

文章编号:1004-9037(2014)02-0198-06

## 基于稀疏编码的鲁棒说话人识别

何勇军<sup>1</sup> 孙广路<sup>1</sup> 付茂国<sup>1</sup> 韩纪庆<sup>2</sup>

(1. 哈尔滨理工大学计算机科学与技术学院, 哈尔滨, 150080; 2. 哈尔滨工业大学计算机科学与技术学院, 哈尔滨, 150001)

**摘要:** 目前的说话人识别系统在噪声环境下性能将急剧下降, 为了解决这一问题, 提出了一种新的基于稀疏编码的说话人识别方法。该方法用一个通用背景字典(Universal background dictionary, UBD)刻画说话人语音的共性, 并为每个说话人和环境噪声训练相应的字典来刻画说话人和环境的特殊变化。这些字典被拼接成一个大字典, 然后将待测试语音稀疏分解在这个大字典上以实现说话人识别。为了提高说话人字典的区分能力, 通过从说话人字典中移除与通用背景字典原子相似的原子来优化说话人字典。为了跟踪变化的噪声, 采用混噪语音在线更新噪声字典。在各种噪声条件下的实验表明, 所提出的方法在噪声环境下具有较强的鲁棒性。

**关键词:** 说话人识别; 稀疏表示; 判别字典; 形态成分分析

中图分类号: R318.04

文献标志码: A

## Robust Speaker Recognition Based on Sparse Coding

He Yongjun<sup>1</sup>, Sun Guanglu<sup>1</sup>, Fu Maoguo<sup>1</sup>, Han Jiqing<sup>2</sup>

(1. School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, 150080, China;

2. School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China)

**Abstract:** Speaker recognition suffers severe performance degradation under noisy environments. To solve this problem, a novel method is proposed based on sparse coding of signals. This method employs a universal background dictionary (UBD) to model common variability of all speakers, a speaker dictionary to model special variability of each speaker and a noise dictionary to model variability of environmental noise. These three dictionaries are concatenated to be a big dictionary, over which test speech is sparsely represented and classified. To improve the discriminability of speaker dictionaries, the speaker dictionaries are optimized by removing speaker atoms which are close to the UBD atoms. To ensure the varied noises can be tracked, an algorithm is designed to update the noise dictionary with the noisy speech. Experimental results under various noise conditions show that the proposed method can obviously improve the robustness of speaker recognition under noisy environments.

**Key words:** speaker recognition; sparse representation; discriminant dictionary; morphological component analysis

## 引 言

说话人识别在过去的几十年里受到了广泛关

注,吸引了大量学者进行深入研究。传统方法比如高斯混合-通用背景模型(Gaussian mixture model-Universal background model, GMM-UBM)<sup>[1]</sup>, 高斯混合-支持向量机(Gaussian mixture model-

**基金项目:** 国家自然科学基金(61305001, 91120303, 91120301)资助项目; 高等学校博士学科点专项科研基金(20132303120003)资助项目; 中国博士后基金(2013M531042)资助项目; 黑龙江省博士后基金(LBH-Z13099)资助项目; 黑龙江教育厅基金(12511096)资助项目; 黑龙江省自然科学基金(F200936)资助项目。

**收稿日期:** 2014-01-23; **修订日期:** 2014-02-13

support vector machine, GMM-SVM)<sup>[2]</sup>和联合因素分析(Joint factor analysis, JFA)<sup>[3]</sup>等,在无噪的理想情况下取得了令人满意的识别率;然而在噪声环境<sup>[4]</sup>下其性能将急剧降低,这严重限制了说话人识别技术走向现实应用。

为解决这一问题,研究者们提出了大量方法,这些方法大致可分为两类。第一类是提取鲁棒特性,例如线性预测倒谱系数(Linear prediction cepstral coefficient, LPCC),梅尔倒谱系数(Mel-frequency cepstral coefficient, MFCC)<sup>[5]</sup>和感知线性预测系数(Perceptual linear predictions, PLPs)。特征类方法相对简单,计算复杂度低,可以集成在识别系统的前端,但其性能有限。因为目前还没有一种特征能有选择地只表示语音而不表示噪声。第二类方法则尝试首先去除语音中噪声,然后从增强后的语音中提取特征,典型的有谱减法和维纳滤波。然而,噪声常常是非平稳的,有些噪声甚至像语音,很难建模和估计。因此这类方法不可避免会造成语音特征的进一步畸变。而现实情况是目前的话人识别方法对这种畸变是非常敏感。因此,希望采用新技术来解决这个问题。

近年来,研究者们对稀疏编码<sup>[6]</sup>进行了广泛而深入的研究,为噪声环境下的说话人识别提供了可能的解决方案。稀疏编码用一组原子(基元信号)来表示信号,所有原子的集合称为原子字典。稀疏编码旨在用少量的原子的线性组合来表示信号。最近,一种被称为形态成分分析(Morphological component analysis, MCA)<sup>[7]</sup>的稀疏编码方法已被成功应用于说话人识别。该方法为每个说话人训练一个字典,并且所有说话人字典连接成一个大字典。在识别过程中,测试语音被稀疏地表示在大字典上。理论上,一个说话人说出的话语只能表示在这个说话人的字典上。因此,稀疏表示可以直接用于分类。

几乎所有基于稀疏编码的说话人识别方法都采用 MCA 的框架。目前常用的方法是首先将说话人的高斯混合模型(Gaussian mixture model, GMM)转变为均值超向量<sup>[8-9]</sup>或全部可变量  $i$ -向量<sup>[10]</sup>,然后把这组向量组成一个大字典,通过在大字典上的稀疏分解来识别说话人。据报道,这些方法具有比传统的 GMM-UBM 和 GMM-SVM<sup>[8]</sup>更好的性能。

尽管目前的方法取得了一定的进展,但仍然存

在两个问题需要进一步解决。首先, MCA 只有在假设成立之时才能取得良好性能,即一个说话人的语音在它自己的字典上是稀疏的,而在其他字典上是稠密的<sup>[7]</sup>。然而,并没有理论或实验上的证据表明这种假设在 GMM 均值超向量域或  $i$ -向量域成立。其次,目前的方法还没有考虑说话人识别的噪声鲁棒性,因而这些方法也无法在噪声环境下取得满意的效果。

为了解决这些问题,提出一种基于 MCA 的说话人识别方法。在这种方法中,信号首先被转换成幅度谱;在这个域中,语音具有稀疏性。此外,字典通过训练的方式获得,与通过样例集合获得字典的方法相比,本文方法能确保语音分解的稀疏性。更重要的是,说话人字典是通过删除与通用背景字典原子相近的原子来进一步优化的,这使得说话人字典具有更强的区分性。此外,噪声字典随输入噪声变化,这使得所提出的方法可以吸收时变噪声。

## 1 形态成分和说话人识别

MCA 最初是被提出用于分离具有不同形态成分的混合信号,目前已经成功用于图像分离。给定由  $K$  个不同的信号  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_K$  混合而成的实数域信号  $\mathbf{Y}$ , 即

$$\mathbf{Y} = \sum_{i=1}^K \mathbf{X}_i \quad (1)$$

存在一个字典  $\Psi = [\Phi_1, \Phi_2, \dots, \Phi_K]$ , 其中  $\Phi_i \in \mathbf{R}^{N \times M_i}$  是信号  $\mathbf{X}_i$  的一个字典,  $K$  是字典的数量,  $M_i$  是  $\Phi_i$  中的原子数量。MCA 假设每个  $\mathbf{X}_i$  在  $\Phi_i$  上稀疏,但是在其他的  $\Phi_j (j \neq i)$  上稠密,对于信号  $\mathbf{Y}$ , MCA 解决如下的优化问题实现稀疏分解

$$\min_{\mathbf{y}} \|\mathbf{y}\|_0 \text{ subject to } \|\mathbf{Y} - \Psi\mathbf{y}\|_2 \leq \epsilon \quad (2)$$

式中:  $\mathbf{y}^T = [x_1^T, x_2^T, \dots, x_K^T]$ ,  $x_i$  为  $\mathbf{Y}$  在  $\Phi_i$  上的表示;  $\epsilon$  为残留误差。在理论上,当 MCA 的假定成立,这一优化任务能成功地分离混合信号,即实现  $\mathbf{X}_i \approx \Phi_i x_i$ <sup>[7]</sup>。

将 MCA 用于说话人识别在理论上是可行的。在说话人识别任务中信号不是混合的,而只是一位说话人的语音,即  $\mathbf{Y} = \mathbf{X}_i$ , 其中  $\mathbf{X}_i$  表示第  $i$  个说话人的语音。如果每个说话人有一个字典  $\Phi_i$ , 在这个字典上该说话人的语音是稀疏的,其他说话人的语音是稠密的,能构造  $\Psi = [\Phi_1, \Phi_2, \dots, \Phi_K]$ 。在理想条件下,求解方程(2)导致  $\mathbf{X}_i \approx \Phi_i x_i$ , 即仅  $\Phi_i$  的原子被使用。因此,可以通过已使用的原子的标

签实现对  $Y$  的分类。在噪声的环境中,可以把噪声看作是一个说话人。满足 MCA 的假设的说话人字典和噪声字典可以通过相应的信号训练得到。因此,在说话人识别任务中应用 MCA 方法是可行的。

从稀疏分解的角度看,噪声会存在两种情况,第一种是噪声本身是稀疏的,即可以找到一个噪声字典稀疏表示噪声信号,那么此时完全可以把噪声看作一个说话人。第二种情况是噪声在噪声字典上不稀疏,虽然此时噪声字典不起作用,但稀疏编码可以有效去噪。

## 2 基于 MCA 的说话人识别方法

从上面的分析中,通过设计一个基于 MCA 的方法得到了良好的性能。在此之前,还有 3 个问题需要考虑:

(1)如何训练一个大字典,使之满足 MCA 的假设。只有当 MCA 的假设在说话人识别中成立时,该方法才能像预期的那样取得好的分类效果。这种假设的基本要求是提高大字典的区分性。

(2)如何应对不同的噪声。噪声在语音中是很难预测的,因为它可能有两种变化。首先是噪声的类型变化,例如噪声从一种类型变成另一种类型。其次,噪声本身可能是时变的。一个固定不变的噪声字典无法刻画变化的噪声,因而影响识别性能。

(3)如何用稀疏表示分类。在理想的条件下,一个说话人的语音只能通过这个说话人的字典来被稀疏表示;然而,在真实的应用环境中,其他说话人的原子也可能被使用,导致错误的分类。

考虑到以上问题,设计了一种新的说话人识别方法(见图 1)。

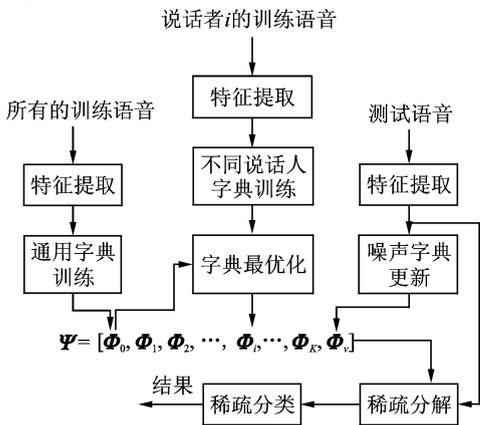


图 1 基于 MCA 的说话人识别步骤

Fig. 1 Procedure of speaker recognition based on MCA

## 2.1 字典准备

### 2.1.1 特征提取

所提出的方法中以幅度谱为输入特征。在特征提取中,语音信号首先被分割成重叠的帧,然后在每一帧上加汉明窗。接下来,对每帧数据作离散傅里叶变换(Discrete Fourier transform, DFT)并计算幅度谱作为输入特征。

### 2.1.2 大字典的结构

设计一种新型结构的大字典,下面是有  $K$  个说话人的识别系统

$$\Psi = [\Phi_0, \Phi_1, \Phi_2, \dots, \Phi_K, \Phi_v] \quad (3)$$

式中:  $\Phi_0$  是一个通用的背景字典,包括所有的说话人的共同的特征。这里借鉴了 GMM-UBM 用 UBM 来为背景建模的方法。  $\Phi_i (i=1, \dots, K)$  用于刻画第  $i$  个说话人的变化性。  $\Phi_v$  用于刻画环境噪声。在  $\Psi$  中的所有原子都被标准化成单位范数向量。

这样的一个大字典结构具有两个显著优势。首先,它使大词典有更强的鉴别能力。其次,每个说话人字典的原子数量可以大幅度减少,这降低了稀疏分解的计算复杂度。

### 2.1.3 字典训练

许多方法被用来训练字典,例如, k-SVD<sup>[11]</sup>, k-means<sup>[12]</sup>。选择 k-SVD 来训练字典。字典训练的问题被描述成

$$\min \|Y - \Phi X\|_2^2 \text{ subject to } \|x_i\|_0 \leq T_0 \quad (4)$$

式中:  $Y = [Y_1, Y_2, \dots, Y_M]$  是训练的数据集,每一个  $Y_i$  都是语音帧的特征向量,  $\Phi$  是字典,  $X = [x_1, x_2, \dots, x_M]$  是一组与  $Y$  相对应的稀疏向量,  $T_0$  是稀疏度的门限。通用的背景字典用大量未标记的不同说话者的语音来训练。每个  $\Phi_i$  都用第  $i$  个说话者的语音来训练,用  $Y$  作初始值。

### 2.1.4 字典的优化

在该方法中,使用  $\Phi_0$  来刻画所有的说话人的共同变化。希望每个说话人的字典只刻画这个说话人和其他说话人之间的不同之处,这样可以提高字典的鉴别能力。为此,提供了一种方法来优化说话人字典(算法 1 如下所示)。

#### 算法 1: 字典优化

Input:  $\Phi_0 = [a_1, a_2, \dots, a_{M_0}]$ ,  $\Phi_i = [b_1, b_2, \dots, b_{M_i}]$ ,  $T$   
 Output:  $\hat{\Phi}_i$   
 initialize:  $w = [1, 1, \dots, 1]$ ,  $\hat{\Phi}_i = []$ ;  
 for  $p=1$  to  $M_i$  do

```

for  $q=1$  to  $M_0$  do
  if ( $|b_p^T a_q| < w[p]$ ) then
     $w[p] = |b_p^T a_q|$ ;
  end if
end for
end for
 $\gamma$  = the Tth smallest element of  $w$ ;
for  $p=1$  to  $M_i$  do
  if  $w[p] < \gamma$  then
     $\hat{\Phi}_i = [\hat{\Phi}_i, b_p]$ ;
  end if
end for

```

当  $|a_i^T b_j|$  接近 1,  $a_i$  和  $b_j$  趋于相同; 如果接近 0, 二者则完全不同的。该算法的主要思想是移除与通用背景字典的原子相近的原子。

## 2.2 识别

### 2.2.1 稀疏分解

稀疏分解通过解式(2)来实现,但这被证明是 NP-难的,不可能通过穷举搜索所有可能的原子集来实现。然而,如果语音是稀疏的或近似稀疏的,那么它可以通过优化式(5)唯一求解<sup>[13]</sup>

$$\mathbf{y} = \arg \min_{\mathbf{y}} \|\mathbf{y}\|_1 + \frac{\lambda}{2} \|\mathbf{Y} - \Psi\mathbf{y}\|_2^2 \quad (5)$$

式中:  $\lambda > 0$  是调节参数。式(5)也称为基追踪去噪法(Basis pursuit de-noising, BPDN),是本文采用的稀疏分解算法。

### 2.2.2 噪声字典更新

由于未知的噪声是时变的,因而不可能用一个固定的字典来刻画。一个可行的方法是用混噪测试语音更新噪声字典。

#### 算法 2: 噪声字典更新

Input:  $\Psi = [\Phi_0, \dots, \Phi_K, \Phi_v]$ , where

$$\Phi_v = [c_1, c_2, \dots, c_{M_v}],$$

$$\gamma = [Y_1, Y_2, \dots, Y_P], \delta$$

Output:  $\Phi_v$

initialize:  $\Gamma = []$ ;  $\zeta = \delta$ ;

while  $\zeta \geq \delta$  do

$\zeta = 0$ ;

$\Gamma = []$ ;

for  $i=1$  to  $P$  do

$\mathbf{y} = \arg \min_{\mathbf{y}} \lambda \|\mathbf{y}\|_0$  subject to  $\|\mathbf{Y}_i - \Psi\mathbf{y}\|_2 \leq \epsilon$ ;

$\zeta = \zeta + \|\mathbf{y}\|_1$ ;

$\Gamma = [\Gamma, (\mathbf{Y}_i - [\Phi_0, \dots, \Phi_K]\mathbf{y}_s)]$ ;

end for

$\Phi_v = \text{K-SVD}(\Gamma, M_v)$ ;

$\zeta = \zeta / P$ ;

end while

更新的方法如算法 2 所示,其中  $\gamma$  是一个被噪声污染的测试语音,  $\delta$  是稀疏分解的误差门限,  $\mathbf{y}_s$  是  $\mathbf{Y}_i$  在  $\Psi = [\Phi_0, \Phi_1, \Phi_2, \dots, \Phi_K]$  上的稀疏表示,  $\Gamma$  用于保存残留噪声样本,  $\text{K-SVD}(\Gamma, M_v)$  表示函数采用 K-SVD 训练算法训练字典,  $M_v$  为字典的原子数,  $\Gamma$  训练数据集。噪声字典以在线的方式训练。噪声语音减去干净语音的残余是训练数据。训练过程重复进行,直到稀疏度收敛为止。

### 2.2.3 稀疏分类

对于一组测试语音帧  $[Y_1, Y_2, \dots, Y_P]$  和其相对的稀疏表示  $[y_1, y_2, \dots, y_P]$ , 根据式(6)实现对测试语音的分类

$$i = \arg \min_{1 \leq i \leq K} \frac{1}{P} \sum_{j=1}^P \|\delta_i(\mathbf{y}_j)\|_1 \quad (6)$$

式中:  $\delta_i(\cdot)$  表示一个实数向量,其中非零项仅仅来自第  $i$  个类别。值得注意的是仅仅在说话人字典上的表示系数被用于范数计算,而噪声字典上的系数则被丢弃。这种方式能有效减小说话人共性和噪声对识别结果的影响。

## 3 实验与分析

### 3.1 实验设置

本文采用“863”中文数据库来测试所提出的方法。该语料库是在国家高新技术项目支持下为开发 ASR 系统而建立,包含 96 269 句 83 名男性和 83 名女性的语音(每个说话人有 520 句)。数据是在无噪环境下通过近距离麦克风录入,采样率 16 kHz,量化位数 16 位。每个说话人的前 10 句语音用于字典训练,接下来的 10 句用于测试。每句测试语音作为一次测试。所有数据以 8 kHz 的采样率进行重采样。用所有训练数据训练一个包含 512 个原子的通用背景字典。用每个说话人的训练数据训练这个人的字典,训练时以 UBD 为初始值,然后通过算法 1 优化。为了获得噪声环境下的测试语音,将 4 种噪声(来自 noisex-92 数据库<sup>[14]</sup>),即 white, f16, babble 和 pink, 分别以 0, 5, 10 和 20 dB 的信噪比人为地加在测试集语音上。这样就有了 4 个噪声版本的测试集。

在预处理中,窗长为 20 ms,窗移为 10 ms。DFT 点的数量是 512,因此每个向量的维数是 257。算法 1 中的参数  $T$  设置为 64,这意味着优化后每个说话人字典中有 64 个原子。噪声字典的原子数量也是 64。式(5)中的  $\lambda$  被设为 0.01,算法 2 中  $\delta$  被设为 7。实验中以 GMM-UBM 方法为基线系统,同时选择 GMM-SVM 和维纳滤波来进行比较。使用的特征是 13 维的 MFCC 系数( $c_0 \sim c_{12}$ )及其一阶和二阶的导数。所有特征都经过倒谱均值和方差的规正处理。采用 EM 算法在所有训练数据集上训练一个混合度为 1 024 的 UBM。每个说话人的 GMM 模型用这个说话人的训练数据来训练,训练算法为最大后验概率自适应。

至于 GMM-SVM,是将每个培训语音转变成 GMM 均值超向量,所有超向量都用来训练 SVM 分类器。另一种用于比较的方法是采用维纳滤波增强语音后提取特征,然后训练模型。实验采用识

别准确率来衡量性能。

### 3.2 结果和分析

结果如表 1 所示,其中 GU 表示 GMM-UBM,GS 表示 GMM-SVM,WF 表示维纳滤波。可以看到,在干净的环境下所有的方法都可以达到满意的效果。GMM-UBM 的性能随着噪声水平的增加而迅速降低。在 10 dB 的白噪声下,其准确率下降到 6.0%。GMM-SVM 的性能比 GMM-UBM 要好,但它对噪声也很敏感。维纳滤波获得的提升是很有限的,在某些情况下,其准确率甚至会下降,如在 Babble 噪声下。尽管语音增强方法可以减少噪声,但它们引起了进一步的畸变,而当前的说话人识别系统对此是非常敏感的。相比之下,本文方法取得了比其他方法更好的性能。当信噪比为 20 dB,精度可以达到高达 90%以上;信噪比下降到 0 dB 时,精度可以保持 60%以上。

表 1 在各种不同噪声环境下的性能比较

Table 1 Performance comparison under various noisy environments

方法	无噪	%															
		White/dB				Pink/dB				F16/dB				Babble/dB			
		0	5	10	20	0	5	10	20	0	5	10	20	0	5	10	20
GU	96.4	6.0	20.5	51.2	78.3	9.0	25.9	49.4	80.1	13.2	31.9	48.8	85.5	21.7	31.9	54.2	87.3
GS	97.8	13.6	37.7	62.8	80.5	19.2	33.4	53.6	83.5	16.6	34.8	51.3	85.0	22.7	36.1	54.8	88.7
WF	94.0	14.5	32.6	55.5	72.3	25.3	27.1	51.2	75.9	26.5	39.8	56.0	82.5	9.6	25.3	60.2	78.3
本文方法	97.6	65.4	74.3	84.3	94.6	70.2	77.1	86.7	93.4	62.4	73.2	88.0	90.3	65.2	69.9	80.7	92.2

## 4 结束语

在现实的应用中,说话人识别系统的性能随信噪比的降低而迅速下降,因此,提高系统的鲁棒性具有重要的作用。目前的研究主要集中在信道失配和场景变化上。虽然目前的文献提出了许多有效的方法;然而,缺乏有效的方法来克服噪声的影响。为了解决这一问题,提出一种基于 MCA 的说话人识别方法。在分析 MCA 的假设基础上,设计了一种字典构造方法使这一假设成立。首先,在频谱域训练字典,这里的语音倾向于稀疏。字典是通过训练的方法而不是收集样例的方法来获取,这进一步确保了语音在字典上稀疏。第二,设计了一种新型的字典结构并通过增强字典的区分性来优化字典。一方面,这种方法使得说话人字典能刻画不同说话人的变化;另一方面,可以大幅度减少原子数,提高系统效率。第三,在大词典中放入噪声字典,并给出了一个根据噪声在线更新噪声字典的算法,这确保了噪声字典可以稀疏表示随时变化的噪

声。最后,在实验中通过在干净语音中人为添加噪声来模拟噪声环境。实验结果表明,该方法对于各种噪声具有鲁棒性。

本方法对环境噪声具有更强的鲁棒性,因此具有更加广泛的实用价值,可用于现实的说话人识别任务。虽然所提出的方法对字典进行了优化,但计算量仍然很大。因此该方法目前不适合用于有实时需求的应用场合。我们将在未来的工作中降低本方法的时间复杂度,提高其运行的效率。

### 参考文献:

- [1] Reynolds D, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal Process, 2000,10:19-41.
- [2] Campbell W M, Sturim D E, Reynolds D A. Support vector machines using GMM supervectors for speaker verification [J]. IEEE Signal Processing Letters, 2006,13:308-311.
- [3] Kenny P, Boulianne G, Ouellet P, et al. Joint factor analysis versus eigenchannels in speaker recognition

- [J]. *IEEE Trans Audio Speech Lang Process*, 2007, 15(4):1435-1447.
- [4] Ming J, Hazen T J, Glass J R, et al. Robust speaker recognition in noisy conditions[J]. *IEEE Trans Audio Speech Lang Process*, 2007, 15(5):1711-1723.
- [5] 王华朋,杨军. 应用似然比框架的法庭说话人识别[J]. *数据采集与处理*, 2013, 28(2):239-243.  
Wang Huapeng, Yang Jun. Application of likelihood ratio for speaker recognition framework[J]. *Journal of Data Acquisition and Processing*, 2013, 28(2):239-243.
- [6] Donoho D L. Compressed sensing[J]. *IEEE Trans on Inf Theory*, 2006, 52(4):1289-1306.
- [7] Bobin J, Starck J-L, Fadili J M, et al. Morphological component analysis: An adaptive thresholding strategy[J]. *IEEE Trans on Image Processing*, 2007, 16(11):2675-2681.
- [8] Naseem I, Togneri R, Bennamoun M. Sparse representation for speaker identification[C]//Proc ICPR, Turkey: Istanbul, 2010:4460-4463.
- [9] Kua J M K, Ambikairajah E, Epps J, et al. Speaker verification using sparse representation classification[C]//Proc ICASSP. Czech, Prague: [s. n.], 2011: 4548-4551.
- [10] Dehak N, Kenny P, Dehak R, et al. Front-end factor analysis for speaker verification[J]. *IEEE Trans Audio Speech Lang*, 2010, 19(4):788-798.
- [11] Aharon M, Elad M, Bruckstein A M. The K-SVD: an algorithm for designing of overcomplete dictionaries for sparse representation[J]. *IEEE Trans on Signal Process*, 2006, 54(11):4311-4322.
- [12] MacQueen J B. Some methods for classification and analysis of multivariate observations [C]//Proc BSMSP. Berkeley: University of California Press, 1967:281-297.
- [13] Chen S, Donoho D, Saunders M. Atomic decomposition by basis pursuit[J]. *SIAM Rev*, 2001, 43(1): 129-159.
- [14] Horman S. Introduction of noisex-92. Available: <http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html>. 1996-08-13

**作者简介:**何勇军(1980-),男,副教授,研究方向:语音识别、模式识别, E-mail: heyongjun@hrbust.edu.cn; 孙广路(1979-),男,教授,研究方向:智能信号处理、信息安全; 付茂国(1988-),男,硕士研究生,研究方向:语音活动检测; 韩纪庆(1964-),男,教授,博士生导师,研究方向:语音识别、说话人识别。

