EasyChair Preprint

№ 6747

Single-Feature and Multi-Feature Fusion Audio Classification for Alzheimer's Disease Based on Convolutional Neural Network

Zhilin Liu, Yanyu Yang, Zhao Yang, Kun Zhao and Wei Xi

October 3, 2021

基于卷积神经网络的单一特征及多特征融合阿尔茨海默症音频分类方法

刘志林* 杨颜瑜* 杨钊 赵鲲 惠维

西安交通大学 陕西省西安市 710061

摘　要：阿尔茨海默症是一种渐进式脑部疾病，随着时间的推移而恶化，因此对阿尔茨海默症进行早期筛查具有非常重要的意义。此前对阿尔茨海默症的识别依赖于各类声学特征和语音转录，而本文仅利用说话人的音频特征，提出了基于卷积神经网络多特征融合的阿尔茨海默症音频识别方法，并结合模型间集成学习方法，利用多次随机采样、语音端点检测等语音处理策略，实现了对阿尔兹海默症患者、轻度认知障碍患者及正常人三类人群音频的有效区分。本文提出的模型在长音频赛道和短音频赛道分别取得了84.87%(排名第四)和83.78%(排名第三)的准确率，此外，本文还探究了在不同网络结构及训练方法上的阿尔兹海默症识别表现。本文的源代码可以在 https://github.com/lzl32947/NCMMSC2021_AD_Competition 中访问。

关键词：阿尔茨海默症识别；音频分类；机器学习；语音处理；多特征融合

Single-feature and Multi-feature Fusion Audio Classification for Alzheimer's Disease based on Convolutional Neural Network

Zhilin Liu* Yanyu Yang* Zhao Yang Kun Zhao Wei Xi

Xi 'an Jiaotong University Xi 'an, Shaanxi Province 710061 China

Abstract: Alzheimer's disease is a progressive brain disease, which worsens over time. Therefore, early screening of Alzheimer's disease is of great significance. Previously, the recognition of Alzheimer's disease relays on various acoustic features and speech transcription. This paper proposes an Alzheimer's disease audio classification method based on convolutional neural network, only uses the speaker's audio features, multi-feature fusion, combined with the ensemble learning method among models, and uses speech processing strategies such as multiple random sampling and speech endpoint detection. It can effectively distinguish the audio of patients with Alzheimer's disease, patients with mild cognitive impairment and normal people. The model proposed in this paper has achieved 84.87% (ranked fourth) and 83.78% (ranked third) accuracy in long audio track and short audio track respectively. In addition, this paper also explores the classification performance of Alzheimer's disease with different network structures and training methods. The source code can be accessed in https://github.com/lzl32947/NCMMSC2021_AD_Competition.

Key words: Alzheimer Detection；Audio Classification ；Machine Learning；Speech Processing；Multi-feature fusion

* 共同作者

1 引言

随着全球人口快速老龄化，阿尔兹海默症(Alzheimer's disease，AD)[1]患病人数逐年增多，它会导致患病者在记忆、沟通等认知领域出现不可逆转的恶化，患者会出现失语、失认、执行功能障碍等症状；而轻度认知障碍(Mild Cognitive Impairment，MCI)[2]则是指正常老化过程与早期阿尔兹海默症之间的一种过渡阶段，表现为轻度的记忆和智能损害，此时患者具有客观的认知损害但日常生活能力尚未受到明显影响。

目前主要通过医疗专家对患者的临床病史和残疾情况、神经心理检查、脑成像和脑脊液检查综合分析进行阿尔兹海默症的诊断，需要大量的人力和时间成本，虽然该病的突出症状是记忆和时空方向的改变，但语言障碍也是目前被证实的一个重要特点[3][4]，故而基于语音和语言的自动检测筛查方法也是阿尔兹海默症识别研究的一个重要领域。

阿尔茨海默症患者在执行利用语义信息的任务时存在困难，具体表现为阿尔茨海默症患者说话更缓慢，停顿时间更长，并花费额外的时间搜索正确的词，从而导致了说话不流利[5]。其中表现最为突出的语言障碍可以总结如下：命名[6]、找词困难[7]、重复[8]、过度使用不明确和模糊的术语[9]和不恰当使用代词[10]。

最近的研究提出利用深度神经网络中自动特征提取的能力进行阿尔兹海默症的辅助诊断。其中，Rohanian[11]等人提出了一个利用了 LSTM 网络层和前馈层的多模态融合模型，从音频、文本以及不流畅特征中进行学习，在通过自动语音进行阿尔茨海默综合症识别竞赛(ADReSS)[12]测试集上准确率达到 79.17%。Chen 等[13]提出了一种基于注意力机制的网络，该网络由 CNN 和 GRU 模块组成使得模型能够分析局部的语言模式和整体的宏观语言功能，对阿尔茨海默症患者的识别准确率为 97.42%。不仅如此，Zargarbashi 等人[14]在基于 N-gram、i-vector 和 x-vector 三个模型的分类准确率分别为 78.2%、75.9%和 75.1%，他们还提出了多特征融合嵌入的方法，三种模型的联合融合精度达到 83.6%。

前人工作表明，语言的辅助能够有效提高基于音频的阿尔兹海默症的识别准确率，但本文的数据集仅仅提供了音频数据，由于受试者方言的存在以及含混不清的语音表述，借助自动化工具获取转录的效果不佳，且由于规则限制不允许人工转录，本文在仅使用音频数据的条件下对阿尔茨海默症的识别进行研究。提出了利用单一特征及多特征融合进行音频分类的方法，同时采用集成学习，并结合语音端点检测以及训练策略，在长短音频赛道的阿尔茨海默症的识别中均达到了较高的准确率。

2 数据集

2.1 数据集概述

本文采用的数据集由第十六届全国人机语音通信学术会议阿尔茨海默综合症竞赛主办方提供，共包含阿尔茨海默综合症患者(AD)、轻度认知功能障碍患者(MCI)、正常人(HC)三类人的语音数据，发音内容则包含看图说话，流畅性测试和自由谈话三种方式，每段语音长度在 30 秒至 60 秒之间，共计包含音频 280 段，不同测试者以编号进行区分，每名测试者可能对应着多段语音数据。

为了更加全面的验证模型效果，本文采用交叉验证的方法进行训练集和测试集的划分，以测试模型的准确性和泛化性能。为了避免同一患者的多段音频被同时划分至训练集和测试集中，本文首先将所有受试者的语音数据按照受试者编号进行合并，合并后共计包含音频 123 段。

2.2 音频特征提取

本文采用声谱图、梅尔声谱图及梅尔频率倒谱系数作为神经网络的输入特征，其提取方式如下：

(1) 声谱图(Spectrogram，Spec)[15]由原始音频经过预加重、分帧、加窗及短时傅里叶变

换(Short-Time Fourier Transform，STFT)等操作得到。在实际应用中，语音信号经过分帧、加窗处理，分割成一帧帧的离散序列，可视为采用短时傅里叶变换，其计算方法见公式(1)。其中K是离散傅里叶变换后的频率点个数，k是频率索引，$0 \le k < K$。$\mathrm{X}[k,l]$建立起索引为lL的时域信号和索引为k的频域信号的关系，对于采样率F_s，相应的索引对应为时间lL/F_s和频率kF_s/K。

$$\mathrm{X}[k,l] = \sum_{\mathrm{n}=0}^{\mathrm{N}-1} x_l[n]\mathrm{e}^{-\frac{j2\pi nk}{K}} = \sum_{\mathrm{n}=0}^{\mathrm{N}-1} w[n]x[n+lL]\mathrm{e}^{-\frac{j2\pi nk}{K}} \tag{1}$$

(2) 梅尔声谱图(Melspectrogram，Melspec)[16]也被称为 FBank，声谱图经过一系列梅尔滤波器组梅尔声谱图，由于相邻滤波器存在重叠，因此梅尔声谱图的特征相关性较高，其计算方法见公式(2)，在提取出声谱图的基础上，求其平方得到能量谱，将每个滤波频带内的能量进行叠加，第k个滤波器输出功率谱$X[k]$；将每个滤波器的输出取对数，得到相应频带的对数功率谱。

$$\mathrm{Y}_{FBank}[\mathrm{k}] = \log X[k] \tag{2}$$

(3) 梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficients，MFCC)[17]在梅尔声谱图的基础上经过反离散余弦变换(Discrete Cosine Transform，DCT)得到，是一种在自动语音识别和说话人识别中广泛使用的特征，其计算方法见公式(3)，其中M是三角滤波器个数，一共产生L个 MFCC 系数。

$$C_n = \sum_{k=1}^{M} \log X[k] \cos\left(\frac{\pi(k-0.5)n}{M}\right), n = 1,2,\cdots,L \tag{3}$$

2.3 语音端点检测与消音

语音端点检测(Voice Activity Detection, VAD)[18]一般用于鉴别音频信号当中的语音出现(Speech Presence)和语音消失(Speech Absence)。由于音频中的背景噪声可能会对特征产生干扰，故对于部分音频，通过检测语音端点并将背景噪声去除可以得到更好的音频特征。

本文中所使用的语音端点检测步骤如下：

(1) 在采样率为 16000 的条件下下读入音频，将音频信号进行分帧、加窗处理并进行归一化；

(2) 为减少背景噪声的影响，增强算法的鲁棒性，在信号中加入高斯白噪声，有助于消除全零值；

(3) 计算每帧的能量，某一帧前后 5 帧都不存在语音信号，则被认为是无语音帧，将该帧删去，否则为有语音帧进行保留。

3 网络结构

3.1 单一特征预测网络

在 2.2 及 2.3 节的基础上，通过对音频进行特征提取，可以得到尺寸为[特征长度，序列长度]的二维向量，通过升维处理，可以将其看作三维向量，从而可以使用二维卷积进行特征提取。同时为了能够有效的使用特征进行训练，本节采用二维卷积神经网络中常用的“特征提取层-全连接层”的网络结构，将单一特征预测网络划分为特征提取部分和全连接层部分，通过组合不同的网络部分达到更好的效果。

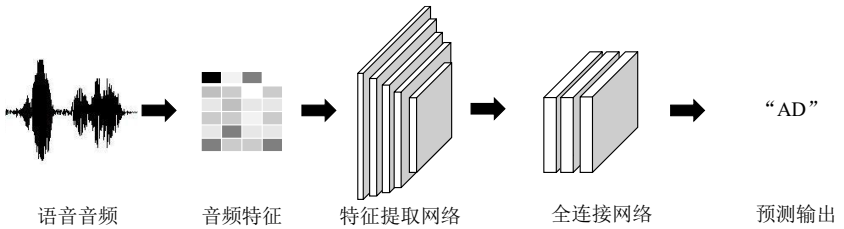


图 1 单一特征预测网络结构图

Fig.1 Structure of single feature prediction network

图 1 给出了该网络的结构，首先由音频提取特定音频特征，如 MFCC 特征、Spec 特征及 Melspec 特征等，提取后的特征经由特征提取网络获得音频的高维特征图，经由全局平均池化展平至指定维度，最终经过全连接层进行预测输出。

此种网络结构有着很大程度上的通用性，不同骨干网络与全连接网络可以相互耦合，从而使得组合后的网络可以对多种不同时长的音频进行训练预测。同时，通过将音频特征堆叠至三维，可以利用预训练好的骨干网络，如 VGG16、ResNet18 等对特征进行提取，从而加速网络训练并提升识别效果。

3.2 多特征融合预测网络

3.1 节中给出的模型由于只采用了一种音频特征，其效果受制于提取的特征，因此较易产生错误，而同时结合多种特征进行预测的方法能够有效避免单一特征带来的不确定性，从而提升预测的准确率。

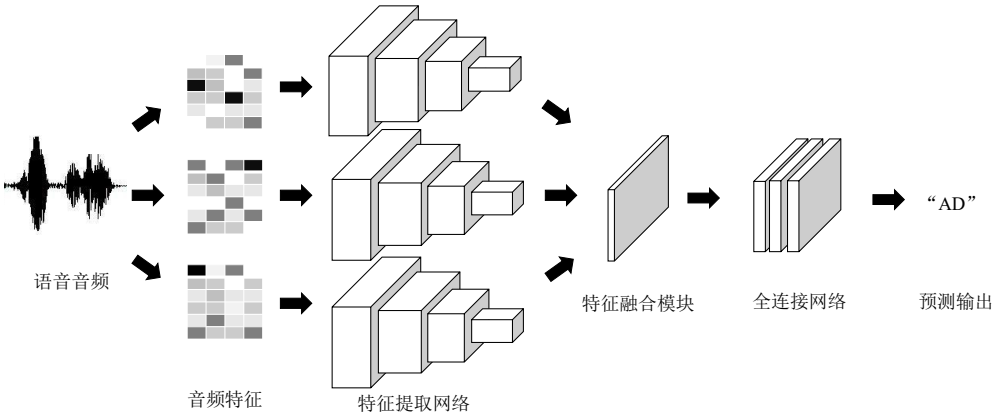
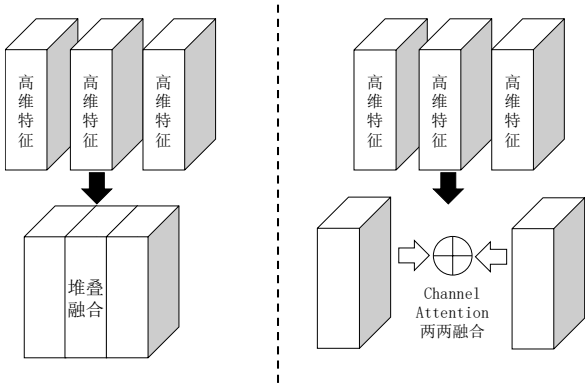


图 2 多特征融合预测网络结构图

Fig.2 Structure of multi-feature fusion prediction network

本文在高维特征图中进行特征融合，模型主体见图 2 所示。在同一段音频中同时提取多种音频特征，并经由骨干网络提取特征，这些骨干网络可以相同，也可以不同，最后在此基础上进行特征融合。

本文采用两种方法进行特征融合，一种是简单堆叠，另一种则是采用 MS-CAM 模块[19]进行特征图融合，两种方法的结构图如图 3 所示，其中堆叠方法(图 3.a)将特征图在首个维度拼接，拼接后首维度变为原维度的三倍，而使用 MS-CAM 模块(图 3.b)则在多种特征间两两融合，最终得到一个融合后的特征图。



(a) 特征堆叠融合示意图　　(b) 特征经由 MS-CAM 模块融合示意图

(a) Structure of stacked feature fusion　　(b) Structure of feature fusion via MS-CAM module

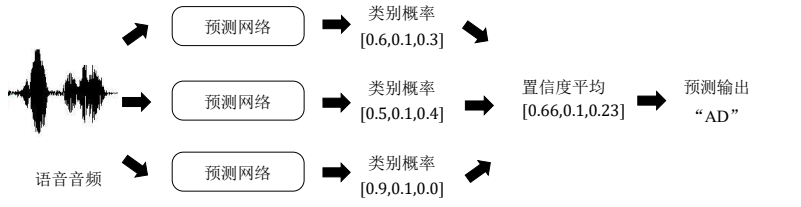
图 3 高维特征融合方法示意图

Fig.3 Structure of high-dimensional feature fusion method

在进行特征融合后，高维特征经由全局平均池化层压缩至指定尺寸，并通过全连接层进行预测并输出。

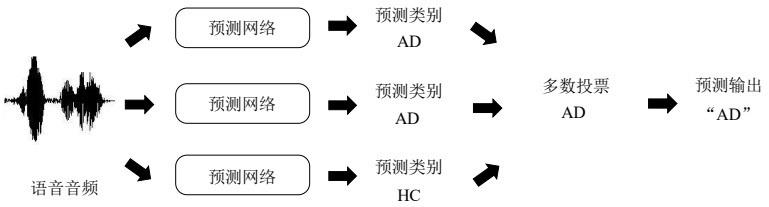
3.3 模型间集成学习

在对 3.1 和 3.2 节模型进行实验时，通过构造分类样本的混淆矩阵，可以发现不同模型对于不同类别有着不同的预测精度，因此通过集成学习的方法，通过综合各个模型的效果，能够进一步提升模型预测精度。



(a) 模型间置信度平均集成学习方法示意图

(a) Structure of ensemble learning with average confidence among models



(b) 模型间投票集成学习方法示意图

(b) Structure of ensemble learning with vote among models

图 4 模型间集成学习结构图

Fig.4 Structure of ensemble learning among models

图 4 给出了模型间集成学习的结构图，本文给出平均置信度及模型投票两种加权方式。平均置信度(图 4.a)将各个模型的预测输出进行平均，并选择平均置信度最大的类别作为该段音频的预测类别，而模型投票(图 4.b)则采用输出结果投票的方式计算最终的输出，如果最大类别的得票相同，则证明该段音频偏差较大，需要重新进行预测。

4 实验

4.1 实验参数

由于数据集内的音频数量较少，因此本文采用 4 折交叉验证方式，采用准确率作为评价模型效果的指标，将数据集拆分成训练集和测试集，每个模型均训练并预测四次，并以输出的平均准确率作为模型的准确率。在短音频任务中，每一段音频均被随机截取 5 秒的片段，每个 epoch 中各个音频均被截取 32 次，在长音频任务中，每一段音频被随机截取 25 秒的片段，每个 epoch 同样截取 32 次。本文选用原音频和端点消音后的音频作为音频输出，在其上提取 MFCC、Spec、Melspec 三类共计 6 种特征作为神经网络的输入特征。在训练中，学习率被设置为 0.001，使用 AdamW 优化器，每个模型训练 20 个 epoch，同时设置学习率衰减以提升模型的训练效果。

4.2 实验结果

在短音频识别任务中，不同单一特征预测网络的实验结果如表 1 所示。

表 1　单一特征预测网络在短音频上的四折交叉验证平均准确率

Tab.1　Average accuracy of four-fold cross validation of single feature prediction network on short audio

骨干网络	使用特征	平均准确率(%)
全卷积网络	Melspec	67.48
全卷积网络	Spec	67.79
全卷积网络	MFCC	68.36
ResNet18	Melspec	71.91
VGG19bn	Melspec	75.78
VGG19bn	Spec	75.93
VGG19bn	MFCC	73.74

同时，多特征融合预测网络的实验结果如表 2 所示，在训练时，对于每种音频特征，均训练一个对应的特征提取网络，此后将训练得到的特征提取网络权重导入至多特征融合网络中，锁定所有特征提取网络的权重，只训练全连接层，得到网络融合后的结果。训练完成后，解除特征提取网络权重的锁定，对融合后的网络进行整体训练，得到融合后微调的结果。

表 2　单一特征训练的特征提取网络及多特征融合预测网络在短音频上的四折交叉验证平均准确率

Tab.2　Average accuracy of four-fold cross validation of single feature trained feature extraction network and multi-feature fusion feature prediction network on short audio

骨干网络	融合方法	使用特征	平均准确率(%)
全卷积网络	/	Melspec	67.48
全卷积网络	/	Spec	67.79
全卷积网络	/	MFCC	68.36
全卷积网络融合	堆叠	MFCC，Spec，Melspec	71.48
全卷积网络融合微调	堆叠	MFCC，Spec，Melspec	75.10
全卷积网络融合	MS-CAM	MFCC，Spec，Melspec	60.52
全卷积网络融合微调	MS-CAM	MFCC，Spec，Melspec	62.77

综合表 1 和表 2 的实验结果，表明多特征融合预测网络相较于单一特征预测网络有着明显的提升，这显示多个音频特征的融合能够有效的提升网络的预测效果。相比较于简单拼接的特征融合，采用 MS-CAM 模块进行融合的特征表现较差，这可能是由于特征之间差异性较大使得维度约减时丢失了部分特征所致。

在长音频识别任务中不同模型的实验结果如表 3 及表 4 所示。

表 3　单一特征预测网络在短音频上的四折交叉验证平均准确率

Tab.3　Average accuracy of four-fold cross validation of single feature prediction network on long audio

骨干网络	使用特征	平均准确率(%)
全卷积网络	Melspec	70.79
全卷积网络	Spec	68.11
全卷积网络	MFCC	71.68
全卷积网络	Melspec(VAD)	67.88
全卷积网络	Spec(VAD)	68.90
全卷积网络	MFCC(VAD)	66.09
ResNet18	Melspec(VAD)	73.77

表 4　多特征融合预测网络在长音频上的四折交叉验证平均准确率

Tab.4　Average accuracy of four-fold cross validation of multi-feature fusion feature prediction network on long audio

骨干网络	融合方法	使用特征	平均准确率(%)
全卷积网络融合	堆叠	MFCC，Spec，Melspec	73.92
全卷积网络融合微调	堆叠	MFCC，Spec，Melspec	73.06
全卷积网络融合	堆叠	MFCC(VAD)，Spec(VAD)，Melspec(VAD)	66.45
全卷积网络融合微调	堆叠	MFCC(VAD)，Spec(VAD)，Melspec(VAD)	67.42

结合短语音任务的表现，长音频相较于短音频，其平均准确率有所下降，这可能是由于模型全局平均池化过程中合并了过多的高维特征表示造成的。与短音频识别任务类似，多特征融合预测网络得到了更好的结果，这也进一步证明了多特征融合方法是有效的。此外，模型在消音后的音频特征上表现相对较差，这可能是由于语音端点检测存在误差使得网络丢失了部分重要音频信息，或消音后长时间的白噪声对模型训练产生了干扰造成的。

此外，本文也对多种模型间的集成学习方法进行了实验，实验结果如表 5 所示。

表 5　集成学习模型实验结果

Tab.5　Experimental results of ensemble learning with different models

网络名称	集成方式	平均准确率(%)
全卷积网络 1(Melspec)	/	70.79
全卷积网络 2(Spec)	/	68.11
全卷积网络 3(MFCC)	/	71.68
全卷积网络 1+2+3	置信度平均	72.33
全卷积网络 1+2+3	结果投票	72.61

由表 5 可以得出，多个模型间的集成学习方法能够一定程度上提升最终模型的预测效果，这是由于各个模型对于相同的音频学习到的特征不同，综合各个模型的结果能够提升整体网络对单个音频的预测效果。另外，置信度平均和投票方法的集成学习结果相差很小，很大程度上是由于模型对分类错误的音频也会给出高概率的输出，因此，基于模型混淆矩阵的置信度加权方法可能能够提升分类的准确率，但受制于时间，本文未做进一步实验。

除了以上模块，为了探究其它可能的结构对特征提取网络产生的影响，本文还以 ResNet18 为骨干网络，在短音频任务中使用 Melspec 特征，针对 Attention 机制、LSTM 网络层进行了组合实验，实验结果如表 6 所示。

表 6　不同网络结构进行实验结果

Tab.6　Experiments results with different network structures

任务	网络模型	平均准确率(%)
短语音识别	ResNet18	73.90
	ResNet18+LSTM	74.70
	ResNet18+Attention+LSTM	72.38

	ResNet18	72.90
长语音识别	ResNet18+LSTM	72.62
	ResNet18+Attention+LSTM	66.34

从表 6 中可以得出，加入 LSTM 层后准确率略微提升，而继续加入 Attention 机制后准确率反而有所下降，其原因可能是是 LTSM 网络层本身就具有一定的记忆能力，再加入 Attention 机制会导致整个网络复杂度过高，或者与分类问题出现不匹配导致准确率有所下降。此外，在加入 Attention 机制后，观察到训练时损失下降较快，能够促进网络快速收敛。

除此之外，本文还尝试了一些数据增强方法，其中时域及频域上的遮罩会造成准确率的降低，通过 Grad-CAM 方法生成的网络注意力区域则表明网络需要依赖于近乎全部音频特征进行预测判断，因此在数据及较少的情况下，遮罩后的特征缺失会使得网络难以学习到特征，从而影响准确率。

在竞赛中，本文综合了以上方法，采用全部训练集进行训练，并且在预测过程中，根据任务不同，首先将从给定音频中随机截取 5 秒或 25 秒的序列，每段音频截取 20 次以基本覆盖音频范围，将此 20 次随机采样的音频提取特征并送入网络进行预测，最终以网络输出的结果进行投票，得票最高的类别作为该网络预测的语音类别。在竞赛测试集中，短音频与长音频三次提交结果的指标如表 7 所示。

表 7　竞赛测试集中各任务三次识别结果

Tab.7　Three identification results of each task on the test set of the competition

任务	准确率(%)	召回率(%)	精度(%)	F1 分数(%)
6s 短语音识别	76.67	75.44	75.84	75.19
	83.78	83.34	83.44	83.36
	80.05	79.11	79.87	79.16
60s 长语音识别	78.15	76.15	79.23	75.36
	84.87	84.49	84.63	84.50
	83.19	83.00	83.21	83.07

5　结论

本文提出了单一特征预测网络和多特征融合预测网络，并通过集成学习方法进一步提升模型的准确率，使得模型能够仅通过音频，有效的区分阿尔茨海默症患者、轻度认知障碍患者和正常人，帮助促进阿尔茨海默症的早期识别及诊断。

在未来的工作中，本文将尝试对音频中讲话人进行词汇标记、不流利标记以及找到对阿尔茨海默症患者识别贡献更大、与语音信息相关性更高的声学特征。此外，本文将尝试其他模型集成的方法进一步提高模型准确度。

参考文献

[1] A. Burns and S. Iliffe, “Alzheimer’s disease,” B M J, vol. 338, no. 7692, pp. 467–471, 2 2009.

[2] “World health organization. dementia: Fact sheet no. 362,” September 2017, 2 (2017).

[3] J. Reilly, J. Troche, and M. Grossman, “Language processing in dementia,” The handbook of Alzheimer’s disease and other dementias, pp. 336–368, 2011.

[4] K. A. Bayles and D. R. Boone, “The potential of language tasks for identifying senile dementia,” Journal of Speech and Hearing Disorders, vol. 47, no. 2, pp. 210–217, 1982.

[5] K. López-de Ipi˜ na, J.-B. Alonso, C. M. Travieso, J. Solé-Casals, H. Egiraun, M. Faundez-Zanuy, A. Ezeiza, N. Barroso, M. EcayTorres, P . Martinez-Lage et al., “On the selection of non-invasive methods based on speech analysis oriented to automatic alzheimer disease diagnosis,” Sensors, vol. 13, no. 5, pp. 6730–6745, 2013.

[6] J. Reilly, J. Troche, and M. Grossman, “Language processing in dementia,” The handbook of Alzheimer’s disease and other dementias, pp. 336–368, 2011.

[7] D. Kempler and E. Zelinski, “Language in dementia and normal aging,” Dementia and normal aging, pp. 331–365, 1994.

[8] B. Croisile, B. Ska, M.-J. Brabant, A. Duchene, Y . Lepage, G. Aimard, and M. Trillet, “Comparative study of oral and written picture description in patients with Alzheimer’s disease,” Brain and language, vol. 53, no. 1, pp. 1–19, 1996.

[9] S. Ahmed, A.-M. F. Haigh, C. A. de Jager, and P . Garrard, “Connected speech as a marker of disease progression in autopsyproven Alzheimer’s disease,” Brain, vol. 136, no. 12, pp. 3727– 3737, 2013.

[10] D. N. Ripich and B. Y . Terrell, “Patterns of discourse cohesion and coherence in Alzheimer’s disease,” Journal of Speech and Hearing Disorders, vol. 53, no. 1, pp. 8–15, 1988.

[11] Rohanian M, Hough J, Purver M. Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech[J]. arXiv preprint arXiv:2106.09668, 2021.

[12] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, “Alzheimer’s dementia recognition through spontaneous speech: The ADReSS Challenge,” 2020. [Online]. Available: https://arxiv.org/abs/2004.06833

[13] J. Chen, J. Zhu, and J. Ye, “An attention-based hybrid network for automatic detection of alzheimers disease from narrative speech,” Proc. Interspeech 2019, pp. 4085–4089, 2019.

[14] S. Zargarbashi and B. Babaali, “A multi-modal feature embedding approach to diagnose alzheimer disease from spoken language,” arXiv preprint arXiv:1910.00330, 2019.

[15] D.H. Klatt and K.N. Stevens, "On the automatic recognition of continuous speech,Implications from spectrogram reading experiment," IEEE Trans. Audio and Electroacoust.1973, 21(3): 210 – 217.

[16] S. O. Arik, G. F. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, “Deep voice 2: Multi-speaker neural text-to-speech,” CoRR, vol. abs/1705.08947, 2017.

[17] MR.D.Mehendale.Speakidentification [J] .Signal&ImageProcessing: AnInternationalJournal, 2011, 78(2): 62 – 69.

[18] 徐鹏进，郭莉，刘书昌． 基于音高与端点联合检测的音符识别算法[J]. 计算机应用, 2011, 31(S2):172-175.

[19] Dai Y, Gieseke F, Oehmcke S, et al. Attentional feature fusion[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 3560-3569.