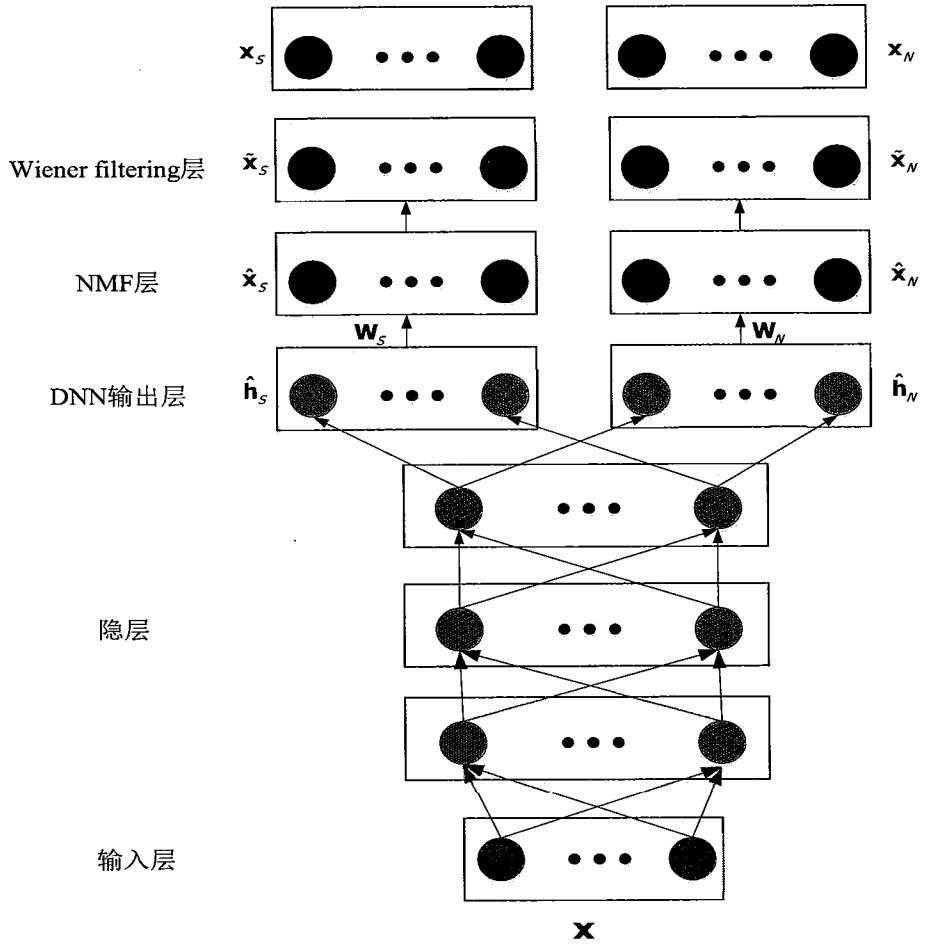

图 5.2: DNN-NMF模型结构图

层，三个隐层，一个输出层，一个NMF层和一个Wiener filtering层。$\boldsymbol{x}$是模型的输入特征，DNN输出层的输出为估计的NMF权重系数，在NMF层估计得到的权重系数和字典矩阵相乘重构出语音和噪声信号的幅度谱，在Wiener filtering层再通过维纳滤波的形式得到最终的语音幅度谱。该模型将DNN和NMF融合到一个框架下，降低了训练误差 [105]。对于NMF层和Wiener filtering层，它们的结构是固定的，不需要训练层与层之间的权重参数，但是这两层的输出被用来计算模型最终的误差函数，然后通过最小化误差函数来训练得到DNN隐层之间的权重参数。

举例而言，如果训练与女性说话人相关的DNN-NMF模型，首先训练得到

字典矩阵$\boldsymbol{W}_S^{(f)}$和$\boldsymbol{W}_N$，然后训练DNN-NMF模型，模型的输入特征为女性说话人和噪声混合信号的幅度谱，DNN输出的权重向量再结合$\boldsymbol{W}_S^{(f)}$和$\boldsymbol{W}_N$重构语音和噪声幅度谱，最后通过维纳滤波的形式得到最终增强后的语音幅度谱。

5.2.2 训练目标

在很多语音增强场景中，由于混合信号中语音和噪声信号具有不同的频谱结构，在增强算法中同时对语音和噪声建模能够有效提高语音增强的效果 [94]。除此之外，在目标函数中加入鉴别性惩罚函数能够在尽可能保留语音成分的同时进一步抑制噪声信号，从而进一步提高增强效果。给定女性说话人对应DNN-NMF模型中DNN的输出$\hat{\boldsymbol{h}}_S$和$\hat{\boldsymbol{h}}_N$，其NMF层的输出为：

$$\hat{\boldsymbol{x}}_S = \boldsymbol{W}_S^{(f)}\hat{\boldsymbol{h}}_S \tag{5.1}$$

$$\hat{\boldsymbol{x}}_N = \boldsymbol{W}_N\hat{\boldsymbol{h}}_N \tag{5.2}$$

其中，$\boldsymbol{W}_S^{(f)}$和$\boldsymbol{W}_N$为女性说话人信号和噪声信号对应的字典矩阵，通过对训练数据分解得到。

通过维纳滤波的形式，Wiener filter层的输出为：

$$\tilde{\boldsymbol{x}}_S = \frac{\hat{\boldsymbol{x}}_S}{\hat{\boldsymbol{x}}_S + \hat{\boldsymbol{x}}_N} \otimes \boldsymbol{x} \tag{5.3}$$

$$\tilde{\boldsymbol{x}}_N = \frac{\hat{\boldsymbol{x}}_N}{\hat{\boldsymbol{x}}_S + \hat{\boldsymbol{x}}_N} \otimes \boldsymbol{x} \tag{5.4}$$

其中，$\boldsymbol{x}$是混合信号的幅度谱，式(5.3)和(5.4)中乘除运算皆为逐点相乘和相除。

训练目标函数采用均方误差准则，定义为：

$$\begin{aligned} J = \frac{1}{2}(||\boldsymbol{x}_S - \tilde{\boldsymbol{x}}_S||_2^2 + ||\boldsymbol{x}_N - \tilde{\boldsymbol{x}}_N||_2^2) - \\ \frac{\lambda}{2}(||\boldsymbol{x}_S - \tilde{\boldsymbol{x}}_N||_2^2 + ||\boldsymbol{x}_N - \tilde{\boldsymbol{x}}_S||_2^2) \end{aligned} \tag{5.5}$$

其中，$\boldsymbol{x}_S$和$\boldsymbol{x}_N$是参考语音和噪声幅度谱，λ是权重参数，用来控制鉴别性惩罚项的相对大小，参数大小可以通过实验来设定。将式(5.1)-(5.4)代入(5.5)，得

到最终的目标函数：

$$\begin{aligned} J = \frac{1}{2}(||\boldsymbol{x}_S - \frac{\boldsymbol{W}_S^{(f)}\hat{\boldsymbol{h}}_S}{\boldsymbol{W}_S^{(f)}\hat{\boldsymbol{h}}_S + \boldsymbol{W}_N\hat{\boldsymbol{h}}_N} \otimes \boldsymbol{x}||_2^2 + \\ ||\boldsymbol{x}_N - \frac{\boldsymbol{W}_N\hat{\boldsymbol{h}}_N}{\boldsymbol{W}_S^{(f)}\hat{\boldsymbol{h}}_S + \boldsymbol{W}_N\hat{\boldsymbol{h}}_N} \otimes \boldsymbol{x}||_2^2) \\ -\frac{\lambda}{2}(||\boldsymbol{x}_S - \frac{\boldsymbol{W}_N\hat{\boldsymbol{h}}_N}{\boldsymbol{W}_S^{(f)}\hat{\boldsymbol{h}}_S + \boldsymbol{W}_N\hat{\boldsymbol{h}}_N} \otimes \boldsymbol{x}||_2^2 + \\ ||\boldsymbol{x}_N - \frac{\boldsymbol{W}_S^{(f)}\hat{\boldsymbol{h}}_S}{\boldsymbol{W}_S^{(f)}\hat{\boldsymbol{h}}_S + \boldsymbol{W}_N\hat{\boldsymbol{h}}_N} \otimes \boldsymbol{x}||_2^2) \end{aligned} \tag{5.6}$$

在本节中，我们采用L-BFGS算法 [129]来通过最小化式(5.6)学习得到模型的参数。

5.2.3　性别鉴定算法

根据本节提出的算法方案，在增强阶段，首先需要对每段测试信号确定其说话人的性别。我们提出了一种基于NMF和组稀疏惩罚项的说话人性别鉴定算法。算法主要分为两个阶段：训练阶段，性别鉴定阶段。

在训练阶段，首先将语音训练数据分为男性和女性说话人训练数据，然后分别对不同性别的语音训练数据进行非负矩阵分解得到对应的语音字典$\boldsymbol{W}_S^{(m)}$和$\boldsymbol{W}_S^{(f)}$。

在性别鉴定阶段，对测试信号的幅度谱$\boldsymbol{X}$进行如下无监督非负矩阵分解：

$$\hat{\boldsymbol{W}}_N, \boldsymbol{H} = \arg\ \min\ d_{\mathrm{KL}}(\boldsymbol{X}||\boldsymbol{W}\boldsymbol{H}) + \beta\Omega(\boldsymbol{H}_S) \tag{5.7}$$

其中，$\boldsymbol{W} = [\boldsymbol{W}_S \quad \hat{\boldsymbol{W}}_N]$，$\boldsymbol{W}_S = [\boldsymbol{W}_S^{(f)} \quad \boldsymbol{W}_S^{(m)}]$是语音字典，$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_S \\ \hat{\boldsymbol{H}}_N \end{bmatrix}$，$\boldsymbol{H}_S = [\boldsymbol{H}_S^{(f)^T}, \boldsymbol{H}_S^{(m)^T}]^T$是语音字典对应的权重矩阵。噪声字典$\hat{\boldsymbol{W}}_N$和权重矩阵$\boldsymbol{H}$通过测试信号分解得到，这样就不需要噪声的训练数据，扩大了算法的适用范围。目标函数的第一项为KL散度，第二项为组稀疏惩罚项$\log/l_1$函数，其定义为：

$$\Omega(\boldsymbol{H}_S) = \sum_i \log(\epsilon + ||\boldsymbol{H}_S^{(i)}||_1) \quad i = f, m \tag{5.8}$$

通过对语音权重矩阵$\boldsymbol{H}_S$进行组稀疏惩罚，找到$\boldsymbol{W}_S$中的$\boldsymbol{W}_S^{(f)}$或者$\boldsymbol{W}_S^{(m)}$来描述测试信号中的说话人语音，β是权重参数。如果说话人为男性，那么稀疏惩罚函数主要选择$\boldsymbol{W}_S^{(m)}$来描述该说话人语音信号，而尽可能抑制另一个字典$\boldsymbol{W}_S^{(f)}$，反之亦然。字典的选择主要通过对应的权重矩阵得到。

在对式(5.7)完成分解之后得到权重矩阵$\boldsymbol{H}_S^{(m)}$和$\boldsymbol{H}_S^{(f)}$，通过比较权重的大小来判断说话人的性别。如果$||\boldsymbol{H}_S^{(m)}||_1 \geq ||\boldsymbol{H}_S^{(f)}||_1$，则说明算法主要采用$\boldsymbol{W}_S^{(m)}$来描述该说话人，因此将该段语音说话人判断为男性，反之则判断为女性。

5.2.4 本节提出的语音增强算法步骤

步骤1)、训练阶段，对训练语音中的男性、女性说话人训练数据，以及噪声训练数据分别进行非负矩阵分解，得到语音字典$\boldsymbol{W}_S^{(m)}$，$\boldsymbol{W}_S^{(f)}$和噪声字典$\boldsymbol{W}_N$。

步骤2)、训练阶段，利用男性说话人训练数据，噪声训练数据，$\boldsymbol{W}_S^{(m)}$和$\boldsymbol{W}_N$，训练与男性说话人相关的DNN-NMF模型；同理，训练与女性说话人相关的DNN-NMF模型。

步骤3)、在增强阶段，首先通过性别鉴定算法判断测试信号中的说话人性别；然后选取对应性别的DNN-NMF模型进行语音增强。

5.2.5 实验评价和讨论

5.2.5.1 实验设置

在实验中，语音数据来自于TIMIT数据集 [134]；噪声数据来自于NOISEX-92数据集 [135]，共包含15种噪声类型。所有的信号首先降采样到16kHz，然后通过STFT到时频域处理，在变换时采用汉宁窗，帧长64ms，帧移16ms。

从TIMIT训练集中随机选择300个说话人，然后从每个说话人中随机选择2句话，共600句话作为语音训练集。从NOISEX-92数据集中随机选择10种噪声作为噪声训练集。将语音训练集和噪声训练集按照不同的信噪比（-6, -3, 0, 3和6dB）进行混合，共得到30000句混合语音训练数据（600*10*5=30000）。从TIMIT训练集中随机选择50个说话人，然后从每个说话人中随机选择1句话，共50句话作为语音开发集，将语音开发集和上述10 种噪声按照不同信噪比（-6, -3, 0, 3, 6dB）进行混合，共得到2500句混合语音开发集数据（50*10*5=2500）。

训练集和开发集使用的语音和噪声数据不重叠。开发集数据主要是为了确定DNN模型的参数。

从TIMIT测试集中随机选取100个说话人，然后从每个说话人中随机选取3句话，共300句话作为测试语音。将这些语音数据和NOISEX-92中的全部15种噪声按照不同的信噪比（-6, -3, 0, 3, 6dB）进行混合得到测试集数据。为了测试算法对不同噪声类型的泛化能力，在测试集中新加入了5种训练集中未出现的噪声类型。

5.2.5.2　评价指标

为了评价算法的增强结果，我们使用了如下几种客观评价指标：

1. SDR、SIR和SAR [131]。这三种评价指标主要是用来衡量算法对于噪声的抑制能力，以及算法自身引入失真的大小。

2. PESQ [132]。该评价指标主要是用来衡量算法增强后的语音质量。

5.2.5.3　算法描述

在实验中，我们比较了以下几种算法的增强结果：

1. 标准的基于NMF的语音增强算法 [57]。

该算法采用KL散度作为目标函数以及乘法迭代准则作为分解算法。语音字典和噪声字典分别包含50个基向量。

2. 基于DNN的语音增强算法 [98]。

该算法中DNN的输出为估计得到的语音幅度谱，这是基于谱映射的增强算法。

3. 基于DNN-NMF的语音增强算法(DNN-NMF-1) [105]。

该算法将DNN和NMF融合到一个框架中，在DNN框架的基础上加入了一个NMF层和一个Wiener filtering层。语音字典和噪声字典分别包含100和300个基向量。

4. 基于DNN-NMF的语音增强算法（DNN-NMF-2） [104]。

在该算法中，DNN被用来估计NMF的权重系数，然后这些权重系数结合语音和噪声字典重构增强后的语音信号。语音字典和噪声字典分别包含100和300个基向量。

5. 本节提出的语音增强算法。

该算法在上述算法的基础上，将说话人性别信息引入进来，对于不同性别分别训练了一个DNN-NMF模型。权重参数$\lambda = 0.5$。男性和女性说话人字典分别包含50个基向量，噪声字典包含300个基向量。

在上述这些算法中，所有的DNN模型都包含三层隐层，每层隐层有1024个节点，隐层中非线性函数采用ReLU激活函数，模型输出层采用线性函数。在进行DNN训练时，模型参数首先随机初始化，然后采用L-BFGS算法进行更新，迭代次数设置为500。DNN模型的输入特征为混合信号的幅度谱，为了进一步描述语音信号的时间相关性，将连续的5帧信号（前两帧，当前帧，后两帧）合并作为输入特征，模型的输出为当前帧的目标值。

5.2.6 结果分析

在结果中我们首先评估了性别鉴定算法的性能。稀疏参数β被设置为256，这样可以获得更加准确的性别鉴定结果。针对测试集中不同信噪比的测试信号，算法鉴定的结果如表5.1所示：

表 5.1: 测试集中不同输入信噪比时的性别鉴定结果

SNR	-6dB	-3dB	0dB	3dB	6dB
准确率	96.4%	97.8%	98.3%	99.6%	98.9%

从结果来看，性别鉴定算法的准确性随着测试信号信噪比的下降而下降。这也符合情理，随着输入信噪比的下降，噪声能量变得更大，导致鉴定说话人性别变得更加困难。具体而言，该算法在低信噪比情况下（-6dB）依然能够获得不错的准确率（96.4%）；其次在信噪比较高时（3dB，6dB），该算法能够获得非常准确的性别鉴定结果（99.6%，98.9%）。

其次我们比较了几种算法在不同输入信噪比时的SDR、SIR、SAR以及PESQ的结果，如图5.3-图5.6所示。从结果来看，对于所有测试情况而言，本节所提算法相比其它算法在SDR和PESQ指标上都有一定提升，这说明无论是对于噪声抑制量还是语音质量该算法都有所提高。具体而言，在大部分情况下该算法能够在保持SAR的情况下显著提高SIR，说明算法可以在保持语音失真不增加的同时，能够有效抑制噪声。这主要是因为所提算法针对不同性别说

表 5.2:　交叉测试时噪声匹配情况下的增强结果

SNR(dB)	SDR	SIR	SAR	PESQ
-6	5.18	7.76	4.36	1.92
-3	6.41	9.39	6.16	2.11
0	7.66	10.82	7.83	2.30
3	8.87	12.18	9.41	2.47
6	10.05	13.43	10.90	2.62
均值	7.63	10.71	7.73	2.28

话人分别训练了不同的DNN-NMF模型，从而使得模型能够学习较少的输入变化，使得模型易于训练，具有更强的鉴别性，更容易区分语音和噪声。然而，从结果来看，本节提出的算法相比文献 [105]的算法在SAR指标上略有下降，这是因为文献 [105]的算法采用了一个较大的字典来描述所有说话人的语音，这样能够对语音有较好的描述能力，导致失真较小，因此SAR指标较高，但是这也导致了语音字典的鉴别性下降，对噪声的抑制能力变弱，导致其SIR和SDR指标较低。

最后，我们对本节提出的算法进行了交叉测试实验，将性别匹配的DNN-NMF模型故意对调。如果性别鉴定算法结果为男性，则选用女性对应的DNN-NMF模型进行增强处理，反之亦然。表5.2给出了在噪声匹配情况下的交叉测试结果。相比模型匹配时的结果，在对调对应性别的DNN-NMF模型之后，所提算法的性能明显下降，这也从侧面说明了所提算法的有效性，通过选择性别匹配的DNN-NMF模型，能够有效提高语音增强效果。

5.3　小结

在本章节中我们提出了一种融合性别信息的基于DNN-NMF的单通道语音增强算法。首先在训练阶段针对不同性别的说话人训练对应的DNN-NMF模型。然后在增强阶段，提出了一种性别鉴定算法用于判断测试信号中说话人的性别，接着选择对应的DNN-NMF模型用于语音增强。实验表明，该算法可以在保持较高语音质量的同时，对噪声有较好的抑制能力。这主要是因为本文算法引入了新一维的先验信息，所以相比其它一些比较算法在增强结果上有进一步提高。

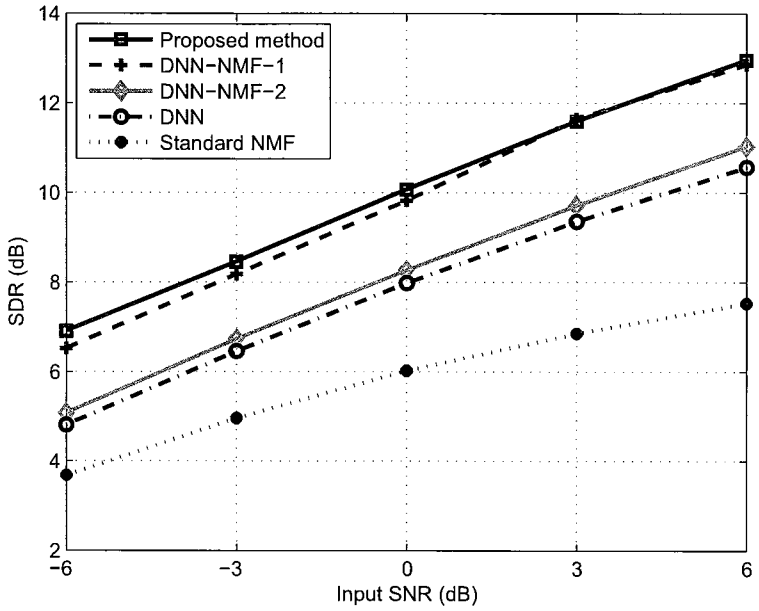

图 5.3: 各种算法在不同输入信噪比时的SDR均值

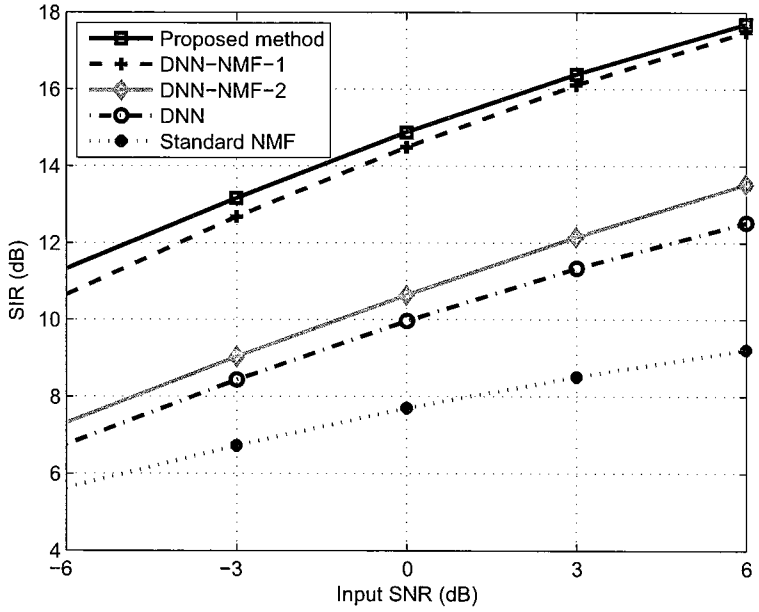

图 5.4: 各种算法在不同输入信噪比时的SIR均值

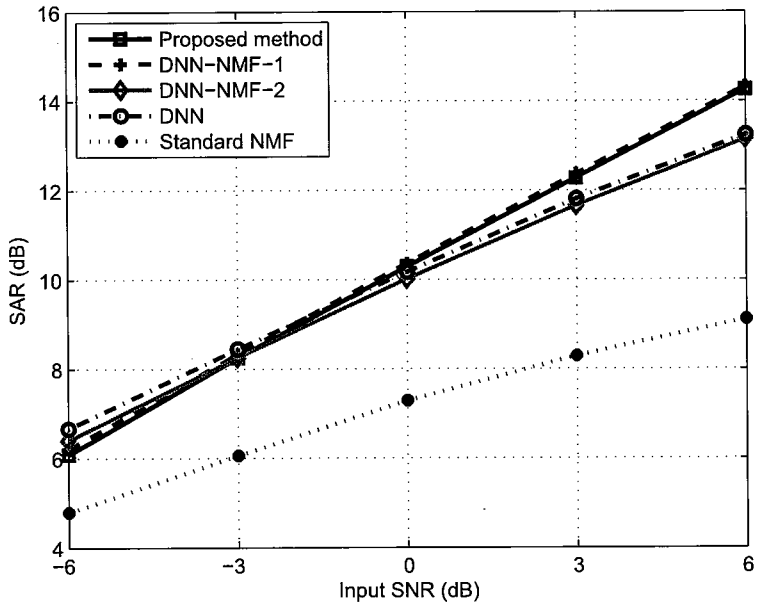

图 5.5: 各种算法在不同输入信噪比时的SAR均值

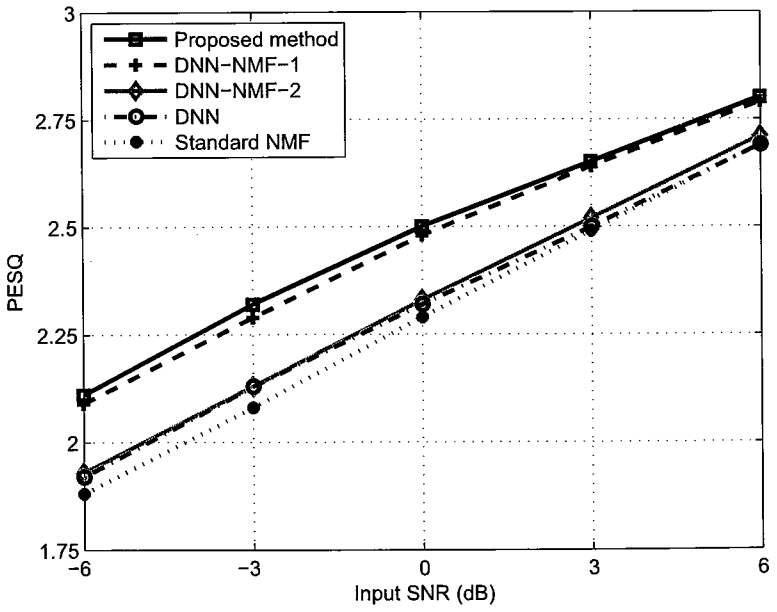

图 5.6: 各种算法在不同输入信噪比时的PESQ均值

第六章　混响环境下的语音增强算法研究

6.1　引言

在前面一章中，我们将非负矩阵分解和深度神经网络结合，并且将性别信息引入进来，提出了一种融合性别信息的基于DNN-NMF的语音增强算法。实验结果表明，所提算法相比一些基于NMF的算法和一些基于DNN的语音增强算法在客观评价指标上有一定提高。然而，这些算法都是扩散场情况下的语音增强，并没有考虑混响的影响。在很多实际场景中，语音信号不仅受到加性噪声的影响，同时也会受到封闭空间内混响的影响。

当语音信号在一个封闭空间内被较远距离的传声器采集时，因为周围墙面和空间内物体的反射，采集到的信号是许多经过延迟和衰减后的语音信号的叠加。一个典型的房间声学冲激响应如图 6.1 所示，主要包括三个部分：直达声，早期反射声和晚期反射声 [137]。房间混响会导致语音信号时间和频谱上的拖尾效应，会同时改变语音信号的包络和精细结构，因此会严重影响语音信号的质量和可懂度 [138]。由于噪声信号和房间声学冲激响应都无法先验知道，因此混响环境下的语音增强是一个更具有挑战性的问题。

近年来，一些基于监督学习的方法被提出来解决混响环境下的语音增强问题，并取得了不错的效果。Han等人提出了基于谱映射的算法同时进行噪声和混响抑制，通过训练一个DNN模型来学习混合信号到纯净语音信号的映射函数 [100, 101]。Jin和Wang通过训练一个多层感知机来估计IBM进行降噪和去混响 [139]。Zhao等人提出了混响情况下基于DNN的IRM估计算法来进行语音增强，主观听觉实验表明能够有效提高听力受损人群的语音可懂度。但是在计算IRM时，该算法将混响语音信号作为目标信号，因此只是对加性噪声进行了抑制 [140]。Bradley等人对混响环境下人耳感知的研究表明，虽然晚期混响会严重影响语音的可懂度，早期混响反而会带来语音可懂度的提高 [125]。Roman和Woodruff根据上述研究结果将IBM定义扩展到混响环境下，将目标语音的直达声和早期混响部分作为期望信号，晚期混合和噪声信号作为残余信号，然后将计算得到的IBM作用于原始混合信号得到增强后的语音信号 [141]。结果表明，经过IBM掩蔽后得到的语音信号相比混合信号具有更高的可懂度。

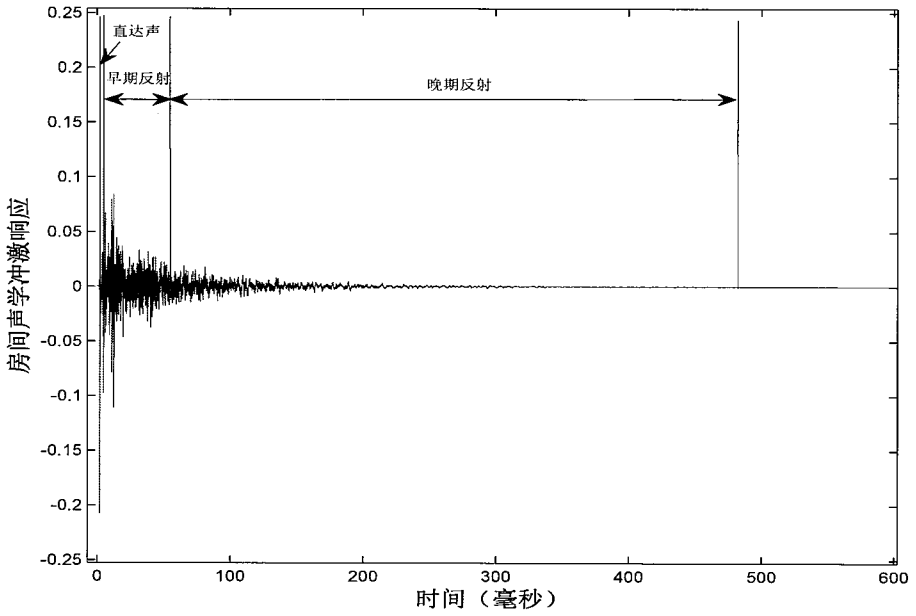

图 6.1: 办公室环境下测量的房间声学冲激响应，混响时间$RT_{60} = 480\text{ms}$

Roman在计算IBM时，采用50ms的阈值来区分早期混响和晚期混响。Li等人又进一步研究了四种不同的早期混响和晚期混响切分时间对于IBM增强后的语音可懂度的影响 [142]。主观测听实验结果表明，不同的早晚期混响切分都能够带来语音可懂度的提高，然而不同的切分时间会造成语音可懂度的不同，而且在某个切分时间区间内，可懂度的变化不大，说明早晚期混响的切分时间并不是某个确定值，而是在某个区间内更为合理。

一个最近的研究表明，在扩散场环境下基于DNN的语音增强算法中，采用IRM作为估计目标相比IBM作为估计目标能够提高语音质量 [91]。因此，在本节中我们将IRM扩展到混响环境下，并且采用DNN模型估计扩展的IRM，然后将估计得到的IRM用于混响带噪的原始信号进行语音增强。同时我们也进行了一系列实验来对噪声和混响对于增强结果的影响进行了研究。

6.2 混响环境下基于DNN的语音增强算法

本节提出的算法框图如图 6.2 所示。算法主要包括两个阶段。在训练阶段，首先从混响带噪的训练数据中提取特征，然后训练一个DNN模型估计IRM。在增强阶段，从测试信号中提取特征，然后送入已经训练好的DNN模型中估

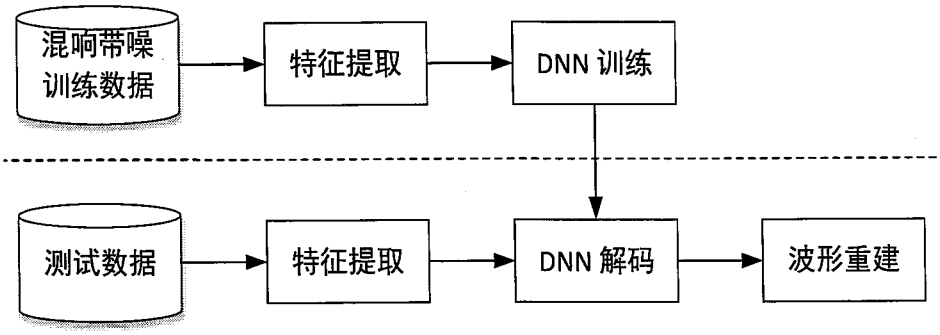

图 6.2: 基于DNN的语音增强算法框图。虚线上方为训练阶段，下方为增强阶段

计IRM，然后用于语音增强。具体的算法实现如下所示。

6.2.1　混响环境下的IRM定义

与IBM在混响情况下的计算方式类似 [141]，本节提出的IRM也是在Cochleagram域进行计算。具体而言，首先通过64通道的Gammatone滤波器组，其中心频率从50到8000Hz等间距分布，然后每个滤波器的输出通过帧长20ms，帧移10ms分解得到一系列二维时频单元。假设$D(k,l)$和$R(k,l)$分别是目标信号和残余信号在第k个频带和第l帧的能量，则IRM定义为：

$$\mathrm{IRM}(k,l)=\sqrt{\frac{D(k,l)}{D(k,l)+R(k,l)}} \tag{6.1}$$

其中，$D(k,l)$包括目标语音的直达声和早期混响部分，通过将语音信号卷积房间冲激响应的直达声和早期反射声得到。残余信号$R(k,l)$通过从混合语音中减去$D(k,l)$得到，主要包括了噪声信号以及语音信号的晚期混响部分。在计算IRM时，我们采用50ms来切分早期混响和晚期混响部分。

6.2.2　特征提取

在采用DNN估计IRM时，我们从混合信号中提取了多种特征用于训练DNN模型。具体而言，主要包括15维的AMS（Amplitude modulation spectrogram）特征，13维的RASTA-PLP（Relative spectral transform PLP）特征，31维的梅尔倒谱系数（Mel-frequency cepstral coefficient, MFCC），64维的GF

（Gammatone feature）特征，以及它们的一阶差分。这样，对于每帧混合信号，共包含246维特征，这些特征具体的描述可以参考文献 [143]。

6.2.3 DNN训练

在训练阶段，通过训练数据得到一个DNN模型用于估计IRM。具体而言，训练的DNN模型包括4个隐层，每个隐层有1024个节点，隐层非线性函数为ReLU函数。由于模型的输出IRM范围是[0,1]，因此输出层函数采用Sigmoid函数。DNN采用反向传播算法进行训练以及采用均方误差准则作为损失函数。Minibatch大小设置为1024，而且为了防止过拟合，采用dropout技术，dropout比例为0.2。同时，在训练模型时采用Adagrad算法 [144]来调整学习率。最大的算法迭代次数设置为300。为了描述语音信号的时间连续性，同时采用连续5帧信号（前2帧，当前帧，后2帧）的特征作为DNN的输入，在输出层同时估计对应5帧的IRM。最后，每帧混合信号对应的IRM由5次估计的IRM平均得到。

6.2.4 波形重建

在增强阶段，首先提取测试信号的特征，然后送入已经训练好的DNN模型中估计IRM，然后将估计得到的IRM再结合测试信号并通过合成步骤来重构出增强后的语音时域信号 [86]。

6.2.5 实验评价和分析

6.2.5.1 实验设置

在实验中，语音数据来自于TIMIT数据集 [134]。从TIMIT训练集中随机选择600条语音作为训练集语料，另外从TIMIT训练集中随机选择100条语音作为开发集语料（和训练集语料不重叠）。

为了生成混响仿真数据，我们使用了Jeub等人提供的房间冲激响应（Room Impulse Responses, RIRs） [145]。这些RIRs是从4个房间中按照不同的音源和传声器距离实录得到，采样率为48kHz。我们从meeting和lecture这两个房间对应的RIRs中各选择两个RIRs（分别对应音源和传声器距离最大和最小），共4个RIRs用来生成训练集混响数据。另外，从meeting和lecture 这两个房间对应的RIRs中各选择一个RIR（和训练集中使用的RIRs不重叠）用于生成

开发集混响数据。因此，在训练集中共包含600×2(Rooms)×2(RIRs) = 2400条混响语音，在开发集中共包含100×2(Rooms)×1(RIR) = 200条混响语音。

在训练集和开发集中，我们使用了babble噪声和speech-shpaed噪声（speech-shaped noise, SSN）生成带噪数据。其中babble噪声是非平稳噪声，而SSN是平稳噪声，将白噪声通过语音数据的平均谱得到。两种噪声数据时长大约4分钟，前3分钟被用来生成训练集和开发集数据，后1分钟被用于生成测试集数据。在进行混合时，首先将语音和噪声信号重采样到48kHz，然后分别和RIRs进行卷积得到带混响的数据，再降采样到16kHz。将混响语音和混响噪声数据按照不同的信噪比（-3，0，3dB）进行混合，得到混响带噪数据。最后，在训练集中共包含2400×3(SNRs)×2(noises)=14400条混响带噪数据，在开发集中共包含200×3(SNRs)×2(noises)=1200 条混响带噪数据。

为了生成测试集信号，从TIMIT测试集中随机选择50条语音作为测试集语料，从meeting，office和lecture这三个房间中各选择一个RIR（和训练集，开发集中RIRs不重叠），共3个RIRs用来生成混响数据。同时使用了3种噪声（babble，SSN和factory）生成带噪数据。所有的语音和噪声信号首先卷积RIRs，然后再按照不同信噪比（-3，0，3dB）进行混合得到混响带噪信号。最后，在测试集中共包含50×3(RIRs)×3(SNRs)×3(noises)=1350条混合带噪数据。

6.2.5.2　评价指标

为了评价算法的增强结果，我们使用了如下客观评价指标：

1. Short-time objective intelligibility (STOI) [146]。

该评价指标主要用于衡量算法增强后的语音可懂度。

2. PESQ [132]。

该评价指标主要是用于衡量算法增强后的语音质量。

6.2.5.3　算法描述

在实验中，我们同时比较了进行IBM估计的增强结果。在估计算法中，将IBM作为估计目标，该IBM是定义在混响情况下 [141]。而且采用DNN估计IBM，具体的估计算法和本节提出的算法类似，除了是采用IBM作为估计目标。

6.2.5.4 结果分析

首先我们比较了几种算法在噪声和混响匹配情况下的增强结果。图 6.3-图 6.4 分别给出了原始混合信号（mixture），经过估计的IRM（eIRM）增强后的语音信号以及经过估计的IBM（eIBM）增强后的语音信号在SSN噪声情况下对应两种混响情况（meeting，lecture）的STOI和PESQ结果。STOI和PESQ的值通过对50条测试语音的结果平均得到。从结果可以看出，当使用eIBM和eIRM用于语音增强时，相比原始混合信号能够有效提高语音可懂度和语音质量。而且比较不同混响环境下的PESQ和STOI可以看出，语音质量和可懂度随着混响时间的增加而下降。除此之外，比较eIBM和eIRM的结果可以看出，虽然eIBM和eIRM得到的STOI大致相同，eIRM能够得到更高的PESQ结果，说明eIRM能够带来更多的语音质量提高，该结果也和文献 [91]的结论是一致的。图 6.5-图 6.6 给出了不同算法在babble噪声情况下对应两种混响环境（meeting，lecture）的STOI和PESQ结果。该结果和图 6.3-图 6.4 的结果类似，eIRM和eIBM相比原始信号都能有效提高STOI和PESQ，以及随着混响时间变大，STOI和PESQ指标下降。

图 6.3(a)和图 6.5(a)是不同算法在meeting混响环境下对应两种噪声（SSN，babble noise）时STOI的结果。从结果可以看出，对于原始混合信号而言，SSN噪声情况下的STOI相比babble噪声对应的STOI指标稍低一些。经过增强之后，SSN对应的eIBM和eIRM相比babble噪声情况能够获得更多的STOI增益，这是因为SSN比较平稳，DNN易于对其进行描述。图 6.3(b)和图 6.5(b)的STOI也显示了相似的结果。

其次我们比较了不同算法在不匹配情况下的结果。图 6.7-图 6.8 给出了不同算法在两种混响情况下（meeting，lecture）对应一个新的噪声类型（factory）的结果。图 6.9-图 6.10 给出了不同算法在两种噪声情况下（SSN，babble noise）对应一个新的混响环境（office room）的结果。从结果可以看出，在新的噪声类型情况下，eIBM和eIRM只能带来微弱的STOI和PESQ提高，而在新的混响情况下，eIBM和eIRM能带来较大的STOI和PESQ提高。这主要是因为测试集中使用的office room的混响时间不够大，因此训练的DNN能够对其较好地建模，而factory噪声频谱和训练集中的两种噪声频谱差异较大，训练的DNN无法对factory噪声准确建模，导致DNN估计不佳。

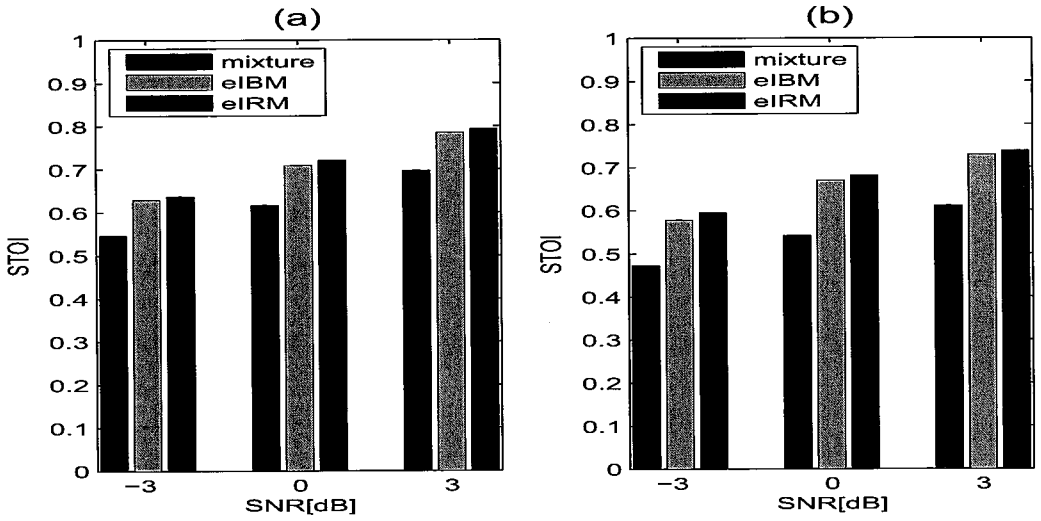

图 6.3: 不同算法在SSN噪声环境下的STOI均值，同时考虑两种混响环境，（a）：meeting room （b）：lecture room

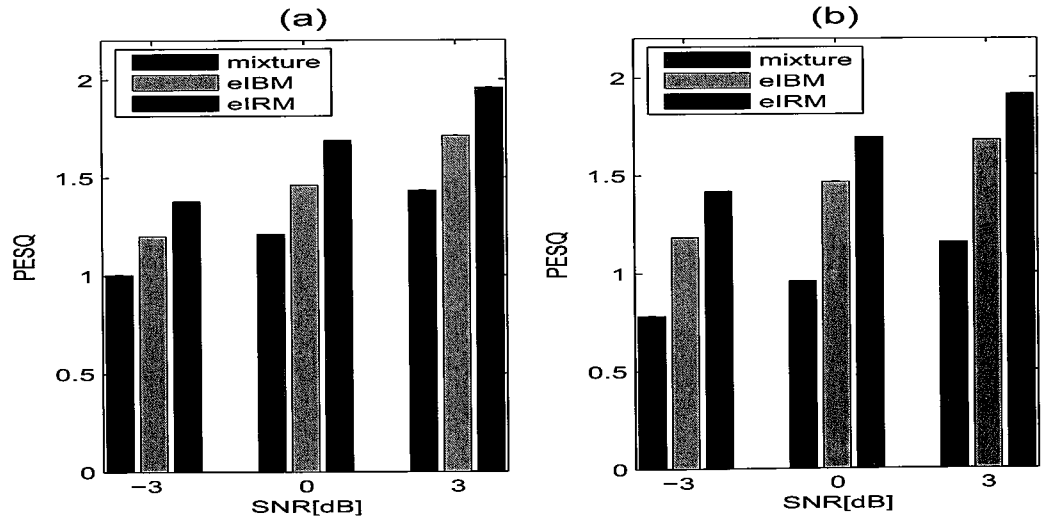

图 6.4: 不同算法在SSN噪声环境下的PESQ均值，同时考虑两种混响环境，（a）：meeting room （b）：lecture room

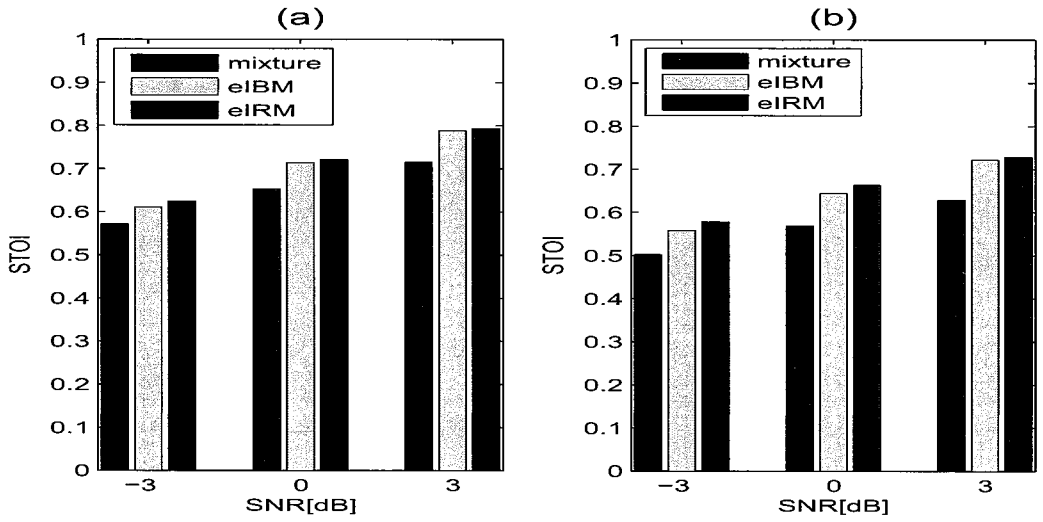

图 6.5: 不同算法在babble噪声环境下的STOI均值，同时考虑两种混响环境，(a)：meeting room (b)：lecture room

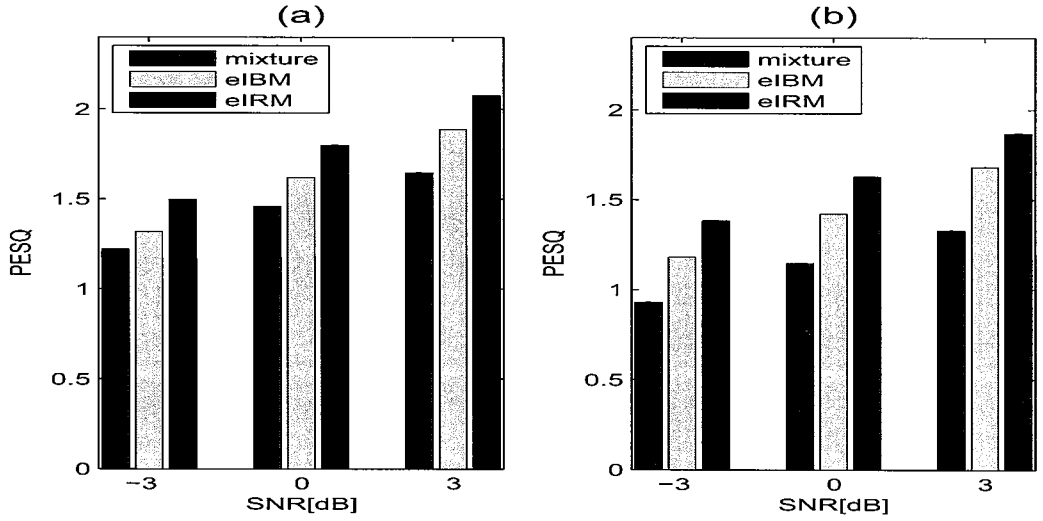

图 6.6: 不同算法在babble噪声环境下的PESQ均值，同时考虑两种混响环境，(a)：meeting room (b)：lecture room

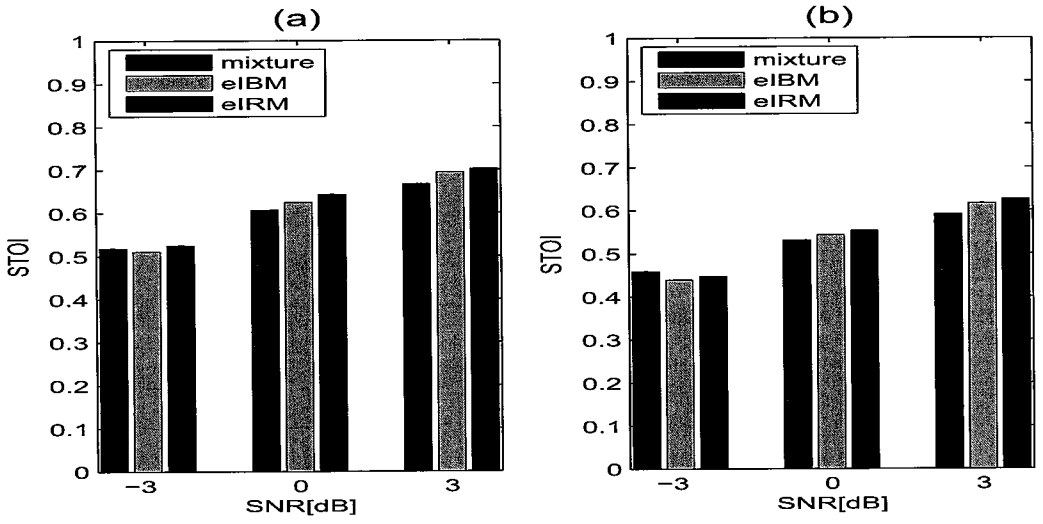

图 6.7: 不同算法在factory噪声环境下的STOI均值，同时考虑两种混响环境，（a）：meeting room （b）：lecture room

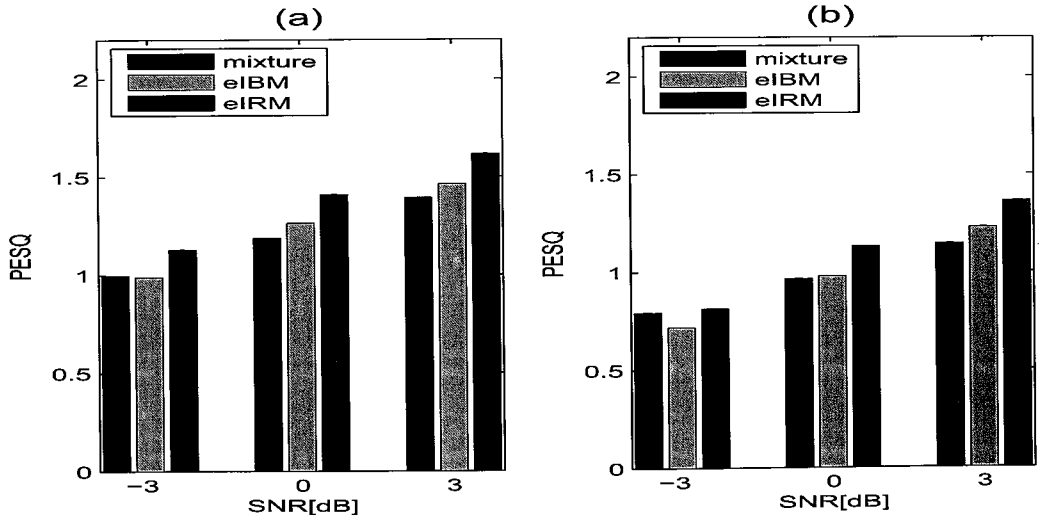

图 6.8: 不同算法在factory噪声环境下的PESQ均值，同时考虑两种混响环境，（a）：meeting room （b）：lecture room

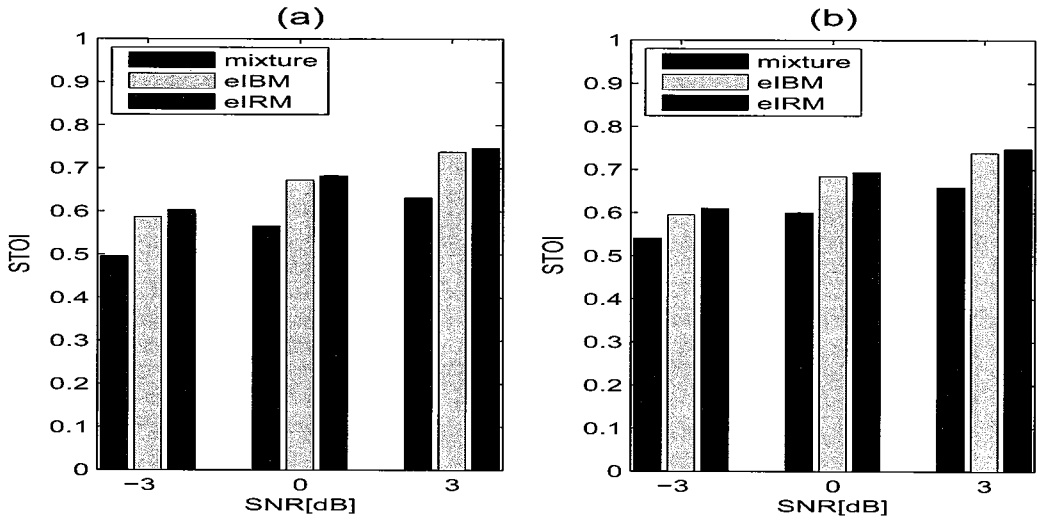

图 6.9: 不同算法在office混响环境下的STOI均值，同时考虑两种噪声环境，(a)：SSN (b)：babble noise

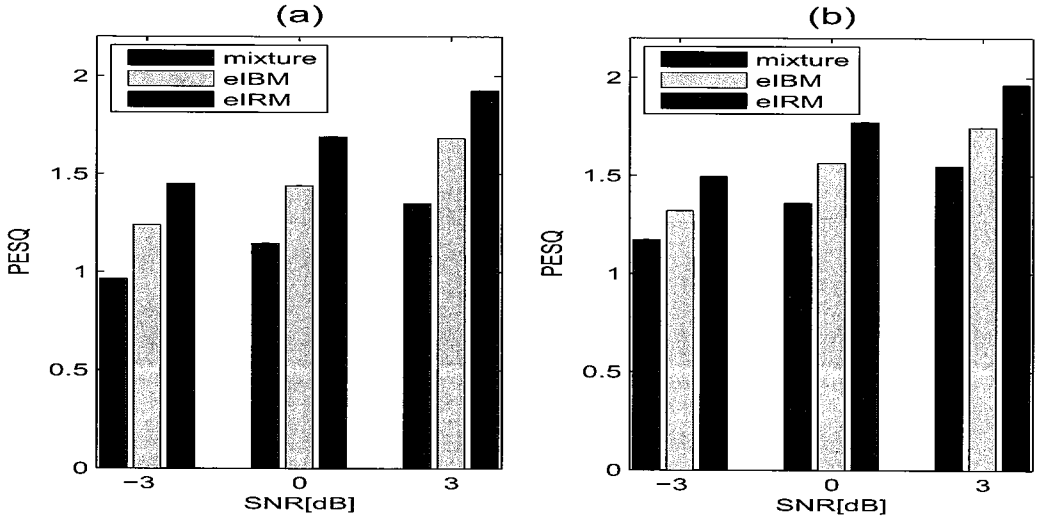

图 6.10: 不同算法在office混响环境下的PESQ均值，同时考虑两种噪声环境，(a)：SSN (b)：babble noise

6.3　小结

在本章节中，我们将IRM扩展到有混响的情况下，提出了一种混响环境下的IRM定义，将目标语音的直达声和早期混响作为目标成分，其余部分作为需要抑制的成分。而且采用DNN模型估计扩展的IRM进行混响环境下的语音增强。实验结果表明，经过估计的IRM增强后的语音信号相比原始混合信号能够有效提高语音质量和可懂度。除此之外，我们也比较了采用IBM和IRM作为训练目标的增强结果，结果显示IRM相比IBM能够进一步提高语音质量。

第七章　总结与展望

随着现代信息技术的发展，电话会议系统、视频通信和VoIP等通信系统不断涌现，语音通信技术得到越来越广泛的应用。另一方面，随着智能可穿戴设备、智能车载和智能家居等领域的兴起，越来越多的消费电子设备具有语音交互功能。此外，随着人口老龄化，听力障碍不仅是一个公共卫生问题，也成为一个社会问题。为了解决听力受损问题，需要借助助听器设备。这些背景使得语音增强技术的研究具有重要意义。

7.1　本文研究内容

在实际的语音增强应用场景中，目标语音容易受到混响以及各种噪声信号的污染，语音通信或交互的体验以及助听器的性能因此而受到影响。本文我们主要对语音增强技术中的噪声抑制算法进行了研究。其次，相比阵列语音增强而言，单通道语音增强技术由于只需要一个传声器、成本较低、没有传声器一致性的要求，对声源位置没有要求以及对混响不太敏感的优势而具有更加广泛的应用前景。然而，由于只有一路混合信号，没有空间信息可以利用，增加了问题的解决难度。本文主要对单通道语音增强算法进行了研究。

传统的单通道语音增强算法一般假设噪声信号是平稳的，然后通过噪声估计算法得到噪声谱，再进行语音增强。然而，在实际场景中很多噪声是非平稳的，导致这类增强算法降噪效果不佳。为了对非平稳噪声有较好的抑制能力，需要引入更多的先验信息。近年来，基于数据训练的语音增强算法被提出来用于语音增强，由于对非平稳噪声有较好的抑制能力而得到了广泛关注和研究，如基于NMF的语音增强算法和基于DNN的语音增强算法。这类算法通过从训练数据中学习一些先验信息，然后再用于语音增强。本文首先对语音信号的时间连续性进行了研究，提出了一种基于NMF和k均值聚类的语音建模方法，通过语音训练数据得到一个状态序列和若干语音字典，能够同时描述语音信号的时间连续性和频谱结构特性。而且，我们将这种语音建模方法和FCRF结合对混合信号的时间动态特性建模，然后用于分离两个说话人的语音信号以及分离语音和噪声信号，相比其它一些算法能够有效提高分离效果。

其次，本文研究了基于NMF的无监督语音增强算法。标准的基于NMF的语音增强算法需要特定说话人或特定噪声类型的训练数据，限制了这类算法在实际场景中的应用。有学者提出了采用全局语音模型的无监督增强算法，通过大量语音训练数据得到一个全局语音模型，然后从测试信号中估计噪声字典，去除了需要匹配训练数据的限制。这种算法很关键的一点是组稀疏惩罚项的应用，为了防止语音模型过拟合，需要对语音模型进行组稀疏惩罚。本文首先提出了基于自适应组稀疏惩罚项的方法，能够自适应选择稀疏参数，增强了算法鲁棒性和噪声抑制能力。其次，本文又提出了基于动态组稀疏惩罚项的方法，通过改进稀疏惩罚项，使得算法能够对于不同的语音帧选择不同的语音字典进行描述，同时为了描述语音连续性，对相邻几帧语音采用同样的字典进行描述，能够同时刻画语音的谱动态变化特性和时间连续性。最后，本文提出了一种与说话人无关的语音模型，并将其用于无监督在线语音增强，能够对输入信号进行逐帧降噪处理，非常适合实际使用。

随后，我们又研究了基于DNN和NMF的语音增强算法。由于NMF能够对语音频谱结构有较好的描述能力，DNN能够学习输入输出之间的非线性关系，DNN和NMF已经被结合用于语音增强。我们在此基础上将说话人性别信息引入进来，提出了一种融合说话人性别信息的增强算法，针对不同性别的说话人分别训练了一个DNN-NMF模型，然后提出了一种性别鉴定算法确定说话人性别，再选择对应的DNN-NMF模型进行语音增强，通过引入更多的先验信息来进一步提高了算法的增强结果。

最后，在很多室内场景中，房间混响也会影响语音通信以及语音交互系统的性能。我们对混响环境下的语音增强算法进行了研究，提出了一种混响环境下的理想浮值掩蔽定义，将目标语音的直达声和早期混响作为目标成分。同时我们采用DNN去估计新定义的IRM，然后将估计得到的IRM用于原始信号进行语音增强。结果表明经过估计的IRM增强后的语音信号相比原始信号能够有效提高语音质量和可懂度。

7.2 下一步研究工作

下一步研究工作主要包括以下几点：

1. 在第三章中我们提出了一种语音建模方法，然后结合FCRF进行分离两个说话人的语音信号。但是这种语音分离算法需要特定说话人的训练数据，限

制了其在实际场景中的使用。因此，研究不需要特定说话人训练数据的语音分离算法是下一步的研究重点。而且，该算法计算复杂度较高，需要研究具有较小计算复杂度的算法，以方便实际应用。

2. 在第四章中，我们研究了基于NMF的无监督语音增强算法。算法首先通过大量的语音数据得到全局语音模型，然而全局语音模型的完备与否直接关系到增强算法的结果，如何获得更加完备的语音模型还有待进一步研究。

3. 在第六章中，我们研究了混响环境下的语音增强算法，将IRM的定义扩展到混响情况下，把目标语音的直达声和早期混响作为期望信号。实验结果表明，在噪声不匹配情况下，DNN估计效果不佳。其次，我们在切分早晚期混响时，采用经验值50ms作为阈值进行早晚期的切分。所以下一步研究工作主要有两方面：一是解决DNN在噪声不匹配情况下的泛化问题；二是研究不同切分时间对于DNN估计的影响。

4. 在本文中，我们都是在前端进行语音增强处理，并没有考虑后端识别器的影响。如何系统考虑复杂环境下的语音识别问题，将前后端进行统一的设计和优化，利用训练数据的先验信息提高语音识别在复杂环境下的鲁棒性是值得深入研究的问题。