人工智能引论实践课实验报告

田宇轩 2100013126

实验一　语音性别分类

目录

1 实验目的

本实验任务是根据语音的声学特征，采用机器学习的方法，对语音来源的性别进行分类。实验数据集是基于 3168 个录制自男性/女性说话者的语音样本，使用 R 的 seewave 和 tuneR 包进行声学分析的预处理（分析频率范围为 $0hz - 280hz$，即人发声频率范围）得到的具有 20 维特征的数据，第 21 维以 label(male/female) 标识性别。作为有监督学习中典型的二分类问题，实验将按照　对特征数据的分析与处理、机器学习中几种分类模型的使用与比较、模型的改进　的研究步骤进行。

2 数据的分析与处理

2.1 特征分析

2.1.1 特征数据的分布分析

加载数据集，得到数据集规模：[3168rowsX21columns]，对前 20 列特征及其数值进行统计，见表 1：

feature	meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	modindx
min	0.0394	0.0184	0.0110	0.0002	0.0429	0.0146	0.1417	2.0685	0.7386	0.0
max	0.2511	0.1153	0.2612	0.2473	0.2734	0.2522	34.726	1309.61	0.9820	0.9323
mean	0.1809	0.0571	0.1856	0.1404	0.2247	0.0843	3.1402	36.568	0.8951	0.1738

表 1：各特征的最大值、最小值、平均值

sfm	mode	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange
0.0369	0.0	0.0393	0.0556	0.0098	0.1031	0.0078	0.0049	0.0078	0.0
0.8429	0.28	0.2511	0.2376	0.2041	0.2791	2.9577	0.4590	21.843	21.844
0.4082	0.1653	0.1809	0.1428	0.0368	0.2588	0.8292	0.0526	5.047	0.1737

表 1（续）

从表 1 可以看出，不同特征的数值范围不同。大多数在 0-0.25 之间；特征 mode（模态频率），maxdom（主频最大值），drange（主频范围）等取值处于 0-10 之间；而 kurt（峰度）的取值最为极端，最大值超过 1300。值得注意的是，kurt 的最大值约为 1309.61，而平均值仅为 36.568，最小值为 2.068，怀疑存在小部分数据的干扰。进一步，用 pandas 的 nlargest 函数求出数据中 kurt 从大到小的前 150 个值，发现其数值迅速地从 1309.61 跌落到 85.6176，这为后

续对初始数据进行预处理提供了思路。就均值而言，17 个 feature 的均值处于 0-1.0 之间，此外还有 3.1402，36.568，5.047，均值的分布也是不均一的。

因此，为提高效率和模型的表现，防止个别特征影响过大、极端数据等的干扰，从而保证准确率，对数据进行归一化和标准化是有必要的，尤其是参数化模型如支持向量机、神经网络等。

2.1.2 特征间的关系分析

对于特征之间的关系进行分析，发现有

IQR（分位数范围）=**Q75**（第三分位数）-**Q25**（第一分位数）

dfrange（主频范围）=**maxdom**（主频最大值）-**mindom**（主频最小值）

等式恒成立，表明这两组特征之间具有绝对的线性关系。同时，声学分析得到的 20 维 feature 之间是相互联系和影响的，这都使得特征的独立性假设难以成立，从而让朴素贝叶斯等算法的表现受到影响。

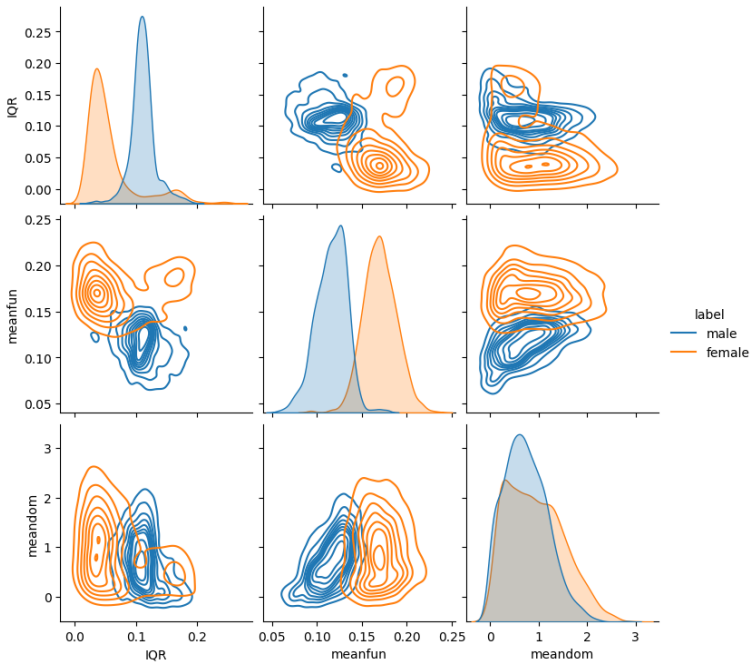

图 1: 用 seaborn 包绘制特征 IQR、meanfun、meandom 的关系图

2.1.3 特征与类别的相关性分析

按 label（性别）将数据划分为两个集合，对 20 个特征，分别绘制出其数值随性别的分布，以直观显示各特征与性别的相关性，如图 2：

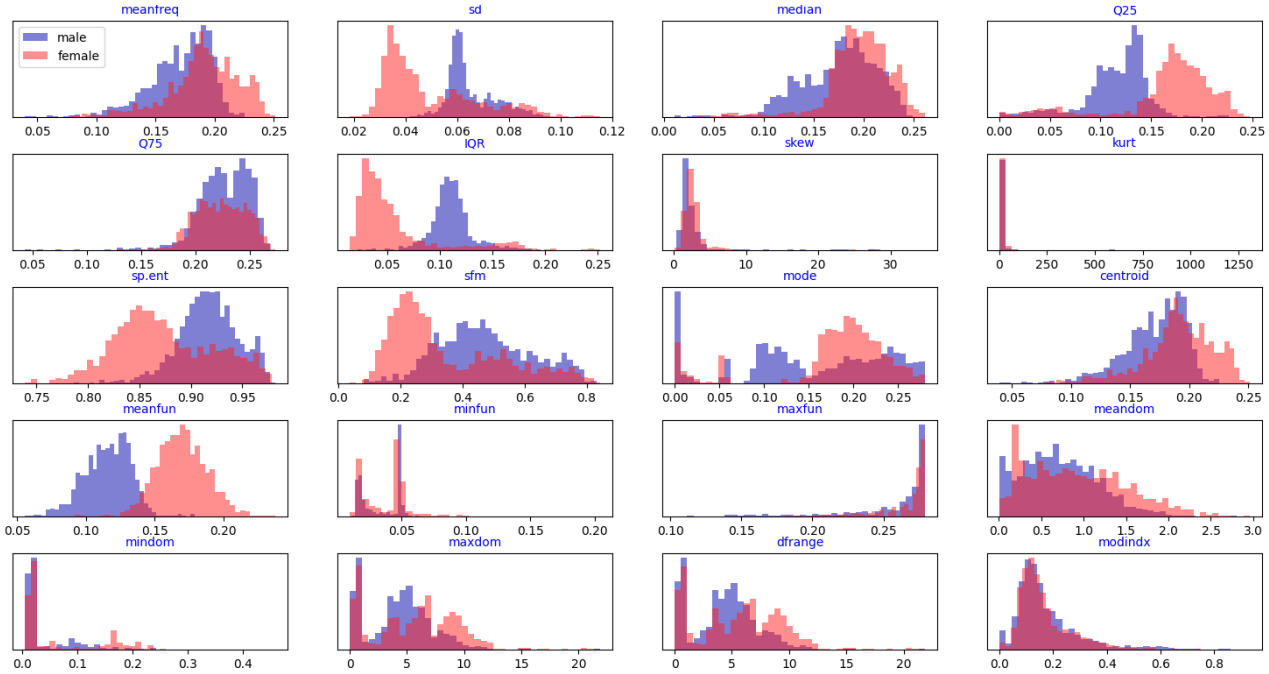

图 2: 不同性别的 feature 数值分布

图像表明，在所有 feature 中 sd（频率的标准偏差）、Q25、IQR、sp.ent（谱熵）、sfm（频谱平坦度）、mode、meanfun（基频均值）随性别的区分，表现出显著差异，而 skew（偏态系数）、minfun（最小基频）、meandom（主频均值）、modinx 的差异不明显。

当基于特征对语音进行分类时，可着重强调/淡化上述特征，挑选一些”好”的特征来训练模型，来减小模型训练时间，也能够提升模型性能。

2.2 特征处理

结合 **2.1 特征分析**的结论，标准归一化能够提高准确率。首先就输入数据进行标准化、归一化对准确率的影响进行实验，得到如表 2、表 3 所示结果：

Model	Decision Tree	Random Forest	Gradient Boosting	Support Vector Machine	Multilayer Perception
Accuracy on training set	1.000	0.998	0.996	0.678	0.950
Accuracy on test set	0.961	0.976	0.975	0.680	0.951

表 2：未进行标准归一化的各模型准确率

Model	Decision Tree	Random Forest	Gradient Boosting	Support Vector Machine	Multilayer Perception
Accuracy on training set	1.000	0.998	0.996	0.985	0.996
Accuracy on test set	0.961	0.976	0.975	0.984	0.983

表 3：标准归一化后各模型准确率

可以看出，使用标准归一化后，模型的准确率有所提升。在决策树以及基于决策树的集成模型，如随机森林、梯度提升树模型上，提升并不明显；在支持向量机、多层感知机模型上，准确率有明显提升，尤其是 SVM。这是因为前者是概率模型（树型模型），不关心特征的值，而关心特征的值和特征之间产生的条件概率；后者是数值化的模型，它们的最优化问题需要标准归一化，类似的，还有 K 最邻近、逻辑回归、K 均值聚类等算法。

3 分类模型

3.1 Decision Tree

3.1.1 模型介绍

与线性模型不同，树型模型是按照一个个特征进行处理，通过根据样本特征在每个节点上的决策，一步步确定其所属类别。树内部的每个节点可看作对一个特征的测试，测试结果产生不同的分支，最终，每个叶节点代表一个类别。树型模型更接近于人的逻辑思维方式，产生的模型（决策树）易于人类理解和解释，并能产生可视化的决策规则，如图 3：

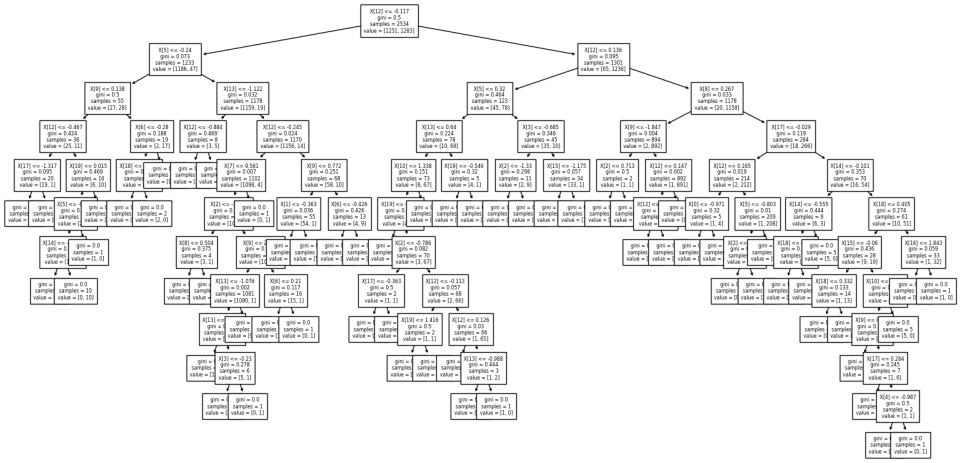

图 3: 本次实验生成的决策树

3.1.2 模型评价与改进

决策树具有不依赖于数据的标准归一化、不错的鲁棒性等优点。但存在缺点：可能陷入局部最优、模型复杂时容易过拟合、对于异或这样的非线性问题不易处理……

为防止过拟合、提高效率，需要进行剪枝 Pruning，包括预剪枝与后剪枝；还可以做正则化 Regularization 改进：对决策树设置约束，限制模型参数，如控制树的最大深度 depth，限制叶节点数量上限，增加分裂一个节点所需的最小样本数/样本数下限，增加分裂节点时依据的最少特征数量……

进一步，在决策树的基础上，许多集成学习技术被开发出来，如：袋法集成学习、随机森林、梯度提升树，对决策树算法进行改进。这是我们下面将要讨论的——

3.2 Random Forests

随机森林建立在决策树的基础上，通过随机选取特征数据，建立多棵树来提高分类的表现 performance，从而防止陷入局部极值问题。随机极度随机树的每棵决策树都是由原始训练样本构建的。在每个测试节点上，每棵树都有一个随机样本，样本中有 k 个特征，每个决策树都必须从这些特征集中选择最佳特征，然后根据一些数学指标 (一般是基尼指数) 来拆分数据。这种随机的特征样本导致多个不相关的决策树的产生

3.2.1 Extra Trees

类似，有极度随机树学习 ExtraTreesClassifier：它将在决策树森林 Forest 中收集的多个去相关决策树的结果聚集起来，输出分类结果。极度随机树比常规随机森林更具随机性（Randomness），因为极度随机树在每个节点分裂或分枝时，随机选择特征子集，并且随即分裂来获取最优的分枝属性和分枝阈值，总体上效果好于随机森林，见表 4：

n_estimators **Accuracy on training/test set**	**n=5**		**n=10**		**n=15**		**n=20**	
Random Forests	0.998	0.976	1.000	0.979	1.000	0.978	1.000	0.976
Extra Trees	1.000	0.972	1.000	0.979	1.000	0.983	1.000	0.984

表 4: 随机森林与极度随机树模型在不同迭代次数下的准确率

3.3 Gradient Boosting

梯度提升，旨在训练模型的过程中，每次迭代都去增加预测结果和真实值相差较大的元素的权重，减小预测结果和真实值相差较小的元素的权重，从而获得更优的预测模型。

3.4 Support Vector Machine

3.4.1 模型介绍

支持向量机解决分类问题可以形象化理解成：按照样本的特征数值将其映射成特征空间中的点集，支持向量机就是在特征空间中寻找满足约束条件的最优分类超平面，将点集一分为二（针对二分类问题）并使到两类的边距最大化。

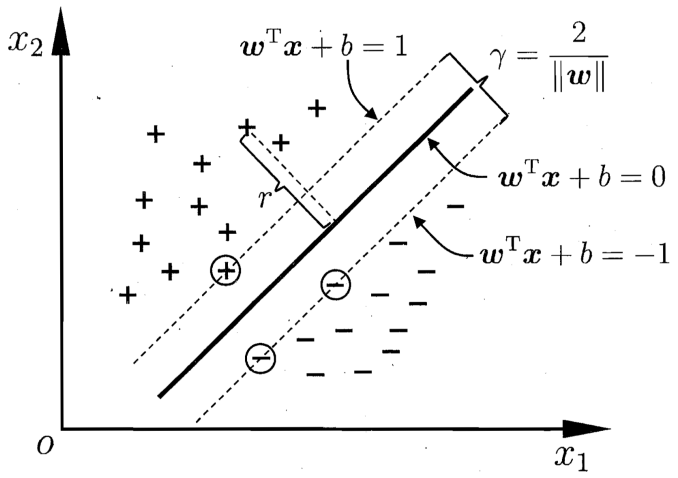

图 4: 二维特征空间中支持向量机所得超平面与支持向量展示

3.4.2 模型评价与改进

支持向量机存在一些问题和局限，比如容易被异常值影响，只对线性可分样本有效等。为处理异常值和非线性数据，可以设置软边界，以容纳小部分异常数据；为正确处理非线性数据，使用核函数 kernel，将非线性可分数据点的特征从相对较低的维度映射到相对较高的维度，便于分类处理，有多项式核，高斯核等

3.5 Multilayer Perception

人工神经网络，以多层感知机为例，通过建立多层神经元的连接模拟脑神经模式。在传统的输入层 InputLayer 和输出层 OutputLayer 之间加入隐藏层 hidden layer。并引入激活函数（非线性函数），实现非线性变换。

4 实验结果分析

在完成针对学习样本的模型训练后，绘制各个模型中的不同 feature 的重要性图进行比较，如图 5：

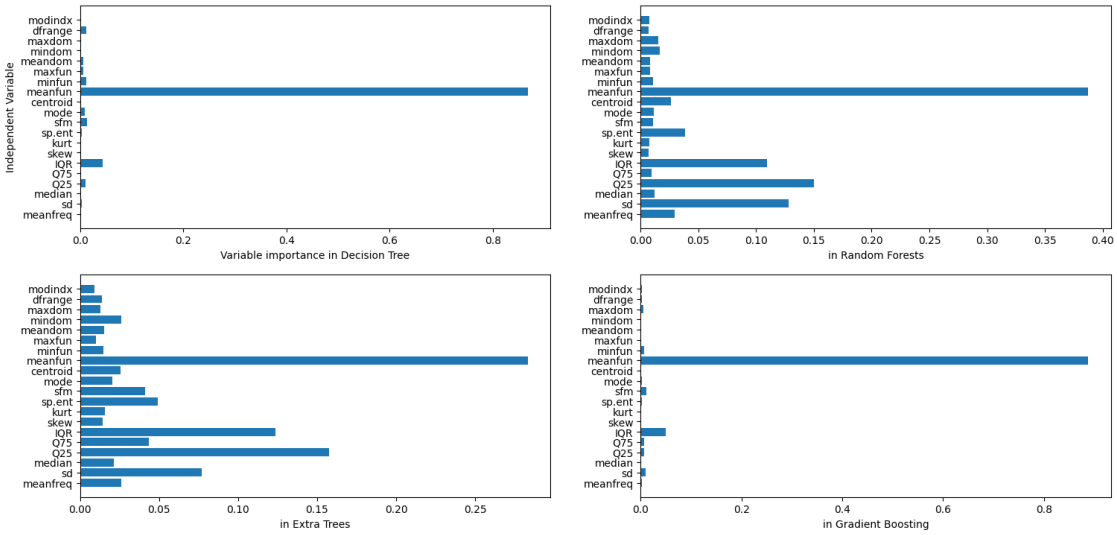

图 5: 四种树型模型中 feature 的重要性图表

从图中可以发现，meanfun 是这四种模型对性别做分类的最主要依据的特征；IQR、Q25、sd 其次，对分类结果有相当的影响；其他特征的重要性较弱，这与我们在 **2.1.3 特征与类别的相关性分析**中的结论很好地符合。同时，在四种模型中的，决策树、梯度提升模型中 meanfun 单个特征重要性最为突出，分类较依赖单个 feature；随机森林、极度随机树对于其他各特征都有所考量。

对于训练出的多层感知机模型，绘制出首层各特征权重矩阵，meanfun 仍是主要影响因素，整体与上述特征重要性分析相符，如图 6：

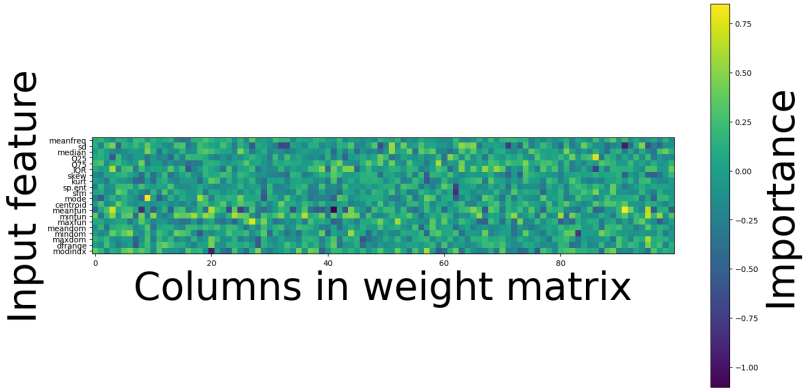

图 6: the heatmap on first layer weights for neural network

5 讨论、总结与思考

从结果可以看出，meanfun（基频均值）是与分类结果相关性最为显著的特征，其次有 IQR（分位数范围）、Q25（第一分位数）、sd（频率的标准偏差）、sp.ent（谱熵）、sfm（频谱平坦度）；与之相对的，skew（偏态系数）、minfun（最小基频）、meandom（主频均值）、modinx（调制指数）、kurt（峰度）对分类影响甚微。我们在实验前的特征数据初步分析（图 2），以及试验后绘制的重要性 importance 图（图 5、图 6），两相印证、共同说明这一点。这也反映实验前的分析是正确且有意义的。

实验也表明，对实验数据的预处理，比如标准化、归一化是相当重要的，能够提高学习效率和准确率，而数值型模型对均值归一化尤其敏感（表 2、表 3）。以上思考，对我们以后更好地完成机器学习任务有指导和借鉴意义。

关于实验模型：就模型分类的过程与依据而言，实验主要采用了树型模型（决策树、随机森林、梯度提升树）和数值模型（支持向量机、多层神经网络）两类机器学习模型；就模型发展而言，神经网络相比传统模型，表现出独有的优越性，而传统机器学习模型在处理此学习样本时也有优良的表现，模型选择与改进措施对准确率的影响是明显的。

关于模型的改进：从分析模型优缺点出发，提出解决方法，进而衍化出新的模型。在异常数值处理、正则化解决过拟合、随机化防止局部极值、预处理不适应的样本、调整参数以更好拟合数据……多方面积累了经验，有助于将来机器学习研究。

关于实验的改进和继续深入：首先，使用到的模型参数大多采用默认参数设置，对于如何调整参数以达到最佳效果可以继续深入实验；其次，筛选部分“好”的特征进行训练的想法没有实施；再者，模型的改进与优化没有进行充分实验，可以更加详细地进行对比与模型升级。

实验二　语音变调

目录

1 实验目的

基频和共振峰频率是区分语音中的元音/浊音的重要特征，通过改变基频和共振峰频率，可以实现语音变调。本实验研究基频和共振峰频率的改变对声音听感（性别、年龄等）的影响，以及参数调整对语音变调效果的影响。

2 实验方法

2.1 人的发声原理

“源-滤波器模型”，将人的发声过程视作声带振动产生的“声门波”（Glottal sound），经过声道、咽腔、口腔、鼻腔等组成的滤波器的调制，最后经唇辐射形成的语音（图 1）。值得注意的是，人的语音还包括一系列无声带振动，仅通过咽、口、鼻动作产生的声音，如鼻音、摩擦音、爆破音等，对于它们，本实验的方法并不适用。我们侧重研究由声带振动激励的脉冲信号经调制形成的语音，它们也是元音、浊辅音最重要的基础和组成部分，改变这部分语音，基本上可以很好地实现语音变调的任务。

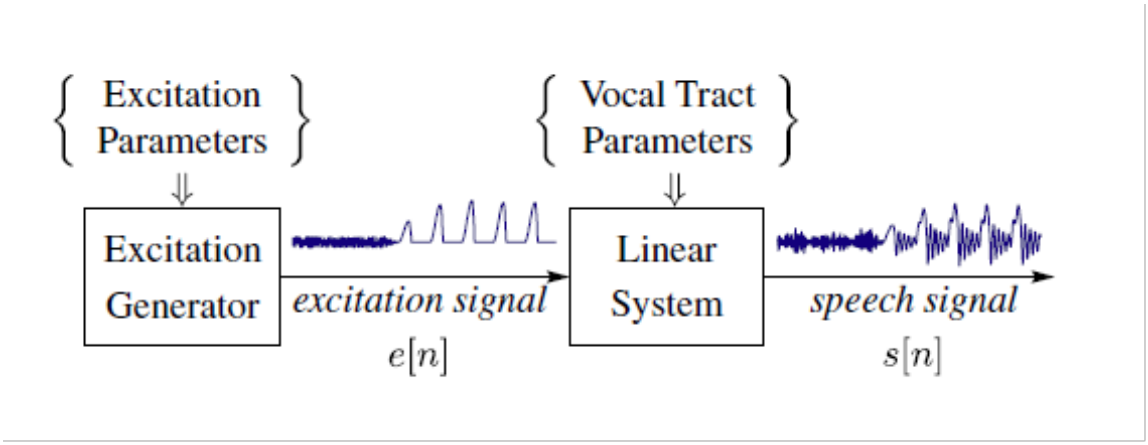

图 1: Source/system model for a speech signal

实现不同音调的语音之间的转换，不仅要改变基频，共振峰频率也应当作出相应的变化。例如，就性别而言，女声的基频一般高于男声（图 2），但同时男女声共振峰也有着差异，因为成年男性的声道一般比女性的声道长，男声的共振峰也相应地低于女声（女声的共振峰通常比男性高 20%）；而儿童的声道更短，基频和共振峰的频率都明显高于成年人。因此，若想通过语音变调进行语音性别的转换，基频与共振峰的频率都要作出改变。

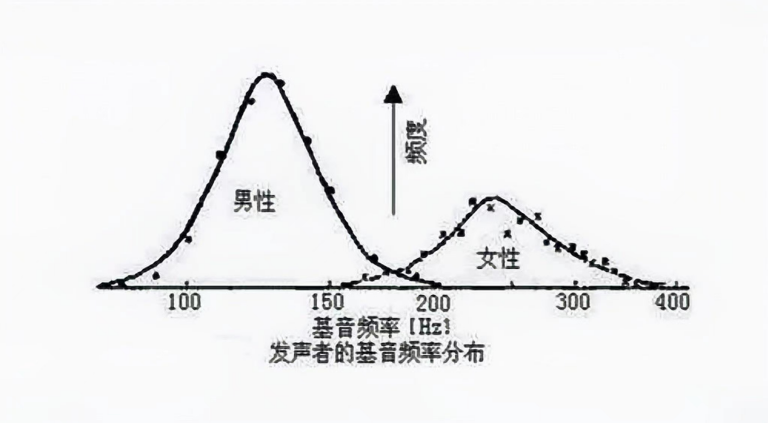

图 2: 男/女声基频分布

2.2 语音变调过程

进行语音变调处理时，假设输入的语音信号是短时平稳的，帧长约 10-30ms，这样在短时内可近似为线性时不变系统。线性预测 LPC 模型，就是在探究声腔（滤波器）的性质，得到对应特征的参数。按照上述帧长分帧后，通过 LPC 分析从预测误差中得到基频、计算出预测系数，进而对共振峰作出估计。用已经提取出来的基频和共振峰频率，按照设置的参数进行变换，求出新的基频，并控制极点的左旋和右旋，调整共振峰频率、合成新的预测系数。最后进行一个反向的过程，控制发出激励脉冲，并添加白噪音，通过滤波器调制最终得到变调后的语音。

以下是 LPC 分析获取基频的原理：，$x(n)$ 为真实信号，$Gu(x)$ 为激励源：

$$x(n) = \sum_{p}^{k=1} a_k x(n-k) + Gu(n) \tag{1}$$

$\hat{x}(n)$ 为预测信号，以线性加权组合近似信号：

$$\hat{x}(n) = \sum_{p}^{k=1} a_k x(n-k) \tag{2}$$

$e(n)$ 为预测误差。根据 $e(n)$ 误差最小化来计算预测系数：

$$e(n) = x(n) - \hat{x} = x(n) - \sum_{p}^{k=1} a_k x(n-k) \tag{3}$$

3 实验过程

1、输入语音样本（收集到的男/女声，童声音频）
2、按照 **2.2 语音变调过程**对样本进行处理，计算基频、预测系数、共振峰频率
3、输入并逐渐调整基频调整倍数和共振峰增加频率
4、计算得到新的基频与共振峰频率，经处理后输出
5、记录输出语音特征，进行分析

4 结果分析

按照如图（图 3）所示的音色与基频、共振峰频率的关系图，对输入样本进行语音变调。发现，对于输入样本，不论 label 是男声/女声，在限制共振峰固定不变（设置共振峰增加频率为 0）的情况下，在基频倍数调整为原来的 0.8-1.4 倍时，语音性别分类没有变化，听感上会变从

低沉逐渐转变为尖锐。超过倍数范围，由于共振峰没有调整，听感上开始脱离人声范围，并且杂音、漏音逐渐明显，缺乏连贯性。

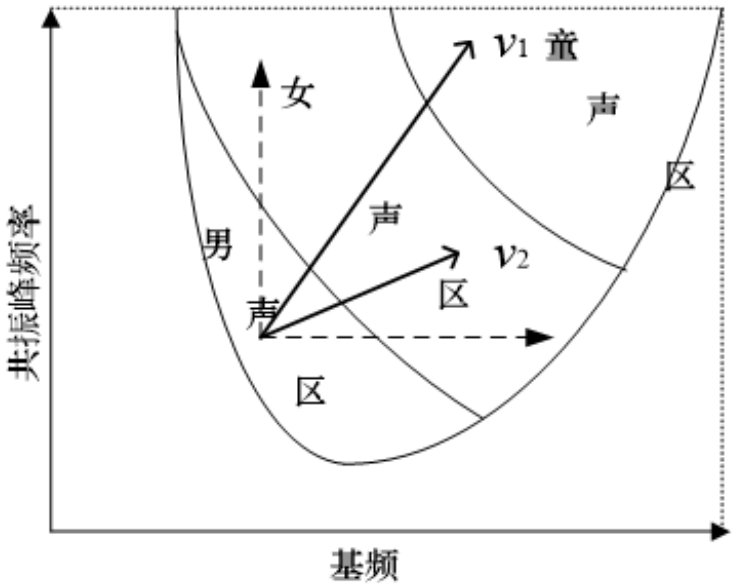

图 3: 男/女声、童声类别随基频、共振峰频率的分布

对于输入的男声样本，在基频调整倍数为 0.6 到 0.8，共振峰增加频率为-100hz 到-50hz 的范围内时，声音能变得低沉并且保持较好的听感，基频越低，声音越低沉。在基频调整倍数从 1 开始增加的过程中，声音先变得细而尖锐；在倍数为 2，接近中性音，难以辨别性别；倍数增加到 2.5 附近时，近似为女声；当基频增加为 3 到 4 倍，声音转变为童声，此时共振峰增加频率为 200hz 以上为佳，过低则声音失真明显；大于 4 倍，则更尖锐而不似人声。

对于输入的女声样本，基频调整倍数小于 1，声音逐渐低沉沙哑，调整倍数为 0.4 到 0.6，声音近似男声；大于 1 倍时，随基频调整倍数的增加，音调逐渐增高，变得尖锐；大于 2 倍，开始向童声转变。

对于输入的童声样本（经挑选的音调较高的音频），将基频和共振峰频率逐渐降低的过程中，确实呈现出先近似于女声，后近似于男声的趋势。但随着基频调整倍数、共振峰频率的降低，失真逐渐明显，听感质量显著下降、出现内容丢失。可见仅通过调整基频、共振峰，对于童声的变调还有所欠缺。

5 讨论、总结与反思

根据对人发声规律的研究得到的“源-滤波器”模型，以及短时平稳假设下，通过 LPC 分析可以较好地得到语音的基频与共振峰频率。对频率进行适当调整后经滤波器输出，可以基本完成语音变调任务。

实验与理论分析共同说明，完成语音变调，基频与共振峰频率都需要进行调整，以符合人发声规律。在固定共振峰的情况下调整基频，会使语音内容损失、声音失真随基频调整幅度的增加而越发明显，不能得到满意的输出。

不论男/女声，基频调整倍数小于 1 并逐渐减小的过程中，声音变得低沉、厚重；基频调整倍数大于 1 并增大的过程中，声音变得尖锐。按照基频、共振峰频率都从小到大的变化，整体呈现出：低沉男声-尖细男声-中性声音-低沉女声-尖锐女声-童声的趋势。

关于实验改进和继续深入：受人力限制，输入样本数量不多，不具有普遍性。对于男/女声、童声的区分，缺乏可量化的标准，由笔者主观推断和判定。综合起来说，对实验数据和结果缺少更加严谨的刻画，主观性较强，有待继续进行充分、严谨的实验。同时，频率变化步长较大，数值与数值范围较粗略。

参考文献

[1] Rabiner,L. R. , Schafer, R. W . Introduction to Digital Speech Processing[J]. Foundations & Trends in Signal Processing, 2007, 1(1–2):1-16.

[2] 宋知用.MATLAB 语音信号分析与合成. 北京. 北京航空航天大学出版社，2017.11

[3]https://blog.csdn.net/qq_43543515/article/details/112505740