음성 합성과 동작 인식 기술을 활용한 CLOVA Dubbing과 Avatar 서비스

□ 배순민 / NAVER Clova

요약

코로나로 인해 사회는 급속한 변화를 겪고 있고, 그 변화의 중심에는 온라인 플랫폼 기업과 서비스가 있다. AI 기술의 발전 속도는 여전히 가속되고 있고, 특히 음성 합성과 실시간 동작 인식, 아바타 생성 기술은 콘텐츠 생성 및 비대면 서비스에서 그 활용이 더욱 기대된다.

Ⅰ. 서 론

인공지능이 처음 등장한 것은 1956년이나 그 이후 꽤 오랜 기간 기대에 부응하지 못했고 명맥만 유지되어왔다. 매사추세츠 공과대학교(MIT)에서는 컴퓨터과학 연구실(The Laboratory of Computer Science)이 1963년 설립된 것에 반해 인공지능 연구실(the Artificial Intelligence Laboratory)은 그보다 4년이나 이른 1959년에 만들어졌으나, 컴퓨터과학이 일상 전반에 빠르게 활용되는 동안 인공지능은 산업계와 학계의 중심이 되는데 60년의 세월이 걸렸다.

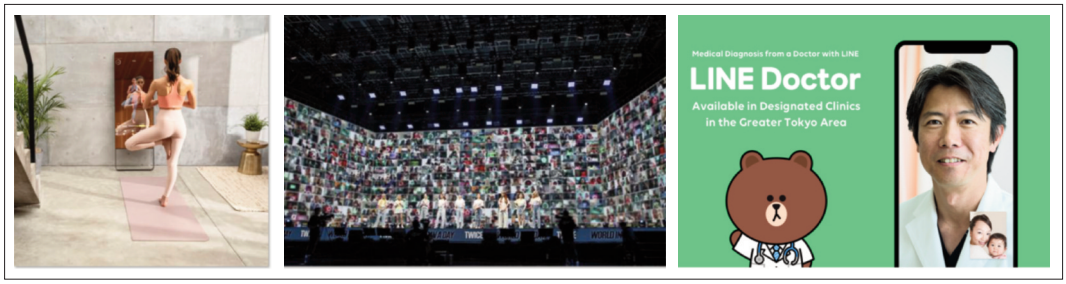
2012년 구글이 3일간 1000만 개의 유튜브 썸네일을 '비지도 학습'(unsupervised learning) 방법으로 학습하여 고양이를 인식해 낸 결과는 모든 컴퓨터 연구자들에게 큰 충격을 안겼을 것이다. 또한 2016년 우리나라에서 열린 이세돌 9단과 알파고(AlphaGo)의 대결은 인간의 지적 능력에 대한 큰 도전이었고, 멋지게 성공하여 인공지능의 가능성을 입증했다. 이미 초기 16만 게임 기보를 사용해서 학습을 시작했고, 5개월 동안 매일 3만 번의 게임을 익혔다고 하는 알파고를 상대로 이세돌 9단이 한 게임을 이긴 것은 인공지능도 예측하지 못했던 0.007% 확률의 신의 한수였다. 1202개의 CPU와 176개의 GPU 정도의 실시간 처리 능력을 한 명의 인간이 혈혈단신 대적했다는 것이 공정하지 않아 보일 지경이다.

이처럼 2012년 이후 딥러닝은 특히 영상 인식 분야에서 그동안 불가능할 거라 여겨졌던 인간 수준의 정확도를 달성하기 시작했다. 또한 음성 인식, 자연어 처리 분야에서도 두각을 나타내어 이제는 인공지능의 정확도에 대한 의심은 사라지고 많은 일상 제품과 서비스에 탑재되고 있다. 특히 네이버 클로바의 경우, 2017년 8월 인공지능 스피커 '웨이브' 출시를 시작으로 2017년 10월 '프렌즈', 그리고 2020년에는 '클락'과 '램프' 출시를 통해서 음악 재생, 날씨, 가벼운 대화(칫챗), 검색, 책 읽기까지 인공지능 서비스의 영역을 넓혀가고 있다.

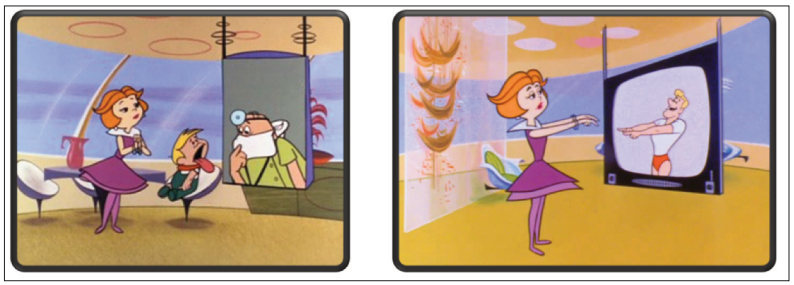
특히 음성 합성 기술은 기존에 텍스트와 화면을 읽어야 하는 모든 사용자 시나리오를 AI 음성으로 들을 수 있게 하여 인터페이스의 혁신을 가져왔다. 파파고 번역기와 네이버 사전의 듣기, 인공지능 스피커, 네이버 지도 길안내, 오상진 목소리로 뉴스를 읽어주는 AI 앵커나 유인나 목소리를 사용한 챗봇 서비스는 네이버 내 인공지능을 활용한 서비스 중에서도 활용 빈도가 가장 높다〈그림 1〉.

또한 코로나로 인한 사회적 변화 속에 기술의 역할이 달라지고, 그 변화의 중심에는 국내에서는 네이버나 카카오, 쿠팡, 배달의 민족, 당근마켓과 같은 온라인 플랫폼 기업이 있다. 비대면에 대한 불편함과 비효율로 인해서 미뤄졌었던 홈오피스, 홈러닝, 홈트레이닝, 홈콘서트, 홈닥터의 경우 선택이 아닌 필수가 되었다〈그림 2〉. 이런 일상적인 비대

<그림 1> 음성 합성 기술을 활용한 네이버 서비스



<그림 2> 홈트레이닝(mirror.co), 홈콘서트(비욘드 라이브-트와이스), 홈닥터(라인닥터)
출처 : https://www.chosun.com/site/data/html_dir/2020/08/10/2020081000556.html



<그림 3> '젯슨 가족(The Jetsons) (1962~1963)'에 이미 등장했던 홈닥터와 홈트레이닝

면 생활이 이미 60년 전 만화에서도 예견되었다는 것은 그리 놀랍지 않다〈그림 3〉.

비대면 시대에 자연스러운 소통과 콘텐츠 생성을 위해서는 영상 인식, 음성 인식, 음성 합성, 문자 인식, 동작 인식, 아바타 생성 기술들이 필요하다. 또한 이 다양한 기술들을 엮는 서비스 개발도 필수다. 특히 최근에 인식 기술에서 생성 기술로 관심이 옮겨가고 있다. 2020년만 해도 게임을 생성해주는 NVIDIA GameGAN[1], 음악을 생성해주는 OpenAI Jukebox[2], 글이나 대화를 생성해 주는 GPT-3[3] 등이 큰 화제가 되었다. 특히 GPT-3의 경우에는 뉴스, 시, 프로그램 코드를 작성하는 높은 수준의 능력으로 범용 AI의 등장이 머지 않았다는 기대와 두려움을 모두 안겨주었다.

음성과 대화 기술 수준이 올라감에 따라, AI를 좀 더 인간의 모습과 가까운 아바타로 형상화하려는 시도도 활발해지고 있다. 머니브레인과 ObEN의 AI 아나운서는 미디어 콘텐츠 생성에 실사(photorealistic) 아바타를 활용하는 것에 대한 기대를 높이고 있다〈그림 4〉. 네이버 제페토는 비실사(non-photorealistic) 아바타 캐릭터 기반 가상 세계(virtual world) 콘텐츠가 패션, 엔터테인먼트, 마케팅 등에 활용 가능하다는 것을 보였다〈그림 5〉. 이를 통해 빅히트엔터테인먼트와 YG엔터테인먼트로부터 총 120억원 규모의 투자를 유치하기도 했다. 일본의 키즈나아이(kizunaai.com)는 비실사 아바타 캐릭터로 이미 SNS, 영상 스트리밍 서비스에서 500만 명 이상의 구독자를 가지고 있다.

<그림 4> 머니브레인(moneybrain.ai)과 ObEN(oben.me)의 실사 아바타 기반 AI 아나운서

<그림 5> 네이버 제페토(naverz-corp.com)가 제작한 BTS와 블랙핑크 비실사 아바타 캐릭터

II. 콘텐츠 생성 기술과 서비스

네이버 클로바는 비대면 시대에 자연스러운 소통과 콘텐츠 생성에 쓰이는 음성 합성 기술과 동작 인식, 아바타 생성 기술 개발에 집중하고 있다.

1. 음성 합성 기술과 더빙 서비스

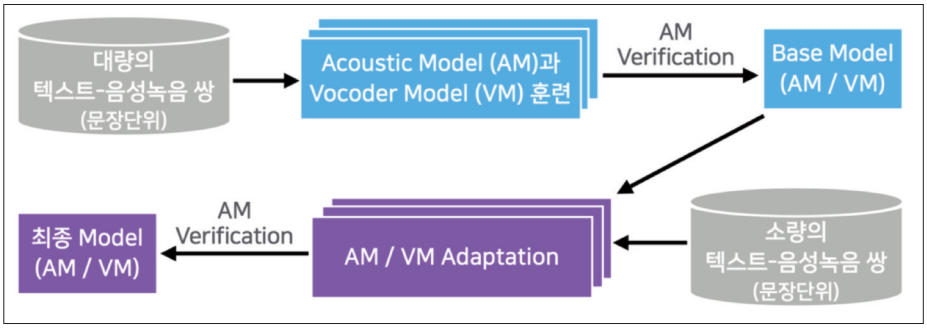
구글이 2018년 발표한 타코트론 2(Tacotron 2) [4]의 경우, 녹음한 실제 음성에 준하는 자연스러운 합성 음성으로 화제가 되었다: 구글 데모 사이트(google.github.io/ tacotron/publications/tacotron2). 음성 합성의 정확도는 '평균 의견 점수'(mean opinion score, MOS)로 측정한다. 5점 만점에서 녹음된 음성의 MOS 점수가 4.582인데 타코트론 2로 생성한 음성은 4.526로 비슷한 점수를 얻었다. 하지만 타코트론 2에서 사용한 웨이브넷(WaveNet) 기반 보코더(vocoder)의 경우, 1초 음성을 생성하는데 5~10분이 걸려서 실시간 서비스에는 맞지 않는다.

네이버 클로바에서는 음질(quality) 저하는 최소로 하고 실시간 생성 속도(speed)는 확보하는 독자 기술[5]을 가지고 있다. 네이버 클로바의 자연스러운 한국어와 일본어 음성 서비스의 경우, 실제 녹음한 음성보다 합성음이 더 자연스럽다는 피드백을 듣곤 한다: 클로바 데모 사이트(clova.ai/voice).

코로나 시대로 인해 원격수업을 준비하기 위해서 대부분의 교사분들이 많은 고민을 하셨을 때, Clova Dubbing(clovadubbing.naver.com)은 적기에 출시되어서 글자를 입력하면 AI 합성음을 생성하여 동영상 또는 PDF 수업 자료 위에 더하는 서비스를 제공했다. 2020년 2월 출시 이후 20만 명의 사용자가 2000만 건의 음성을 생성하였다.

또한 나눔 AI 보이스(clovadubbing.naver.com/event/nanumvoice)는 누구나 집에서 핸드폰을 통해서 40분 가량 400문장을 녹음하면 그 음성 화일을 사용해서 개인화 보이스 폰트를 만들 수 있게 하였다[6]. 기존 보이스 폰트의 경우, 성우와 전문 스튜디오에서 40~100시간을 녹음하는 작업이 필요했던 것을 감안하면 개인화 AI가 현실로 다가왔다는 것을 알 수 있다.

NES(Natural End-to-End Speech Synthesis)의 개인화 보이스 폰트 개발 과정은 〈그림 6〉과 같다: 각 화자 별로 음향 모델(acoustic model)과 보코더 모델(vocoder model) 한 쌍을 만드는 것을 최종 목적으로 한다. 먼저 수만 문장-음성 쌍으로 학습하여 베이스 모델(base model)을 선정한다. 이때 여러 모델을 비교해서 음향 모델 검증(acoustic



<그림 6> 커스텀 보이스 모델 제작 과정

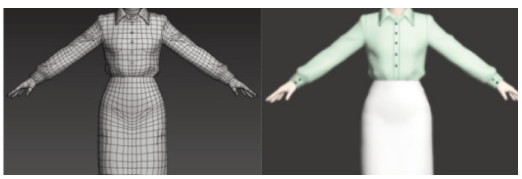
model verification) 과정을 통해서 가장 정확한 베이스 모델을 선택한다. 그 후에 소량의 화자 별 녹음 데이터로 세부 튜닝(fine tuning)을 수행하는 모델 적응(model adaptation) 과정을 거쳐 최종 모델을 확보한다.

2. 실시간 동작 인식을 통한 아바타 생성 기술

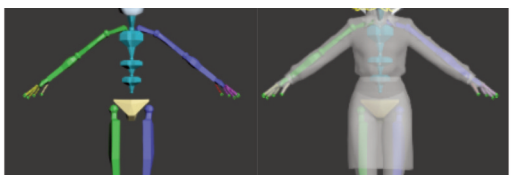
3D 아바타 캐릭터 콘텐츠 제작 과정은 3D 캐릭터 모델링(modeling)〈그림 7〉-애니메이션(animation)-렌더링(rendering) 과정을 거친다. 애니메이션의 경우, 각 프레임(frame) 별로 캐릭터 관절 위치 정보가 필요하다. 사용자 움직임을 따라서 움직이는 캐릭터 콘텐츠를 만들고자 한다면, 특별한 슈트(suit) 또는 마커(marker)를 사용하거나 복수 개의 카메라가 설치된 특별히 고안된 스튜디오에서 촬영하는 방법이 있다. 하지만 이 측정 방법은 고가의 장비와 세팅이 필요하다.

네이버 클로바는 iOS와 Android 모바일에서 실시간 30fps 이상의 속도로 관절을 추출하고 자연스러운 동작을 생성하는 기술을 개발했다[7,8]. Clova HPSDK(Human Pose Software Development Kit)〈그림 9〉를 연동하면 누구나 모바일을 통해 자신의 동작을 따라하는 아바타 콘텐츠를 제작할 수 있다〈그림 10〉. 별도 센서 없이 RGB 영상 정보에서 관절 위치를 추출하고, 그 위치 정보에 맞게 리타게팅(retargeting)을 통해 뼈대를 움직일 때 리깅(rigging)되어 있는 캐릭터의 메쉬(mesh)들은 따라 움직인다〈그림 8〉.

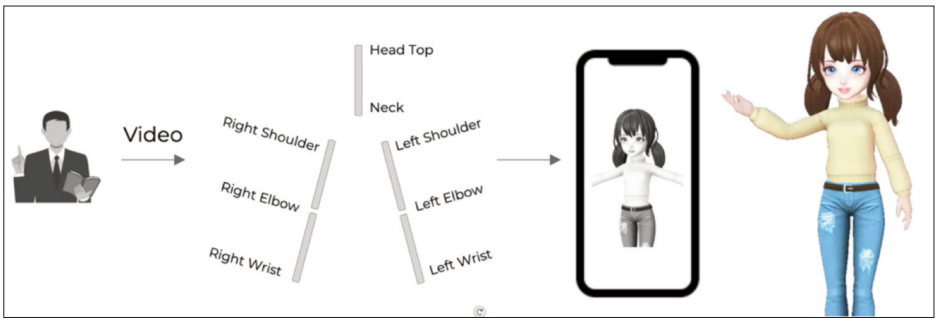
실시간 동작 인식이나 자연스러운 동작 생성 기술은 화상 회의나 영상 통화, 또는 아바타 생성 서



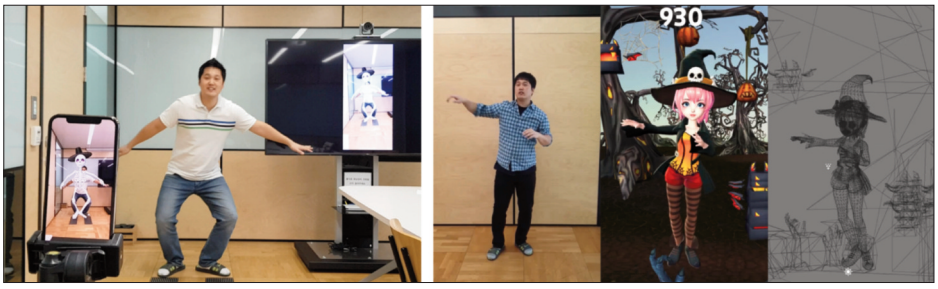
<그림 7> 캐릭터 모델링 designed by 이호진



<그림 8> 캐릭터 뼈대 리깅(Rigging) credit: 이호진



<그림 9> Clova HPSDK : 모바일에서 카메라 입력으로부터 상반신 관절 실시간 출력



<그림 10> 실시간 모션 캡쳐

비스에 적용되어 사용자의 동작을 따라하는 개인화 아바타를 만들 수 있게 한다. 또한 홈닥터, 홈트레이닝, 키오스크, 챗봇 등 사용자 응대에서도 요구되는 기술이다.

Ⅲ. 결 론

코로나로 인해서 비대면의 중요성이 갑작스럽게 대두되었고, 코로나가 종식된 후에도 이 흐름은 한동안 이어져서 새로운 인터페이스와 소통 수단에 대한 혁신은 계속될 것이다. AI 기술 발전의 속도는 그 가운데 늦춰지지 않았고, 더욱더 새로운 가능성을 제시하고 있다. 입력에서 패턴을 인식하던 지적 능력에서, 모방해서 생성하는 창조 능력까지 갖추게 되었고, 그 결과물에 대해 인간이 어색함을 느끼지 않는 수준에 이르렀다. 오히려 높은 품질에 감동을 하고 있다. 아직 범용 AI에 가까워졌다고 하기에는 부족하지만, 특정 도메인에서는 인간의 업무를 대신하거나 개인화된 AI를 가질 수 있는 시대는 이미 도래했다. 네이버 클로바의 음성 합성 기술이나 동작 인식, 아바타 생성 기술은 이미 그 여정을 시작했고, 앞으로가 더 기대된다.

참 고 문 헌

[1] https://blogs.nvidia.com/blog/2020/05/22/gamegan-research-pacman-anniversary/
[2] https://openai.com/blog/jukebox/
[3] https://github.com/openai/gpt-3
[4] TACOTRON2: Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions, Jonathan Shen and Ruoming Pang and Ron J. Weiss and Mike Schuster and Navdeep Jaitly and Zongheng Yang and Zhifeng Chen and Yu Zhang and Yuxuan Wang and RJ Skerry-Ryan and Rif A. Saurous and Yannis Agiomyrgiannakis and Yonghui Wu. (ICASSP 2018)
[5] Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram, Ryuichi Yamamoto and Eunwoo Song and Jae-Min Kim. (ICASSP 2020)
[6] 누구나 만드는 내 목소리 합성기 II (커스텀 보이스 파이프라인) https://deview.kr/2020/sessions/354 (DEVIEW 2020)
[7] 나를 따라하는 아바타: 모델 개발부터 모바일에 적용하기까지 https://deview.kr/2020/sessions/395 (DEVIEW 2020)
[8] Lightweight 3D Human Pose Estimation Network Training Using Teacher-Student Learning, Dong-Hyun Hwang and Suntae Kim and Nicolas Monet and Hideki Koike and Soonmin Bae. (WACV 2020)

필 자 소 개

배순민

– 2003년 : KAIST Computer Science 학사
– 2005년 : MIT EECS / CSAIL 석사
– 2009년 : MIT EECS / CSAIL 박사
– 2010년 ~ 2017년 : 삼성 테크윈 로봇사업부
– 2018년~ : NAVER Clova