文章编号：1003-0530(2023)12-2254-11

基于多源语音信息融合的帕金森病辅助检测方法

季 薇[1] 王传瑜[1] 李 云[2] 郑慧芬[3]
（1. 南京邮电大学通信与信息工程学院，江苏南京 210003；2. 南京邮电大学计算机学院，江苏南京 210023；
3. 南京医科大学附属老年医院，江苏南京 210024）

摘 要： 在病程早期，帕金森病患者由于发声器官的灵活协调能力下降，会出现发音困难、发音不稳定等症状。为分析受试者的言语能力，专家基于上述生理现象设计了包括持续元音、重复音节以及情景对话在内的多类型语料。已有的帕金森病语音检测研究大多基于单类型语料，可评估受试者部分声学器官的协调能力，但无法全面地反映受试者的发声状况，且易受采集环境、个体差异等因素的影响。针对上述问题，本文提出一种用于帕金森病辅助检测的多源语音信息融合模型，旨在充分利用多类型语料获得的多源语音数据，提取丰富全面的病理信息，抵御非病理性因素的影响。所提模型由编码器模块、解码器模块和分类器模块组成。其中，编码器模块通过多个支路分别学习各单源语音数据中的特有信息，并通过一个基于多头注意力机制的多源信息融合分支实现更细粒度的信息交互，学习多源语音数据的共有信息，从而全面提取多源数据所携带的病理信息；解码器模块帮助编码器模块实现信息压缩去冗余；分类器模块根据编码器输出完成帕金森病检测，并辅助编码器模块学习紧凑的病理信息表示。为进一步确保特有信息和共有信息的提取，模型对特有信息和共有信息实施了正交约束。本文在包含 340 个语音样本的自采数据集上进行了多个对比实验。实验结果显示，所提模型在帕金森病检测的准确率、敏感度和 F1 分数等各项性能指标上相较于基于单源语音数据的模型分别提高了 6%、3%、6%；同时，共有信息与特有信息的有效整合，也使得所提模型相较于其他信息融合模型在准确率指标上提高了 2.8% 以上。

关键词： 帕金森病；语音信号处理；多源信息融合；深度学习
中图分类号： TP391.4 **文献标识码：** A **DOI：** 10.16798/j.issn.1003-0530.2023.12.012

引用格式： 季薇，王传瑜，李云，等. 基于多源语音信息融合的帕金森病辅助检测方法[J]. 信号处理，2023，39(12)：2254-2264. DOI：10.16798/j.issn.1003-0530.2023.12.012.

Reference format： JI Wei，WANG Chuanyu，LI Yun，et al. Auxiliary detection method of Parkinson's disease based on multi-source speech information fusion[J]. Journal of Signal Processing，2023，39(12)：2254-2264. DOI: 10.16798/j.issn.1003-0530.2023.12.012.

Auxiliary Detection Method of Parkinson's Disease Based on Multi-Source Speech Information Fusion

JI Wei[1] WANG Chuanyu[1] LI Yun[2] ZHENG Huifen[3]
（1. School of Communications and Information Engineering，Nanjing University of Posts and Telecommunications，Nanjing，Jiangsu 210003，China；2. School of Computer Science，Nanjing University of Posts and Telecommunications，Nanjing，Jiangsu 210023，China；3. Affiliated Geriatric Hospital of Nanjing Medical University，Nanjing，Jiangsu 210024，China）

Abstract: In the early stages of Parkinson's disease，patients develop symptoms such as difficulties in pronunciation and unstable articulation due to a decrease in the flexible coordination ability of the vocal organs. To analyze the speech ability of

收稿日期：2023-04-28；修回日期：2023-06-21
基金项目：江苏省高校基础科学(自然科学)重大项目(21KJA520003)

the subjects, the experts design a multi-type corpus, including sustained vowels, repetitive syllables, and situational dialogues, based on the aforementioned physiological phenomena. Existing researches on Parkinson's disease speech detection mostly rely on single-type corpora, which can evaluate the coordination ability of certain acoustic organs in the subjects but cannot comprehensively reflect the subjects' vocal conditions and are susceptible to factors such as the collection environment and individual differences. To address the aforementioned issues, a multi-source speech information fusion model for assisting Parkinson's disease detection was proposed in this paper. The aim was to fully utilize the multi-source speech data obtained from diverse types of corpora, extract comprehensive and rich pathological information, and counteract the influence of non-pathological factors. The proposed model consists of an encoder module, a decoder module, and a classifier module. In the encoder module, multiple branches are employed to learn the specific information from each individual source of speech data. Through a multi-head attention mechanism-based fusion branch, finer-grained information interaction is achieved, enabling the learning of common information present in the multi-source speech data, thereby comprehensively extracting the pathological information carried by the multi-source data. The decoder module assists the encoder module in information compression and redundancy elimination. The classifier module detects Parkinson's disease based on the output of the encoder module, while also aiding the encoder module in learning compact representations of pathological information. To further ensure the extraction of specific and common information, the model imposes orthogonal constraints on these information components. Multiple comparative experiments were conducted, based on a self-collected dataset containing 340 speech samples. The experimental results demonstrated that the proposed model outperformed the models based on single-source speech data in terms of accuracy, sensitivity, and F1 score for Parkinson's disease detection, with improvements of 6%, 3%, and 6% respectively. Moreover, the effective integration of common and specific information enabled the proposed model to achieve more than a 2. 8% improvement in accuracy compared to other information fusion models.

Key words: Parkinson's disease; speech signal processing; multi-source information fusion; deep learning

1　引言

帕金森病(Parkinson's Disease, PD)是一种中脑黑质多巴胺能神经元变性死亡引发的慢性进展性疾病[1]。由于大脑中多巴胺能神经元的进行性损失，帕金森病患者将无法稳定控制发声器官，常伴有无法稳定发音，口腔、声带、喉咙等发声器官的灵活协调能力下降等症状[2]。为分析受试者的言语能力，领域内的专家基于上述生理现象设计了包括持续元音发音(如/a/、/i/、/u/等)、重复音节(/pakala/)、情景对话等在内的多类型语料[3-4]。其中，持续元音发音涉及到声带和声道中各种肌肉的组合，能够很好地评估受试者的发音能力[5-6]；重复音节发音，能够很好地分析受试者移动齿龈、下颌和舌头等发音器官的协调能力[7-8]；情景对话朗读能够判断受试者能否正确的发出语料所暗含的语气与语调[9-10]。受试者在医学专家的指导下，根据不同类型的语料进行发音，生成用于受试者言语能力分析的原始语音数据。

近年来，基于帕金森病患者的言语能力分析开展帕金森病检测成为一种有效的辅助诊疗手段。文献[11-13]基于持续元音语音数据提取了频率微扰、振幅微扰、谐波噪声比等发音类特征，并利用帕金森病患者和健康人在这些声学特征上存在的差异，结合传统的机器学习分类模型(随机森林(Random forest, RF)、支持向量机(Support vector machine, SVM)等)实现了帕金森病的检测，准确率最高可达89%。文献[14-15]基于重复音节语音数据提取了梅尔倒谱系数(Mel Frequency Cepstral Coefficient, MFCC)、巴克带能量等发声类特征，结合机器学习分类模型(SVM、卷积神经网络等)进行帕金森病的检测，准确率最高可达90%。文献[16]基于情景对话提取了与韵律相关的特征，结合机器学习模型(K近邻、SVM等)，实现了帕金森病的检测，准确率最高可达85%。然而，单类型语料数据无法全面地表征受试者的构音能力，且易受噪声、采集环境等因素的影响导致语音质量下降。为实现多角度分析受试者构音能力，去除非病理性因素的影响，有学者尝试探索基于多类型语料获得的多源语音数据。如Bocklet等人[17]将多个单源语音数据中提取的特征进行简单的拼接实现融合，再送入分类模型进行帕金森病的分类检测。实验结果显示，结合多源语音数据的检测性能反而不如单源语音数据与分类模型相结合的

情况。其原因在于文献[17]所述的多源信息融合方式不足以充分利用多源语音数据带来的信息优势，反而造成了无关信息的累积，强化了无关信息对模型的影响，从而造成性能的下降。

由于多源语音数据来源不一致（朗读的语料不同），且每种语音的发声机理不一致，反映的言语能力不同，可将它们作为多模态数据来看待[18]。因此可借助多模态信息融合技术，解决上述信息融合问题。当前多模态信息融合技术根据融合的时机可大致分为早期融合、后期融合、混合融合[18]。早期融合的方式，通常为每个模态设计预处理网络提取单模态的高级特征，然后通过加权求和、直接拼接等操作实现多模态数据在特征层融合。文献[19]提出一种基于自编码器改进的多模自编码器，通过多个子网络完成单模态信息提取，然后在特征层拼接作为多模态融合信息。文献[20]提出一种基于多核学习的信息融合方式，通过将多模态数据经过不同的核处理，再进行核函数的加权组合实现信息融合。后期融合也称决策层融合，其通过多个独立的推断模型处理不同的单模态数据，然后整合推断结果实现多模态数据的融合。文献[21]使用了一层神经网络对来自不同模态的输出进行整合，输出最终的决策结果。前述两种融合方式均存在多模态信息交互不足的情况，因此研究人员提出了混合融合方式，旨在通过在多层级（特征层、决策层）的模态交互，充分实现信息的融合。例如许多基于多头自注意力机制的多模态融合模型，在图文结合[22]、情感语义识别[23]、机器翻译[24]等领域表现出了优异的性能，成为多模态信息融合领域内的一个主流方向。然而，这些基于多头自注意力机制技术的模型都聚焦于多模态数据间共有信息的学习，对单模态特有信息的学习缺少关注。

本文关注的基于语音的帕金森病检测这一特定任务有如下特点：一方面，帕金森病患者的语音数据不易采集，数据集规模相对较小[25]；另一方面，基于语音数据提取的声学特征维数较高并且存在信息冗余问题。这些特点导致已有的多模态信息融合模型在面对高维小样本数据时易出现过拟合现象，且大量冗余特征的存在会给模型带来更多的无效信息，干扰模型的决策，增加计算开支[26-27]。此外，前述的多模态融合模型，缺乏对单模态特有信息的关注。因此，前述各种的多模态信息融合模型无法直接应用于多源语音数据的帕金森病检测。

基于此，本文提出一种多源语音信息融合模型（Multisource Data Fusion Autoencoder，MSFAE），旨在对多源语音数据携带的病理信息进行全面整合，过滤由多个数据源融合带来的无效信息，实现病理信息的准确表达。考虑到基于情景对话语料的帕金森病语音数据，容易受到受试者的文化水平、地域性口音等无关因素的影响，而引入更多的无效信息，增强过拟合风险。所以，本文在选择多源语音数据时，着重考虑持续元音发音（/a/）以及重复音节（/pakala/）这两种语音数据。该模型包含如下几个模块：（1）编码器模块。该模块由多个并行支路（即3个子编码器）组成，其中两条支路分别提取两个单源语音数据的特有信息（对应于特有信息表征学习子模块）；一条支路作为多源信息融合子模块实现多源数据共有信息的提取。（2）解码器模块。解码器模块帮助编码器模块实现信息压缩去冗余；（3）分类器模块。分类器模块根据编码器输出完成帕金森病检测，并辅助编码器模块学习紧凑的病理信息表示。本文在自采数据集上进行了多个对比实验进行方法有效性验证。实验结果表明，所提模型在帕金森病检测的准确率、敏感度和F1分数等各项性能指标上相较于基于单源语音数据的模型分别提高了6%、3%、6%。同时所提模型相较于其他信息融合模型在准确率指标上提高了2.8%以上。

本文所提方法的主要贡献在于：（1）利用了多源语音数据带来的信息优势；（2）引入基于自注意力机制的Transformer编码块用于多源语音数据的共有信息提取，并与两个单源语音数据表征学习块一起共同完成多源语音数据的表征学习；（3）采用多步信息融合方式，实现多源数据更细粒度的特征交互；（4）联合训练病理表征学习模块（包含编码器模块和解码器模块）和病情检测模块，实现端到端的信息融合与决策。

2 相关工作

2.1 自注意力机制

自注意力机制（Self attention，SA）可用于对序列数据的建模[28]，将每个实例的原始输入特征表示为一串特征向量序列 $\boldsymbol{x}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_l,\cdots,\boldsymbol{x}_L]$，其中

$\boldsymbol{x}_l \in \mathbf{R}^d$，$d$ 为每个特征向量的维度，L 为序列长度。将特征向量 $\boldsymbol{x}_l$ 分别与 3 个随机初始化的投影矩阵相乘，得到对应的 query_l、key_l、value_l 向量。根据不同 $\boldsymbol{x}_l$ 间 query_l 和 key_l 向量的相关性，得到权重系数 $b_{l,r}$，根据权重系数更新每个特征向量：

$$\boldsymbol{x}_l = \sum_{r=1}^{r=L} b_{l,r} \times \text{value}_r \tag{1}$$

最终，获得的每个特征向量都是与其他特征向量信息交互后的融合信息。因此，采用自注意力机制能够更加充分地学习特征向量间的交互。

2.2　多模态信息融合模型

多模态数据是对同一对象的多角度描述，每个模态间可能存在互补关系。多模态信息融合技术旨在通过对来自多个模态的信息进行关联整合，获取目标对象更完备的特征表示。

随着多模态信息融合技术的快速发展，基于多种模态的融合方式早已变得灵活多变，涌现出许多简单高效的融合模型。具有代表性的工作有：基于多模变分自编码器的多模态融合模型（multimodal variant auto-encoder，MVAE）[29]使用多子网络学习单模态特征，并基于变分思想学习多模态特征的潜在分布，实现对图片和文本数据的多模态完备信息提取；基于 Transformer 模型提出的多模态融合模型 ViLT（vision and language transformer）[22]借助多头注意力机制实现视觉特征和文本特征的信息交互，完成了多模态信息深度交互融合；基于张量外积的信息融合方式[30]通过多模态数据的张量外积，实现情感语义识别领域信息的交互融合；生成式模型 CPM-NET（Cross partial multi-view networks）[31]通过在假设空间随机搜索的方式，寻找匹配多模态数据的完备表征，从模态生成的角度为多模态信息融合提供了新的思路。

3　本文方法

3.1　多源语音信息融合模型概述

本文针对帕金森病检测任务和帕金森病患者的多源语音数据，提出了一种多源语音信息融合模型（MSFAE）。该模型包含编码器、解码器以及帕金森病检测 3 个模块，整体框架如图 1 所示。其中，编码器模块由多个并行支路组成，一条支路通过引入自注意力机制的 Transformer 编码块[28]实现多源语音数据共有信息的提取，还有两条支路通过多层前馈神经网络提取单源语音数据的特有信息，多条支

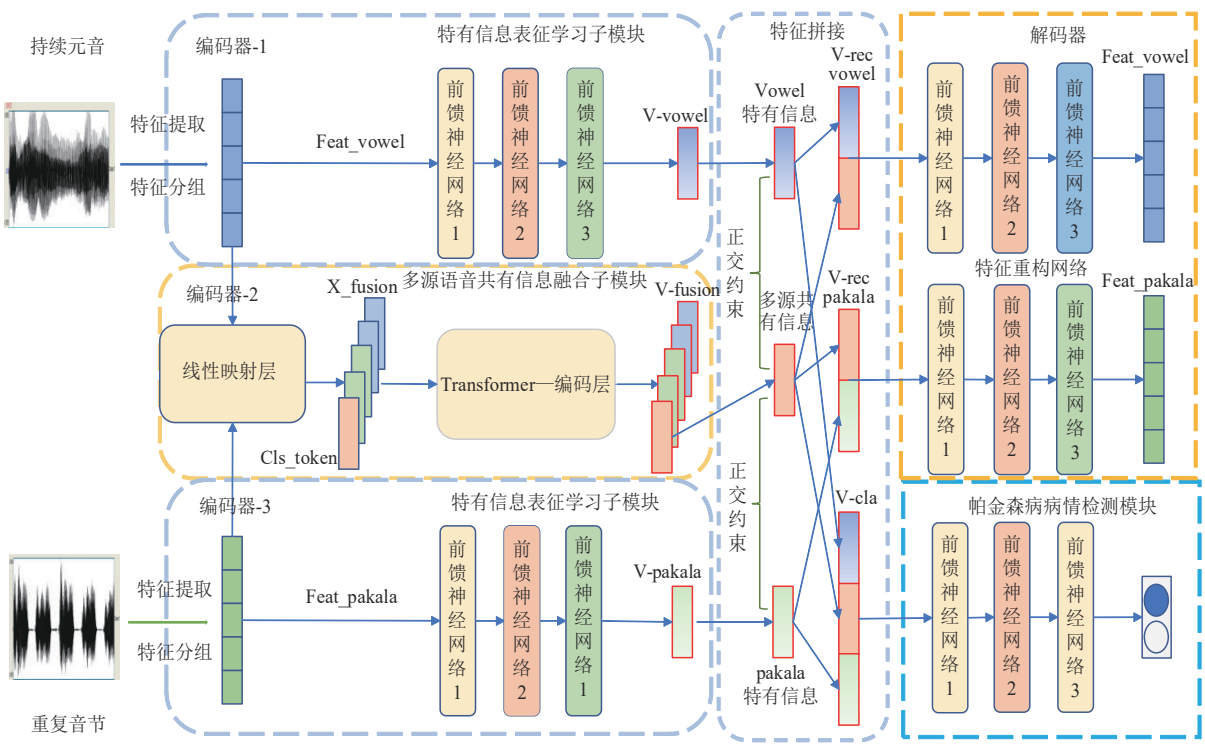

图 1　系统框图

Fig. 1　The system block diagram

路共同完成对来自多个语音数据源中所含病理信息的全面提取。编码器多条支路的输出将以3种不同方式进行特征拼接,以获得3个不同的隐层表征。其中,2个隐层表征将作为多支路解码器的输入,负责完成多个单源数据的重构;另外1个隐层表征则将作为分类器模块的输入实现高效的帕金森病检测。后续各小节将详细描述各模块功能。

3.2 数据预处理

3.2.1 声学特征提取

针对每一个受试者,我们分别采集了持续元音的语音样本和重复音节的语音样本。其中,持续元音数据提取了如频率微扰、振幅微扰、谐波噪声比等发音类特征[11-14];重复音节数据提取了梅尔倒谱系数、巴克带能量等发声类特征[13-15]。

3.2.2 特征分组

文献[32-33]发现,从单源语音数据中提取的声学特征往往存在较大的特征冗余。为了更细粒度地分析数据的特征,我们在特征层面对提取的特征进行了相关性分析,使用均分K-means方法[34]对从单源语音数据中提取的特征集进行了相关聚类分析,并依据组内特征的相关性尽可能大、组间特征的相关性相对较弱的原则对特征进行分组,且每个组的特征数一致。

分组后,第i个受试者的第m个单源语音样本上提取的特征表示为:

$$\boldsymbol{x}_{i,m}=\left[x_{i,m}^{1,1},\cdots,x_{i,m}^{d,1},x_{i,m}^{1,2},\cdots,x_{i,m}^{d,2},\cdots,x_{i,m}^{1,p},\cdots,x_{i,m}^{d,p}\right] \tag{2}$$

其中,d代表每个子组的特征维数,p代表特征的分组标识。m=1时,$\boldsymbol{x}_{i,m}$是持续元音特有信息表征学习模块的输入,对应图1中的Feat_vowel;m=2时,$\boldsymbol{x}_{i,m}$是重复音节特有信息表征学习模块的输入,对应图1中的Feat_pakala。

3.3 编码器模块

如图1所示,本文所提的MSFAE模型的编码器模块由3个并行支路(即编码器-1、编码器-2、编码器-3)组成,其中两条支路分别提取两个单源语音数据的特有信息;一条支路作为多源信息融合子模块实现多源数据共有信息的提取。

3.3.1 单源语音特有信息表征学习子模块

两个单源语音特有信息表征学习子模块是两个并行的分支,一个用于处理从持续元音中获取的声学特征Feat_vowel,一个用于处理重复音节中获取的声学特征Feat_pakala。

单源语音的特有信息表征学习子模块$\mathrm{Enc}_{\mathrm{spc_vowel}}$和$\mathrm{Enc}_{\mathrm{spc_pakala}}$的主要功能在于:从高维的低阶语义特征中学习具备高级语义表达的单源语音特有病理表征。由于Feat_vowel和Feat_pakala对应的声学特征中已经包含了许多丰富的临床病理信息,单源语音特有信息表征学习模块不需要太过于复杂的深度神经网络即可学得有意义的单源语音特有病理信息。这里,单源语音的特有信息表征学习子模块设计成一个具有3个隐藏层的深度神经网络。每层神经网络由55个神经元组成,激活函数为ReLu;针对两种不同的单源语音数据中提取的特征,可学习的权重参数分别为$\boldsymbol{W}_{\mathrm{vowel}}$和$\boldsymbol{W}_{\mathrm{pakala}}$。此外,为方便模型后续的优化处理,加速网络学习,网络的输入端还增加一个批归一化操作,对输入数据作归一化处理。

该模块的输出V_{vowel}、V_{pakala}表示为:

$$V_{\mathrm{vowel}}=\mathrm{Enc}_{\mathrm{spc\_vowel}}\left(\mathrm{Feat\_vowel};\ \boldsymbol{W}_{\mathrm{vowel}}\right) \tag{3}$$

$$V_{\mathrm{pakala}}=\mathrm{Enc}_{\mathrm{spc\_pakala}}\left(\mathrm{Feat\_pakala};\ \boldsymbol{W}_{\mathrm{pakala}}\right) \tag{4}$$

3.3.2 多源语音数据共有信息融合子模块

从多源语音数据提取的声学特征,存在较大的冗余性,且所提特征可能不是处于同一语义层级。如果采用文献[17]中的简单拼接方式,将会引入大量的无效信息,进而影响模型的性能。为避免上述问题,本文采用多步融合的方式,实现多源数据的冗余信息剔除和跨数据源的特征交互融合,具体实现如图2所示。

在多步融合前,为匹配共有信息提取支路的输入形式,对公式(2)所述的$\boldsymbol{x}_{i,m}$进行重新表示:

$$\boldsymbol{x}_{i,m}=\left[\boldsymbol{x}_{i,m}^{1},\boldsymbol{x}_{i,m}^{2},\cdots,\boldsymbol{x}_{i,m}^{p},\cdots,\boldsymbol{x}_{i,m}^{P_m}\right] \tag{5}$$

其中,P_m为第m个单源数据的特征分组个数,$\boldsymbol{x}_{i,m}^{p}=[\boldsymbol{x}_{i,m}^{1,p},\boldsymbol{x}_{i,m}^{2,p},\cdots,\boldsymbol{x}_{i,m}^{d,p}]$为每个特征子组。

本文所提的多步融合操作如下:首先,将每个特征子组$\boldsymbol{x}_{i,m}^{p}$通过一个线性映射层,实现对组内高相关性特征的共有信息提取,并去除冗余性,并使得每个子特征组的信息在特征空间上的对齐。每个特征子组完成线性映射后,所有的输出$\boldsymbol{x}_{i}^{1},\boldsymbol{x}_{i}^{2},\cdots,\boldsymbol{x}_{i}^{P}$组成一个$U\times P$的矩阵,$P$为特征子组总的个数,等于两个单源语音特征的特征子组个数之和,U为特征子组经过线性映射后的特征维度。

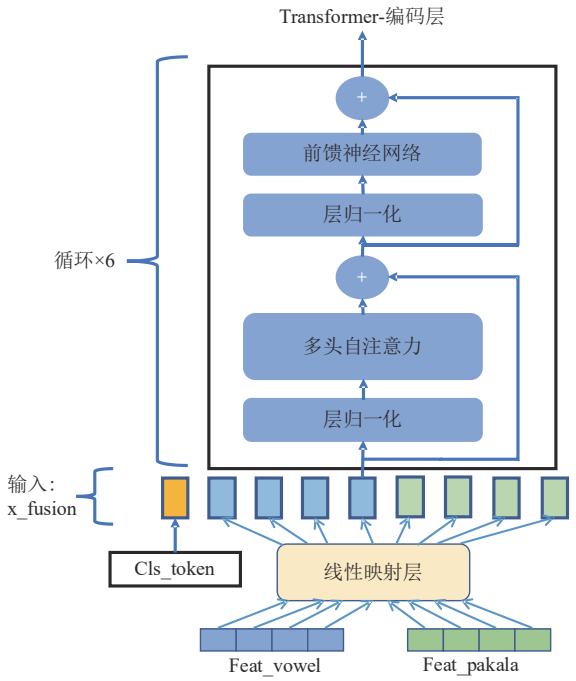

图2　共有信息融合模块框图

Fig. 2　Block diagram of common information fusion module

然后，随机初始化一个 U 维的分类向量 $\boldsymbol{x}_i^{\text{cls\_token}}$，将其与上述矩阵拼接在一起，输入一个多层的 Transformer 编码器模块[28]。拼接后的特征矩阵表示为：

$$\boldsymbol{x}_{\text{fusion}}^{\text{T}} = \left[\boldsymbol{x}_i^{\text{cls\_token}};\ \boldsymbol{x}_i^1;\ \boldsymbol{x}_i^2;\ \cdots;\ \boldsymbol{x}_i^P\right] \tag{6}$$

借助 Transformer 编码块中蕴含的自注意力机制，模型将会学习到各个特征子组间的融合交互信息，完成跨特征子组的信息融合，同时将分类信息汇集在 $\boldsymbol{x}_i^{\text{cls\_token}}$ 上。

Transformer 编码器模块[28]由多头自注意力机制模块（Multihead self-attention，MSA）和前馈神经网络模块（Feedforward neural network，FNN）交替组成。为了加速网络的训练，还在每个块的输入前引入层归一化（LayerNorm，LN）进行数据的归一化处理，在每个块的输出后进行残差连接操作。

信息融合实现的方式是：

$$\boldsymbol{e}_0 = \boldsymbol{x}_{\text{fusion}} = \left[\boldsymbol{x}_i^{\text{cls\_token}}, \boldsymbol{x}_i^1, \boldsymbol{x}_i^2, \cdots, \boldsymbol{x}_i^P\right] \tag{7}$$

$$\boldsymbol{e}_j' = \text{MSA}\left(\text{LN}\left(\boldsymbol{e}_{j-1}\right)\right) + \boldsymbol{e}_{j-1};\ j = 1, 2, \cdots, J \tag{8}$$

$$\boldsymbol{e}_j = \text{FNN}\left(\text{LN}\left(\boldsymbol{e}_{j-1}\right)\right) + \boldsymbol{e}_o';\ j = 1, 2, \cdots, J \tag{9}$$

其中，$\boldsymbol{e}_0$ 代表多头注意力机制的初始输入，$\boldsymbol{e}_j'$ 代表经过 j 次多头注意力机制后的输出，$\boldsymbol{e}_j$ 为 $\boldsymbol{e}_j'$ 经过层归一化后的输出，J 代表编码器网络中 MSA 和 FNN 的迭代次数。公式（8）对应的多头自注意力机制的具体实现为：

$$\text{MSA}(\boldsymbol{X}) = \text{Concat}\left(\text{head}_1, \cdots, \text{head}_h\right)\boldsymbol{W}_j \tag{10}$$

其中，$\boldsymbol{X}$ 为输入 MSA 的特征序列，$\text{head}_1, \cdots, \text{head}_h$ 为多头自注意力机制中的注意机制块。公式（10）通过一个权重为 $\boldsymbol{W}_j$ 的线性映射网络，可将 h 个注意力机制块的输出进行信息汇集。head_h 是信息融合的核心模块，其由 2.1 节所述自注意力机制网络 SA 组成，计算方式如下所示：

$$\text{head}_h = \text{SA}\left(\boldsymbol{X}\boldsymbol{W}_{h,Q}, \boldsymbol{X}\boldsymbol{W}_{h,K}, \boldsymbol{X}\boldsymbol{W}_{h,V}\right) \tag{11}$$

其中，$\boldsymbol{W}_{h,Q}$、$\boldsymbol{W}_{h,K}$、$\boldsymbol{W}_{h,V}$ 为 head_h 的三个投影矩阵的参数，负责将输入的特征序列映射到 query、key、value 向量空间。经过前述的多头自注意力机制后，共有信息提取支路的最终输出为：

$$\boldsymbol{V}_{\text{fusion}} = \text{LN}\left(\boldsymbol{e}_J\right) = \left[\boldsymbol{v}_i^{\text{cls\_token}}, \boldsymbol{v}_i^1, \cdots, \boldsymbol{v}_i^P\right] \tag{12}$$

$\boldsymbol{V}_{\text{fusion}}$ 即为编码器对多源数据分布融合获得的共有信息表征，$\boldsymbol{v}_i^{\text{cls\_token}}$ 为获得的分类表征，包含了从所有输入向量 $\boldsymbol{x}_i^p$ 中提取的目标信息，$\boldsymbol{v}_i^p$ 为 $\boldsymbol{x}_i^p$ 与其他向量经过特征交互后的输出向量。

3.4　特征拼接

我们将从编码器共有信息融合支路模块获取的表征 $\boldsymbol{V}_{\text{fusion}}$ 中取出分类表征 $\boldsymbol{v}_i^{\text{cls\_token}}$，将其与来自两个单源语音表征学习模块的输出 $\boldsymbol{V}_{\text{vowel}}$ 和 $\boldsymbol{V}_{\text{pakala}}$ 进行拼接。拼接的结果作为融合表征 $\boldsymbol{V}_{\text{cla}}$，以实现对多源语音信息的完整表达，将其作为帕金森检测模块的输入。我们还将获取多模数据共有信息的融合特征 $\boldsymbol{v}_i^{\text{cls\_token}}$ 分别与相应单源语音特有信息表征 $\boldsymbol{V}_{\text{vowel}}$ 或者 $\boldsymbol{V}_{\text{pakala}}$ 进行拼接。拼接后的特征向量 $\boldsymbol{V}_{\text{rec\_vowel}}$ 和 $\boldsymbol{V}_{\text{rec\_pakala}}$ 分别作为解码器两条重构单源语音特征支路的输入。

3.5　特征正交化

为进一步确保编码器能够对多源语音数据中共有信息和特有信息的提取，我们对编码器获得的共有信息表征和两个特有信息表征，进行正交约束，降低共有信息表征和特有信息表征间的信息冗余。记矩阵 $\boldsymbol{H}$ 为由多源语音数据共有信息的融合特征 $\boldsymbol{v}_i^{\text{cls\_token}}$ 作为行构成的矩阵，矩阵 $\boldsymbol{S}_m$ 为由第 m 个单源语音数据中提取的单源语音特有信息表征 $\boldsymbol{V}_{\text{vowel}}$ 或者 $\boldsymbol{V}_{\text{pakala}}$ 作为行构成的矩阵，通过正交约束计算得到特征间的差异损失如下：

$$L_{\text{diff}}(\boldsymbol{H}, \boldsymbol{S}_1, \boldsymbol{S}_2) = \left\|\boldsymbol{H}^{\text{T}}\boldsymbol{S}_1\right\|_{\text{F}}^2 + \left\|\boldsymbol{H}^{\text{T}}\boldsymbol{S}_2\right\|_{\text{F}}^2 \tag{13}$$

其中，$\|\cdot\|_{\mathrm{F}}^{2}$为平方Frobenius范数。

3.6 帕金森病检测模块

融合表征$\boldsymbol{V}_{\mathrm{cla}}$作为帕金森病检测模块Cla的输入，通过相应的分类器实现帕金森病的检测。帕金森病检测模块由具有三个隐藏层的神经网络组成。每一层神经元的个数分别为32、16和2，采用ReLu作为激活函数。样本真实标签$y\in[0,1]$，y为0时表示受试者不患病，y为1时代表受试者患有帕金森病。检测模块的分类输出为：

$$\hat{y}=\mathrm{Cla}\left(\boldsymbol{V}_{\mathrm{cla}};\ \boldsymbol{W}_{\mathrm{cla}}\right) \tag{14}$$

其中，$\boldsymbol{W}_{\mathrm{cla}}$为模块参数，分类损失的计算我们将通过预测值$\hat{y}$与真实标签值$y$之间的交叉熵损失来定义。

3.7 解码器

解码器由两个特征重构支路组成：持续元音重构模块$\mathrm{Dec}_{\mathrm{vowel}}$用于重构来自持续元音中提取的特征向量，重复音节重构模块$\mathrm{Dec}_{\mathrm{pakala}}$用于重构重复音节提取的声学特征向量。重构模块网络由3层前馈神经网络组成，使用ReLu激活函数。输出为对该单源语音的原始声学特征$x_{i,m}$的重构，可表示为：

$$\boldsymbol{x}_{\mathrm{vowel}}=\mathrm{Dec}_{\mathrm{vowel}}\left(\boldsymbol{V}_{\mathrm{rec\_vowel}};\ \boldsymbol{W}_{\mathrm{rec\_vowel}}\right) \tag{15}$$

$$\boldsymbol{x}_{\mathrm{pakala}}=\mathrm{Dec}_{\mathrm{pakala}}\left(\boldsymbol{V}_{\mathrm{rec\_pakala}};\ \boldsymbol{W}_{\mathrm{rec\_pakala}}\right) \tag{16}$$

其中，$\boldsymbol{W}_{\mathrm{rec_pakala}}$和$\boldsymbol{W}_{\mathrm{rec_vowel}}$模块的参数，$x_{\mathrm{rec_pakala}}$和$x_{\mathrm{rec_vowel}}$为重构的特征向量，模块使用Smooth L1-loss损失函数对重构误差进行计算，其表达式为：

$$\mathrm{loss}\left(\boldsymbol{x},\boldsymbol{x}_{\mathrm{rec}}\right)=\begin{cases}0.5\times\left(\boldsymbol{x}-\boldsymbol{x}_{\mathrm{rec}}\right)^{2};\ \left|\boldsymbol{x}-\boldsymbol{x}_{\mathrm{rec}}\right|<1\\ \left|\boldsymbol{x}-\boldsymbol{x}_{\mathrm{rec}}\right|-0.5;\ \text{otherwise}\end{cases} \tag{17}$$

其中，$\boldsymbol{x}$、$\boldsymbol{x}_{\mathrm{rec}}$分别为原始特征和模型重构网络的输出。其最终的重构损失为：

$$\mathrm{Loss}_{\mathrm{rec}}=\frac{1}{N}\sum_{i=0}^{N}\begin{cases}0.5\times\left(\boldsymbol{x}_{i}-\boldsymbol{x}_{i,\mathrm{rec}}\right)^{2};\ \left|\boldsymbol{x}_{i}-\boldsymbol{x}_{i,\mathrm{rec}}\right|<1\\ \boldsymbol{x}_{i}-\boldsymbol{x}_{i,\mathrm{rec}}-0.5;\ \text{otherwise}\end{cases} \tag{18}$$

其中$\boldsymbol{x}_{i}$、$\boldsymbol{x}_{i,\mathrm{rec}}$分别代表第$i$个样本的特征表示和重构网络输出的重构特征，$N$为总的样本数。

3.8 多损失优化

本文所提的MSFAE模型由多个子模块组成，其中帕金森病检测模块将采用交叉熵损失函数，特征重构模块将采用Smooth L1-loss函数。为充分利用数据集中的标签信息，本文将联合训练帕金森病检测模块和用于特征学习的编解码模块。最终的模型损失为：

$$L_{\mathrm{total}}=\lambda_{v}L_{\mathrm{rec\_vowel}}+\lambda_{p}L_{\mathrm{rec\_pakala}}+\lambda_{c}L_{\mathrm{cla}}+\lambda_{\mathrm{di}}L_{\mathrm{diff}} \tag{19}$$

其中，$L_{\mathrm{rec_vowel}}$为重构持续元音语音的损失，$L_{\mathrm{rec_pakala}}$为重构重复音节语音的损失，L_{cla}为帕金森病检测模块的分类损失，L_{diff}为共有信息表征和特有信息表征间的差异损失。这里，由于单源语音特征重构损失明显比帕金森病检测模块的损失大得多，为避免多个损失共同优化的过程中出现由于尺度不一致导致模型偏向大损失的方向优化，导致其他模块的性能下降。我们预设了4个超参数λ_{v}、λ_{p}、λ_{c}、λ_{di}，通过对各个损失进行加权，减小尺度不一致对模型的影响。

值得说明的是，为避免参数更新时，所提模型专注于优化特征重构损失而忽略帕金森病检测模块，帕金森病分类模块和解码器中特征重构支路的输入是有区别的，如图1所示。通过上述这些设计能够避免优化过程中的权重不平衡问题，也能共同帮助所提模型学习到更为紧凑的融合表示。

4 实验

4.1 数据集

为开展基于多源语音融合的帕金森病检测研究，本文研究团队与南京医科大学附属老年医院的神经内科展开长期合作。本文所使用的多源语音数据集，即由该医院帕金森病及运动障碍专病门诊筛选出的68名患者和17名健康人的语音数据构成。需要说明的是，在现有的帕金森病语音公开数据集中，尚未发现符合本文研究需求的多源语音数据。自采的帕金森病多源语音数据集中的受试者信息统计见表1。其中，男性受试者57人（含帕金森病患者（PD）49人，健康人（HC）8人），年龄从46岁到88岁不等；女性受试者为28人（含帕金森病患者19人，健康人9人），年龄从56岁到84岁不等。表中提供了患者发病时间和病变程度（HY（Hoeh & Yahr）分期）数据，其中，HY分期3期以前属于轻中度，3期以后症状越来越严重。

受试者在安静环境下接受语音采集（环境噪声低于20 dB）。采集时，受试者的唇部位于距拾音麦克风十厘米以内的范围，在听到专业人员的指令后，开始发声。考虑到不同母语的发音习惯带来的差异，避免由语种带来的混淆因素，让研究成果更好地服务于国内外研究人员，我们仅考虑以下两种

表1　自采帕金森病多源语音数据集信息统计

Tab. 1　Self-collected Parkinson's disease multi-source speech dataset information statistics

	男性		女性		合计	
受试者类别	PD	HC	PD	HC	PD	HC
受试者人数	49	8	19	9	68	17
平均年龄及统计方差	69.3(9.5)	66.5(7.2)	69.8(8.2)	65.3 (6.8)	69.4(9.2)	65.9(7.0)
年龄分布	46~88	58~77	56~84	53~74	46~88	53~77
平均病情持续时间及统计方差	5.9(3.6)	0	5.4(3.1)	0	5.8(3.4)	0
HY分期	1~4	0	1~4	0	1~4	0

方式采集受试者的语音：(1)以稳定的声音进行持续元音/a/发音；(2)以尽可能快的速度进行重复音节发音，即发出/pakala/。每个患者的语音记录经剪辑后共计340个样本，以48 kHz采样率和.wav格式存储。语音采集完成后，由在场的医务人员对受试者的患病与否及严重程度进行标注。

4.2　实验设置

本文实验使用python语言实现，通过多组对比实验从多个角度验证模型的性能。所有的实验均在4.1节所述的自采数据集上进行，实验结果采用了十折交叉验证，使用准确率(ACC)、敏感度(SEN)和F1分数作为实验结果的评估准则。

准确率表示准确区分帕金森病患者和健康人的概率，敏感度代表正确检测出帕金森病患者的概率，F1分数衡量模型的总体预测性能，其计算公式分别如下所示：

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{20}$$

$$\mathrm{SEN}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{21}$$

$$\mathrm{F1}=2\Big/\left(2+\frac{\mathrm{FP}}{\mathrm{TP}}+\frac{\mathrm{FN}}{\mathrm{TP}}\right) \tag{22}$$

其中，TP表示分类正确的帕金森病样本数，TN表示分类正确的健康人样本数，FP表示将健康人样本误分类成帕金森病样本的数量，FN表示将帕金森病样本误分类成健康人样本的数量。

模型的参数设置如表2所示。

表2　MSFAE模型参数设置

Tab. 2　MSFAE Model parameters setting

网络结构参数	参数值
X_vowel	320
X_pakala	320
信息融合模块Q, K, V向量维度	64
信息融合模块多头注意力	2
单模态特有前馈神经网络1-3	[55,55,64]
特征重构前馈神经网络1-3	[100,200,320]
帕金森病检测模块	[32,16,2]
特征分组映射层	16
周期数	30
学习率	0.001
批大小	36
优化器	Adam
Dropout	0.1

4.3　与多个单源语音基线模型的性能比较

为论证多源语音数据融合的优势，本节将所提模型与基于单源语音数据的基线模型进行了性能比较。参与比较的单源语音基线模型有：随机森林(RF)，支持向量机(SVM)以及深度学习模型孪生网络(Siamese-net)[35]。实验结果如表3所示。

表3　与单源语音模型的性能比较

Tab. 3　Performance comparison with single source speech model

Model	ACC(%)	SEN(%)	F1(%)
MSFAE	**92.68**	**86.51**	**90.21**
RF(vowel)	83.86	81.41	83.88
RF(pakala)	86.62	83.32	84.26
SVM(vowel)	83.52	77.53	79.85
SVM(pakala)	84.33	79.23	80.26
Siamese-net(vowel)	77.81	66.67	75.79
Siamese-net(pakala)	79.70	71.29	76.79

从实验结果中可以看到，基于多源语音的MSFAE模型能够比单源语音数据在各个指标上有较大的提升。实验结果验证了，多源语音数据在结合多个数据源数据的信息之后，能够实现更高的检测准确率。

4.4　与其他信息融合模型的性能比较

本节对MSFAE模型以及其他前文所提及的信

息融合模型进行了性能比较。参与比较的模型有：TFN[30]，CPM-NET[31]，Vilt[22]，MKL[20]。实验结果如表4所示。

表4 与其他信息融合模型的性能比较

Tab. 4 Performance comparison with other information fusion models

Model	ACC(%)	SEN(%)	F1(%)
MSFAE	**92.68**	86.51	**90.21**
TFN	89.86	**87.81**	89.44
CPM_NET	89.35	83.65	84.52
Vilt	88.65	83.76	84.24
MKL	86.92	74.85	76.79

从实验结果中可知，我们的模型在与多个多模态信息融合模型相比较，在准确率上分别有2.82%、3.33%、4.03%、5.76%的提升，在敏感度指标上与最优的TFN模型相近，高于其他比较模型，同时F1分数相较其他比较模型也有提升。其原因在于，我们通过同时结合了多源数据的共有信息和特有信息，实现了更加全面的信息提取。同时在共有信息抽取时，通过多步融合方式，避免直接对提取的声学特征拼接带来的语义鸿沟以及噪声冗余。

4.5 消融实验

为进一步探究所提模型的性能，本节通过消融实验来检测子模块的性能，重点考察特征分组线性映射模块，以及基于注意力机制融合的信息融合模块对模型的贡献。实验的详细结果如表5所示。

表5 消融实验

Tab. 5 Ablation experiments

Model	ACC(%)	SEN(%)	F1(%)
MSFAE	**92.68**	**86.51**	**90.21**
MSFAE(without fusion)	83.62	80.32	81.62
MSFAE(without feat_group)	90.27	83.15	89.91
MSFAE(without spec_feat)	87.22	79.71	82.31

由实验结果可知，模型在没有使用多源语音数据信息融合模块时(MSFAE(without fusion))，性能受到较大的影响，模型此时缺乏对多源语音的低阶语义信息融合，仅在单源语音经过表征学习块提取高阶语义信息后进行了拼接，无法实现多源语音数据的互补互增益。模型在缺失特征分组时(MSFAE(without feat_group))，由于缺失对原始输入数据的更细粒度的信息冗余去除，为模型引入更多的噪声信息，从而使得模型性能少许下降。模型在缺失单源语音数据特有信息表征学习模块时(MSFAE(without spec_feat))，性能也出现了较大的性能下降，其原因是特征融合模块的主要作用是同时最大化多源语音数据的共有信息，单源语音数据特有信息表征模块的加入，能够弥补对单源语音数据特有信息的关注。

5 结论

本文提出一种多源语音信息融合模型，解决了单源语音数据无法全面表征受试者构音能力的问题。其中，采用多步信息融合方式，并引入多头自注意力技术实现多源数据更细粒度的特征交互，有效解决了信息冗余问题，避免多源数据融合过程中的噪声累积。通过多分支网络，提取多源数据的特有信息和共有信息，并引入正交约束，有效实现多源数据中病理信息的提取。实验结果显示，本文所提的MSFAE模型与单源语音数据基线模型比较，在各个指标上均有较大程度的性能提升。与其他信息融合模型相比，所提模型在帕金森病检测任务上有独特的优势。在此基础上，我们将进一步研究多源语音数据在受损情况下的帕金森病检测方案。

参考文献

[1] BENBA A, JILBAB A, SANDABAD S, et al. Voice signal processing for detecting possible early signs of Parkinson's disease in patients with rapid eye movement sleep behavior disorder[J]. International Journal of Speech Technology, 2019, 22(1): 121-129.

[2] 沈珺，张天宇，黄菲菲，等. 帕金森病构音障碍声学特点的初步探索[J]. 中华神经科杂志，2019，52(8)：613-619.
SHEN Jun, ZHANG Tianyu, HUANG Feifei, et al. Study of voice disorder based on acoustic assessment in Parkinson's disease[J]. Chinese Journal of Neurology, 2019, 52(8): 613-619. (in Chinese)

[3] OROZCO-ARROYAVE J R, ARIAS-LONDOÑO J D, VARGAS-BONILLA J F, et al. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease[C]//In Proceedings of the Ninth International Conference on Language Resources and Evaluation. Reykjavik, Iceland: Association for Computational Linguistics, 2014: 342-347.

[4] SAKAR B E, ISENKUL M E, SAKAR C O, et al. Col-

lection and analysis of a Parkinson speech dataset with multiple types of sound recordings[J]. IEEE Journal of Biomedical and Health Informatics, 2013, 17(4): 828-834.

[5] SUPHINNAPONG P, PHOKAEWVARANGKUL O, THUBTHONG N, et al. Objective vowel sound characteristics and their relationship with motor dysfunction in Asian Parkinson's disease patients [J]. Journal of the Neurological Sciences, 2021, 426:117487.1-117487.8.

[6] 季薇,杨茗淇,李云,等.基于掩蔽自监督语音特征提取的帕金森病检测方法[J/OL].电子与信息学报:1-9[2023-06-19]. http://kns. cnki. net/kcms/detail/11.4494.TN.20230117.1257.002.html.
JI Wei, YANG Mingqi, LI Yun, et al. Parkinson's disease detection method based on masked self-supervised speech feature extraction[J/OL]. Journal of Electronics & Information Technology: 1-9 [2023-06-19]. http://kns.cnki.net/kcms/detail/11.4494.TN.20230117.1257.002.html. (in Chinese)

[7] NOVOTNÝ M. Automated assessment of diadochokinesis and resonance in dysarthrias associated with basal ganglia dysfunction[D]. Czech Republic: Czech Technical University, 2016.

[8] OROZCO-ARROYAVE J R, HÖNIG F, ARIAS-LONDOÑO J D, et al. Automatic detection of Parkinson's disease in running speech spoken in three different languages[J]. The Journal of the Acoustical Society of America, 2016, 139(1): 481-500.

[9] KLUMPP P, VÁSQUEZ-CORREA J C, HADERLEIN T, et al. Feature space visualization with spatial similarity maps for pathological speech data[C]//Interspeech 2019. ISCA: ISCA, 2019: 3068-3072.

[10] 蹇梦.汉语帕金森病患者的功能语调产出研究[D].南京:南京师范大学,2019.
JIAN Meng. A study on the output of functional intonation in Chinese patients with Parkinson's disease[D]. Nanjing: Nanjing Normal University,2019. (in Chinese)

[11] TSANAS A, LITTLE M A, MCSHARRY P E, et al. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease[J]. IEEE Transactions on Biomedical Engineering, 2012, 59(5): 1264-1271.

[12] OROZCO-ARROYAVE J R, VDSQUEZ-CORREA J C, HÖNIG F, et al. Towards an automatic monitoring of the neurological state of Parkinson's patients from speech [C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China. IEEE, 2016: 6490-6494.

[13] OROZCO-ARROYAVE J R, BELALCAZAR-BOLAÑOS E A, ARIAS-LONDOÑO J D, et al. Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases[J]. IEEE Journal of Biomedical and Health Informatics, 2015, 19(6): 1820-1828.

[14] OROZCO-ARROYAVE J R, VÁSQUEZ-CORREA J C, VARGAS-BONILLA J F, et al. NeuroSpeech: An open-source software for Parkinson's speech analysis[J]. Digital Signal Processing, 2018, 77: 207-221.

[15] OROZCO-ARROYAVE J R, HÖNIG F, ARIAS-LONDOÑO J D, et al. Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinson's disease[C]//Interspeech 2015. ISCA: ISCA, 2015: 95-99.

[16] SZTAHÓ D, TULICS M G, VICSI K, et al. Automatic estimation of severity of Parkinson's disease based on speech rhythm related features[C]//2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). Debrecen, Hungary:IEEE, 2018: 011-016.

[17] BOCKLET T, STEIDL S, NTH E, et al. Automatic evaluation of Parkinson's speech-Acoustic, prosodic and voice related cues[C]//14th Annual Conference of the International Speech Communication Association. Baixas, France: ISCA, 2013:1148-1152.

[18] BALTRUŠAITIS T, AHUJA C, MORENCY L P. Multimodal machine learning: A survey and taxonomy [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2): 423-443.

[19] HONG Chaoqun, YU Jun, WAN Jian, et al. Multimodal deep autoencoder for human pose recovery [J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2015, 24(12): 5659-5670.

[20] ALTHLOOTHI S, MAHOOR M H, ZHANG Xiao, et al. Human activity recognition using multi-features and multiple kernel learning[J]. Pattern Recognition, 2014, 47(5): 1800-1812.

[21] GLODEK M, TSCHECHNE S, LAYHER G, et al. Multiple classifier systems for the classification of audio-visual emotional states[C]//Affective Computing and Intelligent Interaction: Fourth International Conference. Berlin: Springer.2011: 359-368.

[22] KIM W, SON B, KIM I. Vilt: Vision-and-language transformer without convolution or region supervision[C]//International Conference on Machine Learning. New York: Curran Associates, 2021: 5583-5594.

[23] TRUONG Q T, LAUW H W. VistaNet: Visual aspect attention network for multimodal sentiment analysis [J].

Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 305-312.

[24] LI Yachao, LI Junhui, ZHANG Min. Deep Transformer modeling via grouping skip connection for neural machine translation [J]. Knowledge-Based Systems, 2021, 234(25): 107556.1-107556.12.

[25] 季薇, 吕艳洁, 林钢, 等. 基于过滤的域适应模型融合的帕金森病情预测[J]. 仪器仪表学报, 2018, 39(6): 104-111.
JI Wei, LV Yanjie, LIN Gang, et al. Filtering-based domain adaptation model fusion method in prediction of Parkinson's disease symptom severity [J]. Chinese Journal of Scientific Instrument, 2018, 39(6): 104-111. (in Chinese)

[26] AZADI H, AKBARZADEH-T M R, KOBRAVI H R, et al. Robust voice feature selection using interval type-2 fuzzy AHP for automated diagnosis of Parkinson's disease [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 2792-2802.

[27] NOVOTNÝ M, RUSZ J, ČMEJLA R, et al. Automatic evaluation of articulatory disorders in Parkinson's disease [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(9): 1366-1378.

[28] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all You need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA. New York: ACM, 2017: 6000-6010.

[29] KHATTAR D, GOUD J S, GUPTA M, et al. MVAE: Multimodal variational autoencoder for fake news detection [C]//WWW'19: The World Wide Web Conference. San Francisco, CA, USA. New York: ACM, 2019: 2915-2921.

[30] ZADEH A, CHEN Minghai, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017: 1103-1114.

[31] ZHANG C, HAN Z, CUI Y, et al. CPM-Nets: Cross Partial Multi-View Networks[C]//The Thirty-third Conference on Neural Information Processing Systems (NeurIPS). Vancouver Convention Center, Vancouver CANADAs: NeurIPS, 2019, 32: 559-569.

[32] JI W, LI Y. Energy-based feature ranking for assessing the dysphonia measurements in Parkinson detection [J]. IET Signal Processing, 2012, 6(4): 300-305.

[33] JI Wei, LI Yun. Stable dysphonia measures selection for Parkinson speech rehabilitation via diversity regularized ensemble [C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China: IEEE, 2016: 2264-2268.

[34] GANGANATH N, CHENG C T, TSE C K. Data clustering with cluster size constraints using a modified K-means algorithm [C]//2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery. Shanghai, China: IEEE, 2014: 158-161.

[35] BHATI S, VELAZQUEZ L M, VILLALBA J, et al. LSTM Siamese network for Parkinson's disease detection from speech[C]//2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP). Ottawa, ON, Canada: IEEE, 2020: 1-5.

作者简介

季 薇 女,1979年生,江苏淮安人。南京邮电大学通信与信息工程学院硕士生导师,教授,主要研究方向为机器学习与信号处理的交叉研究、无线通信与通信信号处理等。
E-mail: jiwei@njupt.edu.cn

王传瑜 男,1997年生,江西赣州人。南京邮电大学通信与信息工程学院硕士研究生,主要研究方向为机器学习与信号处理的交叉研究。
E-mail: c_y_wangfit@163.com

李 云 男,1974年生,安徽安庆人。南京邮电大学计算机学院博士生导师,教授,主要研究方向为机器学习、特征选择、信息安全。
E-mail: liyun@njupt.edu.cn

郑慧芬 女,1973年生,江苏无锡人。南京医科大学附属老年医院,主任医师,主要研究方向为帕金森病及相关运动障碍性疾病。
E-mail: zhenghuifen@163.com