DOI：10.7652/xjtuxb201606007

用于孤立数字语音识别的一种组合降维方法

宋青松，田正鑫，孙文磊，吴小杰，安毅生
（长安大学信息工程学院，710064，西安）

摘要：针对孤立数字语音识别的噪声鲁棒性问题，提出了一个组合降维方法。该方法由梅尔频率倒谱系数(MFCC)特征提取、线性降维、受限玻尔兹曼机(RBM)、Softmax分类器4个功能模块依次组成；基于主成分分析(PCA)基本原理对MFCC特征向量实现了降维并且统一维度的目的；通过RBM对降维后的特征向量进行学习，改善了后端Softmax分类器的分类性能，RBM的预训练由对比散度算法完成，微调过程使用共轭梯度算法。采用TI-46孤立数字语音库和NOISEX-92典型噪声数据库对方法进行了测试，实验结果表明，该方法可以获得96.09%的正确识别率，相对于常规神经网络识别方法，噪声鲁棒性得到了提高。

关键词：语音识别；主成分分析；受限玻尔兹曼机
中图分类号：TP301.6 **文献标志码**：A **文章编号**：0253-987X(2016)06-0042-05

Combined Dimension Reduction Method for Isolated Digital Speech Recognition

SONG Qingsong, TIAN Zhengxin, SUN Wenlei, WU Xiaojie, AN Yisheng
(School of Information Engineering, Chang'an University, Xi'an 710064, China)

Abstract: A combined dimension reduction method is proposed to improve the noise-robustness in isolated digital speech recognition. The method consists of four functional modules in sequence: a Mel frequency cepstrum coefficient (MFCC) module for feature extraction, a linear dimension reduction module, a restricted Boltzmann machine (RBM) module, and a Softmax classifier module. The dimension of the MFCC feature vector is reduced and its dimensionality is unified based on the basic principle of the principal component analysis (PCA); the obtained reduced features are learned by RBM in order to improve the classification performance of the end Softmax classifier module. The pretraining of the RBM is completed by the contrastive divergence algorithm and the finetuning process is fulfilled by the conjugate gradient algorithm. The proposed method is verified on the TI-46 isolated digital speech corpus and the NOISEX-92 noise datasets. The experimental results and comparisons with the conventional feedforward neural network methods show that the proposed method achieves at a 96.09% recognition accuracy and obtains improved noise robustness.

Keywords: speech recognition; principal component analysis; restricted Boltzmann machine

孤立数字语音识别有着广阔的研究和应用价值，诸如动态时间规整(dynamic time warping,

收稿日期：2015-11-30。 作者简介：宋青松(1980—)，男，副教授。 基金项目：国家自然科学基金资助项目(61201406)；中国博士后科学基金资助项目(2013M531998)；中央高校基本科研业务费专项资金资助项目(310824162022，310824162021)。
网络出版时间：2016-04-15 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20160415.1612.008.html

DTW)、隐马尔科夫(hidden markov model，HMM)、矢量量化(vector quantization，VQ)、主成分分析(principal component analysis，PCA)、人工神经网络(artificial neural network，ANN)等方法用于求解孤立数字语音识别问题[1-3]。DTW 算法基于动态规划的思想解决发音长短不一的模板匹配问题，但是存在运算量大、识别性能依赖端点检测精度等不足。VQ 算法基于聚类识别，运算量小但是最优码书较难得到。PCA 算法可以实现数据降维，并且能够统一数据维数，但本质上是一种基于最优正交变换的线性降维方法，对于非线性问题难以得到满意的结果。ANN 算法特别是 Hinton 等提出的受限波尔兹曼机(restricted Boltzmann machine，RBM)及其快速学习算法，在模式识别与分类问题中表现出良好的非线性降维与特征表征能力，但是通常需要适当的特征参数提取等预处理手段配合使用[4]。常用的数字语音信号特征通常是高维的，分类前需要对数据进行降维处理。因此，为改善数字语音识别效果，本文基于 PCA 线性降维和 RBM 特征学习基本原理，提出了一种用于孤立数字语音识别的组合降维方法，待分类的数字语音信号依次经过线性降维和 RBM 非线性特征表征处理，最终识别性能得到改善。

首先阐述组合降维识别方法涉及的线性降维、RBM、Softmax 分类器等功能模块，然后给出用于 RBM 预训练和微调的学习算法，最后在 TI-46 数据库和 NOISEX-92 噪声数据集上验证了所提算法的先进性。

1　组合降维识别方法

1.1　功能模块组成

组合降维识别方法由梅尔频率倒谱系数(Mel-frequency cepstral coefficients，MFCC)特征提取、线性降维、RBM、Softmax 分类器 4 个功能模块组成，如图 1 所示。首先提取 MFCC 及其一阶差分作为原始语音信号的特征参数，然后对 MFCC 进行线性降维，再将降维后的特征参数输入 RBM 进行特征学习，学习的结果作为后端 Softmax 分类器模块的输入，Softmax 输出分类结果。

1.1.1　MFCC 特征提取　MFCC[5]是将人耳的听觉特性与语音产生机制相结合的一种特征参数，在语音识别领域具有广泛应用。标准 MFCC 参数只反映语音参数的静态特性，MFCC 差分则反映语音参数的动态特性，在语音特征中加入表征语音动态

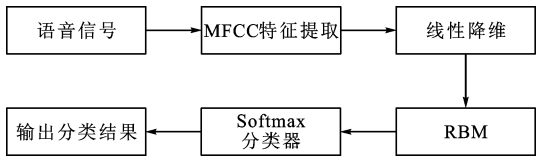


图 1　组合降维语音信号识别方法的功能模块

特性的 MFCC 差分，通常能提高系统的识别性能。因此，本文提取标准 MFCC 及其一阶差分共同作为待识别语音信号的特征参数。MFCC 特征提取的结果是得到一个 F 行 24 列大小的特征向量矩阵 $\boldsymbol{T}$，F 为当前语音信号的帧数。

1.1.2　线性降维　MFCC 特征提取结果存在两个问题：一是每个语音信号由不同数量的帧组成，导致矩阵 $\boldsymbol{T}$ 大小不同；二是 F 取值大导致矩阵 $\boldsymbol{T}$ 过大，存在降维计算需要。因此，基于 PCA 基本原理对特征矩阵 $\boldsymbol{T}$ 作进一步变换，实现其降维并且大小一致的目的。使用的方法是将 $\boldsymbol{T}$ 转置，再与原矩阵 $\boldsymbol{T}$ 相乘，得到 24×24 的方阵 $\boldsymbol{S}$；求 $\boldsymbol{S}$ 的特征值并从大到小排序，取前两个特征值对应的特征向量并串接，得到一个 48 维的特征向量，作为线性降维后当前语音信号的特征向量。

1.1.3　受限波尔兹曼机　降维后的特征向量输入 RBM 模块进行特征学习，学习结果输出到后端 Softmax 分类器中。

RBM 本质上是通过无监督学习最大可能地对输入数据进行特征表征。RBM 由可见层和隐含层构成，如图 2 所示。可见层由一组可见单元 v 构成，用于输入数据；隐含层由另一组隐藏单元 h 构成，用于输出无监督学习获得的对输入数据的特征表示。RBM 的特点是层内无连接，层间全连接。

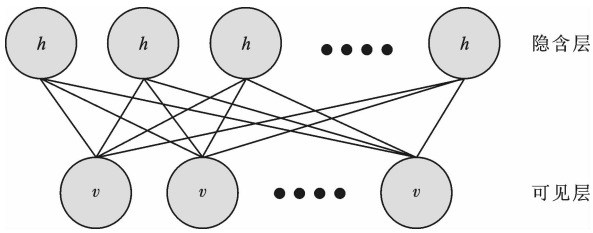


图 2　RBM 结构示意图[6]

1.1.4　Softmax 分类器　采用 Softmax 分类器实现 RBM 输出特征分类。记类标 y 可以取 r 个不同的值，对于训练集 $\{(\boldsymbol{x}^{(1)},y^{(1)}),\cdots,(\boldsymbol{x}^{(m)},y^{(m)})\}$，类标签为 $y^{(n)}\in\{1,2,\cdots,r\}$，$r$ 为分类数。对于给定的输入 $\boldsymbol{x}^{(n)}$，用假设函数 $h_\lambda(\boldsymbol{x}^{(n)})$ 针对每一个类 k 估算出概率值 $p(y^{(n)}=k\mid\boldsymbol{x}^{(n)})$，$k=1,\cdots,r$。$h_\lambda(\boldsymbol{x}^{(n)})$ 输出一个 r 维的列向量(和为 1)，每行表示为当前类

的概率。

定义假设函数 $h_{\lambda}(\boldsymbol{x}^{(n)})$[7] 为

$$h_{\lambda}(\boldsymbol{x}^{(n)}) = \begin{bmatrix} p(y^{(n)}=1 \mid \boldsymbol{x}^{(n)};\boldsymbol{\lambda}) \\ p(y^{(n)}=2 \mid \boldsymbol{x}^{(n)};\boldsymbol{\lambda}) \\ \vdots \\ p(y^{(n)}=r \mid \boldsymbol{x}^{(n)};\boldsymbol{\lambda}) \end{bmatrix} = \left(\sum_{k=1}^{r} \mathrm{e}^{\boldsymbol{\lambda}_k^{\mathrm{T}}\boldsymbol{x}^{(n)}}\right)^{-1} \begin{bmatrix} \mathrm{e}^{\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{x}^{(n)}} \\ \mathrm{e}^{\boldsymbol{\lambda}_2^{\mathrm{T}}\boldsymbol{x}^{(n)}} \\ \vdots \\ \mathrm{e}^{\boldsymbol{\lambda}_r^{\mathrm{T}}\boldsymbol{x}^{(n)}} \end{bmatrix} \tag{1}$$

式中：$\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2,\cdots,\boldsymbol{\lambda}_r$ 是模型参数。将 $\boldsymbol{x}^{(n)}$ 分为第 k 类的概率记为

$$p(y^{(n)}=k \mid \boldsymbol{x}^{(n)};\boldsymbol{\lambda}) = \mathrm{e}^{\boldsymbol{\lambda}_k^{\mathrm{T}}\boldsymbol{x}^{(n)}} / \sum_{l=1}^{r} \mathrm{e}^{\boldsymbol{\lambda}_l^{\mathrm{T}}\boldsymbol{x}^{(n)}} \tag{2}$$

对于样本 $\boldsymbol{x}^{(n)}$，选择概率 $p(y^{(n)}=k \mid \boldsymbol{x}^{(n)};\boldsymbol{\lambda})$ 值最大的对应的类别 k 作为当前样本的分类标签，并与样本本身的标签做比对，如果一致则分类正确，否则分类错误。

1.2 学习算法

组合降维识别方法的学习分为 RBM 预训练和微调两部分。

1.2.1 RBM 预训练 预训练的目的是对线性降维后的特征向量作无监督学习，以获取更好的特征表征。鉴于可见层节点语音特征向量服从高斯分布的特点，使用高斯-伯努利 RBM，定义能量函数[6]

$$E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{M}(v_i-a_i)^2 - \sum_{i=1}^{M}\sum_{j=1}^{N} v_i w_{ij} h_j - \sum_{j=1}^{N} h_j b_j \tag{3}$$

式中：$\boldsymbol{\theta}=\{a_i,b_j,w_{ij}\}$ 是 RBM 模型参数；a_i 和 b_j 分别是可见层节点 i 和隐含层节点 j 的偏置；w_{ij} 是可见层节点 i 和隐含层节点 j 之间的连接权值。当参数确定时，可以得到联合概率分布

$$P(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}) = \exp(-E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}))/Z \tag{4}$$

其中 $Z = \sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}}\exp(-E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}))$ 称为配分函数。模型关于可见层节点的边缘概率分布为

$$P(\boldsymbol{v};\boldsymbol{\theta}) = \sum_{\boldsymbol{h}}\exp(-E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}))/Z \tag{5}$$

RBM 的模型参数使用最大似然准则通过无监督训练得到，训练的目标函数为

$$\boldsymbol{\theta}^{*} = \arg\max_{\boldsymbol{\theta}} \log P(\boldsymbol{v};\boldsymbol{\theta}) \tag{6}$$

对目标函数求偏导，可以得到权值的更新公式

$$\Delta w_{ij} = E_{\mathrm{data}}(v_i h_j) - E_{\mathrm{model}}(v_i h_j) \tag{7}$$

式中：$E_{\mathrm{data}}(v_i h_j)$ 是训练集数据对应的可见层和隐含层状态的期望值；$E_{\mathrm{model}}(v_i h_j)$ 是对所有可能的 $(\boldsymbol{v},\boldsymbol{h})$ 的模型期望值。

$E_{\mathrm{model}}(v_i h_j)$ 直接计算很困难，通常采用对比散度进行近似计算[4]。可见层单元的状态被设置为任取一个训练样本，算法开始，通过一步吉布斯采样获得“重构”的可见单元状态 $\langle v_i\rangle_{\mathrm{recon}}$，再用 $\langle v_i\rangle_{\mathrm{recon}}$ 更新隐含层单元状态，得到 $\langle h_j\rangle_{\mathrm{recon}}$。学习率 ε 大使收敛速度快，但过大会引起算法不稳定，ε 小可消除不稳定，但会减慢收敛速度，为克服该矛盾，在更新参数时增加动量项 c，使得本次参数修改的方向由上一次参数修改方向和本次的梯度方向一起决定，而不是完全由当前样本下的似然函数梯度方向决定。因此，各参数的更新准则为

$$\Delta w_{ij} = c\Delta w_{ij} + \varepsilon(\langle v_i h_j\rangle_{\mathrm{data}} - \langle v_i h_j\rangle_{\mathrm{recon}}) \tag{8}$$

$$\Delta b_i = c\Delta b_i + \varepsilon(\langle v_i\rangle_{\mathrm{data}} - \langle v_i\rangle_{\mathrm{recon}}) \tag{9}$$

$$\Delta a_j = c\Delta a_j + \varepsilon(\langle h_j\rangle_{\mathrm{data}} - \langle h_j\rangle_{\mathrm{recon}}) \tag{10}$$

使用重构误差对 RBM 进行评估。重构误差就是以训练数据作为初始状态，根据 RBM 的分布进行一次吉布斯采样所获得的重构样本与原始样本的差异。

1.2.2 微调 RBM 预训练完成之后，对 RBM 和 Softmax 进行微调。为改善学习效率，在微调开始的前 5 次，只对 Softmax 分类器的模型参数进行有监督学习，从第 6 次开始对 RBM 和 Softmax 的全部参数进行学习。

代价函数定义为

$$J(\boldsymbol{\lambda}) = -\left[\sum_{n=1}^{m}\sum_{k=1}^{r} 1\{y^{(n)}=k\}\log\left(\mathrm{e}^{\boldsymbol{\lambda}_k^{\mathrm{T}}\boldsymbol{x}^{(n)}} \Big/ \sum_{l=1}^{r}\mathrm{e}^{\boldsymbol{\lambda}_l^{\mathrm{T}}\boldsymbol{x}^{(n)}}\right)\right] \tag{11}$$

式中：$1\{\cdot\}$ 是一个指示性函数，当 $\{\cdot\}$ 中的值为真时，该函数值为 1，否则为 0。采用 PRP 共轭梯度算法求解 $\min J(\boldsymbol{\lambda})$ 无约束最优化问题[8]。

微调结束后得到 RBM 和 Softmax 最终的模型参数。给定任意的孤立数字语音信号，依次通过图 1 所示的各个功能模块，可以输出分类结果。

2 实验设计与结果分析

2.1 实验设计

组合降维识别方法的性能测试在 TI-46 数字语音数据库上进行，语音信号的采样频率为 12.5 kHz，16 b 量化。选择 3 000 个样本作为训练集，0～9 共 10 个数字各 300 个样本，选择另外的 1 000 个样本

作为测试集，每个数字各 100 个[9]。

MFCC 特征提取模块中帧长取 256，帧移为 80，窗函数使用汉明窗。RBM 预训练过程中，可见层输入数据归一化到(0，1)之间，连接权重初始化为正态分布 $N(0,0.01)$随机数，可见层和隐含层的偏置均初始化为 0。将数据集分成小批量进行预训练，每个批量为 50 个。学习率 ε 为 0.001，最大训练次数为 50 次，动量项 c 在前 5 次训练中取 0.5，之后取 0.9。微调过程 PRP 共轭梯度算法中线性搜索步长为 3，微调次数为 200 次。

计算机配置为内存 4 GB、双核 i5、处理器 2.67 GHz、GPU 为 NVIDIA GT540。

设计一个 3 层前馈神经网络(feedforward neural network，FNN)取代图 1 中 RBM 和 Softmax 分类器两个功能模块，采用相同的训练集和测试集，相同的 MFCC 特征提取模块和线性降维模块，训练采用经典的误差反向传播算法作对比实验。通过交叉验证确定隐层神经元数量为 78 的 FNN 对应最佳识别性能，即 FNN 模型结构取为 48-78-10。记录 FNN 识别结果，与本文方法结果作对比。

2.2　结果分析

本文方法与 FNN 方法各自独立完成 10 次实验，测试结果见表 1。在无噪声情形下，FNN 方法正确识别率平均为 93.07%，而本文方法为 96.09%，优于前者。图 3 给出了无噪声情形下本文方法和 FNN 方法针对 0～9 单个数字语音信号的正确识别率及其标准差，针对数字 0、1、3、5、6、7、8、9，本文方法正确识别率均高于 FNN 方法，而且正确识别率的标准差均小于 FNN 的，表明无噪声情形下本文方法与 FNN 方法相比，不仅正确识别率高而且性能更加平稳。

对测试集以 20 dB 的信噪比(signal-noise ratio，SNR)分别加入白噪声、汽车噪声、工厂噪声及 F16 机舱噪声等 4 类典型噪声用于评价方法的噪声鲁棒性[10]，结果见表 1，FNN 方法 4 类噪声情形下正确识别率的平均结果由 93.07%降低为 91.44%，降低了 1.63%，而本文方法的正确识别率从 96.09%降低为 95.08%，降低了 1.01%，小于前者，表明有噪声情形下本文方法性能下降慢于 FNN，并且降低后本文方法的正确识别率为 95.08%，仍然高于 FNN 的 91.44%。

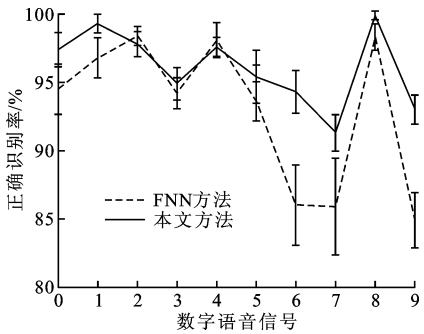
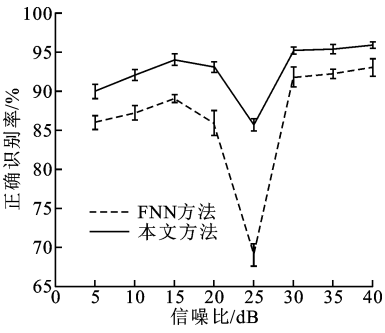


图 3　本文方法与 FNN 方法针对 10 个孤立数字语音信号的性能测试结果

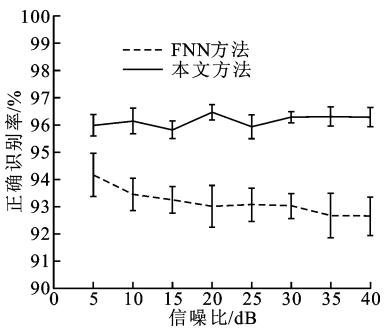
表 1　本文方法与 FNN 方法正确识别率测试结果

噪声类型	正确识别率/%	
	FNN 方法	本文方法
无噪声	93.07	96.09
白噪声	85.98	93.30
汽车噪声	93.02	96.40
工厂噪声	93.73	96.31
F16 机舱噪声	93.02	94.32
平均值	91.44	95.08

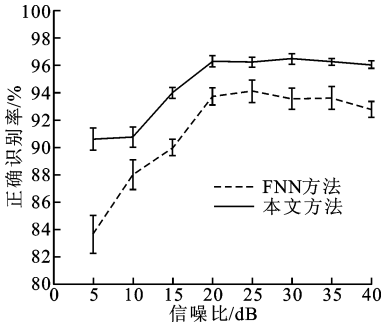
图 4 给出了 5～40 dB 信噪比范围内本文方法和 FNN 方法在上述 4 类典型噪声情形下正确识别率的测试结果。如图 4a～图 4c 所示，白噪声、汽车噪声、工厂噪声 3 种情形下，本文方法的正确识别率均高于 FNN 方法，而且前者的正确识别率标准差比后者要小，说明本文方法的性能更加平稳。图 4d 表明 F16 机舱噪声情形下两种方法在 10～20 dB 范



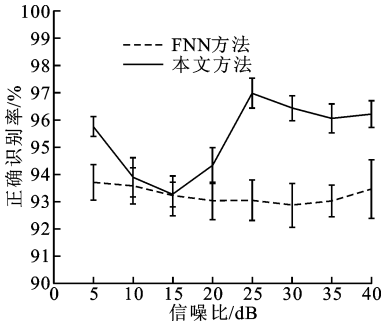
(a)白噪声情形



(b)汽车噪声情形



(c)工厂噪声情形



(d)F16 机舱噪声情形

图 4 典型噪声情形下本文方法与 FNN 方法性能测试结果

围内的正确识别率无明显差别,但是本文方法获取的正确识别率标准差更小,性能更平稳。

上述实验结果表明,针对孤立数字语音识别问题,在有、无噪声两种情形下,本文方法均能够获得优于 FNN 方法的正确识别率,具有一定的噪声鲁棒性,并且性能平稳。

3 结 论

针对孤立数字语音识别问题,基于 PCA 线性降维和 RBM 特征学习基本原理,提出一种组合降维语音识别方法。该方法具有 MFCC 特征提取、线性降维、RBM 特征自动表征等方法的综合优势,特别地,基于 PCA 基本原理对 MFCC 特征向量实现了降维并且统一维度的目的,通过 RBM 非线性特征学习,改善了后端 Softmax 分类器的分类性能。基于 TI-46 孤立数字语音库和 NOISEX-92 典型噪声数据库的测试结果表明,本文方法能够获得优于常规前馈神经网络的正确识别率,并且识别性能更平稳,具有改善的噪声鲁棒性。

参考文献:

[1] SCHAFER P B, JIN D Z. Noise-robust speech recognition through auditory feature detection and spike sequence decoding [J]. Neural Computation, 2014, 26(3): 523-556.

[2] SLOIN A, BURSHTEIN D. Support vector machine training for improved hidden Markov modeling [J]. IEEE Transactions on Signal Processing, 2008, 56(1): 172-188.

[3] TAKIGUCHI T, ARIKI Y. PCA-based speech enhancement for distorted speech recognition [J]. Journal of Multimedia, 2007, 2(5): 13-18.

[4] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.

[5] FANG Z, ZHANG G, SONG Z. Comparison of different implementations of MFCC [J]. Journal of Computer Science and Technology, 2001, 16(6): 582-589.

[6] 张春霞, 姬楠楠, 王冠伟. 受限波尔兹曼机 [J]. 工程数学学报, 2015(2): 159-173.
ZHANG Chunxia, JI Nannan, WANG Guanwei. Restricted Boltzmann machines [J]. Chinese Journal of Engineering Mathematics, 2015(2): 159-173.

[7] SALAKHUTDINOV R, HINTON G E. Replicated Softmax: an undirected topic model [C]// Proceedings of the Advances in Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2009: 1607-1614.

[8] 黄海, 林穗华. 一个 PRP 型共轭梯度法的收敛性 [J]. 西南大学学报: 自然科学版, 2012, 34(3): 28-31.
HUANG Hai, LIN Suihua. Convergence of a PRP type conjugate gradient method [J]. Journal of Southwest University: Natural Science Edition, 2012, 34(3): 28-31.

[9] DODDINGTON G R, SCHALK T B. Speech recognition: turning theory to practice [J]. IEEE Spectrum, 1981, 18(9): 26-32.

[10] VARGA A, STEENEKEN H J M. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems [J]. Speech Communication, 1993, 12(3): 247-251.

(编辑 武红江)