基于时域分析技术的语音识别*

赵 超, 庞 立, 彭宣尧, 龙雅琴, 林逸阳
（西安交通大学 自动化科学与工程学院, 陕西 西安 710049）

摘 要 在人工智能和机器学习的大背景下，更加快捷高效的语音识别和分析技术层出不穷。本文基于时域分析方法，对语音信号进行分帧、加窗等预处理后，综合比较不同的机器学习方法对语音处理的效果，最终使用集成学习的方法对语音信号进行分类，得到了相对较好的语音识别效果。本文通过正确率、查准率等多个指标对实验的结果进行了具体的评估，对不同预处理方法和分类方法进行对比总结，得到了较好的语音信号预处理方法和分类算法，验证了我们的模型的有效性。

关键词 语音识别，时域分析，集成学习，窗函数

Speech Recognition Based on Time Domain Analysis Technology

Chao Zhao,Li Pang,Xuanyao Peng,Yaqin Long,Yiyang Lin
(College of Automation Science and Technology, Xi'an Jiaotong University, Xi'an Shaanxi 710049, China)

Abstract In the context of artificial intelligence and machine learning, faster and more efficient speech recognition and analysis technologies emerge one after another. This paper is based on the time domain analysis method. After preprocessing the speech signal by framing and windowing, and comprehensively comparing the effects of different machine learning methods on speech processing, this paper finally uses the integrated learning method to classify the speech signal, and obtains a relatively good speech recognition effect. This paper specifically evaluates the results of the experiment through multiple indicators such as accuracy rate and precision rate, then compares and summarizes different preprocessing methods and classification methods, and finally obtains relatively good speech signal preprocessing methods and classification algorithms, which verifies the effectiveness of our model.

Key Words speech recognition, Time Domain Analysis, Ensemble Learning, window function

0 引言

语音识别技术起源于20世纪30年代，Homer Dudley提出并研制成功第一个声码器，奠定了语音产生模型的基础。随着计算机的出现，现代语音分析技术可以通过计算机上进行，大大提高了识别效

*收稿日期: XXXX−XX−XX. 基金项目：国家自然科学基金资助项目 (51685168)

率和精确性。如今，语音信号处理无论是在基础研究或在技术应用，都已取得了突破性进展。现在语音信号可分为三个主要分支，即语音编码，语音识别和语音合成技术[1]。

如今，语音信号处理是一个新兴的交叉学科，与认知科学、心理学、语言学、计算机科学、模式识别和人工智能学科有着密切的联系。语音信号处理技术的发展依赖于这些学科的发展，语音信号处理技术的进步也将促进这些领域的进展。

语音信号处理目的是得到一些语音特征参数，以便高效的传输或存储，或通过某种处理以达到诸如语音合成等特定目的，辨识出讲话者、识别出讲话的内容等。随着现代科学技术和计算机技术的发展，除了人与人的自然语言的沟通，人机对话和智能机领域也开始使用语言[2]。由此可见，对语音信号进行时域分析是语音信号处理的基础，其重要性不言而喻。

本文在整体框架上主要分为四个部分进行论述。在第一部分，我们对介绍了语音信号时域分析的理论基础，并对信号数据进行预处理与双门限法端点检测；第二部分，我们对各种传统与集成分类器进行介绍与比较，为后续实验部分奠定理论基础；第三部分，我们使用不同分类器与不同加窗方式，对语音信号精确度进行评估与比较；第四部分，我们在理论支撑与重复对比实验后，对实验结果进行总结。最后列出此次实验参考文献并将本文使用的主要代码段置于附录。

1 数据预处理

1.1 语音采集

在MATLAB环境中，本文采用audiorecorder函数从麦克风中进行语音录制。为了建立数字0-9的语音库，本文对每个数字采集了10组样本。录制结束后，本文利用audiowrite函数进行语音格式的转化，将语音信号另存为.wav格式。该格式文件便于提取音频特征如采样率、每秒采样字节数等，通过编程可以对其中的数据字段进行读取。采集语音信号结束后，可得语音信号如图1：

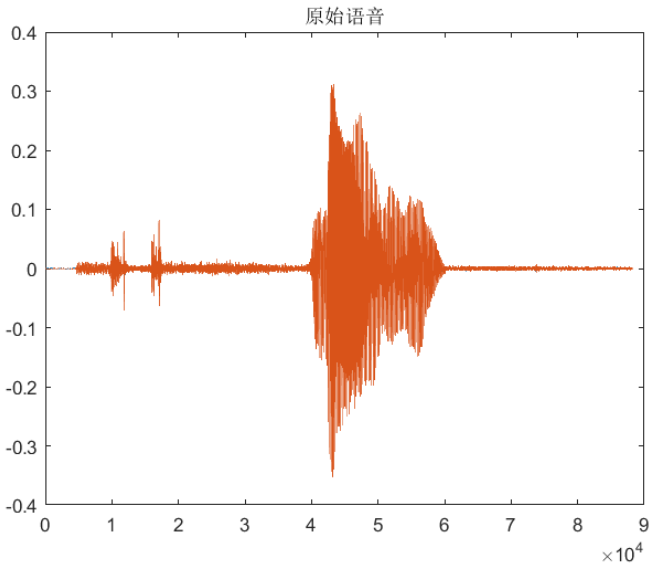


图 1: 原始信号

1.2 语音信号分帧

语音信号是一个非平稳态过程，不能使用处理平稳信号的数字处理技术对其进行处理。而在一段极小的时长范围内，我们可以认定该语音信号是平稳的，即其特征参数恒定。因此，为了利用短时时域分析技术，我们将语音信号分割为一帧一帧来对其特征参数进行分析，其中我们取帧长10-30s。对

语音信号进行短时处理后，我们可以分析出每一帧语音信号的特征参数，从而将其组合成特征参数时间序列。

分帧[3]是利用可移动有限长度窗口进行加权的方法来实现的，即用窗函数w(n)来乘s(n)，从而形成加窗信号

$$s_w = s(n) * w(n) \tag{1}$$

本小组采集的语音信号为双声道，因此在进行分帧前本小组对双声道参数做了平均处理，以方便后续的加窗。

1.3　加窗

在语音数字信号处理中，本文采用三种常用的窗函数[4]，并以数字“1”为例，比较分析不同加窗对应的不同效果。

1.矩形窗

矩形窗数学表达式如公式2

$$w(n) = \begin{cases} 1, 0 \le n \le N-1 \\ 0, \text{其它} \end{cases} \tag{2}$$

矩形窗的时域波形和幅度特性如图2，对原始语音信号进行矩形加窗后得信号图如图3。

2.汉明窗

汉明窗数学表达式如公式3

$$w(n) = \begin{cases} 0.54 - 0.46cos[2\pi n/(N-1)], 0 \le n \le N-1 \\ 0, \text{其它} \end{cases} \tag{3}$$

汉明窗的时域波形和幅度特性如图4，对原始语音信号加汉明窗后得信号图如图5。

3.海宁窗

海宁窗数学表达式如公式4

$$w(n) = \begin{cases} 0.5[1 - cos(2\pi n/N-1)], 0 \le n \le N-1 \\ 0, \text{其它} \end{cases} \tag{4}$$

海宁窗的时域波形和幅度特性如图6，对原始语音信号加汉明窗后得信号图如图7。

通过观察比较可得，频域分析中，矩形窗主瓣窄，频率分辨率较高，但旁瓣也较高，加窗前后时域波形未发生明显变化，而幅值发生了改变；而汉明窗和海宁窗所呈现的效果相似，过滤效果不佳。各种加窗方式具体效果将在实验部分讨论。

1.4　时域分析

语音信号是一种非平稳的时变信号，它所携带的各种信息，可用于语音编码、语音合成、语音识别和语音增强等加工处理。因此，为了对语音信号进行分析，提取特征参数，用于后续处理，我们在短时间范围内将该语音信号视为准稳态过程并对其进行短时时域分析。本文重点提出短时能量、短时平均幅值和短时过零率。

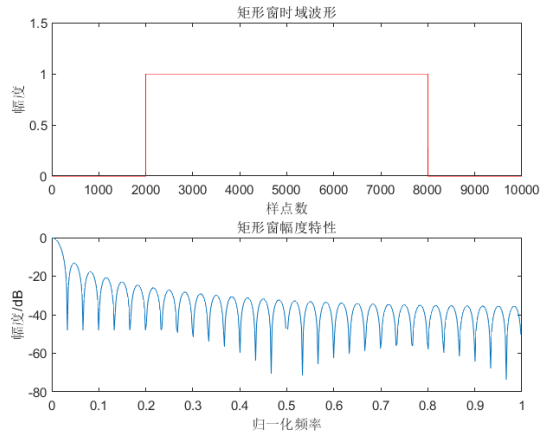


图 2: 矩形窗的时域波形和幅度特性

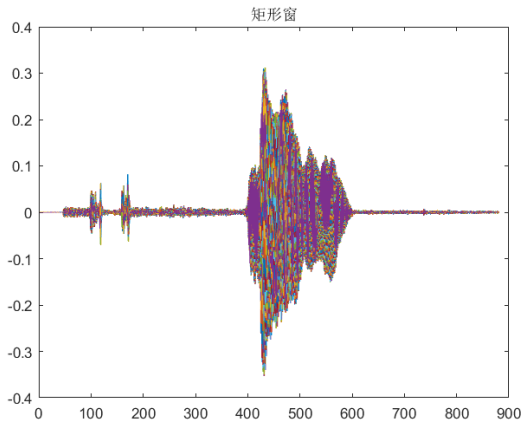


图 3: 矩形窗处理后的信号

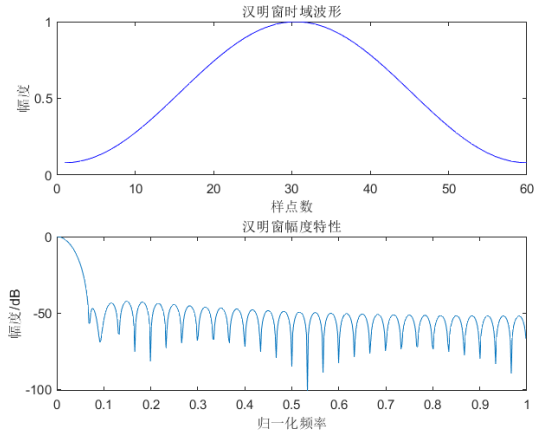


图 4: 汉明窗的时域波形和幅度特性

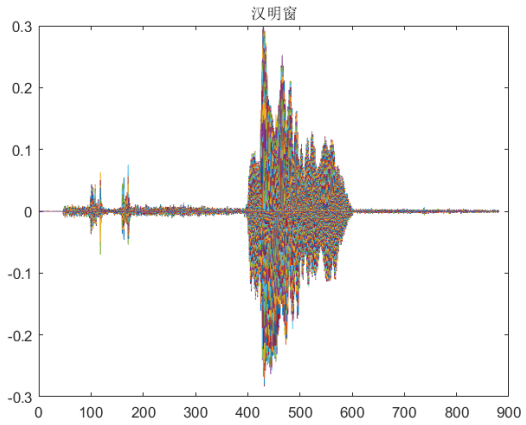


图 5: 汉明窗处理后的信号

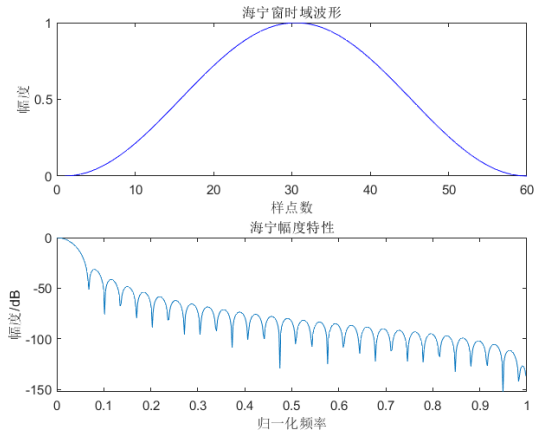


图 6: 海宁窗的时域波形和幅度特性

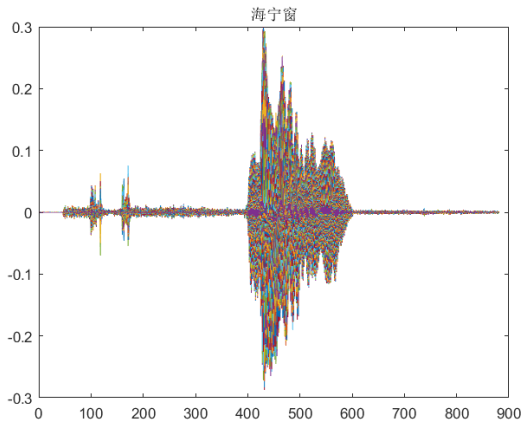


图 7: 海宁窗处理后的信号

1.4.1　短时能量与短时平均幅值函数

短时能量E_n是一个度量语音信号幅度值变化的函数，第n帧语音信号$x_n(m)$的短时能量计算公式如下：

$$E_n = \sum_{m=0}^{N-1} x_n^2(m) \tag{5}$$

但是短时能量E_n对高电平比较敏感。为解决短时能量的高敏感性，可以采用另一个度量语音信号能量大小的短时平均幅值函数M_n，其定义为：

$$M_n = \sum_{m=0}^{N-1} |x(m)| \tag{6}$$

短时平均幅值函数M_n在计算小取样和大取样时，不会因为平方而造成较大差异，对高电平敏感度较低。

1.4.2　短时过零率

短时过零率表示一帧语音中语音信号波形穿过横轴（即零电平）的次数。对于连续语音信号，过零定义为意味着时域波形通过时间轴;而对于离散信号，如果相邻的取样值改变符号则称为过零。由此可以定义第n帧语音信号$x_n(m)$的短时过零率Z_n，计算公式如下：

$$Z_n = \frac{1}{2} \sum_{m=0}^{N-1} |sgn[x_n(m)] - sgn[x_n(m-1)]| \tag{7}$$

由此分析，过零率就是样本改变符号的次数。实质上，过零率是信号频谱分布在时域的一种最简单的体现，即高频分量丰富的信号其过零率也一般较高。

1.5　端点检测

1.5.1　基于短时能量和短时过零率的双门限端点检测原理

语音端点检测是指从一段语音信号中确定语音的起点和结束点的位置。在语音识别系统中，语音端点检测的正确率直接影响着语音识别的准确率，因此，设计一种高性能的端点检测算法对于语音识别系统至关重要。

为了克服传统的端点检测算法的缺点，我们采用一种基于短时能量和短时过零率的双门限语音端点检测算法:**在上述时域分析的基础上**，首先为短时能量和短时过零率分别确定两个门限，一个为较低的门限，对信号的变化比较敏感，另一个是较高的门限。当低门限被超过时，很有可能是由于很小的噪声所引起的，未必是语音的开始，当高门限被超过并且在接下来的时间段内一直超过低门限时，则意味着语音信号的开始。

由此可知，双门限端点检测算法核心是利用短时能量检测浊音，短时过零率检测清音，两者配合从而确定语音的端点。

1.5.2　双门限端点检测实验分析

利用上述原理，我们对语音信号进行初次端点检测实验，设置参数帧长wlen=20ms，步长step=10ms。对高、低门限阈值ITR、IZCT设置后，结果并不理想，实验结果如图8。

反复检查和思考之后，我们在高、低门限阈值参数中引入3σ原则，即公式8，9

$$ITR = max\{\overline{E_n} + 3\sigma\} \tag{8}$$

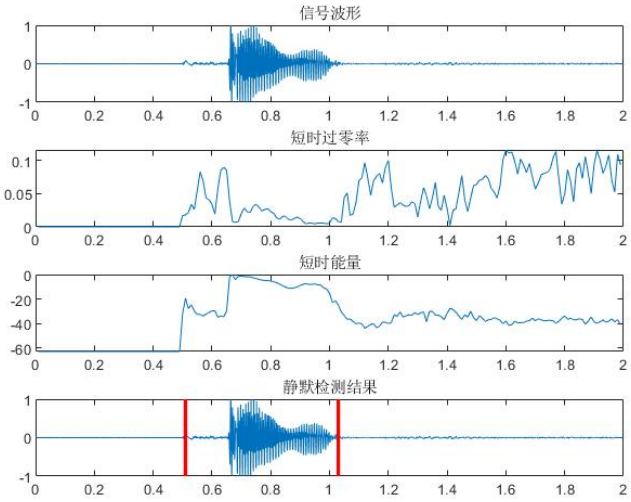


图 8: 传统双门限端点检测实验结果

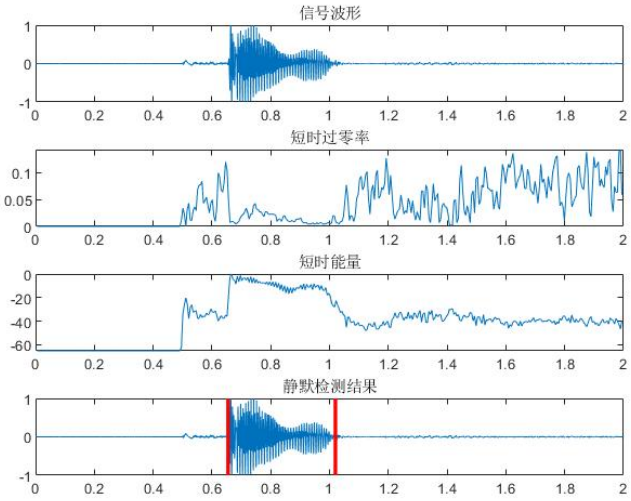


图 9: 引入3σ原则的双门限端点检测实验结果

$$IZCT = max\{\overline{Z_n} + 3\sigma\} \tag{9}$$

重复实验后，实验结果如图9。

通过对比可以发现，引入3σ原则的双门限端点检测效果优于传统双门限端点检测法。利用基于短时能量和短时过零率的3σ双门限端点检测方法对200组语音样本进行端点检测，结果成功率高达98%。

2　语音识别部分

2.1　传统分类器介绍

2.1.1　KNN（K-Nearest Neighbor）

KNN算法[5]的核心思想是，如果一个样本在特征空间中的K个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。该方法在确定分类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。KNN方法在类别决策时，只与极少量的相邻样本有关。

K即是我们判定相邻与否的距离，在应用中，通常使用的是欧氏距离来定义：

$$\rho = \sqrt{(x_2 - x_1)^2 - (y_2 - y_1)^2} \tag{10}$$

当拓展到高维空间：

$$d(x, y) := \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \tag{11}$$

每当录入测试点后，通过计算，找到在限定范围内的值之内的数据集中的点，然后通过比较不同数据集中的点的多少来判定该点是属于哪一个类别。

KNN算法比较适用于样本容量比较大的类域的自动分类，而那些样本容量较小的类域采用这种算法比较容易产生误分。

2.1.2　SVM（support vector machines)

支持向量机[6]是一种二分类模型，它的基本模型是定义在特征空间上的**间隔最大的线性分类器**，利用超平面，对数据集进行最大间隔划分，得到该超平面的表达式，进而利用其对测试数据进行分类。

2.1.3　决策树(Decision Tree）

决策树[7]是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树。

2.2　集成学习介绍

集成学习算法[8]是通过构建并结合多个分类器来完成学习任务，通常拥有较高的准确率，不足之处就是模型可能比较复杂。

我们以KNN为基分类器，构造集成学习分类器。对于训练集数据，我们通过训练若干个个体学习器，通过一定的结合策略，就可以最终形成一个强学习器。集成学习有两个主要的问题需要解决，第一是如何得到若干个个体学习器，第二是如何选择一种结合策略，将这些个体学习器集合成一个强学习器。

弱学习器的选择有两种。第一种就是所有的个体学习器都是同质的，比如都是决策树个体学习器。第二种是所有的个体学习器是异质的，整体模型由不同的基分类器组合而成。

本次实验采用的是随机对原数据的特征值进行随机采样，然后再送到若干个KNN学习器中进行决策判断，最后将所有结果进行投票处理，得到最终的流程如图10所示。

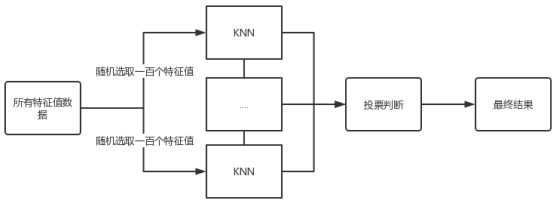


图 10: 集成学习流程图

2.3 语音识别过程

本次语音识别的具体过程是：

采集数据集和预处理：先通过分帧和加窗处理，得到语音信号的特征值。在特征值的基础上，对语音信号进行端点检测得到，得到语音信号的有效部分，接下来，有效的语音部分也通过分帧和加窗的方法进行预处理，得到特征值。

语音识别：通过对语音信号特征值数据集进行了整理和归纳后，通过集成学习方法，对测试的语音信号进行时域的分析和识别。具体框图如图11所示：

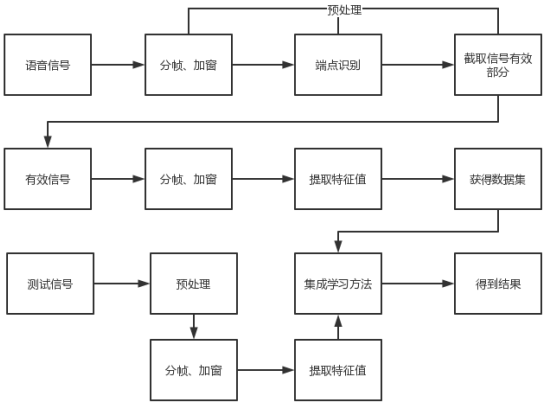


图 11: 语音处理流程框图

2.4 方法优缺点介绍

2.4.1 优点

由图10可以得到此方法的优点有：

1) 性能更好：与任何单个模型的贡献相比，集成可以做出更好的预测并获得更好的性能；

2) 更加合理的边界：弱分类器间存在一定差异性，导致分类的边界不同。多个弱分类器合并后，就可以得到更加合理的边界，减少整体的错误率，实现更好的效果;

3) 适应不同样本体量：对于样本的过大或者过小，可分别进行划分和有放回的操作产生不同的样本子集，再使用样本子集训练不同的分类器，最后进行合并；

4) 易于融合：对于多个异构特征数据集，很难进行融合，可以对每个数据集进行建模，再进行模型融合。

2.4.2 缺点

此方法存在的缺点有：

1) 此方法在预测的时候会添加模型的偏差分量，模型对输入和输出之间映射的函数形式的假设强度较低；

2) 尽管bagging算法能够减少总体方差，但随着训练的迭代，总体方差减少的幅度在逐步减少。

3) 模型整体存在样本数目过少，特征向量种类不多等问题

3 实验部分

3.1 评估指标

表 1: 变量定义

变量	定义	解释
TP	真正例	真实数据集的正样本被预测为正样本
TN	真负例	真实数据集的负样本被预测为负样本
FP	假正例	真实数据集的负样本被预测为正样本
FN	假正例	真实数据集的正样本被预测为负样本

本文中选择的评估指标为AUC（Area Under Curve）、准确率（Accuracy）、查准率(Precision)、第一类错误率(Type Ⅰ error)、第二类错误率(Type Ⅱ error)、F-score。AUC为ROC（Receiver Operating Characteristic）曲线下面积，综合考虑了灵敏度和特异度，可以完整地描述分类效果，其范围从0（无判别力）到1（完全判别力），AUC的值越接近1证明分类效果越好。ROC曲线的横坐标为FPR，即假阳性率，代表将真是数据集的负样本预测为正样本的数量占所有真实负样本数量的比例，纵坐标为TPR，即真阳性率，也称灵敏度或召回率（Recall），代表将真实的正样本预测为正样本的数量占所有真实正样本数量的比例。FPR和TPR可分别表示为：

$$FPR = \frac{FP}{FP + FN} \tag{12}$$

$$TPR = Recall = \frac{TP}{TP + FN} \tag{13}$$

准确率的代表预测正确的样本占总样本的比例，使用公式可表示为：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

查准率代表预测的正样本中真实的正样本所占的比例，使用公式可表示为：

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

第一类错误率的定义为将真实的负样本错误预测为正样本的数量占所有真实负样本数量的比例，误分类代价较大，使用公式可表示为：

$$TypeI = \frac{FP}{TN + FP} \tag{16}$$

第二类错误率定义为将真实的正样本错误预测为负样本的数量占所有真实正样本数量的比例，使用公式表示为：

$$TypeII = \frac{FN}{TP + FN} \tag{17}$$

F-score也称F-measure，是基于查准率和召回率的综合指标，F-score越接近1证明分类效果越好，使用公式可表示为：

$$F - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{18}$$

在评估模型，对模型的五种评估指标有表2的评估方法。

表 2: 评估指标

指标	准确率	查准率	第一类错误率	第二类错误率	F-score
性质	越大越好	越大越好	越小越好	越小越好	越大越好

3.2 不同分类器的效果比较

我们将自己的音频数据分别通过KNN、SVM、决策树、基于KNN的集成学习（Bagging）四类分类器，我们得到表3数据.

依据表2的评估指标，我们集成KNN在集中分类器中准确率、查准率和F-score最高，且第二类错误率最低，得出集成KNN模型的表现最好。

表 3: 不同分类器的效果比较

	准确率	查准率	第一类错误率	第二类错误率	F-score
KNN	0.32	0.33	1.00	0.80	0.25
SVM	0.42	1.00	0.00	0.67	0.50
Tree	0.37	0.50	0.25	0.50	0.50
集成学习KNN	0.44	0.60	0.67	0.25	0.67

3.3 不同加窗方式的效果比较

为考察加窗时不同窗函数对实验结果的影响，我们对每个语音信号分别用矩形窗、汉明窗和海宁窗进行处理, 其余操作相同，最后得到三种不同加窗方式处理后的语音信号，并将其分为三组分别进行实验，最终得到实验结果如表4。从表中可以看出，矩形窗和海宁窗测试时准确率相近，且均高于汉明窗的准确率。F值能够综合考虑精确率(Precision)和召回率(Recall)，矩形窗与汉明窗F值相同，且均高于海宁窗。综合上述因素，可以得出矩形窗处理后的信号语音识别效果最好

3.4 消融实验

为了更好的证明集成学习KNN模型在集中模型中性能的优良性，我们进行了消融实验，在前面的实验中我们取的三个特征向量分别是音频的幅值、过零率和能量，在这一部分，分别只采用幅值

表 4: 加窗方式对结果的影响

窗口种类	准确率	查准率	第一类错误率	第二类错误率	F-score
矩形窗	0.44	0.6	0.67	0.25	0.67
汉明窗	0.39	1	0	0.5	0.67
海宁窗	0.45	0.67	1	0.6	0.5

特征向量、过零率特征向量、能量特征向量、幅值+过零率特征向量、幅值+能量特征向量、过零率+能量特征向量在不同分类器进行分类，相关名词解释和定义如表5。

表 5: 名词解释

变量	定义
幅值特征向量	A
过零率特征向量	Z
能量特征向量	E
幅值+过零率特征向量	AZ
幅值+能量特征向量	AE
过零率+能量特征向量	ZE
幅值+过零率+能量特征向量	AZE

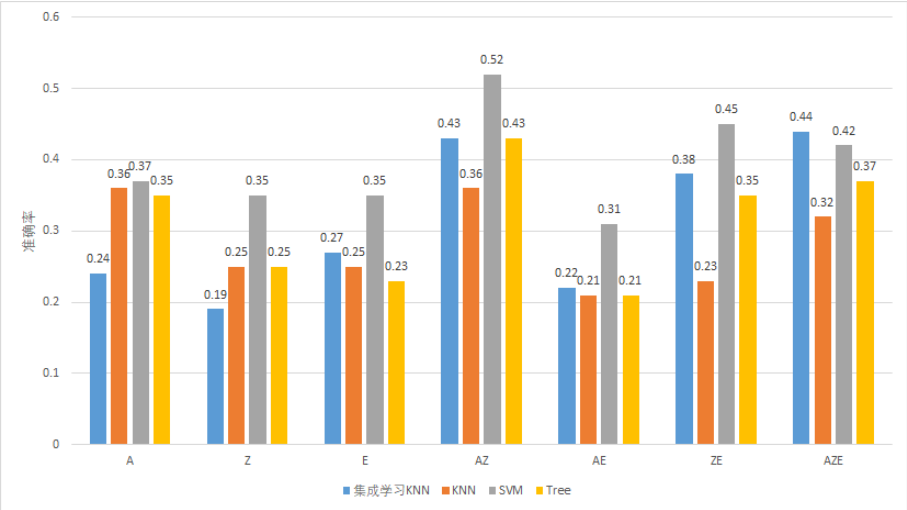


图 12: 消融实验结果

得到的得分率如图12所示，我们可以通过观察结果得到，在特征向量种类较少的时候SVM分类器在准确率上的表现最好，而且在只用幅值和过零率特征向量进行分类时，SVM分类器得到的结果有着超过50%的准确率，达到了本次实验最高的准确率；集成学习KNN在准确率上的表现一般，但是随着特征向量的种类逐渐增加到三个时，集成学习KNN在准确率上的表现最好，SVM随着特征向量的增多，准确率有着一定程度的下降。

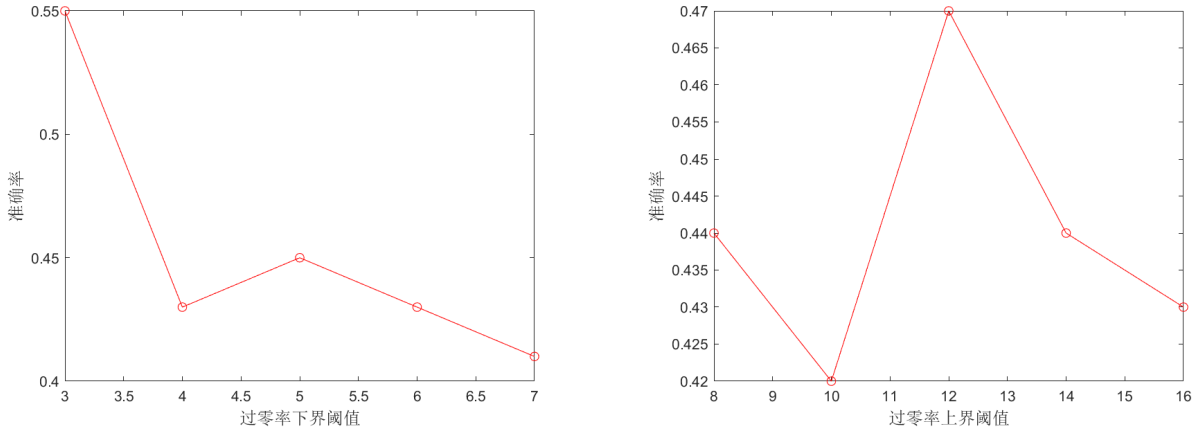
3.5 参数测试

在语音数据处理过程中有四个参数：短时能量门限和过零率门限，以及帧长和帧移值的大小，为探究这些参数对结果的影响，我们每次选择一个参数为变量，并保持其他参数不变，以考察参数变化对结果的影响。下面将依次讨论每个参数对分类准确率的影响。

从图13可以看出，随着过零率门限下界的增大，分类准确率呈下降趋势，可见该值若过大可能导致部分清音无法识别，从而导致准确率下降；随着过零率门限上界的增大，分类准确率先上升后下降，在上界值为12的时候准确率达到峰值，由此可见过零率门限上界值在取特定值时能达到最好效果。

从图14可以看出，随着能量门限下界的增大，分类准确率呈上升趋势，可见能量门限下界的增大有助于噪音的去除，提升分类准确率；随着能量门限上界的增大，分类准确率先上升后下降，可见能量上界门限太高，会导致有效语音信息丢失；上界门限太低，增大了噪音的影响，由此可见能量门限上界值在取特定值时能达到最好效果。

从图15可以看出，帧长为5和20时分类准确率较高，大于20时准确率大幅下降；帧移为20和30 时分类准确率较高。另外可以看出帧长和帧移的不同取值对结果影响较大，需要多加测试以找到最优参数。

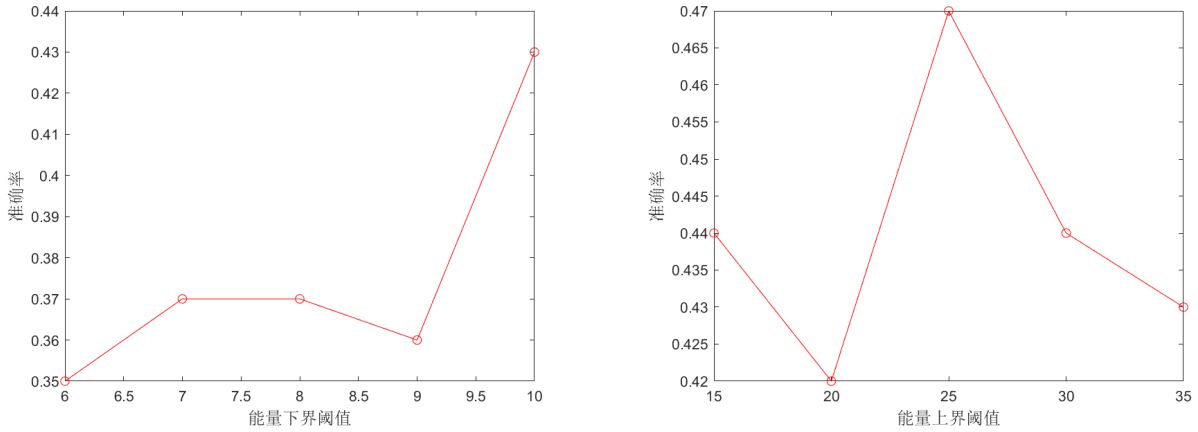


(a) (b)

图 13: 过零率门限对分类准确率影响

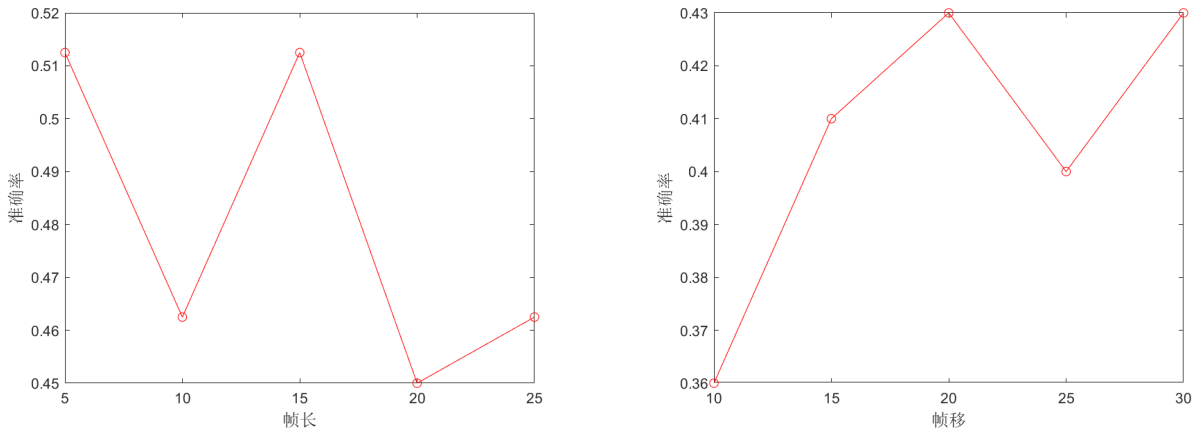
4 结论

本文通过分帧、加窗等方法对收集到的语音信号进行预处理，得到语音信号的特征向量后，使用KNN、SVM等多种机器学习方法对语音信号进行分类识别，最终通过集成学习模型获得了较好的孤立字语音识别效果。在实验部分通过准确率、查准率等多个指标对模型效果进行评估，设计了基于幅值、过零率、能量特征的消融实验，不同分类器的对比实验和参数测试实验，验证了我们模型的有效性。在未来可以收集更多样本训练并设计更多特征组合方式，以增强模型的鲁棒性。



(a)　　(b)

图 14: 能量门限对分类准确率影响



(a)　　(b)

图 15: 帧长、帧移对分类准确率影响

参考文献

[1] 魏伟华.语音合成技术综述及研究现状[J].软件,2020,41(12):214-217.

[2] 廖锎.浅析语音识别技术的发展及趋势[J].科技传播,2010(17):216+206.

[3] 郑南宁.数字信号处理简明教程（第二版）[N].西安：西安交通大学出版社.2019.

[4] 潘文，钱俞寿，周鹗. 基于加窗插值FFT的电力谐波测量理论——(Ⅰ)窗函数研究. 《电工技术学报》，1994

[5] Peterson L E . K-nearest neighbor[J]. Scholarpedia, 2009, 4(2):1883.

[6] Hearst M A , Dumais S T , Osman E , et al. Support vector machines[J]. IEEE Intelligent Systems & Their Applications, 1998, 13(4):18-28.

[7] Gehrke J , Ramakrishnan R , Ganti V . RainForest—A Framework for Fast Decision Tree Construction of Large Datasets[J]. Data Mining & Knowledge Discovery, 2000, 4(2-3):127-162.

[8] Skurichina M , Duin R P W . Bagging, Boosting and the Random Subspace Method for Linear Classifiers[J]. Pattern Analysis & Applications, 2002, 5(2):121-135.