智能麦克风阵列语音分离和说话人跟踪技术研究

杜 江[1], 朱 柯[2]
(11 电子科技大学通信学院, 四川成都 610054; 21 韩国三星电子数字媒体技术研究所)

摘 要: 本文介绍一种新的基于麦克风阵列的语音分离和说话人跟踪技术. 该技术使用麦克风阵列, 形成一个指向感兴趣说话人的波束来增强信号, 并通过方向置零来抑制其他说话人的声音和噪声, 同时用自适应算法跟踪说话人的方位变化. 仿真验证了该技术的有效性. 与常规的自适应算法相比, 该算法不需训练序列, 具有显著的优势.

关键词: 麦克风阵列; 语音分离; 说话人跟踪; 波束形成
中图分类号: TN912 文献标识码: A 文章编号: 0372-2112 (2005) 02-0382-03

Smart Microphone Arrays for Speech Sources Separation and Speaker Tracking

DU Jiang[1], ZHU Ke[2]
(1. Institute of Communication and Information Engineering, UESTC, Chengdu, Sichuan 610054, China;
2. Connectivity Lab, R&D Center, Digital Media, Samsung Electroncs Co. Ltd. Korea)

Abstract: A new speech sources separation and speaker tracking technique is introduced based on microphone arrays. By means of spatial property of the received speech signals from microphone arrays, this method utilizes beamforming to estimate the DOA of the speaker of interest, and attenuates unwanted voices by nulling other directions. Considering the speech environments where the speaker may freely move and the background voices exist, an adaptive algorithm is used to track the movements and the source direc2tion variations automatically. Computer simulations validate the effectiveness of the technique. Compared with the conventional meth2ods, the scheme needs no training sequence, and have great potential practical advantages.

Key words: microphone arrays; speech separation; speaker tracking; beamforming

1 引言

选择性增强感兴趣的语音信号并同时压制噪声和干扰有相当重要的实用价值[1]. 涉及到的关键技术之一是信源分离, 即在多个混合声音信号中, 选择并放大某个或几个声源. 但是, 在嘈杂的背景下, 由于各种声音混迭和说话人位置改变, 用常规的时频域处理技术几乎不可能有效的跟踪和分离出感兴趣的声音. 基于上述事实, 本文利用麦克风阵列对接收的信号进行空时处理, 介绍的算法在空域为线性的, 在时域为非线性, 利用空间信号的位置和每个声源的独立不相关统计特性提取和跟踪感兴趣的说话人的声音[2,3]. 本技术主要包括以下几个方面: (1) 使用改进的 MUSIC 算法, 实现麦克风阵列接收范围的信源数目和方位; (2) 介绍了一种基于神经网络的自适应盲源分离算法, 它是对文献[2, 4] 的信源分离方法的改进. 该算法用神经网络作为信号分离的约束条件, 构成一个最优化盲算法准则. 它能自适应语音通信环境, 使分离的信号保持统计独立条件下的最优; (3) 对空间分离后的每路语音在时频域采用谱抵消技术进一步抑制噪声[5].

2 麦克风阵列的近场声音传播模型

当信源离阵列很近时, 麦克风阵列处理必须采用更精确的球面波前模型, 要考虑声波波前在传播过程中发生的幅度衰减, 其衰减因子与传播距离呈反比[6]. 对于一个长度为 L 的阵列, 如果信源与阵列的距离 $r < 2L^2/K$, K 为声波的波长, 则该信源位于近场之内.

如图 1 定义一个参考麦克风作为三维向量空间的原点. 设位于 $(r_s, H_s, <_s)$ 的信源 S 的空间位置向量为 P_s. 其中, r_s 为信源与参考麦克风的距离, H_s 和 $<_s$ 分别表示信源的方位角和仰角. 在该坐标系中, 同样定义麦克风的空间位置向量 $p_i (i = 1, 2, , , N)$. 从信源 S 到第 i 个麦克风的欧氏距离为: $d_i = \| P_s - p_i \|$,

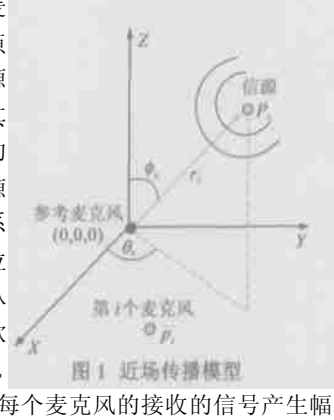

图 1 近场传播模型

这种麦克风阵元的距离差使每个麦克风的接收的信号产生幅度相位差. 在有 N 个麦克风, M 个信源的阵列中, 第 i 个麦克风的接收信号为:

$$x_i(t) = \sum_{k=1}^{N} a(r_k, H_k, <_k)\, s_k(t) + n_i(t) \qquad (1)$$

式中, $s_i(t)$ 为信号, $n_i(t)$ 为加性噪声, $a(r_k, H_k, <_k) = [A_1 e^{-j2\pi f \tau_1}, , , A_M e^{-j2\pi f \tau_M}]^T$ 是阵列响应向量, $A_k = d_1/d_k$ 为幅度衰减因子. 式 (1) 表示为矩阵形式:

收稿日期: 2003-10-16; 修回日期: 2004-10-16

$$x(t)= H(t)s(t)+ n(t) \quad (2)$$

其中, 时变的声音传播效应综合到矩阵. $H(t) = [a(r_1, H_1, <_1), a(r_2, H_2, <_2), , , a(r_M, H_M, <_M)]$.

信源分离是求一个逆矩阵 W, 使 WH= I, 即:

$$Wx(t)= W[H(t)s(t)+ n(t)] U s(t) \quad (3)$$

3　信源数目的确定

进行信源分离和跟踪之前, 需确定在麦克风阵列接收范围内的信源个数和方位. 由于近场情况下多径反射很少, 可以假设各个信源是不相关的, 可以用 MUSIC 算法估计信源个数和方位; 另一方面, 近场声音传播模型与远场的平面波前模型不同, 但是描述方程相似; 因此, 可以利用改进的 MUSIC 算法来估计信源个数和方位. 这里, 更关心的是信源的数目.

利用改进的MUSIC 算法求得阵列的空间谱为:

$$P_{MUSIC}(r, H, <)= 1/ + a(r, H, <) G+ ^2 \quad (4)$$

其中, G 为接收信号相关矩阵的噪声子空间矩阵, $a(r, H, <)$ 为改进的近场阵列响应向量. 在空间谱上, 峰值的个数对应信源的个数, 峰值的方位对应信源的方位.

4　语音信号盲自适应分离与提取

独立信源恢复与分离是信号处理经典但又很困难的问题, 要求仅由接收的信号去估计信道特性, 恢复出原始的信源信号. 目前, 已有的文献主要采用以下两种方法: (1) 常规的离散信号处理, 利用信号的统计特性在离散域进行信号分离[3], 其计算量很大, 无法应用于实时应用系统; (2) 基于神经网络的自适应算法, 由 Heranlt 和 Jutten 于 1991 年介绍, 现称为 HJ 算法[2]. HJ 算法要求传递通道矩阵为满秩, 矩阵的元素为常数, 并且需进行矩阵求逆运算. 然而, 在麦克风阵列应用中, 信道是时变的, 信源的位置变化, 空气变化或是室内家具布设的变化都会导致信道矩阵的元素也随之不断变化. 基于这些事实, 本文介绍了一种新的基于麦克风阵列的语音分离技术, 弥补以上缺陷.

设 $y_i(n)$ 为 M@1 输出向量, $W_p(n)$ 为波束形成器的权系数构成的 M@N 维矩阵, $x_i(n)$ 为接收信息信号. 波束形成器的输入输出关系的操作数形式可表示为:

$$x_i(n)= H(z)[s(n)] \quad (5)$$

$$y_i(n)= \sum_{p=1}^{K} W_p(n)x_i(n- p) = W(z, n)[x_i(n)]= C(z, n)[s(n)] \quad (6)$$

其中, $W(z, n)= \sum_{p=1}^{K} W_p(n)z^{- p}$, $H(z)= \sum_{p=1}^{K} H_p z^{- p}$, $C(z, n)= W(z, n)H(z)$, 分别表示信道、波束形成器以及通道与波束形成器组合的冲击响应的 Z 变换.

在盲算法中, 波束形成器的目标是调整 W(z, n) 使得:

$$\lim_{n y]} C(z, n)= PD(z) \quad (7)$$

其中, P 是一个M@M 的交换矩阵(permutation matrix), 任一行或列仅含有一个为 1 的非零项; D(z) 是对角阵, 第(i, i) 个元素为 $c_i z^{- \$i}$, c_i 是非零复标量权值, $\$i$ 是整数延迟值.

根据式(7) 可以推导出几种典型的 Bussgang 盲算法[7]. 但是, Bussgang 盲算法含有高阶统计量, 在一维信道估计存在收敛速度慢和所需样本数据过多的缺点, 所以, 这种算法在多通道估计和均衡方面不能得到推广和应用.

本算法不直接求解方程式(6) 中的矩阵 W(z, n), 而是用无训练序列的信息信号的统计特性, 在信息信号时间内用神经网络表示矩阵 W(z, n) 的行向量的非线性关系, 用最大信息量准则来训练神经网络的权值进而自适应地分离出原发射信号. 唯一的假设条件是源信号是统计独立同分布的, 这在语音系统中很容易满足.

MIMO 神经网络的输出 $y(n)= [y_{1, d}(n), y_{2, d}(n), , , y_{N, d}(n)]^T$, $d= \{1, 2, , , 52\}$ 为信息信号在 OFDM 符号中的位置的表达式为:

$$y(n)= g(u), u(n)= Wx(n)+ w_0(n) \quad (8)$$

其中, g(#) 表示神经网络的启动函数.

本文介绍的自适应盲算法的最优化准则为确定神经网络参数(W, w_0), 使得 Y 和 X 之间的互信息量 I(Y, X) 最大, 即:

$$\max\{I(Y, X)\}= H(Y)- H(Y| X) \quad (9)$$

其中, $H(Y)= - E f_Y(y)\log f_y(y)= - E[\log f_Y(y)]$ 是 Y 的熵. 由于 Y 与 X 之间总存在一个确定性系统, 所以 H(Y| X) = 0. 这样最优化准则简化为确定(W, w_0) 使 H(Y) 最大.

由式(8) 可以求得 Y 的概率密度函数为:

$$f_Y(y)= f_X(x)| J| ^{- 1} \quad (10)$$

其中, | J| 是雅可比矩阵的行列式, 将式(10) 代入到式(9), 得到:

$$H(Y)= - E[\log f_Y(y)]= E[\log| J|]- E[\log f_X(x)] = E[\log| J|]+ H(X) \quad (11)$$

H(X) 与(W, w_0) 无关, 对式(11) 求导数时只需求 $\frac{9E[\log| J|]}{9W}$ 和 $\frac{9E[\log| J|]}{9w_0}$, 用时间平均代替集平均, 可进一步推导出:

$$\begin{cases} \acute{y} W= \frac{9E[\log| J|]}{9W}= W^{- H}+ (I- 2y)x^H \\ \acute{y} w_0= \frac{9E[\log| J|]}{9w_0}= I- 2y \end{cases} \quad (12)$$

利用最陡下降法, 得到 n+ 1 时刻的自适应更新方程:

$$\begin{cases} W^{(n+ 1)}= W^{(n)}+ L\acute{y} W^{(n)} \\ w_0^{(n+ 1)}= w_0^{(n)}+ L\acute{y} w_0^{(n)} \end{cases} \quad (13)$$

这就是信道矩阵的波束形成器自适应盲分离方程. 该更新方程形式上与熟知的 LMS(最小均方准则) 相同, 因此具有与 LMS 相似的收敛特性.

5　计算机仿真与结果分析

计算机仿真方案如图 2 所示. 由于在很多情况存在大量的背景噪声, 用 MUSIC 锁定一个或几个感兴趣的说话人的声音信号, 这些信号然后被自适应盲算法跟踪.

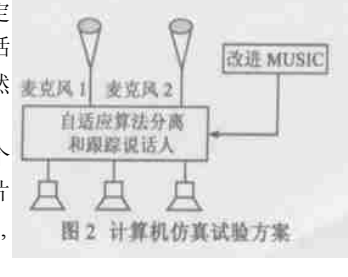

图 2　计算机仿真试验方案

在仿真实验中, 用三个近距离录制的声音信号片段表示/ 纯净0 的期望信号,

采样率为 8KHz, 一个是男声, 一个是女声, 另一个是乐队合奏音乐. 用一个时变的通道矩阵来模拟空间声音传播介质对声音的影响, 改变矩阵的元素相当于改变了声源的空间位置. 麦克风接收的是混合后的信号. 设在某一时刻通道矩阵为:

$$H(t)=\begin{pmatrix}0.2309 & 0.4764 & 0.3709\\ 0.5839 & -0.6475 & 0.2338\\ 0.8436 & -0.1886 & 0.1795\end{pmatrix}$$

首先, 对混合信号用 MUSIC 进行信源估计; 然后, 用自适应 BSS 算法进行信号的分离, 可分离的信号数目一般应小于或等于混合的信号数, 对式(13)取 L= 0.01, 初始逆矩阵 $W^{(0)}$ = I, 迭代 15 次, 得到分离矩阵 W 与通道矩阵乘积:

$$W(15)H(t)=\begin{pmatrix}0.202 & -0.0343 & 2.2048\\ -0.1126 & -2.4561 & -0.0722\\ 1.6283 & 0.0849 & -0.0164\end{pmatrix}$$

可见对角线的元素比其他元素大得多, 说明经 W 抵消后, 该算法能很好地分离出三个声音信号. 处理前后的结果对比如图 3. 从图中清晰可见期望的语音被有效地分离出来, 通过扬声器分别播放可明显听清楚这两个男女说话人的声音和背景音乐. 如果采用常规的谱分离方法根本无法分离这些声音信号, 因为声音信号在主要频带内是相互重叠的. 这里, 假设背景音乐为噪声, 用分段信噪比法(SSNR)可以估计出本方案使得信噪比提高了 20dB.

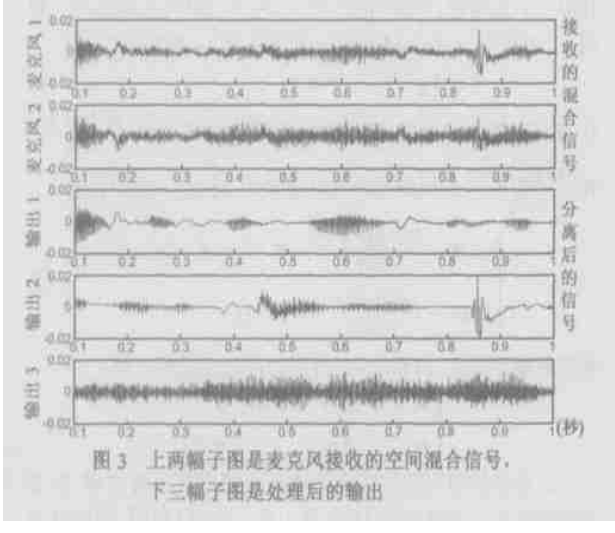

图 3 上两幅子图是麦克风接收的空间混合信号, 下三幅子图是处理后的输出

6 结论

本文介绍一种新的基于麦克风阵列的语音分离和自动跟踪技术, 该方案具有信号选择, 信号提取, 信号跟踪和分离后的语音信号的后处理等功能, 计算机仿真试验验证了该方案的可行性, 并在 TI TMS320C6711 数字处理器上的测试了该技术的实时性. 这种技术可广泛应用于与环境无关的高质量语音识别与编码、助听器、记者采访等语音通信, 具有广泛的市场应用前景和巨大的经济效益.

参考文献:

[1] G Erten, F M Salam. Voice extraction by on2line signal separation and recovery[J]. IEEE Transactions on Circuits and Systems2II: Analog and Digital Signal Processing, July 1999, CAS- 46(7): 912- 922.

[2] C Jutten, J Herault. Blind separation of sources, Part I: an adaptive al2gorithm based on neuromimetic architecture[J]. Signal Processing, July 1991, 24(1): 1- 10.

[3] E Weinstein, M Feder, A V Oppenheim. Mult2channel signal separa2tion by decorrelation[J]. IEEE Trans. Speech Audio Processing, Oct. 1993, 1: 405- 413.

[4] C Jutten, J Herault. Blind separation of sources, Part II: problems state2ment[J]. Signal Processing, July 1991, 24(1): 11- 20.

[5] R Zelinski. A microphone array with adaptive post2filtering for noise re2duction in reverberant rooms[A]. In Proceedings of ICASSP288[C]. IEEE. New York, April 1988. 2578- 2580.

[6] Iain A McCowan, Darren C Moore, S Sridharan. Near2field adaptive beamformer for robust speech recognition[J]. Digital Signal Process2ing, 2002, 12: 87- 106.

[7] Matthew R. Bielefeld and Lynn M. Supplee. Developing a test program for the dod 2400 bps vocoder selection process[A]. In Proceedings of I2CASSP296[C]. 1996. 1141- 1144.

作者简介:

杜 江 男, 1969 年出生于四川南充, 电子科技大学通信与信息工程学院博士生, 主持和参加与信号处理相关的项目 8 项, 已发表论文 10 余篇, 研究兴趣为盲算法和通信信号处理, 特别是高阶统计信号处理、阵列信号处理和自适应信号处理, 目前的研究方向为空时多载波的移动无线通信系统. E2mail: dujiang. du@samsung. com.

朱 柯 男, 1977 年生, 2000 年毕业于复旦大学, 获得电子工程学博士学位, 目前工作于韩国三星电子数字媒体技术研究所, 主要研究方向是静态/动态图像编码及其 VLSL 的结构研究; 下一代的 DVD 和数字电视系统研究.