第 9 卷 第 1 期
2020 年 1 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 9 No. 1
Jan. 2020

引文格式：
王小晨, 朱明星, 杨子健, 等. 基于高密度肌电的对称位置发音肌肉对语音识别贡献的研究［J］. 集成技术, 2020, 9(1): 55-65.
Wang XC, Zhu MX, Yang ZJ, et al. The study on the left/right contributions of articulatory musclesin speech recognition using high-density surface electromyography［J］. Journal of Integration Technology, 2020, 9(1): 55-65.

基于高密度肌电的对称位置发音肌肉对语音识别贡献的研究

王小晨[1,2] 朱明星[1,2] 杨子健[1] 汪 鑫[1,2] 黄剑平[1] 陈世雄[1] 李光林[1]

[1](中国科学院深圳先进技术研究院 深圳 518055)
[2](中国科学院大学深圳先进技术学院 深圳 518055)

摘 要 说话是人类正常生活中最重要的技能之一，是发音相关肌肉在神经中枢的控制下协调运动的结果。表面肌电图法(Surface Electromyography，sEMG)是目前采集肌肉电信号的常用方法，能检测到可靠的肌肉电生理信息。用肌电信号进行语音分类时，所选的电极位置对分类精度有重大作用。但目前基于 sEMG 的语音识别方法选取电极位置及数量时没有一个客观的指标，也不清楚发音相关的面、颈部左右两侧对称位置电极对肌电语音识别的贡献是否冗余。该文使用 120 通道电极(关于面中、颈中对称)采集了 8 名发音正常的受试者分别发 5 个中文单词和 5 个英文单词时的面、颈部 sEMG，考察了面、颈部左右两侧对称位置 sEMG 对语音识别的贡献。结果表明，发音过程中面、颈部左右两侧肌肉活动有相似的变化规律，但整个活动过程中面部对称位置的相关性比颈部低；使用颈部左侧、右侧的肌电信号进行语音分类的分类精度区别不大，而使用面部左、右两侧肌电信号的分类结果差异较明显。因此，颈部对称位置的 sEMG 信号对语音识别贡献程度具有一致性，而面部则不具有，这为后续研究减少电极数量和选择最佳通道提供了新思路。

关键词 语音识别；表面肌电法；支持向量机
中图分类号 R 318 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20191124001

收稿日期：2019-11-24 **修回日期**：2019-12-10
基金项目：国家自然科学基金项目(61771462)；广州市科技计划项目(201803010093)
作者简介：王小晨，硕士研究生，研究方向为发音功能评估与构音障碍康复；朱明星，工程师，研究方向为吞咽功能评估与吞咽障碍治疗；杨子健，助理工程师，研究方向为生物医学电子；汪鑫，硕士研究生，研究方向为听力检测、生物医学信号处理；黄剑平，助理工程师，研究方向为探索外周神经功能；陈世雄(通讯作者)，博士，副研究员，研究方向为听力检测与听觉功能的康复，E-mail：sx.chen@siat.ac.cn；李光林，博士，研究员，研究方向为神经康复工程、生物医学信号处理、生物医学仪器等。

The Study on the Left/Right Contributions of Articulatory Muscles in Speech Recognition Using High-Density Surface Electromyography

WANG Xiaochen[1,2] ZHU Mingxing[1,2] YANG Zijian[1] WANG Xin[1,2] HUANG Jianping[1]
CHEN Shixiong[1] LI Guanglin[1]

[1](*Shenzhen Institutes of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[2](*Shenzhen College of Advanced Technology*, *University of Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)

Abstract Speech is one of the most important skills in human normal life. It is the result of the coordinated movement of the articulation-related muscles under the control of central cervous system. Surface electromyography (sEMG) is a commonly used method for collecting electrical signals of muscles, which can detect reliable electrophysiological information. When using electromyographic signals on speech classification, the selected electrode position plays an important role in classification accuracy. However, the current sEMG-based speech recognition method does not have an objective index for selecting the position and number of electrodes, and it is still unclear whether the contribution of the articulation related symmetrical position electrodes on the left and right sides of the face and neck to speech recognition is redundant. In this study, the facial and neck sEMG of 8 subjects with normal pronunciation were collected by using a 120-channel electrode (about facial and neck symmetry) when they pronounced 5 Chinese words and 5 English words respectively. The contribution of sEMG in the symmetrical position of left and right sides of facial and neck to speech recognition was investigated. The results show that the muscles of the left and right sides of the face and neck had similar variation, but the correlation between the symmetrical positions of the face and neck was lower than that of the neck. There was little difference in classification accuracy between the left and right sEMG signals of the neck, but significant difference between the left and right SEMG signals of the face. Thus, sEMG signals from symmetrical positions in the neck are consistent in their contribution to speech recognition, whereas facial signals are not, which might provide useful clue to reduce the electrode number and select the optimal location of channels for speech recognition.

Keywords speech recognition; surface electromyography; support vector machine

1 引 言

说话是人类特有的表达情感、传递信息、参与社会活动的交流方式[1]，是人类正常生活中最重要的技能之一。无论是在生活还是在工作中，都不可避免地需要通过说话与他人交流。说话是一个非常复杂的面颈部多块肌肉在中枢神经系统的控制下协同收缩运动的过程，这伴随着肌肉电信号的产生[2-3]。发不同的音时，发音肌肉的收缩模式、收缩力量和协同方式是不同的，对应的肌肉电信号特征也会不同[4]。表面肌电图法（Surface Electromyography，sEMG）是目前采集肌电信号的常用方法，能通过无创、简单、稳定的操作，检测到可靠的肌肉电生理信息[5]，因此被广泛用于肌电语音识别研究。

早在 1985 年，第一个使用肌电信号进行语音识别的研究就在 Sugie 和 Tsunoda[6]的实验室展开，他们采集口腔附近的肌电信号对 5 种日语元音字母进行分类。1989 年，Morse 等[7]提取 sEMG 信号幅值、方差等特征值分类 10 个英文

单词，分类精度达到了 60%。2018 年，Srisuwan 等[8]在受试者的面颈部共 6 个位置贴上肌电电极，以评估 14 个特征评估标准及 4 种分类器对单个泰语单词进行分类时的性能，并找到一种接近最佳的标准和分类算法。Janke 等[9]对从受试者发音时面颈部 6 个位置采集到的肌电数据进行研究，捕捉到从发音肌肉运动时产生的 sEMG 信号到语音波形的映射。Jong 和 Phukpattaranont[10]招募 7 名健康受试者和 5 名构音障碍受试者开发了一个语音识别系统。该系统使用从 12 名受试者脸部和颈部共 5 个通道里记录的 sEMG 信号对 9 个泰国音节进行分类。Diener 等[11]使用 sEMG 技术在语音识别方面做了大量工作，提出了使用深度神经网络实现从表面肌电信号到目标声学语音输出的映射。

上述研究中，设置的电极数量均较少，且选取的电极位置都不同，分类结果也有显著差异。由于发音过程涉及到的肌肉多达 30 余块[12]，使用肌电信号进行语音分类时，电极的位置和数量会对分类准确性产生重要影响[13]。而目前基于 sEMG 的语音识别方法选取电极位置及数量时没有一个客观的指标，也不清楚与发音相关的面颈部左右两侧对称位置电极对肌电语音识别的贡献是否存在冗余[14-15]。

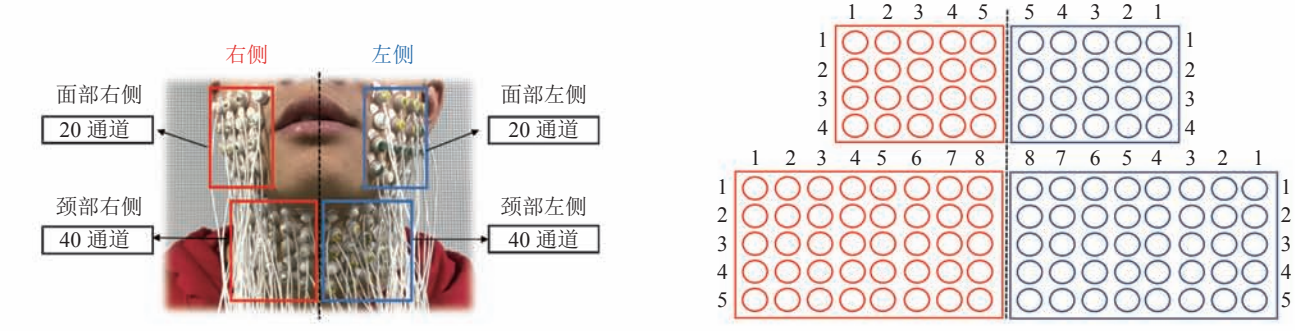
为解决后一个问题，本文提出使用几乎覆盖全部发音肌肉的高密度肌电电极，探究面、颈部左右两侧对称位置电极对肌电语音识别的贡献。首先，使用关于面中部、颈中部对称的共 120 通道电极采集 8 名发音正常的受试者的表面肌电信号。其中，发音测试为 5 个中文单词和 5 个英文单词。然后，对信号预处理后分组提取 4 种时域特征输入支持向量机(Support Vector Machine，SVM)分类器，进行语音分类。最后，对分类结果进行分析，比较面、颈部对称位置肌电信号在语音识别时的贡献程度。

2　实验方法

2.1　信号采集方法

本研究共招募 8 名健康受试者(sub1～8)，其中 6 名男生、2 名女生，年龄为 22～26 岁(平均年龄为 24 岁)。所有受试者均未患有可能影响实验结果的说话和吞咽问题。实验开始前，受试者均阅读知情同意书并签字，且允许出于科学目的公开发表他们的照片和数据。

本实验使用荷兰 TMS 公司研发的高密度肌电采集系统(REFA 128-model system)，以 2 048 Hz 采样率采集面、颈部共 120 通道高密度肌电信号。其中，电极对称放置于受试者面、颈部，分为面部左侧(20 个通道)、面部右侧(20 个通道)、颈部左侧(40 个通道)、颈部右侧(40 个通道)4 个区域，如图 1(a)所示。通道以面、颈部



(a) 四大区域　　(b) 对称排列

图 1　高密度表面肌电电极在面颈部左右两侧的分布

Fig. 1　Distribution of the high density sEMG electrodes on the left and right sides of the face/neck regions

中间位置为对称轴左右对称放置，行、列编号如图 1(b)所示。实验前，使用酒精棉擦拭电极位置，清除皮肤表面的油脂和角质。实验在屏蔽房中进行，以保证测试过程相对安静，受试者发音不被影响。整个实验过程符合中国科学院深圳先进技术研究院人体实验伦理道德规范(审批编号为 SIAT-IRB-170815-H0178)。

2.2 实验过程

实验时，受试者调整舒服的姿势坐在椅子上，保持 40 s 的静息状态(不说话、也不做任何身体运动)，记录下此时的肌电信号作为基线(P11)。随后，受试者按照平时说话的音量及音调进行 10 组发音任务，包含英文 5 组单词：“Thanks”(P1)、“Yes”(P2)、“No”(P3)、“Hello”(P4)和“Goodbye”(P5)，以及对应着相同含义的中文 5 组日常短语：“谢谢”(P6)、“是的”(P7)、“不是”(P8)、“你好”(P9)和“再见”(P10)，具体如表 1 所示。每组任务包括 1 s 的发音过程和 3 s 的休息，两过程交叉连贯，共重复 6 次，以采集整个过程的表面肌电信号。

表 1 5 组英文和 5 组中文发音任务

Table 1 Speaking tasks of five Chinese words and five English words

序号	发音任务	音标
P1	Thanks	/θæŋks/
P2	Yes	/jes/
P3	No	/nəʊ/
P4	Hello	/hə'ləʊ/
P5	Goodbye	/gʊd'baɪ/
P6	谢谢	/xiexie/
P7	是的	/shide/
P8	不是	/bushi/
P9	你好	/nihao/
P10	再见	/zaijian/

2.3 信号处理

由于采集到的原始肌电信号不够干净，即混杂着心电、运动伪迹、工频等各种噪声，故分析肌电特征前需对信号做预处理工作。首先，使用 30～500 Hz 的巴特沃斯带通滤波器滤除大量心电干扰和面部伪迹；然后，设置 50 Hz 及其倍数频率的陷波滤波器去除工频干扰，得到较为干净的肌电信号(数据维度为：120×信号长度)。

使用长度为 250 ms 的分析窗口对滤波后的各通道信号计算均方根(Root Mean Square，RMS)，再利用计算出的高密度表面肌电信号的最大和最小 RMS 值对所有通道的 RMS 值进行归一化得到归一化均方根(Normalized Root Mean Square，NRMS)，并画出左右对称位置的 NRMS 叠加图。由于一段发音过程持续时间较短，保留的特征点不足，故首先根据肌电信号原始波形，确定每段发音过程的起始点与结束点，对滤波后的信号进行人工截取，得到 14 段发音活动的肌电信号；然后，将这些信号进行拼接，得到整段全为发音过程的肌电信号，处理过程如图 2 所示；最后，提取零交叉(Zero Crossing，ZC)、斜率符号变化(Slope Sign Change，SSC)、波形长度(Waveform Length，WL)和平均绝对值(Mean Absolute Value，MAV)4 个特征[16]，得到 11×4×N 维度的数据。其中，11 为类别数；4 为特征数；N 为通道数。四种特征的定义如下：

(1)零交叉(ZC)是在时域中定义 EMG 信号频率信息的度量，为单位时间窗口内信号通过零幅值的次数，其定义如公式(1)～(2)所示。

$$ZC=\sum_{i=1}^{N-1}\left[\operatorname{sgn}\left(x_i\times x_{i+1}\right)\cap\left|x_i-x_{i+1}\right|\geqslant \text{阈值}\right] \tag{1}$$

$$\operatorname{sgn}(x)=\begin{cases}1, & x\geqslant \text{阈值}\\ 0, & \text{其他}\end{cases} \tag{2}$$

其中，x_i 为 i 点的 EMG 信号；N 为 EMG 信号的长度。

(2)斜率符号变化(SSC)记录了在单位时间窗内 EMG 信号斜率的改变次数，其定义如公式(3)～(4)所示。

$$SSC=\sum_{i=2}^{N-1}\left[f\left(x_i-x_{i-1}\right)\times\left(x_i-x_{i+1}\right)\right] \tag{3}$$

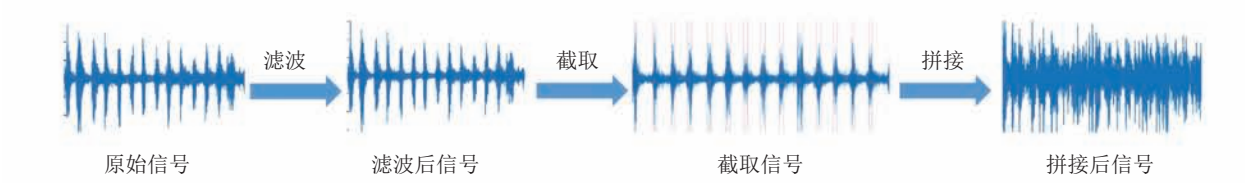


图 2　肌电信号处理过程

Fig. 2　sEMG signal processing

$$f(x)=\begin{cases}1, & x\geqslant \text{阈值}\\ 0, & \text{其他}\end{cases} \tag{4}$$

(3) 波形长度 (WL) 是 EMG 波形在某个时间段上的累计长度积分，其定义如公式 (5) 所示。

$$WL=\sum_{i=1}^{N-1}\left|x_{i+1}-x_i\right| \tag{5}$$

(4) 平均绝对值 (MAV) 是 EMG 信号分析中最常用的一种时域特征，表示一段 EMG 信号幅值绝对值的平均值，可反映肌电的强度，其定义如公式 (6) 所示。

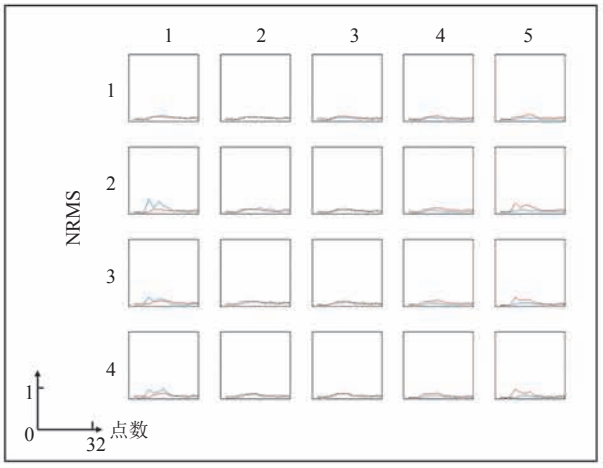
$$MAV=\frac{1}{N}\sum_{i=1}^{N}\left|x_i\right| \tag{6}$$

将特征值输入 SVM 分类器作分类，使用 5 倍交叉验证方法来减少生成训练和测试数据的可变性。其中，SVM 是一种二分类模型，有两大主要优势：更高的速度、用更少的样本 (千以内) 取得更好的表现[17]。这使得该算法非常适合本文分类问题。另外，使用统计方法比较面部和颈部肌肉左右两侧对称阵列的分类精度。

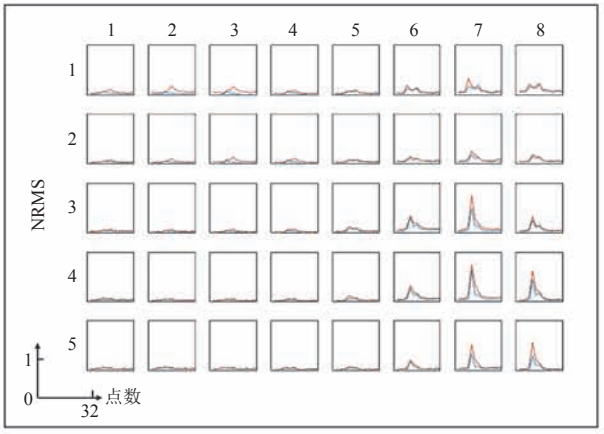
3　实验结果

3.1　发音时面、颈部左右两侧肌肉间的相关程度

图 3 展示了一名受试者说一次“Hello”时的面部 (a)、颈部 (b) 左右对称通道叠加的 NRMS 波形。图中蓝线表示面部/颈部左侧的 NRMS 波形，而红线则表示面部/颈部右侧的 NRMS 波形。从图 3 可以看出，所有的 NRMS 波形均呈现相同的特点：随着发音过程慢慢上升，达到峰值后开始下降直至静息时的水平，面部整体峰值低于颈部。面部和颈部左侧的 NRMS 波形与右侧的波形以相似的速率变化。



(a) 面部



(b) 颈部

图 3　面、颈部左右两侧高密度肌电信号的 NRMS 波形叠加图

Fig. 3　Superposition of NRMS waveforms of high density sEMG recordings from the left and right sides of the facial and neck muscles

相关系数是用于反映变量之间相关关系密切程度的统计指标，能够刻画两个变量之间的相关

程度，*P* 值可以描述相关程度计算结果的“显著程度”[18]。分析面部、颈部左右两侧肌肉之间的相关性有助于理解发音过程中面、颈部对称位置的运动模式相似程度。使用相关系数、*P* 值计算公式对面部、颈部对称通道的 NRMS 波形相似性进行统计，结果如表 2 和 3 所示。表中的序号下标分别对应图 3 波形叠加图的行、列编号，如 F11 表示图 3(a)中 1 行 1 列。可以看出，面部左、右两侧对称通道 NRMS 波形的相关系数范围为 0.395 5～0.929 5，平均值和标准偏差为 0.714 9±0.165 3；颈部对称通道 NRMS 波形的相关系数范围为 0.464 2～0.988 5，平均值和标准偏差为 0.840 5±0.150 6。显然，面部左右两侧的相关性比颈部低，但整体上存在相关性。同时，只有 F21、F31、F41、N12、N13 这 5 个靠近面、颈部边缘对称位置的 NRMS 波形间无显著相关，其余位置的 NRMS 波形间均显著相关。

3.2 使用面部左右两侧对称位置肌电信号的语音分类精度对比

表 4 和 5 分别为使用一名受试者面部左右两侧对称位置的肌电信号进行语音分类的结果。表中对角线上加粗的数据是正确分类的精度，而其余数值则是误识别为其他发音任务的概率。

从表 4 可以看出，静息状态(P11)的分类精度最高，为 1；P1、P3 和 P4 的分类准确率较高，均超过 0.9；而 P6 的分类准确率最低，为 0.647 9。面部左侧的平均分类精度和标准偏差值为 0.823 8±0.106 6。从表 5 可以看出，使用面部右侧肌电信号的分类精度仅在 P1、P3 和 P11 识别任务处高于 0.8；有 6 个识别任务的分类精度

表 2 面部左右两侧对称通道 NRMS 波形的相关系数

Table 2 Correlation coefficients of NRMS waveforms on the left and right sides of the facialmuscles

通道序号	NRMS 波形相关系数	通道序号	NRMS 波形相关系数	通道序号	NRMS 波形相关系数	通道序号	NRMS 波形相关系数
F_{11}	0.648 4**	F_{21}	0.469 8	F_{31}	0.395 5	F_{41}	0.502 0
F_{12}	0.781 8***	F_{22}	0.704 1**	F_{32}	0.571 4*	F_{42}	0.603 3*
F_{13}	0.874 0***	F_{23}	0.857 9***	F_{33}	0.769 2***	F_{43}	0.929 5***
F_{14}	0.828 6***	F_{24}	0.868 3***	F_{34}	0.858 4***	F_{44}	0.912 6***
F_{15}	0.760 2**	F_{25}	0.581 7*	F_{35}	0.726 6**	F_{45}	0.554 2**

注：*P<0.05, **P<0.01 , ***P<0.001

表 3 颈部左右两侧对称通道 NRMS 波形的相关系数

Table 3 Correlation coefficients of NRMS waveforms on the left and right sides of the neckmuscles

通道序号	NRMS 波形相关系数	通道序号	NRMS 波形相关系数	通道序号	NRMS 波形相关系数	通道序号	NRMS 波形相关系数	通道序号	NRMS 波形相关系数
N_{11}	0.540 6*	N_{21}	0.790 1***	N_{31}	0.852 0***	N_{41}	0.816 7***	N_{51}	0.833 6***
N_{12}	0.464 2	N_{22}	0.598 6*	N_{32}	0.836 0***	N_{42}	0.838 2***	N_{52}	0.602 1*
N_{13}	0.484 0	N_{23}	0.609 2*	N_{33}	0.851 4***	N_{43}	0.820 8***	N_{53}	0.553 7*
N_{14}	0.847 2***	N_{24}	0.785 8***	N_{34}	0.912 1***	N_{44}	0.928 0***	N_{54}	0.837 1***
N_{15}	0.897 8***	N_{25}	0.938 1***	N_{35}	0.947 6***	N_{45}	0.939 1***	N_{55}	0.946 0***
N_{16}	0.874 0***	N_{26}	0.866 1***	N_{36}	0.977 8***	N_{46}	0.979 2***	N_{56}	0.968 8***
N_{17}	0.792 8***	N_{27}	0.900 5***	N_{37}	0.990 8***	N_{47}	0.966 0***	N_{57}	0.972 9***
N_{18}	0.906 1***	N_{28}	0.985 9***	N_{38}	0.996 4***	N_{48}	0.988 2***	N_{58}	0.988 5***

注：*P<0.05, **P<0.01 , ***P<0.001

低于 0.7，最低为 0.664 5(P9)；无声模式的分类精度同样为 1。面部右侧的平均分类精度和标准偏差为 0.752 8±0.108 8。由此可见，面部右侧的平均分类精度较左侧低，左右两侧的偏差范围均大于 0.1，波动非常大。比较表 4 和 5 也能看出，除了 P6 外，其余发音任务的分类精度都是面部左侧较高。

由于大部分受试者静息状态(P11)的分类精度几乎都达到了 1，故将它排除在外后，再对所有受试者的其余 10 个发音任务的分类精度进行统计，结果如图 4 所示。图中柱状图的高度代表受试者的 10 个发音任务分类精度的平均值，上下的垂直误差条表示标准偏差范围。蓝色柱状图表示受试者面部左侧 20 个通道肌电信号；红色柱状图表示受试者面部右侧 20 个通道肌电信号。从图 4 可以看出，使用面部左侧肌电信号

表 4　使用面部左侧肌电信号的 11 种语音分类精度

Table 4　Classification accuracies of 11 speaking tasks using the left side of facial SEMG signals

实际发音任务	识别发音任务										
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
P1	**0.914 1**	0.093 5	0	0	0.031 0	0	0.014 3	0.020 0	0.026 7	0	0
P2	0.018 2	**0.720 7**	0	0	0	0.051 0	0	0	0.076 2	0.052 8	0
P3	0	0	**0.902 4**	0.040 0	0.092 7	0	0.013 3	0.049 1	0	0	0
P4	0	0	0.022 2	**0.930 0**	0.015 4	0	0.032 5	0	0.026 7	0	0
P5	0.014 3	0.032 1	0.040 0	0	**0.787 4**	0.011 1	0.027 6	0	0.072 8	0	0
P6	0	0	0	0	0.029 7	**0.647 9**	0.027 6	0	0.016 7	0.118 5	0
P7	0.023 5	0.049 7	0.020 0	0	0	0.033 3	**0.803 2**	0.073 6	0.067 9	0.022 2	0
P8	0	0	0.015 4	0.016 7	0	0	0.081 5	**0.848 3**	0	0	0
P9	0.029 9	0.047 4	0	0.013 3	0.044 0	0.078 6	0	0.009 1	**0.713 1**	0.011 8	0
P10	0	0.056 7	0	0	0	0.178 1	0	0	0	**0.794 7**	0
P11	0	0	0	0	0	0	0	0	0	0	**1**

注：加粗的数据表示某一发音任务被正确分类的精度

表 5　使用面部右侧肌电信号的 11 种语音分类精度

Table 5　Classification accuracies of 11 speaking tasks using the right side of facial SEMG signals

实际发音任务	识别发音任务										
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
P1	**0.898 5**	0.159 0	0.022 2	0	0	0.011 1	0.024 3	0.013 3	0.059 7	0.062 1	0
P2	0.075 4	**0.694 9**	0	0.014 3	0.025 0	0.066 8	0.014 3	0.065 0	0.111 1	0.064 3	0
P3	0	0	**0.821 2**	0.112 3	0.081 3	0	0.020 0	0.093 3	0	0	0
P4	0	0.015 4	0.121 2	**0.738 6**	0.071 7	0.018 2	0.150 9	0.025 8	0	0	0
P5	0	0.040 4	0.010 0	0.053 0	**0.730 2**	0.054 5	0.042 2	0.063 3	0.074 3	0	0
P6	0.014 3	0.012 5	0	0	0.036 8	**0.688 6**	0.040 8	0	0.034 7	0.125 4	0
P7	0	0.012 5	0	0.067 5	0.024 3	0.053 3	**0.676 8**	0.046 7	0.018 2	0.013 3	0
P8	0	0	0.015 4	0	0.015 4	0	0.015 4	**0.692 5**	0	0	0
P9	0.011 8	0.025 0	0.010 0	0.014 3	0.015 4	0.036 4	0.015 4	0	**0.664 5**	0.059 7	0
P10	0	0.040 4	0	0	0	0.071 1	0	0	0.037 5	**0.675 3**	0
P11	0	0	0	0	0	0	0	0	0	0	**1**

注：加粗的数据表示某一发音任务被正确分类的精度

的平均分类精度中只有 sub6 超过 0.8；面部右侧肌电信号的平均分类精度中 sub4、sub7 都高于 0.8。sub7 的垂直误差条的长度最短，标准偏差不超过 0.1；其余的受试者波动范围都大于 0.1。此外，sub2、sub4、sub6、sub7、sub8 的左右平均分类精度高度差较大，均高于 0.5。使用 *t* 检验方法对面部两侧平均分类精度进行比较发现，sub2、sub4、sub6、sub7、sub8 的左右两侧间均存在显著性差异。

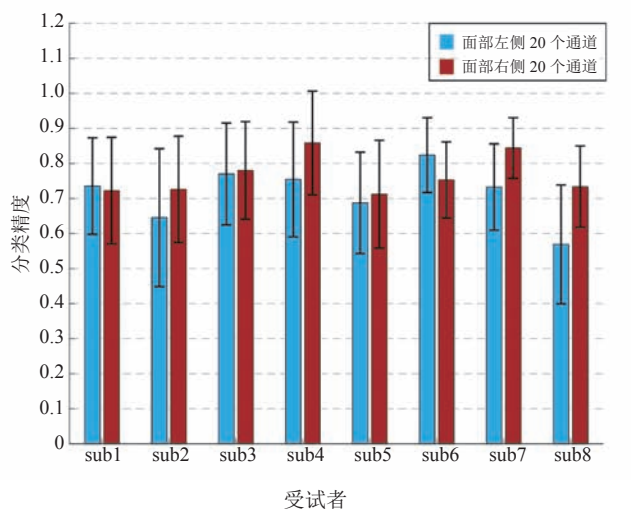


图 4 所有受试者使用面部左、右侧不同通道组合的平均分类精度与标准偏差

Fig. 4 Average classification accuracy and standard deviation for all subjects using different channels of facial muscles

3.3 使用颈部左右两侧对称位置肌电信号的语音分类精度对比

使用同一受试者颈部左、右两侧对称位置的肌电信号进行语音分类的结果如表 6 和 7 所示。与面部相同，静息状态的分类精度同样为 1，可见静息状态与发音时的肌电特征有着显著区别。从表 6 可以看出，使用颈部左侧肌电信号对 11 类发音任务进行分类时，所有的单词分类精度都超过 0.8，且 P3、P4 的分类精度超过 0.9；颈部左侧的平均分类精度和标准偏差值为 0.877 9±0.059 8。从表 7 可以看出，使用发音任务 P6、P9 和 P10 在颈部右侧处的肌电信号的分类精度较低，小于 0.8，而 P2、P4 的分类精度高于 0.9；颈部右侧的平均分类精度和标准偏差为 0.858 7±0.071 9。

与面部相似，使用受试者颈部左右两侧对称位置的高密度表面肌电信号的同一单词的分类精度并不完全相同。在 P1、P3、P6、P7、P9 和 P10 中，使用颈部左侧通道信号的分类精度高于使用颈部右侧通道的分类精度，其余单词则相

表 6 使用颈部左侧肌电信号的 11 种语音分类精度

Table 6 Classification accuracies of 11 speaking tasks using the left side of neck SEMG signals

实际发音任务	识别发音任务										
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
P1	**0.861 6**	0.078 9	0	0	0.026 8	0	0	0	0	0	0
P2	0.058 4	**0.824 4**	0.027 6	0	0	0.015 4	0	0	0.079 3	0.011 8	0
P3	0	0.016 7	**0.972 4**	0	0	0	0	0	0	0	0
P4	0.013 3	0	0	**0.908 9**	0.101 9	0.015 4	0.025 9	0.043 3	0	0	0
P5	0.033 3	0.026 7	0	0.059 3	**0.845 5**	0	0.036 4	0.023 8	0.011 8	0	0
P6	0	0	0	0	0	**0.825 8**	0.066 3	0	0	0.140 7	0
P7	0	0	0	0.022 2	0.013 3	0.041 0	**0.840 6**	0.074 6	0	0	0
P8	0.020 0	0	0	0.009 5	0.012 5	0	0.015 4	**0.857 8**	0.020 0	0	0
P9	0.013 3	0.053 3	0	0	0	0	0.015 4	0	**0.889 0**	0.016 7	0
P10	0	0	0	0	0	0.102 5	0	0	0	**0.830 9**	0
P11	0	0	0	0	0	0	0	0	0	0	**1**

注：加粗的数据表示某一发音任务被正确分类的精度

表 7　使用颈部右侧肌电信号的 11 种语音分类精度

Table 7　Classification accuracies of 11 speaking tasks using the right side of neck SEMG signals

实际发音任务	识别发音任务										
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
P1	**0.849 9**	0.027 9	0.038 9	0	0.014 3	0.032 1	0.013 3	0	0.065 8	0.013 3	0
P2	0.038 2	**0.922 1**	0	0.012 5	0.037 5	0	0.031 5	0.028 6	0.012 5	0.085 3	0
P3	0.015 4	0	**0.853 6**	0	0.055 4	0.015 4	0.012 5	0	0.017 4	0	0
P4	0	0.012 5	0.036 4	**0.940 4**	0.037 5	0	0.018 2	0	0.008 7	0.018 2	0
P5	0.025 4	0	0.033 3	0.035 3	**0.855 3**	0	0	0.031 0	0.018 2	0	0
P6	0	0	0	0	0	**0.799 5**	0.080 7	0	0.012 5	0.100 6	0
P7	0	0	0	0.011 8	0	0.061 5	**0.806 3**	0.072 9	0.061 2	0.026 7	0
P8	0	0	0	0	0	0	0	**0.867 6**	0.008 7	0	0
P9	0.053 0	0	0.037 7	0	0	0	0	0	**0.795 0**	0	0
P10	0.018 2	0.037 5	0	0	0	0.091 5	0.037 5	0	0	**0.756 0**	0
P11	0	0	0	0	0	0	0	0	0	0	**1**

注：加粗的数据表示某一发音任务被正确分类的精度

反。但颈部的分类精度整体高于面部，且颈部两侧分类精度的差异略小于面部。

对所有受试者颈部左、右两侧的肌电信号进行分类，得到的分类精度如图 5 所示。从图 5 可以看出，只有使用 sub8 颈部右侧的表面肌电信号进行分类时，平均分类精度低于 0.8；sub4 的两侧、sub5 的左侧以及 sub8 的右侧标准偏差均略高于 0.1，且只有 sub8 左右平均分类精度高度差大于 0.5，差异比较明显。

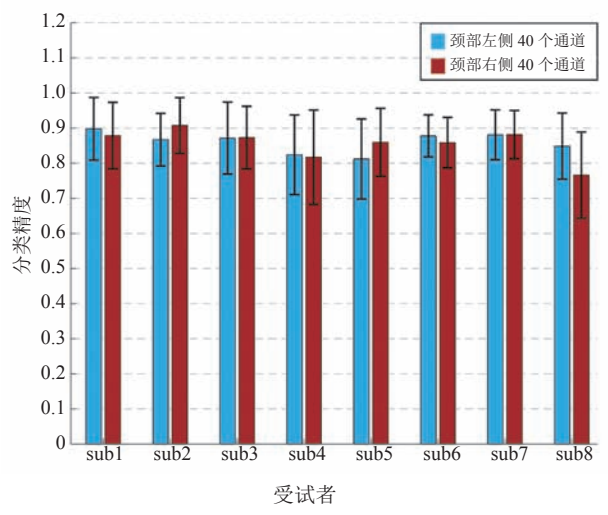


图 5　所有受试者使用颈部左、右侧不同通道组合的平均分类精度与标准偏差

Fig. 5　Average classification accuracy and standard deviation for all subjects using different channels of neck muscles

比较图 4 和 5 可以发现，颈部的平均分类精度比面部高，且标准偏差更低、差异更小，分类效果更稳定。使用 *t* 检验方法对颈部左右两侧平均分类精度进行比较发现，只有 sub8 左右两侧间存在显著性差异。

4　讨　论

语音的产生是一个面部和颈部肌肉共同运动的过程，而肌肉活动产生肌电信号[19]。因此，分析肌电信号对了解语音产生过程中肌肉活动的详细信息非常有帮助。前人已经使用 sEMG 技术在语音识别方面做了大量工作，但实验设置的电极数量较少，选取的电极位置依赖实验操作者的经验，分类结果也具有显著差异[6-11]，最高的分类精度是 Jong 和 Phukpattaranont[10]在 2019 年的研究中对健康受试者的泰语识别，为 0.945。由于面颈部肌肉结构复杂，少数几个电极不能完整覆盖发音肌电活动。为精准量化电极数量、确定电

极位置，本文利用高密度电极对发音相关的面、颈部左右两侧对称位置肌电在语音识别中的贡献进行了初步考察。

本研究使用关于面中部、颈中部对称的共 120 通道高密度表面肌电电极采集 8 名发音正常的受试者分别发 5 个中文单词和 5 个英文单词时的表面肌电信号。首先，对面部、颈部左右对称位置的 NRMS 波形进行分析比较发现，面部和颈部左右两侧的 NRMS 波形具有相同的变化特性，但面部的波形相关性比颈部低。这说明面颈部左右两侧肌肉发音的规律是相同的，但面部左右差异更大。这可能与颈部肌肉活动是被动的，而面部肌肉可以主观控制有关。然后，将不同通道肌电信号按照分布区域分为 4 组，提取 ZC、SSC、WL 和 MAV 四种特征值，并将其输入 SVM 分类器进行 11 种语音模式的分类。结果显示，所有通道电极的平均分类精度均可达 0.98，高于 Jong 和 Phukpattaranont[10]研究成果中最高的 0.945，表明高密度电极相较于少数凭经验放置的电极能提升分类精度。同时，使用同一受试者面部左右两侧通道（各 20 个）信号对相同单词的分类精度存在明显差异，而颈部两侧的差异则略小。所有受试者面部、颈部左右两侧不同通道组合的平均分类精度与标准偏差显示，颈部左右两侧的分类精度差异相较于面部对称位置是比较小的，表明颈部对称位置肌肉电活动对语音识别的一致性更高。因此，使用颈部对称位置的 sEMG 信号进行语音分类时的贡献具有一致性。

5 结 论

本研究提出使用面、颈部对称位置的高密度肌电信号对 11 种语音模式进行分类，以比较面、颈部对称位置肌电信号在语音识别时的贡献程度。结果表明，面、颈部左右两侧肌肉发音的规律是相同的，但面部左右两侧间差异更大。单独使用颈部左右两侧的肌电信号分类结果差异不大，但单独使用面部左右位置的肌电信号分类精度差异较明显。因此，颈部对称位置的 sEMG 信号对语音识别贡献程度具有一致性，而面部则不具有。该实验结果有助于减少记录电极的数量，为选择语音识别通道的最佳位置奠定了基础。

参 考 文 献

[1] Redenbaugh MA, Alan RR. Surface EMG and related measures in normal and vocally hyperfunctional speakers [J]. Journal of Speech and Hearing Disorders, 1989, 54 (1): 68-73.

[2] Anderson R, Wiryana F, Ariesta MC, et al. Sign language recognition application systems for deaf-mute people: a review based on input-process-output [J]. Procedia Computer Science, 2017, 116: 441-448.

[3] 戴立梅, 姚晓东, 王蓓, 等. EMG 在语音信号识别中的应用 [J]. 电子技术应用, 2005, 31(1): 5-7.

[4] 许佳佳, 姚晓东. 基于 EMG 信号的无声语音识别应用及实现 [J]. 计算机与数字工程, 2006, 34(5): 133-136.

[5] Meltzner GS, Heaton JT, Deng Y, et al. Development of sEMG sensors and algorithms for silent speech recognition [J]. Journal of Neural Engineering, 2018, 15(4): 046031.

[6] Sugie N, Tsunoda K. A speech prosthesis employing a speech synthesizer-vowel discrimination from perioral muscle activities and vowel production [J]. IEEE Transactions on Biomedical Engineering, 1985, 7: 485-490.

[7] Morse MS, Day SH, Trull B, et al. Use of myoelectric signals to recognize speech [C] // Proceedings of the Annual International Engineering in Medicine and Biology Society, 1989: 1793-1794.

[8] Srisuwan N, Phukpattaranont P, Limsakul C. Comparison of feature evaluation criteria for speech recognition based on electromyography [J]. Medical & Biological Engineering & Computing, 2018, 56(6): 1041-1051.

[9] Janke M, Wand M, Nakamura K, et al. Further

investigations on EMG-to-speech conversion [C] // 2012 IEEE International Conference on Acoustics, Speech and Signal Processing, 2012: 365-368.

[10] Jong NS, Phukpattaranont P. A speech recognition system based on electromyography for the rehabilitation of dysarthric patients: a Thai syllable study [J]. Biocybernetics and Biomedical Engineering, 2019, 39(1): 234-245.

[11] Diener L, Janke M, Schultz T. Direct conversion from facial myoelectric signals to speech using deep neural networks [C] // 2015 International Joint Conference on Neural Networks, 2015: 1-7.

[12] Wand M, Schultz T. Session-independent EMG-based speech recognition [C] // Proceedings of the International Conference on Bio-inspired Systems and Signal Processing, 2011: 295-300.

[13] Chan ADC, Englehart K, Hudgins B, et al. Hidden markov model classification of myoelectric signalsin speech [J]. IEEE Engineering in Medicine and Biology Magazine, 2002, 21(5): 143-146.

[14] Janke M, Diener L. EMG-to-speech: direct generation of speech from facial electromyographic signals [J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2017, 25(12): 2375-2385.

[15] Balata PMM, da Silva HJ, de Araújo Pernambuco L, et al. Normalization patterns of the surface electromyographic signal in the phonation evaluation [J]. Journal of Voice, 2015, 29(1): 129.e1-129.e8.

[16] Phinyomark A, Phukpattaranont P, Limsakul C. Feature reduction and selection for EMG signal classification [J]. Expert Systems with Applications, 2012, 39(8): 7420-7431.

[17] Vishwanathan SVM, Murty MN. SSVM: a simple SVM algorithm [C] // Proceedings of the 2002 International Joint Conference on Neural Networks, 2002: 2393-2398.

[18] Focquet A, Péréon Y, Ségura S, et al. Diagnostic and prognostic contribution of laryngeal electromyography in unilateral vocal-fold immobility in adults [J]. European Annals of Otorhinolaryngology, Head and Neck Diseases, 2017, 134(1): 13-18.

[19] Bu N, Tsuji T, Arita J, et al. Phoneme classification for speech synthesiser using differential EMG signals between muscles [C] // The 27th Annual Conference of IEEE Engineering in Medicine and Biology Society, 2005: 5962-5966.