doi:10.3969/j.issn.1000-436x.2014.01.021

复杂环境下基于时延估计的声源定位技术研究

张大威，鲍长春，夏丙寅
（北京工业大学 电子信息与控制工程学院 话音与音频信号处理研究室，北京 100124）

摘 要：为了改善在复杂环境下声源定位算法的性能，提出了一种新的时延估计（TDE）方法，即基于传递函数比的统计模型方法（ATFR-SM）。该方法采用统计模型去除噪声对传递函数（ATF）的影响，在计算传递函数时对功率谱密度（PSD）进行平滑和“白化”，以去除混响对传递函数的影响。同时，算法中引入话音激活检测（VAD）去除对求取传递函数无用的噪声段，以提高时延估计的准确性。此外，将所提时延估计方法与线性定位法相结合，构成一套完整的声源定位方法。实验结果表明，在复杂环境下，时延估计方法具有更低的异常点百分比（PAP）和均方根误差（RMSE），且明显优于传统的参考算法，同时声源定位方法具有更高的定位精度。
关键词：时延估计；传递函数比；VAD；统计模型；声源定位
中图分类号：TN912.3 **文献标识码**：A **文章编号**：1000-436X(2014)01-0183-08

Source localization based on time delay estimation in complex environment

ZHANG Da-wei, BAO Chang-chun, XIA Bing-yin
(Speech and Audio Signal Processing Lab, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China)

Abstract: In order to improve the performance of source localization in noisy and reverberant environments, a novel time delay estimation (TDE) method was proposed. This method is called acoustical transfer function ratio based on statistical model (ATFR-SM). In the proposed algorithm, the noise reduction method based on the statistical model was adopted to reduce the effect of noise on acoustical transfer Function (ATF). In the ATF method, the power spectral density (PSD) was smoothed and whitened to reduce the effect of reverberations. voice activity detection (VAD) was used to distinguish the speech period from the noise period, and the TDE was performed in the speech period to improve the estimation accuracy. Moreover, the proposed TDE method and the linear closed-form method for source localization were combined to constitute a source localization system. The results of performance evaluation show that, in both the noisy and reverberant conditions, the lower percentage of abnormal points (PAP) and lower root mean square error (RMSE) can be achieved by the proposed TDE method than those of the reference methods. Meanwhile, the source localization has higher accuracy than the reference methods.
Key words: TDE; ATF ratio; VAD; statistical model; source localization

1 引言

随着信息技术的发展，声源定位技术在当今生活的很多领域都有着越来越广泛并且极为重要的应用。比如，在视频会议系统中，话音的识别技术、麦克风阵列话音增强和助听装置等方面。在选用麦克风传感器时，相比单个传感器来说，采用多个麦克风阵列使阵列信号处理有着很多的优点，它能够克服单个传感器信息量少的缺点，并利用各阵元信号之间存在的相关性对输入数据

收稿日期：2013-06-28；**修回日期**：2013-12-06
基金项目：北京市教育委员会科技发展计划重点基金资助项目(KZ201110005005)；国家自然科学基金资助项目(61072089)
Foundation Items: The Natural Science Foundation Program and Scientific Research Key Program of Beijing Municipal Commission of Education(KZ201110005005); The National Natural Science Foundation of China(61072089)

进行融合处理以实现对待测参数的估计。由于利用了冗余的有效信息，显著提高了信噪比，因此，由阵列信号处理所得到的估计结果往往具有较高的精度。

随着计算机技术的迅速发展和人机交互需求的快速增加，声源定位这个课题日益显现其重要性。基于时延估计的声源定位技术是利用麦克风阵列确定声源方位的众多方法中最常用的一类[1~4]。该方法分两步执行，首先利用不同麦克风的接收信号进行被动时延估计（TDE，time delay estimation），然后结合时延信息和阵元的几何布局信息估算出目标声源的方位（远场情况）或位置（近场情况）。时延估计是其中至关重要的技术。由于噪声和混响的存在，使时延估计很困难。

目前，人们对复杂环境下的近场声源定位研究比较少，其中，广义互相关法(GCC，generalized cross correlation)[5]是最常用的一种时延估计方法。它需要计算两路信号的互相关函数，时延值就是两路信号互相关函数的峰值位置。然而此种方法假定与每个麦克风信号声源相关的声学传递函数（ATF，acoustical transfer function）是一个完全的时延值。但是在混响环境中，这种近似是不准确的[6]。此外，在低信噪比情况下 GCC 方法不能区分说话人和方向性的干扰信号。因此，它只适合于估计强信号的时延。

之后由 GANNOT 等提出了基于传递函数比的时延估计方法[7,8]。这种方法把环境的影响等效成若干未知传递函数，并将互相关函数的峰值搜索转换成对传递函数比的峰值搜索问题。在安静环境下 ATF 可以达到非常好的性能，然而一旦置于复杂环境，存在各种噪声、混响或干扰信号时，估计性能便会急剧下降。

文献[9]提出了一种频域内使用滤波器长度限制的最小均方算法，此方法对混响的抑制能力效果较好，但没有考虑噪声的影响。由于环境噪声的影响会导致传递函数比的峰值严重下降，因此在时延估计方法中减少噪声的影响十分重要。

基于统计模型的话音增强方法是一类广泛应用的话音增强算法，根据实际要求对话音和噪声的特性设定不同的统计假设，在一定的误差准则下可以得到不同的增强话音幅度谱估计器。这类算法由于应用了信号中包含的某些先验特征，通常可以得到较好的增强效果。统计模型方法的研究始于 Ephraim 和 Malah 在 1984 年提出的最小均方误差(MMSE，minimum mean square error)短时幅度谱估计器(STSA，short-time spectral amplitude)[10]，在高斯统计模型条件下，通过使估计值与真实值的均方误差最小得到增强话音的幅度谱估计，达到消除噪声的目的。基于加权欧式失真测度（WEDM，weighted euclidean distortion measure）的算法通过将幅度谱估计的误差平方用幅度谱的倒数进行加权，达到在谱谷处分配较多的失真，而在谱峰处分配失真较小的效果，有效的压缩了增强话音谱谷处易被人耳感知的残留噪声。

鉴于此，本文提出了一种新的时延估计算法，基于传递函数比的统计模型方法（ATFR-SM，acoustical transfer function ratio based on statistical model）。其结合了统计模型、基于传递函数比和话音激活检测（VAD，voice activity detection）的方法。基于统计模型的话音增强方法对于噪声的去除具有良好的效果。这些方法根据噪声统计特性进行去噪，即这些算法都需要先进行噪声功率谱的估计。这些噪声估计算法是针对话音信号提出来的，又因为话音信号大部分是静音段，因此进行噪声估计较为容易。而传递函数法具有良好的抗混响性能，本文提出的算法将统计模型与改进的传递函数比方法相结合，并加入了 VAD 检测和后处理算法。

2 时延估计新方法

图 1 为本文所提的 ATFR-SM 时延估计算法的原理框图。下面将详细介绍各个主要模块，图 1 中各符号的意义也将在对应模块的介绍中说明。首先两路信号分别加窗并经过快速傅里叶变换(FFT, fast Fourier transform)后，利用最小值控制递归平均算法(MCRA，minima-controlled recursive averaging)噪声估计算法获得先验信噪比和后验信噪比，采用 WEDM 短时幅度谱估计器获取增强后的话音信号，然后使用基于对数似然比(LLR, logarithmic likelihood ratio)和谱熵(SE, spectral entropy)相结合的 VAD 方法计算功率谱密度(PSD, power spectral density)函数，同时加入平滑和“白化”以获取最优的功率谱密度函数，接下来利用改进的传递函数比方法估计传递函数比，求得时延估计初值，最后使用后处理方法去除时延估计结果中的扰动，得到最优时延估计值。

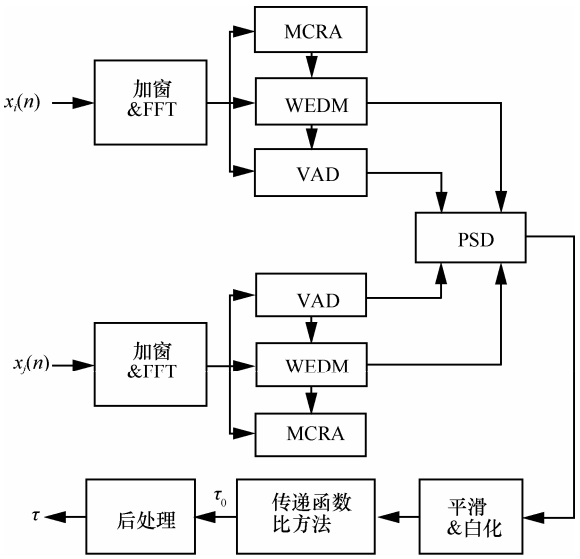

图 1　ATFR-SM 时延估计算法的原理

2.1　噪声估计原理

文献[11]提出了一种基于话音存在不确定度的噪声估计算法，该算法首先计算当前时刻含噪话音的功率谱输入与此时功率谱最小值的比值，通过将该比值与一阈值相比，进而达到估计话音存在不确定度的目的。由于该方法主要通过功率谱最小值估计话音存在不确定度，因此称其为最小值控制递归平均算法。MCRA 方法的步骤如下。

1) 计算平滑含噪功率谱密度 $S(\lambda,k)$ 为

$$S(\lambda,k)=\alpha_s S(\lambda-1,k)+(1-\alpha_s)S_f(\lambda,k) \tag{1}$$

其中，

$$S_f(\lambda,k)=\sum_{i=-L_w}^{L_w} w(i)\left|Y(\lambda,k-i)\right|^2 \tag{2}$$

其中，λ 为帧号，k 为频点，α_s 为平滑因子，L_w 为帧长，$w(i)$ 为哈明窗，$\left|Y(\lambda,k-i)\right|^2$ 为含噪话音功率谱密度，$S_f(\lambda,k)$ 为在频域内的平滑含噪功率谱密度。

2) 采用最小值控制跟踪算法求取含噪功率谱密度的最小值 $S_{\min}(\lambda,k)$ [8]。

3) 计算局部话音存在概率 $P(\lambda,k)$ 为

$$P(\lambda,k)=\begin{cases}1, & S_r(\lambda,k)>\delta\\ 0, & \text{其他}\end{cases} \tag{3}$$

其中，$P(\lambda,k)$ 为话音存在概率，$S_r(\lambda,k)$ 为当前帧含噪话音的功率值与此时功率谱最小值的比值，δ 为阈值。

4) 计算平滑因子 $\alpha_d(\lambda,k)$ 为

$$\alpha_d(\lambda,k)=\alpha+(1-\alpha)P(\lambda,k) \tag{4}$$

其中，α 为固定常数。

5) 更新噪声功率谱密度 $\hat{\sigma}_d^2(\lambda,k)$ 为

$$\hat{\sigma}_d^2(\lambda,k)=\alpha_d(\lambda,k)\ \hat{\sigma}_d^2(\lambda-1,k)+\left[1-\alpha_d(\lambda,k)\right]\left|Y(\lambda,k)\right|^2 \tag{5}$$

其中，$\hat{\sigma}_d^2(\lambda,k)$ 为当前分析帧中频点 k 处的噪声估计功率谱，$\left|Y(\lambda,k)\right|^2$ 表示含噪话音功率谱密度。

2.2　基于统计模型的话音增强方法

后验信噪比定义为

$$\gamma_k=Y_k/\lambda_d(k) \tag{6}$$

其中，$\lambda_d(k)$ 为噪声功率谱密度，Y_k 为含噪话音功率谱。

采用基于判决的方法确定先验信噪比估计 $\hat{\xi}_k$ 为

$$\hat{\xi}_k(\lambda)=\alpha\frac{\hat{X}_k^2(\lambda-1)}{\lambda_d(k,\lambda-1)}+(1-\alpha)\max\left[\gamma_k(\lambda)-1,0\right] \tag{7}$$

其中，$0<\alpha<1$ 为权因子，$\hat{X}_k^2(\lambda-1)$ 为前一帧的振幅估计。

加权欧式失真测试 WEDM 短时幅度谱估计器的表达式如下所示

$$d(X_k,\hat{X}_k)=(X_k-\hat{X}_k)^2/X_k \tag{8}$$

其中，X_k 和 $\hat{X}_k$ 分别为纯净话音的原始幅度谱和估计幅度谱。

利用式(8)中的代价函数，可以令贝叶斯风险函数最小，获得基于统计模型的 WEDM 短时幅度谱估计器[12]。

$$\hat{X}_k=\frac{\sqrt{v_k}}{\sqrt{\pi}\gamma_k}\frac{\exp(v_k/2)}{I_0(v_k/2)}Y_k \tag{9}$$

其中，

$$v_k=\xi_k\gamma_k/(1+\xi_k) \tag{10}$$

$I_0(\cdot)$为 0 阶修正的贝塞尔函数。

2.3　VAD 检测原理

本文的 VAD 检测原理采用对数似然比[13]和谱熵[14]相结合的方法。此 2 种方法充分利用了之前求得的先验信噪比和后验信噪比。

频点 k 处的似然比定义为

$$\Lambda_k=\frac{1}{1+\xi_k}\exp\left\{\frac{\gamma_k\xi_k}{1+\xi_k}\right\} \tag{11}$$

判决准则为

$$\log\Lambda_k = \frac{1}{L}\sum_{k=0}^{L-1}\log\Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{12}$$

其中，η 为阈值，L 表示帧长，H_1 表示话音，H_0 表示非话音。

利用后验信噪比 γ_k 得到第 k 帧的谱熵

$$E(k) = -\sum_{k=1}^{N} P(\gamma_k^2)\log(P(\gamma_k^2)) \tag{13}$$

其中，

$$P(\gamma_k^2) = \gamma_k^2 / \sum_{k=1}^{N}\gamma_k^2 \tag{14}$$

其中，N 为傅里叶变换长度，即窗长。

将噪声帧内谱熵的一阶平滑作为参考，得到谱熵的门限 E_{TH}，即

$$E_{\mathrm{noise}}(k) = \alpha_E E_{\mathrm{noise}}(k-1) + (1-\alpha_E)\ E(k) \tag{15}$$

$$E_{\mathrm{TH}} = \gamma_E(k) E_{\mathrm{noise}}(k) \tag{16}$$

其中，α_E 为平滑因子，$\gamma_E(k)$ 偏差修补因子。

从 NTT 数据库中选取八段纯净话音，分别加入不同信噪比的 babble 噪声，构成总长度为 64 s 的测试话音。将对数似然比方法和谱熵方法相结合，组成 VAD 检测方法。当两者均判决当前帧为话音时，判决当前帧为话音。

VAD 检测通过漏检率（MDR，miss detection rate）和虚警率（FAR，false acceptation rate）测试其检测性能，表 1 为 VAD 算法的性能评价结果。

表 1　　VAD 检测算法性能评价

信噪比/dB	准确率/%	MDR/%	FAR/%
20	96.05	3.95	5.15
15	94.90	5.10	7.11
10	93.66	6.34	8.15
5	91.28	8.72	8.58
0	84.57	15.43	8.70
−5	71.15	28.85	9.74

2.4　改进的传递函数比方法

为了同时解决噪声和混响对时延估计精度的影响，ATF 算法从混响环境下的信号模型出发

$$z_m(t) = a_m(t) * s(t) + b_m(t) * n(t), m = 1, \cdots, M \tag{17}$$

其中，$z_m(t)$ 为第 m 个麦克风的接收信号，$s(t)$ 为纯净信号源，$n(t)$ 为干扰信号，$a_m(t)$ 为目标信号和接收信号之间的传递函数的冲击响应；$b_m(t)$ 为干扰信号和接收信号之间的传递函数的冲击响应，M 为麦克风的个数。

将式(17)两边进行傅里叶变换，得到信号的自功率谱密度与互功率谱密度函数为

$$\boldsymbol{\Phi}_{z_m z_m}(w) = |A_m(w)|^2\ \boldsymbol{\Phi}_{ss}(w) + |B_m(w)|^2\ \boldsymbol{\Phi}_{nn}(w) \tag{18}$$

$$\boldsymbol{\Phi}_{z_m z_1}(w) = A_m(w)A_1^*(w)\boldsymbol{\Phi}_{ss}(w) + B_m(w)B_1^*(w)\boldsymbol{\Phi}_{nn}(w) \tag{19}$$

上式利用了话音与干扰信号之间的不相关性。互功率谱函数满足如下关系为

$$\boldsymbol{\Phi}_{z_m z_1}(w) = \frac{A_m(w)}{A_1(w)}\boldsymbol{\Phi}_{z_1 z_1}(w) + \boldsymbol{\Phi}_{b_1}(w) + \boldsymbol{\varepsilon}(w) \tag{20}$$

其中，$\boldsymbol{\varepsilon}(w)$ 为由环境噪声引起的功率谱误差信号，方向性干扰相关的偏置分量为

$$\varphi_{b_1}(w) = \left[\frac{B_m(w)}{B_1(w)} - \frac{A_m(w)}{A_1(w)}\right]|B_1(w)|^2\ \boldsymbol{\Phi}_{nn}(w) \tag{21}$$

定义目标信号的声学传递函数比为

$$H_m(w) = \frac{A_m(w)}{A_1(w)} \tag{22}$$

利用话音信号的非平稳特性，可以将话音分块之后得到线性方程组为

$$\boldsymbol{\Phi}_{z_m z_1}(n,w) = H_m(w)\boldsymbol{\Phi}_{z_1 z_1}(n,w) + \varphi_{b_1}(w) + \boldsymbol{\varepsilon}(n,w) \tag{23}$$

利用最小二乘法 LS 求解线性方程组，直接得到 $H_m(w)$的估计值。将 $H_m(w)$经傅里叶反变换得到时域函数 $\alpha_{ji}^{\text{ATF-LS}}(\tau)$，并对其进行峰值搜索，得到时间延迟估计 τ_{ij}。

将原始信号加窗经过傅里叶变换后得到的信号称为周期图，通过对周期图进行进一步的平均或者递归平滑，所得到的谱估计方法称为 Welch 方法。设加窗后的信号为

$$x_{iw}(n) = x_i(n)w(n),\ x_{jw}(n) = x_j(n)w(n) \tag{24}$$

定义第 L 帧修正的周期图为

$$P_{x_i x_j}(l,w) = \frac{1}{U}F[x_{iw}(l,n)]\cdot F^*[x_{jw}(l,w)] \tag{25}$$

其中，

$$U = \sum_{n=0}^{N-1} w^2(n) \tag{26}$$

经 L 帧数据平均即得到估计的互功率谱为

$$\hat{\boldsymbol{\Phi}}_{x_i x_j}(w)=\frac{1}{L}\sum_{l=1}^{L}P_{x_i x_j}(l,w) \tag{27}$$

首先利用 Welch 算法，每 L 帧计算求得一个互功率谱和自功率谱后，再利用最小二乘法每 D 个数据块(1 个数据块包含 1 个互功率谱和 1 个自功率谱)求得一个传递函数比。

在计算互功率谱时加入了平滑算法，以减少前后帧互功率谱的波动影响

$$\boldsymbol{\Phi}_{z_m z_1}(\lambda,k)=\begin{cases}\boldsymbol{\Phi}_{z_m z_1}(\lambda,k), & \lambda=1\\ \alpha\cdot\boldsymbol{\Phi}_{z_m z_1}(\lambda-1,k)+(1-\alpha)\cdot\boldsymbol{\Phi}_{z_m z_1}(\lambda,k), & \lambda>1\end{cases} \tag{28}$$

其中，$0<\alpha<1$ 为平滑因子。

为了消除互功率谱的幅度影响并保持相位不变，加入了“白化”算法，为

$$\boldsymbol{\Phi}'_{z_m z_1}(w)=\frac{\boldsymbol{\Phi}_{z_m z_1}(w)}{\left|\boldsymbol{\Phi}_{z_m z_1}(w)\right|} \tag{29}$$

2.5　时延估计的后处理

在获得时延估计的初值后，为去除时延估计中的异常点，将平滑作为时延估计中的后处理模块

$$\tau_i=\begin{cases}\overline{\tau}, & \dfrac{\tau_i-\overline{\tau}}{\overline{\tau}}\geqslant\sigma\\ \tau_i, & \text{其他}\end{cases} \tag{30}$$

其中，τ_i 为第 i 个时延估计值，$\overline{\tau}$ 为 3 帧时延估计值的平均值，σ 为时延估计值与平均值的相对误差的阈值。

3　时延估计性能评价

为验证本文提出的时延估计算法的性能，本文通过仿真实验分别在混响、噪声和混响噪声同时存在的 3 种不同环境下，测试算法性能。其中，本文选择相位变换加权广义互相关函数法（GCC-PHAT，generalized cross correlation-phase transform）和声学传递函数比方法 ATF 作为参考算法。同时采用 Image 模型产生不同混响时间下的房间冲激响应，得到不同的含有混响的接收信号[15]。

3.1　仿真实验环境

时延估计性能测试采用办公室环境作为仿真环境，如图 2 所示，房间大小为 8 m×6 m×3 m 。混响时间从 200 ms~700 ms 变化，采用 P.56 工具加入噪声，噪声输入信噪比由 20 dB 降至−5 dB。加入的噪声为会议室环境中常见的 babble 噪声，噪声源选自于 Noisex92 噪声数据库[16]。

两路麦克风的坐标分别为 M_1(4 m,1 m,1.5 m)和 M_2(4 m, 1.06 m, 1.5 m)，目标声源坐标分别位于 S_1(0.72 m, 3 m, 1.5 m)，S_2(3.02 m, 3 m, 1.5 m)，S_3(4.18 m, 3 m, 1.5 m)和 S_4(6.18 m, 3 m, 1.5 m)，时延估计的结果以采抽样点数衡量，对应的时延估计抽样点数分别为 1、2、3 和 2。从 NTT 中文子数据库中选取 8 段话音作为每个目标声源的测试话音，测试话音总长度为 64 s。以平均结果作为衡量算法性能的最终指标，话音抽样率为 16 kHz。

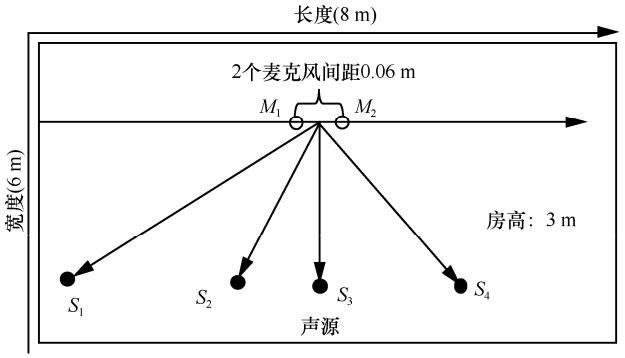

图 2　会议室仿真环境

3.2　性能指标

本文中采用异常点百分比（PAP，percentage of abnormal point）和均方根误差（RMSE，root mean square error）衡量时延估计算法的性能。其中，异常点百分比 PAP 用 P_τ 表示，定义为

$$P_\tau=\frac{1}{N_\tau}\sum_{i=1}^{N_\tau}T(\tau_i-\tau_0),\ T(x)=\begin{cases}0, & |x|<2\\ 1, & |x|\geqslant 2\end{cases} \tag{31}$$

其中，τ_0 为真实时延值，N_τ 为时延估计值总数。

本文中认为时延估计值与真实值相差 2 个或 2 个以上抽样点时，将此估计值计算为一个异常点。

均方根误差 RMSE 定义为

$$\sigma_\tau=\sqrt{\frac{1}{N_\tau}\sum_{i=1}^{N_\tau}(\tau_i-\tau_0)^2} \tag{32}$$

用来衡量时延估计值在真实值附近的分布，这个值越小，表明估计时的时延值分布在真实时延值附近，即表示此方法效果越好。

3.3　仿真实验结果

提出方法 ATFR-SM 和参考方法的异常点百分比和均方根误差如图 3~图 6 所示。

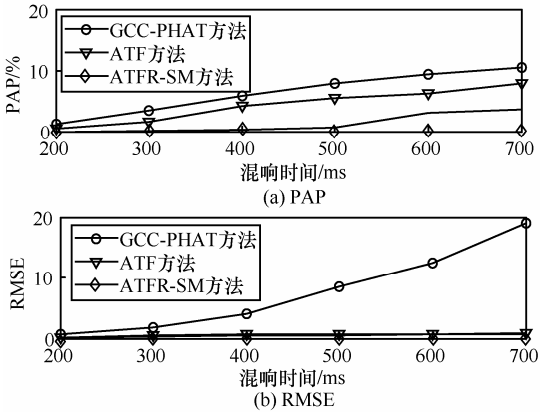

图 3 不同混响环境下各种时延估计算法的 PAP 和 RMSE

图 3 为在混响无噪环境下的测试结果，随着混响时间增加，提出方法和参考方法的异常点百分比均逐渐增加，其中，ATFR-SM 方法要优于参考方法 GCC 和 ATF。GCC 方法的均方根误差随着混响时间的增加而增加，但 ATF 和 ATFR-SM 方法的均方根误差则无明显变化。这是因为 ATF 和 ATFR-SM 方法均采用了传递函数比方法，在传递函数的比值中包含了所有的时延和混响信息，可以通过估计声源到 2 个麦克风的传递函数比，将传递函数中表示混响的部分通过比值去除。

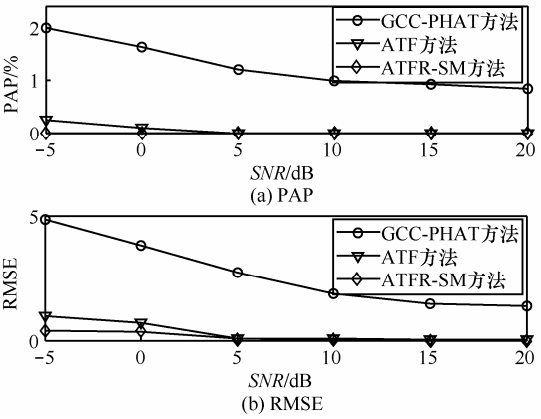

图 4 不同 babble 噪声环境下各种时延估计算法的 PAP 和 RMSE

图 4 为在 babble 噪声无混响环境下的测试结果，随着信噪比逐渐增加，参考方法 GCC 和 ATF 以及提出方法 ATFR-SM 的异常点百分比和均方根误差均逐渐降低，本文提出的方法引入了统计模型，对环境噪声有更好的顽健性，因此，在所有噪声环境下本文所提方法的性能最好。

图 5 和图 6 为在噪声和混响同时存在环境下的测试结果，从图中更可以看出提出方法依然具有更加优异的性能。尤其在强噪声强混响混的复杂环境下，ATFR-SM 方法通过统计模型和改进的传递函数比方法进一步提升了信号间的相关性，在复杂环境下仍然具有很高的时延估计精度。

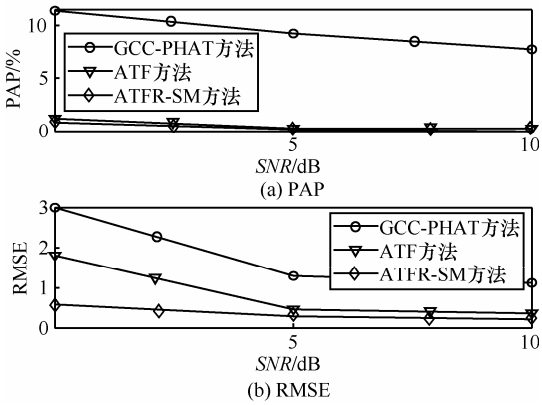

图 5 混响时间为 200 ms，不同 babble 噪声环境下，各种时延估计算法的 PAP 和 RMSE

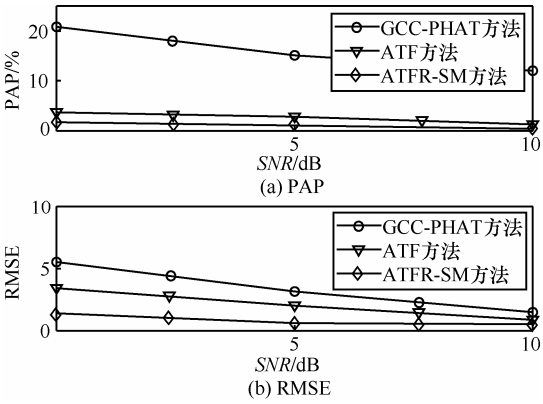

图 6 混响时间为 400 ms，不同 babble 噪声环境下，各种时延估计算法的 PAP 和 RMSE

4 线性定位法

第 i 路麦克风和第 1 路参考麦克风（位于坐标原点）以及声源的位置关系如图 7 所示。

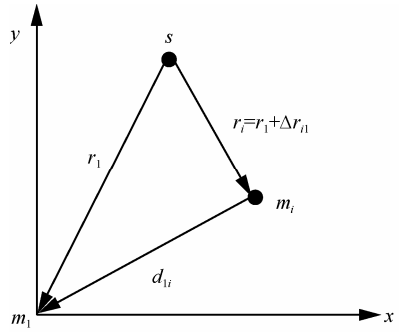

图 7 球形定位法麦克风与声源位置关系

假设麦克风阵列由 N+1 个麦克风组成，非参考麦克风坐标 $m_i=[x_i,y_i,z_i]^{\mathrm{T}}$，参考麦克风坐标 $m_1=[x_1,$

$y_1, z_1]^{\mathrm{T}}$，声源坐标为 $s=[x_s, y_s, z_s]^{\mathrm{T}}$，非参考麦克风和声源分别到原点的距离为 d_{1i} 和 r_1 [17]。

麦克风 m_i 和 m_1 分别到声源 $\boldsymbol{s}$ 的距离差为

$$\Delta r_{i1}=r_i-r_1=\|m_i-s\|-\|m_1-s\| \tag{33}$$

由上式可得

$$r_i^2-r_1^2=\|m_i-s\|^2-\|m_i-s\|^2 \tag{34}$$

此外，声源到参考麦克风和第 i 路麦克风的传输距离的平方差还满足

$$r_i^2-r_1^2=(r_1+\Delta r_{1i})^2-r_1^2=\Delta r_{1i}{}^2-2r_1\Delta r_{1i} \tag{35}$$

由式(33)~式(35)联合并化简可得

$$\begin{aligned}\Delta r_{1i}{}^2-2r_1\Delta r_{1i}=&x_i^2-x_1^2-2x_s(x_i-x_1)+y_i^2-y_1^2-\\&2y_s(y_i-y_1)+z_i^2-z_1^2-z_s(z_i-z_1)\end{aligned} \tag{36}$$

将式(36)中所有已知量结合在一起放在等式右边，未知量放在等式左边，并且令

$$w_{1i}=\frac{1}{2}(\Delta r_{1i}{}^2-x_i^2+x_1{}^2-y_i^2+y_1{}^2-z_i^2+z_1{}^2) \tag{37}$$

于是有

$$r_1\Delta r_{i1}-x_s(x_i-x_1)-y_s(y_i-y_1)-z_s(z_i-z_1)=w_{1i} \tag{38}$$

式(37)中未知量为 x_s、y_s、z_s 和 r_1，且上式中变量均是线性的，4 个未知量需要 4 个方程，写成线性方程组的形式

$$\begin{aligned}\boldsymbol{A}\cdot\boldsymbol{\theta}&=\begin{bmatrix}x_1-x_2 & y_1-y_2 & z_1-z_2 & \Delta r_{21}\\ x_1-x_3 & y_1-y_3 & z_1-z_3 & \Delta r_{31}\\ \vdots & \vdots & \vdots & \vdots\\ x_1-x_N & \mathrm{y}_1-\mathrm{y}_N & \mathrm{z}_1-\mathrm{z}_N & \Delta r_{N1}\end{bmatrix}\cdot\begin{bmatrix}x_s\\ y_s\\ z_s\\ r_1\end{bmatrix}\\&=\begin{bmatrix}w_{12}\\ w_{13}\\ \vdots\\ w_{1N}\end{bmatrix}=\boldsymbol{b}\end{aligned} \tag{39}$$

如果式(39)中的矩阵为非奇异矩阵，则声源坐标以及声源到参考麦克风的距离可采用最小二乘法求解[17]。

5　声源定位性能评价

仿真环境与时延估计相似，采用平面阵列，每行麦克风数目为 N=8，相邻麦克风的间距 d=0.06 m，共有 4 行，行间距为 0.06 m，目标声源位于 S[4.05 m 1.21 m 1.5 m]处。选取 8 段纯净目标话音作为测试话音，话音总长度为 64 s，话音抽样率为 16 kHz。环境噪声为 babble 噪声。提出方法为 ATFR-SM+线性定位法，参考方法 1 为 GCC+线性定位法，参考方法 2 为 ATF+线性定位法。

表 2　混响环境下的近场声源定位方法实验结果

混响时间	方法名称	估计坐标与真实坐标的绝对误差			估计坐标与真实坐标的标准差		
		x	y	z	σ_x	σ_y	σ_z
200 ms	参考方法 1	0.043	0.035	0.031	0.231	0.216	0.172
	参考方法 2	0.005	0.002	0.004	0.046	0.043	0.033
	提出方法	**0.002**	**0.001**	**0.002**	**0.030**	**0.030**	**0.020**
400 ms	参考方法 1	0.065	0.052	0.047	0.472	0.356	0.241
	参考方法 2	0.005	0.004	0.003	0.049	0.045	0.035
	提出方法	**0.002**	**0.002**	**0.002**	**0.032**	**0.030**	**0.020**

从表 2 中可以看出，在混响环境下，不管是 200 ms 还是 400 ms 的混响时间，3 种算法均体现出良好的定位精度，这主要是因为本文提出的 ATFR-SM 时延估计算法在混响环境下良好的顽健性，使时延估计的误差较小，从而使定位算法表现出较好的性能。

表 3　babble 噪声环境下的近场声源定位方法实验结果

SNR	方法名称	估计坐标与真实坐标的绝对误差			估计坐标与真实坐标的标准差		
		x	y	z	σ_x	σ_y	σ_z
10 dB	参考方法 1	0.051	0.032	0.046	0.325	0.531	0.743
	参考方法 2	0.005	0.003	0.004	0.045	0.065	0.078
	提出方法	**0.001**	**0.001**	**0.001**	**0.019**	**0.012**	**0.023**
0 dB	参考方法 1	0.531	0.424	0.421	1.235	1.873	1.432
	参考方法 2	0.042	0.033	0.031	0.426	0.923	0.565
	提出方法	**0.021**	**0.012**	**0.011**	**0.234**	**0.476**	**0.254**

从表 3 中可以看出，在噪声环境下，当信噪比较大时，如 10 dB，时延估计的误差较小，此时 3 种方法的精度都较高。但是当信噪比降低时，比如 0 dB 的噪声环境，参考方法的定位精度开始下降，标准差开始增加。由于本文提出的 ATFR-SM 时延估计方法中加入了基于统计模型的增强方法，大大减少了噪声对算法性能的影响，因此在声源定位算法噪声环境测试中也表现出了优异的性能。

从表 4 中可以看出，在噪声和混响同时存时，环境更加复杂。随着信噪比降低，混响时间变长，3 种方法的定位精度也随之降低。但由于时延估计方法 ATFR-SM 算法的结果在复杂环境下具有很强的顽健性，因此对时延估计误差的补偿性能更好，

定位精度也比参考方法更高。

表 4　混响和 babble 噪声同时存在环境下的实验结果

测试环境	方法名称	估计坐标与真实坐标的绝对误差			估计坐标与真实坐标的标准差		
		x	y	z	σ_x	σ_y	σ_z
SNR=10 dB *RT*=200 ms	参考方法 1	0.123	0.056	0.065	0.756	0.519	0.352
	参考方法 2	0.022	0.003	0.005	0.054	0.043	0.025
	提出方法	**0.005**	**0.001**	**0.002**	**0.023**	**0.020**	**0.014**
SNR=0 dB *RT*=200 ms	参考方法 1	0.212	0.124	0.235	0.785	2.318	0.893
	参考方法 2	0.045	0.032	0.040	0.513	1.021	0.432
	提出方法	**0.013**	**0.016**	**0.023**	**0.232**	**0.544**	**0.246**
SNR=10 dB *RT*=400 ms	参考方法 1	0.122	0.121	0.054	0.312	0.213	0.175
	参考方法 2	0.021	0.020	0.008	0.057	0.051	0.046
	提出方法	**0.002**	**0.004**	**0.003**	**0.021**	**0.030**	**0.011**
SNR=0 dB *RT*=400 ms	参考方法 1	0.322	0.265	0.098	0.878	1.765	1.543
	参考方法 2	0.042	0.045	0.032	0.633	1.189	1.026
	提出方法	**0.021**	**0.012**	**0.013**	**0.221**	**0.544**	**0.863**

6　结束语

本文提出了一种在强噪声强混响环境下的时延估计新方法，并将其应用于近场声源定位方法中。在时延估计方法 ATFR-SM 中，统计模型方法被用于去除噪声对传递函数的影响，平滑和“白化”被用于减少混响对传递函数的影响。采用对数似然比和谱熵相结合的方法构成 VAD 检测算法，以去除对时延估计无用的噪声段。此外，本文将所提时延估计新方法与线性定位法相结合构成一套完整的声源定位方法。实验结果显示，在混响和噪声环境下，无论是时延估计还是声源定位，本文所提方法的性能均优于参考方法。

参考文献：

[1] SOUDEN M, BENESTY J, AFFES S. Broadband source localization from an eigenanalysis perspective[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(6): 1575-1587.

[2] GEDALYAHU K, ELDAR Y C. Time-delay estimation from low-rate samples: a union of subspaces approach[J]. IEEE Transactions on Signal Processing, 2010, 58(6): 3017-3031.

[3] SO H C, CHAN Y T, CHAN F K W. Closed-form formulae for time-difference-of-arrival estimation[J]. IEEE Transactions on Signal Processing, 2008, 56(6): 2614-2620.

[4] LUI K, CHAN F, SO H C. Semidefinite programming approach for range-difference based source localization[J]. IEEE Transactions on Signal Processing, 2009, 57(4): 1630-1633.

[5] KNAPP C, CARTER G. The generalized correlation method for estimation of time delay[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1976, 24(4): 320-327.

[6] CHAMPAGNE B, BÉDARD S, STÉPHENNE A. Performance of time-delay estimation in the presence of room reverberation[J]. IEEE Transactions on Speech and Audio Processing, 1996, 4(2): 148-152.

[7] DVORKIND T G, GANNOT S. Approaches for time difference of arrival estimation in a noisy and reverberant environment[A]. Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC'03)[C]. Kyoto, Japan, 2003. 215-218.

[8] CORNELIS B, DOCLO S, VAN DAN BOGAERT T, *et al*. Theoretical analysis of binaural multimicrophone noise reduction techniques[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(2): 342-355.

[9] SALVATI D, CANAZZA S. Adaptive time delay estimation using filter length constraints for source localization in reverberant acoustic environments[J]. Signal Processing Letters, 2013, 20(5):507-510.

[10] EPHRAIM Y, MALAH D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1984, 32(6): 1109-1121.

[11] COHEN I, BERDUGO B. Noise estimation by minima controlled recursive averaging for robust speech enhancement[J]. Signal Processing Letters, IEEE, 2002, 9(1): 12-15.

[12] LOIZOU P C. Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum[J]. IEEE Transactions on Speech and Audio Processing, 2005, 13(5): 857-869.

[13] SOHN J, KIM N S, SUNG W. A statistical model-based voice activity detection[J]. Signal Processing Letters, IEEE, 1999, 6(1): 1-3.

[14] 李光源, 崔慧娟, 唐昆. 一种基于噪声估计的话音激活检测算法[J]. 信息技术, 2011 (10): 5-8.
LI G Y, CUI H J, TANG K. An algorithm of voice activity detection based on noise estimation[J]. Information Technology, 2011, (10):5-8.

[15] ALLEN J B, BERKLEY D A. Image method for efficiently simulating small-room acoustics[J]. The Journal of the Acoustical Society of America, 1979, 65: 943.

[16] VARGA A, STEENEKEN H J M. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems[J]. Speech Communication, 1993, 12(3): 247-251.

[17] GILLETTE M D, SILVERMAN H F. A linear closed-form algorithm for source localization from time-differences of arrival[J]. Signal Processing Letters, IEEE, 2008, 15: 1-4.

作者简介：

张大威（1987-），男，北京人，北京工业大学硕士生，主要研究方向为麦克风阵列声源定位和话音增强。

鲍长春[通信作者]（1965-），男，内蒙古赤峰人，博士，北京工业大学教授、博士生导师，主要研究方向为话音与音频信号处理。E-mail:chchbao@bjut.edu.cn。

夏丙寅（1986-），男，北京人，北京工业大学博士生，主要研究方向为话音增强。