基于盲源分离理论的麦克风阵列信号有音/无音检测方法

马晓红 梁丽丽 殷福亮
(大连理工大学电子与信息工程学院 大连 116023)

摘 要：该文提出一种在方向性噪声场中多路麦克风信号同时进行有音/无音检测(VAD)的方法。在方向性噪声场中，由于各个麦克风接收信号中的噪声彼此之间相关，因而，可以利用盲源分离理论将方向噪声与语音源信号分离，从而获得相对比较纯净的语音源信号。对分离出的语音源信号进行有音/无音检测，获得 VAD 结果，同时估计出各个麦克风信号相对于该信号的时延值。以相对纯净语音源信号的 VAD 检测结果为参考，将其分别平移相应的时延值，即可同时获得多路麦克风信号的 VAD 结果。计算机模拟结果表明，在方向性噪声场的多种情况下，该方法对具有加性噪声的多路麦克风信号均具有较好的有音/无音检测能力。

关键词：有音/无音检测；盲源分离；时延估计；广义互相关；四阶统计量
中图分类号：TN912.3 **文献标识码**：A **文章编号**：1009-5896(2007)03-0589-04

A Voice Activity Detection Method Based on Blind Source Separation for Microphone Array Signals

Ma Xiao-hong Liang Li-li Yin Fu-liang
(*School of Electronic and Information Engineering*, *Dalian University of Technology*, *Dalian* 116023, *China*)

Abstract: A Voice Activity Detection (VAD) method for microphone array signals in directional noise field is proposed. As the noises received by different microphones are correlated with each other in directional noise field, relatively pure speech can be derived from any two array signals by using Blind Source Separation (BSS) method. The generalized correlation method is used to estimate time delay between this relatively pure signal and every channel signals of microphone array. In the same time, a long-term speech information method is applied to the relatively pure speech signal to obtain its VAD result. Then this VAD result is used as reference to produce those of all array signals by the time shifting of it according to each time delay values. Simulation results illustrate the validity of the proposed method.

Key words: Voice Activity Detection(VAD); Blind source separation; Time delay estimation; Generalized correlation; Fourth-order statistics(kurtosis)

1 引言

麦克风阵列由在空间中按一定几何尺寸排列的若干个麦克风组成。麦克风阵列比单麦克风有许多优越性。例如，利用麦克风阵列提供的空域和时/频信息，可以实现声源的自动定位与跟踪[1]；还可以有效抑制房间混响、散射噪声以及其他说话人话音的干扰[2]等，因此，近年来麦克风阵列技术已广泛应用于电话会议、视频会议等系统中。在这些应用中，经常需要同时对多路麦克风信号进行有音/无音检测。

现有的多路麦克风信号VAD方法大都是基于单路信号的VAD技术，即每一路麦克风信号利用现有的单路VAD方法分别进行检测。单路信号的VAD方法包括：基于过零率、倒谱特征、自适应噪声模型、周期估计、上述参数的不同组合[3]以及长时语音信息[4]等方法。另外，文献[5]利用多路麦克风信号进行了单路信号的有音/无音检测。

本文提出一种基于盲源分离理论的多路麦克风信号 VAD 方法。该方法首先利用盲源分离技术对混有噪声的任意两路麦克风信号进行分离，获得相对比较纯净的语音源信号；然后采用一种单路 VAD 方法对较纯净的语音源信号进行有音/无音检测，同时利用广义互相关法估计该信号分别与各路麦克风信号之间的时延值；最后以较纯净语音源信号的 VAD 检测结果为参考，将该 VAD 结果分别平移相应的时延值，即可同时获得多路麦克风信号的 VAD 结果。实验结果表明，在方向性噪声场的多种情况下，本文提出方法均具有较好的有音/无音检测能力。

2 麦克风信号产生模型

理想情况下， M 个传声器接收到的信号 $x_i(t)$ $(i=1,2,\cdots,M)$ 可以表示为

$$x_i(t)=\alpha_i s(t-\tau_i)+n_i(t) \tag{1}$$

这里 $s(t)$ 为语音源信号， α_i 是语音源信号传播的衰减因子， τ_i 是语音源信号传播到各个传声器所需要的时间， $n_i(t)$ 为干扰，且 $s(t)$， $n_i(t)$ 和 $n_j(t)$ $(1\le i,j\le M, i\ne j)$ 之间彼此不相关。

2005-06-20 收到，2006-12-05 改回
国家自然科学基金(60575011，60372082)资助课题

在某些情况下，由于室内空调和投影等设备的存在，使得干扰具有方向性，即式(1)中 $n_i(t)$ 和 $n_j(t)$ 彼此之间相关，但相互之间有一定时延，此时信号模型为

$$x_i(t)=\alpha_i s(t-\tau_i)+\beta_i n(t-\tau_i') \tag{2}$$

式中 $n(t)$ 为方向性噪声，β_i 是方向性噪声传播的衰减因子，τ_i' 是方向性噪声传播到各个传声器所需要的时间。由于语音源信号与干扰一般是由不同的信源产生，因此可以认为二者之间相互独立。

3 基于盲源分离理论的多路麦克风信号 VAD 方法

基于盲源分离理论的多路麦克风信号 VAD 方法由盲源分离，利用四阶统计量确定语音源信号，时延估计和有音/无音检测 4 部分组成。其基本思想是：首先对任意两路麦克风信号进行盲源分离，得到相对比较纯净的语音源信号 $\hat{s}(t)$ 和方向噪声 $\hat{n}(t)$；其次计算语音源信号 $\hat{s}(t)$ 和方向噪声 $\hat{n}(t)$ 的四阶统计量，根据四阶统计量的大小判断出哪一路为语音源信号 $\hat{s}(t)$；然后对 $\hat{s}(t)$ 进行有音/无音检测，同时估计该信号分别与各路麦克风信号 $x_i(t)\ (i=1,2,\cdots,M)$ 之间的时延值；最后将 $\hat{s}(t)$ 的 VAD 结果分别平移相应的时延值，即可同时获得多路麦克风信号的 VAD 结果 $R_i\ (i=1,2,\cdots,M)$。该方法的系统框图如图 1 所示。

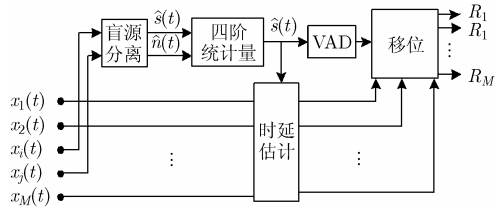

图 1 基于盲源分离理论的多路麦克风信号 VAD 方法系统框图

3.1 盲源分离

瞬时混合的盲源分离(BSS)可以用下面的混合方程描述[6,7]

$$\boldsymbol{x}=\boldsymbol{A}\cdot\boldsymbol{s} \tag{3}$$

式中 $\boldsymbol{s}=[s_1,s_2,\cdots,s_n]^{\mathrm{T}}$ 为 n 个源信号构成的向量，$\boldsymbol{x}=[x_1,x_2,\cdots,x_m]^{\mathrm{T}}$ 为 m 维观测数据向量，$m\times n$ 维矩阵 $\boldsymbol{A}$ 称为混合矩阵。

盲源分离的目的是在混合矩阵 $\boldsymbol{A}$ 和源信号 $\boldsymbol{s}$ 都未知的情况下，只根据观测向量 $\boldsymbol{x}$ 确定出分离矩阵 $\boldsymbol{W}$，使得变换后的输出 $\boldsymbol{y}$ 为源信号向量 $\boldsymbol{s}$ 的估计，即

$$\boldsymbol{y}=\boldsymbol{W}\cdot\boldsymbol{x} \tag{4}$$

式(2)示出的信号模型满足盲源分离问题对信源统计独立性的要求，因而可以利用盲源分离理论将两个源分开。由于式(2)中有时延存在，因而严格地说，应属于卷积盲源分离的范畴，利用卷积分离算法，可以得到与语音源信号比较接近的估计；但当时延值不是很大的情况下，时延可以忽略，此时利用瞬时分离算法，也可以得到语音源信号估计，只是估计值与真值之间有一定的时延存在。由于本文将分离得到的语音源信号 $\hat{s}(t)$ 及其 VAD 结果作为参考信号使用，允许时延存在，因而本文使用独立分量分析方法(ICA)[8]来进行盲源分离。

设白化后的信号为 $\tilde{\boldsymbol{x}}$，则经白化处理后，观测信号 $\boldsymbol{x}$ 变为具有单位方差的信号向量 $\tilde{\boldsymbol{x}}$，且 $\tilde{\boldsymbol{x}}$ 中各信号分量相互正交。对 $\tilde{\boldsymbol{x}}$ 进一步处理，即寻找分离矩阵 $\tilde{\boldsymbol{W}}$ 以实现独立分量的提取。分离过程是一迭代逼近过程，用变量 l 表示迭代步数，令 $y_i^{(l)}$ 为 $\boldsymbol{y}^{(l)}$ 中的某一分量，$\tilde{\boldsymbol{w}}_i(l)$ 为分离矩阵 $\tilde{\boldsymbol{W}}$ 中与 $y_i^{(l)}$ 对应的某一列向量，即

$$y_i^{(l)}=\tilde{\boldsymbol{w}}_i^{\mathrm{T}}(l)\cdot\tilde{\boldsymbol{x}},\quad l=1,2,3,\cdots \tag{5}$$

分离过程中，可用 Hyvrinen 提出的方法对分离结果 $y_i^{(l)}$ 的非高斯性进行度量，并对 $\tilde{\boldsymbol{w}}_i(l)$ 进行调整，即

$$Ng(y_i)\propto\left[E\{G(y_i)\}-E\{y_{\mathrm{Gauss}}\}\right]^2 \tag{6}$$

其中 y_i 是 BSS 的一个输出，y_{Gauss} 是与 y_i 具有相同方差的高斯分布随机量，$G(\cdot)$ 可取 $G_1(u)=\dfrac{1}{a_1}\lg\cos(a_1u)$ 或 $G_2(u)=-\exp(-u^2/2)$ 等函数。

本文采用FastICA算法[9]，其相应的调整公式为

$$\tilde{\boldsymbol{w}}_i(l+1)=E\left\{\tilde{\boldsymbol{x}}G'\left(\boldsymbol{w}_i^{\mathrm{T}}(l)\tilde{\boldsymbol{x}}\right)\right\}-E\left\{G''\left(\boldsymbol{w}_i^{\mathrm{T}}(l)\tilde{\boldsymbol{x}}\right)\right\}\cdot\tilde{\boldsymbol{w}}_i(l) \tag{7}$$

当相邻两次的 $\tilde{\boldsymbol{w}}_i(l)$ 无变化或变化很小时，即可认为 $y_i^{(l)}\approx y_i$，迭代过程结束。每次迭代后，都要对 $\tilde{\boldsymbol{w}}_i(l)$ 进行归一化处理，即 $\tilde{\boldsymbol{w}}_i(l)=\tilde{\boldsymbol{w}}_i(l)/\left\|\tilde{\boldsymbol{w}}_i(l)\right\|$，以确保式(5)的分离结果具有单位能量。对于多个独立分量，可重复使用上述过程进行分离，但每提取出一个独立分量后，要从观测信号中减去这一独立分量。如此重复，直至所有独立分量完全分离。

3.2 利用四阶统计量确定语音源信号[10]

盲源分离模块有两个输出，一个是语音源信号，另一个是干扰噪声。但是，通常我们并不知道它们中的哪个是语音源信号，哪个是干扰噪声。为了解决这一问题，考虑到语音信号服从拉普拉斯分布，属于超高斯情况，而大部分噪声属于类高斯信号，因此通过计算四阶统计量 (kurtosis)对信号的非高斯性进行度量，kurtosis 值大的那一个即为语音源信号。kurtosis 值 ξ 的计算公式为

$$\xi=\frac{E\left[|y_i|^4\right]-2E^2\left[|y_i|^2\right]-\left|E^2\left[y_i^2\right]\right|}{\sigma_{y_i}^4} \tag{8}$$

式中 $E[\cdot]$ 表示期望，$|\bullet|$ 表示绝对值，y_i 是 BSS 的一个输出，$\sigma_{y_i}^2$ 是 y_i 的方差。

3.3 时延估计[11]

经过四阶统计量计算之后，获得相对比较纯净的语音源信号 $\hat{s}(t)\approx\alpha_i's(t-t_0)$，该信号与语音源信号之间存在一定的时移，由于该信号含噪声成分较少，因而，利用一般的时延估计方法即可分别估计出该信号 $\hat{s}(t)$ 与 M 个传声器信号 $x_i(t)\ (i=1,2,\cdots,M)$ 之间的时延值。

本文采用广义互相关法进行时延估计。由于 $s(t)$ 和 $n(t)$ 之间彼此不相关，因此，信号 $\hat{s}(t)\approx\alpha_i's(t-t_0)$ 和 $x_i(t)$ 之间的互相关函数 $R_i(\tau)\ (i=1,2,\cdots,M)$ 可表示为

$$R_i(\tau)=\alpha_i\alpha_i{}'R_{ss}(\tau-(\tau_i-t_0)) \tag{9}$$

当 $\tau=\tau_i-t_0$ 时， $R_i(\tau)$ 取最大值，该最大值所对应的 $\hat{\tau}_i=\tau_i-t_0$ 值，即为信号 $\hat{s}(t)$ 与各个传声器信号 $x_i(t)$ 之间的时延值。

3.4 有音/无音检测方法

经过四阶统计量计算之后，获得相对比较纯净的语音源信号 $\hat{s}(t)\approx\alpha_i{}'s(t-t_0)$ 。由于该信号与多路麦克风信号相比纯净得多，因而大大降低了对 VAD 算法性能的要求，同时其 VAD 结果也将更加准确。本文采用文献[4]中介绍的长时语音信息方法进行 VAD 检测。由于语音和噪声的频谱特性有很大差异，即噪声频谱在各个频带之间变化较平缓，而语音频谱在各频带之间变化较剧烈，根据这一特征，通过对比长时谱包络和平均噪声谱，即可做出有音/无音判决。

由式(2)可以看出，各路麦克风信号的 VAD 结果之间大致变换规律相同，相互之间仅有一定的时延存在，因此，通过对该相对纯净语音源信号 $\hat{s}(t)$ 进行一次 VAD，将该 VAD 结果分别平移相应的时延值，即可同时获得多路麦克风信号的 VAD 结果 R_i $(i=1,2,\cdots,M)$ 。

4　计算机仿真实验结果

为了检验本文方法的性能，我们进行了计算机仿真实验。实验中使用了 3 个间距为 30cm 麦克风($M=3$)构成的麦克风阵列。假设方向噪声场中仅有 1 个语音源和 1 个噪声源存在，这样源信号向量的维数 $n=2$ ；又由于从 3 路数据中任取 2 路进行盲源分离，因此 $m=2$ 。根据语音源、噪声源及每个麦克风所处的空间位置，对图 2 示出的一段 8kHz 采样率、数据长度为 30000 点的纯净语音信号和 4 种不同类型的噪声(白噪声、有色噪声、幅度逐渐增加的白噪声和有色噪声)进行相应的移位，并分别在 3 路信号中以一定信噪比(SNR)加入了 4 种不同类型的噪声，图 3 示出了加入 4 种不同类型噪声的 3 路带噪信号中的 1 路。图 2 和图 3 中横轴为时间 t 的采样点 N，纵轴为信号幅度 A 。

对图 3 示出的 4 种 30000 点带噪语音信号用 FastICA 方法进行分离，获得较纯净的语音源信号 $\hat{s}(t)$ 和方向噪声 $\hat{n}(t)$ ，用四阶统计量进行判断，确定出 4 种噪声情况下的 $\hat{s}(t)$ 信号，

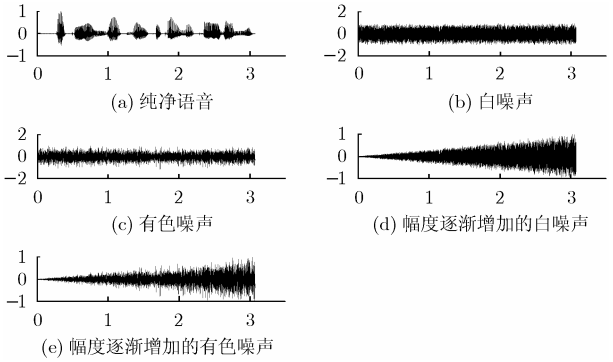

图 2　纯净语音和 4 种不同类型的噪声

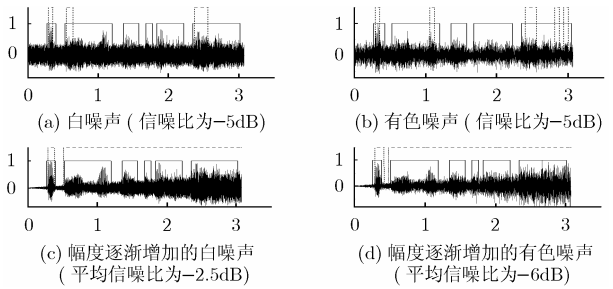

图 3　加入 4 种不同类型噪声的带噪语音信号

对它们分别用文献[4]中的方法进行有音/无音检测。分离出的相对纯净的语音源信号 $\hat{s}(t)$ 及其 VAD 结果如图 4 所示，可以看出，分离效果比较理想，其 VAD 检测结果也比较准确。以图 4 中的 VAD 检测结果为参考，将其分别根据图 4 中信号与图 3 中对应信号的时延值进行移位，从而获得了在 4 种噪声情况下带噪语音的 VAD 检测结果，如图 3 中实线所示；图 3 中虚线示出了直接用文献[4]中方法对带噪语音进行 VAD 检测的结果。由图 3 可以看出，对带噪语音信号直接进行 VAD 检测的结果非常不准确，而本文提出的方法在各种噪声情况下都可以获得比较理想的 VAD 结果。

由于分离算法采用的是 FastICA，其对两路 30000 点数据的分离时间较短，大约为 1.2 秒，因此，本文方法引入的延时影响主要取决于采用的 VAD 算法的复杂性。

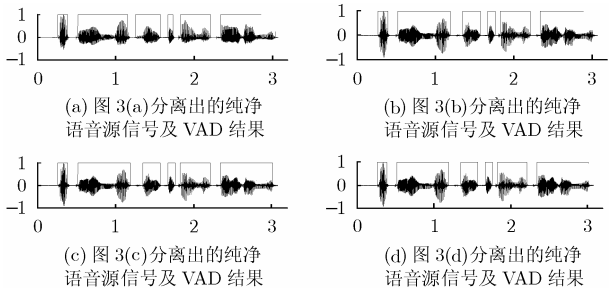

图 4　4 种噪声情况下分离出的语音源信号及 VAD 结果

5　结束语

本文提出一种基于盲源分离理论的多路麦克风信号 VAD 方法。盲源分离技术用于对混有方向噪声的任意两路信号进行分离，获得相对比较纯净的语音源信号；然后采用一种单路 VAD 方法对较纯净的语音源信号进行有音/无音检测，同时利用广义互相关法估计该信号分别与各路麦克风信号之间的时延值；最后以较纯净语音源信号的 VAD 检测结果为参考，将该 VAD 结果分别平移相应的时延值，即可同时获得多路麦克风信号的 VAD 结果。计算机仿真实验结果验证了该方法的有效性。

参 考 文 献

[1] Gustafsson T, Rao B D, and Trivedi M. Source localization in reverberant environments: modeling and statistical analysis. *IEEE Trans. on Speech and Audio Processing*, 2003, 11(6): 791–803.

[2] Gannot S and Cohen I. Speech enhancement based on the general transfer function GSC and postfiltering. *IEEE Trans.*

on Speech and Audio Process., 2004, 12 (6): 561–571.

[3] Tanyer S G and Özer H. Voice activity detection in nonstationary noise. *IEEE Trans. on Speech and Audio Process.*, 2000, 8 (4): 478–482.

[4] Ramírez J, Segura J C, and Benítez C, *et al.*. Efficient voice activity detection algorithms using long-term speech information. *Speech Communication*, 2004, 42 (3-4): 271–287.

[5] Chen J F and Ser W. Speech detection using microphone array. *Electronics Letters*, 2000, 36(2): 181–182.

[6] Cao X R and Liu R W. General approach to blind source separation. *IEEE Trans. on Signal Processing*, 1996, 44(3): 562–571.

[7] Cardoso J F. Blind signal separation: Statistical principles. *Proce. IEEE*, 1998, 86(10): 2009–2025.

[8] Comon P. Independent component analysis, A new concept? *Signal Processing*, 1994, 36(3): 287–314.

[9] Hyvarinen A and Oja E. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 1997, 9(7): 1483–1492.

[10] Siow Yong Low, Nordholm S, and Togneri R. Convolutive blind signal separation with post-processing. *IEEE Trans. on Speech and Audio Processing*, 2004, 12(5): 539–548.

[11] Knapp C and Carter G. The generalized correlation method for estimation of time delay. *IEEE Trans. on Acoustics, Speech, and Signal Process.*, 1976, 24(4): 320–327.

马晓红： 女，1967年生，副教授，在职博士生，从事语音信号处理和阵列信号处理的理论与应用研究工作.

殷福亮： 男，1962 年生，教授，博士生导师，主要从事语音信号处理和阵列信号处理的理论与应用研究工作.