Advances in Applied Mathematics 应用数学进展, 2021, 10(5), 1455-1464
Published Online May 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.105155

Hans汉斯

键盘字母排序的频率理论

张冰冰，吴媛梦，王宇辰，李　洁，杨　澜

河南科技大学数学与统计学院，河南　洛阳
Email: 18336898502@163.com

收稿日期：2021年4月17日；录用日期：2021年5月2日；发布日期：2021年5月19日

摘　要

当今网络信息时代，人们通过电脑、手机、触摸屏等多种设备与网络世界相联。键盘作为人机交互的重要工具，在我们的工作、学习和生活中发挥着极大作用。普遍使用的键盘是传统的QWERTY键盘，依赖英文输入字母的使用频率情况设计构建。这种键盘是否适合中国人的使用，取决于英文输入字母频率与中文拼音输入的字母频率是否存在显著差别。本文研究中文全拼录入中字母的使用频率情况，使用抽样调查方式搜集具有代表性的中文材料，采用统计学中文本挖掘的算法，在Python中实现相关算法。结果显示，中文全拼录入中字母的使用频率与英文的字母频率存在显著性差异。

关键词

键盘，抽样，字母频率

Frequency Theory of Keyboard Alphabetic Sorting

Bingbing Zhang, Yuanmeng Wu, Yuchen Wang, Jie Li, Lan Yang

School of Mathematics and Statistics, Henan University of Science and Technology, Luoyang Henan
Email: 18336898502@163.com

Received: Apr. 17th, 2021; accepted: May 2nd, 2021; published: May 19th, 2021

Abstract

In the network information age, people connect with the network world through computers, mobile phones, touch screens and other devices. As an important tool of human-computer interaction, keyboard plays an important role in our work, study and life. The most commonly used keyboard

文章引用: 张冰冰, 吴媛梦, 王宇辰, 李洁, 杨澜. 键盘字母排序的频率理论[J]. 应用数学进展, 2021, 10(5): 1455-1464.
DOI: 10.12677/aam.2021.105155

is the traditional QWERTY keyboard, which relies on the frequency of input letters in English. Whether the keyboard is suitable for Chinese use depends on whether there is a significant difference in the frequency of letters between English and Chinese Pinyin. In this paper, we study the frequency of the letters in the Chinese complete spelling input, collect the representative Chinese materials by sampling survey, and implement the relevant algorithms in Python by using the text mining algorithm in statistics. The results show that there is a significant difference between the frequency of Chinese characters and English characters.

Keywords

Keyboard, Sampling, Letter Frequency

1. 引言

键盘在我们生活中无处不在：电脑键盘、手机输入法键盘、超市收银一体机键盘等，完成工作和学习中的任务或者进行生活中的线上聊天交际等文字输入性活动，对我们的工作、学习和生活极其重要，离开它将改变我们的生活。我们现在使用的键盘是由克里斯托夫·拉森·肖尔斯(Christopher Latham Sholes)发明的 QWERTY 键盘[1]，来自于打字机，其字母排序最初是按照字母顺序 ABC 排列的。而打字机是全机械结构的打字工具，如果打字速度过快，某些键的组合很容易出现卡键问题，为了解决这个问题，克里斯托夫·拉森·肖尔斯将最常用的几个字母安置在相反方向，最大限度增大重复敲键的时间间隔，因此避免了卡键。之后有很多对键盘字母布局的研究，如来自美国华盛顿的德沃拉克(DVORAK)为了使左右手能交替击打更多的单词发明了 DVORAK 键位布局的键盘字母排序[2]。随着科技水平的提高，打字机卡键问题已不复存在，因此，需要从效率角度重新审视键盘的字母排序问题。此外，键盘字母排序的设计主要是依据英文的字母使用规律设计的，中文拼音的字母规律和英文的字母规律是否有着显著的差异，如果存在显著性差异，那么现有排序对于拼音输入速率有着较大的阻碍作用，研究出一个更适宜拼音输入的键盘字母排序是迫切需要的。

为了设计一个高效的键盘，首先考虑字母的使用频率。字母是拼音的重要组成部分，字母频率[3]就像词频，不同作者或写作主题的作品中往往各不相同，研究字母的使用频率一方面可以探索某个作者的写作风格以及语言规律：字母、双字母组、三字母组、单词频率、单词长度和句子长度，这些都可以经统计后用以证明或反驳某一作品是某作者所写，甚至可以鉴别一个作品与一位作者的写作风格是否相近。另一方面，字母频率在键盘布局的设计上也显现出极为重要的作用：字母使用频率是设计键盘字母排序的重要参考，常用的字母放在手指不易接触的位置这显然会影响打字速率。字母使用频率是设计键盘字母排序的基础，在获得字母使用频率后结合手指作业能力等其他方面的因素进行键盘字母排序的设计[4]。

本文主要利用抽样调查的相关知识对新闻、网络小说和网络推文进行抽样调查，从而获得拼音字母使用频率，将此字母频率与前人统计的英文字母使用频率进行对比分析，让人们发觉现有键盘字母布局的不合理性，通过数据探索更合理的字母布局，从而提高工作效率；还可以为键盘设计的有关部门提供参考，极具参考价值和现实意义。

2. 研究方法

2.1. 抽样方法获取数据

2.1.1. 抽样原因

为了研究键盘上字母的排序问题，我们需要获取字母的使用频率，因此需要获得大量的网上使用文字。因为文字数量极大，不可能进行全面的调查，所以我们就采用抽样来获得有关网上使用文字的了解。抽样调查能够节约调查的人力、物力和财力，而且有利于提高调查数据的质量。

2.1.2. 抽样思路

本次抽样准备采用分层抽样法。分层抽样法，也叫类型抽样法，就是将总体单位按其属性特征分成若干类型或层，然后在类型或层中随机抽取样本单位。这种方法的优点是一方面由于通过划类分层，减小了各抽样层变异性的影响，增大了各类型中单位间的共同性，容易抽出具有代表性的样本，抽样误差比较小[5]。另一方面该方法适用于总体情况复杂，各单位之间差异较大，单位较多的情况。缺点是抽样手续较简单随机抽样还要繁杂些。分层抽样的目的是：把总体各单位分成两个或两个以上的相互独立的完全的组，从两个或两个以上的组中进行抽样，样本相互独立[6]。总体各单位按主要标志加以分组，分组的标志与关心的总体特征相关。因为我们要对网络上使用的文字进行调查，初步判别，文字使用方面新闻的使用文字与网络推文和小说的不同，那么用途和使用范围应是划分层次的适当标准。所以我们可以将总体分为小说、新闻报刊以及网络推文这三个类别。

针对小说类别，我们采用二阶段抽样法。二阶段抽样也称二级随机抽样，就是在抽取样本时分两个阶段来进行，根据网站分类，将小说总体共分为 49 个类别，由于不同类别小说内容有偏差，使用频率高的文字也不尽相同，因此第一阶段是从总体 49 个小说类别中利用 matlab 随机随机生成 10 个随机数的方法抽取 10 个类别，称为初级单位。然后在第二阶段从这些初级单位中又随机抽取若干个小说样本单位，称为基本单位或最终单位，最后根据所抽的基本单位组成共 50 万字的样本进行调查，用取得的样本资料来推断总体。二阶段抽样一方面保持了整群抽样的样本比较集中，便于调查，节省费用等优点，另一方面因为小说文本总体数目过大，所以采用二阶段抽样又避免了对群内单元的过多调查造成的浪费，充分发挥了抽样调查的优点。该抽样目的是在繁杂的小说文本中抽取有代表性的样本进行调查。

针对新闻报刊这一类别，我们采用构造周抽样法。构造周抽样法是在总体中从不同的星期里随机抽取周一到周日的样本，并把这些样本构成“一个周”(即构造周)。构造周抽样法基于报纸的内容结构，在以星期为单位周期变化的前提下，通过在不同星期中抽取星期一至星期日的样本来反映总体，它避免了简单随机抽样中产生的“周期性偏差”，同时考虑了时间因素。有关的研究成果表明，一年抽取 2 个构造周的样本便能可靠地反映总体。我们首先要确定具有代表性的抽样来源，然后再进行抽样。人民网是世界十大报纸之一《人民日报》建设的以新闻为主的大型网上信息交互平台，也是国际互联网上最大的综合性网络媒体之一，是国家重点新闻网站的排头兵，其影响力和代表性很强，因此我们可以从人民日报－人民网中采集文本。人民日报非周末版的版面结构基本是固定的，而周末版无论在内容上还是在信息量上都与非周末版不同。若简单随机抽样，就会出现周末版的样本偏多或偏少的情况，产生“周期性偏差”，使样本不能有效地反映总体而失去代表性。因此，我们可以采用构造周抽样法从人民日报中进行抽样。

针对网络推文类别，因为网络在生活中占比日益增多，所以在对网络推文部分的数据采集中，根据经验以及资料选取网民所使用最多的 APP 进行文字采样，使所采集到的文本都具有广泛性与代表性。然后采用简单随机抽样法对 APP 中的文字进行选取。简单随机抽样是从含有 N 个个体的总体中逐个抽取且

每个个体被抽到的概率相等的一种抽样方法。由于网络瞬息万变，文字变化性极强，因此采用简单随机抽样具有一定的随机性，所采样的数据也有一定的代表性。但又由于不同时段网络信息的不同，为保证样本随机性，我们将定时进行简单随机抽样。将每日的 24 小时分成 24 个时段，利用 matlab 从 1~24 中随机生成两个随机数作为被抽取的两个时段。又考虑到工作日与周末网民浏览内容的不一致性，因此一周七天均进行两次随机抽样。

2.2. 数据的处理

如图 1 所示，整个数据处理过程概述的流程大致经过了转化、统计、作图三个步骤实现。我们需要得到字母的频数来看出，哪些字母是最常用的，以及他们的使用情况。同时我们还需要计算出相应的频率来看看，对于不同种类的汉字使用领域而言，频率变化情况是否相同。在我们后续的示例验证中，通过频率折线图的绘制，可以很明显的看出，在不同的领域中，字母的使用频率变化情况基本一致。这与我们所提出的假设是基本相符的。同时，通过数据处理后观察到的频数，我们可以很直观的了解到有些字母在汉字拼音中使用的重要性。同时我以为键盘的设计提供充分的证据证明这些字母位置摆放所需要考虑的重要性。

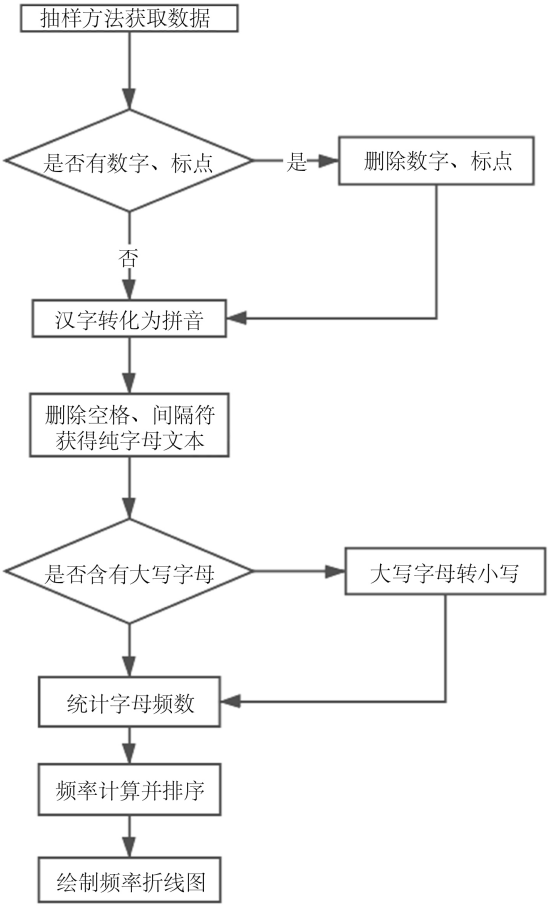

Figure 1. Data processing flowchart
图 1. 数据处理流程图

在转化过程中，对于所采样的数据将其转化为拼音后，对其频数进行统计。本文主要利用 python 语

句，编写程序来完成频数的统计，先通过来完成汉字到拼音的转化，然后对英文字母的频率进行概率统计。在转化程序中，引用了 xpinyin.py 包来完成汉字到拼音的转化，转化过程中需要注意对多音字的处理以及提前需要将文本中的标点数字删除。编者自行完成的程序虽然能够完成转化，但通过引用 xpinyin.py 模块，可以使程序更加完备，能够完成对多音字等等各方面的需求。

统计程序较为简单，可以将文本转化为字符串进行去重统计。但由于所统计的文本量较大，因而可以通过将文本转化为字符数组的方式来完成频数的合计。需要提前载入 re 模块以及 collections 模块来完成该程序。

数据处理以及作图用 excel 处理较为简单，我们先通过 excel 中的 match 函数完成了对不同字母频数的索引，然后以所有种类文章总字母频数由高到低顺序为基准排列顺序，并将总字母频数、小说字母频数、新闻字母频数、网文字母频数四类字母频数按照相同顺序放在一起进行比较，并通过公式编辑计算出频率，将四类的频率分布折线图绘制出来放在一起进行比较。从而较为直观的看出四类频率的差别与变化情况。

频数又称“次数”，指变量值中代表某种特征的数(标志值)出现的次数。按分组依次排列的频数构成频数数列，用来说明各组标志值对全体标志值所起作用的强度。对于本文频数即为所统计的样本中不同字母所出现的次数，将字母频数大小按照从高到低排序即可统计出中文中字母使用频率的大小，频数越大就说明该字母在键盘输入时的使用频率就越高，最终可由样本频数的大小推断中文中字母频数使用的大小。

3. 实例证明

3.1. 抽样步骤

3.1.1. 现代小说

小说共有 49 个类别，从中随机选取 10 个类别，每个类别下抽取 5 万字，总共抽取约 50 万字抽选文字作为调查对象，具体抽选方法如下：

(1) 将每个类别进行标号 1~49；

(2) 用 MATLAB 从 1~49 从中产生 10 个随机数，分别为：40，7，21，25，45，39，48，33，2，42，46；

(3) 在随机数对应的类别下，分别随机抽取 5 万字。

抽取结果见下表 1：

Table 1. Selection of novel categories
表 1. 小说类别选择

序号	类别	是否选中	序号	类别	是否选中
1	现代言情	0	26	影视小说	0
2	都市	1	27	历史	0
3	豪门总裁	0	28	推理	0
4	社会生活	0	29	悬疑	0
5	校园	0	30	盗墓	0
6	玄幻	0	31	甜宠	0
7	名著经典	1	32	古代言情	0

Continued

8	武侠	0	33	中国名著	1
9	灵异	0	34	兵王	0
10	精进提升	0	35	腹黑	0
11	玄幻言情	0	36	战争	0
12	成功励志	0	37	女强	0
13	重生	0	38	学霸	0
14	游戏动漫	0	39	娱乐圈	1
15	生活	0	40	经济管理	1
16	外国名著	0	41	社会生活	0
17	美食	0	42	科幻	1
18	天才	0	43	洪荒	0
19	科幻	0	44	宠物	0
20	穿越	0	45	二次元	1
21	诗歌散文	1	46	历史文化	1
22	直播	0	47	王妃	0
23	皇后	0	48	大唐	1
24	精灵	0	49	社会科学	0
25	快穿	1			

3.1.2. 新闻报刊

我们采用构造周抽样法从人民日报中进行抽样。

具体抽样步骤如下：

(1) 在2019年上、下半年中各抽取一周作为样本。因考虑到报纸内容以星期为单位周期性变化，将总体按时间分段，在不同时间段里进行抽样。

(2) 2019年上半年有26周，将前2周分给星期日，将后24周均分给星期一至星期六。从前两周随机抽取星期日的样本，随后每4个星期随机抽取星期一至星期六的样本，构成一周。最后选取1月6号(星期日)、1月28号(星期一)、2月19号(星期二)、3月13号(星期三)、4月25号(星期四)、5月31号(星期五)、6月22号(星期六)为一个构造周。

(3) 2019年下半年有26周，同理，依例取7月7号(星期日)、7月22号(星期一)、8月20号(星期二)、9月25号(星期三)、10月24号(星期四)、11月22号(星期五)、12月28号(星期六)为第二个构造周。

3.1.3. 网络推文

根据搜狐网的《中国人最常用的150个APP排行榜》一文中，筛选出中国居民最常用的具有代表意义的四个APP——微信、微博、知乎、豆瓣。其中，微信订阅号的推文，微博的热搜评论及文章，知乎的回帖及文章、豆瓣的影评及文章等都具有广泛性与代表性。具体抽样步骤如下：

(1) 从微信“看一看”的精选热门文章中抽取微信推文，从微博的实时热搜以及其评论、相关文章选取微博文字，从知乎的“热榜”中的各大类选取热门帖进行选取知乎文字，从豆瓣的“热点文章”以及知名电视剧、电影的评论文章中选取豆瓣文字。

(2) 在时间方面，考虑到工作日与休息日人们进行网络活动的不一致性，其中周一到周五代表工作日，周末则为休息日，故在周一至周日均进行文字选取。又考虑到不同时间段人们的阅读不同，利用 matlab 软件从 1~24 中随机生成两个随机数作为每日抽取的两个不同时间点进行文字选取，具体时间见下表 2：

Table 2. Article extraction time
表 2. 文章抽取时间

抽取时间 / 日期	第一次	第二次
周一	20 点	14 点
周二	22 点	23 点
周三	4 点	24 点
周四	22 点	4 点
周五	16 点	24 点
周六	3 点	23 点
周日	7 点	12 点

3.2. 结果分析

经过抽样，抽取了现代小说约 50 万个字，新闻报刊约 130 万字，网络推文抽取约 120 万字，利用第三方软件将文本资料转化为拼音形式，并利用自主设计的 Python 程序统计出 26 个字母的使用频数，计算得出每个字母的使用频率，按出现频率从高到低排序，结果如表 3 和图 2 所示：

在现代小说和新闻中前三名依次是 i，a，n，出现次数最少是 v。在网络推文中，虽然 i，a，n 排序发生变化，但是依然排在前三并且 v 出现的次数同样最少。

经过资料查询得到(《汉语拼音方案》) [7]：在汉字拼音中 i 和 a 是 ji、yi、za、ha、zhi、cha 等的单元音，同时还做介音和韵尾，n 直接构成所有鼻音韵尾，所以 i，a，n 的出现的频次高。而 v 是专门拼写外来语、少数民族语言和方言的，这些词或音出现的次数较少故 v 的频次较少。由此可以说明实例抽样结果具有较高的可信度。

Table 3. Letter frequency table
表 3. 字母频数表

字母	全部样本字母频数表		新闻字母使用频数表		网络推文字母频数表		现代小说字母频数表	
I	1081496	0.1396	452193	0.1379	444924	0.1417	184379	0.1391
N	922315	0.1191	420372	0.1282	357405	0.1138	144538	0.1091
A	814722	0.1052	331768	0.1011	331613	0.1056	151341	0.1142
U	614827	0.0794	260911	0.0795	251565	0.0801	102351	0.0772
H	555644	0.0717	242490	0.0739	219835	0.0700	93319	0.0704
E	539486	0.0697	212128	0.0647	232395	0.0740	94963	0.0717
G	534063	0.0690	258196	0.0787	198172	0.0631	77695	0.0586
O	455780	0.0588	183200	0.0558	185502	0.0591	87078	0.0657
D	254133	0.0328	97419	0.0297	111124	0.0354	45590	0.0344

Continued

Z	252216	0.0326	114276	0.0348	95835	0.0305	42105	0.0318
S	229857	0.0297	88784	0.0271	97713	0.0311	43360	0.0327
Y	226084	0.0292	90620	0.0276	95303	0.0304	40161	0.0303
J	187253	0.0242	92195	0.0281	69503	0.0221	25555	0.0193
X	149034	0.0192	62670	0.0191	60839	0.0194	25525	0.0193
L	138444	0.0179	53217	0.0162	57221	0.0182	28006	0.0211
B	110234	0.0142	40443	0.0123	48327	0.0154	21464	0.0162
C	108650	0.0140	51603	0.0157	40793	0.0130	16254	0.0123
T	89635	0.0116	36189	0.0110	35400	0.0113	18046	0.0136
W	89467	0.0116	35957	0.0110	36713	0.0117	16797	0.0127
M	81786	0.0106	29118	0.0089	35724	0.0114	16944	0.0128
Q	79849	0.0103	34538	0.0105	32692	0.0104	12619	0.0095
R	71510	0.0092	26746	0.0082	31667	0.0101	13097	0.0099
F	68634	0.0089	32460	0.0099	27210	0.0087	8964	0.0068
K	50527	0.0065	17071	0.0052	24439	0.0078	9017	0.0068
P	32248	0.0042	13047	0.0040	14885	0.0047	4316	0.0033
V	7517	0.0010	2624	0.0008	3019	0.0010	1874	0.0014

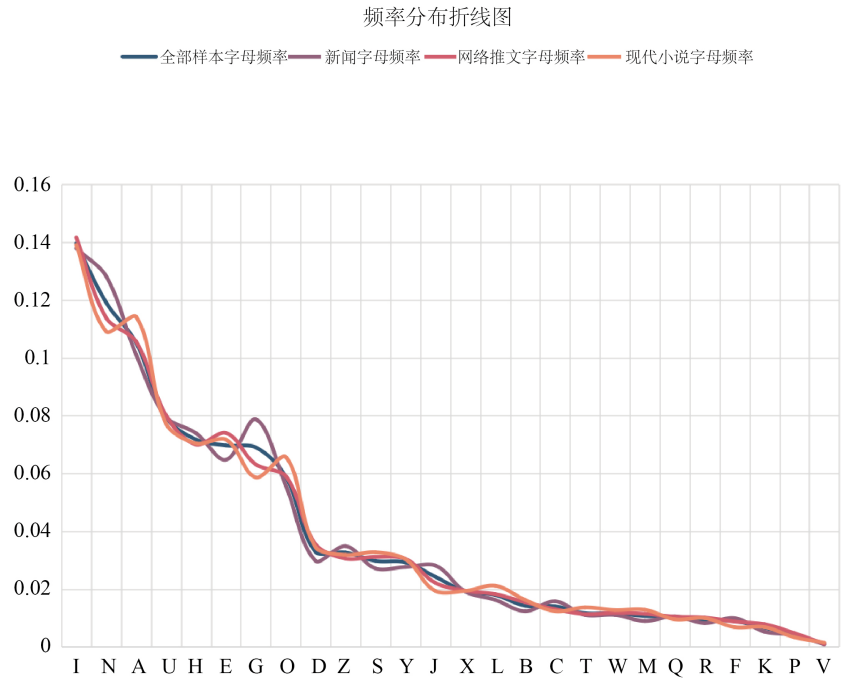

Figure 2. Line graph of frequency distribution
图 2. 频率分布折线图

4. 结论

通过上述分析可知，中文字母的使用频率不同于英文的使用频率。中文中使用频率最高的五个字母为：i、n、a、u、h (见表 3)；英语中最常用的前五个字母是 e、t、a、o、I (见表 4)。这种差异说明中文和英文在字母使用上存在较大差异。但是，e、a、o 三个元音音标的使用频率无论在中文还是英文中的使用频率都很高。

Table 4. Frequency of use of English letters
表 4. 英文字母的使用频率

字母	使用频率	字母	使用频率	字母	使用频率
E	0.1268	L	0.0394	P	0.0186
T	0.0978	D	0.0389	B	0.0156
A	0.0788	U	0.0280	V	0.0102
O	0.0776	C	0.0268	K	0.0060
I	0.0707	F	0.0256	X	0.0016
N	0.0706	M	0.0244	J	0.0010
S	0.0634	W	0.0214	Q	0.0009
R	0.0594	Y	0.0202	Z	0.0006
H	0.0573	G	0.0187		

这种差异可能在于现代汉语拼音的设计，i、n、a 构成了拼音的韵母成分，出现在大部分的汉字中，成为使用频率最高的字母。另一种可能是由于中文的改变，可能是字词的改变，也可能是发音的改变，如从单音节变为双音节。

我们认为形成以下差异可能有以下几个原因

(1) 现代汉语拼音的设计方式常用这些字母

首先，可以观察汉语拼音字母表的构成，包括单韵母 a、o、e、i、u、ü，其中这六个字母除ü外，都为调查结果中排序较为考前的字母，观察前鼻韵母、后鼻韵母表不难看出，这些韵母都是由单韵母之间或与 n 结合构成，所以可以认为出现这样的字母常用度顺序与汉语拼音的设置模式息息相关，同样的例子还比如字母 h，他除了本身作为声母 h，同时还参与构成了整体认读音节 zh、ch、sh，从而提高了字母利用率[8]。

(2) i、a、u、o 等字母音节简单，易于发音，常用于口语

通过观察常用字母 i、a、u、o 的发音方式，如 a (“啊”)，我们在发此音时，只需“嘴唇自然张大，舌放平，舌头自然放置，声带颤动”可以说只需要靠嘴唇声带结合发生，发音难度低，但例如一频率较低音节“z”，则需要“舌尖抵住上门齿背，阻碍气流，让较弱的气流冲开舌尖阻碍，从窄缝中挤出，摩擦成音。”，同时需要牙齿、舌头、最初、声带多处配合完成发音，发音难度相对难度有所提高，倒排结论的顺序，不难发现出现频率较低的字母，大多数发音都较为复杂，需要多个口部器官配合[9]。

(3) 汉字读音发展史影响

在中国汉字读音发展史过程中，原本是没有拼音的，最初采用的是直音和切音的方式进行汉字认读，直音即为直接引用它字读音的方式，如“跟”同“根”、“畔”同“叛”等，这种音同字不同的发音命名方式直接提高了本身就较为常用音节的出现频率。同理，切音的读音命名方式，其实也间接提高了常

用音节的出现频率[10]。例如“冬，都宗切”，因为被命名字的读音其实也是源于原始的两个字。因而，汉字古时的读音方式就决定了未来字母的出现频率会使部分字母使用率偏高，而部分字母很少出现。

参考文献

[1] 为何电脑键盘按键按“QWE”顺序排列? [J]. 黑龙江科技信息, 2013(4): 316-317.

[2] 蝌蚪五线谱. 电脑键盘排序充满偶然性[J]. 科学大观园, 2013(3): 4-5.

[3] 百度百科. 字母频率[EB/OL]. https://baike.baidu.com/item/%E5%AD%97%E6%AF%8D%E9%A2%91%E7%8E%87/9669044?fr=aladdin, 2015-7-7.

[4] 魏欣, 骆玉梅, 吴文凤, 廖斌. 基于汉字拼音输入法和手指作业能力的柯蒂键盘布局评价和改善研究[J]. 人类工效学, 2020, 26(3): 64-66+73.

[5] 金勇进, 杜子芳, 蒋妍. 抽样技术(第四版) [M]. 北京: 中国人民大学出版社, 2015.

[6] 茆诗松, 程依明, 濮晓龙. 概率论与数理统计教程(第二版) [M]. 北京: 高等教育出版社, 2011.

[7] 中华人民共和国教育部. 汉语拼音方案[Z]. 1958-02-01.

[8] 许诺. 汉语声母韵母习得问题研究[J]. 安徽冶金科技职业学院学报, 2013, 23(3): 57-59.

[9] 罗佳. 浅析汉语普通话韵母的发音和训练[J]. 现代语文(学术综合), 2011(4): 140-144.

[10] 陈歌. 基于汉字发展史的汉字简化方法研究[D]: [硕士学位论文]. 曲阜: 曲阜师范大学, 2013.