Text to Speech and Dubbing Liberation:Research on Affordances of Dubbing Technology for Audio and Video Platforms

[1*]Fan Xiao, 2Wenxuan Yin
1 Television School,Communication University of China,Beijing,China
2 School of Journalism & Communication,Shanghai University,Shanghai,China
[*]Corresponding author email: xf2668171884@163.com

Abstract: From broadcasting to the establishment of audio platforms to the booming ear economy, sound has found its home in the era of short video by virtue of its own irreplaceability. And the development of intelligent text to speech technology lowers the entry threshold of dubbing industry. On this basis, the theory of affordances provides insight into the impact of technology on users and their interrelationships. The study found that the frequency of use of text to speech technology has a greater effect on users, while the length of each use has no significant effect.

Key words: Text to Speech; Technology affordances; Audio and video; Dubbing

1 the introduction

With the development of artificial intelligence technology, intelligent voice products are more and more diversified, and the application scenarios of sound are more and more extensive. The number of online audio users in China reached 640 million in 2021 and is expected to rise to 690 million in 2022 [1]. At the same time, the operation of text to speech is becoming more and more convenient, and the cost is getting smaller and smaller. The entry threshold of intelligent speech industry is constantly lowering, which provides technical conditions for the emergence and development of short videos, films and audio books.

Text to Speech, also known as TTS, uses mechanical and electronic methods to produce artificial Speech. In layman's terms, text to speech technology is to give computers the ability to speak as freely as humans. At present, in the fast-paced, high-pressure social environment, the public has become an active choice to access entertainment audio products. Audio programs participate in the public life in the form of fragmentation, or just before going to bed, and the public also relieve pressure and get cured with the power of sound."Accompaniment" makes audio transform individuals in a subtle way, and auditory culture based on auditory practice is also formed and developed in the process of "moistening things silently".

In this context, the audio and video of text to speech itself has a high degree of confusion and concealment, and it is difficult to distinguish the true and false of the synthetic audio and video. This grafting of text to speech has produced an interesting collage effect in the field of new media research. But on closer inspection, a large number of "talking movies" or audio-books dubbing work are not based on the traditional corporate multi-manual production model, but on personal editing and publishing, and some bloggers can even update several pieces a day. However, this is far more than social labor necessary to understand speech and voice dubbing. Under the background of the rapid development of audio and video platforms, this paper attempts to investigate how sound bloggers and film bloggers improve labor productivity of dubbing rapidly based on the theory of technology affordances. And what kind of influence and consequence does it have to the related industry?

2 Research Review

2.1 Literature review of text to speech

"Over the past 100 years, the so-called 'emerging field of sound research has shown a strong trend of emerging, but never ending," says academic Michele Hilmes in The Study of Sound Culture: Is there such a Field?[2] In "Introduction: Listening to American Studies," Kara Keeling notes, "In the West, the study of sound has always been considered a second-rate option.Vision has traditionally been associated with reason, knowledge, science, truth and rationality. Sound, on the other hand, is' seen 'as fleeting, contingent, subjective and contingent. The former provides evidence, the latter

is hearsay."[3]

Since modern times, with the occurrence and development of modernity, there has been a key turning point of sound technology transformation, which means the arrival of the era of sound preservation, reproduction, transfer and transmission. This is what Schafer calls "schizophonia." [4]The turn of the century changed the way humans touch and experience sound, and decisively changed our understanding of the nature of sound.

Some scholars believe that sound acts as an intermediary between people and the environment, through which people perceive the world and get involved in the world. In the audio transformation, broadcasting has obvious advantages of content, users and anchors. Some scholars believe that text to speech is expected to become a breakthrough for broadcast media to lead mobile audio platforms.[5]Some scholars believe that visual shift can help audio media to grasp the wind and get rid of the shackles of a single form of expression.[6]

Both at home and abroad, there are few researches on text to speech, and most of them are literature review. Chinese researcher Yu Guoming analyzed the different propagation effects of synthetic speech and human speech from the perspective of hot and cold media. Electroencephalography (EEG) technology was employed to explore the conditions of different sound sources (human speech and text to speech) and different voice genders (male and female). Audience's user experience when using voice news products.[7] Other studies on text to speech tend to be technical, analyzing the hardware support of intelligent technology for text to speech, but there is no research on text to speech from the perspective of technology affordances.

2.2 Technical affordances and text to speech

The concept affordances originates from the field of ecological psychology, referring to "what the environment affordances, whether good or bad."[8] To explain the specific correlation between living things and their environment. Later, based on the possibility technology affords for social interaction, Gaver (1996) proposed the "Technology affordances", pointing out that the material attributes of technology can

shape social culture and communicative behavior, making affordances further become an effective theoretical framework for user-centered technology investigation.[9]

In 2015, Schrock proposed the affordances of communication, which discusses the interaction between the subjective perception of utility and the objective nature of technology and how to change communication practices or habits. He believes that mobile media has portability, affordances, positioning and multimedia.[10 At the same time, affordances exists on multiple levels, including infrastructure, devices, applications, functions, etc. When people interact with technology, they will strategically move, merge or replace among these levels according to their own situation.

Although the definition of affordances has different emphases in different fields, its fundamental attribute has been widely emphasized, that is, affordances is a relational attribute, which exists in the interaction between subject and technology, including perceptions, actions, communications, etc. Nagy and Neff (2015) advocated attaching importance to the role of imagination and proposed the concept of imagined affordances to investigate the interaction between people and technology. It points out that users deconstruct and reconstruct the meaning of technology from three dimensions: mediated experiences, materiality and influence.[11] Users have specific expectations about the media technologies they use, and these expectations may shape how they approach, interact with, and use the media.

From the perspective of technology affordances, the form of synthetic voice has significantly lowered the threshold for the release of entertainment products such as news and audio books, and it can be quickly produced by using words without human voice dubbing. A large number of Himalayan audiobooks make use of synthetic speech, reducing the amount of labor required and repidly increasing the productivity of individual creation.Moreover, synthetic speech significantly reduces the sound conditions required for a voice recording, allowing ordinary bloggers with heavy dialects and imperfect pronunciation to participate in dubbing. In addition, synthetic speech is not easy to make mistakes, and it is easy to anchor the error point for modification. Compared with the traditional mode of finding the time point after the error to edit and record, its fault tolerance rate is very high.

However, most of the research on technology affordances is focused on digital news and big data algorithms, and there is no study on how the mediation of communication technology affects actors and the environment in combination with text to speech. Although from the perspective of the platform as a whole, synthetic voice will further expand the contact surface between UGC creators and AIGC creation from the technical level, so that a large number of platform users can participate in the production and expand the platform flow. But on the negative side, this Text to Speech is more about creating a bigger pool of traffic for platforms to exploit.In order to understand the formation mechanism of contradiction behind text to speech, technology affordances are an unavoidable theoretical perspective.

2.3 Problem Presentation

Therefore, this paper attempts to answer the following questions from the perspective of technology affordances theory:

1. How do audio and video platform bloggers quickly improve the labor productivity of dubbing?
2. What is the generation mechanism behind it?
3. What are the impacts and consequences on relevant industries?

3 Research Methods

3.1 Questionnaire survey

3.1.1 Hypothesis Establishment:

Based on the above questions, the corresponding hypotheses are proposed here:

H1: The more Text to Speech technology is used, the more audio and video platform bloggers are able to overcome imperfect pronunciation

H2: The longer Text to Speech technology is used, the more audio and video platform bloggers are able to overcome imperfect pronunciation

H3: The more voice synthesis technology is used, the more productivity will be improved for audio and video platform bloggers

H4: The longer the use of voice synthesis technology, the more productivity of

audio and video platform bloggers

In order to prove the hypothesis, this study adopts the questionnaire survey method to sample the population and make the questionnaire.

3.2 In-depth interview method

The interview questions are mainly divided into three parts. The first part mainly involves the basic information of the interviewees, including their names, ages and regions. In the second part, the dimension of research questions is reduced to interview questions.It includes their views on text to speech, specific technology affordances strategy and usage of text to speech technology. The third part, preparation, deals mainly with questions that may be dug deep.If the interviewees do not want to answer any questions or dissatisfaction during the interview, they can keep silent or quit during the interview. The interview materials that the interviewees declared they did not want to make public were not included and cited in this study.

4 Research findings and discussion points

Based on the above analysis of the implementation background and technical affordances of text to speech, the author predicts the implementation effect, but the prediction needs to be further confirmed by the data results of the questionnaire survey. According to the early feedback adjustment, after four times of modification and improvement, the author sent 30 questionnaires to the Text to Speech chat group for pre-test. The analysis results of SPSS software showed that the α reliability coefficient of the questionnaire was 0.994 (>0.9). In the process of questionnaire distribution, the author distributed questionnaires through destination sampling and snowball sampling through wechat moments, questionnaire star sample database, Cremado sample database and other channels. Finally, 308 valid questionnaires were obtained.SPSS software analysis showed that the α reliability coefficient value of the questionnaire was 0.932 (>0.9), ievaluating a high reliability level.

4.1 Sample

Table 4-1

Frequency analysis results				
The name	options	frequency	The percentage (%)	Cumulative Percentage (%)
Your gender is	male	161	52.27	52.27
	female	147	47.73	100.00
What's your age?	Under the age of 18,	36	11.69	11.69
	18 to 25 years old	119	38.64	50.32
	26 to 30 years old	129	41.88	92.21
	31 years of age or older	24	7.79	100.00
What is your educational background?	Primary school	12	3.90	3.90
	Junior high school	12	3.90	7.79
	High school	61	19.81	27.60
	Junior college/secondary school	152	49.35	76.95
	Undergraduate course	55	17.86	94.81
	A graduate student	16	5.19	100.00
Where is your permanent residence?	city	180	58.44	58.44
	Cities and towns	92	29.87	88.31
	rural	36	11.69	100.00
A combined	308	100.0	100.0	

Through descriptive analysis of the recovered sample data, the above table is obtained. As can be seen from Table 4-1, male samples accounted for a larger proportion than female samples in this questionnaire survey.In terms of age, most of the samples are between 26 and 30 years old, reaching 41.88%. In terms of education distribution, the sample with the largest education level of junior college and technical secondary school accounted for the highest proportion, nearly half. In terms of geographical distribution, 58.44% of the samples lived in cities for a long time.

4.2 Variables

Table 4-2

Frequency analysis results				
The name	options	frequency	The percentage (%)	Cumulative Percentage (%)
How many hours do you use the Text to Speech platform at a time?	More than 9 hours (including 9 hours)	27	8.77	8.77
	6-9 hours including 6 hours	130	42.21	50.97
	3-6 hours (including 3 hours)	107	34.74	85.71
	Less than 3 hours	44	14.29	100.00
How often do you personally use voice synthesis platform to edit and publish audio and video content?	Once a month at most	12	3.90	3.90
	About once a month	39	12.66	16.56
	About once a week, but more than once a month	57	18.51	35.06
	About once a day, but more than once a week	136	44.16	79.22
	More than 3 times a day	47	15.26	94.48
	More than 5 times a day	17	5.52	100.00
A combined	308	100.0	100.0	

A descriptive analysis of the distribution of the frequency and time of using synthetic techniques in electronic media was conducted to obtain the figure above.It can be seen from the figure that most people synthesize speech once a day or so, accounting for 44.16%.The usage time was 6-9 hours, accounting for 42.21%. Thus, most people use text to speech technology more frequently and for a longer time.

Frequency analysis results				
The name	options	frequency	The percentage (%)	Cumulative Percentage (%)
Using the Text to Speech platform, I published more Text to Speech works	Totally disagree	17	5.52	5.52
	Basically disagree	33	10.71	16.23
	neutral	70	22.73	38.96
	Basic agree	100	32.47	71.43
	Totally agree with	88	28.57	100.00
A combined	308	100.0	100.0	

In terms of labor productivity, when asked whether the use of text to speech can produce more text to speech works, the proportion of "basically agree" and "completely agree" is more than 50%, indicating that most people believe that text to speech technology can effectively improve the production efficiency of works.

4.3 Hypothesis Verification

Table 4-3-1

Basic indicators						
The name	Sample size	The minimum value	The maximum	The average	The standard deviation	The median
Imperfect pronunciation	308	1.000	5.000	3.841	1.166	4.000
The tone is not pleasant	308	1.000	5.000	3.847	1.147	4.000

Table 4-3-1 shows the pronunciation of sample text to speech.As shown in the table, the average score of "imperfect pronunciation" and "bad timbre" is 3.8, indicating that most people are not satisfied with their pronunciation.When performing audio and video dubbing operations, most people will pay attention to their own "pronunciation" and "timbre" problems, and feel there is room for improvement.

H1: The more Text to Speech technology is used, the more audio and video platform bloggers are able to overcome imperfect pronunciation

Table 4-3-2

Cross (chi-square) analysis results										
The title	The name	How often do you personally use voice synthesis platform to edit and publish audio and video content?(%)						Total	chi-square	p
		1.0	2.0	3.0	4.0	5.0	6.0			
Imperfect pronunciation	1.0	1 (8.33)	1 (2.56)	6 (10.53)	7 (5.15)	2 (4.26)	1 (5.88)	18 (5.84)	35.648	0.017 *
	2.0	0 (0.00)	8 (20.51)	9 (15.79)	4 (2.94)	4 (8.51)	1 (5.88)	26 (8.44)		
	3.0	3 (25.00)	10 (25.64)	11 (19.30)	14 (10.29)	9 (19.15)	4 (23.53)	51 (16.56)		
	4.0	3 (25.00)	9 (23.08)	18 (31.58)	51 (37.50)	18 (38.30)	6 (35.29)	105 (34.09)		
	5.0	5 (41.67)	11 (28.21)	13 (22.81)	60 (44.12)	14 (29.79)	5 (29.41)	108 (35.06)		
Total	12	39	57	136	47	17	308			
The tone is not pleasant	1.0	2 (16.67)	2 (5.13)	3 (5.26)	6 (4.41)	3 (6.38)	0 (0.00)	16 (5.19)	30.569	0.061
	2.0	0 (0.00)	5 (12.82)	5 (8.77)	9 (6.62)	5 (10.64)	4 (23.53)	28 (9.09)		
	3.0	4 (33.33)	10 (25.64)	13 (22.81)	11 (8.09)	8 (17.02)	3 (17.65)	49 (15.91)		
	4.0	4 (33.33)	14 (35.90)	19 (33.33)	50 (36.76)	17 (36.17)	5 (29.41)	109 (35.39)		
	5.0	2 (16.67)	8 (20.51)	17 (29.82)	60 (44.12)	14 (29.79)	5 (29.41)	106 (34.42)		
Total	12	39	57	136	47	17	308			
* * * p < 0.05 p < 0.01										

Table 4-3-2 is obtained by cross analysis of "frequency of personal use of text to speech platform to clip and release audio and video content", "imperfect personal pronunciation" and "bad sound of personal pronunciation". As can be seen from the table, P =0.017, there is a significant difference between imperfect pronunciation and release frequency, H1 is valid.

H2: The longer Text to Speech technology is used, the more audio and video platform bloggers are able to overcome imperfect pronunciation

Table 4-3-3

Cross (chi-square) analysis results								
The title	The name	How many hours per day do you use the Text to Speech platform?(%)				Total	Chi-square	P
		1.0	2.0	3.0	4.0			
Imperfect pronunci-ation	1.0	4 (14.81)	7 (5.38)	3 (2.80)	4 (9.09)	18 (5.84)	17.478	0.132
	2.0	3 (11.11)	10 (7.69)	5 (4.67)	8 (18.18)	26 (8.44)		
	3.0	4 (14.81)	18 (13.85)	20 (18.69)	9 (20.45)	51 (16.56)		
	4.0	8 (29.63)	47 (36.15)	39 (36.45)	11 (25.00)	105 (34.09)		
	5.0	8 (29.63)	48 (36.92)	40 (37.38)	12 (27.27)	108 (35.06)		
Total	27	130	107	44	308			
The tone is not pleasant	1.0	4 (14.81)	5 (3.85)	3 (2.80)	4 (9.09)	16 (5.19)	12.891	0.377
	2.0	4 (14.81)	10 (7.69)	9 (8.41)	5 (11.36)	28 (9.09)		
	3.0	4 (14.81)	19 (14.62)	17 (15.89)	9 (20.45)	49 (15.91)		
	4.0	7 (25.93)	52 (40.00)	37 (34.58)	13 (29.55)	109 (35.39)		
	5.0	8 (29.63)	44 (33.85)	41 (38.32)	13 (29.55)	106 (34.42)		
Total	27	130	107	44	308			
* * * $p < 0.05$ $p < 0.01$								

Table 4-3-3 is obtained by cross analysis of "personal use time of text to speech platform", "personal pronunciation is not perfect" and "personal pronunciation tone is not pleasant". It can be seen from the table that all P are >0.05, indicating that there is no significant difference and H2 is not established

H3: The more voice synthesis technology is used, the more productivity will be improved for audio and video platform bloggers

Table 4-3-4

Cross (chi-square) analysis results										
The title	The name	How often do you personally use voice synthesis platform to edit and publish audio and video content?(%)						Total	chi-square	p
		1.0	2.0	3.0	4.0	5.0	6.0			
Using the Text to Speech platform, I published more Text to Speech works	1.0	2 (16.67)	3 (7.69)	3 (5.26)	7 (5.15)	1 (2.13)	1 (5.88)	17 (5.52)	42.288	0.003 * *
	2.0	2 (16.67)	1 (2.56)	11 (19.30)	9 (6.62)	6 (12.77)	4 (23.53)	33 (10.71)		
	3.0	3 (25.00)	13 (33.33)	17 (29.82)	23 (16.91)	9 (19.15)	5 (29.41)	70 (22.73)		
	4.0	4 (33.33)	16 (41.03)	14 (24.56)	41 (30.15)	22 (46.81)	3 (17.65)	100 (32.47)		
	5.0	1 (8.33)	6 (15.38)	12 (21.05)	56 (41.18)	9 (19.15)	4 (23.53)	88 (28.57)		
Total	12	39	57	136	47	17	308			
* * * p < 0.05 p < 0.01										

Table 4-3-4 is obtained by cross analysis of "frequency of personal use of Text to Speech platform to clip and release audio and video content" and "whether more Text to Speech works can be released after using Text to Speech platform". According to the table, p=0.003, then H3 is valid

H4: The longer the use of voice synthesis technology, the more productivity of audio and video platform bloggers

Table 4-3-5

Cross (chi-square) analysis results								
The title	The name	How many hours per day do you use the Text to Speech platform?(%)				Total	chi-square	p
		1.0	2.0	3.0	4.0			

Using the Text to Speech platform, I published more Text to Speech works	1.0	3 (11.11)	7 (5.38)	4 (3.74)	3 (6.82)	17 (5.52)	17.042	0.148
	2.0	6 (22.22)	10 (7.69)	9 (8.41)	8 (18.18)	33 (10.71)		
	3.0	3 (11.11)	33 (25.38)	22 (20.56)	12 (27.27)	70 (22.73)		
	4.0	9 (33.33)	37 (28.46)	41 (38.32)	13 (29.55)	100 (32.47)		
	5.0	6 (22.22)	43 (33.08)	31 (28.97)	8 (18.18)	88 (28.57)		
Total	27	130	107	44	308			
* * * p < 0.05 p < 0.01								

Make a cross analysis between "the daily time of using the Text to Speech platform" and "whether more works can be released after using the text to speech platform", and get the table 4-3-5.As can be seen from the table, P=0.148, indicating that the significance is not obvious, that is, hypothesis H4 is not tenable.

5 conclusion

To sum up, the hypothetical demonstration results are shown in the table below.

Assuming that	content	The results of
H1	The more text to speech technology is used, the better audio and video bloggers are able to overcome imperfect pronunciation	Set up
H2	The longer the technology is used, the better audio and video bloggers are able to overcome imperfect pronunciation	Don't set up
H3	The more voice synthesis technology is used, the more productivity is improved for audio and video bloggers	Set up
H4	The longer voice synthesis technology is used, the more productivity audio and video bloggers can achieve	Don't set up

Through the combination of quantitative research and qualitative research, this paper studies the use of voice synthesis technology on audio and video platforms with

the theory of technology affordances, and tries to answer how sound bloggers and film bloggers can quickly improve the labor productivity of dubbing based on the theory of technology affordances.And what kind of influence and consequence does it have to the related industry? It is found that H1 and H3 are true, while H2 and H4 are not. This is mainly because the frequency of use of text to speech technology has a greater impact on audio and video bloggers, while the time of use of text to speech technology has a smaller impact on audio and video platform bloggers." After the use of more, I can obviously feel the sound quality of the work is better."(Respondent No. 6) In other words, the more audio and video bloggers use text to speech technology, the more they can overcome their "imperfect pronunciation" and improve the labor productivity of audio and video works.

This study presents the application of text to speech technology in audio and video platforms from the perspective of technology affordances, reflecting the application and development of text to a speech phenomenon in audio and video platforms.The research shows that when the voice synthesis technology is used more often, it can effectively improve the voiceover status of voiceover bloggers and increase the frequency of their works, which greatly liberates the subject of dubbing and improves the production efficiency." After the platform gave me the voice, which I didn't have to change much. It felt much better than my own voice."(Respondent No. 2)

On the other hand, it also provides theoretical and practical support for audio and video synthesis in the context of technology affordances in modern China. The use of text to speech technology is more frequency-biased "practice makes perfect", and with more and more use, the improvement effect becomes more obvious.

However, it is worth noting that this study also found that it takes some time to get used to voice synthesis technology. "I am not familiar with the operational interface when I first get used to it, which leads to a long time for the first few times I use voice synthesis technology."(Respondent No. 5) In addition, excessive use of voice synthesis technology will bring negative effects of excessive reliance on technology. "After using voice synthesis technology, I no longer want to dub myself" (Respondent No. 1). All these are worthy of vigilance and reflection.

However, due to the regional reasons why the survey sample, this study still has some shortcomings, such as the survey scope is not large enough and the number of respondents is not large enough. In future research, researchers in the field of technology affordances may need to pay more attention to the phenomenon of text to speech and dubbing liberation, and expand the research and practice of text to speech on audio and video platforms from the perspective of technology affordances.

Reference

[1] Ai Media Net. Research Report of China online Audio Industry in 2020-2021. https://www.iimedia.cn/c1020/77802.html.

[2] Michele,H.(2005). "Is There a Field Called Sound Culture Studies? And Does It Matter?", *American Quaterly,* p.249.

[3] Kara,K.(2011). "Introduction: Listening to American Studies," American Quarterly, Supl. Special Issue: Sound Clash: Listening to American Studies; *College Park,* pp.445-459, 858-859

[4] R.M urray Schafer. (1994). *The Soundscape: Our Sonic Environment and the Tuning of the World*, Rochester:Destiny Books, pp.88-91.

[5] Lai, L.J & Lv, Y.L.(2021). Short Audio: Broadcast leads the breakthrough of mobile audio platform. *Chinese radio* (03), 38-42.

[6] Wang,J. (2018). Visual exploration of broadcast programs under the impact of new media. *Sound Screen World* (12),42-44.

[7] Feng,F. Wang,W.X. Xiu,L.C. & Yu,G.M.(2020). Hot and cold media: Different propagation effects of synthetic speech and human speech: Experimental evidence based on EEG. *Journalism and Communication Studies* (12),5-20+126.

[8] Gibson, J.J. (1986). *The Ecological Approach to Visual Perception*. New York: Psychology Press. 127.

[9] Gaver, W. (1991). Technology affordances. In: Proceedings of the SIGCHI conference on Human factors in computing systems, *ACM*, pp. 79-84.

[10]Hutchby, I.(2001). Conversation and Technology: From the Telephone to the Internet. Cambridge, UK: Polity.

[11] Nagy, P. & Neff, G. (2015). Imagined Affordance: Reconstructing a Keyword for Communication Theory.