李孜涵, 黄玮. 卡通人物合成语音与自然语音的声学对比分析——以喜羊羊和灰太狼为例[J]. 智能计算机与应用,2026,16(6):75-82. DOI:10.20169/j.issn.2095-2163.24102502

卡通人物合成语音与自然语音的声学对比分析——以喜羊羊和灰太狼为例

李孜涵, 黄 玮
(南开大学 文学院, 天津 300071)

摘 要: 本文以喜羊羊和灰太狼的自然语音和合成语音为语料,从谱重心、离散程度等 4 个谱矩参数和基频微扰、振幅微扰等 8 个嗓音参数这 2 个维度,考察自然语音和合成语音的差异。实验表明,在谱矩参数上,喜羊羊和灰太狼 2 个卡通人物的谱重心和离散程度的自然语音的平均值大于合成语音的平均值,偏度和峰度的合成语音的平均值大于自然语音的平均值。配对样本 T 检验显示,2 个卡通人物的自然语音和合成语音的谱重心、离散程度、峰度、偏度均存在统计上的显著差异。在嗓音参数上,2 个卡通人物的自然语音具有更大的 jitter 和 shimmer 平均值、及较小的 HNR 和 CPP 平均值。配对样本 T 检验显示,2 个卡通人物的 jitter、shimmer、HNR、CPP、H1-A2、H1-A3 均具有统计上的显著差异。总体来说,喜羊羊和灰太狼的自然语音和合成语音既有联系又有区别。

关键词: 卡通人物; 合成语音; 谱矩参数; 嗓音参数; 声学分析
中图分类号: TP912.3 **文献标志码**: A **文章编号**: 2095-2163(2026)06-0075-08

Acoustic comparative analysis of cartoon characters′ synthesized speech and natural speech——A case study of Pleasant Goat and Big Big Wolf

LI Zihan, HUANG Wei
(College of Liberal Arts, Nankai University, Tianjin 300071, China)

Abstract: In this paper, the natural and synthetic speech of Pleasant Goat and Big Big Wolf are used as the corpus to investigate the differences between the natural and synthetic speech in two dimensions: four spectral moment parameters such as spectral centre of gravity and degree of dispersion, and eight voice parameters such as fundamental frequency perturbation and amplitude perturbation. The experiments show that, in terms of spectral moment parameters, the mean values of spectral centre of gravity and degree of dispersion of the two cartoon characters, Pleasant Goat and Big Big Wolf, are larger than those of the synthetic speech, and the mean values of skewness and kurtosis of the synthetic speech are larger than those of the natural speech. Paired-sample T-tests shows that there are statistically significant differences between the spectral centre of gravity, degree of dispersion, kurtosis and skewness of the natural and synthetic voices of the two cartoon characters. In terms of voice parameters, the natural speech of the two cartoon characters had larger mean values of jitter and shimmer, and smaller mean values of HNR and CPP. Paired-sample T-tests shows that there are statistically significant differences in jitter, shimmer, HNR, CPP, H1-A2, and H1-A3 between the two cartoon characters. Overall, the natural and synthetic speech of Pleasant Goat and Big Big Wolf are both related and different.

Key words: cartoon characters; synthetic speech; spectral moment parameters; voice parameters; acoustic analysis

0 引 言

如何构建可以像人类一样说话的智能机器是数百年来学者们感兴趣的研究课题。语音合成就是这样一种利用机械或电子手段产生人造语音的技术[1]。人们试图搭建语音合成机器可以追溯到 1779 年,Kratzenstein 构造了一组类似于人类声管模型的声学共振腔[2]。21 世纪以来,语音合成技术不断进步,基于隐马尔可夫模型(HMM)的语音合成系统和基于神经网络的语音合成系统成为研究主流。

随着深度学习的发展,如何传递语音中的情感得到了学者们的关注,其核心在于利用深度学习模

作者简介: 李孜涵(2001—),女,硕士研究生。主要研究方向:实验语音学,语言实验。Email:lizihansmile@163.com; 黄 玮(1997—),男,博士研究生。主要研究方向:实验语音学,语言实验。

收稿日期: 2024-10-25

型和神经网络模拟人类的语音特征和情感表达。比如,Masanobu[3]选择小说、广告、百科全书三个不同领域且具有清晰风格的文本作为语料,通过时长、能量、共振峰、频谱倾斜等参数,发现频谱和韵律参数可反映说话风格的特性。卡通人物是一种风格独特的语音,语气更明显,情感更丰沛,表达更夸张。对卡通人物进行自然语音和合成语音的对比具有一定的探索意义。

对于合成语音的技术方法、特点特征、实际应用、判断检测等方面,学者们已有诸多研究。张立[4]采用不同方法对合成语音与自然语音进行对比。周峻林[5]就伪装语音和真实语音的声学特性差异展开分析,对比两者在基频、音强、频谱上的表现差异。实验发现,伪造语音的共振峰起伏程度、交错数量小于真实语音,可以将共振峰形态特征作为区分伪造语音与真实语音的依据之一。自然语音和合成语音的对比研究较少,尤其是较少关注到两者嗓音参数上的差别。人的语音具有丰富的情感表达,而合成的语音常常较为机械,缺少自然流畅的表达。本文选取情感丰富的卡通人物的声音,进行合成语音与合成语音的声学对比分析。

用于声学分析的参数十分多样。频谱是语音分析的基础,谱矩参数是一种量化频谱差异的有效方法。谱矩分析是将数学中“矩”的概念引入频谱分析之中,一阶矩、二阶矩、三阶矩、四阶矩分别对应谱重心、离散程度、偏度、峰度。冉启斌[6]测量了北京话塞擦音的谱重心、离散程度等数据,来构建声学格局。冯勇强等学者[7]提取频谱均值、标准偏差、偏度、峰度参数,分析成年口吃者在流畅朗读中的塞音表现,口吃组发音部位在这 4 个参数上均呈现显著影响。曾晨刚[8]采用谱重心、标准差等 9 项参数考察塞擦音的声学界标,发现谱重心和偏度可以作为“后”特征的声学界标。胡涵等学者[9]测算谱重心、离散度等参数,观测成人依恋风格对情绪语音声学特征的影响。嗓音参数也被广泛运用在研究之中。蔡青等学者[10]在研究不同元音的嗓音表现时,采用 jitter、shimmer、NNE 等参数进行测算。韩仲明[11]认为 jitter 和 shimmer 是嗓音声学分析的一类重要参数。雷科等学者[12]测算老、中、青三组被试的 jitter、shimmer 等 5 个指标,发现多组指标在年龄、性别等方面存在差异,应多参数采样分析。本文的声学对比分析计划从谱矩参数和嗓音参数两个层面入手,多维度探索合成语音与自然语音的差异表现。

1 材料与方法

1.1 语音材料

为更好地进行卡通人物的合成语音与自然语音的声学对比分析,本文选择了声音特点明显的喜羊羊和灰太狼作为例子,自然语音来自喜羊羊和灰太狼系列中的 7 部电影和 48 集动画片。

首先,本文使用 NVIDIA NeMo 将音频提取出来。为提取到干净自然的卡通人物声音,本文先对音频去除背景声音,并进行降噪。然后,人工将其中清晰的 2 个人物的音频裁剪出来,并进行分类。去除声音模糊、声音过短等无效音频后,共得到喜羊羊的声音文件 952 个,灰太狼的声音文件 1 354 个。

本文采取百度飞桨 AI Studio 进行语音合成工作,AI Studio 是基于百度深度学习开源平台飞桨的人工智能学习工具,该模型合成的音频可具有高音频质量和音色相似度。研究将通过百度飞桨 AI Studio 训练喜羊羊、灰太狼两个语音克隆模型,步骤如下。首先,将原始的喜羊羊和灰太狼的语音文件采样率转换成所需的 24 000 Hz,随后将音频进行检验和预标注,最后分别进行 2 000 轮的微调训练,训练完成以后输入对应文本,进行语音合成。在使用百度飞桨合成之后,本文还使用 so-vits-svc 项目进一步训练喜羊羊和灰太狼的声音模型,以提高合成声音的质量。

最终,本文就 2 个不同的人物:喜羊羊和灰太狼,随机选取 200 句进行语音合成,从而进行自然语音和合成语音的对比。

1.2 参数提取

本文利用多维参数来分析音色之间的区别,主要分为 2 个部分:一部分是谱矩参数,包括谱重心(Center of gravity)、离散程度(Dispersion)、偏度(Skewness)、峰度(Kurtosis);另一部分是嗓音参数,包括基频微扰(jitter)、振幅微扰(shimmer)、谐波噪音比(Harmonics-Noise Ratio,HNR)、倒频谱峰值突出(Cepstral Peak Prominence,CPP)、谐波差(H1-H2)、谐波和共振峰差(H1-A1、H1-A2、H1-A3),其中 H1 表示第一谐波,H2 表示第二谐波,A1、A2、A3 分别表示第一、第二、第三共振峰。

频谱是语音分析的基础,谱矩参数是一种数字量化声音频谱表现的有效方式,谱矩分析是将数学中“矩”的概念引入频谱分析之中,一阶矩、二阶矩、三阶矩、四阶矩分别对应谱重心、离散程度、偏度、峰度。

谱重心是频率成分的重心,是在一定频率范围

内通过能量加权平均的频率，体现频谱能量的集中分布区域。主要体现发音部位，尤其是擦音，擦音的发音部位越靠前，舌腭收紧点的缝隙越小，能量越强，谱重心相应就越高[13]。离散程度反映的是声音的频谱能量分布的差异程度，本质上是计算方差，方差越大，离散程度越大，频率偏离重心程度越大。偏度体现的是声音数据分布的不对称性，偏度正值，则能量向负倾斜，能量集中在低频区域[14]。峰度衡量的是声音数据分布的陡缓程度，表现频谱能力的平坦或是尖锐的状态，峰度数值的大小与频谱能量集中区域的能量变化剧烈程度呈现正相关关系[8]。本文先使用 Praat 软件来标注合成语音和自然语音的有声段并进行人工校对，以确保参数提取的有效性。然后在 Praat 中，利用脚本，生成频谱，并分别提取 4 个参数。

研究音质时需要多参数综合分析，嗓音参数也是对比自然语音和合成语音的重要参数之一。基频微扰（jitter）、振幅微扰（shimmer）和谐噪比（HNR）主要关注的是对嗓音的短时精准控制能力，常常用于嗓音质量的测定。

基频微扰（jitter）和振幅微扰（shimmer）是常用的 2 个短时抖动参数，分别表示基频和振幅的微小变化。基频微扰描述相邻周期之间声波基本频率的变化，与声带振动的不规律性相关。振幅微扰表示相邻周期之间声波振幅的变化，与声带振动的不稳定性相关。数学定义公式如下：

$$\text{Jitter (local)} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |T_i - T_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \quad (1)$$

$$\text{Shimmer (local)} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |A_i - A_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \quad (2)$$

谐波噪音比（HNR）表示谐波总能量与噪音总能量的比值，这是一种检测病态嗓音和测试嗓音嘶哑程度的客观指标[15]。HNR 可以在一定程度上反映闭合气声状况，声门闭合越差、通过声门溢出的气流越多，声门噪声越大，也就是说，声音的噪音成分越多，HNR 值也就越小。杨式鳞[16]在研究中测定了 42 例正常人的谐噪比，结果表明正常男性的谐噪比大致在 8.3~17.0 dB，正常女性的谐噪比大致在 7.0~14.6 dB。HNR 可由如下公式进行计算：

$$\text{HNR} = 10 \times \log_{10} \frac{\text{AC}_V(T)}{\text{AC}_V(0) - \text{AC}_V(T)} \quad (3)$$

针对卡通人物声音情感丰富的特点，本文还使用了倒频谱峰值突出（CPP），该参数指倒频谱峰值与趋势线上直接低于峰值的相应值之间的振幅差，用于测量声波周期性较弱的发声态，反映信号的周期性。CPP 值越高，表明信号的周期性强。计算公式具体如下：

$$\text{CPP}_{\text{dB}} = 20 \times \log_{10} \frac{C(\tau_p)}{\hat{C}(\tau_p)} \quad (4)$$

谐波差可以反映嗓音的紧张程度，应用最广泛的是由 Bickley 引入语言学研究的 H1-H2，这一参数的值与开商相关[17]。孔江平[18]在进行哈尼语松紧元音发声原理研究时，采用多维嗓音分析，测量了第一谐波与第二谐波的能量差和第一谐波与第一共振峰的能量差。孟晓红等学者[19]认为，常用的第一与第二谐波差（H1-H2）最能够区分不同语言中的多种发声态，但在声调调域等方面，H1-H2 不能达到理想区分效果。Whitehead 利用了 H1-A1 来区分吴语中的气化元音和普通元音，H1-A2 对于塞音的分析具有重要意义，Jessen 和 Roux 利用 H1-A3 分析豪萨语的气嗓音[20]。

本文首先分别对喜羊羊和灰太狼的原始语音和合成语音各 200 条，在 Praat 中进行标注，并提取上述谱矩参数和嗓音参数。

2　实验结果

首先，对喜羊羊和灰太狼的声学参数总体情况进行研究，在表格中以“o”表示自然语音，以“t”表示合成语音。

2.1　喜羊羊声学参数

喜羊羊谱矩参数平均值表现情况见表 1，对比小提琴图如图 1 所示。自然语音谱重心和离散程度的平均值大于合成语音，偏度和峰度则是合成语音的平均值大于自然语音。

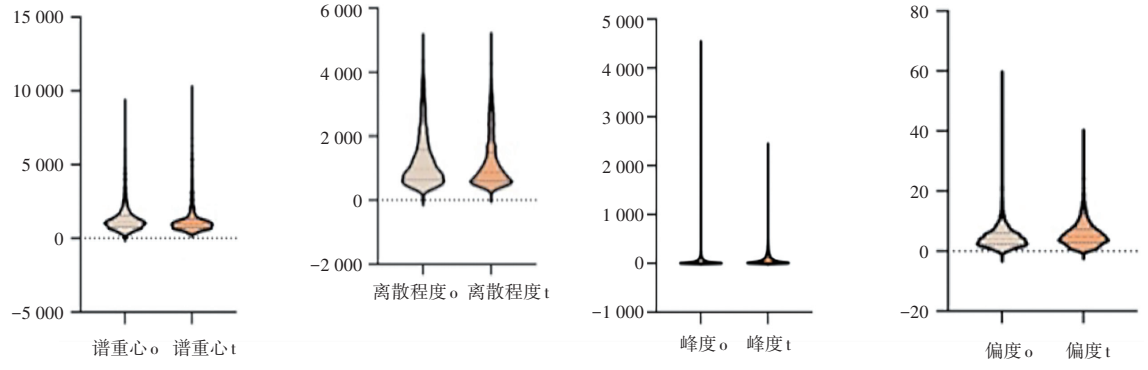
表 1　喜羊羊谱矩参数平均值

Table 1　Mean values of spectral moment parameters of Pleasant Goat

类别	谱重心	离散程度	偏度	峰度
自然语音	1 366.02	1 237.49	4.66	54.91
合成语音	1 274.13	1 173.82	5.55	85.06

标准偏差上，4 个参数的标准偏差均为自然语音大于合成语音，可见自然语音的频谱能量分布范围更广，声音的可调控性更大。喜羊羊 2 种语音偏

度的平均值均为正值,这说明两者的频谱能力集中在低频区域,且喜羊羊合成语音的偏度正值更大,说明合成语音的能量向负倾斜更多。



(a) 对比 1 (b) 对比 2 (c) 对比 3 (d) 对比 4

图 1 喜羊羊自然语音(o)与合成语音(t)谱矩参数对比小提琴图

Fig. 1 Violin plot comparing the spectral moment parameters of Pleasant Goat's natural speech (o) and synthesized speech (t)

在 SPSS 中,分别对合成语音和自然语音的 4 个谱矩参数做了皮尔逊相关性分析。

2 个变量之间的皮尔逊相关系数定义为 2 个变量之间的协方差和标准差的商,常用希腊小写字母 ρ 作为代表符号。数学定义公式为:

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} \tag{5}$$

估算样本的协方差和标准差,从而得到皮尔逊相关系数,这经常用英文小写字母 r 代表,范围在-1 到 1 之间,r 值越接近 1 或-1、表示相关性越强,接近 0 表示相关性弱。数学定义公式为:

$$r=\frac{\sum_{i=1}^{n}((X_i-\bar{X})(Y_i-\bar{Y}))}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{6}$$

显著性检验采用 t 分布,通过查阅 t 分布表,可以得到对应的 p 值,从而判断相关系数的显著性,t 统计量的计算公式如下:

$$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{7}$$

结果显示,喜羊羊的自然语音和合成语音的谱重心、离散程度、偏度、峰度均呈现显著的正相关关系($r=0.276,p<0.01$;$r=0.334,p<0.01$;$r=0.145,p<0.01$;$r=0.126,p<0.01$)。从内部来说,喜羊羊自然语音的谱重心和自然语音的偏度($r=-0.550,p<0.01$)、自然语音的离散程度和自然语音的偏度($r=-0.602,p<0.01$)、合成语音的偏度和合成语音的峰度($r=0.894,p<0.01$)、合成语音的离散程度和合成语音的偏度($r=-0.632,p<0.01$)呈现出强相关关系。

喜羊羊噪音参数总体情况见表 2。在嗓音参数上,jitter 和 shimmer 的自然语音的平均值大于合成语音,HNR 和 CPP 的自然语音的平均值小于合成语音。H1-H2、H1-A1、H1-A2、H1-A3 四组均为合成语音的值大于自然语音的值。从标准偏差来看,除了 H1-A1 之外,jitter、shimmer、HNR、CPP、H1-H2、H1-A2、H1-A3 的自然语音的标准偏差要大于合成语音的标准偏差。这说明自然语音的 jitter 和 shimmer 值较高但分布较为分散,合成语音的 HNR 和 CPP 值分布较高且分布较为集中。总体来看,除 H1-A1 外,自然语音分布均比合成语音的分散范围大。

研究中利用 SPSS,对每个嗓音参数对应的自然语音和合成语音进行对照皮尔逊相关性分析。8 组嗓音参数中,jitter、HNR、CPP、H1-A2 的自然语音与合成语音之间存在显著的正相关关系($r=0.321,p<0.01$;$r=0.354,p<0.01$;$r=0.331,p<0.01$;$r=0.203,p<0.01$)。另外,又分别对自然语音内部和合成语音内部进行了皮尔逊相关性分析。结果显示,自然语音的 jitter 和自然语音的 shimmer ($r=0.552,p<0.01$)、自然语音的 jitter 和自然语音的 HNR ($r=-0.519,p<0.01$)、合成语音的 jitter 和合成语音的 shimmer ($r=0.596,p<0.01$)、合成语音的 jitter 和合成语音的 HNR、CPP ($r=-0.673,p<0.01$;$r=-0.377,p<0.01$) 呈现强相关关系。两者的 jitter 和 shimmer、HNR 之间的关系十分密切。

配对样本 T 检验是一种有效的比较 2 个相关参数的数据检验方法,可以推断 2 个相关样本总体上是否存在显著性差异。本文在 SPSS 中采用配对样本 T 检验的方法,对各自自然语音和合成语音的声学参数和嗓音参数进行配对检验。

喜羊羊声学参数的配对样本 T 检验结果见表 3，除 H1−H2 外，其他参数在自然语音和合成语音的比较上均具有显著性差异（$p < 0.01$）。说明自然语音和合成语音在频谱上是可以进行区分的，存在声学差异。

表 2　喜羊羊嗓音参数总体情况

Table 2　Overall voice parameters of Pleasant Goat

嗓音参数	声音来源	平均值	标准偏差	最小值	最大值
jitter	自然语音	0.021 51	0.005 59	0.005 90	0.039 18
	合成语音	0.016 66	0.005 06	0.007 41	0.046 68
shimmer	自然语音	0.108 87	0.021 21	0.039 47	0.184 20
	合成语音	0.093 77	0.014 98	0.060 58	0.143 29
HNR	自然语音	5.421 26	1.871 75	0.874 54	14.466 05
	合成语音	7.902 16	1.689 33	2.798 27	12.324 86
CPP	自然语音	28.429 75	3.677 03	19.370 71	37.543 92
	合成语音	32.801 66	2.825 31	24.970 51	39.195 76
H1−H2	自然语音	0.344 81	6.418 51	−32.323 32	21.794 31
	合成语音	1.580 32	3.988 87	−25.065 12	13.833 12
H1−A1	自然语音	−0.135 00	7.050 38	−34.916 69	16.938 78
	合成语音	4.989 67	7.080 37	−17.647 66	24.348 78
H1−A2	自然语音	6.942 77	7.941 52	−27.085 34	26.253 94
	合成语音	12.251 87	7.040 57	−18.498 59	26.351 51
H1−A3	自然语音	12.113 89	6.491 30	−13.560 41	28.670 59
	合成语音	17.774 14	5.735 62	−7.169 54	29.823 60

表 3　喜羊羊声学参数配对样本 T 检验结果

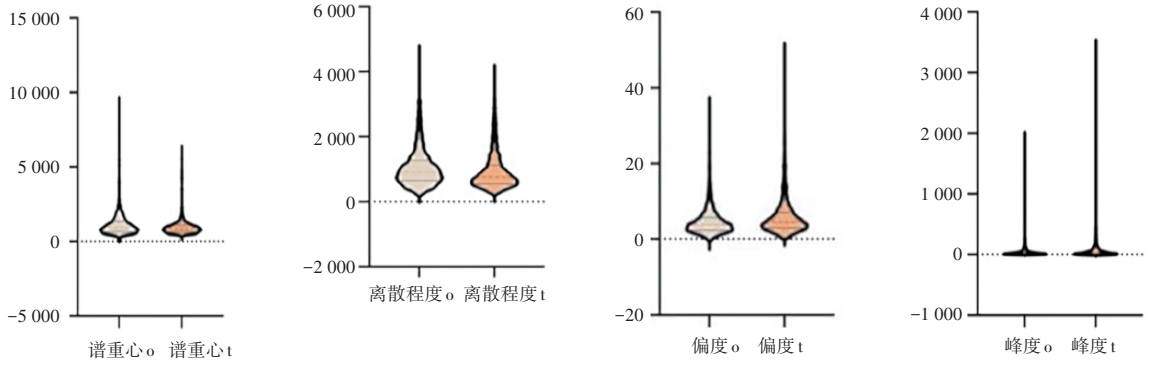
Table 3　Acoustic parameters paired samples T−Test results of Pleasant Goat

参数配对	平均值	标准偏差	t	自由度	Sig.（双尾）
谱重心 o−谱重心 t	91.898 000	1 260.899 000	4.610	3 999	<0.001
离散程度 o−离散程度 t	63.675 000	930.574 000	4.328	3 999	<0.001
偏度 o−偏度 t	−0.891 521	4.979 516	−11.323	3 999	<0.001
峰度 o−峰度 t	−30.153 870	209.690 760	−9.095	3 999	<0.001
jittero−jittert	0.004 850	0.006 220	11.016	199	<0.001
shimmero−shimmert	0.015 100	0.025 260	8.454	199	<0.001
HNRo−HNRt	−2.480 910	2.029 630	−17.287	199	<0.001
CPPo−CPPt	−4.371 910	3.825 570	−16.162	199	<0.001
H1−H2o−H1−H2t	−1.235 510	7.329 200	−2.384	199	0.018
H1−A1o−H1−A1t	−5.124 670	9.404 740	−7.706	199	<0.001
H1−A2o−H1−A2t	−5.309 100	9.484 800	−7.916	199	<0.001
H1−A3o−H1−A3t	−5.660 250	8.433 090	−9.492	199	<0.001

2.2　灰太狼声学参数

谱矩参数结果显示，在平均值上，谱重心和离散程度的自然语音大于合成语音，偏度和峰度合成语音大于自然语音，这与喜羊羊的声音表现一致。在标准偏差上，谱重心和离散程度为自然语音大于合成语音，偏度和峰度为合成语音大于自然语音。可见，自然语音的谱重心和离散程度分布重心高且分布较为分散，自然语音的偏度和峰度分布值小且分布较为集中。

如图 2 所示，灰太狼 4 个参数的平均值相差不大，但分布差异较大，自然语音的谱重心和离散程度分布范围均大于合成语音，自然语音的偏度和峰度的分布范围均小于合成语音。灰太狼自然语音和合成语音的峰度均为正值，且合成语音的正值更大，合成语音频谱能量更加集中在低频区域。



(a) 对比 1 (b) 对比 2 (c) 对比 3 (d) 对比 4

图 2 灰太狼自然语音(o)与合成语音(t)谱矩参数对比小提琴图

Fig. 2 Violin plot comparing spectral moment parameters of Big Big Wolf's natural speech (o) and synthesized speech (t)

研究中分别对灰太狼两种语音的 4 个谱矩参数做了皮尔逊相关性分析。结果显示,自然语音和合成语音的谱重心、离散程度、偏度、峰度之间均呈现显著性($r = 0.187, p < 0.01$; $r = 0.266, p < 0.01$; $r = 0.206, p < 0.01$; $r = 0.153, p < 0.01$)。在自然语音内部,谱重心和离散程度呈现显著的正相关关系($r = 0.725, p < 0.01$),谱重心和偏度、峰度呈现显著的负相关关系($r = -0.542$, $p < 0.01$; $r = -0.047$, $p < 0.001$)。偏度和峰度呈现显著的正相关关系($r = 0.891, p < 0.01$)。离散程度和偏度、峰度呈现显著的负相关关系($r = -0.596, p < 0.01$; $r = -0.413, p < 0.001$)。在合成语音内部,谱重心和离散程度呈现显著的正相关关系($r = 0.710, p < 0.01$),谱重心与偏度、峰度呈现显著的负相关关系($r = -0.458, p < 0.01$; $r = -0.281, p < 0.001$),离散程度和偏度、峰度呈现显著的负相关关系($r = -0.598, p < 0.01$; $r = -0.409, p < 0.001$)。

灰太狼噪音参数总体情况见表 4。在嗓音参数上,jitter 和 shimmer 的自然语音的平均值大于合成语音,说明灰太狼自然语音中的声音信号变化多于合成语音。HNR 和 CPP 的自然语音的平均值小于合成语音的平均值,说明自然语音的声门噪音更多。在标准偏差上,灰太狼的 jitter、shimmer、HNR 和 CPP 的自然语音的标准偏差要大于合成语音的标准偏差。其中,自然语音的 jitter 和 shimmer 数值较大且分布较为分散,合成语音的 HNR 和 CPP 数值较大且分布集中。从谐波差和共振峰差来看,灰太狼自然语音的数值分布更加广泛。总体来看,灰太狼的自然语音的 jitter、shimmer、HNR 和 CPP 数值分布均比合成语音的分散范围更大。

表 4 灰太狼嗓音参数总体情况

Table 4 Overall voice parameters of Big Big Wolf

参数	声音来源	平均值	标准偏差	最小值	最大值
jitter	自然语音	0.026 21	0.006 05	0.013 19	0.046 58
	合成语音	0.015 89	0.002 74	0.010 41	0.025 50
shimmer	自然语音	0.123 90	0.022 10	0.070 75	0.193 54
	合成语音	0.101 29	0.012 80	0.067 30	0.133 71
HNR	自然语音	4.026 39	1.878 15	−0.703 95	9.997 62
	合成语音	7.010 76	1.107 63	4.035 95	10.007 08
CPP	自然语音	27.231 06	3.044 35	20.795 08	36.153 61
	合成语音	33.284 71	2.832 05	25.930 61	39.728 32
H1−H2	自然语音	−4.182 02	6.524 93	−31.456 35	13.128 31
	合成语音	−5.379 38	3.864 53	−28.922 78	13.046 41
H1−A1	自然语音	−2.659 53	7.490 58	−34.610 41	17.284 34
	合成语音	−2.700 72	6.644 40	−35.953 55	16.045 17
H1−A2	自然语音	4.537 00	8.106 74	−29.010 78	21.064 22
	合成语音	6.866 69	6.961 22	−32.511 28	19.400 53
H1−A3	自然语音	7.508 70	7.478 41	−28.308 64	24.069 97
	合成语音	10.778 68	4.986 74	−23.416 05	22.055 69

然后，将 2 种语音的 8 个嗓音参数两两对照，结果发现，8 组嗓音参数内部均呈现正相关关系。其中，jitter、shimmer、HNR、H1-A1、H1-A2 有显著性表现（$r = 0.165, p < 0.05$；$r = 0.309, p < 0.01$；$r = 0.302, p < 0.01$；$r = 0.140, p < 0.05$；$r = 0.250, p < 0.01$）。

配对样本 T 检验（结果见表 5）显示，灰太狼自然语音和合成语音的谱重心、离散程度、偏度、峰度、jitter、shimmer、HNR、CPP 均呈现统计学意义上的显著性差异（$p < 0.01$）。这说明，灰太狼自然语音和合成语音在声学上是可以区分的。

表 5　灰太狼声学参数配对样本 T 检验

Table 5　Acoustic parameters paired samples T-test results of Big Big Wolf

参数配对	平均值	标准偏差	t	自由度	Sig.（双尾）
谱重心 o-谱重心 t	122.504 000	970.259 000	7.985	3 999	<0.001
离散程度 o-离散程度 t	117.616 000	679.205 000	10.952	3 999	<0.001
偏度 o-偏度 t	-1.183 091	5.151 420	-14.525	3 999	<0.001
峰度 o-峰度 t	-45.855 244	200.054 249	-14.497	3 999	<0.001
jittero-jittert	0.010 320	0.006 210	23.501	199	<0.001
shimmero-shimmert	0.022 620	0.021 850	14.640	199	<0.001
HNRo-HNRt	-2.984 360	1.870 020	-22.569	199	<0.001
CPPo-CPPt	-6.053 650	3.998 800	-21.409	199	<0.001
H1-H2o-H1-H2t	1.197 360	7.420 160	2.282	199	0.024
H1-A1o-H1-A1t	0.041 190	9.290 760	0.063	199	0.95
H1-A2o-H1-A2t	-2.329 690	9.270 160	-3.554	199	<0.001
H1-A3o-H1-A3t	-3.269 980	8.289 890	-5.578	199	<0.001

3　讨论

本文通过提取谱矩参数和嗓音参数来观测 2 位卡通人物的自然语音和合成语音的特点。

3.1　表现一致性

本文结果表现出一定的一致性，即 2 个卡通人物的谱重心和离散程度的自然语音的平均值大于合成语音的平均值，偏度和峰度合成语音的平均值大于自然语音的平均值。这说明自然语音更加明亮，合成语音带有沙哑的音色特点，这与合成语音的自然度有关。但是合成语音较自然语音偏离程度稍小，说明合成语音与自然语音的相似度较高，合成效果较好。从分布差异来看，两者的谱重心和离散程度的自然语音数值分布比合成语音的分布更加广泛。从频谱能量集中分布来看，喜羊羊和灰太狼的 2 种语音的峰度均为正值，且均表现为合成语音的正值更大，这说明两者的合成语音频谱能量更加倾向于集中在低频区域。

谐波噪音比（HNR）体现声音中的噪音成分。喜羊羊和灰太狼的自然语音的 HNR 都是小于合成语音的，自然语音由于受到录制环境、背景音乐、他人声音等因素的影响，自然语音的噪音大于合成语音的噪音。喜羊羊和灰太狼自然语音的 CPP 值较小，其声音中气化的语音信号多，这也是合成语音较难克隆自然语音的原因之一，自然语音的情感更加丰富，语气词使用更加流畅。

多种因素会对嗓音产生影响。声带干燥与否会影响嗓音的稳定性，干燥寒冷的空气环境往往导致 jitter 等数据的增大[21]，声带干燥发出的嗓音，jitter 和 shimmer 较正常嗓音有明显的增高[22]。本文数据显示，合成语音的 jitter、shimmer 的值较小，HNR 的值较大，说明合成语音的嗓音表现较好。自然语音往往受到发音人自身嗓音条件、录制环境的影响，尤其是卡通人物的声音需要刻意变化，这些因素可能导致了本文自然语音嗓音变化的不稳定性有所上升。而合成语音是通过 AI Studio 训练合成的，其受外界环境等因素影响较小，表现较为稳定。

通过皮尔逊相关性分析，更加清楚地看到了自然语音和合成语音的谱矩参数之间的相关性。谱重心、离散程度这 2 个参数多呈正相关变化，偏度、峰度之间多呈正相关变化，而 2 组参数之间呈现的是负相关关系。

3.2　语音差异性

通过数据分析，发现 2 位卡通人物自然语音和合成语音之间存在一定的差距。

配对样本 T 检验显示自然语音和合成语音之

间的差距。2 个卡通人物的自然语音和合成语音的谱重心、离散程度、峰度、偏度均存在显著差异（$p < 0.01$）。说明 2 种语音在声音的明亮沙哑、声音能量的分布区域等方面是有区别、但也有联系的。

在嗓音参数的皮尔逊相似性的检验上，2 位卡通人物的表现各异。喜羊羊 4 个嗓音参数整体表现较为稳健，具有统计意义上的显著性差异，而灰太狼的 4 个嗓音参数表现显著性不是很强。在配对样本 T 检验上，2 位卡通人物的 jitter、shimmer、HNR、CPP、H1-A2、H1-A3 均具有统计上的显著差异（$p < 0.01$）。

4 结束语

本文使用了喜羊羊和灰太狼两个卡通人物作为个例研究对象，研究范围较小，以后可以扩大研究对象，选择更多的具有代表性的卡通人物，通过更多的数据观测其合成语音与自然语音的差异。

另外，喜羊羊在动画片和电影中表现为男性角色，但是由女演员进行配音，那么性别是否能对自然语音和合成语音的关系产生一定影响仍是领域中备受关注的研究方向。以往研究中，学者们普遍认为男女语音存在差异。Trudgill[23]认为，男性和女性在语言上存在差异。由于声带等生理特征的显著差别，男性女性的基频差异很大。王治平等学者[24]对基音参数的性别差异进行了分析，并对基音参数进行基于性别差异的规整。毛佩文[25]从性别差异的角度，提取时长和基频曲线，发现女性话语包含的情绪比男性更容易被感知。但是，也有学者的声学实验显示，不同性别在嗓音参数上无显著差异。如，于萍等学者[26]对比分析嗓音障碍患者和嗓音正常者两组的 jitter、shimmer、HNR 等参数在性别中的差异性，结果显示没有显著性差异。易彬等学者[27]在研究吴语方言嗓音障碍指数正常值（DSI）时，发现性别对 DSI 正常值范围无显著影响。是否应该根据不同性别建立不同的嗓音检测模式，性别能否对合成语音产生影响等问题仍需后续进一步的探索研究。

参考文献

[1] 吕士楠，初敏，许洁萍，等. 汉语语音合成：原理和技术[M]. 北京：科学出版社，2012.

[2] LAWRENCE R，RABINER R，SCHAFER W . 数字语音处理理论与应用[M]. 刘加，译. 北京：电子工业出版社，2016.

[3] MASANOBU A. 说话风格：使用文语转换系统进行统计分析和合成[M]. 蔡莲红，译. 北京：机械工业出版社，2005.

[4] 张立. 计算机合成语音与自然语音鉴别技术的研究[D]. 宁波：宁波大学，2017.

[5] 周峻林. 基于声学特性的伪造语音识别方法研究[D]. 北京：中国人民公安大学，2023.

[6] 冉启斌. 北京话塞擦音的声学格局分析[J]. 中国语文，2017(4)：459-467.

[7] 冯勇强，严芊，高兴龙，等. 成年口吃者流利朗读中塞音的声学分析[J]. 声学学报，2013，38(4)：509-516.

[8] 曾晨刚. 普通话塞擦音声学界标研究[J]. 语言科学，2019，18(5)：488-502.

[9] 胡涵，顾文涛. 个体依恋风格对亲密话语韵律及嗓音特征的影响[J]. 声学学报，2022，47(2)：276-286.

[10] 蔡青，陶泽璋，杨强. 发不同元音时嗓音声学分析参数的比较[J]. 临床耳鼻咽喉科杂志，2001(4)：167-168.

[11] 韩仲明. 嗓音声学检测分析[J]. 中国耳鼻咽喉头颈外科，2006(5)：351-353.

[12] 雷科，杨旭，沈建中，等. 成人嗓音声学参数正常参考值的研究[J]. 临床耳鼻咽喉科杂志，2000(6)：255-257.

[13] 郑玉玲. 试论语音声学参数[C]//第九届中国语音学学术会议论文集. 北京：中国社科院民族学与人类学所，2010：1-7.

[14] 马富榆. 天津津南方言擦音的声学特性研究[J]. 现代语言学，2023，11(9)：3967-3975.

[15] 叶志腾. 应用 Praat 软件分析成人嗓音声学参数[D]. 福州：福建医科大学，2009.

[16] 杨式麟. 嗓音的声学检测 [J]. 听力学及言语疾病杂志，2001(4)：255-258.

[17] KIRK P，LADEFOGED P，LADEFOGED J. Using a spectrograph for measures of phonation types in natural language [J]. UCLA Working Papers in Phonetics，1984，59：102-113.

[18] 孔江平. 哈尼语发声类型声学研究及音质概念的讨论[J]. 民族语文，1996(1)：40-46.

[19] 孟晓红，张梦翰. 发声态 45 种声学特征的综合效应分析[J]. 南开语言学刊，2017(2)：32-38.

[20] 熊睿. 祁阳方言塞音声母发声态研究[D]. 长沙：湖南大学，2015.

[21] 冉启斌. 嗓音的南北差异与汉语声调产生的地域先后[J]. 语言研究，2020，40(4)：46-53.

[22] 李暐，屈季宁，周涛，等. 声带干燥对嗓音基频微扰和振幅微扰的影响[J]. 听力学及言语疾病杂志，2013，21(3)：252-254.

[23] TRUDGILL P. Sociolinguistics：An introduction[M]. London：Penguin Books，2001.

[24] 王治平，赵力，邹采荣. 基于基音参数规整及统计分布模型距离的语音情感识别[J]. 声学学报，2006(1)：28-34.

[25] 毛佩文. 基于性别差异的汉语情绪语音韵律分析[J]. 文教资料，2017(Z2)：31-34.

[26] 于萍，韩冰，杨伟炎，等. 不同性别嗓音客观检测参数的对比分析[J]. 听力学及言语疾病杂志，2004(6)：390-392.

[27] 易彬，KIM H，高少华，等. 上海（吴语）方言区正常青年嗓音障碍指数正常值及性别差异研究[J]. 听力学及言语疾病杂志，2021，29(5)：487-490.