基于 AutoEncoder DBN-VQ 的说话人识别系统

刘俊坤,李燕萍,凌云志
(南京邮电大学 通信与信息工程学院,江苏 南京 210003)

摘 要: 基于矢量量化的说话人识别算法,通过描述说话人语音特征的不同分布进行说话人识别。在说话人数量较多,训练语音时长较短时,系统识别率不高。模型训练一般在纯净语音条件下进行,在实际有噪声环境下进行识别时,系统性能会急剧恶化。为改善系统识别性能,提出一种基于自动编码深度置信网络与矢量量化结合的说话人识别方法。该方法采用深度置信网络对说话人语音数据进行学习和挖掘,在语音时长较短时可以更好地捕获说话人的个性特征;同时采用自动编码器有去噪声的特点,构造自动编码深度置信网络,使网络模型可以对有噪语音数据进行有效地噪声过滤。实验结果证明,该方法在说话人训练语音时长有限时,以及对说话人有噪语音进行识别时,系统识别率都有很大提升。

关键词: 说话人识别;深度置信网络;自动编码器;矢量量化
中图分类号: TP302 **文献标识码:** A **文章编号:** 1673-629X(2018)02-0045-05
doi:10.3969/j.issn.1673-629X.2018.02.011

Speaker Recognition System Based on AutoEncoder Deep Belief Network and Vector Quantization

LIU Jun-kun,LI Yan-ping,LING Yun-zhi
(School of Communications & Information Engineering,Nanjing University of Posts and Telecommunications,Nanjing 210003,China)

Abstract: The speaker recognition system using vector quantization works by describing the different characteristics of the speaker's speech features.When the number of speakers are large and training speech length is short,the recognition rate of the system is not high.For the model is usually trained under the condition of pure speech,the performance of the system will be poor when it is used in the actual environment.In order to improve the recognition performance of the system,we propose a method of speech recognition based on the combination of AutoEncoder deep belief network and vector quantization.It adopts the deep belief network to model and learn for speech data,so speaker's personality characteristics in speech can be better captured when the speech length is short.In the meantime,it structures AutoEncoder deep belief network,which is effective on noise filtering for noisy speech data.The experiment show that the proposed method can improve the recognition rate greatly when there is only a small amount of speaker training data and speech is noisy.

Key words: speaker recognition; deep belief network; AutoEncoder; vector quantization

0 引 言

说话人识别(speaker recognition,SR),又称话者识别[1],是利用说话人语音中的个性特征进行身份鉴定的一种认证技术。基于矢量量化(vector quantization,VQ)的说话人识别模型[2-3]是基于不同说话人的语音特征矢量具有不同分布这一假设,然后采用最小化失真原则对不同说话人特征矢量进行编码识别。该算法直接采用语音的梅尔倒谱参数(Mel frequency cepstral coefficients,MFCC)作为模型训练或识别的特征参数。实际应用时该方法存在两方面的问题:一方面,在说话人数量较多,且每个说话人语音数据较少时,该模型对说话人个性特征得不到充分学习,导致系统的正确识别率达不到期望值;另一方面,系统的模型训练一般是在干净无噪语音条件下,采用说话人有噪语音或是应用在有噪条件下进行识别时,会出现模型训练数据和测试数据不匹配现象,从而系统的识别结果会受到很大影响或者识别结果直接崩溃。

2006 年,Hinton 等[4]提出深度学习的概念,深度

收稿日期: 2017-03-31 **修回日期:** 2017-07-11 **网络出版时间:** 2017-11-15
基金项目: 国家自然科学基金(61401227);江苏省博士后基金(1402067B)
作者简介: 刘俊坤(1991-),男,硕士研究生,研究方向为说话人识别;李燕萍,博士,副教授,研究生导师,通讯作者,研究方向为语音转换和说话人识别。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20171115.1438.080.html

置信网络(deep belief network,DBN)是由多层受限玻尔兹曼机(restricted Boltzmann machine,RBM)堆叠构成的多层深度网络。DBN 网络采用贪婪逐层训练学习算法,通过逐层预训练和整体微调的方法,可以从少量数据中充分学习数据中的潜在特征,挖掘数据中深层表示,并且克服了传统多层神经网络易陷入局部最优解、需要大量数据标记等问题。深度置信网络被证明对自然界中的实际信号建模,比传统浅层结构的建模方法强[5],可以更好地对实际信号进行建模学习。1986 年,Rumelhart 提出自动编码器的概念[6],自动编码器采用这样一个思想:原始输入 x 经过加权(W、b)、映射(Sigmoid)之后得到 y,再对 y 反向加权映射回来成为 z。通过反复迭代训练两组加权系数(W、b),使得误差函数最小,尽可能保证 z 近似于 x,即实现重构 x。自动编码器可以获得代表良好输入的特征,可以实现数据的编码重构,并且训练完成的模型对输入带噪数据具有噪声过滤能力。基于自动编码器的这种优势,文中构造自动编码深度置信网络[7],利用其对不同说话人语音特征编码[8],使网络模型对说话人个性特征进行深度学习和挖掘,然后通过网络模型实现数据重构,在对输入带噪语音提取说话人有效个性特征的同时,有效地过滤噪声。

在基于矢量量化的传统说话人识别方法的基础上,采用自动编码深度置信网络(AutoEncoder DBN)与矢量量化结合的说话人识别方法(AutoEncoder DBN-VQ)。采用 AutoEncoder DBN 对说话人语音特征进行特征编码与重构,将网络输出作为 VQ 的模型训练或识别的输入。结合深度置信网络和自动编码器的优势,AutoEncoder DBN 具备对少量说话人个性特征数据进行深度学习和挖掘,进而提取有效个性特征信息的能力,同时通过模型重构可以过滤说话人语音中的干扰噪声数据。

1 基于矢量量化的说话人识别系统

VQ 是很重要的信号处理方法,具有运算量少,速度快,原理简单等优点,广泛应用于图像和语音等领域。VQ 的原理是把输入的矢量数据空间划分为不同的小区域,每个小区域寻找一个合适的矢量,该矢量用来代表落入到该小区域的所有矢量,用所有的代表矢量即码本来表示整个训练数据。VQ 说话人识别系统在模型训练时将说话人训练语音特征进行聚类,形成码书,每一位说话人对应一个码书。文中采用的码书生成算法是 LGB 算法[9-10],LGB 算法是最常用的也是比较简单的码书生成算法。在识别阶段,采用矢量量化方法计算待识别语音特征与码本之间的失真测度,根据失真测度判定该语音属于哪位说话人。VQ 说话人识别系统中常见的失真测度有欧氏距离、加欧氏距离、Itakura-Saito 距离等,文中采用欧氏距离测度。

基于 VQ 算法的原理,VQ 说话人识别系统存在两个问题:系统采用说话人的 MFCC 特征参数,为不同说话人训练不同的矢量分布,MFCC 参数中包含说话人多种信息,在说话人语音数据有限时,会使量化码本学习不充分,即得到的每个小区域的量化值代表性较弱,影响系统识别准确性;模型训练一般在纯净语音条件下,当待识别语音数据中有噪声时,会因为模型对训练数据和测试数据无法匹配导致系统识别率崩溃。

2 AutoEncoder DBN-VQ 说话人识别系统

2.1 深度置信网络

为了解决 VQ 说话人识别系统在说话人语音数据不足条件下的模型学习不充分、系统识别率下降等问题,采用 DBN 网络对说话人语音进行特征学习,DBN 网络可以有效提取说话人的个性特征信息[11-12]。DBN 相比于传统神经网络,有着更多层非线性映射结构[13],可以完成更复杂的数据学习。该网络是由 RBM 模块堆叠而成的深层网络结构[14-15]。典型的 RBM 是由可见层和隐含层构成二部图模型,可见层或隐含层层内没有连接,只有可见层和隐含层节点间存在连接。

RBM 是一个能量模型,其能量函数表示为:

$$E(v,h) = -v^{\mathrm{T}}Wh - a^{\mathrm{T}}v - b^{\mathrm{T}}h = -\sum_{i=1}^{V}\sum_{j=1}^{H}W_{ij}v_ih_j - \sum_{i=1}^{V}a_iv_i - \sum_{j=1}^{H}b_jh_j \tag{1}$$

其中,v_i 和 h_j 表示可见层第 i 个节点状态和隐含层第 j 个节点状态;W_{ij} 为第 i 个可见层节点和第 j 个隐含层节点的连接权重;a_i 和 b_j 分别为可见层节点和隐含层节点的偏置。

可见层 v 和隐含层 h 的联合概率分布为:

$$p(v,h) = \frac{1}{Z}\exp(-E(v,h)) \tag{2}$$

其中,Z 为分配函数,或称归一化常量,可以通过所有隐含层单元和可见层单元分配能量计算得到,表示如下:

$$Z = \sum_{v}\sum_{h}\exp(-E(v,h)) \tag{3}$$

由于 RBM 在训练时,同一层中具有条件独立性,条件概率分布如下:

$$p(h_j = 1 \mid v) = f(a_i + \sum_{i} w_{ij}v_i) \tag{4}$$

$$p(h_j = 0 \mid v) = 1 - p(h_j = 1 \mid v) \tag{5}$$

$$p(v_i = 1 \mid h) = f(b_j + \sum_{j} w_{ij}h_j) \tag{6}$$

$$p(v_i = 0 \mid h) = 1 - p(v_i = 1 \mid h) \tag{7}$$

其中，函数 f 为 sigmoid 函数，$f(x)=1/(1+e^{-x})$。可以得到 RBM 的更新公式:

$$\Delta w_{ij}=(\partial \ln p(v)/\partial w_{ij})=\varepsilon(\langle v_i,h_j\rangle_{data}-\langle v_i,h_j\rangle)_{model} \quad (8)$$

$$\Delta v_i=\varepsilon(\langle v_i^2\rangle_{data}-\langle v_i^2\rangle)_{model} \quad (9)$$

$$\Delta h_j=\varepsilon(\langle h_j^2\rangle_{data}-\langle h_j^2\rangle)_{model} \quad (10)$$

其中，ε 为学习率；$\langle\rangle_{data}$ 为数据期望；$\langle\rangle_{model}$ 为模型期望。

模型期望计算比较复杂，它需要随机初始化可见层状态然后经过长时间采样，可通过对比散度算法[16]求解。

多层 RBM 堆叠，依次将 RBM 隐含层单元的输出数据作为更高层 RBM 输入层数据，通过学习下一层 RBM 对输出数据的 RBM 隐藏单元的显著依赖关系进行建模，则构成 DBN[17]，这种层层递进的深层网络结构可以有效挖掘语音数据中说话人的深层个性特征，提取出更具代表性的特征向量。DBN 网络模型训练首先进行逐层 RBM 预训练，每层的 RBM 预训练方式和 RBM 训练方式相同，经过多次迭代得到节点间权重和偏置，多层网络依次预训练完毕，然后根据误差反向微调整个网络。DBN 网络可以实现数据有监督或非监督式学习，并且可以提取数据高层特征实现数据特征升降维度。DBN 优化权值的学习算法克服了传统神经网络无法求出最优解等缺点，并有更强的数据建模能力。其模型结构如图 1 所示。

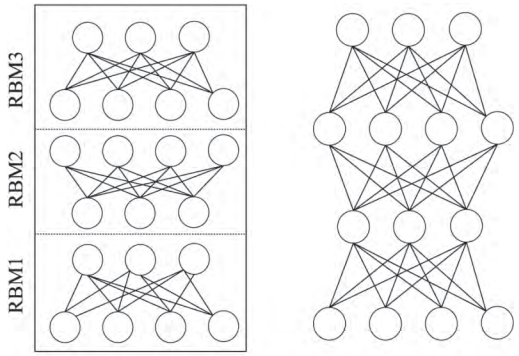

图 1　DBN 模型结构

将 DBN 网络应用在说话人语音识别中，采用少量的说话人语音数据进行 DBN 网络逐层 RBM 训练，可以有效学习和挖掘到语音中的潜在特征，更好地捕获到说话人个性信息，从而在说话人语音数据不充分的条件下大大改善系统识别能力。

2.2　系统描述

为进一步解决噪声环境下系统识别性能不好的问题，结合自动编码器的去噪特点，应用 DBN 网络构造 AutoEncoder DBN 网络，实现深层自动编码网络。网络训练首先采用贪婪学习算法对 DBN 逐层预训练，得到编码网络的初步训练参数，然后由得到的参数反转重构其对称网络，最后通过 BP 算法反向微调整个网络，完成整个网络训练。AutoEncoder DBN 网络前半部分可以实现对输入数据的特征提取和数据编码，后半部分通过深层的特征数据实现对输入数据重构。AutoEncoder DBN 网络结构如图 2 所示。

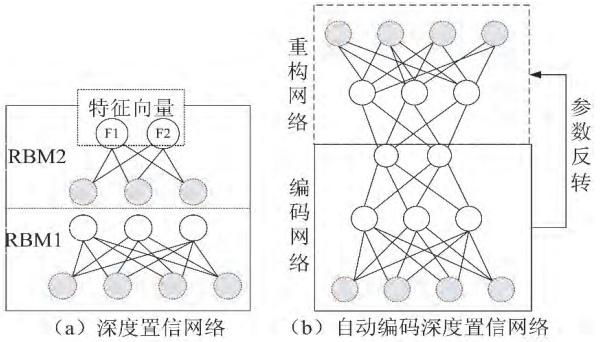

图 2　AutoEncoder DBN 网络结构

在文中说话人识别系统中，AutoEncoder DBN 网络首先采用纯净语音特征数据根据其训练算法进行网络训练，得到 AutoEncoder DBN 网络参数。模型训练完成后，一段有噪语音数据输入网络时，首先经过图 2(b) 中的编码网络对数据进行编码，获得说话人语音深层特征。由于网络训练数据为纯净语音数据，编码网络会捕获语音数据中说话人有效的语音特征数据，过滤掉语音中的噪声数据，得到的特征可以代表说话人语音去噪后深层个性特征，然后经过图 2(b) 中的重构网络，利用得到的深层个性特征重构输出数据，便得到去噪后并且代表说话人的有效个性特征向量。采用 AutoEncoder DBN 网络不仅可以在少量说话人语音数据中捕获高质量的说话人个性特征，还可以对输入的有噪语音数据进行噪声过滤，在提高系统识别率的同时增强了系统鲁棒性。

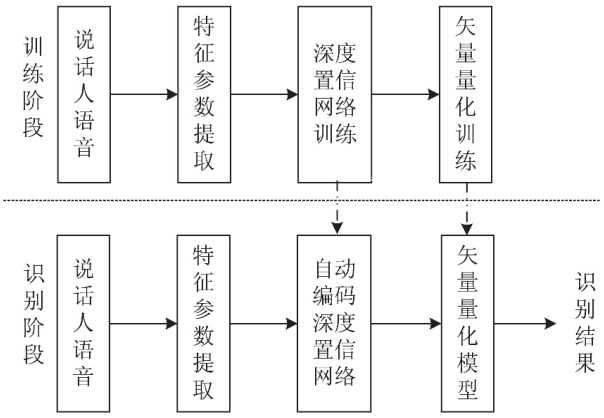

图 3　AutoEncoder DBN-VQ 说话人识别系统流程

整个说话人识别系统流程如图 3 所示。首先需要对说话人语音进行预处理，并提取网络模型训练数据。网络训练时，对特征数据根据不同说话人进行标记，将所有说话人标记过的数据输入网络进行有监督的模型

训练。AutoEncoder DBN 训练完成后,分别将不同说话人无标签特征数据经过 AutoEncoder DBN 编码重构,得到经过 AutoEncoder DBN 网络挖掘和重构的说话人数据,重构数据再作为 VQ 模型训练输入数据,进行 VQ 模型训练。说话人识别时,一段语音过来,经过预处理,提取该语音特征,提取的语音特征数据经过训练好的 AutoEncoder DBN 编码重构,然后输入 VQ 进行说话人身份识别。

3 实验结果与分析

实验运行环境为 MATLAB2014a。采用 TIMIT 语音数据库进行实验。TIMIT 是一个全英文语音数据库,由麻省理工 MIT、斯坦福研究院 SRI 和德州仪器 TI 共同设计。该数据库每位话者在安静环境下录制 10 句话,声音采集频率是 16 000 Hz,采样位数为 16 位。实验选取该语音库 200 名录音人,其中男 128 名,女 72 名。实验将每个人 10 句语音分为互不交叉的训练语音集和测试语音集,每句话平均时长 3 s 左右。实验中采用的噪声信号取自 NoiseX-9 噪声数据库。实验分为两部分,一部分是测试纯净语音条件下说话人语音数据有限时系统性能,另一部分是测试在语音加入噪声情况下算法的正确识别率。

AutoEncoder DBN-VQ(简称 AEDBN-VQ)中初始 DBN 网络结构设置为 3 层,每层节点数为 1024-1024-1024,模型学习率为 0.000 2。训练数据提取说话人语音 40 维 MFCC 参数,去除代表直流分量的第一维数据,然后依次取每帧前后各两帧拼接,形成 195(5*(40-1))维的超帧。采用的 VQ 模型编码长度为 32,码本设计采用 LBG 算法。

文中提出的算法是在 VQ 方法上改进的,首先与该方法进行系统性能对比。基于矢量量化方法的实验设置为:说话人语音特征数据同样提取 40 维 MFCC 参数,去除第一帧直流分量,直接进行连续 5 帧拼接构成 195 维超帧,矢量量化编码长度是 32,码本设计采用 LBG 算法。基于高斯混合模型方法(GMM)的基本原理是用多个高斯模型来拟合说话人语音信号。该方法在说话人识别领域是研究热点,同样选择该方法进行系统性能对比。基于高斯混合模型方法的设置为:语音特征数据采用 20 维 MFCC 参数,高斯混合度设为 16。

3.1 纯净语音条件下的测试

在说话人语音时长有限(不超过 10 s)时,测试模型训练语音和测试语音都为纯净语音条件下的系统识别性能。表 1 和表 2 是模型训练语音时长每人 2 句话(时长约 6 s)和 3 句话(时长约 9 s),测试语音时长为 1~3 句话下的系统识别结果。

表 1 纯净语音条件下每人训练 2 句话的正确识别率 %

算法	测试 1 句话	测试 2 句话	测试 3 句话
AEDBN-VQ	86.5	95.0	97.5
VQ	68.5	79.0	86.0
GMM	74.5	83.5	83.5

表 2 纯净语音条件下每人训练 3 句话的正确识别率 %

算法	测试 1 句话	测试 2 句话	测试 3 句话
AEDBN-VQ	93.5	98.0	99.5
VQ	76.0	87.5	90.0
GMM	85.0	89.5	89.5

在训练语音时长为 2 句话时,AEDBN-VQ 识别率达到 97.5%,另外两种算法识别率不到 90%;训练语音增加到 3 句话时,AEDBN-VQ 识别率基本达到性能最优,识别率是 99.5%,其他两种算法识别率还有很大提升空间,与 AEDBN-VQ 相差 10%左右。另外,在测试语音时长为 1 句话这种极端条件下,AEDBN-VQ 系统在训练时长 2 句话时识别率达到 86.5%,训练为 3 句话时识别率达到 93%,比另外两种算法高出平均 10%。实验结果表明,在纯净语音及说话人训练和测试语音时长有限的条件下,AEDBN-VQ 系统可以更好地捕获说话人个性特征,进行准确识别,系统性能明显高于 VQ 算法与传统 GMM 算法。在每人训练 2 句话和 3 句话的条件下,测试时语句由测试 1 句话到测试 3 句话时长增加,AEDBN-VQ 系统的识别率也有一定的改善,进一步说明了 AEDBN-VQ 说话人识别系统的稳定性。

3.2 有噪语音条件下的实验

有噪语音条件下的实验是测试系统对带有噪声的语音或者模拟实际有噪声环境下的系统识别情况。实验中每位说话人选取的训练语音时长为 3 句话(时长约 9 s),每人剩余语句数都用来进行识别测试。实验设计加入的噪声类型分别是 white 噪声、factory1 噪声、babble 噪声、pink 噪声。AEDBN-VQ 算法网络模型训练数据采用说话人的纯净语音,识别时,由网络模型对待测有噪语音数据进行编码重构,然后输入下一模型进行说话人身份识别。

传统 VQ 算法和 GMM 算法采用纯净语音数据进行模型训练,然后对带噪语音识别时由于训练环境和测试环境不匹配,导致系统识别率急剧恶化,所以在模型训练时在训练语音数据中加入和测试语音中相应的噪声。

表 3~5 分别是在测试语音信噪比为 10 dB、5 dB、0 dB 时三种算法的识别结果。

表 3　信噪比为 10 dB 时三种算法的正确识别率　%

噪声类型	white	factory	babble	pink
AEDBN-VQ	87.0	95.0	97.0	94.5
VQ	65.5	73.5	88.0	63.0
GMM	72.5	74.5	85.5	78.0

表 4　信噪比为 5 dB 时三种算法的正确识别率　%

噪声类型	white	factory	babble	pink
AEDBN-VQ	79.5	81.5	91.5	82.5
VQ	53.0	56.0	74.0	53.5
GMM	62.0	61.0	73.5	66.5

表 5　信噪比为 0 dB 时三种算法的正确识别率　%

噪声类型	white	factory	babble	pink
AEDBN-VQ	56.0	53.5	66.0	62.5
VQ	38.0	29.5	48.0	33.0
GMM	49.5	35.0	43.5	49.5

由表中数据可以看出,在信噪比为 10 dB 时,平稳噪声(white 噪声) 条件下 AEDBN-VQ 算法的正确识别率高出另外两种算法 15%之多,可达到 87%; 非平稳噪声条件下,AEDBN-VQ 算法正确识别率在 95%左右,同样高出另外两种算法平均 10%之多。信噪比为 5 dB 时,AEDBN-VQ 算法正确识别率能稳定在 80%左右,相比另外两种算法系统性能也平均高出 15%。信噪比在 0 dB 时三种算法的识别率都变得很差,但是 AEDBN-VQ 识别率还可以在 50%之上。在测试语音数据中加入噪声,VQ 和 GMM 说话人识别系统的识别率大幅降低,AEDBN-VQ 算法实验结果仍可达到期望识别效果。

实验数据表明,AEDBN-VQ 算法中的自动编码深度置信网络对输入的有噪语音数据确实具有挖掘有效说话人个性信息以及进行有效噪声过滤的作用,使说话人识别系统具有了一定的鲁棒性。

4　结束语

在传统矢量量化方法的基础上,提出深度置信网络与矢量量化方法相结合的算法。应用深度置信网络构造自动编码深度置信网络,实现对说话人语音数据个性特征深度学习,改善了当说话人语音时长有限或不足时传统算法模型训练不充分、识别率不高等问题; 更进一步,结合自动编码器对数据编码重构可以实现数据噪声过滤的优势,使网络模型具备对有噪语音进行噪声过滤的能力,提升了系统的鲁棒性,确保该算法在有噪声环境下也能具备稳定的系统性能。实验结果表明,在纯净语音和有噪语音条件下,该算法比传统算法有更好的识别结果。当然,在 0 dB 等这种极端噪声环境下,该算法的识别率还无法保持在一个可以接受的正确识别率之上,仍然需要进一步探索和完善。

参考文献:

[1] QUATIERI T F.离散时间语音信号处理: 原理与应用 [M]. 北京: 电子工业出版社,2004.

[2] MARTINEZ J,PEREZ H,ESCAMILLA E,et al. Speaker recognition using Mel frequency cepstral coefficients (MFCC) and vector quantization (VQ) techniques [C] //International conference on electrical communications and computers. [s.l.]: IEEE,2012: 248-251.

[3] HUANG C C, GONG W, FU W L,et al. A research of speaker recognition based on VQ and MFCC [J]. Applied Mechanics and Materials,2014,644-650: 4325-4329.

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

[5] YU D,SELTZER M L. Improved bottleneck features using pretrained deep neural networks [C] //Conference of the international speech communication association. [s.l.]: [s.n.], 2011: 237-240.

[6] RUMELHART D E,HINTON G E,WILLIAMS R J. Learning representations by back-propagating errors [J]. Nature, 1986,323(6088) : 533-536.

[7] 曲建岭,杜辰飞,邸亚洲,等.深度自动编码器的研究与展望 [J].计算机与现代化,2014(8) : 128-134.

[8] VINCENT P,LAROCHELLE H,LAJOIE I,et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion [J]. Journal of Machine Learning Research,2010,11(12) : 3371-3408.

[9] 赵　力.语音信号处理 [M].北京: 机械工业出版社,2009.

[10] 丁艳伟,戴玉刚.基于 VQ 的说话人识别系统 [J].电脑知识与技术,2008,4(5) : 1181-1183.

[11] 田　垚,蔡　猛,何　亮,等.基于深度神经网络和 Bottleneck 特征的说话人识别系统 [J].清华大学学报: 自然科学版,2016,56(11) : 1143-1148.

[12] 王山海,景新幸,杨海燕.基于深度学习神经网络的孤立词语音识别的研究 [J].计算机应用研究,2015,32(8) : 2289-2291.

[13] LIU Y,ZHOU S,CHEN Q. Discriminative deep belief networks for visual data classification [J]. Pattern Recognition, 2011,44(10) : 2287-2296.

[14] HINTON G E,OSINDERO S,TEH Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation,2006, 18(7) : 1527-1554.

[15] HINTON G E. Learning multiple layers of representation [J]. Trends in Cognitive Sciences,2007,11(10) : 428-434.

[16] MOHAMED A,DAHL G E,HINTON G. Acoustic modeling using deep belief networks [J]. IEEE Transactions on Audio, Speech,and Language Processing,2012,20(1) : 14-22.

[17] SALAKHUTDINOV R. Learning deep generative models [D]. Toronto: University of Toronto,2009.