ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 2,Mar. 2015,pp. 365－371
DOI:10.16337/j.1004-9037.2015.02.015

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

# 深度神经网络在维吾尔语大词汇量连续语音识别中的应用

麦麦提艾力·吐尔逊[1,2]　戴礼荣[1]

(1.中国科学技术大学语音及语言信息处理国家工程实验室,合肥,230027; 2.新疆师范大学计算机科学技术学院,乌鲁木齐,830054)

**摘　要**:研究将深度神经网络有效地应用到维吾尔语大词汇量连续语音识别声学建模中的两种方法:深度神经网络与隐马尔可夫模型组成混合架构模型(Deep neural network hidden Markov model, DNN-HMM),代替高斯混合模型进行状态输出概率的计算;深度神经网络作为前端的声学特征提取器提取瓶颈特征(Bottleneck features, BN),为传统的GMM-HMM(Gaussian mixture model-HMM)声学建模架构提供更有效的声学特征(BN-GMM-HMM)。实验结果表明,DNN-HMM模型和BN-GMM-HMM模型比GMM-HMM基线模型词错误率分别降低了8.84%和5.86%,两种方法都取得了较大的性能提升。

**关键词**:深度神经网络;维吾尔语;GMM-HMM;瓶颈特征

**中图分类号**:TP391.4　　**文献标志码**:A

## Deep Neural Network based Uyghur Large Vocabulary Continuous Speech Recognition

Maimaitiaili Tuerxun[1,2], Dai Lirong[1]

(1. National Engineering Laboratory of Speech and Language Information Processing, University of Science and Technology of China, Hefei, 230027, China; 2. School of Computer Science and Technology, Xinjiang Normal University, Urumqi, 830054, China)

**Abstract**: Two methods are proposed by employing deep neural network for Uyghur large vocabulary continuous speech recognition: Hybrid architecture models are established with deep neural network (DNN) and hidden Markov model (HMM) for replacing Gaussian mixture model (GMM) in GMM-HMM to compute the state emission probabilities; DNN is facilitated as a front-end acoustic feature extractor to extract bottleneck feature(BN) to provide more effective acoustic features for the traditional GMM-HMM modeling framework (BN-GMM-HMM). The experimental results show that DNN-HMM and BN-GMM-HMM reduce word error rate(WER) by 8.84% and 5.86% compared with the GMM-HMM baseline system, which demonstrates that the two methods accomplish significant performance improvements.

**Key words**: deep neural network; Uyghur; GMM-HMM; bottleneck feature

## 引　　言

现在大部分语音识别技术还是使用隐马尔可夫模型(Hidden Markov model,HMM)来建模语音信

**基金项目**:新疆师范大学网络信息安全与舆情分析实验室公开项目(WLYQ2012103)资助项目。
**收稿日期**:2015-01-16;**修订日期**:2015-02-04

号的时序性,高斯混合模型(Gaussian mixture model,GMM)来建模一帧短时输入声学语音特征与隐马尔可夫模型每一个状态的匹配程度。声学输入特征主要采用 Mel 频率倒谱系数(Mel-frequency cepstral coefficients,MFCC)和感知线性预测(Perceptual linear predictive coefficients,PLP)[1]参数。

最近一种含有多个隐层的前向神经网络即深层神经网络(Deep neural network,DNN)用多帧声学特征作为输入,替代 GMM,输出 HMM 每一个状态的后验概率,且使用新的训练方法在英文等主要语种大词汇量连续语音识别(Large vocabulary continuous speech recognition,LVCSR)任务中性能胜过 GMM,成为主流的声学模型状态输出概率建模方法[2]。

另外,将 DNN 作为前端的声学特征提取器,提取一种瓶颈特征(Bottleneck features,BN)为传统的 GMM-HMM 声学建模架构提供更有效的声学特征。在英文和中文大词汇量连续语音识别任务中性能也优于以传统 MFCC 或 PLP 特征作为声学特征的 GMM-HMM 系统的性能[3]。

尽管 DNN 在英文中文等主要语种的语音识别任务上的优势得到实验验证,但 DNN 在维吾尔语大词汇量连续语音识别任务中的应用尚未深入研究。为此,本文通过研究将深度神经网络两种方式有效地应用到维吾尔语大词汇量连续语音识别声学建模当中,以验证 DNN 技术在维吾尔语大词汇量连续语音识别任务中的有效性:(1)DNN 与 HMM 组成混合架构模型 DNN-HMM,代替 GMM 进行状态输出概率的计算;(2)DNN 作为前端的声学特征提取器,为传统的 GMM-HMM 声学建模架构提供更有效的声学特征。并通过实验验证了两种方法都取得了明显的性能提升。

## 1 基于深层神经网络的声学模型

基于深度神经网络 DNN-HMM 声学模型[4-5]被提出并成功应用于语音识别,并且在英文、中文等语种语音识别任务上一致性地取得相比于传统 GMM-HMM 模型较大性能提升。DNN 相比于 GMM 的优势:(1)使用 DNN 估计 HMM 的状态的后验概率分布不需要对语音数据分布进行假设;(2)DNN 的输入特征可以是多种特征的融合,包括离散或者连续的;(3)DNN 可以利用相邻的语音帧所包含的结构信息。在文献[6]中的研究表明,DNN 的性能提升主要是归功于第 3 点。

图 1 为语音识别系统的各个组成部分示意图。从图 1 可看出,在计算 HMM 状态最大似然估计时使用 GMM 构成 GMM-HMM 模型,使用 DNN 构成 DNN-HMM 模型。GMM 的输入是在时间 $t$ 的一帧特征向量,而 DNN 的输入是在时间 $t$ 的一帧和左右相邻多个帧拼接而成的特征向量。GMM 直接输出 HMM 状态最大似然度 $p(\boldsymbol{o}_t \mid q_j)$,而 DNN 的输出是每一个 HMM 状态的后验概率 $p(q_j \mid \boldsymbol{o}_t)$,这个与 GMM 有所不同,但应用贝叶斯公式,可以把后验概率转化为观察似然度的估计 $p(\boldsymbol{o}_t \mid q_j)$。

对于给定的一个观察 $\boldsymbol{o}_t$(特征向量),DNN 计算 HMM 状态 $j$ 的后验概率 $p(q_j \mid \boldsymbol{o}_t)$。但是,对于 HMM 的观察概率函数 $b_j(\boldsymbol{o}_t)$,需要观察似然度是 $p(\boldsymbol{o}_t \mid q_j)$。可以从 $p(q_j \mid \boldsymbol{o}_t)$来计算 $p(\boldsymbol{o}_t \mid q_j)$。

$$p(q_j \mid \boldsymbol{o}_t) = \frac{p(\boldsymbol{o}_t \mid q_j)\, p(q_j)}{p(\boldsymbol{o}_t)} \tag{1}$$

即

$$\frac{p(\boldsymbol{o}_t \mid q_j)}{p(\boldsymbol{o}_t)} = \frac{p(q_j \mid \boldsymbol{o}_t)}{p(q_j)} \tag{2}$$

式(2)右边的两项可以直接通过 DNN 来计算,分子为 DNN 的输出,分母为给定状态的全部概率,也就是所有的观察加起来的概率,即 $\sigma_j(t)$所有 $t$ 的总和。这样,尽管不能直接计算 $p(\boldsymbol{o}_t \mid q_j)$,但可以由 $\bar{p}(\boldsymbol{o}_t \mid q_j) = \frac{p(\boldsymbol{o}_t \mid q_j)}{p(\boldsymbol{o}_t)}$计算按比例的似然度。实际上这个按比例的似然度就像正规的似然度一样好,因为在识别过程中观察概率 $p(\boldsymbol{o}_t)$是一个常数,它不妨碍应得的结果。

训练 DNN 时误差反向传播(Back-propagation,BP)算法要求对于每个观察 $o_t$ 要有正确状态标记

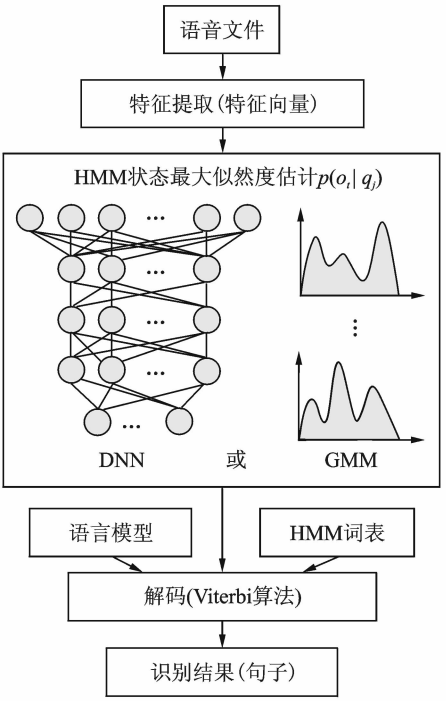


图 1 简化的语音识别总体结构图

Fig. 1 Simplified overall structure of speech recognition

$q_j$,用 GMM-HMM 模型对训练数据进行强对齐(Forced alignment,FA)可以得到状态级标记。总状态数目等于 GMM-HMM 模型中最终状态绑定后的三音素状态数。给定一个观察训练集和正确的标记,随机梯度下降 (Stochastic gradient descent, SGD) 算法反复调整 DNN 中的权值,以减少训练集的错误。DNN 权值的初时化很重要[7]。

## 2 瓶颈特征

DNN 还有一种方式有效地应用于声学建模当中。利用 DNN 提取 BN 替代传统的语音特征,用于训练传统的 GMM-HMM 模型。对输入特征进行多次非线性变换(对应于多层结构,每一层可以看做一个非线性变换),得到区分性更强的声学特征,然后进行 GMM-HMM 的声学模型建模。这种方法结合了 DNN 深度挖掘输入特征的能力以及 GMM-HMM 快速高效的计算能力和相对较为成熟的技术体系,同样能够取得与 DNN-HMM 相当的识别性能。中文实验结果表明基于 BN-GMM-HMM 框架的语音识别系统可以取得和 DNN-HMM 可以比较的性能[8]。

BN 层的节点比其他层少,通常等于普通一帧声学特征向量的长度。图 2(a)是传统的 BN 特征提取网络(BN 层在中间隐层)。BN 网络的训练过程跟普通 DNN 的训练过程一样。训练完 BN 网络以后,抛弃 BN 层后面的隐层,提取 BN 特征,然后使用这些 BN 特征训练 GMM-HMM 模型。文献[9]中提出了更好的 BN 特征网络,把 BN 层后移到最后一个隐层,如图 2(b)所示。在实验中发现这种网络结构的 BN 特征跟传统结构 BN 特征有了明显的性能提升,在维吾尔语大词汇量连续语音识别中也得到了类此的实验结果。

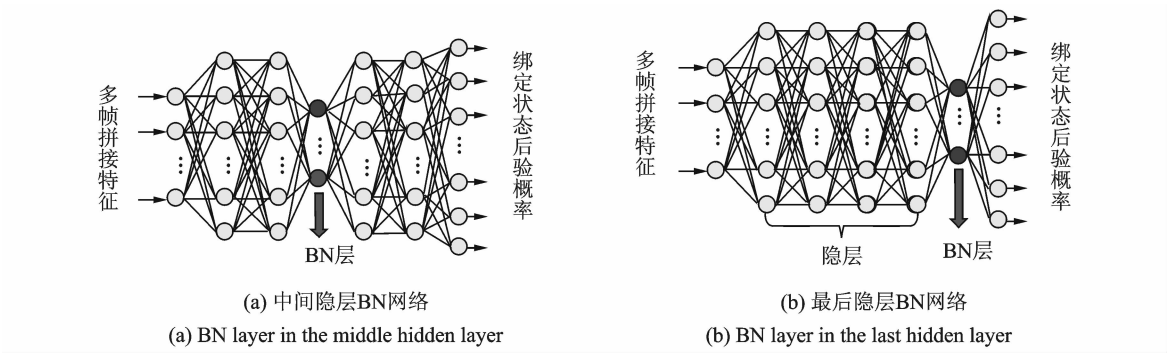


图 2 BN 特征提取网络结构图

Fig. 2 Bottleneck features extraction architecture

## 3 激活函数

DNN-HMM 声学模型中 DNN 网络的激活函数通常都为 Sigmoid 函数。在文献[10-11]中提出采用一种称为 ReLUs(Rectified linear units)激活函数代替 Sigmoid 激活函数，如图 3 所示，两种激活函数为

$$\text{Sigmoid}:\sigma(x)=\frac{1}{1+\mathrm{e}^{-x}}$$

$$\text{ReLUs}:\ f(x)=\max(0,x)$$

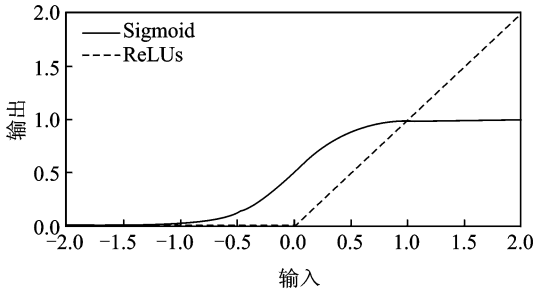


图 3 ReLUs 和 Sigmoid 函数图形

Fig. 3 ReLUs and Sigmoid function

实验结果表明，使用 ReLUs 激活函数的 DNN，不但可随机初始化，而且可以缩短训练时间，取得更好的语音识别性能。本文中使用基于 ReLUs 的 DNN 网络。

## 4 维吾尔语简介

维吾尔语属于阿勒泰语系突厥语族的西匈奴语支。维吾尔语有 8 个元音，24 个辅音，一共有 32 个音素。维吾尔语是字形与音位相互对应的语言，构造发音字典比较简单。和汉语一样，维吾尔语中能够感受的最小语音片段是音节，每个单词由一个或多个音节构成。维吾尔语音节结构是：(起音)＋领音＋(收音)。大部分音节中起音和收音是一个辅音，领音是一个元音，少量音节中起音和收音可以由两个辅音构成，领音也可以由两个元音构成。音节中可以没有起音和收音，但不能没有领音，而且，领音必须是元音。维吾尔语形态结构比较丰富，单词由词干和词缀组成。词缀有前缀和后缀，其中后缀的数量比较多。一个词干连接一个或多个词缀可以构造多个单词，所以容易出现较高的未登录词(Out of vocab-

ulary,OOV) 问题。

## 5 实验与分析

### 5.1 语料库

本文使用科大讯飞公司提供的维吾尔语标准口音朗读式语音语料库。语音语料库朗读文本由三音素覆盖率比较平衡的 2 443 个句子构成,主要来源于维吾尔文新闻、词典例句、小说和网站。朗读人有 163 人(女 87,男 76),每人平均朗读 125 个句子,每个句子平均朗读 9 次,一共 20 600 个句子,总时长大约 50 h。语音文件采样率 16 kHz、采样位 16 bit。从语音数据库中挑选了 100 句子(包含 832 个单词),然后朗读这 100 个句子的 1 099 个句话(时长约 1.5 h)做了测试集,其余(时长 48.5 h)的做了训练集。因为本实验的重点是声学模型性能改进,所以挑选测试集文本时尽量挑选了没有 OOV 的句子。

### 5.2 语言模型

目前,绝大部分 LVCSR 系统均采用基于 n-gram 的统计语言模型用于解码器中计算语言模型概率。本文中使用训练集中的 2 343 个句子文本和另外 6 592 个词典例句文本来训练词级三元语言模型。为了解决数据稀疏的问题,采用了 Good-Turing 平滑技术。语言模型训练文本中的 23 143 个不重复的单词来生成发音字典。

### 5.3 GMM-HMM 模型

HMM 进行声学模型建模,首先需要根据语音的特点选择合适的建模单元,每个建模单元都对应不同的 HMM。根据识别任务不同的难度和规模,通常可以选用词(Word)、音节(Syllable)或者音素(Phoneme)作为声学建模单元,而音素建模单元往往是 LVCSR 任务的首选。考虑到连续语音中的协同发音(Co-articulation)现象,采用上下文相关(Context-dependent)的音素建模,本文实验中使用了跨词三音素(Tri-Phone)建模。对普通的三音素单元使用自左向右的无状态间跨越的三状态 HMM,而对静音模型和停顿模型等特殊模型,则分别采用状态间可跨越的三状态 HMM 和单状态 HMM,每个 HMM 拓扑结构前后有一个开始状态和一个结束状态。训练时利用发音方式的相似性设计好的问题集来进行状态绑定,根据训练语料的规模绑定状态数设置为 3 000,加上静音 3 个状态和停顿 1 个状态,一共设置了 3 004 个状态。

GMM-HMM 模型用最大似然估计准则(Maximum likelihood estimation, MLE)[12]来训练,使用 39 维特征,包括 12 维 MFCC 特征和能量特征,以及它们的一阶和二阶差分,并使用倒谱均值方差归一化处理。帧长设置 25 ms,帧移设置 10 ms。HMM 中每个状态设置 20 个独立的高斯分量。语言模型规整因子设置为 12,声学模型规整因子设置为 1。

### 5.4 DNN-HMM 模型

DNN-HMM 模型中使用基于 ReLUs 的 5 个隐层 DNN 网络,每个隐层包含 2 048 节点。输入层包含拼接 11 帧(5+1+5)39 维 MFCC 特征,有 429 节点,输出层节点数跟最终绑定状态数一样,有 3 004 节点。用 BP 算法对 DNN 进行训练,由 DNN 计算得到的预估概率分布之间的交叉熵(Cross entropy, CE)作为目标函数。在 BP 算法中,随机梯度下降的 Mini-batch 大小设置成 1 024。BP 过程所使用的绑定状态标注利用 MLE 训练的 GMM-HMM 模型对训练集进行强制对齐得到。网络权重使用文献[13]的方法随机初始化,因为使用 ReLUs 激活函数,没有使用无监督的预训练。为了防止权重值无限涨大,对初始值进行了动态范围的控制处理,并且使用了较小的学习率。初始学习率是 0.02,保持前 4 个迭代,后续迭代中每次迭代减半。

表 1 列出了 GMM-HMM 和 DNN-HMM 模型的性能对比，可以看到，DNN-HMM 的识别性能提升非常显著，相比于 MLE 训练的 GMM-HMM 模型，它使词错误率下降了 8.84%。

**表 1 GMM-HMM 和 DNN-HMM 模型性能**

**Table 1 Performance of GMM-HMM and DNN-HMM models**

| 模型 | 词错误率/% |
|---|---|
| GMM-HMM | 21.82 |
| DNN-HMM | 12.98 |

### 5.5 BN_GMM-HMM 模型

BN 特征网络使用基于 ReLUs 的 5 个隐层 DNN 网络，BN 层分别定义在中间和最后一个隐层，每个 BN 层的节点数和 MFCC 特征维数一样有 39 个节点。它的训练所涉及到的一些参数配置与 DNN 的训练一致。

表 2 中列出了 BN-GMM-HMM 模型性能，可以看到 BN 层后移的 BN-GMM-HMM 模型取得了较好的识别性能，相比于 MLE 训练的 GMM-HMM 模型，它使词错误率绝对下降了 5.86%，比较接近 DNN-HMM 模型性能。这次试验也再次证明 BN 层后移的 BN-GMM-HMM 模型取得了较好的识别性能。

**表 2 BN-GMM-HMM 模型性能**

**Table 2 Performance of BN-GMM-HMM model**

| 网络结构 | 词错误率/% |
|---|---|
| 429-4 * 2048-39-3004 | 15.96 |
| 429-2 * 2048-39-2 * 2048-3004 | 18.76 |

BN 特征是由原始 MFCC 特征拼接形成的高维向量经过 DNN 多个隐层的非线性处理而得到，其声学空间相比于原始 MFC 特征必然发生了较大的变化，因此，在利用基于 BN 特征训练得到的 GMM-HMM（即 BN-GMM-HMM）模型进行识别解码时，为了能使 BN 特征取得优越的性能，使用文献[14]的方法进行了一些启发式的尝试。重新设置声学规整因子，用来重新调整解码过程中声学模型得分和语言模型得分之间的相对重要性，通常是按照式$[\xi \cdot \log p(\boldsymbol{o}|w) + \log p(w)]$来调整，其中，$\xi$ 为声学规整因子，$w$ 为句子中的某一个词，$\boldsymbol{o}$ 为 $w$ 所对应的一串声学特征向量，而 $\log p(\boldsymbol{o}|w)$ 和 $\log p(w)$ 则分别为声学模型得分和语言模型得分。对于原始 MFCC 特征，声学规整因子一般设置成 1.0，而对于 BN 特征，一般设置成小于 1.0，本实验中声学模型规整因子等于 0.4 时取得了最好的性能。

## 6 结束语

本文首先介绍了基于深度神经网络的语音识别模型和普通 GMM-HMM 模型的区别，然后将深度神经网络两种方式有效地应用到维吾尔语大词汇量连续语音识别声学建模当中，实验结果表明，DNN-HMM 模型和 BN-GMM-HMM 模型分别比 GMM-HMM 基线模型词错误率降低了 8.84%和 5.86%，两种方法都取得了明显的性能提升。

本实验为了减少 OOV 单词对声学模型性能测试的影响，使用了没有 OOV 的测试集。维吾尔语形态结构比较丰富，在严格的测试集中可能会出现较高的未登录词。后续收集更多的语言模型训练文本，训练更好的语言模型，进一步验证本文方法的有效性。

后续计划是在最小化帧级交叉熵准则下训练得到的 DNN 基础上，使用相应的 DNN 基线模型对训练集重新进行强制对齐，获得更为精准的绑定状态标注，利用序列区分性训练（Sequential discriminative

training,SDT)算法进行进一步的调优,获得更多的性能提升。

**参考文献:**


[1] Hermansky H. Perceptual linear predictive (PLP) analysis of speech[J]. The Journal of the Acoustical Society of America, 1990,87(4):1738-1752.

[2] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. Signal Processing Magazine, IEEE, 2012,29(6):82-97.

[3] Grezl F, Karaat M, Kontar S, et al. Probabilistic and bottle-neck features for LVCSR of meetings[C]//ICASSP. Honolulu: IEEE,2007:757-760.

[4] Mohamed A, Dahl G E, Hinton G. Acoustic modeling using deep belief networks[J]. Audio, Speech, and Language Processing, IEEE Transactions, 2012,20(1):14-22.

[5] Mohamed A, Dahl G E, Hinton G E. Deep belief networks for phone recognition[C]//NIPS Workshop on Deep Learning for Speech Recognition and Related Applications. Hyatt Regency Vancouver, Canada: [s. n.], 2009:1-9.

[6] Pan J, Liu C, Wang Z, et al. Investigation of deep neural networks(DNN)for large vocabulary continuous speech recognition: Why DNN surpasses GMMS in acoustic modeling[C]//Chinese Spoken Language Processing(ISCSLP), 2012 8th International Symposium on. Hong Kong, China:IEEE,2012:301-305.

[7] Daniel J, James H M. 自然语言处理综论[M]. 冯志伟,等,译. 北京:电子工业出版社, 2006:158-168.
Daniel J, James H M. Speech and language processing[M]. Beijing: Publishing House of Electronics Industry, 2006:158-168.

[8] 戴礼荣,张仕良.深度语音信号与信息处理:研究进展与展望[J].数据采集与处理,2014(02):171-179.
Dai Lirong, Zhang Shiling. Deep speech signal and information processing: Research progress and prospect[J]. Journal of Data Acquisition and Processing. 2014(02):171-179.

[9] Tuerxun M M T A L, Zhang S L, Bao Y B, et al. Improvements on bottleneck feature for large vocabulary continuous speech recognition[C]//Proceedings of IEEE 12th International Conference on Signal Processing (ICSP). Hangzhou, China: IEEE, 2014:516-520.

[10] Dahl G E, Sainath T N, Hinton G E. Improving deep neural networks for LVCSR using rectified linear units and dropout [C]//ICASSP. British Columbia: IEEE,2013:8609-8613.

[11] Zeiler M D, Ranzato M, Monga R, et al. On rectified linear units for speech processing[C]//ICASSP. British Columbia: IEEE, 2013:3517-3521.

[12] Bahl L R, Jelinek F, Mercer R L. A maximum likelihood approach to continuous speech recognition[J]. IEEE Trans on Pattern Analysis and Machine Intelligence (PAMI), 1983,5(2):179-190.

[13] Glorot X, Bengio Y. Understanding the difficulty of training deep feed forward neural networks[C]//13th International Conference on Artificial Intelligence and Statistics (AISTATS). Sardinia: [s. n.], 2010:249-256.

[14] Bao Y B, Jiang H, Liu C, et al. Investigation on dimensionality reduction of concatenated features with deep neural network for LVCSR Systems[C]//Proceedings of IEEE 11th International Conference on Signal Processing (ICSP). Beijing, China: IEEE, 2012:562-566.


作者简介:**麦麦提艾力·吐尔逊**(1980-),男,博士研究生,研究方向:语音识别,E-mail: mamatali@mail. ustc. edu. cn;**戴礼荣**(1962-),男,教授,博士生导师,研究方向:语音识别、语音合成和说话人识别等。