大连理工大学

硕士学位论文

学 科、专 业：信号与信息处理

硕　士　生：周浩洋

指 导 教 师：殷福亮　教授

论文答辩日期：2002 年 3 月

大连理工大学硕士学位论文题目名称

基于麦克风阵列的声源定位方法研究

计：学位论文 51 页
表　　格 3 个
插　　图 33 幅

学位论文完成日期 2002 年 3 月 13 日

评阅人 王承训 教授

马晓红 副教授

指导教师 殷福亮 教授

教研室主任 郭成安 教授

系主任 刘军民 教授

摘 要

基于麦克风阵列的声源定位技术在视频会议、声音检测及语音增强等领域有重要的应用价值。但由于噪声和混响的存在，使得现有语音定位算法的定位精度较低。此外，已有的声源定位方法的运算量较大，难以实时处理。为解决这些问题，本文系统地研究了基于麦克风阵列的声源定位方法，主要做了以下几方面工作：

1) 详细描述了基于可控波束形成和时延估计-几何定位的声源定位方法，并给出了性能分析。
2) 对延迟-相加波束形成的定位方法作了详细推导，并针对该方法的缺点加以改进，使其在混响环境下也有较高的精度。
3) 讨论了广义互相关时延估计法，自适应时延估计法以及基于人耳特点的时延估计法。在理论上对它们进行性能分析，给出各种算法的适用场合和优缺点，并进一步改进了 GCC-PHAT 时延估计方法。
4) 归纳比较角度距离定位法，球形插值法及线性插值法的优缺点，并对主要定位方法给出了计算机仿真结果。
5) 提出一种可实时实现的定位系统。该系统不仅能够较好的抑制混响和噪声的影响，而且运算量比较低。文中给出了该系统的实现框图及计算机仿真结果。

关键词：麦克风阵列；声源定位；时延估计；波束形成；广义互相关函数；自适应滤波

Abstract

Microphone array can be employed for speaker localization in videoconference. However, background noise and room reverberations in enclosure greatly degrade the effectiveness of acoustic source localization. In addition to high accuracy, the location estimator must be computationally no demanding in order to be useful for real-time tracking and beamforming. Aiming at these difficulties, two methods for speaker localization using microphone array are discussed in this paper: delay and sum beamforming, time delay estimation and geometry localization. Then a new method by combining beamforming with GCC is presented to efficiently suppress the noise and reverberation.

What is more, the improvement to GCC-PHAT algorithm is also developed, specifically usefulness to background noise and room reverberation elimination.

In the end, a new system for microphone array speaker localization, which can suppress noise and reverberation, are presented in the paper. The result of computer simulation shows that the new algorithm outperforms other algorithm such as beamforming or LMS adaptive filter.

Key words:

Microphone array; Acoustic source localization; Time delay estimation; Beamforming; Generalized cross correlation; Adaptive filter

致 谢

经过一年的努力，终于完成了硕士学位论文。回顾自己经历过的风风雨雨的日子，有许多值得回忆的事和感谢的人。

首先要感谢的是导师殷福亮教授。殷老师严谨的治学态度、渊博的专业知识、求实的科研作风给我留下了深刻的印象，并将使我终生受益。

读研期间，我的师兄陈喆、刘兴立、高可攀、房德新给了我很多帮助，特向他们表示感谢。还要感谢顾巨峰、朱健华、魏建强等同学以及教研室的师弟师妹们。

最后，谨以此文献给我的父母及女友，感谢他们对我人生追求的支持。希望他们一生幸福。

目 次

第一章 概 述

1.1 麦克风阵列简述

在无噪声、无混响的情况下，距离声源很近的高性能、高方向性的单麦克风可以获得高质量的声源信号。但是，这要求声源和麦克风之间的位置相对固定，如果声源位置改变，就必须人为地移动麦克风。若声源在麦克风的选择方向之外，则会引入大量的噪声，导致拾取信号的质量下降。而且，当麦克风距离声源很远，或者存在一定程度的混响及干扰的情况下，也会使拾取信号的质量严重下降。为了解决单麦克风系统的这些局限性，人们提出了用麦克风阵列进行语音处理的方法。麦克风阵列系统就是由一组按一定几何结构摆放的单向麦克组成的系统。麦克风阵列系统较之单麦克风系统具有许多优点，其优越性表现在：

1) 高方向性的单麦克风通常只能拾取一路信号，而麦克风阵列系统可以采集多路信号。虽然麦克风阵列是对单个目标的数据采集，但由于各麦克位置的不同，它采集的数据在时间或者空间上必然存再某些差异。从而通过多路信号的数据融合技术，就可以提取出所需要的信息。
2) 麦克风阵列系统具有空间选择特性。它以“电子瞄准”的方式使所形成的波束对准声源，这抑制了其他说话人的声音和环境噪声，从而获得高品质的声源信号。

20 世纪 80 年代以来，传感器阵列信号处理技术得到迅猛的发展，并在雷达、声纳及通信中得到广泛的应用。这种阵列信号处理的思想后来应用到语音信号处理中。在 1985 年 Flanagan 将麦克风阵列引入到大型会议的语音增强应用中，并开发出多种实际产品。之后，Silverman 和 Brandstein 将其应用于语音识别和声源定位中。进入九十年代以来，基于麦克风阵列的语音处理算法正逐渐成为一个新的研究热点。现有的麦克阵列系统已经有了许多的应用，这些应用包括语音识别、强噪声环境下的语音获取、大型场所的会议记录、声音检测和助听装置等。特别是将麦克风阵列应用在视频会议系统中，用于确定和实时跟踪说话人的位置。

我国在这方面的研究工作起步较晚，目前尚未见到有相应的论文发表。

1.2 基于麦克风阵列的定位方法

基于麦克风阵列的定位问题简而言之就是利用一组按一定几何位置摆放的麦克定出声源的空间位置。

基于麦克风阵列的声源定位方法大体上可分为三类：(*a*) 基于最大输出功率的可控波束形成技术。该方法对麦克风阵列接收到的语音信号进行滤波、加权求和，然后直接控制麦克风指向使波束有最大输出功率的方向；(*b*) 基于高分辨率谱估计的定向技术。该方法利用求解麦克信号间的相关矩阵来定出方向角，从而进一步定出声源位置；(*c*) 基于到达时间差(TDOA)技术。该方法首先求出声音到达不同位置麦克的时间差，再利用该时间差求得声音到达不同位置麦克的距离差，最后用搜索或几何知识确定声源位置。

(1) 基于最大输出功率的可控波束形成定向方法

该方法对麦克风所接受到的声源信号滤波并求加权和来形成波束，进而通过搜索声源可能的位置来引导该波束，最终使波束输出功率最大的点就是声源的位置。在文献[1,2]中最早提出该方法的理论基础，在文献[3]中进一步得出可控定位的理论和实际上的方差。并在文献[4]中将该方法应用于多声源的定位。

可控波束形成技术本质上是一种最大似然估计，它需要声源和环境噪声的先验知识。而在实际使用中，这种先验知识往往很难获得。此外，最大似然估计是一个非线性优化问题，这类目标函数往往有多个极点，且该方法对初始点的选取也很敏感，因此使用传统的梯度下降算法往往容易陷于局部极小点，从而不能找到全局最优点。如果采用别的搜索方法，若要力求找到全局最优点，就会极大地增加计算复杂度，从而不可能被用于实时处理系统。

(2) 基于高分辨率谱估计技术的定向方法

该方法来源于一些现代高分辨率谱估计技术(如 AR 模型，MV 谱估计，MUSIC 算法，特征值分解等)。虽然该方法成功地应用于一些阵列信号处理的应用，但在说话人定位中的效果不佳。原因有以下四点：

1) 该方法需要通过时间平均来估计各麦克信号之间的相关矩阵，这就需要信号是平稳的，且估计的参数是固定不变的。而语音信号是一个短时平稳过程，它往往不能满足这个条件，因此该方法效果和稳定性不如可控波束形成法。此外，该方法往往假设理想的信号源、相同特性的麦克等这些在实际中不可行的条件。虽然可以通过某些方法减弱这些因素的影

响，但这往往需要成倍的增加运算量[5]。

2) 由于房间的混响作用，使信号和噪声有一定的相关性，这也会降低该方法的有效性。
3) 该方法还需假定声源离麦克的距离比较远，且麦克是一个线性阵列，这样声波可以近似看成平面波。而这对需近距离定位的系统是不可行的[6]。
4) 高精度谱估计技术往往针对窄带信号，而语音信号是宽带信号，这也需要以增加运算量为代价来提高定位精度[7,8]。

(3) 基于时延的定位方法

基于时延的定位方法在导航系统、声纳系统等领域都有广泛的应用。该方法主要是估计各麦克间的相对时延，适合于单个声源的定位。由于每对麦克时延唯一对应一个双曲面，因此多个麦克对就可以确定多个双曲面。双曲面之间的交集从某种意义上就是声源的次最优估计。基于此原理产生出许多定位方法，各有自己的优缺点。

就已经获得的一组麦克时延，大体上可以有两种方法定位出声源。其一是以获得的时延求得一个目标函数，通过搜索的方法来确定声源的位置；其二用次最优的方法通过几何插值的方法估计出声源的位置。

基于时延的定位方法在运算量上远远小于可控波束形成和谱估计法，可以在实际中实时实现。但是该方法也有不足之处，其一是估计时延和定位分成两阶段来完成，因此在定位阶段用的参数已经是对过去时间的估计，这在某种意义上只是对声源位置的次最优估计；其二时延定位的方法比较适合于单声源的定位，而对多声源的定位效果就不好；其三在房间有较强混响和噪声的情况下，往往很难获得精确的时延，从而导致第二步的定位产生很大的误差。虽然如此，由于时延估计定位方法的运算量比较低，而且在适当改进后，在一定的噪声和混响下有比较好的定位精度，因此适合于在实际中实时应用。本文将主要研究该方法。

1.3 麦克风阵列定位系统简介

传统的视频会议系统通过人来控制摄像机使其对准说话人的位置。该方法不仅精度低，而且带来极大的不便。本文通过麦克风阵列确定房间内说话人位置，从而控制摄像机自动地对准说话人。因此该麦克风阵列定位系统主要是集中在室内。图 1-1 详细描绘了一个麦克风阵列定位系统的实际情况。

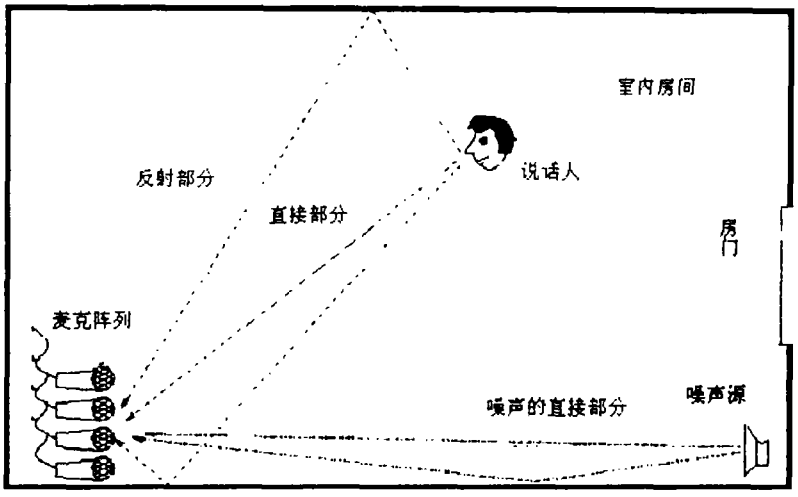


图 1-1 麦克风阵列定位系统描述

从图 1-1 可知，一个麦克风阵列系统定位声源的精度受多方面因素的影响。第一是噪声和反射的噪声；第二是声源的多重反射；第三是说话者与麦克的相对位置。

假定声音传播满足线性波动方程，且房间内的环境在一段时间内不变，则从声源到麦克之间可看成线性时不变系统。设声源信号为 $s(n)$，到达麦克的信号为 $x_i(n)$，则

$$x_i(n) = h_i(n) * s(n) + w_i(n) \tag{1-1}$$

其中 $h_i(n)$是房间的单位冲激响应，$w_i(n)$是高斯白噪声。图 1-2 是一个典型房间的单位冲激响应

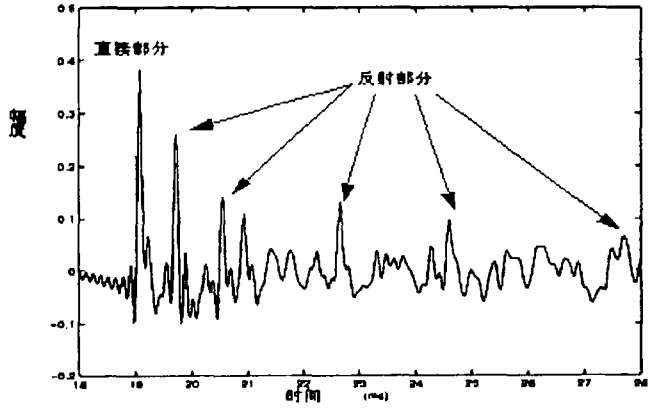


图 1-2 房间的单位冲激响应

由图 1-1 可知，一个高精度声源定位系统所面临的难点主要有：

1) 定位系统不仅受到噪声的干扰，而且由于墙的反射作用，会产生相关噪声。这样各麦克间噪声的互相关函数就不等于零，从而增大了定位的难度。
2) 由于墙的反射，麦克不仅收到声源的直接部分还收到反射部分。而声音的反射会导致互相关函数或者波束的尖峰扩展，难以确定最大值，加大了定位的误差。
3) 对于单面墙上摆放的线性麦克，当声源在方向角比较小的位置时，会影响几何近似的精度，从而给麦克与声源间距离的确定引入了比较大的误差。因此声源与麦克的相对位置也会极大的影响定位的精度。
4) 麦克的摆放。对于一个定位系统而言，麦克的数量越多，麦克的相对位置越多样化，提供的空间信息量越大，从而具有较高的定位精度。而在实际系统中，麦克的摆放位置比较固定，数量也比较少。因此难点就在尽量少的麦克和固定摆法条件下，提供高的定位精度。

1.4 模拟仿真数据的产生和参数设置

为了检验方法的性能，本文设计了许多仿真实验，这些实验测试数据的产生和参数设置在此一并说明。

(1) 房间混响的仿真

由上一节的讨论可知，房间混响的仿真主要是确定 $h_i(n)$。本文用 IMAGE 模型[9]产生该单位冲激响应。$h_i(n)$的计算公式如下

$$h(n,x,x')=\sum_{p=0}^{1}\sum_{r=-\infty}^{\infty}\beta_{x1}^{|n-q|}\beta_{x2}^{|n|}\beta_{y1}^{|l-j|}\beta_{y2}^{|l|}\beta_{z1}^{|m-k|}\beta_{z2}^{|m|}\times\frac{\delta(n-|R_p+R_r|/c)}{4\pi|R_p+R_r|} \quad (1\text{-}2)$$

其中 $\beta_{x1},\beta_{x2},\beta_{y1},\beta_{y2},\beta_{z1},\beta_{z2}$ 是各面墙的反射系数，x，x' 分别指声源坐标和麦克坐标，c 指声速，L_x,L_y,L_z 为房间的长宽高，$R_r=[2nL_x \quad 2lL_y \quad 2mL_z]$ 和 $R_p=[(x-x'+2qx') \quad (y-y'+2yj') \quad (z-z'+2qk')]$ 表示声源的镜像向量。

房间的混响时间 T 与反射系数 β 的关系为

$$\beta = \exp(-13.82 / [c(L_x^{-1} + L_y^{-1} + L_z^{-1})T]) \quad (1\text{-}3)$$

这样由要求的混响时间就可以求出房间的反射系数，然后由麦克和声源的坐标就可以求出从声源到麦克的单位冲激响应 $h_i(n)$。假定采样率为 8KHZ，并以采样间隔为基本距离单位。图 1-3 是房间大小为 100*120*80，声源坐标为(60，45，60)，麦克坐标为(30，0，60)，反射系数为 0.927 的房间单位冲激响应图

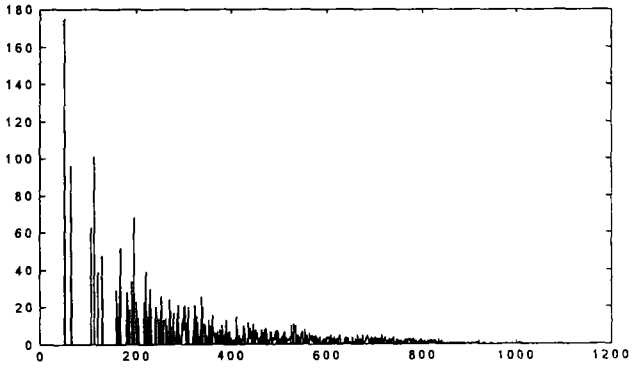


图 1-3 IMAGE 模型产生的单位冲激响应

(2) 文中的噪声信号 $w_i(n)$是高斯白噪声，由 MATLAB 中的 Randn()函数直接产生

(3) 语音信号处理前先要进行模数转换，设定采样率为 16KHZ，所加窗为汉宁窗，窗长为 512 个采样点，其中前后语音帧重叠 50%。

1.5 本论文的主要工作

1) 对各种麦克风阵列语音定位方法进行总结分类，详细描述一些有代表性的算法。本论文着重介绍两类麦克风阵列定位方法，它们分别是：基于可控波束形成的定位方法，时延估计-几何定位方法。

2) 对延迟-相加波束形成的定位方法作了详细推导。并针对该方法的缺点加以改进，使其在混响环境下也有较高的精度。

3) 讨论了各种时延估计方法的优缺点，本论文着重介绍了三种时延估计的方法，它们分别是：广义互相关时延估计法，自适应时延估计法，基于人耳特点的时延估计法。在理论上对各类方法进行性能分析，给出各种算法的适用场合和优缺点。并进一步改进了 GCC-PHAT 时延估计方法。

4) 归纳比较了各种定位方法，它们分别是：角度距离定位法，球形插值法，线性插值法。并对主要定位方法给出了计算机仿真结果。
5) 提出一种可实时实现的定位系统，该系统不仅能够较好的抑止混响和噪声的影响，而且具有比较低的运算量。在文中给出了该系统的实现框图，及计算机仿真结果。

第二章 基于可控波束形成的定位方法

基于可控波束形成的定位方法[1-4]是出现比较早、并在实际中得到应用的一种定位方法。基于可控波束形成的定位主要分为延迟累加波束法和自适应波束法。前者运算量小，信号失真小，但抗噪性低，需要较多的麦克才有比较好的效果。后者因为加了自适应滤波，所以运算量比较大，而且输出信号有一定程度的失真，但需要的麦克数目相对较少，在没有混响时有比较好的效果。本章详细讨论了延迟累加波束定位法的原理，并提出了一种改进方法。

2.1 延迟累加可控波束形成定位法

可控波束形成定位方法在满足最大似然准则的前提下，以搜索的方式使麦克风阵列所形成的波束对准信号源，从而获得最大输出功率。

假设 M 个麦克成线性排列，声源搜索点的位置用 θ 表示。用 $s_i(t)$表示声源信号，$w_i(t)$表示零均值的高斯白噪声，其中 $s_i(t)$和 $w_i(t)$是不相关的宽平稳高斯随机过程。 τ_i 表示声源搜索点到麦克的延迟时间(τ_i 可由声源搜索点的位置 θ 得到)。在不考虑混响和衰减的情况下，第 i 个麦克所接收到的信号为

$$x_i(t) = s(t-\tau_i(\theta)) + w_i(t) \qquad (i = 1,2,\cdots,M) \tag{2-1}$$

对(2-1)式两边求傅立叶变换得

$$X_i(\omega) = \mathrm{e}^{-j\omega\tau_i} S(\omega) + W_i(\omega) \qquad (i = 1,2,\cdots,M) \tag{2-2}$$

因为 $s_i(t)$和 $w_i(t)$是不相关的宽平稳高斯随机过程，所以由高斯随机过程的性质可知，$x_i(t)$、$X_i(\omega)$也是宽平稳的高斯随机过程。

对于具有 M 个麦克的阵列系统，在频率ω_l处，式(2-2)的向量形式可以表示为

$$X(\omega_l) = V(\omega_l)S(\omega_l) + W(\omega_l) \tag{2-3}$$

其中 $X(\omega_l) = [X_0(\omega_l),\cdots,X_{M-1}(\omega_l)]^T$， $W(\omega_l) = [W_0(\omega_l),\cdots,W_{M-1}(\omega_l)]^T$，

$V(\omega_l) = [\mathrm{e}^{-j\omega\tau_0},\cdots,\mathrm{e}^{-j\omega\tau_{M-1}}]^T$

因为$X(\omega_l)$是宽平稳的高斯随机过程，所以其条件概率分布为

$$p(X \mid \theta) = (\frac{1}{\pi^M \det P}) \exp\{-X^H P^{-1} X\} \tag{2-4}$$

其中$P(\omega_l)$是麦克输入信号$x_i(t)$的互谱密度阵

$$P(\omega_l) = E\{X(\omega_l) X^H(\omega_l)\} \tag{2-5}$$

将(2-3)代入(2-5)可得

$$P(\omega_l) = R_s(\omega_l) V(\omega_l) V^H(\omega_l) + R_W(\omega_l) \tag{2-6}$$

其中$R_W(\omega_l) = E[W(\omega_l) W^H(\omega_l)]$，$R_S(\omega_l) = E[S^2(\omega_l)]$。

对式(2-4)的两边求对数得

$$\ln(p(X(\omega_l) \mid \theta)) = -\ln(\pi^M \det P) + Q \tag{2-7}$$

其中

$$Q = -X^H P^{-1} X \tag{2-8}$$

最大似然参数估计就是选择搜索点参数θ使$p(X \mid \theta)$取最大值，也就是使(2-7)式最大，此时θ即为声源位置的估计值。经过变换，求(2-7)式最大等效于求

$$P(\omega_l) = |H(\omega_l)|^2 |Z(\omega_l)|^2 \tag{2-9}$$

最大，其中

$$H = \left[R_s^{-1} + V^H R_W^{-1} V\right]^{-1} \tag{2-10}$$

$$Z(\omega_l) = V^H(\omega_l) R_W^{-1}(\omega_l) X(\omega_l) \tag{2-11}$$

上式只求出在ω_l处频率的功率，实际上应对整个频带的功率求和，此时声源θ的估计值可由下式得出

$$\hat{\theta} = \arg\max_{\theta}\{\int P(\omega)\, \mathrm{d}\omega\} \tag{2-12}$$

式(2-12)的物理意义可以由图(2-1)表示

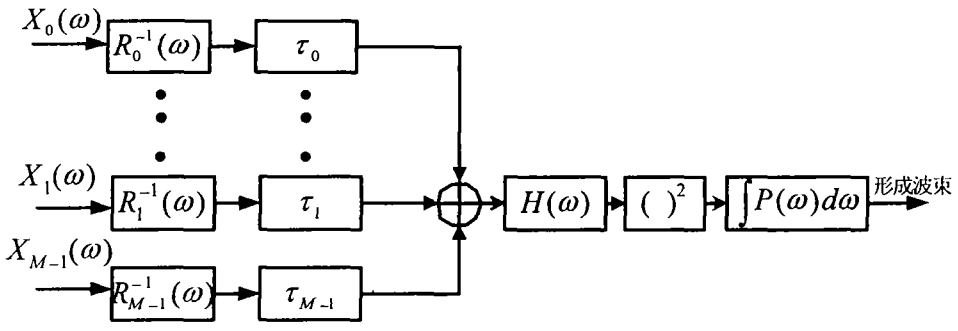


图 2-1 延迟累加可控波束形成定位法的原理框图

在图 2-1 中，首先对输入麦克信号 $x_i(\omega)$加延迟 $\tau_{i,}$，然后对它们累加求和，最后再滤波。其中 $R_i^{-1}(\omega)$ 是噪声功率谱的倒数。它实质上是一种频域加权，对噪声功率大的频率给予小权值，而对噪声功率小的频率给予大权值，这样起到了抑制噪声的作用。 τ_i 表示声源到麦克的延迟时间，通过此时延的补偿可以使各输入信号同步，从而才能使输出功率最大。$H(\omega)$是信噪比滤波，实际也是起压制噪声的作用。

由以上推导可看出，该方法的优点是一步完成定位，且具有最大似然意义上的最优，同时对不相关的噪声有抑制作用。但其缺点也是明显的，首先看图(2-2)描述了一个典型房间的波束功率分布图。x 代表房间宽的方向，y 代表房间长的方向。搜索点的高度为 1.8m。图(2-2a)是该分布的三维图，图(2-2b)是该分布的等值线图。

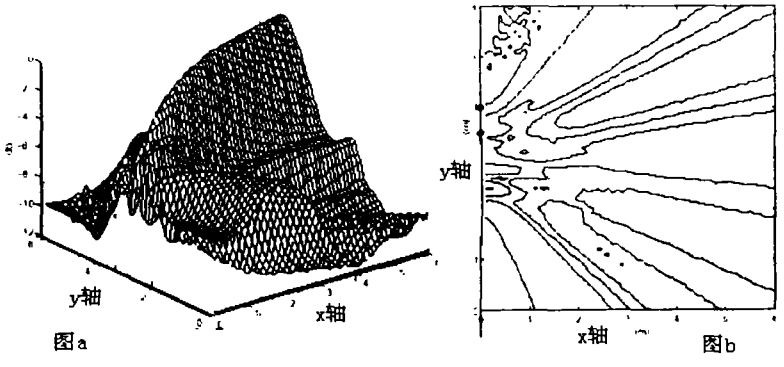


图 2-2 最大似然函数的三维图和等值线图。

从图(2-2)可以看出，该波束分布是具有多个极点的非线性多模态函数，只有通过全局优化搜索方法才能找到声源的最佳估计值，因此需要较大的运算量。

其次在推导过程的假设中，需要信号和噪声的先验知识，在实际中这是不可能的。此外语音信号也不满足平稳性。还有从式(2-1)可以看出，该方法没有考虑混响的影响，因此在房间有较大混响存在时，会降低该方法的定位精度。

由于对信号的相关函数加以一定处理后，可以在一定程度上抑制噪声和混响的影响，因此，可以把求麦克信号的延迟累加和改为求各麦克对相关函数的累加和。

2.2 改进的可控波束形成定位法

从上一节的推导得出，基于可控波束形成的定位方法可以简单地理解为求波束 $J(t)$的功率最大值

$$J(t)=\sum_{i=1}^{M-1}x_i(t+\tau_i(s))$$

其中 $x_i(t)$是第 i 个麦克所接受到的语音信号，$\tau_i(s)$是声音从空间某搜索点 s 到第 i 个麦克的所需的时间。显然，不同的搜索点 s 对应不同的 τ_i，而不同的 τ_i 则对应着不同的波束 $J(t)$。这样使 $J(t)$的功率取最大值的搜索点就对应着所估计声源的位置。从式(2-1)可知，第一节推导并未考虑声音混响的影响。因此当房间中存在比较强的混响时，波束 $J(t)$对应的峰值变的很不明显，这会降低该方法的定位精度。下面对第一节的结论做进一步推导。

对(2-11)式做傅立叶反变换，得到

$$z(t)=\sum_{i=0}^{M-1}g_i(t+\tau_i(\theta)) \tag{2-13}$$

其中 $g_i(t)=\left[R_i^{-1}(\omega)X_i(\omega)\right]^{-1}$

从图 2-1 可以得出，波束的输出功率可以表示为

$$p(t)=(z(t)*f(t))^2=\left\{\sum_{i=0}^{M-1}q_i(t-\tau_i)\right\}^2 \tag{2-14}$$

其中 $f(t)=[H(\omega)]^{-1}$，$q_i(t-\tau_i)=g_i(t-\tau_i)*f(t)$

由于语音信号需分段处理，因此定义短时波束能量为

$$\int_{-\frac{T}{2}}^{\frac{T}{2}} p(t)dt = \int_{-\frac{T}{2}}^{\frac{T}{2}} \left\{ \sum_{i=0}^{M-1} q_i(t-\tau_i) \right\}^2 \mathrm{d}t \tag{2-15}$$

由于语音信号在短时内是平稳的，因此麦克信号间的互相关函数可以表示为

$$\hat{R}(\tau_i - \tau_j) = \int_{-\frac{T}{2}}^{\frac{T}{2}} q_i(t-\tau_i) q_j(t-\tau_j) \mathrm{d}t \tag{2-16}$$

取 $\tau_{ij} = \tau_i - \tau_j$， τ_{ij} 表示麦克 i 相对于麦克 j 的延迟，因此式(2-15)可以表示为

$$\int_{-\frac{T}{2}}^{\frac{T}{2}} p(t)dt = \sum_{i=0}^{M-2} \sum_{j=i+1}^{M-1} \hat{R}_{ij}(\tau_{ij}(\theta)) \tag{2-17}$$

由(2-12)和(2-17)可以得出，使上式取最大的 θ 就是声源的估计值。

图(2-3)是两个麦克信号的互相关函数图

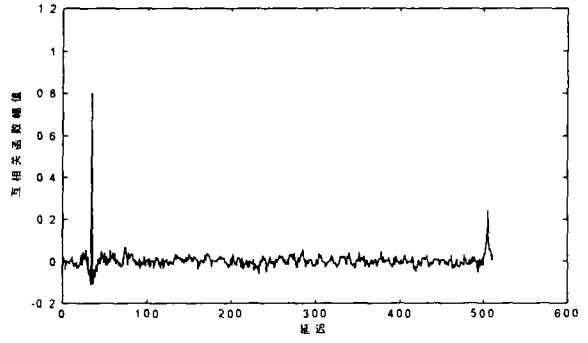


图 2-3 邻近麦克信号间的互相关函数

由图 2-3 可以看出，麦克克的互相关函数在两个麦克的相对时延处有尖锐的峰值。信号间的广义互相关函数(GCC)[10](对信号的互功率谱函数在频域内给于一定的加权)可以抑制噪声和混响的影响。因此用麦克间广义互相关函数的和来形成波束，就可以抑止混响的影响，突出波束功率的峰值。信号间的广义互相关函数可以表示为

$$R_{ij}(\tau) = \int_{\Omega} \psi_{ij}(\omega)\ X_i(\omega)\ X_j^*(\omega) \mathrm{e}^{j\omega\tau_{ij}} \mathrm{d}\omega \tag{2-19}$$

其中 $X_i(\omega)$ 和 $X_j(\omega)$ 分别是 $x_i(t)$ 和 $x_j(t)$ 的傅立叶变换，τ_{ij} 是麦克 i 和 j 间的相对延迟，因此由(2-17)式得出广义互相关函数波束和为

$$P(\theta)=\sum_{i=0}^{M-2}\sum_{j=i+1}^{M-1}R_{ij}(\tau_{ij}(\theta)) \tag{2-20}$$

再由(2-12)式得声源的估计位置为

$$\hat{\theta}=\arg\max_{\theta}P(\theta) \tag{2-21}$$

在实际运用时，可根据噪声和混响的具体情况，在频域内调节 $\psi_{lk}(\omega)$，从而起到抗噪和抑制混响的作用。从以上的推导过程可以看出，该方法在保持原有运算量的前提下，可以起到抗噪声和混响的影响。

2.3 实验结果

定位系统的房间大小设为 4×7×3m，采样率为 16KHZ，信号窗采用汉宁窗。信噪比为 12dB，混响为 200ms。麦克的数量为 8 个，成双线性排列，如图(2-4)所示

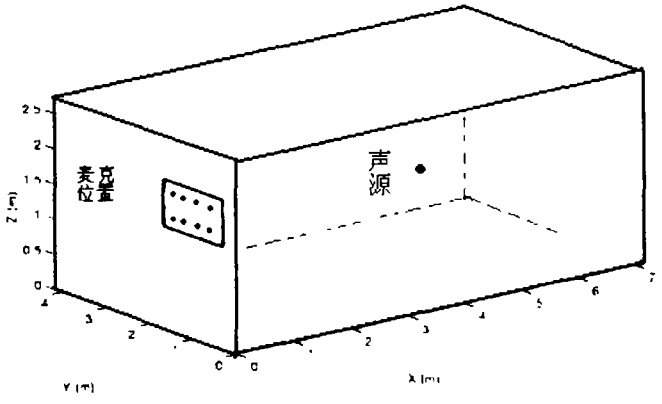


图 2-4 房间和麦克的三维视图

改进后的可控波束形成法采用的加权因子为

$$\psi_{ij}(\omega)=\frac{1}{\left|X_i(\omega)\ X_j^{*}(\omega)\right|} \tag{2-22}$$

采用 Matlab 中的 fminsearch()函数来优化(2-17)式和(2-19)式。在模拟实验中，原点定位在两列麦克的中心。实际声源离原点的距离为 5 米，用极坐标 $(5,\alpha,\beta)$ 表示声源搜索点 θ。不同的 θ 对应不同的波束功率值，在图中用灰度来表示。图(2-5)表示用(2-17)式优化的结果，图(2-6)表示用(2-19)式优化的结果。图中的圆形亮点表示实际声源位置，叉号表示这两种方法所估计的声源位置。九幅图依次表示连续处理的九帧语音，从而可以从平均意义上来衡量定位的效果。

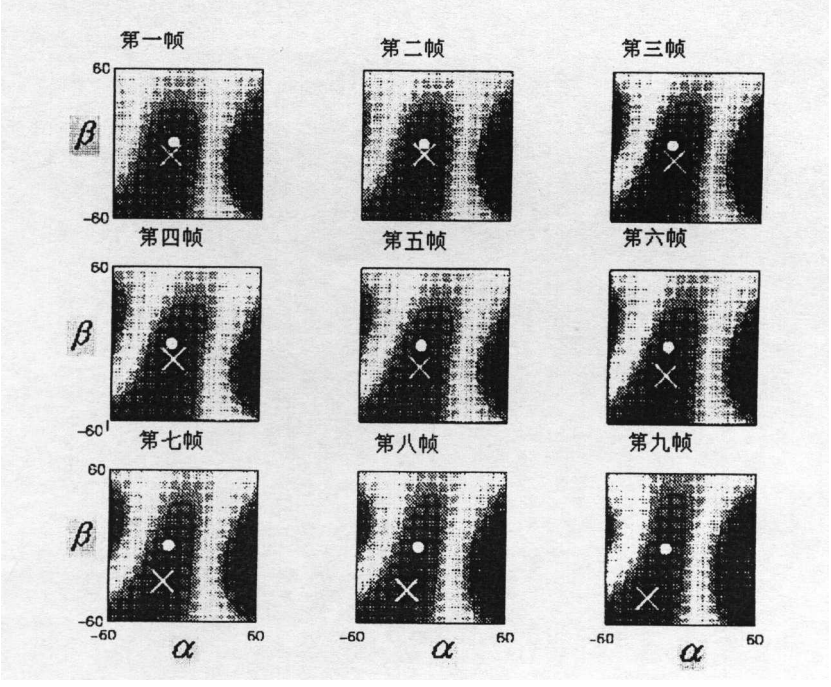


图 2-5 原始可控波束形成法的定位结果

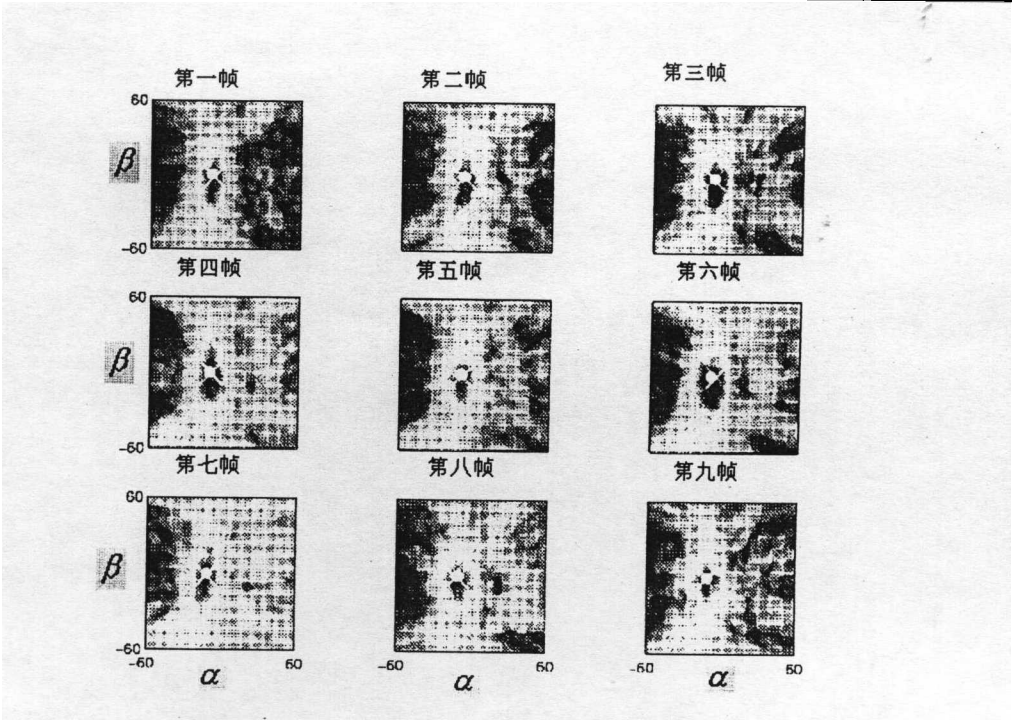


图 2-6 改进可控波束形成法的定位结果

从图 2-5 和图 2-6 明显可以看出，改进后的可控波束形成法所形成的波束峰值尖锐，声源位置的估计值接近声源的实际位置，定位比较精确。而未加改进的可控波束形成法由于噪声和混响的影响，其波束的峰值被明显扩展，很难辨别出最大值的位置，因此导致定位误差明显大于改进后的可控波束形成法。

第三章 时延估计的方法

时延估计-几何定位两阶段法是近几年发展起来的一种可实时实现的定位方法。该方法的关键就是时延估计的精确度。本章论述了两种时延估计的方法，即广义互相关时延估计法(GCC)[10]和自适应时延估计法[11,12]，并对广义互相关时延估计法(GCC)中的相位变换加权法(GCC-PHAT)[13,14]加以改进。此外还介绍了一种源于人耳定位原理的时延估计新方法。

3.1 麦克信号产生模型

假设两麦克 m_1 和 m_2 间距为 D，在没有混响的情况下，两麦克接受到的信号 $x_1(n)$和 $x_2(n)$为

$$x_1(n) = \alpha_1 s(n-\tau_1) + w_1(n) \tag{3-1}$$

$$x_2(n) = \alpha_2 s(n-\tau_2) + w_2(n) \tag{3-2}$$

其中 $s(n)$为声源信号，$w_1(n)$和 $w_2(n)$是不相关的高斯白噪声，$s(n)$和 $w(n)$也是不相关的随机信号。 τ_1 和 τ_2 是声波从声源到麦克的传播时间， α_1 和 α_2 是声波衰减系数。

当房间内存在混响时，两麦克接受到的信号 $x_1(n)$和 $x_2(n)$为

$$x_1(n) = h_1(n) * s(n-\tau_1) + w_1(n) \tag{3-3}$$

$$x_2(n) = h_2(n) * s(n-\tau_2) + w_2(n) \tag{3-4}$$

其中 $h_1(n)$和 $h_2(n)$是房间的单位冲激响应

$\tau_{12} = \tau_1 - \tau_2$ 就是两麦克 m_1 和 m_2 间的时延。一个好的时延估计算法不仅能在低信噪比和较强的混响下精确地估计出时延,而且应该具有比较低的运算量。下面各节依次讨论各种时延估计算法。

3.2 广义互相关时延估计法

声源到两麦克的信号 $x_1(n)$和 $x_2(n)$的互相关函数 $R_{12}(\tau)$可表示为

$$R_{12}(\tau) = E(x_1(n)x_2(n-\tau)) \tag{3-5}$$

将式(3-1)和(3-2)代入(3-5)式，得

$$\begin{aligned} R_{12}(\tau) = {} & \alpha_1\alpha_2 E(s(n-\tau_1)s(n-\tau_2-\tau)) + \alpha_1 E(s(n-\tau_1)w_2(n-\tau)) + \\ & \alpha_2 E(s(n-\tau_2-\tau)w_1(n) + E(w_1(n)w_2(n-\tau)) \end{aligned} \tag{3-6}$$

因为 $w_1(n)$和 $w_2(n)$是不相关的高斯白噪声，$s(n)$和 $w(n)$也是不相关的随机信号，所以由式 (3-6)可得

$$R_{12}(\tau) = E(\alpha_1\alpha_2 s(n-\tau_1)s(n-\tau_2-\tau)) = \alpha_1\alpha_2 R_s(\tau-(\tau_1-\tau_2)) \tag{3-7}$$

由相关函数的性质得，当 $\tau_{12}=\tau_1-\tau_2$ 时 $R_{12}(\tau)$取最大值。因此求得 $R_{12}(\tau)$的最大值对应的 τ就是两麦克间的时延 τ_{12}。

由互相关函数与互功率谱的关系可得

$$R_{12}(\tau) = \int_0^{\pi} G_{12}(\omega)e^{-j\omega\tau}d\omega \tag{3-8}$$

其中 $G_{12}(\omega)$为麦克信号 $x_1(n)$和 $x_2(n)$间的互功率谱。

但是噪声和语音的短时处理导致 $R_{12}(\tau)$的峰值不明显，降低了时延 τ_{12} 估计的精度。为了锐化 $R_{12}(\tau)$的峰值，可以根据信号和噪声的先验知识，在频域内，给互功率谱 $G_{12}(\omega)$一定的加权来抑制噪声和混响的影响。此时再反变换到时域，得到的互相关函数就是广义互相关函数(GCC)，即

$$R_{12}^{g}(\tau) = \int_0^{\pi} \psi_{12}(\omega)\ G_{12}(\omega)e^{-j\omega\tau}d\omega \tag{3-9}$$

在实际中，针对不同的噪声和混响情况，可以选择不同的加权函数$\psi_{12}(\omega)$，使 $R_{12}^{g}(\tau)$具有比较尖锐的峰值。但是此时，由于低信噪比和有限窗长往往使这种分析不稳定，因此加权函数$\psi_{12}(\omega)$的选择是个难点。以下讨论加权函数。

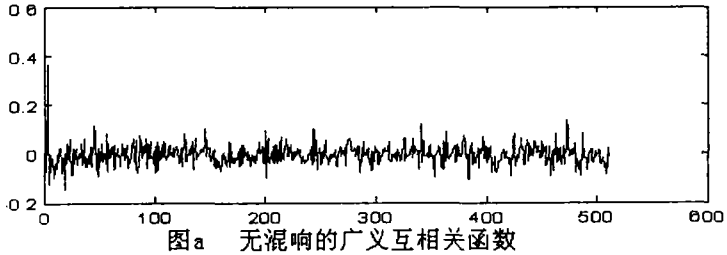
3.2.1 最大似然加权函数

最大似然加权函数为[10](具体推导见文献)

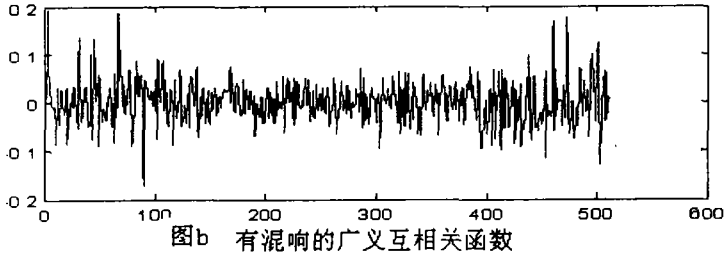
$$\psi_{12}(\omega)=\frac{|\gamma(\omega)|^2}{|G_{12}(\omega)|(1-|\gamma(\omega)|^2)} \tag{3-10}$$

其中$|\gamma(\omega)|^2$为两麦克接收到的信号$x_1(n)$和$x_2(n)$的模平方相干函数。最大似然加权函数实质是一个频域信噪比函数。它对信噪比大的频段给于大权值，而对信噪比小的频段给于小权值，从而比较好的抑制了噪声的影响。

但是该加权函数只考虑了噪声，如果有混响的影响，则会极大的影响时延估计的精度。在10dB的信噪比下，图(3-1)a是无混响的广义互相关函数，图(3-1)b是混响下的广义互相关函数



图a　无混响的广义互相关函数



图b　有混响的广义互相关函数

图 3-1 无混响和有混响下的广义互相关函数

可以看出，在混响干扰下的广义互相关函数的最大值不明显，会给时延估计带来很大的误差。一种改进办法就是在信号前端加上预处理，以去除混响的影响，在本节中采用倒谱技术来去除混响的影响[26]。

此时采用的信号模型为式(3-3)和(3-4)，对(3-3)两边做倒谱变换，得

$$\hat{X}(k)=\hat{h}(k)+\hat{s}(k)+\hat{\eta}(k) \tag{3-11}$$

此处$\hat{X}(k)$，$\hat{h}(k)$，$\hat{s}(k)$分别是$x(n)$，$h(n)$，$s(n)$的倒谱。

单位冲激响应$h(n)$的傅立叶变换为$H(\omega)$，它可以分为最小相位部分(MPC)

和全通部分(APC)，即下式

$$H(\omega) = H_{ap}(\omega) \cdot H_{\min}(\omega) \tag{3-12}$$

其中$|H_{ap}(\omega)|=1$。

单位冲激响应$h(n)$的倒谱可以表示为

$$\hat{h}(k) = \hat{h}_{ap}(k) + \hat{h}_{\min}(k) \tag{3-13}$$

其中$\hat{h}_{\min}(k)$和$\hat{h}_{ap}(k)$分别对应最小相位部分(MPC)和全通部分(APC)的倒谱。由最小相位的性质[40]可得倒谱的最小相位部分为

$$\hat{h}_{\min}(k) = \begin{cases} 0, & k<0 \\ \hat{h}(0), & k=0 \\ \hat{h}(k)+\hat{h}(-k), & k>0 \end{cases} \tag{3-14}$$

倒谱的全通部分为

$$\hat{h}_{ap}(k) = \begin{cases} \hat{h}(k), & k<0 \\ \hat{h}(0), & k=0 \\ -\hat{h}(-k), & k>0 \end{cases} \tag{3-15}$$

因此，输入信号$x(n)$的倒谱可以表示为

$$\hat{X}(k) = \hat{h}_{\min}(k) + \hat{h}_{ap}(k) + \hat{s}(k) + \hat{\eta}(k) \tag{3-16}$$

对于输入信号$x(n)$而言，时延τ_{12}在频域相当于乘上因子$e^{-j\omega\tau_{12}}$，该因子只影响全通部分(APC)，而不影响最小相位部分(MPC)。但如果直接从信号倒谱$\hat{X}(k)$中减去该信号的最小相位倒谱$\hat{X}_{\min}(k)$部分，然后再做广义互相关 GCC，发现信号的相关性变得很差。而如果只减去$\hat{h}_{\min}(k)$部分，则不仅能较好的消除混响的影响，而且相关函数具有比较尖锐的峰值。因此在倒谱预滤波方法中，需要先估计出$\hat{h}_{\min}(k)$，然后从$\hat{X}(k)$中减去$\hat{h}_{\min}(k)$部分，得到$\hat{X}'(k)$；再将$\hat{X}'(k)$

变换到时域，最后做互相关。经过这些处理的 GCC 时延估计就能降低混响的影响，而且也能有一定的抗噪性能。

$\hat{h}_{\min}(k)$ 可以通过式(3-17)估计，即

$$\hat{h}_{\min}(k;m)=\begin{cases}\hat{X}_{\min}(k;m), & m=1\\ (1-\mu)\hat{h}_{\min}(k;m-1)+\mu\hat{X}_{\min}(k;m), & m>1\end{cases} \tag{3-17}$$

其中 μ 的大小可以调节 $\hat{h}_{\min}(k;m)$ 的收敛速度。

该方法虽然可以抑制噪声和混响的影响，但也有自身的缺陷。主要有以下几点：首先，倒谱的估计需要比较长的窗长(200ms)，同时还需要几帧的平滑，因此需要将近 1s 的语音，从而导致了 1s 的延迟。此外，倒谱的计算需多做 2 次 FFT 和一次求对数运算，增加了额外的计算。其次，由于开门、人的走动等一些因素都会导致房间混响的变化，因此需要重新估计 $\hat{h}_{\min}(k)$，这样也会极大地增加运算量。这样该方法虽然效果比较好，当如果要实时处理还比较困难。

3.2.2 相位变换(PHAT)加权函数

相位变换(PHAT)的加权函数为

$$\psi_{12}(\omega)=\frac{1}{|G_{12}(\omega)|} \tag{3-18}$$

从表达式可以看出，相位变换加权函数实质是一个白化滤波器，使信号间的互功率谱变得平坦，从而锐化广义互相关函数。$G_{12}(\omega)$ 是两麦克接受到的信号 $x_1(n)$和 $x_2(n)$的互功率谱。由式(3-7)可得其表达式为

$$G_{12}(\omega)=\alpha_1\alpha_2G_{ss}(\omega)\mathrm{e}^{-j\omega\tau_{12}} \tag{3-19}$$

通过 PHAT 加权函数$\psi_{12}(\omega)$的加权，式(3-9)变为下式

$$R_{12}^{g}(\tau)=\alpha_1\alpha_2\delta(\tau-\tau_{12}) \tag{3-20}$$

上式表明，PHAT 加权函数起到了很好的锐化作用。而且经过 PHAT 加权的互功率谱近似于单位冲激响应的表达式，因此在混响比较弱时，该方法本身就

对混响有一定的抑制作用。但是在实际环境中，由于噪声的存在以及噪声本身的相关性，噪声互功率谱$G_{w1w2}(\omega)$不为零，而且房间还存在比较强的混响，这都会大大影响了 PHAT 加权函数的效果。

在本节中提出针对上述问题的改进。首先针对$G_{w1w2}(\omega)$不为零的情况，在实际中可在无音段估计出噪声互功率谱$G_{w1w2}(\omega)$，然后从信号的互功率谱中减去噪声谱，从而减弱噪声的影响。

其次针对房间的混响，$G_{12}(\omega)$可以分成两部分：直接部分和反射部分

$$G_{12}(\omega)=\alpha_1\alpha_2 G_{ss}(\omega)\mathrm{e}^{-j\omega\tau_{12}}+G_{ss}(\omega)\sum_{i=1,j=1}^{\infty,\infty}\alpha_i\alpha_j\mathrm{e}^{-j\omega t_{ij}} \tag{3-21}$$

由于 PHAT 加权函数的分母$|G_{12}(\omega)|$不等于$\alpha_1\alpha_2 G_{ss}(\omega)$，从而不能保证对$e^{-j\omega\tau_{12}}$部分的白化作用。基于此，我们对 PHAT 加权函数加以修正，乘上一个加权因子γ，使$\gamma|G_{12}(\omega)|=\alpha_1\alpha_2 G_{ss}(\omega)$，这样就可以起到比较好的抗混响效果。实际上加权因子γ是直接分量在功率谱中的比例。可以根据房间的混响时间估算出直接部分在整个功率谱中的比例，从而确定加权因子γ。

PHAT 加权函数$\psi_{12}(\omega)=\dfrac{1}{|G_{12}(\omega)|}$还存在一个问题。当信号能量比较小的时候，$|G_{12}(\omega)|$接近于零。当在定点 DSP 上运行时，就会把$|G_{12}(\omega)|$近似为零。此时 PHAT 加权函数$\psi_{12}(\omega)$就会变得很大，从而带来错误结果。因此，在信噪比低时，可以加上一个加权因子，如下所示

$$\psi_{12}(\omega)=\frac{1}{\gamma|G_{12}(\omega)|+(1-\gamma)|W_{12}(\omega)|^2} \tag{3-22}$$

也就是说，在信噪比低时，就对频域给予一定的信噪比加权(类似于最大似然ML)。

就两路信号求一个时延而言，PHAT 加权函数互功率谱法只需计算 3 次 FFT。因此相对而言运算量比较低。而且其对混响和噪声都有一定的抑制作用，因此改进后的 PHAT 加权函数法适合于在实际定位系统中应用。

3.3 自适应时延估计法

因为式(3-1)和(3-2)中的信号 $x_1(n)$和 $x_2(n)$的产生模型从本质上都未考虑混响，所以在混响比较强的时候，广义互相关时延估计法的效果都会比较差。而自适应时延估计法从时域出发采用(3-3)和(3-4)式的信号产生模型，通过自适应滤波产生 $h_1(n)$和 $h_2(n)$的估计，再从 $h_1(n)$和 $h_2(n)$中估计出时延。该方法从混响入手来估计时延，因此从根本上抑制了混响的影响。下面详细讨论两种自适应滤波法的原理和性能。

3.3.1 自适应时延估计法[11]

图(3-2)是自适应时延估计法的原理框图

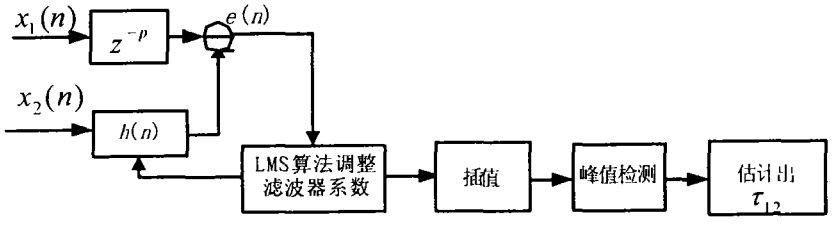


图 3-2 自适应时延估计法的原理框图

对 $x_1(n)$加 z^{-p} 是为了保证因果性而加入 p 个采样周期的延迟，以保证该结构能适应正延迟和负延迟两种情况。

从图(3-2)可以看出，用自适应滤波求时延时，将两麦克所接受到的信号 $x_1(n)$和 $x_2(n)$分别当作目标信号和输入信号，用 $x_2(n)$去逼近 $x_1(n)$。自适应时延估计法的算法如下

$$e(n) = x_1(n-p) - y(n) \tag{3-23}$$

$$y(n) = \sum_{m=-p}^{p} h_m(n) x_2(n-m) \tag{3-24}$$

$$h_m(n+1) = h_m(n) + \mu e(n) x_2(n-m) \qquad m = -p, -p+1, \cdots, 0, 1, \cdots, p \tag{3-25}$$

根据最小均方误差(LMS)准则，当滤波器系数为

$$h(n) = R_{22}^{-1}(n) R_{12}(n) \tag{3-26}$$

时，信号 $x_1(n)$和 $x_2(n)$之间的均方误差 $E\{e^2(n)\}$取最小值。此时滤波器系数 $h(n)$收敛，找出 $h(n)$中最大值对应的 m，再减去 p，就求得信号 $x_1(n)$和 $x_2(n)$间的时延 τ_{12}。

对 $h(n)$插值可以求得连续时间表示的时延，如(3-27)所示

$$\hat{h}(t) = \sum_{m=-p}^{p} h_m(n) \mathrm{sinc}(t-m) \tag{3-27}$$

$\hat{h}(t)$最大值对应的时间 t 再减去 p 即为 $x_1(n)$和 $x_2(n)$间的时延。

式(3-26)的傅立叶变换为

$$H(\omega) = \frac{G_{12}(\omega)}{G_{22}(\omega)} \tag{3-28}$$

式(3-28)相当于加权函数取 $\frac{1}{G_{22}(\omega)}$ 的 GCC 法，两者的区别在于 GCC 方法基于信号和噪声的先验知识，这需由数据估计出来。但在实际中，往往只用一帧数据就获得信号的功率谱和互功率谱的估计，因此该估计的精度不高。而 LMS 自适应滤波则通过一定的误差准则，在收敛的情况下给出时延估计，因此其对功率谱和互功率谱的估计相对来说更精确些。此外，自适应时延估计法还可以处理时变信号，根据信号统计特性的变化自动调节滤波器系数 $h(n)$。

自适应时延估计法也有自己的缺陷，首先它的运算量远远大于 GCC；此外，由于信号 $x_1(n)$和 $x_2(n)$都是通过房间的反射形成的，因此用 $x_2(n)$直接去逼近 $x_1(n)$而得到两者的关系是比较困难的。

3.3.2 改进的自适应时延估计法[12]

针对自适应时延估计法的缺点，特征值分解法考虑了信号 $x_1(n)$和 $x_2(n)$的混响特性，即公式(3-3)和(3-4)中 $h_1(n)$和 $h_2(n)$的影响。

在不考虑噪声影响的情况下，公式(3-3)和(3-4)可以表示为

$$x_1(n) = h_1(n) * s(n) \tag{3-29}$$

$$x_2(n) = h_2(n) * s(n) \tag{3-30}$$

将(3-29)式代入 $x_1(n) * h_2(n)$，可得

$$x_1(n) * h_2(n) = [h_1(n) * s(n)] * h_2(n) \tag{3-31}$$

根据卷积的性质，得

$$[h_1(n) * s(n)] * h_2(n) = [s(n) * h_2(n)] * h_1(n) = x_2(n) * h_1(n)$$

因此

$$x_1(n) * h_2(n) = x_2(n) * h_1(n) \tag{3-32}$$

当有干扰噪声存在的时候，上式不相等，其误差为

$$e(n) = x_1(n) * h_2(n) - x_2(n) * h_1(n) \tag{3-33}$$

也就是说，将自适应时延估计法中误差 $e(n) = x_1(n) - x_2(n) * h(n)$ 中一个混响的逼近变为对两个混响的逼近。图(3-3)为改进自适应时延估计法原理框图。

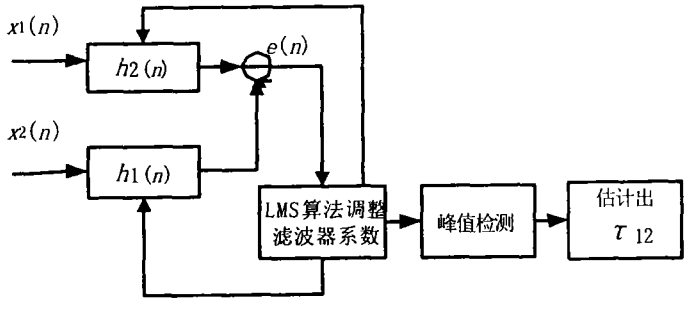


图 3-3 改进自适应时延估计法原理框图

在实际算法中，为了实现方便，我们采用的结构还是原始的 LMS 自适应滤波器结构，但其输入信号和目标信号的取法不同。输入信号将两个麦克信号 $x_1(n)$ 和 $x_2(n)$整和在一起，组成 $X=[x_1(n)\ x_2(n)]$，而对应的目标信号为零。图(3-4)是其实现框图

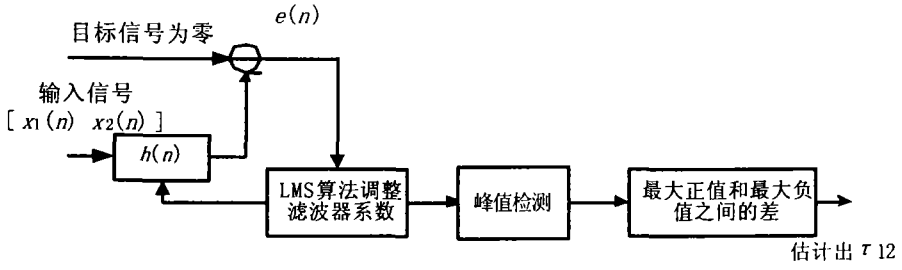


图 3-4 改进自适应时延估计法的实现框图

由以上分析可知，输入信号 X 经过改进的 LMS 自适应滤波器收敛后，对应的滤波器系数 $h(n)$应该为$[h_2(n)\ \ -h_1(n)]$，其中 $h_2(n)$的最大正值对应 $x_2(n)$的单位冲激响应的直接分量，而$-h_1(n)$的最大负值对应 $x_1(n)$的单位冲激响应的直接分量。提取这两个最大值所对应 n 之间的差即是信号 $x_1(n)$和 $x_2(n)$之间的时延 τ_{12}。

下面分析改进的自适应时延估计法的性能。由于 LMS 自适应滤波在时域内要作卷积，因此其运算量会很大。而当一个长信号 $x(n)$与一个有限长的信号 $h(n)$作卷积时，可以用 FFT 变换来实现信号的快速卷积[40]。一个无限长信号 $x(n)$与一个有限长的信号 $h(n)$的线性卷积，等效于将 $x(n)$分解为半重叠的信号块与 $h(n)$作圆周卷积。因此，$y(n)=x(n)*\ h(n)$可以用 $\mathrm{FFT}^{-1}[x_i(n)*\ h(n)]$来实现，其中 $x_i(n)$和 $x_{i-1}(n)$为半重叠的相邻信号块。针对这个结构特点，在仿真实验中，我们采用了频域 LMS 自适应滤波方法[34]。

特别需要注意的是，在仿真实验中，不需要精确的求出 $h_1(n)$和 $h_2(n)$，而只需求出其直接分量，因此在仿真中初始化十分重要。在仿真实际中，取 $h_1(n)$的中点为-1，其余分量为零。在滤波器系数收敛时，$h_2(n)$必然有个最大的正峰值，找出这个最大值与中点的偏离数，即为延迟 τ_{12}。实验中的语音参数同第一章的约定，同时对语音加上了 200ms 的混响，信噪比为 9dB，收敛因之 μ 为 0.03，遗忘因子 α=0.06，信号的初始功率 $P^2(0)$ 设定为第一帧语音的平均功率。

图(3-5)是算法收敛时 $h_2(n)$的时域图。

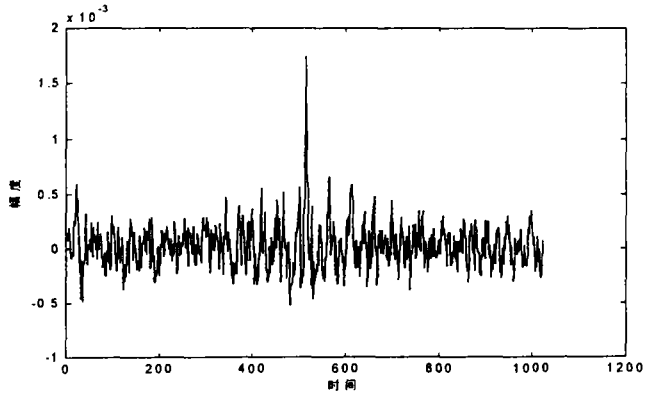


图 3-5 改进自适应时延估计法收敛时 $h_2(n)$的时域图

虽然该方法对混响能有很好的抑制作用，但是改进自适应时延估计法还是有以下缺点：改进自适应时延估计法的运算量虽然通过频域变换变小了，但是还是需要 7 次 FFT，比起 GCC-PHAT 的三次还是大一倍，而且还需要几帧的平滑，因此增大了计算量；改进自适应时延估计法主要估计房间单位冲激响应 $h_1(n)$和 $h_2(n)$的峰值位置，所以 $h_1(n)$和 $h_2(n)$其余分量的估计就不精确，因此无法通过对 $h_1(n)$和 $h_2(n)$插值提高时延估计的精度。

3.4 基于人耳定位原理的时延估计法

根据神经生物学，人耳利用两耳间强度差(IID)和两耳间时间差(ITD)来确定声源的位置[15]。人在有混响的房间里也能正确的辨认出声源的位置，这主要是利用了声音的超前效应[16]，即声音的直接分量总是先于反射分量到达人耳，也就是说人耳利用了未被反射污染的声音段来定位，这段声音称为初始段。而在求时延时，通过提取这段声音求 GCC，就能较好的抑制混响的影响。

此外，由于语音信号从小段来看，明显呈周期性，而只有语音的包络体现了该语音段的特点，因此，如果提取出语音信号的包络来求解信号的相关函数，就会取得比较好的效果。

3.4.1 包络和初始段相关算法[39]

两个麦克信号为 $x_1(n)$和 $x_2(n)$，则两信号的包络分别为

$$env_1(n) = \max[\beta \cdot env_1(n-1), |x_1(n)|] \tag{3-34}$$

$$env_2(n) = \max[\beta \cdot env_2(n-1), |x_2(n)|] \tag{3-35}$$

其中 β 是包络衰减因子($0<\beta<1$)。包络信号的特点是快速响应信号幅度的上升，而衰减比较慢，从而可以掩盖反射部分。β 的选择应该遵循这样的原则：即使包络的衰减比反射部分的衰减慢，这样就能有效的抑制反射部分。

接下来，从包络信号中提取初始段，主要是提取包络的上升部分，即

$$onset_1(n) = \max[0, env_1(n) - env_1(n-1)] \tag{3-36}$$

$$onset_2(n) = \max[0, env_2(n) - env_2(n-1)] \tag{3-37}$$

图(3-6)通过单位冲激响应信号描述了信号包络和初始段的提取过程。图(3-6-1)代表单位冲激信号。图(3-6-2)中的第二个脉冲是该单位冲激经过衰减和墙的反射作用形成的。将包络提取公式应用于图(3-6-2)中的信号，得到了该单位冲激响应信号的包络。再将初始段提取公式应用于图(3-6-3)中的包络信号，主要是提取上升段，即得到单位冲激响应信号的初始段。从图(3-6-4)可以看出，初始段信号去除了经反射形成的第二个脉冲。

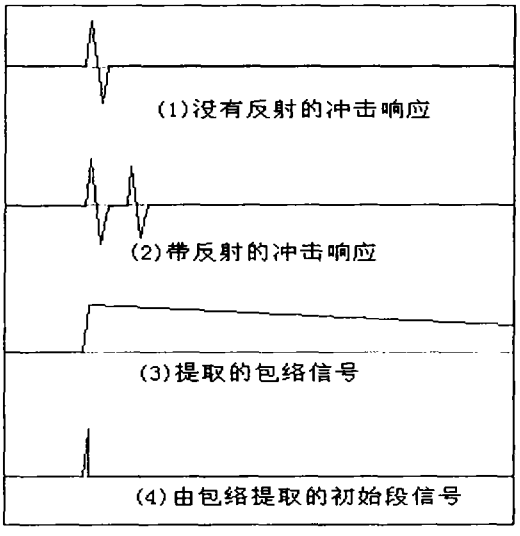


图 3-6 信号包络，初始段的提取

3.4.2 实验结果

假设墙的反射系数为 0.5，衰减系数 β 为 0.995，采样率为 8KHZ，原始的语

音信号为 $x_1(n)$和 $x_2(n)$(由于篇幅所限，并未给出)。图(3-7)表示加混响的语音信号 $x_{1r}(n)$和 $x_{2r}(n)$

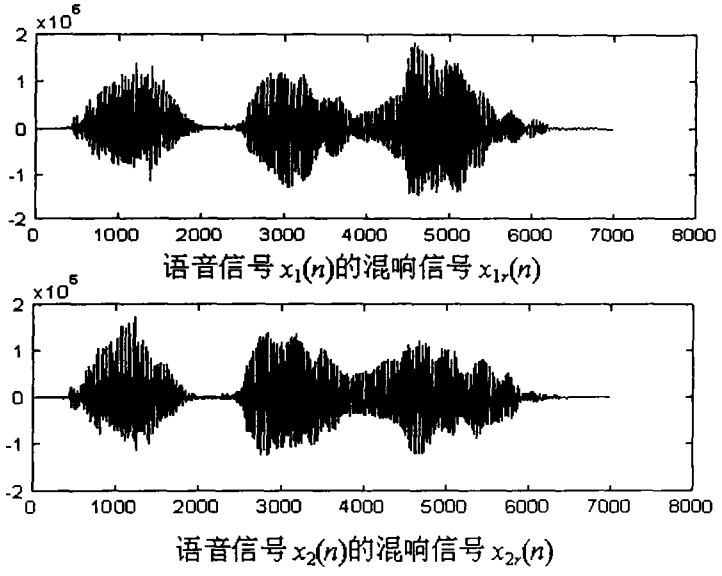


图 3-7 加混响的语音信号 $x_{1r}(n)$和 $x_{2r}(n)$

图(3-8)是混响语音信号 $x_{1r}(n)$和 $x_{2r}(n)$的包络

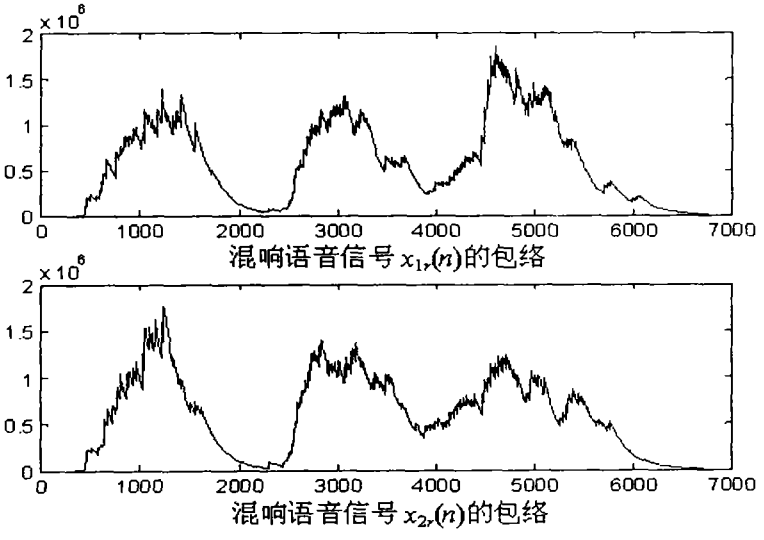


图 3-8 混响语音信号 $x_{1r}(n)$和 $x_{2r}(n)$的包络

图(3-9)是混响语音信号 $x_{1r}(n)$和 $x_{2r}(n)$的包络初始段

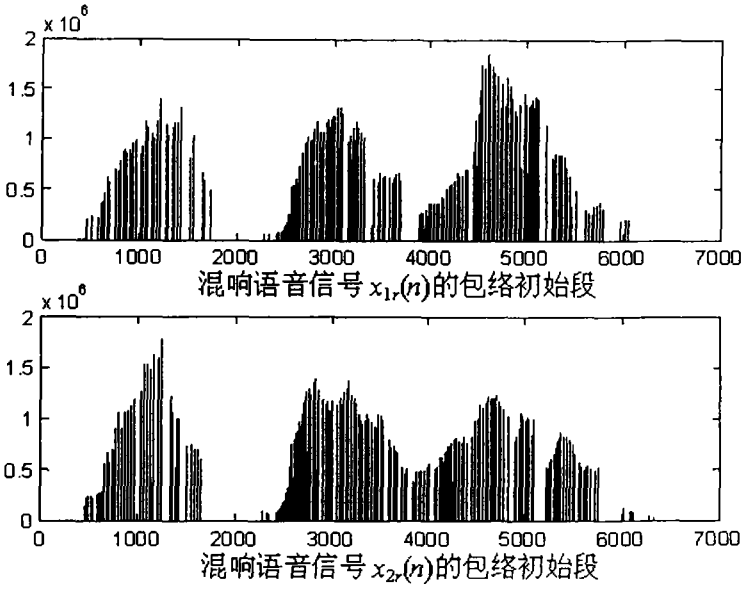


图 3-9 混响语音信号 $x_{1r}(n)$和 $x_{2r}(n)$的初始段

图(3-10)中的 a 图是混响语音信号 $x_{1r}(n)$和 $x_{2r}(n)$的互相关函数，图(3-10)中的 b 图是混响语音信号 $x_{1r}(n)$和 $x_{2r}(n)$的包络初始段的互相关函数。可以明显看出，b 图的互相关函数有更为明显的峰值；而 a 图的互相关函数在多处时延处都有峰值，在房间混响比较强时，就会给真正时延的确定引入较大的误差。

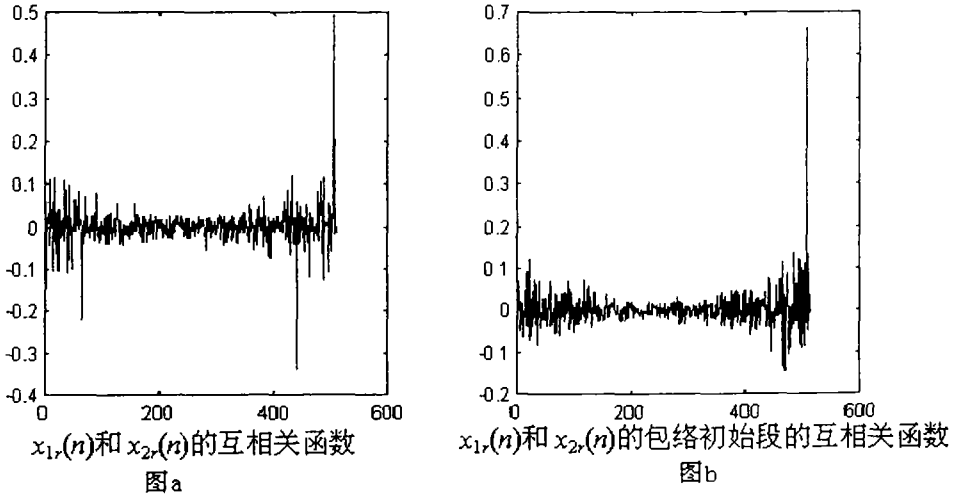


图 3-10 互相关函数的比较

第四章 基于时延的定位方法

上一章详细讨论了时延估计的几种方法，本章主要讨论如何根据时延来确定声源的位置。定位的方法主要分为两大类，即搜索的方法和几何定位法[17]。本章首先讨论麦克和声源的几何模型，然后根据麦克的数量和麦克的几何位置，分别讨论角度距离定位法，球形插值法(SI)[18]和线性插值法(LI)[19,20,21]。

4.1 麦克和声源的几何模型

声源和麦克的坐标位置如图 4-1 所示

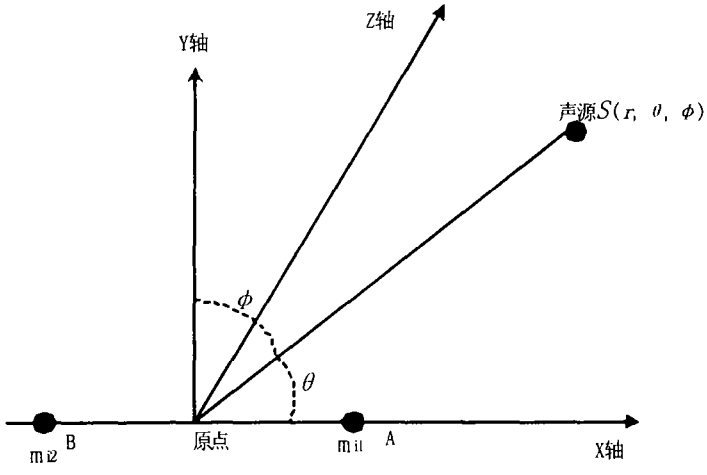


图 4-1 声源和麦克的坐标位置

假设第 i 对麦克 A 和 B 连线的中点为原点，它们的连线为 x 轴，声源到这两个麦克间的时间差是τ_i。用矢量 $\boldsymbol{m}_{i1}$ 和 $\boldsymbol{m}_{i2}$ 表示这两个麦克的位置，用矢量 $\boldsymbol{r}_s$ 表示声源的位置，则声源 S 应该满足矢量方程

$$\|r_s - m_{i1}\| - \|r_s - m_{i2}\| = \tau \cdot c \tag{4-1}$$

其中 c 为声速。由双曲面的定义可得，满足该方程的 S 必落在双曲面上。

由于声源 $s(r,\theta,\phi)$ 是极坐标形式，将其转化为直角坐标形式，可得

$$\boldsymbol{r}_s = (r\cos\theta, r\cos\phi, r\sqrt{1-\cos^2\theta-\cos^2\phi})$$

将上式和 $\boldsymbol{m}_{i1}=(|\boldsymbol{m}_{i1}-\boldsymbol{m}_{i2}|/2,\ 0,\ 0)$、$\boldsymbol{m}_{i2}=(-|\boldsymbol{m}_{i1}-\boldsymbol{m}_{i2}|/2,\ 0,\ 0)$代入(4-1)式，两边平方可得

$$\frac{\cos^2(\theta)}{(c\cdot\tau_i)^2}-\frac{\sin^2(\theta)}{|m_{i2}-m_{i1}|\cdot(c\cdot\tau_i)^2}=\frac{1}{4r^2} \tag{4-2}$$

当声源离麦克比较远时(即 r 变得很大时，$\frac{1}{4r^2}$ 趋近与零)，式(4-2)可以近似为

$$\theta=\cos^{-1}\left(\frac{\tau_i\cdot c}{|m_{i1}-m_{i2}|}\right) \tag{4-3}$$

所以当已知麦克间的时延和麦克间的距离，可以近似求得图(4-1)中的 θ 角。也就是说，当声源离麦克比较远时，可以用以 θ 为顶角的圆锥面来近似代替声源可能的位置。该替换过程可以由图(4-2)来表示

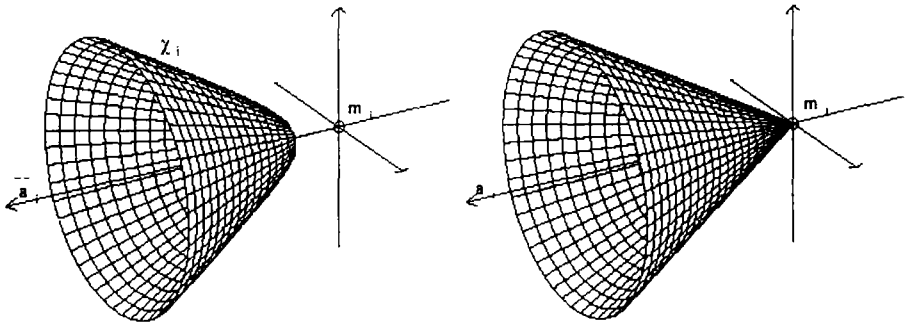


图 4-2 声源的双曲面和锥面近似图

4.2 角度距离定位法

在视频会议系统中，可提供的麦克数少，而且摆放比较固定。图 4-3 就是一种由四个麦克组成的典型麦克摆放方法。

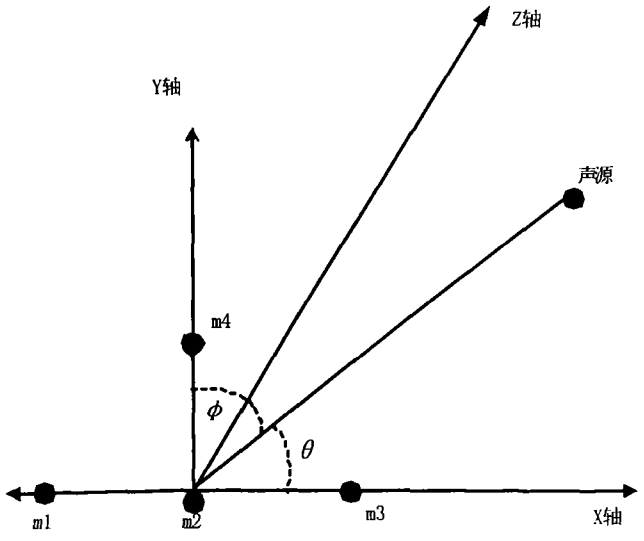


图 4-3 角度距离定位法麦克摆放

假定麦克 2 在原点，麦克之间的间距为 a，声源到各麦克对的距离差为 d_{12}，d_{32}，d_{13}，d_{42}(距离差等于时延乘以声速)。根据第一节中用锥形代替双曲面的几何近似，可以求得声源相对原点和 $\overrightarrow{m_1m_3}$ 的水平角 θ_{azimuth} 为

$$\theta_{\text{azimuth}}=\cos^{-1}\left(\frac{d_{13}}{2a}\right) \tag{4-4}$$

声源相对原点和 $\overrightarrow{m_2m_4}$ 的仰角 $\phi_{\text{elevation}}$ 为

$$\phi_{\text{elevation}}=\cos^{-1}\left(\frac{d_{24}}{a}\right) \tag{4-5}$$

声源相对原点的距离 r 可以由下推出。假定声源坐标为(x，y，z)，则麦克 1 和 3 的坐标是(-a，0，0)和(a，0，0)，可得

$$\sqrt{(x+a)^2+y^2+z^2}-\sqrt{x^2+y^2+z^2}=d_{12} \tag{4-6}$$

$$\sqrt{(x-a)^2+y^2+z^2}-\sqrt{x^2+y^2+z^2}=d_{32} \tag{4-7}$$

将 $r=\sqrt{x^2+y^2+z^2}$ 移到方程的右边并平方，得

$$2xa + a^2 = d_{12}^2 + 2rd_{12} \tag{4-8}$$

$$-2xa + a^2 = d_{32}^2 + 2rd_{32} \tag{4-9}$$

两式相加可得

$$r = \frac{a^2 - (d_{12}^2 + d_{32}^2)/2}{d_{12} + d_{32}} \tag{4-10}$$

在实际定位中，只需将摄像机放在原点，首先转到 θ_{azimuth}，再上下调节仰角 $\phi_{\text{elevation}}$，最后根据 r 调整焦距即可对准说话人。下面给出该方法的性能分析。

时延估计和几何定位法是分成两部分来完成的。为了便于仿真和检验几何定位法的性能，在实验中，用实际声源到两麦克间的距离差来代替时延估计，并加上一个正态分布的随机数。通过调节该正态分布随机数的方差，就可以仿真时延估计偏差的大小。

选择 a=20cm，根据实际声源所在位置，随机产生 100 组与之对应的距离差 d_{12}，d_{32}，d_{13}，d_{42}。再根据这些距离差并利用本节方法估计出声源位置。实验根据声源位置和距离差方差的不同分四种情况测试。

1) 实际声源在(10，200，5)，距离差 d_{12}，d_{32}，d_{13}，d_{42} 的标准差为 0.1
2) 实际声源在(10，200，5)，距离差 d_{12}，d_{32}，d_{13}，d_{42} 的标准差为 0.25
3) 实际声源在(125，200，5)，距离差 d_{12}，d_{32}，d_{13}，d_{42} 的标准差为 0.1
4) 实际声源在(125，200，5)，距离差 d_{12}，d_{32}，d_{13}，d_{42} 的标准差为 0.25

实验 1 和实验 2 的实际声源位置相同，但是所提供距离差的准确度不同。其在实际系统中的意义可以理解为实验 2 中的噪声和混响强于实验 1。

表 4-1 角度距离定位法性能测试

估计声源位置与实际声源位置的偏差				估计声源位置标准差		
	$\hat{\theta}-\theta$ (度)	$\hat{\phi}-\phi$ (度)	$\hat{r}-r$ (cm)	σ_θ	σ_ϕ	σ_r
1	0.81	-1.07	10	0.35	1.39	17.9
2	-1.11	1.56	-49	0.63	2.55	46.12
3	-0.98	-1.5	-50	0.49	1.79	36.58
4	-2.16	2.06	-53	0.72	2.72	55.29

由表 4-1 可知，当方向角 θ 变小时，距离 r 的偏差明显变大，主要原因是 θ 变小时，d_{12} 和 d_{32} 的值变得很接近，因此 $d_{12}+d_{32}$ 的值接近于零。因为 $d_{12}+d_{32}$ 在分母上，所以放大了距离差的偏差。总得来说，该方法算法简单，在 θ 比较大时，效果还可以。不过可以增加一个纵向的麦克来提高距离估计的精度。

4.3 球形插值法

球形插值法根据多个麦克对的时延求得一组方程，并在满足最小均方误差准则下解这个方程组。下面给出详细的推导过程。

首先给出麦克 m_i、m_j 和声源 S 的几何关系图(4-4)

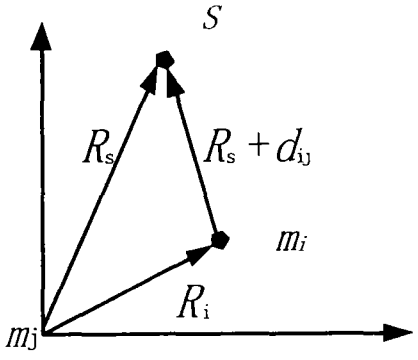


图 4-4 麦克 m_i 及 m_j 和声源的几何关系图

其中 $R_i=|r_i|$，$R_s=|r_s|$，r_i 是 m_j 到 m_i 的矢量，r_s 是 m_j 到声源 S 的矢量，d_{ij} 是声源 S 到麦克 m_i 和 m_j 间的距离差(由时延算法估计出来)

由矢量几何和三角形三边关系可得[18]

$$(R_s + d_{ij})^2 = R_i^2 - 2r_i^T r_s + R_s^2 \tag{4-11-a}$$

将上式展开并整理得

$$R_i^2 - d_{ij}^2 - 2R_s d_{ij} - 2r_i^T r_s = 0 \tag{4-11-b}$$

由于 d_{ij} 是通过估计时延得到的，自然 d_{ij} 与实际值相比有一个偏差，因此上式不为零，其误差为

$$\varepsilon = R_i^2 - d_{ij}^2 - 2R_s d_{ij} - 2r_i^T r_s \tag{4-12}$$

假设有 M 个麦克，记为(0,1, ••• ,M-1)，则可以估计出第(1, ••• ,M-1)个麦

克到第 0 个麦克的距离差，从而根据式(4-12)得到 M-1 个方程，将这些方程写成矩阵形式，可得下式

$$\varepsilon = \delta - 2R_s d - 2Sr_s \tag{4-13}$$

其中

$$\delta = \begin{bmatrix} R_1^2 - d_{10}^2 \\ R_2^2 - d_{20}^2 \\ \vdots \\ R_{M-1}^2 - d_{(M-1)0}^2 \end{bmatrix},\quad d = \begin{bmatrix} d_{10} \\ d_{20} \\ \vdots \\ d_{(M-1)0} \end{bmatrix},\quad S = \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \vdots & \vdots & \vdots \\ x_{M-1} & y_{M-1} & z_{M-1} \end{bmatrix}$$

由于 R_s 和 r_s 的非线性关系，这样若将 R_s=|r_s|代入(4-13)式，就不能给出线性最小均方意义下的声源的估计值 $\hat{r}_s$。而如果给定 R_s，则该方程相对于 r_s 是线性的；反之，如果给定 r_s，则该方程相对于 R_s 也是线性的。

因此该方程的求解必需分为两步，首先假设给定 R_s，可以求得当 $r_s = \frac{1}{2} S_w^* (\delta - 2R_s d)$ 时，式(4-13)的均方误差为最小，其中 $S_w^* = (S^T S)^{-1} S^T$。然后将 r_s 代入(4-13)式，可求得 R_s。将 R_s 代入 $r_s = \frac{1}{2} S_w^* (\delta - 2R_s d)$ 中，即可得到线性最小均方意义下声源的估计值 $\hat{r}_s$。

上述的估计过程求 R_s 和 $\hat{r}_s$ 分成了两阶段。通过对其改进，我们可以只用一步就可求出 $\hat{r}_s$，从而大大降低了运算量。

首先将方程(4-13)改写为下式

$$\varepsilon = A\theta - b \tag{4-14}$$

其中 $A = [S \mid d]$，$\theta = \begin{bmatrix} r_s \\ R_s \end{bmatrix}$，$b = \frac{1}{2}\delta$。

根据矩阵广义逆的原理，方程(4-14)的解为

$$\hat{\theta} = (A^T A)^{-1} A^T b \tag{4-15}$$

把上式写为矩阵分块形式

$$\theta = \begin{bmatrix} S^T S & S^T d \\ d^T S & d^T d \end{bmatrix}^{-1} \begin{bmatrix} S^T \\ d^T \end{bmatrix} b \tag{4-16}$$

进一步简化得

$$\begin{bmatrix} S^T S & S^T d \\ d^T S & d^T d \end{bmatrix}^{-1} = \begin{bmatrix} Q & v \\ v^T & k \end{bmatrix} \tag{4-17}$$

其中 $v = -(S^T S - \frac{S^T d d^T S}{d^T d})^{-1} \frac{S^T d}{d^T d}$

$$Q = (S^T S)^{-1}[I - (S^T d)v^T]$$

$$k = \frac{1 - (d^T S)v}{d^T d}$$

定义投影矩阵 P_d 为

$$P_d = I - \frac{d d^T}{d^T d} \tag{4-18}$$

将上式代入(4-16)，化简得

$$\hat{r}_{s,Sl} = (S^T P_d S)^{-1} S^T P_d b \tag{4-19}$$

利用上式，我们就可一步给出声源位置的估计值，从而大大降低了运算量。

球形插值算法估计误差可以由下式给出

$$J(r_s) = \left\| P_d b - P_d S r_s \right\| \tag{4-20}$$

接下来给出球形插值算法的性能分析。在仿真实验中，采用 9 个麦克，各麦克的间距为 a。时延估计和几何定位法是分成两部分来完成的。为了便于仿真和检验几何定位法的性能，在实验中，用实际声源到两麦克间的距离差来代替时延估计，并加上一个正态分布的随机数。通过调节该正态分布随机数的方差，就可以仿真时延估计偏差的大小。

球形插值算法的麦克摆放如图(4-5)所示

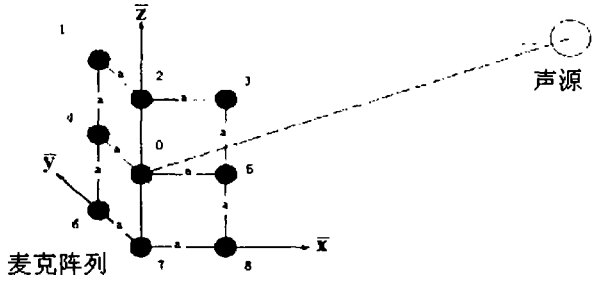


图 4-5 球形插值算法的麦克摆放

假定麦克间距 a=20cm，麦克 0 在原点(0,0,0)。根据实际声源所在位置，随机产生 100 组与之对应的距离差(d_{10}，d_{20}，•••，$d_{(M-1)0}$)。再根据这些距离差利用本节方法估计出声源位置。实验根据声源位置和距离差方差的不同分四种情况测试。

1) 实际声源在(10，200，5)，距离差(d_{10}，d_{20}，•••，$d_{(M-1)0}$)的标准差为 0.1
2) 实际声源在(10，200，5)，距离差(d_{10}，d_{20}，•••，$d_{(M-1)0}$)的标准差为 0.25
3) 实际声源在(125，200，5)，距离差(d_{10}，d_{20}，•••，$d_{(M-1)0}$)的标准差为 0.1
4) 实际声源在(125，200，5)，距离差(d_{10}，d_{20}，•••，$d_{(M-1)0}$)的标准差为 0.25

表 4-2 球形插值算法测试结果

估计声源位置与实际声源位置的偏差				估计声源位置标准差		
	$\hat{x}-x$ (cm)	$\hat{y}-y$ (cm)	$\hat{z}-z$ (cm)	σ_x	σ_y	σ_z
1	0.9	12.5	0.29	0.59	7.3	0.17
2	3.68	14.5	0.27	5.3	39.4	0.8
3	7.3	11.8	0.29	5.11	8.24	0.2
4	19.4	30.3	0.09	10.9	23.05	0.52

由以上测试结果可知，球形插值算法的效果明显优于上一节的方法，而且对于水平角的变化不敏感，也就是说，无论说话者站在哪里，都不会太大的影响定位的精度。

4.4 线性插值法

线性插值法[19,20,21]的基本麦克摆放如图(4-6)所示：

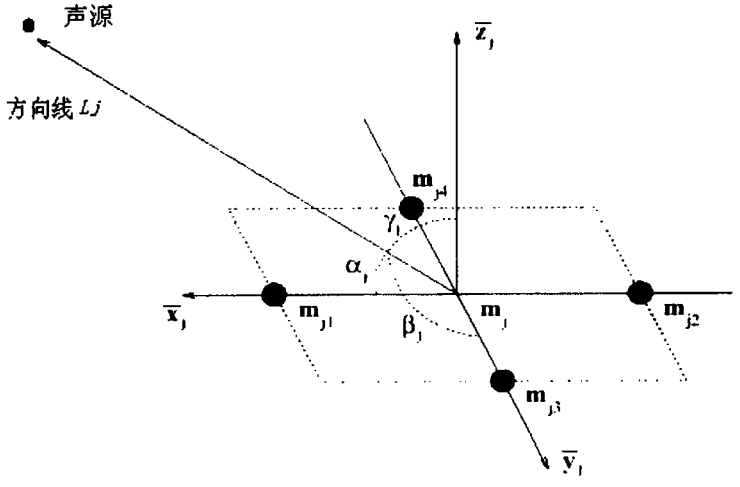


图 4-6 线性插值法的麦克摆放结构

其中麦克 m_{j1} 和 m_{j2} 的连线与麦克 m_{j3} 和 m_{j4} 的连线相互垂直平分，原点为 m_j。当声源离麦克比较远的时候，根据第 1 节公式(4-3)可得，$\cos\alpha_j = \frac{d_{j1j2}}{D_{m_{j1}m_{j2}}}$ 和 $\cos\beta_j = \frac{d_{j3j4}}{D_{m_{j3}m_{j4}}}$，其中 d_{j1j2} 是声源到麦克 m_{j1} 和 m_{j2} 的距离差(等于时延乘以声速)，d_{j3j4} 是声源到麦克麦克 m_{j3} 和 m_{j4} 距离差，$D_{m_{j1}m_{j2}}$ 是麦克 m_{j1} 和 m_{j2} 间的距离，$D_{m_{j3}m_{j4}}$ 是麦克 m_{j3} 和 m_{j4} 间的距离。由空间解析几何可知，声源就在由角 α_j 和角 β_j 唯一确定的直线 L_j 上。

用线性插值法实现声源的定位需要多组如图(4-6)摆放的麦克，如图(4-7)所示：

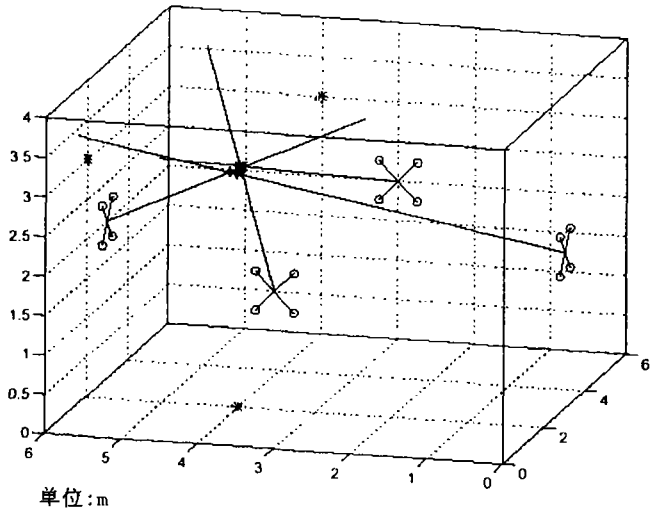


图 4-7 线性插值法定位的三维表示

从图(4-7)可以看出，当有多个这样的麦克对时，就可以得出多条这样的方向线，这些直线的交点就应该是声源的位置。但由于d_{j1j2}和d_{j3j4}估计不准和采样率的限制，直线往往不能相交于一点。假设有两条这样不相交的方向线 l_i 和 l_j，在 l_i 上应该有一点到 l_j 上某一点的距离最近，把这两点记为 S_{ij} 和 S_{ji}，这两点可以看作声源的近似估计点。此外根据时延估计的方差，可设定 S_{ij} 和 S_{ji} 的权值为 W_{ij} 和 W_{ji}，通过这些点的线性插值即可估计出声源的位置 $\hat{S}_{LI}$，即

$$\hat{S}_{LI}=\frac{\sum_{j=1}^{M}\sum_{\substack{k=1\\k\neq j}}^{M}w_{jk}S_{jk}}{\sum_{j=1}^{M}\sum_{\substack{k=1\\k\neq j}}^{M}w_{jk}} \tag{4-21}$$

通过适当的改进，线性插值法也可用于多个声源的定位。对由线性插值法产生的点 S_{ij}，可用聚类的算法(ISODATA 算法)来确定多个声源的位置。ISODATA 算法与 K 均值算法有相似之处，即聚类中心同样是通过 S_{ij} 均值的迭代运算来决定的，但 ISODATA 算法还加入了一些试探步骤，能自动的进行类的合并和分裂，从而得到类数较合理的聚类结果，从而实现了多声源定位。

第五章 一种实际可行的定位方法

通过前几章对各种定位方法的详细讨论，本章提出一种适合于实时实现，且具有较高精度的定位算法。该方法采用时延估计、几何定位两阶段策略。在时延部分采用第三章的改进 GCC-PHAT 法估计时延，然后用降低运算量的球形插值法定出声源的位置。本章首先讨论具体的参数选择，然后给出时延估计的实现框图及实验结果，最后再用球形插值法定位。

5.1 改进的 GCC-PHAT 的实现框图和性能分析

对输入语音用 16K 的采样率，采用半重叠的 512 点汉宁窗。麦克采用第四章球形插值定位法中的摆放方法，麦克位置在墙的一角，其中麦克 1 为坐标原点的位置。

对 9 个麦克一共要求 8 对时延，即 2~9 号麦克对麦克 1 的时延 d_{i1}。提取麦克 1 作为是否有语音的检测麦克。任取其中一对麦克，采用框图(5-1)估计时延。

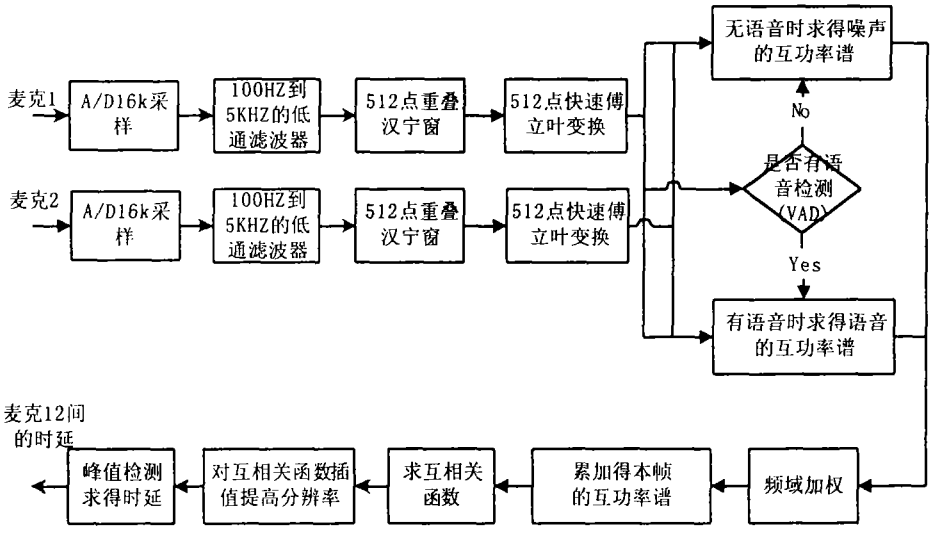


图 5-1 改进的 GCC-PHAT 法估计时延的实现框图

如上图所示，首先由麦克 1 和 2 获得说话人的语音信号，再经过 16k 采样率的 A/D 变换器和低通滤波器，最后得到待处理语音信号，分别记为 $x_1(n)$和 $x_2(n)$。

用 512 点的半重叠汉宁窗对 $x_1(n)$和 $x_2(n)$加窗得

$$X_{1w}(n)= x_1(n)*win(n) \tag{5-1}$$

$$X_{2w}(n)= x_2(n)*win(n) \tag{5-2}$$

其中

$$win(n)=\begin{cases}0.54+0.64\cos\left\{\left[\left(\frac{2n}{N-1}\right)-1\right]\pi\right\}, & n=0,1,\cdots,N-1\\ 0, & \text{其它}\end{cases} \tag{5-3}$$

接着对 $X_{1w}(n)$和 $X_{2w}(n)$做快速傅立叶变换，得

$$X_{1w}(k)=\mathrm{FFT}(X_{1w}(n)) \tag{5-4}$$

$$X_{2w}(k)=\mathrm{FFT}(X_{2w}(n)) \tag{5-5}$$

然后求得 $x_1(n)$和 $x_2(n)$的互功率谱

$$P_{12}(k)=X_{1w}(k)X_{2w}^{*}(k) \tag{5-6}$$

进行有音无音检测(VAD)。当无音时，噪声的互功率谱为

$$W_{12}(k)=P_{12}(k) \tag{5-7}$$

而当有音时，麦克 1 和 2 间的互功率谱即为 $P_{12}(k)$。

为了减弱噪声和混响的影响，根据第三章的公式(3-22)，频域加权因子应为

$$\psi_{12}(k)=\frac{1}{\gamma|P_{12}(k)|+(1-\gamma)|W_{12}(k)|^2} \tag{5-8}$$

这样可以求得加权后的互功率谱为

$$C_{12}(k)=\psi_{12}(k)\times P_{12}(k) \tag{5-9}$$

为了进一步突出峰值，在改进相位变换加权法中，对麦克信号间的互功率谱加了平滑，因此当前帧的加权平滑互功率谱 a_{12}^{m} 为

$$a_{12}^{m}=\begin{cases}C_{12}^{1}, & m=1\\ a_{12}^{m-1}+C_{12}^{m}, & m>1\end{cases} \quad (m\text{ 为帧数}) \tag{5-10}$$

接下来对 a_{12}^{m} 求傅立叶反变换，即可得麦克 1 和 2 间的广义互相关函数 $R_{12}(\tau)$

$$R_{12}(\tau)=\mathrm{FFT}^{-1}(a_{12}^{m}) \tag{5-11}$$

为了提高分辨率，对互相关函数进行插值，得

$$R_{12}^{'}(\tau) = h(\tau) * R_{12}(\tau) \tag{5-12}$$

其中 $h(\tau)$ 为插值滤波器，可以根据实际情况选择。

插值后，$R_{12}^{'}(\tau)$ 的峰值位置即为麦克 1 和 2 间的时延。

在图 5-1 的实现过程中，要注意以下问题：

1) 采用 VAD 技术可有效的检测出是否有人说话，从而省略了不必要的计算。也可控制摄像机在静音时固定在一个位置。此外，还可以在静音时估计出噪声的互功率谱，从而可以在语音信号互功率谱中减去该分量，大大提高了语音信号互功率谱估计的精度。
2) 语音信号采用半重叠的汉宁窗可以提高语音信号互功率谱估计的精度，对语音信号的互相关函数的锐化有一定的作用。
3) 对连续各帧的互功率谱累加求和(平滑)，可以在一定程度上消除噪声的影响，提高互功率谱的估计精度。在实际中发现其效果明显好于单帧估计。
4) 几何定位法需要高精度的距离差。例如角度距离定位法要求距离差的最小单位为 $0.1cm$。所以对 $16k$ 的语音信号，需要对互相关函数 $R_{12}(\tau)$往上插值 20 倍，这里产生的运算量极大。因为三角形两边之差小于第三边，所以声源到麦克间的距离差小于麦克的间距。这样估计出来的时延也是小于 D/c(D 为麦克的间距，c 为声速)，所以对 $R_{12}(\tau)$插值时只需计算 D/c 以内的点。此外，在找峰值时，也只需找 D/c 以内的点，从而可以大大地降低运算量。

仿真实验：

下面给出 GCC-PHAT 法改进前后的比较。其中房间混响由 IMAGE 模型产生，反射系数由第一章给出的公式计算。仿真实验中的参数选择为：$\gamma = 0.85$ (根据房间的具体情况确定)，功率谱平滑帧数 m=4，$h(\tau)$ 由 Matlab 中的 *interp*()函数产生并完成插值。

最后的仿真结果由互相关函数来表示。通过互相关函数的峰值尖锐程度，可以显示该方法估计时延的精度。

未加改进的 GCC-PHAT 法估计的互相关函数为每幅图中的图 a，改进后 GCC-PHAT 法估计的互相关函数为每幅图中的图 b。

实验 1：信噪比 30dB 时两种方法的比较

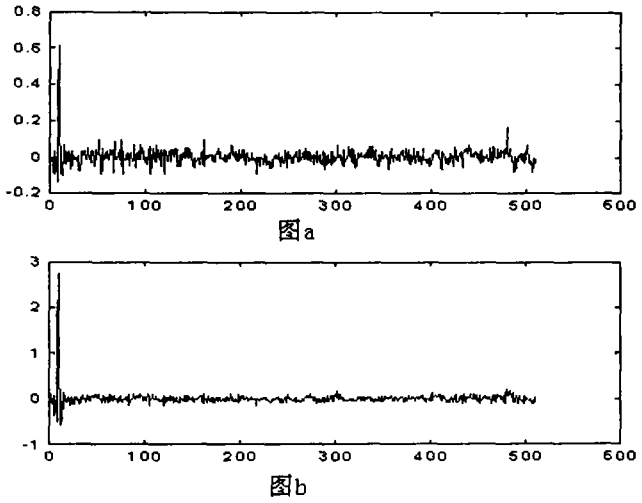


图 5-2：信噪比 30dB 时两种方法的互相关函数比较

由图(5-2)可以看出，在信噪比比较高时，两种方法互相关函数的峰值都比较尖锐，时延估计的精度相差不大。

实验 2：信噪比 10dB 时两种方法的比较

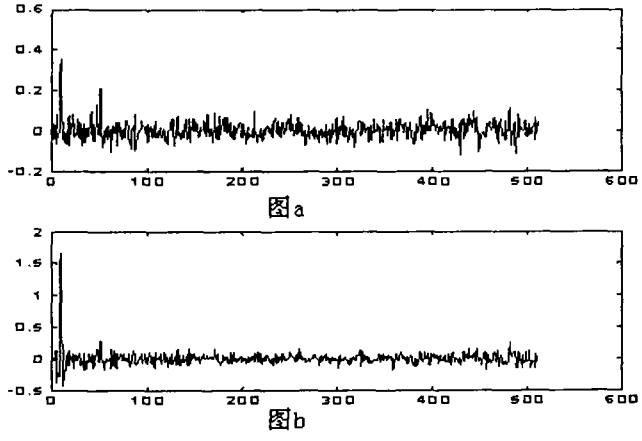


图 5-3：信噪比 10dB 时两种方法的互相关函数比较

由图(5-3)可以看出，在信噪比降低时，方法 1 互相关函数的峰值开始拓展，而且出现比较多的干扰峰值，而方法 2 互相关函数的峰值一样尖锐。

实验 3：信噪比 0dB 时两种方法的比较

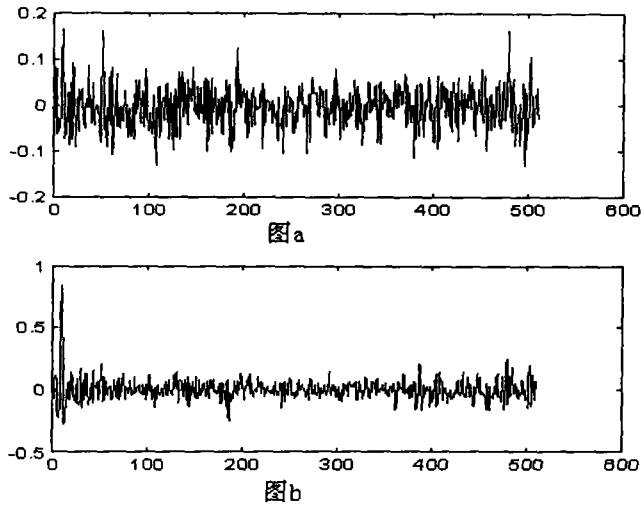


图 5-4：信噪比 0dB 时两种方法的互相关函数比较

由图(5-4)可以看出，在信噪比比较低时，方法 1 互相关函数的峰值已经完全淹没在干扰峰值中，从而根本无法确定时延。而方法 2 互相关函数的峰值还是比较尖锐。

以下测试加上房间混响的影响。为了便于比较，在相同信噪比(10dB)下给出未加混响和加混响时互相关函数的对比图。

实验 4：混响为 200ms，信噪比为 10dB 时两种方法的比较

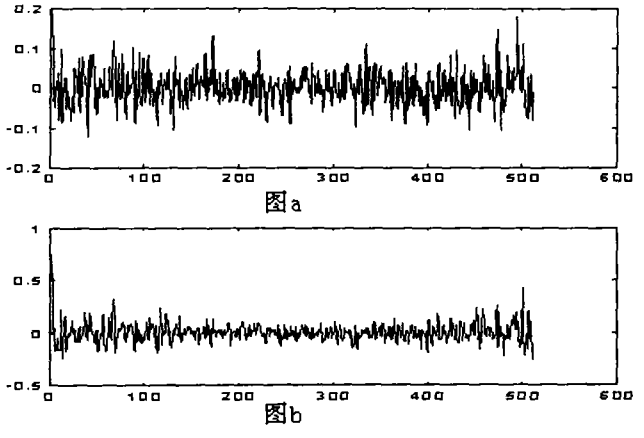


图 5-5：混响为 200ms，信噪比为 10dB 时两种方法的互相关函数比较

由图(5-5)可以看出，在有一定噪声情况下，再加入比较低的混响时，方法 1

互相关函数的峰值已经变得比较难以辨认。而方法 2 互相关函数的峰值还是比较明显的。

实验 5：混响为 400ms，信噪比为 10dB 时两种方法比较

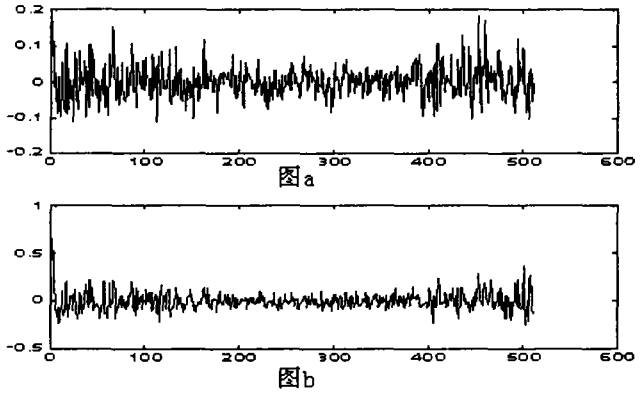


图 5-6：混响为 400ms，信噪比为 10dB 时两种方法的互相关函数比较

由图(5-6)可以看出，在有一定噪声情况下，再加入适中的混响时，方法 1 互相关函数的峰值变得几乎难以辨认。而方法 2 互相关函数的峰值相对来说强于方法 1。

实验 6 混响为 700ms，信噪比为 10dB 时两种方法的比较

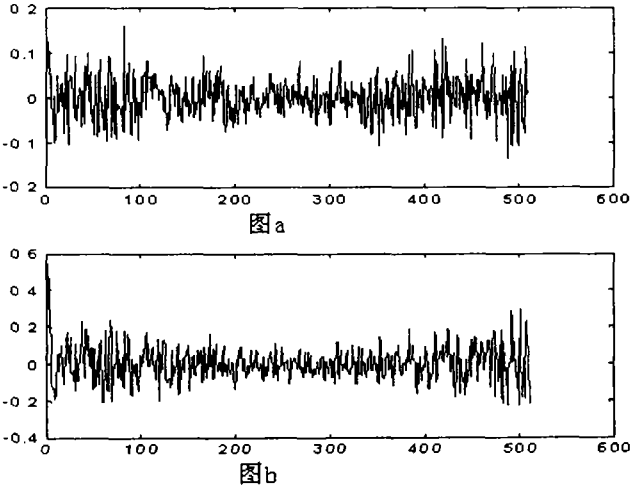


图 5-7：混响为 700ms，信噪比为 10dB 时两种方法的互相关函数比较

由图(5-7)可以看出，在有一定噪声情况下，再加入高混响时，方法 1 互相

关函数的峰值变得无法辨认。而方法 2 互相关函数的峰值还是明晰可见。

从实验结果可以看出，改进的 GCC-PHAT 法在噪声和混响下，都有相对比较尖锐的峰值，因此在估计时延时有很高的精度。

此外，改进的 PHAT 法的运算量远远小于波束法和自适应滤波法，它估计一对时延只需 3 次 FFT，因此完全可以在实时系统中实现。

5.2 基于时延的球形插值定位法

由上节改进的 GCC-PHAT 法，可求得 2~9 号麦克相对于第 1 号麦克的 8 对时延。再乘以声速，就获得 8 对距离差，记为(d_{21}，d_{31}，•••，d_{91})。而 2~9 号麦克相对于第 1 号麦克的距离记为 R_{21}，R_{31}，•••，R_{91}。设麦克 1 在坐标原点，再由麦克位置摆放图 4-5，可求得各麦克的坐标为：(0, 0, -40)，(0, 0, 40)，(40, 0, 0)，(40, 0, 40)，(40 , 0, -40)，(0, 40, 0)，(0, 40, 40)，(0, 40, -40)。

下面给出第四章球形插值定位法中声源估计公式 $\hat{r}_{s,SI}=(S^TP_dS)^{-1}S^TP_db$ 各矩阵的确定。麦克位置矩阵 S 由各麦克的坐标组成

$$S=\begin{bmatrix}0 & 0 & -40\\ 0 & 0 & 40\\ 40 & 0 & 0\\ 40 & 0 & 40\\ 40 & 0 & -40\\ 0 & 40 & 0\\ 0 & 40 & 40\\ 0 & 40 & -40\end{bmatrix} \tag{5-13}$$

由 8 对距离差组成距离差矢量

$$d=[d_{21},\ d_{31},\ \cdots,\ d_{91}]^T \tag{5-14}$$

再由 2~9 号麦克相对于第 1 号麦克的距离组成距离矢量

$$R=[R_{21},\ R_{31},\ \cdots,\ R_{91}]^T \tag{5-15}$$

由球形插值的推导过程可知

$$b=\frac{1}{2}(R.*R-d.*d) \tag{5-16}$$

$$P_d = I - \frac{d * d^T}{d^T * d} \tag{5-17}$$

至此，可由公式$\hat{r}_{s,SI} = (S^T P_d S)^{-1} S^T P_d b$估计出声源的位置。

仿真实验：

为了仿真强噪声和混响，取距离差(d_{21}，d_{31}，•••，d_{91})的标准差为 0.4。根据实际声源所在位置，随机产生 100 组与之对应的距离差(d_{21}，d_{31}，•••，d_{91})。再根据这些距离差利用式(5-17)估计出实际声源位置，用直角坐标($\bar{x}$，$\bar{y}$，$\bar{z}$)表示。

表 5-1：球形插值定位法性能分析

	声源估计位置的均值			声源估计位置的标准差		
实际声源位置	$\bar{x}(cm)$	$\bar{y}(cm)$	$\bar{z}(cm)$	σ_x	σ_y	σ_z
(5,200,5)	4	164	4	0.52	23	0.57
(140,140,5)	125	125	4	46	46	2.1
(200,5,5)	207	6	5	29	2.2	0.68
(5,300,10)	10	299	10	8.6	57	2.0
(212,212,10)	204	204	9	71	71	4
(300,5,10)	323	10	10	48	7.9	1.7
(5,400,-5)	6	366	-5	2.2	104	1.3
(280,280,-5)	267	267	-5	66	66	1.3
(400,5,-5)	350	2	-4	106	1.2	2.3

由表 5-1 可以看出，球形插值定位法在噪声和混响都比较强的情况下，仍然能比较准确的定位出声源的位置。一般各坐标的误差在 40*cm* 内，但估计的方差还是比较大。视频会议的环境噪声和混响一般比较弱，因此定位效果一定强于表 5-1 的结果。因此，该方法可以考虑在实际中应用。

第六章 总结和展望

本文详细地论述了基于麦克阵列的定位技术。首先，在第二章中讨论了波束形成法，并提出了改进，使其在相同计算量下，有比较好的抗噪性，并可抑制一定的混响影响。然后，在第三章和第四章中，讨论了时延估计-几何定位两阶段法，并改进了 GCC-PHAT 法。此外，还介绍了基于语音初始段的时延估计方法。在第四章中，重点论述了各种定位方法，主要有搜索定位法和几何定位的方法，其中着重论述了几何定位的方法。角度距离定位法的算法简单，但效果不好。相对而言，球形插值定位法运算量不大，而且在时延估计有一定的误差时，也能比较精确地定位。线性插值法的定位效果近似于球形插值，运算量也不大，而且不用对高阶矩阵求逆，但是其麦克摆放很不灵活，不适合实际应用。

通过对各种方法的总结和比较，本文提出了一种运算量不大，而且在一定噪声和混响条件下，也能比较准确定位的方法。在第五章中，给出了该定位系统的实现框图，并给出了注意事项和实验结果。从最后的实验结果可以看出，该方法在一定的噪声和混响的干扰下，可以准确的确定声源的位置。

有关基于麦克阵列定位技术的研究尽管已经取得了重要进展，但在理论和应用上还有很多难点需要深入研究解决。首先是麦克的摆放问题。麦克的间距和相对位置可以极大地影响定位的精度。此外，也可利用麦克的空间几何关系来化简定位的几何表达式。作者认为，如果能提出一种优化准则来确定麦克结构，一定会有助于定位算法的设计。其次，需要在强混响和低信噪比环境下准确地进行定位。目前的定位方法往往只考虑噪声或混响的影响，而对两者同时存在的情况，研究的相对较少。当然，采用自适应波束形成技术是解决该问题的一条途径，但如果要得到好的处理效果还需要做大量的工作。由于时间所限，本文并未涉及多声源和移动声源的情况，但这也是声源定位的一个难点。作者建议多声源定位可以考虑从高分辨率谱估计和波束形成方法，而移动声源的定位可以从降低时延估计的运算量入手。

作者在查阅文献和研究工作中深深地体会到，每一项科研成果，都凝聚着一代又一代科学工作者的辛勤汗水，科学的发展离不开积累和创新。限于学识和时间，作者只查阅了很有限的关于麦克风阵列和数字信号处理的文献和研究成果，本论文仅介绍了麦克风阵列声源定位的基本问题以及几种常用的技术，而对其它方法本文没作相应的研究和介绍，感兴趣者可查阅相关文献。

参考文献

[1] G. Carter. "Variance bounds for passively locating an acoustic source with a symmetric line array". Journal of Acoustical Society of America, 1977, Vol.62(4):922-926

[2] W. Hahn, S.Tretter. "Optimum processing for delay vector estimation in passive signal arrays". IEEE Transactions on Information Theory, 1973, Vol.19(1):608-614

[3] W. Hahn. "Optimum signal processing for passive sonar range and bearing estimation". Journal of Acoustical Society of America, 1975, Vol58. (1) :1210-1217

[4] M.Wax, T.Kailath. "Optimum localization of multiple sources by passive arrays". IEEE Transactions on Acoustics, Speech and Signal Processing, 1983, Vol.31(5):1210-1217

[5] T. Shan, M. Wax, T.Kailath. "On spatial smoothing for direction of arrival estimation in coherent signals". IEEE Transactions on Acoustics, Speech and Signal Processing, 1985, Vol.33(8):806-811

[6] R.Schimdt. "A signal subspace approach to multiple emitter location and spectral estimation". PhD thesis, 1981, Stanford university, Stanford, CA,

[7] H.Wang, M.Kaveh. "Coherent signal subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources". IEEE Transactions on Acoustics, Speech and Signal Processing, 1985, Vol.33(4):823-831

[8] K. Buckley, L. Griffiths. "Broad-band signal subspace spatial spectrum estimation". IEEE Transactions on Acoustics, Speech and Signal Processing, 1988, Vol.36(7):953-964

[9] J.B. Allen, D.A. Berkley. "Image method for efficiently simulating small room acoustics" Journal of Acoustical Society of America, 1979, Vol.65(4):943-950

[10] C.H. Knapp, G. C. Carter. "The generalized correlation method for estimation of time delay". IEEE Transactions on Acoustics, Speech and Signal Processing, 1976, Vol.24(4):320-327

[11] D.H. Youn, N. Ahmed, G. C. Carter. "On using the LMS algorithm for time delay estimation". IEEE Transactions on Acoustics, Speech and Signal Processing, 1982, Vol.30(5):798-801

[12] F.A. Reed, P.L. Feintuch, N. J. Bershad. "Time delay estimation using the LMS adaptive filter-static behavior". IEEE Transactions on Acoustics, Speech and Signal Processing, 1981, Vol.29:561-571

[13] M. Omologo, P. Svaizer. "Acoustic event localization using a crosspower spectrum phase based technique". Proceedings of ICASSP, Australia, 1994, pp. 273-276.

[14] M. Omologo, P. Svaizer. "Acoustic source location in noisy and reverberant environment using CSP analysis". Proceedings of ICASSP, Atlanta, GA, 1996, pp. 921–924.

[15] E.I. Knudsen, M. Konishi. "Mechanisms of Sound Localization in the Barn Owl". Journal of Comparative Physiology, 1979, Vol.133:13-21.

[16] P.M. Zurek. "The Precedence Effect and Its Possible Role in the Avoidance of Interaural Ambiguities". Journal of the Acoustical Society of America, 1980, Vol.67:952-964.

[17] H.C. Schau, A. Z. Robinson. "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences". IEEE Transactions on Acoustics, Speech and Signal Processing, 1987, Vol.35(8):1223-1225.

[18] J.O. Smith, J.S. Abel. "Closed-form least-squares source location estimation from range difference measurements". IEEE Transactions on Acoustics, Speech and Signal Processing, 1987, Vol.35(12):1661-1669.

[19] J. Adcock, J. DiBiase, M. Brandstein, H.F. Silverman. "Practical issues in the use of a frequency-domain delay estimator for microphone-array applications". Proceedings of the Acoustical Society of America, Austin 1994.

[20] M.S. Brandstein. "A framework for speech source localization using sensor arrays". Dissertation, 1995 Providence, R. I.: Brown University.

[21] M.S. Brandstein, J.E. Adcock, H.F. Silverman. "A practical time delay estimator for localizing speech sources with a microphone array". Computer, Speech, and Language, 1995, Vol9(2):153-

[22] M.S. Brandstein. "A pitch-based approach to time-delay estimation of reverberant speech". Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 1997, New Paltz, NY.

[23] P. G. Georgiou, C. Kyriakakis, P. Tsakalides. "Robust time delay estimation for sound source localization in noisy environments". Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 1997, New Paltz, NY.

[24] H. Wang, P. Chu. "Voice source localization for automatic camera pointing system in video-conferencing". Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 1997 New Paltz, NY.

[25] B. Champagne, S. Bedard, A. Stephenne. "Performance of time-delay estimation in the presence of room reverberation," IEEE Transactions on Acoustics, Speech and Signal Processing, 1996, Vol.4:148-152.

[26] Stephenne, B. Champagne. "A new cepstral prefiltering technique for time delay estimation under reverberant conditions". Signal Processing. 1997, Vol.59:253-266.

[27] D. R. Morgan, V. N. Parikh, C. H. Coker. "Automated evaluation of acoustic talker direction finder algorithms in the varechoic chamber". Journal of Acoustical Society of

America, 1997, Vol.102:2786–2792 .

[28] H. F. Silverman, S. E. Kirtman. "A two-stage algorithm for determining talker location from linear microphone array data". Computer, Speech and Language, 1992, Vol.6:129-152.

[29] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A closed-form location estimator for use with room environment microphone arrays". IEEE Transactions on Acoustics, Speech and Signal Processing, 1997, Vol.5: 45–50.

[30] D.V. Rabinkin, R.J. Renomeron, A. Dahl, J.C. French, J.L. Flanagan. "A DSP implementation of source location using microphone arrays". Proceedings of SPIE,1996 2846, pp. 88-99.

[31] P.C. Ching, Y. T. Chan, K. C. Ho. "Constrained adaptation for time delay estimation with multipath propagation". Proceedings of IEEE Workshop on Radar Signal Processing,1991, vol.138:453-458.

[32] J.Benesty, F. Amand, A. Gilloire, Y. Grenier. "Adaptive filtering algorithms for stereophonic acoustic echo cancellation". Proceedings of ICASSP, Detroit, MI, 1995, pp. 3099–3102.

[33] Y. Huang, J. Benesty, G. W. Elko. "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization system". Proceeding of IEEE ICASSP, 1999.

[34] D. Mansour, A. H. Gray, Jr. "Unconstrained frequency-domain adaptive filter". IEEE Transactions on Acoustics, Speech and Signal Processing, 1982, Vol.30:726-734.

[35] G. C. Carter, C. H. Knapp, A. H. Nuttall. "Estimation of the magnitude-squared coherence function via overlapped fast Fourier transform processing". IEEE Transactions on Audio Electro-acoustic,1973, Vol.21(4):337-344.

[36] P. M. Peterson. "Simulating the response of multiple microphones to a single acoustic source in a reverberant room". Journal of Acoustical Society of America, 1986, Vol.80:1527-1529.

[37] S. Chiu. "Fuzzy Model Identification Based on Cluster Estimation," Journal of Intelligent & Fuzzy Systems, 1994, Vol.2(3).

[38] Y. T. Chan, K. C. Ho. "A simple and efficient estimator for hyperbolic location". IEEE Transactions on Signal Processing, 1994, Vol.42(8):1905–1915.

[39] S. G. Goodridge. "Multimedia Sensor Fusion for Intelligent Camera Control and Human-Computer Interaction". Dissertation, North Carolina State University.

[40] A.V.Oppenheim, R.W. Schafer, "Digital signal processing", NJ,Englewood Cliffs , Prentice-hall, 1975, Chapter 10