ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 2,Mar. 2015,pp. 328－335
DOI:10.16337/j.1004-9037.2015.02.010

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

人机交互中的声源定位与增强方法

殷瑞祥　李文东

(华南理工大学电子与信息学院,广州,510640)

摘　要:基于人机交互的实际应用场景,提出了一种交互目标声源三维定位与语音增强算法。该算法首先在广义相关法的基础上提出一种声达时延差的估计方法,通过由6个麦克风构成的平行均匀线阵接收模型,实现对目标声源的三维定位;然后在交互目标声源定位的基础上,通过调整时延波束叠加的权值,实现对目标语音的增强。仿真实验结果表明,文中提出的算法能够准确定位目标声源并对目标语音进行有效增强。在信噪比大于1.5 dB的环境中,该算法可使目标声源达到98%以上定位精度,5 dB以上信噪比改善,同时算法运算量小且易于硬件实现。

关键词:声源定位;语音增强;人机交互;麦克风阵列;近场语音模型

中图分类号:TN912.3　　**文献标志码**:A

Acoustic Source Localization and Speech Enhancement in Human-Computer Interaction

Yin Ruixiang, Li Wendong

(School of Electroic and Information Engineering, South China University of Technology, Guangzhou, 510640, China)

Abstract: Focusing on the human-computer interaction(HCI) real scenarios, a method for acoustic source three-dimensional localization and speech enhancement is proposed. Combining with the receiver model of six-microphone parallel uniform linear array (ULA), the target acoustic source is located in three dimensions based on the method of time-difference of arrival (TDOA) estimation improved from generalized cross correlation (GCC). On the basis of positioning the target acoustic source, the target speech enhancement is implemented by the method of delay-and-sum beamforming (DSBF). Simulation results show that the method can position the target acoustic source accurately and enhance the target speech effectively. In the context of SNR greater than 1.5 dB, the positioning accuracy of the target acoustic source can reach more than 98% and improvement of SNR can reach 5 dB with less computation cost and easier hardware implementation.

Key words: acoustic source localization; speech enhancement; human-computer interaction; microphone array; near-field voice model

引　　言

随着互联网、云计算、移动智能终端等的快速发展,人们更加注重用户体验,对人机交互技术提出了更高的要求。以鼠标和键盘为代表的交互技术已无法满足人们的体验需求,并严重阻碍了计算机系统

收稿日期:2014-03-25;**修订日期**:2014-06-13

拟人化、微型化以及随身化的发展[1];以人为中心,自然、高效是发展新一代人机交互方式的主要目标。近年来,基于语音、手写、姿势、视线以及表情等新型的交互技术成为研究热点。其中,语音作为人类使用最多、最自然的交互方式,既不需要接触或佩戴数据设备,也不存在视觉盲点,是人机交互的最理想交互方式之一[2-3]。

在实际的人机语音交互环境中,语音信号总会受到外界环境噪声乃至其他说话人的干扰,这些噪声的污染导致麦克风接收到信号的信噪比严重下降,不仅恶化语音通信的质量,也降低了交互的效率。为提升用户体验,提高交互效率,人们相继提出了维纳滤波法[4-5]、谱减法[6-7]、最小均方误差方法[8]和子空间方法[9]等语音增强方法。这些方法在理想条件下可以获得良好的噪声抑制性能,然而在实际交互环境中,特别是存在其他说话人语音干扰时,利用单麦克风捕捉相对纯净的语音非常困难。随后人们将麦克风阵列引入到语音信号处理中,提出了首先定位目标声源,然后麦克风阵列在期望方向上有效地形成一个波束去拾取波束内的目标信号,并消除波束外的噪声,从而达到同时提取声源和抑制噪声的目的。基于麦克风阵列的声源定位和语音增强技术已经成为国内外的研究热点。在声源定位方面,文献[10-13]对多信号分类算法(Multiple signal classification,MUSIC)进行改进,结合不同麦克风摆放阵型,实现声源的二维或三维的定位。然而此类方法要事先估计声源数,并对麦克风数量有一定的要求,而且算法计算量较大,实现较为困难[14]。

在人机交互的实际场景中,交互的目标声源通常只有一个,而且目标语音在嘈杂的环境中以主导的成分存在,本文基于这种人机交互的实际场景,仅仅利用声源的传播特性,提出了一种目标声源的三维定位方法,由 6 个麦克风组成的平行均匀线阵接收模型,利用声达时延差(Time-difference of arrival,TDOA)的估计进行目标声源的三维定位,在定位基础上,通过延时波束加权叠加实现交互语音的增强。

1 近场二维声源定位算法

麦克风阵列处理的对象主要为近场宽带语音信号,即声源与阵列距离 r 满足

$$r < \frac{2L^2}{\lambda} \tag{1}$$

式中:L 为阵列尺度;λ 为声波波长。本文根据语音的声学特性,采用了一种基于球面波假设的近场宽带信号模型[15]。

图 1 为三麦克风均匀直线阵列,相邻麦克风之间的距离为 Δd,阵列中心的麦克风为参考点,目标声源距参考点的距离为 r_0。

根据图 1 的几何关系可知目标声源与三麦克风的距离关系满足

$$\begin{cases} r_1 - r_0 = \tau_{10} v, r_1 = \sqrt{(r_0 \sin\theta)^2 + (-\Delta d + r_0 \cos\theta)^2} \\ r_2 - r_0 = \tau_{20} v, r_2 = \sqrt{(r_0 \sin\theta)^2 + (\Delta d + r_0 \cos\theta)^2} \end{cases} \tag{2}$$

式中:τ_{10},τ_{20} 分别为目标声源到达阵列两侧的麦克风与参考麦克风的时延差,v 为声速。由式(2)可得目标声源相对于阵列参考点的极坐标 (r_0,θ)

$$\begin{cases} r_0 = \dfrac{2\Delta d^2 - (\tau_{10}^2 + \tau_{20}^2) v^2}{2(\tau_{10} + \tau_{20}) v} \\ \theta = \arccos\left(\dfrac{(\tau_{20} - \tau_{10})(\tau_{10}\tau_{20} v^2 + \Delta d^2) v}{\Delta d (2\Delta d^2 - \tau_{10}^2 v^2 - \tau_{20}^2 v^2)} \right) \end{cases} \tag{3}$$

由式(3)可知,只要获知 τ_{10},τ_{20} 的值,则可确定目标声源的唯一极坐标。由此可见,对时间延迟进行准确地估计是实施麦克风阵列声源定位需要解决的首要问题,时延估计的精度越高,定位误差就越小。目前基于时延估计的算法很多,其中最为经典、运用最为广泛的为广义互相关函数法[16-18](Generalized cross correlation,GCC)。在来自同一信源的信号存在一定的相关性而信号与噪声之间不相关的基础之

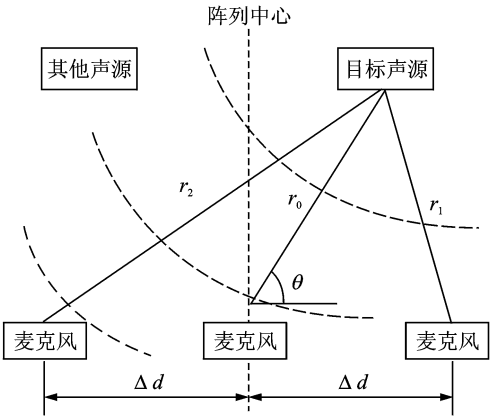

图 1 三麦克风均匀线阵近场信号模型

Fig. 1 AULA near-field signal receiver model based on three microphones

上，GCC 算法通过对接收信号进行频域加权，突出相关的信号而抑制噪声和干扰，再经离散傅里叶逆变换到时域，根据两信号互相关函数的峰值位置估计两信号之间的相对时延。然而实际应用中，噪声和信号的先验知识并不能预先得到的，特别在存在其他说话者干扰的人机语音交互环境中，GCC 算法并不能取得理想的效果。

为了减弱噪声对时延估计的影响，采用先对接收信号进行阈值判决再做相关的 TDOA 估计。实际的人机语音交互主要以单目标声源交互为主，目标语音虽然在传播中易受环境噪声与其他说话者的干扰，但目标语音波形中幅度较大的信号在进行噪声叠加时，相对变化较小，保持原有的时域特征，则可以通过设定合适的阈值对接收的信号进行筛选，忽略小于阈值的信号，以大于阈值的信号为基准通过计算互相关函数估计两通道信号的相对时延差。由图 1 的均匀线阵近场信号模型可知，第 i 麦克风接收到的信号

$$x_i(t)=\partial\frac{1}{r_i}s(t-\Delta t_i)+n_i(t)\qquad i=1,2,3 \tag{4}$$

式中：∂ 为声音在大气中传播的幅度衰减因子；r_i 为第 i 麦克风与目标声源之间的距离；$s(t)$ 为目标声源信号；$n_i(t)$ 为包括其他说话人等干扰源在内的各种噪声的总和。

$$x'_i(n)=\begin{cases}x_i(n) & x_i(n)\geqslant A_0\\ 0 & \text{其他}\end{cases} \tag{5}$$

首先，为了确保时延差的估计精度，对麦克风接收到的信号进行插值处理得到 $x_i(n)$。然后，设定一个较大的阈值 A_0，根据式(5)对插值后的 $x_i(n)$ 进行阈值判决处理。由于目标语音在接收的信号中以主导的成分存在，所以在选取合适的阈值进行判决后，留下的信号大多由目标语音决定，而只受少量的干扰影响。即由式(5)中 $x'_i(n)$ 之间的时延差可估计式(4)的 $x_i(t)$ 之间的时延差。

阈值的选取极其关键，过小的阈值不但会增加计算量，而且在阈值判决后会残留大量的噪声干扰，进而影响 TDOA 的估计，导致定位误差的增大；而过大的阈值会导致大量的目标语音信息在阈值判决中丢失，而且易受突发强噪声的干扰，从而导致错误地估计 TDOA。本文通过二次均值操作完成阈值 A_0 的选取，首先获取观察时间段内信号绝对值均值，然后选取观察时间段内大于信号绝对值均值的信号，求其均值作为阈值 A_0 的值。

$$R_{ij}(\tau)=E[x_i(n)x_j(n+\tau)] \tag{6}$$

最后,由式(6)对来自不同麦克风经过阈值判决后的信号进行相关处理,通过互相关函数的峰值检测完成 TDOA 的估计,再由式(3)实现目标声源的二维定位。

2 声源三维定位算法

在实际情况下,语音信号位于三维空间。当干扰源与目标声源位于以麦克风直线阵为中心的同一圆周上时,干扰声源与目标声源相对均匀线阵参考点具有相同的极坐标,麦克风均匀线阵模型就无法区分目标声源与干扰源,进而严重影响阵列有方向性地增强目标声源的信号。因此,需要对图 1 麦克风均匀线阵模型加以改进。

基于图 1 的均匀线阵,本文提出一种平行均匀线阵三维信号接收模型如图 2 所示。阵列模型由两个平行的三麦克风均匀线阵组成,两子阵列间距为 d,且都以中间的麦克风为参考点。把阵列 1 的参考点作为原点,建立三维坐标系,假设目标声源的球坐标为 (r,θ,φ)。根据三麦克风均匀线阵近场信号模型,可以完成目标声源坐标中 (r,θ) 的确定所以只需确定方位角 φ 即可完成目标声源的三维定位。

首先,分别在两个子阵列中通过上述的近场二维声源定位算法,得到目标声源相对两个子阵列参考点的极坐标 (r,θ) 与 (r',θ'),根据阵列的几何关系,(r,θ) 与 (r',θ') 的关系可表示为

$$\begin{cases} r' = \sqrt{m^2 + n^2 + (r\cos\theta)^2} \\ \theta' = \arccos\left(\dfrac{r\cos\theta}{r'}\right) \end{cases} \tag{7}$$

式中:$m = r\sin\theta\cos\varphi$,$n = d - r\sin\theta\sin\varphi$。

由式(7)可知,给定一个方位角 φ',就能由 (r,θ) 唯一地确定一组 (r'',θ'')。引入误差函数 $\mathrm{Err}(\varphi)$,在[$-90°$ $90°$]的范围内遍历方位角 φ 值,取使得误差函数 $\mathrm{Err}(\varphi)$ 最小 φ 为目标声源方位角的最佳估值。

$$\mathrm{Err}(\varphi) = \frac{1}{2}\left(\left(\frac{r'' - r'}{r'}\right)^2 + \left(\frac{\theta'' - \theta'}{\theta'}\right)^2\right) \tag{8}$$

3 交互语音增强算法

在获取交互目标声源位置信息的条件下,通过加权延时叠加波束形成法[14,19](Delay and sum beamforming, DSBF)实现交互目标语音的增强。首先利用交互目标声源的位置信息,计算交互目标语音到达麦克风阵列各阵元的时延;然后对各个阵元接收到的信号进行时延补偿,使各阵元接收的信号关于交互目标语音达到同步,而此时各通道接收的噪声和来自其他声源的语音并不同步;最后对各通道补偿后的语音信号进行加权累加平均,由于各通道中交互目标语音是同步的,累加平均后被完整保留下来,而不同步的干扰噪声随着累加平均而被减弱,进而达到改善信噪比的效果。加权延时叠加波束形成法的结构如图 3 所示。

现有的延时叠加波束形成法中的加权较为固定,例如固定波束法加权通常取麦克风总数的倒数。由式(4)可知,声音在传播中不但存在时延而且还有一定的幅度衰减,特别是声源位于近场时,各通道接收信号的幅度差异非常明显,各个麦克风接收信号的信噪比也有一定的差异。为了更好地改善信噪比,增强目标交互语音,对高信噪比的通道应给予较大权重,低信噪比通道给予较小权重。在实际的交互环境中,距离交互目标声源较近的麦克风接收到较高信噪比的信号,应给予相对较大的权重,经过大量的仿真实验,提出一种易于实现的经验权重配置方案

$$\omega_i = \begin{cases} 2.0 & r_i < r_0 \\ 1.0 & r_i = r_0 \\ 0.5 & r_i > r_0 \end{cases} \tag{9}$$

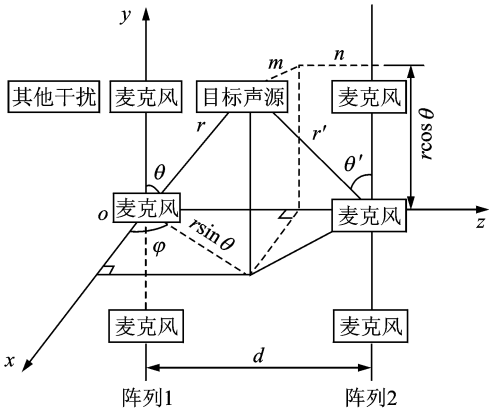

图 2 平行均匀线阵三维信号接收模型

Fig. 2 Receiver model of six-microphone's parallel uniform linear array

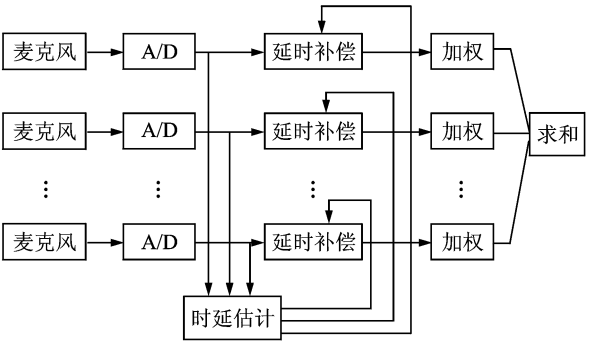

图 3 延时叠加波束形成法结构图

Fig. 3 Structure diagram of DBSF

式中：ω_i 为对第 i 麦克风信号加权的权重；r_i 为第 i 麦克风与目标声源之间的距离；r_0 为信号接收模型原点处麦克风与目标声源的距离。

4 仿真实验

仿真实验采用 Matlab 语言。使用如图 2 所示的平行均匀线阵三维信号接收模型，每个子阵列由 3 个全向麦克风构成，麦克风间距为 15 cm，阵列间距为 30 cm，采用事先录制好的一段语音信号作为目标声源，除了目标声源外，还有 6 段事先录制好的不同于目标语音的信号作为干扰声源分布在目标声源周围，同时空间存在一定强度的白噪声。声音在空气中的传播速度定为 340 m/s，信号的采样频率为 16 kHz。

$$\text{阈值 } A_0 = \text{信号绝对值均值} + \mu * (\text{信号峰值} - \text{信号绝对值均值}) \tag{10}$$

图 4 对比了本文提出的 TDOA 估计算法与广义互相关法在不同程度噪声和干扰下 TDOA 估计的性能。仿真实验中，本文算法采用 10 倍插值，通过二次均值获取阈值，而广义互相关法同样采用 10 倍插值，使用 PATH 函数加权。由图 4 可知，本文算法通过阈值判决减弱了噪声的干扰，在同等噪声干扰下，比广义互相关法取得较精确的时延估计，更加有利于后续的声源定位。

图 5 分析了阈值大小对本文算法的 TDOA 估计性能的影响。在麦克风接收平均信噪比为 3 dB 环境下，阈值按式(10)选取。由图 5 可知，当 $\mu < 0.55$ 时，随着 μ 的逐渐增大，阈值也不断增大，对噪声的抑制也越为明显，TDOA 估计的相对误差呈现缩小趋势；当 $\mu > 0.55$ 时，随着 μ 的逐渐增大，过大的阈值导致大量的目标语音信息在阈值判决中丢失，而导致 TDOA 估计的相对误差的增大。本文采用的二次均值操作获取的阈值大都处于 $\mu < 0.5$ 附近，是较为理想的选择。

在声源定位实验中，对目标声源位于阵列近场中的不同位置进行仿真。图 6 为目标声源位于(0.7 m，45°，45°)坐标处，不同程度干扰下的定位效果，图 6 中的信噪比为各麦克风信噪比的均值，相对误差为 20 次定位结果的均值。由图 6 可知，麦克风接收信号的平均信噪比大于 1.5 dB 时，本文算法即可达到 98%的定位精度。虽然当平均信噪比小于 0 dB 时，算法的定位误差相对较大，但是此时干扰已完全淹没目标声源，此类情况在实际人机交互场景中以不存在有效的交互信号处理。

在后续的语音增强中，采用延时叠加波束形成算法，在各通道中目标语音信号同步后，通过式(9)对各通道信号进行加权。图 7 为目标声源中的一段目标语音增强的效果图，从上至下分别是目标语音信

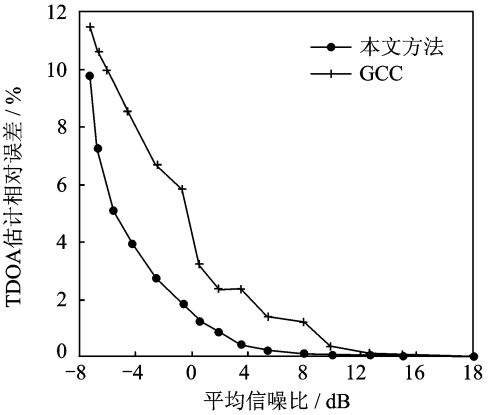

图 4 本文 TDOA 估计方法与 GCC 对比图

Fig. 4 Comparison between TDOA estimation method proposed in this paper and GCC

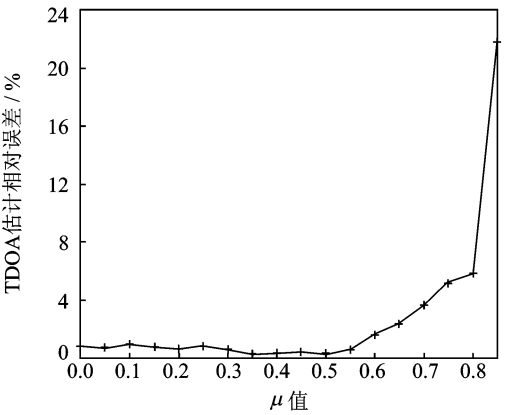

图 5 阈值大小对本文 TDOA 估计算法的影响

Fig. 5 Impact of the threshold size on the algorithm of TDOA estimation in this paper

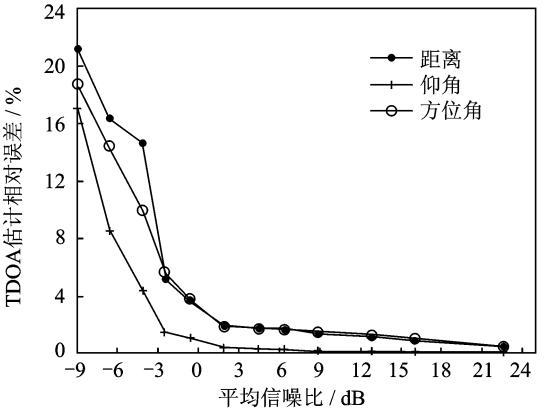

图 6 目标声源定位误差图

Fig. 6 Positioning error map of the target acoustic

号,参考点麦克风接收的包括目标声源、6 个其他干扰语音以及白噪声的混合信号,最后是增强后的信号。由图 7 可知通过延时叠加波束形成算法达到 5 dB 以上的增强效果。

图 8 对比了特殊场景下目标声源二维定位与三维定位对目标语音增强的影响。仿真中目标声源与其他 6 个干扰声源都分布在以图 2 的阵列 1 为中心的圆周上,即各个声源在球坐标系中具有相同的距离 r 和仰角 θ,只有方位角 φ 不同。图 8(a)为参考点麦克风接收到的混合信号;图 8(b)为仅采用由 6 个麦克风构成的单均匀阵列进行二维声源定位后增强的效果,由图 8(a,b)对比可知,此场景中二维定位无法区分出目标声源与干扰声源,延时波束加权叠加无法消除干扰噪声,达到增强目标语音的目的;图 8(c)为通过图 2 模型声源三维定位后增强的效果。由图 8 的(b,c)对比可知,通过本文提出的平行均匀线阵对目标声源的三维定位弥补了单线阵在此场景中无法区分出目标声源与干扰声源的缺陷,达到有效的目标语音增强。

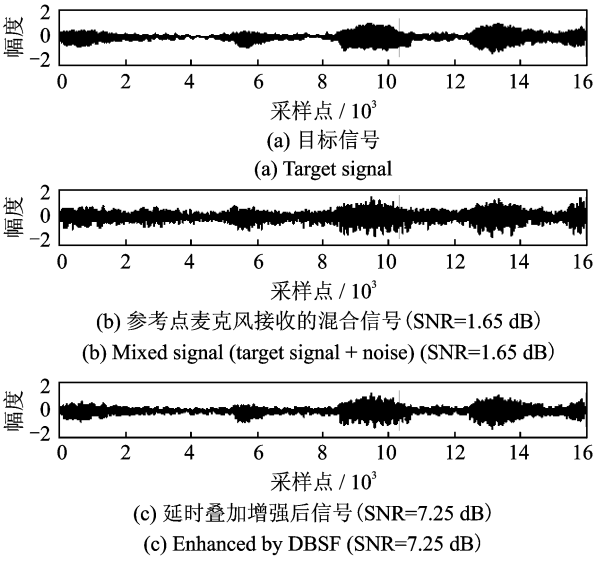

图 7 目标声源一段语音的增强效果

Fig. 7 Enhancement of a piece of signal from the target acoustic source

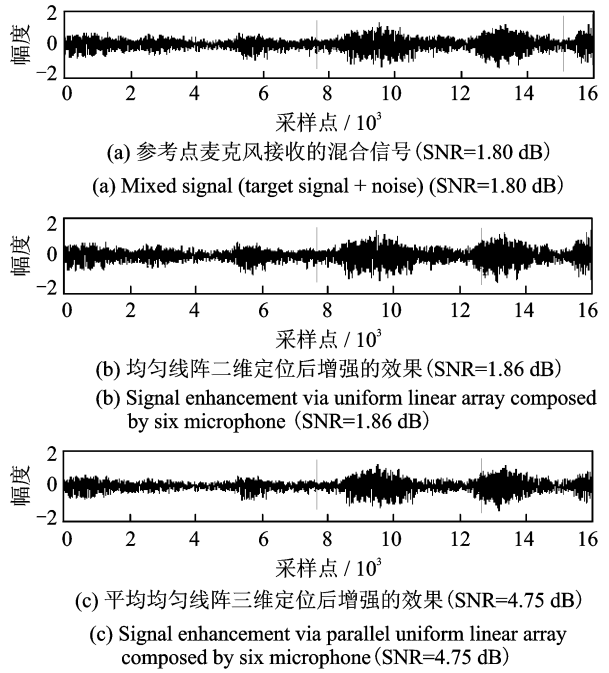

图 8 不同接收模型对语音增强的影响

Fig. 8 Impact on speech enhancement of the different receiver model

5 结束语

在基于语音的人机交互场景中,为了降低噪声干扰对交互性能和效率的影响,本文提出了一种交互目标声源三维定位与语音增强算法。在声源定位方面,基于广义相关法提出了一种简单有效的声达时

延差估计方法,通过平行均匀线阵接收模型,实现目标声源的三维定位;在语音增强方面,基于交互目标声源的定位信息,通过调整时延波束叠加的权值,实现对目标语音的增强。实验结果表明本中提出的算法能够准确定位目标声源并对目标语音进行有效增强,在信噪比大于 1.5 dB 的环境中,该算法可使目标声源达到 98%以上定位精度,5 dB 以上信噪比改善,同时算法结构简单、运算量较小,硬件实现较为容易。

参考文献:

[1] Liu B. Research and implementation of the speech recognition technology based on DSP[C]// IEEE International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC). [S. l.]:IEEE, 2011: 4188-4191.

[2] Cho H, Choi J, Ko H. Robust sound source localization using a Wiener filter[C]// IEEE Conference on Emerging Technologies & Factory Automation (ETFA). Cagliari, Italy:IEEE, 2013: 1-6.

[3] Zhao Z, Ma X. Active learning for speech emotion recognition using conditional random fields[C]// ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). Honolulu, HI:IEEE, 2013: 127-131.

[4] Lim J S, Oppenheim A V. Enhancement and bandwidth compression of noisy speech[J]. Proceedings of the IEEE, 1979, 67(12): 1586-1604.

[5] Scalart P, Filho J V. Speech enhancement based on a priori signal to noise estimation[C]// IEEE International Conference on Acoustics, Speech, and Signal Processing. Atlanta, GA:IEEE, 1996: 629-632.

[6] Boutaleb R, Meraoubi H, Ykhlef F, et al. Comparative performance study between spectral subtraction and discreet wavelet transform for speech enhancement[C]// ACS International Conference on Computer Systems and Applications. Ifrane, Morocco:IEEE, 2013:1-1.

[7] Wan Y L, Zhang T Q, Wang Z C, et al. Robust speech recognition based on multi-band spectral subtraction[C]// IEEE International Congress on Image and Signal Processing. Hangzhou, China:IEEE, 2013:36-40.

[8] 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.

[9] Ephraim Y, Van Trees H. A signal subspace approach for speech enhancement[J]. IEEE Transactions on Speech and Audio Processing and Audio Processing,1995, 3(4): 251-266.

[10] Sugimoto Y, Miyabe S, Yamada T, et al. Employing moments of multiple high orders for high-resolution underdetermined DOA estimation based on MUSIC[C]// IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz, NY:IEEE, 2013:1-4.

[11] Liang J, Liu D. Passive localization of mixed near-field and far-field sources using two-stage MUSIC algorithm[J]. IEEE Transactions on Signal Processing, 2010, 58(1): 108-120.

[12] Liu C L, Hang H M. Direction of arrival estimation of speech signals using ICA and MUSIC methods[C]// IEEE Conference on Industrial Electronics and Applications (ICIEA). Taichung, China:IEEE, 2010: 1768-1773.

[13] Nakamura K, Nakadai K, Ince G. Real-time super-resolution sound source localization for robots[C]// IEEE International Conference on Intelligent Robots and Systems (IROS). Vilamoura, Portugal:IEEE, 2012: 694-699.

[14] Benesty J, Chen J, Huang Y. Microphone array signal processing[M]. Berlin: Springer,2008:181-217.

[15] Zhi W, Chia M Y W. Near-field source localization via symmetric subarrays[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Honolulu, HI:IEEE,2007: 1121-1124.

[16] Cetnarowicz D, Dabrowski A. Enhancement of time-delay of arrival estimation[C]// Signal Processing: Algorithms, Architectures, Arrangements, and Applications(SPA). Poznan, Poland:IEEE,2013: 323-326

[17] 周康辉,董万胜,刘恒毅,等.利用二次相关改进的广义互相关时延估计算法[J].数据采集与处理,2013,28(6):801-806.
Zhou Kanghui, Dong Wansheng, Liu Hengyi, et al. Improved generalized cross correlation method for time delay estimation by using second correlation[J]. Journal of Data Acquisition and Processing, 2013, 28(6):801-806.

[18] 王瑞,陈砚圃,张恒.声源定位中波达角一致性检测方法[J].数据采集与处理, 2010, 25(6):756-760.
Wang Rui, Chen Yanpu, Zhang Heng. Method for detecting DOA homogeneity in sound source localization[J]. Journal of Data Acquisition and Processing, 2010, 25(6):756-760.

[19] Król D, Wielgat R. Enhancement of loudspeaker impulse response measurement using beamforming methods[C]// International Conference on Signals and Electronic Systems. Wroclaw, Poland:IEEE, 2012, 12: 18-21.

作者简介:**殷瑞祥**(1960-),男,教授,研究方向:信号处理的理论与技术、模式识别、智能信息处理和大规模集成电路设计,E-mail:etrxyin@scut.edu.cn;**李文东**(1989-),男,硕士研究生,研究方向:智能信息处理,E-mail:564163890@qq.com。