딥보이스를 이용한 보이스피싱의 대응방안을 위한 음성 주파수 분석

*김지후, *류세민, *한채림, 이창의, 이종현, 이민재, 오민규, 박세진
계명대학교 컴퓨터공학전공
e - mail : jihu0210@naver.com, tpals3799@naver.com, cofla0904@naver.com, rhljh7410@gmail.com, jhyeon2450@gmail.com, qzqzcaraz00@gmail.com, alsrb4719@gmail.com, baksejin@kmu.ac.kr

Voice frequency analysis for response to voice phishing using deep voice

*Ji-Hu Kim, *Se-min Ryu, *Chae-Rim Han,
Chang-Ui Lee, Jong-Hyeon Lee, Min-Jae Lee, Min-Gyu Oh, Se-Jin Park

Department of Computer Engineering
Keimyung University

Abstract

With the development of deep learning along with artificial intelligence, ethical problems that abuse it while providing convenience to humans have increased. While the problem of voice phishing intensifies with deep voices synthesized using artificial neural networks, this paper analyzes the similarities and differences between real voices and voices coated with voice synthesis technology to explore factors that can identify voices. As a result of the paper, it was confirmed that the characteristics of the experimenter's voice were maintained in certain areas even though they were synthesized as deep voices. Through this area, it was possible to specify the voice before synthesis in the deep voice.

I. 서론

최근 인공지능(Artificial Intelligence)과 함께 딥러닝(Deep Learning) 기술이 발전함에 따라서 인간에게 편리함을 제공해 주지만, 이를 악용하는 윤리적 문제 또한 증가하였다.[1]

인공지능 중에서도 인공신경망을 사용하여 음성 합성을 한 딥보이스(Deep Voice)는 목소리를 위조하는 신종 보이스 피싱 문제를 증가시켰다.[2] 범죄 증가에 따라서 대검찰청은 '자동 화자 확인 시스템' 이라는 이름으로 증거 녹취록의 음성을 입력하면 어떤 용의자의 목소리와 일치하는지 자동으로 비교 및 분석을 해주는 프로그램을 개발하였다.

그러나 이러한 시스템은 인공지능이 만들어 내는 가짜 목소리에 대해서 제대로 인식하지 못하기에 첨단 신종 범죄에 대한 대응 방안의 필요성을 느꼈다.

본 논문에서는 실험자의 목소리에 인공 지능의 음성 합성 기술을 이용하여 유명인의 목소리로 합성해 보았다. 이를 실험자 목소리와 실제 유명인의 목소리로 비교 분석을 진행했다.

II. 본론

2.1 TTS

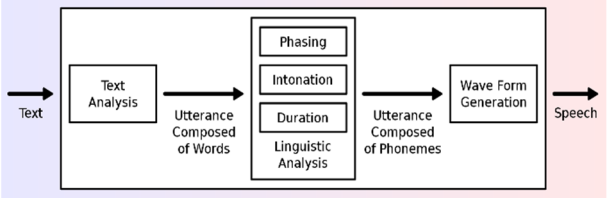

그림 1 음성 합성 기술(TTS) 전체 구조

음성 합성 기술의 기본적인 원리는 TTS(Text-to-Speech)이다. 텍스트를 오디오로 바꿔주는 이 기술은 텍스트를 통해서 출력으로 Waveform을 생성하여 음성으로 바꾸어 준다.

음성 모델로는 WaveNet[3]이나 WaveRNN 같은 모델이 있다. 텍스트 분석에서는 입력한 텍스트를 처리하여 발음을 결정하고 어절 강세, 억양의 정보를 추출한다. 음성 합성 중 텍스트 분석 부분에서 생성된 멜 스펙트로그램이 Vocoder에 입력되어 최종 오디오 출력이 만들어진다.

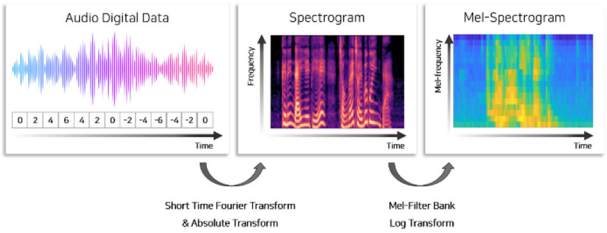

그림 2 오디오 데이터 전처리 과정

음성데이터로부터 멜 스펙트로그램[4]을 추출하기까지 3가지 전처리 과정을 거친다. 오디오 데이터 안에 있는 여러 개의 오디오를 분리한다.

두 번째 단계에서는 멜 스펙트로그램에 있는 비선형 함수를 적용하여 저주파(row frequency)영역을 확대한다. 사람의 귀는 고주파의 소리보다 저주파의 소리에서 민감하다. 저주파의 영역 확대와 동시에 고주파 영역을 축소하여 명료한 음성과 사람이 쉽게 알아들을 수 있도록 scale로 변환한다. 마지막 단계에서 log scaling을 진행하여 생성된 모델을 활용한다.

본 논문에서는 TTS 기술에서 파생된 개별적인 기술 조합인 STS(Speech-to-Speech)를 다룬다. 텍스트에서 음성이 아닌, 음성을 다른 음성으로 변환하는 시스템을 이용한다.

2.2 딥보이스

딥보이스는 인공지능 중에서도 음성 합성 기술을 이용하여 간단하게 조작할 수 있기에 이를 악용하여 사이버범죄 등에 이용될 수 있다.[5] 유사 범죄로 보이스피싱이 있으며, 해외와 국내를 가리지 않고 딥보이스를 악용한 범죄가 나타나고 있다.

2.3 분석

실제 음성을 변조하기 위한 음성 합성 방법으로는 voice.ai[6]를 사용하였다. voice.ai에는 다양한 음성 샘플들이 존재하기에 음성 합성에 어려움이 없고, 실제 목소리와 유사하다고 판단하여 선택하게 되었다.

딥보이스 합성 과정에서 사용된 음성과 비교에 사용된 대상의 음성 등을 분석하기 위해 주파수 분석 프로그램 Praat을 사용하였다.

Praat은 음성 분석으로 Pitch(음높이)와 Formant 형태를 파악할 수 있다. Formant는 성대의 울림에 의해 성도가 공진 되는 주파수로 수치가 비슷할수록 유사한 음색을 지닌다.[7] 그 외에도 Intensity(강도 분석), Speech Synthesis(음성 조합, 음성 합성) 등과 같은 정보들을 스펙트로그램과 같이 보여주기 때문에 분석에 용이하였다.[8]

표 1 음성 주파수 분석을 위해 사용한 인물과 대사

인물	대사
조 바이든	“ We are ready"
버락 오바마	“ Yes we can ”

2명의 인물을 설정하고 Praat로 음성 주파수 분석을 진행하였다.

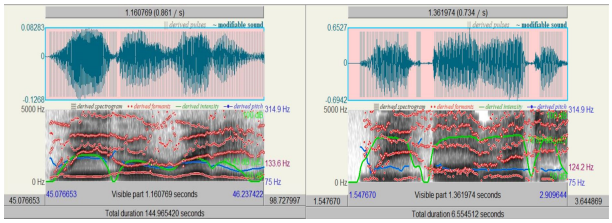

그림 3 조 바이든 목소리와 딥 보이스 주파수 분석

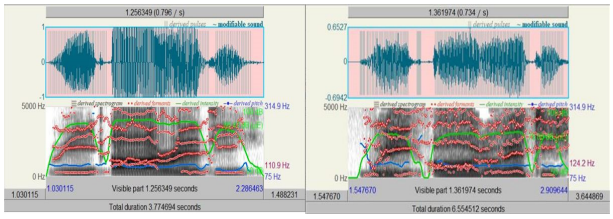

그림 4 실험자 목소리와 조 바이든 딥보이스 주파수 분석

표 2 조 바이든의 음성 주파수 분석 수치

	Pitch Mean [Hz]	Pitch Minimum [Hz]	Pitch Maximum [Hz]	Intensity [dB]	First Formants [Hz]	Second Formants [Hz]
조 바이든 목소리	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
실험자 목소리	110.863	92.853	291.973	86.466	578.862	1800.633

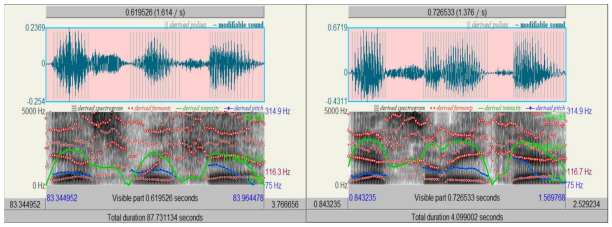

그림 5 버락 오바마 목소리와 딥 보이스 주파수 분석

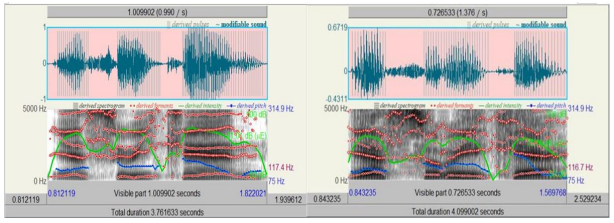

그림 6 실험자 목소리와 버락 오바마 딥보이스 주파수 분석

표 3 버락 오바마의 음성 주파수 분석 수치

	Pitch Mean [Hz]	Pitch Minimum [Hz]	Pitch Maximum [Hz]	Intensity [dB]	First Formants [Hz]	Second Formants [Hz]
오바마 목소리	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
실험자 목소리	117.384	87.655	86.894	81.005	637.008	1982

표2와 표3에서 실험자 목소리는 딥보이스가 적용되기 이전의 실험자 목소리를 말한다.

표3에서 버락 오바마와 딥보이스의 Pitch와 Formants, Intensity가 유사한 수치를 갖는 것을 확인할 수 있었다. 반면에 표2에서는 딥보이스의 First Formants 수치가 조 바이든과는 372Hz, 실험자 목소리와는 192Hz의 차이가 났다.

표2에서 딥보이스와 실험자 목소리의 First Formants 값의 차이가 딥보이스와 조 바이든 목소리의 차이보다 적다는 결과를 얻을 수 있다. 이를 통해 딥보이스로 합성을 하였음에도 특정 영역에서 실험자 목소리의 특징이 유지된다는 것을 유추할 수 있다.

First Formants는 '아', '에'와 같은 저모음에 대한 수치이다.[9] 아래에서는 First Formants를 통해서 유추한 것을 검증하기 위해 저모음이 포함된 'Ready'과 'Yes' 발음으로 영역을 특정하고 더 자세한 주파수 분석을 진행했다.

표 4 조 바이든의 대사 중 'Ready' 발음 분석

	Pitch Mean [Hz]	Intensity [dB]	First Formants [Hz]	Second Formants [Hz]
조 바이든 목소리	122.445	61.368	455.864	1331.997
딥 보이스	116.156	81.188	694.381	721.701
합성 전 목소리	115.343	88.051	718.154	991.407

표 5 오바마 대사 중 'Yes' 발음 분석

	Pitch Mean [Hz]	Intensity [dB]	First Formants [Hz]	Second Formants [Hz]
오바마 목소리	115.781	72.895	872.929	1910.349
딥 보이스	107.016	76.165	487.122	1522.748
실험자 목소리	112.052	86.582	492.734	1682.691

표4와 표5는 조 바이든 대사의 'Ready' 단어 음절 'rɛd'와 오바마 대사의 'Yes' 단어 음절 'jɛs'에 대하여 각각 분석한 결과이다. 그 결과 Pitch와 Intensity는 여전히 유의미한 차이를 볼 수 없었지만 음색을 나타내는 Formants에 대해 비교해 보았을 때, 표4의 First Formants 수치에서 조 바이든 목소리는 455Hz, 딥보이스는 694Hz로 어느 정도 차이를 보였음에 반해 실험자 목소리는 718Hz로 딥보이스와 굉장히 근접한 수치를 보였다. Second Formants에서도 조 바이든 목소리는 1331Hz, 딥보이스는 721Hz로 굉장한 차이를 보였지만 실험자 목소리는 991Hz로 조 바이든 목소리보다 딥보이스에 더 근접한 수치를 보였다.

표5에서 오바마의 대사 음절의 Formants를 분석한 경우에도 자료의 수치를 확인해 봤을 때, 오바마 목소리와 딥보이스를 비교한 경우보다 딥보이스와 실험자 목소리를 비교한 경우에 더욱 수치가 근접하였다. 분석 결과로 특정 영역에서 실험자 목소리의 특징이 유지되는 것을 확인했다.

III. 결론

본 연구에서는 기존의 목소리와 딥보이스의 차이점을 분석하였다. 또한, 실험자 목소리와 딥보이스의 유사점을 찾아 딥보이스에서 실험자 목소리를 특정할 수 있는 요소에 대해 탐구하였다. 결과로 특정 영역에서 실험자 목소리의 특징이 유지되는 것을 확인했다.

적은 양의 음성샘플로 분석을 진행해 결과를 일반화함에 있어서는 무리가 있다. 추후에 더 많고 다양한 표본으로 분석할 것을 제안한다.

위와 같은 연구 결과가 향후에 용의자 음성 식별 기술 발전의 도움을 기여하기를 기대하며 다양한 연구가 이루어질 것이다.

참고문헌

[1](N.d.).https://www.etnews.com/20230518000198

[2](N.d.).http://stock.mk.co.kr/news/view/176363

[3] Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A.,

Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016, September 19). WaveNet: A Generative Model for Raw Audio.

[4] Shen, J., Pang, R., Weiss, ron J, Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Ryan, R. S., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018, February 16). Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions.

[5] 김소운, 이성택. (2022). 딥보이스를 악용한 보이스 피싱 피해방지 서비스 개발. 한국통신학회논문지

[6](N.d). https://voice.ai/

[7](N.d.).http://www.ktword.co.kr/test/view/view.php?m_temp1=3745

[8](N.d.). https://www.fon.hum.uva.nl/praat/

[9] 조단, 박지연, 성철재. (2023). 한국어 이중모음 활음과 핵모음의 공조음 상황에서의 음향 특성. 한글, 84(1), 5 - 43, 10.22557/HG.2023.3.84.1.5