一种基于模式识别的多路盲语音提取方法*

徐 舜 刘郁林 柏 森
(重庆通信学院DSP实验室 重庆 400035)

摘要 盲分离算法能在缺少混合系统参数的条件下仅由观测信号估计初始源,但分离信号存在固有的排列模糊性,这往往导致两次批处理过程中同一信号"对不准",因此很难获得连续的源信号。本文针对盲声源分离中存在的相同问题,根据语音和其他音频信号的特征差异,提出一种修正的自相关函数并以其值作为一个特征基元来表征声音信号的时序相关特性,同时用平均声门波形状参数作为另一个特征基元来表征语音产生的生理效应。以这两个参数作为识别不同音频信号的二维模式特征,采用一种模糊聚类算法提取多路盲分离语音。本方法有效克服了批处理盲声源分离中的信号排列顺序的不确定性,并通过选择合适的阈值提取多路连续语音。仿真给出了5路混合音频信号中盲提取两路连续语音的实验结果。

关键词 盲分离,模式识别,语音

An approach to multiple blind-speech-abstraction based on pattern recognition

Xu Shun Liu Yu-Lin Bai Sen
(*DSP Lab, Chongqing Communication College, Chongqing* 400035)

Abstract Classic blind source separation algorithm can estimate the original sources without given parameters of the mixture system by only using observation signals, but the inherent permutation ambiguity always leads to a signal not being "end to end" during the two batch processings, and so it is difficult to obtain the consecutive sources. Aiming at the same question in the blind audio source separation, this paper proposes a pattern recognition approach for its solution. From the difference between the characteristics of speech and other audio signals, a modified autocorrelation function is proposed and its value is taken as a characteristic basis to express the time sequence correlation of different audio signals, while the mean glottal flow shape parameter is taken as the other characteristic basis to express the physiological effect that speech produces. The above parameters are selected as two-dimension pattern characteristic to distinguish different audio signals, and a fuzzy clustering algorithm is used to extract multiple speeches blindly separated. This approach effectively solves the permutation ambiguity in batch blind audio source separation and by selecting appropriate threshold can extract multiple consecutive speeches. Simulation shows the experiment results of blind extracting two consecutive speeches among five mixture acoustic signals.

Key words Blind signal separation, Pattern recognition, Speech

2006-08-28 收稿; 2007-06-01 定稿

* 国家自然科学基金资助项目(No. 60672157, No. 60672158)

作者简介:徐舜(1980—),男,湖北沙市人,工学硕士。研究方向为:盲分离算法及实现,语音信号处理。
刘郁林(1971—),男,教授,工学博士,硕士生导师。 柏森(1963—),男,教授。

† 通信联系人 E-mail: xushun@sohu.com

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1 引言

在批处理盲声源分离过程中，需要对分离出的音频信号帧快速进行扫描识别，特别是在战场环境中，要识别出分离的信号哪些是干扰信号(如枪炮声，飞机飞行声，坦克行进声等)，哪些是我们需要的指挥员的语音，然后将识别出的语音帧拼接成连续的语音信号。而盲分离过程中的排列不确定性[1]导致了连续语音获取的困难，因此可以考虑用模式识别的方法来判断音频信号的类型，从而达到提取连续语音的目的。

所谓模式识别是根据研究对象的特征或属性，利用以计算机为中心的机器系统运用一定的分析算法认定它的类别，系统应使分类识别的结果尽可能地符合真实。自20世纪30年代Fisher提出统计分类理论，奠定了统计模式识别的基础到现在，这门学科涉及的理论与技术已相当广泛，目前，主流的技术[2,3]有：句法模式识别、模糊数学方法、神经网络法、人工智能方法等。这些理论和技术已成功的应用于工业、农业、国防、科研、公安、生物医学、气象、天文学等许多领域。

对于盲分离语音信号的识别，[4]中介绍了一种基于相邻频点幅度相关特性的语音信号获取方法，这种方法利用不同信号相邻频点之间的相关性差异来进行分离信号位置上的重新排列，从而解决了盲分离问题中的排列不确定问题，但该方法需要的数据量大，只能离线处理，频域变换使得计算的复杂度增加，不便于硬件实现。[5]中提出用时频掩蔽的方法来识别盲分离后的语音信号，这种方法利用了听觉系统的心理声学特性，避免了固有的盲分离排序矛盾，可以直接将语音信号提取出来，运算效率比较高，但该方法同样要利用域变换，并且掩蔽门限的选择需要前端的训练，增加了处理的时间和冗余量。考虑到以上的问题以及语音和其他音频信号的特性差异，本文提出了一种新的方法从时域来解决盲声源分离中连续语音获取的问题，该方法首先根据语音的短时平稳性将批处理盲分离算法中的每一帧信号作分窗处理，然后定义两个参量—修正的自相关函数值和平均声门波形状参数，用它们分别反映语音和噪声之间的时序结构差异以及不同语音之间生理特性，并以此作为二维模式特征来区分不同的音频信号，最后对各帧分离信号的二维特征矢量进行模糊聚类判别，识别不同语音。由于算法在时域中处理并且每次运算数据量较少，通过验证，识别的时间少于盲分离算法所需的时间，因此可以将各帧分离信号进行无缝拼接，获得连续语音。另外，基于此方法的硬件实现也给出了良好的识别效果。

2 问题描述

盲分离问题模型通常可以表述为图1的形式，其中，$A \in R^{mxn}$ 是未知的列满秩混合矩阵，$s(k)=[s_1(k),s_2(k),\cdots,s_n(k)]^T$，是彼此独立的 n 维源信号矢量。$x(k)=[x_1(k),x_2(k),\cdots,x_m(k)]^T$ 是 m 维观测信号矢量；$v(k)=[v_1(k),v_2(k),\cdots,v_m(k)]^T$ 为与源信号不相关的加性噪声矢量，其中"T"表示向量的转置。为从混合信号 $x(k)$ 中分离出源信号 $s(k)$，就是要通过寻找一个 $n\times m$ 阶的满秩线形变换(或分离)矩阵 W，以便由 $y(k)=Wx(k)$ 定义的输出矢量 $y=[y_1,y_2,\cdots,y_n]^T$ 的各分量尽可能相互独立，其独立性可用基于 K-L 散度或稀疏度、光滑度、线形预测性等信息论损失函数来度量，这样就可以获得关于随机向量 $x(k)=[x_1(k),x_2(k),\cdots,x_m(k)]^T$ 的盲分离和混合矩阵 A 的辨识。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

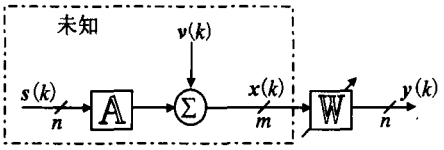

图 1　盲分离模型

$$x(k) = As(k) + v(k) \tag{1}$$

$$y(k) = Wx(k) \tag{2}$$

需要指出的是,在"盲"的范畴里,不可能实现混合矩阵 A 的完全辨识,即盲分离存在模糊和不确定性,这种不确定和模糊性可以看作是对被估计源信号的伸缩、排序或时滞。

设批处理盲分离算法每进行一次分离运算得到的音频信号为 $\hat{s} = [\hat{s}_1, \cdots, \hat{s}_n]^T \in R^{n \times L}$,$L$ 为每帧分离信号的长度,$\hat{s}_{d_1}, \cdots, \hat{s}_{d_k}$ $(1 \leqslant d_1, \cdots d_k \leqslant n)$ 为 k 帧语音信号,其余帧为非语音的音频干扰,那么提取语音信号帧的识别系统可以描述为图 2。

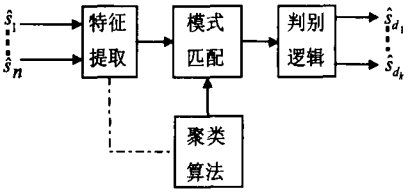

图 2　语音识别系统模型

需要说明的是,为了保证提取连续的语音,模式识别的时间应少于一次批处理盲分离算法完成的时间,这样在硬件实现过程中就不会因为两次批处理中间的识别算法而产生数据的延时,同时也不会占用过多缓存空间。这就要求模式的选择必须尽可能的少但最能反映音频信号的差异,而且聚类算法要足够简单有效。

3　特征选择

语音是一个时变的、非平稳的随机过程,人类发声系统的生理结构的变化速度是有一定限度的,在一段短时间内(10 ~ 30ms)人的声带和声道形状有相对稳定性,可以认为其特征是不变的,因此语音信号具有短时平稳性[6]。而在战场环境中,噪音除了宽带白噪声就是如机器轰鸣声、枪炮声等周期性噪声或脉冲噪声。因此这里将音频信号大致归为三种典型的类别:语音、非白噪声和宽带白噪声。在后面的讨论中我们将会以这三种类别作为研究的模式。人工观察噪声和语音信号的实时波形,它们是有明显差别的,是能够确切地识别的,这表明语音信号和噪声虽然在频域上要占相同的频段,但在时域上却有明显的差异。自相关函数正是信号在时域最基本、最重要的特征,因此,可以利用语音信号与噪声短时自相关函数值的差异来提取模式特征。另外,为了同时区分不同的语音,我们选择语音的平均声门波形状参数 α_m 作为另一个模式特征,因为不同说话人的声门特征分布可以用形状参数 α 明显地分开。

3.1　修正的自相关函数

设一次批处理盲分离算法获得的一帧源信号为 $\hat{s}_i = [\hat{s}_i(1), \cdots \hat{s}_i(L)], 1 \leqslant i \leqslant n$,取窗长 $T(T \ll L$ 一般为语音信号基音周期)进行分块处理,设窗移为 L_0,两窗重叠 L_1,因此有 $L_0 + L_1 \equiv T$,窗数 $N = \dfrac{L - L_1}{T - L_1}$,源信号在第 d 个窗内的自相关函数为:

$$R_T(d,p) = \frac{1}{T-P}\sum_{k=p+1}^{T} \hat{s}((d-1)T + k) \cdot \hat{s}((d-1)T + k - p) \tag{3}$$

这里,p 表示时延。将自相关函数进行归一化处理,定义为 $r_T(d,p)$,

$$r_t(d,p) = \frac{R_T(d,p)}{R_T(d,0)} \times 100 \tag{4}$$

为了反映语音的非平稳特性和短时平稳性,我们定义一种修正的自相关函数:

$$\hat{R} = R_s(N,T) = \frac{1}{N}\sum_{d=1}^{N} \rho(N,d) \cdot r_T(d,p) \tag{5}$$

其中,ρ 是跟随信号非平稳变化的遗忘因子,$0 \leqslant \rho \leqslant 1, d = 1, \cdots N$。文中我们定义 $\rho(N,d) = \lambda^{N-d}$,典型的,$\lambda$ 取 0.98。这个定义

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

的优势如下：

文献[7]通过计算(4)式来作为识别音频信号差异的模式特征，不同的是它没有进行重叠分窗，而是直接计算100个样点(抽样频率 11025Hz) 的归一化自相关函数值 $r_T(p)$。我们发现，这种方法用来区分语音信号和宽带白噪声效果是显著的。但由于语音可以分为静音、清音和浊音，浊音具有短时平稳的周期性而清音的时域特征类似于白噪声，如果仅用 $r_T(p)$ 来识别语音和其他非白噪声，特别是周期性噪声或脉冲噪声是不稳健的，如图3所示，

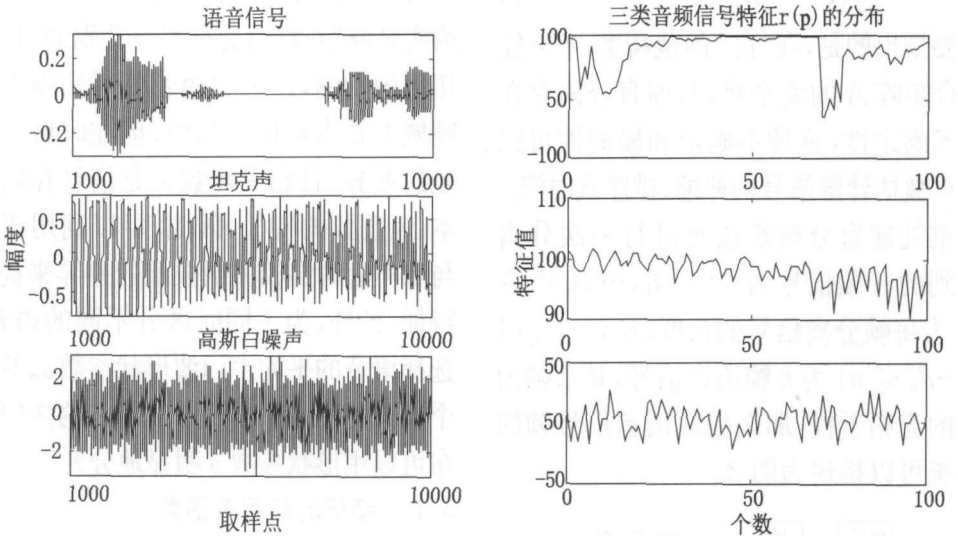

图3 三类音频信号的时域波形(样点数:10000)(左)
以及对应的100个特征 值曲线(右)

从图3中可以看出，取100个抽样点的归一化短时自相关函数值 $r_T(p)$ 作为模式特征，语音信号在 $-40 \sim 100$ 之间，坦克声信号在 $90 \sim 101$ 之间，高斯白噪声在 $-20 \sim 30$ 之间，它们都不具有长时的谱平坦性，同时特征值有很大的重叠，因此单用[7]中的方法不能良好的判断出语音信号，必须选择更能反映音频模式差异的特征。

当我们选择合适的帧长、时延、窗长和窗移时，利用修正的归一化自相关函数值，可以显著区分各类音频信号。同时考虑到用最少的数据量获得良好的分离和识别性能，减少算法运算量和便于硬件实现的需要，经反复实验权衡得到：当抽样频率为 11025Hz，$L = 2500$，$T = 100$，$L_0 = \frac{1}{4}T$，$p = 3$ 时，$\hat{R}$ 作为识别不同音频信号的模式特征是有效的。

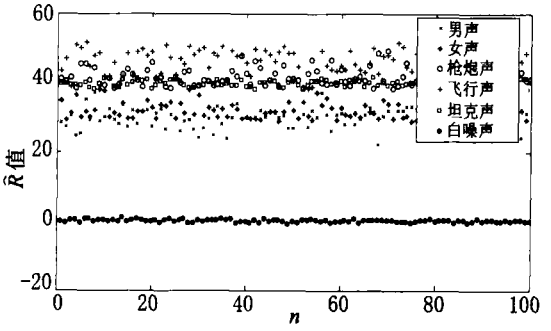

图4 100次实验得出的三类(6种)
不同声音的 值分布

图4中画出了具有典型性的三类共6种音频信号分别经过100次实验计算出的 $\hat{R}$ 值分布情况。从图中可以清楚地看到，语音同白噪声、脉冲噪声或周期性噪声等非白噪声通过定义的特征值 $\hat{R}$ 显著的分离开来，这说明当我们选择合适的阈值(本文设定的语音

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

阈值为 32) 后,通过基于最小距离原则的聚类算法能从噪声中识别出语音。但不同的语音没法根据 $\hat{R}$ 值来区分。

3.2 声门波形状参数

由语音学知识我们知道,不同说话者口腔和鼻腔的长度、声道截面积、声带的质量和形状等是不一样的,也就是说可以根据这些生理参数建立语音生成和感知模型,提取不同说话者的特征,从而分辨不同的语音。迄今比较成熟的用于识别不同语音的特征参数主要是语音的谱特征,比如 Mel 倒谱和子倒谱等。本文利用声门波形状参数作为一个特征基元主要考虑到:第一,声门波导数的估计是基于声源和声道的特性,是时域上的,而不是基于频谱的,可以有效的和前一个模式特征在识别域上统一起来;第二,实验已经证明,声门特征分布,特别是形状参数能有效区分不同说话人的语音特征。基于以上两点,文中选择了声门波形状参数 α 来作为盲声源分离后识别不同语音的模式特征。

根据语音产生的声学理论和零-极点语音模型,声门波导数的"粗糙结构"可表示为由 7 个参数组成的分段函数的 Liljencrants-Fant(LF) 模型,该模型由牛顿-高斯型非线性估计方法获得[6]。

LF 模型中,单个声门周期内声门波导数的粗糙分量表示为:

$$V_{LF}(t)=\begin{cases}0 & 0\leqslant t<T_0\\ E_0e^{\alpha(t-T_0)}\sin[\Omega_0(t-T_0)] & T_0\leqslant t<T_e\\ E_1[e^{-\beta(t-T_e)}-e^{-\beta(T_c-T_e)}] & T_e\leqslant t<T_c\end{cases} \tag{6}$$

其中 $E_1=\dfrac{E_e}{1-\exp[-\beta(T_c-T_e)]}$,$E_0=\dfrac{E_e}{e^{\alpha}(T_e-T_0)\sin[\Omega_0(T_e-T_0)]}$,$E_e$ 是负峰的绝对值。时间原点 $t=0$ 是闭合相的起始时刻(也是前一个相邻声门周期的返回相的终止时刻),T_0 是开启相的起始时刻(也是闭合相的终止时刻),T_e 是返回相的起始时刻(也是开启相的终止时刻以及声门波的持续时间),T_c 是返回相的终止时刻(也是下一个声门周期闭合相的起点)。E_0,Ω_0 和 α 三个参数用来描述开启相期间声门波形状。E_0 和 β 两个参数用来描述返回相期间声门波的形状。模型中 4 个形状参数和 3 个脉冲分量定时参数的具体含义见表 1:

表 1　LF 模型表示的声门波导数波形参数

T_0	声门开启时刻
α	决定 E_e 与声门波导数的正部峰值高度之比的因子
Ω_0	决定声门波左边脉冲导数曲率的频率;也决定了零相交和 T_e 之间的时间间隔
T_e	声门波的最大负值时刻
E_e	脉冲导数在 T_e 时刻的值
β	指数时间常数,决定脉冲导数在 T_e 时刻之后返回零值的速度。
T_c	声门闭合时刻。

为了说明声门波形状参数 α 与说话人的相关性,我们选择 4 段长度为 20 秒的不同语音,其中两段为男生,两段为女生,图 5 用直方图表示了 α 的分布情况,从图中可以看到,不同的说话者其声门波形状参数有很大的差异,能够用 α 进行良好的区分。因此我们取一帧盲分离信号的平均声门波形状参数 α_m 作为识别不同语音的模式特征。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

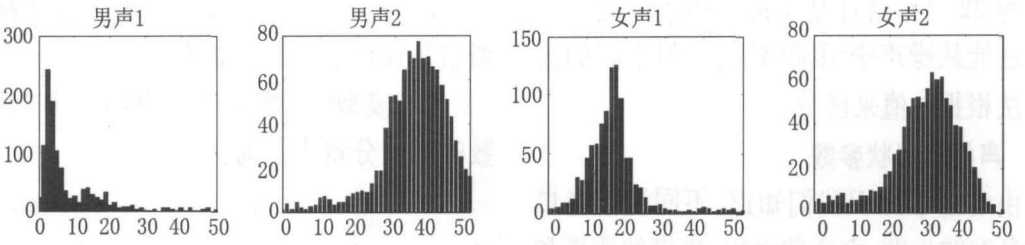

图5 声门波形状参数 的直方图比较

4 识别算法

上述可知，利用二维模式特征参数[$\hat{R}$，α_m]能简单有效地反映盲分离音频信号的特征差异，而要准确迅速地将各帧分离信号进行拼接，同时提取其中的连续语音还需要利用聚类分析进行识别。本文根据相似性阈值和最小距离原则采用模糊聚类来识别不同语音，这种算法较以往类心不变的算法，如$C-$均值算法[3]等，能更有效地对信号进行动态分类识别。

设按文中方法提取的二维特征模式集为$X=\{x_1,x_2,\cdots,x_N\}$，其中的每一个元素表示一帧分离信号的特征参数，要将这N个二维特征矢量分成c类，分类结果用分类矩阵$U=(u_{ij})_{cxN}$表示。这个矩阵的阵元u_{ij}表示目标x_j隶属于ω_i类的程度，它满足：

(1)$u_{ij}\in[0,1]$；

(2)$0<\sum_{j=1}^{N}u_{ij}<N,\forall i$，即任一类都不是确定的空集，总是有一些模式以不同的程度隶属于它，同时它也不是确定的全集X；

(3)$\sum_{j=1}^{N}u_{ij}=1,\forall j$即$X$中的每一个模式$x_j$属于各类的程度总和为1。

算法在迭代寻优过程中，不断更新各类的中心及分类矩阵各元素的值，直到逼近下列准则函数最小值

$$J_a(U,Z)=\sum_{j=1}^{N}\sum_{j=1}^{c}u_{ij}^{a}d_{ij}^{2} \qquad (7)$$

式中，$Z=\{v_1,v_2,\cdots,v_c\}$，v_i为ω_i类的中心矢量，权重$a\in(1,\infty)$，d_{ij}为欧式距离$\|x_j-v_i\|_2$。具体算法步骤如下：

(1) 确定类数c、权重a以及终止条件ε。文中c取估计源的个数，$<1<\leqslant 3$，ε取一个大于零的适当小数。

(2) 置定初始分类矩阵U^0，令$\eta=0$。

(3) 计算$U^{(\eta)}$时的$\{v_i^{(\eta)}\}$：

$$v_i^{(\eta)}=\sum_{j=1}^{N}u_{ij}^{a}x_j/\sum_{j=1}^{N}u_{ij}^{a},i=1,2,\cdots,c \qquad (8)$$

(4) 对j从1到N，计算$Ij=\{i\mid d_{ij}=0\}$，$\bar{I}_j=\{1,2\cdots,c\}-I_j$。

(5) 计算x_j的新隶属度从而更新$U^{(\eta)}$为$U^{(\eta+1)}$：如果$I_j=\phi$，那么

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{a-1}}}$$

$u_{ij}=0,\forall i\in\bar{I}_j$，并取$\sum_{k\in Ij}^{c}u_{ij}=1$。

(6) 如果$\|U^{\eta}-U^{(\eta+1)}\|<\varepsilon$，停止；否则，$\eta=\eta+1$，返回(3)。

当算法终止，比较分类矩阵U每列的最大值从而确定各特征矢量隶属的类别。由于从二维模式特征中的$\hat{R}$值可以首先判别语音和噪声类，然后由α_m区分不同语音，因此由($\hat{R}$，α_m])构成的二维空间中语音与语音之间以及语音与噪音之间可以形成相互分离的的簇结构，也即聚类过程能解决分离音频信号的排列不确定性问题，使各帧语音正确拼接起来，提取出每个不同的语音。又因为特征维数低，聚类算法简单，经实验验证处理时间小于一帧混合信号的分离时间，因此最后能获得连续的语音信号。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

5　实验仿真

我们以频率11025Hz分别采样了5段时长2秒的不同声音信号,如图6(左)所示,从上往下分别是男声1、坦克行进声、飞机飞行声、男声2和枪炮射击声。经未知混合系统后得到如图6(右)所示的5路混合信号,用*SOCW*批处理盲分离算法[8]进行混合信号的分帧分离(每帧2500个样点,时长约0.23秒),其波形如图7所示,图7中每列5个子图分别代表分离出的第1至第5路源信号中的一帧。同时利用本文的方法边分离边识别,提取连续语音。因为只是验证所提出方法的有效性,因此只取了三帧来识别(见图7,8)。

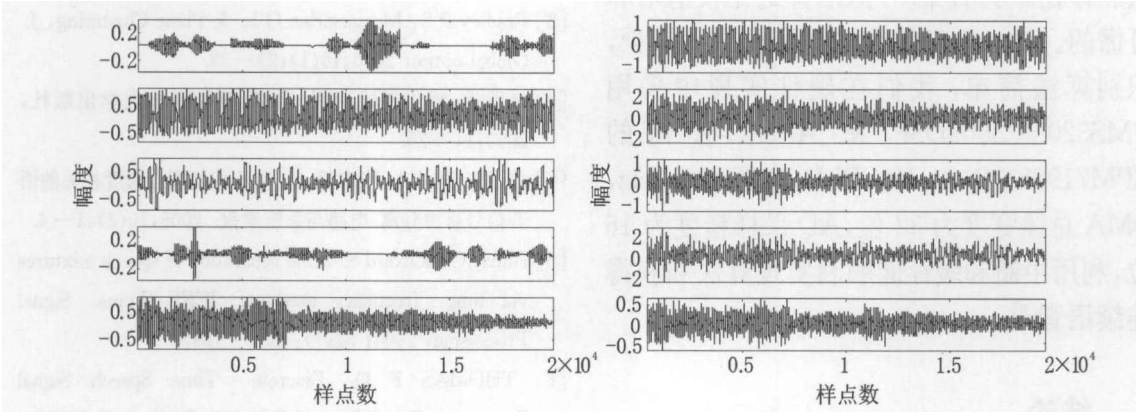

图6　源信号(左)和混合信号(右)波形(样点:20000)

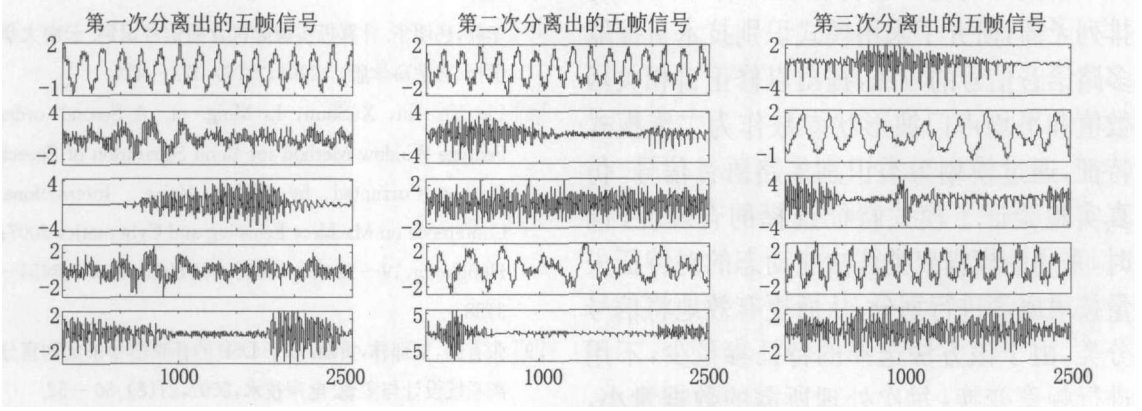

图7　连续三次盲分离分别得到的源信号帧

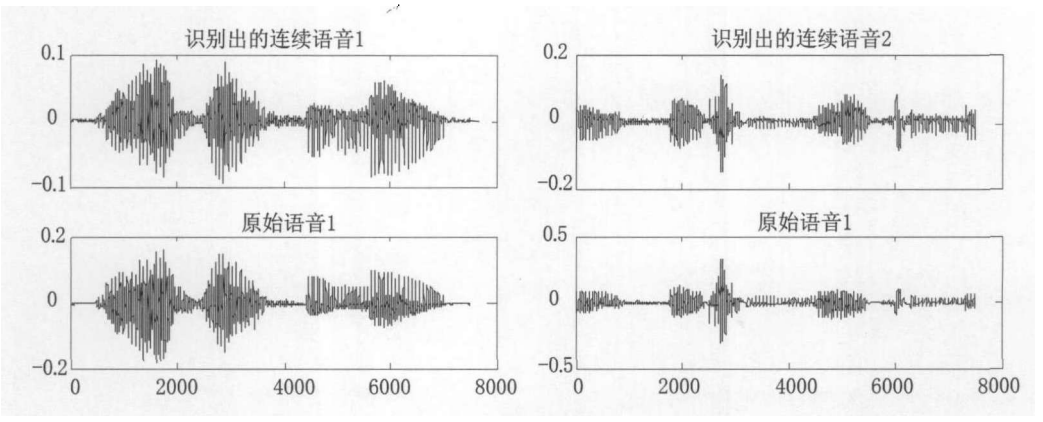

图8　识别出的语音信号和源语音信号的对比(前三帧7500个样点)

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

从图7可以看到，经典批处理盲分离算法每次所分离出的音频信号排列位置具有不确定性，第一次分离出的语音帧在第3路和第5路，第二次则在第2路和第5路，而第三次分离出来后在第1路和第3路。利用本文提出的方法将混杂在三路音频噪声中的两路语音进行了良好的拼接识别，试听效果虽然存在微弱的噪声，但语音是比较清晰和可懂的。另外，由于该方法需要的数据量少，识别算法简单，我们在硬件实现中采用TMS320VC5509DSP和Altera公司的EPM7192SQC160-10 CPLD，主频120MHz，DMA总线宽度为32位，AD采样精度为16位，利用中断和缓存能顺利实现算法和获得连续语音[9]。

6 结论

本文针对盲声源分离算法存在的固有排列矛盾，研究了采用模式识别技术盲提取多路语音信号的方法，提出以修正自相关函数值和平均声门波形状参数作为二维模式特征，通过模糊聚类识别多路语音信号。仿真实验验证了模式特征选择的合理性，同时，采用的模糊聚类算法能动态的对特征矢量按隶属度进行划分，从而更有效地将信号分类。由于该方法选择的特征参量少，不用进行频率变换，每次处理所需的数据量小，结合批处理的SOCW盲分离算法，较好地实现了多路连续语音的盲提取。

参 考 文 献

[1] Choi S, Cichocki A et al. Blind source separation and independent component analysis: A review. Neural Information Processing - Letters and Reviews, 2005, 6(1):1－57.

[2] Bradley P S, Mangasarian O L. k-Plane Clustering, J. Global optim, 2000,16(1):23－32.

[3] 孙即祥. 现代模式识别. 北京：国防科技大学出版社，2002:14－25.

[4] 姜卫东，陆佶人，张宏滔，等. 基于相邻频点幅度相关的语音信号盲源分离. 电路与系统学报. 2005,10(3):1－4.

[5] Yilmaz O, Rickard S. Blind separation of speech mixtures via time - frequency masking. IEEE Trans. Signal Processing, 2004; 52(7):1830－1847.

[6] THOMAS F Q. Discrete - Time Speech Signal Processing:Principles and Practice. BeiJing: Publishing House of Electronics Industry, 2004; 393－417,172－180.

[7] 李莉，杨明华. 计算机实现随机音频信号识别. 云南大学学报(自然科学版)，2001,23(6):422－424.

[8] Liu Yu - lin, XuShun, Li Ming - qi. A Second - order Feature Window Method for Blind Separation of Speech Signals Corrupted by Color Noise. International Conference on Machime Learning and Cybernetics 2007; Hongkong,19－22 August 2007;Volume 6 of 7:3454－3458.

[9] 张有鹏，刘郁林，黄磊. 基于DSP的音频信号采集和盲分离系统设计与实现. 电声技术，2007,31(8):50－52.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net