第 35 卷第 4 期　　声　学　学　报　　Vol. 35, No. 4
2010 年 7 月　　ACTA　ACUSTICA　　Jul., 2010

卷积噪声环境下语音信号鲁棒特征提取*

吕　钊[1,2]　吴小培[1]　张　超[1]　李　密[2]
(1　安徽大学 计算智能与信号处理教育部重点实验室　合肥　230039)
(2　空军第一航空学院 航空电子工程系　信阳　464000)

2009 年 6 月 30 日收到
2009 年 9 月 9 日定稿

摘要　提出了一种基于独立分量分析 (ICA) 的语音信号鲁棒特征提取算法，用以解决在卷积噪声环境下语音信号的训练与识别特征不匹配的问题。该算法通过短时傅里叶变换将带噪语音信号从时域转换到频域后，采用复值 ICA 方法从带噪语音的短时谱中分离出语音信号的短时谱，然后根据所得到的语音信号短时谱计算美尔倒谱系数 (MFCC) 及其一阶差分作为特征参数。在仿真与真实环境下汉语数字语音识别实验中，所提算法相比较传统的 MFCC 其识别正确率分别提升了 34.8% 和 32.6% 。实验结果表明基于 ICA 方法的语音特征在卷积噪声环境下具有良好的鲁棒性。

PACS 数：43.60, 43.72

Robust speech features extraction in convolutional noise environment

LÜ Zhao[1,2]　WU Xiaopei[1]　ZHANG Chao[1]　LI Mi[2]
(1　*The Key Laboratory of Intelligent Computing & Signal Processing, Anhui University*　Hefei　230039)
(2　*The First Aeronautical College of Air-Force*　Xinyang　464000)

Received Jun. 30, 2009
Revised Sept. 9, 2009

Abstract　To resolve the mismatch between training features and testing features in convolutive noise environment, a robust speech features extraction algorithm based on Independent Component Analysis (ICA) is proposed. Noisy speech signals are firstly converted from time-domain to frequency-domain via short time Fourier transform, then a complex ICA algorithm is used to acquire short-time spectrum of speech signal from that of noisy speech signal, furthermore, Mel Frequency Cepstral Coefficients (MFCC) and its 1-order differential coefficients are computed in accordance with the separated speech signals frequency spectrum. Simulation and real environment experiments on different noisy Chinese digit recognition are carried out. The results show that the recognition ratio of the proposed algorithm obtains the relative increasing of 34.8% and 32.6% compared with conventional MFCC, which reveal that the speech features based on ICA have a good robust performance.

引言

语音识别系统在特征级别上的鲁棒性研究已成为语音信号处理领域的研究热点[1−3]，对于语音识别技术走向实用化具有非常重要的意义。目前基于 HMM 的语音识别系统常用美尔频率倒谱系数 (MFCC) 作为特征参数，这是因为 MFCC 特征参数充分考虑了人的听觉特性，而且没有任何前提假设，所以该参数具有良好的识别性能[4] 。但是当语音信号受噪声干扰时，其 MEL 滤波器输出必然包含着大量的噪声频谱，这些噪声频谱将会引起训练特征与测试特征的不匹配，从而使识别器识别错误[5] 。当噪声为加性或乘性噪声时，一些语音特征增强算法[1,6−8] 可以有效地抑制噪声干扰。

然而，在实际使用过程中由于空间物体反射和吸收作用，传感器接收的信号中存在空间内不同物体对声音的反射波，这就使得传感器所采集到的声音信号不再是各声源信号的简单混合，而是多声源信号的时延叠加[9−10] 。因此，在卷积噪声环境下上述去噪算法并不理想。为了提取卷积噪声环境下语

* 国家自然科学基金 (60771033) 和博士点基金 (200803570002) 资助项目。

音信号的鲁棒特征，提高语音识别系统识别率，本文采用独立分量分析 (ICA) 的方法，在频域内对语音信号频谱与噪声信号频谱进行分离，然后根据分离后的语音信号频谱提取语音信号的 MFCC 及其一阶差分作为特征参数，实现了卷积噪声环境下的鲁棒特征提取。

1 鲁棒特征提取

1.1 算法概述

在双入双出 (TITO) 条件下，基于独立分量分析 (ICA) 的卷积噪声环境下语音信号鲁棒特征提取算法的流程如图 1 所示。

从图 1 中可以看到，算法首先对带噪语音信号进行短时短时傅里叶变换，然后通过复值 ICA 方法将纯净语音信号的短时谱与噪声信号的短时谱进行分离，并对分离后的频谱进行频域预加重处理，最后提取 MFCC 及其一阶差分作为特征参数。由于算法没有改变 MFCC 参数所具有的人耳听觉特性，因而保证了该参数的优点。同时，前端采用了 ICA 方法对语音和噪声的短时谱进行了分离，因此可以较大程度的去除噪声对语音信号频谱的干扰，实现了卷积噪声环境下鲁棒特征的提取。

1.2 短时傅里叶变换

为了在某一频点上获取充足的观测数据，同时满足语音信号的短时平稳特性，算法首先对带噪语音信号进行短时傅里叶变换 (STFT)。设在 TITO 条件下，观测信号 $x(t)$ 可以表示为：

$$x_i(t)=\sum_{j=1}^{2}\sum_{k=0}^{P-1}a_{ij}(k)s_j(t-k),\quad i=1,2 \tag{1}$$

其中，P 为混合滤波器阶数，a_{ij} 为第 j 个源到第 i 个传感器的冲激响应。通过对上式进行 L 点短时傅里叶变换，可以得到：

$$X_i(f_l,\tau)=\sum_{t=0}^{L-1}x_i(t)\mathrm{win}(t-\tau)\exp(-\mathrm{j}2\pi f_l t),\ i=1,2 \tag{2}$$

其中，$l=0,\cdots,L$–1; $f_l=(l/L)f_s$ 为对应的 L 个频点，f_s 为采样率，$\mathrm{win}(t)$ 为窗函数，$\tau=\tau_0,\tau_1,\cdots,\tau_{M-1}$ 为每次滑动窗函数开始的时间位置。可以看出，经过短时傅里叶变换后，输入的时域观测信号变为一 $L\times M$ 点的频域观测信号矩阵。

1.3 基于负熵的复 ICA 固定点算法

由于 STFT 的结果是复值的，本文采用了复值负熵 ICA 算法进行噪声和语音信号短时谱的分离。相对于其他类型的复值 ICA 算法，负熵极大 ICA 算法在稳定性和计算复杂度等方面具有一定的优势。

负熵是度量信号非高斯性的一种常用准则，但由于计算负熵时需要事先得到概率密度函数，因而直接使用较为复杂。实际计算时往往使用非多项式函数对负熵进行近似，其计算过程如下：

$$J(x)\propto\left[E\{G(x)\}-E\{G(v)\}\right]^2, \tag{3}$$

式中 v 是一标准的高斯随机变量，因此能够保证当 x 为高斯分布随机变量时，负熵的估值为零。G 函数可参照 x^4 的波形和变化趋势进行选择，即要求所选的也是下凸的偶对称函数，且比 x^4 的增长速度要慢。我们采用梯度算法对式 (3) 定义的目标函数进行优化，算法可简单描述为：

$$\begin{cases} J(\widetilde{\boldsymbol{w}})=\left[E\left(G(\widetilde{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{z})\right)-E\left(G(v)\right)\right]^2, \\ \widetilde{\boldsymbol{w}}(n+1)=\boldsymbol{w}(n)+\mu\nabla_{\widetilde{w}}J, \\ \|\widetilde{\boldsymbol{w}}\|=1, \end{cases} \tag{4}$$

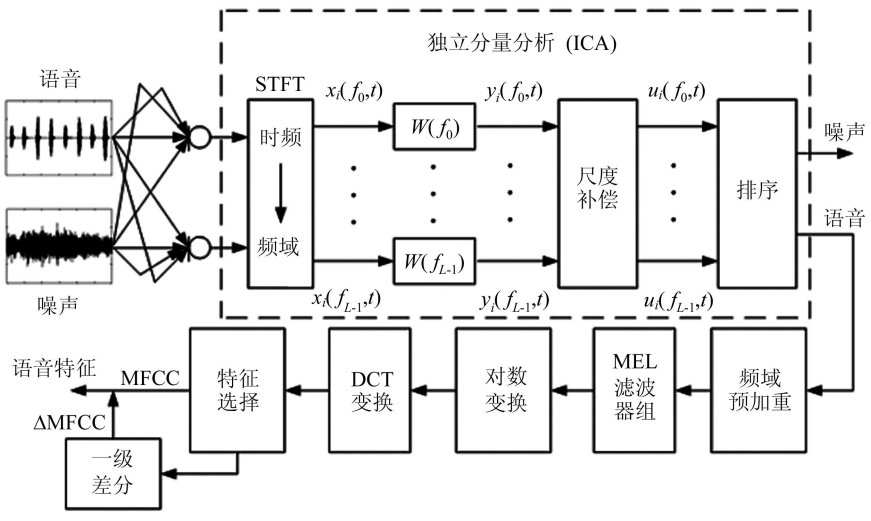


图 1 卷积噪声环境下语音信号鲁棒特征提取算法框图

式 (4) 中：

$$\nabla_{\widetilde{\boldsymbol{w}}} J = \left(E\left\{G(\widetilde{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{z})\right\} - E\left\{G(v)\right\}\right) E\left\{\boldsymbol{z} g(\widetilde{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{z})\right\}. \tag{5}$$

式 (5) 中，$\boldsymbol{z}$ 是对观测信号 x 进行白化预处理后所得到的新的观测信号，$\widetilde{\boldsymbol{w}}$ 是 n 维复向量并且 $E\{(\widetilde{\boldsymbol{w}}^{\mathrm{T}}z)^2\} = 1$, v 为零均值单位方差高斯分布随机变量。$G(\cdot)$ 是一光滑非多项式数，$g(\cdot)$ 是 $G(\cdot)$ 的导数。两种不同的 $G(\cdot)$ 函数可供选择：

$$G_1(y) = \frac{1}{a}\log\cosh(ay),\ g_1(y) = \tanh(ay), \tag{6}$$

$$G_2(y) = -\exp\left(-\frac{y^2}{2}\right),\ g_2(y) = y\exp\left(-\frac{y^2}{2}\right). \tag{7}$$

一般取 $1 \le a \le 2$ 。上述两种方法中，G_1 的普适性较强。G_2 对超高斯信号效果较好并能提供更为稳健的估计。

在复数域内，根据 (4) 、(5) 两式改写复值固定点迭代 ICA 算法如下 [11]:

$$\begin{cases} \widetilde{\boldsymbol{w}} = E\left\{z\left(\widetilde{\boldsymbol{w}}^{\mathrm{T}}z\right)^* g\left(\left|\widetilde{\boldsymbol{w}}^{\mathrm{T}}z\right|^2\right)\right\} - \\ \qquad E\left\{g\left(\left|\widetilde{\boldsymbol{w}}^{\mathrm{T}}z\right|^2\right) + \left|\widetilde{\boldsymbol{w}}^{\mathrm{T}}z\right|^2 g'\left(\left|\widetilde{\boldsymbol{w}}^{\mathrm{T}}z\right|^2\right)\right\}\widetilde{\boldsymbol{w}} \\ \widetilde{\boldsymbol{w}} = \dfrac{\widetilde{\boldsymbol{w}}}{\|\widetilde{\boldsymbol{w}}\|}, \end{cases} \tag{8}$$

式中，$\boldsymbol{z}$ 是经白化处理后的复观测数据，$G(|y|^2)$ 是非线性函数，$g(\cdot)$ 是其对应的一阶导数。

为了验证该算法对卷积混合语音分离的有效性，我们进行了仿真实验，实验结果如图 2 所示。对比分离以后语音信号的频谱与和原始信号的频谱可见，基于负熵的复值固定点算法可以对复信号进行较好地分离。

1.4 尺度和排序不确定问题的解决方法

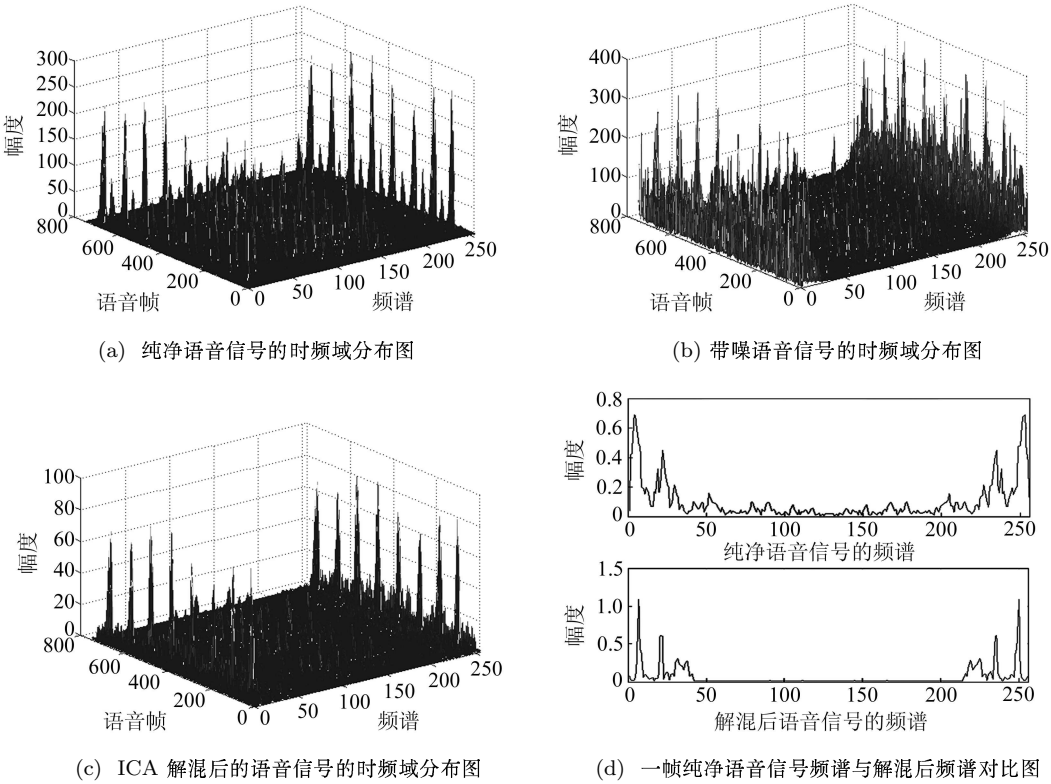
在频域盲解卷积问题中，由于涉及到多个频点的分离，因此需要对各频点的分离结果进行重新组合以便由各频点分离矩阵形成反卷积滤波器，因而必须保证各频点分离信号对应同一源信号，若来自不同源的信号被错误归类组合将对最终分离质量产生巨大影响。另一方面，在各频点混合和解混过程中不同频点的信号可能获得不同的增益，这使得尺度不确定同样成为影响频域 ICA 算法性能的重要因素。

1.4.1 尺度不确定性的补偿

假设 $W(f_k)$ 为所估计的某频点分离矩阵，则在该频点上的混合矩阵可以表示为：

$$A(f_k) = W^{-1}(f_k). \tag{9}$$

利用所得混合矩阵 $A(f_k)$ 系数对各频点独立分量进行补偿，即：



(a) 纯净语音信号的时频域分布图　(b) 带噪语音信号的时频域分布图

(c) ICA 解混后的语音信号的时频域分布图　(d) 一帧纯净语音信号频谱与解混后频谱对比图

图 2 基于负熵的复值固定点 ICA 算法实验结果

$$\begin{bmatrix} V_{1j}(f_k,\tau) \\ \vdots \\ V_{ij}(f_k,\tau) \\ \vdots \\ V_{Nj}(f_k,\tau) \end{bmatrix} = \begin{bmatrix} A_{11}(f_k) & \cdots & A_{1j}(f_k) & \cdots & A_{1N}(f_k) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ A_{N1}(f_k) & \cdots & A_{Nj}(f_k) & \cdots & A_{NN}(f_k) \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ Y_j(f_k,\tau) \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} A_{1j}(f_k)Y_j(f_k,\tau) \\ \\ A_{ij}(f_k)Y_j(f_k,\tau) \\ \\ A_{Nj}(f_k)Y_j(f_k,\tau) \end{bmatrix}, \quad (10)$$

其中 $Y_j(f_k,\tau)$ 表示尺度补偿前已分离的第 j 通道独立分量，$V_{ij}(f_k,\tau)=A_{ij}(f_k)\cdot Y_{ij}(f_k,\tau)$ 则表示经过尺度补偿后，第 j 个独立分量在第 i 个观测信号中的真实成分。由上述分析不难看到，使用式 (10) 对某频点 f_k 的独立分量进行尺度补偿后，一个频域独立分量将产生 N 个补偿后的输出，这 N 个结果经过消除排序模糊、不同频点的组合、逆变换等后续处理，最终可以得到 N 个来自同一声源的纯净信号。它们相当于在 N 不同位置上的拾音器分别获得的来自某一声源的纯净信号。我们可以从中选择一个作为最终的输出，也可以将它们进行适当的平滑处理后作为最终输出。

1.4.2 排序不确定性补偿

了解决频域 ICA 解卷积中出现的排序不确定性问题，文献 12 对相邻频点的分离矩阵进行平滑处理，文献 13 提出利用各频点包络的相关性进行排序，文献 14 提出使用波束形成进行信号到达方向辨识，还可使用对时域分离滤波器长度进行限制等方法。

本文中，我们采用高阶累计量进行排序以降低系统运算量。高阶累计量定义如下：

通过衡量各频点不同通道信号和第一频点某一通道信号的高阶累积量，可判断两信号的同源或非同源特性。其排序函数如下 [15]：

$$\begin{aligned} C(Y_j(f,t)) = {} & E\left[|Y_i(0,t)|^2 |Y_j(f,t)|^2\right] - \\ & E\left[|Y_i(0,t)|^2\right] E\left[|Y_j(f,t)|^2\right] - \\ & -\left|E\left[Y_i(0,t)Y_j(f,t)^{\mathrm{T}}\right]\right|^2 - \\ & |E\left[Y_i(0,t)Y_j(f,t)\right]|^2, \end{aligned} \quad (11)$$

式 (11) 中，若信号 $Y_j(f,t)$ 和 $Y_i(0,t)$ 同源，则 $C(Y_j(f,t))$ 将近似为零，i,j 表示信号通道号。

1.5 频域预加重

频域预加重的作用是对分离后语音信号的高频成分进行提升，使频谱变得平坦，其表达形式为：

$$H(Z)=1-\mu z^{-1}, \quad (12)$$

式中，μ 为预加重系数，取值范围为 $0.9\le\mu\le1.0$。

1.6 获取鲁棒语音特征

为了求取 MFCC 参数，算法接着将预加重过的语音信号频谱通过 MEL 滤波器组并计算每组能量输出并进行对数运算；最后对输出的结果做离散余弦变换 (DCT)，即得到 MFCC 特征参数。

另一方面，为了获取语音信号的动态信息，算法进一步将 MFCC 参数做一阶差分。假设当前所获得的特征倒谱参数是 R 维，那么一阶差分特征的计算如式 (13) 所示 [16]。当 i 从 1 到 $R-1$(分析阶数) 时

$$d_{\mathrm{cep}}(i,t)=\alpha\sum_{k=1}^{K}k\left[C_{\mathrm{mfcc}}(i,t-k)-C_{\mathrm{mfcc}}(i,t+k)\right], \quad (13)$$

式中，d_{cep} 表示动态特征，C_{mfcc} 表示信号倒谱，K 是求差分的帧的范围，α 为因子，用来换算这些特征。经过差分后得到 R 维 MFCC 特征参数，最后将其与 MFCC 参数组合成混合特征矢量作为语音信号特征参数。

2 实验和结果

实验中所使用的语音数据库采样率为 8 kHz、16 bit 量化，录制时间为 12 s(96000 个样本点)。训练语音包含来自 20 个讲话者共 1000 个数据，其内容为 0~9 共 10 个阿拉伯数字，测试语音为 11 个讲话者共 660 个数据。实验采用自左向右连续 HMM 模型，其状态数为 6，高斯混合数是 3，最大迭代次数为 40 次，结束迭代概率门限为 5e−6。为提取特征参数，实验首先对语音信号作预加重和加 Hamming 窗处理，其中窗长为 32 ms，预加重系数为 0.97，窗折叠率为 50%。从每帧语音信号中提 12 维 MFCC 和 12 维 MFCC，共 24 维系数作为特征参数。实验噪声来源于 NOISEX-92 数据库，噪声数据的采样率量化比特与纯净语音一致。实验所添加的噪声依次为：白噪声 (white)，说话人噪声 (babble)，高频噪声 (HF)，喷气式发动机座舱噪声 1(buc1)，喷气式发动机座舱噪声 2(buc2)，驱逐舰机房噪声 (desops)，驱逐舰工作房背景噪声 (desengine) 及坦克噪声 (m109)。

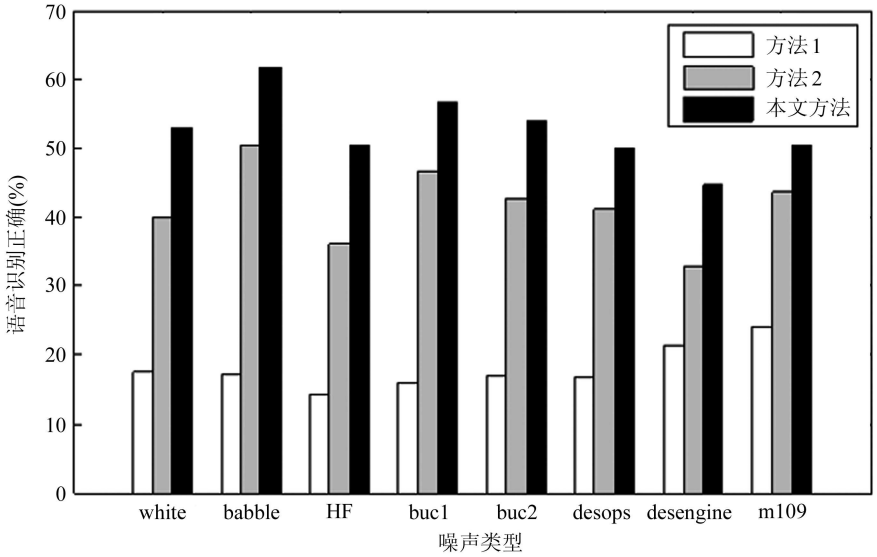


图 3　仿真条件下不同噪声的语音识别结果

2.1　仿真条件下语音识别实验

纯净语音与噪声通过 8 阶混合滤波器进行仿真混合，混合滤波器为：

$$H = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix},$$

其中：

$$h_{11} = [0.10.55 - 0.430.730.26 - 0.380.120.75],$$
$$h_{12} = [0.43 - 0.260.880.030.630.460.22 - 0.11],$$
$$h_{21} = [-0.280.140.54 - 0.340.190.250.620.48],$$
$$h_{22} = [0.410.120.36 - 0.870.710.95 - 0.330.44].$$

不同噪声环境下的语音识别结果如图 3 所示。

图 3 中，“方法 1” 表示对混合后的语音信号未做任何处理，直接提取 MFCC 后送入识别器进行识别的结果；“方法 2” 表示把各频点分离后经过排序和尺度补偿的信号进行逆短时傅里叶变换 (ISTFT)，并按照前期加窗短时傅里叶变换时的逆过程进行截取组合，最终形成对原时域信号的估计，并对这个恢复时域的信号提取 MFCC 后送入语音识别器进行识别；“本文方法” 表示文章所提出的基于独立分量分析的特征域增强后的识别结果。

从实验结果中可以看出 “方法 1” 的平均识别正确率仅为 17.8%，这是因为语音信号的频谱被噪声所干扰，从而引起测试语音与训练语音特征间的失配，最终导致识别结果的不正确。“方法 2” 可以使得平均识别正确率有较大改善 (识别正确率达到 41.7%)，但是由于在恢复时域信号时需要对相邻帧之间重叠部分的信号进行平均，所以必然会带来计算误差，这种误差将限制了识别正确率的进一步提升。而本文所提算法，由于直接利用分离后语音信号的短时谱提取特征参数，避免了恢复时域时所带来的误差，因此识别正确率为 52.6%，相比较 “方法 1” 提升了 34.8%，“方法 2” 提升了 10.9% 。

2.2　真实环境下语音识别实验

实验是在一个长度为 5.1 m，宽度为 3.0 m，高度为 3.6 m 的房间内进行。声源高度约为 1.5 m，为避免 8 kHz 采样率条件下出现空间混叠现象，传感器高度设为 1.3 m，间距为 4.1 cm，说话人在语音传感器正前方约 60 cm 的距离。不同噪声环境下的语音识别结果见图 5 。实验中语音传感器与播放背景噪声的扬声器具体位置见图 4 所示。

从实验结果中可以看出，相比较仿真混合环境下的语音识别，“方法 2” 与本文所提算法的正确率均有所下降，这是因为在真实环境下，由于房间内障碍物的反射，使得传感器所采集到的语音不再是各声源的简单混合，而是多声源信号时延的叠加，同时再加上传感器产生的信道畸变及录音设备自身所产生的干扰，使得真实环境下的混合滤波器比仿真条件下的混合滤波器要复杂得多，不利于 ICA 算法对语音信号的短时谱进行分离，所以识别正确率必然有所下降。然而，从真实环境下语音识别的实验结果依然可以看出，本文所提出算法的识别率相比较 “方法 1” 提升了 32.6%，“方法 2” 提升了 10.5%，结果证明了所提算法的有效性。

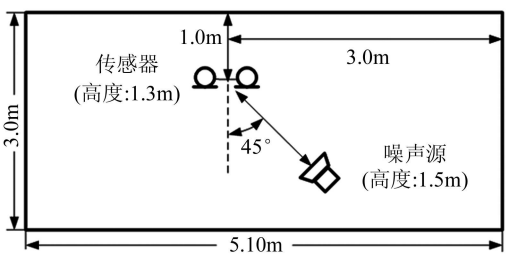


图 4　实验中噪声源与语音传感器的分布图

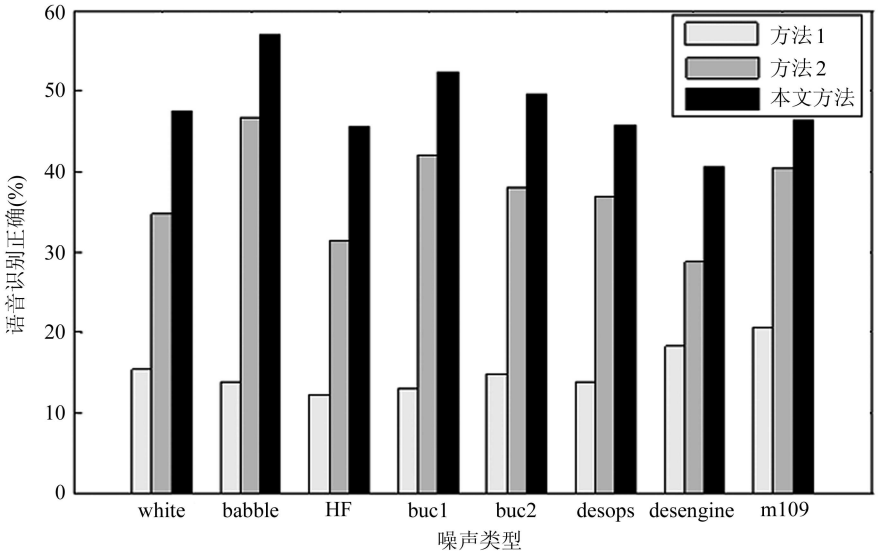


图 5 真实环境下不同噪声的语音识别结果

3 结论

提出一种语音信号鲁棒特征的提取算法，利用独立分量分析的方法解决了卷积噪声环境下训练特征与测试特征不匹配的问题。该算法使用 ICA 方法实现了频率域内语音信号的短时谱与噪声短时谱的分离，并根据所提取出的语音信号短时谱计算鲁棒特征参数，实验结果表明所提算法在卷积噪声环境下能够有效地抑制噪声干扰并提取出正确的语音信号特征，相比较传统的 MFCC 提取算法，语音识别正确率上有了较大程度的提升。另外，由于在实验中存在一些未知因素的干扰，所以算法的稳定性有待进一步提高；同时考虑到实际噪声中不仅含有卷积噪声，还包括传感器的信道畸变、加性或乘性噪声，为了进一步提高语音识别系统的实际应用能力，与其它噪声抑制方法相结合是以后需要深入研究的内容。

参 考 文 献

1 LI Deng, Jasha Droppo, Alex Acero. Log-domain speech feature enhancement using sequential MAP noise estimation and a phase-sensitive model of the acoustic environment. In I CSLP, 2002

2 Ikbal, Misra S, Bourlard H, H. Phase autocorrelation (PAC) derived robust speech features. In: Proc. ICASSP, 2003(2): 133-136

3 Bhiksha Raj, Michael L. Seltzer, Richard M. Stern. Reconstruction of missing features for robust speech recognition. *Speech Communication*, Issue 4, 2004(3): 275—296

4 ZHEN Bin, WU Xihong. On the importance of components of the MFCC in speech and speaker recognition. Acta Scientiarum Naturalium Universitatis Pekinensis, 2001; **37**(3): 371—378

5 Han J, Han M, Park G B. Relative mel-frequency cepstral coefficients compensation for robust telephone speech recognition. In: Proc. European Conf. on Speech Communication and Technology, 1997(3): 1531—1543

6 Alejandro Acero. Acoustical and environmental robustness in automatic speech recognition. PhD thesis, Carnegie Mellon University, 1990

7 王智国，吴及，戴礼荣等. 一种对加性噪声和信道函数联合补偿的模型估计方法. 声学学报， 2008; **33**(3): 238—243

8 Atal B. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of Acoustic Society of America*, 1974; **55**: 1304—1312

9 张华，冯大政，庞继勇. 卷积混迭语音信号的联合块对角化盲分离方法. 声学学报， 2009; **34**(2): 168—174

10 徐舜，陈绍荣，刘郁林. 基于非线性时频掩蔽的语音盲分离方法. 声学学报， 2007; **32**(4): 375—381

11 Bingham E, Hyvarinen A. A fast fixed-point algorithm for independent component analysis of complex valued signals. *Int. J.Neural Systems*, 2000; **10**: 1—8

12 Smaragdis P. Blind separation of convolved mixtures in the frequency domain. *Neurocomput.*, 1998; **22**: 21—34

13 Ikeda S, Murata N. A method of ICA in time-frequency domain. In: Proc. ICA99, 1999: 365—370

14 Ikram M Z, Morgan D R. A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation. In: Proc. ICASSP, 2002: 881—884

15 杨福生，洪波. 独立分量分析的原理与应用. 北京：清华大学出版社， 2006

16 Yan Y H. Development of an approach to language identification based on language-dependent phone recognition: (Ph. D. Thesis). Department of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, 1995