UHD World Association
世界超高清视频产业联盟

Technical requirements for quality grading of 3D digital human

（Version NO. 1.0）

Release Time
2023-08-30

UHD World Association (UWA)
T/UWA 019-2023

Contents

1 Scope........1
2 Normative references........1
3 Terms and definitions........1
3.1 3D digital human........1
3.2 Fidelity........1
3.3 Detail........2
3.4 Lip sync........2
3.5 Naturalness........2
3.6 Frame rate........2
3.7 Resolution........2
3.8 Facial Animation Parameters (FAP)........2
3.9 Body Animation Parameters (BAP)........2
4 Abbreviations........2
5 Metrics........3
5.1 Character effect........3
5.2 Recognition and perception........5
5.3 Interaction and decision-making........5
6 3D digital human rating........6
6.1 3D digital human grading principle........6
6.2 3D Digital Human Segmentation Index Grading Standard........6
6.3 Overall grading standard of 3D digital human........8
Appendix A........10

Technical requirements for grading the quality of 3-Dimensional digital human

1 Scope

This document specifies the grading method of 3D human image digital human vision and interaction effect.

This document is applicable to grading the application effect of 3D real person image digital human, and is used to provide reference for both the supplier and the demander to select digital human products according to the scene demand.

2 Normative references

The contents in the following documents, through normative references in the text, constitute indispensable provisions of this document. Among them, only the version of the dated reference document is applicable to this document; For undated references, the latest version (including all amendments) is applicable to this document.

GB/T 21023-2007 General Technical Specification for Chinese Speech Recognition System

GB/T 36464.4-2018 Information Technology - Intelligent Voice Interaction System - Part 4: Mobile Terminal

GY/T 307-2017 Ultra High Definition Television System Program Production and Exchange Parameter Values

T/UWA 012.6-2022 Technical Requirements for Network Transmission of "One Hundred Cities, One Thousand Screens" UHD Video and Audio Transmission System

3 Terms and definitions

The following terms and definitions are applicable to this document.

3.1 3D digital human

Digital characters with human appearance, voice, action and interaction ability created by 3D technology in the digital world.

Note 1: 3D digital human can interact with users through voice, posture, facial expression and other ways. It has intelligent question and answer, emotion recognition, speech synthesis and other functions, and can simulate real human behavior and reaction.

Note 2: 3D digital human can be used in virtual reality, augmented reality, video games, entertainment, education and other fields.

Note 3: Based on digital technology, a three-dimensional digital person with anthropomorphic shape in the digital world.

Note 4: 3D digital human has been applied in more and more fields, such as virtual actors, virtual marketing, virtual tour guides, virtual doctors, virtual teachers, etc.

3.2 Fidelity

A measure used to evaluate the realism and verisimilitude of computer images, 3D models, animations or games in the virtual world. It reflects the similarity between the digital world and the real world.

3.3 Detail

It refers to the quantity and quality of details in an image, model or scene. In the field of computer graphics and computer vision, fineness is one of the important indicators to measure rendering quality and model accuracy.

3.4 Lip sync

It refers to the simulation and performance of the lip movement and pronunciation process of the characters in the animation and special effects production, so that it can synchronize with the voice content and enhance the realism and expression ability of the characters.

3.5 Naturalness

Naturalness refers to the degree to which a speech or behavior conforms to nature, truth and sincerity. In the field of artificial intelligence, naturalness can be used to evaluate the communication ability of robots or speech assistants and the quality of human-computer interaction. An intelligent robot with high naturalness can communicate with users in a smooth and natural way, answer questions with appropriate language and expression, and give users a satisfactory experience. Naturalness not only involves the accuracy of language and speech, but also includes the emotional expression ability of robots, the adjustment of intonation and speech speed, the understanding of context and appropriate response.

3.6 Frame rate

It refers to the number of image frames displayed in one second, and is an important indicator to measure the smoothness and smoothness of images.

3.7 Resolution

Resolution refers to the amount of detail that can be displayed in an image or video, usually expressed in pixels.

3.8 Facial Animation Parameters (FAP)

It refers to the key feature points that control the face. These feature points are used to generate animated vision and facial expressions, as well as head and eye movements. These feature points are part of the face definition parameters defined by MPEG-4. FAP represents 66 displacements and rotations of feature points relative to neutral facial positions, which are defined as mouth closure, eyelid tangency with iris, line of sight and head direction in front, tooth contact, and tongue contact with teeth. These FAPs are designed to be closely related to the movement of human facial muscles.[1]

3.9 Body Animation Parameters (BAP)

Key feature points that control body joints. These feature points are part of the face definition parameters defined by MPEG-4. MPEG-4 defines 168 body animation parameters to describe almost all possible body postures, of which 12 parameters describe the motion of each arm, and 29 parameters describe the motion of each hand.

4 Abbreviations

The following abbreviations are applicable to this document.

[1] MPEG-4 Facial Animation: The Standard, Implementation and Applications. Wiley. pp. 17 – 55. ISBN 978-0-470-84465-6.

3D 3-Dimensional
FAP Face Animation Parameter

BAP Body Animation Parameter
FPS Frame Per Second

5 Metrics

With the development of computer graphics and animation technology, 3D digital humans have been widely used in film, game, virtual reality and other industries. However, the presentation and interaction quality of digital humans are quite different, and there is a lack of universal evaluation criteria to evaluate their experience effects. Therefore, it is necessary to develop methods for comprehensive evaluation of digital human effects to ensure its high quality and availability, and to enhance user experience in various applications while promoting the development of digital human industry.

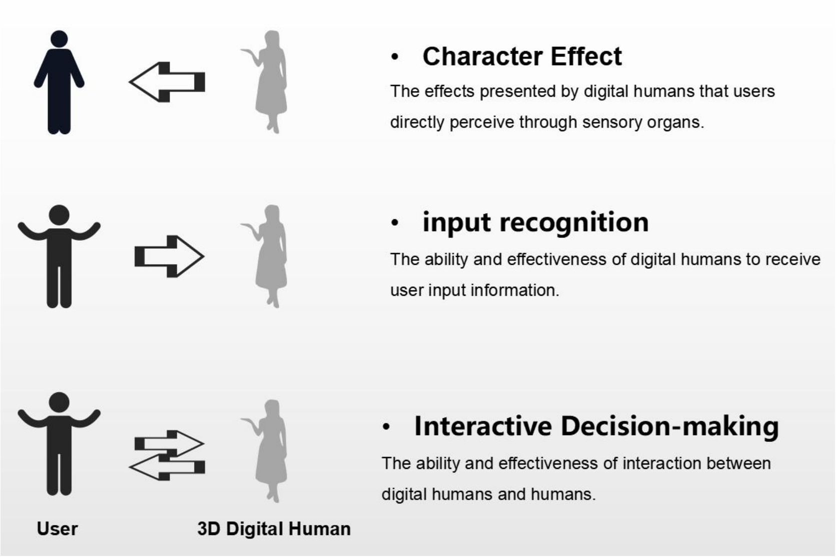
• Character Effect
The effects presented by digital humans that users directly perceive through sensory organs.
• input recognition
The ability and effectiveness of digital humans to receive user input information.
• Interactive Decision-making
The ability and effectiveness of interaction between digital humans and humans.
User
3D Digital Human

Figure 1 Three dimensions of digital human based on user experience quality

Based on user experience, the metrics can be divided into three dimensions. Character effect: digital human shape features that users can perceive, currently mainly including visual and auditory aspects; Recognition perception: digital figure recognition and the ability to detect the input information of users and the external environment, such as the accuracy rate of voice to text, face recognition rate, emotion recognition rate, etc; Interactive decision-making: the ability of digital figures to interact "independently" with users, such as the completion rate of dialogue interaction, the correct rate of expression feedback, the correct rate of limb feedback and other indicators.

5.1 Character effect

5.1.1 Facial realism

It is used to represent the degree of personification of 3D real human image digital human face. It refers to the degree to which the facial expression, structure, action and other anthropomorphic features of 3D digital people or virtual characters are similar to real humans. The range is 1-100%.

The larger the value, the higher the similarity.

5.1.2 **Visual fineness**

It is used to represent the fineness of 3D real person image and digital person image. Including the degree of detail presentation of hair, teeth, skin and other details of the 3D human image digital human model. The range is 1-100%. The larger the value, the richer the visual content and the higher the fineness.

5.1.3 **Dynamic appearance effect**

It is used to represent the richness of expressions that can be presented by 3D real person image digital people. It is quantified by the number of facial animation parameters. The more the number is, the richer the expressions can be expressed. The number is 0~66.

5.1.4 **Lip synchronization**

It is used to represent the dynamic effect of lips when 3D digital people speak. Focus on the analysis of lip movement and pronunciation process, so that it can synchronize with the dubbing content, and enhance the authenticity and expression ability of the characters. The range is 1-100%. The larger the value, the better the lip movement effect.

5.1.5 **Accuracy of text to speech**

It is used to represent the accuracy of 3D human image digital human voice synthesis and playback. The range is 1-100%. The larger the value, the higher the accuracy. The calculation method is shown in Formula (1):

$$\textbf{accuracy of text to speech} = \frac{\text{the number of correct samples}}{\text{totlal number of samples}} *100\%\text{.............. (1)}$$

5.1.6 **Speech naturalness**

It is used to represent the perceived naturalness of 3D digital human voice synthesis or real human pronunciation. Refer to Chapter 5.2.3 of GB/T 36464.4-2018 Information Technology Intelligent Voice Interaction System Part 4: Mobile Terminal for evaluation. The evaluation method is subjective evaluation, and the value is 1-5, of which 5 is the best.

5.1.7 **Physicalmovement effect**

It is used to represent the action richness of 3D real digital human. It is measured by the number of body animation parameters. The more the number, the higher the action richness. The value range is 0~168.

5.1.8 **Naturalness of combined limb movements**

It is used to represent the natural fluency of 3D digital human's combined body movements. The range is 1-100%. The larger the value, the better the lip movement effect.

5.1.9 **Frame rate**

It is used to represent the fluency of 3D real person image digital presentation image. It is expressed in frames per second (FPS).

5.1.10 **Resolution**

It is used to represent the details of 3D real person images. The number of pixels representing the width and height of the image in pixels.

5.2 Recognition and perception

5.2.1 Speech recognition accuracy

It refers to the performance of 3D digital human for speech recognition of users. Its performance is evaluated according to 5.2.1, 5.2.2 and 5.2.3 in GB/T 21023-2007 according to the different uses of speech recognition in the system.

5.2.2 Face recognition accuracy

It refers to the performance of 3D digital human for speech recognition of users. The probability representation of correct recognition of face by 3D digital human in the process of face recognition, ranging from 0 to 100%.

5.2.3 Accuracy of emotion recognition

It is used to represent the accuracy of 3D digital people in recognizing users' positive, negative and neutral emotions. The calculation method is to correctly identify the probability of the number of user conversation emotion samples accounting for the total number of samples, with a range of 0-100%.

5.3 Interaction and decision-making

5.3.1 Completion rate of dialogue interaction

It is used to represent the ability of 3D digital people to communicate with users. The calculation method is the ratio that the 3D real person image digital person can correctly understand the user's intention and give corresponding answers when interacting with the user, ranging from 0-100%.

5.3.2 Expression feedback accuracy

It is used to represent the ability of 3D digital human to correctly understand the user's expression and feedback the correct expression in the process of interaction with the user's expression. The calculation method is the ratio that the 3D real person image digital person can correctly understand the user's expression and feedback the correct expression when interacting with the user's expression, ranging from 0 to 100%.

5.3.3 Accuracy feedback accuracy

It is used to represent the ability of 3D digital human to correctly understand user actions and feedback correct body actions in the process of interaction with user expressions. The calculation method is the ratio that the digital person can correctly understand the user's actions and feedback appropriate actions when the user interacts with the 3D real person image digital person's limbs, ranging from 0 to 100%.

5.3.4 Consistency of dialogue, expression and physical action

It is used to assess the matching degree of the three items of voice, capacity and line in the digital person's fixed duration (unit: s) video. If the audio is in advance, the expression is redundant, missing, the limb action is delayed, and the limb action is misaligned, it will be deemed as mismatching. In the standard evaluation time, EN represents the duration of errors such as audio advance, redundant expression, missing, limb movement delay, and limb movement dislocation, T represents the total duration, and F represents the total score of 100. The calculation method is shown in Formula (2):

$$S = F \times (1 - \frac{EN}{T}) \times 100\% \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (2)$$

5.3.5 Average stutter duration

It is used to assess the fluency of interaction between 3D digital human and users. The average duration of stuttering during interaction with digital human includes video image stuttering, perception stuttering due to mismatched voice actions, and voice stuttering. The calculation method is to obtain the single stuck time length (recorded as ti, that is, the time difference between the beginning of this stuck time and the end of this stuck time) and the stuck times within the standard evaluation time. The calculation method is shown in Formula (3), and the unit is seconds/time:

$$\text{Average stutter duration} = \frac{\Sigma(\text{ti})}{\text{totlal number of } \textbf{stutter}} \cdots\cdots\cdots\cdots\cdots\cdots \quad (3)$$

6 3D digital human rating

6.1 3D digital human grading principle

With the continuous development of digital human technology, more and more application scenarios require digital human to have a high degree of realism and interactivity. Therefore, the evaluation of 3D digital human vision and interaction effects has become an important part of the development of digital human technology. By establishing unified evaluation standards and methods, we can standardize the digital human industry, improve the quality of digital human, and promote the wide application of digital human technology in virtual reality, games, education, medical care, entertainment and other fields.

This chapter grades the 3D digital human based on the detailed description of the previous 3D digital human indicators and the calculation method suggestions provided in the appendix. Before the overall grading, complete the grading of three dimensions of 3D digital human segmentation indicators: character effect, recognition perception, and interactive decision-making. These three dimensions cover different aspects of digital human respectively and can provide more detailed and specific evaluation results. By dividing digital people into these three dimensions and grading them, we can more accurately evaluate their performance and capabilities in different aspects, help users choose digital people that meet their needs, and provide targeted feedback and suggestions to improve user experience.

6.2 3D Digital Human Segmentation Index Grading Standard

6.2.1 Grading requirements for 3D digital characters

Character effects mainly evaluate the appearance and performance of digital human. This includes appearance, voice, action and other aspects, as well as the similarity and realism with real people. The quality of character effects will affect users' acceptance of digital people. It is mainly applied to the rating of digital people in non interactive scenes such as film and television production, advertising, and digital people broadcasting. Reference Table 1 for specific classification requirements:

Table 1 Grading Requirements for Digital Figure Effects

Primary indicators	**Secondary indicators**	**Grading (range)**				
		Level 1	**Level 2**	**Level 3**	**Level 4**	**Level 5**

Character effect	Facial realism	0-39%	40% -59%	60% -79%	80% -89%	90% -100%
	Visual fineness	0-39%	40% -59%	60% -79%	80% -89%	90% -100%
	Number of basic expressions	0-29	30-39	40-49	50-65	66
	Lip synchronization	0-5	6-8	9-11	12-14	15
	Accuracy of text to speech	0-39%	40% -59%	60% -79%	80% -98%	99% -100%
	Speech naturalness	1	2	3	4	5
	Number of physical movements	0-75	76-115	116-155	156-195	196
	Physical movement naturalness	0-39%	40% -59%	60% -79%	80% -89%	90% -100%
	Frame rate (FPS)	0-24	25-39	40-49	50-119	Over 120
	Resolving	Less than 1920 * 1080/2048 * 1080,	1920 * 1080/2048 * 1080	4K	8K	Above 8K

6.2.2 Requirements for perception grading of 3D real person image digital person recognition

Recognition perception refers to the digital human's perception of the surrounding environment and users. It includes voice recognition, image recognition, emotion recognition and other technologies. In the process of interaction with users, it can accurately understand the needs and instructions of users. It is mainly applied to digital human based on environment recognition perception scene, such as intelligent input assistant, smart home, etc. See Table 2 for specific classification requirements:

Table 2 Requirements for Perception Grading of Digital Human Identification

Primary indicators	**Secondary indicators**	**classification**				
		Level 1	**Level 2**	**Level 3**	**Level 4**	**Level 5**
Recognition and perception	Face recognition accuracy	0-39%	40% -59%	60% -79%	80% -98%	99% -100%

	Speech recognition accuracy	0-39%	40% -59%	60% -79%	80% -98%	99% -100%
	Emotion recognition accuracy	0-39%	40% -59%	60% -79%	80% -98%	99% -100%

6.2.3 **Grading requirements for interactive decision-making of 3D real person image digital person**

Interactive decision-making refers to the ability of digital people to respond and make decisions according to user needs in specific situations. This includes analyzing and reasoning from the information provided by users, making appropriate responses or providing appropriate suggestions, etc., and scenarios such as intelligent customer service and digital staff for major users. See Table 3 for specific classification requirements:

Table 3 Digital Human Interaction Decision Grading Requirements

Primary indicators	Secondary indicators	classification				
		Level 1	Level 2	Level 3	Level 4	Level 5
Interaction and decision-making	Completion rate of dialog interaction	0-39%	40% -59%	60% -79%	80% -98%	99% -100%
	Expression feedback accuracy	0-39%	40% -59%	60% -79%	80% -98%	99% -100%
	Action feedback accuracy	0-39%	40% -59%	60% -79%	80% -98%	99% -100%
	Consistency of dialogue, expression and body feedback	0-39%	40% -59%	60% -79%	80% -89%	90% -100%
	Average stutter duration (seconds)	Greater than 10	7-10	4-6	2-3	0-1

6.3 **Overall grading standard of 3D digital human**

According to its application scenarios and specific requirements, the evaluation indicators included in the digital human system will also vary. Specific index requirements are as follows.

Table 4 Grading Standards for Digital People

Primary indicators	Secondary indicators	Overall grading				
		Level 1	Level 2	Level 3	Level 4	Level 5

Character effect	Facial realism	●	●	●	●	●
	Visual fineness	●	●	●	●	●
	Number of basic expressions	●	●	●	●	●
	Lip synchronizationt	●	●	●	●	●
	Accuracy of text to speech	●	●	●	●	●
	Speech naturalness	●	●	●	●	●
	Number of　physical movements	●	●	●	●	●
	Physical movement naturalness	●	●	●	●	●
	Frame rate	●	●	●	●	●
	Resolving	●	●	●	●	●
Recognition and perception	Face recognition accuracy	○	○	○	●	●
	Speech recognition accuracy	○	○	○	●	●
	Emotion recognition accuracy	○	○	○	●	●
Interaction and decision-making	Completion rate of dialog interaction	○	○	○	●	●
	Expression feedback accuracy	○	○	○	●	●
	Action feedback accuracy	○	○	○	●	●
	Consistency of dialogue, expression and action feedback	○	○	○	●	●
	Average stutter duration	○	○	○	●	●

Note: "●" means necessary index item, and "○" means optional index item

Appendix A

(Informative appendix)

Suggestion on Calculation Method of 3D Digital Human Grading Parameters

See the following table for suggestions on the calculation method of digital person grading indicators:

Table A.1 Technical Requirements for Digital Human Grading

Primary indicators	**Secondary indicators**	**Subdivision indicators**	**Calculation method and parameter description (appendix)**
Character effect	Image effect	Facial realism	Extract N pictures of digital people
			① The discriminator in the StyleGan generation countermeasure network trained by LFW face data is used to score the digital human face. 1000 examples are used as the sub box standard to sub box the face fidelity of the final image[2][3]
			② Judge k (k>=10) testers with professional background
			③ Subjective and objective consistency ratio=number of digital person pictures with the same algorithm and test personnel's box division results/N * 100%
		Visual fineness	Extract N pictures of digital people
			① Calculate the fineness of pictures at different positions of digital human through image definition algorithm, and calculate the sharpness of different positions according to the importance of vision as the final visual fineness, and

[2] http://vis-www.cs.umass.edu/lfw/

[3] https://github.com/NVlabs/stylegan

			then divide the final weighted fineness score into boxes to get the score
			② Judge k (k>=10) testers with professional background
			③ Subjective and objective consistency ratio=number of digital person pictures with the same algorithm and test personnel's box division results/N * 100%
	Expression effect	Dynamic facial richness	Refer to MPEG-4 Facial Animation: The Standard, Implementation and Applications Wiley Pp. 17 – 55 ISBN 978-0-470-84465-6
		Lip synchronization	Evaluate the video clips of digital people talking, and count them by dots. The initial value of N is 0. For words with poor lip movement effect, N=N+1. According to the statistical results, the calculation accuracy=(total number of words - N)/total number of words * 100%
	Voice effect	Accuracy of text to speech	The digital human video clips using speech synthesis technology are evaluated and counted by point counting. The initial value of N is 0, the words with inaccurate pronunciation and wrong pronunciation are recorded as N=N+1, and the accuracy rate=(total number of words - N)/total number of words *

			100%
		Speech naturalness	Refer to GB/T 36464.4-2018 Information Technology Intelligent Voice Interaction System Part 4: Mobile Terminal
	Action effect	Body movement richness	Refer to MPEG-4 Facial Animation: The Standard, Implementation and Applications Wiley Pp. 17 – 55 ISBN 978-0-470-84465-6
		Physical movement naturalness	Extract N pictures of digital people
			① Using the 3D human body estimation method, based on the Human3.6M action library, identify the 3D dynamic digital human, track their actions, and observe the number of points in the mold to calculate and analyze the number of degrees of freedom of each digital human's activities. Use 1000 examples as the box dividing standard, and divide the number of degrees of freedom into boxes to get the score[4]
			② Judge k (k>=10) testers with professional background
			③ Subjective and objective consistency ratio=number of digital person pictures with the same algorithm and test personnel's box division results/N * 100%

[4] http://vision.imar.ro/human3.6m/description.php

	Rendering effect	Frame rate	GY/T 307-2017 Basic Parameters of UHD TV Program Production and Exchange T/UWA 012.6-2022 6. Technical Requirements for Video Signal of Ultra HD Broadcasting Programs in the Technical Requirements for Program Broadcasting of "One Hundred Cities, One Thousand Screens" Ultra HD Video and Audio Broadcasting System
		Resolving	
Recognition and perception	Face recognition	Face recognition accuracy	LFW2 face data set based on benchmark is used for digital person recognition, and 1000 examples are used as the sub box standard to determine the recognition accuracy, and the recognition percentage is sub box
	speech recognition	Speech recognition accuracy	Refer to GB/T 21023-2007 General Technical Specification for Chinese Speech Recognition System
			Refer to ISO 9241-154:2013: Ergonomics of human system interaction Part 154: Interactive voice response (IVR) applications
	Emotional recognition	Emotion recognition accuracy	① Based on the multi language emotion analysis algorithm and the Universal Dependency Treebanks database, emotion recognition is carried out for user utterances after the voice is converted into text. There are three

			categories: positive, negative and neutral5 ② Give k (k>=10) testers with professional background knowledge to further judge the response emotion in combination with the context ③ Subjective and objective consistency ratio=the number of texts whose algorithm is the same as the test result/the total number of database corpora * 100%
Interaction and decision-making	Dialogue interaction	Completion rate of dialog interaction	Through algorithms such as the Dialog State Tracking Challenge proposed by Cambridge, 11 evaluation indicators and 3 evaluation opportunities, the basic values are tested based on the corpus and divided into boxes6 Give further box division judgment to k (k>=10) testers with professional background Subjective and objective consistency ratio=number of conversation rounds with the same algorithm and tester's judgment result/total number of conversation rounds * 100%
	Facial interaction	Expression feedback accuracy	① Give expression recognition and classification through facial action

[5] https://universaldependencies.org

[6] https://paperswithcode.com/dataset/dialogue-state-tracking-challenge

			coding system, such as FACS[7] ② Give further judgment to k (k>=10) testers with professional background ③ Subjective and objective consistency ratio=the number of expressions with the same judgment result of the algorithm and the tester/the total number of feedback * 100%
	Limb interaction	Action feedback accuracy	① The discrimination box is given through the real-time human posture recognition method in single depth image proposed by Microsoft[8] ② Give further box division judgment to k (k>=10) testers with professional background ③ Subjective and objective consistency ratio=the number of limb feedbacks with the same algorithm as the tester's judgment result/total number of feedbacks * 100%
	Interaction quality	Consistency of dialogue, expression and body feedback	It is used to assess the matching degree of the three items of tone, content and line in the digital person's fixed duration (unit: s) video. If the audio is in advance, the expression is superfluous, or missing, the total score of this indicator is 100, and the calculation formula is: S=F * (1-EN/T * 100%), in which EN

[7] https://web.cs.wpi.edu/~matt/courses/cs563/talks/face_anim/ekman.html

[8] https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/BodyPartRecognition.pdf

			represents the duration of errors such as the audio is in advance, the expression is superfluous, or missing, the limb action is delayed, or the limb action is misplaced, T represents the total duration, T>600s, F represents the total score of 100
		Average stutter time	Average length of stuck time during interaction