글로벌 표준화 세미나

AI 기반 디지털 휴먼

강지수, CRO, 클레온

INDEX

KLleon

Make digital communication sincere

클레온은 디지털 휴먼을 사용해 인간의 소통을 혁신하고자 합니다. 인간의 소통은 시공간적 한계가 있습니다.

1. 시간적 공간적 한계로 인해 나와 대화할 수 있는 사람은 제한적
2. 과거에 돌아가신 위인과 대화하는게 불가능
3. 전세계적으로 인지도 있는 사람과 대화 어려움

KLleon’s Digital Human

INDEX

The Rapid Development of AI Tech

ALEX Net

Professor Geoffrey Hinton's ALEX Net wins the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) with an overwhelming performance.

2012

2014

GAN

Generative Adversarial Network (GAN) developed by Professor Ian J. Goodfellow opened the horizon for generative models.

ALPHAGO

Google DeepMind's AlphaGo catches the world's attention by defeating the world's top professional knight, Sedol Lee.

2016

2018

Deep Fake

Deepfake-based video production has become a hot topic, and various deepfake videos are spreading on the web.

Examples

Various Digital Human

딥페이크 기술을 활용해 다양한 디지털 휴먼이 제작되고 있습니다. 하지만 딥페이크 기술은 몇가지 문제점이 있습니다.

1. 많은 데이터(1000+)를 필요
2. 디지털 휴먼을 생성할 때마다 학습이 필요
3. 디지털 휴먼 제작을 위한 툴을 다룰 수 있어야 함

Reenactment

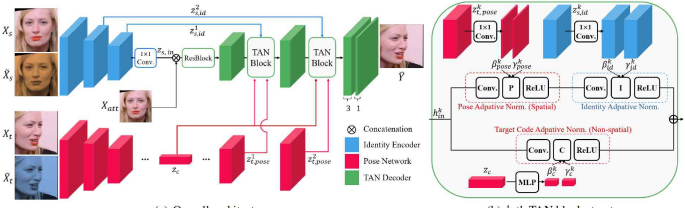
Only one image..

단 한장의 사진만으로 디지털 휴먼의 얼굴을 만들 수 있습니다.

How

Like a painter

마치 화가와 같이 다양한 사람의 얼굴을 그리다 보면, 정면 사진 한장만으로
도 다양한 표정 및 각도의 얼굴 예측이 가능합니다.



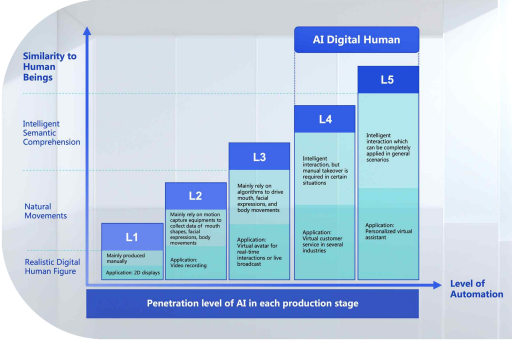
(a) Overall architecture (b) k-th TAN block structure

Question

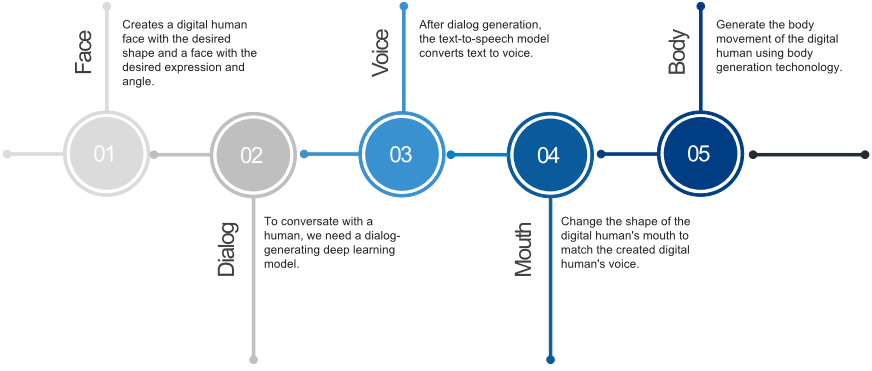
Digital Human?

얼굴 변환 기술만을 사용해 제작된 가상인간을 우리가 진짜 가상인간이라고 할 수 있을까요? Sense Time에서는 디지털 휴먼의 5단계를 정의했습니다.

1. 수작업을 통한 디지털 휴먼 외형 제작
2. 다량의 데이터를 수집해 데이터 기반 디지털 휴먼 제작
3. 데이터 수집 등의 수작업 없이 알고리즘만을 사용한 디지털 휴먼 외형 제작
4. **지능을 갖고 제한적 환경에서 대화가 가능한 디지털 휴먼**
5. 완전한 대화가 가능한 디지털 휴먼



Generation Process



Face

Face Generation Process

AI기술 기반으로 디지털 휴먼을 제작하기 위해서는 총 세가지 기술이 필요합니다.

1 Virtual Face

We create a digital human face with the look customers want.

2 3D Face

Create a face from any angle and expression using a single digital human photo.

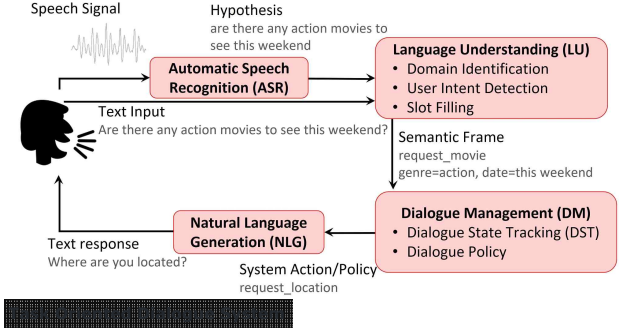
3 Face + Body

Naturally combines the generated face with the body of a digital human.

Dialog

Dialog Generation Process

디지털 휴먼의 얼굴이 제작된 이후 사람과 소통을 위한 대화 생성이 필요합니다. 이때 NLP기술을 활용해 대화를 생성합니다.



TTS

Text-to-Speech Generation Process

생성된 대화가 디지털 휴먼의 목소리로 발화되기 위해 TTS기술이 필요합니다. 현재 30초의 대상 음성만으로 디지털 휴먼의 목소리 제작이 가능합니다.

Voice Model

Only 30 sec voice

Few-Shot TTS

Result

Voice-Lip Synchronization

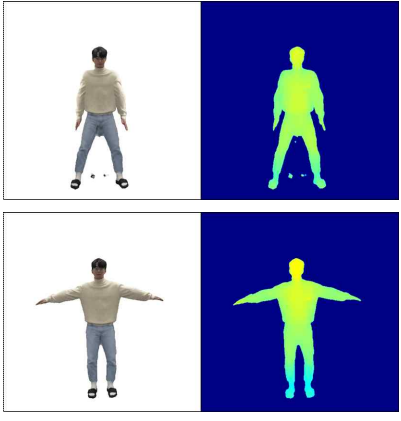
Lip-Sync Process

디지털 휴먼이 발화하는 목소리와 입술 모양의 싱크가 맞아야 자연스러운 디지털 휴먼을 만들 수 있습니다.

Body Generation

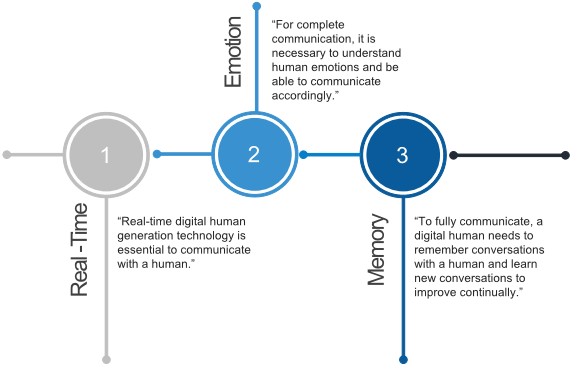
Body Generation Process

디지털 휴먼의 몸을 자유롭게 움직이도록 체형 움직임 생성 기술이 필요합니다.



INDEX

Road Map

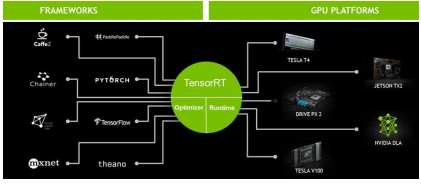

Emotion
"For complete communication, it is necessary to understand human emotions and be able to communicate accordingly."
1
2
3
Real -Time
"Real-time digital human generation technology is essential to communicate with a human."
Memory
"To fully communicate, a digital human needs to remember conversations with a human and learn new conversations to improve continually."

Real-Time

Streaming

NNStreamer

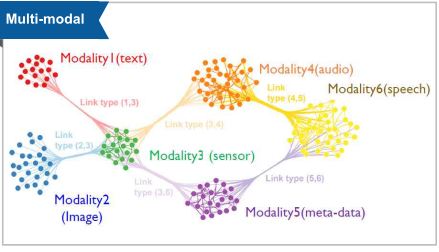
TRT



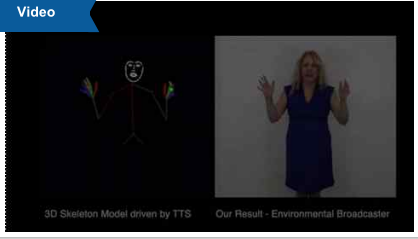
- 스트리밍: 소통을 위해서는 실시간 디지털 휴먼 움직임이 가능해야 합니다.
- 최적화: GPU 최적화를 통해 딥러닝 모델이 효율적으로 inference되어야 합니다.
- 스케듈링: 여러가지 딥러닝 모델이 활용되기 때문에 GPU연산에서의 스케듈링이 필요합니다.
- 경량화: 딥러닝 모델 자체의 경량화를 통해 inference시간을 줄여야 합니다.

Converation with Emotion

Multi-modal

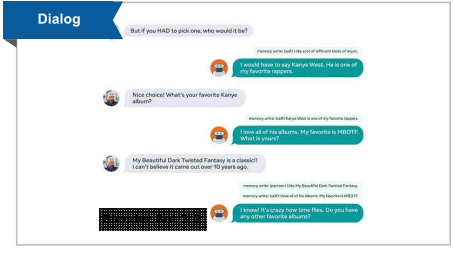


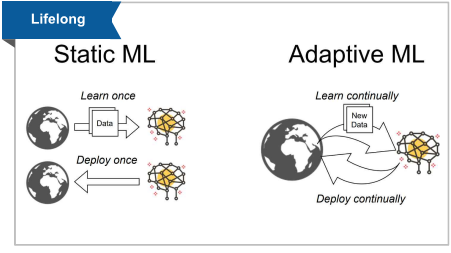
Video



- Multi-modal: 사람의 감정을 이해하기 위해서는 다양한 modality의 데이터를 이해할 수 있어야 합니다.
- NLP: 대화 상대의 감정에 맞는 대화를 생성할 수 있어야 합니다.
- Voice: 생성된 대화의 감정에 맞는 음성이 제작되어야 합니다.
- Video: 생성된 대화와 맞는 행동 및 표정을 만들 수 있어야 합니다.

Conversation with Memory

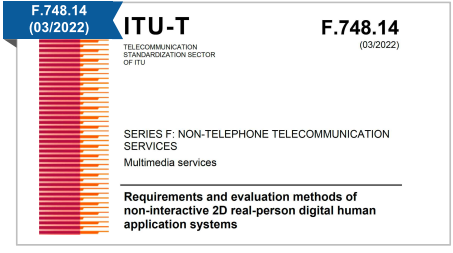


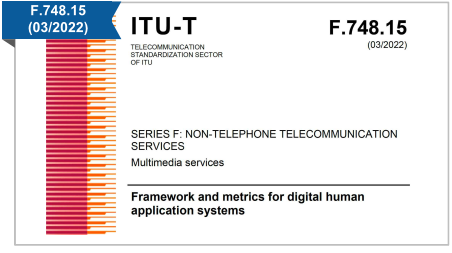


- 메모리: 기존 대화에 대한 내용을 기억하고, 대화 생성에 활용할 수 있어야 합니다.
- Lifelong: 새로운 대화를 바탕으로 딥러닝 모델이 발전할 수 있어야 합니다.

INDEX

Standards about Digital Human





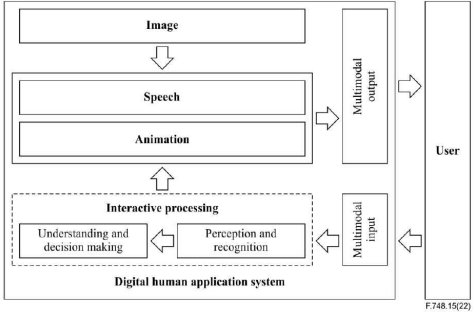
- ITU-T F.748.14는 이미지, 음성, 움직임, 디스플레이 등의 측면에서 **비대화형 2차원(2D) 디지털 휴먼 애플리케이션 시스템에 대한 요구 사항** 및 평가 방법을 지정합니다.
- ITU-T F.748.15는 디지털 인간 응용 시스템에 대한 프레임워크를 지정하고 이미지, 음성, 애니메이션, 대화형 처리 및 **다중 모드 입력/출력의 차원에 대한 해당 주관적 및 객관적 메트릭을 제안**합니다.

Standard about Digital Human

F.748.15

지능을 갖고 있는 디지털 휴먼 프레임워크 및 객관적 메트릭을 제안합니다.

1. Multi-modal 데이터를 입력으로 분석 및 대화를 생성
2. Image, Speech, Animation, Interactive Processing, Multi-modal Input, Multi-modal Output Module에 대한 메트릭을 제안



Standard about Digital Human

F.748.14

비대화형 2D 디지털 휴먼 애플리케이션 시스템

1. 비 대화형 2D 디지털 휴먼을 정의
2. F.748.15표준에서 제안한 메트릭을 사용해 성능을 평가하는 방식 제안

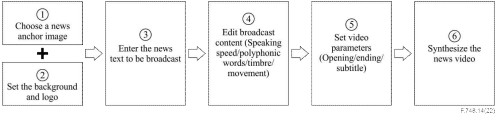


Table 1 – Subjective scoring rules for speech comfort

Evalustion dimension	Description	5	4	3	2	1
Pronunciation and intonation	Is the pronunciation standard as a whole?	Quite standard	Relatively standard	Basically standard	Individually standard	Quite nonstandard
	Is pronunciation clear?	Quite clear	Relatively clear	Basically clear	Not quite clear	Quite unclear
	Is word segmentation and sentence pause proper?	Quite proper	Relatively proper	Basically proper	Not quite proper	Quite improper
	Is tone and intonation natural?	Quite natural	Relatively natural	Basically natural	Not quite natural	Quite unnatural
	Is accent and pronunciation proper?	Quite proper	Relatively proper	Basically proper	Not quite proper	Quite improper
	Is speech speed expression proper?	Quite proper	Relatively proper	Basically proper	Not quite proper	Quite improper
Fluency and coherence	Is speech expression fluent?	Quite proper	Relatively proper	Basically proper	Not quite proper	Quite improper
Emotional fullness	According to the semantics and content of the text, is the emotional expression proper?	Quite proper	Relatively proper	Basically proper	Not quite proper	Quite improper
Anthropomorphic comfort	Is voice as anthropomorphic as a real human?	Completely indistinguishable	Relatively similar and slightly different from real voice	Basically similar	Not quite the same	Quite different
	When listening to the sound, do you feel happy?	Quite happy	Relatively happy	Just so so	Not quite happy	Quite unhappy
	Are you willing to have this voice serve you?	Quite willing	Relatively willing	Just so so	Not quite willing	Quite unwilling

Our Standard

Requirement of communication services for digital human

ITU-T, SG16, Q24에서 디지털 휴먼을 사용한 communication service에 대하여 표준 기고문을 제안했습니다.

INTERNATIONAL TELECOMMUNICATION UNION
TELECOMMUNICATION STANDARDIZATION SECTOR
STUDY PERIOD 2022-2024

SG16-TD16/WP2
STUDY GROUP 16
Original: English

Question(s):	24/16	Geneva, 17-28 October 2022

TD

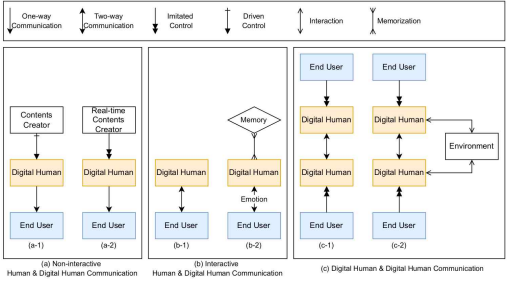
Source:	Editor F.CSDH	
Title:	F.CSDH "Requirements of communication services for digital humans" (New): Updated draft (Geneva, 17-28 October 2022)	
Contact:	Jisu Kang KLleon Republic of Korea	Tel: + 82-10-2724-5144 E-mail: jisu.kang@klleon.io
Contact:	Seunggeun Baek KLleon Republic of Korea	Tel: +82-10-7550-4329 E-mail: seunggeun.baek@klleon.io
Contact:	Miran Choi ETRI Republic of Korea	Tel: + 82-42-860-6197 E-mail: miranc@etri.re.kr

Our Standard

Requirement of communication services for digital human

총 세가지 커뮤니케이션 서비스 타입을 정의합니다.

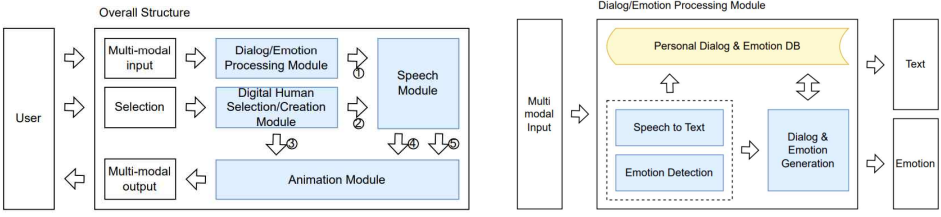
1. Non-interactive Communication Service
2. Interactive Communication Service
3. Digital-Human Digital-Human Communication Service



Our Standard

Requirement of communication services for digital human

디지털 휴먼 외형 제작 모듈, 감정 및 메모리 기반 대화 생성 모듈을 포함합니다.



Future Standard

Future Digital Human Technology

미래 디지털 휴먼 기술관련 표준화 기고문이 작성되어야 합니다.

1. 디지털 휴먼의 스트리밍 기술 관련 표준화
2. 감정 기반 대화가 가능한 디지털 휴먼 시스템
3. 메모리 기반 대화가 가능한 디지털 휴먼 시스템

Thank you

강지수, CRO, 클레온
jisu.kang@klleon.io