Regular paper | Journal of the KIECS. pp. 519-524, vol. 16, no. 3, Jun. 30. 2021, t. 107, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2021.16.3.519

인공지능 기반 영어 발음 인식에 관한 연구

이철승* · 백혜진**

A Study on the Recognition of English Pronunciation based on Artificial Intelligence

Cheol-Seung Lee* · Hye-Jin Baek**

요 약

최근 4차 산업혁명은 주요 선진국을 중심으로 세계의 국가들의 관심을 갖는 분야가 되고 있다. 4차 산업혁명 기술의 핵심기술인 인공지능기술은 다양한 분야에 융합하는 형태로 발전하고 있으며, 에듀테크 분야에도 많은 영향을 미치고 있으며 교육을 혁신적으로 변화하기 위해 많은 관심과 노력을 하고 있다.

본 논문은 DTW 음성인식 알고리즘을 이용하여 실험환경을 구축하고 다양한 원어민 데이터와 비원어민 데이터를 딥러닝 학습하고, CNN 알고리즘과의 비교를 통해 영어 발음의 유사도를 측정하여 비원어민이 원어민과 유사한 발음으로 교정할 수 있도록 연구한다.

ABSTRACT

Recently, the fourth industrial revolution has become an area of interest to many countries, mainly in major advanced countries. Artificial intelligence technology, the core technology of the fourth industrial revolution, is developing in a form of convergence in various fields and has a lot of influence on the edutech field to change education innovatively.

This paper builds an experimental environment using the DTW speech recognition algorithm and deep learning on various native and non-native data. Furthermore, through comparisons with CNN algorithms, we study non-native speakers to correct them with similar pronunciation to native speakers by measuring the similarity of English pronunciation.

키워드

AI, DTW, CNN, English Pronunciation, English Language Training
인공지능, 동적 시간 워핑, 합성곱 신경망, 영어 발음, 영어 교육

Ⅰ. 서 론

4차 산업혁명 시대 ICT 기술은 여러 학문분야에 융합하는 형태로 지속적으로 발전하고 있다. 이중 인공지능 AI(: Artificial Intelligence) 기술은 4차 산업혁명 시대의 핵심기술로 발전하고 있으며, 미래사회를 준비하기 위해 많은 연구가 이루어지고 있다. 인공지능 기술은 미래 교육 분야인 Edu-Tech 분야에도 큰 영향을 줄 것으로 예상

* 교신저자: 광주여자대학교 AI융합학과
광주여자대학교 교양과정부
· 접 수 일 : 2021. 04. 14
· 수정완료일 : 2021. 05. 16
· 게재확정일 : 2021. 06. 17
· Received : Apr. 14, 2021, Revised : May. 16, 2021, Accepted : Jun. 17, 2021
· Corresponding Author : Cheol-seung Lee, Hye-jin Baek
Dept. of AI Convergence, Kwangju women's University
Dept. of Liberal Arts, Kwangju women's University
Email : cyberec@kwu.ac.kr, hjpaik81@kwu.ac.kr

되며, 주요 선진국에서 진행되고 있다[1-3].

인공지능 기반의 영어 말하기 교육 및 학습은 발성의 내용을 이해하고 발음, 문법 및 대화 내용의 적합성을 확인해 교육적인 피드백을 제시하는 인공지능 원어민 교사의 역할을 수행하는 환경을 말한다. 본 논문의 인공지능 기반 영어 발음 인식에 관한 연구는 다양한 원어민 학습자 데이터와 비 원어민 학습자 데이터를 기반으로 음성인식 알고리즘인 DTW(: Dynamic Time Wrapping)를 이용하여 설계하고, CNN(: Convolutional Neural Network) 알고리즘과 비교하여 영어 발음 유사도를 측정하여 학습자의 발음이 원어민과 흡사한지 판단하고 교정을 도와 영어 발음을 개선하기 위한 기초연구에 그 목적이 있다.

II. 영어 음성 데이터 조사

한국어는 음절 박자 언어(syllable-timed language)의 특징을 지니고 있어서 영어의 강세 박자 언어(stress-timed language)의 특징인 리듬, 강세, 억양 등의 초 분절음소 자질을 습득하는데 많은 어려움이 있다[4]. 인공지능 학습을 하기 위해서는 일반인들도 쉽게 접근할 수 있는 다양하고 많은 데이터가 확보되었을 때 정확한 분석이 가능해진다. 또한 영어 발음의 교정을 위해서는 발음의 표준이 될 영어 발음 음성데이터가 필요하게 된다. 허나 구글 번역기의 음성데이터 및 네이버와 같은 포털에서 제공하는 발음데이터는 기계식으로 가공된 발음 데이터이므로, 동일 환경에서 제공된 데이터가 아니기 때문에 비교판단의 근거로 사용할 수는 없다.

따라서 본 연구에서 영어음성 데이터의 경우 다국적 발음 가이드 사이트인 Forvo에서 발음을 수집하였으며, 0부터 9까지의 영어숫자 발화 데이터는 미국 원어민(2명), 프랑스 억양 벨기에인(1명), 그리스 억양 그리스인(1명), 독일인(2명)등 총 6명의 화자에게서 수집하였다. 최종적으로 미국 원어민(2명) 데이터를 확보하였고, 각각 숫자에 100개의 음성을 확보하여 총 1,000개의 숫자 음성을 수집하였다, A부터 Z까지 11개의 데이터를 수집하여 입력하였으며, 11개의 데이터에는 원어민 남녀노소의 음성데이터로 총 26×11의 286개의 영어 발화 데이터를 이용하였고, 이를 인공지능 시스템에 딥러닝을 했을 경우 단어분석, 문장분석, 즉 자연어 처리가 가능하게 된다[5].

III. 음성인식 알고리즘

3.1 DTW 알고리즘

DTW란 시계열 데이터 간 비교를 최적의 index 매칭을 추정하는 알고리즘이다. 알고리즘은 Distance matrix를 구성하여 각 index의 값 간의 유사도(distance)를 계산한다[6].

	Index	0	1	2	3	4	5	6	7	8	9
Index	Data	1	3	5	7	6	8	9	10	8	7
0	1	**0**	2	4	6	5	7	8	9	7	6
1	2	1	**1**	3	5	4	6	7	8	6	5
2	6	5	3	**1**	1	0	2	3	4	2	1
3	5	4	2	**0**	2	1	3	4	5	3	2
4	7	6	4	2	**0**	**1**	1	2	3	1	0
5	8	7	5	3	1	2	**0**	**1**	**2**	**0**	**1**

그림 1. DTW 유사도 행렬
Fig. 1 DTW distance matrix

DTW는 그림 1과 같이 첫 index부터 마지막 index까지 최소 cost(distance)로 이동할 수 있는 최단 경로를 찾는다. 최단 경로에 포함된 노란색 셀이 최적의 index 매칭 유사도 distance 이다. 하지만 DTW는 음의방향으로 이동하지 못하는 단점이 존재한다.

python의 fastdtw함수의 출력값은 두 데이터의 Manhattan distance([0]와 최적의 path([1])로 path: [(0, 0), (1, 1), (2, 2), (2, 3), (3, 4), (4, 4), (5, 5), (6, 5), (7, 5), (8, 5), (9, 5)].를 통한 최적의 path로 데이터를 warping하는 방식이다.

3.2 CNN 알고리즘

CNN[7][8]은 딥러닝 알고리즘 중 가장 널리 알려진 알고리즘으로, 주로 이미지나 영상 데이터를 사용할 때 사용하는 신경망 모델 중 하나다. 간단하게 강아지와 고양이를 분류하는 알고리즘을 설계한다는 가정 하에 다음과 같은 프로세스를 거친다.

사람은 고양이를 봤을 때 고양이의 특징적 요소를 파악하여 고양이를 고양이로 인식할 수 있다. 귀가 두개 있고, 눈이 두 개 있고, 코가 하나 있고, 네 다리로 걷는 생명체라고 하면 강아지도 마찬가지다. 하지만 고양이의 눈, 코, 귀, 크기 등이 강아지와 다르기 때문에 고양이와 강아지를 헷갈리지 않는 것이다.

딥러닝은 이런 성질을 이용한다. 인풋값에서 특징(feature)을 찾아내고, 각 특징에 가중치를 자동적으로 부여하여 보다 더 정확하게 인풋값의 인식이 가능하다[9-10].

Ⅳ. 인공지능 기반 영어발음 인식 시스템 구현 및 실험

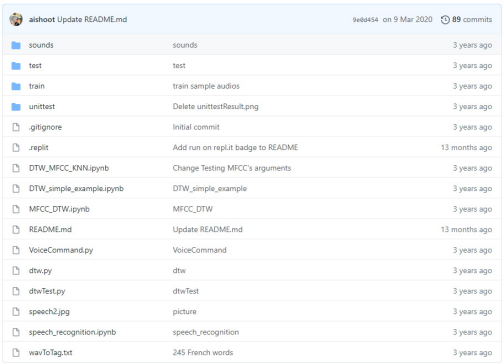

그림 2. 깃 허브의 DTW 오픈소스
Fig. 2 DTW open source of Github

그림 2는 실험에 사용한 깃 허브의 오픈소스 및 데이터이며, 발음 평가를 위한 DTW를 중심적으로 조사한 후 연구를 진행하였다. 또한, 실험을 위해 0~9까지의 숫자 음성 데이터를 활용하였고, DTW 알고리즘의 경우, speech DTW 오픈 소스를 활용하였다.

4.1. 영어발음 구성요소 및 고려사항

DTW는 시간적 변동을 고려할 수 있는 속성을 활용한 인토네이션을 비교할 수 있게 된다. 음성에서 가장 세밀한 정도가 낮고 범위적이기 때문에 실제로 음성 파형의 구성요소를 살펴보면 음의 길이, 에너지, 피치, 음소, 강세 등 다양한 소정 요소들이 존재한다. 만약 이 각각의 요소들을 모두 고려할 경우, 다음과 같은 두 가지 고려사항이 발생한다.

첫 번째, 범위가 좁아 전체 음성구간에서 한정적인 부분에 대해서만 알고리즘이 적용될 수 있다. 만약 음소 하나하나를 고려하여, [f] 발음에 대한 발음 유사도를 측정한다고 가정한다면, 전체 음성에서 [f]발음이 있는 구간에만 해당 알고리즘을 적용해야 하며, 만일 이 과정에서 오차가 발생할 경우 음성 전체에 낮은 인식률을 드러내게 된다.

이러한 특징은 두 번째 문제와도 연결되는데, 범위가 좁은 만큼 식이 복잡해진다. 영어는 24개의 자음 발음과, 16개의 모음 발음으로 구성이 된다. 이 각기의 음소를 고려할 경우, 아래의 두 가지 프로세스를 따라야 한다. 첫째는 총 40개의 음소 패턴을 모두 고려한 발음 교정기를 개발해야 한다. 둘째, 비원어민이 발화를 하였을 때 음성을 작은 음소로 분리해야 한다. 보다 세밀하게 발음 인식을 할 수 있겠으나, 그만큼 고도의 프로세스를 요구하게 된다.

4.2. DTW 알고리즘 설계

DTW 알고리즘은 두 개의 시계열 데이터가 있다고 가정했을때, 두 개의 간격의 유사도를 알아내기 위한 알고리즘 중 하나다.

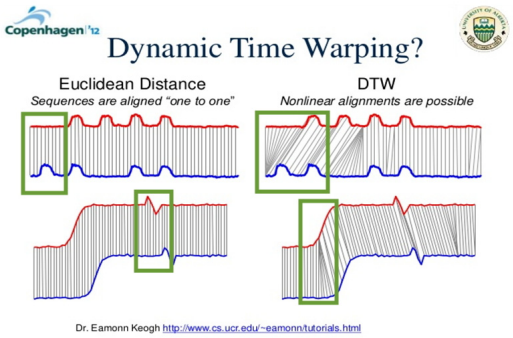

그림 3. DTW 알고리즘
Fig. 3 DTW algorithm

그림 3의 상단과 하단의 붉은색 파형과 푸른색 파형의 유사도를 비교하는 경우를 생각한다. 가로축은 시간이며, 세로축은 세기이다. 직관적으로 두개의 파형이 비슷하다고 파악할 수 있지만, 세기가 다르고 특정 패턴이 나타나는 시간이 다르다는 사실이 명확히 알 수 있다. 붉은색 파형과 푸른색 파형의 유사도 척도 평균제곱 오차를 사용하는 식은 다음과 같다.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

오차의 거리로, 유클리디안 거리를 사용할 경우, 그림 3의 좌측에서 볼 수 있는 것 처럼 동일 시간대의 데이터를

매칭하여 서로 비교하게 된다. 이와 같은 방식으로 비교할 경우, 그림 3의 녹색으로 표시된 지점을 살펴보았을 때 두 파형의 패턴은 일치하지 않으며, 실제로도 유사한 그래프일지라도 오차가 크게 발생한다. 이러한 부분을 보완하기 위해 개발된 것이 DTW 알고리즘이고, 각각의 시계열 데이터의 시간을 인덱스로 사용하여, 그림 3의 우측처럼 유사한 구간끼리 인덱스를 연결한다.

DTW 알고리즘은 비슷한 형태를 지닌 파형끼리 인덱스를 매칭한 후 유사도를 산출해내는 방식으로 길이 및 세기가 다른 파형끼리 유사도를 비교할 수 있다. 그러나 이를 수행하기 위해서는 비교 척도로 사용될 기준 패턴을 다수 수집하고 생성해야 한다. 기준 패턴이 필요한 이유는 굉장히 명확하다. 하지만 원어민 억양이라는 것은 절대적인 발음이 아니고, 같은 발음을 할지언정 목소리 톤이나 말의 속도, 구강 구조 등에 따라서 조금씩 다른 소리를 만들게 된다. 이런 이유로 기준 패턴은 한 가지가 아니라 여러 가지를 생성해야 하며, 비원어민의 발음을 분석할 때 가장 유사한 기준 패턴을 선택하여 비교하는 것이 합리적이다.

CNN을 통한 발음 교정의 경우, 원어민 음성의 특징과 비원어민 음성의 특징을 추출한 후, 각 특징점의 분포를 비교하는 방식으로 접근한다[11].

4.3 실험 및 평가

영어 발음 교정 알고리즘 실험은 DTW 코드는 speech DTW 오픈소스를 활용하였고, 실험과정에 사용한 코드는 깃 허브에 첨부 하였다. 그림 5, 6과 같은 실험환경 UI의 녹음시작 버튼을 누리면 2초간 음성을 녹음하여 ../data/test/test.wav로 파일을 저장하고, CNN 알고리즘을 활용하여 분석한 발음과 DTW 알고리즘을 활용하여 발음 평가를 실시한다. 본 연구는 DTW 알고리즘의 효용성을 평가하기 위해 작성되었으며, DTW 알고리즘으로 시계열 데이터를 비교했을 때, 비슷한 인토네이션을 지녔다면, 두 음성 사이의 간격은 균일하다. 하지만 인토네이션의 차이가 발생할 경우 두 음성은 균일하지 않는 것으로 판단한다.

그림 4는 파란색, 빨간색, 초록색 파형 사이의 회색선을 연결하여, 길이를 측정한 결과 파란색과 빨간색은 회색선의 길이가 비슷하기 때문에, 인토네이션 차이가 크지 않으며, 빨간색과 초록색은 인토네이션 차이가 크므로 음성 사이의 간격이 균일하지 않는 점을 착안하여 DTW 알고리즘을 적용하였다. 인터페이스에서 그래프는 기준 패턴이 되는 음성과, 자신이 녹음한 음성의 거리를 비교하여 시각화 해준다. 그래프의 형태가 일정한 값을 지닌다면 상대적으로 좋은 발음이라 볼 수 있으며, 높낮이가 불규칙 하다면 발음이 나쁘다고 판단할 수 있다.

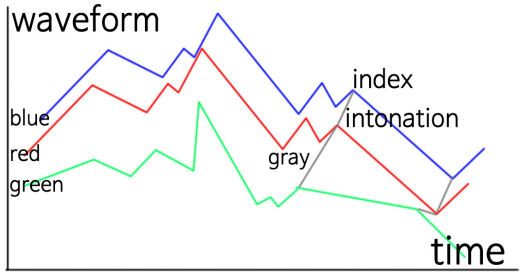

그림 4. DTW 알고리즘 적용
Fig. 4 DTW algorithm applied

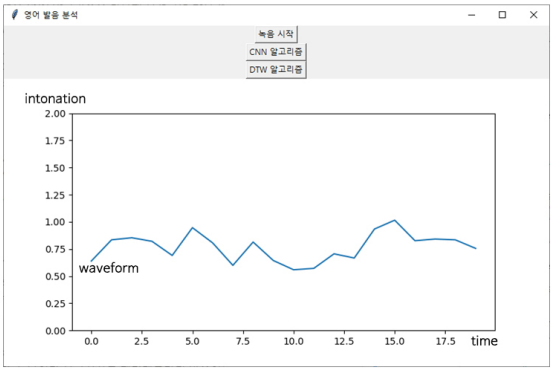

그림 5. 기준패턴 : Jackson(미국 원어민) / 테스트 음성 : Theo(미국 원어민)
Fig. 5 Reference Pattern: Jackson (Native American speaker) / Test voice: Theo (Native American speaker)

그림 5는 실험에 사용한 기준 패턴인 미국 원어민 Jackson의 음성을 선택하였다. 테스트 데이터로는 미국인 Theo, 독일인 George, 그리고 비원어민인 음성을 가지고 'zero' 발음 테스트를 진행하였고, 네 명의 음성은 github의 data파일에 기록되어 있어 열람이 가능하다.

미국 원어민인 Theo의 경우, 0.5~1 사이의 그래프 형태를 띄며, George는 0.5~1.5 사이의 변동을 갖는다. 그림 6은 비 원어민의 경우이며, 0.25~1.5 사이의 변동 폭을 가지므로, 인토네이션이 기준 패턴과 일관되지 않는 양상을 보인다.

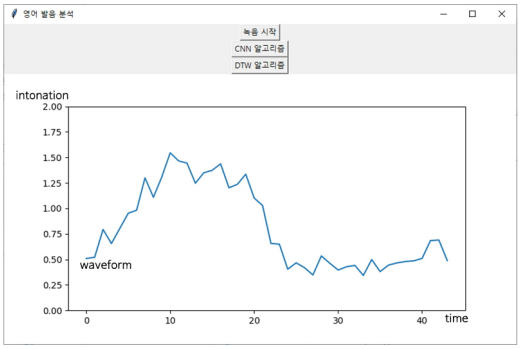

그림6. 기준패턴 : Jackson(미국 원어민) / 테스트 음성 : 비 원어민

Fig. 6 Reference Pattern: Jackson (Native American speaker) / Test voice: Non-native speaker

본 연구는 그래프가 지닌 절대적 수치보다는 변동폭에 중심을 두어야 한다. 변동 폭이 일정할 경우 인토네이션이 기준 패턴과 유사해 좋은 발음이며, 일정하지 않을 경우 좋지 않은 발음으로 볼 수 있다. 따라서, 그래프의 시각적 분석 이상으로 확률 개념의 도입이 필요할 것으로 보인다. 가령, 표준편차를 도입해, 인토네이션의 분산 정도를 수치적으로 변환하면 더욱 개선된 결과를 보일 수 있을 것이다.

Ⅴ. 결 론

본 논문에서는 영어 발음의 구성요소 중 인토네이션을 베이스로 한 발음 교정을 제안하고 있다. 교정을 위해서는 발음 평가를 진행하여 발화자 본인의 발음이 얼마나 부정확한지 수치적으로 제안할 필요가 있다. 그 방법론으로서 크게 DTW와 CNN을 제안했으며, 본 연구에서는 DTW에 초점을 맞추어 실험을 진행하였다.

DTW는 유사도를 측정해줄 수는 있으나, 형태의 유사도 외의 요소(발화자 목소리의 크기나 톤)를 고려하기는 어렵다. 따라서 이번 실험에서는 절대적 수치를 사용하는 것이 아니라 얼마나 그 차이가 균일한지에 대해 초점을 맞추어 진행하였으며, 유의미한 수치를 내보일 수 있었다.

향후, 음성의 다른 구성요소를 고려하여 알고리즘을 설계하기보다는, 인토네이션에 초점을 두어 알고리즘을 강화하고 확장하는 것이 정확도를 증대시키는 방향으로 연구를 진행하는 것을 제안하고자 한다. DTW 알고리즘을 강화하는 것은 고전적 방식에 대한 재도전이므로 유의미한 결과를 낼 수 있을 것이다. 또한 CNN 알고리즘을 통한 발음 평가 실험을 진행할 경우, 사전에 원어민과 비원어민의 음성데이터를 대량 확보하고, 과거의 음성인식이 지닌 단점을 보완할 수 있기에, CNN을 사용한 접근도 충분히 정확도 증진을 위한 방법이 될 것이다.

감사의 글

"본 연구결과는 2021학년도 광주여자대학교 교내연구비 지원에 의하여 연구되었음".

(KWUI21-027)

References

[1] H. Jeon, H. Chung, B. Kang, and Y. Lee, "Survey of Recent Research in Education based on Artificial Intelligence," *J. of the Korea Institute of Electronics and Telecommunications Trends (ETRI)*, vol. 36, no. 1, Feb. 2021, pp. 71-80.

[2] T. Kim, M. Ryu, and S. Han, "Framework Research for AI Education for Elementary and Middle School Students," *J. of The Korean Association of Artificial Intelligence Education*, vol. 1, no. 1, 2020, pp. 31-42.

[3] M. Cho, "A Study on the History, Classification and Development Direction of Artificial Intelligence," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 16, no. 2, Apr. 2021, pp. 307-312.

[4] J. Kim, "The Development and Application of the Web-based English Pronunciation Learning Material Focused on Suprasegmental Features," *J. of the Ewha Education*, vol. 13, 2003, pp. 443-457..

[5] E. Kim, and W. Son, "STT, Scheduling, Automatic subtitle generation, STT post-processing, Sentence analysis," *J. of the Korea Institute of Electronic Communication*

Sciences, vol. 16, no. 1, Feb. 2021, pp. 81-88.

[6] S. Kim, and M. Park, " A Study on Time-series Clustering Analysis based on Dynamic Time Warping," *J. of the Korean Data Analysis Society*, vol. 20, no. 5, 2018, pp. 2319-2332.

[7] K. Heo, and D. Lim, "Noise reduction using patch-based CNN in images," *J. of the Korean Data Analysis Society*, vol. 20, no. 5, 2018, pp. 2319-2332.

[8] Y. Jeong, and G. Choi, "Efficient iris recognition using deep-learning convolution neural network(CNN)," *J. of the Korean Data & Information Science Society*, vol. 30, no. 2, June 2019, pp. 349-363.

[9] W. K. Leung, X. Liu, and H. Meng, "CNN-RNN-CTC Based End-to-end Mispronunciation Detection and Diagnosis," In *Proc. IEEE Intf. Conf. Acoustics, Speech and Signal (ICASSP)*, Brighton, UK, May 2019.

[10] B. Kang and O. Kwon, "DNN-based acoustic modeling for speech recognition fo native and foreign speakers," *J. of the Phonetics and Speech Sciences*, vol. 9, no. 2. 2017, pp. 95-101.

[11] F. Nazir, M. N. Majeed, M. A. Ghazanfar, and M. Maqsood, "Mispronunciation Detection Using Deep Convolutional Neural Network Features and Transfer Learning-Based Model for Arabic Phonemes," *J. of the IEEE Access*, vol. 7, Apr. 2019, pp. 52589-52608.

저자 소개

이철승(Cheol-Seung Lee)

2001년 광주대학교 공과대학 컴퓨터학과 졸업 (공학사)
2003년 조선대학교 대학원 컴퓨터공학과 졸업 (공학석사)
2008년 조선대학교 대학원 컴퓨터공학과 졸업 (공학박사)
2012년 ~ 광주여자대학교 AI융합학과 교수
※관심분야 : RFID, AI, Android Security Wireless Network Security

백혜진(Hye-Jin Baek)

2008년 조선대학교 대학원 영어영문학과 졸업 (문학석사)
2019년 조선대학교 대학원 영어영문학과 졸업 (문학박사)
2012년 ~ 광주여자대학교 교양과정부 교수
※관심분야 : 영어교육, 영문학