# 基于时域分析技术的语音识别实验研究

任翌玮，李相宜，马茂原

## 摘要

本文介绍了一种基于时域分析技术的语音识别实验方法，该方法利用语音信号的短时能量、短时平均幅度和短时过零率等特征参数，结合基于双门限法的端点检测技术，提取语音特征向量，并运用不同的分类器算法实现自动语音判别。本文详细描述了实验的设计、实现和结果分析，并对各个分类器算法的性能进行了比较和评价。

**关键词：**语音识别；时域分析；特征提取；分类器

## 引言

语音识别是指将人类的语音信号转换为文本或命令的技术，它是人机交互的重要手段之一，也是人工智能领域的热门研究方向之一[1]。语音识别的难点在于语音信号的复杂性和多样性，不同的说话人、环境、口音、情感等都会影响语音信号的特征。因此，如何有效地提取语音信号的特征并进行分类是语音识别的关键问题。本文采用了一种基于时域分析技术的语音识别实验方法，该方法主要包括以下几个步骤：（1）对语音信号进行采集、读取和预处理；（2）对语音信号进行分帧和加窗处理；（3）分析和提取语音信号的时域参数；（4）运用基于双门限法的端点检测技术，提取语音特征向量；（5）

选用不同的分类器算法，实现自动语音判别。本文以 0-9 这十个数字作为实验对象，采用 MATLAB 编程实现了上述步骤，并对实验结果进行了分析和评价。

## 语音信号的采集、读取和预处理

本文使用 MATLAB 进行了语音信号的自动化采集工作，采集 0-9 这十个数字的 wav 文件，每个类别包含十组样本。对本小组的三人的语音信号全部进行了采集数据。通过互联网调研以及 MATLAB 编程，找到并理解其中与本任务密切相关的字段，如采样率、频率、频道数、位数等，使用 MATLAB 实现了编程对其中语音数据字段的读取功能，并且对语音信号进行了预处理工作。预处理工作包括去直流化、归一化和端点检测等。

1.语音信号的分帧和加窗处理

为了方便对语音信号进行时域分析，需要将连续的语音信号划分为若干个短时帧，并在每个帧上加上一个窗函数。窗函数可以减少帧与帧之间的不连续性，提高时域参数的计算精度。本文分别运用了三种窗函数（矩形窗[2]、汉明窗[3]和海宁窗[4]），对语音进行分帧和加窗处理，并且分析了三种窗函数的优劣。

2.语音信号的时域参数分析和提取

本文选用了三个最基本的时域参数作为语音信号的特征，分别是短时能量[5]、短时平均幅度和短时过零率。短时能量反映了语音信号的强度，短时平均幅度反映了语音信号的振幅，短时过零率反映了语音信号的频率。本文首先按照时间顺序统计采样点数值符号变号的次数，然后将上述计述出的次数针对序列时长进行归一化操作，获取了语音信号的三个时域参数。

首先，我们用 matlab 中的函数 audiorecorder 记录声音为 wav 格式，并且对语音信号进行分帧处理。之后，我们对每一帧进行加窗处理，即 $s_{w(n)} = s(n) * w(n)$

其中,

矩形窗 $w(n) = \begin{cases} 1, & 0 < n < N-1 \\ 0, & 其他 \end{cases}$

汉明窗 $w(n) = \begin{cases} 0.54 - 0.46 * \cos\left[\frac{2\pi n}{N-1}\right], & 0 \le n \le N-1 \\ 0, & 其他 \end{cases}$

海宁窗 $w(n) = \begin{cases} 0.5(1 - \cos\left[\frac{2\pi n}{N-1}\right]), & 0 \le n \le N-1 \\ 0, & 其他 \end{cases}$

最后我们对音频数据进行短时时域特性分析，包括短时能量$E_n$和短时平均幅度$M_n$，过零率$Z_n$

其中，

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

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

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

2. 语音信号的端点检测和特征向量提取

为了区分语音段和静音段，需要对语音信号进行端点检测。本文采用了基于双门限法[6]的端点检测技术，该技术通过分别为短时能量和过零率确定两个门限，分别判断出浊音和静音，静音和清音，提取出语音特征向量。语音特征向量是由若干个时域参数组成的向量，它可以表示语音信号的特征。

## 实验方法

1. 语音特征向量的归一化和聚类

由于每个语音特征向量的值没有进行归一化，不能直接利用每个数字的特征向量进行运算，需要根据每个特征进行归一化。每个数字的特征向量的长度不同，不能直接利用每个数字的特征向量进行分类，需要对特征向量进行聚类操作，从而使得每个数字的特征向量的维度相同。

首先，我们对每个数字的特征进行数据清洗，清除由于测量而导致粗大误差的数据。然后，我们对每个数字的特征进行数据归一化，避免了“大数吃小数”的情况，充分利用获得的特征数据。最后，我们对特征向量进行聚类操作，从而使得每个数字的特征向量的维度相同，为后续的分类工作做好铺垫，最终我们将每个语言的特征向量用聚类所获得的两个点的坐标(x1, y1, z1)和（x2, y2, z2）进行表示，每个特征向量为六维，为(x1, y1, z1, x2, y2, z2)。我们将经过数据预处理之后得到的每个语音特征向量进行了可视化，如图 1 所示。

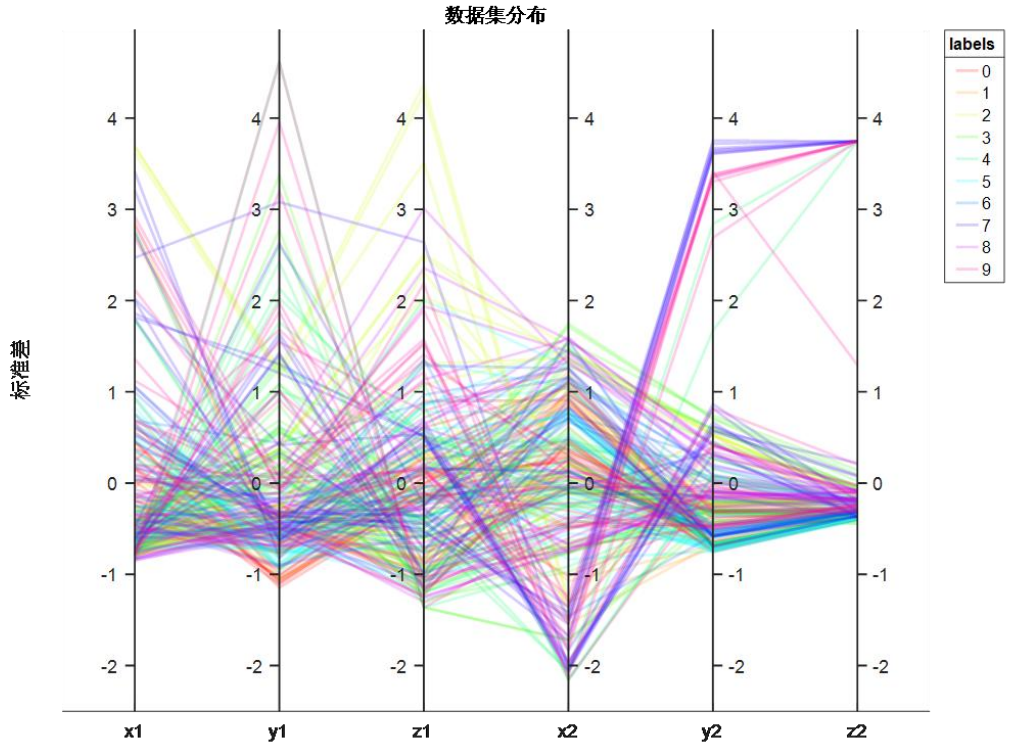


图 1 经过数据预处理之后的每个语音特征向量的可视化

2. 语音信号的自动判别和分类器算法比较

为了实现对 0-9 这十个数字的自动判别，需要选用合适的分类器算法来对语音特征向量进行分类。本文选用了四种常用的分类器算法，分别是线性判别[7]、决策树[8]、支持向量机[9]和最近临分类器[10]。本文通过一定数量的实验结果分析各个算法的性能，并将语音识别的准确度、AUC 等各种指标进行统计分析和对比。

## 对比实验与消融实验

1. 对比实验

首先，我们将不同的分类器方法进行对比。我们统一选用矩形窗，分别使用线性判别、决策树、支持向量机和最近临分类器对语音特征向量进行分类，实验结果图 2-6 所示。

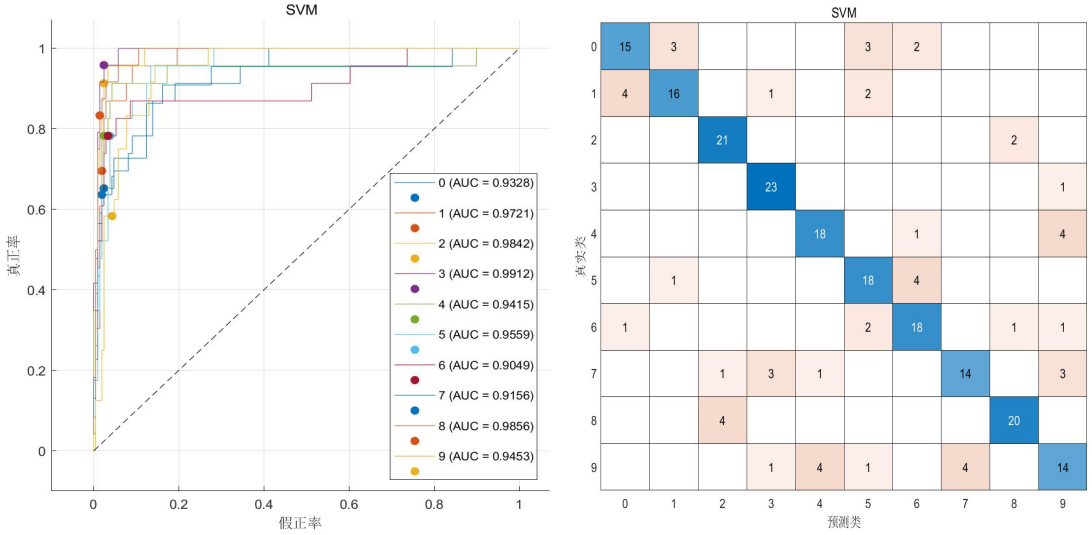


图 2 使用 SVM 分类器得到的 ROC 曲线和混淆矩阵

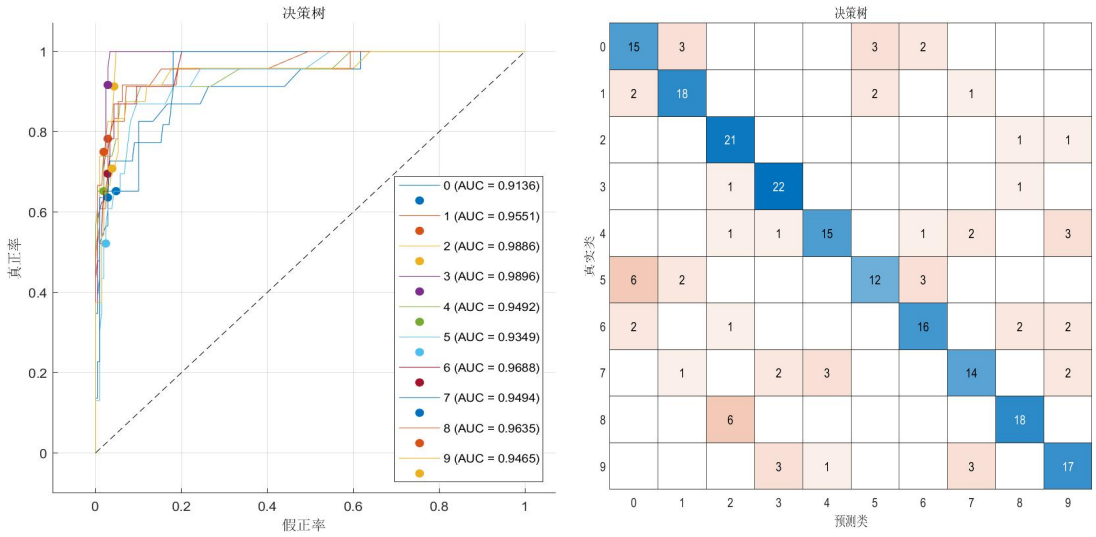


图 3 使用决策树分类器得到的 ROC 曲线和混淆矩阵

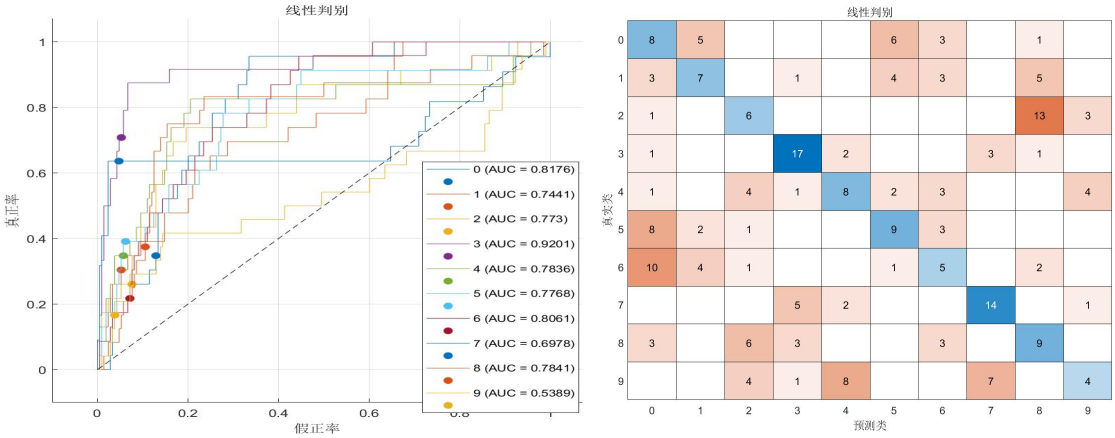


图 4 使用线性判别分类器得到的 ROC 曲线和混淆矩阵

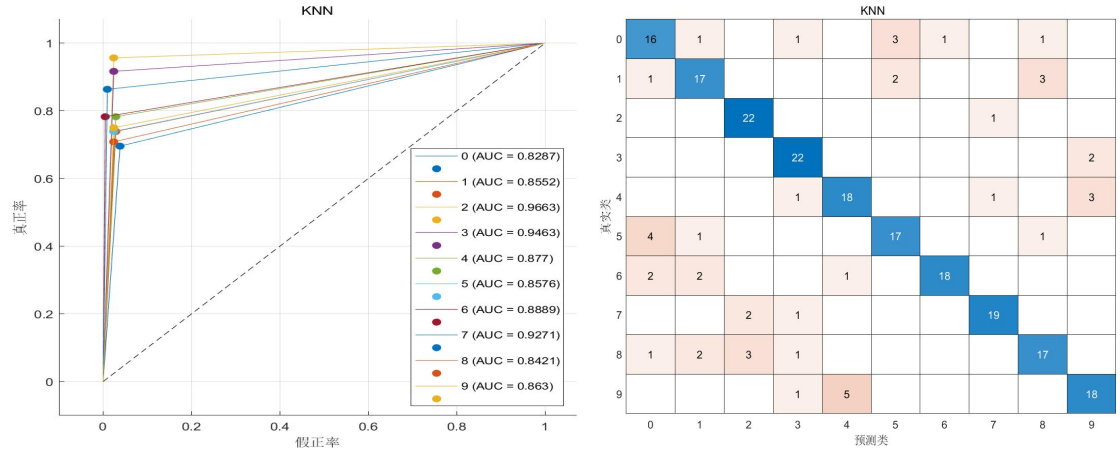


图 5 使用 KNN 分类器得到的 ROC 曲线和混淆矩阵

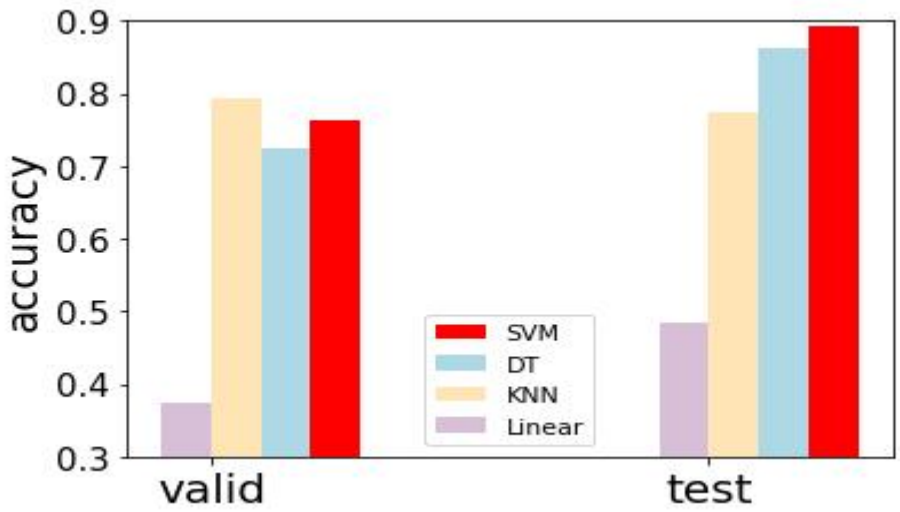


图 6 验证集和测试集中四种分类器的准确度

由图 2-6 可以看出，SVM 分类器在 AUC 和准确度上综合评价性能最佳，之后的实验中我们均采用 SVM 分类器。

其次，我们将不同的窗函数进行对比。我们统一选用 SVM 分类器，分别使用矩形窗、汉明窗和海宁窗作为窗函数，实验结果图 7-9 所示。

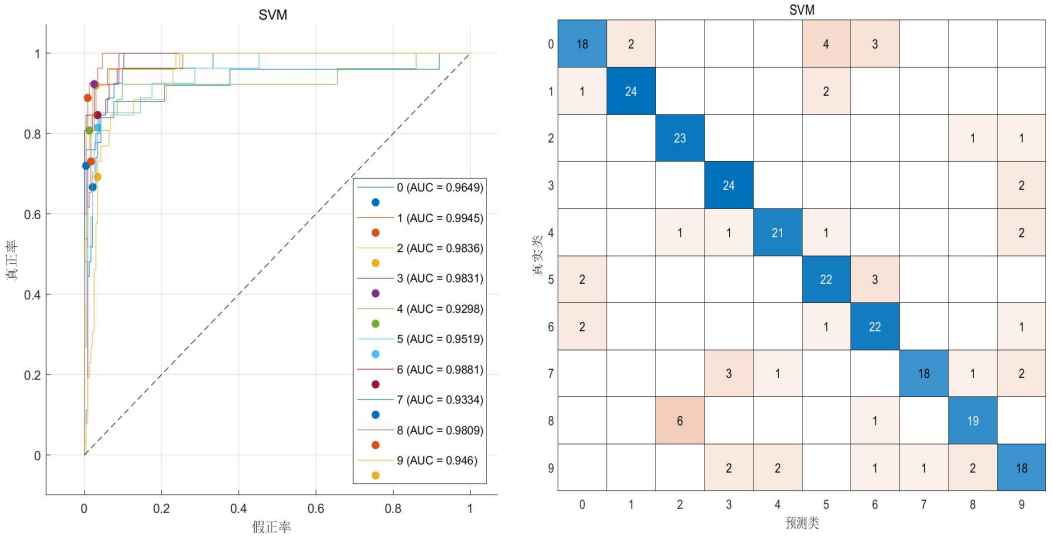


图 7 使用汉明窗得到的 ROC 曲线和混淆矩阵

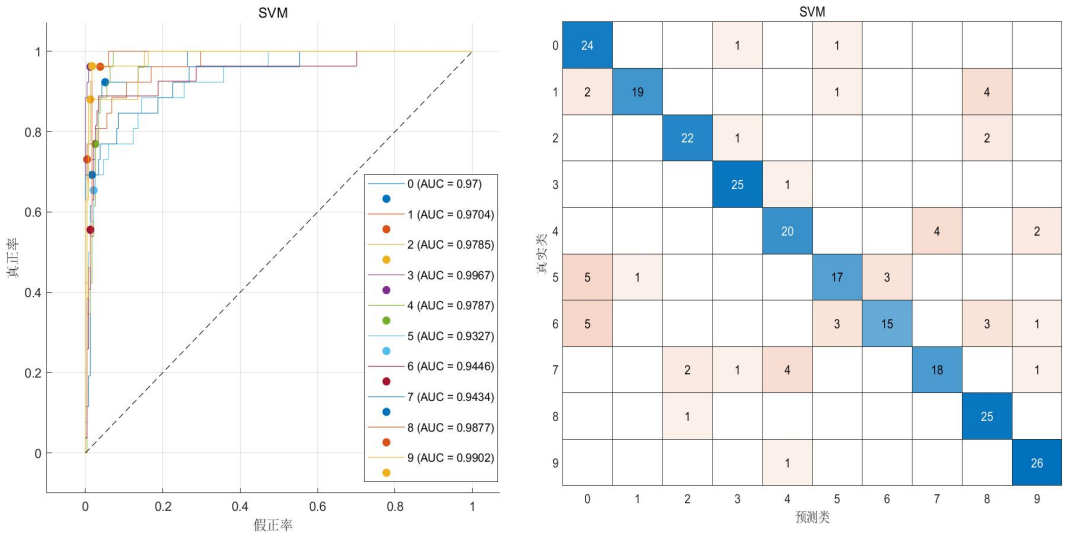


图 8 使用海宁窗得到的 ROC 曲线和混淆矩阵

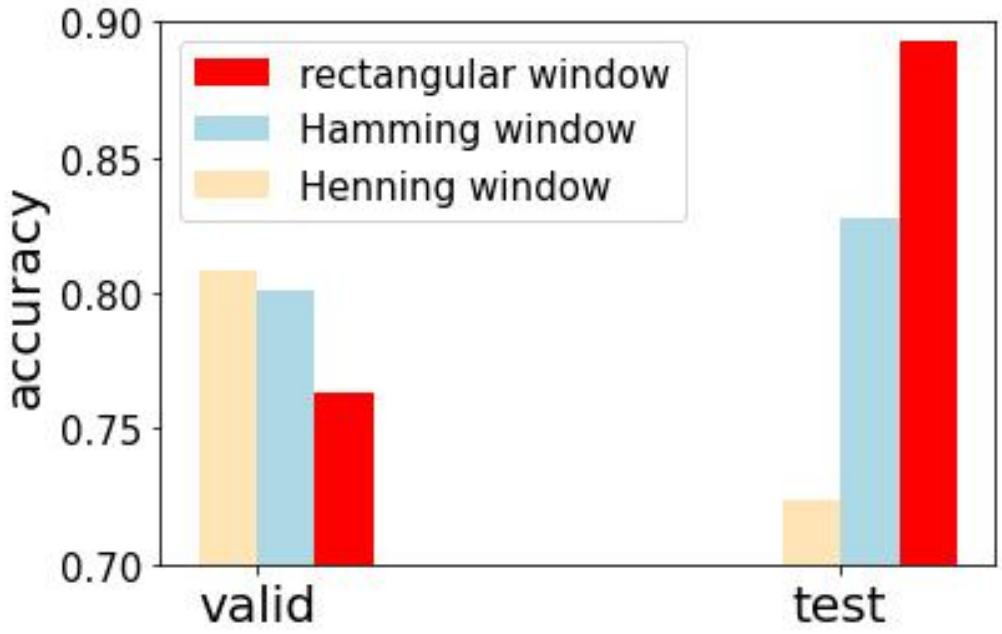


图 9 验证集和测试集中三种窗函数的准确度

由图 7-9 可以看出,矩形窗在 AUC 和准确度上综合评价性能最佳,是在测试集上的表现尤为突出。之后的实验中我们均采用矩形窗作为窗函数。

2. 消融实验

我们有三个语音信号的特征参数，分别是短时能量、短时平均幅度和短时过零率，为了明确这三个特征参数的重要性程度，我们依次排除每一个特征参数，再对每个语音的特征向量进行分类，根据上文的实验结果,我们均采用矩形窗作为窗函数,均采用 SVM 作为分类器。

实验结果如图 10-13 所示。

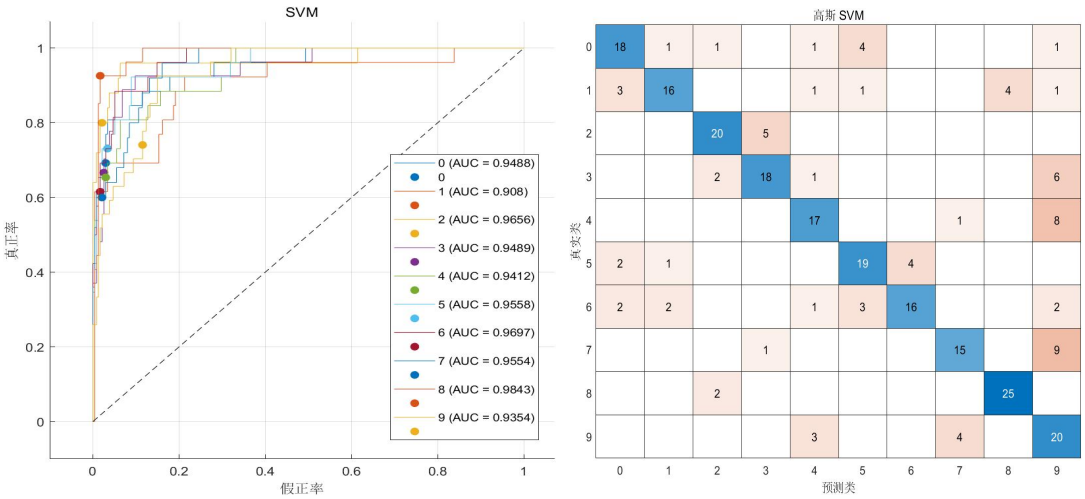


图 10 缺少短时平均幅度所得到的 ROC 曲线和混淆矩阵

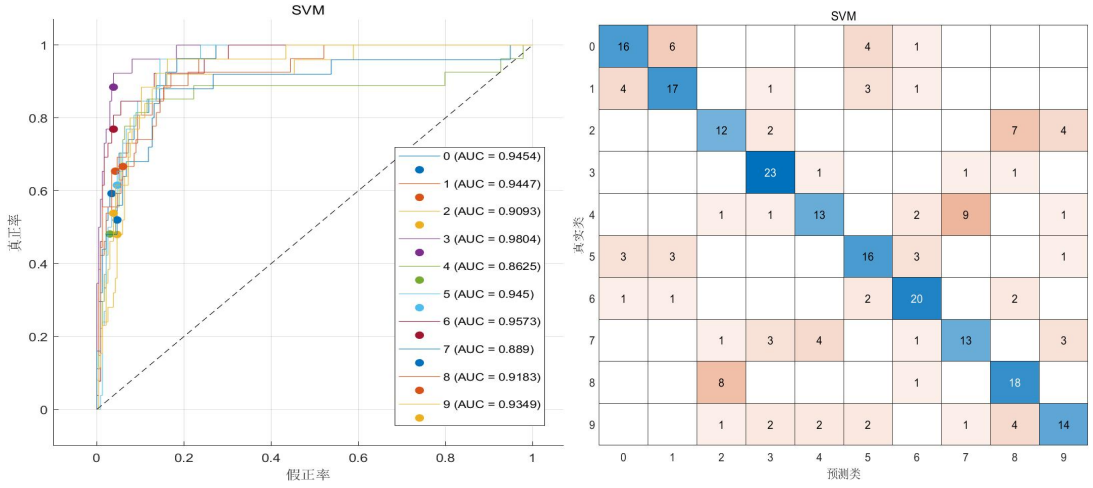


图 11 缺少短时能量所得到的 ROC 曲线和混淆矩阵

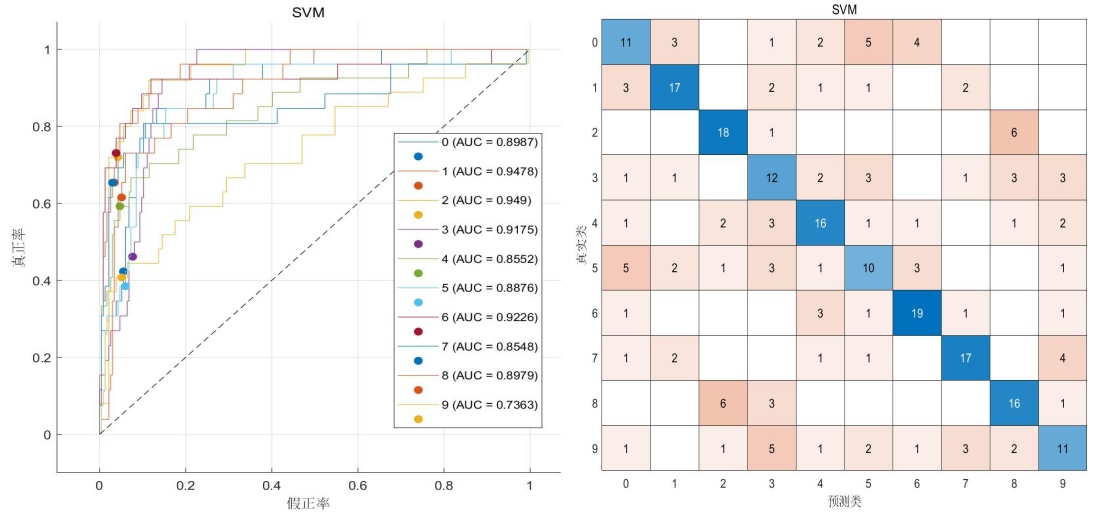


图 12 缺少短时过零率所得到的 ROC 曲线和混淆矩阵

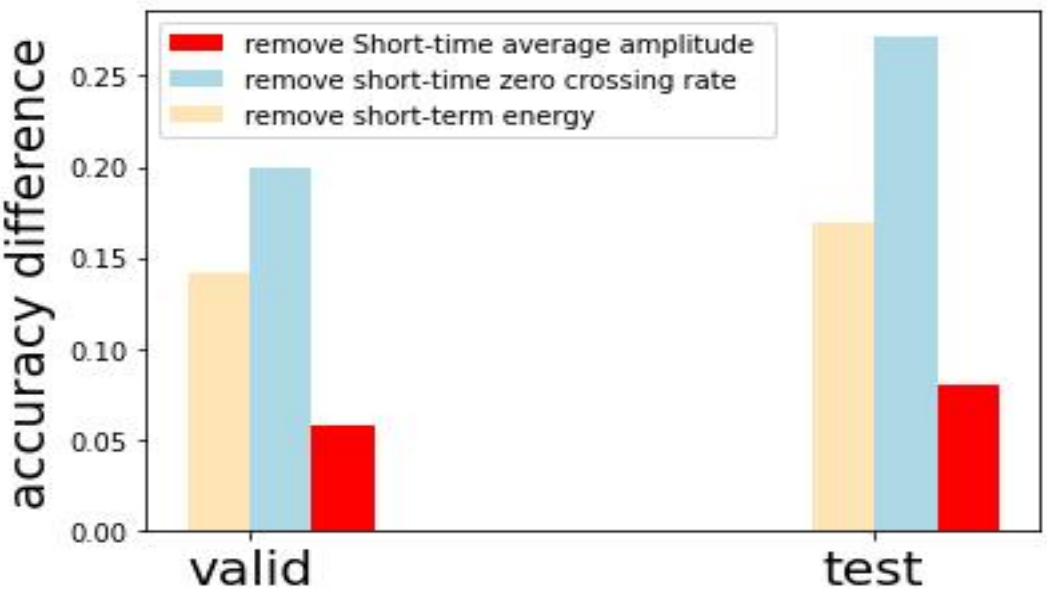


图 13 缺少某个特征参数后的准确率差值

由图 10-13 可以看出，短时能量、短时平均幅度和短时过零率任意一个的缺失，均会对分类的效果产生影响，所以短时能量、短时平均幅度和短时过零率均为语音信号的重要特征参数。其中短时能量、和短时过零率对结果的影响较大，短时平均幅度对结果的影响较小。

## 总结

本文介绍了一种基于时域分析技术的语音识别实验方法，该方法利用语音信号的短时能量、短时平均幅度和短时过零率等特征参数，结合基于双门限法的端点检测技术，提取语音特征向量，并运用不同的分类器算法实现自动语音判别。本文详细描述了实验的设计、实现和结果分析，并对各个分类器算法的性能、窗函数的选择和特征参数的选择进行了对比实验和消融实验。

## 参考文献

[1] P. Vary, “An adaptive filter-bank equalizer for speech enhancement,” *Signal Process.*, vol. 86, no. 6, pp. 1206–1214, Jun. 2006, doi: 10.1016/j.sigpro.2005.06.020.

[2] E. Loweimi, S. M. Ahadi, T. Drugman, and S. Loveymi, “On the Importance of Pre-emphasis and Window Shape in Phase-Based Speech Recognition,” in *Advances in Nonlinear Speech Processing*, T. Drugman and T. Dutoit, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013, pp. 160–167. doi: 10.1007/978-3-642-38847-7_21.

[3] A. Testa, D. Gallo, and R. Langella, “On the Processing of harmonics and interharmonics: using Hanning window in standard framework,” *IEEE Trans. Power Deliv.*, vol. 19, no. 1, pp. 28–34, Jan. 2004, doi: 10.1109/TPWRD.2003.820437.

[4] M. Mottaghi-Kashtiban and M. G. Shayesteh, “New efficient window function, replacement for the Hamming window,” *IET Signal Process.*, vol. 5, no. 5, pp. 499–505, Aug. 2011, doi: 10.1049/iet-spr.2010.0272.

[5] F. Jabloun, A. E. Cetin, and E. Erzin, “Teager

energy based feature parameters for speech recognition in car noise," *IEEE Signal Process. Lett.*, vol. 6, no. 10, pp. 259–261, Oct. 1999, doi: 10.1109/97.789604.

[6] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 3, pp. 146–157, Mar. 2002, doi: 10.1109/TSA.2002.1001979.

[7] O. Siohan, "On the robustness of linear discriminant analysis as a preprocessing step for noisy speech recognition," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, May 1995, pp. 125–128 vol.1. doi: 10.1109/ICASSP.1995.479289.

[8] P.-F. WONG and M.-H. SIU, "Decision tree based tone modeling for Chinese speech recognition," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2004, p. I–905. doi: 10.1109/ICASSP.2004.1326133.

[9] B. A. Sonkamble and D. D. Doye, "An overview of

speech recognition system based on the support vector machines," in *2008 International Conference on Computer and Communication Engineering*, May 2008, pp. 768–771. doi: 10.1109/ICCCE.2008.4580709.

[10] T.-L. Pao, W.-Y. Liao, and Y.-T. Chen, "Audio-Visual Speech Recognition with Weighted KNN-based Classification in Mandarin Database," in *Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2007)*, Nov. 2007, pp. 39–42. doi: 10.1109/IIHMSP.2007.4457488.