문학석사 학위논문

인공지능 자동음성인식기들의 한국인의 영어 발음 산출 훈련 적합성 평가

2022 년 8 월

서울대학교 대학원

언어학과

이 여 진

인공지능 자동음성인식기들의 한국인의 영어 발음 산출 훈련 적합성 평가

지도 교수 이 호 영

이 논문을 문학석사 학위논문으로 제출함

2022 년 8 월

서울대학교 대학원

언어학과 음성학 전공

이 여 진

이여진의 문학석사 학위논문을 인준함

2022 년 8 월

위 원 장	정 민 화	(인)
부위원장	이 호 영	(인)
위 원	황 효 성	(인)

초 록

이 연구는 문맥 독립적 환경에서 인공지능 자동음성인식기들의 한국인 학습자의 영어 발음에 대한 인식 성능을 평가하여, 발음 산출 훈련(production training) 프로그램을 개발할 때 인공지능 자동음성인식 시스템을 활용할 수 있는지 검증하고, 있다면 어떤 방식으로 활용하면 좋을지 논의하는 것을 목적으로 한다.

문맥 정보와 독립적으로 발음을 정확하게 인식하는지 평가하기 위해 실험 자료는 개별 단어 형태와 문장틀에 넣은 형태를 활용하였으며, 목표음의 위치에 따라 성능이 달라지는지 알아보기 위해 목표음을 문두, 문중, 문미, 어두, 어중, 어말에 위치시켜 인식률을 알아보았다. 한국인 학습자의 발음 오류와 음성인식기의 오인식을 구분하고자 원어민이 한국인 학습자의 발음을 듣고 판단한 것과 자동음성인식기의 응답결과를 직접 비교하는 방법을 사용했다. 발음오류는 원어민이 발음오류라고 판단한 것이며, 오인식은 원어민의 판단과 자동음성인식기의 결과가 서로 다른 것으로 규정했다. 실험 단어로는 한국인 학습자가 발음하기 어려워하는 최소대립쌍을 활용하여 자동음성인식기가 음을 얼마나 정확하게 구분하는지 파악하였다. 상용화된 오픈 API를 제공하는 Google Cloud Speech-to-Text, Microsoft Azure Speech Service, IBM Watson Speech to Text, Amazon Transcribe, Naver CLOVA Speech 와 음소인식기반 음성인식기를 포함하여 총 6가지의 자동음성인식기를 서로 비교하였다.

한국인 학습자의 영어 발음에 대한 각 인공지능 자동음성인식기와 원어민의 판단을 직접 비교하여 그 일치율을 확인한 결과, 음소인식기반 음성인식기의 일치율이 약 77%로 가장 높게 나타났다. 입력형태가 문장일 때 전반적으로 일치율이 높으며, 목표음이 문중과 어중 위치에 있을 때 일치율이 높게 나타났다. 자음의 일치율과 모음의 일치율 사이의 차이는 미미했다. 또한 전반적으로 분절음 [b, f, p]의 일치율이 높고 [dʒ, s, ʌ]의 일치율이 낮게 나타났다.

이 연구는 산출 훈련의 관점에서 문맥 독립적으로 여러 인공지능 자동음성인식기들의 성능을 평가하였다는 점에서 의미가 있다. 산출 훈련 프로그램 개발을 위해서는 음소인식기반 자동음성인식기가 가장 성능이 좋지만, 최소대립어를 활용한 산출 훈련에는 단어인식기반 자동음성인식기도 꽤나 높은 성능을 보임을 알 수 있었다. 자동음성인식기가 약 80%의 인식률로 발음의 정오를 판가름하기 때문에 발음 교육에 활용할 경우 긍정적인 효과를 예상할 수 있었으며, 산출 훈련 프로그램을 제작함에 있어서 언어학적 단위별로 어떻게 구성해야 오인식을 최대한 피할 수 있을지 확인할 수 있었다.

주요어 : 자동음성인식(ASR), 성능 평가, 한국인 학습자, 영어 발음, 인식률, 인공지능, 산출 훈련, 발음 교육

학 번 : 2015-22455

목 차

표 목차

그림 목차

1. 서 론

1.1. 연구의 목적과 필요성

외국어 발음 교육은 학습자에게 매우 중요한 문제이다. 발음 수준은 외국어로 의사소통하는 것에 대한 자신감과 직접적으로 연결되어 있으며(Evers & Chen, 2020) 발음이 좋지 않으면 진학과 취업, 직무 수행 등에서 불이익을 받을 수 있기 때문이다. 외국어 학습자의 발음은 모국어의 영향에서 자유로울 수 없기 때문에 발음 오류가 빈번하게 발생하는 특성이 있다. 잘못된 외국어 발음은 모국어의 영향으로 쉽게 굳어져 많은 학습자들에게 고질적인 문제가 되므로 효율적인 발음 교육이 필요하다.

외국어 학습자를 대상으로 발음 교육을 할 때 지각 훈련(perception training)만으로는 발음 개선 효과가 부족하기 때문에 산출 훈련(production training)을 병행하는 것이 필요하다(Golestani & Pallier, 2007; Lopez-Soto & Kewley-Port, 2009; Peperkamp & Bouchon, 2011; Huensch, A., & Tremblay, A., 2015). Lopez-Soto & Kewley-Port(2009)의 연구에서 스페인어 화자에게 영어 종성(coda) 발음에 대해 지각 훈련을 진행한 결과, 지각 능력은 향상되었지만 학습자의 종성 발음 산출 능력은 지각 훈련 전후가 거의 비슷하게 나타났다. 또한 Peperkamp & Bouchon(2011)의 연구에서 프랑스어-영어 이중언어 화자에게 영어 모음 /i/–/ɪ/ 대립에 대해 지각 훈련을 진행한 결과, 지각 능력과 산출 능력 사이에 관계가 없는 것으로 나타났다. 하지만 반대로 산출 훈련은 지각 능력과 산출 능력을 모두 향상시키는 데 도움을 준다(Hirata, 2004; Kartushina et al., 2015). Kartushina et al.(2015)에서 프랑스어 화자에게 덴마크어 모음에 대해 산출 훈련을 실시한 결과 산출 능력과 지각 능력 두 가지가 모두 개선된 것으로 나타났다. 지각 훈련으로 지각 능력은 개선되지만 산출 능력은 개선되는 데 한계가 있으며, 산출 훈련으로는 지각 능력과 산출 능력 모두

개선되기 때문에, 외국어 학습자의 발음이 향상되기 위해서는 산출 훈련을 적극적으로 하는 것이 필요하다.

산출 훈련에서는 교육자가 학생의 발음이 맞았는지 틀렸는지를 확인해주고, 틀렸다면 발음을 어떻게 수정해야 하는지에 대해 방법을 제시해주는 교정 피드백(Corrective Feedback)이 중요하다. 원어민의 발음을 들려주는 청각적 피드백과 조음 방법을 보여주는 시각적 피드백을 제공하는 산출 훈련이 발음을 개선하는 데 효과적이라는 연구들이 있다(Hazan et al., 2005; Massaro et al., 2008;). 하지만 청각적 피드백은 발음 향상에 효과가 크지 않으며, 시각적 피드백은 학습자들이 조음 기관 그림에 익숙하지 않거나 발음 시 조음상의 시각적 대립이 뚜렷하지 않은 경우 효과가 떨어지는 한계가 있다. 이에 산출 훈련에서는 학생의 조음을 듣고 직접 교정해주는 교정 피드백이 최선의 방법이 될 수 있다. 조음 방법에 대한 설명만으로는 효과가 미미하며 직접적인 교정 피드백을 제공하는 것이 산출 훈련에 효과가 있다는 점이 입증되어 있다(Kartushina et al., 2015; Saito, K., & Lyster, R. ,2012). 따라서 외국어 음성에 대한 지각 능력과 산출 능력이 모두 향상되기 위해서는 교정 피드백과 함께 산출 훈련을 받는 것이 중요하다.

하지만 교정 피드백과 평가를 제공하는 것은 교육자에게 가장 시간이 오래 걸리는 작업 중 하나이기 때문에 일상적으로 하기 어려운 일이며(Carrier, 2017; Ashwell & Elam, 2017), 교육자가 음성학적으로 발음 지도를 하는 데 익숙하지 않은 경우도 많다. 교육자가 음성학적 지식을 가지지 않아 음성학적으로 발음 지도를 하는 데 익숙하지 않은 경우도 많고, 설령 음성학적 지식을 가졌더라도 수업 현장에서 학생 개개인에게 교정 피드백과 평가를 제공하는 일은 시간이 오래 걸리기 때문에 현실적으로 매우 어렵다. 이에 교육 현장에서 발음을 교육하는 방법은 대부분 교사의 발음을 듣고 따라하는 방식으로 이루어지고 있으며, 오류에 대한 정확한 분석이나 피드백 없이 기계적인 연습으로 이루어지고 있다(김효진, 2018).

이러한 현실적 문제를 해결하기 위해 자동음성인식기술이 외국어 발음 산출 훈련과 진단 평가 등 다방면으로 활용될 수 있을 것으로 기대된다. 고성능의 인공지능 자동음성인식기를 활용한다면 학습자의 발화에 대해 자동으로 즉각적인 평가와 교정 피드백을 제공하여 교육자의 시간과 부담을 줄일 수 있을 것이다. 최근 음성인식 기술이 성숙하면서 정확도가 높아져 자동음성인식기를 외국어 교육에 활용하는 것의 긍정적 효과에 대한 연구가 진행되고 있다(Carrier, 2017; Golonka et al., 2014). 고성능의 자동음성인식기술을 활용한 프로그램이 학습자의 발음을 듣고 발음의 정확도에 대한 평가와 교정 피드백을 제공하는 것이 가능해지는 것이다. 이처럼 인공지능 자동음성인식기술을 통해 교육자를 대신하거나 보완할 수 있는 저비용 고효율의 교육 및 평가 프로그램을 개발할 수 있을 것으로 보인다.

자동음성인식기를 활용한 발음 교육 프로그램을 개발하기 위해서는 우선적으로 자동음성인식기의 성능이 담보되어야 한다. 자동음성인식기의 인식 결과를 바탕으로 하여 발음을 평가하거나 피드백을 제공하기 때문에 자동음성인식기의 정확도가 이후 단계의 품질을 좌우한다. 따라서 발음 훈련의 관점에서 자동음성인식기의 정확도를 점검하고, 다양한 종류의 자동음성인식기의 성능을 비교하여 어느 것을 선택하는 것이 좋을지 기준을 마련하는 것이 필요하다.

자동음성인식기는 단어인식기반 자동음성인식기와 음소인식기반 자동음성인식기로 나눌 수 있다. 단어인식기반 자동음성인식기의 경우 공개된 API 를 제공하여 접근성이 높은 프로그램들이 여럿 있으며, 음소인식기반 자동음성인식기의 경우 이를 발음 오류 검출 및 진단(Mispronunciation Detection and Diagnosis, MDD)에 활용하는 연구가 활발히 진행되고 있다(Wu et al., 2021; Korzekwa et al., 2021; Ye et al., 2022;). 음소인식기반 자동음성인식기의 경우 단어인식기반 자동음성인식기와는 달리 단어를 기준으로 인식하는 것이 아니라 개별 음을 기준으로 인식하기 때문에 발음을 정확히 구분하여 인식하는 과업에 더욱 적합하다고 볼 수 있다. 따라서 음소 기반의 발음 교정 피드백을

제공하기 위해서는 음소인식기반 자동음성인식기를 활용하여 프로그램을 개발할 필요가 있다. 그러나 음성인식 개발자가 아니라면 이를 직접 개발하여 사용하기 어려우며, 개발한다 하더라도 시간이 오래 걸린다는 한계가 있다. 빠른 시간 안에 최첨단의 기술을 활용하여 음성인식 시스템을 구축해야 하는 외국어 교육자나 실험음성학자, 교육 프로그램 개발자라면 시중에 나와 있는 단어인식기반 자동음성인식기를 활용하는 것이 최선의 대안이 될 수 있다. 그러나 자동음성인식기를 발음 교육의 측면에서 평가한 연구는 거의 없기 때문에 발음 교육을 위한 인식기 선택에는 기준이 없는 현실이다. 따라서 발음 교육의 측면에서 단어인식기반 자동음성인식기와 음소인식기반 자동음성인식기의 최소대립어 인식 성능을 모두 평가해 볼 필요가 있다. 어느 자동음성인식기가 최소대립어를 활용한 발음 교육 프로그램 개발에 도움이 될지 점검해보고, 특히 단어인식기반 자동음성인식기를 발음 산출 훈련에 활용할 수 있을 것인지 평가하고자 한다.

발음 교육 현장에서는 발음 산출 훈련을 위해 주로 최소대립어를 활용한다. “heat - hit”과 같은 최소대립어의 발음이 서로 잘 구분이 되어야 외국어의 발음을 제대로 습득했다고 할 수 있으며, 기능부담량이 높은 쌍을 잘 구분해서 발음하는 것이 중요하기 때문이다. 실제 수업에서는 개별 음소 단위가 아니라 최소대립어를 발음 산출 교육에 사용하기 때문에 음소 수준이 아닌 단어 수준에서의 자동음성인식기를 사용하는 것이 효율적일 수 있다. 음성인식기 개발자가 아닌 교육자나 실험음성학자가 최소대립어를 활용하여 발음 산출 교육을 하고자 할 때, 음소인식기반 자동음성인식기보다 단어인식기반 자동음성인식기가 빠르고 유용하게 쓰일 가능성이 있다. 따라서 음소인식기반 자동음성인식기의 성능과 대비하여 단어인식기반 자동음성인식기의 성능을 평가할 필요성이 있다.

단어인식기반 자동음성인식기의 경우, 이들의 인식률을 비교하여 평가한 다양한 연구가 진행된 바 있다(유현재 외, 2020; 노희경 & 이강희, 2017; IANCU, 2019; Këpuska & Bohouta, 2017). 하지만 많은 연구에서

자유발화 상황에서의 인식률을 평가하였다는 문제가 있다. 단어인식기반 자동음성인식기의 경우 주변 단어들을 고려하여 통계적으로 어느 단어일 확률이 가장 높은지 계산하여 인식하기 때문에 문맥 정보가 인식에 주요한 영향을 미친다. 따라서 통제되지 않은 자유발화 상황에서의 인식률을 평가할 경우, 문맥에 의해 발음이 유추된다는 문제가 있다. 발음을 정확히 구분해서 인식하여 이를 바탕으로 발음 교정 피드백을 제공하는 프로그램을 개발하기 위해서는, 자동음성인식기의 성능을 평가할 때 문맥 정보와 독립적으로 통제된 환경에서 성능을 평가할 필요가 있다. 이에 이 연구에서는 개별단어 형태와 동일한 문장틀에 넣은 형태를 사용하여 인식 성능을 평가하고자 한다.

한편, 음소인식기반 자동음성인식기를 활용한 발음오류 검출 및 발음교육 프로그램 개발에 대한 연구도 다수 이루어져 왔다(Tepperman & Narayanan, 2008; Li et al., 2016; 류혁수 & 정민화, 2016a; 류혁수 & 정민화, 2016b; Xie et al., 2020; Korzekwa et al., 2021; Hirschi et al., 2020). 하지만 이러한 연구에 활용된 음소인식기반 자동음성인식기의 성능에 대한 정확한 언급이 없다는 한계가 있으며, 성능에 대한 언급이 있더라도 효과적인 산출 훈련을 위해 언어학적 단위별로 성능을 자세히 살펴본 연구는 없다. 이에 이 연구에서는 자동음성인식기의 성능 평가를 단어와 문장, 자음과 모음, 목표음의 위치와 같이 언어학적 단위별로 나누어 다방면으로 진행하고자 한다. 효율적인 발음 산출 훈련 프로그램 개발을 위해서는 언어학적 단위별로 인식률을 나누어 평가하는 것 역시 필요하다. 학습 또는 평가하고자 하는 목표음의 언어학적 형태와 위치가 다양하게 나타날 수 있고, 형태와 위치에 따라 인식률이 달라질 경우 학습 결과에 영향을 미칠 수 있기 때문이다. 자동음성인식기를 활용하여 발음 산출 훈련 프로그램을 개발하기 위해서는 자동음성인식기가 이러한 다양성에 어느 정도로 대응하는지를 자세히 알아보는 것이 필요하다. 그러나 자동음성인식기의 성능을 언어학적 단위별로 나누어 살펴본 연구는 없다. 이에 이 연구에서는 단어와 문장, 자음과 모음, 단어 내에서는 어두,

어중, 어말, 문장 내에서는 문두, 문중, 문미와 같이 언어학적 단위별로 인식률에 차이가 있는지 확인하고자 한다.

자동음성인식기를 외국어 교육에 활용하기 위하여, 외국어 학습자의 발화에 대한 자동음성인식기의 성능을 평가하는 연구가 이루어져 왔다. 양병곤(2017), 윤정희(2014), 박향숙, 이예식, & 윤정희(2018)는 한국인 학습자의 영어 발화에 대한 자동음성인식기의 성능을 평가하고, 자동음성인식기가 잘 인식하지 못하는 발음의 음성언어학적 특징을 살펴보았다. 그러나 학습자의 발음 오류와 자동음성인식기의 오인식을 구분하여 성능을 살펴보지 않았다는 한계점이 있다. 학습자의 틀린 발음을 인식기가 틀렸다고 인식했을 경우 인식기가 발음 오류를 제대로 감지했다는 뜻인데도 이를 오인식에 포함시켜 인식률을 계산하였다. 이 연구에서는 원어민의 판단과 자동음성인식기의 인식결과를 서로 비교하여 둘 사이의 일치율을 평가함으로써, 원어민이 발음 오류로 판단한 것과 자동음성인식기의 오인식을 구분하고자 한다.

이 연구에서는 발음 산출 교육 프로그램 개발의 측면에서 인공지능 자동음성인식기의 활용가능성을 살펴보기 위하여, 한국인 학습자의 영어 최소대립어 발화에 대해 인공지능 자동음성인식기가 문맥 독립적으로 발음을 인식하는 성능을 평가하고자 한다. 자동음성인식기의 성능이 괄목할 만한 향상을 보이는 현재, 발음 산출 교육의 측면에서 자동음성인식기가 비원어민 화자의 발음을 정확히 구분할 수 있는지 살펴보기 위해서는 문맥 독립적인 환경에서의 인식률을 테스트하는 것이 필요하다. 문맥 정보로부터 영향을 받지 않고 자동음성인식기의 성능을 평가하기 위하여 자극음을 개별 단어 형태 및 동일한 문장틀에 넣은 형태를 활용하여 실험을 진행한다. 동일하게 통제된 환경에서 목표음의 위치별로 인식률이 달라지는지 세부적으로 파악하기 위하여 목표음을 문두, 문중, 문미, 어두, 어중, 어말에 위치시켜 성능을 평가한다. 평가 방법에 있어서는 학습자의 발음 오류와 자동음성인식기의 오인식을 구분하여 성능을 평가하기 위하여, 원어민이 한국인 학습자의 발음을 듣고 판단한 것과 자동음성인식기의 응답결과를 직접 비교하여 둘

사이의 일치율을 평가한다. 자극음은 한국인 학습자가 발음하기 어려워하는 최소대립쌍 단어들을 활용하여 단어인식기반 음성인식기와 음소인식기반 음성인식기가 자모음을 얼마나 정확하게 구분하는지 파악하고자 한다. 이를 통해 인공지능 자동음성인식기의 최소대립어 인식 성능을 서로 비교하고, 자동음성인식기가 발음 산출 교육에 활용할 만한지, 활용할 만하다면 어떻게 활용하면 좋을지 확인하고자 한다.

이러한 과업을 인공지능 자동음성인식기가 어느 정도까지 정확하게 수행할 수 있는지 알아보고자, Google Cloud Speech-to-Text, Microsoft Azure Speech Service, IBM Watson Speech to Text, Amazon Transcribe, Naver CLOVA Speech 이렇게 단어인식기반 자동음성인식기 5 가지와 음소인식기반 자동음성인식기 1 가지(Baevski et al., 2020), 총 6 가지의 인공지능 자동음성인식기들을 서로 비교한다. [1] 단어인식기반 자동음성인식 프로그램들은 모두에게 공개된 Open API 를 제공하고 있어서 접근과 활용이 용이하며 유사한 선행연구를 참고하여(유현재 외, 2020; 노희경 & 이강희, 2017; 최승주 & 김종배, 2017; Këpuska & Bohouta, 2017; Kodish-Wachs et al., 2018) 실험대상으로 선정했다. 이와 같이 현재 개발된 인공지능 자동음성인식 시스템의 성능을 비교 분석함으로써 발음 산출 교육의 관점에서 인공지능 음성인식 시스템을 활용하기 위한 실증적 근거를 제공하고자 한다.

이 연구를 통해 아래와 같은 질문에 답하고자 한다.

1. 한국인 학습자의 영어 발음에 대해 원어민 화자와 각 인공지능 자동음성인식기가 서로 얼마나 일치하는가? 단어인식기반 자동음성인식기의 최소대립어 인식 성능이 음소인식기반 자동음성인식기의 최소대립어 인식 성능에 얼마나 근접하는가?
2. 자동음성인식기가 발음 산출 교육에 활용할 만한가? 활용할 만하다면 어떤 방식으로 활용하는 것이 좋은가?

[1] Kakao I Newton Voice 의 경우, 베타 버전으로 한국어 음성인식만 지원하여 제외하였다. ETRI AI 음성인식 API 의 경우, 음성인식 결과가 대부분 불일치하여 논외로 하였다

1.2. 선행연구 검토

1.2.1. 자동음성인식기의 성능에 관한 연구

자동음성인식(Automatic Speech Recognition, ASR)은 인간의 음성언어를 의미 있는 텍스트로 변환하는 인공지능 기술의 한 분야로, 지난 40 여 년 동안 현저한 발전을 이뤄왔다(Juang & Rabiner, 2005). 자동음성인식 시스템은 잡음 등의 신호를 개선하는 전처리 단계, 음성신호의 특징을 추출하여 텍스트로 출력하는 음향모델 단계, 특정 단어열이 주어졌을 때 다음에 나올 단어의 확률을 추정하는 언어모델 단계, 음향모델과 언어모델 등으로 이루어진 탐색 네트워크에서 최적의 경로를 찾는 단계인 디코딩 네트워크의 과정을 통해 음성 데이터를 텍스트로 만들어낸다. 이러한 자동음성인식 시스템은 학습데이터의 종류와 양, 방법 등에 따라 차별적인 결과를 만들어낸다(유현재 외, 2020).

자동음성인식 시스템의 성능을 평가하는 표준적인 방법은 단어오류율(Word Error Rate, WER)을 측정하는 것이다(Jurafsky & Martin, 2009). 단어오류율은 자동음성인식기가 반환한 단어들이 원본과 얼마나 다른지에 기반하여 측정된다. 단어오류율의 계산식은 아래와 같으며, 단어오류율이 낮을 수록 음성인식기가 정확하다는 의미이다. 유사하게 음소인식기반 음성인식기의 경우에는 음소오류율(Phone Error Rate, PER)을 사용하여 성능을 평가할 수 있다.

$$\textit{단어오류율}\,(\%) = \frac{\textit{대체} + \textit{탈락} + \textit{삽입된 단어 수}}{\textit{원본 내 전체 단어 수}} \times 100$$

(Jurafsky & Martin, 2009)

자동음성인식기 API 를 활용하여 성능을 비교한 연구들은 크게 단어오류율을 기준으로 성능을 평가한 연구와 자체 기준으로 성능을 평가한 연구로 나뉜다. 단어오류율을 기준으로 자동음성인식기의 성능을 서로 비교하여 평가한 연구는 다음과 같다.

Këpuska & Bohouta(2017)는 다양한 코퍼스로부터 추출한 미국식 영어 음성 자료에 대해 Microsoft API, Google API, CMU Sphinx 세 개의 자동음성인식기의 성능을 단어오류율을 기준으로 비교한 결과, Google 의 단어오류율이 9%로 가장 우수하게 나타났다.

Kodish-Wachs, Agassi, Kenny, & Overhage(2018)는 임상 대화 자료를 대상으로 하여 단어오류율을 기준으로 Bing Speech API, Google Cloud Speech API, IBM Speech to Text, Microsoft Azure MAVIS1, MAVIS2, Nuance, Amazon Transcribe, Mozilla DeepSpeech 총 8 개 자동음성인식기의 성능을 비교하였다. 그 결과 자동음성인식기의 단어오류율이 평균 약 50%대로 높게 나왔으며 Microsoft Azure Mavis2 의 단어오류율이 35%로 가장 낮게 나왔다.

IANCU(2019)는 루마니아어 음성자료에 대해 단어오류율(WER)을 기준으로 Google Speech-to-Text API 의 성능을 평가했다. 그 결과 단어오류율이 30.96%로 다소 높게 나타났다. 동영상을 출처로 한 음성자료를 사용함으로 인해 오디오의 음질과 소음 문제로 인하여 자동음성인식기의 인식률에 영향을 끼치는 한계가 있었다.

유현재, 김명화, 박상길, & 김광용(2020)은 뉴스 음성자료에 대해 단어오류율을 기준으로 Google, Amazon, IBM, MS, Naver, Kakao, ETRI 총 7 개의 자동음성인식기 성능을 비교하였다. 그 결과 자동음성인식 엔진은 Kakao 의 정확도가 94%로 전체적으로 가장 좋은 성능을 보였으며, 뉴스 음성자료의 분야별로 자동음성인식 엔진마다 정확도의 차이를 보이는 특성이 있음을 확인했다.

단어오류율(WER)을 측정하는 것의 한계도 존재한다. 단어오류율 계산식에서는 모든 단어가 동일한 가중치를 갖고 계산이 되는데, 실제 문장내의 모든 단어가 동일하게 중요한 역할을 담당한다고 보기는 어렵다. 내용어의 경우 기능어보다 가중치를 높게 하여 계산하는 것이 문장내 단어의 중요도를 반영하는 대안이 될 수 있다. 그러나 어휘별로 가중치를 조정하면서 모든 자동음성인식기에 적용될 수 있는 계산식을 만드는 데 어려움이 있는 실정이다(Jurafsky & Martin, 2009). 또한 외국어

학습자의 음성에는 원본과는 다른 발음오류가 포함되어 있을 수 있는데, 발음오류와 오인식의 구분이 어렵다는 문제가 있다.

자체적인 기준으로 자동음성인식기의 성능을 비교하여 평가한 연구는 다음과 같다.

최승주 & 김종배. (2017)의 연구에서 Google, 카카오, 네이버 세 자동음성인식기의 성능을 숫자음, 한국어 음절, 문장의 세 가지 범주별로 10 회 반복하여 오인식률을 비교한 결과, 전반적으로 카카오의 성능이 우수하게 나타났다.

노희경 & 이강희(2017)는 성별, 나이, 방언별로 자유발화한 문장을 대상으로 원본과 틀린 개수를 기준으로 Google, 카카오, 네이버 세 개의 자동음성인식기의 인식률을 살펴보았다. 그 결과 표준어에서 전체적인 정확도는 Google 이 가장 높게 나타났으며, 방언마다 음성인식 정확도가 달라졌다.

대부분의 연구에서 자유발화한 음성자료를 대상으로 실험을 진행했다. 그러나 발음훈련 프로그램을 개발하는 데 있어서는 자유발화를 대상으로 한 인식률 평가가 얼마나 신뢰성이 있는지 알 수 없다. 발음의 정오와 상관 없이 문맥 정보가 음성인식에 영향을 끼쳤을 수 있기 때문이다. 자동음성인식기 시스템은 주변 단어열을 고려하는 언어 모델을 활용하기 때문에, 자유발화한 문장 속 단어의 경우 발음의 정오와 상관 없이 주변 단어열을 통해 해당 단어가 확률적으로 유추된다. 발음훈련 프로그램 개발을 위해서는 자동음성인식기가 발음의 정오를 정확히 구분할 줄 아는 음성학적 엄밀함을 갖춰야 하기 때문에, 자동음성인식기가 이러한 과업이 가능한지 파악하기 위해서는 실험 자료 역시 문맥 정보와 독립적으로 목표 단어와 그 주변 단어열을 통제하여 살펴볼 필요가 있다. 이처럼 개별 단어 차원에서 통제된 실험 자료를 활용하여야 자동음성인식기가 발음의 정오를 구분하고 있는지 정확하게 파악할 수 있을 것이다.

이에 이 연구에서는 외국어 산출 교육을 위한 자동음성인식 소프트웨어 연구의 기초를 마련하기 위하여, 외국어 학습자의 발음을

대상으로 하여 문맥 정보와 독립적으로 통제된 환경에서 자동음성인식기의 성능을 평가하고자 한다.

1.2.2. 외국어 학습자의 발화에 대한 자동음성인식기의 성능에 관한 연구

윤정희(2014)는 초등학생 영어 학습자의 발화에 대한 Google Voice Actions 의 인식률을 평가하였다. 초등학교 고학년 학생 16 명을 대상으로 초등학교 권장어휘 219 개 단어들을 녹음하고 자동음성인식기에 들려준 결과 평균인식감도는 73.18 점으로 나타났다. 또한 인식감도가 낮게 나타난 단어들의 발음 저해요인과 인식 저해요인을 분석하였다. 발음 저해요인으로는 한국어 음운체계에 없는 자음군(Consonant Cluster)과 이중모음(Diphthong)의 비중이 크게 나타났다. 인식 저해요인으로는 단어의 길이가 영향을 미치는 것으로 설명하였다. 단어의 길이가 길수록 데이터베이스 내에 있는 비교단어와의 유사성을 발견할 확률이 높아지기 때문에, 긴 단어는 몇 가지 음운을 부정확하게 발음하더라도 짧은 단어에 비해 더 정확히 인식할 가능성이 높아진다. 반면 짧은 단어는 음성 데이터베이스에서 비교할 수 있는 단어의 특성이 적기 때문에 음운의 발음이 조금이라도 부정확할 경우 인식확률이 급격히 낮아지게 된 것으로 분석했다.

양병곤(2017)은 한국인 영어 학습자의 발화에 대한 구글 자동음성인식기의 인식률을 평가하였다. 영어교육을 전공하는 대학생 33 명을 대상으로 영문 텍스트를 녹음하고 Google Speech Recognition Soundwriter 자동음성인식기에 인식시킨 결과 평균 73%의 인식률을 보였다. 양병곤(2017)은 자동음성인식기의 오인식률을 높인 요인으로 기능어와 마찰음을 들었다. ‘for, can, will’ 같은 기능어가 높은 오인식 빈도수를 보이는 것으로 나타났다. 또한 마찰음으로 시작하는 단어의 오인식률이 높은 것으로 나타났다. 특히 [s]로 시작하는 단어가 눈에 띄게 오인식률이 높았는데, 그 이유는 [s]의 음향학적 특징으로 인해 인식기가 제대로 인식하지 못하는 것으로 설명했다. [s]는 음향적으로 공명도가 다른 자음에 비해 낮고, 마찰음 스펙트럼에서 4000~8000Hz 사이에 절단주파수가 나타나 이 부분이 자동음성인식기에 제대로 입력되지 않아 인식 오류를 가져왔을 것으로 보았다.

박향숙, 이예식, & 윤정희(2018)는 한국인 영어학습자들이 발음 오류를 산출할 것으로 예상되는 영어 단어의 음운 요소와 친숙도가 음성인식 감도에 미치는 영향을 조사했다. 그 결과 발음 오류가 예상되는 음운 요소가 많을 수록 인식 감도가 낮아지며, 친숙도가 높을 수록 인식 감도가 높아지는 것으로 나타났다. 그러나 초등학생만을 대상으로 연구를 진행했기 때문에 영어 학습 기간이 긴 학습자를 대상으로 특히 어려운 음운 요소들을 선별하여 연구를 진행할 필요가 있음을 시사했다.

Ashwell & Elam(2017)은 일본인 영어 학습자의 발화에 대한 Google Web Speech API 의 인식 정확도를 측정했다. 실험은 단순한 영어 문장을 대상으로 일본인 영어 학습자의 발화와 원어민의 발화를 비교하였다. 그 결과 원어민의 경우 정확도가 평균 89.4%로 나타났으며, 일본인 영어 학습자의 경우 65.7%로 나타났다. 원어민과 비원어민의 발화를 비교했다는 점에서 의미가 있으나, 정확도를 측정하는 방법에 대해 명확하게 서술하지 않은 점이 한계이다.

Kim(2006)은 한국인 영어 학습자의 발음을 평가함에 있어서 자동음성인식 소프트웨어 Fluspeak 이 평가한 점수와 원어민 화자가 평가한 점수를 비교하여 그 상관관계를 밝혔다. 그 결과 단어 차원에서의 평가 점수 상관계수는 0.56($p<0.01$)로 높지 않게 나타났으며 특히 억양 차원에서의 평가 점수 상관계수는 0.06($p<0.05$)으로 거의 0 에 가깝게 낮게 나타났다. Kim(2006)의 연구는 원어민 화자와 자동음성인식기의 판단 능력을 직접적으로 비교하고자 했다는 데 의의가 있다. 그러나 해당 연구가 진행된 시점에 비해 현재의 음성인식기술이 많이 발전하여 새로 연구를 진행할 필요가 있으며, FluSpeak 라는 하나의 소프트웨어를 대상으로 한 것 이외에 현재 시판되고 있는 더욱 다양한 자동음성인식 소프트웨어를 대상으로 비교 연구를 진행할 필요가 있다.

Guskaroska(2019) 역시 원어민 화자의 판단과 자동음성인식기의 인식 결과를 직접 비교한 연구이다. 이 연구는 모음 훈련에 있어서 Siri 와 같은 모바일 자동음성인식기의 유용성을 검증해보고자 자동음성인식기를

활용한 발음 훈련의 효과와 자동음성인식기의 인식 정확도, 학습자의 태도 세 가지 측면을 살펴보았다. 모음 대립 4 가지 쌍 /i-ɪ/, /æ-ɛ/, /u-ʊ/, /ɑ-ʌ/에 대해 최소대립쌍 단어를 실험 자료로 활용하였다. 자동음성인식기의 성능은 비원어민 화자에 대해 인식률이 약 57%로 나타났으며, 자동음성인식기를 활용하여 발음 훈련을 진행한 결과 실험군의 발음 정확도가 향상되었고 학습자들이 자동음성인식기에 대해 긍정적인 태도를 보인 것으로 나타났다. 이 연구는 자동음성인식기의 성능을 점검함과 동시에 훈련 효과를 검증하였다는 점에서 의미가 있다.

Escudero-Mancebo et al.(2015) 역시 임의의 숫자음이나 자유발화를 실험 자료로 하지 않고 언어학적으로 의미 있는 최소대립쌍을 실험 자료로 자동음성인식기의 성능을 평가했다는 점에서 의미가 있다. 최소대립쌍을 구분하는 것은 자동음성인식기에게 어려운 작업이지만, 언어학적 의미 전달에 있어서 매우 중요한 작업이기 때문에 자동음성인식기가 최소대립쌍을 어느 정도로 변별하는가를 판단하는 것은 중요하다. Escudero-Mancebo et al.(2015)은 최소대립쌍 단어들을 대상으로 자동음성인식기의 인식 성공률을 원어민과 고급학습자, 초급학습자 세 수준별로 비교하여 인식 성공률과 발음 유창성이 서로 관련이 있는 것으로 밝혔다. 또한 ‘wreathe, luff, wader’ 등 일부 단어들에서 주로 나타난 오인식의 원인이 단어의 음성학적 구조 때문이 아니라 언어 모델에서 나타나는 빈도가 낮기 때문으로 분석했다. 즉, 자동음성인식 시스템에서 참조로 하는 언어 모델에 해당 단어의 빈도수가 낮으면 제대로 인식이 되지 않는 것으로 보았다. 이를 통해 자동음성인식기의 성능을 판단하기 위해서는 빈도수가 낮지 않은 단어를 활용해야 함을 알 수 있다.

대부분의 연구가 학습자의 발음 오류와 자동음성인식기의 오인식을 구분하지 않고 있어서 발음의 정확도에 대한 인식률을 파악하는 데 어려움이 있다. 오인식된 것이 곧 발음 오류라고 판단하거나, 발음 오류가 있는데도 제대로 인식한 것을 인식 정확도에 포함시킴으로써 실제 발음 오류 정도와 인식 정확도의 구분이 어렵다는 한계가 있다.

이에 이 연구에서는 원어민이 직접 학습자의 발음의 정오를 판단함으로써 실제 발음 오류와 자동음성인식기의 오인식을 구분하고자 한다. 발음 오류는 원어민이 한국인 학습자의 발음을 듣고 목표음과 틀린 것으로 판단한 것으로 규정하며, 오인식은 자동음성인식기가 원어민의 판단과 다른 결과를 산출한 것으로 규정할 수 있다. 이를 통해 자동음성인식기가 옳은 발음은 옳게 인식하는지, 틀린 발음은 오류로 제대로 인식하는지 정확히 구분하여 파악함으로써, 인공지능 자동음성인식기가 발음 진단을 얼마나 정확하게 수행할 수 있는지 확인하고자 한다.

1.2.3. 음소인식기반 자동음성인식기를 활용한 발음 오류 검출 및 발음 교육에 관한 연구

음소인식기반 자동음성인식기를 활용하여 발음 오류를 검출하고 진단하는 연구로는 다음과 같다.

Tepperman & Narayanan(2008)은 조음 자질을 활용하여 비원어민의 발음에서 분절음 오류를 검출하였다. 이 때 조음기관의 움직임을 턱, 입술, 혀의 움직임 등으로 범주화하여 조음 특성을 모델링함으로써 발화 오류를 검출하고자 했다. 그 결과 비원어민 화자의 발음에서 분절음 레벨의 오류를 검출해내는 능력이 향상됨을 확인하였다. 이 연구는 조음 자질이라는 새로운 접근방식을 통해 외국어 학습자가 음 레벨의 오류를 산출하는 것을 자동으로 검출해내어 향후 외국어 교육을 위한 프로그램 개발에 기여했다는 점에서 의미가 있다.

류혁수 & 정민화(2016a)는 한국인 영어 학습자가 발화한 낭독체 문장에 대해 조음 자질 기반의 사후 확률인 조음 발음 적합 점수(articulatory Goodness-Of-Pronunciation, aGOP)를 통해 자음 발음 오류를 자동으로 검출하였다. 한국인 영어 학습자 발화에서 GOP 와 aGOP 를 계산한 후, 이를 이용하여 오류 검출 모델을 구성한 결과, aGOP 자질을 함께 활용하였을 때 오류 검출 성능이 향상되었다. 이 연구는 발음 오류 자동 검출 모델링에 조음 자질을 반영함으로써 추후 학습자에게 조음 교정 피드백을 제공할 수 있도록 하였다는 점에서 의미가 있다.

Korzekwa et al.(2021)은 발음의 불확실성을 고려하여 발음 오류 검출의 정밀도를 향상시켰다. 보통의 발음 오류 자동 검출 방식은 원어민의 예상 발음과 학습자의 발음을 비교하는 것인데, 여기에는 음소 인식의 불확실성과 발음의 변이를 고려하지 않는다는 문제가 있다. 이에 저자는 음소 인식 단계에서의 불확실성을 고려하고, 하나의 문장일지라도 여러 가지 발음이 가능한 점을 고려한 새로운 접근 방식을 제안했다.

Wu et al.(2021)은 발음 오류 검출과 진단(Mispronunciation Detection and Diagnosis, MDD)을 위해 두 개의 트랜스포머 기반 아키텍처를 도입하여 성능을 향상시켰다. 첫 번째 트랜스포머 아키텍처는 크로스 엔트로피 로스(Cross Entropy loss)를 포함한 인코더, 디코더로 이루어진 표준적인 방식이며, 두 번째 아키텍처는 wav2vec 2.0 을 기반으로 한다. 첫 번째 아키텍처의 경우 PER 이 8.69%, 두 번째 아키텍처의 경우 PER 이 5.97%로 나타나 이전의 모델보다 향상된 성능을 보였다. MDD 의 성능을 점검하기 위해 FRR(False Rejection Rate), FAR(False Acceptance Rate), DER(Diagnosis Error Rate) 등 다양한 지표를 사용했다는 점에서 의미가 있다.

Ye et al.(2022)은 음소 단위에서 주석 처리가 된 대용량의 훈련 데이터를 구하기 어려워서 MDD 의 성능 향상에 제약이 있는 점을 고려하여, 단어 단위에서 주석 처리가 된 대용량의 데이터로 훈련된 ASR 모델로부터 추출한 음성 임베딩을 사용하였다. 저자들은 음향, 음성, 언어학적(Acoustic, Phonetic and Linguistic, APL) 임베딩을 사용하여 더욱 강력한 MDD 시스템을 개발하는 것을 제안하였고, PER 이 16.96%로 나타나 베이스라인보다 향상된 성능을 보였다.

발음 오류를 검출한 것에서 나아가 피드백 및 평가를 제공하는 연구는 다음과 같다.

Li et al.(2016)은 지식 기반 및 데이터 기반 결정 트리를 활용하여 외국어 학습자의 발음 오류를 검출하고 피드백을 제공하는 프로그램을 개발하였다. 분절음 레벨의 오류를 검출하고 조음 방법과 조음 위치에 기반하여 조음 레벨의 피드백을 제공하기 위해 조음 자질을 기반으로 하는 결정 트리를 제안하였다. 이 연구는 전통적인 점수 기반의 시스템과는 달리, 외국어 학습자에게 발음 오류가 왜 일어났으며 어떻게 교정할 수 있는지 알려준다는 점에서 의미가 있다.

류혁수 & 정민화(2016b)는 한국인 학습자의 영어 발화를 자동으로 평가하는 프로그램을 개발함에 있어서 조음 자질인 aGOP 를 평가 자질로 제안하였다. 이 때 조음 자질은 조음 위치와 조음 방법, 유성성을

포함하여 언어학적으로 거의 모든 변별적 자질을 활용하였다. 기존의 발음 평가 연구에서 사용된 길이, 속도와 같은 자질들을 포함하여 변별적 자질을 표현하는 aGOP 를 더해 평가 모델의 발음 점수 예측 성능을 살펴본 결과, aGOP 자질들을 포함할 때 평가 점수 예측 성능이 향상되었다. 이를 통해 기존의 발음 평가 연구에서는 주로 길이, 속도와 같은 자질들을 사용한 것에 더해 음성학 및 음운론적 자질을 활용하여 발음 평가 성능을 향상시켰다는 점에서 의미가 있다.

Xie et al.(2020)은 중국어 학습자를 위한 발음 오류 검출 및 교정 피드백을 제공하는 앱을 개발하였다. 이 앱은 학습자가 단어를 발음하면 음성 단위에서 발음 오류를 검출하여 화면에 표시하고 전반적인 발음 점수도 보여준다. 발음 오류가 있는 부분에 대해서는 표준 발음을 들어볼 수 있고 조음 기관의 움직임을 볼 수 있다. MDD 의 성능은 발음을 맞게 감지한 것의 비율인 DA(Diagnostic Accuracy)가 79% 정도로 나타났다. 이 앱을 활용하여 6 주간 발음 훈련을 진행한 결과, 학습자들이 스스로 발음 오류를 약 83% 정도 교정할 수 있었다. 이 논문은 MDD 의 성능과 함께 자체 개발한 앱을 이용한 발음 훈련의 효과를 동시에 보여주었다는 점에서 의미가 있다.

Xie et al.(2020)이 분절음 수준의 발음 교육 및 평가를 진행했다면, Hirschi et al.(2020)은 초분절음 수준의 발음 교육 및 평가를 진행했다. Hirschi et al.(2020)은 Novo Play 라는 모바일 앱을 활용하여 제한적인 영어 유창성을 갖는 화자들에 대해 어휘적 강세(lexical stress)와 두드러짐(prominence) 교육을 약 2 주간 실시했다. 앱을 통해 실시한 각 레슨에서는 어휘 강세가 있는 부분을 시각적으로 표시해서 보여주고 원어민의 음성 샘플도 들려주었으며, 학습자가 틀린 발음을 하면 피드백을 보여주었다. 교육 진행 전과 진행 후에 수집한 발화에 대해 숙련된 평가자들이 학습자의 운율을 평가한 결과, 운율 교육이 이해 가능성(comprehensibility)에 있어서는 효과적이었지만 악센트 정도(accentedness)에 있어서는 효과가 거의 없는 것으로 나타났다.

이처럼 음소인식기반 자동음성인식기를 활용하여 발음을 인식하여 오류를 검출하고 발음을 평가하는 등 발음 교육적 측면에서 자동음성인식기의 활용 가능성을 살펴볼 수 있다. 그러나 이들 연구에 사용된 자동음성인식기의 성능에 대해서는 제대로 검증되어 있지 않은 경우가 많고, 검증되어 있더라도 언어학적 단위별로 나누어 세부적인 성능은 알 수가 없다. 발음 오류 검출 및 발음 평가를 하기 이전에 자동음성인식기의 성능을 세부적으로 검증하는 단계가 필요하나 이에 대한 언급이 없다는 한계가 있다. 이에 음소인식기반 자동음성인식기의 성능을 단어인식기반 자동음성인식기와 함께 확인하고자 한다. 발음 교육적 측면에서 성능을 확인하기 위하여 언어학적 단위별로 나누어 다방면에서 성능을 평가하고자 한다.

1.2.4. 한국인 영어 학습자의 발음 오류에 관한 연구

Lee & Hwang(2016)은 한국인 영어 학습자에게 지각훈련을 진행하여 영어 분절음의 학습용이성(learnability) 위계를 선정했다. 학습용이성은 훈련 전 지각 정확도와 훈련 후 지각 향상도를 모두 포함하는 개념으로, 학습용이성이 높은 음은 훈련 효과가 크며 학습용이성이 낮은 음은 발음 오류가 쉽게 고쳐지지 않는다는 것을 뜻한다. 초등학생을 대상으로 고변이 음성훈련(High Variability Phonetic Training)을 진행하여 한국인이 어려워하는 음소 쌍에 대한 학습용이성을 측정한 결과, 모음에서는 /u-ʊ/, 자음에서는 /ʒ-dʒ/ 쌍의 학습용이성이 낮은 것으로 나타났다.

유혜배 & 윤한나(2009)는 한국인 영어학습자의 자유발화를 녹음하여 대화상에 나타난 분절음 발음 오류를 살펴보았다. 그 결과, 영어의 이중모음을 단음화하여 발음하는 오류가 빈번했으며, 한국어 자음에는 없는 유성음과 마찰음, 유음 /r/ 발음에서 난점을 보였다. 피험자가 영어 학습기간이 긴 영어 전공자임에도 불구하고 이러한 발음 오류들을 빈번하게 산출하였다는 점에서 의미가 있다.

박향숙, 이예식, & 윤정희(2018)는 한국어와 영어의 음운체계 차이를 바탕으로 한국인 영어학습자가 발음하기 어려워하는 음운 요소를 제시했다. 첫째, 자음에 있어서 변별하는 자질이 다르고 마찰음의 개수가 확연히 차이 난다. 영어의 자음은 크게 유성음과 무성음이 서로 변별되지만 한국어의 장애음은 무성음으로 발음되어 유성성으로 변별되지 않는다. 영어의 마찰음은 /f, v, s, z, ʃ, ʒ, θ, ð, h/로 9 가지이지만 한국어의 마찰음은 /s, sʷ, h/로 3 가지이다. 따라서 한국인 영어학습자들은 한국어에 없는 유성음과 마찰음과 같은 영어 자음을 비슷한 한국어 자음으로 대체해서 발화하는 경향이 있다. 둘째, 모음에 있어서 영어는 긴장도가 변별적이다. 영어에서는 긴장-이완 모음의 구분이 확연하지만 한국어는 그렇지 않다.

이에 이 논문에서는 한국인 영어 학습자의 발음의 정오에 대하여 인공지능 자동음성인식기와 원어민의 판단을 비교하기 위해, 한국인

영어학습자의 발음 오류를 의도하여야 하므로 모음은 긴장-이완 모음 대립을 중심으로 하고, 자음은 마찰음과 유음, 유성음을 중심으로 실험 자료를 구성하고자 한다.

2. 실　　험

2.1. 실험 방법

한국인 학습자가 발음한 영어 단어 및 문장을 대상으로 인공지능 자동음성인식기의 인식 결과와 영어 원어민의 판단을 비교하여 인식률을 확인함으로써, 인공지능 자동음성인식기를 발음 산출 교육에 사용하려 할 때 어느 정도의 신뢰도를 보이는지 밝히고자 한다.

2.1.1. 피험자

(1) 한국인 영어 학습자

한국인 영어 학습자는 영어 단어와 문장 녹음을 수행하며, 이 때 녹음된 음성을 자동음성인식기와 원어민이 듣고 발음의 정오를 판단한다. 한국인 영어 학습자는 한국어를 모국어로 하는 20~30 대 남녀 각 10 명씩 총 20 명을 모집했다. 한국인 영어 학습자의 영어 수준은 TOEIC 점수를 기준으로 평균은 약 875 점, 표준편차는 약 98 점이었다.

(2) 영어 원어민 화자

영어 원어민 화자는 비교군으로서 영어 단어와 문장을 녹음하고, 한국인 영어 학습자가 녹음한 단어와 문장을 듣고 발음 오류를 판단한다. 영어 원어민 화자는 북미 영어 화자 20~30 대 남자 1 명과 여자 2 명으로 총 3 명을 모집했으며, 뉴욕주와 워싱턴주, 캘리포니아주 출신으로 구성되었다.

2.1.2. 실험 자료

실험 자료가 될 영어 단어 최소대립쌍은 개별 단어 형태와 문장 형태를 모두 활용한다. 단어들을 (1) 단어 단독으로 발음하게 하고, 문장틀에 넣어 (2) 문두 "____ she said", (3) 문중 "Say ____ again", (4) 문미 "I said ____" 각각의 위치에서 발음하게 한다. 이를 통해 문맥 정보가 주어지지 않았을 때 자동음성인식기가 얼마나 발음을 정확하게 인식하는지 확인하고자 한다.

실험 자료가 될 자음과 모음 쌍은 선행연구에서 한국인 학습자가 구분하여 발음하기 어려워하는 것들로 선정하였다. 자극음은 자동음성인식기가 단어로 인식하지 못할 가능성을 배제하기 위해 모두 유의미어를 사용했다. 또한 단어의 빈도수가 인식률에 영향을 끼친다는 선행연구를 참고하여(Escudero-Mancebo et al, 2015) 최소대립쌍 단어들은 빈도수가 낮지 않은 단어로 선정하였으며, 테스트를 거쳐 자동음성인식기가 비교적 일관되게 잘 인식하는 단어로 선정했다.

(1) 모음

5 개의 최소대립쌍을 대상으로 총 10 개의 실험 단어를 선정하였다. 선행연구를 참고하여 한국인 학습자가 발화할 때 오류를 내기 쉬운 자음은 최대한 제외하여 모음 오류만을 통제할 수 있도록 단어쌍을 선정하였다.

<표 1> 자극음으로 선정한 모음 최소대립쌍

모음 대립쌍	최소대립쌍 단어
iː – ɪ	heat-hit
æ – ɛ	pan-pen
uː – ʊ	Luke-look
ɔː – oʊ	pause-pose
ɑː – ʌ	cop-cup

(2) 자음

총 6 개 최소대립쌍을 대상으로 총 24 개 단어를 선정하였다. 자음쌍은 선행연구에서 한국인 학습자가 구분하여 발음하기 어려워하는 것들로 선정하였으며, 특히 상급 학습자도 발음하기 어려워 하는 6 개의 자음쌍으로 선별하여 발음 오류를 살펴보고자 한다. 또한 선행연구에서(Lee & Hwang, 2016) 한국인 학습자가 발화할 때 오류를 내기 쉬운 모음은 최대한 제외하여 자음 오류만을 통제할 수 있도록 단어쌍을 선정하였다. 자음의 경우 어두와 어중, 어말 3 가지 위치에서 인식률이 달라지는지 확인하기 위해, [p-f], [b-v], [d-ð] 3 개의 자음쌍에 대해 세 가지 위치 모두에서 변별하는 최소대립쌍을 선정하였다. 2 개의 자음쌍 [l-r], [s–θ]의 경우, 어두에서 변별하는 것으로 선정하였고, [ʒ-dʒ]의 경우에는 어두에서 변별하는 최소대립쌍을 찾을 수 없어서 어중에서 변별하는 쌍으로 선정하였다.

<표 2> 자극음으로 선정한 자음 최소대립쌍

자음 대립쌍	어두 대립	어중 대립	어말 대립
p – f	pat-fat	coffee-copy	beef-beep
b – v	base-vase	rebel-revel	curb-curve
d – ð	day-they	header-heather	breed-breathe
l – r	lay-ray	-	-
s – θ	sink-think	-	-
ʒ – dʒ	-	pleasure-pledger	-

2.1.3. 실험 절차

자세한 실험 절차는 다음과 같다.

(1) 한국인 영어 학습자에게 영어 단어 및 문장을 각각 1 번씩 발음하게 했다. [2] 녹음은 방음시설이 있는 서울대학교 음성실험실에서 진행했으며 16bit, 44100Hz 로 녹음했다.

(2) 영어 원어민 화자에게 동일한 영어 단어 및 문장을 각각 1 번씩 발음하게 했다. 마찬가지로 녹음은 음성실험실에서 진행했으며 16bit, 44100Hz 로 녹음했다.

(3) 영어 원어민 화자가 한국인 학습자가 발음한 단어 및 문장을 어떤 단어 및 문장으로 인지하는지 지각실험을 진행했다. 피험자는 헤드폰을 통해 한국인 영어 학습자가 발화한 음성을 듣고 제시된 최소대립쌍 단어들 중에서 적합한 것을 고르는 구별 과제(identification test)를 수행했다.

(4) 녹음된 영어 발화 음성을 각 자동음성인식기가 어떤 단어 및 문장으로 인식하는지 보기 위해, 자동음성인식기에 자극음 음성을 입력하고 어떤 텍스트를 생성해내는지 확인했다.

각 화자가 발음해야 하는 단어는 모음 최소대립쌍 10 개 단어와 자음 최소대립쌍 24 개 단어를 합하여 34 개이다. 이것을 개별 단어 형태와 문두 “____ she said”, 문중 “Say ____ again”, 문미 “I said ____”에 넣은 형태로 발화하여 한 명의 화자가 총 136 개(34 개 * 4)의 단어 및 문장을 발화하였다. 한국인 학습자 20 명에게서 총 2720 개(136 개 * 20 명)의 녹음 데이터를 확보하였다. 원어민 화자가 발음하는 단어 및 문장 수는 한국인 학습자와 동일하게 136 개이며, 총 3 명의 화자에게서 408 개(136 개 * 3 명)의 녹음 데이터를 확보하였다.

[2] <부록 1: 녹음 스크립트> 참조

한 명의 영어 원어민 화자가 듣게 되는 한국인 학습자가 발화한 음성의 수는 총 2720 개(136 개 * 20 명)이며, 실험결과로 생성되는 전체 분석 대상은 총 8160 개(2720 개 * 원어민 피험자 3 명)가 된다.

인공지능 자동음성인식기가 한국인 학습자의 영어 발화를 어떤 음으로 인식하는지 보기 위해, 자동음성인식기에 한국인 학습자가 녹음한 자극음 2720 개와 영어 원어민 화자가 녹음한 자극음 408 개, 총 3128 개를 입력하고 어떤 텍스트를 생성하는지 확인하였다. 6 가지의 자동음성인식기에서 총 18768 개(자극음 3128 개 * 6 가지 음성인식기)의 분석 데이터를 확보하였다.

이 실험에서는 Google 3.0.0 버전, Microsoft azure-cognitiveservices-speech 1.18.0 버전, ibm-watson 5.2.3 버전, AWS Transcribe boto3 1.18.48 버전, Naver CLOVA Speech 1.4.0 버전을 활용하여 실험을 진행하였다. 음소인식기반 음성인식기는 wav2vec 2.0 Large (Baevski et al., 2020) 모델을 TIMIT 데이터로 훈련시킨 것으로, 여기서 TIMIT 데이터 세트의 PLU 세트는 Lee & Hon(1989)에서 제시한 방법대로 39 개로 합쳐서 진행했다. 훈련 결과 이 음소인식기반 자동음성인식기는 TIMIT 테스트 세트에 대해서 7.77%의 음소오류율(Phone Error Rate, PER)을 보였다.

2.1.4. 분석 절차

한국인 학습자의 영어 발음에 대한 인공지능 자동음성인식기의 인식 결과와 원어민의 판단을 비교하여 서로 어느 정도로 일치하는지 파악하기 위해 다음과 같이 분석하였다.

(1) 한국인이 발음한 목표음과 음성인식기가 반환한 텍스트의 발음이 일치하면 1, 불일치하면 0 으로 표시하였다.
(2) 원어민이 구별 과제를 수행하여 한국인 학습자의 발음이 목표음과 일치하면 1, 불일치하면 0 으로 표시하였다.[3]
(3) 1 과 0 으로 표시되어 있는 음성인식기의 응답 결과와 각 원어민의 응답 결과를 비교하여, 서로 일치하면 1, 불일치하면 0 으로 표시하였다.
(4) (3)에서 표시한 전체 항목 대비 일치(1)하는 것의 개수를 계산하여 인공지능 자동음성인식기와 원어민의 일치도를 구하였다.

(1)의 경우, 음성인식기가 생성하는 텍스트가 목표 단어나 문장과 달리 다양하게 나타날 수 있기 때문에, 발음의 일치와 불일치에 대한 세부적인 표시 기준을 아래와 같이 정하였다.

(1) 동음이의어인 경우
목표 단어와 다르지만 음이 같으므로 일치하는 것으로 보았다.
예) (목표 단어-인식 결과) base-bass, ray-REI, sink-sync
(2) 문장틀이 다른 경우
목표 단어는 제대로 인식했지만 문장틀을 제대로 인식하지 못한 경우 일치하는 것으로 보았다.

[3] 원어민 전사자간 일치도는 Fleiss's Kappa 상관계수가 0.58 로, 신뢰도는 보통 정도로 나타났다.

예) “Base she said”-“Base is she said”, “I sad pen”-“As a pen”, “Say pat again”-“C pat again”

(3) 목표음 외에 다른 음운이 차이 나는 경우

가장 흔한 경우로, 최소대립쌍에서 목표음을 제대로 인식했다는 점에서 일치하는 것으로 보았다.

예) curb-carb, breed-read, think-thank, pose-post, sink-seeing

(4) 목표음은 일치하지만 단어 경계가 다른 경우

단어 경계가 다른 경우에는 불일치하는 것으로 보았다.

예) “Say cup again”-“Fake up again”, “Coffee she said”-“Cough fish is set”, “heather”-“Hi there”

(5) 인식 결과가 없는 경우

음성인식기가 생성하는 텍스트가 없는 경우에는 불일치하는 것으로 보았다.

이러한 과정을 거쳐 분석 대상이 되는 데이터의 모습은 다음과 같다. 지면의 편의상 구글 자동음성인식기 하나와 원어민 한 명을 예시로 들었다. 가장 오른쪽 열에서 자동음성인식기와 원어민의 일치 여부를 파악할 수 있다.

<표 3> 데이터 분석 예시

발음	목표음	구글	구글응답결과	원어민응답결과	구글_원어민
heat	heat	heat	1	1	1
hit	heat	hit	0	0	1
they	they	day	0	1	0
day	they	D	0	0	1

2.2. 결과

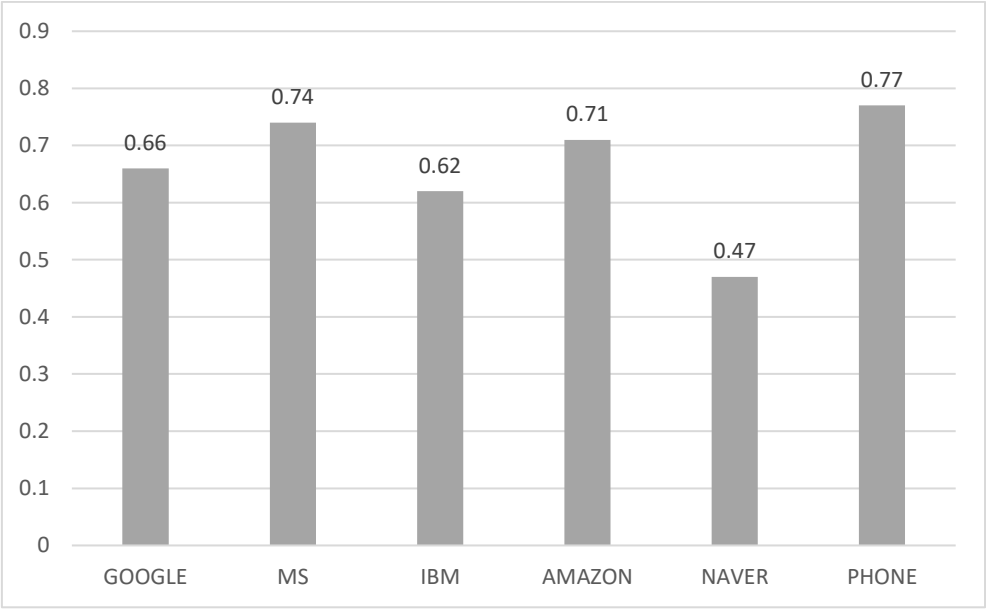
문맥 정보가 주어지지 않았을 때 인공지능 자동음성인식기의 성능을 평가한 결과는 다음과 같다. 여기서 자동음성인식기의 인식률은 자동음성인식기와 원어민의 응답 일치율과 같은 의미로 사용하였다.

원어민 화자와 각 인공지능 자동음성인식기의 응답 결과 중 서로 일치하는 것의 비율을 계산하여, 한국인 학습자의 영어 발음에 대한 원어민 화자와 각 인공지능 자동음성인식기의 응답 결과가 서로 얼마나 일치하는지를 확인한 결과 <표 4>와 <그림 1>과 같이 나타났다. 음소인식기반 음성인식기의 일치율이 0.77 로 가장 높게 나타났으며, 그 다음으로 Microsoft Azure Speech Service, Amazon Transcribe, Google Cloud Speech-to-Text, IBM Watson Speech to Text 순서였으며 Naver CLOVA Speech 가 가장 낮게 나타났다. Naver CLOVA Speech 의 경우에는 개별 단어 형태에 대한 음성 인식 결과가 없는 경우가 대부분이었기 때문에 다른 음성인식기와 달리 유독 일치율이 낮게 나타났다.

원어민이 직접 한국인 학습자의 발음을 판단함으로써 발음 오류와 오인식을 구분하여 살펴볼 수 있었다. 발음 오류는 원어민이 한국인 학습자의 발음을 듣고 목표음과 다르다고 체크한 것을 기준으로 하였으며, 한국인 학습자의 발화 중 약 12%가 발음 오류로 나타났다. 발음 오류가 나타난 발화 중에서 자동음성인식기가 발음 오류를 제대로 파악한 비율은 Google 은 0.7, Microsoft 는 0.73, IBM 은 0.73, Amazon 은 0.74, Naver 는 0.45, 음소인식기반 음성인식기는 0.74 로, Naver 를 제외하고는 발음 오류 파악 비율이 거의 유사하게 나타났다. 오인식은 자동음성인식기의 판단이 원어민의 판단과 불일치한 것으로, 오인식률은 1 에서 일치율을 뺀 값이 된다. 오인식률은 Naver CLOVA Speech 가 0.53 으로 가장 높게 나타났다.

<표 4> 각 인공지능 자동음성인식기와 원어민의 응답 일치율

자동음성인식기	일치율
GOOGLE	0.66
MS	0.74
IBM	0.62
AMAZON	0.71
NAVER	0.47
PHONE	0.77



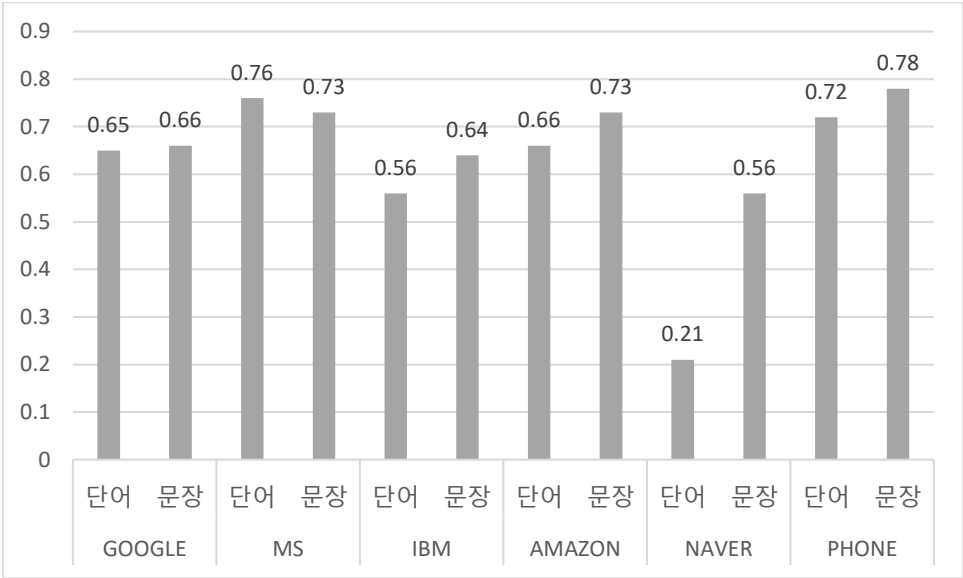
<그림 1> 각 인공지능 자동음성인식기와 원어민의 응답 일치율

입력 형태가 개별 단어 형태인지 문장 형태인지에 따라 원어민과 인공지능 자동음성인식기의 판단의 일치율에 차이가 있는지 확인해본 결과, 6 개 중 4 개의 음성인식기에서 문장 형태가 입력되었을 때 원어민과의 일치율이 더 높은 것으로 나타났다. Google 의 경우, 카이제곱 검정 결과 x^2 (1, N = 8160) = 0.8472, p = 0.4 로 단어와 문장 인식에 있어서 차이가 없는 것으로 나타났다. Microsoft 의 경우, 카이제곱 검정 결과 x^2 (1, N = 8160) = 2.6547, p = .008 로 개별 단어 형태가 입력되었을 때

일치율이 유의미하게 더 높은 것으로 나타났다. IBM 의 경우, 카이제곱 검정 결과 $x^2(1, N = 8160) = 5.9297, p < .001$ 로 문장 형태가 입력되었을 때 일치율이 유의미하게 더 높은 것으로 나타났다. Amazon 의 경우, 카이제곱 검정 결과 $x^2(1, N = 8160) = 6.599, p < .001$ 로 문장 형태가 입력되었을 때 일치율이 유의미하게 더 높은 것으로 나타났다. Naver 의 경우, 카이제곱 검정 결과 $x^2(1, N = 8160) = 27.0688, p < .001$ 로 문장 형태가 입력되었을 때 일치율이 유의미하게 더 높은 것으로 나타났다. 특히 Naver 의 경우, 단어가 입력되었을 때 인식 결과가 없는 경우가 대부분이어서 다른 음성인식기보다도 입력 형태에 따른 차이가 두드러졌다. 음소인식기반 음성인식기의 경우, 카이제곱 검정 결과 $x^2(1, N = 8160) = 5.5971, p < .001$ 로 문장 형태가 입력되었을 때 일치율이 유의미하게 더 높은 것으로 나타났다.

<표 5> 입력 형태에 따른 일치율

자동음성인식기	입력 형태	일치율	*p* 값
GOOGLE	단어	0.65	$p = .4$
	문장	0.66	
MS	단어	0.76	$p = .008$
	문장	0.73	
IBM	단어	0.56	$p < .001$
	문장	0.64	
AMAZON	단어	0.66	$p < .001$
	문장	0.73	
NAVER	단어	0.21	$p < .001$
	문장	0.56	
PHONE	단어	0.72	$p < .001$
	문장	0.78	



<그림 2> 입력 형태에 따른 각 인공지능 자동음성인식기의 일치율

문장 형태를 입력했을 때 목표음의 문장 내 위치에 따라 원어민과 인공지능 자동음성인식기의 판단의 일치율에 차이가 있는지 확인해본 결과, 음소인식기반 자동음성인식기를 제외한 모든 자동음성인식기에서 문장 내 위치에 따른 일치율 차이가 있는 것으로 나타났다. 6 개 중 3 개의 자동음성인식기에서 목표음이 문중에 위치할 때의 일치율이 가장 높게 나타났으며, 또한 6 개 중 4 개의 자동음성인식기에서 문두의 일치율이 가장 낮게 나타났다.

Google 의 경우, 문미 > 문중 > 문두 순서로 일치율이 높게 나타났으며 카이제곱 검정 결과 $x^2(2, N = 5814) = 21.3556$, $p < .001$ 로 차이가 유의미한 것으로 나타났다. Microsoft 의 경우, 문미 > 문두 > 문중 순서로 일치율이 높게 나타났으며 카이제곱 검정 결과 $x^2(2, N = 5814) = 22.6953$, $p < .001$ 로 차이가 유의미한 것으로 나타났다. IBM 의 경우, 문중 > 문미 > 문두 순서로 일치율이 높게 나타났으며 카이제곱 검정 결과 $x^2(2, N = 5814) = 11.4707$, $p = .003$ 로 차이가 유의미한 것으로 나타났다. Amazon 의 경우, 문중 > 문미 > 문두 순서로 일치율이 높게 나타났으며 카이제곱 검정 결과 $x^2(2, N = 5814) = 17.0324$, $p < .001$ 로

차이가 유의미한 것으로 나타났다. Naver 의 경우, 문중 > 문두 > 문미 순서로 일치율이 높게 나타났으며 카이제곱 검정 결과 x^2 (2, N = 5814) = 55.1523, p < .001 로 차이가 유의미한 것으로 나타났다. 음소인식기반 자동음성인식기의 경우, 문미 > 문두 = 문중 순서로 일치율이 높게 나타났으며 카이제곱 검정 결과 x^2 (2, N = 5814) = 0.782, p = 0.68 로 이러한 차이가 유의미하지 않은 것으로 나타났다.

<표 6> 문장 내 위치에 따른 일치율

자동음성인식기	문장 내 위치	일치율	*p* 값
GOOGLE	문두	0.62	p < .001
	문중	0.66	
	문미	0.69	
MS	문두	0.71	p < .001
	문중	0.7	
	문미	0.77	
IBM	문두	0.6	p = .003
	문중	0.65	
	문미	0.64	
AMAZON	문두	0.7	p < .001
	문중	0.76	
	문미	0.73	
NAVER	문두	0.54	p < .001
	문중	0.63	
	문미	0.52	
PHONE	문두	0.75	p = .68
	문중	0.75	
	문미	0.76	

<그림 3> 문장 내 위치에 따른 각 인공지능 자동음성인식기의 일치율

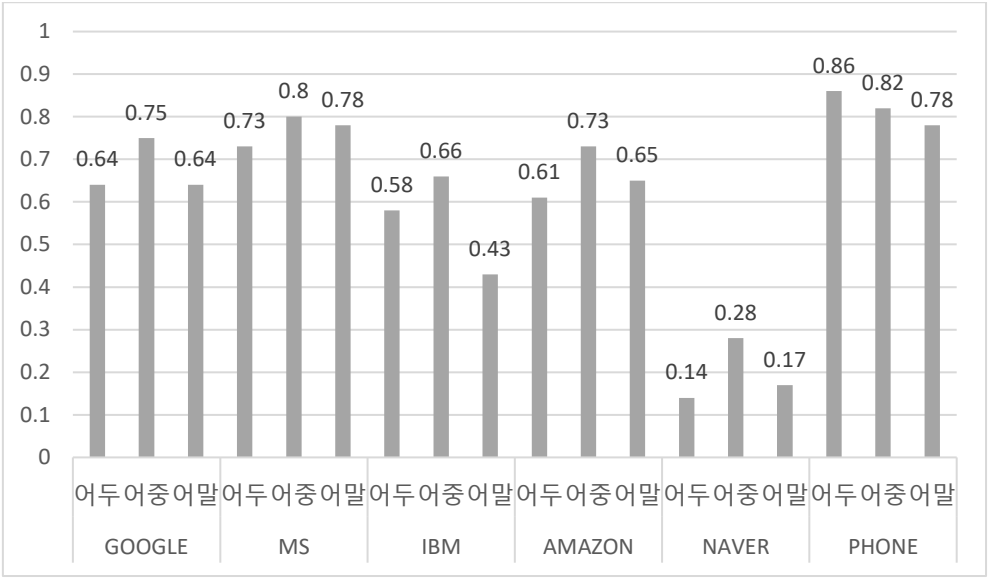
목표음이 자음인 경우에 목표음의 단어 내 위치에 따라 원어민과 인공지능 자동음성인식기의 판단의 일치율에 차이가 있는지 확인해본 결과, 음소인식기반 자동음성인식기를 제외한 모든 자동음성인식기에서 목표음이 어중에 위치했을 때 일치율이 가장 높은 것으로 나타났다. 또한 6 개 중 4 개의 자동음성인식기에서 목표음이 어두에 위치했을 때 일치율이 가장 낮게 나타났다.

Google 의 경우, 어중 > 어말 > 어두 순서로 일치율이 높았으며 카이제곱 검정 결과 x^2 (2, N = 1440) = 19.5169, p < .001 로 차이가 유의미한 것으로 나타났다. Microsoft 의 경우, 어중 > 어말 > 어두 순서로 일치율이 높았으며 카이제곱 검정 결과 x^2 (2, N = 1440) = 8.928, p = .012 로 차이가 유의미한 것으로 나타났다. IBM 의 경우, 어중 > 어두 > 어말 순서로 일치율이 높았으며 카이제곱 검정 결과 x^2 (2, N = 1440) = 43.8441, p < .001 로 차이가 유의미한 것으로 나타났다. Amazon 의 경우, 어중 > 어말 > 어두 순서로 일치율이 높았으며 카이제곱 검정 결과 x^2 (2, N = 1440) = 17.2566, p < .001 로 차이가 유의미한 것으로 나타났다. Naver 의 경우, 어중 > 어말 > 어두 순서로 일치율이 높았으며 카이제곱 검정 결과 x^2 (2, N = 1440) = 39.0077, p < .001 로 차이가 유의미한 것으로 나타났다. 음소인식기반 자동음성인식기의 경우, 어두 > 어중 > 어말 순서로

일치율이 높았으며 카이제곱 검정 결과 $x^2(2, N = 1440) = 10.0111$, $p = .007$로 이러한 차이가 유의미한 것으로 나타났다.

<표 7> 단어 내 위치에 따른 일치율

자동음성인식기	단어 내 위치	일치율	*p* 값
GOOGLE	어두	0.64	$p < .001$
	어중	0.75	
	어말	0.64	
MS	어두	0.73	$p = .012$
	어중	0.8	
	어말	0.78	
IBM	어두	0.58	$p < .001$
	어중	0.66	
	어말	0.43	
AMAZON	어두	0.61	$p < .001$
	어중	0.73	
	어말	0.65	
NAVER	어두	0.14	$p < .001$
	어중	0.28	
	어말	0.17	
PHONE	어두	0.86	$p = .007$
	어중	0.82	
	어말	0.78	

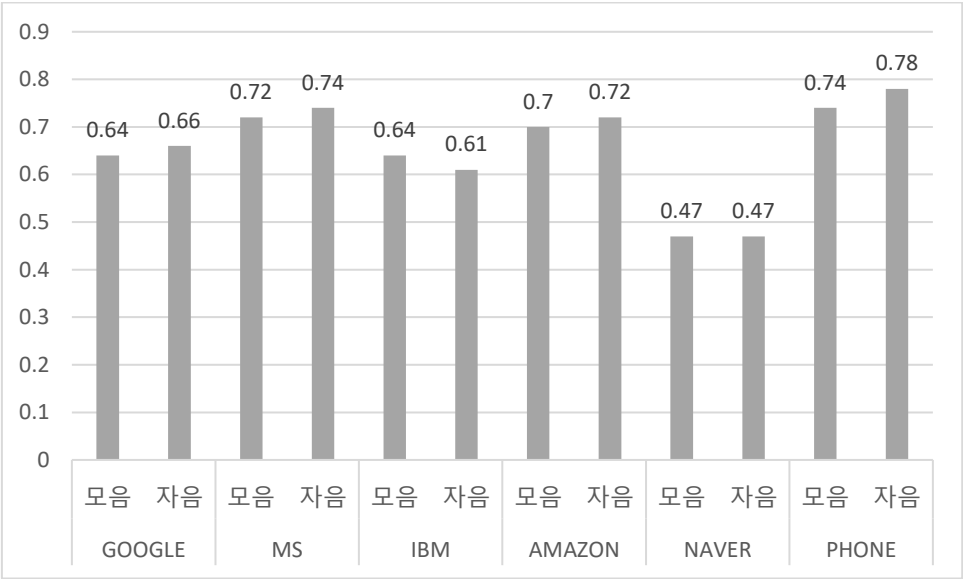


<그림 4> 단어 내 위치에 따른 각 인공지능 자동음성인식기의 일치율

목표음이 자음인지 모음인지에 따라 원어민과 인공지능 자동음성인식기의 판단의 일치율에 차이가 있는지 확인해본 결과, 전반적으로 자음에 대한 일치율이 높지만 자모음에 따른 차이는 미미한 것으로 나타났다. Google 의 경우, 카이제곱 검정 결과 x^2 (1, N = 8160) = 2.2311, p = .026 로 자음을 더 잘 인식하는 것으로 나타났다. Microsoft 역시 마찬가지로, x^2 (1, N = 8160) = 2.4458, p = .014 로 자음을 더 잘 인식하는 것으로 나타났다. 음소인식기반 자동음성인식기 역시 x^2 (1, N = 8160) = 3.5185, p < .001 로 자음을 더 잘 인식하는 것으로 나타났다. 반대로 IBM 의 경우, x^2 (1, N = 8160) = 2.6048, p = .009 로 모음을 더 잘 인식하는 것으로 나타났다. Amazon 은 x^2 (1, N = 8160) = 1.2607, p = .207 로 나타났고, Naver 는 x^2 (1, N = 8160) = 0.1947, p = .846 으로 나타나 Amazon 과 Naver 모두 p 값이 귀무가설을 기각하지 않아 자음과 모음 인식에 있어서 차이가 없는 것으로 나타났다. 따라서 자동음성인식기를 활용함에 있어서 자음과 모음에 대한 인식 정도의 차이는 고려하지 않아도 될 것으로 보인다.

<표 8> 자모음에 따른 일치율

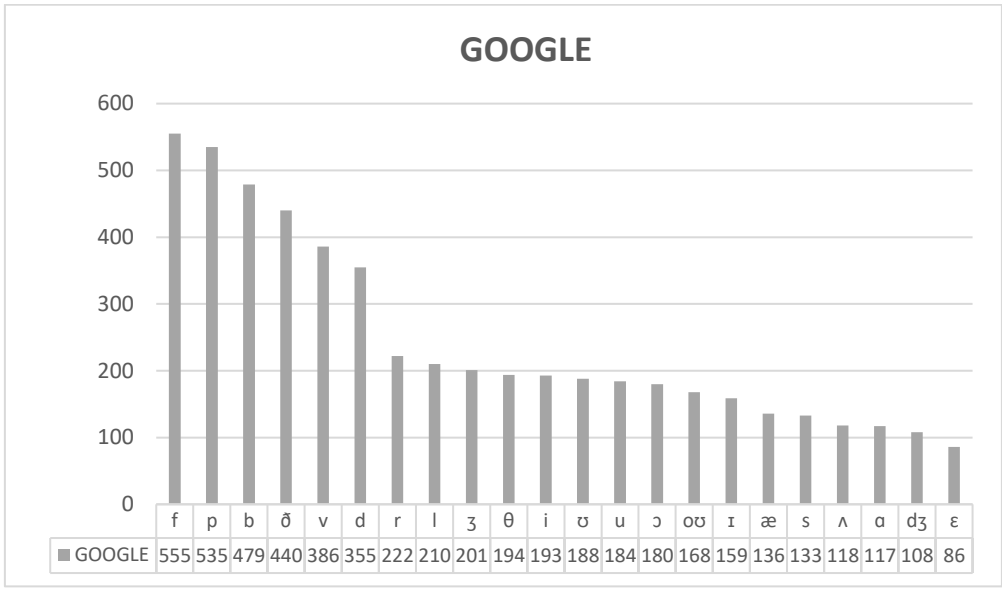
자동음성인식기	자모	일치율	*p* 값
GOOGLE	모음	0.64	$p = .026$
	자음	0.66	
MS	모음	0.72	$p = .014$
	자음	0.74	
IBM	모음	0.64	$p = .009$
	자음	0.61	
AMAZON	모음	0.7	$p = .207$
	자음	0.72	
NAVER	모음	0.47	$p = .846$
	자음	0.47	
PHONE	모음	0.74	$p < .001$
	자음	0.78	



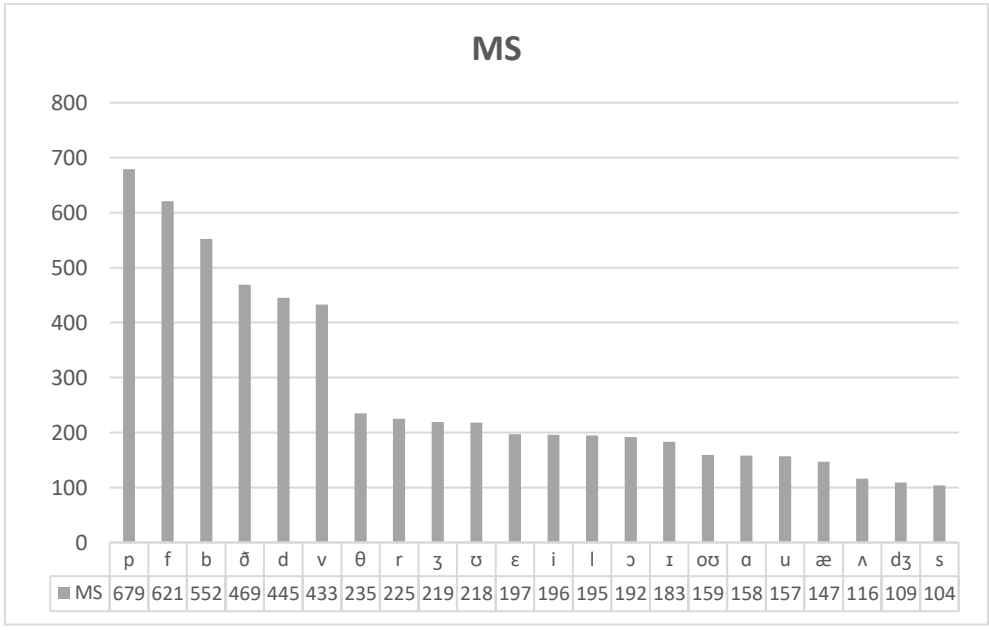
<그림 5> 자모음에 따른 각 인공지능 자동음성인식기의 일치율

각 인공지능 자동음성인식기마다 원어민의 판단과의 일치도가 높은 분절음과 낮은 분절음을 확인해본 결과, 인식기마다 상이한 결과가

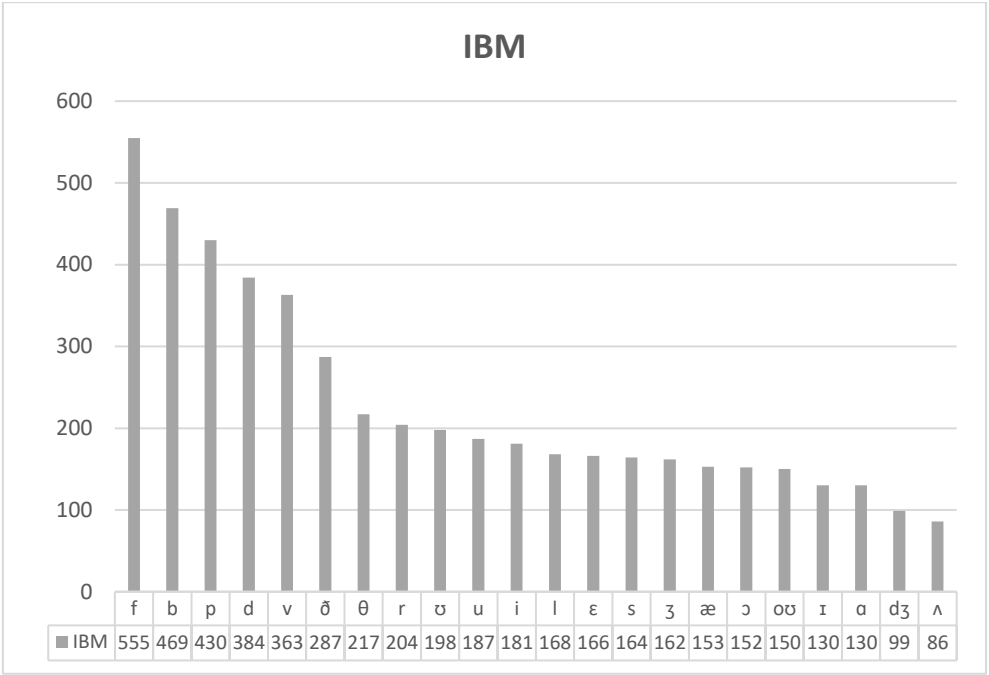
나타났으나, 전반적으로 [b, f, p]의 일치도가 높고 [dʒ, s, ʌ]의 일치도가 낮게 나타났다.



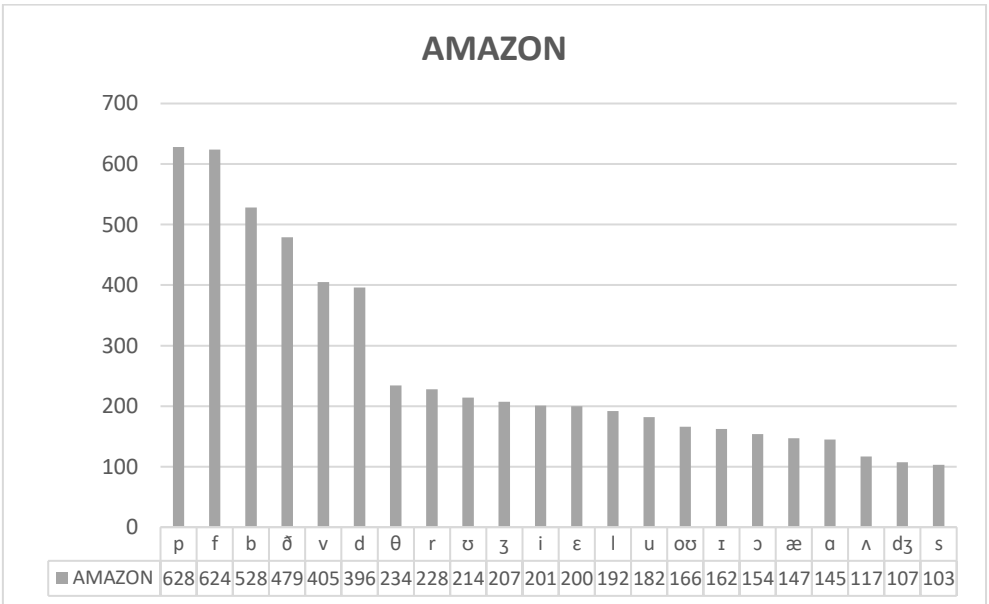
<그림 6> Google 자동음성인식기의 분절음에 따른 원어민과의 일치도



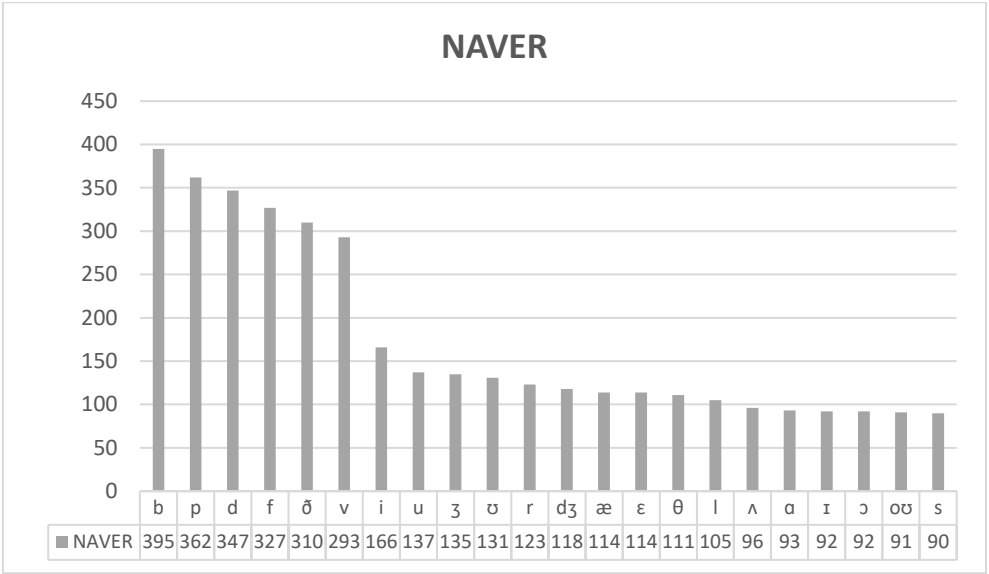
<그림 7> Microsoft 자동음성인식기의 분절음에 따른 원어민과의 일치도



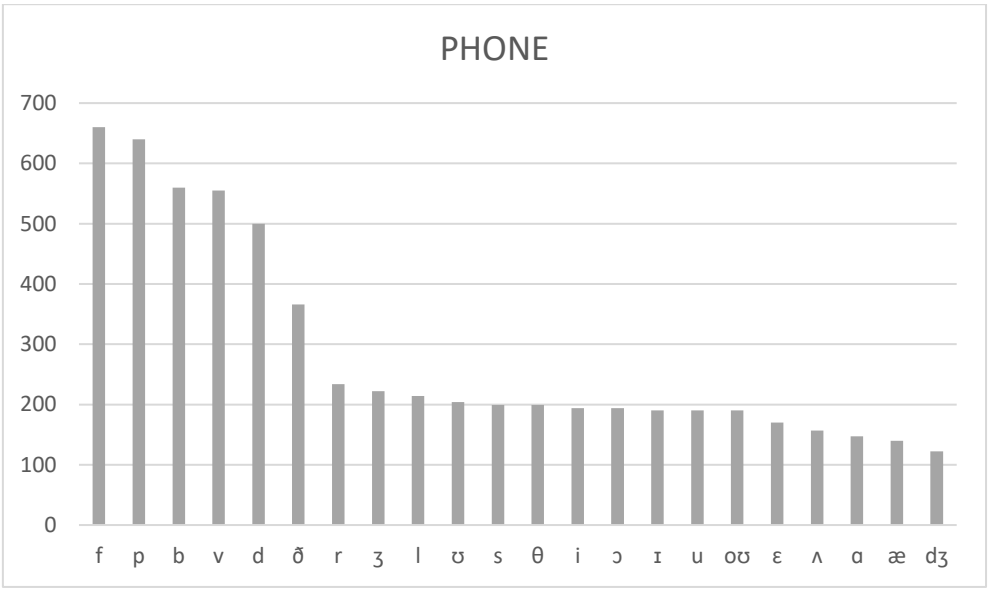
<그림 8> IBM 자동음성인식기의 분절음에 따른 원어민과의 일치도



<그림 9> Amazon 자동음성인식기의 분절음에 따른 원어민과의 일치도



<그림 10> Naver 자동음성인식기의 분절음에 따른 원어민과의 일치도



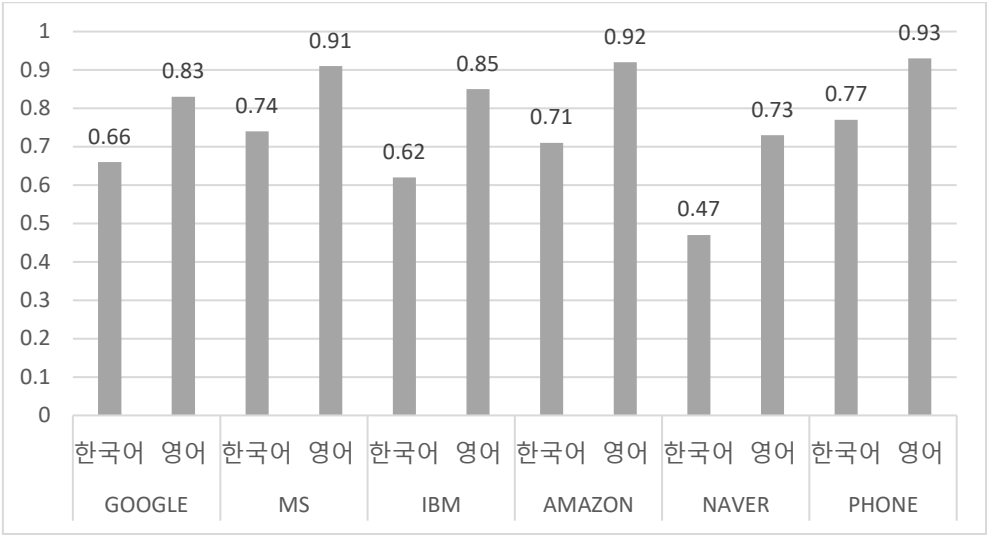
<그림 11> 음소인식기반 자동음성인식기의 분절음에 따른 원어민과의 일치도

모국어에 따라 인공지능 자동음성인식기의 판단의 일치율에 차이가 있는지 확인해본 결과, 모든 자동음성인식기에서 카이제곱검정 결과 $p < .001$ 으로 한국인학습자의 발음보다 원어민의 발음에 대한 일치율이

유의미하게 높게 나타났다. 특히 Microsoft 와 Amazon, 음소인식기반 자동인식기의 경우, 원어민을 대상으로 했을 때 발음 인식률이 90%대로 높게 나타났다. 카이제곱 검정 결과, Google 은 x^2 (1, N = 9384) = 12.438, p < .001 로, Microsoft 는 x^2 (1, N = 9384) = 13.1892, p < .001 로, IBM 은 x^2 (1, N = 9384) = 16.1127, p < .001 로, Amazon 은 x^2 (1, N = 9384) = 15.5102, p < .001 로, Naver 는 x^2 (1, N = 9384) = 16.7766, p < .001 로, 음소인식기반 자동음성인식기는 x^2 (1, N = 9384) = 13.4192, p < .001 로 나타났다.

<표 9> 모국어에 따른 일치율

자동음성인식기	모국어	일치율	*p* 값
GOOGLE	한국어	0.66	p < .001
	영어	0.83	
MS	한국어	0.74	p < .001
	영어	0.91	
IBM	한국어	0.62	p < .001
	영어	0.85	
AMAZON	한국어	0.71	p < .001
	영어	0.92	
NAVER	한국어	0.47	p < .001
	영어	0.73	
PHONE	한국어	0.77	p < .001
	영어	0.93	



<그림 12> 모국어에 따른 각 인공지능 자동음성인식기의 일치율

3. 논의 및 결론

이 연구에서는 발음 산출 훈련에서 인공지능 자동음성인식기가 최소대립어의 발음을 어느 정도로 정확하게 파악하는지 확인하고자, 문맥 독립적 환경에서 한국인 학습자의 영어 발음에 대한 인공지능 자동음성인식기의 성능을 평가하였다. 여섯 가지의 인공지능 자동음성인식기의 인식 결과와 원어민의 인식 결과를 직접적으로 비교하여 그 일치율을 다방면에서 확인해보았다. 인공지능 자동음성인식기가 여러 기업과 연구기관에서 개발되고 있으며 실생활에서 활용되고 있는 현재, 이 연구는 특히 발음 교육에의 쓰임을 목적으로 하여 문맥 독립적으로 인공지능 자동음성인식기의 성능을 면밀히 점검해본 데 의의가 있다.

이 논문에서 인식률이 높은 것으로 확인된 자동음성인식기들은 학습자의 발음을 정확하게 지각하고 교정 피드백을 제공함으로써 발음 교육에 유용하게 쓰일 수 있을 것이다. 특히 음소인식기반 자동음성인식기의 경우에는 발음을 음소 단위로 구분해서 인식하며 인식률도 가장 높게 나왔기 때문에, 음소 단위의 정교한 발음 진단과 교정 피드백 제공에 유용할 것으로 보인다. 더군다나 음소인식기반 자동음성인식기는 훈련 데이터 등을 조정함으로써 최소대립어 인식에 더 적합하게 튜닝할 수 있으므로 더 높은 성능을 기대할 수 있다. 따라서 음성인식 시스템 개발자가 있는 기업이나 학교에서는 발음 산출 교육 프로그램 제작을 위해 음소인식기반 자동음성인식기를 활용하는 것이 적합할 것이다.

단어인식기반 자동음성인식기 중 인식률이 가장 높게 나온 Microsoft Azure Speech Service 와 그 다음으로 높게 나온 Amazon Transcribe 의 경우에는 음소인식기반 자동음성인식기에 비해 인식률은 떨어지지만, 그 차이가 미미한 정도라는 점이 주목할 만하다. 최소대립어를 활용하여 발음 산출 교육을 진행하는 교육 현장에서는 단어 단위에서의 산출

훈련이 이루어지기 때문에 단어인식기반 자동음성인식기를 사용하는 것이 효율적일 것으로 보인다. 단어 단위에서 음소인식기반 자동인식기와 단어인식기반 자동음성인식기의 인식률 차이가 크지 않으므로, 음소인식기반 자동음성인식기를 직접 개발할 필요 없이 단어인식기반 자동음성인식기 API 를 활용하여 산출 훈련 프로그램을 개발하는 것이 가능하다. 학습 대상이 되는 목표 음소별로 최소대립어가 인식되었을 때 어떤 교정 피드백을 제공할지 미리 프로그래밍해 둔다면, 온라인 환경에서도 충분히 학습자에게 교정 피드백을 제공하는 산출 훈련을 진행할 수 있을 것이다. 예를 들어, 목표 단어가 “pan”인데 학습자의 발화에서 “pen”이 인식되었다면 “pen 보다 입을 더 벌리고 발음하세요.”라는 교정 피드백이 가능할 것이다. 물론 최소대립어가 아닌 단어의 경우에는 음소 단위별로 맞고 틀림을 파악하는 과업에는 활용할 수 없는 한계가 있겠으나, 단어인식기반 자동음성인식기의 경우 공개된 API 로 서비스되고 있기 때문에 손쉽게 단어 수준에서의 발음 연습 프로그램을 만들어서 과제나 수업 중 학습 활동 등으로 활용할 수 있을 것이다. 공개된 API 를 활용하여 발음 교육 프로그램을 개발할 필요성이 있는 상황이라면 음소인식기반 음성인식기의 좋은 대안이 될 것으로 판단된다.

이 논문에서 점검한 자동음성인식기보다 인식률이 낮은 자동음성인식기를 활용하여도 발음 훈련의 효과가 입증되었다는 점에서, 인식률이 더 높은 자동음성인식기를 활용할 경우 교육 효과가 있을 것으로 예상할 수 있다. Guskaroska(2019)에서 비원어민 화자의 발화에 대한 인식률이 57%인 자동음성인식기를 활용하여 발음 훈련을 했을 때 실험군의 발음이 향상되는 효과가 있었던 점으로 보았을 때, 비원어민 화자의 발화에 대한 인식률이 약 80%에 도달하는 음소인식기반 자동음성인식기나 약 70%에 달하는 단어인식기반 자동음성인식기의 경우, 발음의 맞고 그름을 판가름하여 발음 산출 교육에 유의미한 도움을 줄 가능성이 있다.

자동음성인식기를 활용한 발음 산출 훈련을 설계할 때, 언어학적 단위별로 달라지는 성능을 고려하면 오인식을 최대한 피할 수 있을 것으로 보인다. 언어학적 단위별로 나누어 인식률을 살펴본 결과, 입력 형태와 목표음의 단어 및 문장 내 위치에 따라 인식률에 차이가 남을 확인할 수 있었다. 음성인식기마다 결과가 상이하게 나타났으나, 전반적으로 문장 형태가 입력되었을 때 원어민 인식과의 일치율이 높아진 것으로 나타났다. 문맥 정보가 주어지지 않았는데도 불구하고 문장 형태의 일치율이 높게 나타난 것은 주목할 만하다. 이는 각 자동음성인식기가 애초에 문장을 인식하는 것을 목표로 개발되었기 때문으로 볼 수 있으며, 목표음 이외의 주변음들이 목표음의 인식 시작점을 잡기 용이하게 하여 인식 실패를 방지했기 때문으로 보인다. 따라서 자동음성인식기를 활용한 발음 산출 교육 프로그램을 개발할 때에는 문장 형태를 입력 형태로 활용하는 것이 인식률을 높이는 방안이 될 수 있을 것이다.

목표음의 위치가 문중 및 어중일 때의 일치율이 높고, 문두 및 어두일 때의 일치율이 상대적으로 낮게 나타난 것은 자동음성인식기가 음성을 인식하는 구조에 의한 것으로 보인다. 음성 특징 파라미터를 추출할 때 음성을 고정된 길이로 나누는 프레이밍을 거치는데, 처음이나 끝에 위치한 음보다 중간에 위치한 음의 경우 이 프레임이 중첩되는 부분이 많아져서 정확도가 높아진다. 이로써 목표음의 앞뒤에 다른 음이 있을 때 목표음의 음향 특징이 대비되어 더욱 잘 처리되는 것으로 보인다. 또한 목표음의 위치에 따라 음성적인 특성이 달라지기 때문에 일치율이 다르게 나타난 것으로 볼 수도 있다. 영어에서 유성자음의 경우 어두 및 문두에서 종종 무성으로 발음되기도 하는 것처럼(Kingston & Diehl, 1994) 어중에 있을 때 본래의 음성적 특징이 더욱 잘 드러나 인식률이 높아진 것으로 볼 수 있다. 따라서 인식률을 높이려면 목표음의 앞뒤로 다른 음을 배치함으로써 일종의 완충지대를 형성하는 것이 필요하다고 해석할 수 있다.

따라서 인공지능 자동음성인식기를 활용한 발음 산출 교육 프로그램을 제작할 때, 문장을 기본적인 입력 형태로 활용하며 목표

발음을 문중 및 어중에 위치시킬 경우 인식 오류가 적어져서 학습자와 교육자가 인식 오류로 인한 어려움을 피할 수 있을 것으로 예측할 수 있다. 예를 들어, 'Say ____ again' 같은 문장틀에 최소대립어 단어를 넣어서 발음 산출 훈련을 설계하면 인식 오류를 최소화하면서 자동음성인식기를 활용할 수 있을 것이다. 이러한 방식으로 인공지능 자동음성인식기를 활용하면 발음 교육에 있어서 오인식을 최대한 피하여 학습자의 의욕을 떨어트리지 않고 교육 프로그램을 설계할 수 있을 것이다.

또한 연구 결과 인식률이 높은 분절음과 낮은 분절음을 확인할 수 있었다. 인식률이 높게 나온 [b, f, p]의 경우 모두 순음이라는 공통점이 있다. 자동음성인식기가 순음의 음향 특징을 다른 분절음의 음향 특징보다 잘 구별되게 학습했기 때문으로 보인다. 순음의 특징인 F2 와 F3 의 급격한 경사가 충분한 음향적 단서가 된 것으로 보인다. 인식률이 낮게 나온 [dʒ]의 경우, 단어 'pledger'가 쌍이 되는 'pleasure'에 비해 음성인식기의 언어 모델 코퍼스에서 빈도수가 압도적으로 낮기 때문에 인식이 제대로 되지 않은 것으로 보인다. 한국인 학습자가 [dʒ]를 제대로 발음한 경우에도 음성인식기가 'pleasure'를 발음한 것으로 인식한 경우가 대부분이었다. 또한 /ʒ-dʒ/의 대립은 기능부담량이 낮기 때문에(Brown, 1988) 자동음성인식기가 서로 구별을 잘 하지 못하는 것은 기능부담량이 높은 음을 중심으로 하는 영어 교육에 있어서 그다지 문제가 되지 않을 것으로 보인다. [s]의 경우, 영어 자음 중에서 가장 인식률이 떨어진다는 선행연구와 동일한 결과를 보인다. 자동음성인식 시스템이 인식 대상의 소음 구간을 처리할 때 마찰음 [s]의 소음을 처리함에 있어서 오류가 발생하는 것으로 보인다. [s]의 경우, 음향적으로 공명도(sonority)가 다른 자음에 비해 상대적으로 낮고 마찰 소음이 4000~8000Hz 사이에 절단주파수(cutoff frequency)가 나타나 이 부분이 인식기에 오류를 불러일으키는 것으로 보인다(양병곤, 2017). [ʌ]의 경우에는 'cup-cop'쌍에서 'cup'을 'cop'으로 인식하는 경우가 많았다. 이는 남성의 경우에 [ɑ]는 F1 이 768Hz, F2 가 1333Hz 정도이고 [ʌ]는 F1 이 623Hz, F2 가

1200Hz 정도로 두 음소의 음향 거리가 가깝기 때문으로 보인다(Hillenbrand et al., 1995).

이처럼 인식률이 낮게 나온 분절음의 경우에는 발음 산출 교육 프로그램을 제작하는 데 있어서 주의가 필요할 것으로 보인다. 자음 교육을 상정하였을 때, Microsoft Azure Speech Service 에서 인식률이 낮게 나온 자음 [dʒ, s]를 제외하고 자음만의 인식률을 살펴본 결과, 0.8 로 기존의 0.74 보다 인식률이 상승하였다. 발음 산출 교육에서 그릇된 교정 피드백이 제공되면 학습자의 신뢰가 떨어질 가능성이 높기 때문에 이처럼 인식률이 낮게 나온 분절음은 피하거나 교정 피드백을 제공하지 않는 것이 필요할 것이다. 자음 [dʒ, s]의 경우에는 한국인 학습자가 산출하는 데 어려운 소리가 아니므로(Lee & Hwang, 2016) 이 두 소리에 대한 교정 피드백을 제공하지 않아도 큰 문제가 되지 않을 것으로 판단된다. 자동음성인식기의 인식 오류를 피하기 위해서는 이처럼 인식률이 낮게 나온 분절음을 피하거나 최소한으로 활용하는 것이 필요할 것이다.

자동음성인식기의 인식률이 100%가 아니라는 점에서 이를 활용하는 것에 의구심을 가질 수 있다. 그러나 이 연구에서도 볼 수 있듯이 원어민의 평가도 항상 100% 서로 일치하는 것이 아니기 때문에 인식 오류라는 단점보다 장점에 더욱 주목하여 이를 활용할 방안을 찾는 것이 더 생산적인 방향일 것이다. 선행연구에서 자동음성인식기를 활용한 학습을 진행하였을 때, 학생들이 소프트웨어를 활용하여 학습하는 경험에 대해 긍정적으로 평가하였으며 외국어를 학습하는 동기부여가 되었고, 학생들의 흥미도와 참여도가 높아졌으며, 외국어를 사용하는 자신감이 향상되었다는 점이 다수 보고되어 왔다(임창근 & 신혜정, 2001; Golonka et al., 2014; 김효진, 2018; Guskaroska, 2019). 외국어 학습에는 학습자가 느끼는 염려나 불안감이 실제 학습에 부정적인 영향을 끼치지만, 자동음성인식 소프트웨어를 활용하면 부담감을 주는 환경으로부터 자유롭게 독립적으로 학습을 수행할 수 있어 긍정적 동기를 제공할 수 있다(옥종석, 2004). 특히 자동음성인식 소프트웨어가

발음에 대한 피드백을 신속하게 제공하는 점이 학습자에게 매우 긍정적으로 작용하였으며, 이는 단순히 발음 훈련에 그치는 것이 아니라 외국어 학습 전체에 좋은 영향을 미쳤다(이경숙, 2018). 이처럼 정서적인 측면과 더불어 훈련 효과에 대한 실증적인 증거들(임창근 & 신혜정, 2001; 이경숙, 2018; Guskaroska, 2019; Xie et al., 2020)을 토대로 볼 때, 자동음성인식기는 외국어 학습 분야에서 가치 있는 요소로 자리매김할 것으로 보인다.

이 연구의 한계는 다음과 같다. 우선, 인공지능 자동음성인식기의 성능을 평가하는 데 기준으로 삼았던 원어민의 평가가 서로 일치하지 않을 수 있다. 원어민끼리의 응답 일치율을 Fleiss's Kappa 상관분석으로 신뢰도를 검증해본 결과, Fleiss's Kappa 상관계수가 0.58 로 신뢰도는 보통 정도로 나타났다. 영어는 사용 지역에 따라 모음의 차이가 크고 자음의 차이는 그다지 크지 않으므로, 영어 원어민은 자음의 오류보다 모음의 오류에 대해 용인도가 높을 수 있다. 또한 외국어 학습자의 발음에 대해 장시간 듣고 평가를 내렸기 때문에 원어민들의 응답 일치율이 보통 정도로만 나타났다. 하지만 이는 원어민의 응답에도 오류가 있을 수 있다는 점에서 오히려 음성인식기의 오류에 대한 우려를 덜게 해주는 부분이기도 하다. 그리고 이 연구에서는 발음 오류가 흔히 일어날 만한 분절음만을 대상으로 하였는데, 영어의 전체 분절음을 대상으로 하면 더욱 흥미로운 결과를 확인할 수 있을 것이다. 이 연구에서는 인공지능 자동음성인식기가 발음 오류를 오류로 제대로 파악하는지 살펴보고자 한국인 학습자가 발음오류를 산출할 것으로 예상되는 분절음만을 대상으로 하였으나, 전체 분절음을 대상으로 한다면 더욱 다양한 결과를 확인할 수 있을 것이다. 또한 이 연구에서 인식률이 높게 나타난 인공지능 자동음성인식기를 활용한 프로그램을 제작하여 실제 외국어 학습자를 대상으로 발음 교육 실험을 하는 후속 연구도 가능할 것이다.

참고 문헌

김효진. (2018). 구글번역기를 사용한 학습이 초등학생의 영어 말하기 능력에 미치는 영향. 부산대학교 외국어교육학과 영어교육학전공 석사학위논문.

노희경, & 이강희. (2017). 구글, 네이버, 다음 카카오 API 활용앱의 표준어 및 방언 음성인식 기초 성능평가. 예술인문사회 융합 멀티미디어 논문지, 38, 819-829.

류혁수, & 정민화. (2016a). 조음 기반의 음소 레벨 사후 확률을 이용한 한국인 영어 학습자의 자음 발음 오류 검출. 한국음성학회 가을 학술대회 발표 논문집, 85-86.

류혁수, & 정민화. (2016b). 조음자질을 이용한 한국인 학습자의 영어 발화 자동 발음 평가. 말소리와 음성과학, 8(4), 103-113.

리졘화. (2019). (The) perception and production of English vowels by Chinese and Korean EFL learners. 원광대학교 영어영문학과 박사학위논문.

박향숙, 이예식, & 윤정희. (2018). 구글 자동음성인식기를 활용한 한국 초등학생들의 영어 단어 발음인식감도 조사. 현대문법연구, 97, 179-201.

안소연. (2019). 한국인 초등 영어학습자의 영어 모음발화와 인지 연구. 부산대학교 영어영문학과 석사학위논문.

양병곤. (2013). 한국인과 미국인이 발화한 영어전설모음의 상대적 거리 비교. 말소리와 음성과학, 5(4), 99-107.

양병곤. (2017). 대학생들이 또렷한 음성과 대화체로 발화한 영어문단의 구글음성인식. 말소리와 음성과학, 9(4), 43-50.

옥종석. (2004). 초등영어교육의 음성인식 소프트웨어 활용과 전망. STEM Journal, 5(2), 85-102.

유현재, 김명화, 박상길, & 김광용. (2020). 클라우드 기반의 음성인식 오픈 API 의 응용 분야별 한국어 연속음성인식 정확도 비교 분석. 한국통신학회논문지, 45(10), 1793-1803.

유혜배, & 윤한나. (2009). 한국인 성인 영어학습자의 대화상에 나타난 분절음 발음오류 연구. 인문학연구, 12, 185-212.

윤정희(2014). Google 음성인식프로그램에 의한 한국 어린이 영어학습자의 영어단어 발음인식 실태분석: 영어학습도우미로봇개발을 목적으로. 경북대학교 교육대학원 영어교육전공 석사학위논문.

이건상, 양성일, & 권영헌. (2001). 음성인식. 한양대학교 출판부.

이경숙. (2018). 음성인식 어플리케이션을 활용한 일본어 음성교육 방안 연구. 한국일본어교육학회, (85), 29-42.

임창근, & 신혜정. (2001). 컴퓨터 음성인식(ASR) 기술을 활용한 영어 말하기 교육. Multimedia-Assisted Language Learning, 4(2), 187.

최승주, & 김종배. (2017). 음성 인식 오픈 API 의 음성 인식 정확도 비교 분석. 예술인문사회융합멀티미디어논문지, 34, 411-418.

Ashwell, T., & Elam, J. R. (2017). How Accurately Can the Google Web Speech API Recognize and Transcribe Japanese L2 English Learners' Oral Production. Jalt Call Journal, 13(1), 59-76.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449-12460.

Brown, A. (1988), Functional load and the teaching of pronunciation. Tesol Quarterly, 22(4), 593-606.

Carrier, Michael. (2017). Automated speech recognition in language learning: potential models, benefits and impact. Training, Language and Culture, 1(1).

Cooper, W. E., & Blumstein, S. E. (1974). A “labial” feature analyzer in speech perception. *Perception & Psychophysics, 15*(3), 591-600.

Escudero-Mancebo, D., Cámara Arenas, E., Tejedor García, C., González Ferreras, C., & Cardeñoso Payo, V. (2015). Implementation and test of a serious game based on minimal pairs for pronunciation training. ISCA Workshop on Speech and Language Technology in Education, 125-130.

Evers, K., & Chen, S. (2020). Effects of an automatic speech recognition system with peer feedback on pronunciation instruction for adults. Computer Assisted Language Learning, 1-21.

Golestani, N. & Pallier, C. (2007). Anatomical Correlates of Foreign Speech Sound Production, *Cerebral Cortex, 17*(4), 929–934

Golonka, E., Bowles, Anita R., Frank, Victor M., Richardson, Dorna L., & Freynik, Suzanne. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. Computer Assisted Language Learning, 27(1), 70-105.

Guskaroska, A. (2019). ASR as a tool for providing feedback for vowel pronunciation practice (Doctoral dissertation, Iowa State University).

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech communication*, *47*(3), 360-378.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical society of America*, *97*(5), 3099-3111.

Hirata, Y. (2004). Computer assisted pronunciation training for native English speakers learning Japanese pitch and durational contrasts, *Computer Assisted Language Learning 17*(3-4), 357-376

Hirschi, K., Kang, O., Cucchiarini, C., Hansen, J., Evanini, K., & Strik, H. (2020). Mobile-Assisted Prosody Training for Limited English Proficiency. Learner Background and Speech Learning Pattern.

Huensch, A., & Tremblay, A. (2015). Effects of perceptual phonetic training on the perception and production of second language syllable structure. *Journal of Phonetics*, *52*, 105-120.

IANCU, Bogdan. (2019). Evaluating Google Speech-to-Text API's Performance for Romanian e-Learning Resources. Informatica Economica, 23(1/2019), 17-25.

Jiang, S. W. F., Yan, B. C., Lo, T. H., Chao, F. A., & Chen, B. (2021). Towards Robust Mispronunciation Detection and Diagnosis for L2 English Learners with Accent-Modulating Methods. arXiv preprint arXiv:2108.11627.

Juang, B. H., & Rabiner, Lawerence R. (2005). Automatic Speech Recognition – A Brief History of the Technology Development. Georgia Institute of

Technology. Atlanta Rutgers University and the University of California, Santa Barbara.

Jurafsky, D., & Martin, J. (2009). Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Upper Saddle River, NJ, USA: Prentice-Hall.

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. The journal of the acoustical society of America, 138(2), 817-832.

Këpuska, Veton, & Bohouta, Gamal. (2017). Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx). International Journal of Engineering Research and Applications, 7(3), 20-24.

Kim, In-Seok. (2006). Automatic Speech Recognition: Reliability and Pedagogical Implications for Teaching Pronunciation. Educational Technology & Society, 9(1), 322-334.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, *70*(3), 419-454.

Kodish-Wachs, Jodi, Agassi, Emin, Kenny, 3rd, Patrick, & Overhage, J Marc. (2018). A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. AMIA ... Annual Symposium Proceedings, 2018, 683-689.

Korzekwa, D., Lorenzo-Trueba, J., Zaporowski, S., Calamaro, S., Drugman, T., & Kostek, B. (2021, June). Mispronunciation detection in non-native (L2) English with uncertainty modeling. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7738-7742). IEEE.

Lee, Ho-Young, & Hwang, Hyosung. (2016). Gradient of learnability in teaching English pronunciation to Korean learners. The Journal of the Acoustical Society of America, 139(4), 1859-1872.

Lee, K. F., & Hon, H. W. (1989). Speaker-independent phone recognition using hidden Markov models. IEEE Transactions on Acoustics, Speech, and Signal Processing, 37(11), 1641-1648.

Li, W., Li, K., Siniscalchi, S. M., Chen, N. F., & Lee, C.-H. (2016). Detecting Mispronunciations of L2 Learners and Providing Corrective Feedback Using Knowledge-guided and Data-driven Decision Trees. Proceedings of INTERSPEECH 2016 (pp. 3127-3131). San Francisco, CA. 8-12 September, 2016.

Lopez-Soto, T., & Kewley-Port, D. (2009). Relation of perception training to production of codas in English as a second language. In Proceedings of Meetings on Acoustics (Vol. 6, 062003). The Acoustical Society of America.

Massaro, D. W., Bigler, S., Chen, T. H., Perlman, M., and Ouni, S. (2008). Pronunciation training: The role of eye and ear, In *The Proceedings of Interspeech 9*, 2623–2626.

McCrocklin, Shannon M. (2016). Pronunciation learner autonomy: The potential of Automatic Speech Recognition. System (Linköping), 57, 25-42.

Munro, M. J., & Derwing, T. M. (2006), The functional load principle in ESL pronunciation instruction: An exploratory study. System, 34(4), 520-531.

Peperkamp, S., & Bouchon, C. (2011). The relation between perception and production in L2 phonological processing. In *The Proceedings of INTERSPEECH 2011*, 161-164.

Saito, K., & Lyster, R. (2012). “Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /ɹ/ by Japanese Learners of English,” Lang. Learn. 62(2), 595–633.

Tepperman, J., & Narayanan, S. (2008). Using Articulatory Representations to Detect Segmental Errors in Nonnative Pronunciation. IEEE Transactions on Audio, Speech, and Language Processing, 16(1), 8-22.

Wu, M., Li, K., Leung, W. K., & Meng, H. (2021). Transformer Based End-to-End Mispronunciation Detection and Diagnosis. *Proc. Interspeech 2021*, 3954-3958.

Xie, Y., Feng, X., Li, B., Zhang, J., & Jin, Y. (2020). A Mandarin L2 Learning APP with Mispronunciation Detection and Feedback. In INTERSPEECH (pp. 1015-1016).

Ye, W., Mao, S., Soong, F., Wu, W., Xia, Y., Tien, J., & Wu, Z. (2022, May). An Approach to Mispronunciation Detection and Diagnosis with Acoustic, Phonetic and Linguistic (APL) Embeddings. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6827-6831). IEEE.

부 록

<부록 1: 녹음 스크립트>

1. heat
2. hit
3. pan
4. pen
5. Luke
6. look
7. pause
8. pose
9. cop
10. cup
11. pat
12. fat
13. coffee
14. copy
15. beef
16. beep
17. base
18. vase
19. rebel
20. revel
21. curb
22. curve
23. day
24. they
25. header
26. heather
27. breed
28. breathe
29. lay
30. ray

31. sink
32. think
33. pleasure
34. pledger

35. heat she said
36. hit she said
37. pan she said
38. pen she said
39. Luke she said
40. look she said
41. pause she said
42. pose she said
43. cop she said
44. cup she said
45. pat she said
46. fat she said
47. coffee she said
48. copy she said
49. beef she said
50. beep she said
51. base she said
52. vase she said
53. rebel she said
54. revel she said
55. curb she said
56. curve she said
57. day she said
58. they she said
59. header she said
60. heather she said
61. breed she said
62. breathe she said

63. lay she said
64. ray she said
65. sink she said
66. think she said
67. pleasure she said
68. pledger she said

69. Say heat again
70. Say hit again
71. Say pan again
72. Say pen again
73. Say Luke again
74. Say look again
75. Say pause again
76. Say pose again
77. Say cop again
78. Say cup again
79. Say pat again
80. Say fat again
81. Say coffee again
82. Say copy again
83. Say beef again
84. Say beep again
85. Say base again
86. Say vase again
87. Say rebel again
88. Say revel again
89. Say curb again
90. Say curve again
91. Say day again
92. Say they again
93. Say header again
94. Say heather again

95. Say breed again
96. Say breathe again
97. Say lay again
98. Say ray again
99. Say sink again
100. Say think again
101. Say pleasure again
102. Say pledger again

103. I say heat
104. I say hit
105. I say pan
106. I say pen
107. I say Luke
108. I say look
109. I say pause
110. I say pose
111. I say cop
112. I say cup
113. I say pat
114. I say fat
115. I say coffee
116. I say copy
117. I say beef
118. I say beep
119. I say base
120. I say vase
121. I say rebel
122. I say revel
123. I say curb
124. I say curve
125. I say day
126. I say they

127. I say header
128. I say heather
129. I say breed
130. I say breathe
131. I say lay
132. I say ray
133. I say sink
134. I say think
135. I say pleasure
136. I say pledger

Abstract

Evaluation of conformity of AI ASR on English production training for Korean learners

Yeojin Lee
Department of Linguistics
The Graduate School
Seoul National University

The aim of this paper is to verify whether an Artificial Intelligence Automatic Speech Recognition(AI ASR) can be used when developing a production training program, and to discuss how to use AI ASR by evalutating the recognition performance of AI ASR on the pronunciation spoken by Korean learners of English in a context-independent environment.

To evaluate pronunciation independently from context information, individual word forms and carrier-sentences were used as experimental data. And to find out whether performance varies depending on the position of the target sound, the recognition rate was identified in sentence-initial, sentence-medial, sentence-final position, and word-initial, word-medial, and word-final position. In evaluating the performance of AI ASR, a method of directly comparing the native speaker's response with the response of the ASR was used to distinguish between Korean learners' pronunciation errors and misrecognition of the ASR. For the stimuli to understand how accurately the ASR systems distinguish sounds, we used minimal pairs that Korean learners of English find difficult to pronounce. The six AI ASR systems were compared with each other: Google Cloud Speech-to-Text, Microsoft

Azure Speech Service, IBM Watson Speech to Text, Amazon Transcribe, Naver CLOVA Speech, and phone-based ASR.

As a result of directly comparing the judgment of each AI ASR and native speakers on the English pronunciation of Korean learners in a context-independent environment, the agreement rate of phone-based ASR was the highest at 77%. When the input form was a sentence, the overall recognition rate was high, and when the target sound was at the sentence-medial or word-medial position, the recognition rate was high. The difference in the recognition rate between consonants and vowels was insignificant. In addition, the overall recognition rate of segments [b, f, p] was high and the recognition rate of [dʒ, s, ʌ] was low.

This study evaluated the context-independent performance of multiple AI ASR systems from the perspective of production training. While a phone-based ASR has the best performance for developing a production training, it was found that word-based ASR systems also showed quite high performance for developing production training using minimal pairs. Since the ASR determines the mispronunciation with a recognition rate of about 80%, a positive effect could be expected when used for pronunciation education and evaluation. When developing a production training and evaluation program, it was confirmed how to compose linguistic units to avoid misrecognition of ASR.

Keywords : ASR, performance evaluation, Korean learners of English, English pronunciation, recognition rate, AI, production training, pronunciation education

Student Number : 2015-22455