中国科学院大学

University of Chinese Academy of Sciences

博士学位论文

双耳语音增强算法研究

作者姓名：方 义

指导教师：冯海泓 （研究员）

中国科学院声学研究所

学位类别：工学博士

学科专业：信号与信息处理

培养单位：中国科学院声学研究所

2018 年 5 月

Research on Binaural speech enhancement

A dissertation submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Doctor of Technical Science

in Signal and Information Processing

By

Fang Yi

Supervisor: Professor Feng Haihong

Institute of Acoustics, Chinese Academy of Sciences

May 2018

中国科学院大学

研究生学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：方义

日　　期：2018.5.22

中国科学院大学

学位论文授权使用声明

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

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：方义　　导师签名：

日　　期：2018.5.22　　日　　期：2018.5.22

摘 要

语音增强算法在助听器，蓝牙耳机等听力设备和手机通讯等设备中一直有着很重要的应用价值。随着人工智能技术的发展，新一代的智能设备，如智能音箱，智能耳机等设备的兴起使得远场语音交互技术再次得到广泛的关注，在远场语音交互中，一个关键的技术就是语音增强技术。远场相较于近场来说信噪比往往更低，这也意味着远场语音交互技术相较于近场的语音增强有着更多的挑战，这些挑战主要包括噪声的干扰，混响的干扰以及多说话人的干扰。而复杂的噪声环境下的语音增强技术无论是对于人耳的语音可懂度还是智能语音交互设备的识别率都有着重要的研究意义。

为了更好的对噪声进行抑制，麦克风阵列成为一种主流的方案，而波束形成技术由于具有计算简单，目标语音失真小等特点，成为远场语音交互中的一种主流的技术方案。双麦克风阵列由于具有安装简单，成本低等特点，也成为一种主流的选择。随着 Google Home 的双麦克风智能音箱的推出，国内外针对双麦克风的语音增强算法展开了广泛的研究，但是由于麦克风数目较小，传统的固定波束形成算法在双麦克风上效果有限。随着深度学习的快速发展，基于深度神经网络的语音增强算法也得到广泛研究，而另一类基于无监督聚类与多通道波束形成结合的方法也取得了显著的效果。尽管近年来众多学者提出一系列的语音混响抑制，语音降噪和语音分离方案，但是在实际环境下算法的鲁棒性还有待提升。为此，本文分别利用传统的信号处理知识和机器学习，深度学习新方法针对实际环境下的双耳语音增强算法展开研究。主要研究内容和创新点包括：

1.针对传统的优先效应模型的缺点，提出一种简单鲁棒的时延差估计算法，用于在噪声和混响环境下进行时延差估计。同时对混响抑制算法中的直达声与晚期混响的比值进行研究，最后利用估计的比值进行混响增益函数的计算。

2.提出一种基于独立向量分析的目标语音协方差矩阵估计算法，现有的基于深度学习的 Mask 估计和基于聚类的 Mask 估计算法中对噪声有着不错的效果，但是却不能解决语音分离任务，为此本文在分别利用 DNN 估计 Mask 结合 IVA 用于干扰加噪声协方差矩阵的获取。最后与多通道维纳滤波器结合，得到最后的增益函数。

3.针对欠定情况下的语音分离，首先针对目标方位已知的条件下提出一种双耳语

音增强算法。随后针对噪声干扰下的多语音分离算法，提出一种降噪与分离系统联合的神经网络结构，该算法不需要声源数目和声源方位等先验信息，能够在噪声干扰的条件下完成多语音分离任务。

关键词：双耳模型，语音降噪，混响抑制，语音分离，深度学习

Abstract

Speech enhancement algorithm has always been very important in application of hearing AIDS, Bluetooth headsets and mobile phone devices. With the development of Artificial Intelligence and the rise of new generation smart devices like smart Audio and smart headset, Far-field voice interaction technology has received wide attention again. In the far filed voice interaction, a key technology is the speech enhancement algorithm. The SNR (Signal -Noise Ratio) is always lower in far-field scenes, which means that far-field voice interaction technology has more challenges than near-field case. These challenges mainly include background noise, reverberation, and multi-speaker interference. The key technology of far-field voice interaction is speech enhancement. Therefore, the speech enhancement technology in the complex noise environment has important research significance both for the human's speech intelligibility and the recognition rate of the voice interaction device.

In order to reduce the noises, microphone array becomes a mainstream solution. And beamforming technology have become a mainstream solution in far-field speech interaction due to their simple calculation and small distortion of the target speech. And the dual microphone array has become a mainstream choice because of its simple installation and low cost. With the launch of Google Home, the dual-microphone speech enhancement algorithms have been widely studied. However, due to the small number of microphones, traditional fixed beamforming algorithm has limited effectiveness. With the rapid development of deep learning, speech enhancement algorithms based on deep neural networks have also been extensively studied, and the method of unsupervised clustering combined with multi-channel beamforming has also achieved remarkable results. Although many scholars have proposed a series of speech dereverberation, speech denoising, and speech separation schemes in recent years, the robustness of these algorithm remains to be improved in the actual environment. Therefore, this paper uses traditional signal processing ,

machine learning and deep learning method to reseach the binaural speech enhancement algorithm in the actual environment. The main research content and innovations include:

1: In order to overcome the drawback of the traditional precedence effect models, a simple and robust algorithm is proposed to estimate the time-delay values in the environment of noise and reverberation. And then the estimated time-delay values were applied to the coherent-to-diffuse power ratio (CDR) estimator, which can be used for reverberation suppression.

2: A speech covariance estimation algorithm based on independent vector analysis is proposed. Existing supervised deep learning and unsupervised clustering-based mask estimation algorithms can effectively suppression noise but cannot solve speech separation tasks. In this paper, we use the DNN and the independent vector analysis to estimated the interference plus noise covariance matrix. Finally, the final gain is obtained by combining with the multichannel Wiener filter.

3: For the case of underdetermined speech separation, a binaural speech enhancement algorithm is first proposed under the condition that the target direction is known. And then a neural network structure combining noise reduction and separation system is proposed for speech separation. This algorithm does not require prior information such as the number and the direction of the sound source, which can be completed speech separation task in a noisy environment.

Key Words: Binaural model, noise reduction, dereverberation, speech separation, deep learning

目 录

第1章 引言

1.1 研究背景

对于听力障碍人士来说，选配助听器和人工耳蜗成为恢复听力的一种有效手段，而听力设备中的噪声抑制则是提高佩戴者舒适度和语音可懂度的一种关键技术，而双耳听力设备的佩戴有助于佩戴者利用空间信息而获得更好的语音可懂度。

对于小型机器人和智能音箱等设备，前端的降噪对于语音识别的影响非常大，随着亚马逊 Echo 和 Google home 等智能音箱设备的推出，远场模式下的语音降噪成为一个研究热点。由于受到成本以及尺寸的限制，两个麦克风的语音降噪方案被广泛的使用。其中 Google home 就使用了两个麦克风。国内的科大讯飞，思必驰，云知声等企业也相继推出两麦克风方案。

图 1.1 Google home 实物图

Figure 1.1 Practicality picture of Google home

尽管在语音增强这一领域有着很长的研究历史，然而复杂场景下的语音分离依然面临着很大的挑战，这些挑战主要是现实环境中音频场景的复杂性，如图 1.2 所示，现实生活中往往同时存在着混响，强噪声和多说话人声的干扰。尽管针对不同的场景有很多相应算法的提出，但是各类算法往往仅在某一种场景下能够取得效果。而且在仅有两个麦克风的情况下，如何在复杂的场景下更加鲁棒的完成语音增强更是一个难点。为了提高算法在实际环境中的鲁棒性，本文分别对双耳的时延差估计，混响抑制，以及声源分离算法进行了研究。并对相关算法处理前后的效果进行了主客观的评价。

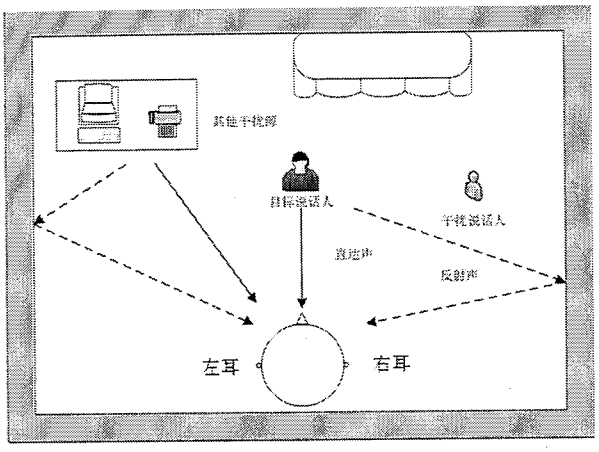


图 1.2 双耳语音增强场景图[1]

Figure 1.2 binaural auditory scene

1.2 研究历史与现状

1.2.1 单声道噪声抑制

传统的语音降噪方法主要有谱减法[2]，维纳滤波器[3]，以及基于统计模型的方法[4]。基于统计模型方法主要包括最大似然估计器，贝叶斯估计器，MMSE 估计器，其中应用最多的要属 log-MMSE 估计器[5]。这些方法假设人耳对语音的相位不敏感，利用语音与噪声在频谱结构上的差异去估计噪声的功率谱，从而获得与信噪比 SNR(Signal-to-noise ratio)相关的增益函数，对不同频点乘以不同的增益，回到时域后获得最后降噪后的信号。假设频域观测信号为：

$$Y_k = S_k + N_k \qquad \text{...（1.1）}$$

其中 S 和 N 分别代表纯净语音和干扰噪声。下表中给出了三类方法各自的准则以及增益函数形式。

表 1.1 传统语音增强算法

Table 1.1 Traditional speech enhancement algorithm

方法		设计准则	增益函数(suppression rule)
谱减法		$\|S_k\| = \|Y_k\| - \|N_k\|$	$G_k = \sqrt{\frac{\xi_k}{\xi_k + 1}}$
维纳滤波		$\arg\min_{w} E\{\|w^H y - s\|^2\}$	$G_k = \frac{\xi_k}{\xi_k + 1}$
统计模型（最大似然，贝叶斯，MMSE）	Log-MMSE	$\arg\min E\{\log S_k^* - \log S_k\}^2$	$G_k = \frac{\xi_k}{\xi_k + 1} \exp\{\frac{1}{2}\int_{\nu_k}^{\infty} \frac{e^{-t}}{t} dt\}$

其中：

$$\xi_k(t) = \frac{|S_k^2|}{\lambda_d} \quad \text{...（1.2）}$$

$$\gamma_k(t) = \frac{|Y_k^2|}{\lambda_d} \quad \text{...（1.3）}$$

$$v_k = \frac{\xi_k}{\xi_k + 1}\gamma_k \quad \text{...（1.4）}$$

这类非线性处理的增益形式往往会给增强后的信号引入“音乐噪声”[6]。为了更好的抑制音乐噪声问题，往往在估计增益值得时候使用判决引导法。

$$\xi_k(t) = \alpha \frac{|S_k(t-1)|^2}{\lambda_d(t-1)} + (1-\alpha)\max[\gamma_k(t) - 1, 0] \quad \text{...（1.5）}$$

其中$\xi_k(t)$和$\gamma_k(t)$分别代表先验 SNR 和后验 SNR，即最后的先验 SNR 为历史的先验 SNR 和当前的先验 SNR 的加权平均。从以上公式中可以看出，这类噪声抑制算法的核心在于噪声功率谱λ_d的准确估计。噪声的功率谱估计一般采用 VAD(Voice Activity Detection)或者语音存在概率等技术。常用的语音存在概率估计主要是基于最小值控制的递归平均算法[7-10]。该类算法的缺点是对非平稳噪声的抑制能力有限，且容易引起目标语音失真。

非负矩阵分解(Non-negative Matrix Factorization)是另一类噪声抑制算法[11]，算法假设原始矩阵 V 可以分解为 WH 两个非负矩阵，以便于对基矩阵 W 进行分类来进行语音增强。通过不同的分类方式，NMF 也分为有监督和无监督形式，其中无监督形式主要利用语音和噪声频谱结构的不同进行聚类，而有监督 NMF 则是离线对语音和噪声的基矩阵 W 进行训练，最后合成一个联合基矩阵。根据代价函数的不同，NMF 可以推导出不同的乘法更新法则。代价函数主要有欧式距(Euclidean Distance)、KL 散度（Kullback-Leibler divergence）、IS 散度(Itakura-Saito divergence)[12]。NMF 既包括了信号处理的知识也运用了机器学习的理论，是一类连接信号处理与模式识别的语音增强算法。该算法存在着计算量较大且鲁棒性有待提高等问题。

基于深度学习的语音降噪算法则是一类完全属于模式识别的算法，近年来，一些学者开始探索深度学习在语音增强领域的应用[13]。根据输入输出的不同，基于深度学习的语音增强算法可以分为以下几类：

表 1.2 深度学习语音增强算法

Table 1.2 Deep learing based speech enhancement algorithm

方法	输入特征	目标输出
分类	每个子带的特征	IBM
回归	当前帧与前后若干帧联合的带噪语音log谱	Soft mask
回归	当前帧与前后若干帧联合的带噪语音log谱	Clean_feature

其中分类方法指的是将原始信号分为若干个频带，在每个频带利用一个DNN(Deep neural network)进行语音或者噪声的分类任务。此时的目标函数称为Ideal binary mask(IBM),即目标函数非 0 即 1.这种 Binary mask 常常会引起语音过于失真。为此，更多的 Soft mask 的目标函数被提出，包括 Ideal Ratio Mask[14]，Spectral Magnitude Mask，以及考虑语音相位的 Complex Mask[15-16]。除了 Mask 的估计外，目标输出也可以直接设为纯净语音的特征，一般为纯净语音的 log 谱（即幅度谱取 log）[17-18]。在 DNN 基础上，模仿传统的语音增强算法，将多个神经网络级联进行语音增强的网络结构[19]和加入跳跃连接(Skip connection)的结构被提出[20]，实验结果证明这些网络结构相较于传统的 DNN 结构具有更好的泛化能力。另外循环神经网络中的 long short-term memory (LSTM)由于适合于处理和预测时间序列，也被广泛用于语音增强中[21]，卷积神经网络（Convolutional Neural Network CNN）也被证明在语音增强中能够取得不错的效果[22]。生成对抗网络 Generative Adversarial Networks 近期也被用于语音增强中[23]，生成对抗网络中的生成器 G 用于获取增强后的信号，判别器 D 用于判别增强后的信号与纯净语音信号的真伪。在该网络中，生成器 G 由卷积神经网络组成的 Encoder-decoder 结构构成，输入信号直接为带噪语音时域信号，输出直接为增强后的时域信号，避免了频域转换过程，是一种端到端的结构[24]。

1.2.2 单声道语音分离

语音分离指的是将多个说话人声独立分解出来。鸡尾酒会效应是一种描述人耳能够在复杂场景中能够选择性的听取某个说话人的技术，解决该问题早期的一个经典算法称为计算听觉场景分析（CASA/computational auditory scene analysis）。计算听觉场景分析即利用计算机技术来模拟人耳听觉的心理和生理特性[25]，例如 Gammtone 滤波器以及 Meddis 模型等，早期的 CASA 方法主要为无监督方法。非负矩阵分解 NMF 也可被用于语音分离任务，在 NMF 中我们可以分别训练不同说话人的基矩阵 W，由于 NMF 是单层线性模型，不容易描述不同说话人数据中的非线性关系。近期一些基于深度学习的单声道语音分离算法被提出。目前基于深度学习的单声道语音分离，一般来说可以分为三类：

Speakerer Separation，目标说话人和干扰说话人都固定

Target dependent，目标说话人固定，干扰说话人可变

Speaker independent，目标说话人与干扰说话人都可变

一个经典的基于深度神经网络的双说话人语音分离结构如图 1.3 所示。其中两个 output 的目标输出为：

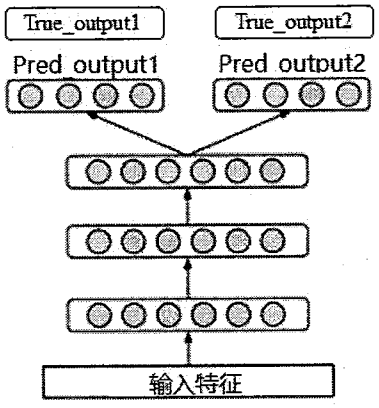


图 1.3 语音分离结构框图

Figure 1.3 Block diagram of speech separation

$$True\_output_1(t,f) = \left| \frac{|s_1(t,f)|}{|s_1(t,f)| + |s_2(t,f)|} \right| Y(t,f) \quad \text{...（1.5）}$$

$$True\_output_2(t,f) = \left| \frac{|s_2(t,f)|}{|s_1(t,f)| + |s_2(t,f)|} \right| Y(t,f) \quad \text{...（1.6）}$$

其中 $Y(t,f)$代表混合信号在 t 帧和 f 频点的幅度。S1，S2 分别代表两个纯净语音的幅度谱。我们定义：

$$\cos t_1 = \sum_t ((pred\_output_1 - true\_output_1) + (pred\_output_2 - true\_output_2)) \quad \text{...（1.7）}$$

$$\cos t_2 = \sum_t ((pred\_output_1 - true\_output_2) + (pred\_output_2 - true\_output_1)) \quad \text{...（1.8）}$$

对于前两种情况，我们可以利用如下的代价函数训练[26-27]：

$$\cos t = \cos t_1 - \gamma \cos t_2 \quad \text{...（1.9）}$$

其中γ为一个小于 1 的常数。随后，俞栋等人提出了基于置换不变性 PIT（Permutation invariant training）的训练方法[28]，该训练方法主要是为了解决语音分离任务中的排序问题，核心思想是计算真实标签和估计的标签之间的不同组合，并找到最小的组合，作为最后的误差值。以两个说话人的分离任务来说，此时的代价函数为：

$$\cos t = \min(\cos t_1, \cos t_2) \quad \text{...（1.10）}$$

基于深度聚类（Deep Clustering, DC）的说话人分离方法是另一种被广泛研究的语音分离模型[29]，该方法基于稀疏假设，即每个频点仅有一个说话人占主导，然后通过神经网络将原始信号投影到一个高维的空间，以便于在高维的空间利用传统的无监督的算法完成聚类。但是该方法的问题是，在事实的语音分离系统中，如果保证排序的正确性。另外，类似 SEGAN 等语音增强算法的结构，GAN 也同样被用于语音分离任务[30]。

1.2.3 多声道语音增强算法

由于单声道信号不能有效的利用空间的信息，对于人声干扰这种与目标语音谱结构类似的情况不能得到很好的抑制，虽然已有一些单声道的语音分离算法，但是实际情况下的应用还有诸多问题。而目前解决方向性点源干扰中最常用的是波束形成技术。传统波束形成技术的一大缺陷是需要考虑阵列一致性和各类估计误差。而对这类误差更加鲁棒的基于无监督聚类或分解等方法近年来成为发展的趋势。

1.2.3.1 固定波束形成算法

波束形成是一种经典的空域滤波算法。其中又可将其划分为固定波束形成，和自适应波束形式。固定波速形成中主要包括传统的延迟相加(DS, Delay and sum)波束形成 ,差分麦克风阵列[31]，以及在 MVDR 理论基础上建立的超指向波束形成。其中 DS 波束形成的白噪声增益最高，差分技术（适合小尺寸的阵列）和超指向波束形成能够获得更高的指向因子，对方向性点源的抑制比 DS 波束形成要好，但是实际使用过程中，需要平衡白噪增益和指向因子两个点。

1.2.3.2 自适应波束形成算法

自适应波束形成能够按照一定的准则自适应的调整权重。其中包括 MVDR,GSC,MAXsnr，MWF（multi-channel wiener filter）等等。MVDR 是一种目标方向无失真通过的同时干扰加噪声功率最小的情况下推导出的增益公式，而 GSC 理论上与 MVDR 等效,可以认为是 MVDR 的一种自适应滤波器(LMS,NLMS 等）的实现形式，经典的 GSC 中包括两个支路，其中上支路为固定波束形成模块指向目标语音，下支路为一个阻塞矩阵，目的是抵消目标语音获取参考噪声[32]。其中阻塞矩阵的设计是算法的核心部分，为了增加 GSC 的鲁棒性，带有约束条件的 LMS 滤波器(coefficient-constrained adaptive filters，CCAF)被加入到阻塞矩阵中，同时在自适应抵消模块中加入（norm-constrained adaptive filter，NCAF)。这类 CCAF-NCAF 结构的 GSC 成为一直简单有效的自适应滤波形式,随后 Cohen 等人首先提出的 TFGSC 结构也被广泛研究与应用[33-34]。MAXsnr 是一种最大 SNR 准则的滤波器，可以通过特征值分解的方法来求得滤波器权重。多通道维纳滤波器 MWF 是一种基于 MMSE 准则的滤波器，理论上该滤波器等效于 MVDR 加上单声道的维纳增益[35]。为了引入语音失真和噪声抑制的权衡，MWF 算法通过在最小均方误差准则中加入加权因子 μ 得到 SDW-MWF 算法[36]。值得一提的是，以上的这些不同准则推导出的增益公式，本质上并没有太大区别，只是各类形式的滤波器在目标语音失真和降噪抑制的平衡上有所不同。更关键的在于公式里面的干扰加噪声的协方差矩阵的求取。为此下文着重对提高 MVDR 鲁棒性的

各类方法进行介绍。

理论上来说，自适应波束形成的滤波效果显著优于固定波束形成，但是由于实际情况下各种误差(包括阵列自身的一致性和安装误差以及算法参数估计的误差)的存在，这类算法的鲁棒性成为一个难点。为了提高波束形成的鲁棒性，各种改进的算法被提出[37-39]。这类方法的本质是对导向向量约束在一定的范围内，从而牺牲空间分辨率来提高鲁棒性。该类方法的滤波器系数可以通过 Matlab 的 CVX 工具箱来求解。这类方法一方面计算复杂，难以实用，一方面对角加载类技术是在牺牲阵增益或空间分辨率的前提下换来的鲁棒性，没有解决本质的问题。

近年来，一些基于协方差矩阵重建的方法被提出。在传统的 MVDR 中，利用接收信号的协方差矩阵代替干扰加噪声的协方差矩阵，这种做法的劣势是对导向向量的失配非常敏感，容易发生目标语音零陷现象。这个问题的根本解决办法是估计出噪声加干扰协方差矩阵。 Gu Y 等人提出一种干扰加噪声协方差矩阵重构的方法，在假设已经知道干扰源方向的前提下，对该干扰方位一定区间内进行 capon 谱的积分，然后再利用最差性能最佳化的方法进行导向向量的约束[40]。该方法相对于直接用接收信号协方差矩阵代替噪声协方差矩阵有一定的改进，但是在导向向量存在误差的前提下，效果提升有限。而借助于最差性能最佳化等凸优化方法计算量太大，难以实时的实现。

基于 CGMM(complex Gaussian mixture model)的导向向量估计方法[41]，通过假设噪声信号与带噪语音的协方差矩阵服从复高斯分布。通过 EM 算法，来求解噪声的协方差矩阵，利用带噪语音协方差矩阵减去噪声协方差矩阵获取语音信号协方差矩阵，取语音信号协方差矩阵的最大特征值对应的特征向量为导向向量，该方法能够显著提高 SNR，且目标语音失真较小。但是，该方法对于语音分离任务的效果有待提升。且 EM 算法存在着对初始值敏感的问题。同样基于深度神经网络的 Time-frequency mask 与 MVDR 结合的方法也被广泛研究[42-43]。该方法与单声道 Mask 预测的方法类似，通过单个声道或每个声道独立进行 Mask 的估计来获取噪声的协方差矩阵，从而与 CGMM 类似的方法获得目标。由于深度学习方

法摒弃了相位信息，仅仅使用幅度信息来进行 Mask 估计，基于深度学习的方法同样不能解决人声干扰的情况。

多声道 NMF 是一种无监督的多声道语音分离算法，该方法将空间协方差矩阵 H 与单声道 NMF 中的谱分解(V=WH)相结合[44]。 Nikunen J 等人[45-46]在空间相关矩阵的基础上假设已知各声源的方位，将方位的先验信息加入到 MNMF 中。文献[47-48]提出一种 Rank-1 spatial model 结合 NMF 的分离算法，该方法假设每个声源的空间矩阵 $H=\alpha*\alpha^{H}$，其中α即导向向量。同时结合(independent vector analysis) IVA 技术解决聚类和排序问题，该方法只能解决 Mic 数目等于声源数目的情况，所以需要借助 PCA 对数据进行预处理。这类无监督方法由于相较于有监督学习方法，能够更好的适应不同的声学环境，在远场语音增强中能够取得比有监督学习更好的效果，但是也存在着计算复杂，对初始值的选取敏感等问题。

1.2.4 双耳声源定位算法

Lord Rayleigh 等人于 1907 年首次在球形人头模型基础上，提出基于耳间线索差的声音定位理论，随后一系列的双耳空间听觉理论知识得到研究[49]。我们知道，当声源在人侧面传来时，由于路径差的原因首先导致了两耳间出现时延差，而由于人头部的遮挡，同时会出现能量差。这两类差别称为耳间时间差 ITD（Interaural time difference）和耳间能量差 ILD（Interaural level difference）。且在小于 1.5K Hz 的低频，ITD 为主要的定位因素。在高于 1.5K Hz 的高频区域，ILD 为主要定位因素[50]。当然除了这两个主要的空间特征，还有单耳的一些定位线索。利用这些定位线索，目前的定位算法主要分为 2 类：

1.2.4.1 基于耳间差的定位

1948 年 Jeffress 首次提出了经典的双耳“巧合假说”，成为沿用至今的定位模型[51]。在该假说的理论基础上，1983 年 Lyon 就开始了声源定位方向的尝试，并提出了联合声源定位和声源分离同时进行的信号处理模型[52]。系统首先将双耳信号进行耳蜗特性模拟，得到不同的子带，在每个子带上求得互相关函数，并将所有子带的互相关函数进行累加，通过累加后的互相关函数的峰值信息即可得到 ITD

信息。时至今日，很多的时延差估计的研究依然是基于该模型。Kim Y I 等人使用直接对比两通道过零点位置的时延差估计方法[53]，并且利用过零点的时延估计统计信息(如方差)等来估计各个频段的信噪比，利用信噪比对各个频点进行加权联合，从而获得最终的结果，但是在混响环境下的表现有待提升。心理声学实验表明，人耳对声音的起始时刻较为敏感，人耳利用起始时刻的特性使得声源定位的能力加强。受此该听觉机理的启发，Braasch 等人提出了双耳互相关差模型[54]。此后很多的利用人耳听觉模型进行定位的算法被提出，但是大多还是在互相关寻找峰值的基础上进行的改进。

1.2.4.2 基于头相关传递函数的定位

ITD 和 ILD 信息仅仅是双耳空间听觉中的两个最主要的特征，实际上人耳的定位过程还包括了其他的很多线索。而头相关传递函数则包括了所有的定位线索，所以理论上说头相关传递函数的方法可以完成 3D 空间的声源定位。这是因为 HRTF 会直接利用人工头进行双耳数据的采集，这些数据包括了所有的双耳，单耳以及人体躯干部分的信息。利用 HRTF 进行声源定位的一般做法是：利用白噪声卷积 HRTF，进行训练。在实际测试时，提取相应的特征并与训练数据匹配进行分类。虽然该方法可以实现全空间的定位，但是由于每个人头部都会存在差别，而用通用的 HRTF 时会存在数据不匹配导致结果不正确的现象，而对每个单独个体进行特定的 HRTF 测量显然又不切实际。

1.2.5 双耳语音增强算法

计算听觉场景分析[55]（computational auditory scene analysis，CASA）很早就被应用于双耳语音增强领域中。该研究就是通过对人耳耳蜗部分功能建模，然后再去进行相应的分析。1990 年，加拿大麦基尔大学著名心理听觉学家 Albert Bregman 首先提出听觉场景分析（ASA）的概念，他认为听觉信息处理的过程就是将声音按各自自身特有的属性分离成“不同的流”，也就是将不同的声源进行归类[56]。Bregman 指出，人耳对混合语音的感知分成两个阶段。第一个阶段为分解（segmentation），多个声源混合在一起的声音首先分解为多个单元，每一个单元

代表了其中某个声源的主要信息。第二个阶段称为组织（Grouping），这个过程将分解的单元中属于同一类的单元进行排序组合在一起。该“分解组合理论”成为近年来 CASA 模型的经典结构[57-58]。2004 年，Hu G 等人提出了一个包络的调制谱信息和基频信息的分离系统，提高了高频区域的分离性能[59]。在国内，陈雪勤等人提出强背景噪声下的基音检测方法，得出相较与传统的自相关法，该方法具有更好的鲁棒性[60]。王磊等人利用听觉外围模型进行水下目标识别与分类[61]。类似的单声道声源分离系统近期也得到很多学者的研究[62-63]。

以上这些算法主要是基于单声道的计算听觉场景分析，基于双耳的 CASA 系统最早在 1983 年由 Lyon 提出，2003 年 Deliang Wang 等人提出基于声源定位的语音分离方法[64]，结果表明相较与单声道语音分离，双耳语音分离效果有明显提高。国内也有学者利用计算听觉场景分析进行双耳语音分离的研究[65]。由于利用了双侧耳的信号，相较与单通道信号，有了双耳的空间特征。

多通道维纳滤波算法是目前广泛研究的一个双耳语音增强算法[66],该方法不仅能够进行噪声的抑制同样能够保留声源的空间信息。但是目前大多的研究都是假设干扰加噪声的协方差矩阵已知的情况下去调整平衡信噪比与失真的最优权重。然而，如何准确的估计噪声加干扰协方差矩阵是个值得研究的问题。基于相干函数的干扰抑制算法[67]利用的是语音相位差信息，来计算左右耳的信噪比，然后计算相应的增益函数用于干扰点源的抑制。另外还有一系列算法利用双耳头相关传递函数（Head Related Transfer Function）来合成双耳散射函数[68]，然后利用超指向波束形成来形成固定指向性。同样利用 VAD 等信息估计干扰加噪声协方差矩阵的双耳 MVDR 波束形成也被提出[69]。尽管双耳语音增强算法可以使用前述的理论来实现。但是，双耳阵列与自由场阵列相比有着一些特殊性，由于头部的存在，使得侧方的声源在两耳之间存在明显的时延差（Interaural time difference，ITD）和能量差（Interaural level difference，ILD）。其中 ILD 信息相较自由场阵列有明显不同。利用这些双耳线索（binaural cues），一些学者提出一系列的双耳声源定位与分离算法[70-73]。其中最常用的思路为假设 ITD 和 ILD 信息

分别服从高斯混合模型，从而利用 EM 算法估计 Time-frequency mask[74]。同时利用深度学习的方法联合 LPS（单耳特征）与 ITD 和 ILD（空间特征）进行语音增强的工作也被很多学者提出[75-76]。但是这类的算法都是一个多输入单输出系统，对于双耳听力设备来说，丢失了原有的空间信息（ITD,ILD 等）。如何利用深度学习的方法在去除噪声的同时对干扰进行抑制是个有待研究的问题。

1.3 本论文主要研究内容

本论文主要研究了双耳语音增强算法，分别从传统的谱估计方法，无监督盲分离方法和基于深度神经网络的方法对双耳声源的定位，混响抑制和语音分离等进行了研究。

第一章阐述了语音增强算法对于听力设备佩戴者和语音识别系统的重要性。分析了远场语音增强系统中的关键点以及难点；对单声道语音降噪算法，单声道语音分离算法，多通道语音增强算法，双耳语音增强算法的研究现状进行了综述。

第二章主要是基础理论和知识的介绍，主要内容包括：人耳的听觉系统和数学模型以及双耳的“优先效应”模型。同时对头相关传递函数(HRTF)和混响理论进行了介绍。算法方法，对固定波束形成的，自适应波束形成的基本理论知识和不同滤波器之间的区别和联系进行了总结。最后，阐述了盲源分离理论和神经网络理论基础。

第三章首先介绍了混响环境下的时延差估计与 CDR 估计问题。我们提出一种利用优先效应模型的简单鲁棒时延差估计方法。在该章节中，我们首先回顾了经典的优先效应模型，并分析各种模型的优劣势，随后提出一种基于相干函数的峰值平滑策略，并在此基础上利用复数相干函数进行时延差的估计。利用仿真实验将本文的时延差估计算法和经典的优先效应模型的时延差估计算法进行了比较。随后，本文将时延差估计算法与 CDR 估计算法有效的结合，利用 CDR 估计值，构造一个维纳滤波器，进行晚期混响的抑制。

第四章，我们对常用的语音分离算法进行了详细分析，并提出目前的主流方向是基于 Mask 估计与波束形成结合的方法，该方法不依赖阵列流形，对阵列的一致

性误差不敏感。对目前主流的基于 CGMM 聚类和深度学习的 Mask 估计中对人声干扰抑制不了的缺点，本文提出一种基于独立向量分析(IVA)的Mask估计算法，独立向量分析(IVA)是在独立成分分析(ICA)基础上解决声源分离中的排序问题的改进算法。本文利用 IVA 估计出的解混矩阵，提出一种 Mask 估计算法，实验结果表明该方法比直接利用解混矩阵来进行分离能够获得更高的信干比。

第五章中，我们将重点研究欠定语音分离算法，首先提出一种在目标声源方位已知条件下利用双耳空间信息完成欠定情况下的目标语音分离算法。为了解决传统的多语音分离中的一系列问题，如噪声环境下算法效果的下降，欠定情况下算法效果的下降，需要预先知道声源数目等问题。提出一种基于深度学习的两阶语音分离系统。该系统分为两个部分，一部分为语音降噪系统，一部分为语音分离系统，在语音分离系统中我们利用置换不变性训练方法进行训练，并同时利用空间信息和谱特征。该系统在背景噪声和混响等干扰下依然能够有效分离出多个纯净语音。

第六章，将对本文进行总结并对目前算法中可能存在的不足点，提出下一步的研究工作。

第 2 章　语音增强理论基础

2.1 人耳听觉理论基础

2.1.1 人耳听觉系统与数学模型

图 1 为人耳听觉系统模型图。人耳听觉系统主要由外耳，中耳和内耳组成。其中外耳起到声源定位和放大的作用，可以将声音的能量集中于鼓膜上。中耳由三块听小骨组成。这三块听小骨互相作用，既能起到放大声压的功能，同时也能防止过大的声音带来的损害。内耳是整个听觉系统的感知部分，其中耳蜗部分是最主要的部分，声信号到神经信号的转换就是在耳蜗中的内毛细胞中进行的[77]。下面就介绍建立耳蜗部分的数学模型。

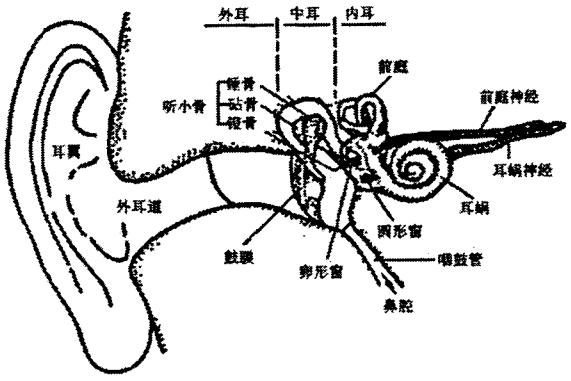


图 2.1　人耳听觉系统模型图

Figure 2.1 Model of human auditory system

2.1.2 耳蜗基底膜频率分解特性模型

耳蜗的基底膜具有频率分解特性，一般采用 gammtone 滤波器实现[78]。常用的 gammtone 听觉滤波器组的时域冲击响应函数见下式：

$$g(f,t)=\begin{cases}t^{l-1}\,exp(-2\pi bt)\,cos(2\pi ft) & if\ t>0\\ 0 & else\end{cases} \quad \text{...(2.1)}$$

其中 f 为滤波器中心频率, l 为滤波器阶数。B 定义为等效矩形带宽(equivalent rectangle bandwidth,ERB)。中心频率 f 与 ERB 一般满足如下关系[79]：

$$ERB(f)=24.7(0.0043f+1.0) \quad \text{...(2.2)}$$

图 2 为一个 8 通道的 gammtone 滤波器的时域和频域响应图。

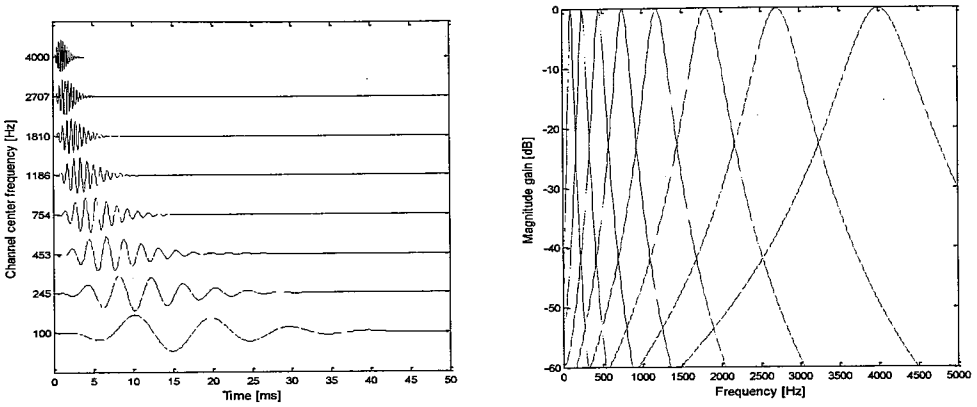


图 2.2　8 通道 gammtone 滤波器的时域和频域响应

Figure 2.2 The time and frequency domain response of the 8 channel gammtone filter

2.1.3 内耳毛细胞发放特性模拟

1986 年 Meddis 提出的内耳毛细胞发放特性模型成为一种沿用至今的经典模型[80]。所谓毛细胞发放特性指的是毛细胞和听觉神经突触之间进行物质交互的过程。语音在耳蜗的基底膜上产生振动后，会引发递质向突触间隙渗透，渗透率 h(t)与于输入信号的振幅有关，见下式：

$$h(t)=\begin{cases}\dfrac{A+stim(t)}{A+B+stim(t)}g, A+stim(t)>0\\ 0, \qquad A+stim(t)<0\end{cases} \qquad \text{...（2.3）}$$

其中:stim(t)是信号的瞬时幅度,A 和 B 代表阈值和最大渗透率。毛细胞发放特性可以一般用以下公式来描述：

$$\frac{dq}{dt}=y[1-q(t)]+rc(t)-h(t)q(t) \qquad \text{...（2.4）}$$

$$\frac{dc}{dt}=h(t)q(t)-lc(t)-rc(t) \qquad \text{...（2.5）}$$

$$p(t)=hc(t)dt \qquad \text{...（2.6）}$$

图 2.3-2.5 分别为一个 500Hz 的纯音信号，以及该信号分别经过 gammtone 滤波器和 Meddis 模型后的输出图，从图中可以明显看出经过 Meddis 模型后，在信号的起始阶段的内毛细胞发放率会瞬间提高，随后会逐渐趋于平稳状态，这与心理声学实验中的“起始主导”现象[81-83]吻合，起始主导现象指的是人耳对声音的起始时刻很敏感，随着声音的持续播放，敏感度趋于平稳。

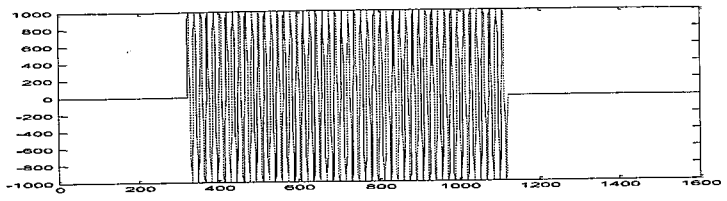


图 2.3　500Hz 纯音原始信号

Figure 2.3 original 500Hz pure tone signal

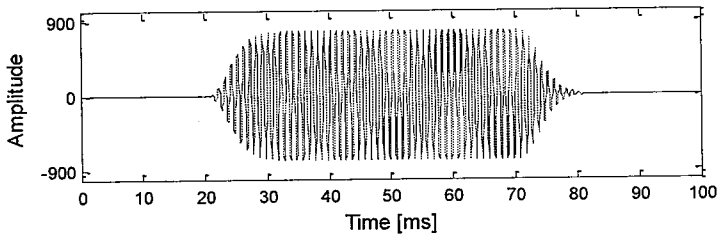


图 2.4 中心频率为 500Hz 通道内的 gammtone 滤波器输出信号

Figure 2.4 The output signal of gammtone filter in the 500Hz

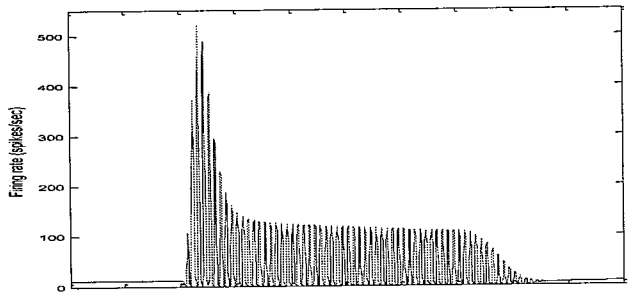


图 2.5　经过 Meddis 模型处理后的信号

Figure 2.5 The signal processed by Meddis model

2.1.4 双耳效应

双耳对于声源的定位一般来说，依靠多个因素，主要包括双耳时间差 ITD，双耳能量差 ILD，耳郭的影响，以及人耳的心理声学特性。由于人头部的遮挡，侧方的声源在两耳之间形成明显的时间差与能量差，其中低于 1500Hz 时，ITD 占主导，高于 1500Hz 时，ILD 占主导。

人耳定位的另外一个线索是“耳郭效应”。耳郭效应是一种单耳线索，指的是在声波到达耳郭后有部分声波在耳郭中经过一段时间不断的反射后才进入耳道，这些反射现象的最终结果是形成频谱上的峰谷。而不同位置，特别是上下方向不同的角度来的声波造成的反射现象有较大区别，所以这些区别有利于帮助人耳对声源头分位的判断[84]。

人耳的定位除了依靠以上的双耳特征，在混响环境下，人耳依然能够准确的进行定位，这主要归因于人耳的“优先效应”。对于优先效应(Precedence Effect)，许多的学者很早就有研究，并将其定义为：尽管在混响环境下声源可以通过多个路径传递至人的双耳，但是首先到达的直达声却在声源定位中起到主导作用[85-89]。该现象的发现与前面所述的“起始主导”现象相吻合。

2.2 头相关传递函数

我们知道声源到接收器之间的信道可以看成是一个滤波器，通过时域卷积的形式可以获取接收信号与发射信号之间的关系。而头相关传递函数指的是声源分别到两耳鼓膜之间的两组滤波器。上节中，我们提到了双耳的一系列定位线索，而HRTF 理论上包括了以上所有的单耳以及双耳特征。

HRTF 函数的获取可以通过 KEMAR 人工头来测量或通过理论模型计算得到。国外的麻省理工学院，德国亚琛工业大学，国内的中科院声学研究所，华南理工大学等机构对 HRTF 进行了深入的研究。目前也有各类的消声室环境或混响环境的开源 HRTF 库。

2.3 混响基础知识

混响指的是声源在室内经过各种物体的反射现象，这些物体主要是包括墙壁，地板，天花板等物体。在一个声源停止发声后，由于这些物体反射的存在，使得声音不是马上停止，而是经过一系列反射后，持续一段时间再停止。其中一般衡量停止时间大小的参数为 T60，T60 的定义为：声源停止发声后，声压级减少到 60dB 的时间称为混响时间[90]，单位为秒。

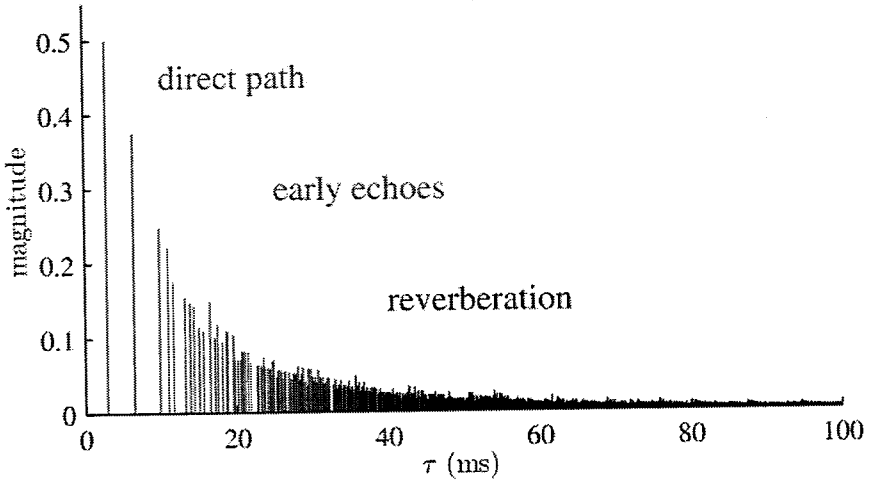


图 2.6 RT60=0.25s 时的房间冲击响应[93]

Figure 2.6 The room impose response function of RT60=0.25s

2.4 波束形成理论基础

麦克风阵列波束形成是一种应用最广泛的空间滤波器，现假设有 M 个阵元，阵列的接收信号定义为：

$$y(t) = ds(t) + n(t) \quad \text{... (2.7)}$$

写成向量形式为：

$$\begin{pmatrix} y_1(t) \\ y_2(t) \\ \cdot \\ \cdot \\ y_{M-1}(t) \\ y_M(t) \end{pmatrix} = \begin{pmatrix} d_1 s(t-\tau_1) \\ d_2 s(t-\tau_2) \\ \cdot \\ \cdot \\ d_{M-1} s(t-\tau_{M-1}) \\ d_M s(t-\tau_M) \end{pmatrix} + \begin{pmatrix} n_1(t) \\ n_2(t) \\ \cdot \\ \cdot \\ n_{M-1}(t) \\ n_M(t) \end{pmatrix} \quad \text{... (2.8)}$$

其中 t 为时间，y(t)为接收信号，s(t)为目标源信号，n(t)为噪声信号，d 为导向向量。转换到频域后有：

$$Y(e^{jw}) = S(e^{jw})d + N(e^{jw}) \quad \text{... (2.9)}$$

d 是一组与阵列流形和声源方位和距离有关的向量，一般在远场假设下忽略幅度的影响，则 d 是一组声源在麦克风之间的时延差有关的向量。此时 d 为：

$$d^T = [\exp(-jW\tau_1), \exp(-jW\tau_2) \ldots, \exp(-jW\tau_M)] \quad \text{... (2.10)}$$

假设定义波束形成的输出即为：

$$Y(e^{jw}) = \sum_{m=1}^{M} W_m(e^{jw}) X_m(e^{jw}) = W^H X \quad \text{... (2.11)}$$

2.4.1 评价标准

麦克风阵列的评价标准主要有阵增益，指向因子(Directivety index，DI)，以及Beampatten 波束图。其中阵增益定义为输出信噪比与输入信噪比的比值[91]：

$$G = \frac{SNR_{\text{out}}}{SNR_{\text{in}}} = \frac{\left|W^H d\right|^2}{W^H R_{nn} W} \quad \text{...（2.12）}$$

其中 Φ_{nn} 为噪声的协方差矩阵。特别指出的是，如果噪声为空间互不相关的白噪声，此时 Φ_{nn} 为单位矩阵 I,即自相关为 1，互相关部分为 0.此时的阵增益称为白噪增益（White Noise Gain,WNG）：

$$WNG = \frac{\left|W^H d\right|^2}{W^H W} \quad \text{...（2.13）}$$

指向因子 DI 指的是，阵列抑制散射噪声(diffuse noise)的能力[92]：

$$DI = 10\log_{10}\frac{\left|W^H d\right|^2}{W^H \Gamma_{vv} W} \quad \text{...（2.14）}$$

其中 Γ_{vv} 指的是散射场函数，一般情况下：

$$\Gamma_{vv} = \text{sinc}(\frac{wf_s d}{c}) \quad \text{...（2.15）}$$

波束图一般指的不同频率的信号在不同方向上的增益，某一角度的增益定义为：

$$\left|H(\theta)\right| = -10\log_{10}\frac{\left|W^H d\right|^2}{W^H \Gamma_\theta W} \quad \text{...（2.16）}$$

计算出所有角度的增益，即可画出波束图。

2.4.2 固定波束形成

延迟相加波束形成(Delay-and-Sum beamforming)是最简单的一类波束形成方法，该方法直接将各麦克风之间的时延差补齐与参考麦克风一致后累加[93]。对于非相关的空间白噪声效果较好，理论上来说，DS 波束形成的最大白噪声增益为 M。但是对于相干噪声场抑制效果则不理想。

为了提高抑制相干噪声场的能力，超指向波束形成技术被提出。超指向波束形成的优势在于在低频拥有比 DS 更高的指向因子。超指向波束形成的理论是假设空间分布着各向同性的噪声，则此时的噪声协方差矩阵可以通过对各个方位的导向向量的积分获取，一般形式为：

$$\Gamma_{vv} = \mathrm{sinc}(\frac{wf_s d}{c}) \quad \text{...（2.17）}$$

通过 MVDR 的公式可以方便得出此时的滤波器权重为：

$$W = \frac{\Gamma_{vv}^{-1} d}{d^H \Gamma_{vv}^{-1} d} \quad \text{...（2.18）}$$

虽然超指向能够在低频获得较高的指向因子，但是白噪声增益却很低。实际使用时，由于阵列自身的误差和空间白噪声的存在，使得超指向波束形成需要考虑白噪增益和指向因子之间的平衡。一般可以通过对角加载来提高超指向波束形成的鲁棒性。

$$W = \frac{(\Gamma_{vv} + \mu I)^{-1} d}{d^H (\Gamma_{vv} + \mu I)^{-1} d} \quad \text{...（2.19）}$$

其中 μ 为对角加载量。

另外一类固定波束形成的设计方法叫做差分麦克风阵列（Differential microphone arrays），差分麦克风阵列一般用于端射结构，最简单的差分为两个麦克风之间的一阶差分。而随着麦克风数目增加，阶数也随着增强，理论上来说，差分麦克风阵列的阶数可以做到麦克风数减 1 阶，但是随着阶数的增加，低频的白噪增益等问题也需要去考虑。下面就以一阶差分为例，来说明差分麦克风阵列技术。下图所示两个麦克风位置呈端射方向。通过设计不同的时延，再相减。可以得出不同的波束图。（Dipole：$\alpha_{1,1}$=0；Cardioid：$\alpha_{1,1}$=-1；Hypercardioid：$\alpha_{1,1}$=-1/2；Supercardioid：$\alpha_{1,1}$=-1/$\sqrt{2}$ ；τ_0=δ /c）

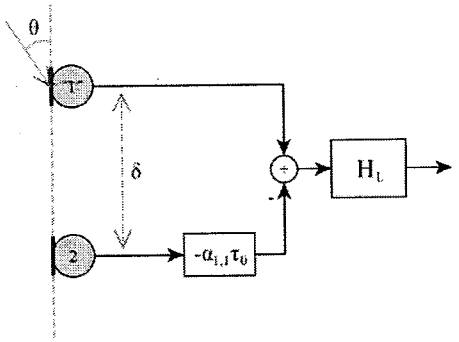


图 2.7 一阶差分麦克风阵列

Figure 2.7 First-order Differential microphone arrays

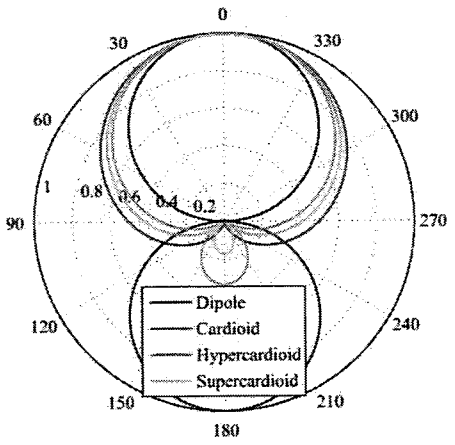


图 2.8 不同延迟时的波束图

2.4.3 自适应波束形成理论

固定波束形成的阵增益有限，为了更大的提升阵增益，一系列的自适应波束形成理论被提出，自适应波束形成能够根据干扰源的方位和能量等信息，自适应的对干扰进行抑制，大大提升阵增益，但是自适应波束形成也有其缺点，主要是实际环境下存在各种误差，导致传统的自适应波束形成算法的性能受到影响。

2.4.3.1 MVDR

Capon 等人于 1969 年提出的 MVDR 波束形成器成为一种应用最广泛的自适应波束形成器[94]。该算法的设计准则为：

$$\min w^H R_{nn} w \quad s.t.\ w^H d = 1 \qquad \text{...(2.20)}$$

即使得目标方向无失真通过的前提下，噪声的输出功率最小。利用拉格朗日算子可以得出 MVDR 的权重为：

$$w = \frac{R_{nn}^{-1} d}{d^H R_{nn}^{-1} d} \qquad \text{...(2.21)}$$

2.4.3.2 GSC

另一类广泛使用的自适应滤波器称为，广义旁瓣抵消器。一个典型的 GSC 结构包括三个部分：一个固定波束形成器(FBF)，一个阻塞滤波器(BM)和一个自适应滤波器(MC)[32]。其中 FBF 是为了增强目标信号，BM 目的是为了阻塞目标信号，从而获取参考噪声，最后 FBF 的输出与 BM 的输出进行自适应滤波，获取最后的信号。在该过程中最为关键的部分是 BM 模块，即 GSC 算法的好坏取决于是

否能够抵消掉目标语音，获取参考噪声。最简单的 GSC 的矩阵矩阵设计为延迟相减的形式，在麦克风矩阵较小的情况下，可以利用差分技术分别形成指向目标语音和零陷目标语音的两个波束，分别构成 GSC 的上下支路[95]。

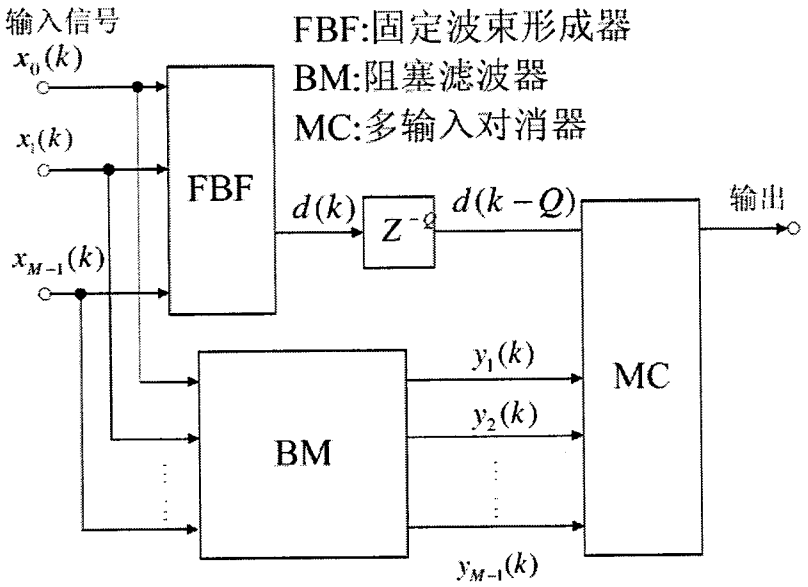


图 2.9 GSC 结构框图

Figure 2.9 Block diagram of GSC

2.4.3.3 多通道维纳滤波器

多通道维纳滤波器（Multi-channel wiener filter，MWF）是基于 MMSE 准则的自适应滤波，理论上等价于 MVDR 波束形成算法级联一个单通道维纳滤波算法。MVDR,GSC 都可以理解为多通道维纳滤波器（PMWF）的一种特殊情况[96]。

假设接收信号为：

$$\mathrm{y}(t,f) = c(f)\mathrm{x}(t,f) + n(t,f) \quad \text{...（2.22）}$$

MWF 方法基于 MMSE 准则求取滤波器系数矢量：

$$W_{MWF} = \arg\min_{w} E\{|w^H y - x_1|\} \quad \text{...（2.213）}$$

为了引入语音失真和噪声抑制的权衡，MWF 算法通过在最小均方误差准则中加入了加权因子 μ 得到 SDW-MWF 算法[97-100]，该算法的准则为：

$$W_{SDW-MWF} = \arg\min_{w} E\{|w^H \mathrm{x} - x_1| + \mu |w^H \mathrm{n}|\} \quad \text{...（2.24）}$$

最终的解为：

$$W_{SDW-MWF} = \left[R_{xx} + \mu R_{nn}\right]^{-1} R_{xx} u_1 \quad \text{...(2.25)}$$

2.4.3.4 最大信噪比波束形成

最大信噪比波束形成是使得信号与噪声加干扰比值最大的准则下提出的[101]。信干噪比可以定义为：

$$\frac{S}{n} = \frac{W^H R_{xx} W}{W^H R_{nn} W} \quad \text{...(2.26)}$$

其中 R_{xx} 为目标信号的协方差矩阵，R_{nn} 为干扰加噪声的协方差矩阵。该滤波器的权重可以通过对$R_{nn}^{-1}R_{xx}$进行特征值分解来获得。

表 2.2 不同准则下的多通道滤波器

Table 2.2 Multi-channel filter under different criteria

滤波器	设计准则	滤波器的增益函数
MVDR	$\min W^H R_{nn} W \quad s.t.\ W^H d = 1$	$W = \frac{R_{nn}^{-1} d}{d^H R_{nn}^{-1} d}$
MWF	$W_{MWF} = \arg\min_W E\left\{\left\|W^H y - x\right\|\right\}$	$W_{SDW-MWF} = \left[R_{xx} + \mu R_{nn}\right]^{-1} R_{xx} u_1$
MAXsnr	$\arg\max \frac{W^H R_{xx} W}{W^H R_{nn} W}$	$\mathrm{eig}(R_{nn}^{-1} R_{xx})$

2.5 神经网络理论基础

神经网络是一类模拟神经元的结构。一个典型的神经网络结果如下图所示。

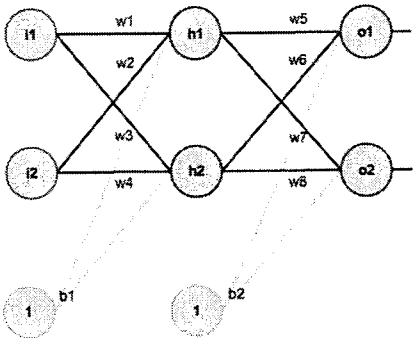


图 2. 10 神经网络基本结构

Figure 2.10 Basic structure of neural network

在有监督的神经网络中，一般需要定义合适的代价函数（cost function）。随后通

过正向的参数传递与反向的误差回传不断调整神经网络中的参数，使得神经网络的预测值与标签值接近。传统的多层神经网络存在诸多问题：如梯度消失(神经网络反向传播过程中，梯度越来越小，主要原因在于以前广泛使用的 sigmiod 激活函数的的梯度小于 1，见图)；过拟合（数据过分拟合训练数据，而不符合测试数据)；同时存在着难以训练，容易陷入局部最优等等。这些问题，使得神经网络一度被其他的机器学习方法（如支持向量机等）所替代。

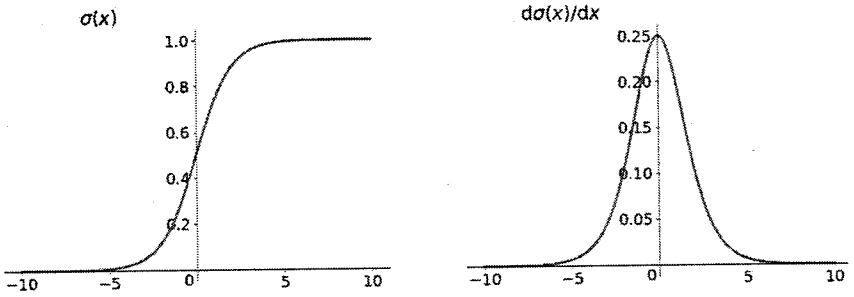


图 2.11 Sigmoid 激活函数(左)与其偏导数(右)

Figure 2.11 activation function of Sigmoid (left) and its partial derivative (right)

随着 Hinton 等人对神经网络的改进，使得深度神经网络得到极大的发展，这些改进包括：新的激活函数的提出，如 Relu，PReLU ，ELU，Swish 等等。各类激活函数的公式为：

表 2.2 不同激活函数及其偏导数

Table 2.2 Different activation functions and their partial derivatives.

激活函数	公式	偏导
Sigmoid	$f(x)=\frac{1}{1+e^{-x}}$	$f'(x)=f(x)(1-f(x))$
ReLU	$f(x)=\begin{cases}0 & \text{for } x<0\\ x & \text{for } x\geq 0\end{cases}$	$f'(x)=\begin{cases}0 & \text{for } x<0\\ 1 & \text{for } x\geq 0\end{cases}$
PReLU	$f(x)=\begin{cases}\alpha x & \text{for } x<0\\ x & \text{for } x\geq 0\end{cases}$	$f'(x)=\begin{cases}\alpha & \text{for } x<0\\ 1 & \text{for } x\geq 0\end{cases}$
ELU	$f(x)=\begin{cases}\alpha(e^{x}-1) & \text{for } x<0\\ x & \text{for } x\geq 0\end{cases}$	$f'(x)=\begin{cases}f(x)+\alpha & \text{for } x<0\\ 1 & \text{for } x\geq 0\end{cases}$

Swish	$f(x) = \frac{x}{1+e^{-x}}$	$f'(x) = f(x) + \frac{1-f(x)}{1+e^{-x}}$

除了激活函数的改进，一系列新的防止陷入局部最优的参数更新策略也被提出，这些方法主要包括：增加了动量的 SGD，带有速率更新的 Adagrad 以及改进的带有二阶动量的 AdaDelta / RMSProp。而 Adam 则结合了以上方法的优点，将一阶动量，二阶动量结合，从而成为目前主流的参数更新策略。

总的来说，传统的神经网络存在梯度消失，过拟合等难以训练的问题。深度学习中采用新的逐层训练机制（2006 年 Hinton 提出）。以及一系列新的激活函数：ReLU，新的权重初始化方法，新的损失函数，新的防止过拟合方法（例如随机丢弃策略 Dropout,正则化等）,新的参数更新方法。这些方法主要都是为了解决传统的多层神经网络的一些不足：梯度消失，过拟合等。这些新的方法的出现也使得目前深度神经网络得到了广泛的应用。

网络结构：现在的深度神经网络中出现很多新的网络结构，包括图像处理中常用的 CNN(卷积神经网络),语言建模中常用的 RNN(循环神经网络),GAN[102](生成对抗网络)

随着深度学习的大热，各大公司也相继推出自己的深度学习开源库函数，主要包括以下几类：

Tensorflow: 由谷歌推出，是目前使用最多的一个库。Tensorflow 的优点是灵活且用户量极大，在 Github 中可以找到大量的深度学习项目实例。

Keras: Keras 基于 python 编写，后端可以是 TensorFlow 或 Theano,可以认为是在 Tensorflow 等结构下的封装。今年刚被 tensorflow 设置为官方接口。Keras 的高度模块化，使得入手极快，但也存在不够灵活的缺点。

Caffe: Caffe 是由 C++的深度学习框架。由 Facebook 推出。优势是速度快，但缺少对 RNN 的支持。另外值得注意的是 Facebook 推出基于 python 语言的 Pytorch。成为一种发展速度极快的开源框架。

另外，Matlab 2017 版本也新增了对深度学习各种库函数的支持。

2.6 盲源分离理论基础

盲源分离是一个经典的解决“鸡尾酒会”问题的方法。如图 2.12 所示，在语音

分离任务中，指的是仅通过观测到的多通道信号，恢复出无法直接观测到的各个原始声源的过程[103]。

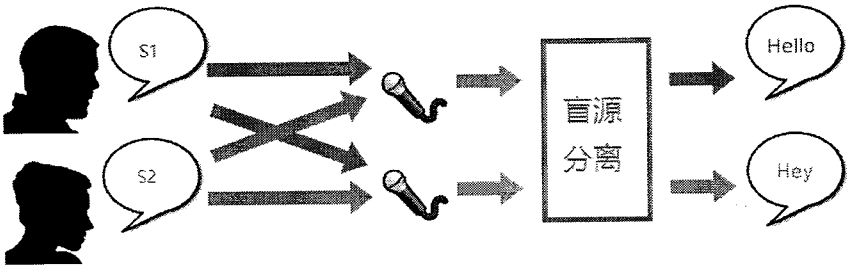


图 2.12 盲源分离系统

Figure 2.12 The system of blind source separation

一个典型的线性模型（瞬时模型）可以表示为：

$$y(t) = As(t) + n(t) \quad \text{...（2.27）}$$

写成矩阵形成：

$$\begin{bmatrix} y_1(t) \\ y_2(t) \\ \cdot \\ \cdot \\ y_{M-1}(t) \\ y_M(t) \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & & a_{1,N} \\ a_{2,1} & a_{2,2} & & a_{2,N} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{M-1,1} & a_{M-1,2} & & a_{M-1,N} \\ a_{M,1} & a_{M,2} & & a_{M,N} \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \\ \cdot \\ \cdot \\ s_{N-1}(t) \\ s_N(t) \end{bmatrix} + \begin{bmatrix} n_1(t) \\ n_2(t) \\ \cdot \\ \cdot \\ n_{N-1}(t) \\ n_N(t) \end{bmatrix} \quad \text{...（2.28）}$$

其中 y 为观测信号，A 为信号到传声器阵列之间的传递函数，n 为噪声。M 为麦克风数目，N 为声源数目。盲源分离的目标即找到一个解混合矩阵 B，能够将各个独立声源分开。当 M≥N 时其中 B 可以通过对 A 求逆运算来获得。对解混合矩阵 B 求解的一个经典方法为独立成分分析（Independent Component Analysis，ICA）。

在 ICA 中一般需要假设各个信号是独立的且为非高斯分布。这是因为 ICA 的估计的其中一个思想中用到了中心极限定理：即多个独立分量的和近似服从高斯分布。在该定理的基础上，我们有如下定理，假设滤波器 B 与 Y 的输出为某一个独立声源 S_i，则 BY 比任何一个 S_i 都要接近高斯分布，除非 BY 等于 S_i，所以 ICA 问题转换为求解 WY 使得每个 S_i 具有最大的非高斯性。从这个角度来说，各个声源的非高斯分布是 ICA 估计算法的关键。度量非高斯性的参数主要有以下几类：峰度/峭度 Kurtosis，KL 散度，熵，互信息量（Mutual Information）等。不

同的代价函数具有不同的迭代公式。在进行 ICA 之前，往往需要利用（Principal Component Analysis，PCA）进行预处理，PCA 一方面用于数据的降维，降低计算量，另一方面也承担解相关操作。为了提高运算效率，在工程中，一般采用频域（Frequency-Domain ICA，FDICA）算法进行语音的盲分离。FDICA 算法有两个缺点：第一个问题是为不能确定信号的能量(方差)，这个问题在 ICA 中一般直接假设各个声源的方差为 1. 第二个问题更为关键-排序问题，在实时的处理过程中，我们需要一帧帧(或 batch)处理信号，且语音属于宽带信号。而 ICA 处理过程中虽然能够在每个频点将声源分开，但是如何将属于同一声源的所有频点序列正确排在一起，是一个难以解决的问题。尽管一系列的解决序列模糊的问题被提出，主要方法为同一声源频域间的相关性，或者利用 DOA 信息。为了彻底解决频域 ICA 排序问题，一种称为独立向量分析（Independent vector Analysis，IVA）的技术被提出[104]。下图 2.13 所示为 IVA 分离框图，在 IVA 中，将整个频段作为一个整体考虑，该方法有效解决了 ICA 的排序问题。

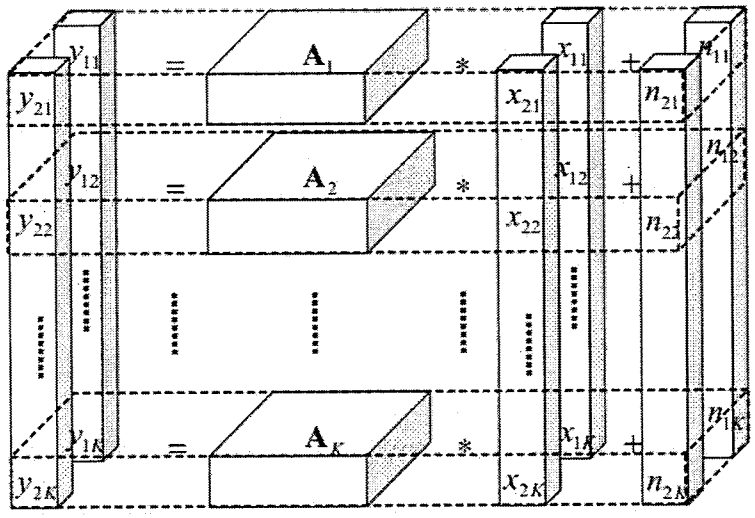


图 2. 13 IVA 结构框图[105]

Figure 2.13 The structure diagram of IVA

第 3 章　双耳混响抑制算法

3.1 引言

在人耳的听觉系统中，ITD 和 ILD 被认为是人耳定位的两个最主要的特征。而大多数的定位算法中是根据 ITD，因为对于方位估计而言，ITD 值的分布相较 ILD 值具有更小的方差。所以，本章着重对混响环境下的 ITD 估计进行详细的研究。ITD 估计中使用最广泛的一类算法是基于 Jeffress 模型，该模型是一个经典的生理学模型，在该模型中假设声源到达耳朵后，神经脉冲开始在左右侧传递，当左右侧的神经脉冲在传递过程中在某一个神经元同时激活时，便产生声源空间信息。

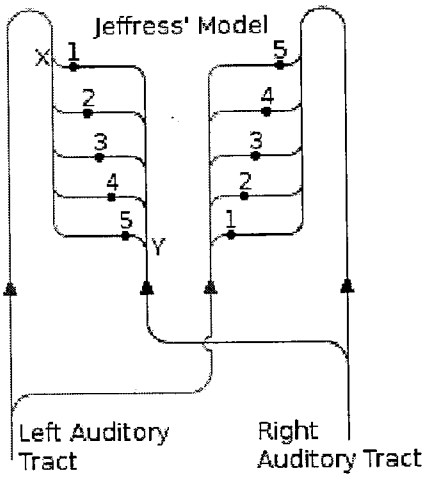


图 3.1 Jeeffress 模型[106]

Figure 3.1 Jeeffress's model

基于该机理，众多算法被提出，其中大多数算法是通过寻找广义互相关的峰值[107-110]来进行时延差估计，这类方法在没有混响和噪声的安静环境下有着很好的表现，但是现有的语音设备往往在存在背景噪声和混响的房间内使用，在这种环境下，传统的方法定位误差急剧增加。为了提高混响环境下的定位准确率，一些“优先效应模型”被提出。优先效应指的是人耳在混响环境下能够通过首先到达人耳的直达声进行声源定位而忽略后续混响声的干扰[111]。Jie huang 等人提出了一种经典的回波分离模型[112]，在该模型中利用房间的冲击响应函数去构建一个直达声与混响声的比值，只利用比值较高的时频段进行时延估计。但是该模型高度依赖房间的冲击响应函数，而每个房间的冲击响应函数又不同，这限制

了该算法的鲁棒性。

Martin 等人提出另外一种模型[113]，该模型是对 Zurek 模型的实现。在该模型中，首先将信号分解到不同的频带，然后通过 Meddis 毛细胞发放模型[80]，最后在直达声起始点处产生一个抑制信号，最后将该抑制信号乘上广义互相关函数来进行时延估计。该模型中的前端信号处理过程计算量很大，且噪声环境下的起始点检测也是一个难点。

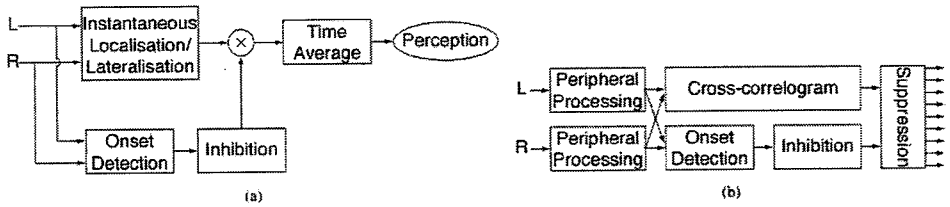


图 3.2 Zurek 模型（a）和 Martin 模型（b）

Figure 3.2 Zurek's model and Martin's model

Faller 和 Merimaa 等人提出一种新的基于频域相关性的优先效应模型[114]，该模型在每个子带计算一个互相关函数，利用互相关函数最大值处的幅度值大小来判断该频段的有效性，最后只利用相关性大的子带进行时延估计。近期的一些心理声学实验中也表明了耳间相关性与声源定位存在着密切的联系[115-116]。尽管这些模型提供了一系列的方法用于在混响环境中提取有用信号，但是这些模型的前端信号处理过程复杂，且存在阈值难以选取的问题。

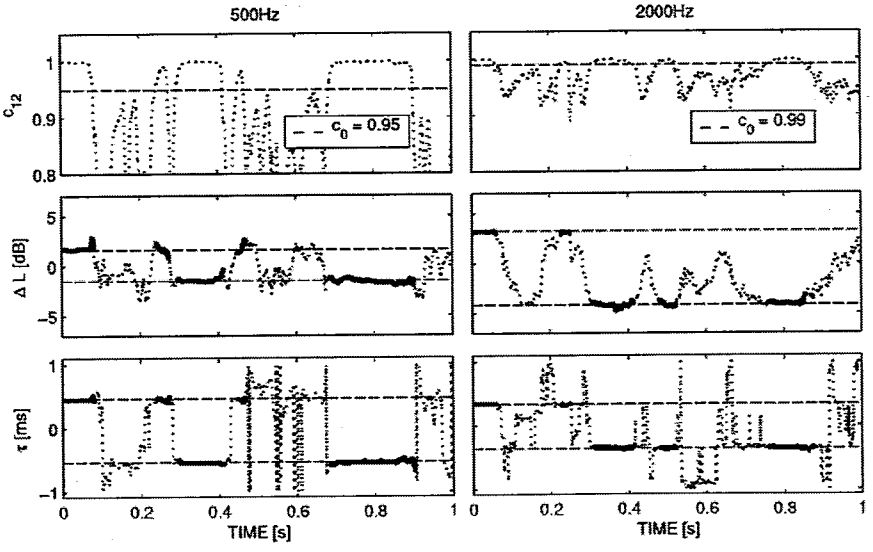


图 3.3 窄带信号的 IC(上)与 ILD(中)和 ITD(下)之间的关系

Figure 3.3 The relationship between the IC(upper) and the ILD(middle), ITD(lower) of the narrow-band signal.

房间混响的存在影响了智能语言设备的声源定位能力，同时对于听力设备佩戴者来说，混响影响了语音的可懂度。为了降低房间混响的影响，众多学者提出了一系列的算法。当前主流的混响抑制方法主要包括有：基于波束形成方法、基于统计模型的方法和基于深度学习的方法。基于波束形成的混响消除方法是一种最基本的混响抑制方法，该方法通过将导向向量指向特定角度来完成抑制非主瓣方向的干扰，由于混响可以认为是四面八方都存在的信号，所以波束形成具有一定的混响抑制作用，但是对于双耳信号来说，由于麦克风数目较少，使用波束形成方法难以取得满意的效果。基于统计模型方法，假设晚期混响部分是由前面语音帧进行联合加权得到的，通过估计相应的滤波器，达到混响抑制的效果。基于深度学习的方法则是通过有监督学习任务，学习混响信号的频谱结构和纯净信号频谱结构之间的映射。

基于相干函数的混响抑制方法在过去几个世纪中已经被广泛的研究[117-119]。这种方法的主要思想是直达声信号在两个麦克风之间相关性较高，而由墙壁等物体反射叠加后的晚期混响信号则被认为是非相干性的散射噪声（diffuse noise）。如图 3.4 所示，图中蓝色部分为晚期混响部分。

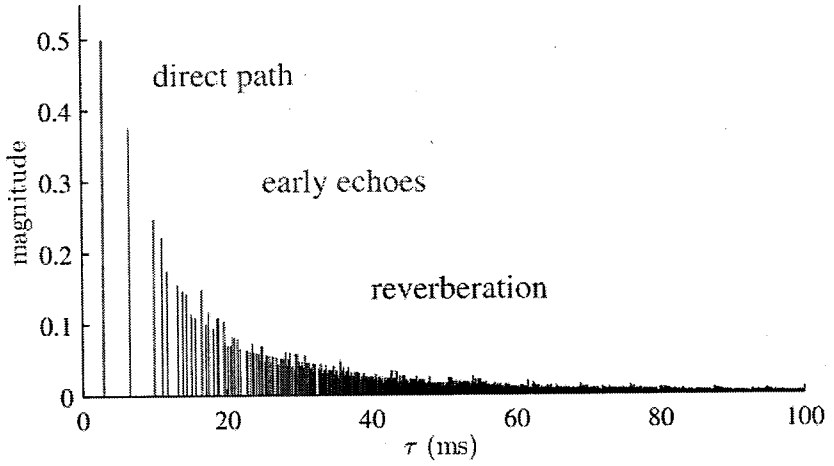


图 3.4 RT60=0.25s 时的房间冲击响应[93]

Figure 3.4 The room impose response function of RT60=0.25s

近期，基于 CDR（Coherent-to-Diffuse power Ratio）的混响抑制算法被广泛研究[120]。Marco Jeub 等人在假设两个麦克风之间没有时延差的情况下提出一种双通道 CDR 估计函数，随后 Marco Jeub 等人又在假设已经知道时延差的情况下，在频域乘以一个与时延差有关的相移因子来获得有时延差时的 CDR 估计函数

[121]。Schwarz 等人汇总了前人的结果并提出一系列的 CDR 无偏估计函数，包括依赖方位信息和不依赖方位信息的估计函数[122]。郑成诗考虑到头部的阴影效应，提出一种双耳的混响抑制估计函数[123]。但是这些方法有着以下的不足。首先是大部分的估计函数是在假设时延差已知的前提下进行的，而在混响环境下的时延差估计本身也是一个难点。

本文提出一种基于相干函数的简单有效的“优先效应”模型，有效的将声源定位与混响抑制结合。本章内容安排如下：第二节详细介绍了本文的信号模型，包含相干函数的介绍以及 CDR 估计的理论知识；第三节和第四节重点介绍了我们提出的方法，包括基于峰值平衡的时延差估计方法以及 CDR 估计方法；第五节为实验结果和评价；最后一节给出本章小结。

3.2 信号模型

我们定义左右两个麦克风接收到的时域信号为：

$$x_i = s_i(t) + n_i(t) \quad i = l, r \qquad \text{...（3.1）}$$

其中 t 表示采样点，$s_i(t)$ 和 $n_i(t)$分别代表目标语音和干扰噪声。变换到频域后：

$$X_i(\lambda,\mu) = S_i(\lambda,\mu) + N_i(\lambda,\mu) \; i = l, r \qquad \text{...（3.2）}$$

其中 λ 和 μ 分别代表帧数和频点。两通道间的频域相干函数定义为：

$$\Gamma_{XlXr}(\lambda,\mu) = \frac{P_{XlXr}(\lambda,\mu)}{\sqrt{P_{XlXl}(\lambda,\mu)\cdot P_{XrXr}(\lambda,\mu)}} \qquad \text{...（3.3）}$$

其中 $P_{XlXl}(\lambda,\mu)$和 $P_{XrXr}(\lambda,\mu)$分别 $X_l(\lambda,\mu)$和 $X_r(\lambda,\mu)$的自功率谱，而 $P_{XlXr}(\lambda,\mu)$ 为 $X_l(\lambda,\mu)$和 $X_r(\lambda,\mu)$间的互功率谱。它们的计算公式为：

$$P_{XiXi}(\lambda,\mu) = \alpha \cdot P_{XiXi}(\lambda-1,\mu) + (1-\alpha)\cdot|X_i(\lambda,\mu)|^2 \quad i = l, r \qquad \text{...（3.4）}$$

$$P_{XlXr}(\lambda,\mu) = \alpha \cdot P_{XlXr}(\lambda-1,\mu) + (1-\alpha)\cdot X_l(\lambda,\mu)X_r^*(\lambda,\mu) \qquad \text{...（3.5）}$$

其中 α 为相邻帧之间的平滑因子。

3.2.1 理想情况下点源的相干函数

在没有混响和各种噪声干扰的情况下，理想的相干函数定义为：

$$\Gamma_{ideal}(\lambda,\mu) = e^{j\cdot\omega\cdot fs\cdot\tau} \qquad \text{...（3.6）}$$

将公式展开为：

$$\Gamma_{ideal}(\lambda,\mu) = \cos(\omega\cdot fs\cdot\tau) + \mathrm{j}\cdot\sin(\omega\cdot fs\cdot\tau) \qquad \text{...（3.7）}$$

其中 τ 为两个麦克风之间的时延差。即理想点源的相干函数实部和虚部分别为：

$$\mathrm{Real} = \cos(\omega\cdot fs\cdot\tau) \qquad \text{...（3.8）}$$

$$\mathrm{Imag} = \sin(\omega \cdot fs \cdot \tau) \quad \text{...（3.9）}$$

我们发现理想相干函数的模值为 1.

$$|\Gamma_{ideal}(\lambda,\mu)| = \sqrt{\cos(\omega \cdot fs \cdot \tau)^2 + \sin(\omega \cdot fs \cdot \tau)^2}=1 \quad \text{...（3.10）}$$

下图所示为不同时延下的理想相干函数的实部和虚部。该图中两个麦克风间距为 0.255m，采样率为 16 kHz，频点数为 256.

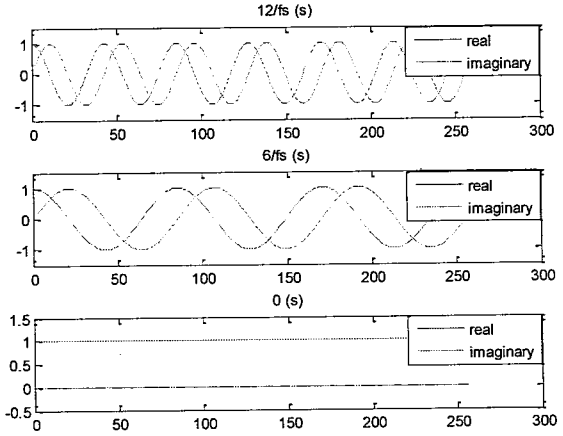


图 3.5 不同时延下的理想相干函数实虚部

Figure 3.5 The Real and Imag of the ideal coherence function at different time-delays

3.2.2 CDR 估计算法理论基础

我们假设一个声源信号在两个麦克风之间没有时延差，且语音和噪声不相关，此时有：

$$P_{slsl}(\lambda,\mu) = P_{srsr}(\lambda,\mu) = P_s(\lambda,\mu) \quad \text{...（3.11）}$$

$$P_{nlnl}(\lambda,\mu) = P_{nrnr}(\lambda,\mu) = P_n(\lambda,\mu) \quad \text{...（3.12）}$$

$$P_{xlxl}(\lambda,\mu) = P_{xrxr}(\lambda,\mu) = P_x(\lambda,\mu) \quad \text{...（3.13）}$$

$$P_x(\lambda,\mu) = P_s(\lambda,\mu) + P_n(\lambda,\mu) \quad \text{...（3.14）}$$

其中$P_s(\lambda,\mu)$和$P_n(\lambda,\mu)$分别代表信号和散射噪声的自功率谱，然后 CDR 定义为：

$$CDR = \frac{P_s(\lambda,\mu)}{P_n(\lambda,\mu)} \quad \text{...（3.15）}$$

又有：

$$\Gamma_s(\lambda,\mu) = \frac{P_{slsr}(\lambda,\mu)}{P_s(\lambda,\mu)} \quad \text{...（3.16）}$$

$$\Gamma_n(\lambda,\mu) = \frac{P_{nlnr}(\lambda,\mu)}{P_n(\lambda,\mu)} \quad \text{...（3.17）}$$

其中$\Gamma_s(\lambda,\mu)$和$\Gamma_n(\lambda,\mu)$分别为语音和噪声的相干函数，最后的 CDR 定义式为：

$$\mathrm{CDR} = \frac{\Gamma_n(\lambda,\mu)-\Gamma_x(\lambda,\mu)}{\Gamma_x(\lambda,\mu)-\Gamma_s(\lambda,\mu)} \quad \text{...（3.18）}$$

3.3 本文提出的算法

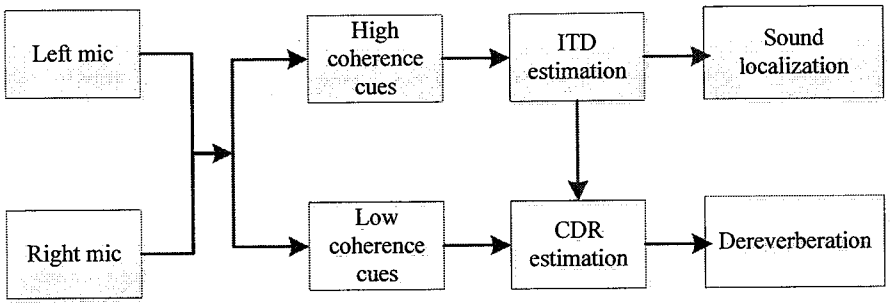


图 3.6 本文基于相干函数线索选择算法框图

Figure 3.6 The proposed coherence based binaural cues selection system

本文算法的总体框架图如上图所示，首先提出一种简单有效的“优先效应”模型来对高相关信号进行提取，随后对高相关性的信号进行 ITD 估计，随后估计的 ITD 联合 CDR 估计算法，进行晚期混响抑制。

3.3.1 基于相干函数的峰值平滑

在本节中我们提出一种利用相干函数实部和虚部的时延估计算法。一个基于峰值平滑策略的双耳线索选择算法在 3.3.1 中被提出，然后利用(K-Nearest Neighbor，KNN)最邻近算法进行最后的定位结果的预测。在 3.2.1 节中我们分析了理想情况下点源的相干函数，并得出结论，此时相干函数的模值为 1. 然而，在实际环境下，由于混响和各种噪声干扰的存在，在每个频点的相干函数的模值为 0-1 之间的数。Faller 和 Merimaa 等人的模型中通过子带相干函数来进行信号可靠性的判断。受该方法的启发，我们提出以下假设：在每个频点，如果相干函数的模值越大则在该频点的直达声占主导地位的可能性越大。我们所提的模型与 Faller 和 Merimaa 等人的模型的区别是利用模值来代替子带的相关函数，因此显著降低了计算量。基于最小值跟踪的噪声功率谱估计算法是一种经典的单声道语音增强算法[124]，该方法假设带噪语音功率谱中较小的值代表了噪声的功率谱[125]。受到该方法的启发，我们提出一种对相干函数进行峰值平滑的策略，即通过相干函数的模值的大小来对复数的相干函数进行更新：

$$Peak(\lambda,\mu)=\begin{cases}\alpha_2\cdot Peak(\lambda-1,\mu) & if\ |\Gamma_{X1X2}(\lambda,\mu)|<|Peak(\lambda-1,\mu)| \\ \alpha_1\cdot Peak(\lambda-1,\mu)+(1-\alpha_1)\cdot\Gamma_{X1X2}(\lambda,\mu) & \text{else}\end{cases} \quad ...(3.19)$$

显然，进行峰值平滑的目的是为了选择更大的相干函数的模值的频点，这些频点被认为是直达声占主导地位的概率更高。

3.3.2 时延差估计算法

在实际的时延估计算法中，我们首先建立一个离线的相干函数库，在线预测中首先对相干函数进行峰值平滑，而后利用 KNN 算法来进行结果的预测。

图 3.7 是离线模型的框图，根据公式 3.6，我们可以获得不同时延条件下的理想相干函数的实部和虚部。对全频带的特征我们可以利用 KNN 进行建模，KNN 是一种简单但有效的分类方法，且比较适合多分类任务。我们将时延差（-12/fs 到 12/fs）平均划分为 25 个类，在我们的 KNN 模型中，K 取值为 25，欧式距离（Euclidean distance，EU）被定义为判决准则。在线预测阶段，我们首先将时域接收信号通过短时傅里叶变换变换到频域，然后利用计算两个通道间的相干函数，并利用 3.3.1 节中所述的峰值平滑技术估计$Peak(\lambda,\mu)$，最后利用 KNN 将$Peak(\lambda,\mu)$与离线模型匹配，得出最后的预测值。

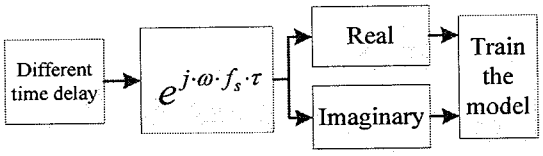


图 3.7 离线模型

Figure 3.7 Offline model

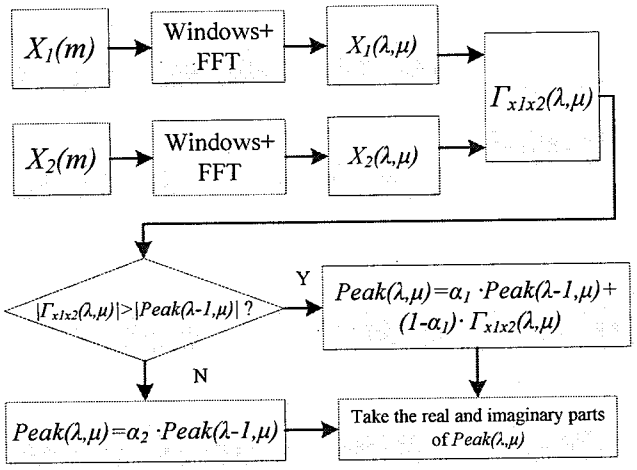


图 3.8 在线预测

Figure 3.8 Online prediction

3.4 CDR 估计算法

在 3.2.2 节中，我们推导了 CDR 的估计公式：

$$CDR=\frac{\Gamma_n(\lambda,\mu)-\Gamma_x(\lambda,\mu)}{\Gamma_x(\lambda,\mu)-\Gamma_s(\lambda,\mu)} \quad \text{...（3.20）}$$

上式中的 CDR 估计公式中假设两个麦克风间的信号没有时延差，对于存在时延差的情况，Schwarz 和 Kellermann 等人提出一种改进算法[122]，该算法将一个相位因子 $e^{-j2\pi f\Delta t}=\Gamma_s^*$ 乘以原来的 CDR 公式中：

$$CDR_{DOA-dependent}\left|\frac{\Gamma_s^*(\Gamma_n-\Gamma_{x1x2})}{Re\{\Gamma_s^*\Gamma_{x1x2}\}-1}\right| \quad \text{...（3.21）}$$

在该方法中，算法依赖先验 ITDs 信息，Thiergart 等人提出在频域利用瞬时相位信息来代 ITD [124]，即 $e^{jargPx1x2}=\Gamma_s^*$.此时 CDR 公式可以表述为：

$$CDR_{Thiergart}=Re\{\frac{\Gamma_n-\Gamma_{x1x2}}{\Gamma_{x1x2}-e^{jarg\Gamma_{x1x2}}}\} \quad \text{...（3.22）}$$

该估计不依赖先验 ITD 信息，这使得该公式能够容易在实际环境下使用，然而实际情况下混合信号的瞬时相位并不等价于纯净直达声的相位。Schwarz 和 Kellermann 等人提出一种改进的无偏估计算法[122]：

$$CDR_{DOA-independent}=\frac{\Gamma_n Re(\Gamma_{x1x2})-|\Gamma_{x1x2}|^2-\sqrt{\Gamma_n{}^2(Re(|\Gamma_{x1x2}|^2)-|\Gamma_{x1x2}|^2+\left(\Gamma_n-Re(\Gamma_{x1x2})\right)^2}}{|\Gamma_{x1x2}|^2-1} \quad \text{...（3.23）}$$

在[123]中，双耳的 CDR 估计函数被提出，目标信号的相干函数基于 ITD 模型而双耳的散射噪声相干函数利用 Lindevald 和 Benade 等人的理论[125]。作者随后对比了自由场和双耳模型下的 PESQ 分数，结果表明在 ITD 信息已知的条件下 $CDR_{DOA-independent}$ 能够取得最好的效果。在 3.3 节中我们已经提出一种鲁棒 ITD 估计算法，根据 ITD 信息我们可以得出目标信号的相干函数。而双耳的散射噪声相干函数我们同样使用 Lindevald 和 Benade 等人的模型：

$$\Gamma_n^{Binaural}=\frac{1}{\sqrt{1+(\pi fd/c)^4}}\frac{\sin(4.4\pi fd/c)}{4.4\pi fd/c} \quad \text{...（3.24）}$$

最后，噪声相干函数和目标信号相干函数代入到最后的公式中：

$$\text{CDR}=\left|\frac{\Gamma_s^*(\Gamma_n^{Binaural}-\Gamma_{x1x2})}{Re\{\Gamma_s^*\Gamma_{x1x2}\})-1}\right| \quad \text{...（3.25）}$$

其中 Γ_s 代表目标信号相干函数，$Re\{\}$代表取复数的实部而*代表复共轭。

3.5 实验评价与讨论

3.5.1 ITD 估计实验

在本实验中，信号采样率为 16 kHz，FFT 点数为 512，前后帧之间重叠率为 75%。

为了评估算法在不同混响时间下的性能，我们用 Image 模型[126]产生一个 4×5×3 的房间。麦克风放置于房间中心位置，麦克风间距为 0.255m。声源与麦克风中心距离为 1m。混响时间 T60 分为为 0.1 到 1s（每 0.1s 为 1 间隔）。我们首先在 IEEE 语料库[127]中选择纯净的语音信号，然后与 Image 模型产生的房间冲击响应函数卷积，加入一定信噪比（分别为 20，10，0dB）的白噪声后，获得最后的麦克风接收信号。2 阵元的麦克风阵列放置于房间中间，声源距离麦克风中心位置为 1m，麦克风阵列与声源高度均为 1.3m。本文算法的参数见表 3.1.

表 3.1 本文算法参数

Table 3.1 Parameter values used in implementation

参数	α	α_1	α_2	d	c
数值	0.5	0.35	0.95	0.255m	340m/s

本文算法的时延差估计算法与传统的 PHAT 加权的广义互相关函数(GCC-PHAT)，Martin 的优先效应模型[113]和 Faller 的基于 IC 的模型[114]. 这两个优先效应模型的部分参数[128]列于表 3.2 中,其余参数与本文算法保持一致。为了证明峰值平滑策略的有效性，我们同样对比了不采用峰值平滑(Without Peak-tracking)与采用峰值平滑(Peak-tracking)的结果。其中不采用峰值平滑的具体做法为：直接利用 KNN 来利用公式 3.3 计算得到的相干函数进行时延差的估计。图 3.9-3.10 所示为 RT_{60}=0.6 s ,SNR=20 dB 的条件下，每一帧的 ITD 估计结果。其中图 3.9 为单个声源大的定位结果，图 3.10 为 3 个声源的估计结果。其中，红色点线代表实际估计的结果，黑色虚线代表的是真实结果。图 3.11-3.13 所示为在 SNR=20 dB 时,不同的混响时间下的平均正确率(100 条语句)，其中允许的误差为 0（即只要结果与正确结果不一样，就判为错）。同样的，图 3.14-3.19 所示分别为 SNR=10 和 0 dB 时的平均正确率。

以上结果表明：当混响时间增加或信噪比降低时，定位结果受到明显的影响。当 SNR 足够高(SNR=20 dB)时,估计结果主要受到混响的影响。从图 3.11-3.13 中可以看出随着混响时间的增加，传统方法正确率显著下降，其中基于 PHAT 加权的广义互相关方法(GCC-PHAT)和不采用峰值平滑(Without Peak-tracking)的方法受混响时间影响最为严重。Martin 和 Faller 的“优先效应”模型都有助于提高正确率，且 Faller 模型表现更好。而本文提出的算法显著优于对比算法，尤其在混

响时间高于 500ms 的情况。在 SNR=10 dB 时的结果与 SNR=20 dB 时的结果类似。当 SNR=0 dB 时，如图 3.17-3.19 所示，传统方法的正确率在低混响时间下都已经显著下降，而本文算法在所有条件下均优于对比算法。

为了更一步的评估估计结果，我们对不同条件下估计结果的均方根误差(Root-Mean-Square Error, RMSE)进行统计并列于表 3.3 中。其中 Proposed1 和 Proposed2 分别代表不采用峰值平滑(Without Peak-tracking)与采用峰值平滑(Peak-tracking)的时延差估计方法。从图中可以看出在时延差为 0 时，所有方法取得最好的结果。且随着混响时间和信噪比的降低，所有方法估计结果均有下降，但是本文提出的方法在任何条件下始终取得最低的 RMSE。

表 3.2 优先效应模型中选取的参数

Table 3.2 Parameter values used in precedence effect models

模型	前处理	频带数目	其它
Martin	Gammatone filter and Meddis model	32	Inhibitory Gain=1 Inhibitory Time Constant=1.5 ms
Faller	Gammatone filter	32	IC Threshold=0.5

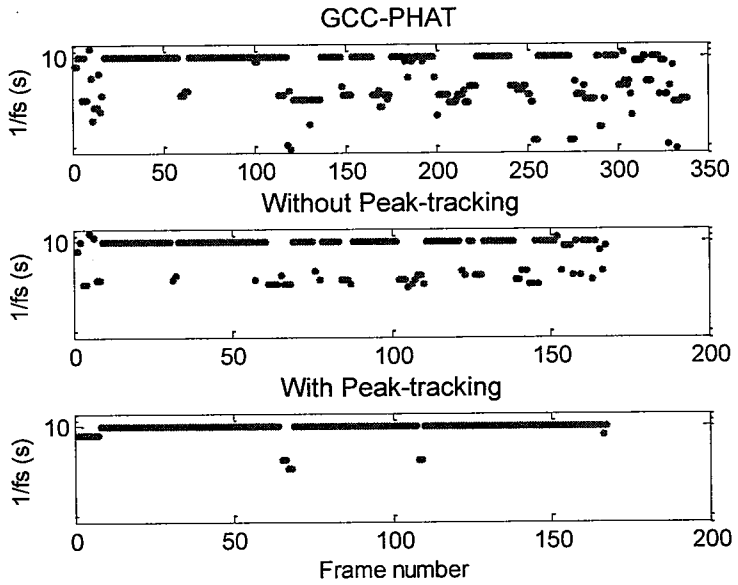


图 3.9 单声源定位结果

Figure 3.9 Localization result of single sound source

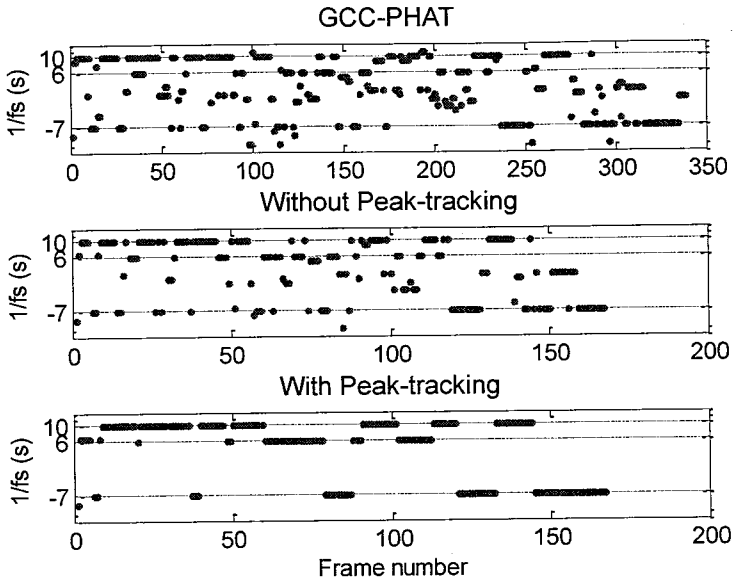


图 3.10 三声源定位结果

Figure 3.10 Localization result of three sound sources

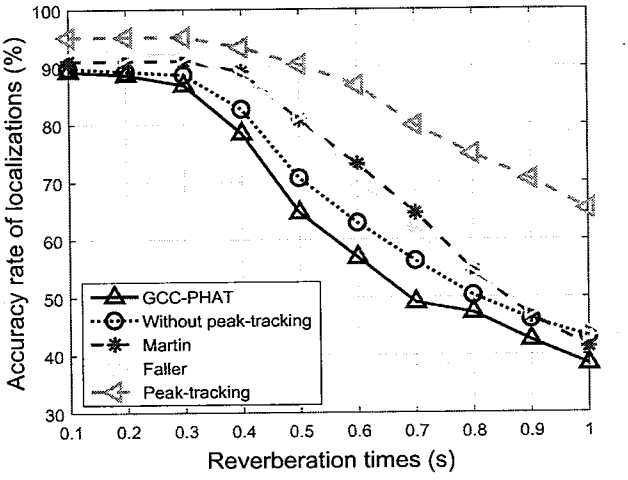


图 3.11 30° (20 dB SNR)时的平均正确率

Figure 3.11 Average accuracy at 30° and 20 dB SNR

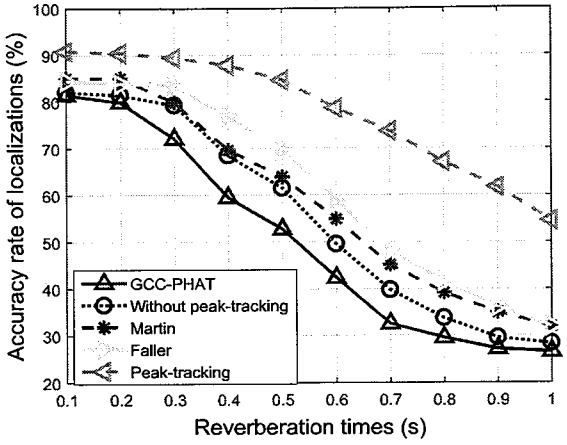


图 3.12 60° (20 dB SNR)时的平均正确率

Figure 3.12 Average accuracy at 60° and 20 dB SNR

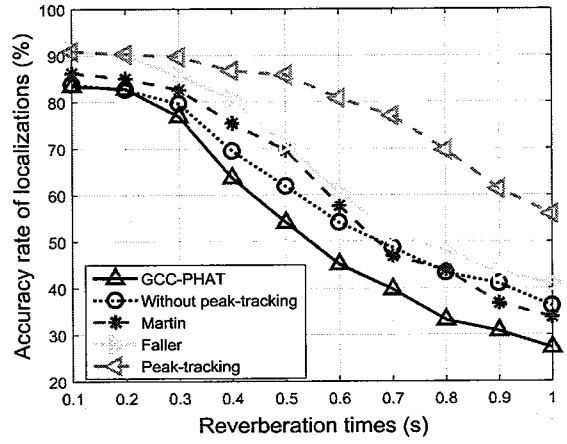


图 3.13 135° (20 dB SNR)时的平均正确率

Figure 3.13 Average accuracy at 135° and 20 dB SNR

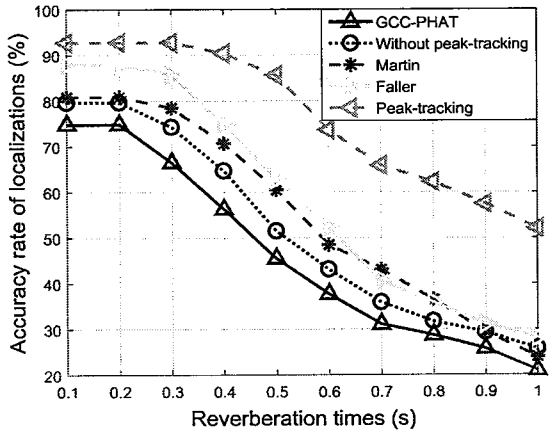


图 3.14 30° (10 dB SNR)时的平均正确率

Figure 3.14 Average accuracy at 30° and 10 dB SNR

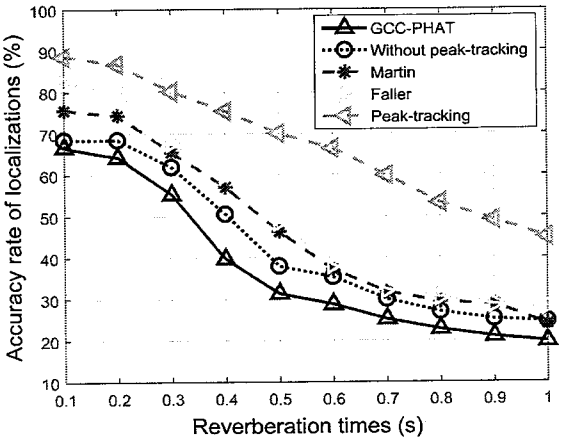


图 3.15 60° (10 dB SNR)时的平均正确率

Figure 3.15 Average accuracy at 60° and 10 dB SNR

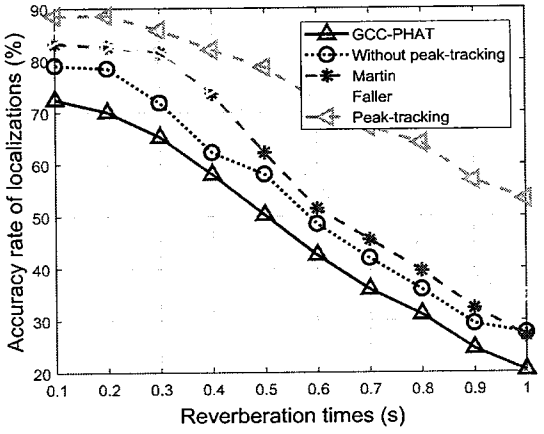


图 3.16 135° (10 dB SNR)时的平均正确率

Figure 3.16 Average accuracy at 135° and 10 dB SNR

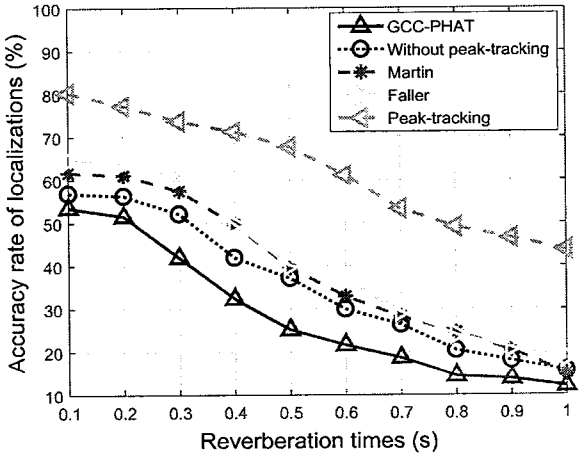


图 3.17 30° (0 dB SNR)时的平均正确率

Figure 3.17 Average accuracy at 30° and 0 dB SNR

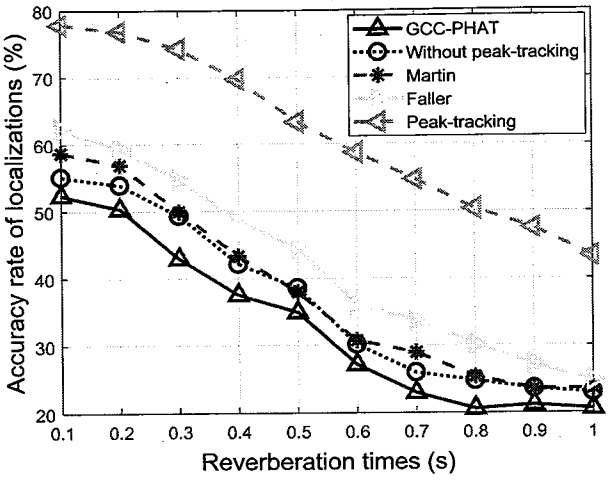


图 3.18 60° (0 dB SNR)时的平均正确率

Figure 3.18 Average accuracy at 60° and 0 dB SNR

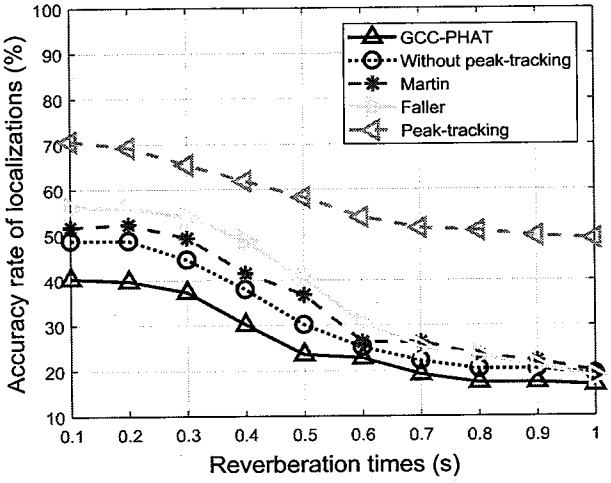


图 3.19 135° (0 dB SNR)时的平均正确率

Figure 3.19 Average accuracy at 135° and 0 dB SNR

表 3.3 时延差估计的均方根误差

Table 3.3 The RMS error (in time-delays) of the algorithm

T_{60}	Algorithm	Time delay (1/fs s) SNR = 20 dB				Time delay (1/fs s) SNR = 10 dB				Time delay (1/fs s) SNR = 0 dB			
		12(0°)	10(30°)	6(60°)	0(90°)	12(0°)	10(30°)	6(60°)	0(90°)	12(0°)	10(30°)	6(60°)	0(90°)
600 ms	GCC-PHAT	6.33	5.43	4.76	1.24	7.37	6.47	4.78	1.66	9.13	8.39	5.00	1.69
	Martin	4.71	4.47	4.05	1.01	6.31	6.45	4.52	1.94	7.67	7.63	5.57	1.63
	Faller	4.17	4.63	3.76	0.64	6.29	6.20	3.54	1.71	7.55	6.93	4.47	1.60
	Proposed1	5.66	5.09	4.70	1.15	7.11	6.43	4.80	1.65	8.25	7.13	5.16	1.73
	Proposed2	**3.47**	**2.92**	**1.59**	**0.58**	**4.84**	**4.09**	**2.74**	**0.85**	**5.34**	**5.15**	**3.18**	**0.96**
700 ms	GCC-PHAT	6.97	6.59	5.51	1.27	7.39	7.31	6.09	1.82	9.57	7.37	6.78	1.93
	Martin	5.92	4.93	3.37	1.21	6.02	6.12	6.07	1.72	8.93	6.72	6.22	1.89
	Faller	5.86	5.03	3.15	1.14	5.12	6.18	4.63	1.40	8.65	6.47	6.05	1.51
	Proposed1	6.67	5.50	3.85	1.28	6.34	6.58	5.53	1.55	8.87	6.69	6.58	1.61
	Proposed2	**3.90**	**3.14**	**1.61**	**0.85**	**4.85**	**4.77**	**3.58**	**1.05**	**6.19**	**5.81**	**4.49**	**1.17**
800 ms	GCC-PHAT	7.55	6.57	5.66	1.26	7.76	7.05	6.62	1.84	9.61	7.98	7.47	1.96
	Martin	6.71	5.60	4.14	1.12	6.84	6.99	6.66	1.73	9.06	7.71	6.75	1.87
	Faller	6.73	5.62	4.03	1.18	6.75	7.01	5.91	1.47	9.17	7.06	6.15	1.76
	Proposed1	6.90	5.76	4.74	1.17	7.88	7.26	6.41	1.47	9.23	7.81	6.58	1.83
	Proposed2	**4.28**	**4.02**	**2.53**	**1.06**	**5.36**	**5.04**	**4.01**	**1.07**	**6.53**	**5.94**	**4.80**	**1.25**
900 ms	GCC-PHAT	7.65	7.05	5.69	1.49	8.57	7.39	6.90	1.86	9.70	7.98	6.95	2.09
	Martin	7.08	6.59	4.38	1.36	7.69	7.59	6.69	1.76	8.78	8.00	6.03	2.23
	Faller	7.21	6.47	4.21	1.38	7.44	7.40	6.11	1.35	8.49	7.94	5.48	1.97
	Proposed1	7.55	6.55	4.92	1.30	8.15	7.56	6.12	1.60	8.81	7.66	6.20	1.95
	Proposed2	**5.90**	**5.43**	**2.84**	**1.07**	**6.05**	**5.48**	**4.16**	**1.26**	**6.89**	**6.20**	**4.88**	**1.47**
1 s	GCC-PHAT	8.65	7.60	6.09	1.38	8.96	8.45	6.92	1.93	9.95	8.97	6.98	2.21
	Martin	7.65	7.65	4.39	1.33	7.60	8.08	7.01	1.88	9.09	8.44	5.79	2.03
	Faller	7.30	7.30	4.31	1.33	7.71	7.84	6.15	1.41	8.85	7.56	5.36	1.99
	Proposed1	8.00	7.32	4.75	1.36	8.35	8.07	6.18	1.44	9.10	7.75	6.26	1.98
	Proposed2	**6.18**	**6.02**	**2.96**	**1.12**	**6.34**	**6.04**	**4.24**	**1.34**	**7.07**	**6.45**	**4.88**	**1.53**

3.5.2 混响抑制实验

在 3.4 节中已经介绍了 CDR 估计算法的内容，将 CDR 估计值构建一个维纳滤波器，就可以用做混响抑制的增益函数：

$$G(\lambda,\mu)=\sqrt{\frac{CDR(\lambda,\mu)}{CDR(\lambda,\mu)+1}} \qquad \text{...（3.26）}$$

图 3.20 所示为利用本文算法进行混响抑制的一个时域图，其中最上面图为纯净语音，中间为纯净语音卷积了一个来自 180 度 HRTF[129]的混响信号，最下面为处理后的语音信号。从时域图可以明显看出本文算法很好的还原了原始时候的波形。为了对语音质量做进一步的客观评价，我们利用 Perceptual Evaluation of Speech Quality (PESQ)[130]分数来评估不同的 CDR 估计函数的表现。PESQ 是一种广泛使用的语音质量感知评价模型。双耳的混响数据选自 Aachen Impulse Response (AIR) 库[129]，该数据库包含了一组不同距离（1m，2m，3,m）不同角度（-90 到 90 度）下的双耳头相关传递函数（HRTF）。纯净语音信号选自 IEEE 语料库[38]，之后卷积上不同的 HRTF，获得最后的双耳信号。我们首先计算原始信号（Unprocessed）的 PESQ 分数，然后本文算法（Proposed）分别与 Schwarz 和 Kellermann 等人提出的不依赖 ITD 信息的方法（Nodoa）；和依赖 ITD 信息的 CDR 估计方法，其中 ITD 估计方法分别为：GCC-PHAT, Martin 模型和 Faller 模型。图 3.21-3.23 所示为不同角度下的平均（100 条语句）PESQ 分数。所有角度

下的平均 PESQ 分数列于表 3.4 中。从结果可以看出：当距离为 1m 时，Martin 模型和 Faller 模型与本文算法的结果类似，而随着间距的加大，本文算法的优势越大。因为随着距离的增大，CDR 值相对较小，所以结果表明本文算法能够取得更好的语音质量尤其在更低的 CDR 环境中。

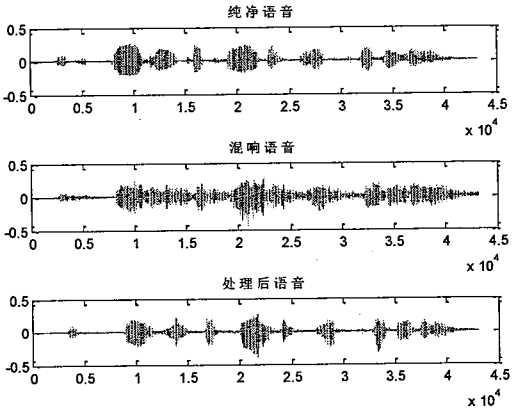


图 3.20 混响抑制算法处理前后时域图

Figure 3.20 Time-domain signal processed by our algorithm

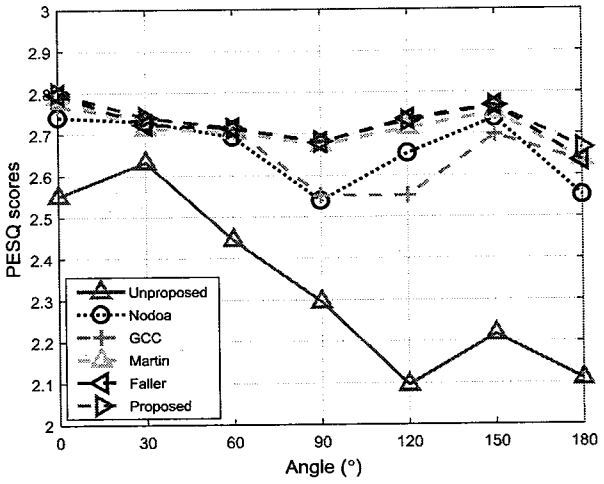


图 3.21 1m 时的平均 PESQ 分数

Figure 3.21 Average PESQ scores at distance of 1m

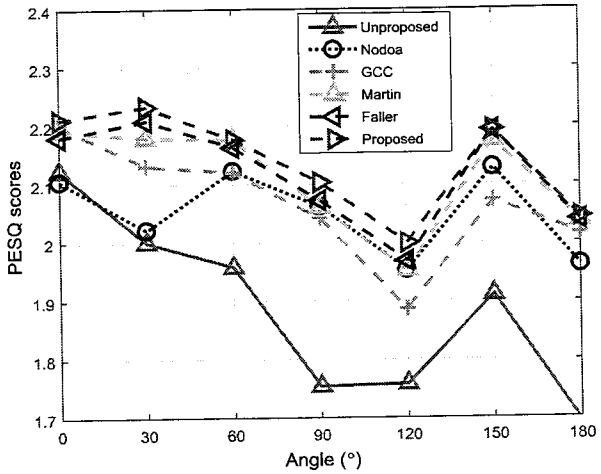


图 3.22 2m 时的平均 PESQ 分数

Figure 3.22 Average PESQ scores at distance of 2m

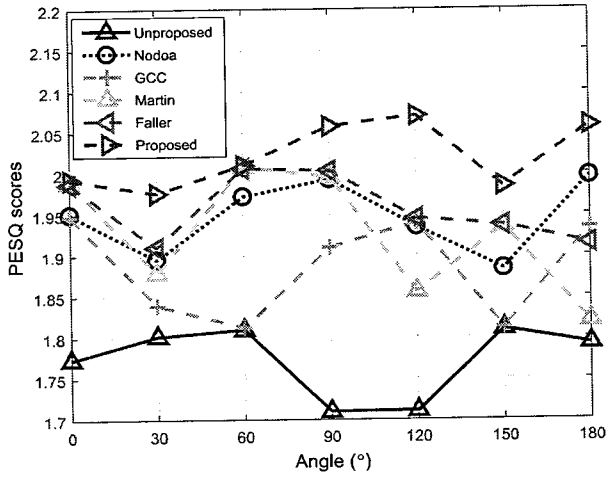


图 3.23 3m 时的平均 PESQ 分数

Figure 3.23 Average PESQ scores at distance of 3m

表 3.4 表标题所有角度下的(0^o-180^o)平均 PESQ 分数

Table 3.4 The average PESQ scores in all directions (0^o-180^o)

Distance	Unprocessed	Non-DOA	GCC	Martin	Faller	Proposed
1m	2.32	2.66	2.66	2.71	2.72	2.73
2m	1.89	2.06	2.07	2.11	2.11	2.13
3m	1.77	1.94	1.88	1.92	1.95	2.02

3.6 小结

在本章中，我们首先提出一种简单有效的“优先效应”模型，该模型通过一种峰值平滑策略进行可靠信息的选取，随后利用峰值平滑后的相干函数的数值，来进行时延差估计。实验结果表明该方法明显优于传统的时延估计方法。随后估计的时延差被用于 CDR 值得估计。实验结果表明该 CDR 估计函数相较于传统方法能够取得更高的 PESQ 分数。

第 4 章　双耳多麦克风语音分离算法

4.1 引言

“鸡尾酒会”问题是一个经典且富有挑战性的问题，所谓“鸡尾酒会”问题指的是正常人可以从多人交谈的背景噪声里选择并识别特定的说话人。该效应揭示了人类双耳听觉系统强大的能力。为了解决这个问题，一系列的算法被广大学者提出。

波束形成算法是一种最为简单的语音分离技术，该方法具有简单有效且滤波后语音失真小的特点，使得这项技术成为目前听力设备和远场语音识别设备的主流方案。波束形成方法又可以分为固定波束形成方法和自适应波束形成方法。固定波速形成中主要包括传统的延迟相加(DS, Delay and sum)波束形成 ,差分麦克风阵列，以及在 MVDR 理论基础上建立的超指向波束形成。这类固定波束形成方法的优点是鲁棒且语音失真小，但是主瓣宽，在麦克风数量较少时，效果欠佳。自适应波束形成能够按照一定的准则自适应的调整权重。其中包括 MVDR,GSC,MAXsnr，MWF（multi-channel wiener filter）等等。需要指出的是，以上的这些不同准则推导出的增益公式，本质上并没有太大区别，只是各类形式的滤波器在目标语音失真和降噪抑制的平衡上有所不同。更关键的在于公式里面的干扰加噪声的协方差矩阵以及导向向量的求取。

另外一种主流的语音分离方法称为独立成分分析（Independent Component Analysis，ICA），传统的频域 ICA 算法存在两个明显的缺点：不能确定信号的能量(方差)和排序问题，尤其是排序问题对语音分离影响很大，在 ICA 处理过程中虽然能够在每个频点将声源分开，但是如何将属于同一声源的所有频点序列正确排在一起，是一个难以解决的问题。为了解决 ICA 排序问题，一些方法被提出，主要思路为利用频率间的相关性或空间信息[131-133]。另外一类更加有效解决频域 ICA 排序问题的方法为独立向量分析（Independent vector Analysis，IVA）的技术被提出[134-136]。IVA 将整个频段作为一个整体考虑，该方法有效解决了 ICA 分离过程中的序列模糊问题。

基于双耳模型的聚类算法是一种在双耳中广泛使用的算法，该算法假设 ITD 和 ILD 的分布服从高斯分布，利用这些特征来联合估计 Mask[137]。近年来，基于 MASK 估计与波束形成方法结合的思路被广泛研究，成为一种主流的研究方法。基于 CGMM(complex Gaussian mixture model)的导向向量估计方法[41]是一种典型的无监督聚类方法，该方法通过假设噪声信号与带噪语音的协方差矩阵服从复高斯分布。通过 EM 算法，来求解噪声的协方差矩阵。但是，该方法对于语音分离任务的效果有待提升，且 EM 算法存在着对初始值敏感的问题。同样基于深度神经网络的 Time-frequency mask 与 MVDR 结合的方法也被广泛研究[42]，该类方法一般为有监督方法。该方法与单声道 Mask 预测的方法类似，通过单个声道或每个声道独立进行 Mask 的估计来获取噪声的协方差矩阵，从而与 CGMM 类似的方法获得目标。由于深度学习方法摒弃了相位信息，仅仅使用幅度信息来进行 Mask 估计，基于深度学习的方法同样不能解决人声干扰的情况，而且由于该类方法属于有监督学习，存在着数据不匹配等问题。

针对多说话人分离的情况下，以上方法的问题，本文提出一种基于 Mask 估计与独立向量分析的协方差矩阵估计方法，并与多通道维纳滤波器结合，用于增强分离后的信号。

本章内容安排如下：第二节回顾了多通道滤波器理论知识和独立向量分析理论基础。在此基础上，提出本文的基于 IVA 的目标语音协方差矩阵获取算法并与多通道维纳滤波器结合。

4.2 多通道滤波器回顾

4.2.1 多通道维纳滤波器理论

多通道维纳滤波器（Multi-channel wiener filter，MWF）是基于 MMSE 准则的自适应滤波，理论上等价于 MVDR 波束形成算法级联一个单通道维纳滤波算法。MVDR,GSC 都可以理解为多通道维纳滤波器（PMWF）的一种特殊情况。

假设接收信号为：

$$y(t, f) = x(t, f) + n(t, f) \quad \text{...(4.1)}$$

其中 y 为接收信号，x 为目标信号，n 为噪声信号，t 和 f 分别表示时间帧和频点，

为了推导方便，下文的公式中中省略 t 和 f。MWF 方法基于 MMSE 准则求取滤波器系数矢量：

$$W_{MWF} = \arg\min_{w} E\{|w^H y - x_1|\} \quad \text{...(4.2)}$$

预测的信号与真实信号的误差表示为：

$$\varepsilon = w^H y - x \quad \text{...(4.3)}$$

该误差的损失函数可以定义为最小二乘的形式，即：

$$J = E\{|\varepsilon|^2\} = (w^H y - x)(w^H y - x^*) \quad \text{...(4.4)}$$

$$= w^H E\{yy^H\} W - 2E\{xw^H y\} + E\{xx^*\} \quad \text{...(4.5)}$$

令梯度为 0，可以得到：

$$\nabla_w J = 2R_{yy} W - 2r_{xy} \quad \text{...(4.6)}$$

其中:

$$R_{yy} = E\{yy^H\} \quad \text{...(4.7)}$$

表示接收信号的协方差矩阵，而：

$$r_{xy} = E\{xy\} \quad \text{...(4.8)}$$

表示接收信号与目标信号件的互相关矩阵。则其解为：

$$W = R_{yy}^{-1} r_{xy} \quad \text{...(4.9)}$$

假设目标信号的导向向量已知，此时信号模型为：

$$y = dx + n \quad \text{...(4.10)}$$

其中 d 为目标信号的导向向量。此时：

$$R_{yy} = E\{yy^H\}$$

$$= \delta_s^2 dd^H + R_n \quad \text{...(4.11)}$$

假设目标信号与噪声不相关:

$$r_{xy} = \delta_s d \quad \text{...(4.12)}$$

此时的滤波器可以写成以下形式：

$$W = [\delta_s^2 dd^H + R_n]^{-1} \delta_s d \quad \text{...(4.13)}$$

根据矩阵求逆引理[138]，(4.13)可以写成如下形式：

$$W = [R_n^{-1} - \frac{\delta_s R_n^{-1} d d^H R_n^{-1}}{1+\delta_s d^H R_n^{-1} d}]\delta_s d \quad \text{...（4.14）}$$

$$=[1 - \frac{\delta_s d R_n^{-1} d^H}{1+\delta_s d^H R_n^{-1} d}]\delta_s R_n^{-1} d \quad \text{...（4.15）}$$

$$=[\frac{\delta_s}{1+\delta_s d^H R_n^{-1} d}] R_n^{-1} d \quad \text{...（4.16）}$$

$$=[\frac{\delta_s}{\delta_s + (d^H R_n^{-1} d)^{-1}}] \frac{R_n^{-1} d}{d^H R_n^{-1} d} \quad \text{...（4.17）}$$

MVDR 波束形成器中噪声的输出功率可以表示为：

$$\delta_n = W_{mvdr}^H R_n W_{mvdr} = \frac{1}{d^H R_n^{-1} d} \quad \text{...（4.18）}$$

而由于在 MVDR 波束形成中，目标语音无失真通过，所以，MVDR 滤波前后目标信号功率不变。为此将(4.18)代入 4.17 中，可以得到：

$$W_{MWF} = [\frac{\delta_s}{\delta_s + \delta_n}] \frac{R_n^{-1} d}{d^H R_n^{-1} d} \quad \text{...（4.19）}$$

可以看到，此时公式的第一项为维纳后置滤波，第二项为标准的 MVDR 波束形成公式。所以，标准的 MWF 可以认为是 MVDR 波束形成器加上单声道的维纳后置滤波器。所以，本质上 MWF 相较于 MVDR，能够提高输出信噪比。为了引入语音失真和噪声抑制的权衡，MWF 算法通过在最小均方误差准则中加入了加权因子 μ 得到 SDW-MWF 算法，该算法的准则为：

$$W_{SDW-MWF} = \arg\min_w E\{|w^H x - x_1| + \mu |w^H n|\} \quad \text{...（4.20）}$$

最终的解为：

$$W_{SDW-MWF} = \left[R_{xx} + \mu R_{nn}\right]^{-1} R_{xx} u_1 \quad \text{...（4.21）}$$

通过加权因子 μ 的不同取值，MWF 可以有不同的变体，而 MVDR,GSC 是 MWF 中的一种特殊情况，由此将以上滤波器联系起来。

4.2.2 基于辅助函数的独立向量分析

考虑一个欠定盲源分离算法，即麦克风数目大于声源数目的情况。为了提高传统梯度下降 IVA 算法的收敛速度和稳定性，基于辅助函数（Auxiliary function）的 IVA 算法被提出[136]。与传统的梯度下降 IVA 算法直接对代价函数操作相比，辅

助函数通过优化辅助变量达到逼近代价函数的最优解。辅助函数技术能够保证代价函数单调下降，为此关键在于如何设计辅助函数。

对于一个辅助函数，应该满足以下条件（假设 D(θ)代价函数，G(θ,α)为 D(θ)的辅助函数）：

1. 容易优化，这个是显而已见的
2. $D(\theta) = \min_{\alpha} G(\theta, \alpha)$
3. 存在θ_1，使得 $D(\theta^{(I)}) = G(\theta^{(I)}, \alpha^{(I+1)})$

下图揭示了辅助函数的工作流程，首先我们有:

$$D(\theta^{(I)}) = G(\theta^{(I)}, \alpha^{(I+1)}) \quad \text{...（4.22）}$$

同时，我们可以得到:

$$G(\theta^{(I)}, \alpha^{(I+1)}) \geq G(\theta^{(I+1)}, \alpha^{(I+1)}) \quad \text{...（4.23）}$$

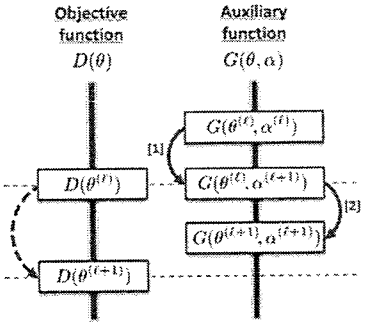
通过条件 2 又有:

$$G(\theta^{(I+1)}, \alpha^{(I+1)}) \geq D(\theta^{(I+1)}) \quad \text{...（4.24）}$$

由此我们得到:

$$D(\theta^{(I)}) \geq D(\theta^{(I+1)}) \quad \text{...（4.25）}$$

即在通过优化辅助函数的过程中，使得原始的代价函数单调下降。



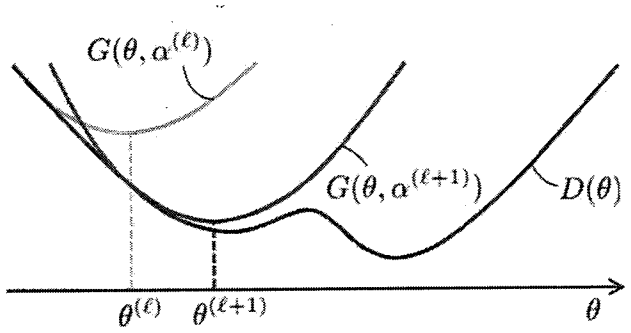


图 4.1 辅助函数原理框图[139]

Figure 4.1 Block diagram of the auxiliary function

在频域中，假设接收信号为：

$$x(\omega) = A(\omega)s(\omega) \quad \text{...（4.26）}$$

其中 $A(\omega)$ 为混合矩阵，而估计的信号表示为：

$$y(\omega) = W(\omega)x(\omega) \quad \text{...（4.27）}$$

其中 $A(\omega)$ 为混合矩阵，$W(\omega)$ 为解混矩阵。

$$W(\omega)= (\mathrm{w}_1(\omega)\cdots\mathrm{w}_k(\omega))^h \quad \text{...（4.28）}$$

其中 h 代表厄米特转置，$s(\omega)$，$x(\omega)$，$y(\omega)$ 分别代表源信号，接收信号和估计的源信号：

$$s(\omega)= (\mathrm{s}_1(\omega)\cdots\mathrm{s}_k(\omega))^t \quad \text{...（4.29）}$$

$$x(\omega)= (\mathrm{x}_1(\omega)\cdots\mathrm{x}_k(\omega))^t \quad \text{...（4.30）}$$

$$y(\omega)= (\mathrm{y}_1(\omega)\cdots\mathrm{y}_k(\omega))^t \quad \text{...（4.31）}$$

其中 k 代表声源的数目，t 代表向量的转置。在 IVA 中，我们基于最大似然准则定义如下代价函数：

$$J(W) = \sum_{k=1}^{K} E[G(y_k)] - \sum_{\omega=1}^{N_\omega} \log\left|\det W(\omega)\right| \quad \text{...（4.32）}$$

其中 E 代表数学期望，y_k 代表估计的信号：

$$y_k = (y_k(1)\cdots y_k(N_\omega)) \quad \text{...（4.33）}$$

而 $G(y_k)$ 代表对照函数且：

$$G(y_k) = -\log p(y_k) \quad \text{...（4.34）}$$

其中 $p(y_k)$ 代表声源的概率密度分布函数。在本文中，对照函数选用球形对照函数[104]:

$$G(y_k) = G_R(r_k) \quad \text{...（4.35）}$$

$$r_k = \|y_k\|_2 = \sqrt{\sum_{\omega=1}^{N_\omega} |y_k|^2} \quad \text{...（4.36）}$$

为了定义适合 IVA 代价函数的辅助函数，我们做如下定义:

定义 1 随机变量 z 的一组实值函数 S_G 定义为:

$$S_G = \{G(z) | G(z) = G_R(\|z\|_2)\} \quad \text{...（4.37）}$$

其中 $G_R(r)$ 是一个连续函数，而 $G_R'(r) / r$ 处处连续且当 r 大于等于 0 时，单调递减。$G_R(r)$ 是一个与声源先验分布有关的量，在本文中，我们取 $G_R(z)=r$ ，其中:

$$r=\|z\|_2 \quad \text{...（4.38）}$$

此时声源的概率密度函数对应着拉普拉斯变换。基于上面的定义，我们引入以下定理来定义辅助函数。

定理 1 对于任意的 $G(z)=G_R(\|z\|_2) \in S_G$

$$G(z) \le \frac{G_R'(r_0)}{2r_0}\|z\|_2^2 + (G_R(r_0) - \frac{r_0 G_R'(r_0)}{2}) \quad \text{...（4.39）}$$

等式成立，当且仅当:

$$r_0 = \|z\|_2 \quad \text{...（4.40）}$$

定理 2 对于任意的 $G(z)=G_R(\|z\|_2) \in S_G$

$$Q(W,V) = \sum_{\omega=1}^{N_\omega} Q_\omega(W(\omega), V(\omega)) \quad \text{...（4.41）}$$

$$Q_\omega(W,V) = \frac{1}{2}\sum_{\omega=1}^{N_\omega} W_k^h V_k(\omega) W_k(\omega) - \log|\det W(\omega)| + R \quad \text{...（4.42）}$$

其中:

$$V_k(\omega) = E[\frac{G_R'(r_k)}{r_k} x(\omega) x^h(\omega)] \quad \text{...（4.43）}$$

其中 r_k 为正的随机变量，$V(\omega)$ 代表对于所有的 k 产生的一系列的 $V_k(\omega)$,而 V 代表

对于所有的 k 和 ω 产生的 $V_k(\omega)$ 向量。而 R 为一个常量。然后：

$$J(\omega) \le Q(W,V) \quad \text{...（4.44）}$$

等式成立当且仅当：

$$r_k = \|y_k\|_2 = \sqrt{\sum_{\omega=1}^{N_\omega} \left|w_k^h(\omega)x(\omega)\right|^2} \quad \text{...（4.45）}$$

更多详细的推导和证明过程见文献[137]。

最后基于辅助函数的 IVA 算法的参数更新过程为：

$$r_k = \sqrt{\sum_{\omega=1}^{N_\omega} \left|w_k^h(\omega)x(\omega)\right|^2} \quad \text{...（4.46）}$$

$$V_k(\omega) = E[\frac{G_R'(\mathrm{r_k})}{\mathrm{r_k}} x(\omega)x^h(\omega)] \quad \text{...（4.47）}$$

$$w_k(\omega) \leftarrow (W(\omega)V_k(\omega))^{-1}e_k \quad \text{...（4.48）}$$

$$w_k(\omega) \leftarrow w_k(\omega)/\sqrt{w_k^h(\omega)V_k(\omega)w_k(\omega)} \quad \text{...（4.49）}$$

4.3 本文算法

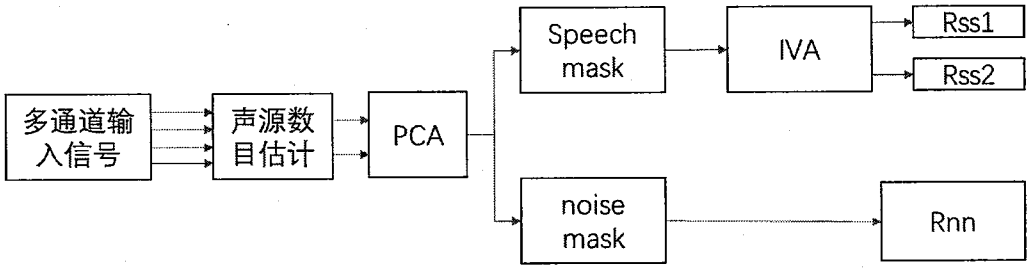


图 4.1 本文算法结构框图

Figure 4.2 Block diagram of the proposed algorithm

以两个声源为例，说明本文算法的结构框图，首先估计语音数目，然后进行 PCA 预处理，在 PCA 处理后的信号中利用 DNN 来估计语音和噪声的 Mask；将语音的 Mask 乘以 PCA 处理后的信号后，进行 IVA 估计解混矩阵，从而获得两个语音信号的协方差矩阵 Rss1 和 Rss2；通过噪声的 Mask 来获取噪声的协方差矩阵 Rnn。最后联合多通道维纳滤波器获得增强后的两个信号。

4.3.1 声源数目估计算法

在声源分离任务中，声源数目是一个需要预先告知的先验信息，现有的大部分算法都假设声源数目已知。信号源数目估计的主要方法有基于直方图方法、信

息论准则 [140]、盖尔圆法（GDE）[141]以及正则相关技术（CCT）[142]。信息论准则、盖尔圆法（GDE）以及正则相关技术等方法虽然提供了一系列的理论来估计声源数目，但是在实际系统中需要考虑计算量等问题，且在干扰噪声环境下的鲁棒性有待提升，这些问题使得这些方法不适合在实际环境下使用。

基于直方图方法的方法是一类最简单的声源数目估计方法，该方法通过将多帧的每个频点的 DOA 估计值进行统计分析，画图直方图，直方图中呈现明显峰值的位置即代表声源出现。通过统计直方图的峰值的个数即可知道声源的数目。该方法虽然简单，但是在噪声和混响环境下峰值可能并不明显，如何确定阈值从而判断是否为声源是一个需要解决的问题。

现假设左右耳各有两个麦克风，即双耳共有四个麦克风，并且假设声源数目小于等于 4。为了在实际环境下估计声源数目，假设各个声源处于不同的方位下，本文利用每个频点估计的 DOA 信息提出一种简单有效的基于直方图的声源数目估计方法。由于混响的存在，如果直接将 DOA 信息绘制直方图的话，会噪声直方图的峰值不尖锐，容易产生声源数目误判现象。在第三章中，我们提出一种简单鲁棒的“优先效应”模型，即通过峰值平滑技术来实现高相关性线索的提取，只是在本章中，为了在很短的数据情况下能够获得较大的特征量，我们在每个频点进行 DOA 信息的判断，本文的声源数目估计算法具体实现步骤如下：

1. 获取无混响环境下的离线 HRTF。提取不同方位下的 ITD 和 ILD 信息，建立离线库；
2. 实际混响环境下，通过对相干函数的峰值平滑，提取可靠的频点，并与步骤 1 中的离线库进行对比，估计每个频点的 DOA 信息；

$$if\ |\Gamma_{X1X2}(\lambda,\mu)| < |Peak(\lambda-1,\mu)|$$

$$Peak(\lambda,\mu) = \alpha_2 \cdot Peak(\lambda-1,\mu)$$

else

$$Peak(\lambda,\mu) = \alpha_1 \cdot Peak(\lambda-1,\mu) + (1-\alpha_1) \cdot \Gamma_{X1X2}(\lambda,\mu)$$

End

DOA estimation in each frequency bin

3. 画出 DOA 信息的直方图 R(n)，找出直方图中的最大值和最小值。将直方图所有数值减去最小值，然后除以最大值。得出概率值 P(n):

$$P(n) = \frac{R(n) - \min R(n)}{\max(R(n))} \quad \text{...（4.50）}$$

4. 设置阈值，超过阈值的被认为是一个声源，从而得出最后的声源数目。

4.3.2 语音与噪声 Mask 估计算法

本实验中利用 DNN 训练一个 Speech mask 和 Noise mask，参考文献[143]训练一个区分性的神经网络。该神经网络的 cost function 为：

$$\text{loss}=(M_s|Y|-|X|)^2+(M_n|Y|-|N|)^2 \quad \text{...（4.51）}$$

其中 X 为纯净语音的幅度谱，而 N 为噪声的幅度谱，Y 为经过 PCA 处理后的 noisy 信号的幅度谱。Ms 和 Mn 分别为本文需要估计的语音和噪声的 Mask。理想 Mask 使用 IRM：

$$IRM=\frac{s(t,w)^2}{s(t,w)^2+n(t,w)^2} \quad \text{...（4.52）}$$

最后，噪声的协方差矩阵为：

$$\therefore R_n=\frac{\sum_{t=L-1}^{t=L} M_n(t,w)X(t,w)x(t,w)^H}{\sum_{t=L-1}^{t=L} M_n(t,w)} \quad \text{...（4.53）}$$

本实验中，语音数据选自 TIMIT 语料库，噪声选自 Noise-92 噪声库，信号采样率为 16 kHz。然后将纯净语音与噪声相加，信噪比随机选取为 0-10dB，从而获得 5w 条带噪的语音数据。我们首先对信号进行分帧加窗处理，帧长为 32 ms，overlap 为 75%，随后利用 STFT 变换到频域。

神经网络输入层为 257 个神经元，三个隐含层为 500 个神经元，两个输出层各包括 257 个神经元。神经网络的优化方法采用 Adam，激活函数采用 PReLU，代价函数为公式(4.51)。

4.3.3 声源分离算法

在引言中，我们回顾了多通道语音分离的常用方法，并指出基于 Mask 预测与多通道滤波器结合的思想成为目前的主流方法，但是无论是基于 CGMM 还是基于神经网络的方法，对于多说话人下的语音分离任务都难以表现欠佳。为此，本文提出一种基于 IVA 的目标信号协方差矩阵估计方法。

在 4.2.2 节中，我们已经对基于辅助函数的 IVA 算法进行了介绍，并得出最后的迭代公式：

$$r_k = \sqrt{\sum_{\omega=1}^{N_\omega} \left| w_k^h(\omega) x(\omega) \right|^2} \quad \text{...（4.54）}$$

$$V_k(\omega) = E[\frac{G_R^{'}(\mathrm{r_k})}{\mathrm{r_k}} x(\omega) x^h(\omega)] \quad \text{...（4.55）}$$

$$w_k(\omega) \leftarrow (W(\omega) V_k(\omega))^{-1} e_k \quad \text{...（4.56）}$$

$$w_k(\omega) \leftarrow w_k(\omega) / \sqrt{w_k^h(\omega) V_k(\omega) w_k(\omega)} \quad \text{...（4.57）}$$

最后估计的第 k 个语音信号为：

$$S_{\mathrm{k}} = Y \cdot W_{\mathrm{k}} \quad \text{...（4.58）}$$

其中 W 为 IVA 估计的解混矩阵，Y 为 PCA 处理后的信号。随后可以获得第 k 个目标语音的协方差矩阵：

$$R_{\mathrm{s}}^k = \frac{\sum_{\mathrm{t=L-1}}^{\mathrm{t=L}} S_{\mathrm{k}}(t, w) S_{\mathrm{k}}(t, w)^H}{L} \quad \text{...（4.59）}$$

则干扰加噪声的协方差矩阵 $R_{\mathrm{i+n}}=R_{\mathrm{n}}+R_{\mathrm{s\neq k}}$ 等于噪声的协方差矩阵加上非目标信号的协方差矩阵。获得两个协方差矩阵后，我们代入到公式(4.21)中得到最后的第 k 个声源的多通道维纳滤波器的增益函数：

$$w_{NMF}^k = [R_s^k + \mu R_{i+n}] R_s^k \mu_1 \quad \text{...（4.60）}$$

式中 μ 取 1.μ_1代表参考麦克风为麦克风 1.

4.4 实验评价

4.4.1 声源数目估计

按照 4.3.1 节中所介绍的基于直方图的声源数目估计算法。利用我们选取混响环境下的双耳 HRTF 数据库[144]合成多个声源的混合信号。以三个声源为例，图 4.3-4.4 为本文算法提出的经过 4.3.1 节中介绍的经过相干函数频点选择后的 DOA 直方图和未经过频点选择的原始信号 DOA 直方图。而图 4.5-4.6 为经过后置处理后的声源存在概率。从图中可以明显看出，本文算法的峰值更为明显。图 4.7-4.10 为两个声源时的结果，同样可以看出本文算法的峰值更加的明显。

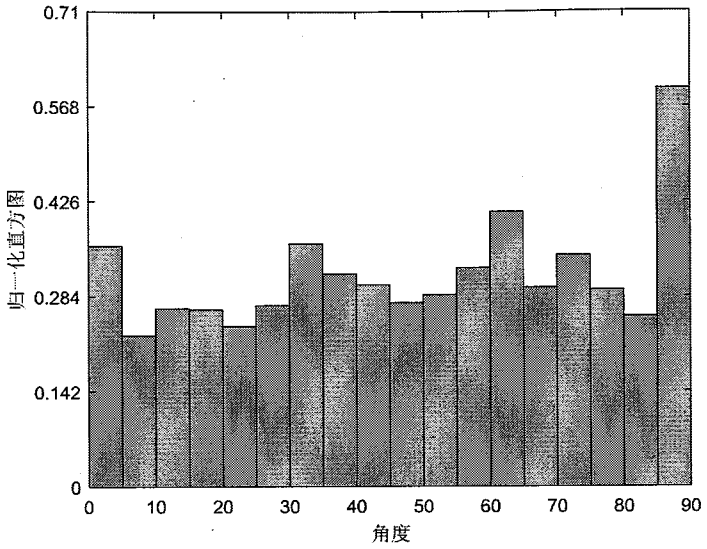


图 4.3 三声源时未经过频点选择的 DOA 直方图

Figure 4.3 DOA histogram of three source without frequency-bin selection

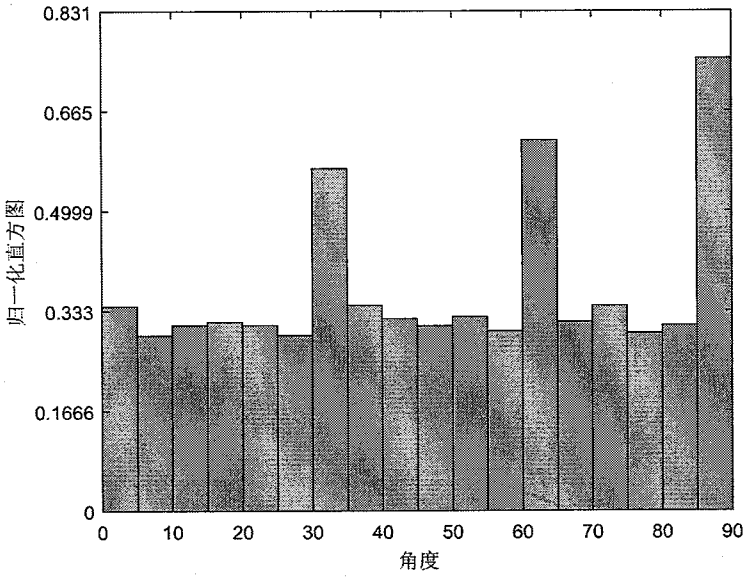


图 4.4 三声源时经频点选择后的 DOA 直方图

Figure 4.4 DOA histogram of three source with frequency-bin selection

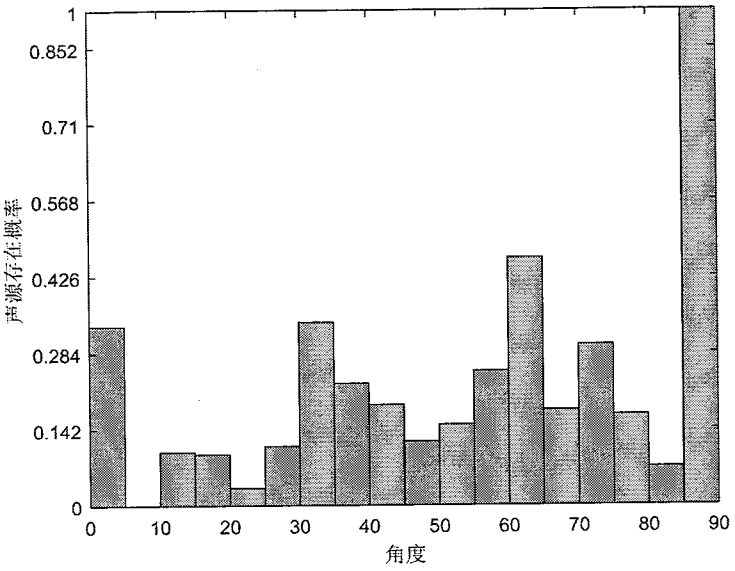


图 4.5　三声源时未经过频点选择的声源存在概率

Figure 4.5 Source presence probability of three source without frequency-bin selection

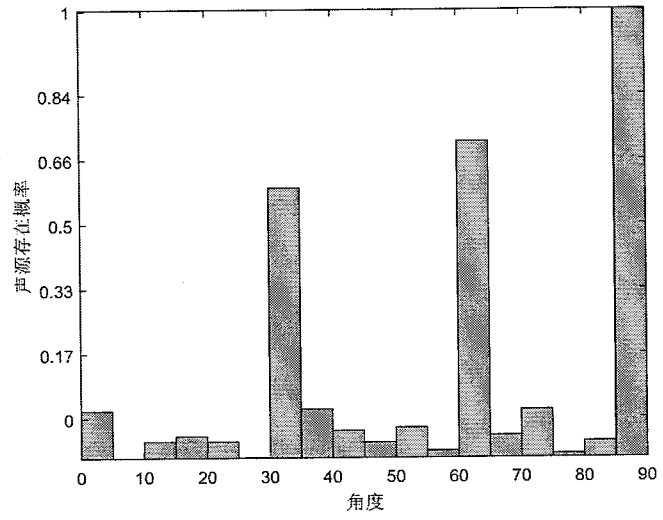


图 4.6　三声源时经过频点选择后的声源存在概率

Figure 4.6 Source presence probability of three source with frequency-bin selection

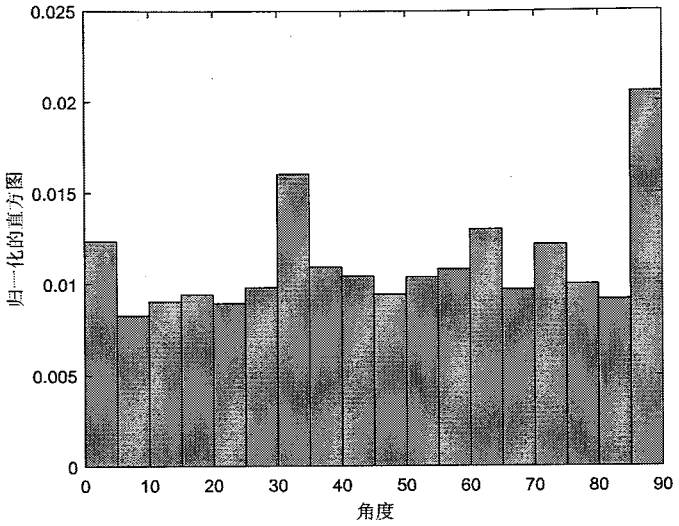


图 4.7　两声源时未经过频点选择的 DOA 直方图

Figure 4.7 DOA histogram of two source without frequency-bin selection

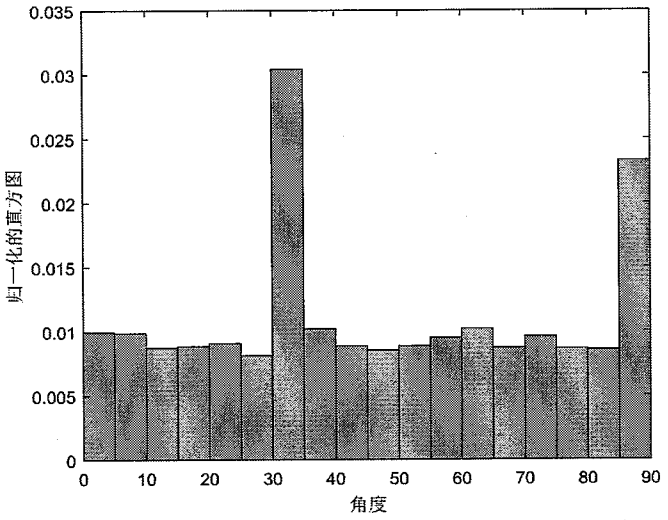


图 4.8　两声源时经频点选择后的 DOA 直方图

Figure 4.8 DOA histogram of two source with frequency-bin selection

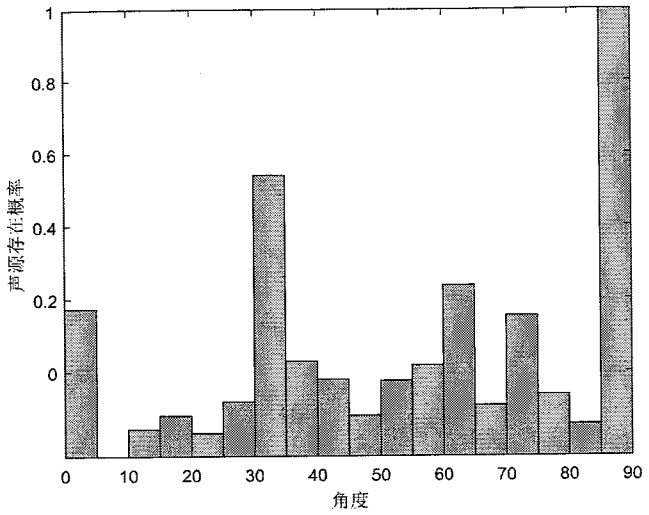


图 4.9　两声源时未经过频点选择的声源存在概率

Figure 4.9 Source presence probability of two source without frequency-bin selection

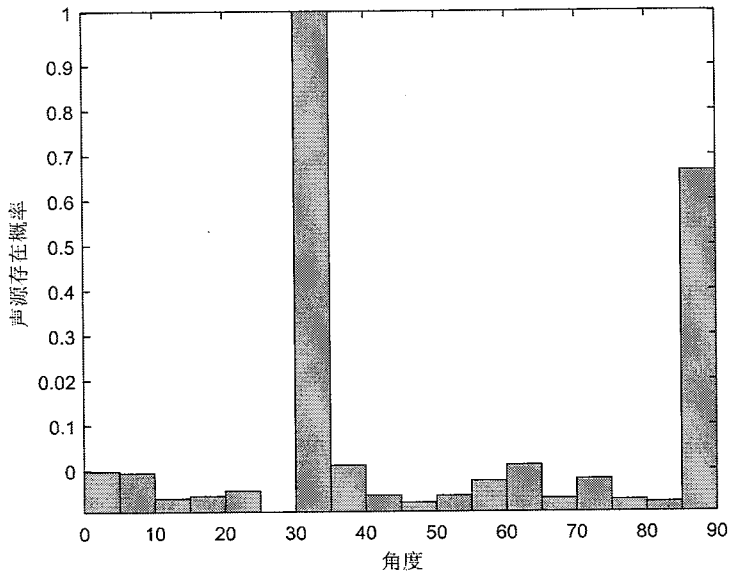


图 4.10　两声源时经过频点选择后的声源存在概率

Figure 4.10 Source presence probability of two source with frequency-bin selection

表 4.1　未经频点选择的分类正确率

Table 4.1 Classification accuracy without frequency-bin selection

混淆矩阵		预测值			
		1	2	3	4
真实	1	89.5%	9.8%	0.7%	0%
	2	9.8%	71%	18%	1.2%

值	3	2%	11%	65%	22%
	4	0.2%	9.8%	32%	58%

表 4.2 经频点选择的分类正确率

Table 4.2 Classification accuracy with frequency-bin selection

混淆矩阵		预测值			
		1	2	3	4
真实值	1	99.3%	0.7%	0%	0%
	2	1.1%	97%	1.9%	0%
	3	0%	2.7%	95%	2.3%
	4	0%	1.5%	10.5%	88%

表 4.1 和 4.2 分别为未经频点选择和经过本文算法处理后的的声源数目估计正确率，其中信噪比为 20dB，最小角度间隔为 10 度，直方图选取阈值选取为 0.4，即超过 0.4 的概率选取为声源。从表中可以明显看出，本文算法处理能够有效提高声源个数识别的正确率。

4.4.2 声源分离实验

为了评估分离的效果，我们选取混响环境下的双耳多麦克风数据[144]。在该 HRTF 库中，存在 8 个麦克风(左右各四个)，我们选取左右侧助听器上面的前面两个麦克风，一共组成 4 个通道的数据。

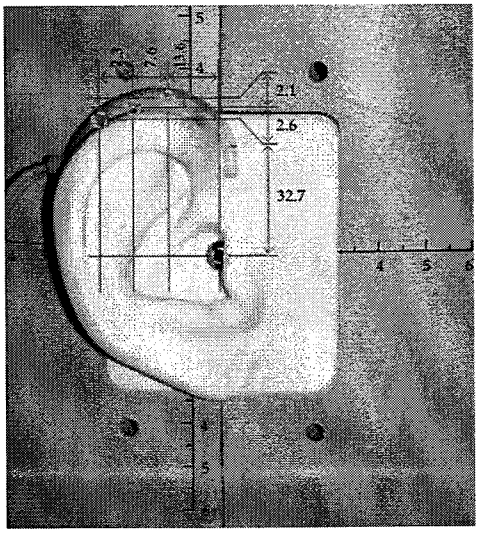


图 4.11 HRTF 库中麦克风位置[144]

Figure 4.11 The microphone position in HRTF database

纯净语音取自于 TIMIT 数据集[145]，通过卷积不同方位的 HRIR，我们获得各个方位的目标声源，通过将不同方位的声源叠加起来，获得多个声源信号。最后添加 diffuse noise 噪声，获得最后的观测信号，其中噪声分别为 Speech Shaped Noise（SNN），Car noise 和 Babble noise。diffuse noise 的获取方法为：将噪声卷积所有方位的 HRIR，累加后进行幅度归一化。在实验中，我们分别选取数据库中的 office Ⅰ与 office Ⅱ两个房间的 HRIR 进行语音分离实验。

1）语音降噪客观评价：由于本文利用了 DNN 来估计噪声的协方差矩阵，为此验证本文算法在降低 diffuse noise 干扰的性能。语音降噪的客观评价指标使用 SNR 和 PESQ。将各算法估计的多个目标语音进行累加，进行 SNR 和 PESQ 的计算，其中参考信号为通道 1 的多语音信号。声源个数分别为 2，3，4，每种条件下各合成 30 条带噪语音，信噪比分别为 0,5,10 dB。统计所有数据的平均 SNR 和 PESQ 分数。结果见表 4.2 中，从表中可知，MESSL 方法本身并没有噪声抑制功能，所以基本没有提升的信噪比和 PESQ 分数。而 IVA 算法本身也不具备噪声抑制功能，但是由于进行了 PCA 预处理，使得该算法能够提升有限的 SNR 和 PESQ 分数，而本文算法较对比算法显著提高 SNR，PESQ 分数也较 IVA 略有提高。

表 4.2 平均 SNR 和 PESQ 分数

Table 4.2 Average SNR and PESQ scores

	SNR(dB)				PESQ			
噪声类型	Noisy	MESSL	IVA	Proposed	Noisy	MESSL	IVA	Proposed
SSN	0	0	1.71	6.82	0.98	0.98	1.05	1.19
	5	5	7.64	12.39	1.73	1.73	1.88	1.99
	10	10	12.26	16.21	1.95	1.95	2.04	2.23
Car	0	0	2.09	6.09	0.89	0.89	0.92	1.01
	5	5	6.35	13.18	1.69	1.69	1.98	1.98
	10	10	12.38	15.40	1.82	1.82	2.00	2.01
Bab	0	0	0.45	4.11	0.62	0.62	0.66	0.71
	5	5	5.69	8.23	1.18	1.18	1.22	1.34
	10	10	11.91	13.45	1.73	1.73	1.79	1.80

2）语音分离客观评价：语音分离客观评价指标选用 BSS 工具箱中常用的 SDR (signal to distortion ratio),SIR (signal to interference ratio)和 SAR(signal to artifact ratio)等三个指标[146]。假设某一个估计的信号包括三个部分：

$$S_{est} = S_t + e_{interf} + e_{artif} \quad \text{...（4.61）}$$

其中 S_{es} 表示估计的源信号，S_t 表示真实的信号，e_{interf} 表示其他干扰源引起的误差信号，e_{artif} 表示算法自身引起的误差。

三个评价方法的公式如下：

$$SAR=10\log10\frac{\left\|S_{est}-e_{artif}\right\|^2}{\left\|e_{artif}\right\|^2} \quad \text{...（4.62）}$$

$$SDR=10\log10\frac{\left\|S_t\right\|^2}{\left\|S_{est}-S_t\right\|^2} \quad \text{...（4.63）}$$

$$SIR=10\log10\frac{\left\|S_t\right\|^2}{\left\|e_{interf}\right\|^2} \quad \text{...（4.64）}$$

本文提出的算法（Proposed）分别与基于双耳模型的聚类算法(MESSL) [137]和基于辅助函数的 IVA 算法(IVA)[136]进行比较。MESSL 是一种经典的利用双耳空间信息（ITD,ILD）进行聚类的算法，文中假设不同声源的 ITD 与 ILD 分别服从高斯混合模型，利用 EM 算法来估计用于分离不同声源的 Mask。图 4.12-4.17 为两个房间下两个声源分离时的平均 SDR,SIR,SAR 值，同样 4.18-4.23 为三个声源的客观评价值（20 条数据的统计平均）。

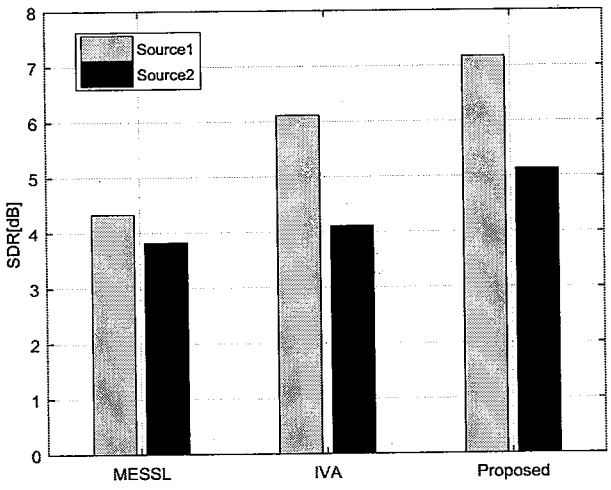


图 4.12　两个声源分离的平均 SDR 值(office Ⅰ)

Figure 4.12 Average SDR values of two source(office Ⅰ)

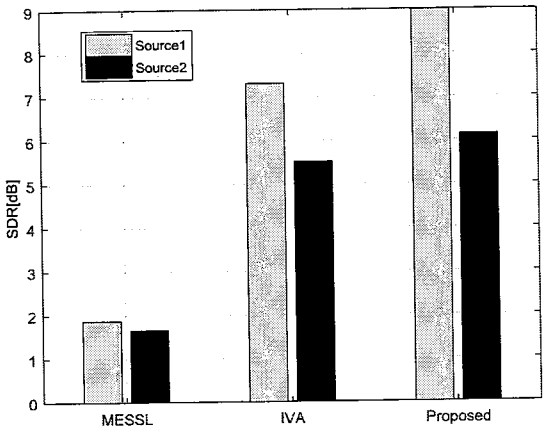


图 4.13 两个声源分离的平均 SDR 值(office Ⅱ)

Figure 4.13 Average SDR values of two source (office Ⅱ)

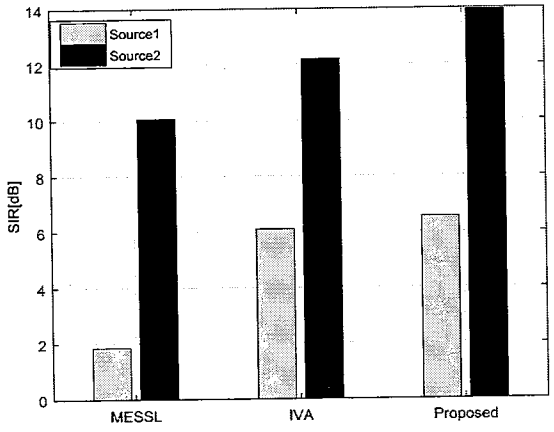


图 4.14 两个声源分离的平均 SIR 值(office Ⅰ)

Figure 4.14 Average SIR values of two source(office Ⅰ)

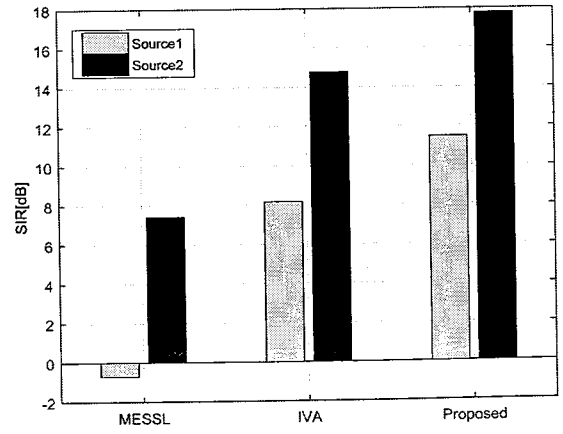


图 4.15 两个声源分离的平均 SIR 值(office Ⅱ)

Figure 4.15 Average SIR values of two source (office Ⅱ)

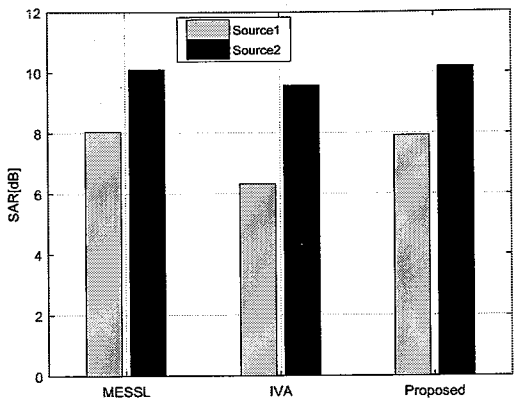


图 4.16　两个声源分离的平均 SAR 值(office Ⅰ)

Figure 4.16 Average SAR values of two source(office Ⅰ)

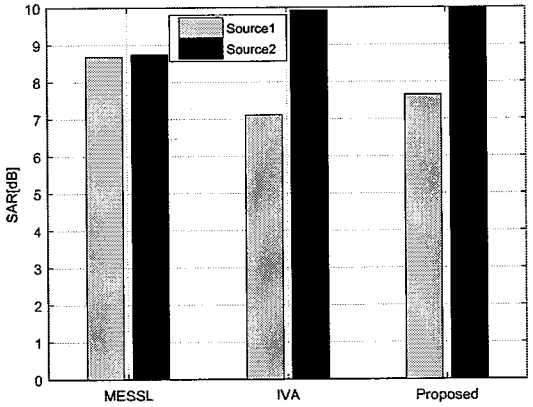


图 4.17　两个声源分离的平均 SAR 值(office Ⅱ)

Figure 4.17 Average SAR values of two source (office Ⅱ)

三个声源时的结果如下：

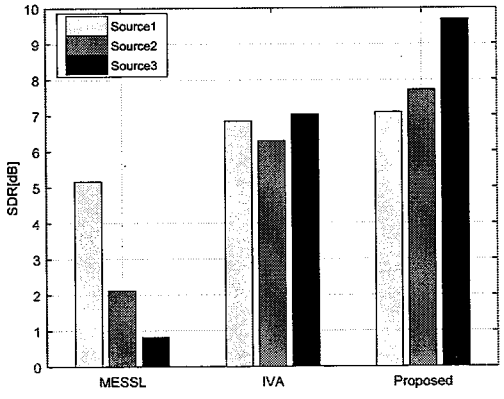


图 4.18　三个声源分离的平均 SDR 值(office Ⅰ)

Figure 4.18 Average SDR values of three source(office Ⅰ)

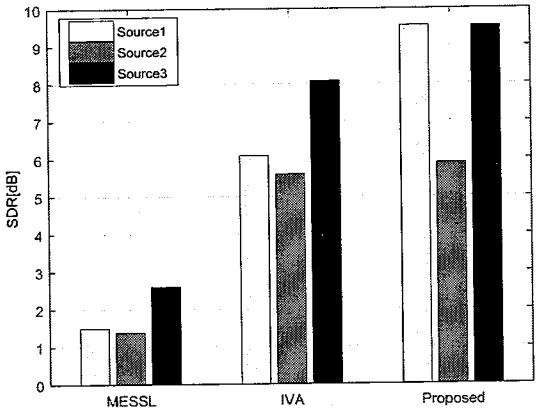


图 4.19 三个声源分离的平均 SDR 值(office Ⅱ)

Figure 4.19 Average SDR values of three source (office Ⅱ)

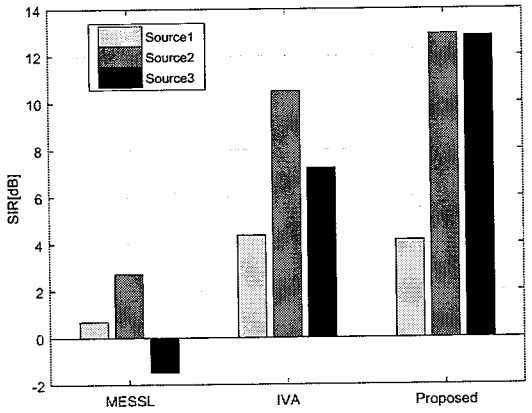


图 4.20 三个声源分离的平均 SIR 值(office Ⅰ)

Figure 4.20 Average SIR values of three source(office Ⅰ)

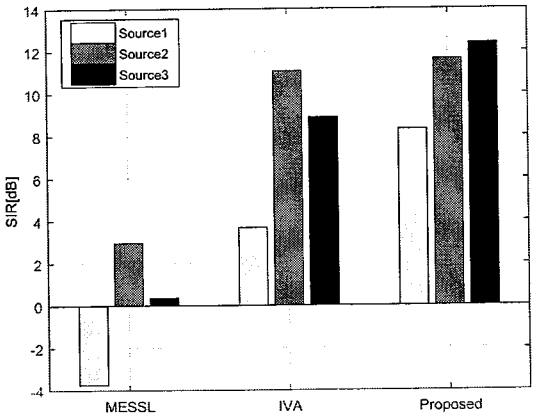


图 4.21 三个声源分离的平均 SIR 值(office Ⅱ)

Figure 4.21 Average SIR values of three source (office Ⅱ)

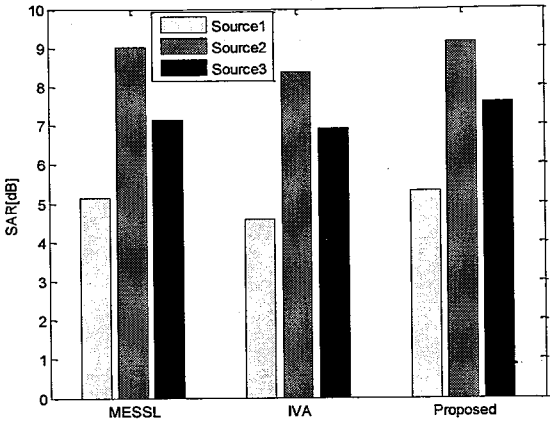


图 4.22 三个声源分离的平均 SAR 值(office Ⅰ)

Figure 4.22 Average SAR values of three source(office Ⅰ)

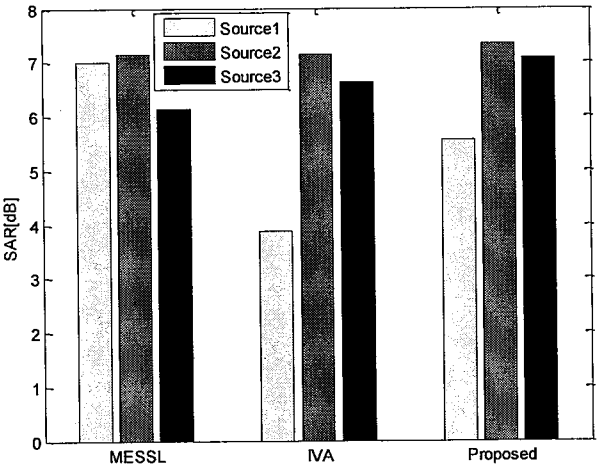


图 4.23 三个声源分离的平均 SAR 值(office Ⅱ)

Figure 4.23 Average SAR values of three source (office Ⅱ)

由以上结果可知，在两种房间环境下，无论是两个声源还是三个声源的情况，本文算法在 SDR 和 SIR 指标上都明显优于对比算法，而在 SAR 指标上三种方法表现无明显差异。这在一定程度上说明本文算法在提高信干扰比方面有明显的优势。另一方面，本文算法同样能够显著提高信噪比，即本文算法能够在噪声环境下完成多语音分离任务。

4.4.3 实录语音实验

为了更一步验证算法在实际环境下的有效性，利用 KEMAR 人工头进行实际环境下的多说话人音频采集，实际环境下与使用 HRTF 库函数的一个最大区别 HRTF 库函数中的传声器经过严格的配对和校准，能够保证传声器之间的一致性，而本

文的采集设备并未对传声器进行校准。实验过程中，先利用扬声器独立播放每个声源信号，保存相应的独立声源，作为参考的纯净声源。随后多个扬声器同时播放声源，获得混合信号。

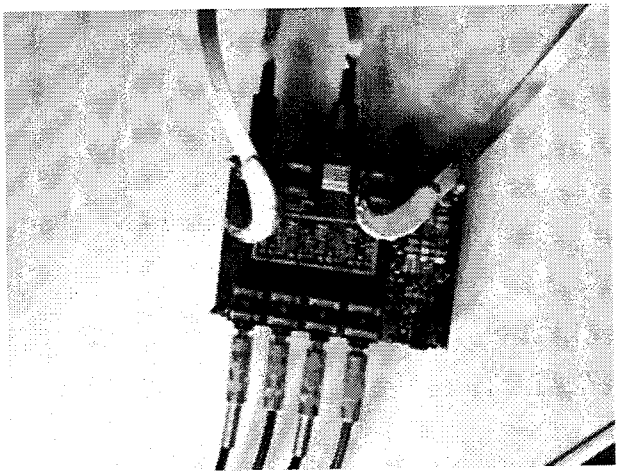


图 4.24　双耳四通道录音板

Figure 4.24 Binaural four-channel recording devices

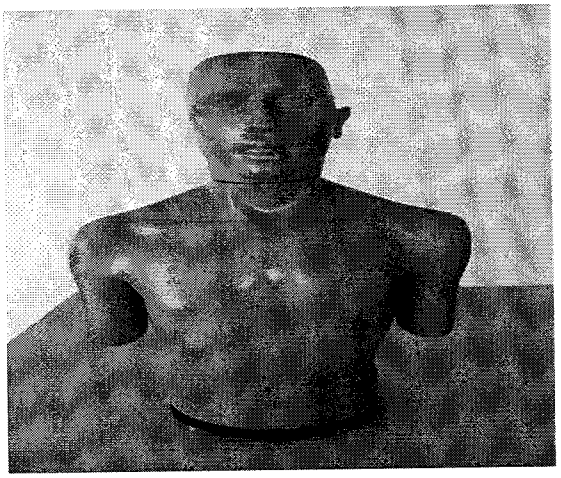


图 4.25　KEMAR 人工头

Figure 4.25 KEMAR artificial head

同样的我们利用 SDR,SIR,SAR 三个指标对实录语音数据进行评价。两个声源和三个声源下的结果分别如图 4.26-4.31 所示，从结果中可以看出，在实际环境下，IVA 算法和本文算法的趋势与仿真结果保持一致，而 MESSL 算法相较于仿真数据有较大下降。原因为 MESSL 是以双耳空间特征 ILD,ITD 进行聚类而进行语音分离的，由于实际录音数据中各个麦克风的一致性没有严格校准而导致性能的下降。为此也可以看出，本文算法对阵列的一致性误差有较好的鲁棒性。

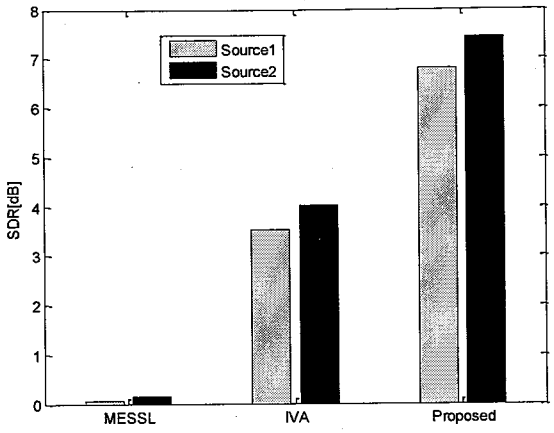


图 4.26　两个声源分离的平均 SDR 值

Figure 4.26 Average SDR values of two source

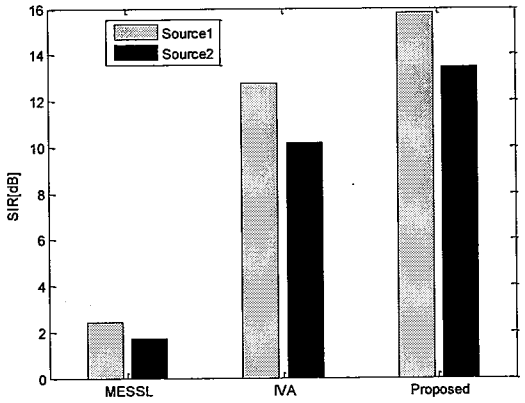


图 4.27　两个声源分离的平均 SIR 值

Figure 4.27 Average SIR values of two source

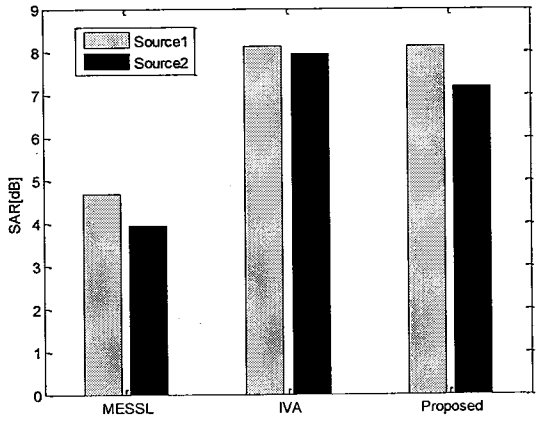


图 4.28　两个声源分离的平均 SAR 值

Figure 4.28 Average SAR values of two source

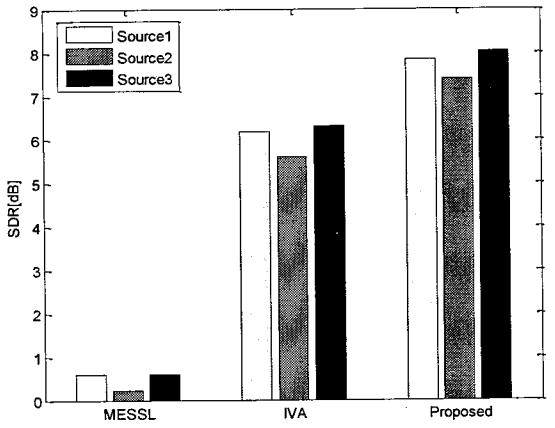


图 4.29 三个声源分离的平均 SDR 值

Figure 4.29 Average SDR values of three source

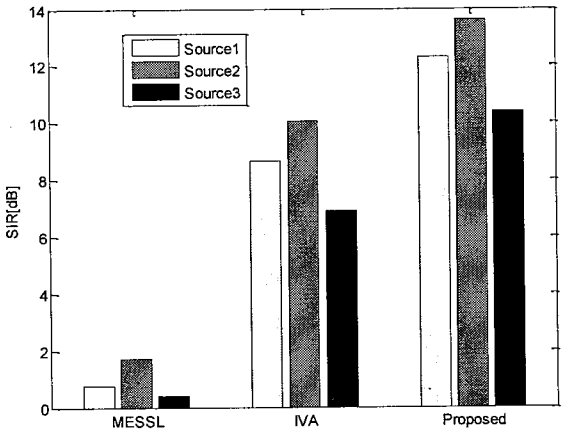


图 4.30 三个声源分离的平均 SIR 值

Figure 4.30 Average SIR values of three source

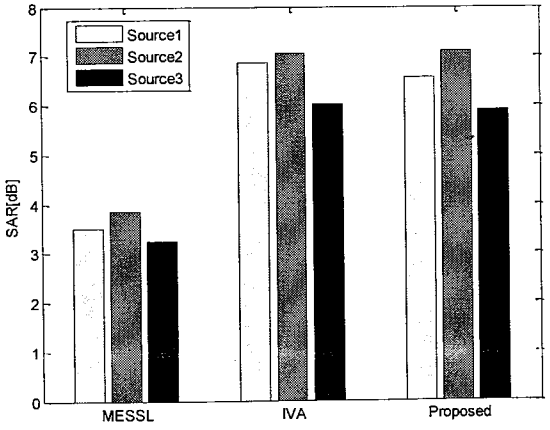


图 4.31 三个声源分离的平均 SAR 值

Figure 4.31 Average SAR values of three source

4.5 总结

针对传统语音分离算法中需要预先知道声源个数信息的限制，本文首先提出一种利用可靠频点选择来构造 DOA 信息直方图的声源数目估计算法，随后利用 DNN 估计噪声协方差矩阵，并结合独立向量分析算法进行目标语音协方差矩阵的获取。最后与多通道维纳滤波器结合，得到最后的分离信号。仿真实验和实录语音实验证明了本文算法能够有效提高信干噪比。

第5章 双耳欠定语音分离算法

5.1 引言

语音分离问题一直成为研究热点，但是对于复杂环境下的分离任务却少有研究。尽管以独立成分分析（ICA）技术为代表的盲源分离技术被广泛的研究，但是在实际情况下，依然存在着一些问题。主要问题有：实际环境下的鲁棒性问题，在实际环境下，各种噪声的存在使得分离效果大大降低；在所有的分离算法中都假设声源数目已知，而实际情况下，如何准则的估计声源数目是需要解决的一个重要问题[147-148]。

近期多声道 NMF 成为一类广泛研究的语音分离算法，文献[149-150]通过引入一个空间相关矩阵(Spatial Correlation Matrix,SCM)，同时迭代公式完成声源的聚类和分离。但是该算法在迭代过程中存在大量的矩阵求逆，计算复杂，且只对谱结构有差别的信号分离效果较好，如音乐和说话人声之间的分离，对于多说话人的分离效果欠佳。文献[151-152]在文献[150]的基本上增加了 DOA 信息，即假设已知各声源的方位，将方位的先验信息加入到 MNMF 中,该方法同样存在着计算量过于复杂且需要各声源的 DOA 信息。文献[153-154]提出一种 Rank-1 spatial model 结合 NMF 的分离算法，该方法假设每个声源的空间矩阵 $H=\alpha*\alpha^{H}$，其中 α 即导向向量。同时结合(independent vector analysis) IVA 技术解决聚类和排序问题。该算法由于将空间分解分解为向量的形式，使得最后的迭代公式中没有矩阵的求逆，计算量大大减少。但是该算法没有考虑噪声干扰的情况，为此文献[155]提出一种基于 NMF 的波束形成算法，该算法首先利用 Rank-1 model 估计导向向量，再用一个稀疏贝叶斯 NMF 估计噪声的 Mask 用于求 Rnn，最后联合导向向量和 Rnn 用于求解最后的波束形成系数。但是这类算法只能解决麦克风数目大于等于声源数目的情况，显然这不适合为双麦克风情况。

DUET(Degenerate unmixing estimation technique)是一种经典的欠定语音分离方法[156]（麦克风数目少于声源数目）。该方法是利用语音信号的稀疏假设进行分离的。通过一个无回声模型，绘制时延差和幅度差信息，来通过直方图构造 Binary mask。但是该方法一方面没有考虑噪声的干扰，另外一方面由于是在一个无回声的假设条件下进行的，而实际环境中存在混响的干扰，这也会降低算法的性能。

随着 2006 年 Hinton 提出深度神经网络的概念后[157]，深度学习在各大领域都获得了很大的成功。同样的在语音增强领域，基于深度学习的语音增强算法也被各大学者所提出来[158]。利用深度学习进行语音降噪，语音混响抑制等算法被相继提出 [159-162]，其中 Y. Zhao 等人认为频谱映射更适合于混响抑制，而 Mask 估计更适合于噪声抑制，为此提出一种两阶 DNN 模型[163]，其中第一阶段利用估计的 Mask 进行噪声的抑制，第二个阶段利用频谱映射完成混响的抑制。另外，结合深度学习与波束形成的语音增强算法也被广泛研究。

对于双耳语音分离系统来说，由于麦克风数目较少(正常情况下只有两个)，为此有必要去研究欠定情况下的语音分离算法。而另一个声源数目估计的问题也是造成实际环境下难以实用的一个关键因素，为此，本文分别提出一种目标声源方位已知的目标语音增强算法和基于深度学习的无任何先验信息的一种降噪与分离结合的多语音分离算法。

5.2 DUET 算法与基于深度学习 Mask 估计算法回顾

5.2.1 DUET 算法

DUET 是在稀疏假设的原理上进行 Mask 估计的方法。假设存在两个声源 s_1 和 s_2，变换到频域后有：

$$S_1(\omega)S_2(\omega) = 0, \forall\omega \qquad \text{...(5.1)}$$

即假设每个频点只有一个声源占主导地位。DUET 算法假设声源位于一个无混响模型中，在该算法中，假设混合信号可以表示为：

$$\begin{bmatrix} X_1(\omega) \\ X_2(\omega) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \alpha_1 e^{-j\omega\delta_1} & \alpha_2 e^{-j\omega\delta_2} \end{bmatrix} \begin{bmatrix} S_1(\omega) \\ S_2(\omega) \end{bmatrix} \qquad \text{...(5.2)}$$

在稀疏假设的前提下，有：

$$\begin{bmatrix} X_1(\omega) \\ X_2(\omega) \end{bmatrix} = \begin{bmatrix} 1 \\ \alpha_1 e^{-j\omega\delta_1} \end{bmatrix} \begin{bmatrix} S_i(\omega) \end{bmatrix} \qquad \text{...(5.3)}$$

为此计算幅度与时延等相关特征：

$$(a_i, \delta_i) = \left(\left\| \frac{X_2(\omega,\tau)}{X_1(\omega,\tau)} \right\|, imag\left(\frac{X_1(\omega,\tau)}{X_2(\omega,\tau)}\right) / \omega \right) \qquad \text{...(5.4)}$$

其中 *imag* 代表取虚部的意思。随后将这个两个特征联合绘制直方图，通过直方图的峰值确定声源的数目。最后通过聚类等算法计算出每个时频点归属于哪一

类，最后估计出每个声源对应的 Mask，完成最后的语音分离过程。该算法简单实用，且适合于欠定情况。

5.2.2 基于深度学习的 Mask 估计算法

在绪论中，我们对传统的语音增强算法进行了介绍，在传统的语音增强算法中，利用信号处理求出每个频点的信噪比，通过信噪比构造增益函数，而基于深度学习的 Mask 估计算法则是通过数据驱动型的增益函数估计算法，深度学习中的 Mask 可以等效认为是传统语音增强算法中的增益函数。这些 Mask 主要包括 Ideal Ratio Mask(IRM) [14]，Spectral Magnitude Mask(SMM)，以及考虑语音相位的 Complex Ideal Ratio Mask (cIRM)[16]。假设带噪语音信号表示为：

$$y(t) = s(t) + n(t) \quad \text{...(5.5)}$$

经过 STFT 变换到频域后为：

$$Y(t,f) = S(t,f) + N(t,f) \quad \text{...(5.6)}$$

则不同 mask 的计算公式为：

$$IBM = \begin{cases} 1, & if\ SNR(t,f) > LC \\ 0, & else \end{cases} \quad \text{...(5.7)}$$

$$IRM = \left(\frac{S(t,f)^2}{S(t,f)^2 + N(t,f)^2}\right)^{\beta} \quad \text{...(5.8)}$$

$$SMM = \frac{|S(t,f)|}{|Y(t,f)|} \quad \text{...(5.4)}$$

$$cIRM = \frac{Y_r S_r + Y_i S_i}{Y_r^2 + Y_i^2} + i\frac{Y_r S_i - Y_i S_r}{Y_r^2 + Y_i^2} \quad \text{...(5.9)}$$

其中 Y_r 和 Y_i 分别表示频域带噪信号的实部和虚部，而 S_r 和 S_i 分别表示频域纯净语音信号的实部和虚部。当然，除了 Mask 估计，另外一种更直接的方式是频谱的映射，即输入时带噪语音的幅度谱，输出为纯净语音的幅度谱。这种形式被证明在混响抑制中表现更好。为此，Y. Zhao 等人认为频谱映射更适合于混响抑制[163]，而 Mask 估计更适合于噪声抑制，为此提出一种两阶 DNN 模型，其中第一阶段利用估计的 Mask 进行噪声的抑制，第二个阶段利用频谱映射完成混响的抑制。

5.2.3 基于置换不变性的语音分离算法

随着深度学习的快速发展，基于深度学习的语音分离算法也被很多学者提出。目前来说，主要的研究方向还是基于单声道的语音分离算法。一个经典的基于深度神经网络的双说话人语音分离结构如图 5.1 所示。其中两个 output 的目标输出为：

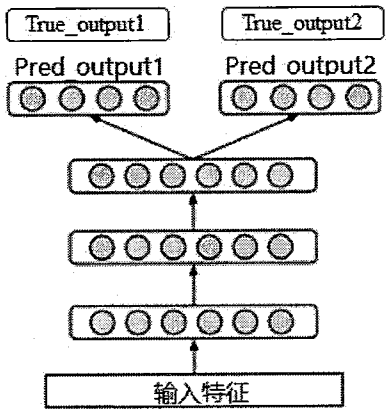


图 5.1 语音分离结构框图

Figure 5.1 Block diagram of speech separation

$$True\_output_1(t,f) = \left| \frac{|s_1(t,f)|}{|s_1(t,f)| + |s_2(t,f)|} \right| Y(t,f) \quad \text{...（5.10）}$$

$$True\_output_2(t,f) = \left| \frac{|s_2(t,f)|}{|s_1(t,f)| + |s_2(t,f)|} \right| Y(t,f) \quad \text{...（5.11）}$$

其中 $Y(t,f)$代表混合信号在 t 帧和 f 频点的幅度。S1，S2 分别代表两个纯净语音的幅度谱。我们定义：

$$\cos t_1 = \sum_t ((pred\_output_1 - true\_output_1)^2 + (pred\_output_2 - true\_output_2)^2) \quad \text{...（5.12）}$$

$$\cos t_2 = \sum_t ((pred\_output_1 - true\_output_2)^2 + (pred\_output_2 - true\_output_1)^2) \quad \text{...（5.13）}$$

在第四章介绍的语音分离任务中，我们就已经提到，排序问题是语音分离系统中必须考虑的一个重要问题。为了解决这个问题，俞栋等人提出了基于置换不变性 PIT（Permutation invariant training）的训练方法[28]，该方法的结构框图如下：

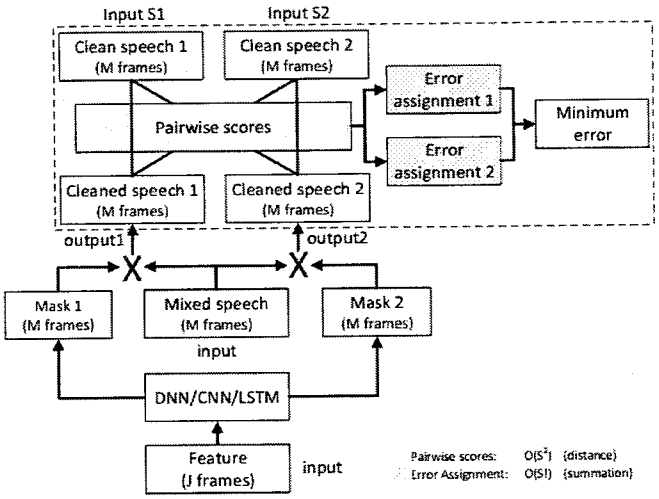


图 5.2 PIT 语音分离结构框图

Figure 5.2 Block diagram of PIT speech separation

该训练方法主要是为了解决语音分离任务中的排序问题，核心思想是计算真实标签和估计的标签之间的不同组合，并找到最小的组合，作为最后的误差值。以一个两个声源分离系统为例，即此时的代价函数为：

$$\cos t = \min(\cos t_1, \cos t_2) \qquad \text{...(5.14)}$$

5.3 目标方位已知的语音分离算法[164]

在助听器，人工耳蜗等听力设备中，往往针对面对面交流场景设计语音增强算法，也就是说在这类设备中假设目标声源位于正前方。前文所介绍的 DUET 算法中，基于信号的稀疏假设，可以利用空间信息进行声源的聚类，而在本节中，我们同样基于稀疏假设，提出一种目标声源已知情况下的双耳语音增强算法。

如图 5.3 所示。假设目标声源位于正前方，干扰噪声源与正前方夹角为 θ。

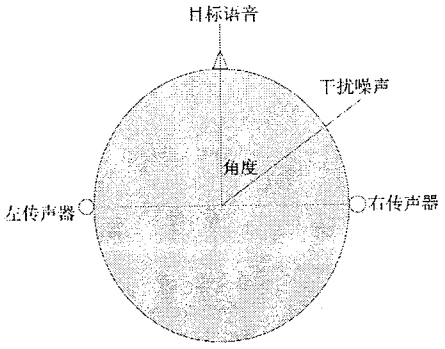


图 5.3 噪声源与双传声器布置图

Figure 5.3 Placement of noise source and dual-microphone

由于人头的影响，侧方的声源在左右耳之间有明显的能量差（ILD）与（ITD）。且在高频区域 ILD 占主导，低频区域 ITD 占主导[163][165]。定义左右耳接收到的时域信号为：

$$x_1(\mathrm{m}) = h_l(\mathrm{m}) * s(\mathrm{m}) + h_{nl}(\mathrm{m}) * n(\mathrm{m}) \quad \text{...（5.15）}$$

$$x_2(\mathrm{m}) = h_r(\mathrm{m}) * s(\mathrm{m}) + h_{nr}(\mathrm{m}) * n(\mathrm{m}) \quad \text{...（5.16）}$$

其中 $s(m)$代表目标语音，$n(m)$代表干扰噪声，m代表时域的时间点。$h_l(\mathrm{m})$，$h_r(\mathrm{m})$为目标语音分别到左右耳间的头相关脉冲响应函数 $h_{nl}(\mathrm{m})$，$h_{nr}(\mathrm{m})$为干扰噪声到左右耳间的脉冲响应函数。将时域信号变换到频域后得到：

$$X_1(n,k) = H_l(n,k) \cdot S(n,k) + H_{nl}(n,k) \cdot N(n,k) \quad \text{...（5.17）}$$

$$X_2(n,k) = H_r(n,k) \cdot S(n,k) + H_{nr}(n,k) \cdot N(n,k) \quad \text{...（5.18）}$$

将公式（5.17）（5.18）改写为：

$$X_1(n,k) = S_1(n,k) + N(n,k) \quad \text{...（5.19）}$$

$$X_2(n,k) = S_2(n,k) + H_{12}(n,k) \cdot N(n,k) \quad \text{...（5.20）}$$

其中 $H_{12} = \dfrac{H_{nl}}{H_{nr}}$ 为干扰噪声分量在左右耳间的头相关传输函数（HRTF）的比值。

求 X_1，X_2的自功率谱，有：

$$PX_1X_1(n,k) = PS_1S_1(n,k) + PNN(n,k) \quad \text{...（5.21）}$$

$$PX_2X_2(n,k) = PS_2S_2(n,k) + \left|H_{12}(n,k)\right|^2 \cdot PNN(n,k) \quad \text{...（5.22）}$$

计算两个通道的能量差有：

$$\Delta PX(n,k) = PX_1X_1(n,k) - PX_2X_2(n,k) = \Delta PS(n,k) + PNN(n,k) \cdot (1 - \left|H_{12}(n,k)\right|^2) \quad \text{...（5.23）}$$

其中 PSS，PNN 分别代表目标语音信号自功率谱，干扰噪声自功率谱。由于假设目标语音来自正前方，$\Delta PS(n,k) \approx 0$，则：

$$PN(n,k) = \frac{\Delta PX(n,k)}{1 - \left|H_{12}(n,k)\right|^2} \quad \text{...（5.24）}$$

根据维纳滤波思想，滤波器增益 $G = \dfrac{PSS}{PSS + PNN}$，并代入（5.24）中，得到：

$$G(n,k) = \frac{PSS(n,k) \cdot (1 - \left|H_{12}(n,k)\right|^2)}{PSS(n,k) \cdot (1 - \left|H_{12}(n,k)\right|^2) + \Delta PX(n,k)} \quad \text{...（5.25）}$$

其中 $PSS(n,k)$和 $H_{12}(n,k)$为未知数。假设目标语音与干扰语音不相关，又有：

$$PX_1X_2(n,k) = PS_1S_2(n,k) + H_{12}(n,k) \cdot PNN(n,k) \quad \text{...（5.26）}$$

$$\text{则 } H_{12}(n,k) = \frac{PX_1X_2(n,k) - PS_1S_2(n,k)}{PX_1X_1(n,k) - PSS(n,k)} \quad \text{...（5.27）}$$

$$\text{又有，} PS_1S_2(n,k) = \Gamma_S(n,k)\cdot PSS(n,k) \quad \text{...（5.28）}$$

其中 $\Gamma_S(n,k) = \dfrac{PS_1S_2(n,k)}{\sqrt{PS_1S_1(n,k)\cdot PS_2S_2(n,k)}}$ 为两个通道目标语音的相干函数。现在将未知量转换为 $\Gamma_S(n,k)$ 和 $PSS(n,k)$（$PSS \approx PS_1S_1 \approx PS_2S_2$）。为了估计这两个未知量，现定义一个相对能量差：

$$\Delta_{PLD}(n,k) = \frac{PX_1X_1(n,k) - PX_2X_2(n,k)}{PX_1X_1(n,k) + PX_2X_2(n,k)} \quad \text{...（5.29）}$$

1. 如果 $\Delta_{PLD}(n,k) < \Phi_{\min}$，则表明能量差很小，表明此频域点位于语音段。由此更新语音功率谱，以及语音的相干函数。

$$\begin{aligned} PSS(n,k) &= \alpha_1 \cdot PSS(n-1,k) + (1-\alpha_1)\cdot |X_1(n,k)|^2 \\ \Gamma_S(n,k) &= \alpha_\Gamma \cdot \Gamma_S(n-1,k) + (1-\alpha_\Gamma)\cdot \Gamma x(n,k) \end{aligned} \quad \text{...（5.30）}$$

2. 如果 $\Delta_{PLD}(n,k) > \Phi_{\max}$，此时表明能量差很大，位于干扰噪声段，此时不更新语音功率谱与语音的相干函数。

$$\begin{aligned} PSS(n,k) &= PSS(n-1,k) \\ \Gamma_S(n,k) &= \Gamma_S(n-1,k) \end{aligned} \quad \text{...（5.31）}$$

3. 如果 $\Phi_{\min} < \Delta_{PLD}(n,k) < \Phi_{\max}$，如果此时能量差介于不大不小之间，则此时的语音谱用能量低的那一侧的信号来估计。

$$PSS(n,k) = \alpha_2 \cdot PSS(n-1,k) + (1-\alpha_2)\cdot |X_2(n,k)|^2 \quad \text{...（5.32）}$$

其中公式（5.29）（5.30）（5.31）中的 α_1、α_2、α_Γ 为平滑因子，n 与 k 分别代表信号所在的帧与该帧上的频率点。通过对 $PSS(n,k)$ 和 $\Gamma_S(n,k)$ 的更新，将最后的 $PSS(n,k)$ 和 $\Gamma_S(n,k)$ 带入到（5.25）（5.27）（5.28）中，得出最后的维纳滤波增益：

$$G_{X1}(n,k) = \frac{PSS(n,k)\cdot(1-|H_{12}(n,k)|^2)}{PSS(n,k)\cdot(1-|H_{12}(n,k)|^2) + \Delta PX(n,k)} \quad \text{...（5.33）}$$

$$G_{X2}(n,k) = \frac{PSS(n,k)\cdot(1-|H_{12}(n,k)|^2)}{PSS(n,k)\cdot(1-|H_{12}(n,k)|^2) + H_{12}(n,k)\cdot\Delta PX(n,k)} \quad \text{...（5.34）}$$

5.3.1 客观评价

在计算机上使用 MATLAB 仿真软件来测试算法的效果。测试用纯净语音为 IEEE 语料库中的标准语音文件，采样率为 16 kHz；干扰噪声分别为 NoiseX-92 噪声库中的 BAB(Babble) 噪声、汽车 (Car) 噪声和 IEEE 语料库中的 CT(Competitive Talker)干扰语音。本文使用的头相关传递函数来自于 MIT 实验室。参考文献[166]中使用相位差来判决目标语音与干扰噪声的二值掩码(Binary mask）方法，和文献[167]中的基于相干函数（Coherence）法作为对照。使用 PESQ（Perceptual Evaluation of Speech Quality，语音质量感知评价）和 SNR

（Signal to Noise Ratio，信噪比）对本文算法（Proposed）进行客观评价，最后统计出 20 条句子的平均分数。

表 5.3 所示为单个噪声源干扰下的客观评价分数，三种单个噪声源分别为 45 度的 CT 噪声、60 度的 Car 噪声、90 度的 Bab 噪声。由于算法是在每个频率点进行操作，当多个噪声源干扰时只要噪声源的频率分布不完全一致的时候算法也应该有一定效果，为此这里验证了算法在三个噪声源干扰下的去噪效果。表 5.4 所示为三个噪声源同时干扰下的分数，其中 45 度的 CT 噪声，60 度的 Bab 噪声，90 度的 Car 噪声这三个噪声源位于双耳的同一侧；另外 45 度的 CT 噪声，60 度的 Bab 噪声，180 度的 Car 噪声这三个噪声源位于双耳的不同侧。从表中结果可以看出无论是单个噪声源或是三个噪声源情况下，本文算法的 SNR 与 PESQ 的分数都优于两种对比算法，一定程度上说明了本文提出的双耳语音增强算法去噪能力优于基于相干函数的双耳语音增强算法和二值掩码法。

表 5.3 单个噪声源干扰下的客观评价分数

Table 1.1 Objective scores of single noise interference

		SNR(dB)				PESQ			
噪声类型	角度	Noisy	Coherence	Binary Mask	Proposed	Noisy	Coherence	Binary Mask	Proposed
CT	45°	-5	0.70	4.13	4.66	1.29	1.82	1.55	1.90
		0	4.17	6.81	6.89	1.53	2.05	1.84	2.12
		5	7.91	9.20	9.74	1.75	2.29	2.12	2.48
Car	60°	-5	1.77	6.09	9.09	1.60	1.71	1.82	2.09
		0	4.09	8.13	11.06	1.65	1.81	1.98	2.28
		5	8.64	10.38	13.78	1.88	2.01	2.25	2.56
Bab	90°	-5	0.24	2.19	5.30	1.01	1.41	1.55	1.81
		0	3.72	5.66	8.01	1.10	1.78	1.75	2.04
		5	8.51	9.42	11.45	1.72	2.01	2.02	2.31

表 5.4 三个噪声源干扰下的客观评价分数

Table 1.1 Objective scores of three noise interference

		SNR(dB)				PESQ			
噪声类型	角度	Noisy	Coherence	Binary Mask	Proposed	Noisy	Coherence	Binary Mask	Proposed
CT	45°	-5	0.81	5.85	6.74	1.08	1.70	1.79	1.98
Bab	60°	0	4.94	8.04	9.09	1.61	1.82	2.14	2.21
Car	90°	5	8.57	11.43	11.90	1.77	2.01	2.44	2.60
CT	45°	-5	-0.10	-1.07	1.18	1.04	1.59	1.36	1.48
Bab	60°	0	5.42	4.32	6.73	1.66	1.81	1.77	1.84
Car	180°	5	8.31	7.62	9.26	1.80	2.00	1.99	2.13

5.3.2 主观评价

主观评价采用言语接受阈（speech reception thresholds, SRT）的测试，在噪声环境测试时，SRT 定义为听者对单词具有 50%的识别正确率时的信噪比。实际测试在隔音室中进行，测试材料使用[168]中的 400 句标准普通话句子，并且将全部句子分为 25 个表，每个表中 16 个句子。其中 Car 噪声选自 NoiseX-92 噪声库,CT 噪声的生产方法是从标准普通话库中随机选取一个句子作为干扰噪声（此句子后续不再使用）。Bab 噪声的产生方法为从标准普通库中随机选取 5 句话（后续不再使用），然后直接相加。

对于每个测试者来说要进行 5 种情况的测试（3 种单个噪声源干扰的情况和 2 种三个噪声源干扰的情况），每种情况下测试 4 组 SRT，这四组分别为未经算法处理的带噪语音（Noisy）、相干函数（Coherence）算法处理后的语音、二值掩码（Binary mask）算法处理后的语音，以及本文算法（Proposed）处理后的语音。将这四组标记为 A、B、C、D，打乱排列顺序进行测试。测试目的是比较未处理前以及不同算法处理后的 SRT 的差别。选定一种噪声干扰的情况后，随机选择一组（A、B、C、D）算法，然后随机选取一个表中的句子，用于测试该组情况下的 SRT。

在每个算法的每组测试中采用一上一下自适应调整信噪比(SNR)的方法[169]。初始 SNR 为 10dB，在第二个反转点（reversal point）前，调整步长为 8dB，在第四个反转点前步长为 4dB，之后反转点为 2dB，直至当前组全部句子测试完毕。其中，在每个试次中（即播放每个句子时），听者可以要求至多再重听两次，当听者复述出多于半数的字时，判定为答题正确，否则为答题错误。每组中最后 8 句话的 SNR 的算数平均值记为当前条件下的 SRT。被试者通过佩戴包耳式耳机来听取双通道语音信号，正式实验开始前会对被试者进行训练和指导。

本次试验共 8 位受试者，均为听力正常人士。性别分别为 4 男 4 女。图 5.4 中为测试结果，图中的 5 种情况的设置与客观评价中的 5 种情况保持一致。其中柱状图代表采用不同降噪算法后 8 位受试者平均提高的 SRT（dB）分数。相应算法提高的 SRT（dB）分数的计算方法为未经算法处理的 SRT（dB）减去相应算法处理后的 SRT （dB）。柱状图上面的误差线为 8 位受试者提高的 SRT（dB）的标准差。从图中可以看出，两个对比算法在不同的噪声干扰情况下各有优势，但是本文的算法在所有情况下相对参考算法都有一定的优势，这证明了本文算法能更好

的提高语音可懂度。

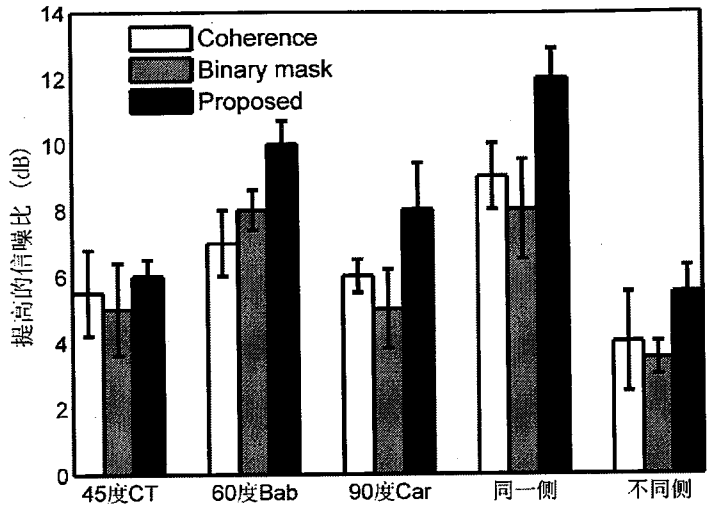


图 5.4 不同噪声干扰情况下提高的 SRT

Figure 5.4 Improved SRT under different conditions

5.4 目标声源方位未知的复杂场景下的多声源分离

无论是传统的语音分离算法和基于深度学习的语音分离算法，都假设处于安静环境下。而在传统的波束形成语音分离算法或利用声源空间信息进行聚类或Mask估计的算法中，假设各个声源处于不同的位置，而现实环境下这些假设显然不合理，为了能够在噪声环境下完成多语音分离的工作。本节提出一种将语音降噪与语音分离联合处理，利用深度学习的方式完成欠定情况下的双耳语音分离。该方法无需假设声源位于不同的方位，且能够在噪声干扰下完成多条语音的分离。

下图为本文的双耳语音降噪与分离算法的结构框图。我们首先利用双耳特征和单声道的特征训练一个语音降噪的神经网络，训练目标为 IRM。训练结束后，保存神经网络的结构和权重。之后，再利用不包含噪声的纯净语音训练一个双耳语音分离系统，其中输入特征为左右通道联合的谱特征和双耳空间特征(包括 ITD,ILD 和 Coherence)。输出为每个声源的左右声道纯净语音特征。训练过程的误差回传中采用 5.2.1 节中所介绍的置换不变性（PIT）方法。

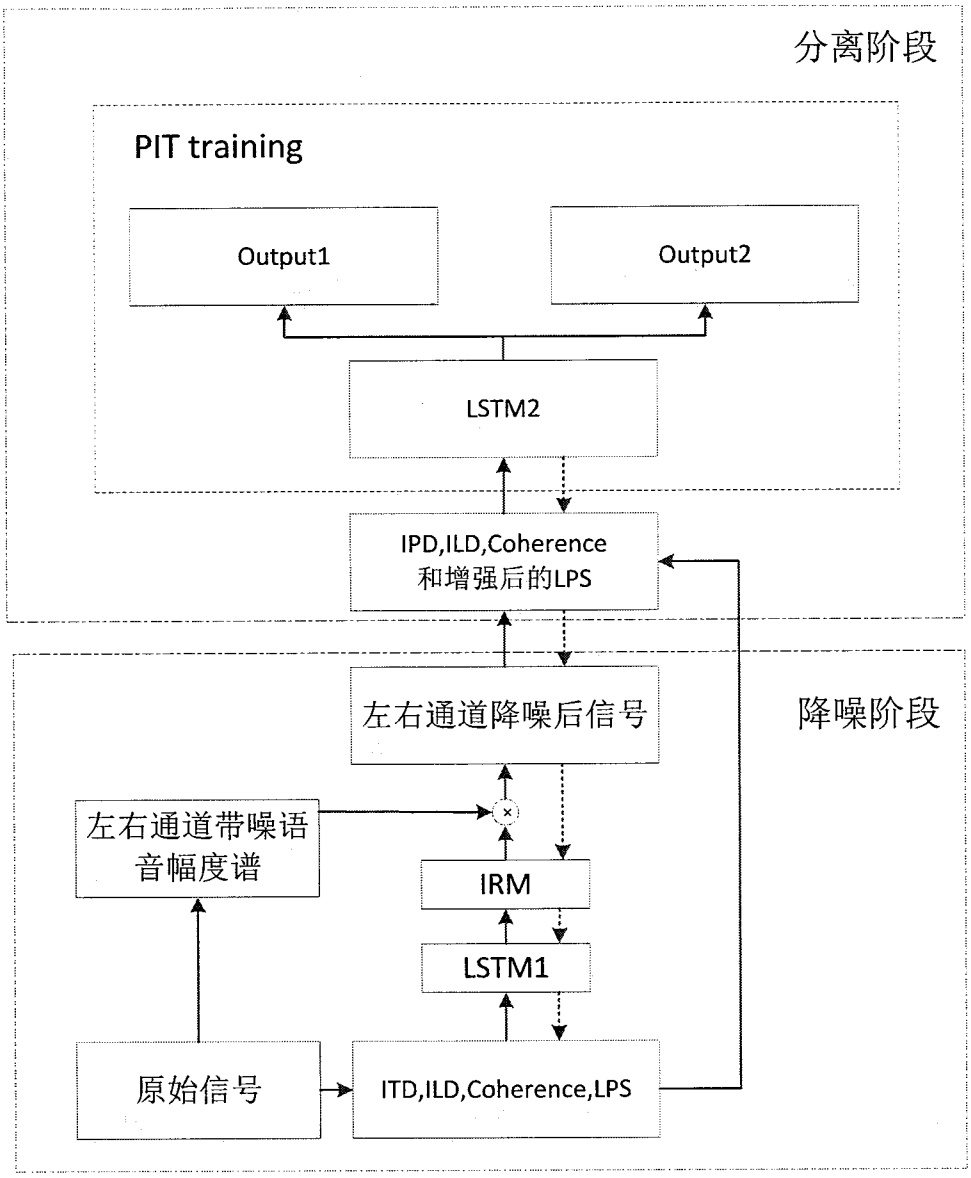


图 5.5 本文的神经网络结构框图

Figure 5.5 Structure diagram of our neural network architecture

数据预处理：我们首先对信号进行分帧加窗处理，帧长为 32 ms，overlap 为 75%。信号采样率为 16 kHz。将每一帧的时域信号通过短时傅里叶变换变换到频域。由于声源可能位于不同方位，也有可能位于相同方位，所以输入特征包括单耳的谱特征和双耳的空间特征。分别计算相应的特征，包括 log 谱(LPS)，相位差(IPD)，能量差(ILD)和相干函数(Coherence)。

纯净的语音数据选自 TIMIT 语料库。双耳背景噪声选自于 ETSI 提供的实录双耳噪声[170]。HRTF 选自 CATT 声学模型库[171]。

两声源和三声源联合训练:传统的语音分离系统中，需要知道声源的数目，为了提高本章所提出的算法的鲁棒性，采用两声源和三声源联合处理，即当只有两声源时，其中某一个目标通道为白噪声。这使得三个输出的神经网络同样能够处理两声源的分离任务。

语音降噪神经网络结构：首先我们通过 HRTF 库获取多说话人语音信号，以三声源为例，这里介绍具体获取方式为：在 TIMIT 语料库[145]中随机选取两条或者三条不重复的纯净语音信号，对应的在 HRTF 库[171]中随机选择 2 个角度或三个角度的可以重复(这就意味着声源可能来自于同一方位)的 HRTF(一共 37 个角度，从 0-360 度)，每个角度的 HRTF 分别卷积一条纯净语音，从而获取双耳多语音数据。当只有两个声源时，则其中某个通道只包含白噪声。

然后将双耳语音数据与双耳噪声相加，信噪比分别为 0,5,10dB,从而获得带噪的多语音数据。我们选取 LSTM 用于语音降噪，并利用 IRM 作为目标输出，cost function 为 MSE。隐含层的激活函数采用 PReLU[172]，输出层的激活函数采用 Sigmoid。训练方法采用 Adam。

输入维数为 LPS (257)+IPD (48)+ILD(209)+Coherence(257)=771 维。输入层与特征维数相同，两个隐含层包括 1024 个节点，输出层为 257 维。该网络训练好之后，保存网络的权重。

语音分离网络结构：与语音降噪的网络类似，此时的双耳多声源数据在 TIMIT 库中随机选取两条或者三条不重复的纯净语音信号，对应的在 HRTF 库中随机选择 2 个角度或三个角度（可重复）的 HRTF(一共 37 个角度，从 0-360 度)，每个角度的 HRTF 分别卷积一条纯净语音，从而获取双耳多语音数据。

输入特征同样为 LPS 和 IPD,ILD,Coherence 等特征，此时目标有三个，分别为三个纯净语音的幅度谱。我们利用两层的 LSTM 作为语音分离结构，激活函数采用 PReLU，训练方法同样采用 Adam。Cost function 采用 PIT 的训练方式，即从所有的配对中找出最小值作为最后的 loss。对于三个声源来说，一共有 6 种情况，分别是：

配对 1：(t1-p1),(t2,p2),(t3,p3)

配对 2：(t1-p1),(t2,p3),(t3,p2)

配对 3：(t1-p2),(t2,p1),(t3,p3)

配对 4：(t1-p2),(t2,p3),(t3,p1)

配对 5：(t1-p3)，(t2,p1)，(t3,p2)

配对 6：(t1-p3)，(t2,p2)，(t3,p1)

该网络训练好之后，保存网络的权重。

联合处理结构：将语音降噪的神经网络和语音分离的神经网络联合起来。网络的初始权重分别使用前面已经独立训练好的网络结构，之后再次对该网络进行训练。此时的数据获取方式与降噪的网络结构一样。

5.4.1 实验结果

将 TIMIT 语音和双耳噪声库合成 5000 条训练数据，进行训练。测试语音选自中文数据库，测试噪声同样选自 ETSI 双耳噪声库，但是与训练数据中的噪声不重合。以下本文算法处理的结果，三个声源下原始信号的语谱图和利用本章节所提出的算法估计的语谱图对比如下：

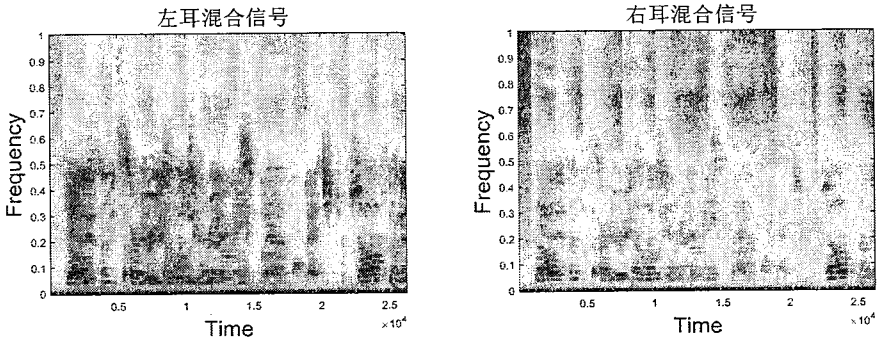


图 5.6 左右耳原始混合信号

Figure 5.6 Original binaural mixed signal

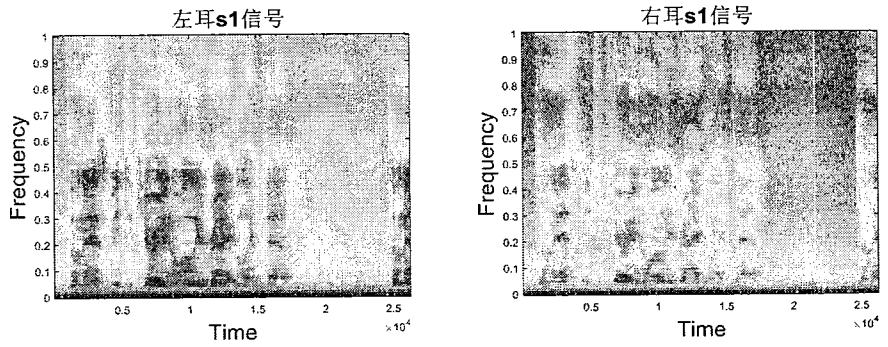


图 5.7 真实的左右耳 s1 信号

Figure 5.7 The real binaural signal s1

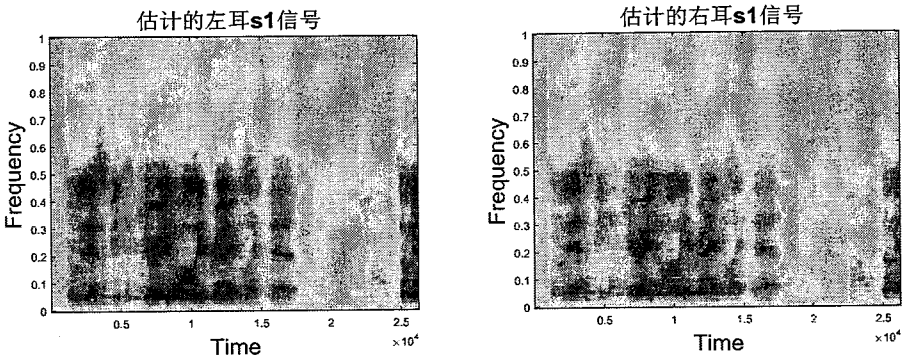


图 5.8 估计的左右耳 s1 信号

Figure 5.8 The estimated binaural signal s1

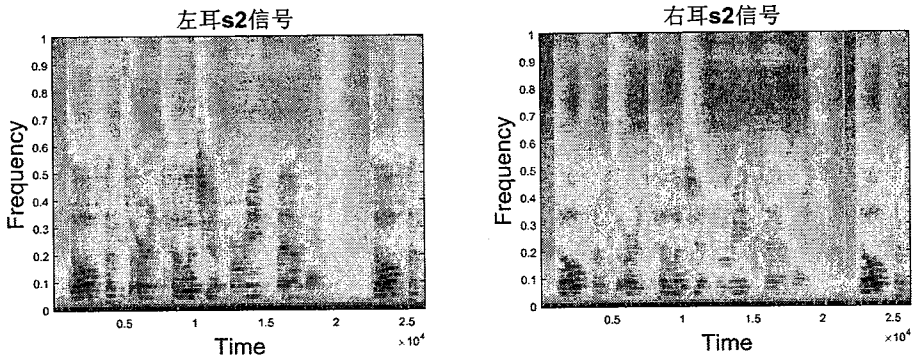


图 5.9 真实的左右耳 s2 信号

Figure 5.9 The real binaural signal s2

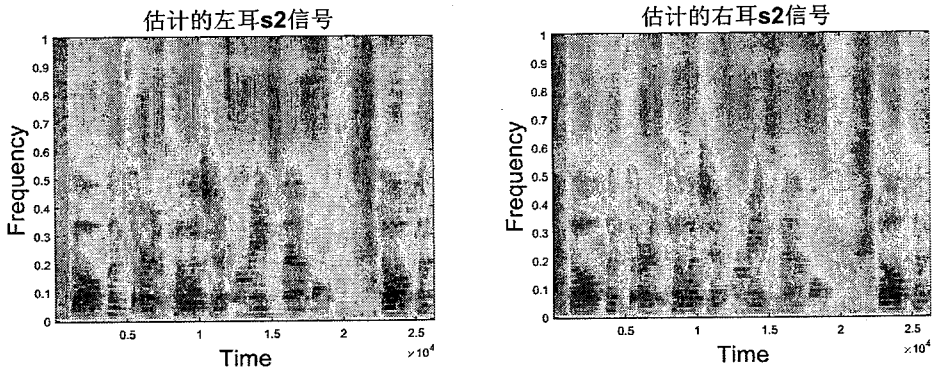


图 5.10 估计的左右耳 s2 信号

Figure 5.10 The estimated binaural signal s2

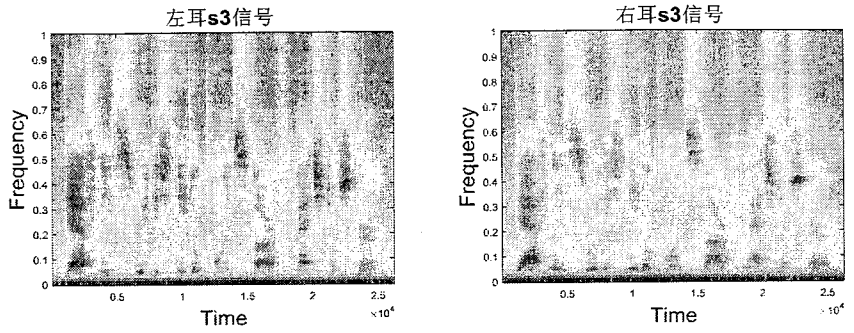


图 5.11 真实的左右耳 s3 信号

Figure 5.11 The real binaural signal s3

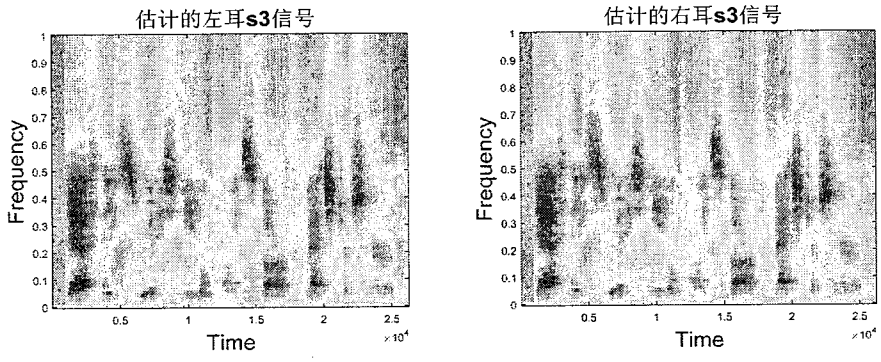


图 5.12 估计的左右耳 s3 信号

Figure 5.12 The estimated binaural signal s3

为了验证两声源时该网络的分离效果，我们观察混合信号和估计的各声源的语谱图，很明显的发现，在估计的第三个声源中，能量非常小，近似为 0，而其他两个声源能够准确的估计出来。

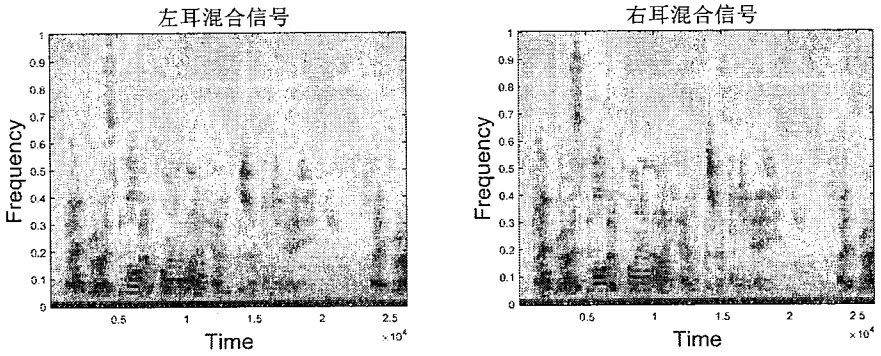


图 5.13 左右耳原始混合信号

Figure 5.13 Original binaural mixed signal

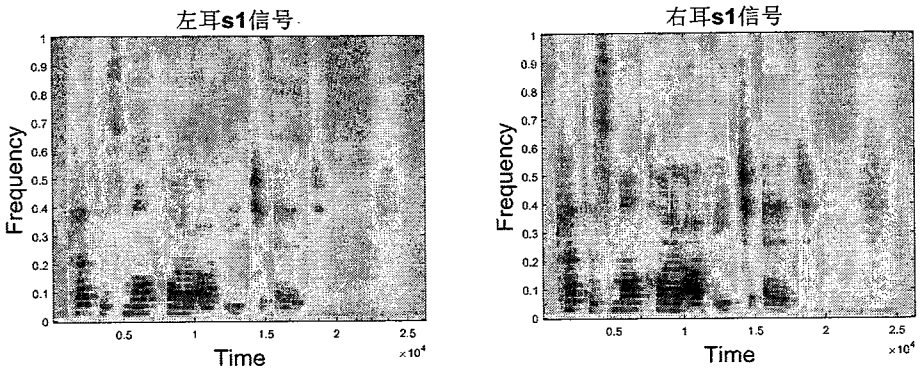


图 5.14 真实的左右耳 s1 信号

Figure 5.14 The real binaural signal s1

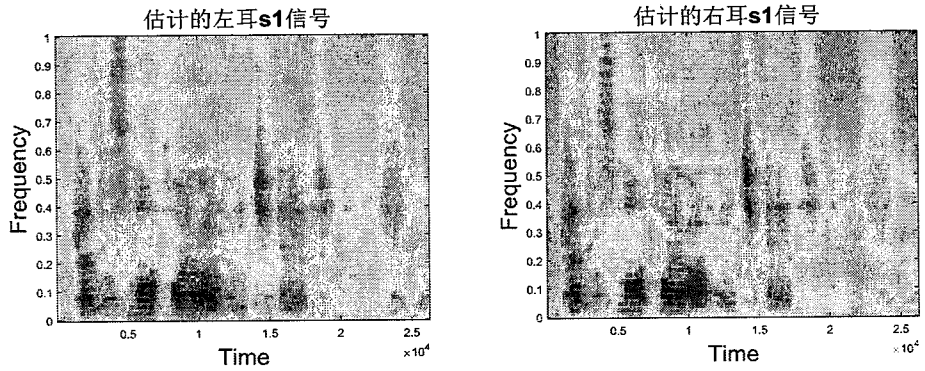


图 5.15 估计的左右耳 s1 信号

Figure 5.15 The estimated binaural signal s1

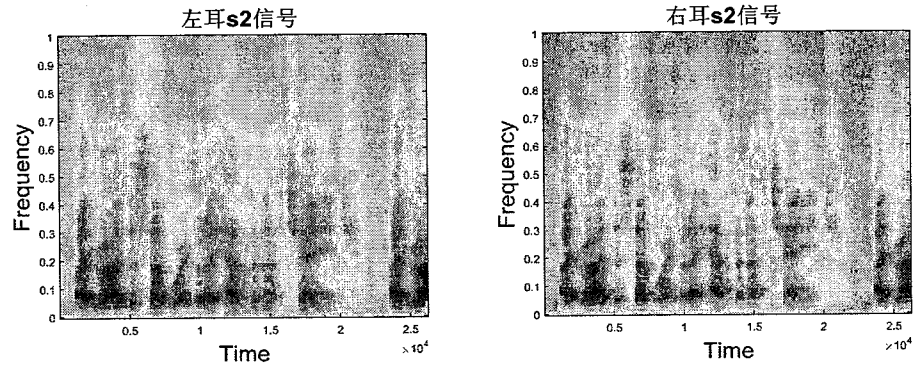


图 5.16 真实的左右耳 s2 信号

Figure 5.16 The real binaural signal s2

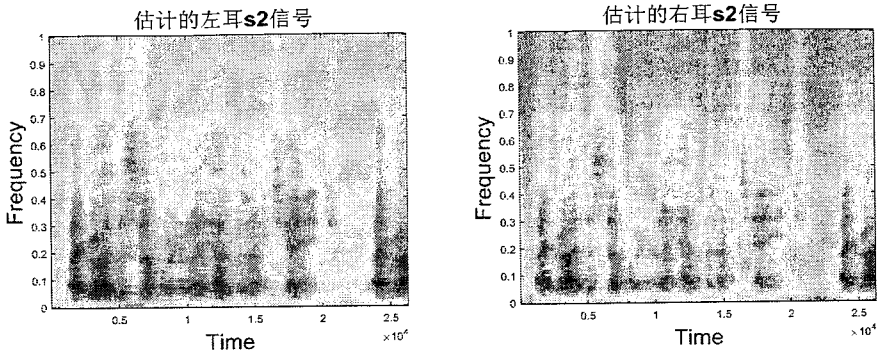


图 5.17 估计的左右耳 s2 信号

Figure 5.17 The estimated binaural signal s2

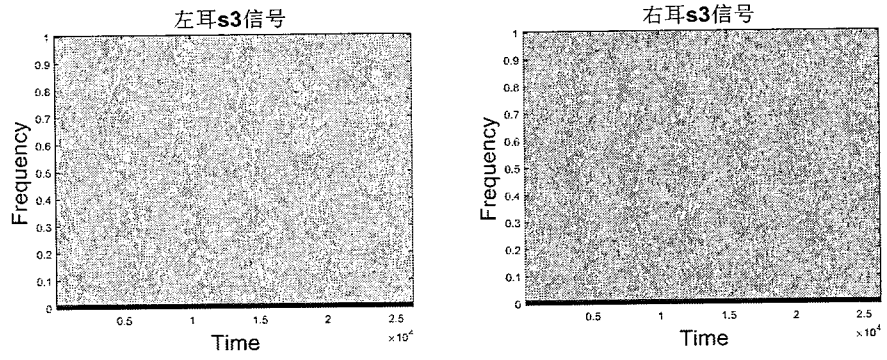


图 5.18 真实的左右耳 s3 信号（无信号）

Figure 5.18 The real binaural signal s2（no signal）

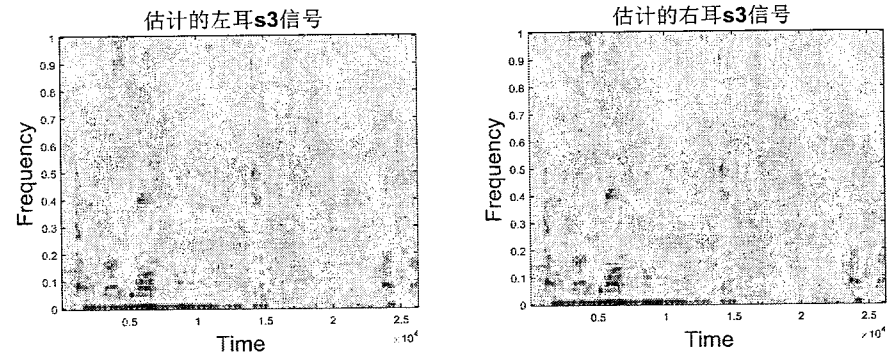


图 5.19 估计的左右耳 s3 信号

Figure 5.19 The estimated binaural signal s3

为了更一步的严重本文算法（模型 2）的优势，我们使用基于双耳模型进行聚类的 MESSL 语音分离算法和直接使用神经网络进行 PIT 训练的模型(模型 1)，即不采用两个阶段的预训练进行比较，结构框图见下图。本文算法相较于模型 1，有着以下优势，优势 1 为 Y. Zhao 等人的两阶系统已经证明了对于降噪系统来说，更加适合于目标为 Mask，而分离和混响抑制更适合直接用纯净信号的频谱作为

目标。为此提出本文的两阶降噪与分离联合系统，第一阶利用 Mask 去除噪声干扰，第二阶段用频谱映射来完成语音分分离。另外一个优势在于，本文预先训练好两个网络，相当于权重初始化的过程，随后的联合过程可以理解为精细调优（fine tune）的过程。相较于直接训练的方式，通过初始化神经网络的权值可以让神经网络更快速的收敛，也能一定程度上防止过拟合等现象的发生。

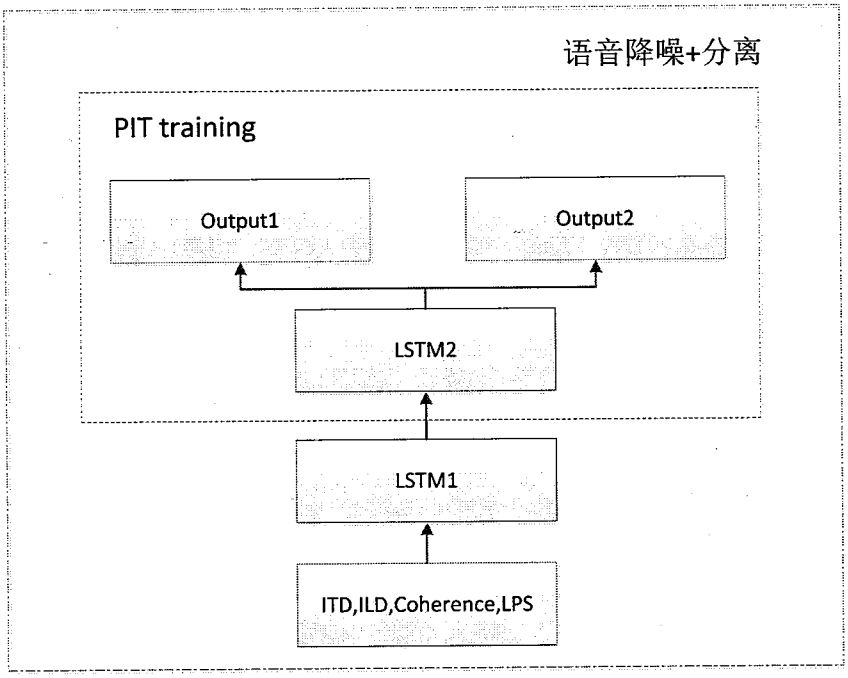


图 5.20 对比模型(模型 1)

Figure 5.20 The compared model (Model 1)

表 5.5 不同算法的 SDR 值

Table 5.5 The SDR values of different algorithms

SNR\ (dB)	2-speaker				3-speaker			
SNR (dB)	未处理	MESSL	模型 1	模型 2	未处理	MESSL	模型 1	模型 2
0	-6.29	-1.05	9.10	9.05	-7.33	-2.11	7.54	7.51
5	-3.16	1.03	8.20	8.21	-4.12	-0.94	6.41	6.40
10	-1.10	2.49	7.84	7.88	-2.08	1.58	5.26	5.27
Avg	-3.51	0.82	8.38	8.38	-4.51	-0.49	6.40	6.39

表 5.6 不同算法的 PESQ 值

Table 5.6 The PESQ values of different algorithms

SNR (dB)	2-speaker				3-speaker			
SNR (dB)	未处理	MESSL	模型 1	模型 2	未处理	MESSL	模型 1	模型 2
0	0.85	0.98	2.34	2.58	0.81	0.94	1.98	2.40
5	0.90	1.31	2.33	2.65	0.89	1.23	2.12	2.54
10	0.97	1.47	2.80	2.82	0.93	1.41	2.28	2.67
Avg	0.91	1.25	2.49	2.68	0.88	1.19	2.13	2.53

从实验结果来看，模型 1 与模型 2 提高的 SDR 值基本一致，模型 2 略有优势。在 PESQ 分数方面，模型 2 有比较明显的优势，各种情况下均为本文算法的分数最高，以上的客观评价分数证明了本文的分析。而 MESSL 方面由于加入了噪声的干扰，在两项评价指标上效果都有限。且 MESSL 方法与其他的传统盲分离或聚类方法类似，需要预先知道声源个数。而本文算法仅需要设置一个多个的 output，通过不同声源数的联合训练，能够同时处理不同声源数目的情况，仿真实验结果表明了本文联合降噪与分离系统的有效性。

5.5 总结

在本章中，我们分别针对目标方位已知，且干扰与目标方位不相同情况下的目标语音分离算法和目标与干扰方位均未知条件下的语音分离算法进行研究。第一类算法利用空间信息估计目标声源的传递比值来构造滤波器，达到目标语音增强的目的。在第二类算法中为了解决传统的语音分离中的一系列问题，如噪声环境下算法效果的下降，欠定情况下算法效果的下降，需要预先知道声源数目等问题。提出一种联合降噪与分离的双耳多声源分离网络，并以三个目标输出为例，分别对比了两个声源和三个声源分离的效果。仿真实验表明：该网络结构在噪声环境下的语音分离性能优于传统的基于模型的聚类算法。且该方法由于不需要准确知道声源的数目，使得在实际环境中能够更加鲁棒的工作。后续的工作中将考虑进一步加大数据量，进行更系统更详细的评价，并对实际环境下的语音分离性能进

行实验验证。

第 6 章 总结和展望

语音增强算法对于提高听力设备佩戴者的语音可懂度或提升智能语音交互设备的后端识别结果都有着重要的研究意义。随着人工智能技术的快速发展，单声道语音增强技术的研究方向也从传统的噪声功率谱估计转向基于深度学习的增益函数预测。多通道语音增强技术从传统的波束形成技术转向基于无监督聚类与有监督学习结合等新方法。本文结合传统的信号处理知识与深度学习理论重点研究基于双耳的语音增强算法，包括双耳的混响抑制，噪声抑制以及语音分离等问题。

6.1 本论文主要研究内容总结

本文围绕双耳的语音增强算法展开细致研究。针对现实环境中出现的混响，噪声干扰，人声干扰等问题，提出一系列的解决方法：

第一章主要阐述了论文的研究内容和研究意义，分析了目前语音增强算法中面临的难点。随后对传统的单声道语音增强算法，基于深度学习的单声道语音增强算法，传统的多声道语音增强算法，基于机器学习/深度学习的多通道语音增强算法进行了回顾并分析各类算法的优势和问题。并对双耳的语音增强算法现状单独进行了综述。最后阐述了本文的研究内容和各章节的结构安排。

第二章主要对人耳的听觉系统和数学模型等基本知识进行了介绍，重点分析了双耳的“优先效应”模型。随后对头相关传递函数(HRTF)和混响理论进行了介绍。算法方法，对固定波束形成的，自适应波束形成的基本理论知识进行了介绍。最后，阐述了盲源分离理论和神经网络基础。这些理论的介绍都是为后续章节的内容做铺垫。

第三章首先介绍了混响环境下的时延迟估计与 CDR 估计问题。传统的时延差估计算法的准确率会受到混响的影响从而急剧下降。而基于优先效应模型的方法是一类模仿人耳的听觉机理去解决混响下时延差估计问题的方法，在该章节中，我们首先回顾了经典的优先效应模型，并分析各种模型的优劣势，随后提出一种基于相干函数的峰值平滑策略，并在此基础上利用复数相干函数进行时延差的估计。利用仿真实验将本文的时延差估计算法和经典的优先效应模型的时延差估计算法进行了比较。随后，本文将时延差估计算法与 CDR 估计算法有效的结合，利用 CDR 估计值，构造一个维纳滤波器，进行晚期混响的抑制，实验结果表明

了算法的有效性。

第四章，我们对语音分离常用算法进行了详细分析，并提出目前的主流方向是基于 Mask 估计与波束形成结合的方法，该方法不依赖阵列流形，对阵列的一致性误差不敏感，针对目前主流的基于 CGMM 聚类和深度学习的 Mask 估计中对人声干扰抑制不了的缺点。本文提出一种基于独立向量分析(IVA)的 Mask 估计算法，独立向量分析(IVA)是在独立成分分析(ICA)基础上解决声源分离中的排序问题的改进算法。本文利用 IVA 估计出的解混矩阵，结合 DNN 的 Mask 估计算法，实验结果表明该方法比直接利用解混矩阵来进行分离能够获得更高的信干噪比提升。

第五章中，我们利用 IVA 进行解混矩阵的估计，但是该方法需要预先知道声源的个数，且声源个数需要小于麦克风数目。对于双耳语音设备来说，有可能仅有两个麦克风，为此有必要对欠定情况下的语音分离算法进行研究。在该章中，我们首先对声源方位已知条件下的语音分离算法进行研究。现有的算法中往往只考虑了安静环境下的语音分离，而实际环境下有可能存在强背景噪声的干扰，为此本章提出一种基于深度学习的两阶语音分离系统。该系统分为两个部分，一部分为语音降噪系统，一部分为语音分离系统，在语音分离系统中我们利用置换不变性训练方法进行训练，并同时利用空间信息和谱特征。该系统在背景噪声和混响等干扰下依然能够有效分离出多个纯净语音。

第六章，将对本文进行总结并对目前算法中可能存在的不足哦，提出下一步的研究工作。

6.2 下一步研究工作

下一步的工作主要包括以下几个方面：

1. 在第三章中，我们主要对基于 CDR 的混响抑制算法进行了研究，但是并未与其它的主流方法进行比较，例如基于加权预测误差(WPE)和基于深度学习的混响抑制算法的对比。另外，对于多个说话人时的混响抑制效果也并未进行实验评价，后续工作中将继续对其它的主流混响抑制算法进行仿真和比较，并评价不同算法在多个说话人声的混响抑制性能。

2. 在第四章中，我们提出基于独立向量分析的语音分离方法，但是在该算法中假设声源数目已知，对于声源数目错误情况下的分离效果并未进行实验，且该算

法计算量偏大。为此在后续的工作中将继续对该方法进行研究，提供声源数目估计算法，并评价在声源数目估计出错时，算法的性能。进一步的，减小计算量，并希望能够用到实时嵌入式系统中去。

3. 在第五章中，我们提出一种两阶语音降噪与分离联合系统，在小数据上进行了训练和测试，而基于深度学习的方法存在着测试数据与训练数据不匹配时，性能下降的问题。为此在后续工作将将进一步考虑提高数据量，并改进网络结构，提升泛华能力。另外，也将尝试一些其它的神经网络形式或结构，例如：卷积神经网络（CNN），循环神经网络（RNN），生成对抗网络（GAN）等。另外一个是本文的语音分离系统完全是数据驱动型的有监督学习，而为了更好的自适应的在各种环境下都取得最好的效果，将在后续考虑有监督学习与无监督学习算法的结合，例如：利用有监督学习对某一个通道信号进行一次 Mask 估计，随后在该 Mask 基础上进行空间特征的聚类，将单声道的 Mask 与空间聚类的 Mask 结合得到最后的 Mask，最后与波束形成算法结合。

4.尽管本文对混响抑制，语音降噪和语音分离问题都进行了研究，但是评价指标都是客观的参数评价和主观评价，并未与语音识别后端结合，或者说这些算法对于语音识别正确率的提升并未进行评价，为此在后续的工作中会继续熟悉语音识别的后端，将语音前端降噪和语音识别后端联合起来。