Artificial Intelligence Speech Naturalness Evaluation and Improvement Strategies

Lingcong Zhang[1], Mingge Zhang[2, *]

[1]School of Journalism and Communication, Jinan University, Guangzhou, Guangdong, China

[2]School of Literature and Communication, Hebei Normal University for Nationalities, Chengde Hebei, China

Abstract

Artificial intelligence speech has been widely used in various fields of society, but the evaluation method of its naturalness has not been unified. Facing the problems of the existing evaluation methods, this paper attempts to conduct a hearing experiment on the application of artificial intelligence speech in documentaries and subjects with high variable frequency were interviewed, and the three-level listening discrimination standard was proposed under the theoretical framework of phonetics and broadcasting, and the spectral characteristics were analyzed. The research results show that the three-level auditory discrimination standard can test the prosody and fluency of artificial intelligence speech to a certain extent. Making up for the lack of fundamental frequency value of artificial intelligence speech, improving the coordination degree of artificial intelligence language phonological ratio, and refining the fineness of tone value combination have become the breakthrough points for improving the naturalness of artificial intelligent speech.

Keywords

Artificial intelligence speech; Naturalness; Experimental method; Standard of listening and discrimination; Strategy.

1. Introduction

In December 2018, the Central Economic Work Conference of the Communist Party of China defined artificial intelligence as "new infrastructure construction" [1]. Artificial intelligence has technical advantages in speech synthesis, image recognition, data processing, etc., and has been widely used in many fields of social life, with good development prospects [2], and has become one of the important research topics [3].

At present, artificial intelligence speech research mainly focuses on media production [4], media consumption [5], media ethics [6], communication effect [7], synthesis technology [8], etc. improving the naturalness and expressiveness of artificial intelligence speech has become a speech synthesis technology. The problem to be solved urgently [9], the WaveNet model fused with phase information avoids the overlapping effect of STFT method synthesis, and greatly improves the naturalness of artificial intelligence speech [10], the use of ACGAN to improve the benchmark model scheme can effectively improve artificial intelligence to improve the sound quality of speech, improve the quality of hearing [11]. However, there is currently no objective evaluation system for testing AI voice quality [12]. MOS evaluation is a commonly used subjective evaluation standard, emphasizing the subjective feelings of the evaluator [13] but the subjective differences of evaluation subjects cause the uneven problem of AI voice quality [14].

Lakoff's study of American women's language found that women have higher speech sensitivity [15], Cassidy JW and Ditty KM measured cochlear sensitivity by the instantaneous arousal cochlear discharge hearing screening procedure, corroborating the finding that women have higher hearing sensitivity than men [16]. Domestic research of Yang Yuejun and Wang Dongbo of the Mandarin proficiency test found that the tester's listening ability decreased with age [17]. Perceptual Assimilation Model proposed by Best emphasizes that the fundamental of speech acquisition lies in perception, and the auditory recognition and perception of second language speech is affected by the acoustic characteristics of the second language speech and its similarity with the native language [18].

The above research proves to a certain extent that the subjects of auditory discrimination are different, and the results become the logical starting point of this experiment. This paper attempts to explore new subjective evaluation methods from the perspective of linguistics to compensate for the differences in evaluation subjects and propose corresponding strategies for the optimization of the current artificial intelligence speech naturalness.

2. Experimental Design

2.1. The purpose of the Experiment

The experiment firstly verifies whether the differences in subjectivity affecting the results of debate and listening are affected by three factors: gender, majors studied, and level of Putonghua. Secondly, conduct in-depth interviews with high-discrimination listening subjects, revise the subjective listening standards, test the artificial intelligence voice, and propose a strategy for improving the naturalness of artificial intelligence voice.

2.2. Experimental Subjects

A total of 150 subjects (75 males and 75 females) native speakers of Chinese participated in the experiment. They came from broadcasting and hosting art majors (60 people), Chinese language and literature majors (30 people), and journalism majors in a college of Hebei Province. (30 people) and music majors (30 people), all the subjects are from the northern dialect area, without language, hearing, neurological damage, can participate in this experiment.

2.3. Experimental Materials

The audio selected in this experiment consists of two parts. The source of the sound is the work of Mr. Li Yi, a CCTV announcer.

The artificial intelligence voice is selected from "Innovative in China". The commentary voice of the documentary was synthesized by iFLYTEK CO.LTD. and Beijing Mu Si Zhou Culture Development CO.LTD. based on the recording materials of Mr. Li Yi during his lifetime. The natural voices of real people are selected from works such as " Planet Earth" and " Marine Expeditionary Brigade ". Documentary episodes and overall text present the overall structure of the overall score, and the style of interpretation varies accordingly. According to the five nodes of the beginning of the text, the end of the text, the beginning of the paragraph, the middle of the paragraph, and the end of the paragraph, the sound was extracted to make a single piece of 30s speech material, a total of 104 pieces to form the artificial intelligence speech recognition speech corpus (AIS).

Randomly select 10 pieces of artificial intelligent speech and 4 pieces of speech screening test audio from the artificial intelligence speech identification and listening speech corpus (AIS) to form the test corpus. The 4 voice screening test audios is pronounced by a 30-year-old male who is engaged in the art industry of broadcasting and hosting. The speaker is from the northern dialect area and has received professional voice training. The recording is performed in a quiet room. The software used is Adobe audition version 3.0 with a sampling rate of 44,100 Hz. Each of 4 recordings is a separate four sentences, which are used as the screening test audio for the artificial intelligence speech listening and discrimination experiment, which are used to detect whether the subjects can listen accurately and autonomously.

2.4. Experimental Steps

2.4.1. High-definition Listening Subjects Selection and Influencing Factors

150 subjects were stimulated with speech in the speech room and made their own choices according to the requirements of the listening questionnaire. The recovered experimental data were cleaned, entered into SPSS version 26.0 to analyze the test results, and a list of the subjects' hearing discrimination was obtained.

Specific research questions and hypotheses:

Q1: To examine the relationship between the four factors of gender, grade, major, and language ability and the results of human natural speech and artificial intelligent speech recognition.

Hypothesis 1: The gender of the subjects will affect the hearing recognition accuracy. Women are more sensitive to voices. We preliminarily assume that the hearing rate of real natural speech of female subjects will be higher than that of men; the artificial intelligent voice of female subjects will be higher. Hearing rate will be higher than that of men.

Hypothesis 2: The subject's professional attributes have an impact on the accuracy of hearing recognition. We preliminarily assume that the students majoring in broadcasting and hosting art have a high hearing rate of human natural speech; high.

Hypothesis 3: The language ability of the subjects has an impact on the accuracy of hearing recognition. We initially assumed that the higher the level of Mandarin, the higher the hearing rate of human natural speech; the higher the level of Putonghua, the higher the hearing rate of artificial intelligence.

Q2: Investigate the relationship between the hearing rate of human natural speech and artificial intelligence and the total hearing rate.

Hypothesis 4: Human natural speech is more closely related to the total listening rate and has a greater impact on it.

Hypothesis 5: AI speech is more closely related to the total listening rate and has a greater impact on it.

2.4.2. The Standard Extraction of High-resolution Listening Subjects

According to the screening results in step 1, a total of 6 subjects with high debating and debating rates were selected, whose natural debating rate and artificial intelligence speech debating rate were both higher than 80% and the total debating rate was higher than 90%. Debate and listening rate. The subject's subjective identification method, and the identification and listening standards were extracted according to the interview content of the six subjects. Specific operation steps: propose and mark the keywords in the interview records of the subjects; combine the keywords with similar items and try to classify them; mark the interview content again to check whether the merged similar items can cover the second time If the marked result can be covered, the proof has been exhausted, and it can be regarded as the extraction of the audiometric standard for the time being. Try to use the phonetic theory to revise and supplement the standard terminology, and reuse the newly formed auditory discrimination standard to extract the interview records of the subjects. If it can cover the subjects' auditory discrimination standards, it can be regarded as a the final hearing standard.

2.4.3. Artificial Intelligent Voice Recognition and Annotation

According to the listening standards extracted from the interview in step 2, a standard table of artificial intelligent speech recognition and listening was made, and 24 people (12 males and 12 females) were randomly selected from 150 subjects for standard training. The complete texts of episodes 1-6 of "Innovation in China" were paired and assigned to the listening and tagging experiment, and the subjects were asked to tag the artificial intelligence speech according to the listening and identification standard table. According to the comparison results of annotation, the voices with high listening rate in "Innovation in China" are extracted, and the spectrogram is generated and analyzed by praat6.1.40.

3. Experimental Results and Analysis

3.1. Screening and Influencing Factors of High-resolution Listening Subjects

3.1.1. Descriptive Analysis

A total of 106 valid samples were obtained through the audio-discrimination experiment on 150 subjects. In the sample distribution, there are 47 males, accounting for 44.3%, and 59 females, accounting for 55.7%; in terms of professional distribution, there are 46 majors in broadcasting and hosting arts, accounting for 43.4%, and 40 majoring in Chinese language and literature, accounting for 37.7%. %, 15 students majoring in journalism, accounting for 14.2%, 5 students majoring in music, accounting for 4.7%; in terms of language ability, 42 students with good Mandarin proficiency, accounting for 39.6%, and 64 students with average Mandarin proficiency, accounted for 60.4%.

Table 1. The experimental results of speech aural discrimination

Category	number of valid samples	minimum value	maximum value	mean	standard deviation
Human voice listening resolution	106	0.00	1.00	0.6520	0.2136
AI voice listening resolution	106	0.00	1.00	0.5950	0.3053
Total listening resolution	106	0.35	1.00	0.6236	0.14246

Table 2. Audiometric performance of different categories of attributes

Factor	category	Human voice listening resolution	AI voice listening resolution	Total listening resolution
Discipline	The art of broadcasting and hosting	0.6519	0.5953	0.6236
	Chinese Language and Literature	0.6650	0.4900	0.5775
	Journalism	0.6510	0.5245	0.5878
	Music	0.6365	0.5270	0.5817
Gender	Male	0.6490	0.5981	0.6236
	Female	0.6519	0.5953	0.6236
language skills	Mandarin in general	0.6672	0.5625	0.6148
	Mandarin is better	0.6286	0.6452	0.6369

The listening results are shown in Table 1: the highest human language listening resolution is 100%, the lowest is 0.00%, and the average is 65.20%; the highest artificial intelligence language listening resolution is 100%, the lowest is 0.00%, and the average is 59.50% , slightly lower than the listening resolution of real language; the overall listening resolution is up to 100%, which can accurately distinguish real natural language and artificial intelligence synthetic language, and the minimum is 35.00%, which can distinguish 35% of real natural language and artificial intelligence to synthesize language.

Further analyze the performance of hearing resolution under different category attribute indicators, as shown in Table 2:

From a professional point of view, the total listening resolution of the subjects in the broadcasting and hosting art is higher than that of the subjects in the other three majors. Four subjects of different majors had little difference in the hearing resolution of human natural language, but the difference in the effect of artificial intelligence language listening directly affects the total listening resolution. In terms of gender, the common sense that women are more sensitive to sound than men has been challenged. Women have a slight advantage in human speech recognition, but their overall listening rate is basically the same as that of male subjects, and they do not show unique characteristics. Some gender advantage. In terms of language ability, subjects with better Mandarin proficiency performed slightly better in total listening resolution than subjects with average Mandarin proficiency.

3.1.2. Correlation and Regression Analysis

First, the independent sample t-test was performed on the data collected in the experiment, and it was found that the categorical variable factor of gender had no significant relationship with the natural speech recognition rate of real people and the speech recognition rate of artificial intelligence ($P>0.05$). Secondly, through correlation test and linear regression analysis, it was found that the human voice hearing resolution was negatively correlated with the artificial intelligence voice hearing resolution and had a significant impact ($P=-0.441$, $sig<0.01$). There is a negative correlation between the major and the artificial intelligence speech hearing resolution, and there is a significant impact ($P=-0.225$, sig.<0.05). There is no significant correlation between artificial intelligence speech hearing resolution and gender and language ability.

It can be concluded that the variable of professionalism has an impact on the artificial intelligence voice listening rate and the human natural voice hearing rate, while gender and Mandarin level have no impact on the artificial intelligence voice hearing resolution and the human natural voice hearing rate. From a professional point of view, the listening resolution of the subjects in the broadcasting and hosting arts majors is generally higher than that of the subjects in the other three majors; It is basically the same as that of male subjects; from the perspective of language ability, Mandarin is better or generally does not show significant difference.

3.2. "Breathe-voice-expression" Three-level Listening Standard

According to the theory of Broadcasting and Hosting Art(BHA), the interview content of the six high-resolution subjects was extracted to form a three-level listening-discrimination judgment standard, which was divided into three index levels, namely, the level of gas use, the level of pronunciation, and the level of expression.

The level of air use mainly includes two indicators: Breathe flow and Ventilation point. The breathe flow mainly recognizes whether the air flow in the sound can be perceived from the perspective of air. The ventilation point mainly judges whether there is a sound of inhalation or ventilation in the sound from a physiological point of view, and grabbing the air and taking the air are the ways of using air for different manuscripts in the art of broadcasting, and are more commonly used in broadcasting and dubbing works.

The phonetic level mainly includes five dimensions: initial consonant, final consonant, tone, speech flow and tone change, and light and heavy format. Examining the naturalness of speech from the changes in sound, rhyme, tone and practical application.

The expression level mainly includes five indicators: pause, stress, tone, rhythm, and emotion. The first four indicators are the four external skills commonly used in the theory of broadcast and host creation, which can be identified by sound. The pause includes pause, connection, and tone. In addition to the attitude of the sentence, it also includes the ups and downs of the tone, and the emotional index refers to the tone and attitude externalized by emotion, which is often reflected in the comprehensive presentation of the use of voice and language (internal and external skills). Internal skills cannot be measured externally and are expressed through external skills, so after comprehensive measurement, emotional indicators are assigned to the expression level as a comprehensive indicator of the expression level.

3.3. Spectral Characteristics of Artificial Intelligence Speech

3.3.1. The Fundamental Frequency Value of Artificial Intelligence Speech with Lower Naturalness Is Missing More

Through the acoustic inspection of the manual annotation results, after comparing the natural voice of the real person with the artificial intelligence voice, and the internal comparison of the artificial intelligence voice, it is found that the natural natural language of the real person and the artificial intelligence voice with high naturalness can basically guarantee the smooth flow of speech. , each word holds the fundamental frequency value and can be measured. For artificial intelligence speech with low naturalness, under the premise that the speech flow is basically smooth, the fundamental frequency value of the word is missing, and it is inversely proportional to the fluency. The higher the fluency, the fewer missing values and the lower the fluency, the higher the missing value.

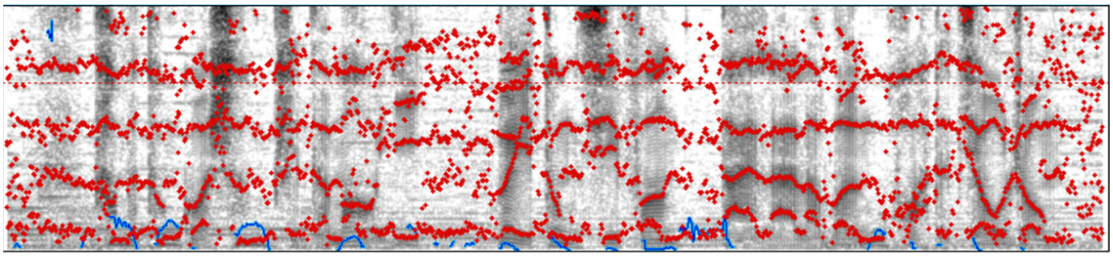

Figure 1. Sentence example of a real person's natural speech

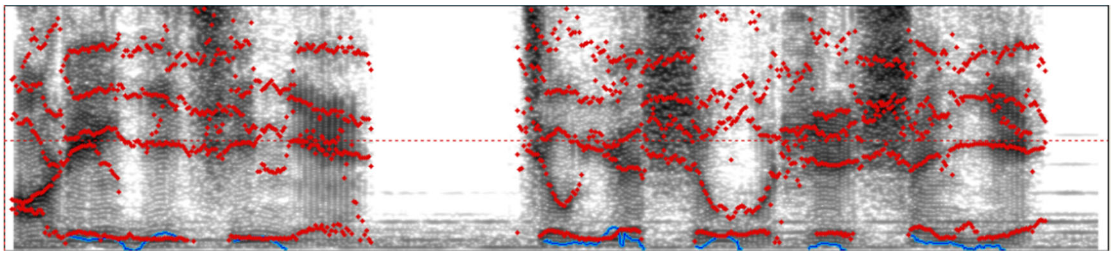

Figure 2. Examples of sentences with high naturalness of artificial intelligence speech

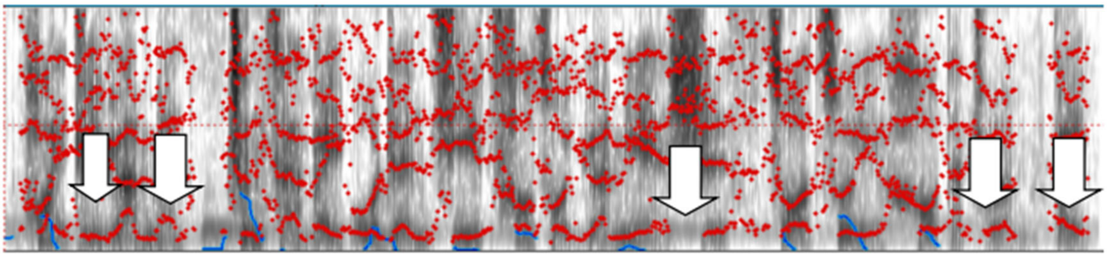

Figure 3. Examples of sentences with low naturalness of artificial intelligence speech

3.3.2. The Proportion of Artificial Intelligence Speech with Low Naturalness Is Out of Proportion

Compared with human natural speech, artificial intelligence speech with low naturalness is "uncomfortable" in hearing sense, mainly due to the problem of the sound-rhyme ratio, which is also the most easily recognized by the sense of hearing in the hearing discrimination experiment. According to the rules of Chinese pronunciation, phonology divides the internal structure of bytes into prefixes, bases and tails. The word abdomen also includes rhyme head and rhyme abdomen. The ratio of rhyme without rhyme is roughly 1:2, and the ratio of rhyme with rhyme is roughly 1:3. Taking the pronunciation of the word "jiang" as an example, j is the beginning of the word, i is the beginning of the rhyme, a is the belly of the rhyme, and ng is the end of the rhyme (the end of the word). As shown in Figure 4, cut the spectrogram of "jiang", H1 is the pronunciation of j, the fundamental frequency is high; S1 is the pronunciation of i and a, the fundamental frequency is low, and the red solid line at the bottom has experienced from The fluctuation from low to high indicates that the pronunciation has experienced sliding from i to a, and the fundamental frequency has changed; T1 is the pronunciation of ng, the fundamental frequency cannot be recognized, and it belongs to the nasal rhyme. Pronunciation of the word, the sound-rhyme ratio (H1:S1+T1) is roughly 1:3, and the ratio is relatively balanced. As shown in Figure 5, in the pronunciation of artificial intelligence speech, H2 is the pronunciation of the prefix j, which accounts for too much in the sound-rhyme ratio, the rhyme a is missing, and the rhyme i and the rhyme end ng are directly combined with nasal rhyme. The pronunciation of ing, S2 and T2 are also directly mixed together and it is difficult to distinguish, the pronunciation of the word, the rhythm ratio (H2:S2+T2) is roughly 1:1, the ratio is out of balance, the pronunciation of jiang becomes the pronunciation of jing.

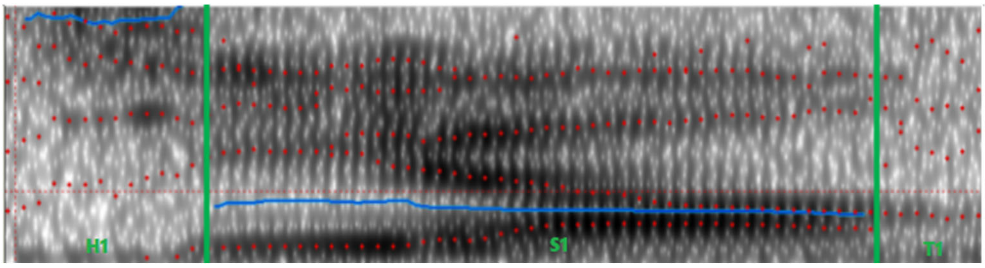

Figure 4. Pronunciation of real natural voice (jiang)

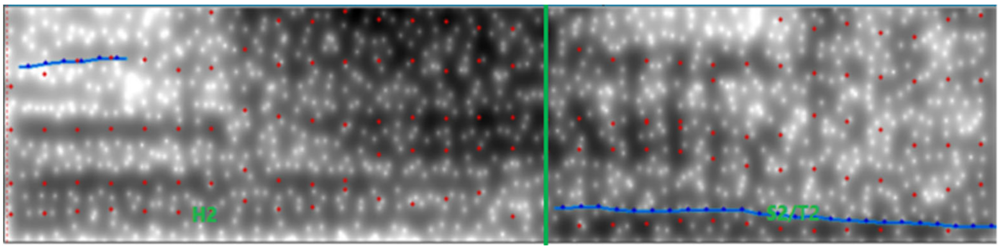

Figure 5. Pronunciation of artificial intelligence voice (jiang)

3.3.3. The Artificial Intelligence Voice with Low Naturalness Has A "Turning Tone" in the Speech Flow

In Mandarin Chinese, under the influence of the phonemes of adjacent syllables, the tone of the initials and finals in the syllables will undergo phonetic changes. Changes of the modal particle "ah". Lack of naturalness in artificial intelligence speech is most reflected in the inaccurate tone value of words, which will lead to the phenomenon of inflection. Taking the word Qian yang as an example, Qian (one tone/55 pitch) and Yang (two tone/35 pitch), in Figure 6, the blue line in the front section represents the tonal change of "qian", while the blue line in the back section represents "Yang"'s tone changes. The part circled by the green box in the middle indicates that the tone ending 5 of "Qian" and the tone beginning 3 of "Yang" will have a natural intonation drop when connecting, which is a typical change in natural speech flow, not a phenomenon of sound change, which belongs to the natural connection between sound and sound; and the pronunciation of “Qian yang” in the synthetic language in Figure 7 seems to be connected better than the real one. But the unbroken blue line in the picture has changed the pitch and key. "Qian" changed from one tone to two tone, and the pitch of "Yang" is affected by the lack of finals and endings, and the endings are obviously not enough. After the combination, it was similar to the pronunciation of "Qian (two tone/35 pitch) Yang (two tone/35 pitch).

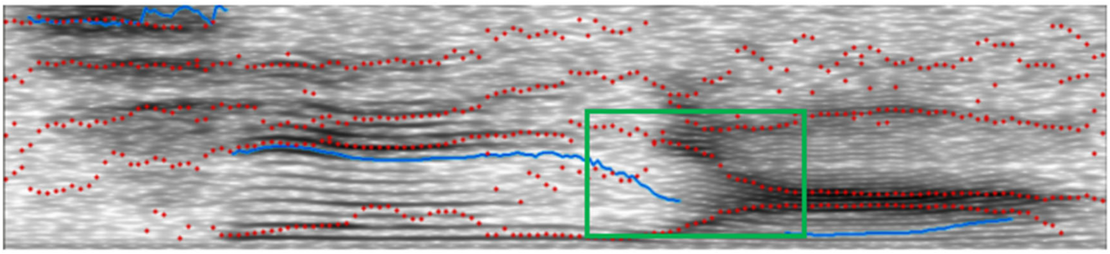

Figure 6. Pronunciation of "Qianyang" in the natural voice of a real person

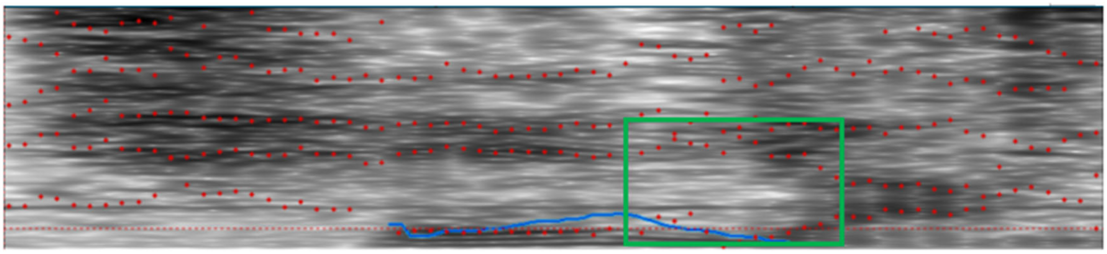

Figure 7. Pronunciation of "Qianyang" in the artificial intelligence voice

4. Conclusion and Discussion

The following conclusions are drawn through listening and discrimination experiments, interviews with high-resolution subjects, and annotation experiments on subjects of different genders, majors, and language abilities:

First, the majority has an impact on the artificial intelligence speech hearing rate and the natural human speech hearing rate. Subjects majoring in broadcasting and hosting art are more likely to conduct human speech and artificial intelligence speech than the subjects majoring in Chinese language and literature, musicology and journalism. Take advantage when listening. Previous studies have found that women are more sensitive to hearing and more sensitive to speech than men. [18] The results of this experiment are inconsistent with the results of this study. The results of this study are that the hearing discrimination rate of female subjects is basically the same as that of male subjects, and gender factors have no effect on the hearing discrimination results. It may be that the 150 subjects in this experiment have a higher rate of language use in daily life and study life. Both men and women are more sensitive to language, failing to highlight the differences in gender in hearing.

Second, the formation of three-level auditory discrimination criteria from the level of gas use, sound level, and expression level can make judgments on the naturalness of artificial intelligence speech. The rhythm and fluency of intelligent speech are tested.

Thirdly, there are three prominent problems in the current application of artificial intelligent speech in acoustic features, that is, artificial intelligence speech with low naturalness lacks more fundamental frequency values, artificial intelligence speech with low naturalness has a disproportionate rhyme ratio and low naturalness. Therefore, when artificial intelligence speech is used to make up for the lack of fundamental frequency value of artificial intelligence speech, improve the coordination degree of artificial intelligence speech rhythm ratio and refine the fineness of tone value combination, it becomes artificial intelligence speech A breakthrough for natural improvement.

At present, the subjects selected in this study only involve 150 students including 4 majors in an undergraduate university. Whether the research results based on 106 samples are more general remains to be verified, and we will try to expand to more a larger sample size is obtained from the schools in the study, which has become the next extension direction of this research topic.

References

[1] Liu Hongkui. Political and Economic Analysis of the New Development Pattern of "Double Cycle"[J]. Journal of Southwest University for Nationalities (Humanities and Social Sciences Edition), 2021,42(07):143-151.

[2] Zhao Yao. A Three-Dimensional Interpretation and Practical Strategy of Xi Jinping's Important Treatise on Artificial Intelligence [J]. Journal of Hainan University (Humanities and Social Sciences Edition), 2021,39(04):32-40.

[3] Pu Dexiang, Huo Huifang. Hotspots, Trends and Prospects of Digital Economy Research [J]. Statistics and Decision, 2021, 37(15): 9-13.

[4] Cui Xiaojing, Wang Congrong. The evolution of the main body of audio language communication in the context of new media [J]. China Radio & TV Academic Journal, 2021(08): 52-54.

[5] Lei Xia. Power control and human initiative in intelligent recommendation of search engines [J]. Modern Communication, 2021,43(05):145-151.

[6] Ren Zihan, Yao Yao, Yu Ren. Application Risks and Prevention Strategies of Voice Interaction Technology in Audiobooks [J]. Editors Monthly, 2021(04):18-23.

[7] Yu Guoming, Wang Wenxuan, Feng Fei, Xiu Lichao. Evaluation of the communication effect of synthetic speech news: The EEG evidence of the effect of speech speed [J]. Chinese Journal of Journalism & Communication ,2021,43(02):6-26.

[8] Zhang Xiaofeng, Xie Jun, Luo Jianxin, Yang Tao. Overview of Deep Learning Speech Synthesis Technology[J]. Computer Engineering and Applications,2021,57(09):50-59.

[9] Wang Yan. CVAE Based Tone Speech Synthesis and It's Application in Portable Translator [D]. Shanghai: Donghua University, 2021.

[10] Zheng Changyan, Yang Jibin, Zhang Xiongwei, Sun Meng. Bone-conducted speech enhancement using WaveNet fused with phase information [J]. Acta Acustica,2021,46(02):309-320.

[11] Li Yanping,Cao Pan,Shi Yang,Zhang Yan,Qian Bo. Voice Conversion Based on Variational Autoencoder and Auxiliary Classifier Generative Adversarial Network in Non-parallel Corpora[J]. Journal of Fudan University (Natural Science),2020,59(03):322-329.

[12] Tang Meng, Zhu Jie. Research on a natural sound evaluation method of synthetic speech based on LSTM [J]. Information Technology, 2019, 43(05): 41-44.

[13] Liu Mengyuan, Yang Jian. Design and implementation of Burmese speech synthesis system based on HMM [J]. Journal of Yunnan University (Natural Science Edition), 2020,42(01):19-27.

[14] Shuai Yuan. Study of Image/Video Super-Resolution Based on Deep Learning [D]. Shanghai: Shanghai University, 2019.

[15] Liu Zhengxia, Chen Yuxiu. A Probe into Language Ideology in Gender Language[J].Journal of Taiyuan Normal University(Social Science Edition),2019,18(05):33-36.

[16] Cassidy J W， Ditty K M. Gender differences among newborns on a transient otoacoustic emissions test for hearing[J].Journal of Music Therapy,2001,38(1):28-35.

[17] Yang Yuejun, Wang Dongbo. Research on the operation of computer-aided Putonghua proficiency test [J]. Education Teaching Forum, 2014(24):267-268.

[18] Hou Wen. Analysis of Second Language Speech Acquisition from the Perspective of Phonetic Perception[J]. Journal of Beijing Institute of Graphic Communication,2021,29(05):78-81.

[19] Cassidy J W， Ditty K M. Gender differences among newborns on a transient otoacoustic emissions test for hearing[J].Journal of Music Therapy,2001,38(1):28-35.