Project 1：语音端点检测

519030910402 包清泉

1. 基于线性分类器和语音短时能量的简单语音端点检测算法

图1展示了以阈值状态机作为分类器为例的工作流，其中分别为预处理与特征提取、分类器以及后处理三个部分。

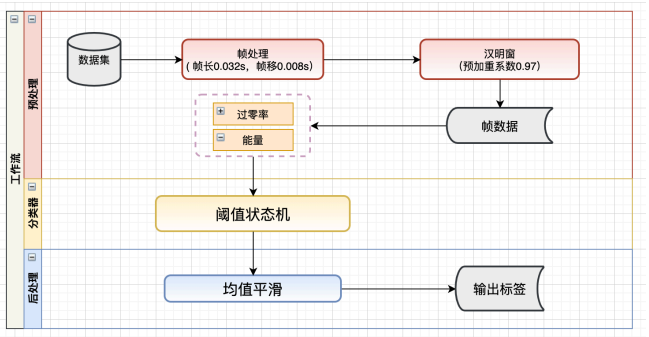

图 1: *Workflow*

1.1. 数据预处理及特征提取

预处理与特征提取分为 3 步骤，分帧、加窗和特征提取。

1.1.1. 分帧

我们假设语音信号具有短时平稳的特性，故而可以通过分帧的方式对于每一帧的信号进行特征提取和分析。在本文中，我们采用帧长 $0.032s$ 和帧移 $0.008s$ 的参数，使得每一帧之间有一定的重叠，更好表现语音的连续变化。

1.1.2. 加窗

对于每帧的数据，我们使用 Hamming 窗进行处理，从而减少窗边界的影响 [2]。Hamming 窗的数学表示如等式1，

$$w(n) = 0.54 - 0.46\cos(2\pi n M - 1) \quad 0 \le n \le M-1 \tag{1}$$

另外，我们还对每一帧的数据进行了预加重处理，见公式2，从而提升信号的信噪比。

$$x_n = x_n - \alpha x_{n-1} \tag{2}$$

其中我们设置预加重系数 $\alpha = 0.97$。

1.1.3. 特征提取

对于每一帧的数据，我们提取了两维的短时特征——短时过零率和短时能量。

短时过零率衡量了一段语音信号中清音或浊音的可能性：一般而言，过零率越高，该短时音频更有可能是清音。其数学表示如公式3，

$$\mathrm{ZCR} = \sum_{n=0}^{N-1} sgn\left(s[n] \cdot s[n+1]\right), \tag{3}$$

其中 s 表示语音序列，N 表示为一帧的数据大小，$sgn(x) \triangleq \mathbf{1}[x \ge 0]$ 是符号函数。

短时能量衡量了一段语言信号强度的大小，能量越大，信号往往更有可能是语言部分，其定义如公式4，

$$\mathrm{Energy} = \sum_{n=0}^{N} s^2[n], \tag{4}$$

其中 x 表示语音序列，N 表示为一帧的数据大小。

1.2. 算法描述

我们引入了两种分类器来进行分类，其一为简单的线性分类器 Logistic Regression，其二为阈值状态机。

1.2.1. 线性分类器

对于二分类问题（本文中即人声部分和非人声部分），最常见的模型便是 Logistic Regression [1]，

表 1: 实验结果

Model	AUC	EER	Accuracy
all 1	0.5	0.9999	0.815
Logistic Regression	0.8575	0.2127	0.902
preSmooth ($winlen = 20$) + Logistic Regression	0.9291	0.1060	**0.9512**
StateMachine + postSmooth ($winlen = 40$)	**0.9453**	**0.0943**	0.9316 (threshold=0.3)

如公式5

$$\hat{y} = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}+\mathbf{b})}}, \tag{5}$$

其中 $\mathbf{w}, \mathbf{b}$ 为参数，$\mathbf{x}$ 为某一帧的特征向量。我们使用开发集来拟合已经提取的特征向量。

1.2.2. 阈值状态机

阈值状态机是一种基于阈值来切换状态的模型。在本文中，我们仅有两个状态，人声和非人声。我们对数据的每个维度人为设定阈值，当每个维度满足一定的阈值条件便进行状态切换。如图2展现了我们使用的状态机模型，我们设定当所有维度均满足一定阈值条件后，方进行状态转移。

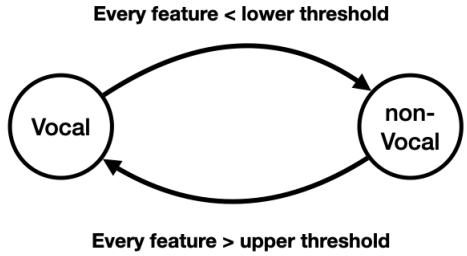

图 2: *State machine based on thresholds*

1.3. 后处理

我们使用了均值平滑作为预测结果的后处理。它的数学形式可以用卷积来表达，如公式6。

$$\mathbf{x} = \mathbf{x} * \mathbf{w} \tag{6}$$

其中 $\mathbf{w}$ 是一个长度为 $winlen$，大小为 $\frac{1}{winlen}$ 的序列。

1.4. 实验结果

我们选用了三个指标来评价我们的模型在开发集上表现的优劣，分别是 Area Under Curve (AUC), Equal Error Rate (EER), 和准确率 (Accuracy)。前二者的指标均是基于 Receiver Operating Characteristic 曲线形成的指标：AUC 越大，其反映了模型的预测值对于各阈值的泛化性能越佳；而 EER 越小，其反映了当模型的误识率和误拒率相等时模型预测出错越少。

如表1，我们对比了 4 种方式。all 1 意思是直接将所有时间点预测为人声 (vocal) 的状态，它将作为基准线。preSmooth + Logistic Regression 先对特征使用了公式6的平滑操作，其中窗长 $winlen = 20$，再进行 Logistic Regression 的拟合，获得了最高的准确率。StateMachine + postSmooth 先使用阈值状态机，再使用窗长 $winlen = 40$ 的均值平滑操作，获得了最高的 AUC 和最低的 EER。

我们最终提交的标签选用了 preSmooth + Logistic Regression 的模型方案，主要是鉴于其高准确率和不差的 AUC 与 EER。

2. References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*.

[2] Fredric J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. In *Proc. IEEE*, pages 51–83, 1978.