|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Pitch Mean [Hz] | Pitch Minimum [Hz] | Pitch maximum [Hz] | Intensity [dB] | First Formants [Hz] | Second Formants [Hz] |
| 원본 | 225.223 | 146.282 | 317.045 | 76.120 | 740.549 | 1708.702 |
| 딥 보이스 | 208.906 | 132.081 | 293.775 | 76.894 | 706.850 | 1527.912 |
| 딥 보이스  이전 목소리 | 128.945 | 76.526 | 182.958 | 81.565 | 520.822 | 1469.361 |
| 제3자 목소리 | 183.218 | 82.09 | 297.786 | 70.535 | 667.95 | 1742.348 |

테일러 스위프트 표

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Pitch Mean [Hz] | Pitch Minimum [Hz] | Pitch maximum [Hz] | Intensity [dB] | First Formants [Hz] | Second Formants [Hz] |
| 원본 | 134.122 | 109.269 | 177.37 | 63.663 | 398.958 | 1594.235 |
| 딥 보이스 | 123.235 | 103.828 | 218.708 | 78.858 | 770.744 | 1865.661 |
| 딥 보이스  이전 목소리 | 110.863 | 92.853 | 291.973 | 86.466 | 578.862 | 1800.633 |
| 제3자 목소리 | 132.662 | 120104 | 157.015 | 81.068 | 309.434 | 1714.905 |

조바이든

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Pitch Mean [Hz] | Pitch Minimum [Hz] | Pitch maximum [Hz] | Intensity [dB] | First Formants [Hz] | Second Formants [Hz] |
| 원본 | 115.172 | 77.075 | 75.082 | 67.885 | 640.336 | 1994.463 |
| 딥 보이스 | 116.716 | 81.726 | 162.124 | 75.763 | 643.534 | 1918 |
| 딥 보이스  이전 목소리 | 117.384 | 87.655 | 86.894 | 81.005 | 637.008 | 1982 |
| 제3자 목소리 | 134.478 | 106.119 | 171.312 | 76.837 | 524.375 | 82.809 |

오바마

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Pitch Mean [Hz] | Intensity [dB] | First Formants [Hz] | Second Formants [Hz] | | 원본 | 122.445 | 61.368 | 455.864 | 1331.997 | | 딥 보이스 | 116.156 | 81.188 | 694.381 | 721.701 | | 딥 보이스  이전 목소리 | 115.343 | 88.051 | 718.154 | 991.407 | | 제3자 목소리 | 103.393 | 82.617 | 808.885 | 1235.9 |   Are 발음   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Pitch Mean [Hz] | Intensity [dB] | First Formants [Hz] | Second Formants [Hz] | | 원본 | 115.781 | 72.895 | 872.929 | 1910.349 | | 딥 보이스 | 107.016 | 76.165 | 487.122 | 1522.748 | | 딥 보이스  이전 목소리 | 112.052 | 86.582 | 492.734 | 1682.691 | | 제3자 목소리 | 102.428 | 71.576 | 305.79 | 1844.283 |   Yes 발음 |

|  |
| --- |
| **\*김지후, \*류세민, \*한채림, 이창의, 이종현, 이민재, 오민규, 박세진**  **계명대학교 컴퓨터공학과**  **e-mail :** [**jihu0210@naver.com**](mailto:jihu0210@naver.com)**,** [**tpals3799@naver.com**](mailto:tpals3799@naver.com)**,** [**cofla0904@naver.com**](mailto:cofla0904@naver.com)  **Artificial Intelligence Imitates Human Voices:**  **The Future of Deepfake Voices**  **\*Ji-Hu Kim, \*Se-min Ryu, \*Chae-Rim Han,**  **Chang-Ui Lee, Jong-Hyeon Lee, Min-Jae Lee, Min-Gyu Oh, Se-Jin Park**  **School of Computer Engineering**  **Keimyung University** |

**Abstract**

영어 서론

I. 서론

딥보이스(Deep Voice) 기술이란 딥러닝(Deep Learning)과 음성(Voice)의 합성어로, 특정 인물의 목소리를 모방하거나 생성하는 기술이다. 정확도가 높은 딥보이스가 만들어지기 위해서는 많은 양의 데이터 수집이 필요하다. 기존의 모델들은 학습 데이터를 수집하는데 많은 시간과 비용이 들어가는 등의 문제점이 있었다.

최근에 이러한 문제점을 해결한 Micro Soft사의 VALL-E라는 모델의 등장으로 기존보다 획기적으로 적은 양의 음성샘플만으로도 특정인물의 목소리와 비슷한 딥보이스를 만들 수 있게 되었다.

기존의 딥보이스 기술에 있어서 특정인의 목소리를 딥러닝 기술로 학습하고 문자를 음성으로 자동 변환하는 TTS(Text To Sound) 기술을 사용하여 목소리를 모방하는 형태로 딥보이스가 만들어 졌었으나, 최근에 들어서 실시간으로 목소리를 특정인의 목소리로 변환하는 STS(Sound To Sound) 기술을 사용하는 프로그램도 등장했다.

딥보이스는 AI 딥러닝 기술이 발전함에 따라 주목해야 할 기술 중 하나이다. 딥보이스로 생성된 페이크 음성은 자연스럽지 않거나 감정을 잘 반영하지 못한다는 문제 등이 있었으나 AI 딥러닝 기술의 발전으로 뛰어난 음질의 음성뿐만 아니라 감정, 녹음된 분위기와 음향환경도 유지할 수 있게 되었다. 이러한 딥보이스 기술의 발전으로 그에 따른 활용도가 높아질 것으로 예상이 된다.

본 논문에서는 VALL-E의 등장으로 발전한 딥보이스 기술과, STS(Sound To Sound)기술을 사용해 생성된 딥보이스와 실제 음성 사이의 정확성 그에 따른 딥보이스의 미래에 대해서 탐구해보고자 한다.

II. 본론

2.1 VALL-E

적은 양의 음성샘플만으로 정확도 높은 딥보이스를 만들어내기 위해서는 해당 음성샘플의 특징을 찾고, 훈련 데이터에서 비슷한 특징을 가진 데이터를 통해 학습해 음성을 합성할 수 있어야 할 것이다. 이를 가능하게 하는 주요 방법 중 하나는 대규모의 음성 데이터를 사용하여 모델을 훈련하는 것이다.

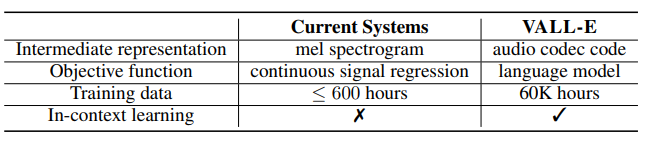


그림 1. VALL-E와 현재 시스템 간의 비교

위 그림에서 볼 수 있듯이 VALL-E는 기존의 시스템 보다 최소 1000배가량 더 많은 훈련 데이터를 가지고 있으며, 최초로 언어 모델을 기반으로 하고 있다. 특히, In-context learning을 통해서 효과적으로 학습하고 활용할 수 있게 되었다. [1]

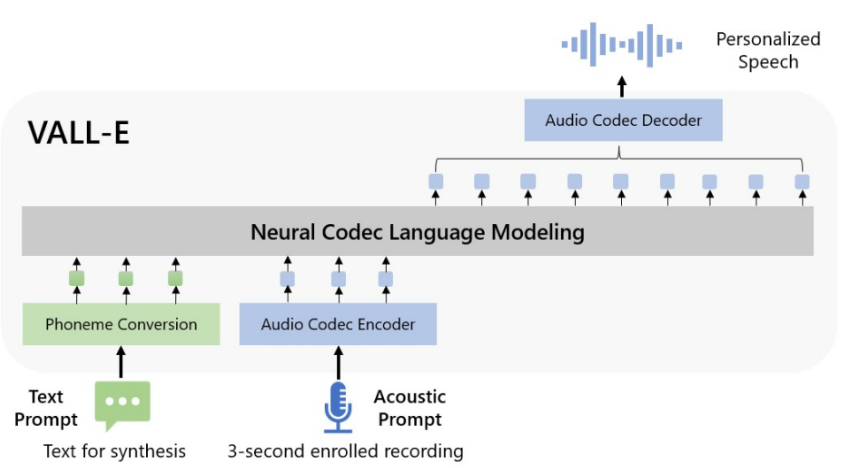


그림 2. VALL-E의 기술 구조

위 그림에서 볼 수 있듯이 음성을 합성하기 위해 VALL-E는 3초 정도의 짧은 음성샘플을 단어의 가장 최소 단위인 음소로 나누고 최종적으로 코덱 디코더를 거치며 waveform을 합성한다. 이 때 코덱 디코더를 거치기 전 inference를 통해서 다양하게 샘플링 하고 최종적으로 다양한 합성 결과를 생성한다. [2]

Ⅲ. 구현

voice.ai를 통해서 실시간으로 목소리를 특정인의 목소리로 변환하는 STS(Sound To Sound) 기술을 사용해 만들어진 딥보이스와 실제 음성 사이의 정확도를 주파수 분석 프로그램 Praat을 통해서 분석했다. [3]

Ⅳ. 결론 및 향후 연구 방향

딥보이스 기술의 발전으로 이제는 들을 수 없는 그리운 사람의 목소리를 더욱 생생하게 듣거나, 언어장애를 가진 이들을 도울 수 있을 것으로 기대가 된다.

이러한 기술의 발전이 가져올 긍정적인 영향은 제한적이지만 그에 따른 오용 또한 많아질 것이라 우려가 된다. 실시간으로 목소리를 변조하고 감정과 음향환경 또한 잘 반영할 수 있다면 보이스피싱으로 인한 피해가 더욱 증가할 것이다.

기술의 발전이 가져올 미래가 기대되지만 그에 따른 오용에 대한 대책 마련 또한 필요할 것이다.

참고문헌

[1] Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers

[2]www.microsoft.com/en-us/research/project/vall-e-x/

[3] voice.ai