Journal of Computer Applications
计算机应用, 2020, 40(10): 3034 - 3040
ISSN 1001-9081
CODEN JYIIDU
2020-10-10
http://www.joca.cn

文章编号:1001-9081(2020)10-3034-07
DOI:10.11772/j.issn.1001-9081.2020020272

基于听皮层神经元感受野的强噪声环境下说话人识别

牛晓可[1,2*],黄伊鑫[1],徐华兴[1,2],蒋震阳[1]
(1. 郑州大学 电气工程学院,郑州 450001; 2. 河南省脑科学与脑机接口技术重点实验室(郑州大学),郑州 450001)
(*通信作者电子邮箱 niuxiaoke@zzu.edu.cn)

摘 要:针对说话人识别易受环境噪声影响的问题,借鉴生物听皮层神经元频谱-时间感受野(STRF)的时空滤波机制,提出一种新的声纹特征提取方法。在该方法中,对基于STRF获得的听觉尺度-速率图进行了二次特征提取,并与传统梅尔倒谱系数(MFCC)进行组合,获得了对环境噪声具有强容忍的声纹特征。采用支持向量机(SVM)作为分类器,对不同信噪比(SNR)语音数据进行测试的结果表明,基于STRF的特征对噪声的鲁棒性普遍高于MFCC系数,但识别正确率较低;组合特征提升了语音识别的正确率,同时对环境噪声具有良好的鲁棒性。该结果说明所提方法在强噪声环境下说话人识别上是有效的。

关键词:听皮层;频谱-时间感受野;梅尔倒谱系数;含噪说话人识别;支持向量机
中图分类号:TP391.4 **文献标志码**:A

Speaker recognition in strong noise environment based on auditory cortical neuronal receptive field

NIU Xiaoke[1,2*], HUANG Yixin[1], XU Huaxing[1,2], JIANG Zhenyang[1]
(1. *School of Electrical Engineering, Zhengzhou University, Zhengzhou Henan* 450001, *China*;
2. *Henan Key Laboratory of Brain Science and Brain-Computer Interface Technology (Zhengzhou University), Zhengzhou Henan* 450001, *China*)

Abstract: Aiming at the problem that speaker recognition is susceptible to environmental noise, a new voiceprint extraction method was proposed based on the spatial-temporal filtering mechanism of Spectra-Temporal Receptive Field (STRF) of biological auditory cortex neurons. In the method, the quadratic characteristics were extracted from the auditory scale-rate map based on STRF, and the traditional Mel-Frequency Cepstral Coefficient (MFCC) was combined to obtain the voiceprint features with strong tolerance to environmental noise. Using Support Vector Machine (SVM) as feature classifier, the testing results on speech data with different Signal-to-Noise Ratios (SNR) showed that the STRF-based features were more robust to noise than MFCC coefficient, but had lower recognition accuracy; the combined features improved the accuracy of speech recognition and had good robustness to noise. The results verify the effectiveness of the proposed method in speaker recognition under strong noise environment.

Key words: auditory cortex; Spectral-Temporal Receptive Field (STRF); Mel-Frequency Ceptral Coefficient (MFCC); noisy speaker recognition; Support Vector Machine (SVM)

0 引言

生物识别技术在过去几十年得到了广泛研究与应用,说话人识别作为仅次于掌纹和指纹识别的第三大生物特征识别技术,目前世界市场占有率为15.8%,并有逐年上升的趋势。相较于指纹和掌纹这些生物特征识别技术,声纹识别技术发展较晚,但在应用上因具备语音提取方便、适合远程身份确认等特点而具有明显优势。该技术的实现原理主要为声纹特征的提取与匹配,即:首先,从与文本不相关的语音片段中提取出说话人的声纹特征;然后,建立对应的说话人模型即声纹数据库,最后,在测试时采用相同特征提取方法与说话人模型,获取被测试说话人的语音特征,并与声纹数据库中的特征进行匹配,根据匹配结果判决说话人的身份。总的来讲,说话人识别技术的研究可概括为声纹特征参数的提取与说话人模型构建(或称为特征匹配/分类)。

在声纹特征参数的提取方面,梅尔倒谱系数(Mel-Frequency Cepstral Coefficient, MFCC)是较为常用的,操作简单、样本量小。MFCC主要描述了声道特征,在没有噪声时有很好的特征表达,但在高噪声存在时鲁棒性会明显降低[1]。针对噪声环境下语音识别系统的鲁棒性问题,目前已经有很多学者提出了不同的方法,典型的方法主要有:感知听觉场景分析、小波变换法、模型补偿法的鲁棒语音识别分析、信号空间的鲁棒语音识别分析和模拟生物听觉感知特性法[2]。感知听觉场景分析能在多噪声环境中清楚分离出目标语音信号,但是会出现一定的信号缺少。王凯龙等[3]基于计

收稿日期:2020-04-12;**修回日期**:2020-06-01;**录用日期**:2020-06-08。 **基金项目**:国家自然科学基金资助项目(11804309)。
作者简介:牛晓可(1987—),女,河南平顶山人,讲师,博士,主要研究方向:生物信息建模、生物信号处理; 黄伊鑫(1994—),男,河南许昌人,硕士研究生,主要研究方向:基于人耳听觉感知特性的鲁棒说话人识别; 徐华兴(1988—),男,河南驻马店人,讲师,博士,主要研究方向:声音信号处理; 蒋震阳(1997—),男,河南洛阳人,硕士研究生,主要研究方向:生物信息建模。

算听觉场景分析理论，对单通道多说话人混合语音分离问题进行了研究，该方法在消除多种典型噪声干扰方面能得到较好的效果。小波变换法具有多分辨率分析的特点，能够通过选择不同的尺度以减小噪声对信号的影响，从而提高对语音信号的特征提取的正确率。而模型补偿法以及针对信号空间的方法中心思想即在信号空间消除噪声的影响，以维纳滤波、谱估计、语音增强为代表。张靖等[4]针对环境噪声的多变性导致训练时无法预测实际应用中的环境噪声的问题，引入环境自学习和自适应思想，通过改进的矢量泰勒级数（Vector Taylor Series，VTS）刻画环境噪声模型和说话人语音模型之间的统计关系，于2020年提出了一种具有环境自学习能力的鲁棒说话人识别算法，该算法在高信噪比（Signal-to-Noise Ratio，SNR）条件下的识别率以及对噪声的鲁棒性均有所提升，但低信噪比条件下的性能仍存在不足。近几年，深度神经网络（Deep Neural Network，DNN）逐渐成为学者们研究的重点，主流的方法有深度信念网络（Deep Belief Network，DBN）、卷积神经网络（Convolutional Neural Network，CNN）和循环神经网络（Recurrent Neural Network，RNN）。顾婷[5]在2019年利用CNN构造了一种CNN融合特征，使识别率有明显提升，但受网络层数影响较大；同年，赵飞[6]提出了一种基于DNN的语音分离和说话人确认联合训练框架，该框架将语音分离部分产生的对噪声具有鲁棒性的特征应用在说话人确认网络，能够显著提高说话人识别的正确率。但是深度学习的方法的缺陷也很明显，即对样本量依赖较大，给该技术在实际应用领域带来一些影响；并且，随着信噪比的降低，识别率的下降较为严重，强噪声环境下该技术的鲁棒性明显降低。但是生物的听觉系统对噪声却具有很强的鲁棒性，即使在信噪比极低的条件下，依然有很高的识别率，因此近些年来模拟生物听觉特性进行语音识别的方法越来越受到研究者的青睐。典型的代表是：Chi等[7]于2005年首次将生物听皮层神经元频谱-时间感受野（Spectra-Temporal Receptive Field，STRF）的概念引用到了简单的语音处理中，并提出了一套神经计算模型，解释了从外部输入的声音信号是如何转换为大脑皮层传递的电信号。2012年，Patil等[8]利用该神经计算框架，模拟了听觉皮层神经元的活动，实现了在不考虑音高和演奏风格的情况下进行稳健的乐器分类，正确率为98.7%。进一步地，2015年，Carlin等[9]从听觉神经生理学的角度出发，构建了一个任务驱动下的STRF在频域的可塑性计算模型，展示了STRF如何在抑制与各种非语音相关的声音的同时，通过调整其时频感受野特性来提高对语音的识别性能，即皮层过滤器从一些“默认”调整到任务最优形状，以增强任务相关特征的神经响应，同时抑制干扰物的神经响应。同年Carlin等[10]又提出了任务驱动的STRF自适应调整策略，可以改善特定语音事件的检测性能，并设计了一个刺激重建任务。通过在干净和加性噪声条件下进行测试对比的结果表明，任务驱动下的STRF自适应模型对语音的处理具备更高的保真度，显著提升了噪声环境下语音信号处理的鲁棒性。另外在其他领域，针对不同的含噪语音信号，该模型也体现出了良好的抗噪声能力。2018年，Emmanouilidou等[11]采用基于STRF的模型，将含噪（包括环境声、心脏杂音、哭声等）的肺音信号投射到频谱-时间特征空间中，对大于1 000例儿童的肺音信号进行识别的结果表明，该方法表现出了对噪声的鲁棒性，能够有效识别病患与健康人之间的肺音信号，正确率高达86.7%。

在说话人建模和模式匹配方面，早在20世纪80年代就提出了动态时间规整、矢量化、隐马尔可夫以及人工神经网络，并成功得到了应用[12-15]。到了20世纪90年代，高斯混合模型（Gaussian Mixture Model，GMM）和支持向量机（Support Vector Machine，SVM）模型相继被提出。2000年以来，林肯实验室的Reynolds等[16]提出一种所需样本较少且花费时间更短的高斯混合模型通用背景模型，使说话人识别向实用领域迈进了一大步。在此模型的基础上，Campbell等[17]于2006年提出了高斯超向量（Supervector）的概念，并应用在了GMM-UBM（GMM-Universal Background Model）和高斯超向量-支持向量机（Gaussian Super Vector-Support Vector Machine，GSV-SVM）的结合模型中。紧接着在2008年，Kenny等[18]在前人超矢量的基础上提出了联合因子分析（Joint Factor Analysis，JFA）方法，已有研究利用联合因子分析算法去除信道干扰，得到与信道无关的说话人因子，减少了多信道条件下对目标语音的干扰。2011年Dehak等[19]又在此基础上提出了I-Vector方法，使文本无关的说话人识别系统的性能有了更大的提升。然而分类器对识别率的提高相对较为有限，说话人识别性能提升的关键在于有效特征参数的提取，因此本文将侧重点放在文本无关语音信号的特征提取上。

针对目前主流的说话人识别算法所存在的问题，即强噪声环境下识别率下降较为严重，本文提出了一种基于STRF与MFCC组合特征的声纹特征提取方法，对噪声环境下说话人语音信号的识别具有较强的鲁棒性。首先，采用对数频谱幅度（Optimally Modified Log-Spectral Amplitude，OM-LSA）[20]语音估计与改进的最小控制递归平均（Improved Minima Controlled Recursive Averaging，IMCRA）[21]噪声估计结合的方法对说话人语音进行降噪等预处理；然后，利用STRF模型将语音信号投射到特定的频谱-时间空间，并进一步提取听觉谱图的二次特征与MFCC系数进行组合；最后，采用常规的支持向量机[22]对声纹特征进行分类与识别。对来自清华大学中文语音库thchs30的36个说话人（每人40段语音）数据进行测试，通过加入不同等级噪声后的对比结果表明，本文方法在低信噪比条件（-10 dB）下仍然能够得到较高的识别正确率（86.68%），从而验证了本文方法对强噪声环境下说话人识别的鲁棒性[23]。

1　算法原理

当外界声音由外耳道传到鼓膜，经鼓膜震动传递到听小骨，经听小骨传到耳蜗，这时听觉感受器接受刺激兴奋，通过感受器官中的向心神经元将神经冲动传到听皮层，引起听皮层神经元产生神经冲动，进而形成听觉感知。诱发听皮层神经元产生神经冲动的刺激区域称为听皮层神经元的感受野。而听皮层神经元的感受野具有一定的频段和时间选择性，因此又称为频谱-时间感受野（STRF），可看作是一个时间和频率上的二维滤波核，反映了神经元对特定频带和特定周期特征声音信号的线性处理特性。一个典型的STRF滤波核如图1所示。

在哺乳动物初级听觉皮层中，STRF对广泛的声学特征表现出详细的敏感性，并对表征自然声音的时域包络和频域特征缓慢变化的频谱-时间能量调制具有选择性，而对没有特定统计特性的环境噪声不敏感，因此经STRF滤波后的声音信号理论上对嘈杂的环境声具有较高的容忍性。

此外，除了其固有的调谐到特定的声音调制的信号，皮层神经元可以动态调整其过滤性能。当认知资源指向一个感兴趣的声音时，认知反馈被认为可以诱导STRF自适应调制的能力，即皮层过滤器从一些“默认”调整到任务最优形状，以增强任务相关特征的神经响应[6]，同时抑制干扰物的神经响应。这种自适应调制的模式在其他生理感觉模式（视觉）中也观察到类似的效应。本文利用了STRF的以上特性，针对特定类型的噪声，通过手动调节STRF模型的相关参数，以获得其对特定类型环境噪声较高的容忍性。

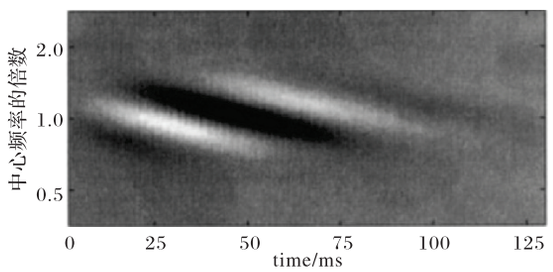


图1 二维STRF滤波器核范例（向下，Ω=1 cyc/oct，ω=16 Hz）

Fig. 1 Example of two dimensional STRF filter kernel (Downward, Ω=1 cyc/oct, ω=16 Hz)

2 本文方法

本文所提出的基于STRF与MFCC组合特征的说话人识别方法主要包括语音信号的预处理、声纹的特征提取与特征分类三个部分，每一部分的具体计算过程如下。

2.1 语音信号的预处理

本文采用OM-LSA与IMCRA噪声估计结合的方法对含噪声的语音信号进行预处理，预流程如图2所示。

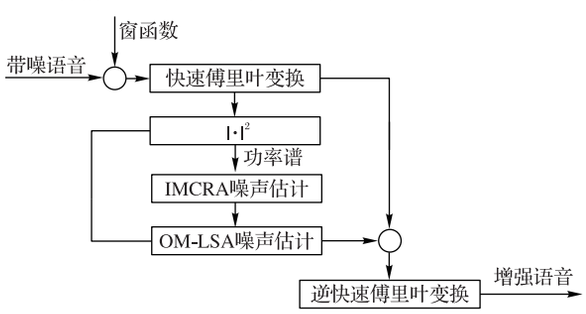


图2 OM-LSA与IMCRA结合的预处理流程

Fig. 2 Flowchart of pre-processing process combining OM-LSA and IMCRA

预处理过程可概括为：

首先，根据IMCRA算法估计含噪语音的时变功率谱分布。

然后，根据估计的功率谱分布，结合OM-LSA算法来增强瞬态噪声和非瞬态噪声成分的差异，并估计瞬态噪声的功率谱分布。另一方面，采用IMCRA算法，从瞬态噪声和语音信号中估计背景噪声的功率谱分布。

最后，将估计的瞬态噪声和背景噪声功率谱分布进行合并，运用OM-LSA算法同时抑制瞬态噪声和背景噪声，得到增强后的语音信号。

2.2 声纹特征提取

2.2.1 基于STRF的声纹特征

基于STRF的声纹特征提取包括三个阶段的处理过程：第一个阶段模拟了生物听觉系统的外周模型，即耳蜗核的处理过程，将输入的语音信号转化为听觉外周的频谱图；第二个阶段是模拟了听皮层神经元感受野的处理过程，将第一阶段输出的频谱图转化为特定尺度的尺度-速率谱图；第三个阶段的处理就是对第二阶段生成的尺度-速率谱图进一步做二次特征提取。

1）听觉外周模型的处理过程。

听觉外周系统的模型处理流程如图3所示。计算过程描述为：

首先，将音频信号$s(t)$通过耳蜗滤波器组，采用式（1）对信号$s(t)$进行仿射小波变换。耳蜗滤波器组的输出用y_C表示。

$$y_C(t,f) = s(t) *_t h(t,f) \tag{1}$$

其中：$h(t,f)$为各滤波器的脉冲响应；$*_t$为时域卷积运算。

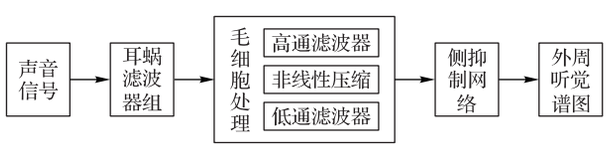


图3 外周听觉系统的模型处理流程

Fig. 3 Model framework of peripheral auditory system

然后，耳蜗输出y_C经过毛细胞的处理转化成听觉神经响应，表示为y_A。毛细胞的处理主要包括高通滤波、非线性压缩$g(\cdot)$和低通滤波器$w(t)$。数学描述如下：

$$y_A(t,f) = g(\partial_t y_C(t,f)) *_t w(t) \tag{2}$$

进一步地，经由耳蜗核的侧抑制网络作用，以模拟耳蜗核的频率选择性。表达式如下：

$$y_{LIN}(t,f) = \max(\partial_f y_A(t,f), 0) \tag{3}$$

利用短窗口函数$\mu(t,\tau)$与$y_{LIN}(t,f)$求卷积，得到第一阶段的输出$y(t,f)$。

$$y(t,f) = y_{LIN}(t,f) *_t \mu(t,\tau) \tag{4}$$

$$\mu(t,\tau) = e^{\frac{-t}{\tau}} u(t) \tag{5}$$

其中：τ是微秒级别的时间常数。

任取一段语音信号，加入不同信噪比噪声（工厂车间噪声，取自NoiseX92数据库），经第一阶段处理后的外周听觉谱图如图4所示。

2）听皮层神经元感受野模型的处理过程。

该阶段的处理是通过模拟听皮层神经元的频谱-时间感受野（STRF）特性来实现，主要采用一组具备不同时频域特征选择性的滤波器模拟，这些特征包括时域中从缓慢变化到快速骤变的节律（rate）和频率域从较窄到较宽的尺度（scale）信息。

该组滤波器的输出是预处理后的声音经过第一阶段处理得到的时频谱图与上述滤波器的卷积。因此，由第一阶段输出的时频谱图如果与某个滤波器所选择的节律和尺度较为吻合，则会在相对应的特征点处输出较大值。由此得出，该阶段的处理结果是经一系列滤波器特征选择后结果的组合。具体该阶段数学描述如下：

首先，构造STRF滤波器。STRF滤波器可看作是空间脉冲响应h_S与时间脉冲响应h_T的乘积。分别定义如下：

$$h_S(f,\omega,\theta) = h_{scale}(f,\omega)\cos\theta + \hat{h}_{scale}(f,\omega)\sin\theta \tag{6}$$

$$h_{\mathrm{T}}(t,\Omega,\phi)=h_{\mathrm{rate}}(t,\Omega)\cos\phi+\hat{h}_{\mathrm{rate}}(t,\Omega)\sin\phi \tag{7}$$

其中:Ω、ω分别表示滤波器的谱密度和速率参数;ϕ、θ表示特征相位。$\hat{h}$表示希尔伯特变换,定义为:

$$\hat{h}(x)=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{h(x)}{z-x}\mathrm{d}z \tag{8}$$

分别采用二阶高斯模型和伽马方程模拟h_{scale}和h_{rate},对应的方程表示为:

$$h_{\mathrm{scale}}(x)=(1-x^2)\mathrm{e}^{-\frac{x^2}{2}} \tag{9}$$

$$h_{\mathrm{rate}}(t)=t^3\mathrm{e}^{-4t}\cos(2\pi t) \tag{10}$$

不同频率和尺度的脉冲响应采用下述方式进行扩展。

$$h_{\mathrm{scale}}(x,\Omega)=\Omega h_{\mathrm{scale}}(\Omega x) \tag{11}$$

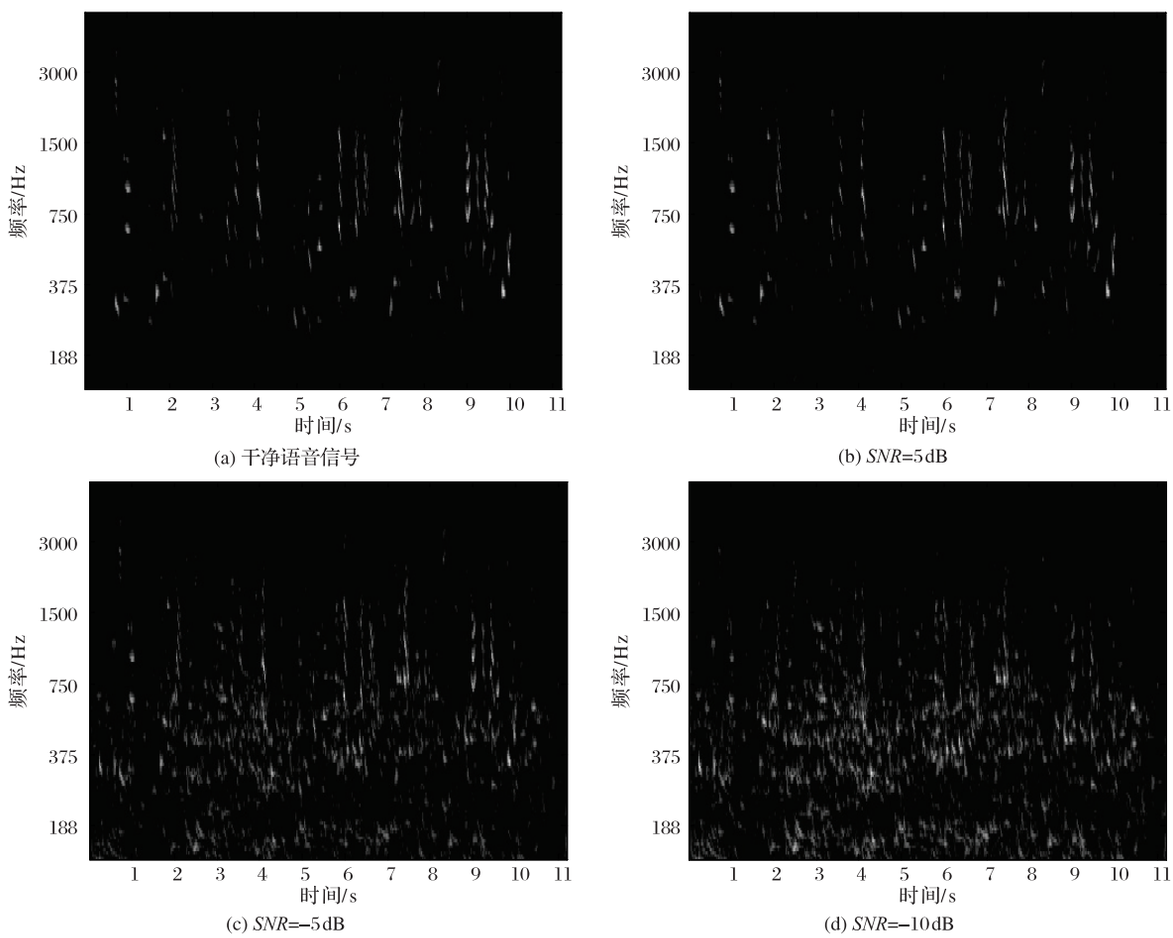
$$h_{\mathrm{rate}}(t,\omega)=\omega h_{\mathrm{rate}}(\omega x) \tag{12}$$

然后,计算该阶段输出的响应,表示为:

$$STRF(t,f,\omega,\phi,\theta)=y(t,f)*_{tf}[h_S(f,\omega,\theta)\cdot h_T(t,\Omega,\phi)] \tag{13}$$

其中:$*_{tf}$为时域和频域的卷积运算。

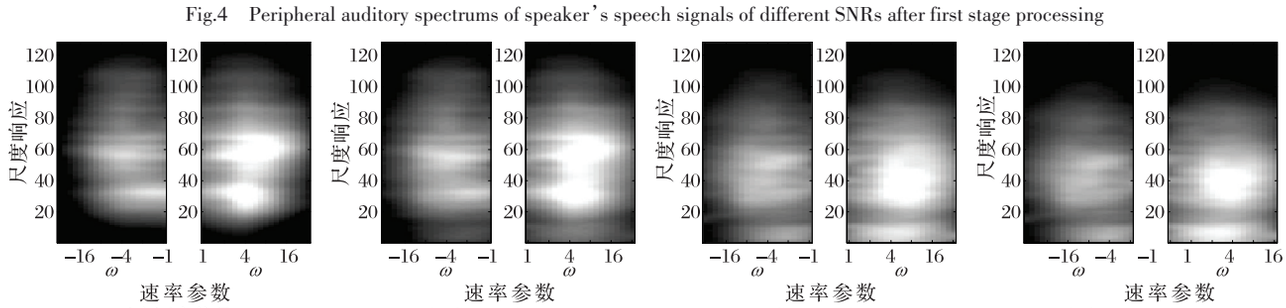
图5为STRF第二阶段输出的时间频谱图,即尺度-速率谱图,横轴代表速率参数ω,纵轴为尺度响应(空间脉冲响应和时间脉冲响应的乘积)。随着图5中响应区域的不同及变化,对应了感受野对该语音信号的兴奋和抑制。所得结果反映了听觉皮层神经元对特定频率和尺度能量选择后的结果,即颜色高亮深色的区域代表了皮层神经元对特定频率和尺度能量的选择识别,极大程度减少随机噪声的影响,保留声纹信号中较为稳定的特征信息。图4所示经第一阶段输出的听觉谱图进一步经由皮质阶段的模型处理后的结果如图5所示。



(a) 干净语音信号　(b) *SNR*=5 dB　(c) *SNR*=−5 dB　(d) *SNR*=−10 dB

图4　不同信噪比说话人语音信号经第一阶段处理后的外周听觉谱图

Fig.4　Peripheral auditory spectrums of speaker's speech signals of different SNRs after first stage processing



(a) 干净语音信号　(b) *SNR*=5 dB　(c) *SNR*=−5 dB　(d) *SNR*=−10 dB

图5　第二阶段输出的尺度-速率谱图

Fig.5　Scale-rate spectrum output at second stage

3)频域-时间谱图的二次特征提取。

本文进一步地从第二阶段听觉皮层模型生成的听皮层谱图中提取了三种基于STRF的二次特征,包括每个尺度的能量S、对数尺度能量S_{L}、对数尺度能量的离散余弦变换(Discrete Cosine Transform,DCT)系数S_{DL}。其中,第一个特征S采用式(14)计算,即将第二阶段输出的时频谱图中所有尺度和速率对应的结果直接叠加。

$$S(t,\omega)=\sum_{f}\sum_{\omega}\left|STRF(t,f,\omega,0,0)\right|;\ \omega=1,2,\cdots,N_{\omega} \tag{14}$$

其中:N_{ω}是比例数。注意,等式中的相位特征ϕ和θ都设置为零。

第二个特征是S_{L}采用式(15)计算,即对第一个特征S进行对数运算。

$$S_{\mathrm{L}}(t,\omega)=\lg(S(t,\omega));\ \omega=1,2,\cdots,N_{\omega} \tag{15}$$

第三个特征S_{DL}是采用式(16),在第二个特征的基础上进行了离散余弦变换。

$$S_{\mathrm{DL}}(t,k)=\sum_{\omega=1}^{N_{\omega}}S_{\mathrm{L}}(t,\omega)\cos\left(\frac{2\pi\omega k}{N_{\omega}}\right);\ k=1,2,\cdots,N_{k} \tag{16}$$

其中:N_k是第三特征$S_{\mathrm{DL}}(t,k)$的特征指数,$N_k\leqslant N_{\omega}$。

2.2.2 MFCC系数

MFCC是基于人耳听觉感知特性的倒谱参数,在频域,人耳听到的声音高低与频率不成线性关系;但在Mel域,人耳感知与Mel频率是成正比的。它与频率的换算关系采用式(17)计算:

$$Mel(f)=2\,595\times\lg\left(1+\frac{f}{700}\right) \tag{17}$$

其中:f为频率,单位Hz。

MFCC系数的提取过程如图6所示,具体概括为:①对语音进行预加重、分帧和加窗;②对每一个短时分析窗,通过快速傅里叶变换(Fast Fourier Transformation, FFT)得到对应的频谱;③将上面的频谱通过Mel滤波器组得到Mel频谱;④在Mel频谱上面进行倒谱分析(即进行取对数和离散余弦变换(Discrete Cosine Transform, DCT)运算);⑤取DCT后的第2个到第13个系数作为MFCC系数。

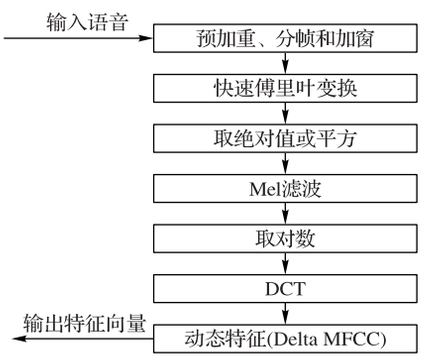


图6 MFCC系数的提取流程

Fig. 6 MFCC coefficient extraction process

2.2.3 基于MFCC和STRF组合特征

本文使用的两种特征分别为MFCC与基于STRF提取的三个二次特征。这两类特征都从不同的侧面反映了不同的说话人信息,通过有效的融合能更加全面地表征出说话人特征,本文将提取的13维MFCC系数与基于STRF的三个二次特征(分别为13维)分别组合,扩展得到三组26维的组合特征。最后在纯净语音下,对比了基于MFCC特征和基于STRF提取的三个二次特征的识别正确率;并在加入不同信噪比等级噪声下,对比了三种组合特征的识别正确率。

2.3 分类器的选取

支持向量机是20世纪80年代提出的一种特征分类方法,在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势,已经在模式识别、函数逼近和概率密度估计等方面取得良好效果。

本文采用带有径向基函数内核的多类支持向量机对说话人数据进行分类,径向基函数内核的γ值设置为2,其他参数选择LIBSVM(LIBrary for Support Vector Machines)工具的默认设置。

3 实验与结果分析

3.1 数据来源

本文采用清华大学thchs30中文语料库作为数据库来源,共选取了其中36个说话人每人40段语音片段做样本,共计1 440个语音片段。将所有语音片段分为8组,随机选取1组,即180段语音片段(每个说话人5段语音片段)作训练集,余下7组语音数据分别加入SNR为-10 dB、-5 dB、5 dB、10 dB、15 dB、20 dB的Babble噪声作测试集,共交叉验证8次,最终的识别正确率以"平均值±标准差"的形式给出。

在实验中,所有语音片段分为16 ms的帧,重叠8 ms,并将汉明窗应用于每个帧。STRF的尺度参数设置为2^{n},n=-5,-4,-3,-2,-1,1,2,3,4,5共10个等级。

3.2 单一特征对识别结果的影响

本文共提取了四个特征,包括MFCC系数特征和基于STRF的三个二次特征(能量总和S、对数运算后的能量S_{L}和离散余弦变换后的S_{DL})。首先,对比了基于单一特征的干净说话人语音识别结果。多次交叉验证的统计结果汇总在表1中。

表1 基于单一特征的说话人识别统计结果

Tab. 1 Statistical results of speaker recognition based on single feature

声纹特征		识别正确率/%
MFCC		94.12±0.10
基于STRF的特征	S	52.49±1.36
	S_{L}	50.77±1.10
	S_{DL}	69.26±0.87

从表1中可以看出,基于MFCC系数特征的识别率最高,平均识别正确率达到94.12%;而基于STRF的二次特征中,经离散余弦变换后的S_{DL}的识别率最高,但是都显著低于基于MFCC系数特征。由此可以看出,对于纯净说话人语音的识别,基于单一STRF的特征并不占优势。

3.3 组合特征对识别结果的影响

接下来尝试将基于STRF的单一特征与MFCC系数特征进行组合,对比STRF的特征是否有助于提升对说话人的识别性能。基于不同组合特征的说话人识别正确率统计结果汇总在表2中。

表2　基于不同组合特征的说话人识别统计结果

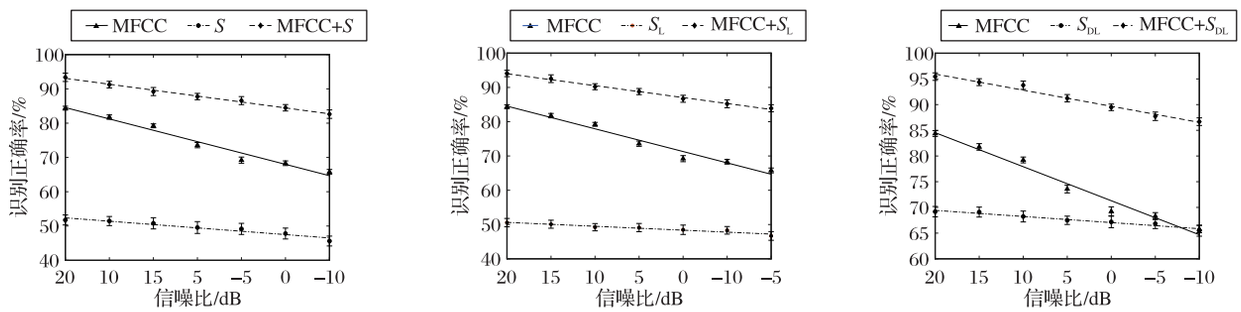
Tab. 2　Statistics of speaker recognition based on different combinations of features

不同组合声纹特征	识别正确率/%
MFCC+S	97.12±0.61
MFCC+S_{L}	96.88±0.64
MFCC+S_{DL}	97.85±0.53

通过对比表1~2的结果可以看出，对于纯净说话人语音信号，所有基于组合特征的识别率均显著高于基于单一特征的识别率。其中，采用经离散余弦变换后的S_{DL}和MFCC系数特征的组合形式取得了最高的识别正确率，高达97.85%。

3.4　不同特征对环境噪声的鲁棒性分析

进一步分析了单一声纹特征以及各种组合声纹特征对环境噪声的鲁棒性。每种组合特征与单一特征的对比结果如图7所示。图7中每个信噪比识别结果为交叉验证识别结果的平均正确率±标准差形式。



(a) MFCC、S和MFCC+S对比　(b) MFCC、S_{L}和MFCC+S_{L}对比　(c) MFCC、S_{DL}和MFCC+S_{DL}对比

图7　基于不同特征的识别率随信噪比的变化趋势

Fig. 7　Trend of recognition rate varying with different SNRs based on different features

从图7可以看出，随着信噪比下降，无论是单一特征还是组合特征，均影响说话人识别性能，其正确率有不同程度下降。其中，对于单一特征而言，基于MFCC的相对识别正确率较高，但是对于噪声容忍性较差，下降较为迅速。

图7中的线条为采用线性函数拟合后的结果。每条拟合直线的斜率表示识别性能受噪声影响的程度，斜率的绝对值越高表示对噪声的鲁棒性越差。不同特征识别率随噪声变化的斜率对比结果如表3所示。

表3　不同声纹特征组合对噪声的鲁棒性对比

Tab. 3　Robustness comparison of different features to noise

MFCC	S	S_{L}	S_{DL}	MFCC+S	MFCC+S_{L}	MFCC+S_{DL}
0.663	0.195	0.114	0.119	0.345	0.347	0.314

从表3可以看出，基于组合特征与STRF的特征对噪声鲁棒性均优于MFCC。因此，将基于STRF特征与MFCC特征进行组合，既能提高总体识别正确率（普遍高于相同信噪比条件下基于单一声纹特征的识别率），同时又能提升对噪声的容忍性，在信噪比低至-10 dB情况下，仍达到86.68%的平均正确率。同时，在与文献[4]中提出的具有环境自学习机制的鲁棒说话人识别算法相比，在低信噪比条件下（0 dB），本文提出的方法的识别率（89.47%）明显高于前者（63.3%）。以上结果说明了本文方法在强环境噪声下的说话人识别上具有一定优势。

4　结语

本文针对说话人识别易受环境噪声影响的问题，提出了基于生物听觉感知机理的声纹特征提取方法，用于说话人识别中，提升了对环境噪声的鲁棒性。首先，采用对数频谱幅度（OM-LSA）语音估计与最小控制递归平均（MCRA）噪声估计结合的方法对说话人语音进行降噪等预处理，在模拟外周听觉系统耳蜗核处理过程的基础上，进一步模拟了STRF对特定频率变化速率与尺度的特征选择性，以获取含噪语音信号中的稳定特征，通过所提出的基于STRF的听觉模型，输出代表说话人信息的频谱图，并通过频谱图进一步提取二次特征；之后与传统的MFCC处理方式相结合，得出三种组合的二次特征，分别是MFCC+S、MFCC+S_{L}、MFCC+S_{DL}；最后采用支持向量机对声纹特征进行分类识别。本文从清华大学thchs30中文语料库里选取了其中36个说话人每人40段语音片段做样本，共计1 440个语音片段。对其加入不同信噪比等级的噪声进行实验。实验重点进行了两个方面的对比分析：一方面比较了单纯基于STRF的特征与MFCC系数的识别正确率，发现前者普遍低于后者，但是前者对噪声的鲁棒性明显优于后者；另一方面，通过将二者进行组合，并与每组单一特征的识别进行比较发现，组合特征的识别正确率普遍高于单一特征，且对噪声的鲁棒性也有显著提高。以上实验结果表明，本文方法能够用于强噪声环境下的说话人识别上，表现出了对环境噪声的强鲁棒性。

参考文献（References）

[1] 李湾湾. 说话人声纹识别的算法研究[D]. 杭州：浙江大学，2017：3.（LI W W. Research on algorithms for speaker recognition [D]. Hangzhou：Zhejiang University，2017：3.）

[2] ZHENG T F，LI L. Speaker recognition：introduction [M]// Robustness Related Issues in Speaker Recognition. Singapore：Springer，2017：1-14.

[3] 王凯龙. 基于计算听觉场景分析的多人语音分离方法[D]. 南京：南京理工大学，2017：5.（WANG K L. Multi-person speech separation method based on computational auditory scene analysis [D]. Nanjing：Nanjing University of Science and Technology，2017：5.）

[4] 张靖，俞一彪. 具有环境自学习机制的鲁棒说话人识别算法[J]. 通信技术，2020，53(3)：618-624.（ZHANG J，YU Y B. Robust speaker recognition algorithm with environment self-learning mechanism [J]. Communications Technology，2020，53(3)：618-624.）

[5] 顾婷. 基于深度特征的说话人辨认技术研究[D]. 南京：南京邮电大学，2019：43-46.（GU T. Research of speaker identification technology based on deep features [D]. Nanjing：Nanjing

University of Posts and Telecommunications, 2019: 43-46.)
[6] 赵飞. 基于深度神经网络的鲁棒性说话人确认方法研究[D]. 呼和浩特:内蒙古大学, 2019: 11. (ZHAO F. Research on robust speaker confirmation method based on deep neural network [D]. Hohhot: Inner Mongolia University, 2019: 11.)
[7] CHI T, RU P, SHAMMA S A. Multiresolution spectrotemporal analysis of complex sounds [J]. The Journal of the Acoustical Society of America, 2005, 118(2): 887-906.
[8] PATIL K, PRESSNITZER D, SHAMMA S, et al. Music in our ears: the biological bases of musical timbre perception [J]. PLoS Computational Biology, 2012, 8(11): No. e1002759.
[9] CARLIN M A, ELHILALI M. Modeling attention-driven plasticity in auditory cortical receptive fields [J]. Frontiers in Computational Neuroscience, 2015, 9: No. 106.
[10] CARLIN M A, ELHILALI M. A framework for speech activity detection using adaptive auditory receptive fields [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(12): 2422-2433.
[11] EMMANOUILIDOU D, MCCOLLUM E D, PARK D E, et al. Computerized lung sound screening for pediatric auscultation in noisy field environments [J]. IEEE Transactions on Bio-Medical Engineering, 2018, 65(7): 1564-1574.
[12] DODDINGTON G R. Speaker recognition — identifying people by their voices [J]. Proceedings of the IEEE, 1985, 73(11): 1651-1664.
[13] SHIKANO K. Text-independent speaker recognition experiments using codebooks in vector quantization [J]. Journal of the Acoustical Society of America, 77(S1): S11-S11.
[14] WAIBEL A. Modular Construction of time-delay neural networks for speech recognition [J]. Neural Computation, 1989, 1(1): 39-46.
[15] REYNOLDS D A, ROSE R C. Robust text-independent speaker identification using Gaussian mixture speaker models [J]. IEEE Transactions on Speech and Audio Processing, 1995, 3(1): 72-83.
[16] REYNOLDS D A, QUATIERI T F, DUNN R B. Speaker verification using adapted Gaussian mixture models [J]. Digital Signal Processing, 2000, 10(1/2/3): 19-41.
[17] CAMPBELL W M, STURIM D E, REYNOLDS D A, et al. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation [C]// Proceedings of the 2006 IEEE International Conference on Acoustics Speech and Signal Processing. Piscataway: IEEE, 2006: I-I.
[18] KENNY P, OUELLET P, DEHAK N, et al. A study of interspeaker variability in speaker verification [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2008, 16(5): 980-988.
[19] DEHAK N, KENNY P J, DEHAK R, et al. Front-end factor analysis for speaker verification [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(4): 788-798.
[20] 刘凤增,李国辉,李博. OM-LSA和小波阈值去噪结合的语音增强[J]. 计算机科学与探索, 2011, 5(6): 547-552. (LIU F Z, LI G H, LI B. Speech enhancement with OM-LSA Incorporating wavelet thresholding [J]. Journal of Frontiers of Computer Science and Technology, 2011, 5(6): 547-552.)
[21] 张建伟,陶亮,周健,等. 基于改进谱平滑策略的IMCRA算法及其语音增强[J]. 计算机工程与应用, 2017, 53(1): 153-157. (ZHANG J W, TAO L, ZHOU J, et al. Improved minima controlled recursive averaging algorithm based on improved spectrum smoothing strategy and speech enhancement [J]. Computer Engineering and Applications, 2017, 53(1): 153-157.)
[22] 周于皓,张红玲,李芳菲,等. 局部关注支持向量机算法[J]. 计算机应用, 2018, 38(4): 945-948. (ZHOU Y H, ZHANG H L, LI F F, et al. Local attention support vector machine algorithm [J]. Computer applications, 2018, 38(4): 945-948)
[23] 崔锐. 噪声环境下鲁棒性说话人识别算法研究[D]. 西安:西安电子科技大学, 2017: 34. (CUI R. Research on robust speaker recognition algorithm in noisy environment [D]. Xi'an: Xidian University, 2017: 34).

This work is partially supported by the National Natural Science Foundation of China (11804309).

NIU Xiaoke, born in 1987, Ph. D., lecturer. Her research interests include biological information modeling, biosignal processing.

HUANG Yixin, born in 1994, M. S. candidate. His research interests include robust speaker recognition based on human auditory perception.

XU Huaxing, born in 1988, Ph. D., lecturer. His research interests include sound signal processing.

JIANG Zhenyang, born in 1997, M. S. candidate. His research interests include biological information modeling.