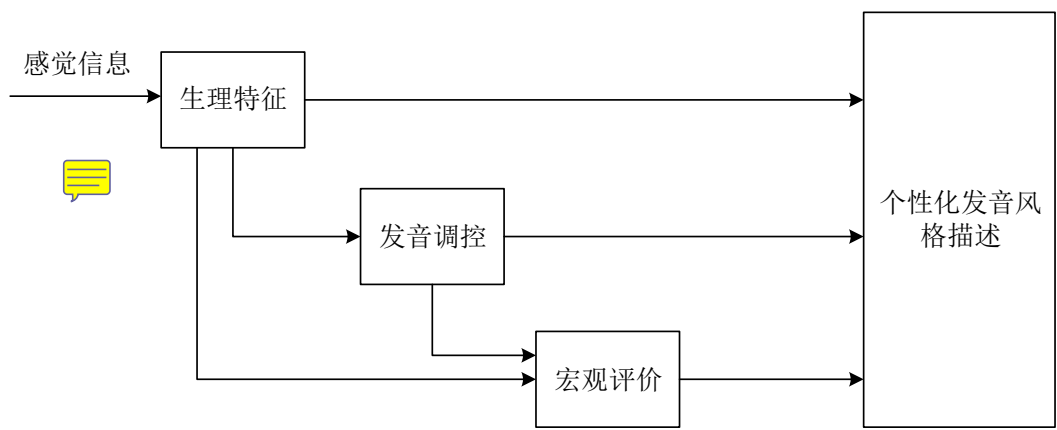


3 个性化发音风格描述体系

人从感觉信息到产生具有丰富情感以及个性风格的发音的因素来自于生理固定特征和认知感受等多个层面。但发音风格总体来讲更多来源于听者对于说话人的发音评价，故应加入听者的感受。因此在生理、心理多方面的发音产生过程的基础上还应加入一个宏观的听觉感受描述特征。图 3-1 给出从感觉信息输入到生成风格化发音的过程及其内部关系。描述体系是由生理特性、发音调控和宏观评价三种成分构成，分别从不同视角刻画发音风格某方面的变化。每个模块都会直接或间接作用于发音风格，前面模块也会影响后面模块，且影响会累积向后传递，最终三种成分相辅相成共同构成个性化发音风格的多视角描述体系。



各视角内容的具体表示采用离散类别表示和层级表示相结合的方式，每个视角形成一个超平面，而超平面的集合则支撑起一个多层结构空间，用于描述发音风格的复杂关系。由于发音风格更侧重于听觉感受，所以本文是先进进行宏观评价的调查，再逐步细化。所以下面将从外层到内层开始介绍各视角内容的具体描述方案。

3.1 宏观评价

宏观评价是站在听者的角度，发音人的发音风格使听者得到的听觉体验。由于是说话发音体验，所以本文列举出了 122 种说话的风格场景以及相关修饰词（基本涵盖生活中常见的所有说话场景）。122 种说话方式能给听者带来多种听觉体验。通过对各种听觉体验的归类总结，列举出了友善、温柔、高冷、冷漠、

冷艳、清爽、做作、单纯、活力、调皮、沧桑、低缓、奸诈、天真、绵软、尖细、弱小、刚强、激情、有力、正义、色气、滑稽、戏谑、成熟、轻佻、庄重、严肃、威严、木楞、高贵、高傲、富贵、可爱、神圣、孱弱、沉稳。这三十七个形容词均是形容发音人发音风格的词语，并且每种发音风格给听者产生的听觉感受均有所不同。由于以上维度会有所交叠，为了探究这些形容维度间的相互关系，进行了以下心理学问卷实验。

3.1.1 问卷设计

我们下载了 104 段不同角色的音频。每个角色都用各自的发音风格念同样的台词。

由于 104 段音频过多，为防止测试者的疲劳我们对 104 段音频进行了分析筛选。经过一个初步的打分测定我们从中选取了 15 个具有代表性的音频来设计问卷。本次调查问卷采用 APP 在线调查，较为方便地解决了音频文件嵌入的问题。问卷设计如图 3-1 与图 3-2 所示。



图 3-1 问卷首页

图 3-2 问卷内容设计

可见，我们在问卷首页采用允许匿名的方式在统计测试者性别的同时使其能够进入系统。在内容设计上，每个页面对应一个音频文件，对于每个选项我们采用连续型数据进行测评。利用星型填充条控件，测试者只要将填充条拉至自己认为合适的地方即可。被试为中北大学安卓实验室的同学，参与人数为 36 人，男女比例是 20:16，以下实验均相同，不再赘述。

3.1.2 相关性分析

通过问卷调查，获取到的有效回答为 509 条（36 人每人 15 个问题，其中有 4 人未回答完所有问题，有 7 条回答为空白）。每条回答均是受试者在听了相应音频后对 37 个维度进行打分，分数越低该维度的表达越低，分数越高该音频对该维度的风格表达得越充分，分数均为浮点数。通过对 509 条回答进行求和、取平均以及相关性分析，得到这几个维度间的相关性分析结果，如图 3-3 所示。

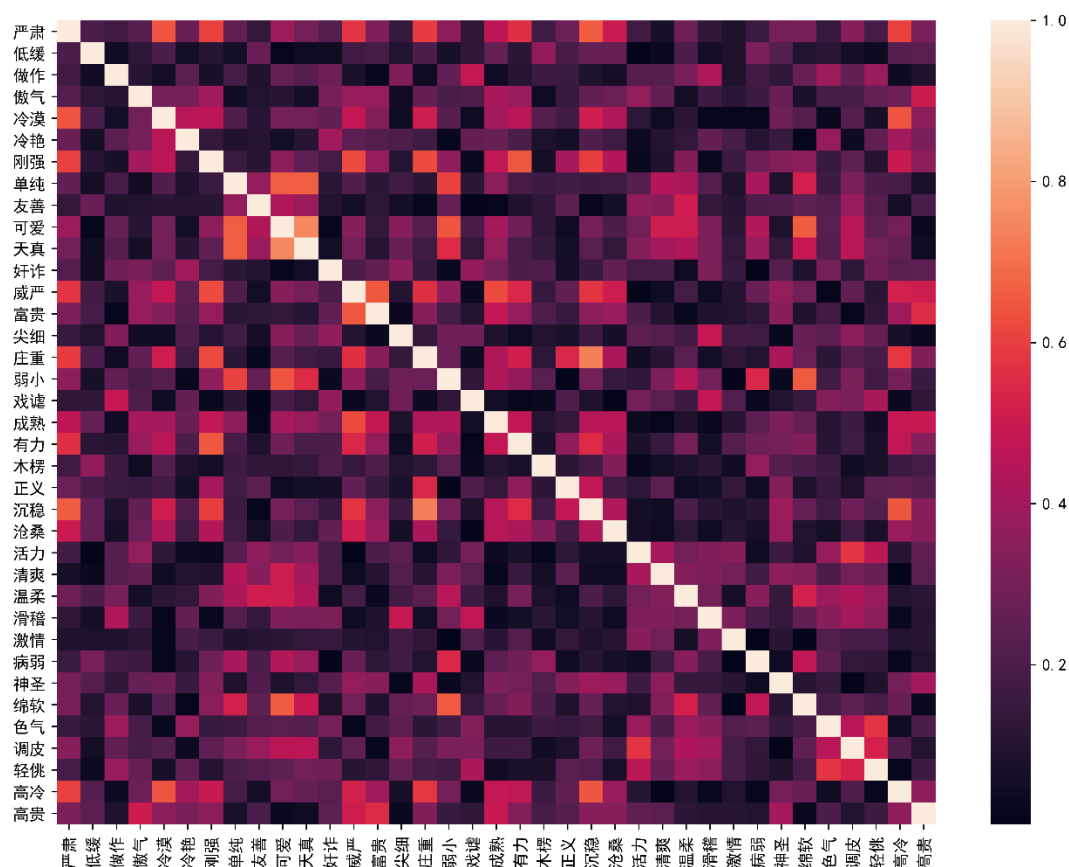


图 3-3 各维度相关性分析图

由于我们是探究维度之间的相关性，为图示简单易懂，这里对所有计算得到的相关系数进行了绝对值处理，以便区分高低相关度区域。图中颜色越深代表相关系数越接近 0，也就是相关性越低，颜色越浅代表相关系数越大，也就是相关性越大。一般来讲相关系数大于 0.5 则为相关性较强的值，图 3-4 对大于 0.5 的区域进行了可视化（白色区域）。

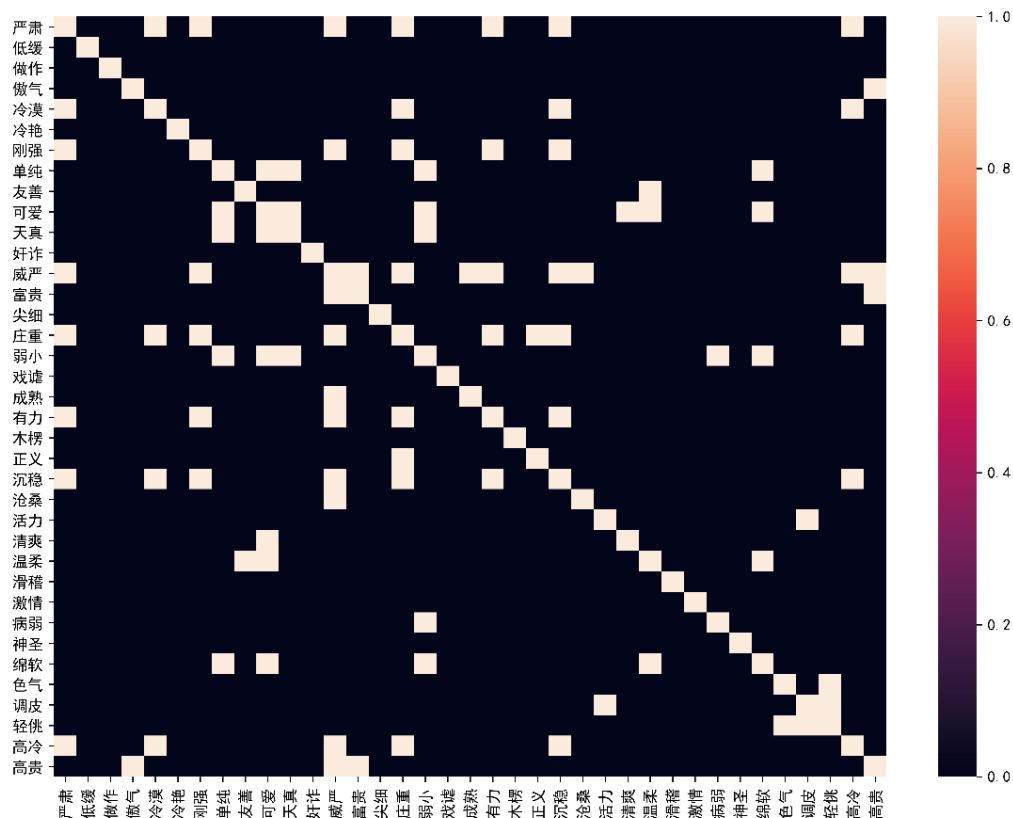


图 3-4 相关性系数大于 0.5 区域

由此可见，整体来讲各维度之间相关性不算太大。但是部分维度之间也存在一定的强相关性。该分析一方面证明了我们选择的维度的可行性，另一方面也提醒我们需要对选择的维度进行合理的合并与解释。

3.1.3 层级聚类

层次聚类是试图在不同层次对数据进行划分，从而形成树形的聚类结构。本次实验采用自顶向下的分拆策略以及 K-means 聚类算法，基于调查问卷得到的数据结果对三十七个形容词维度进行树形聚类。

(1) K-means 算法

K-means 算法是广泛应用于科学和工业诸多聚类算法中有效的算法之一。其工作机理是把 n 个样本点分为 k 个簇，各簇内的样本点具有较高的相似性，而各簇间的样本点相似程度较低，相识度的计算是依据一个簇中样本点的平均值来进行。算法的具体流程如下：

- i. 在样本数据 D 中选择 k 个样本点，将 k 个样本点分别当作初始聚类中心 $(C^{(1)}_1, C^{(1)}_2, \dots, C^{(1)}_k)$ ；
- ii. 在第 j 次迭代时，对样本点 D 中的所有点 $p_t (t=1, \dots, n)$ ，依次计算到各簇中心 $C^{(j)}_i$ 的欧氏距离 $d(t, i)$

$$d(t, i) = \sqrt{(p_t - C^{(j)}_i)^2} \quad (1)$$

- iii. 通过比较欧式距离，找出离 p_t 最近的簇中心，并将其归入到其中；
- iv. 更新各簇的聚类中心

$$C^{(j+1)}_i = \frac{1}{n_i} \sum_{t=1}^{n_i} p_{it} \quad i=1, 2, \dots, k \quad (2)$$

- v. 计算数据集 D 中所有点的平方误差 E_i ，并与前一次误差 E_{i-1} 比较。

$$E_i = \sum_{i=1}^k \sum_{t=1}^{n_i} |p_{it} - C^{(j+1)}_i| \quad (3)$$

若 $|E_{i+1} - E_i| < \sigma$ 则算法结束，否则进入 ii 再一次迭代。

(2) Gap Statistic

为确定每次使用 K-means 算法时 k 值的大小，本此实验采用 Robert 教授提出的 Gap Statistic 方法。该方法每次在样本空间里按照均匀分布随机地产生和原始样本数一样多的随机样本，并对这些随机样本做 K-means 聚类，再按公式 (3) 计算误差，如此反复对随机点进行 k 值取值为最大可接受值 k_{\max} 以下的 K-means 聚类。可以得到 $k_{\max}-1$ 个误差。利用每个 k 值对应的误差值计算 Gap Statistic 值：

$$\text{Gap}_n(k) = E_n^* \{\log(E_k)\} - \log(E_k)$$

最终取 Gap Statistic 值最大所对应的 k 值作为最佳簇类数量。

(3) 层次聚类

本实验分析层次聚类算法是无需人工干预的自动分层聚类方法，各节点聚类采用 Gap Statistic、K-means 算法，可以得到聚类对象的树形结构，具体实施通过 Python 代码实现：

- i. 构建 Node 类，类中包含属于该节点的数据集和孩子节点；
- ii. 先将所有数据点当作一个簇，也就是创建一个 Node 类对象，并将所有数据点放到该 Node 对象里；
- iii. 判断包含数据数量是否小于 2，若不小于 2 则使用 Gap Statistic 算法计算出最佳分类簇数，若小于 2 则结束算法；
- iv. 按照计算出来的分类簇数使用 K-Means 算法对数据集进行聚类，将每个簇放到一个新的 Node 对象里，并将 Node 对象的地址赋值给父节点的孩子节点；
- v. 对每个子节点执行返回 iii 步骤循环执行。

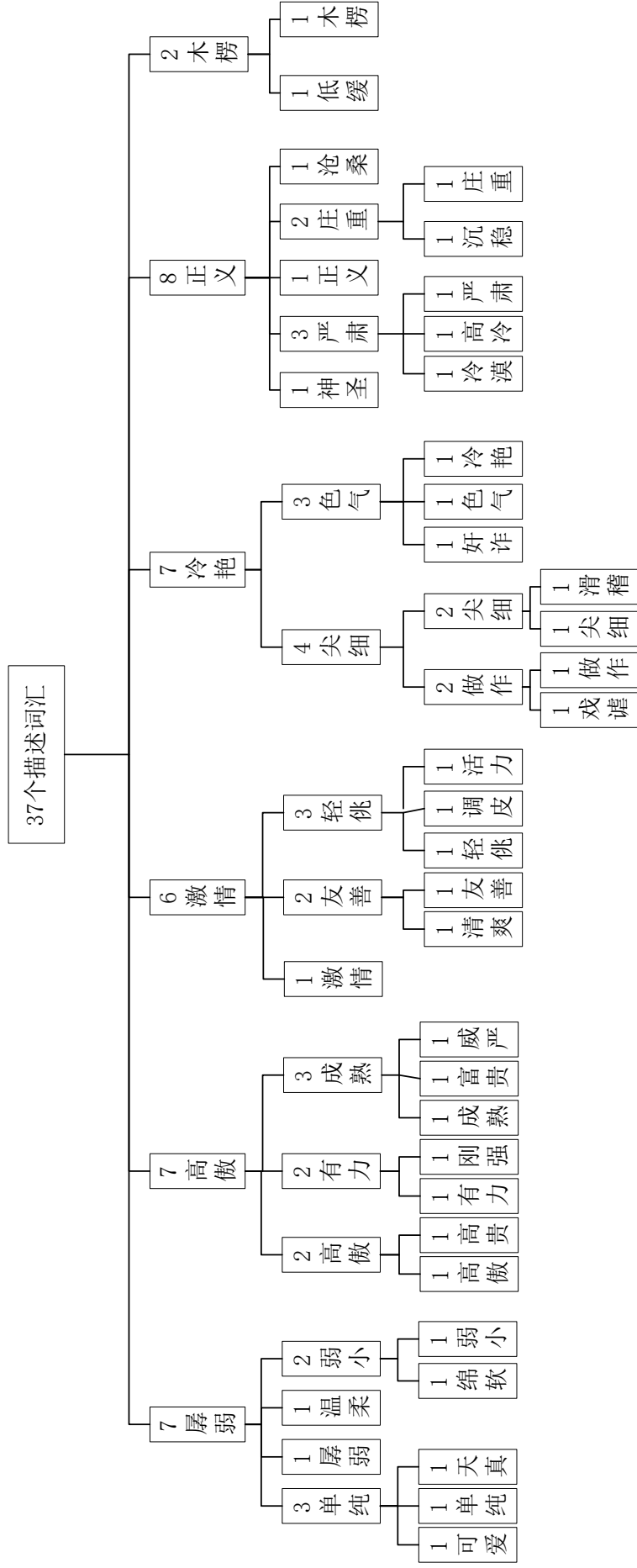


图3-5 层次聚类结果

图 3-5 即层次聚类算法生成的聚类树，从图中可以看到聚类结果呈一种由粗到细的层级分布；在第一层通过 Gap Statistic 算法可得将 37 个描述词划分为 6 类是最为合适的，这 6 类的中心分别是：孱弱、高傲、激情、冷艳、正义、木楞。其中以木楞为中心的簇中只有两个数据，分析原因是形容人说话比较呆的词汇相对其他的较少，所以木楞这一簇只有木楞和低缓；反观其他簇，数量都比较均匀，而且每一类词语意思都有理可循。

从上往下按簇分析，本文将 37 个发音风格形容词按第一次聚类结果分为 6 大分支：孱弱、高傲、激情、冷艳、正义、木楞，每个分支下又有不同的小分支。由于我们是寻找发音风格的描述系统，这里均站在听者角度来对各分支进行解释：孱弱簇下的词语主要给予听者一种单纯、柔弱、让人怜爱的感觉；高傲簇下的词语主要给予听者一种自信、高贵、成熟的感觉；激情簇下的词语主要给予听者一种有活力、积极的感觉；冷艳簇下的词语主要给予听者一种戏谑、引诱的感觉；正义簇下的词语主要给予听者一种严肃、冷静、沉着的感觉；木楞簇下的词语主要给听者一种说话迟缓、呆呆的感觉。

结合聚类树的结果对 6 大分支下的小分支进行修剪，将过细的分支合并，以达到合理地去掉相关性较高的描述词的目的。根据词语给听者所带来的感受进行合并。在孱弱簇内，将可爱、天真、单纯合并为单纯，绵软、弱小合并为绵软；在高傲簇下，将高傲、高贵融合为高傲，有力、刚强合并为有力，并将富贵删除；在激情簇下，去掉调皮；在冷艳簇下，删除做作、色气、奸诈；在正义簇下，删除冷漠、严肃和沉稳。再结合图 3-3 各词语维度之间的相关系数来看，删除了威严、绵软、友善。最终得出 20 个维度描述词，分别是：温柔、单纯、孱弱、有力、成熟、高傲、清爽、活力、激情、轻佻、尖细、滑稽、戏谑、冷艳、高冷、沧桑、神圣、正义、低缓、木楞。筛选后的 20 个维度描述词之间的相关系数如图 3-6 所示。

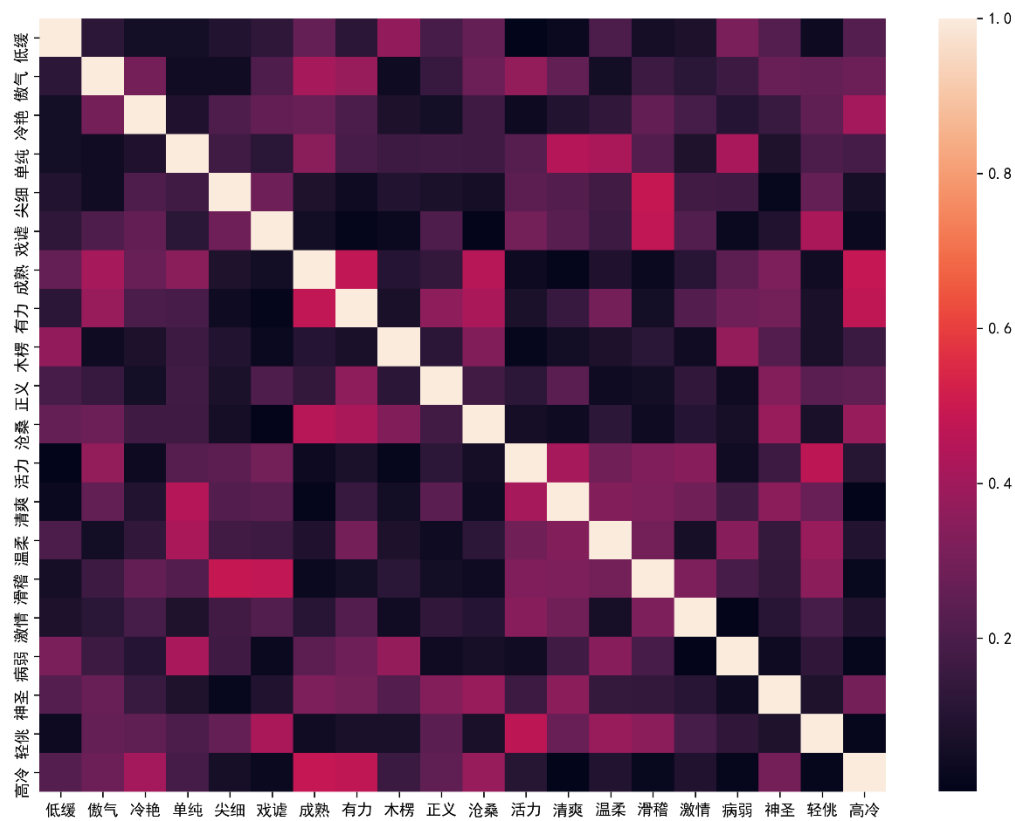


图 3-6 20 维度相关系数

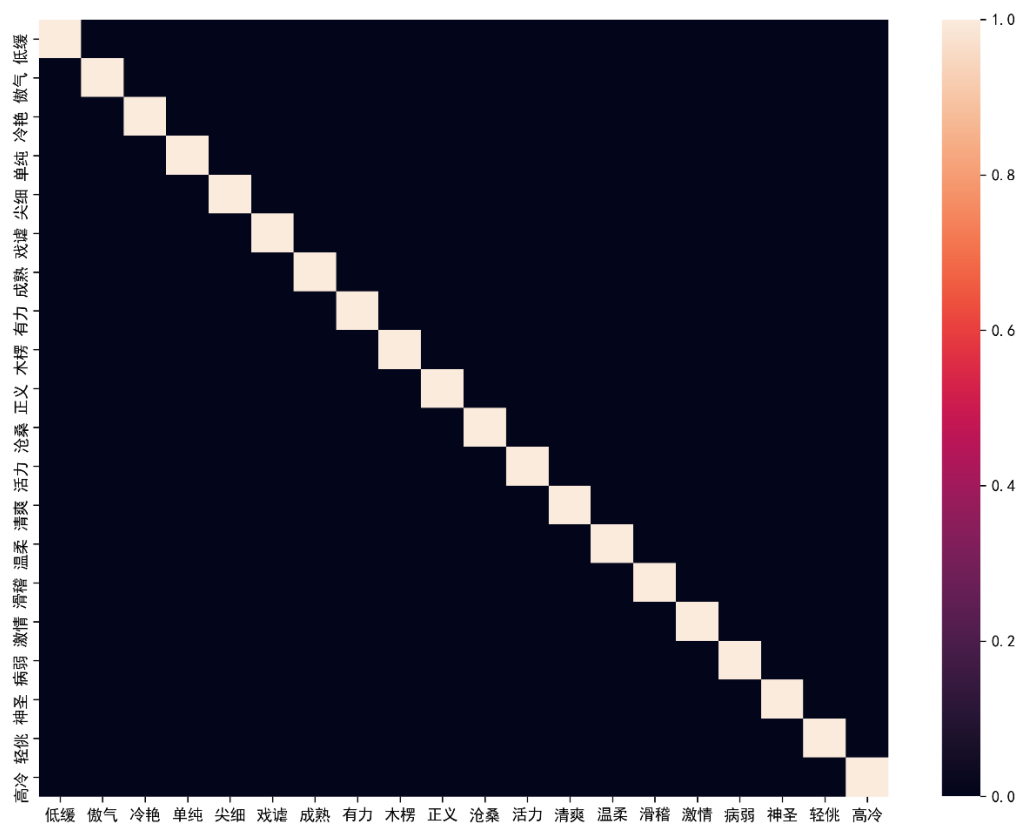


图 3-7 20 维度相关系数大于 0.5 的(图中白色区域为大于 0.5)

为了使数据可视化更加明显突出，对图像进行了条件显示，当相关系数大于 0.5 时为白色，小于为黑色，如图 3-7 所示。可见经过聚类树以及相关性筛选合并后，各维度之间相关系数均无大于 0.5 的，满足基本的各维度独立性的要求。

3.2 发音调控

发音调控是发音者通过对感知的信息理解加工来得到发音的细节过程。根据心理学和朗读学中对该过程的描述，本文将该过程又分为 2 个阶段，分别为咬字和韵律，并先将各个阶段进行了初步的细化，如下所示：

- 咬字阶段：咬字清晰度、单字发音到位程度、相邻字发音黏着程度；
- 韵律阶段：平均音高、音高变化程度、音高变化频率、语调的平顺性、停顿频度、语速控制、节奏的稳定性。

为调查阶段细化的合理性，我们进行了相应的问卷调研。

3.2.1 问卷设计

该问卷目的是审视我们发音调控阶段提出的描述维度的合理性并对其做合理的筛选。由于发音调控是对宏观评价的一个补充，所以本次实验通过各维度与宏观评价的关系来进行分析。所以本次问卷将一些宏观维度的词汇(为保证情感的丰富性选取了 38 个描述词)作为问题，发音调控各维度描述词作为问题，让受试者根据发音调控各维度与宏观评价的维度的相关性进行打分。问卷设计如图 3-8 所示。

中午12:38 0.9K/s 78%

一个小小的问卷调查

你的性别: ☐ 男 ☒ 女

给自己起个名字 (默认会随机给个) 29:12:12:06871144890

填上一次的名字可以让你接着你上次答的位置答

信息填好了, 开始吧!

下午4:09 0.2K/s 100%

活力

声音的描述 --反向-----中间线-----正向---

| 维度 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|------------|---|---|---|---|---|---|---|
| 停顿次数 | ★ | ★ | ★ | ★ | ★ | ★ | ★ |
| 语速 | ★ | ★ | ★ | ★ | ★ | ★ | ★ |
| 节奏稳定 | ★ | ★ | ★ | ★ | ★ | ★ | ★ |
| 发音干净 | ★ | ★ | ★ | ★ | ★ | ★ | ★ |
| 单字发音到位 | ★ | ★ | ★ | ★ | ★ | ★ | ★ |
| 相邻字发音之间的黏着 | ★ | ★ | ★ | ★ | ★ | ★ | ★ |
| 平均音调高低 | ★ | ★ | ★ | ★ | ★ | ★ | ★ |
| 语调变化程度 | ★ | ★ | ★ | ★ | ★ | ★ | ★ |
| 语调变化频率 | ★ | ★ | ★ | ★ | ★ | ★ | ★ |
| 语调切换的圆润度 | ★ | ★ | ★ | ★ | ★ | ★ | ★ |

提交, 并开始下一个

图 3-8 发音调控问卷设计

可见受试者可以根据宏观评价维度与发音调控维度的相关性进行打分，反向为逆相关，正向为正相关，零为需要正常水平，若该维度无关则可以点击为红色。

3.2.2 数据分析

和宏观维度分析类似，首先对数据的相关性进行分析。图 3-9 为本次调研得到数据的相关性系数示意图。图 3-10 为相关系数大于 0.5 的示意图。

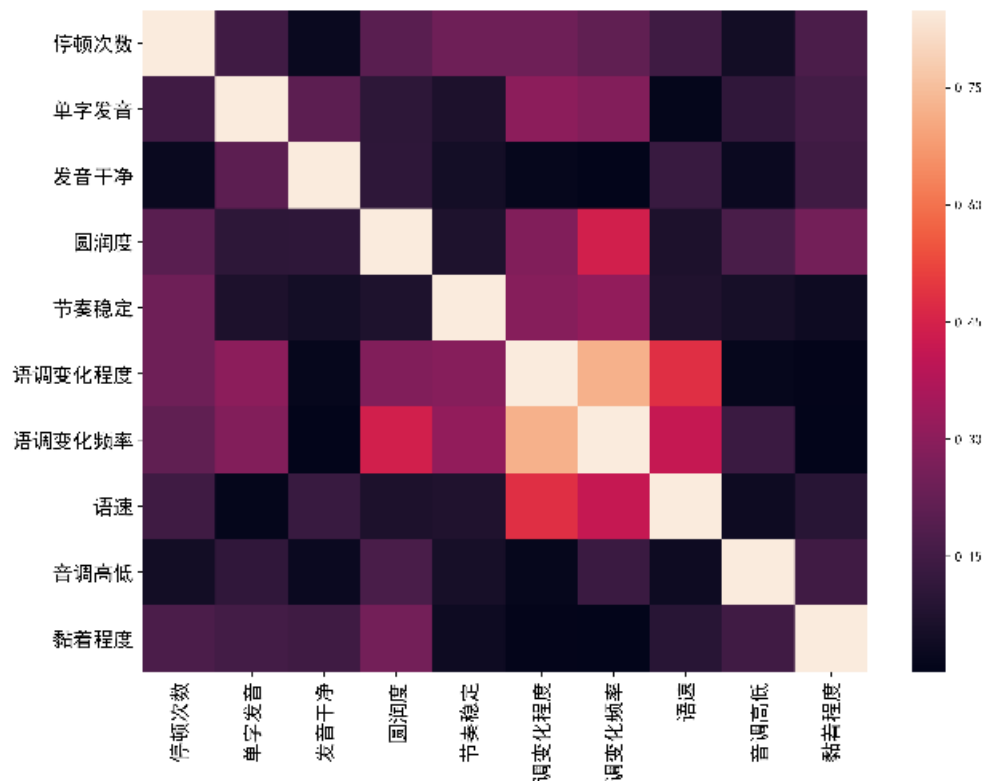


图 3-9 发音调控各维度之间的相关系数

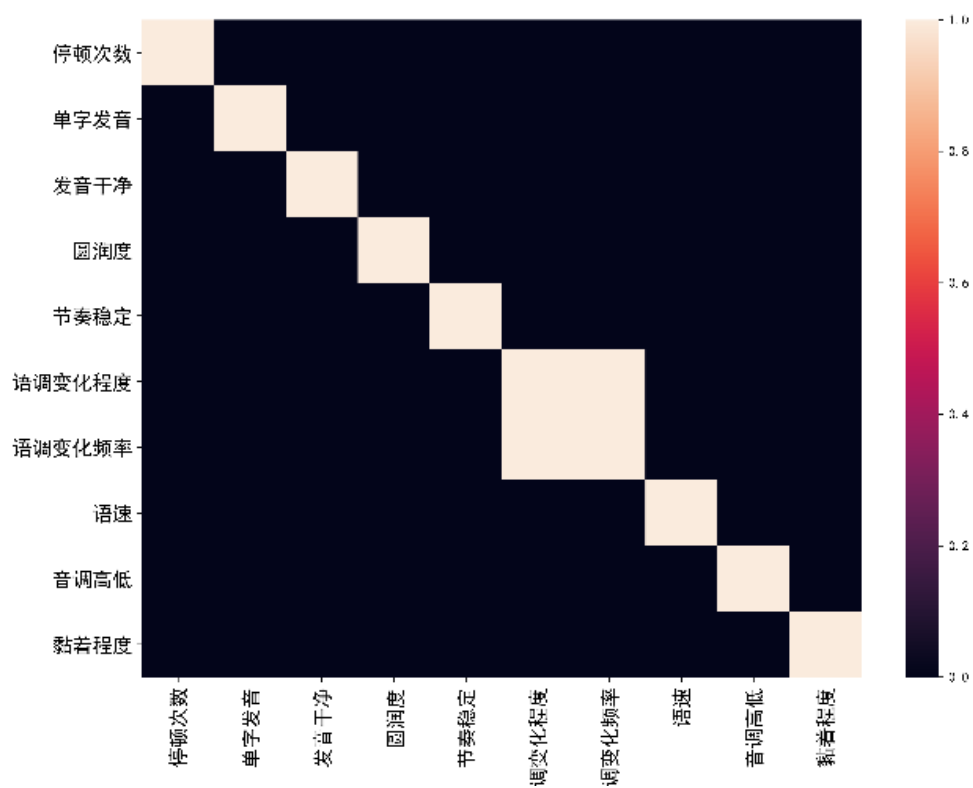


图 3-10 发音调控各维度之间的相关系数大于 0.5 的示意图（大于 0.5 为白色）

由此可见，各维度相关性不高。唯一一组超过 0.5 的是音高变化程度-音高



变化频率，并且该组的相关性系数达 0.8，所以将二者合并为音高变化。

维度之间有较多无关数据，如果打分为无关则表示该发音调控的维度对于宏观评价影响不大。图 3-11 则是各描述维度中被评价为无关的数量的示意图。

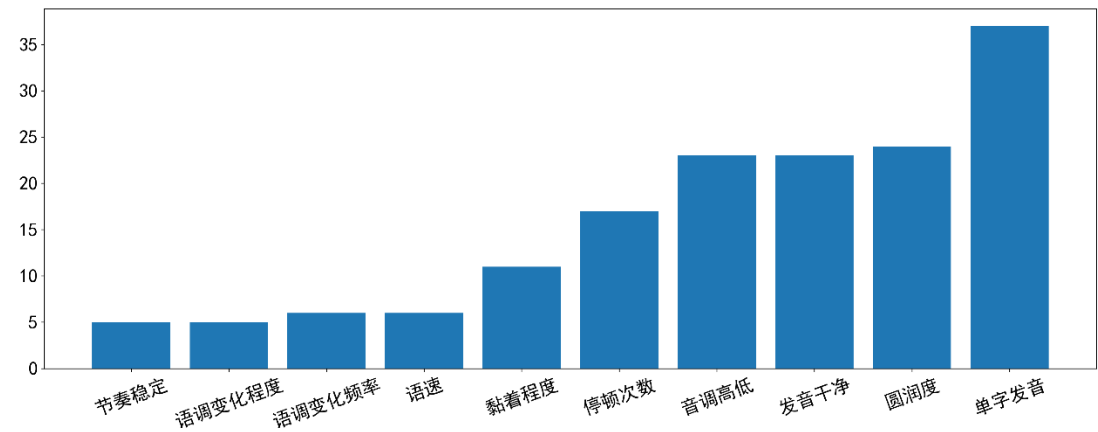


图 3-11 发音调控各维度被评为“无关”的数量

从图中可以清楚得看到，无关数量最多的是单字发音到位程度。超过半数无关的有平均音高程度、咬字清晰度和语调的圆润程度。由此可见单字发音到位程度与发音风格的联系性不太紧密，故将其删除。

经过分析后，最终得到的发音调控描述维度为：

咬字阶段：咬字清晰度、相邻字发音黏着程度；

韵律阶段：平均音高、音高变化、语调的平顺性、停顿频度、语速控制、节奏的稳定性。

3.3 生理特性

在人的认知调控之下，由生理发音学可知，肺、气管、声带、口和鼻等器官与个性化发音风格也有着紧密的联系。而除开病变等特殊情况，器官的发育与人的成长以及基因有关。由于描述系统更多是站在听者的角度描述较为宏观的个体，所以本文在生理特性方面提出年龄感、性别感、清脆度的三个评价维度，各维度均有五个程度，分别为：

- 年龄感：幼年音-少年音-青年音-中年音-老年音；
- 性别感：男子力-女子力；
- 清脆度：暗哑-明亮；

其中性别感维度在本描述方案中对于发音的听觉体验更加偏重，考虑到部分

人群独特的发音听觉感受，采用男女子力的程度来进行描述。出于尽可能全面刻画评价内容的考虑。以上各维度之间可能存在交叠现象。为了探究这些维度间的相互关系，进行了和发音调控类似的心理学调查实验（由于实验方法类似，故不再赘述）。

和发音调控数据分析类似，首先分析各维度之间的相关性系数。如图 3-12 所示。

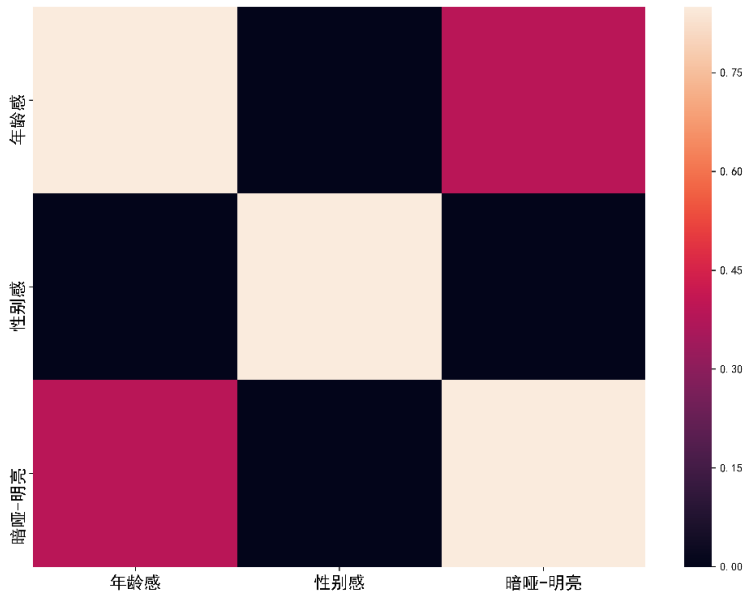


图 3-12 生理特性维度相关系数

从相关系数可以看出，三个维度之间相关性比较低，性别感和其他各个维度相关性都为 0。图 3-13 为各维度中与测试词之间关系被标记为“无关”的数量示意图。

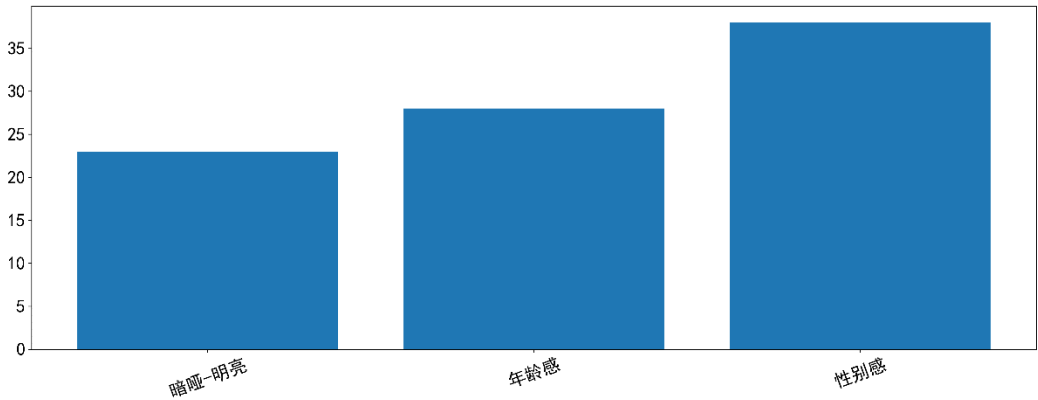


图 3-13 与测试词之间关系“无关”数量示意图

从图中可以看到。性别感和被测发音风格描述词“无关”数量有 38 个，其

余也都超过了半数。由此可见生理特性对于发音风格来说，只有部分具有相关特色的风格才会受其影响。此处将明亮-暗哑、年龄感的详细均分可视化出来，如图 3-14 所示。

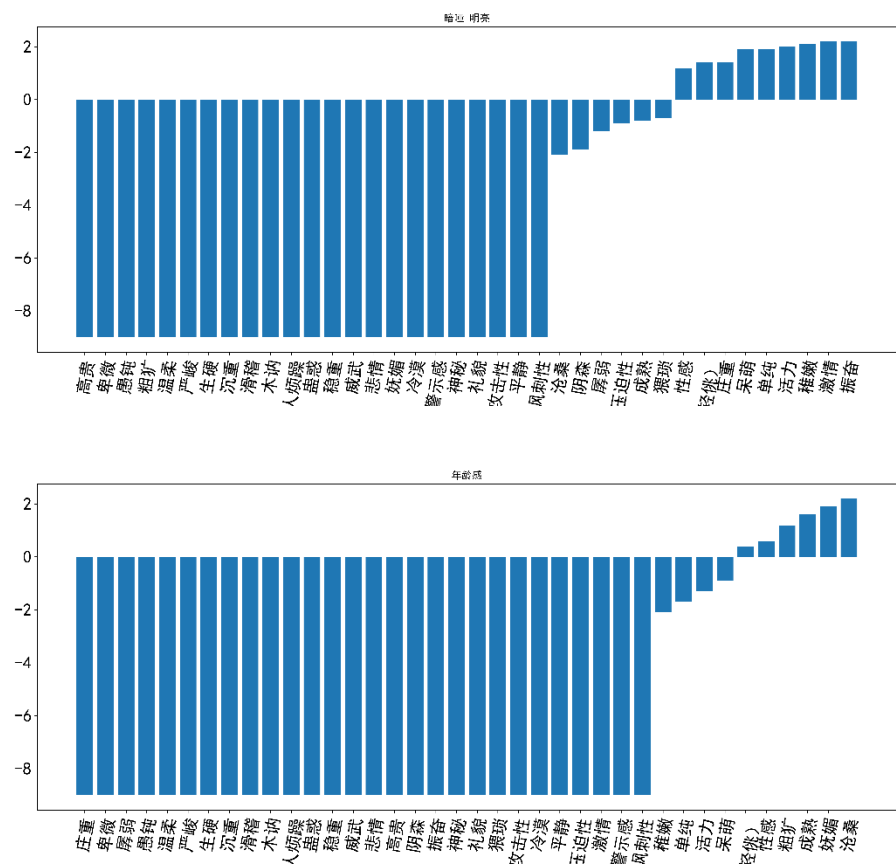


图 3-14 明亮-暗哑、年龄感均分详细示意图

从图可以看出，与生理特性相关的宏观维度如成熟、单纯、呆萌等，均是对身体本身的特性有一定要求。反观性别维度，站在发音风格的角度，不管是男是女他们均能表现出各自的风格。根据分析结果，年龄感、性别感和清脆度基本符合要求，能够较好地描述生理特性。

3.4 本章小结

本章工作主要围绕个性化发音风格描述问题展开：


(1) 以心理学和朗读学为主的发音过程研究为参考。将发音者基于感知事物的风格发音生成及衍化并被听者感知的过程概括为生理特性、发音调控和宏观评价等几步。各步骤之间又有相互联系，前者为后者打下基础，彼此配合完成这一

过程。

(2) 基于发音者感知信息到听者听到的过程, 提出了个性化发音风格描述体系, 包含生理特性、发音调控和宏观评价三种成分, 分别从不同视角解读个性化发音风格的不同方面, 各视角互为补充、缺一不可, 共同组成了发音风格的分布式表达。

(3) 经过对生理特征、发音调控和宏观评价三个模块的问卷调查, 最终得出了各个模块的描述维度, 如下:

生理特征:

- 年龄感: 幼年音-少年音-青年音-中年音-老年音;
- 性别感: 男子力-女子力;
- 清脆度: 暗哑-明亮; 

发音调控:

- 咬字阶段: 咬字清晰度、相邻字发音黏着程度;
- 韵律阶段: 平均音高、音高变化、语调的平顺性、停顿频度、语速控制、节奏的稳定性;

宏观评价:

- 该模块为 20 个描述词: 温柔、单纯、孱弱、有力、成熟、高傲、清爽、活力、激情、轻佻、尖细、滑稽、戏谑、冷艳、冷漠、沧桑、神圣、正义、低缓、木楞。

共计 31 个描述维度。

4 发音风格数据库建立

基于第 3 章提出的个性化发音风格描述体系，我们构建了发音风格数据库，一方面用于验证发音风格描述体系的合理性，一方面为下一章的情感预测模型的训练提供数据支持。

4.1 概述

在情感语音研究中，按采集方式的不同将情感语言数据分为自然语音、诱导语音和表演语音。

自然语音是指采集人与人日常生活中的自然对话来作为训练数据，其真实性和自然度都是最好的，但是获取起来非常困难。对于采集数据后的数据分析来说，对语音的清晰度与质量也有较高的要求，所以要想采集到清晰且高质量的自然语音是极难实现的。

诱导语音是指说话人利用带有情感刺激性的诱导因素来产生相应的感情，从而发出带有相应感情的语音信号。但该方法由于个体的认知体验的不同会出现情感可控性不强的情况。

表演语音是指专业的人员通过某种场景的联想进而对某些感情进行主观模仿。获取表演语音的方式主要有两种：一是对电影、广播或游戏的音频文件进行剪辑截取需要的类型，这种方式可以保证情感的自然度，但是由于内容较广泛容易出现语音内容和语音质量不佳的情况；另一种方式是请专业人员来对准备好的文本素材进行朗读录制，该方法也是目前建立语音数据库最常用的方法，该方法获得的数据在语音质量、情感自然以及语音环境等条件都非常容易满足。

由于本文是研究不同发音风格之间的区别，再加上研究条件有限，本文采用方式一来获取表演语音。为避免数据不规整现象的出现，本文尽量获取无杂音的纯声语音；为减少除发音风格以外的影响因素，本文选取的数据语料要求几乎一致，并且力求各个发音角色的风格有所不同。

数据库的搭建流程包含以下几步：

- (1) 语音录制和切割阶段：将游戏中 48 个角色的语音内容录制下来，并将每个角色的语音按台词内容切割；

(2) 个性化发音风格信息标注：基于本文提出的个性化发音风格描述体系对语音片段进行信息标注；

(3) 标注结果处理及分析：对标注的结果进行分析和筛选。

接下来分别对每个过程进行具体介绍。

4.2 语音录制和切割阶段

本文对游戏中不同角色的语音信号进行了采集。一共获取到了该游戏中 48 个不同的角色的语音内容，每个角色均有约 25 句台词且各个角色的台词都几乎相同，这一定程度上解决了内容所带来的影响。音频的录制采用手机录频软件，再利用处理软件将视频中的语音分离出来。最终成功获取到了 48 个不同角色的全语音内容。每个角色 25 句台词，平均每段音频持续时长为 2 分钟左右。获取到的数据如图 4-1 所示。

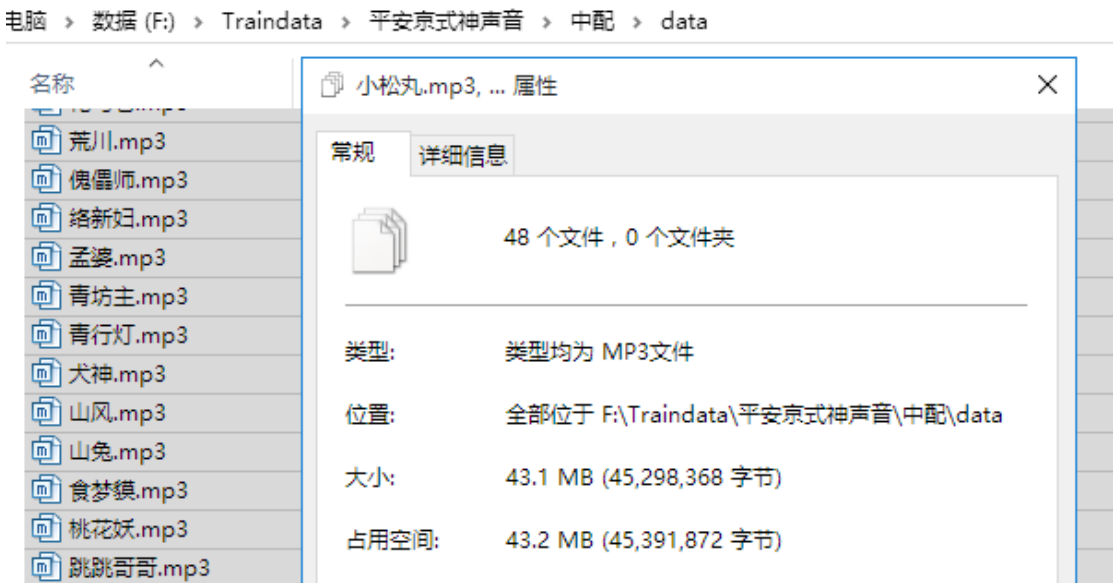


图 4-1 48 个角色所有台词录音信息图

由于每个角色各自的发音风格有所不同，并且同一角色不同台词的发音风格也有所区别。所以本文利用剪切软件根据台词对每个角色的语音文件进行了分割，如图 4-2 所示。

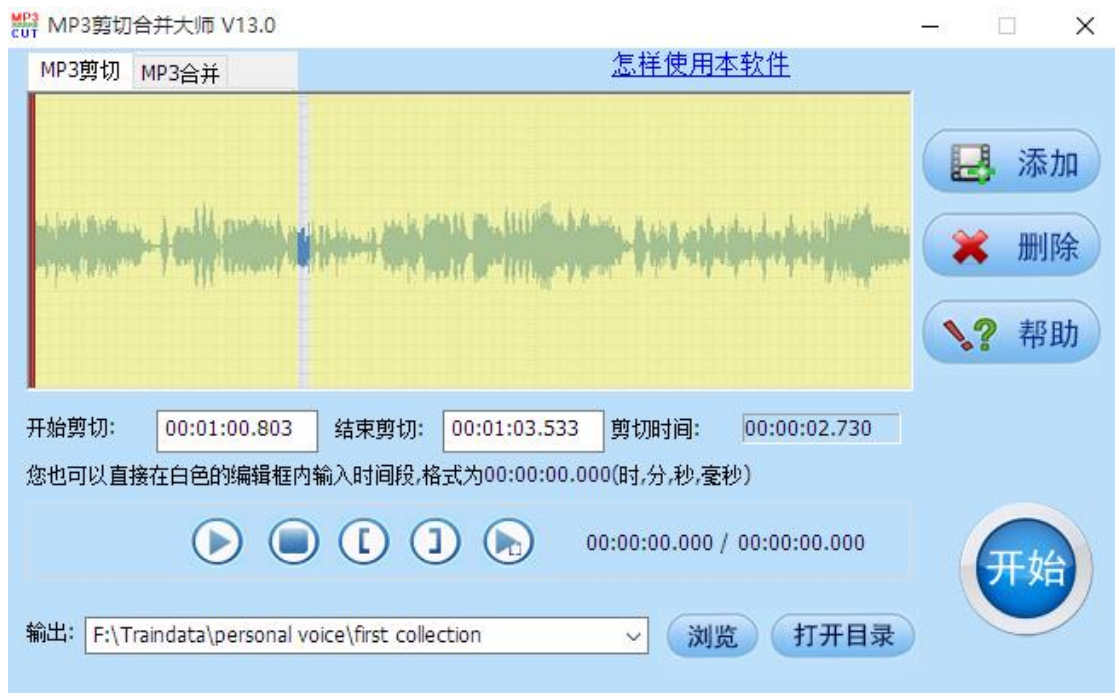


图 4-2 根据台词剪切音频示意图

将每个角色的语音分为 25 个不同台词的语音文件，由于其中有 5 条台词持续时间过短，故将其删除。最终获得 860 个不同发音风格的音频文件，每个音频文件为一段台词的朗读，平均每段音频持续时间为 2 秒左右，如图 4-3 所示。

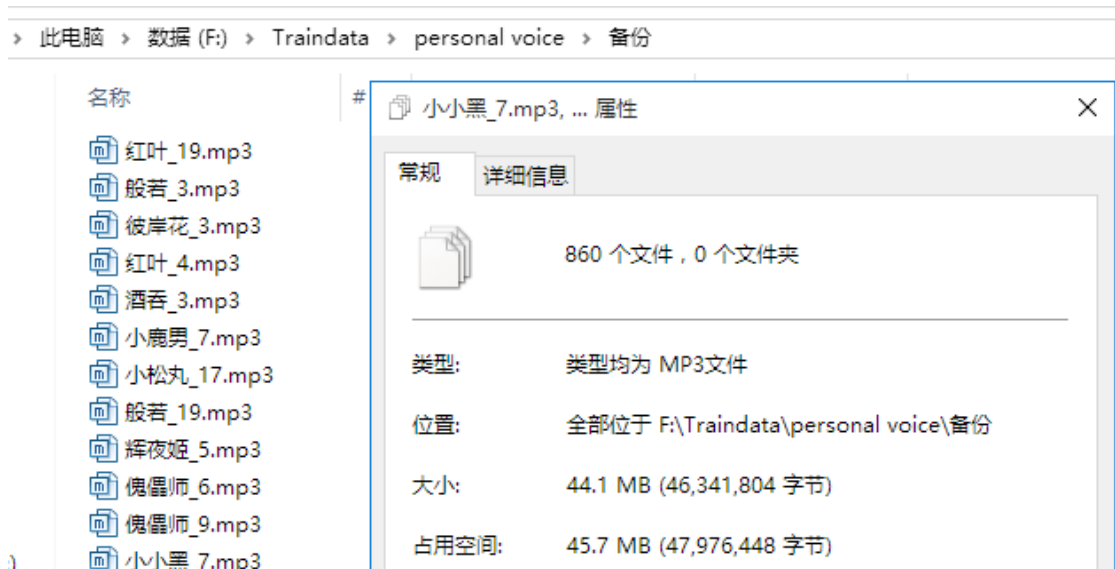


图 4-3 860 个音频数据信息图

4.3 个性化发音风格信息标注

对上述 860 段音频进行人工标注。本文根据标注内容对第 3 章所用的 APP 进行了轻微的修改,将其问题修改为了本文提出的个性化发音风格描述体系的各个维度标签。为减少维度之间标注时的影响,将原本的三层结构的维度描述随机打乱排列,如图 4-4 所示。



图 4-4 发音风格信息标注 APP 示意图

每个维度最大评分刻度为 7,最低为 0,通过填充星型控件来控制分数。为防止标注者审美疲劳,要求标注者每隔一小时标记 20 个,一天最多标记 140 个数据。

4.4 标注结果处理和分析

通过一周的时间,完成了对 860 个音频数据的标注,如图 4-5 所示。经过筛查发现,其中有 5 条数据由于录制和剪切的原因,有一定的背景杂音,故将其删除,最终留下 855 条有效数据。

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | AA | AB | AC | AD | AE | AF | |
|----|----|------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|
| 1 | | audio_name | 低缓 | 停顿 | 冷漠 | 冷艳 | 单纯 | 发音 | 孱弱 | 尖细 | 平均 | 年龄 | 性别 | 戏谑 | 有力 | 木楞 | 正义 | 沧桑 | 活力 | 清爽 | 清脆 | 温柔 | 滑稽 | 激情 | 相邻 | 神圣 | 节奏 | 语调 | 语调 | 语调 | 语调 | 轻佻 | 高傲 |
| 2 | 0_ | (1).mp3 | 2.8 | 2.4 | 0.7 | 5.7 | 0.8 | 4.9 | 0.7 | 5.7 | 5.1 | 2.7 | 6.6 | 6.3 | 1.9 | 0.5 | 2.4 | 1.6 | 3.5 | 4.5 | 0.7 | 2.9 | 4.2 | 2.5 | 4.5 | 0.6 | 0.9 | 5.1 | 4.6 | 3.6 | 5.8 | 5.4 | |
| 3 | 1_ | (2).mp3 | 1.6 | 2.5 | 3.3 | 0.4 | 6.2 | 5.9 | 0.7 | 1.6 | 1.9 | 2.6 | 2.1 | 0.6 | 2.5 | 2 | 5.5 | 0.5 | 2.7 | 5.3 | 4.1 | 5.7 | 0.3 | 2.2 | 1.4 | 1.3 | 5.8 | 1.3 | 0.8 | 2.6 | 0.4 | 3 | |
| 4 | 2_ | (3).mp3 | 5.9 | 3.4 | 1.7 | 6.3 | 1 | 2.7 | 0.6 | 1.5 | 0.8 | 3.2 | 5.4 | 5.9 | 0.7 | 0.7 | 0.7 | 2.8 | 0.3 | 0.8 | 1.1 | 5.9 | 0.9 | 0.6 | 5.8 | 0.4 | 3.5 | 5.5 | 4.5 | 2 | 2.5 | 4.4 | |
| 5 | 3_ | (4).mp3 | 2.7 | 2.1 | 5 | 6.3 | 0.4 | 3.8 | 0.7 | 5.4 | 4.5 | 3 | 6 | 6.2 | 2.5 | 0.7 | 0.8 | 0.5 | 3.8 | 3.5 | 4.7 | 3.9 | 0.7 | 0.6 | 3 | 0.7 | 2.6 | 5.9 | 4.5 | 0.7 | 6.8 | 6.8 | |
| 6 | 4_ | (5).mp3 | 0.8 | 2.6 | 0.5 | 0.6 | 1.6 | 5 | 0.9 | 0.8 | 2.8 | 3.6 | 1.3 | 0.7 | 5.5 | 0.7 | 6.4 | 3.8 | 3 | 4.9 | 0.5 | 4.5 | 0.8 | 0.6 | 1.7 | 3.5 | 0.6 | 2.3 | 1.2 | 3.5 | 0.4 | 6.3 | |
| 7 | 5_ | (6).mp3 | 0.6 | 4.5 | 0.7 | 0.4 | 6 | 4.6 | 0.6 | 1 | 0.4 | 2.1 | 2.3 | 0.6 | 5.6 | 0.5 | 6.7 | 0.3 | 1.4 | 6.2 | 5 | 5.5 | 0.6 | 3.5 | 1.9 | 2.8 | 4.8 | 2.4 | 1.7 | 3.9 | 0.7 | 3.9 | |
| 8 | 6_ | (7).mp3 | 2.6 | 2.7 | 0.6 | 0.4 | 6.5 | 4.6 | 6.6 | 6.3 | 4.9 | 0.4 | 4.7 | 0.6 | 0.9 | 0.6 | 4.6 | 0.3 | 6.8 | 4.5 | 5.7 | 3.7 | 0.5 | 2.2 | 4.7 | 0.7 | 0.8 | 4.3 | 5.5 | 6 | 4.4 | 0.6 | |
| 9 | 7_ | (8).mp3 | 5.2 | 2.3 | 0.5 | 0.5 | 4.9 | 3.3 | 5.1 | 1.5 | 0.8 | 2.3 | 1.8 | 0.6 | 2.6 | 1 | 4.4 | 0.7 | 1.6 | 3.8 | 5.2 | 5.2 | 0.4 | 0.6 | 1.9 | 1.2 | 5.7 | 1.7 | 2 | 3.5 | 0.9 | 1.2 | |
| 10 | 8_ | (9).mp3 | 6.5 | 1.8 | 0.3 | 6.7 | 6.3 | 3.2 | 6.6 | 3.4 | 1.4 | 2 | 6.1 | 0.5 | 0.3 | 5.3 | 3.7 | 0.3 | 0.6 | 0.6 | 0.4 | 6.6 | 0.4 | 0.4 | 4.7 | 1.8 | 4.6 | 5.1 | 2.6 | 0.8 | 1.6 | 6.3 | |
| 11 | 9_ | (10).mp3 | 6.9 | 1.6 | 6.6 | 6 | 3.3 | 0.7 | 6.5 | 2.2 | 0.3 | 2 | 5.3 | 0.5 | 0.3 | 5.7 | 0.6 | 2.6 | 0.5 | 0.5 | 0.5 | 5.7 | 0.5 | 0.4 | 6.6 | 0.5 | 6 | 4.7 | 0.6 | 0.4 | 0.4 | 0.6 | |

图 4-5 标注结果示意图

对整合后的标注数据进行统计学分析，下面分别从分布和样本标签相关性的角度对本数据库展开探讨。

4.4.1 数据分布

通常在训练模型前需要对数据集进行分布分析，进而改变策略解决隐患。本文按照该思路将数据标注结果分布进行了可视化，如图 4-6 所示。

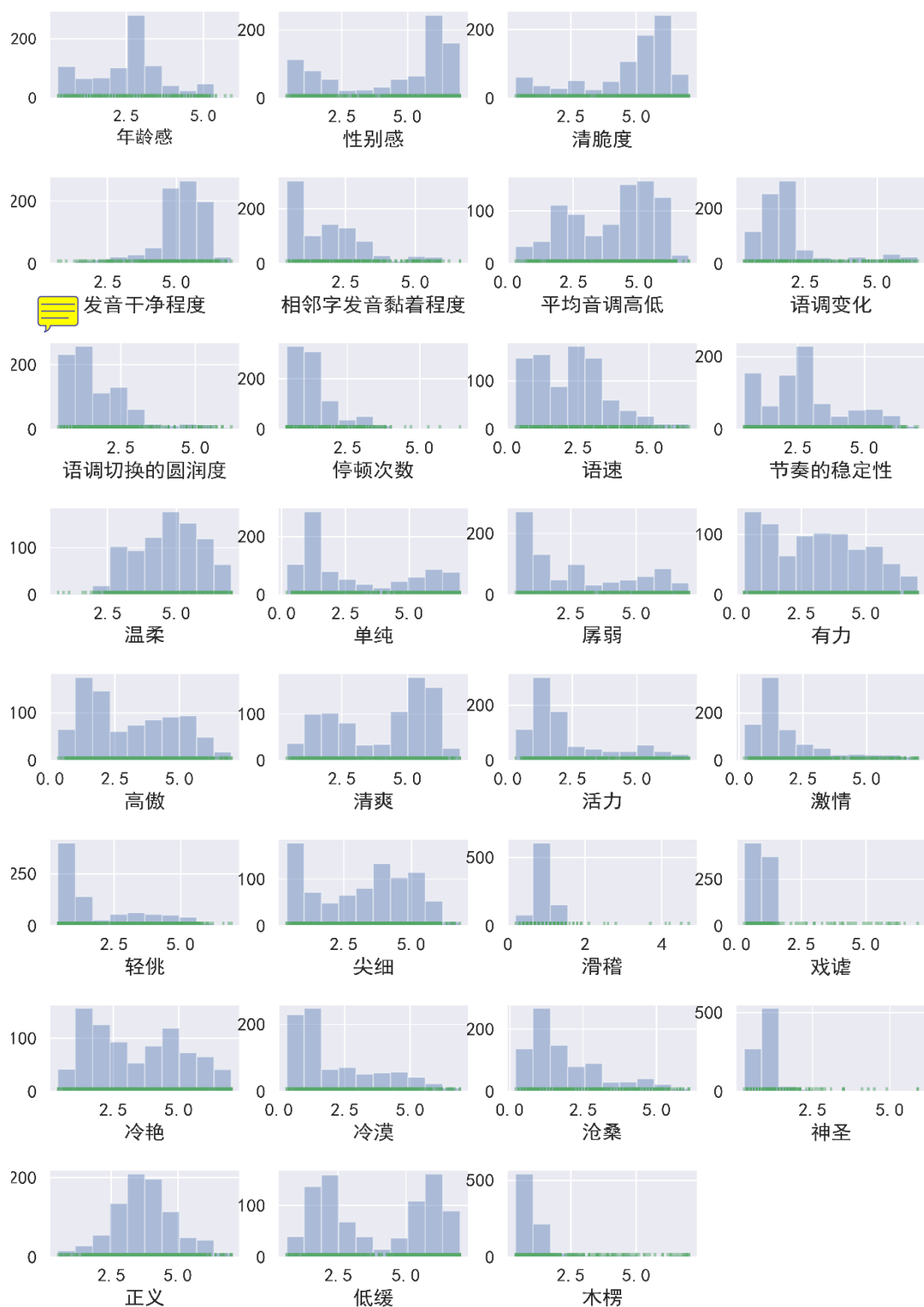


图 4-6 发音风格数据库标注分布示意图

通过对数据的可视化分析，发现数据部分维度有较强的非均衡性。

第一排是数据集在生理特性这一模块的数据分布。可以看到该数据集在年龄维度分布略微靠前，分析其原因是数据源年老的角色较少；性别感呈两级分布，

较为合理；清脆度分布略微靠后，这一定程度也肯定了年龄感分布的理由，发音者年龄相对较为年轻所以清脆值较高的数据更多。二三排是发音调控模块的数据分布。可以看到由于均为较为专业人员配的音，所以发音都较为干净；由于每句台词不长，所以相邻字发音黏着度都不高，音高音高变化不多以及其切换平顺性偏低，停顿频度也因此偏少，语速和节奏的稳定性都略低。剩下的均为宏观评价模块的数据分布。可以看到滑稽、戏谑、神圣、木楞等几个维度都有较强的不均匀现象。分析其原因是采取的音源较为正式，故滑稽、戏谑和木楞风格的数据较少。而神圣发音的数据多为有较强的回音和背景杂音，故相关数据偏少。

针对数据不均匀的问题，在之后的模块中会根据算法进行调整。

4.4.2 相关性分析

为探究本文提出的发音风格描述体系在该数据集上的表达程度，对标注数据进行了相关性统计，如图 4-7 所示。

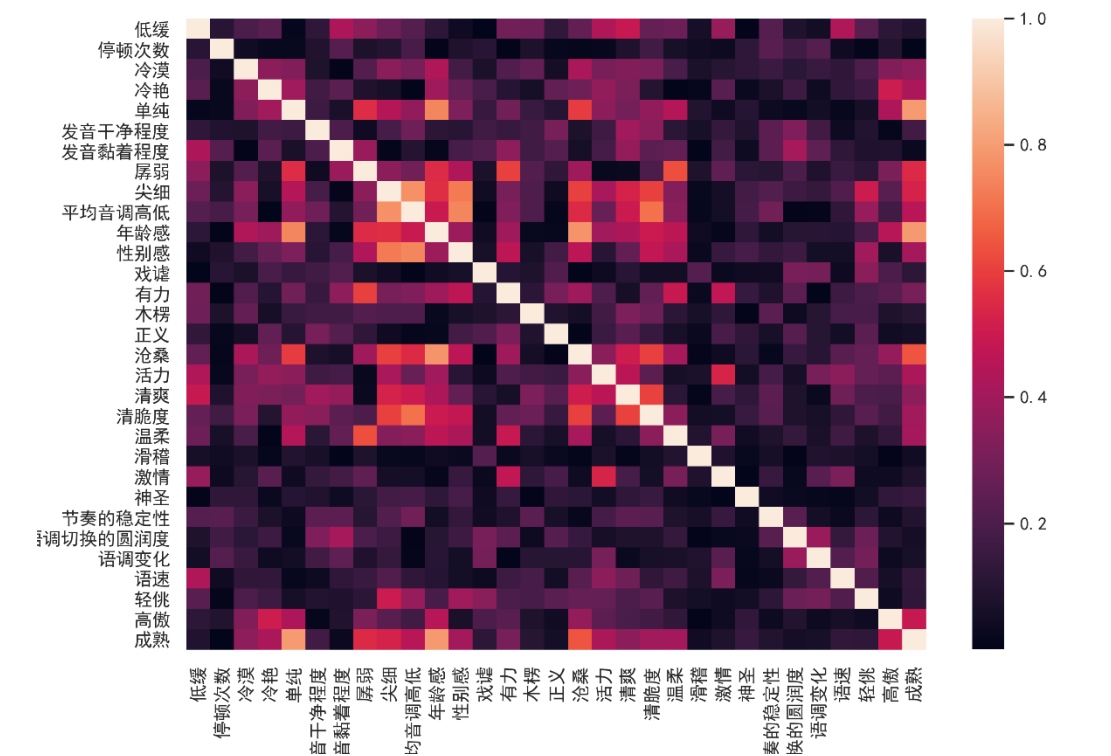


图 4-7 根据标注计算的各维度相关系数

从图中可以看到整体来讲，在该数据集上各维度相关系数较小，绝大部分都小于 0.5。其中相关系数超过 0.5 的大多数由年龄感、音调高低、成熟度、孱弱

度、尖细度、沧桑度、性别感、清脆度等和器官发育相关的维度。分析其原因是数据源的发音者们大多是以正常音线进行发音，所以发音的尖细程度以及成熟等会和生理特性有较强的关系。

4.5 本章小结

本章主要工作是采用表演语音的形成构建发音风格数据库。建立流程包括语音录制和切割、个性化发音风格信息标注和标注结果处理和分析几部，得到的结论有：

- (1) 总共收集了 860 段音频文件；
- (2) 对 860 段音频进行发音风格信息标注；
- (3) 最后通过分布和相关性两个方面分析了标注结果，得出数据在部分维度分布不太均匀以及部分维度之间相关性系数大于 0.5 的问题，在之后的应用中会采取相应的措施进行调整。

建模以及实验

生理特征

I 年龄感

I 性别感

I 清脆度

- 数据准备
利用自己采集的860个短音频进行训练
对每个音频进行最大值为7的label标注
- 特征提取
MFCC特征
生理特征主要与音高、谐波以及音高抖动有关，所以采用MFCC可以对音高、谐波等有具体的表现
- 模型选择
采用卷积神经网络，由于MFCC特征提取后特征有较强的横纵向、局部相关性所以采用卷积核对其进行处理。
经过卷积池化降维后，利用全连接层对提取到的特征进行权重结合，最终预测出结果
- 最终结果

年龄感误差：0.70065402681 / 7.0

性别感误差：1.13084420650 / 7.0

清脆度误差：1.15585780881 / 7.0

发音调控

I 咬字阶段：咬字清晰度、相邻字发音黏着程度；

I 韵律阶段：平均音高、音高变化、语调的平顺性、停顿频度、语速控制、节奏的稳定性

- 数据准备
利用自己采集的860个短音频进行训练
对每个音频进行最大值为7的label标注
利用腾讯智聆api对860个音频做了音节边界标记，用于音节分割训练
- 特征提取
 - pitch
音高是能直接提取得到的，所以通过直接提取音频的音高来获得平均音高、音高变化
 - MFCC
黏着度与包络连贯性有关、咬字干净与谐波组成有关、不同音节在频谱上的表现也有所不同。所以本模块也用了mfcc作为发音清晰度、相邻字发音黏着度、语调平顺性。

- Fbank

和MFCC相比缺少了最后的离散余弦变化一步，更多的是希望符合声音信号的本质，拟合人耳接收的特性。不同的音节能在Fbank特征上有较好的体现

- 音节提取

语速、节奏稳定性、停顿等都需要用到音节边界这一特征

- 模型选择

- 发音清晰度、黏着度、平顺性三个直接利用MFCC特征训练的采用的是和生理特征模块相同的模型——卷积神经网络

最终结果为

```
发音清晰度平均误差:    0.7744751781850437 / 7.0
发音黏着程度的平均误差:  1.0075982218 / 7.0
语调的平顺性的平均误差:  0.92381925050 / 7.0
```

- 音节分割

采用vad方法检测出silence

将fbank特征与pitch按照时域拼接

利用卷积神经网络对特征进行处理训练，最终结果为

| | F1 | HR | OS | R |
|-------|------------|------------|------------|------------|
| count | 860.000000 | 860.000000 | 860.000000 | 860.000000 |
| mean | 0.812844 | 79.654628 | -3.560585 | 0.822803 |

R-value采用动态region方法进行测评

基本满足需求

- 平均音高、音高变化

平均音高采用 pitch的（中位数+均值）/ 2

音高变化采用pitch的方差

- 停顿频度

音节之间间隔次数/时间

- 语速

音节发音时间的均值

- 节奏的稳定性

各音节发音时间的方差

宏观（抽象）评价

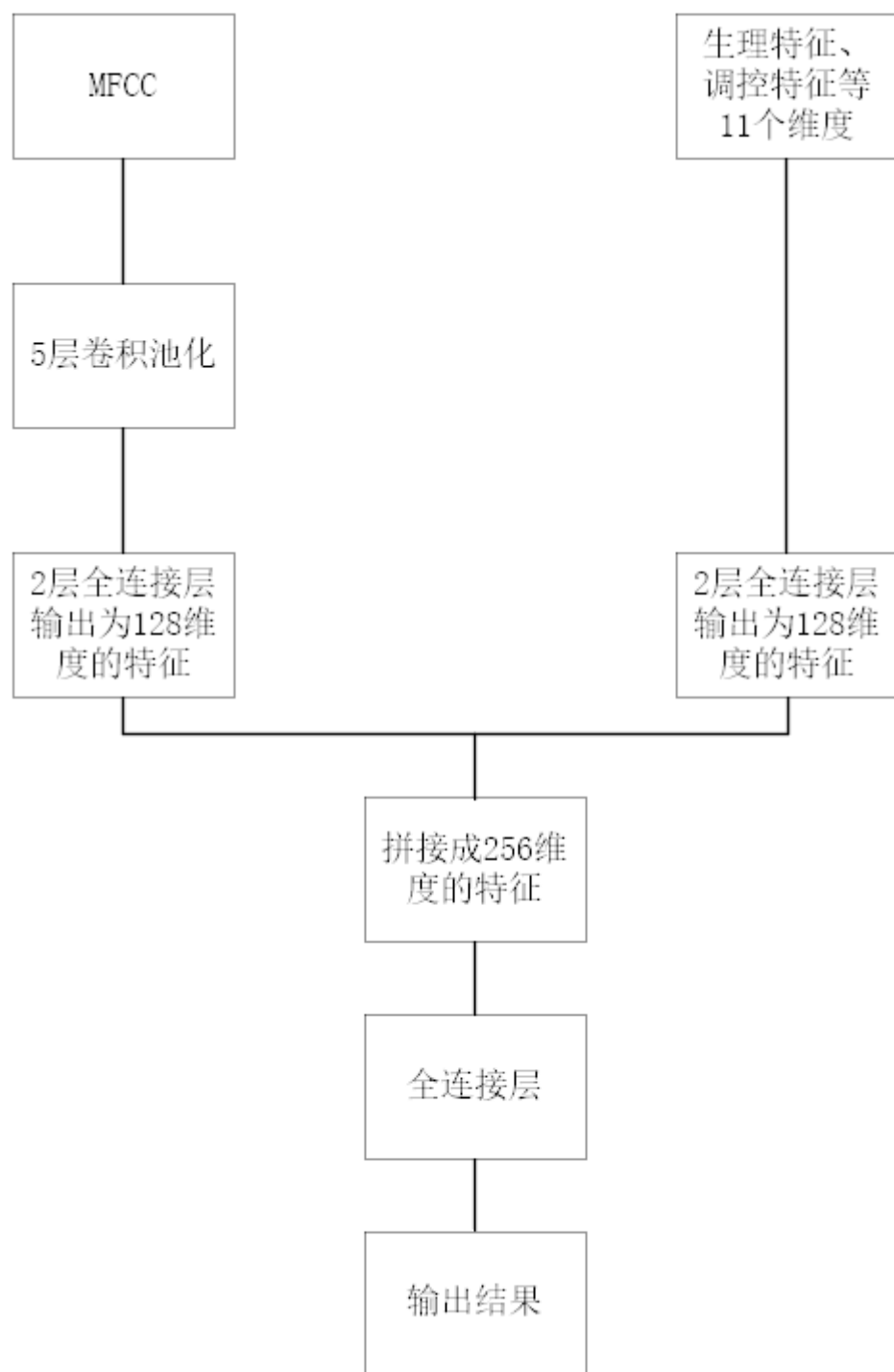
温柔、单纯、孱弱、有力、成熟、高傲、清爽、活力、激情、轻佻、尖细、滑稽、戏谑、冷艳、高冷、沧桑、神圣、正义、低缓、木楞

- 数据准备

利用自己采集的860个短音频进行训练

对每个音频进行最大值为7的label标注

- 特征提取
 - 生理特征
 - 年龄感
 - 性别感
 - 清脆度
 - 发音调控特征
 - 咬字阶段：咬字清晰度、相邻字发音黏着程度；
 - 韵律阶段：平均音高、音高变化、语调的平顺性、停顿频度、语速控制、节奏的稳定性
 - MFCC
- 模型选择
 - 采用多特征结合思想，先使用卷积对MFCC进行特征再提取，然后利用全连接将生理特征、调控特征组合提取，最后利用逻辑回归将拼接起来的特征进行组合得出结果



○ 最终结果

整体平均误差 : 1.036830073627828

| | |
|----|---------------------------|
| 低缓 | loss : 1.7395625419453187 |
| 高冷 | loss : 1.7860195660912903 |
| 冷艳 | loss : 1.014093239664892 |
| 单纯 | loss : 1.2696723516554393 |
| 孱弱 | loss : 1.6063810307783024 |
| 尖细 | loss : 0.9161219021014009 |
| 戏谑 | loss : 0.6128652089974802 |
| 有力 | loss : 1.110721804090578 |

| | |
|----|---------------------------|
| 木楞 | loss : 1.5987329244573887 |
| 正义 | loss : 1.4404893855803216 |
| 沧桑 | loss : 1.360141946073011 |
| 活力 | loss : 0.7689765134076854 |
| 清爽 | loss : 1.535077632810883 |
| 温柔 | loss : 1.932533746203615 |
| 滑稽 | loss : 0.8968183388750208 |
| 激情 | loss : 0.5312550728594192 |
| 神圣 | loss : 1.196321327394942 |
| 轻佻 | loss : 1.5537195912105972 |
| 高傲 | loss : 0.7374732748721218 |
| 成熟 | loss : 1.3160406374418765 |