ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 5,Sep. 2021,pp. 932－938
DOI:10. 16337/j. 1004-9037. 2021. 05. 008

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

融合声学特征和深度特征的语音文档分类

刘　谭，郭　武

（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）

摘　要：传统的语音文档分类系统通常是基于语音识别系统所转录的文本实现的，识别错误会严重影响到这类系统的性能。尽管将语音和识别文本融合可以一定程度上减轻识别错误的影响，但大多数融合都是在表示向量层面融合，没有充分利用语音声学和语义信息之间的互补性。本文提出融合声学特征和深度特征的神经网络语音文档分类，在神经网络训练中，首先采用训练好的声学模型为每个语音文档提取包含语义信息的深度特征，然后将语音文档的声学特征和深度特征通过门控机制逐帧进行融合，融合后的特征用于语音文档分类。在语音新闻播报语料集上进行实验，本文提出的系统明显优于基于语音和文本融合的语音文档分类系统，最终的分类准确率达到97.27％。

关键词：神经网络；语音文档分类；语音识别；深度特征；门控机制

中图分类号：TN911　　**文献标志码**：A

Spoken Document Classification Based on Fusion of Acoustic Features and Deep Features

LIU Tan, GUO Wu

(National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Heifei 230027, China)

Abstract: Traditional speech document classification systems are usually completed through the transcribed text from speech recognition systems, which suffer from the recognition errors. Although the fusion of speech and recognized text can reduce the impact of recognition errors to some extent, the fusion that is made at the level of representation vector does not take full advantage of the complementarity between speech and text information. A neural network spoken document classification system based on the fusion of acoustic feature and deep feature is proposed in this paper. In the training procedure of the neural network, a trained acoustic model is first adopted to generate deep feature that contains semantic information for each document. Then acoustic feature and deep feature of each spoken document are fused frame by frame through the gating mechanism. Finally, the fused feature is used for spoken document classification. The proposed system is evaluated on a speech news broadcast corpus. The experimental result showed that the proposed system was obviously superior to the spoken document classification systems based on the fusion of speech and text, and the final accuracy reached 97.27％.

Key words: neural network; spoken document classification; automatic speech recognition; deep feature; gating mechanism

收稿日期：2021-01-12；**修订日期**：2021-07-12

引　言

语音文档分类旨在自动将大量的语音文档按照内容的主题进行分类。随着互联网和信息技术不断发展,语音文档分类技术在信息检索中扮演着愈发重要的角色。

传统的语音文档分类系统通常由语音识别(Automatic speech recognition,ASR)模块和文本文档分类(Textual document classification,TDC)模块组成。ASR模块首先将语音识别为文本,TDC模块再根据识别文本的内容进行主题分类。近几年来,ASR和TDC技术都已取得了很大的进展。对于ASR,目前主流的系统有两种,一种是基于隐马尔可夫模型(Hidden Markov model, HMM)的语音识别系统,另一种则是端到端ASR系统[1-2]。基于HMM的ASR系统一般由声学模型、语言模型、发音词典等多部分组成,训练过程复杂。端到端ASR系统直接将输入的语音特征序列转化成文本,相比于基于HMM的ASR系统,其结构更加简单,且准确率可以达到甚至超越基于HMM的ASR系统。基于链接时序分类(Connectionist temporal classification,CTC)[3-4]的ASR系统就是一种典型的端到端结构。对于TDC而言,关键技术就是如何准确地构建文本文档的表示向量。目前常用的构建文档表示向量的方法有概率隐语义分析(Probabilistic latent semantic analysis, PLSA)[5]和隐含狄利克雷分布(Latent Dirichlet allocation, LDA)[6]。在获得文档表示向量后,便可以使用分类器(例如支持向量机(Support vector machine, SVM)[7]等)对这些表示向量进行分类。此外,鉴于神经网络(Neural network,NN)在许多任务上都取得了令人满意的效果,一些基于NN的文本分类方法也已经被提出。Kim[8]提出将卷积神经网络(Convolutional neural network, CNN)用于文本分类。CNN能够捕获相邻词的语义特征,通过多个不同尺度的滤波器来提取不同层面的语义信息。此外,Yang等[9]根据文档的结构信息,采用层级注意力网络(Hierarchical attention network, HAN)依次构建句子的表示向量和文档的表示向量,进一步提高了文档分类的准确率。

显而易见,在这种串联型结构的语音文档分类系统中,ASR错误会降低系统的准确率,尤其是在嘈杂环境中,由于噪声和混响的干扰,ASR错误率会明显增加。为降低识别错误带来的影响,Gogate等[10]提出将语音和识别的文本进行融合,利用语音信息改善识别错误带来的影响。Yang等[9]使用CNN分别提取语音信息和文本信息用于构建语音表示向量和文本表示向量,然后将这两种表示向量拼接用于情感分类。和文献[10]相似,文献[11]采用长短期记忆网络(Long short-term memory, LSTM)[12]网络分别构建语音表示向量和文本表示向量,然后通过注意力机制将这两种表示向量融合,用于口语语言分类。尽管将语音和识别文本融合后,系统的性能有所提高,但是由于语音信息和文本信息只在表示向量层面进行融合,语音和文本的互补性没有被充分利用。鉴于此,本文提出一种融合声学特征和深度特征的系统用于语音文档的分类。首先采用一个训练好的LSTM-CTC声学模型[13]为每个语音文档提取深度特征(Deep feature),LSTM输出的隐状态即为本文所描述的深度特征。然后将语音文档的声学特征(Acoustic feature)和深度特征分别输入到声学特征编码器和深度特征编码器,并将声学特征编码器和深度特征编码器的输出通过门控机制逐帧融合得到融合特征,最后将融合特征用于语音文档的分类。

1　基于语音和识别文本融合的语音文档分类系统

1.1　基于CTC的连续语音识别

CTC是目前端到端ASR的主流算法之一,本文采用LSTM来训练神经网络声学模型。训练好LSTM-CTC之后,可以和语言模型结合用于语音的识别解码得到更准确的结果,也可以将这个LSTM-CTC用于获得深度特征。

对于ASR任务，输入序列为人工提取的声学特征序列$\boldsymbol{x}=\{x_1,x_2,\cdots,x_T\}$，对应的输出标签序列为$\boldsymbol{y}=\{y_1,y_2,\cdots,y_M\}$，通常$M\ll T$。CTC引入了一个空白标签blank，用来表示无标签时的空白映射。CTC的核心是建立中间标签序列$\boldsymbol{\pi}=\{\pi_1,\pi_2,\cdots,\pi_T\}$，该中间序列允许标签的重复出现，从而建立中间序列和输出序列的多对一映射。所有可能映射到输出标签序列的中间序列集合为$\Phi(\boldsymbol{y}')$，CTC的训练目标是最大化输出序列的概率$P(\boldsymbol{y}|\boldsymbol{x})$

$$P(\boldsymbol{y}|\boldsymbol{x})=\sum_{\pi\in\Phi(y')}P(\boldsymbol{\pi}|\boldsymbol{x}) \tag{1}$$

式中：$\boldsymbol{y}'$为经过插入blank及重复标签单元等操作而得到的映射序列，最终的输出是对中间序列合并连续重复单元及去除blank得到。

在CTC准则中，输出单元之间是假设独立的，则$P(\boldsymbol{\pi}|\boldsymbol{x})$可由式(2)得到。

$$P(\boldsymbol{\pi}|\boldsymbol{x})=\prod_{t=1}^{T}P(\pi_t|\boldsymbol{x})=\prod_{t=1}^{T}P(\pi_t^{l_t}) \tag{2}$$

式中$P(\pi_t^{l_t})$为输出在t时刻对应标签为l_t的概率。

对于$P(\pi_t^{l_t})$的计算，通过LSTM、全连接层和Softmax层来得到。将声学特征序列$\boldsymbol{x}$输入LSTM中，得到对应的隐状态序列$\boldsymbol{h}=\{h_1,h_2,\cdots,h_T\}$，然后将隐状态序列输入全连接层和Softmax层，得到每个时刻对应的标签概率分布。

1.2 基于语音和识别文本融合的语音文档分类

ASR系统在训练完成后，便可以将所有语音文档识别为文本，但是识别错误导致语音文档分类系统性能不佳，而将语音和识别文本进行融合可以提高语音文档分类的准确率。基于语音和识别文本融合的语音文档分类系统结构如图1所示。该系统结构主要由3部分组成：文本编码器、声学特征编码器和表示向量融合层。语音信息和文本信息分别以声学特征和字向量的形式输入到系统中。本文采用fbank特征作为声学特征，并且采用预训练的word2vec[14]模型将每个字处理成固定维度的字向量。

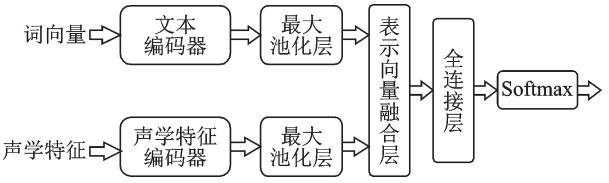


图1 基于语音和识别文本融合的语音文档分类系统结构图

Fig.1 Architecture of spoken document classification system based on fusion of speech and recognized text

声学特征编码器和文本编码器分别用来构建语音表示向量和文本表示向量。由于语音和文本都属于序列结构的信息，因此本文采用LSTM作为声学特征编码器和文本编码器，并且通过在时间维度进行最大池化得到语音表示向量和文本表示向量。

表示向量融合层用于将提取的语音表示向量和文本表示向量进行融合，本文通过基于注意力机制[15-16]将这两种表示向量进行融合。注意力机制动态地为这两种表示向量分配注意力权重，再将其加权求和，得到融合后的表示向量，计算式为

$$\boldsymbol{u}_i=\tanh(\boldsymbol{W}\boldsymbol{v}_i+\boldsymbol{b})\quad i\in[1,2] \tag{3}$$

$$\boldsymbol{\alpha}_i=\frac{\exp(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u})}{\sum_{j=1}^{M}\exp(\boldsymbol{u}_j^{\mathrm{T}}\boldsymbol{u})} \tag{4}$$

$$\boldsymbol{v}_{\mathrm{atten}}=\sum_{i=1}^{2}\boldsymbol{\alpha}_i\boldsymbol{v}_i \tag{5}$$

式中：$\boldsymbol{v}_1$、$\boldsymbol{v}_2$分别表示语音表示向量和文本表示向量；$\boldsymbol{W}$、$\boldsymbol{b}$和$\boldsymbol{u}$均为可学习的参数，若表示向量的维度为

d,则 $\boldsymbol{W}\in\mathrm{R}^{d\times d}$,$\boldsymbol{b}\in\mathrm{R}^{d\times 1}$,$\boldsymbol{u}\in\mathrm{R}^{d\times 1}$;$\boldsymbol{\alpha}_i$表示的注意力权重;$\boldsymbol{v}_{\mathrm{atten}}$表示加权融合得到的表示向量。

2　融合声学特征和深度特征的语音文档分类系统

本文提出的融合声学特征和深度特征的语音文档分类系统结构如图2所示。该系统主要由4个模块组成:声学特征编码器,深度特征编码器,门控单元(Gate)以及融合特征编码器。

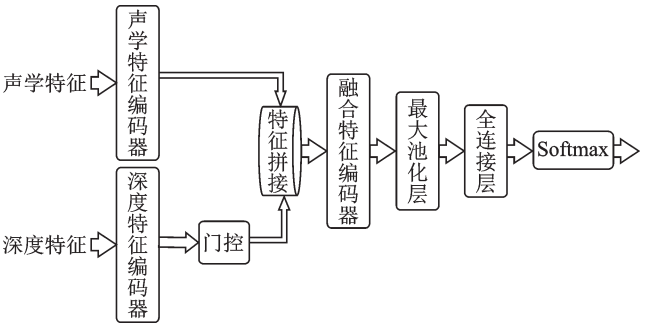


图2　融合声学特征和深度特征的语音文档分类系统结构

Fig.2　Architecture of spoken document classification system based on fusion of acoustic features and deep features

在ASR的声学模型中,深度特征经过全连接层和输出层(Softmax)后可以得到对应字分布概率,本文中将全连接层的前一层隐状态序列作为深度特征。采用训练完成的LSTM-CTC声学模型(如1.1节所述)作为深度特征提取器。将语音文档的声学特征序列输入到LSTM,最后一个LSTM层的输出就是对应的隐状态序列 $\boldsymbol{h}$,$\boldsymbol{h}$ 即为本文所描述的深度特征序列。因此,深度特征可以看作字在另一维度空间的表示。相对于原始的声学特征,深度特征不仅包含更高级的声学信息,还包语义信息,因此可以用来进行语音文档的分类。相比于识别文本,深度特征具有更强的泛化能力,即每个深度特征不表示为具体的某个字,而表示声学特征相似的字的集合,这在一定程度上缓解了识别错误带来的影响。

在得到每个语音文档的深度特征序列后,将声学特征和深度特征分别输入声学特征编码器和深度特征编码器中。由于深度特征序列和声学特征序列具有相同的帧数,因此可以将其逐帧进行融合。相对于语音和文本在表示向量水平的融合,逐帧融合进一步利用了信息之间的互补性。这是深度特征相对于识别文本的另一优点。本文通过门控机制将这两种特征序列逐帧进行融合,假设声学特征编码器的输出为 $\boldsymbol{a}=\{a_1,a_2,\cdots,a_T\}$,深度特征编码器的输出为 $\boldsymbol{d}=\{d_1,d_2,\cdots,d_T\}$,融合过程由式(6~8)得到。

$$\boldsymbol{d}_{\max}=\mathrm{maxpooling}(\boldsymbol{d}) \tag{6}$$

$$\boldsymbol{g}_i=\mathrm{sigmoid}(\boldsymbol{W}_1\boldsymbol{d}_i+\boldsymbol{W}_2\boldsymbol{d}_{\max}) \tag{7}$$

$$\boldsymbol{f}_i=[\boldsymbol{a}_i,\boldsymbol{g}_i\cdot\boldsymbol{d}_i] \tag{8}$$

式中:maxpooling()表示将在时间维度最大池化操作,$\boldsymbol{g}_i$ 用来控制引入多少深度特征信息,[·]表示将向量进行拼接,$\boldsymbol{f}_i$ 即为第 i 帧融合特征。在得到融合特征后,将其输入融合特征编码器中,并通过在时间维度最大池化来构建最终的语音文档表示向量。

3　实验结果与分析

3.1　数据集

本文采用Aishell-1数据集来训练ASR系统,并采用一个普通话新闻播报语料集来训练和测试所有的语音文档分类系统。Aishell-1数据集和新闻播报语料集均以16 kHz采样率,16 bit量化的格式存储。该新闻播报语料集共包含12 447条语音文档,涉及6个主题,分别为“娱乐”“财经”“军事”“体育”“科技”“天气”,每条语音文档都涉及其中一个主题。在实验中,选择9 957条语音作为训练集,1 244条语音作

为验证集，1 246条语音作为测试集。本实验采用的声学特征是108维的fbank特征，由36维的fbank特征结合其一阶差分和二阶差分所组成。此外，对于ASR系统，以字为建模单元，共有4 294个单元。本文以pytorch、kaldi[17]作为实验平台，比较不同模型的实验结果，验证所提出方法的性能。

3.2 模型

本实验总共测试了6个语音文档分类模型：Speech only (SO), Text only (TO), Deepfeature only (DO), Fusion of speech and text(ST), Attention based late fusion of speech and deepfeature (ALF), Fusion of speech and deepfeature(SD)。SO模型仅使用声学特征进行语音文档的分类。SO模型首先将语音文档的fbank特征输入到一个基于LSTM的声学特征编码器中，然后通过最大池化操作将编码器的输出压缩成固定维度的表示向量，该表示向量中包含了语音文档主题的相关信息，最后将该表示向量输入到全连接层和Softmax层来预测语音文档的主题，SO模型常用于端到端的口语理解[18]。TO模型和DO模型分别只使用识别的文本和深度特征进行语音文档分类，其结构组成和SO相同。TO模型采用搜狗新闻语料预训练的word2vec模型，每个字都首先被映射为300维的字向量，即每个语音文档的识别文本可以用一个$N \times 300$的矩阵表示(N表示总的字数)，然后再被输入到一个基于LSTM的文本编码器中[19]。ST模型即为1.2节介绍的基于语音和识别文本融合的语音文档分类系统，如图1所示。同时，为了验证环境噪声对于实验结果的影响，本文为每条语音文档添加了信噪比为20 dB的高斯白噪声，并用ST(clean)和ST(noisy)分别表示使用干净语音和加噪语音的ST模型。ALF模型采用目前常用的特征融合框架[20]，其结构和ST模型相似。 ALF模型包含两个编码器，分别使用声学特征和深度特征进行构造语音文档的表示向量，然后通过注意力机制将这两种表示向量进行融合得到最终的表示向量用于分类，所采用的注意力计算方式和式(3～5)相同。 SD模型即为本文所提出的融合声学特征和深度特征的系统，如图2所示。

3.3 参数设置

LSTM-CTC声学模型结构与文献[10]中的声学模型结构相同，双向LSTM隐藏节点数为512，因此提取的深度特征维度为1 024。对于语音文档分类系统，每个模型的参数都是调节到最好的。SO模型、TO模型和DO模型中的编码器均是由2层双向LSTM实现，隐藏节点数均为512。ST模型中的声学特征编码器由一个2层双向LSTM实现，文本编码器采用2层双向LSTM，隐层节点数均为512。ALF 模型中的编码器均由2层隐层节点数为512的LSTM实现。SD模型中的声学特征编码器采用2层双向LSTM，深度特征编码器采用1层双向LSTM。

3.4 实验结果与分析

本文采用语音文档的分类准确率(Accuracy rate, ACC)作为模型评价指标。不同模型的实验结果如表1所示。

表1 不同模型的实验结果

Table 1 Results of different models

模型	准确率 / %
SO	86.84
TO	83.79
DO	92.46
ST (noisy)	93.26
ST (clean)	95.43
ALF	95.88
SD	97.27

从表1可以看出，本文提出的SD模型实现最高的准确率97.27%，相比于ST(clean)模型，准确率提高了1.84%，验证了1.2节所述的深度特征相对于识别文本的优点，并且相比于目前主流的特征融合模型ALF，SD模型的准确率提高了1.39%，验证了该模型的有效性。同时，相比于ST(noisy), ST(clean)的准确率提高了2.17%，验证了环境噪声对于实验结果的影响。此外，ST模型的准确率要高于SO和TO模型，说明语音和文本信息融合有助于语音文档的分类。最后，DO模型的准确率高于SO模型和TO模型，这是因为深度特征既包含声学信息，又包含语义信息。

为了验证本文提出的声学特征和深度特征的融合方式的有效性,还另外构建了两个对比系统。首先考虑门控机制对于模型性能的影响,本文设计了语音和深度特征的无门控融合(Ungated fusion of speech and deepfeature, USD)系统。USD模型没有采用门控机制,直接将声学特征和深度特征逐帧进行拼接,其他参数设置和SD相同。第二个对比系统是语音和深度特征相加(Addition of speech and deepfeature, ASD)系统,该系统将SD模型的声学特征和深度特征的融合方式变成了逐元素相加(Element-wise add)方法,即将式(8)的拼接换成逐元素相加。实验结果如表2所示,可以看出SD模型的性能优于USD和ASD系统,这也证明了本文提出的模型结构的有效性。

表2　消融实验结果

Table 2　Results of ablation experiments

模型	准确率/%
USD	96.95
ASD	96.79
SD	97.27

4　结束语

本文提出了一个融合声学特征和深度特征的语音文档分类系统,首先采用一个训练好的LSTM-CTC声学模型每个语音文档提取深度特征,然后将声学特征和深度特征通过门控机制逐帧融合,最后使用融合特征构建语音文档的表示向量用于分类。本文在一个新闻播报语料集上进行实验,实验结果表明,相比于基于语音和文本融合的语音文档分类系统,该系统的准确率提升了1.84%,验证了该系统的有效性。

参考文献:

[1] PHAM N Q, NGUYEN T S, NIEHUES J, et al. Very deep self-attention networks for end-to-end speech recognition[EB/OL]. (2019-04-30) [2021-01-05]. https://arxiv.org/1904.13377.

[2] GRAVES A, JAITLY N. Towards end-to-end speech recognition with recurrent neural networks[C]//Proceedings of International Conference on Machine Learning. Beijing, China: JMLR, 2014.

[3] GRAVES A, FERNÁNDEZ S, GOMEZ F, et al. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd International Conference on Machine Learning. New York: Association for Computing Machinery, 2006.

[4] SOLTAU H, LIAO H, SAK H. Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition[EB/OL]. (2016-10-31) [2020-12-20]. https://arxiv.org/abs/1610.09975.

[5] HOFMANN T. Unsupervised learning by probabilistic latent semantic analysis[J]. Machine Learning, 2001, 42(1/2): 177-196.

[6] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[7] NOBLE W S. What is a support vector machine?[J]. Nature Biotechnology, 2006, 24(12): 1565-1567.

[8] KIM Y. Convolutional neural networks for sentence classification[EB/OL]. (2014-08-25) [2020-12-10]. https://arxiv.org/abs/1408.5882.

[9] YANG Z, YANG D, DYER C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: Association for Computational Linguistics, 2016.

[10] GOGATE M, ADEEL A, HUSSAIN A. Deep learning driven multimodal fusion for automated deception detection[C]//Proceedings of 2017 IEEE Symposium Series on Computational Intelligence (SSCI). [S.l.]: IEEE, 2017: 1-6.

[11] GU Y, YANG K, FU S, et al. Hybrid attention based multimodal network for spoken language classification[C]//Proceedings of the Conference of Association for Computational Linguistics Meeting. [S.l.]: NIH Public Access, 2018: 2379.

[12] OLAH C. Understanding lst networks[EB/OL]. (2015-08-30) [2021-03-05]. http://colah. github. io/posts/2015-08-Understanding-LSTMs/.

[13] DING F, GUO W, GU B, et al. Adaptive speaker normalization for CTC-based speech recognition[J]. Proc Interspeech 2020, 2020: 1266-1270.

[14] RONG X. Word2vec parameter learning explained[EB/OL]. (2014-11-11) [2020-11-03]. https://arxiv.org/abs/1411.2738.

[15] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. (2014-09-01) [2020-12-05]. https://arxiv.org/abs/1409.0473.

[16] HORI C, ALAMRI H, WANG J, et al. End-to-end audio visual scene-aware dialog using multimodal attention-based video features[C]//Proceedings of ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2019: 2352-2356.

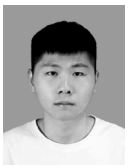
[17] MIAO Y, GOWAYYED M, METZE F. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding[C]//Proceedings of 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). [S.l.]: IEEE, 2015: 167-174.

[18] SERDYUK D, WANG Y, FUEGEN C, et al. Towards end-to-end spoken language understanding[C]// Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2018.

[19] NOWAK J, TASPINAR A, SCHERER R. LSTM recurrent neural networks for short text and sentiment classification[C]// Proceedings of International Conference on Artificial Intelligence and Soft Computing. Cham: Springer, 2017: 553-562.

[20] CHEN H, HU G, LEI Z, et al. Attention-based two-stream convolutional networks for face spoofing detection[J]. IEEE Transactions on Information Forensics and Security, 2019(15): 578-593.

作者简介:



刘谭(1997-),男,硕士研究生,研究方向:语音识别,E-mail: liutan@mail. ustc. edu.cn。

郭武(1973-),**通信作者**,男,副教授,研究方向:语音识别、声纹识别,E-mail: guowu@ustc.edu.cn。

(编辑:张彤)