COMMUNICATION SCIENCES & DISORDERS

Original Article

| ISSN 2288-0917 (Online) | Commun Sci Disord 2023;28(1):30-38

https://doi.org/10.12963/csd.23952

Comparison of Automatic Speech Recognition System for School-aged Children's Narratives: Naver Clova Speech and Google Speech-to-Text

Hui Jae Yang[a], Eun-Byel Oh[a], Jung-Mee Kim[b]

[a]*Department of Communication Disorders, The Graduate School of Korea Nazarene University, Cheonan, Korea*
[b]*Department of Communication Disorders, Korea Nazarene University, Cheonan, Korea*

Correspondence: Eun-Byel Oh, MA
Department of Communication Disorders, Korea Nazarene University, 48 Wolbong-ro, Seobuk-gu, Cheonan 31172, Korea
Tel: +82-41-570-1411
Fax: +82-41-570-7846
E-mail: oeb8312@naver.com

Jung-Mee Kim, PhD
Department of Communication Disorders, Korea Nazarene University, 48 Wolbong-ro, Seobuk-gu, Cheonan 31172, Korea
Tel: +82-41-570-1411
Fax: +82-41-570-7846
E-mail: jmkim@kornu.ac.kr

Received: January 20, 2023
Revised: February 24, 2023
Accepted: February 24, 2023

Objectives: Language sample analysis (LSA) is a critical component of child language assessment. However, most clinicians consider LSA to be time consuming work. In particular, transcription is seen as an overwhelming task. Due to rapid technological advances, various automatic speech recognition systems have been developed. This study aimed to investigate the accuracy and the characteristics of two automatic speech recognition programs, Naver Clova Speech (Naver Clova) and Google Speech-to-Text (STT). **Methods:** A total of 40 school-aged children with typical development (TD) and children with language learning disabilities (LLD) participated in the study. Each child was asked to generate two fictional narratives. In total, 72 narratives produced by 36 children were used. To examine the accuracy of Naver Clova and Google STT, syllable error rate was analyzed and compared to reference transcripts. For the detailed analysis, types of error such as substitution, deletion and insertion were examined. **Results:** Results showed that Naver Clova was significantly lower than Google STT in error rate of transcription. But the transcription error rate of the two child groups was not significantly different. Additionally, the Naver Clova error rate was higher in substitution, deletion, and insertion respectively. The Google STT error rate, on the other hand, was higher in deletion, substitution and insertion respectively. **Conclusion:** Naver Clova were more accurate than Google STT in transcribing children's narratives. But the transcription accuracy of two child groups was not different. This suggests that recently developed automatic speech recognition systems have clinical utility. These systems can reduce clinician's workload in regards to LSA and this would contribute to qualitatively enhanced language assessment.

Keywords: Transcription, Automatic speech recognition system, Syllable error rate

임상가가 아동의 언어 능력을 평가하기 위해서는 두 가지 방법을 사용할 수 있다. 하나는 표준화된 언어 검사를 사용하는 것이고 다른 하나는 아동의 발화를 수집하여 분석하는 것이다. 표준화된 언어 검사는 또래 아동과의 비교를 통해 언어 문제의 유무와 정도 그리고 전반적인 변화를 평가하는 데에 유용하다. 그러나 표준화 검사는 중재 목표를 세우거나 중재 후 진전을 세밀하게 모니터하기에는 어려움이 있으며 아동이 일상생활에서 사용하는 언어에 대한 자세한 정보를 얻기 어렵다(Kim, 2014; Westerveld & Classen, 2014). 반면에 발화 분석은 매일의 의사소통 상황에서 아동의 언어 산출 능력을 평가하는 최적의 기준으로 알려져 있다(Constanza-Smith, 2010). 발화 분석은 임상가가 살펴보고 싶은 언어의 영역을 자세하게 분석할 수 있어 아동의 언어 능력에 대해 표준화 검사가 제공할 수 없는 중요한 정보를 제공해 준다. 또한 발화 분석은 자연스러운 맥락에서 아동이 산출하는 발화를 수집하기 때문에 아동이 평소 사용하는 언어를 평가한다는 의미에서 생태학적 타당도가 높다(Botting, 2002).

발화를 수집하기 위해 대화, 이야기, 설명하기 등 다양한 방법을 사용할 수 있는데, 복잡한 구문 능력을 보이는 학령기 아동에게는

이야기를 사용하는 것이 유용하다. 이야기는 짧은 시간 안에 가장 복잡한 구문을 효과적으로 이끌어낼 수 있는 과제로 알려져 있다(Nippold et al., 2014; Scott & Windsor, 2000; Yang & Kim, 2021). 좋은 이야기 산출을 위해서는 복잡하고 정교한 언어 능력과 인지 능력의 협응이 요구되기 때문에 이야기는 학령기 아동의 언어 능력을 살펴보기 위한 가장 좋은 도구로 사용된다(Veneziano & Nicolopoulou, 2019; Wagner, Nettelbladt, Sahlén, & Nilholm, 2000).

미국에서는 Systematic Analysis of Language Transcripts, SALT; Miller & Iglesias, (2010)나 Computerized Language Analysis, CLAN; MacWhinney, (2010)과 같은 다양한 언어 분석 프로그램이 개발되어 임상가들이 쉽게 발화 분석을 할 수 있게 되었다. 그럼에도 불구하고 일부 임상가들은 임상 현장에서 발화 분석을 규칙적으로 사용하지 않고 있다(Heilmann, 2010). Pavelko, Owens Jr, Ireland와 Hahs-Vaughn (2016)이 1,399명의 학교 임상가들을 대상으로 발화 분석 사용에 관한 조사를 실시한 결과, 임상가들 중 2/3만이 전년도에 발화 분석을 사용했으며, 그중 55%의 임상가들은 일 년 동안 10개 미만의 발화 자료를 분석하였다고 보고하였다. 또한 언어 분석을 사용하지 않는 이유로 79%의 임상가들이 시간 소요의 문제를 꼽았다. 많은 연구에서도 발화 분석의 걸림돌이 발화 자료를 전사하는 데에 필요한 시간이라고 지적하고 있다(Fox, Israelsen-Augenstein, Jones, & Gillam, 2021; Heilmann, 2010; Tomasello & Stahl, 2004; Westerveld & Claessen, 2014). Heilmann (2010)은 1시간 발화 자료를 전사하는 데에 5시간이 걸린다고 보고하였으며, Tomasello와 Stahl (2004)은 10-20배의 시간이 걸린다고 보고하였다. 최근에는 발화 분석에 방해가 되는 시간 문제에 관심이 증가하면서 이를 개선하기 위한 노력들이 이루어지고 있다(Scott, Gillon, McNeill, & Kopach, 2022).

최근 첨단기술이 발전함에 따라 인공지능이 스마트폰, TV, 네비게이션 등 다양한 일상생활 영역에 도입되면서 우리의 편의성과 삶의 질이 높아지고 있다. 그중 인공지능의 자동음성인식(automatic speech recognition) 시스템은 발화 전사를 위한 도구로 사용될 수 있는 큰 잠재력을 지니고 있다(Fox et al., 2021). Google, Naver, Microsoft를 비롯한 많은 기업들은 개발자가 아니더라도 누구나 쉽게 자동음성인식 프로그램을 개발할 수 있도록 클라우드 기반의 Speech-to-Text (STT) Open API를 제공하고 있다. 따라서 자동음성인식 프로그램을 이용한 연구들이 국내외적으로 활발히 이루어지고 있으며 여러 분야에서 응용 사례들이 제시되고 있다. 국내에서 자동음성인식 프로그램을 기업별로 비교한 연구에서는 Kakao (94%)의 정확도가 가장 높고 Google (77%)이 가장 낮은 것으로 나타났다(Yoo, Kim, Park, & Kim, 2020). 또한 ETRI STT 시스템의 자동음성인식 프로그램을 학습시킨 연구에서는 음성인식 정확도가 약 90%로 Naver (약 77%)와 Google (약 65%)보다 인식률이 높게 나타났다(Choi et al., 2020). 그러나 국내 연구들은 뉴스, 시사, 다큐 등에서 수집한 성인 발화만을 사용했기 때문에 아동의 음성인식에 관한 정보는 제시되어 있지 않다. 아동은 성인에 비해 발음이 불분명하며 말의 가변성이 높고 문법적인 오류가 많아 성인의 말과는 다른 특징을 갖는다. 그러나 아직 국내에서는 아동 음성에 초점을 둔 자동음성인식 연구를 찾아보는 것이 힘든 실정이다.

아동 음성에 대한 자동음성인식 연구는 국내보다 국외에서 활발히 이루어지고 있다. Kennedy 등(2017)은 만 4-5세 아동의 발화를 Google, Microsoft, CMU sphinx의 음성인식률로 비교한 결과, Google이 가장 높게 나타났으나 인식률이 40%를 넘기지 못했으며, 아동의 말에서 나타난 문법적 오류로 인해 인식률이 저하되었다고 보고했다. 음성인식의 정확도를 단어 오류율(Word Error Rate)로 살펴본 연구들에서는 오류율이 15-60%로 일관적이지 않은 것으로 나타났다(Booth, Carns, Kennington, & Rafla, 2020; Fox et al., 2021; Lileikyte, Irvin, & Hansen, 2020; Scott et al., 2022). 연구자들은 인공지능을 학습시킴으로써 음성인식 오류율을 개선하고자 하였다. Lileikyte 등(2020)은 음향 및 데이터 증강(acoustic & data augmentation) 실험으로 인공지능을 훈련시켰으나, 아동 발화 인식에 대해 약 60%의 높은 단어 오류율을 보고하였다. 반면 Scott 등(2022)은 발화 전사 및 분석을 위한 플랫폼을 개발하여 Azure의 STT로 인공지능을 훈련시킨 결과, 만 5-6세의 발화에 대한 단어 오류율이 15-20%로 나타났음을 보고했다. Fox 등(2021)은 학령기 언어발달장애 아동의 음성을 Google STT와 실시간 전사자와 비교한 결과 Google STT의 단어 오류율은 30%였으나, SALT의 4가지 측정 항목(TNU, MLU-w, NDW, TNW)과 높은 신뢰도를 보여 자동음성인식 시스템의 임상적인 유용성에 대한 큰 잠재력을 보고하였다.

본 연구는 자동음성인식 프로그램을 이용하여 학령기 일반 아동과 언어학습장애 아동이 산출한 이야기 자료의 음성인식 정확도를 비교·분석하고자 하였다. 자동음성인식 프로그램은 Naver Clova Speech (Naver Clova)와 Google Speech-to-Text (Google STT)를 이용하였으며, 두 프로그램의 선정 이유는 접근성이었다. 국내는 Kakao, ETRI 등 API를 제공하는 여러 기업들이 있으나, Kakao는 기업을 대상으로 하여 일반인들의 접근성이 떨어지며 ETRI는 음성 파일의 시간이 1분 미만에 한하여 지원이 가능하다는 한계가 있다. 반면 Naver는 한국어 기반의 자동음성인식 플랫폼인 Clova를 제공하고 있고 파일의 크기와 상관없이 파일 업로드가 가능하며 접근이 용이하고 비용이 적게 들기 때문에 본 연구에서 선정되

었다. Google은 Naver와 달리 STT Open API를 이용하여 자동음성인식 프로그램을 직접 개발해야 하는 번거로움이 있다. 그러나 IBM, Amazon 등의 다른 해외 기업들에 비해 국내에서 인지도가 높고 접근이 용이하기 때문에 본 연구에서 선정되었다. Naver Clova와 Google STT의 음성인식 정확도는 음절 오류율로 측정되었다. 한국어는 영어와 달리 단어 단위보다는 음절 단위의 구분이 쉬워 구분 일치도가 높다. 이러한 한국어의 특성을 고려하여 음절 단위의 자동음성인식 오류율을 비교하였으며, 각 자동음성인식 프로그램에 나타난 전사 오류 유형을 분석하였다. 본 연구의 연구문제는 다음과 같다. 첫째, 자동음성인식 프로그램과 아동 집단에 따라 음성인식 정확도에 차이가 있는가? 둘째, 각 자동음성인식 프로그램이 보이는 오류 유형은 어떠한가?

연구방법

연구대상

본 연구의 대상자는 초등학교 1, 2학년의 일반 아동 20명과 언어학습장애 아동 20명으로 총 40명이었다. 연구에 참여한 아동은 한국 비언어지능검사 2판(Korean Comprehensive Test of Nonverbal Intelligence-second edition, K-CTONI-2; Park, 2014)의 도형 척도에서 80 이상에 속하였으며, 부모나 교사에 의해 신체, 정서, 청력상의 문제가 없다고 보고되었다. 일반 아동의 경우 한국어 읽기 검사(Korean Language Reading Assessment, KOLRA; Pae et al., 2015)의 읽기 지수 2 (해독+읽기이해+읽기 유창성)가 91 이상이고, 학령기 아동 언어검사(Language Scale for School-aged Children, LSSC; Lee, Heo, & Jang, 2015)의 언어지수가 85 이상인 아동으로 선정하였다. 언어학습장애 아동의 경우 KOLRA의 읽기 지수 2가 90 이하이고, LSSC의 언어지수가 85 미만인 아동으로 선정하였다. 각 대상자들의 연령, 지능, 읽기 능력, 언어 능력에 대한 평균 및 표준편차는 Table 1에 제시하였다.

이야기 자료

이야기 자료는 Yang과 Kim (2021) 그리고 Jang (2022)의 연구에서 수집된 이야기를 사용하였다. 이야기 과제는 자발적으로 산출한 꾸며말하기(generation) 과제인 도깨비 이야기(Kim, Hwang, & Kim, 2018)와 놀이동산 이야기(Kim, Kim, & Han, 2015)를 사용하였고, 40명의 아동에게 각각 두 가지 이야기 발화 자료를 수집하여 총 80개의 이야기 자료가 수집되었다.

전체 이야기 자료는 이야기를 수집한 연구자들에 의해 2차 전사된 자료였다. 본 연구의 분석을 위해 제 1연구자와 제 2연구자가 함께 3차로 이야기를 전사하였다. 전체 이야기의 10%를 무선으로 추출하여 제 1연구자와 제 2연구자가 전사한 결과 98.95%의 신뢰도가 나타났다. 두 연구자의 전사가 일치하지 않는 부분은 제 3연구자와 함께 논의 후 수정하였다. 모두 4차에 걸쳐 수정된 최종 자료는 두 자동음성인식 프로그램의 전사 정확도를 산출하기 위한 준거자료로 사용되었다. 분석에 사용할 이야기 자료는 총 80개였으나, 전반적으로 음질이 낮았던 언어학습장애 아동 4명의 음성 파일을 제외하여 최종적으로 총 36명의 72개 자료를 사용하였다.

연구 절차

각각의 이야기 자료는 클라우드 기반의 음성인식 Open API (Application Programming Interface)로 국내에서 대표적으로 사용하는 Naver Clova와 Google STT를 사용하여 자동전사 하였다.

Naver Clova

Naver Clova는 포털에서 상품 이용을 신청한 후 도메인을 생성하고 음성 파일과 결과 파일을 저장할 object storage를 생성하였다. 도메인의 Clova Speech 빌더를 실행하여 인식 작업을 요청하였고, 각 음성 파일을 object storage에 업로드 하였다. 버킷을 통해 업로드한 파일은 인식 작업 요청 후에 자동전사 결과를 .smi 파일로 저장하였다.

Google STT

Google STT는 클라우드에서 프로젝트를 생성한 후 API키를 발급받았다. 연구자들이 의뢰한 개발자에 의해 구현된 Google STT는 .NET Framework 기반으로 개발되었으며 Google 클라우드 서비스(Google Cloud Service) API 중 STT 서비스를 활용하였다. Google STT는 파일의 스트리밍 시간이 1분 미만이거나 10 MB 미만인 버퍼에 대해서는 다이렉트로 서비스 호출이 가능하나, 본 연구에 사용된 이야기 자료는 대부분 1분 이상에 해당되어 Google 스토리지 서비스(Google Storage Service)를 통한 API 호출이 필요

Table 1. Participant's characteristics

	TD (N=20)	LLD (N=20)
Age (month)	92.45 (8.16)	91.35 (7.30)
K-CTONI-2 nonverbal IQ[a]	111.35 (18.46)	104.05 (11.79)
KOLRA standard score[b]	98.70 (6.95)	63.25 (17.90)
LSSC language quotient[c]	103.70 (9.34)	74.30 (7.92)

Values are presented as mean (SD).
TD = typically developing children; LLD = language learning disability.
[a]Korean Comprehension Test of Nonverbal Intelligence-second edition (K-CTONI-2; Park, 2014), [b]Korean Language Reading Assessment (KOLRA; Pae et al., 2015), [c]Language Scale for School-aged Children (LSSC; Lee et al., 2015).

https://doi.org/10.12963/csd.23952

했다. 이후 Google 콘솔에서 버킷을 생성하여 자동전사하였다.

자료처리 및 분석

준거 자료와 자동 전사 자료의 음절 일치의 기준은 다음과 같다. (1) 모든 전사는 기본적으로 철자 전사를 원칙으로 하였다. (2) 자동 전사에서 준말로 전사된 경우 준거 자료와 일치하는 것으로 처리하였다. 예를 들어, 준거 자료의 "아이"를 자동 전사가 "애"로 전사한 경우 두 단어의 의미를 동일한 것으로 보아 일치하는 것으로 처리하였다. (3) 사람 이름과 같은 고유 명사가 연음법칙과 같은 음운변동으로 인하여 전사가 달라질 경우 일치하는 것으로 처리하였다. 예를 들어 준거 자료의 "솜이"를 자동 전사가 "소미"라고 전사한 경우 명확한 식별이 어려운 것으로 보고 일치하는 것으로 처리하였다. (4) 단, 고유 명사가 접미사 "-이"와의 결합으로 연음으로 전사된 경우, 한국어 이름이 대부분 2음절로 구성되어 있는 특성을 고려하여 준거 자료와 일치하지 않는 것으로 처리하였다. 예를 들어, "양갈이가"를 자동 전사가 "양가리가"라고 전사했을 경우 불일치하는 것으로 처리하였다.

음성인식 정확도는 준거 자료를 기준으로 Naver Clova 와 Google STT의 음절 오류율로 산출하였다. 음절 오류율은 음절의 대치(Substitution), 생략(Deletion), 삽입(Insertion)의 수를 합쳐 준거 자료의 총 음절 수로 나누어 계산하였다. 아동의 발음이 불명료하여 전사하기 어려운 부분은 총 음절 수에서 제외하였다.

$$\text{음절 오류율} = \frac{\text{대치+생략+삽입}}{\text{총 음절 수}} \times 100$$

오류 유형을 살펴보기 위해 모든 오류 음절은 대치, 생략, 삽입 중 하나로 분류하여 오류율을 산출하였다.

통계

집단과 자동음성인식 프로그램(Naver Clova, Google STT)에 따라 음절 오류율에 차이가 있는지 살펴보기 위해 집단을 개체 간 변수로 하고 자동음성인식 프로그램을 개체 내 변수로 하여, 반복측정 분산분석(repeated measures ANOVA)을 실시하였다. 통계분석을 위해 IBM SPSS ver. 22를 사용하였다.

Table 2. Descriptive statistics of syllable error rate (%)

	Naver Clova	Google STT
TD (N=20)	12.61 (6.33)	26.19 (14.17)
LLD (N=16)	17.70 (8.95)	34.04 (16.11)

Values are presented as mean (SD).
TD = typically developing children; LLD = language learning disability.

연구결과

자동음성인식 프로그램과 집단에 따른 음절 오류율의 차이

자동음성인식 프로그램과 집단의 따른 음절 오류율을 살펴본 결과는 다음과 같다. Naver Clova에서는 일반 아동의 음절 오류율이 12.61%, 언어학습장애 아동의 음절 오류율이 17.70%로 나타났고, Google STT에서는 일반 아동의 음절 오류율이 26.19%, 언어학습장애 아동의 음절 오류율이 34.04%로 나타났다. 즉 Naver Clova는 Google STT보다 더 낮은 음절 오류율을 보였으며, 일반 아동 집단에서의 음절 오류율이 언어학습장애 아동 집단에서의 음절 오류율보다 낮았다. 아동 집단과 자동음성인식 프로그램에 따른 음절 오류율의 기술통계는 Table 2, Figure 1에 제시하였다.

자동음성인식 프로그램과 집단에 따른 주효과와 상호작용 효과를 살펴보기 위해 반복측정 분산분석을 실시한 결과, 자동음성인식 프로그램에 따른 주효과는 유의미한 것으로 나타났으나($F_{(1.34)}$ = 110.59, p = .000), 집단에 따른 주효과는 유의미하지 않은 것으로 나타났다($F_{(1.34)}$ = 3.50, p = .07). 즉 Naver Clova는 Google STT보다 통계적으로 유의하게 음절 오류율이 낮은 것으로 나타났으며 이는 Naver Clova가 Google STT보다 아동의 이야기를 더 정확하게 전사했다는 것을 의미한다. 그러나 일반 아동과 언어학습장애 아동의 음절 오류율에서는 통계적으로 유의미한 차이가 나타나지 않았다. 자동음성인식 프로그램과 집단에 따른 상호작용 효과 또한 유의하지 않은 것으로 나타났다($F_{(1.34)}$ = 0.94, p = .9).

자동음성인식 프로그램에 따른 오류 유형

Naver Clova와 Google STT의 오류 유형을 살펴본 결과, Naver Clova에서는 대치(59.21%), 생략(34.84%), 삽입(5.95%) 순으로 높게 나타났고, Google STT에서는 생략(56.04%), 대치(40.59%), 삽입(3.40%) 순으로 높게 나타났다. 즉 Naver Clova는 대치 오류가 가장

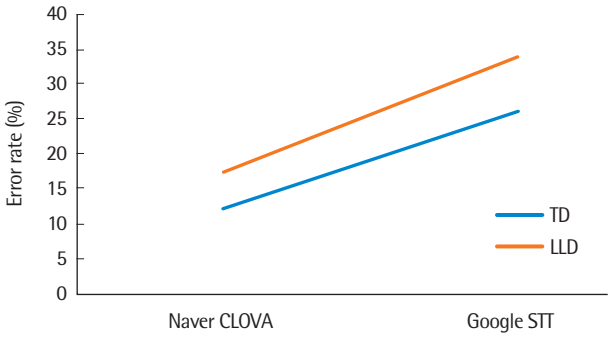

Figure 1. Syllable error rate of two groups and automatic speech recognition program.
TD = typically developing children; LLD = language learning disability.

Table 3. The percentage of error types in the automatic speech recognition program

	Naver Clova	Google STT
Substitution	59.21 (1.31)	40.56 (.77)
Deletion	34.84 (0.78)	56.04 (1.30)
Insertion	5.95 (.19)	3.40 (.08)

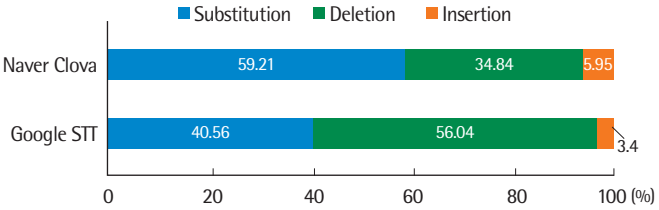

Figure 2. The percentage of error types in the automatic speech recognition program.

높은 반면, Google STT는 생략 오류가 가장 높은 것으로 나타났다. 두 자동음성인식 프로그램에서 삽입 오류는 가장 낮은 것으로 나타났다. 각 프로그램에 따른 대치, 생략, 삽입의 비율(%)은 Table 3과 Figure 2에 제시하였다.

논의 및 결론

본 연구는 Naver Clova와 Google STT가 학령기 일반 아동과 언어학습장애 아동의 음성을 얼마나 정확하게 인식하는지 살펴보았고, 자동음성인식 프로그램에서 나타난 오류 유형을 분석하여 각 프로그램이 갖는 특징을 살펴보았다.

첫 번째로 자동음성인식 프로그램과 집단에 따른 음절 오류율의 차이가 있는지 살펴본 결과, Naver Clova는 Google STT보다 음절 오류율이 유의하게 낮았다(p = .000, p < .05). 즉 국내의 자동음성인식 프로그램이 국외의 자동음성인식 프로그램보다 한국어를 더 잘 인식하고 전사하였다. 이러한 결과는 국내의 선행연구에서 Naver Clova가 Google STT에 비해 성인 음성을 더 잘 인식한다는 연구결과와 일치한다(Choi et al., 2020; Yoo et al., 2020). Naver Clova는 코퍼스 97%를 한국어로 구성하여 한국어에 특화된 언어 모델을 구축하였다(https://naver-ai-now.kr/). 그러므로 많은 한국인의 음성 자료를 기초로 한 Naver Clova가 한국어 음성인식에 유리하게 작용한 것으로 보인다.

반면 일반 아동 집단과 언어학습장애 아동 집단의 음절 오류율은 통계적으로 유의한 차이가 나타나지 않았다(p = .07, p > .05). 이러한 결과는 본 연구의 참여자가 학령기 아동이었기 때문에 나타난 결과로 보인다. Booth 등(2020)은 자동음성인식 프로그램이 더 어린 아동의 음성을 인식하는 것이 어려웠다고 보고하였다. 어린 아동의 말은 가변성이 높고 비문법적인 발화의 비율이 높아 자동음성인식 프로그램의 음성 인식률을 저하시키게 된다. 아동은 연령이 높아질수록 언어 산출을 위한 운동 기능이나 언어 계획 능력이 발달하면서 말 산출이 정교해지고 명료도가 높아진다. Booth 등(2020)의 연구에서 1학년 집단에 비해 2-3학년 집단의 전사 일치율이 높았던 것을 고려하면, 본 연구에서도 높은 말 명료도와 안정적인 음향학적 특징으로 인해 자동음성인식은 장애 집단과 관계없이 학령기 아동의 음성을 잘 인식했을 것으로 보인다.

두 번째로 자동음성인식 프로그램의 오류율을 유형별로 비교하여 분석한 결과, Naver Clova의 오류 유형은 대치, 생략, 삽입 순으로, Google STT는 생략, 대치, 삽입 순으로 높게 나타났다. 먼저 대치 유형을 살펴보면, 대치는 Naver Clova에서 59.21%로 가장 높았고, Google STT에서는 40.59%로 두 번째로 높은 비율이었다. Naver Clova와 Google STT는 모두 고유명사를 인식하는 데 어려움이 있었다. 예를 들어, 두 프로그램은 등장인물의 이름인 '지연이'를 발화 전체에서 '지연이', '지윤이', '지은이' 등 비일관적으로 전사하였다. 이는 자동음성인식 프로그램이 동일한 고유명사를 인간처럼 일관성 있게 예측하거나 유추해내는 것이 어렵다는 것을 보여준다. 인간은 발화가 불명료한 경우에도 맥락을 고려하여 단어를 일관적으로 인식할 수 있는 반면 자동음성인식 프로그램은 아직 그 기능이 미흡한 것으로 보인다.

'고유명사+조사' 형태에서 Naver Clova는 고유명사의 음절 대치가 주로 나타난 반면 Google STT는 고유명사의 품사 자체가 바뀌는 것이 종종 나타났다. 예를 들어 Naver Clova는 '강성이랑 곰곰이가'를 '간성이랑 궁금이가'로 전사하여 고유명사라는 형태 안에서의 음절 오류만 나타났으나, Google STT는 '강 서방이랑 궁금합니다'로 전사하여 고유명사가 일반명사나 동사로 바뀌는 것이 관찰되었다. 이는 Naver Clova가 등장인물의 이름은 고유명사로, 조사는 조사로 인식하는 것, 즉 한국어 품사를 더 잘 구분하는 것으로 해석할 수 있으며, 조사를 전사하는 면에서도 비교적 수행력이 좋다고 볼 수 있다.

대치는 일반명사와 동사에서도 많이 나타났다. 일반명사에서는 Naver Clova가 '회전목마'와 '바이킹'을 '회장 엄마, 바이크 등'으로, Google STT는 '세종 엄마, 파이팅 등'으로 대치시켰다. 아동이 명사의 산출 오류를 보였을 때에는 자동음성인식 프로그램이 해당 단어를 신조어로 해석하여 의미있는 단어로 대치시키기도 하였다. 예를 들어 아동이 '놀이공원'을 '놀이공영'이라고 산출했을 때 Naver Clova는 '놀이공 밤', Google STT는 '놀이 공룡'으로 대치시켰고, '도깨비'를 '도빼기'라고 산출했을 때 Naver Clova는 '도배기, 독백

https://doi.org/10.12963/csd.23952

이', Google STT는 '도깨비 김, 뚝배기'로 대치시켰다. 동사에서는 두 프로그램 모두에서 어간 대치(예: 말했습니다 → 맛있습니다, 만졌는데 → 맞았는데 등)와 어미 대치(예: 했습니다 → 했으니까, 라라잖아 → 라라 자나 등)가 자주 나타났다. 이러한 결과들 역시 자동음성인식 프로그램이 선행 정보를 활용하여 유추하는 능력이 인간에 비해 미흡한 것으로 보인다.

다음으로 생략은 Naver Clova에서 34.84%로 두 번째로 많은 오류 유형이었고, Google STT에서는 56.04%로 가장 많은 오류 유형으로 나타났다. Google STT는 Naver Clova보다 훨씬 더 많은 생략을 보였는데, 이는 자동음성인식 프로그램이 언어적 비유창성(mazes)을 무시하여 나타난 결과라고 해석된다. 발화 중간에 약간의 쉼이나 연장이 생기거나, 반복, 수정 등 언어적 비유창성 요소가 포함될 경우, 자동음성인식 프로그램은 언어적 비유창성 요소들을 전사 제외 항목으로 인식하여 자체적으로 생략하였다. 그 예로 아동이 '걸어 (쉼) 서'를 산출했을 때 Naver Clova는 '그 수'로, Google STT는 '서'로 전사했으며, 아동이 '소 소리'라고 1음절을 반복했을 때 Naver Clova와 Google STT는 모두 '소리'라고만 전사하였다. 반면 자동음성인식 프로그램이 언어적 비유창성 요소를 제대로 인식하지 못한 경우도 있었다. 예를 들면 '방망 (쉼) 이를 들고'에 대한 자동전사에서 Naver Clova는 '엄마 이제 들고'로 전사했고 Google STT는 '애들 두고'로 전사했다. 이는 인공지능이 해당 단어를 코퍼스 내에서 음운적으로 가장 유사한 단어로 찾아내는 것으로 보인다.

Naver Clova와 Google STT는 모두 명사, 대명사, 접속부사, 동사 등 다양한 품사의 단어를 통째로 생략하기도 했다. 생략 빈도는 Naver Clova보다 Google STT에서 월등히 높았는데, 이는 인공지능의 코퍼스 크기에 따른 차이로 해석할 수 있다. Lileikyte 등(2020)은 인공지능을 학습시키기 위해 많은 아동의 음성 코퍼스가 필요함을 언급한 바 있다. 앞서 설명한 바와 같이 Naver Clova는 충분한 양의 한국인의 음성 자료를 확보하여 한국어에 특화된 모델로 인공지능을 학습시켰으나, 반면 Google STT는 한국어 코퍼스 크기가 작아 한국어를 인식하여 적절한 단어를 찾아내는 것이 더 어려웠을 것으로 보인다.

Naver Clova와 Google STT에서 나타난 공통적인 특징은 두 프로그램 모두 단어를 철자 전사 원칙에 따라 전사하는 경향을 보였다는 것이다. 예를 들어 '탈라고 → 타려고', '몰르고 → 모르고' 등과 같은 구어적 표현은 모두 표준어로 자동전사 되었다. 그러나 아동이 산출 오류를 보였을 때에도 철자 전사 원칙이 적용되어 오류를 오류가 아닌 것으로 전사하는 문제가 발생했다(예: 두둘렸어요 → 두드렸어요, 만들으지만 → 만들었지만 등). 발화 분석 시 아동이 산출하는 오류는 임상가들에게 유용한 정보를 제공한다. 그러나 철자 전사 원칙에 따라 표준어로 전사되는 자동음성인식 프로그램을 사용할 때는 프로그램의 이러한 특성을 고려하여 부분적인 수정이 필요하다.

이러한 결과는 언어적 비유창성에서도 나타났다. 두 프로그램은 모두 아동의 수정, 간투사, 다시 시작하는 발화 등의 여러 비유창성 요소 대부분을 제외 항목으로 인식하였다. 발화에 포함된 수정 발화를 비교적 더 전사한 프로그램은 Naver Clova로, '그러 그리고'가 '그 그리고'로, '거짓말로 했 거짓말'이 '거짓말로 했지 거짓말'로 전사되었다. 반면 Google STT에서는 이러한 발화의 경우 아예 전사되지 않았다. 간투사에서는 위와 다른 특징을 보였는데, 자동음성인식 프로그램이 모두 간투사를 인식하여 전사에서 제외시켰다는 점은 동일했으나 간투사를 처리하는 방식에서는 약간의 차이가 있었다. Naver Clova는 단어 중간에 삽입되는 '음'을 간투사로 인식하여 간투사를 제외한 단어 형태로 전사할 수 있었다(예: 앉아 음 서 → 앉아서). 반면 Google STT는 단어 내 간투사를 처리하지 못해 의미있는 한 단어로 전사하지 못했다(예: 앉아 음 서 → 앉아 잘). 몇몇 경우에는 특정 간투사가 의미있는 단어로 대치되었는데, 비교적 Google STT에서 간투사 '음'이 대답의 기능인 '응'이나 명사 '물'로 전사되는 것이 많이 관찰되었다.

마지막으로 본 연구에서 아동 음성의 자동음성인식 정확도는 선행연구에서 보고한 성인 음성에 비해서 높은 오류율을 보였으나(Choi et al., 2020; Yoo et al., 2020), 자동음성인식 프로그램이 실제 전사보다 시간을 단축시켰다는 점에서 임상적 유용성이 있다. Scott 등(2022)은 연구자들이 개발한 인공지능 플랫폼을 이용하면 수집에서 분석까지 소요되는 시간을 크게 줄일 수 있으나, 낮은 전사 정확도로 인해 인간이 전사 내용을 다시 확인하여 아동의 발화를 수동으로 식별하고 전사해야 하는 노력이 필요하다고 보고했다. 본 연구에서 자동음성인식 프로그램이 1분 녹음된 음성 파일을 전사하기 위해 소요된 시간은 Naver Clova에서 약 4초, Google STT에서 약 24초로 나타나 시간이 매우 단축되었다. 그러나 선행연구와 유사하게 자동전사가 이루어진 후 다시 인간에 의한 수동 전사가 필요한 것은 불가피한 것으로 나타났다. 그럼에도 불구하고 이러한 점은 그동안 발화 전사에 드는 시간으로 기피했던 부분들이 인공지능을 통해 보완이 되고, 임상가들이 발화 전사를 위해 자동음성인식 프로그램을 사용할 수 있다는 가능성을 보여준다. 또한 이는 언어 평가의 질을 높여 이후 질 높은 중재에도 기여할 수 있을 것으로 사료된다.

결론적으로 자동음성인식 프로그램에서 나타난 특징을 비교·분석한 결과는 다음과 같다. 첫째, 발화 내 언어적 비유창성 요소가

적고 음성 파일의 음질 문제가 없는 경우에는 Naver Clova의 전사 정확도가 Google STT에 비해 더 높았다. 둘째, 두 프로그램은 모두 철자 전사를 원칙으로 하여 아동의 산출 오류를 표준어로 수정하는 특징을 보였다. 셋째, 자동음성인식 프로그램은 전사 시간을 매우 단축시킬 수 있어 전사 도구로서의 실용성과 임상에서의 유용성이 매우 클 것으로 예상된다.

본 연구에서 나타난 제한점은 다음과 같다. 첫째, 연구에 참여한 아동은 총 36명으로, 대상자 수가 적어 연구결과를 일반화시키기에 충분하지 않았다. 따라서 후속연구는 대상자 수를 늘려 더 많은 발화 자료를 통해 자동음성인식 프로그램의 유용성을 입증할 필요가 있다. 둘째, 아동의 연령이 학령기로 제한되어 있어 학령전기 아동의 음성인식 정확도를 살펴볼 수 없었다. 학령전기 아동은 학령기 아동과 다른 말 특성을 보이므로 학령전기 아동을 대상으로 연구할 경우 본 연구와 다른 연구결과가 나타날 수 있다. 따라서 후속연구에서는 학령전기 아동을 대상으로 한 자동음성인식 프로그램의 전사 정확도를 비교해 보는 것이 필요할 것으로 사료된다.

REFERENCES

Booth, E., Carns, J., Kennington, C., & Rafla, N. (2020). Evaluating and improving child-directed automatic speech recognition. *In Proceedings of the 12th Language Resources and Evaluation Conference,* 6340-6345.

Botting, N. (2002). Narrative as a tool for the assessment of linguistic and pragmatic impairments. *Child Language Teaching & Therapy, 18*(1), 1-21.

Choi, M. A., Kim, S. H., Jo, M. A., Park, D. Y., Kim, Y. H., &Yoon, J. H. (2020). Development and enhancement of automatic caption generation system based on speech-to-text for the hearing impaired. *Proceedings of The Korean Institute of Broadcast and Media Engineers Summer Conference,* 343-346.

Costanza-Smith, A. (2010). The clinical utility of language samples. *Perspectives on Language Learning & Education, 17*(1), 9-15.

Fox, C. B., Israelsen-Augenstein, M., Jones, S., & Gillam, S. L. (2021). An evaluation of expedited transcription methods for school-age children's narrative language: automatic speech recognition and real-time transcription. *Journal of Speech, Language, & Hearing Research, 64*(9), 3533-3548.

Heilmann, J. J. (2010). Myths and realities of language sample analysis. *Perspectives on Language Learning & Education, 17*(1), 4-8.

Jang, H. B. (2022). *Use of cohesive devices in narratives of children with language learning disabilities.* Nazarene University, Cheonan, Korea.

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., ..., & Belpaeme, T. (2017). Child speech recognition in human-robot interaction: evaluations and recommendations. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction,* 82-90.

Kim, J. M, Kim, H. S., & Han, D. S. (2015). Picture narrative assessment: content validity of sequential picture task. *Proceedings in Spring Conference on the Korean Society of Speech Sciences,* 57-58.

Kim, J. M., Hwang, S. E., & Kim, H. S. (2018). Narrative macrostructure of school-aged children under different picture tasks. *Communication Sciences & Disorders, 23*(2), 255-269.

Kim, Y. T. (2014). *Assessment and intervention of child language disorders.* Seoul: Hakjisa.

Lee, Y., Heo, H., & Jang, S. (2020). *Language scale for school-aged children* (LSSC). Seoul: Hakjisa.

Lileikyte, R., Irvin, D., & Hansen, J. H. (2020). Assessing child communication engagement via speech recognition in naturalistic active learning spaces. *Proceedings of the Odyssey 2020 Speaker and Language Recognition Workshop,* 396-401.

MacWhinney, B. (2000). *CLAN* [Computer software]. Pittsburgh, PA: Carnegie Mellon University.

Miller, J. F., & Iglesias, A. (2010). *Systematic analysis of language transcripts (SALT),* research version [Computer Software]. WI: SALT Software, LLC.

Nippold, M. A., Frantz-Kaspar, M. W., Cramond, P. M., Kirk, C., Hayward-Mayhew, C., & MacKinnon, M. (2014). Conversational and narrative speaking in adolescents: examining the use of complex syntax. *Journal of Speech, Language, & Hearing Research, 57*(3), 876-886.

Pae, S., Kim, M., Yoon, H., & Jang, S. (2015). *Korean Language-based Reading Assessment* (KOLRA). Seoul: Hakjisa.

Park, H. W. (2014). *Korean version of comprehensive test of nonverbal intelligence-second edition (K-CTONI-2).* Seoul: Mindpress.

Pavelko, S. L., Owens Jr, R. E., Ireland, M., & Hahs-Vaughn, D. L. (2016). Use of language sample analysis by school-based SLPs: results of a nationwide survey. *Language, Speech, & Hearing Services in Schools, 47*(3), 246-258.

Scott, A., Gillon, G., McNeill, B., & Kopach, A. (2022). The evolution of an innovative online task to monitor children's oral narrative development. *Frontiers in Psychology, 13,* 1-10.

Scott, C. M., & Windsor, J. (2000). General language performance measures in spoken and written narrative and expository discourse of school-age children with language learning disabilities. *Journal of Speech, Language, & Hearing Research, 43*(2), 324-339.

https://doi.org/10.12963/csd.23952

Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech: how much is enough?. *Journal of Child Language, 31*(1), 101-121.

Yang, H. J., & Kim, J. M. (2021). Comparison of syntactic ability of children with and without language learning disabilities in narratives. *Journal of Speech-Language & Hearing Disorders, 30*(4), 43-52.

Yoo, H. J., Kim, M. W., Park, S. K., & Kim, K. Y. (2020). Comparative analysis of Korean continuous speech recognition accuracy by application field of cloud-based speech recognition open API. *The Journal of Korean of Communications & Information Sciences, 45*(10), 1793-1803.

Veneziano, E., & Nicolopoulou, A. (2019). *Narrative, literacy and other skills: studies in intervention.* Philadelphia: John Benjamins Publishing Company.

Wagner, C. R., Nettelbladt, U., Sahlén, B., & Nilholm, C. (2000). Conversation versus narration in pre-school children with language impairment. *International Journal of Language & Communication Disorders, 35*(1), 83-93.

Westerveld, M. F., & Claessen, M. (2014). Clinician survey of language sampling practices in Australia. *International Journal of Speech-Language Pathology, 16*(3), 242-249.

국문초록

학령기 아동의 이야기 전사를 위한 자동음성인식 프로그램 비교: Naver Clova와 Goolge STT를 중심으로

양희재[1] · 오은별[1] · 김정미[2]

[1]나사렛대학교 일반대학원 언어치료학전공, [2]나사렛대학교 언어치료학과

배경 및 목적: 대부분의 임상가들은 발화 분석 시 전사에 많은 시간이 소요되는 것에 부담을 가진다. 최근 기술의 발전으로 자동음성인식 프로그램들이 개발되었으며, 본 연구는 Naver Clova와 Google STT 프로그램의 전사 정확도를 비교하고, 각 프로그램이 보이는 오류 특성을 체계적으로 살펴보고자 하였다. 방법: 초등학교 1, 2학년의 일반 아동 20명과 언어학습장애 아동 20명을 대상으로 80개의 이야기를 수집하였다. 4차 전사된 이야기 자료는 Naver Clova와 Google STT의 전사 정확도를 비교하는 준거 자료로 사용하였으며 전사 오류를 유형별로 살펴보았다. 결과: 전사 오류율에 있어 Naver Clova (15.17%)는 Google STT (30.11%)보다 통계적으로 유의하게 낮은 것으로 나타났다. 그러나 일반 아동과 언어학습장애 아동의 이야기 자료의 전사 오류율에서는 차이가 나타나지 않았다. 또한 각 프로그램의 오류를 유형별로 분석한 결과, Naver Clova의 오류는 대치가 많은 반면, Google STT의 오류는 생략이 많았다. 논의 및 결론: 결론적으로 아동의 이야기 자료 전사에서 Naver Clova가 Google STT보다 정확도가 높았으나 아동 집단 간 차이는 나타나지 않았다. 실제로 본 연구에서 산출된 자동전사의 낮은 오류율은 자동음성인식 프로그램이 임상적 유용성이 있음을 보여준다. 또한 자동음성인식 프로그램의 사용은 발화 분석에 방해가 되는 시간적 부담의 문제를 경감시키며, 이는 질적으로 높은 언어평가를 가능하게 할 것으로 사료된다.

핵심어: 전사, 자동음성인식 시스템, 음절 오류율

참고문헌

김영태 (2014). **아동언어장애의 진단 및 치료**. 서울: 학지사.

김정미, 김효선, 한다솜 (2015). **그림 이야기 평가: 연속그림 내용 타당도 연구**. 2015 한국음성학회 봄 학술대회 자료집, 57-58.

김정미, 황성은, 김효선(2018). 이야기 유도 과제에 따른 학령기 일반아동의 이야기 대형구조 특성. **Communication Sciences & Disorders, 23**(2), 255-269.

박혜원 (2014). **한국비언어지능검사(Korean comprehensive Test of Nonverbal Intelligence-Second Edition, K-CTONI-2)**. 서울: 마인드프레스.

배소영, 김미배, 윤효진, 장승민 (2015). **한국어읽기검사(Korean Language-based Reading Assessment, KOLRA)**. 서울: 학지사.

양희재, 김정미 (2021). 언어학습장애 아동과 일반 아동의 이야기에 나타난 구문 능력 비교. **언어치료연구, 30**(4), 43-52.

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

이윤경, 허현숙, 장승민 (2015). **학령기아동언어검사(Language Scale for School-aged Children, LSSC)**. 서울: 학지사.

장혜빈 (2022). **초등학교 1, 2학년 언어학습장애 아동의 이야기에 나타난 결속표지 사용 특성**. 나사렛대학교 대학원 석사학위논문.

최미애, 김승현, 조민애, 박동영, 김용호, 윤종후 (2020). 청각장애인을 위한 음성-자막 자동 변환 시스템 개발 및 음성 인식률 고도화. **한국방송미디어공학회 학술발표대회 논문집**, 343-346.

ORCID

양희재(제1저자, 박사과정 https://orcid.org/0000-0003-1141-9803); **오은별**(공동교신저자, 박사과정 https://orcid.org/0000-0003-4173-1005)
김정미(공동교신저자, 교수 https://orcid.org//0000-0003-2420-9434)

 https://doi.org/10.12963/csd.23952