Brain's Pick

음성은 미래의 인터페이스다

1) 지난해 종결한 KBS2 드라마 '고백부부'에는 오랜 기간에 걸쳐 쌓인 오해와 갈등으로 이혼을 결심한 진주(배우 장나라)와 반도(배우 손호준) 커플이 등장한다. 어느 날 갑자기 1999년 스무 살 대학 시절로 시간 여행을 떠난 두 사람은 젊은 날 자신들의 모습을 되돌아보며 서로의 소중함을 깨닫는다. 극 중에서 갑작스럽게 과거로 돌아간 진주는 신장염으로 투석 치료를 받다 돌아가신, 꿈에서도 그리던 엄마 은숙(배우 김미경)과 재회한다. 현재로 돌아와서도 엄마를 기억하고 싶었던 진주는 노래방에서 노래하는 엄마의 목소리를 녹음한 테이프를 책상 서랍 안에 넣어 고이 간직한다.

2) 세계 최고의 성악가로 칭송받던 루치아노 파바로티(Luciano Pavarotti). 아쉽게도 지난 2007년 췌장암으로 세상을 떠났으나[*1], 팬들은 콘서트 영상이나 음반 CD를 통해 11년째 그를 추억하고 있다. 전 세계적으로 큰 사랑을 받았던 파바로티도 피해 가지 못한 질타가 있었다. 바로 자신의 모국어인 이탈리아어로 쓰여진 작품으로만 활동했다는 점[*2]이다. 이탈리아어, 프랑스어, 독일어, 영어로 작성된 오페라 작품이 주를 이루는 만큼 다국어에 능통하지 못한 성악가는 치명적인 핸디캡을 가질 수 밖에 없다. 파바로티의 경쟁자였던 플라시도 도밍고(Placido Domingo)가 모국어인 스페인어는 물론 프랑스어, 이탈리아어, 독일어 등 세계 각국의 언어로 쓰여진 작품 속 배역을 맡았던 것과는 대조적인 모습이다.[*3]

만약 진주가 엄마의 목소리로 부른 노래를 수십, 수백 곡 가지고 있었다면 엄마를 향한 그리움을 달래는 데 큰 도움이 됐을 것이다. 파바로티도 마찬가지다. 이탈리아어로 작곡된 작품에 한정해 음악 활동을 펼치던 파바로티가 세계 각국 오페라 작품 속 배역을 맡았다면 활동 스펙트럼을 더 넓힐 수 있었을 것이다. 이런 발칙한 상상을 해볼 수 있는 배경에는 바로 인공지능을 활용한 음성 변환(voice conversion) 기술이 있다.

글 | 이수경 samantha.lee@kakaobrain.com

2016년 3월 알파고와 이세돌 9단이 펼치는 세기의 대결을 두 눈으로 목도한 이후 인공지능을 제대로 공부해봐야겠다고 결심했습니다. 인공지능 본진이자 연구소인 카카오브레인으로 걸어 들어온 이유입니다. 인공지능 기술과 이로 인해 바뀔 미래 사회를 다루는 글을 통해 사람들과 소통하고 싶습니다.

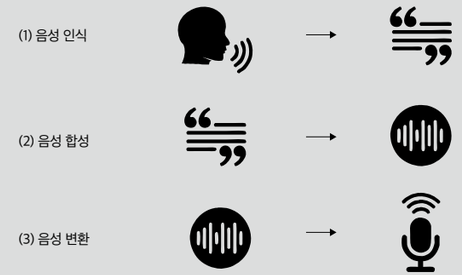
글/감수 | 안다비 avin.hero@kakaobrain.com

언뜻 보면 예쁜 한글 이름 같지만, 많고 크다(多丕)는 뜻을 가지고 있습니다. '욕심 많은' 이름 그대로 매사에 의욕이 넘치지만 가끔은 생각에 압도 당하기도 합니다. 컴퓨터 과학을 6년 간 공부하고 엔지니어로서 5년을 살았습니다. 또 어렸을 적부터 인공지능과 같은 최첨단 기술을 좋아했습니다. 전세계를 놀라게 한 알파고가 인공지능을 핫키워드로 만드는 것을 보며 시대를 잘 타고 났다고 느끼는 중입니다. 구글 검색, 넷플릭스처럼 지구 상의 많은 이들이 좋아할 유명 서비스를 만들고 싶은 자칭 야망가. 적당한 때를 기다리며 엔지니어로서 계속해서 열정적으로 일하고 있습니다.

음성 변환이 유력한 4개 분야

음성 변환을 이해하려면 음성 인식(speech recognition)과 음성 합성(voice synthesis)에 대한 이해가 전제되어야 한다. [그림 1]은 음성 인식과 음성 합성, 음성 변환의 차이를 보여준다. 음성 인식은 특정인이 말한 내용을 텍스트 형태로 이해하는 것이며, 음성 합성은 텍스트를 음성으로 변환함을 의미한다. 음성 합성과 음성 변환 모두 음성을 결과물로 생성한다. 그러나 음성 변환은 음성을, 음성 합성은 텍스트를 입력값으로 활용한다는 차이가 있다.

[그림 1] 음성 인식과 음성 합성, 음성 변환



언뜻 보면 음성 변환은 A라는 사람의 음성을 텍스트로 인식한 뒤, B의 목소리로 말하도록 하는 흐름처럼 보인다. 자칫 '음성 인식+음성 합성=음성 변환'으로 인식될 수 있다는 의미다. 사실은 이와 다르다. 음성 합성의 중간 매개체는 반드시 텍스트다. 반면 음성 변환에서는 /ㅁ/, /ㅂ/, /ㅍ/와 같은 소리의 최소 단위인 음소(phoneme)와 같은 추가 매개 정보를 활용하거나, 혹은 어떤 중간 매개 정보도 활용하지 않고 사람 A의 음성에서 사람 B의 음성으로 직접 변환하기도 한다.

카카오브레인 연구팀은 특정인의 목소리를 제 3자의 것으로 손쉽게 변환할 수 있다면 이 음성변환 기술을 활용할 수 있는 어플리케이션의 잠재적 가치가 클 것으로 분석하고 있다.

첫 번째 활용 분야는 오디오 북(audio books)이다. 원하는 사람의 목소리로 책[*4]이나 편지를 듣고 싶은 사람들이 고객이 될 수 있다. 또는 자녀 교육에도 활용해볼 수 있다. 잠자리에서 아이에게 동화책을 읽어주는 게 중요하다는 것은 잘 알지만[*5], 이를 실천하기란 막상 쉽지 않다. 부모 모두 다음날 출근 때문에 마음의 여유가 없다면 매일 밤 동화를 읽어주는 게 여의치 않을 수 있기 때문이다. 부모의 빈 자리를 대신하기에 현재의 텍스트 음성 변환(text to speech, TTS)은 지나치게 인위적이다. 만약 엄마나 아빠의 목소리로 동화를 읽어주는 인공지능이 있다면, 부모는 자녀에게 매일 책을 읽어줘야 한다는 부담으로부터 해방될 수도 있을 것이다.

두 번째 분야는 엔터테인먼트다. 자신이 좋아하는 가수가 자신이 좋아하는 노래를 부르도록 하는 것이다. 카카오브레인 연구팀은 개그맨 박명수가 가수 이승철처럼 노래 부르는 유튜브 영상[*6]이 큰 인기를 끄는 것을 본 이후, 이 기능이 새로운 어플리케이션이 될 가능성을 발견했다.

세 번째 분야는 언어교육 서비스다. 음성 변환 기술을 활용해 자신이 원어민처럼 외국어를 구사한다면 어떤 느낌인지 미리 들어보며 발음을 교정해나갈 수 있다.

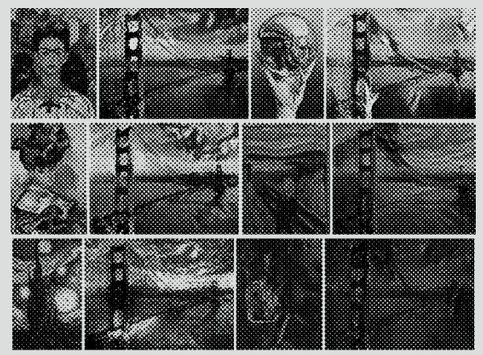
마지막 분야는 외국어 더빙이다. 현재는 극중 배우와 다른 목소리의 현지 성우의 목소리가 덧입혀지는데 그친다. 만약 음성 변환을 활용한다면 극중 배우의 목소리로 다른 나라 언어를 구사하는 것이 가능해진다. 예를 들어, 한국어를 배우지 않은 영화 배우 라이언 고슬링(Ryan Gosling)이 한국어로 배역을 소화하는 장면을 볼 수 있다.

음성 스타일 변환(Voice Style Transfer) 프로젝트

카카오브레인 안다비 연구원과 박규병 연구원은 지난해 하반기부터 음성 합성 기술을 활용한 음성 스타일 변환 프로젝트를 진행했다. 깃허브(GitHub)에 프로젝트 관련 내용을 일목요연하게 정리한 README[*7]문서로 예상을 뛰어넘는 흥행에 성공했다. 레딧(Reddit)의 기계 학습(machine learning) 카테고리에서 1위를 기록했던 것이다.[*8]

카카오브레인 연구팀은 프로젝트가 이처럼 높은 관심을 받을 수 있었던 배경을 3가지로 분석했다. 첫 번째는 TTS와 같은 음성 분야가 전 세계적으로 인기를 끌기 시작한 시점에 적절히 깃허브 링크를 공개했다는 점이다. 두 번째, 2015년도 이미지 영역에서 스타일 변형(style transfer)[*9]이 주목을 받으며 영상뿐만 아니라 음성의 스타일을 변형하는 구현 사례들도 속속들이 등장하며 관련 기술이 받기 시작했다. 세 번째, 프로젝트의 대상 화자 목소리 주인공인 할리우드 배우 케이트 윈슬렛(Kate Winslet)이 가진 명성도 일정 부분 영향을 미쳤을 것이라는 분석이다.

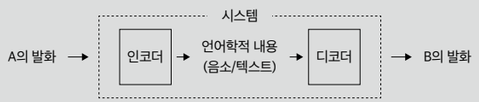
[그림 2] 기계 학습을 통한 이미지 변환 예시[*10]



음성 변환의 학습 모델 구성

[그림 3]에서 보듯이, 음성 변환의 학습 모델은 크게 인코더(encoder)와 디코더(decoder)로 구성된다. 인코더는 원 화자(source speaker)의 음성을 인식하여 무슨 말을 하는지 알아내는 역할을 한다. 디코더는 인코더로 파악된 특성을 기초로 대상 화자(target speaker)의 목소리를 생성한다. 다시 말하면 인코더는 무슨 말을 하는지(linguistic feature), 어떻게 말하는지(acoustic feature)와 같은 특징(feature)을 인식한다. 디코더는 이런 특징을 활용해 음성을 생성하는 역할을 담당한다.

[그림 3] 카카오브레인이 진행했던 케이트 윈슬렛 음성 변환 구조



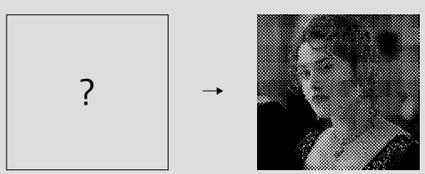
원 화자(source speaker)는 실제로 말을 하는 사람을 지칭하고, 대상 화자(target speaker)는 변환하고자 하는 목소리 스타일을 보유한 사람을 뜻한다.

이렇게 설계한 이유는 원 화자가 누구냐에 관계 없이 대상 화자의 목소리로 변환할 수 있어서다. 즉, 해당 구조는 다대일(many-to-one) 시스템이다. 한 명의 화자 목소리를 한 명의 대상 화자로 전환하는 일대일(one-to-one) 구조와 비교하면 응용 분야가 더 넓다. 더 많은 사용자에 대응할 수 있기 때문이다.

현실에서 구하기 어려운 병렬 데이터셋(parallel dataset)이 필요 없다는 것도 이 모델의 장점이다. 여기서 병렬 데이터셋이란 쌍을 이루는 데이터 집합을 의미한다. 예를 들어, <원 화자의 발화, 대상 화자의 발화>의 쌍 혹은 <대상 화자의 발화, 대상 화자가 말한 내용을 옮겨 적은 텍스트> 쌍과 같은 데이터가 없어도 대상 화자의 목소리 데이터셋 만으로 모델을 가르칠 수 있다.

카카오브레인 연구팀은 케이트 윈슬렛이 녹음한 오디오북을 학습 데이터로 활용했다. 목소리의 품질을 원하는 수준으로 높이고자 모델의 형태와 설정을 수없이 바꿔가며 실험을 진행했다.

[그림 4] 원 화자의 음성을 케이트 윈슬렛의 목소리로 바꾸는 것이 프로젝트의 핵심[*11]



그런데 모델을 학습시키고 추론(inference) 단계에 접어들었을 때 원 화자와 대상 화자 간 데시벨 차이가 클수록 좋은 결과를 얻지 못하는 경향이 있었다. 이를 해결하기 위해 카카오브레인 연구팀은 추론 단계에서 원 화자의 음성 데시벨을 조정하거나 모델 학습 단계에서 데이터 어그멘테이션(data augmentation)[*12]을 통해 여러 데시벨의 데이터를 가르쳤다.

데시벨의 값 범위를 0에서 1 사이의 값으로 재조정해 본 결과, 학습 속도나 학습 안정성, 성능 면에서 크게 도움이 됐다. 모델 내 한 레이어로부터 나온 값을 정규화(normalization)하는 것은 물론, 최초의 입력 데이터 자체의 정규화가 주효했던 것으로 분석된다.

인코더와 디코더 학습

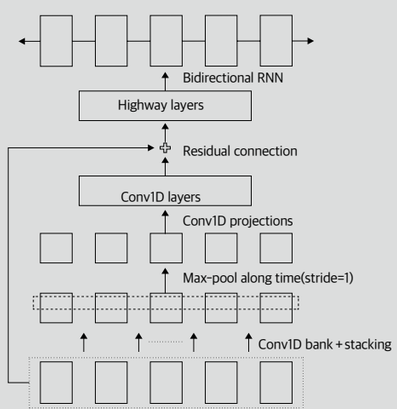
인코더와 디코더는 각각 데이터로부터 독립적으로 학습한다. [그림 3]과 같이 인코더는 지도 학습(supervised learning)[*13]을 통해 A(원 화자)가 말하는 소량의 (발화, 음소) 쌍으로부터 언어적 내용을 배운다. 인코더의 정확도는 80% 수준이다. 한편 디코더는 대립 학습(adversarial training)[*14]의 한 모델인 WGAN(Wasserstein GAN)으로 B(대상 화자)의 발화 데이터를 학습한다. 그중에서 생성자(generator) 부분은 CBHG 모듈을 부분적으로 활용하고 있다. 하지만 모델 학습 시에 기울기 폭발(gradient explosion)[*15]이 자주 발생한다. 값에 최대값 또는 최소값을 설정하는 기울기 클리핑(gradient clipping) 등의 기법을 활용할 수는 있지만, 이는 완벽한 솔루션이 아니라는 결론을 내렸다.[*16]

CBHG

2017년 구글이 발표한 TTS 논문인 타코트론(Tacotron)에서 사용한 신경망 블록이다. convolutional 1-D filters, bank, highway networks, gated recurrent unit bidirectional의 첫글자를 따서 지어진 이름이다. 단어가 나열된 순서대로 각 블록을 쌓아 만든 구조이기 때문이다.

부분적인 특징을 잡아내는데 유용한 CNN(convolutional neural network)과 연속적인(sequential) 데이터를 생성하기 좋은 RNN(recurrent neural network)의 특징을 모두 활용했다. 수초짜리의 발화(utterance) 데이터를 정해진 크기만큼 순차적으로 학습하고, 목소리 특징을 뽑아내는(CNN) 동시에, 뽑혀진 부분 특징을 전체적으로 보면서 결과를 만들어 낸다(RNN).

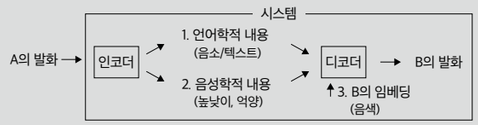
[그림 5] CBHG 구조[*17]



프로젝트의 한계와 개선 사항

범용적인 서비스로 나아가기 위해서는 카카오브레인 연구팀의 프로젝트가 개선되어야 할 한계점은 분명히 있다. 현재는 한 명의 대상 화자 목소리로만 전환할 수 있다. 케이트 윈슬렛 목소리를 학습한 모델은 레오나르도 디카프리오(Leonardo DiCaprio) 목소리를 흉내낼 수 없다는 의미다. 전 세계 모든 음성 합성 연구가 부딪히는 한계점이기도 하다. 아울러 프로젝트 모델에서 인코더는 화자가 전달하려는 내용(linguistic content)만 추출하고, 원 화자의 말하기 특징(acoustic feature)에 대한 정보는 담지 못한다. 관련된 예시로 말의 높낮이(intonation)가 여기에 해당한다.

[그림 6] 원 화자의 말하기 특징(높낮이, 억양, 음색) 학습을 위해 개선한 모델 구조



현재 카카오브레인 연구팀은 원 화자의 말하기 특징을 반영하는 모델로 구조를 개선하고 있다. 처음 고안한 버전은 [그림 6]처럼 원 화자가 전달하려는 메시지와 말하기 특징(언어학적 내용), 대상 화자의 음색(음성학적 내용)을 명확히 구분하는 것이 핵심이다. 다시 말하면 음성에서 원하는 특징을 추출하고, 이를 토대로 임의의 음성을 완벽하게 생성하는 것이 최종 목표라고 할 수 있다.

이 구조는 인코더와 디코더, 둘 중 하나라도 에러율이 높으면 모델의 성능이 전반적으로 저하된다는 한계를 지닌다. 이를 극복하기 위해 이미지 스타일 변형에서 얻은 통찰력을 활용해 인코더와 디코더를 단일 네트워크로 통합하기 위한 작업을 진행 중이다. 입력층에 가까운 쪽에서는 사람의 말하기 특징을 추출하고, 출력층에 가까운 쪽에서는 컨텐츠 정보를 추출하는 식이다.

또 다른 한계는 원하는 수준의 목소리 품질을 얻으려면 잡음(noise)이 없는 깨끗한 음질을 다량 확보해야 한다는 것이다. 한 사람 목소리로 적어도 24시간 이상 분량의 녹음 데이터가 필요한데, 현실에서는 이런 데이터를 쉽게 얻을 수 없고 얻을 수 있더라도 스튜디오 녹음 및 편집 비용이 많이 든다. 기존 보유하고 있는 데이터셋은 크기 자체가 부족해, 배우지 않은 발음과 억양이 생길 수밖에 없다. 따라서 어느 정도 잡음이 있는 데이터로도 깨끗한 목소리를 생성할 수 있는 강인한(robust) 인공지능을 만드는 연구도 진행될 필요성이 있다.

카카오브레인 연구팀은 올해 신규 아키텍처를 기반으로 상용화가 가능한, 높은 수준의 음성변환 시스템을 구축하고자 한다. 자가 학습을 통해 음성에서 특징을 추출할 수 있도록 하고자 오랜 시간과 노력을 투자할 계획이다.

가까운 미래에 여러 화자의 목소리를 흉내내는 기술이 출현할 것으로 기대된다. 카카오브레인 연구팀은 오랜 시간 반복 학습을 하지 않고도 원 화자의 말하기 스타일을 단시간 내에 익혀서 이를 곧잘 따라하는 인공지능 구현을 최종 목표로 하고 있다. 이 목표에 도달하기 위해서는 한 사람이 특정 언어를 자연스럽게 구사하는 데 필요한 최소 단어 세트의 범위를 정의하는 과제를 선결적으로 해결해야 한다.

음성 합성 관련 추천 논문

카카오브레인 연구팀은 음성 합성 프로젝트를 진행하기에 앞서 선행 연구 조사 및 재현을 진행했다. 그런 뒤 여러 연구로부터 얻은 기법과 통찰을 최대한 활용해 카카오브레인 만의 모델 구조를 설계하는 데 집중했다. 자연어 처리(NLP) 관련 주요 논문은 박규병 카카오브레인 연구원이 일목요연하게 정리한 깃허브 README[*18] 문서를 참고할 것을 추천한다.

***1** 참고 | http://v.media.daum.net/v/20171013111117091 ***2** 참고 | 파바로티는 오로지 이탈리아, 스페인어 가사만?, 메트 사상 가장 많은 커튼콜을 받은 스타는? 프랑스어 오페라 전곡 녹음으로는 '연대의 딸(La fille du regiment)'이 유일하다. 독일 오페라는 아예 없다. 모차르트의 오페라 중 이도메네오(Idomeneo)를 부른 적이 있으나 이는 애초에 이태리어로 작곡된 작품이다. ***3** 참고 | 시대를 앞서간 예술상인, 도서 '빅 파바로티' ***4** 참고 | 오디오북 '100인의 배우, 우리 문학을 읽다' ***5** 참고 | 잠자리에서 아이에게 읽어주는 동화책의 중요성 ***6** 참고 | https://www.youtube.com/watch?v=C8tW-1xVINw 실제로는 이승철이 부른 음성 파일의 키를 낮춘 것인데, 마치 박명수가 이승철을 모사하는 것처럼 들린다. ***7** 참고 | https://github.com/andabi/deep-voice-conversion/blob/master/README.md ***8** 참고 | 레딧은 우리나라의 '인공지능계 클리앙'과 같은 글로벌 커뮤니티다. 전세계 머신러닝 관련 글과 링크가 이곳에 공유되며 사람들이 좋아요(upvote)한 콘텐츠의 순위가 올라간다. 등수는 흥행도를 뜻한다. ***9** 참고 | https://github.com/jcjohnson/neural-style ***10** 참고 | https://github.com/jcjohnson/neural-style ***11** 참고 | http://ew.com/movies/2017/12/19/kate-winslet-titanic-20th-anniversary/ ***12** 참고 | 원 데이터(source data)를 활용해 새로운 데이터를 생성하는 것을 의미한다. 예를 들어, 음성 데이터에 화이트 노이즈를 추가하거나, 볼륨을 임의로 높이거나 낮추고, 음의 높낮이를 조절하는 것이 여기에 해당한다. ***13** 참고 | 정답이 라벨링된 데이터를 이용해 학습 모델을 찾는 과정을 의미한다. 그러고 나면 학습하지 않은, 완전히 새로운 데이터가 주어졌을 때 자동으로 답을 추론한다. ***14** 참고 | [카카오AI리포트]Do you know GAN?(½) 새로운 데이터를 생성하는 생성자(generator)와 이 데이터를 평가하는 식별자(discriminator)가 서로 대립하며 각각의 성능을 높이는 방식으로 학습하는 것을 의미한다. BEGAN, WGAN 등 다양한 모델이 있다. ***15** 참고 | 오차 함수의 기울기가 커지면 무한대로 접근한다. ***16** 참고 | 이같은 이유로 카카오브레인 연구팀은 최근 구글이 발표한 타코트론2(Tacotron2)에서는 CBHG 블록이 배제된 것으로 분석한다. ***17** 논문 | Wang, Y. et al. (2017) Tacotron: Towards end-to-end speech synthesis. doi: arXiv:1703.10135v2 ***18** 참고 | https://github.com/Kyubyong/nlp_tasks/blob/master/README.md#voice-conversion