의학박사 학위논문

진료실 대화에 대한 음성인식 솔루션의 정확도 평가

2023년 2월

서울대학교 대학원
의학과 의공학전공

이 승 화

진료실 대화에 대한
음성인식 솔루션의
정확도 평가

지도 교수 최진욱

이 논문을 의학박사 학위논문으로 제출함
2023년 1월

서울대학교 대학원
의학과 의공학전공
이 승 화

이승화의 의학박사 학위논문을 인준함
2023년 2월

위원장	김 희 찬
부위원장	최 진 욱
위 원	김 영 수
위 원	윤 형 진
위 원	김 인 영

초 록

배경: 임상의사의 진료기록 작성을 보조하기 위한 음성인식 시스템은 1980년대부터 의료영역에서 사용 중이다. 그러나 클라우드 기반 음성인식 솔루션의 진료실 대화에 대한 인식 정확도에 대한 연구는 부족한 상황이다. 본 연구에서는 클라우드 기반 음성인식 솔루션을 사용하여 진료실 대화 인식을 통해 진료기록 작성 자동화의 가능성을 알아보고자 하였다. 이를 위해 현재 상용화된 클라우드 기반 음성인식 솔루션들의 한국어 의료대화의 인식률에 대하여 비교연구하였다.

방법: 삼성서울병원 순환기내과 외래 진료를 위해 방문한 환자와 의사와의 실제 진료대화를 녹음하여 현재 사용 가능한 클라우드 기반 음성인식 솔루션의 음성인식 정확도를 비교 분석하였다. 이에 더하여 aihub.or.kr 에서 제공하는 인공지능 훈련을 위한 데이터 셋을 사용하여 의료진의 질문에 대한 음성인식 정확도를 추가로 비교분석하였다.

Results: 의학용어의 음성인식 정확도 연구를 위하여 총 112명의 환자-의사간 대화가 녹음되었으며 Naver Clova SR, Google Speech-to-text, Amazon Transcribe의 총 세종류의 클라우드 기반 음성인식 솔루션을 사용하여 음성인식 작업을 시행한 후 실제 대화를 토대로 작성한 대본에서 나온 의학용어와 각각의 솔루션에서 인식한 의학용어의 인식률을 비교 분석하였다. 세가지 음성인식 솔루션 중 Naver Clova SR이 가장 높은 의학용어의 인식 정확도를 보여주었다. (75.1% vs. 50.9% vs. 57.9%, $P < 0.001$) 추가적으로 Amazon Transcribe가

Google Speech-to-text와 비교하였을 때 통계적으로 유의미한 높은 음성 인식 정확도를 보여주었다. 하위 분석에서 Naver Clova SR은 전반적으로 높은 음성인식 정확도를 보여주었으며, 5글자 이상의 의학용어에서는 구글 음성인식이 더 높은 인식 정확도를 보여주었으나 통계적으로 유의미하지는 않았다.

의료진의 질문의 음성인식 정확도 분석을 위해서 총 500개의 문장을 aihub.or.kr 데이터에서 추출하였으며 Naver Clova SR, Kakao API Speech-to-text, Google Speech-to-text의 세 종류의 음성인식 솔루션을 사용하여 음성인식을 수행 후 비교분석 하였다. 임상의의 판정 및 자동화 측정 지표를 사용하여 음성인식의 정확도를 비교하였으며 Naver Clova SR이 다른 두 솔루션에 비해 유의미하게 높은 음성인식 정확도를 보여주었다. 임상의의 판정에서는 각각 Naver Clova SR (94.7%), Kakao API Speech-to-text (83.8%), Google Speech-to-text (76.7%)의 음성인식 정확도를 보여주었으며 ($p<0.001$), 자동화 측정지표에서는 Bleu-1 (0.654 vs. 0.578 vs. 0.535, $p<0.001$), Bleu-2 (0.557 vs. 0.463 vs. 0.418, $p<0.001$), CIDEr (4.18 vs. 3.39 vs. 3.02, $p<0.001$)로 Naver Clova SR이 가장 높은 음성인식 정확도를 보여주었다.

결론: 현재 상용화된 클라우드 기반 음성인식 솔루션은 환자-의사간 대화의 인식에 있어서 한계점이 있었으며, 실제 의료현장에서 진료실 대화의 음성인식을 통한 의무기록 작성시 바로 적용할 수 없음을 알 수 있었다. 다양한 음성인식 솔루션 중에서는 국내기업에서 제작한 솔루션이 가

장 높은 인식률을 보여주었으며 각각의 음성인식 솔루션 별로 서로 다른 단어 영역에서 음성인식의 강점이 있음을 보여주었다. 추가적으로 실제 대화의 녹음보다는 인공지능 훈련을 위해 제작된 데이터의 음성인식 정확도가 높음을 알 수 있었다.

본 연구의 결과는 차후 음성인식 기술발전을 통하여 의료대화의 음성인식 정확도의 발전이 가능함을 보여주었다. 현재 클라우드 기반 음성인식 솔루션의 인식 정확도 개선을 위해서는 더 많은 의료산업에 적용 목적으로 정제된 훈련 데이터셋이 필요하다.

주요어: 음성인식 (Speech recognition), 인공지능 (artificial intelligence), 의무기록 (medical documentation), 전자의무기록 (electronic health record)

학번: 2020-36376

본 박사논문은 저자가 2022년 5월에 Journal of Korean Medical Science에 출판한 Accuracy of Cloud-Based Speech Recognition Open Application Programming Interface for Medical Terms of Korean 논문을 기반으로 작성하였음.

목 차

표 목차

그림 목차

제 1 장. 서론

1.1 연구 배경

1.1.1 의무기록 작성시 음성인식 기술의 필요성

의료 서비스를 위한 음성 인식 시스템은 1980년대에 최초로 상용화된 후 현재까지 사용되어지고 있다.[1] 의료용 음성인식 시스템은 의료진 혹은 환자의 음성을 텍스트로 번역하거나 사용자가 음성으로 프로그램 등을 제어할 수 있도록 함으로써 의무기록 작성을 지원해주는 입력 시스템이다.[2] 음성인식은 전자의무기록 (Electronic health record, EHR) 시스템을 사용하는 데 있어 가장 많은 시간 및 비용을 소모하는 요소로 알려져 있는 의무기록 작성의 효율성을 재고할 수 있는 기술로 기대되어 왔다.[3] 최근 연구에서는 미국에 소재 중인 의료 기관의 90% 이상이 음성인식 애플리케이션을 의무기록 작성에 채택하거나 음성인식 기능을 확장할 계획임을 보고하였다.[3] 음성인식 시스템을 전자의무기록 작성에 적용할 수 있다면 임상의의 업무 시간을 절약하고 이를 통해 환자의 정확한 진단 및 적절한 치료에 보탬이 될 수 있다.[1,4–6]

1.1.2 COVID-19, 원격의료와 음성인식

인공지능 (artificial intelligence, AI), 머신 러닝 및 빅데이터에 기반한 컴퓨팅 기술 분야의 최근 발전은 COVID-19 팬더믹 상황에서 환자의 감시, 진단, 진료 및 치료제 개발 등 다양한 영역에서 도움을

줄 수 있을 것으로 기대되었다.[7] 또한, 대면진료에 의한 감염의 위험을 감소시키기 위하여 원격의료는 팬더믹 기간동안 환자, 임상의 및 지역사회에 뚜렷한 혜택을 제공하였다.[8,9] AI 기술이 접목된 음성인식 기술은 환자의 증상 및 질병상태를 확인할 뿐 아니라 환자의 추적관찰 자동화 프로그램에 적용되는 등 팬더믹 기간 동안 다양한 목적을 위해 사용되었다.[10,11] Covid-19의 유행 후 이의 전파를 막기위해서 국내에서는 원격진료가 한시적으로 허용되었으며 다양한 원격진료 플랫폼이 사용가능한 현황이다. 원격진료는 개인 컴퓨터, 스마트 디바이스 등을 사용하여 이루어지는 점을 고려하였을 때 대면진료에 비해 음성인식 기술의 적용이 비교적 용이할 뿐 아니라 임상의 의 업무부담 완화 및 진료의 질 향상에 큰 역할을 할 수 있을 것으로 사료된다.[12]

음성인식의 정확도를 높이기 위해서는 음성녹음의 질이 보장되어야 함은 명약관화한 사실이다. 스마트폰 등을 통해 직접 음성녹음을 수행할 수 있다는 사실을 고려하여 본 연구자는 음성인식을 통한 자동 의무기록 작성의 가능성을 알아보고자 본 연구를 계획하였다. 개인 컴퓨터, 스마트폰을 통한 원격진료 시행 전 대기시간에 간단한 환자의 현 병력, 과거력, 약물사용이력 등을 조사 및 녹음을 시행 후 이를 토대로 음성인식 및 자연어 처리를 통하여 진료 전 필요한 정보를 임상의사에게 제공할 수 있다면 진료시간의 단축 및 효율성을 제고할 수 있으며 이를 통해 정확한 진단 및 치료, 궁극적으로는 의료의 질 향상을 불러올 수 있을 것이다.

1.1.3 의료산업에서 음성인식 기술 적용 현황

비록 음성인식 기술은 의료산업에 1980년대부터 사용이 시작되었지만 아직까지 이 기술은 일부 진료과에서만 적용되고 있다.[4] 낮은 음성인식 기술 수준과 높은 인식 오류 발생율은 음성인식 기반 의무기록 작성 기술을 실제 임상현장에 적용하는데 중요한 장애 요인이었다. 그러나 지난 20년간 음성인식 알고리즘 및 음성인식 정확도는 지속적으로 개선되어 왔다. 이에 더불어 최근 AI의 급속한 발전과 클라우드 컴퓨팅 기술의 활용에 힘입어 음성 인식 시스템은 급격한 성능 향상을 보이고 있다.[13,14] 의료 산업은 클라우드 컴퓨팅의 이점인 데이터 수집, 데이터 저장 및 데이터 전송 측면에서 유연성, 확장성 및 편재성을 통한 빠른 발전이 가능한 분야 중 하나로 인식되고 있다.[5]

현재, 음성 인식 시스템은 서구의 병원에서 그리고 응급실에서 환자의 분류, 병리학과, 영상의학과에서 판독 소견을 녹음한 후 기록지를 작성하는 업무 등 실제 임상에서 사용 중에 있다.[1,2,5] 실제로 영상의학과에서 음성인식 시스템을 판독업무에 도입한 이후 판독기록의 작성시간이 15.7 시간에서 4.7 시간으로 단축되었음을 보고한 연구도 존재한다.[15] 그러나 실제 음성인식과 타이핑을 통한 의무기록을 작성한 비교 연구에서는 음성인식이 더 빠른 의무기록 작성을 가능하게 한다는 참가자들의 인식에도 불구하고 음성인식이 의무기록을 작성하는데 더 효율적이거나 정확하다는 증거는 발견되지

않았으며, 기존 EHR 시스템 내 통합의 효과에 대한 우려만 확인할 수 있었다.[6]

1.1.4 클라우드 기반 음성인식 개방형 API

클라우드 기반 음성인식 개방형 애플리케이션 프로그래밍 인터페이스(application programming interfaces, API) 솔루션은 음성인식 시스템 구축 시 시간과 인력, 비용을 효과적으로 절감시켜주는 장점을 가지고 있어 최근 영화 자막, 실시간 번역 등의 다양한 분야에서 사용 중이다.[16] 최근 이러한 음성인식 솔루션을 의료 산업에 접목하려는 다양한 시도가 이루어지고 있다.[10,11,17–19] 하지만 다양한 시도에도 불구하고, 현재 상용화된 의료 서비스 사용 목적의 음성인식 API는 Amazon® Transcribe Medical 또는 Nuance® Dragon Medical One 등 소수의 솔루션만이 존재한다.

Amazon® Transcribe Medical은 배치 워크로드와 실시간 음성-텍스트 애플리케이션을 모두 제공해주는 공용 API 세트로서 심장내과, 신경과, 산부인과, 소아과, 종양내과, 영상의학과 및 비뇨기과와 같은 주요 진료 분야에 대한 의무기록 작성 기능을 제공하지만 오직 영어로만 사용이 가능하다.[20]

Nuance® Dragon Medical One은 클라우드 기반 호환 음성 인식 솔루션으로 이를 사용하려면 임상의가 음성인식 프로그램, 전용 마이크 등의 장치를 구입하고 사용료를 지불해야 한다. Nuance® Dragon Medical One 역시 Amazon® Transcribe Medical과

마찬가지로 영어로만 사용이 가능하다.[21] 또한 두 프로그램 모두 의료 대화의 인식시에 단어 인식 오류율에서 좋은 결과를 보여주지 못하고 있는 실정이다.[22]

1.1.5 환자 병력 수집의 중요성 및 저해 요소

환자의 정확한 진단과 적절한 치료를 수행하고 임상의 간의 의견 조율을 위해 충분하고 정확한, 포괄적인 의무 기록의 수집 및 작성은 필수적이다.[3,23,24] 진료시간이 연장될 수록 환자에 대한 정확한 진단 및 질병 상태가 개선되고 약물 처방의 수가 감소한다는 사실은 널리 알려져 있다.[25] 의료 현장에서 전자의무기록이 보급된 이후 컴퓨터를 통한 의무기록 작성 등의 업무 부담이 증가함에 따라 실제 환자와 의사간의 대면 진료시간은 지속적으로 감소하고 있다.[26] 게다가, 전자 의무 기록 시스템을 다양한 의료 영역에 적용하면서 임상의의 업무량이 늘어났음이 보고되었으며 이의 결과로 감정적인 피로, 몰개인화, 낮은 개인 성취감의 3가지 요소에 의해 발생하는 업무 관련 증후군인 번아웃이 점점 더 늘어나고 있다. [27,28] 의료진의 번아웃을 해결하기 위한 다양한 방책중의 하나로 음성인식을 이용한 의무기록 작성의 자동화가 제시되고 있으며 다양한 방법이 연구되고 있다.[29]

1.1.6 음성인식 기술의 의무기록 작성에 적용시 장점

개인건강정보 유출 등의 위험성 및 민감정보 제공에 대한 환자의 우려 등으로 인한 데이터 수집의 어려움과 양질의 데이터 셋의

부족은 의사-환자의 진료실 대화 음성인식의 정확도에 대한 연구를 진행하는 데 있어 가장 큰 제한점이다. 이 제한점은 영어 혹은 라틴어 기반 언어가 아닌 다른 언어에서 더 두드러지는데, 이는 AI 번역 시스템 훈련은 훈련 데이터의 양에 크게 의존하지만, 한국어와 같은 사용자가 많지 않은 언어는 훈련 데이터 세트가 절대적으로 부족함에 기인한다.[30] 만약에 클라우드 기반 음성인식 솔루션이 환자 의사간의 대화를 정확하게 인식할 수 있다면 이를 전자의무기록 시스템에 적용하여 자동화된 의무기록의 작성을 가능하게 하고, 이를 통해 임상의의 업무 부담을 줄이고 환자와의 대면 및 문진 시간을 늘리는 데 큰 도움을 줄 수 있을 것이다. 그러나 현재 클라우드 기반 음성인식 솔루션의 의사-환자간 대화에 대한 인식 정확도 및 어떤 음성인식 솔루션이 의료산업에 적용하기 적절한 지에 대한 연구는 존재하지 않고 있는 실정이다. 국내 음성인식 오픈 API와 해외 API의 한국어에 대한 인식률을 비교한 연구가 존재하지만 이는 일반적인 음절이나 문장을 인식한 결과였다.[16,31] 기계 학습을 통한 음성인식 기술을 기반으로 환자의 증상을 기록하고자 하는 시도도 존재하나 미국에서 시행하였으며 아직은 개념적인 연구에 불과하였다.[32] 최근 국내 연구에서는 자연어 처리를 활용하여 응급실 환자의 분류 자동화를 하고자 하는 시도를 하였으나 이는 의무기록 작성과는 관계없는 내용이었다.[33] 이를 종합하여 보았을 때 아직 국내에서 클라우드 기반 음성인식 솔루션의 의무기록 자동화 적용 가능성에 대한 연구는 수행된 바가 없었다.

이러한 연구들에 근거하여 본 연구자는 클라우드 기반 음성인식 솔루션의 정확도를 평가하기 위하여 국내 최초로 실제 의사-환자가 진료실에서 행하는 대화의 녹음 데이터를 수집하여 다양한 클라우드 기반 음성인식 솔루션을 사용하여 인식 작업을 수행 후 각 솔루션 간의 인식 정확도를 비교하였다. 첫번째 연구를 수행 후 단어단위가 아닌 문장단위 인식의 연구 및 녹음 파일의 질이 보장된 데이터셋의 필요성을 절감하여, AI 훈련을 위해 구축되고 공개된 Aihub.or.kr의 원격 의료를 위한 의사-환자간 음성 데이터 세트를 사용하여 다양한 클라우드 기반 음성인식 솔루션의 의료진 대화에 대한 인식 정확도를 비교했다. 본 연구의 결과를 통해서 실제 의료현장에서 환자-의사간의 대화 녹음을 이용한 클라우드 기반 음성인식 솔루션 적용을 통해 의무기록 작성업무 자동화의 가능성에 대해 예측해 볼 수 있을 것이다. 이에 더불어 한국어 같은 사용자가 적은 언어의 음성인식 솔루션을 의료 산업에 적용하기 위한 다양한 방법론 및 해결책을 제시해 줄 수 있을 것으로 기대한다.

1.2 연구 목적

본 연구의 목적은 실제 환자-의사 간 대화에 상용화된 클라우드 기반 음성인식 솔루션을 사용해 인식된 단어 및 문장의 결과를 이용하여 자연어 처리 기술을 적용하여 자동화된 의무기록 작성 가능성을 평가하는 것이다. 이 연구 목적을 달성하기 위해 실제 의사-환자 대화 녹음 데이터 및 AI 훈련 전용 데이터 세트의 음성과 대본을 사용하여 한국어로 된 의학용어와 진료실 대화에 대해서 클라우드 기반 음성인식 솔루션의 음성인식 정확성을 평가 및 분석하였다.

제 2 장. 연구 방법

2.1 의학용어에 대한 음성인식 정확도 분석

2.1.1 대상 환자 포함 및 제외 기준

삼성서울병원 순환기내과 외래를 내원한 환자를 대상으로 진료실 음성 녹음을 수집하였다. 20세 이상의 초진 환자 중 녹음에 동의한 환자를 대상으로 환자-의사간 대화를 녹음하였다. 20세 미만이거나 언어 장애의 존재, 알츠하이머 병 등의 인지장애가 있거나 녹음을 거절한 환자는 수집대상에서 제외되었다. 2021년 4월부터 2021년 7월까지 총 112 건의 환자-의사 대화를 녹음하였다.

2.1.2 녹음 방법

외래진료실에서 음성 녹음은 YouTube® 등의 개인 방송 녹화시에 가장 많이 사용하는 것으로 알려져 있는 PCM-A10 레코더 (Sony, 도쿄, 일본)를 사용하여 시행하였다. (Figure 1)

Figure 1. PCM-A10 recorder (image from Amazon.com)

녹음기를 외래 진료실의 임상의와 환자 사이의 탁자 위에 위치시킨 후 녹음을 진행하였다. 외래 진료실에서 환자-의사간 대화의 녹음은 다음과 같은 순서에 따라 시행하였다. 첫번째, 환자의 개인정보가 녹음 파일에 저장되는 것을 방지하기 위해 녹음 전 환자의 이름, 환자번호, 생년월일 및 사전동의 여부에 대해 다시 확인한 후 녹음을 시작함을 환자에게 알린다. 두번째, 임상의가 직접 녹음기를 작동하여 진료실 대화의 녹음을 시작한다. 셋째, 병력 청취는 임상의가 환자의 주소, 현병력, 과거력, 동반질환, 입원, 수술력 순서로 진행하였다. 병력 청취 중에 발화된 임상의, 환자 및 환자 보호자의 질문과 답변을 모두 녹음하였다. 넷째, 병력 청취가 끝난 후 임상의의 평가, 진단 및 치료에 대한 설명을 마친 후 외래 환자 진료가 끝났음을 환자에게 알린 후에 녹음을 종료하였다.

레코더의 녹음 모드는 압축되지 않은 오디오 정보를 디지털 인코딩할 때 사용하는 방법인 96 kHz/24 bit의 linear pulse code modulated audio 로 설정하였으며 wav. 파일로 저장하였다, 모든 녹음 파일을 연구를 위해 마련한 저장 장치로 이동한 후에 레코더에 저장되어 있는 진료실 녹음 파일들을 즉시 삭제하였다. 연구 흐름도는 Figure 2 로 정리하였다.

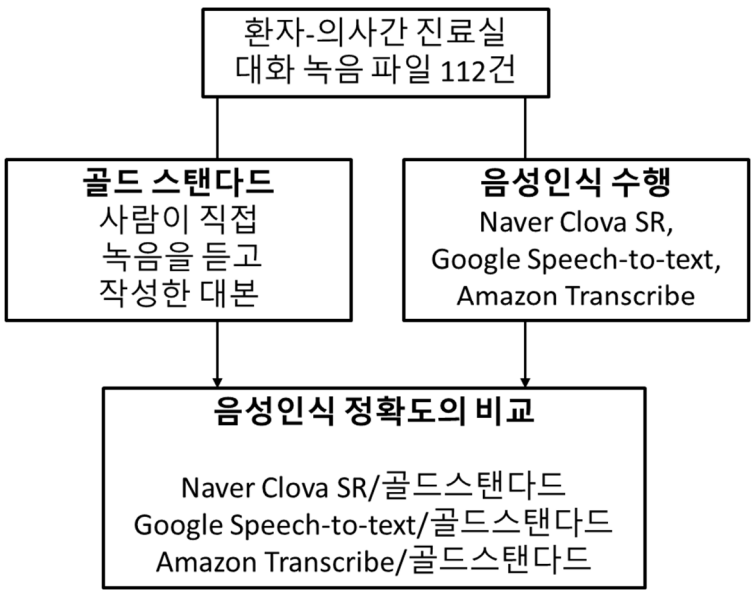

Figure 2. Flowchart of the analysis for medical terminology

2.1.3 골드 스탠다드 및 음성인식 솔루션 선택

음성인식 정확도 분석을 위해 설정한 골드 스탠다드는 각 녹음 파일을 직접 듣고 작성한 대본으로 이 대본은 연구와 독립적인 간호학과 학생과 의사가 녹음을 듣고 작성하였으며 담당 외래 임상의가 다시 검증하였다. Naver Clova SR (Naver, Seongnam,

Korea), Google Speech-to-text (Alphabet Inc., Mountain View, CA, USA), Amazon Transcribe (Amazon, Seattle, WA, USA)의 3가지 클라우드 기반 음성 인식 솔루션을 선정하여 음성인식 작업을 수행하였다. Naver Clova SR은 국내 기업에서 제공하는 음성인식 솔루션이며 나머지 두 솔루션은 해외 업체에서 제공하는 솔루션이다. 모든 음성파일을 각 업체가 제공하는 클라우드에 업로드 한 후 업체가 제공하는 API (Naver Clova SR and Amazon Transcribe) 혹은 Python 3 (Google Speech-to-text)를 이용하여 음성인식 작업을 수행하였다.

2.1.4 음성인식 작업 프로토콜

Naver Clova SR의 음성인식 작업 프로토콜은 다음과 같다. 우선, 네이버 클라우드에 도메인을 생성하고 Clova Speech 빌더를 이용해 사용자 인터페이스 환경에서 음성인식 작업을 요청한다. 두번째, 생성한 네이버 클라우드 도메인에 진료실 대화를 녹음한 wav. 파일을 업로드한다. 세번째, 네이버에서 제공하는 REST API를 사용하여 음성인식 작업을 수행한다. 최종적으로, 음성인식 작업 수행 후 출력된 임상의와 환자의 음성에서 메타데이터를 포함한 텍스트를 Excel 파일로 저장했다. 메타데이터 파일 중 오직 의료진과 환자, 환자 보호자의 음성인식 내용만을 텍스트로 저장하였다. 네이버에서 제공하는 음성인식 사용자 가이드의 URL은 다음과 같다; https://guide.ncloud-docs.com/docs/naveropenapiv3-speech-

recognition-api.

Amazon Transcribe로 음성 인식 작업을 수행하기 위해서는 먼저 특정 언어 모델을 선택해야 하며 본 연구에서는 한국어 KR(ko-KR)을 선택하였다. Amazon S3에서 버킷을 생성한 후 의사-환자 대화 녹음 wav 파일을 버킷에 업로드한다. 다음으로 사용자가 지정한 S3 버킷에 출력 데이터 버킷을 배치하고 음성인식 작업을 수행하였다. 최종적으로 인식처리된 의사 환자간 대화의 대본은 txt파일로 저장하였다. 아마존에서 제공하는 Amazon Transcribe 사용자 가이드의 URL은 다음과 같다; https://ap-northeast-1.console.aws.amazon.com/transcribe/home?region=ap-northeast-1#createJob.

Google Speech-to-text의 음성인식 작업의 수행을 위해 본 연구에서는 Anaconda, Jupyter Notebook, Python 3의 프로그램을 사용하였다. 먼저, Google Cloud 프로젝트에서 Speech-to-text를 활성화하고 인증 환경 변수를 저장하고 wav 파일을 저장하기 위한 새로운 Google Cloud Storage 버킷을 생성한다. 둘째, Jupyter Notebook과 Python 3을 사용하여 클라이언트 라이브러리를 설치하고 Google Speech-to-text에서 제공하는 Python 3 코드로 음성인식 요청작업을 수행한다. 마지막으로 음성인식 작업 후 출력된 텍스트 대본을 txt 파일로 저장하였다. Google에서 제공하는 Google Speech-to-text 사용자 가이드의 URL은 다음과 같다; https://cloud.google.com/speech-to-text/docs/transcribe-

client-libraries. 음성인식 작업 후 출력된 모든 텍스트 대본은 txt 파일로 저장하였으며, 클라우드에 업로드하였던 모든 오디오 파일은 각 음성인식 작업 수행 직후 클라우드에서 제거하였다.

2.1.5 추출 및 주석 정의

모든 녹음파일에 대한 대본 텍스트 파일을 생성한 후 각 대본에서 명사로 된 의학용어를 추출했다. 추출된 각 의학용어는 7개의 분류 중 하나로 각각 정의하였다. 7개의 정의는: 1) 진료과; 2) 질병의 상태 또는 이름을 나타내는 것으로 간주되는 신체적 또는 정신적 특징을 의미하는 질병; 3) 일반적으로 스스로 작용하며 특정한 중요한 기능 또는 특정한 위치를 가진 유기체의 부분을 의미하는 장기; 4) 혈액 또는 의료기기를 사용하여 수행하는 검사; 5) 질병 또는 손상에 대해 환자에게 제공되는 치료; 6) 치료를 위해 사용되는 약품; 7) 제품 또는 특정 의약품의 성분 이름을 나타내는 성분명, 상표명으로 분류하였다. Table 1에서 각각의 예시를 확인할 수 있다. 단어의 길이는 한국어 음절에 따라 계산하였으며 한국어가 아닌 단어도 한국어 발음에 따라서 계산하였다.

Table 1. Classification of medical terms and examples

분류	예시
진료과	순환기내과, 외과, 비뇨의학과, 피부과
질병	흉통, 고혈압, 암, 출혈, 통증
장기	심장, 가슴, 피, 팔
검사	피검사, 심전도, 내시경
치료	수술,입원, 방사선치료, 마취
약품	약,항혈전제,혈압약, 당뇨약
성분명, 상표명	아스피린,오메가쓰리,리피토,클로피도그렐

2.1.6 주석 작업 및 통계처리

의학용어 추출 작업은 두 명의 연구와 독립적인 의사들이 수행하였으며 추출 후 교차 검증을 시행하였다. 단어의 총 개수와 발생 빈도, 단어의 길이, 단어가 한국어인지 외래어인지에 대해서도 평가하였다. 단어의 총 개수는 출현 빈도에 관계없이 한 번이라도 출현한 단어의 개수를 의미한다. 발생 빈도는 반복에 관계없이 대본에 출현한 총 단어 수를 의미한다. 단어의 길이는 음절의 수를 의미한다. 정의에 따라 SR 대본에서 의학용어도 추출작업을 수행하였으며, 각 음성인식 대본의 의학용어의 오타의 종류 및 빈도도 측정하였다. 오타는 세가지 종류로 분류하여 정의하였다: 1) 생략 (단어의 삭제); 2) 철자 오류 (맞춤법이 틀리지만 단어의 원래 의미를 이해할 수 있는 경우); 3) 잘못된 단어: (완전히 다른 의미를

가진 단어).[34] 두 명의 의사가 오타에 대한 주석 작성 작업을 수행하고 교차 확인했다. 각 원본 대본 및 음성인식 대본에 대해 각 클라우드 기반 음성인식 솔루션의 정확도 (원본 대본의 의료 용어 수 당 음성 인식 대본의 의학 용어 수의 비율)를 분석했다. 추가적으로, 단어의 종류, 길이, 외래어 여부에 따른 인식 정확도를 분석했다. 우리는 각각의 대본에 따른 의학용어에 대한 엑셀 파일을 만들었고 단어의 발생 빈도와 오타 횟수를 저장하였다.

통계 작업을 수행하기 위해 Jupyter Notebook 과 Python3의 Pandas 패키지를 사용하여 전체 엑셀 파일 내에 존재하는 단어에 따라 각각 단어의 빈도, 오타 값을 추출하고 정의, 글자수 등의 변수를 추출하여 정확도 분석을 수행하였다. 추출 작업은 다음과 같은 방법으로 수행하였다. 먼저, 각 스크립트에서 의학용어를 엑셀 스프레드시트로 추출하여 저장하였으며 이 과정을 통해 전체 대본의 수인 112장의 스프레드시트를 생성하였다. 두번째, 각 스프레드시트의 의학용어는 파이썬 3 코드를 사용하여 Jupyter Notebook을 통해 추출 및 재배열을 수행하고 빈도 수를 계산하였다. Table 2는 단어 길이에 따른 추출 작업 및 계산 작업을 수행하는 Python 3 코드의 예시이다.

Table 2. Pseudocode of Python 3 for calculating word number by word length

```
# 목록 'total_list'를 8개의 요소로 초기화 (모두 0으로 설정)
total_list = [0 for i in range(8)]
# 목록 ‘naver_list를 8개의 요소로 초기화 (모두 0으로 설정)
naver_list = [0 for i in range(8)]
# 목록 'google_list'를 8개의 요소로 초기화 (모두 0으로 설정)
google_list = [0 for i in range(8)]
목록 'aws_list'를 8개의 요소로 초기화 (모두 0으로 설정)
aws_list = [0 for i in range(8)]

# 'i'를 0부터 'k'까지 반복(exclusive)
for i in range(0, k):
    # 인수가 'filename' 및 'i'인 'SpeechToText' 유형의 새 개체 'a' 생성
    a = SpeechToText(filename, i)

    # 변수 "total"에 인수 "length" 및 "total"을 사용하여 객체 "a"에 호출 메서드 "wordCount"의 결과를 할당
    total = a.wordCount("length", "total")
    # 변수 "naver"에 인수 "length" 및 "naver"을 사용하여 객체 "a"에 호출 메서드 "wordCount"의 결과를 할당
    naver = a.wordCount("length", "naver")
    # 변수 "google"에 인수 "length" 및 "google"을 사용하여 객체 "a"에 호출 메서드 "wordCount"의 결과를 할당
    google = a.wordCount("length", "google")
    # 변수 "aws"에 인수 "length" 및 "aws"을 사용하여 객체 "a"에 호출 메서드 "wordCount"의 결과를 할당합니다
    aws = a.wordCount("length", "aws")

    # 목록 'total'과 'total_list'를 압축하고 반복한 후 해당 요소를 추가하고 결과를 'total_list'에 할당
    total_list =[ (i+j) for i, j in zip(total, total_list) ]
    # 목록 naver과 'naver_list'를 압축하고 반복한 후 해당 요소를 추가하고 결과를 'total_list'에 할당
    naver_list =[ (i+j) for i, j in zip(naver, naver_list) ]
    # 목록 ‘google’과 'google_list'를 압축하고 반복한 후 해당 요소를 추가하고 결과를 'total_list'에 할당    google_list =[ (i+j) for i, j in zip(google, google_list) ]
    # 목록 'aws'과 'aws_list'를 압축하고 반복한 후 해당 요소를 추가하고 결과를 'total_list'에 할당
    aws_list =[ (i+j) for i, j in zip(aws, aws_list) ]
```

```
# key가 "total", "naver", "google" 및 "aws"이고 값이 "total_list", "naver_list", "google_list" 및 "aws_list"인 사전 "final_dict"를 각각 생성
final_dict = {"total" : total_list, "naver" : naver_list, "google" : google_list, "aws" : aws_list}

# 'final_dict'에서 새 pandas 데이터 프레임 'df'를 생성
df = pd.DataFrame(final_dict)

# "total" column의 합, "naver" column의 합, "google" column의 합, "aws" column의 합을 사용하여 "sum_list_length" 목록을 만듭니다
sum_list_length = [df["total"].sum(), df["naver"].sum(), df["google"].sum(), df["aws"].sum()]

# 값이 'sum_list_length'인 인덱스 -1의 'df'에 새 행 추가
df.loc[-1] = sum_list_length

# "df"의 인덱스를 목록 ["1", "2", "3", "4", "5", "6", "7", "8", "sum"]에 설정
df.index = ["1", "2", "3", "4", "5", "6", "7", "8", "sum"]

# Return df
df
```

Table 3은 본 연구에서 사용한 모든 기본 변수를 나타내고 있다. 모든 변수를 엑셀 파일로 저장하였으며 .csv 파일로 변환한 후 R programming을 사용하여 통계분석을 시행하였다.

Table 3. Variables of transcriptions

Variables
Purpose of outpatient clinic visiting
Recording time
Total word count of original script
Total word count recognized by Naver Clova SR
Total word count recognized by Amazon Transcribe
Total word count recognized by Google Speech-to-text
Number of Non-Korean medical terms
Number of medical terms
Length of medical terms
Class of medical terms
Total count of medical terms
Total count of recognized medical terms by Naver Clova SR
Total count of recognized medical terms by Amazon Transcribe
Total count of recognized medical terms by Google Speech-to-text
Total count of missed medical terms by Naver Clova SR
Total count of omitted medical terms by Naver Clova SR
Total count of misspelled medical terms by Naver Clova SR
Total count of wrong words by Naver Clova SR
Total count of missed medical terms by Amazon Transcribe
Total count of omitted medical terms by Amazon Transcribe
Total count of misspelled medical terms by Amazon Transcribe
Total count of wrong words by Amazon Transcribe
Total count of missed medical terms by Google Speech-to-text
Total count of omitted medical terms by Google Speech-to-text
Total count of misspelled medical terms by Google Speech-to-text
Total count of wrong words by Google Speech-to-text

2.2 진료실 대화에 대한 음성인식 정확도 분석

2.2.1 데이터 수집

진료실 대화의 음성인식 정확도를 평가하기 위해 Aihub.or.kr에 연구 목적으로 공개된 데이터세트를 사용하였다. Aihub는 한국정보화진흥원이 구축한 AI 통합 플랫폼으로 이에 접속하여 연구자는 AI 기술과 제품·서비스 개발에 필요한 AI 인프라를 활용하고 참여할 수 있다. Aihub의 원격진료를 위한 의사-환자간 음성 데이터셋은 의료 제공자와 환자의 wav, text, json 파일로 구성되어 있으며 33,000개의 의료 제공자의 텍스트와 150,000개의 환자의 텍스트로 구성되어 있으며 다음의 주소에서 접속 가능하다 (https://aihub.or.kr/aidata/27769, accessed on 27 December 2021). Table 4는 데이터 셋에서 제공하는 메타파일(json)의 예시이다.

Table 4. Example of json metadata file of the dataset

{"기본정보":{"Language":"KOR","Version":"N/A","ApplicationCategory":"
N/A","NumberOfSpeaker":"N/A","NumberOfUtterance":"N/A","DataCateg
ory":"mariaDB",
"RecordingDate":"2021-01-11
16:11:14","FillingDate":"N/A","RevisionHistory":"N/A","Distributor":"Med
iazen"},
"음성정보":{"SamplingRate":"48000","ByteOrder":"N/A","EncodingLaw":"
SignedIntegerPCM","NumberOfBit":"16","NumberOfChannel":"1","SignalT
oNoiseRatio":"N/A"},
"전사정보":{"LabelText":"집안에 고혈압인 분 있나요?"},
"화자정보":{"Gender":"Female","Age":"30~39","Region":"서울/인천/경기","
Dialect":"경상"},
"환경정보":{"RecordingEnviron":"가정","NoiseEnviron":"가정","RecordingD
evice":"노트북"},
"파일정보":{"FileCategory":"Audio","FileName":"HA_0001-1-01-02-
F-05-
A.wav","DirectoryPath":"/nia/HA/data/HA_0001","HeaderSize":"44","File
Length":"3.36",
"FileFormat":"PCM","NumberOfRepeat":"1","TimeInterval":"0","Distance"
:"30"},
"기타정보":{"QualityStatus":"Good"}}

Aihub에서 제공하는 데이터 셋의 구축 목적은 원격 의료의 활성화를 위하여 의료 영역의 음성 인식 성능을 개선하는 것이었다. 총 500개의 임상의의 질문 파일을 데이터 셋에서 추출하여 음성인식 작업을 수행하였다.

2.2.2 음성인식 솔루션 선정 및 수행 프로토콜

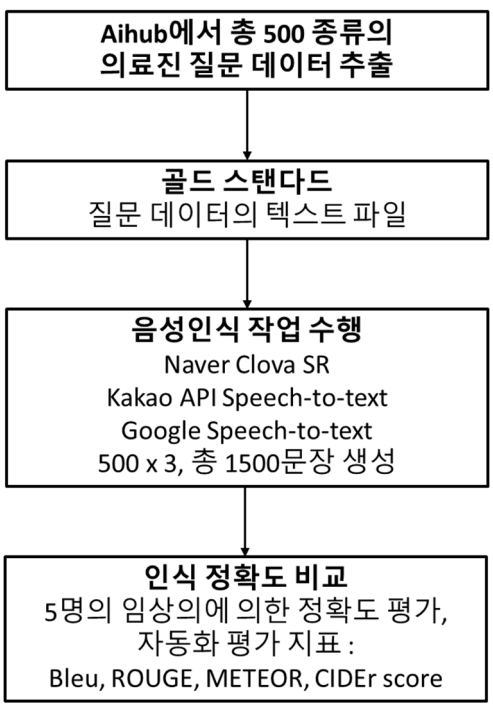

Figure 3. Flowchart of the analysis for medical speech

Figure 3 은 본 연구의 연구 흐름도이다. 골드 스탠다드는 데이터셋에서 제공하는 원본 텍스트 파일로 설정하였다. Naver Clova SR (Naver, Korea), Kakao API Speech-to-text (Kakao Corp., Jeju, Korea), Google speech-to-text (Alphabet Inc., Mountain View, CA, USA) 의 총 세가지 음성인식 솔루션을 연구에

사용하기로 결정하였다. 앞의 두 솔루션은 국내 기업에서 제공하는 것이며 마지막 솔루션은 해외기업에서 제공하는 것이었다. 이전 분석에서 국내 업체의 음성인식 엔진이 해외 업체보다 높은 인식 정확도를 보여주었기 때문에 국내 업체의 음성인식 엔진 2개를 선정하였다.[35]

음성인식 작업은 각 음성인식 솔루션 별로 차이가 있었다. Naver Clova SR 와 Google Speech-to-text의 음성 인식 작업은 앞에서 이미 설명하였다. Kakao API Speech-to-Text는 오직 웹브라우저를 통한 데모버전을 제공하였기 때문에 wav. 파일을 PC 환경에서 재생한 후 Voice Meeter Ver. 1.0.8.2 (VB-Audio Software, V.Burel©)프로그램을 사용하여 가상 마이크를 설정 후 직접 웹브라우저로 음성을 입력하여 음성인식 작업을 수행하였다.

2.2.3 사람에 의한 음성인식 정확도 평가 방법

사람에 의한 음성 인식 정확도는 다음과 같은 방법으로 측정하였다. 연구와 독립적인 임상의 5명을 모집한 후, 3개의 서로 다른 SR 엔진에 의해 500개의 문장에서 음성인식 작업 후 생성된 1,500개의 문장 각각에 대하여 인식 정확도를 측정하였다. 음성인식의 정확도는 1) 완전히 동일한 정확함; 2) 완전히 동일하지 않고 유사하지만 정확한 의미를 추론할 수 있는 유사함; 3) 완전히 다른 의미를 나타내는 틀림의 3가지로 정의하였다. 음성인식에 실패하여 텍스트가 출력되지 않은 경우에는 틀림으로 정의하였다. Table 5는 각 정의에

따른 문장의 예시를 보여주고 있다.

Table 5. Example of each SR case of medical speech

정의	원문	음성 인식 문장
정확함	암 검사를 주기적을 받고 있나요	암 검사를 주기적을 받고 있나요
	헛기침이 멈추지 않나요?	헛기침이 멈추지 않나요.
	배가 이유 없이 가스찬적 있나요?	배가 이유 없이 가스 찬 적 있나요.
유사함	자주 메슥거리세요	자주 매스거리세요
	헛기침이 멈추지 않나요?	학 기침이 멈추지 않아요
	배가 이유 없이 가스찬적 있나요?	배가 이유 없이 가스한 적이 있나요
틀림	최근 물 같은 변을 보시기도 했을까요	최근 물 같은 면을 보시기도 했을까요
	헛기침이 멈추지 않나요?	기침이 멈추지 않아요
	배가 이유 없이 가스찬적 있나요?	배가 이유없이 가스 천장 있나요

2.2.4 음성인식 정확도 자동화 평가 지표

음성인식 정확도의 평가를 위해 BiLingual Evaluation Understudy (Bleu) score, Consensus-based Image Description Evaluation (CIDEr) score, Metric for Evaluation of Translation with Explicit Ordering (METEOR) score, Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score의 4종류의 자동화 평가 지표를 선정하였다.[36-39]

Bleu score 와 ROUGE score는 자연어 처리 작업에 가장 자주 사용되는 지표로 알려져 있다. [40] METEOR metric 은 기계가 생성한 번역과 인간이 생산한 참조 번역 사이의 유니그램 매칭이라는 개념을 기반으로 하여 기계 번역의 성능 평가를 위해 개발된 평가 지표이다.[38] CIDEr score는 비교적 자동화된 평가 지표로, 내용, 문법, 돌출성을 측정하는 데 임의적인 작업을 수행하지 않고 "인간과 유사함"을 기반으로 기계 번역 방식의 객관적인 비교를 평가하도록 개발된 지표이다.[37] Bleu, METEOR, ROUGE score의 결과 값은 항상 0에서 1사이이며 숫자가 1에 가까울수록 높은 성능을 나타낸다. 예를 들어, Bleu score가 0.6보다 크면 기계 번역 품질이 인간보다 우수하다는 것을 의미한다. CIDEr score의 점수 범위는 0에서 10으로 점수가 높을 수록 번역 품질이 우수하다는 것을 의미한다.

모든 평가지표는 Python 3의 pycocoevalcap 패키지를 사용하여 측정하였다. 원본 텍스트 파일의 문장을 기준 값으로 정의하였으며, 4개의 자동화된 평가 지표를 사용하여 각 음성인식 솔루션의 인식

정확도를 비교하였다. Bleu score의 unigram, bigram 및 trigram score를 모두 계산하였다.

2.3 통계분석 방법

연속변수 간의 차이점 분석은 t-test 및 가능한 경우 Mann-Whitney U test를 사용하였으며 데이터는 평균 ± 표준편차나 중앙값과 사분위수를 사용하여 작성하였다. 범주형 데이터는 숫자와 퍼센트로 표시하였으며 χ^2 나 Fisher's exact test를 사용하여 분석하였다. 원본 대본의 단어 수를 기준으로 하여 각 음성 인식 솔루션의 인식 정확도를 비교했다. 또한 각 클라우드 기반 음성인식 솔루션 간 인식 정확도의 차이를 각각 분석하였다. *post hoc* adjustment를 위하여 Bonferroni 검정을 시행하였다. 모든 통계 분석은 Python 3 와 R version 4.1.2 (R Foundation for Statistical Computing, Vienna, Austria)을 사용하여 시행하였다. 모든 검사는 two-tailed로 수행하였으며, P<0.050을 통계적으로 유의미한 것으로 설정하였다.

2.4 연구 윤리

의학용어의 인식 정확도 분석을 위한 연구는 삼성서울병원 기관심의위원회 (2021-03-123-001)의 승인을 받았으며, 관련 환자 전원으로부터 서면에 의한 사전동의를 받았다. 본 전향적

연구는 헬싱키 선언의 원칙에 따라 수행되었다. 의료진 대화의 정확성을 평가 연구의 경우 공개된 데이터 세트를 사용한 연구로 기관심의위원회에 의한 승인이 면제되었다.

제 3 장. 분석 결과

3.1 의학용어 음성인식 정확도 분석 결과

3.1.1 원본 파일의 기본 정보

Table 6 은 원본 파일의 기본 정보를 보여주고 있다. 연구를 수행한 순환기내과 의사의 세부전공이 수술 전 심장평가와 심장질환 예방이었기 때문에 환자의 내원 목적이 수술 전 평가인 경우가 79명이었으며 나머지 환자는 심장질환 진단이나 치료를 위한 방문이었다. 평균 녹음 시간은 328초였다. 대본 당 나타난 평균 의학용어의 종류 수는 25.3개였고, 대본당 총 의학용어의 출현 빈도의 평균은 65.4개였다. 외래어의 출현은 대본당 1.88개로 낮은 빈도를 보여주었다.

Table 6. Baseline characteristics of the original transcriptions

변수	값
방문 목적	
수술 전 평가	79 (70.5)
질환 진단 및 치료	33 (29.5)
대본 당 녹음 시간 (초)	328 ± 161
대본 당 추출된 의학용어 종류의 수	25.30 ± 7.48
대본 당 추출된 총 의학용어 수	65.40 ± 26.89
대본 당 추출된 외래어로 된 의학용어 수	1.88 ± 1.71

Data are presented as number (%) or mean ± standard deviation values.

3.1.2 의학용어의 분류 별 음성인식 정확도 분석

Table 7은 의학용어의 분류에 따른 각각의 음성인식 솔루션의 인식 정확도의 분석 결과를 보여주고 있다. Naver Clova SR 이 가장 높은 인식 정확도를 보여주었으나 절대적인 수치는 비교적 낮은 편이었다(75.1%). 의학용어의 분류에 따른 분석에서, Naver Clova SR 이 의학용어의 분류에 관계없이 가장 높은 인식 정확도를 보여주었으나 질병 (78.9% vs. 53.5% vs. 64.7%; p <0.001, 0.005, 0.008, respectively) 과 장기 (84.1% vs. 57.1% vs. 72.9%; p <0.001, 0.003, 0.700, respectively)분류에 속하는 단어의 인식률에서만 다른 솔루션들과 비교하였을 시 통계적으로 유의미한 차이를 보여주었다.

세 종류의 음성인식 솔루션 모두 진료과 혹은 성분명, 상품명을 나타내는 의학 용어에 대한 인식 정확도가 가장 낮았다. Table 8에서 단어의 분류에 따라 출현 빈도가 상위 5개에 속하는 단어의 인식률을 제시하였다. 연구에 참여한 임상의의 전공은 심장학과 수술 전 심장 기능 평가였던 관계로 대부분의 단어가 심장이나 수술과 연관된 의학 용어였다.

Table 7. Accuracy of speech recognition by classes

변수명	원본	Naver	Google	Amazon
총 단어 수	7,319	5,493 (75.1) *,†	3,726 (50.9)	4,237 (57.9)
단어 정의				
진료과	276	145 (52.5)	141 (51.1)	128 (46.4)
질병	1,343	1,060 (78.9) *,†	718 (53.5)	869 (64.7)
장기	1,935	1,627 (84.1) *,†	1,104 (57.1)	1410 (72.9)
검사	1,160	799 (68.9)	601 (51.8)	587 (50.6)
치료	1,251	944 (75.5)	522 (41.7)	605 (48.4)
약품	1,139	840 (73.7)	569 (50.0)	589 (51.7)
성분명, 상품명	215	79 (36.7) *	71 (33.0)	49 (22.8)

Data are presented as number (%).

*, $p < 0.05$ compared to Google; †, $p < 0.05$ compared to Amazon

Table 8. Top 5 most frequent word according to classes

정의	의학용어
진료과, 276	순환기내과, 57 (20.7) 정형외과, 37 (13.4) 응급실, 27 (9.8) 내과, 26 (9.4) 외과, 19 (6.9)
질병, 1,343	고혈압, 141 (10.5) 부정맥, 116 (8.6) 증상, 89 (6.6) 당뇨, 88 (6.6) 협심증, 83 (6.2)
장기, 1,935	심장, 643 (33.2) 혈관, 114 (5.9) 가슴, 106(5.5) 피, 103 (5.3) 신장, 80 (4.1)
검사, 1,160	혈압, 375 (32.3) 심전도, 154 (13.3) 심초음파, 101 (8.7) 피검사, 92 (7.9) 초음파, 73 (6.2)
치료, 1,251	수술, 852 (68.1) 입원, 146 (11.7) 스텐트, 50 (4.0) 시술, 40 (3.2) 항암, 24 (1.9)
약품, 1,139	약, 754 (66.2) 혈압약, 129 (11.3) 고지혈증약, 47 (4.1) 당뇨약, 35 (3.1)

	고혈압약, 29 (2.5)
성분명, 상품명, 215	아스피린, 72 (33.5) 비타민, 11 (5.1) 플라빅스, 11 (5.1) 와파린, 10 (4.7) 인슐린, 8 (3.7)

Data are presented as number (%).

3.1.3 의학용어의 길이 별 및 외래어 음성인식 정확도 분석

의학용어의 길이에 따른 정확도를 분석하였을 때 Naver Clova SR은 다른 음성인식 솔루션과 비교하였을 때 3자 미만의 단어에서 높은 인식 정확도를 보여주었으며 단어 길이가 길어지면 정확도의 차이가 줄어들었다 (Table 9). Naver Clova SR은 2글자 이내 의학용어 인식에서 80% 이상의 정확도를 보여주었다. 외래어로 된 의학용어의 인식 정확도 역시 Naver Clova SR (58.6%)이 Google Speech-to-text (35.5%), Amazon Transcribe (30.9%)에 비해 높은 인식 정확도를 보여주었으나 통계적으로 유의미함을 보여주지는 못하였다.

Table 9. Accuracy of speech recognition by word length and non-Korean

변수	Total	Naver	Google	Amazon
단어 길이				
1	1,108	894 (80.7) *,†	542 (48.9)	658 (59.4)
2	3,695	3,049 (82.5) *,†	1,874 (50.7)	2,387 (64.6)
3	1,468	955 (65.1)	749 (51.0)	740 (50.4)
4	659	408 (61.9)	337 (51.1)	305 (46.3)
5	325	171 (52.6)	183 (56.3)	119 (36.6)
6	61	15 (24.6)	39 (36.9)	27 (44.3)
7	1	1 (100.0)	1 (100.0)	1 (100.0)
8	2	1 (50.0)	1 (50.0)	0
외래어	459	269 (58.6)	163 (35.5)	142 (30.9)

Data are presented as number (%).

*, $p < 0.05$ compared to Google; †, $p < 0.05$ compared to Amazon

3.1.4 Google Speech-to-text 와 Amazon Transcribe간 인식 정확도 비교

Google Speech-to-text와 Amazon Transcribe의 인식 정확도를 비교하였으며, 전반적인 인식 정확도는 Amazon Transcribe가 유의미하게 높았으나 인식 정확도 퍼센테이지의 차이는 크지 않았다 (57.9% vs. 50.9%; P <0.001). Amazon Transcribe는 질병으로 분류된 의학 용어의 인식에서 구글에 비해 더 높은 정확도를 보여주었으나 (64.7% vs. 53.5%; P = 0.008) 성분명, 상품명의 인식에서 낮은 인식 정확도를 보여주었다 (22.8% vs. 33.0%; P = 0.010). 또한, 3글자 미만의 의학용어 인식에서는 Amazon Transcribe가 높은 정확도를 보여주었으나, 4글자 이상의 단어는 Google Speech-to-text의 정확도가 높았다.

3.1.5 민감도 분석

민감도 분석의 시행을 위해서 가장 출현 빈도수가 높은 10개의 단어를 추출하여 인식 정확도의 차이를 분석하였으며, Naver Clova SR의 음성 인식률이 다른 두 음성 인식 솔루션에 비해 상당히 높은 정확도를 보여주었으나, Google Speech-to-text와 Amazon Transcribe의 비교에서는 통계적으로 유의미한 차이를 보여주지 못하였다 (Table 10). 이에 더하여 총 100회 이상 추출된 의학 용어를 대상으로 하여 음성 인식 정확도를 비교 분석하였다. 총 13종류의 단어가 100회 이상 추출되었으며, Naver Clova SR은 9개의 단어에서, Google Speech-to-text는 1개의 단어에서, and Amazon Transcribe는 4개의 단어에서 80% 이상의 인식 정확도를 보여주었다. (Table 11). 심장초음파를 제외한 모든 단어의 길이는 3글자 이내였다.

Table 10. Recognition accuracy for top 10 most frequent words

	전체 수	Naver	Google	AWS
전체 단어	3473	81.7 (2837) *,†	53.2 (1847)	62.6 (2175)
수술	852	76.1 (648)	42.4 (361)	48.0 (409)
약	754	81.7 (616)	54.4 (410)	58.5 (441)
심장	643	93.6 (602)	65.9 (424)	84.4 (543)
혈압	424	83.5 (354)	46.0 (195)	70.3 (298)
심전도	154	44.2 (68)	74.7 (115)	42.2 (65)
입원	146	87.7 (128)	58.9 (86)	64.4 (94)
고혈압	141	95.7 (135)	54.6 (77)	87.2 (123)
혈압약	129	66.7 (86)	45.0 (58)	42.6 (55)
부정맥	116	77.6 (90)	57.8 (67)	45.7 (53)
혈관	114	96.5 (110)	47.4 (54)	82.5 (94)

Data are presented as % (n).

*, $p < 0.05$ compared to Google; †, $p < 0.05$ compared to Amazon

Table 11. Accuracy rate over than 80% among words appeared over than 100 times according to SR platforms

	단어	원본 단어 수	인식 단어 수	비율
Naver	혈관	114	110	96.5
	고혈압	141	135	95.7
	심장	643	602	93.6
	피	104	97	93.2
	가슴	106	97	91.5
	심장초음파	101	90	89.1
	입원	146	128	87.7
	혈압	424	354	83.5
	약	754	616	81.7
Google	심장초음파	101	88	87.1
Amazon	고혈압	141	123	87.2
	가슴	106	92	86.8
	심장	643	543	84.4
	혈관	114	94	82.5

3.1.6 음성인식 솔루션 별 오타 발생 비교 분석

Table 12는 음성인식 솔루션에 따른 오타의 유형을 추가로 분석하여 정리한 표이다. Google Speech-to-text, Amazon Transcribe와 비교하였을 때 Naver Clova SR 은 틀린 단어로 인식하는 비율이 높았으나 (69.0% vs. 34.2% vs. 30.8%, P <0.001, respectively), Naver Clova SR과 Google Speech-to-text의 비교에서는 통계적으로 유의미한 차이를 보여주지 못하였다 (P =

0.180). Google Speech-to-text와 Amazon Transcribe는 Naver Clova SR과 비교 시 의학용어의 생략이 높은 비중으로 발생하였다 (13.5% vs. 61.0% vs. 55.6%, P <0.001, respectively). 또한 100회 이상 추출된 의학용어 중 인식 정확도가 50% 미만인 단어를 분석하였으며 Naver Clova SR에서는 심장초음파의 1개 단어만이 50% 미만의 인식 정확도를 보여주었다 (Table 13).

Table 12. Error rate according to the classification of typos

	Naver	Google	Amazon
오타	24.9 (1826) *,†	49.1 (3593)	42.1 (3082)
대본 당 오타 수	16.3 (±7.5) *,†	32.1 (± 13.6)	27.5 (± 12.2)
오타 종류			
생략	13.5 (246) *,†	61.0 (2191)	55.6 (1714)
철자 오류	18.1 (330) *,†	4.8 (173)	13.6 (420)
틀린 단어	69.0 (1260) †	34.2 (1229)	30.8 (948)

Data are presented as % (n) or mean (± standard deviation).

*, p < 0.05 compared to Google; †, p < 0.05 compared to Amazon

Table 13. Accuracy rate less than 50% among words appeared over than 100 times according to SR platforms

	단어	원본 단어 수	인식 단어 수	비율
Naver	심전도	154	68	44.2
Google	혈관	114	54	47.4
	피	104	48	46.2
	혈압	424	195	46.0
	혈압약	129	58	45.0
	수술	852	361	42.4
Amazon	수술	852	409	48.0
	부정맥	116	53	45.7
	혈압약	129	55	42.6
	심전도	154	65	42.2

3.2 진료실 대화 음성인식 정확도 분석 결과

3.2.1 원본 파일의 기본 정보

Table 14는 원본 음성 파일의 기본 정보에 대하여 보여주고 있다. 모든 파일은 pulse code modulated audio, 48 kHz/16 bit로 녹음되었다. 평균 파일의 길이는 3.16 (±0.68)초였으며 평균 글자 수는 4.32 (±1.42)개였다.

Table 14. Baseline characteristics of original speech files

파일 포멧	PCM
헤더 사이즈	44
비트 레이트	16
샘플링 레이트	48000
평균 녹음 시간 (초)	3.16 (±0.68)
평균 단어 수 (개)	4.32 (±1.42)

3.2.2 임상 의사에 의한 음성인식 정확도 평가

임상 의사의 판단에 의한 인식 정확도의 평가 결과를 Table 15에 정리하였다. 문장 인식 실패는 Naver Clova SR에서만 9건이 발생하였다. 정확함과 유사함를 합한 인식 정확도를 비교하였을 때 Naver Clova SR이 Kakao API Speech-to-text, Google Speech-to-text에 비해서 높은 인식 정확도를 보여주었다 (94.7% vs. 83.8% vs. 76.7%; p <0.001, respectively). 이에 더하여 음성인식이 정확함으로 판정된 비율도 Naver Clova SR 이 가장 높았다 (89.7% vs. 77.2% vs. 66.0%; p <0.001, respectively). Kakao API Speech-to-text와 Google Speech-to-text의 인식 정확도를 비교하였을 시 Kakao API Speech-to-text가 더 높은 인식 정확도를 보여주었다 (83.8% vs. 76.7%; p <0.001). Google Speech-to-text는 Naver Clova SR (23.3% vs. 5.3%; p <0.001), Kakao API Speech-to-text (23.3% vs. 16.2%; p <0.001)와 비교하여 틀린 문장으로 인식하는 비율이 제일 높았다.

Table 15. Recognition accuracy judged by clinicians

	Naver	Kakao	Google
정확함 혹은 유사함	2367 (94.7) *,†	2095 (83.8)	1916 (76.7)
정확함	2243 (89.7) *,†	1930 (77.2)	1649 (66.0)
유사함	124 (5.0) *,†	165 (6.6)	267 (10.7)
틀림	133 (5.3) *,†	405 (16.2)	584 (23.3)

Values are n(%), p-values are adjusted with bonferroni correction

*, $p < 0.05$ compared to Google; †, $p < 0.05$ compared to Kakao

3.2.3 자동화 측정 지표를 이용한 음성인식 정확도 평가

Table 16은 자동화 측정 지표를 이용하여 각 음성인식 솔루션의 인식 정확도를 평가한 결과를 보여주고 있다. Naver Clova SR은 다른 두 음성인식 솔루션과 비교하였을 시 가장 높은 Bleu-1 score를 보여주었다 (Naver, 0.654 vs. Kakao, 0.578 vs. Google, 0.535; p <0.001 respectively). Naver Clova SR의 Bleu-2 (0.557 vs. 0.463 vs. 0.418; p <0.001; respectively)와 Bleu-3 score (0.389 vs. 0.306 vs. 0.262; p <0.001; respectively)도 다른 두 솔루션에 비해 높았다. Kakao API Speech-to-text와 Google Speech-to-text의 비교에서는 Bleu-1, 2, 3 scores 모두 Kakao API Speech-to-text가 유의미하게 높았다. CIDEr score, METEOR score, ROGUE score 모두 같은 결과로 Naver Clova SR 이 가장 높은 인식 정확도를 보여주었다.

Table 16. Recognition accuracy by automated methods

	Naver	Kakao	Google
Bleu-1	0.654 (0.199) *,†	0.578 (0.216)	0.535 (0.238)
Bleu-2	0.557 (0.280) *,†	0.463 (0.293)	0.418 (0.323)
Bleu-3	0.389 (0.356) *,†	0.306 (0.328)	0.262 (0.323)
ROUGE	0.661 (0.192) *,†	0.592 (0.208)	0.553 (0.231)
CIDEr	4.18 (2.52) *,†	3.39 (2.32)	3.02 (2.34)
METEOR	0.600 (0.125) *,†	0.560 (0.089)	0.542 (0.115)

Values are mean(standard deviation), p-values are adjusted with bonferroni correction,

*, $p < 0.05$ compared to Google; †, $p < 0.05$ compared to Kakao

제 4 장. 고찰

4.1 연구의 주요 결과

의학용어에 대한 음성인식 정확도 분석의 주요 결과는 다음과 같다: 1) 3종류의 클라우드 기반 음성인식 솔루션 중 국내 기업에서 제공하는 솔루션이 가장 높은 인식 정확도를 보여주었다; 2) 현재 서비스 중인 클라우드 기반 음성인식 솔루션은 의무기록 작성에 적용하기에는 비교적 낮은 의학용어에 대한 인식 정확도를 보여주었다; 3) 각각의 음성인식 솔루션 별로 강점을 가지는 음성인식 영역이 존재한다.

진료실 대화에 대한 음성인식 정확도 분석의 주요 결과는 다음과 같다: 1) 세 종류의 클라우드 기반 음성인식 솔루션 중 Naver Clova SR이 Kakao API Speech-to-text 및 Google speech-to-text와 비교하였을 때 가장 높은 음성 인식 정확도를 보여주었다; 2) 국내 기업에서 개발한 음성인식 솔루션에 해외 기업의 솔루션과 비교하여 더 높은 음성인식 정확도를 보여주었다; 3) 임상의사에 의한 음성인식 정확도 평가나 자동화 평가 지표를 사용한 정확도 평가 모두 비슷한 결과를 보여주었다.

본 연구의 결과는 현재 상용화된 클라우드 기반 음성인식 솔루션을 실제 진료 현장에서 의무기록 작성에 적용하기 위해서는 아직 기술 수준이 충분하지 않음을 보여주었으며 앞으로 의료용 음성인식 솔루션의 개발 및 연구에 대한 방향성을 제시하고 있다.

4.2 음성인식 기술과 의료 산업

의료 산업 영역에서 음성인식 기술은 현재 코로나 바이러스로 자가 격리된 환자를 대상으로 한 원격 증상 확인부터 응급실에 방문한 환자의 분류에 이르는 다양한 영역에서 사용되는 중이다.[33,41] 전자의무기록 시스템이 병원에 보급된 이후 임상의들은 진료기록의 문서 작성에 가장 많은 업무시간을 소모하는 것으로 알려져 있다.[42] 음성인식 기술을 의무기록 작성에 적용할 시에 의사들이 의무기록 작성에 소모하는 시간이 감소함은 이미 잘 알려져 있으며, 이를 통해 의료의 질 향상, 생산성 재고 및 비용-효율성에 긍정적인 영향을 끼칠 수 있음을 다양한 연구에서 보고하였다.[1,6,43]

그러나, 실제 전자의무기록 시스템에 의무기록 작성시 음성인식 기능을 적용하였을 시에 실제적으로 이점에 발생한다는 근거를 제시하는 연구는 아직 부족하며, 음성인식 시스템을 의무 기록 작성에 적용할 시 선행되는 음성인식 시스템 사용 방법에 대한 교육, 기존 전자 의무 기록 시스템과의 상호 운용성의 문제로 인하여 명확한 이익에 대한 근거도 아직까지는 중립적이다.[1,44] 현재 대부분의 의료용 음성인식 시스템은 영상의학과, 병리학 보고서와 같은 의료진의 판독의 기록 영역에서 사용 중에 있으며, 의사-환자 간의 대화에 적용하여 의무기록을 작성하는 영역에 대한 음성인식 솔루션은 부족한 상황이며 아직 개념증명 연구정도만이 존재한다.[1,18]

이에 더하여 음성 인식 시스템을 전자 의무 기록 시스템에 적용할 때 발생하는 추가적인 비용은 상기한 음성인식 솔루션을 이용한

의무기록 작성의 임상적인 장점을 상쇄시킬 수 있다.[6]

4.3 클라우드 기반 음성인식 솔루션의 장점

기존의 음성인식 시스템과 비교하여 클라우드 기반의 음성인식 솔루션은 다양한 이점을 가지고 있다. 진료실 대화를 수집하여 클라우드 기반 음성 인식 솔루션을 이미 사용중인 의무기록 작성 시스템에 적용할 수 있다면 기존에 존재하던 음성인식 시스템을 새롭게 적용시에 추가되는 비용 부담과 추가적인 사용자 교육의 필요성을 줄일 수 있다. 클라우드 기반 음성인식 솔루션은 인공지능과 클라우드 컴퓨팅 기술의 발전에 힘입어 괄목할 만한 발전을 이루었으며, 음성인식 시스템의 개발 기간, 적용 시 요구되는 비용을 절감시켜 주었다.[16] 또한 클라우드 기반 음성인식 솔루션의 특성 중 하나인 다른 시스템 적용의 간편함, 즉 상호 운용성은 병원에서 사용중인 전자의무기록 시스템에 음성인식 기능을 적용시 요구되는 시스템 개발 및 구현에 소요되는 시간과 비용을 줄일 수 있으며, 최종적으로 Covid-19로 인한 재앙적인 전염병 상황에서 의사의 번아웃 증후군을 줄이는 역할을 할 것으로 기대되고 있다.[7]

본 연구에서는 국내에서 널리 사용되는 클라우드 기반 음성인식 솔루션 간의 의학용어에 대한 인식 정확도를 비교분석 하였으며 국내 기업(Naver Clova SR)이 구축한 음성인식 솔루션이 나머지 2가지 솔루션과 비교하여 가장 높은 음성인식 정확도를 보여주었다. 이전 연구에서도 한국의 국내 기업들이 국제 기업들에 비해 한국어의

음성인식에서 더 큰 성과를 달성했다는 결과를 보고한 적이 있다.[16,31] AI의 성능은 훈련용 데이터베이스의 품질과 양에 크게 좌우됨을 고려하였을 시, 국내 기업에서 제작한 음성인식 솔루션은 모국어로 된 데이터 수집이 용이하며 이를 토대로 상대적으로 높음 음성인식 정확도를 보임을 유추할 수 있다.

4.4 클라우드 기반 음성인식 솔루션의 의학용어 인식 한계점

연구 결과에서 클라우드 기반 음성인식 솔루션의 의학용어 인식 정확도는 상대적으로 낮은 80% 미만의 결과를 보여주었으며, 이는 실제 의사-환자의 대화의 인식에 적용을 위해서는 현재 음성인식 솔루션의 인식 정확도가 향상되어야 함을 시사한다. 현재 상용화된 클라우드 기반 음성인식 솔루션의 낮은 의학용어 인식 정확도에 대해서는 다음과 같은 원인들을 고려할 수 있다.

첫번째로 외래 진료실은 양질의 의사-환자 간 음성 데이터를 수집하기에 적합한 환경이 아니라는 점이다. 진료 중 이학적 검사 시행, 혹은 문진 중 컴퓨터 사용, 기구의 사용에 의해 발생하는 소음이나 간호사나 간병인 등 진료 중 참여하는 환자와 의사 이외의 인원들이 발화하는 음성들로 인하여 양질의 녹음 데이터를 수집하는 데에 어려움이 존재한다. 게다가, 의사-환자 간의 대화는 단순한 질문과 응답의 반복으로 구성되어 있지 않으며, 대화 중에 언제나

환자와 의사 양쪽에 의해 질문, 답변 중에 중단, 간섭, 새로운 질문이 발생할 수 있다. 그러나 이들 문장의 발화자를 구분하기 위한 화자 인식 기능은 현재 상용화 된 클라우드 기반 음성인식 솔루션에서 한계가 있다.

의학용어 자체의 특성도 낮은 음성인식 정확도에 대한 원인이 될 수 있다. AI의 성능 향상을 위해는 수많은 훈련용 데이터베이스가 필요하지만 일반적인 대화문에서는 의학용어가 사용되는 빈도가 상대적으로 낮다. 이런 측면을 고려하였을 때 의학용어의 음성인식 훈련은 일반 언어의 훈련에 비해 난이도가 높은 작업임을 유추할 수 있다. 이에 더하여 전반적인 음성인식의 정확도는 Naver Clova SR이 가장 높았으나, 다른 음성 인식 솔루션들도 긴 단어나 특정 단어의 인식에서 더 큰 정확도를 보여주었다. 예를 들어 Naver Clova SR은 “순환기내과” 라는 단어를 “술 한잔”으로 “심방”을 “신방”으로 인식하는 경우가 많이 발생하였으며, 이는 Naver Clova SR의 AI 훈련 시에 사용한 데이터셋에 관련 의학용어가 많이 포함되지 않았음을 반영한 것일 가능성이 있다. 이의 결과로 “순환기내과”는 전체 대본에서 총 57회 등장하였으나 Google Speech-to-text는 19회, Amazon Transcribe는 20회 인식한 반면에 Naver Clova SR은 오직 2건만 인식하였다. 이러한 결과는 딥러닝 알고리즘 훈련 데이터베이스에 내재된 편향성을 반영하는 것이라 할 수 있다.[45] 이와 같은 점을 고려하였을 때 의료 산업에 음성 인식 솔루션을 접목하기 위해서는 Amazon Medical Transcribe와 같은 의료 서비스용으로

구축된 데이터셋으로 훈련을 시행한 의료 서비스 전용으로 구축된 음성인식 솔루션이 필요함을 알 수 있다.[20]

또한 각각 다른 서비스 제공자의 클라우드 기반 음성인식 솔루션별로 어휘나 단어 길이에 따라 각각 높은 인식 정확도를 보이는 분야가 존재하였다. 이를 고려하였을 시 각 솔루션이 강점을 보이는 인식 분야의 알고리즘을 결합할 수 있다면 의료용 음성인식 솔루션의 성능을 향상시킬 수 있을 것이다. 앞으로 클라우드 기반 음성인식 솔루션을 의료 산업에 도입하기 위해서는 이러한 사항들을 반드시 고려해야 할 것이다.

비록 의학용어에 대한 음성 인식 정확도의 분석에서 현재 상용화된 음성인식 솔루션의 인식 정확도는 한계점을 보여주었으나, 진료실 대화의 인식을 시행한 두번째 분석에서는 음성인식 정확도가 만족할 만한 결과를 보여주었다. 두 분석 사이의 가장 큰 차이는 각각 서로 다른 데이터셋, 실제 진료실에서 녹음한 음성 데이터와 인공지능 훈련을 위해 제작된 전용 데이터를 사용했다는 점이다. 또한 첫번째 분석은 의학용어에만 초점을 맞추어 분석하였으나, 두번째 분석은 단순한 의료진의 질문문의 인식에 초점을 맞추어 분석을 시행하였다. 이와 같은 측면을 고려하였을 때 현재 상용화된 클라우드 기반 음성인식 솔루션을 의무기록 작성을 위해 직접 의료서비스에 적용하기에는 중대한 한계가 존재하며, 이의 성능 개선을 위해 의료 대화 인식을 위해 구축한 전용 AI 훈련 데이터 세트가 더 필요하다는 것을 유추할 수 있다.

추가적으로 한국어와 라틴어 기반 언어의 차이를 고려해야 한다. 일반적으로 의사-의사 또는 의사-간호사 같은 의료진의 대화는 영어, 라틴어 기반의 원어로 된 의학용어를 사용한다. 그러나 환자-의료진 간의 대화에서는 CT, 아스피린 같은 단순하거나 유명한 단어를 사용하는 경우 이외에는 대부분 한국어로 번역된 의학용어를 사용한다. 따라서 의료 서비스 영역에서 사용 가능한 클라우드 기반 음성인식 솔루션을 구축하기 위해서는, 특히 라틴어 기반이 아닌 언어를 사용하는 국가에서는 모국어로 된 의학용어의 인식만이 아닌 의료진 간의 대화에서 일반적으로 사용하는 원어로 된 의학용어의 인식 정확성 또한 보장되어야 한다. 게다가, 한국어는 라틴어 기반 언어들과는 확연히 다른 접착어이다; 한국어의 단어들은 의미를 결정하기 위해 의미상으로 관련이 있는 두가지 이상의 형태소를 포함하여 구성되는 경우가 많으며, 이러한 측면은 음성인식 솔루션의 한국어 인식 정확도를 향상시키는 훈련을 어렵게 만드는 요인 중 하나이다.[46] 이와 같이 세계적으로 정보기술산업이 발달한 나라중의 하나인 한국에서도 음성인식 기술을 의료 산업에 사용하기 위한 기술력이 아직은 부족함을 알 수 있다. 이와 같은 측면을 고려하였을 시 서로 다른 언어 시스템을 사용하는 많은 국가들은 음성 인식 기술의 발전에 의한 이득을 동등하게 향유할 수 없을 것이다.

이러한 사실은 AI에 의한 의료의 발전이 건강 불평등을 개선하는데 기여해야 한다는 현대 의학의 중요한 화두와 연계된다.[19] 환자의 치료 질 향상 및 의사-환자 관계 개선에 대해 AI의 기여에 의한 의료의

발전은 이러한 상황에서, 특히 선진국이 아닌 나라에서 균등하게 향유되지 않을 가능성이 높으며, 이에 대해 진지한 논의가 필요하다.[47,48] 다른 언어에서 영어로의 AI 번역 알고리즘을 향상하는 것도 이 문제에 대한 또 다른 해답일 수 있다. AI에 의한 정확한 영어 번역이 보장된다면 현재 영어로만 이용 가능한 의료 음성인식 시스템을 다른 언어를 사용하는 환자에게도 사용할 수 있을 것이다.

4.5 음성인식 솔루션 간 의학용어 인식 성능 차이

연구 결과에서 흥미로운 발견 중 하나는 국내 기업과 국제 기업에서 제공하는 음성인식 솔루션 간 오타발생의 경향 차이이다. Naver Clova SR은 틀린 단어의 비중이 더 높았으며, 나머지 두 개의 솔루션은 생략의 비중이 더 높았다. 이 경향성은 각 회사의 음성 인식 알고리즘의 고유한 특징을 반영하는 것이라 할 수 있다. 부정확한 음성을 인식할 시 Naver Clova SR은 유사한 다른 단어로 대체하려는 경향을 보였으나 다른 솔루션들은 해당 단어를 그냥 건너뛰었다고 추정할 수 있다. 의료 서비스 분야에서는 명확하지 않은 단어를 다른 단어로 대체하는 것이 그냥 건너뛰는 것보다 더 심각한 문제를 야기할 수 있다. 예를 들면, 심장, 신장처럼 한 글자의 차이가 완벽히 다른 의미를 지닐 수 있다. 앞으로 의료 서비스를 위한 음성인식 솔루션을 구축시에는 이와 같은 점 또한 고려해야 할 것이다.

4.6 임상의와 자동화 측정 지표에 의한 음성인식 정확도 분석

3개의 SR 엔진의 진료실 대화에 대한 인식 정확도를 분석하기 위해서 본 연구에서는 5명의 임상의에 의한 평가와 4개의 자동화 측정 지표를 사용하였다. 자동화 측정 지표는 AI 음성인식 및 번역 성능을 평가하는데 널리 사용되는 방법이지만, 지금까지 진료실 대화 또는 의무기록에 이 방법론을 적용하는 것은 연구된 바가 없다. 본 연구의 결과에서 음성인식 솔루션의 정확도 측정 시, 임상의의 평가와 자동화 측정 지표 사이에서 유사한 경향성을 보여주었다. 이를 고려하였을 시, Bleu, ROUGE, METEOR, CIDEr score 등의 자동화 측정 지표는 진료실 대화의 정확도 측정 지표로서 신뢰할 수 있는 방법임을 알 수 있었다.

4.7 전자의무기록 작성용 음성인식 솔루션의 구축을 위한 제언

연구 결과에 비추어 볼 때, 전자의무기록 작성에 적용 가능한 음성인식 시스템을 구축 및 개선하기 위해 몇 가지 필요한 요소를 알 수 있었다.

첫번째, 진료실 대화에 대한 음성인식 기능을 향상시키기 위해서는 진료과와 의약품 명칭의 인식 정확도를 향상할 필요가 있다. 의학

용어에 대한 인식 정확도를 분석했을 시, 이 두 분류에 포함된 단어들이 다른 분류의 단어들에 비해 낮은 정확도를 보여주었다. 이 두 분류의 단어는 환자의 진단과 치료방법에 대해 알기 위해서 가장 중요한 의학용어 분류이다. 진료과를 통하여 우리는 환자의 질병 및 방문 목적에 대해 유추할 수 있다. 또한, 약품명은 환자의 질병과 치료 내용을 확인하는데 필수적이다. 특히 같은 성분의 약품이라도 복제약 등의 존재로 인하여 다양한 상품명이 존재하며, 대부분의 상품명은 성분명을 따라서 명명되는 관계로 대부분 외국어, 특히 영어로 유사하게 이름 지어지는 경우가 많다. 그러므로 더욱 발전된 기능을 가진 음성을 통하여 의무기록 작성을 가능하게 하는 의료용 음성인식 솔루션의 구축을 위해서는 이 분류에 해당하는 단어들의 인식 능력을 향상시킬 수 있는 훈련이 필수적이다.

두번째, 높은 인식 정확도를 보인 의학용어는 대부분 3자 미만의 길이였다. 앞으로 구축될 음성인식 솔루션은 3자 이상의 긴 단어에 대한 인식 정확도의 개선이 필요하다. 흉골-쇄골-유돌기근 같은 단어의 예처럼 조합을 통한 수많은 새로운 단어의 생성은 의학용어의 독특한 특성중의 하나이며, 이 특성으로 인하여 의학용어는 비교적 긴 단어의 비중이 높다. 따라서, 의료 산업에 적용 가능한 음성인식 시스템의 발전을 위해서는 긴 단어에 대한 정확도를 향상시키는 것이 매우 중요하다.

세번째, 화자의 구분 기술이 필수적이다. 현재 대부분의 음성인식 솔루션은 화자의 구별 기능을 제공하지 못하고 있다. 진료실 대화는

임상의와 환자 간 상호 작용을 통해 이루어지는 관계로 화자의 구분 기능의 부재는 음성인식 시스템을 의료 산업에 적용하는 데 있어 주요한 한계점 중 하나라고 할 수 있다. 임상의의 질문의 의미에 따라 환자들의 답변의 의미가 달라진다. 만약에 음성인식 솔루션에서 화자의 구분 기능의 추가 혹은 향상이 힘들다면 외래에서 대화를 녹음할 때 다채널 마이크 혹은 화자별로 별개의 마이크를 사용하거나 다채널 마이크를 이용하여 녹음을 수행하는 방법을 고려해 볼 수 있을 것이다.

마지막으로 전자의무기록 시스템에서 의무기록 작성에 적용 가능한 음성인식 솔루션의 성능 향상을 위해서는 단순한 문장의 모음이 아닌, 환자와 의사간의 대화 혹은 원격진료를 상정한 질문 답변 형식의 훈련용 데이터셋을 구축해야 한다. 질의 응답 형식의 문장구조는 AI 훈련시에 다음 답변 혹은 질문을 유추하는 데에 중요한 단서를 제공해 줄 수 있다. 예를 들어, 임상의가 혈당이 상승했다고 환자에게 이야기를 한다고 가정할 시, 저혈당 환자에서는 개선된 결과를 의미하는 소견이지만 당뇨병 환자에서는 악화된 결과를 의미한다. 음성인식 AI 훈련시에 문장의 의미를 정확하게 판단하고 인식하기 위해서는 이와 같이 의사, 환자 간의 상호 대화의 맥락을 판단할 수 있는 데이터 셋이 필요하다. 본 연구에서는 적어도 환자-의사가 3회에서 5회 정도 질문과 대답을 주고받는 내용으로 설정된 데이터 셋을 구축해야 할 것으로 제시한다. 실제 진료실 대화는 그보다는 더 많은 대화로 구성되지만 많은 양의 대화문이 존재할 시에 AI 훈련이

효과적으로 이루어지지 못할 가능성이 높다. 다만 상대적으로 적은 내용의 문답으로 이루어진 데이터셋의 한계를 극복하기 위해서 메타데이터에 문답의 내용에 대한 정보를 삽입하는 것이 필요할 것이다. 예를 들면 각 문답 데이터의 주제를 과거력, 현병력, 약물력 등으로 한정 지어서 구축한다면 의료진 대화에 대한 AI 훈련을 효과적으로 수행할 수 있을 것이다.

4.8 연구의 한계점

본 연구 결과를 해석할 시에는 다음과 같은 한계점을 고려하여야 한다. 첫번째, 본 연구는 한국에서 수행되었기 때문에 한국어가 아닌 다른 언어에서의 음성인식 정확도를 평가할 수 없었다. 두번째, 현재 상용화된 음성인식 솔루션은 화자의 구분을 정확하게 시행하지 못하였다. 세번째, 앞의 연구는 한 명의 순환기내과 전문의의 진료실에서 녹음된 데이터를 바탕으로 수행하였기 때문에 환자의 질병의 종류가 심장관련 영역으로 제한되어 있다. 그러나 절반 이상의 환자의 외래 방문 목적이 수술 전 심기능 평가였던 관계로 비교적 다양한 의학 용어가 사용되었다. 네번째, 본 연구의 결과에서 상용화된 음성인식 솔루션이 상당히 낮은 인식 정확도를 나타낸 관계로 어떤 음성인식 솔루션이 의무기록 작성에 가장 적합한지에 대해 제시할 수 없었다. 마지막으로 의료 대화에 대한 음성인식 정확도 평가는 임상의사의 질문만을 대상으로 수행하였기 때문에

환자의 답변에 대한 인식 정확도를 평가할 수 없었다. 그러나, 대부분의 환자는 의료인이 아닌 일반인이고 의학용어보다는 일반적인 단어를 사용한다는 것을 고려하였을 시에 임상의의 질문에 대한 인식 정확도가 환자 답변의 정확도에 비해 낮을 것이라는 예측이 가능하다.

이러한 한계에도 불구하고, 본 연구는 최초로 진료실에서 이루어지는 실제 의사-환자 음성을 수집하여 클라우드 기반 음성인식 솔루션의 인식 정확도에 대한 결과를 제공하였으며, 이를 통해 의료 서비스에 클라우드 기반 음성인식 솔루션의 적용을 위한 앞으로의 연구에 필요한 통찰력을 제공하였다.

제 5 장. 결론

이번 연구의 결과를 통하여 실제 의료현장에서 이루어지는 환자-의사간의 대화의 인식에 있어서 현재 상용화된 클라우드 기반 음성인식 솔루션을 적용하는 것에는 한계가 존재함을 알 수 있었다. 의료용어에 대한 음성인식 정확도 분석에서는 클라우드 기반 음성인식 솔루션은 아직까지 의무기록 작성에 직접 적용하기에는 상대적으로 낮은 인식 정확도를 보여주었다. 의료 대화의 인식 정확성에 대한 분석에서는 높은 인식률을 보여주었지만, AI 훈련을 위해 구축된 전용 데이터세트와 단순하고 짧은 문장의로 인식을 수행한 결과였다. 그러므로 클라우드 기반 음성인식 솔루션의 실제 의료현장의 적용을 위해서는 아직은 많은 발전이 필요하다.

세 종류의 상용화된 클라우드 기반 음성인식 솔루션 중에서 국내업체에서 구축한 Naver Clova SR이Google Speech-to-Text와 Amazon Transcribe와 비교했을 시에 한국어로 된 의학용어의 인식에서 가장 높은 인식 정확도를 보여주었다. 3종류의 음성인식 솔루션 모두 질병, 장기로 분류된 의학용어에 대해 높은 인식 정확도를 보여주었다. 특히 Naver Clova SR의 경우 2글자 이하의 의학용어 인식에 있어서 80%가 넘는 인식률을 보여주었다. Naver Clova SR은 진료실 대화의 인식에 있어서도 가장 높은 인식 정확도를 보여주었으며, 진료실 대화의 인식 정확도 평가시에

임상의에 의한 판정이나 자동화 평가 지표 모두 유사한 결과를 보여주었다.

연구의 결과를 통하여, 의무기록 작성을 위한 의료 목적 음성인식 솔루션의 성능을 향상시키기 위해 차후 연구 및 개발에서 요구되는 몇가지의 필요사항을 알 수 있었다. 의학용어 중 진료과와 성분명, 상품명에 대한 인식 정확도의 개선이 필요하며, 또한 앞으로의 음성인식 시스템은 긴 의학용어에 대한 인식 정확도를 개선해야 할 것이다. 추가적으로 화자 구분 기술이 절실히 필요하다. 마지막으로, 의료 산업에 적용하기 위한 음성인식 솔루션의 훈련을 위해서는 대화형식으로 구성된 훈련용 데이터 세트가 절실히 필요하다. 아직까지 의무기록작성을 위한 음성인식 솔루션에는 많은 한계점이 존재하지만, 본 연구의 결과를 참고하여 이 유망한 기술을 앞으로 더욱 더 고도화 할 수 있을 것으로 사료된다.

참고 문헌

1. Hodgson, T. & Coiera, E. Risks and benefits of speech recognition for clinical documentation: A systematic review. *Journal of the American Medical Informatics Association* vol. 23 e69–e179 Preprint at https://doi.org/10.1093/jamia/ocv152 (2016).
2. Paulett, J. M. & Langlotz, C. P. Improving language models for radiology speech recognition. *J Biomed Inform* **42**, 53–58 (2009).
3. Zhou, L. *et al.* Analysis of Errors in Dictated Clinical Documents Assisted by Speech Recognition Software and Professional Transcriptionists. *JAMA Netw Open* **1**, e180530 (2018).
4. Hammana, I., Lepanto, L., Poder, T., Bellemare, C. & Ly, M.-S. Speech Recognition in the Radiology Department: A Systematic Review. *Health Information Management Journal* **44**, 4–10 (2015).
5. Johnson, M. *et al.* A systematic review of speech recognition technology in health care. *BMC Medical Informatics and Decision Making* vol. 14 Preprint at https://doi.org/10.1186/1472-6947-14-94 (2014).
6. Blackley, S. v. *et al.* Physician use of speech recognition versus typing in clinical documentation: A controlled observational study. *Int J Med Inform* **141**, (2020).
7. Swayamsiddha, S., Prashant, K., Shaw, D. & Mohanty, C. The prospective of Artificial Intelligence in COVID-19 Pandemic. *Health and Technology* Preprint at https://doi.org/10.1007/s12553-021-

00601-2 (2021).

8. Kang, S. *et al.* Oculoplastic video-based telemedicine consultations: Covid-19 and beyond. *Eye (Basingstoke)* vol. 34 1193–1195 Preprint at https://doi.org/10.1038/s41433-020-0953-6 (2020).

9. Liao, C. te, Chang, W. T., Yu, W. L. & Toh, H. S. Utility of telemedicine in the COVID-19 era. *Rev Cardiovasc Med* **21**, 583 (2020).

10. Lee, S.-W. *et al.* CareCall: a Call-Based Active Monitoring Dialog Agent for Managing COVID-19 Pandemic. *ArXiv* (2020) doi:10.48550/arXiv.2007.02642.

11. Iqbal, M. Z. & Faiz, M. F. I. Active Surveillance for COVID-19 Through Artificial Intelligence Using Real-Time Speech-Recognition Mobile Application. in *2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)* 1–2 (IEEE, 2020). doi:10.1109/ICCE-Taiwan49838.2020.9258276.

12. Fernandes, J. G. Artificial Intelligence in Telemedicine. in *Artificial Intelligence in Medicine* 1219–1227 (Springer International Publishing, 2022). doi:10.1007/978-3-030-64573-1_93.

13. Hinton, G. *et al.* Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Process Mag* **29**, 82–97 (2012).

14. Chiu, C.-C. *et al.* State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2018). doi:10.1109/ICASSP.2018.8462105.

15. Callaway, E. C., Sweet, C. F., Siegel, E., Reiser, J. M. & Beall, D. P. Speech recognition interface to a hospital information system using a self-designed visual basic program: Initial experience. *J Digit Imaging* **15**, 43–53 (2002).

16. Yoo, H. J., Seo, S., Im, S. W. & Gim, G. Y. The performance evaluation of continuous speech recognition based on Korean phonological rules of cloud-based speech recognition open api. *International Journal of Networked and Distributed Computing* **9**, 10–18 (2021).

17. Hossain, M. S. & Muhammad, G. Cloud-Assisted Speech and Face Recognition Framework for Health Monitoring. *Mobile Networks and Applications* **20**, 391–399 (2015).

18. Kulkarni, S., Torse, D. A. & Kulkarni, D. A Cloud based Medical Transcription using Speech Recognition Technologies A Cloud based Medical Transcription using Speech Recognition Technologies. *International Research Journal of Engineering and Technology* **07**, (2020).

19. Muhammad, G. Automatic speech recognition using interlaced derivative pattern for cloud based healthcare system. *Cluster Comput* **18**, 795–802 (2015).

20. Guide of Amazon Transcribe Medical.

21. Guide of Nuance Dragon Medical One.

22. Kodish-Wachs, J., Agassi, E., Kenny, P. & Overhage, J. M. A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. *AMIA Annu*

Symp Proc **2018**, 683–689 (2018).

23. Spinazze, P., Aardoom, J., Chavannes, N. & Kasteleyn, M. The Computer Will See You Now: Overcoming Barriers to Adoption of Computer-Assisted History Taking (CAHT) in Primary Care. *J Med Internet Res* **23**, e19306 (2021).

24. Kreps, G. L. & Neuhauser, L. New directions in eHealth communication: Opportunities and challenges. *Patient Educ Couns* **78**, 329–336 (2010).

25. Elmore, N. *et al.* Investigating the relationship between consultation length and patient experience: a cross-sectional study in primary care. *British Journal of General Practice* **66**, e896–e903 (2016).

26. Sinsky, C. *et al.* Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. *Ann Intern Med* **165**, 753 (2016).

27. Kannampallil, T., Abraham, J., Lou, S. S. & Payne, P. R. O. Conceptual considerations for using EHR-based activity logs to measure clinician burnout and its effects. *J Am Med Inform Assoc* **28**, 1032–1037 (2021).

28. Arndt, B. G. *et al.* Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. *The Annals of Family Medicine* **15**, 419–426 (2017).

29. Wang, J., Lavender, M., Hoque, E., Brophy, P. & Kautz, H. A patient-centered digital scribe for automatic medical documentation. *JAMIA Open* **4**, (2021).

30. Koehn, P. & Knowles, R. Six Challenges for Neural Machine Translation. in *Proceedings of the First Workshop on Neural Machine Translation* 28–39 (Association for Computational Linguistics, 2017). doi:10.18653/v1/W17-3204.

31. Choi, S. J. & Kim, J.-B. Comparison Analysis of Speech Recognition Open APIs' Accuracy. *Asia-pacific Journal of Multimedia services convergent with Art, Humanities, and Sociology* **7**, 411–418 (2017).

32. Rajkomar, A. *et al.* Automatically Charting Symptoms from Patient-Physician Conversations Using Machine Learning. *JAMA Intern Med* **179**, 836–838 (2019).

33. Kim, D. *et al.* Automatic Classification of the Korean Triage Acuity Scale in Simulated Emergency Rooms Using Speech Recognition and Natural Language Processing: a Proof of Concept Study. *J Korean Med Sci* **36**, 1–13 (2021).

34. Basma, S., Lord, B., Jacks, L. M., Rizk, M. & Scaranelo, A. M. Error Rates in Breast Imaging Reports: Comparison of Automatic Speech Recognition and Dictation Transcription. *American Journal of Roentgenology* **197**, (2011).

35. Lee, S.-H., Park, J., Yang, K., Min, J. & Choi, J. Accuracy of Cloud-Based Speech Recognition Open Application Programming Interface for Medical Terms of Korean. *J Korean Med Sci* **37**, (2022).

36. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* –

ACL '02 311 (Association for Computational Linguistics, 2001). doi:10.3115/1073083.1073135.

37. Vedantam, R., Zitnick, C. L. & Parikh, D. CIDEr: Consensus-based Image Description Evaluation. *Arixiv* (2014).

38. Banerjee, S. & Lavie, A. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* https://aclanthology.org/W05-0909 (2005).

39. Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. (2004).

40. Blagec, K., Dorffner, G., Moradi, M., Ott, S. & Samwald, M. A global analysis of metrics used for measuring performance in natural language processing. *ArXiv* (2022).

41. Kim, S. *et al.* Building a Korean conversational speech database in the emergency medical domain. *Phonetics and Speech Sciences* **12**, 81–90 (2020).

42. Starren, J. B. *et al.* A retrospective look at the predictions and recommendations from the 2009 AMIA policy meeting: did we see EHR-related clinician burnout coming? *J Am Med Inform Assoc* **28**, 948–954 (2021).

43. Koivikko, M. P., Kauppinen, T. & Ahovuo, J. Improvement of report workflow and productivity using speech recognition - A follow-up study. *J Digit Imaging* **21**, 378–382 (2008).

44. Ghatnekar, S., Faletsky, A. & Nambudiri, V. E. Digital scribe utility and barriers to implementation in clinical practice: a scoping review. *Health Technol (Berl)* **11**, (2021).

45. Kaushal, A., Altman, R. & Langlotz, C. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. *JAMA* **324**, (2020).

46. Choi, D. H., Park, I. N., Shin, M., Kim, E. G. & Shin, D. R. Korean Erroneous Sentence Classification with Integrated Eojeol Embedding. *IEEE Access* (2021) doi:10.1109/ACCESS.2021.3085864.

47. Leslie, D., Mazumder, A., Peppin, A., Wolters, M. K. & Hagerty, A. Does 'aI' stand for augmenting inequality in the era of covid-19 healthcare? *The BMJ* **372**, (2021).

48. Aminololama-Shakeri, S. & López, J. E. The doctor-patient relationship with artificial intelligence. *American Journal of Roentgenology* **212**, 308–310 (2019).

Abstract

Background: There are limited data on the accuracy of cloud-based speech recognition (SR) solutions for medical conversations. The purpose of present research was to evaluate the applicability of current SR systems to real world doctor-patients conversation for future research. To achieve the purpose, we aimed to evaluate the accuracy of cloud-based SR solutions in discerning medical conversation both medical terminology and clinician's question presented in Korean, a non-Latin-based language, using records and transcriptions of real doctor-patient conversations and dedicated dataset for AI training and to find out the applicability to real world doctor-patients conversation recording.

Methods: We analyzed the SR accuracy of currently available cloud-based SR solutions using real doctor-patient medical terms recordings collected from an outpatient clinic at a large tertiary medical center in Korea. After first analysis, we analyzed the accuracy of current SR engines about doctor's speeches using dedicated dataset available at aihub.or.kr. For each original and SR transcription, we analyzed the accuracy rate of each cloud-based SR solutions by clinicians' judge and evaluation metrics (Bleu, CIDEr, ROUGE and METEOR score).

Results: The results of accuracy for medical terms as follows. A total

of 112 doctor-patient conversation recordings were converted with three cloud-based SR solutions (Naver Clova SR from Naver Corporation, Seongnam, Korea; Google speech-to-text from Alphabet Inc., Mountain View, CA, USA; and Amazon Transcribe from Amazon.com, Seattle, WA, USA), and each transcription was compared. Naver Clova SR (75.1%) showed the highest accuracy with the recognition of medical terms compared to the other solutions (Google speech-to-text, 50.9%, $P < 0.001$; Amazon Transcribe, 57.9%, $P < 0.001$), and Amazon Transcribe demonstrated higher recognition accuracy compared to Google speech-to-text ($P < 0.001$). In the sub-analysis, Naver Clova SR showed the highest accuracy in all areas according to word classes, but Google speech-to-text showed the highest recognition accuracy of words longer than five syllables, without statistical significance.

In the aspect of SR accuracy for medical speech, we extracted a total of 500 doctor's questions from "dataset for speech of health care provider and patient for telemedicine" of aihub.or.kr. The extracted doctor's questions were converted with three cloud-based SR soultions (Naver Clova SR from Naver Corporation, Seongnam, Korea; Kakao API, Speech-to-Text (demo) from Kakao Corp., Jeju, Korea; and Google speech-to-text from Alphabet Inc., Mountain View, CA, USA), and comparisons of accuracy were evaluated via clinicians'

judge and automated methods. Naver Clova SR showed the highest accuracy judged by clinicians (Naver, 89.7% vs. Kakao, 77.2% vs. Google, 66.0%; p< 0.001; respectively). In the aspects of automated methods, Bleu-1 score of Naver was the highest among three SR engines (Naver, 0.654 vs. Kakao, 0.578 vs. Google, 0.535; $p < 0.001$ respectively). Moreover, Bleu-2 (0.557 vs. 0.463 vs. 0.418; $p < 0.001$; respectively) and Bleu-3 score (0.389 vs. 0.306 vs. 0.262; $p < 0.001$; respectively) of Naver were the highest compared to Kakao and Google. Kakao showed higher Bleu-1, 2, and 3 scores compared to Google with statistical significance. CIDEr, METEOR, and ROGUE scores presented the same results that Naver Clova SR showed, the highest SR accuracy among three SR engines.

Conclusions: Current cloud-based SR solutions have limitations in the recognition of medical terminology. The SR solutions manufactured by a domestic company showed the highest recognition accuracy among the three solutions assessed in this study. Meanwhile, SR accuracy rate of medical conversation using dedicated database for AI training showed acceptable accuracy. There is still room for improvement of this promising technology and consideration about construction of dataset of low-resource language is needed.