논문 20-45-10-19 The Journal of Korean Institute of Communications and Information Sciences '20-10 Vol.45 No.10
https://doi.org/10.7840/kics.2020.45.10.1793

클라우드 기반의 음성인식 오픈 API의 응용 분야별 한국어 연속음성인식 정확도 비교 분석

유 현 재◆, 김 명 화*, 박 상 길**, 김 광 용°

Comparative Analysis of Korean Continuous Speech Recognition Accuracy by Application Field of Cloud-Based Speech Recognition Open API

Hyun-Jae Yoo◆, Myung-Wha Kim*, Sang-Kil Park**, Kwang-Yong Kim°

요 약

음성인식은 딥러닝 기술의 적용과 클라우드 컴퓨팅 등장으로 성능이 크게 개선되었다. 개선된 음성인식은 차량, 로봇, 헬스케어, 콜센터 등 다양한 분야에서 응용되고 있다. 본 논문은 클라우드 기반의 음성인식 Open API에 대한 응용 분야별 한국어 연속음성인식 성능을 비교하였다. 실험은 국내와 국외 클라우드 기업 7개를 하였다. 한국어 연속음성데이터는 국내 3개 방송사의 뉴스 데이터를 활용하였다. 수집 방법은 총 10개 분야로 나누어 분야별 15개 문장, 총 150문장을 수집하였다. 실험 결과 분야별 전체 음성인식 정확도는 카카오가 가장 좋았고, IBM이 가장 낮았다. 분야별로는 카카오는 6개분야, Amazon와 Microsoft은 2개분야, ETRI는 1개 분야에서 좋은 성능을 보였다. 실험 결과 클라우드 컴퓨팅 기업이 지원하는 음성인식 엔진이 특정 분야에 우수한 성능을 보이는 특성이 있음을 확인하였다. 본 연구는 클라우드 기반의 음성인식 Open API를 지원하는 기업들의 음성인식 엔진의 분야별 성능 개선에 기여하기를 바란다. 그리고 음성인식 개발자에게는 응용 음성인식 시스템을 개발하는데 해당 응용 분야에 가장 적합한 음성인식 Open API를 선택하는데 도움이 되기를 기대한다.

Key Words : Speech recognition, Acoustic model, Language model, Open API, Cloud computing

ABSTRACT

Speech recognition has significantly improved performance through the application of deep learning technology and the emergence of cloud computing. The improved speech recognition has been applied in various fields such as vehicles, robot, healthcare, and call center. This paper compares the performance of continuous Korean speech recognition by application field for the cloud-based speech recognition Open API. The experiment was conducted with 7 domestic and foreign cloud companies. Korean continuous speech data was used by news data from three domestic broadcasters. The collection methods were divided into a total of 10 fields and collected 15 sentences by sector, a total of 150 sentences. As a result of the experiment, the overall speech recognition accuracy by field was the highest in Kakao and the lowest in IBM. By field,

◆ First Author : Graduate School of IT Policy and Management, Soongsil University, callnet0@daum.net, 정회원
° Corresponding Author : Soongsil University, gygim@ssu.ac.kr, 정회원
* Graduate School of IT Policy and Management, Soongsil University, beauhwa1@naver.com
** WAREBIZ Co., Ltd., highroad@warebiz.co.kr
논문번호 : 202007-154-0-SE, Received July 11, 2020; Revised August 20, 2020; Accepted August 24, 2020

Kakao showed good performance in 6 fields, Amazon and Microsoft in 2 fields, and ETRI in 1 field. As a result of the experiment, it was confirmed that the speech recognition engine supported by the cloud computing company exhibits excellent performance in a specific field. This study is hoped to contribute to improving the performance of speech recognition engines for companies that support cloud-based speech recognition open APIs. In addition, for speech recognition developers, it is expected that it will help to select the most suitable voice recognition open API for the application field in developing an applied voice recognition system.

Ⅰ. 서 론

클라우드 컴퓨팅 환경의 등장과 클라우드 기반 음성인식 오픈 API제공은 음성인식 시스템 개발의 애로사항인 음성데이터 수집의 어려움, 고가의 고성능 컴퓨터의 필요성, 음성인식 시스템 개발을 위한 많은 시간과 노력 등 문제를 해결하여 주었다. 클라우드 기반의 음성인식 오픈 API는 음성인식 시스템 개발자가 아니더라도 누구나 쉽고, 빠르게 응용 음성인식 시스템을 개발 할 수 있도록 지원을 하였다. 콜센터 음성인식, 음성인식 개인비서, AI 스피커, 언어번역 앱, 로봇, 의료 음성인식, 차량용 음성인식 등 여러 분야에서 응용 사례가 증가하고 있다. 음성인식 시스템은 전처리, 음향모델, 언어모델, 디코딩 네트워크의 과정을 통해서 음성 데이터를 텍스트 문장으로 만들어 내는데, 학습데이터의 종류, 양, 방법 등에 따라 차별적인 결과를 만들어 낸다. 따라서 음성인식 시스템은 어떤 알고리즘과 어떤 데이터를 가지고 학습을 했느냐에 따라 음성인식 시스템의 특징으로 정의할 수 있다. 본 연구는 이러한 음성인식 시스템의 특징을 실험을 통해서 알아보고 클라우드 기반의 음성인식 Open API를 활용하기 위한 기준을 제시하고자 한다.

본 연구는 클라우드 기반의 음성인식 오픈 API를 이용한 한국어 연속음성인식의 정확도를 알아보았다. 논문의 구성은 2장에서는 음성인식의 개요, 음성인식 기술의 분류, 클라우드 음성인식 Open API, 음성인식 관련 선행연구를 기술하였고, 3장은 실험 방법과 실험 결과에 대한 내용으로 음성데이터 수집 방법, 연구 방법 및 절차, 실험 환경, 실험 결과에 대한 내용을 기술하였다. 마지막으로 4장에서는 본 연구의 결론으로 결과에 대한 평가와 의미를 정리하였다. 그리고 본 연구의 아쉬운 점을 정리하고 그에 대한 향후 과제에 대하여 기술하였다.

Ⅱ. 본 론

2.1 음성인식 개요

음성인식이란 인간의 말을 문자로 자동 변환하는 기술이다. 한국정보통신기술협회(TTA) 정보통신용어 사전에서는 음성인식을 "음성으로부터 언어적 의미 내용을 자동으로 식별하는 것으로, 보다 구체적으로는 음성파형을 입력하여 단어나 단어열을 식별하고 의미를 추출하는 처리 과정이다"라고 설명하고 있다[1]. 음성인식의 기술은 음성 입력 장치를 통해서 얻은 음성 신호를 분석하여 단어나 문장으로 변환시키는 기술로 식(1)과 같이 정의할 수 있다.

$$arg_W max P(W|O) \fallingdotseq \underset{\text{디코딩}}{arg_W max} \underset{\text{음향 모델}}{P(O|W)} \underset{\text{언어 모델}}{P(W)} \quad (1)$$

식(1)[3,26] 좌측의 정의는 W는 N개의 단어들로 이루어진 문장이고, O는 소리가 주어졌을 때 관측된 값이다. 식(1) 우측의 정의는 좌측의 정의를 Bayesian rule을 적용하여 우측의 식을 유도할 수 있다. 음성인식 시스템의 주요 구성은 식(1) 우측의 정의에 따라 음향모델 P(O|W), 언어모델 P(W), 디코딩 네트워크 $arg_W max$ 로 구성된다. 음성인식이란 식(1) 정의에 따라 관측된 값 O가 주어졌을 때 P(W|O)의 값

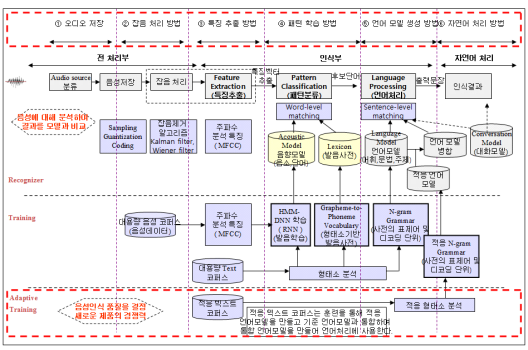


그림 1. 음성인식시스템의 구성도
Fig. 1. Composition diagram of speech recognition system

을 최대로 해주는 문장 W를 찾는 것이 음성인식의 결과이다.

음성인식시스템의 처리과정은 그림 1[7]음성인식시스템의 구성도 에서 보는 것과 같이 전처리부와 인식부로 나누어서 처리된다.

2.1.1 음성인식 전처리

음성인식 전처리는 음성인식 처리에 사용되는 음성 데이터를 잘 가공하는 작업이다. 음성 신호는 음성 발화 환경, 음성 전달 매체 등의 형태에 따라 다양한 특징을 보인다. 그 특징은 잡음, 음향간섭 등으로 음성 신호가 왜곡되어 음성인식의 성능을 크게 저하 시킨다. 이러한 특징을 갖는 음성신호를 개선하기 위한 기술이 음성인식 전처리 기술이다. 전처리 기술에는 음성신호 저장, 잡음제거, 음성 특징 파라미터 추출 등의 기술이 있다.

2.1.2 음향모델

음향모델은 음성인식 구성에서 음성을 인식해서 글자로 변환하는 가장 핵심적인 역할을 한다. 음향모델에 입력은 음성신호의 특징을 추출한 음성데이터가 입력되어 문자로 출력 된다. 음성인식 기술의 기초가 되고, 음향모델에서 가장 널리 사용되고 있는 기술은 HMM(Hidden Markov Model)이다.

현재의 가장 좋은 성능을 보이는 시스템은 Hinton 교수가 제안한 DNN-HMM을 기반으로 하고 있다. Hinton 교수에 의하면 GMM-HMM 기반 음성인식 대비, DNN-HMM 기반 음성인식의 성능은 상대적으로 약 20% 정도 좋은 것으로 정리 하였다[2].

그러나 DNN은 추정을 해야 하는 파라미터가 많기 때문에 학습 시간이 많이 소요되고, DNN에서 모델의 구분 단위, 대용량 음성 코퍼스로부터 모델 구분 단위별 학습 자료 음소를 자동으로 분할하여 HMM경로 중에서 확률이 가장 높은 HMM경로를 찾는 것과 모델 결합을 통한 문장 인식 확장성, 즉 디코딩 네트워크에서는 HMM의 해결 방안을 그대로 사용한다. 최근에는 HMM을 접목한 GMM-HMM, DNN-HMM 등의 하이브리드 방법이 복잡한 음성인식 구조로 되어있고, 학습하는데 많은 시간이 소요 된다는 문제점이 있어[3], 오디오 신호의 특징을 입력 받아 바로 단어 또는 문장으로 출력하는 End-to-End 학습 방법을 많이 제안 하고 있다[4].

2.1.3 언어모델

언어 모델은 특정 단어열이 주어졌을 때 다음에 나올 단어들의 확률을 추정하는 모델이다. 언어모델은 식(2)[3][26]에서 P(W)에 해당되는 것으로, T개의 단어로 구성된 문장 W(w0 … wT)에 대하여 문장 생성확률 P(W) = P(w0 … wT)를 계산하는 것이다.

$$\underbrace{arg_w max}_{\text{디코딩}} \underbrace{P(O|W)}_{\text{음향 모델}} \underbrace{P(W)}_{\text{언어 모델}} \quad (2)$$

언어모델은 단어 별로 분해를 한 후, history(wk-1 wk-2… w0) 로부터 다음 단어 (wk)를 예측한다. T개의 단어로 구성된 문장 W(w0 … wT)에 대하여 문장 생성확률은 식(3)[27]과 같이 계산된다.

$$P(W) = \prod_{k=1}^{T} P(w_k|w_{k-1}w_{k-2}\dots w_0) \quad (3)$$

단어를 구분하는 단위는 형태소(morpheme), 어절, 음절(syllable)이 있다. 형태소(morpheme)는 국립국어원 표준국어대사전[5]에서 “형태소는 뜻을 가진 가장 작은 말의 단위이다.”[5]라고 정의한다. 형태소는 더 쪼개면 전혀 의미가 없는 단어가 된다. 어절은 국립국어원 표준국어대사전[5]에서 “어절은 문장 성분의 최소 단위로서 띄어쓰기의 단위가 된다.”라고 정의한다. 음절(syllable)은 국립국어원 표준국어대사전[5]에서 “음절은 하나의 종합된 음의 느낌을 주는 말소리의 단위. 몇 개의 음소로 이루어지며, 모음은 단독으로 한 음절이 되기도 한다.”라고 정의한다.

2.1.4 디코딩 네트워크

디코딩(Decoding) 네트워크는 음성에 대하여 음향모델, 언어모델, 발음사전, 어휘사전 등으로 만들어진 탐색 네트워크에서 가장 최적의 경로를 찾아 단어열을 추정한다. 디코딩 네트워크의 기술로는 Lexical Tree 기반 search와 Weighted Finite State Transducer (WFST) 의 두가지 방법이 있다.

Lexical Tree 기반 search는 디코딩 구성 요소인 HMM, Context dependency, Lexicon, Grammar가 별도로 구현되어 있으며, 직관적이고 수동으로 수정이 가능하며, 쉽게 만들 수 있다는 장점이 있다. 단점은 탐색하는 속도가 늦다. Weighted Finite State Transducer (WFST)은 디코딩 구성 요소인 HMM, Context dependency, Lexicon, Grammar를 결합하여 미리 네트워크를 구현하였으며, 탐색하는 속도가 빠르다. 단점은 컴퓨터 시스템의 메모리가 많이 소요 되고, 어휘나 학습 텍스트가 약간만 바뀌어도 다시 구성해야 하는 번거로움이 있다. 최근 음성인식 기술에서는

Weighted Finite State Transducer (WFST)의 디코딩 네트워크 방법을 많이 사용하고 있다[6].

2.2 음성인식의 기술 분류

음성인식의 기술 분류는 일반적으로 인식대상 화자, 발성의 형태, 인식대상 단어로 구분하여 그 기술이 분류된다. 발성의 형태에 따른 분류는 고립단어, 연결단어, 연속어, 핵심어 인식으로 분리되며, 인식대상 화자에 대한 분류는 화자종속, 화자독립, 화자적응 인식으로 분류 된다. 그리고 인식대상 단어에 따른 분류는 고정단어, 가변단어 인식으로 분류된다. Table 1[7]은 이러한 음성인식 기술 분류에 따른 특징을 설명하고 있다. 음성인식은 새롭고 성능을 높이는 기술의

표 1. 음성인식 기술의 구분 및 특징
Table 1. Classification and characteristics of speech recognition technolog

구 분		특징
발성의 형태	고립단어 인식	- 고립된 단어만을 인식하는 기술 - 음성인식의 가장 기초적인 기술
	연결단어 인식	- 여러 개의 단어를 인식하는 기술 - 고립단어 인식보다 난이도 있음.
	연속어 인식	- 연속음성인식으로 자연스럽게 음성을 인식하는 기술 - 현재 인식률을 높이기 위한 많은 연구가 이루어지고 있다.
	핵심어 인식	- 연속적으로 입력되는 음성신호에서 미리 정해진 핵심단어만을 인식하는 기술. - 연속어 인식의 전단계의 기술.
인식 대상 화자	화자종속 인식	- 특정 화자의 음성을 미리 인식 시스템에 훈련시켜 등록해 놓고 해당 화자만 인식하는 기술. - 비교적 구현이 간단한 기술.
	화자독립 인식	- 특정 화자에 종속되지 않고 대량의 음성데이터로 훈련을 하여 어떤 화자든 상관없이 인식하는 기술. - 특정 화자의 훈련이 필요 없이 인식할 수 있는 특징이 있으며 많은 사용제품들이 채택하고 있다.
	화자적응 인식	- 화자종속 및 화자독립의 절충한 기술. - 화자독립 인식의 방법에 특정 화자의 인식률을 높이기 위하여 해당 화자의 음성을 별도로 훈련시켜 인식하는 기술.
인식 대상 단어	고정단어 인식	- 고정된 단어를 인식하는 기술. - 인식해야 할 대상단어를 교체할 경우 음성모델을 구축해야 함. - 시간과 비용이 많이 소모됨.
	가변단어 인식	- 인식 대상단어를 수시로 갱신함. - 대상단어가 갱신될 경우, 기존의 음소에 대한 정보를 이용하여 인식 대상 모델을 생성함.

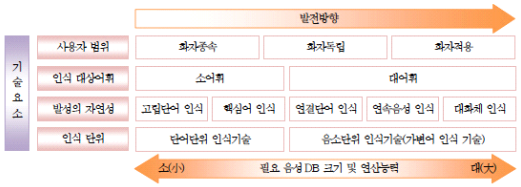


그림 2. 음성인식 기술의 발전 방향
Fig. 2. Direction of development of speech recognition technology

발전이 꾸준하게 이루어지고 있으며, 그에 따른 많은 연구를 하고 있다. 그러나 아직도 음성인식 기술의 수준은 많은 기술적인 한계를 가지고 있다. 음성인식의 궁극적인 목표인 인간이 발성하는 모든 음성을 자연스럽게 인식할 수 있는 수준까지는 아직도 많은 해결해야 할 과제들이 많다. 그림 2[8]은 음성인식 기술 분류를 기술 발전 방향을 나타낸 것이다. 음성인식 기술은 더 많은 화자, 더 많은 어휘, 더 자연스러운 대화체를 인식하는 방향으로 발전해 가고 있다.

2.3 클라우드 음성인식 Open API

2.3.1 클라우드 컴퓨팅

클라우드 컴퓨팅은 위키피디아에서 "클라우드(인터넷)을 통해 가상화된 컴퓨터의 시스템 리소스(IT 리소스)를 요구하는 즉시 제공(on-demand availability)하는 것이다"[9]라고 설명하고 있다. 클라우드 컴퓨팅은 수도 또는 전기를 필요할 때마다 사용할 수 있는 것과 같이 컴퓨터 자원(하드웨어, 소프트웨어, 스토리지, 등)을 원격으로 제공하고, 이용자는 시간과 장소에 상관없이 컴퓨터 자원을 원격으로 주문하여 사용하며, 그 사용량에 대하여 비용을 지불하는 방식을 의미한다. 클라우드 컴퓨팅은 주문형 사용, 유비쿼터스 접근, Multitenancy, Elasticity, 사용량 측정, Resiliency 등의 특성을 가지고 있다.

클라우드 컴퓨팅의 대표적인 장점은 첫째, 컴퓨터 리소스에 대한 투자비용이 적고, 그에 따른 유지비용이 절감된다. 둘째, 컴퓨터 리소스의 확장성이 좋다. 즉, 필요할 때 자원을 늘리고 줄일 수 있는 편리성을 제공한다. 셋째, 시스템 구축 기간이 단축된다. 넷째, 시스템 구축에 대한 조직의 의사결정이 빠르다. 여섯째, 시스템의 가용성과 신뢰성이 높다. 즉, 장애와 유지보수 중에도 중단 없는 서비스가 가능하다. 반면 클라우드 컴퓨팅은 단점도 있다. 첫째, 보안 취약점이 증가한다. 즉, 중요한 데이터의 외부환경에 저장한다는 점과 외부업체에 데이터의 안정성을 위임해야하는

부담이 있다. 둘째, 시스템에 대한 운영통제 관리가 감소한다. 셋째, 기존 업체와의 계약 종료 후 다른 서비스 업체로의 데이터를 옮기는 것이 어렵다. 넷째, 다양한 지역의 규정, 규제 및 법적 문제로 특정 정부기관이 데이터의 공개를 요구한다.[28,29]

2.3.2 음성인식 Open API(Aplication Programming Interface)

클라우드 기반의 음성인식 Open API는 음성인식 시스템을 개발할 수 있도록 클라우드 컴퓨팅 기업들이 공개적으로 제공하는 API이다.

클라우드 기반의 음성인식 Open API의 제공은 클라우드 컴퓨팅의 장점을 통하여 음성인식 개발자가 아니라도 응용 음성인식 시스템을 개발할 수 있도록 지원한다. 클라우드 컴퓨팅 기업은 클라우드의 특성을 통하여 다량의 음성데이터를 수집하고, 고성능 컴퓨터로 학습을 통하여 음성인식 엔진을 만들었다. 음성인식 개발자들은 기업들이 제공하는 Open API를 통하여 이렇게 만들어진 음성인식 엔진을 이용할 수 있게 되었다. 음성인식 Open API는 음성인식 시스템을 개발하기 위한 시간과 노력을 줄여 주었고, 누구나 쉽고, 빠르게 원하는 응용 음성인식 시스템을 개발할 수 있도록 하였다. 클라우드 기반의 음성인식 Open API를 제공하는 기업들은 국내의 Naver Clova Speech Recognition[10], Kakao Speech-to-Text system[11], ETRI STT[12], KT GiGA Genie 음성인식[13], SKT NUGU[14] 등이 대표적이고, 국외는 Google Cloud Speech-to-Text[15], IBM Watson Speech to Text[16], MicroSoft Azure Cognitive Speech Service[17], Amazon Transcribe[18] 등이 대표적인 기업이다.

2.3.3 음성인식 Open API 지원 내용 비교

Table 2는 클라우드 기반 음성인식 오픈 Open API의 제공하는 기업들의 지원 내용을 비교하였다.

표 2. 음성인식 Open API 비교
Table 2. Comparison of speech recognition Open API

구 분	언어지원	서비스 방식	오디오 형식	API지원	가격
Google Cloud Speech-to-Text [15]	120개 언어	- 스트리밍 - 비스트리밍	LINEAR16, FLAC, MULAW, AMR, AMR_WB, OGG_OPUS, SPEEX_WITH_HEADER_BYTE, MP3	- SDK (C#, GO, 자바, Node.js, PHP, Python, Ruby) - OS 지원 (Linux, macOS, Windows) - REST API 지원 - RPC API 지원	- 최초 60분이내 무료 - 표준모델:$0.006/15초 - 프리미엄:$0.009/15초 - 표준모델:$0.004/15초 - 프리미엄: $0.006/15초 (로깅 있음)
IBM Watson Speech to Text[16]	7개 언어	- 스트리밍 - 비스트리밍	alaw, u-law(mu-law), flac, g729, linear 16, mp3, mpeg, ogg, wav, webm	- HTTP REST API - WebSocket - 비동기 HTTP	- Lite:월500분 무료 - Standard:월100분 무료, Tire Level에 따라 $0.01/MINUTE ~ $0.03/MINUTE - Premium: 언어모델의 사용자 정의 기능, 별도문의
MicroSoft Azure Cognitive Speech Service[17]	43개 언어	- 스트리밍 - 비스트리밍	WAV(16KHz 또는 8Khz, 16비트 및 mono PCM), mp3, mpeg, OPUS/OGG, ALAW, MULAW	- SDK (GO, C#, C++, Java, JavaScript, Python) - OS (Linux, macOS, Windows) - REST API	- 무료(웹/컨테이너 1개 동시 요청) : 월 5시간 무료 - Standard(웹/컨테이너 20개 동시 요청) : 월 1시간당 1$
Amazon Transcribe [18]	32개 언어	- 스트리밍 - 비스트리밍	FLAC, MP3, MP4, or WAV file format	- SDK (.NET, GO, Java, Javascript, PHP, Python, Ruby)	- 초당 0.0004 USD, 사용량은 1초 단위로 청구되며 요청당 15초의 최소 요금이 부과. - 프리티어: 12개월 매월 60분 무료

구 분	언어지원	서비스 방식	오디오 형식	API지원	가격
Naver Clova Speech Recognition[10]	4개 언어	- 스트리밍 (모바일 SDK), - 비스트리밍 (녹음파일: REST API)	mp3, aac, ac3, ogg, flac, wav	- 모바일 SDK (안드로이드 2.3.4 - API 레벨 10), iOS 8이상 - REST API	- 15초당 4원
kakao Speech-to-Text system[11]	1개 언어	- 스트리밍 (모바일 SDK), - 비스트리밍 (녹음파일: REST API)	Mono channel, 16KHz samplerate, 16bit RAW PCM	-모바일 SDK (안드로이드 4.6 - API 레벨 14, 23), iOS 9이상 - REST API	- Beta 서비스중. 무료
ETRI STT[12]	9개 언어	- 비스트리밍	16Khz Linear PCM 지원	- REST API 지원	- 비상용서비스, 무료

2.3.4 음성인식 Open API 지원 내용 비교

Table 3의 음성인식 오픈 API의 선행 연구자료를 보면, 2017년 3월 연구에서 Google이 우수한 성능을 보였다[19]. 2017년, 8월, 10월, 12월의 연구자료를 보면, 실험 데이터의 종류가 차이는 있어도 여전히 Google의 성능이 우수한 것으로 나타났다[20-22]. 그러나 2018년 12월 연구에서 Google은 INTERMEDIATE[23], 2019년에는 한글에서는 ETRI, Naver, Microsoft보다 성능이 떨어졌다[24]. 실험의 시점에 따라 음성인식 오픈 API의 성능이 변화하고 있음을 확인할 수 있었다. 또한 실험 음성데이터의 종류도 숫자, 한글, 영어, 단어, 문장, 방언, 사투리 등 다양했다. 이러한 실험 음성데이터의 종류에 따라 음성인식 오픈 API의 성능의 차이도 보였다. 본 연구는 기존의 음성데이터와 달리 응용분야별 음성데이터를 수집하여 실험하였다. 응용 분야별 음성데이터의 실험은 응용 음성인식 시스템 개발

표 3. 음성인식 Open API 선행 연구
Table 3. Prior study of speech recognition Open API

논문주제	실험내용	실험결과	출판 연도
Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx) [19].	오류율 (WER) (Microsoft API, Google API And CMU Sphinx)	Google이 가장 우수하게 나타났다.	March 2017
Comparison Analysis of Speech Recognition Open APIs' Accuracy [20].	Google, 카카오, Naver 3개 API의 숫자, 한글, 문장 세가지에 대한 오류율 측정.	- 숫자: 카카오, 네이버, 구글순으로 낮음 - 한글: 네이버, 카카오, 구글순으로 낮음 - 문장: 네이버, 카카오는 같았고, 구글이 다음으로 낮았다.	August 2017
A comparisonof cloud-based speech recognition engines [21].	정확성과 성능 평가 (Bing Speech API, Google Cloud Speech 및 IBM Watson Speech to Text)	정확도는 Google, 가장 높았고, 음성인식 처리 속도는 Google 가장 느렸다[21].	Oct 2017
A Basic Performance Evaluation of the Speech Recognition APP of Standard Language and Dialect using Google, Naver, and Daum KAKAO APIs [22].	- 구글, 네이버, 다음, 카카오, 각각의 API를 이용하여 앱을 만듦. - 성별, 나이별, 지역별 표준어, 방언에 대한 음성 인식 정확도 비교[22].	- 표준어 : 띄어쓰기 → 네이버, 구글, 다음 순 : 받침 → 네이버, 다음, 구글 순 : 조사 → 구글, 네이버, 다음 순 : 단어의 오류 → 구글=네이버, 다음 순 : 전체 정확도는 Google이 가장 높음. 음성인식 처리 속도는 Google 가장 느림[22]. - 방언 : 충청도 → 구글, 네이버, 다음 순 : 전라도 → 구글, 네이버, 다음 순 : 경상도 → 억양과 음의 높낮이 차이가 큰 문장은 이해 못함. 생소한 경상도 방언 단어는 잘 인식함[22].	Dec 2017

논문주제	실험내용	실험결과	출판 연도
Comparison of the Top Speech Processing APIs [23].	EXCELLENT, GOOD, INTERMEDIATE 로 평가 (Google, IBM, Microsoft, Amazon, Twilio, Speech -matics, Nexmo)[23].	- GOOG : Amazon, IBM, MS - INTERMEDIATE : Google	Dec 2018
PERFORMANCE COMPARISON OF OPEN APIS FOR SPEECH RECOGNITION [24].	선정한 문장을 1m, 3m, 5m에서 녹음한 후 인식율 실험 (Kakao 뉴톤, KT GiGA Genie : 한국어, Amazon Transcribe, Microsoft Azure Speech Service : 영어 ETRI Open API, Google Cloud Speech-to-Text, IBM Cloud API, Naver Clova : 한국어 영어)	- 한국어의 경우 ETRI Open API 가 72%, 1m 에서 ETRI Open API 가 57%, 3m 에서 Naver Clova 가 23%, 5m 에서 Naver Clova 가 7%로 가장 높은 인식률을 보였고, - 영어의 경우 원본 파일에서 Microsoft Azure Speech Service 가 91%, 1m 에서 Microsoft Azure Speech Service 가 45%, 3m 에서 Amazon Transcribe 가 16%, 5m 에서 ETRI Open API 가 7%로 가장 높은 인식률을 보였다	2019

에 실제적인 도움을 줄 수 있다는 것에 연구의 필요성과 의의가 있다.

Ⅲ. 실 험

3.1 실험 방법

음성인식 Open API를 이용한 실험 시스템은 기업들이 지원하는 선택 사항들을 고려하지 않았다. 실험은 기업별 같은 음성데이터를 입력하고, 결과로 나온 출력 텍스트 문장에 대하여 분석하였다.

음성데이터는 공용방송 3사(KBS, MBC, SBS) 뉴스를 10개의 분야(문화, 경제, 부동산, 의료, 군사, 정치, 과학, 사회, 스포츠, 날씨)로 구분하여 분야별 15개, 총 150개의 음성데이터를 수집하였다. 실험은 기업에서 제공하는 선택 사항 없이 기본적으로 제공되는 Open API를 이용하였다. 음성데이터는 남녀 구분을 하지 않았다. 실험 대상 음성인식 Open API 기업은 국내 3개사(네이버, 카카오, ETRI)와 국외 4개사(구글, 아마존, IBM, MS)를 포함한 총 7개 기업을 선정하였다. 개발언어는 PHP를 이용하였고, 서비스 방식은 비스트리밍 방식, 디바이스는 데스크톱 컴퓨터, 실험 프로그램은 웹환경에서 개발하였다. 평가방법은 어절단위로 식(4)의 방법으로 정확도를 측정하였다. 결과로 나온 출력 텍스트 문장은 기업마다 표현 방식을 통일시키기 위하여 다음과 같이 교정 후 분석을 하였다. 첫째, 숫자(서수, 기수)의 표기는 입력 전사 문장에 통일 시켰다.(예:457주년 → 사백오십칠주년, 11개 → 열한개, 80.4 → 팔십점사)

둘째, 기호는 모두 없앴다.(예:쉼표, 마침표, 인용부호) 셋째, 영어발음 표기는 통일 시켰다. (예:SUV → 에쓰유브이) 넷째, 영어 대소문자는 통일시켰다. 다섯째, 띄어쓰기는 한국어 맞춤법검사기를 이용하여 교정을 보았다[25].

정확도 계산 방법은 식(4)로 정의한 것과 같이 오류율(WER)을 구하여 계산하였다. 오류율(WER:Word Error Rate)은 치환오류(S:Substitution), 삽입오류(I:Insertion), 삭제오류(D:Deletion)의 합을 입력 어절수(N)로 나눈 것으로 WER=(S+I+D)/N으로 표현한다.

$$\text{정확도(Accuracy)}=(1-\text{WER})\text{X}100 \quad (4)$$

3.2 실험 결과

실험 결과는 Table 4에서 와 같이 분야별 구분 없

표 4. 전체 음성인식 Open API 정확도 비교
Table 4. Comparison of speech recognition Open API accuracy

구 분	Google	Amazon	IBM	MS	Naver	Kakao	ETRI
전체 어절수	2199	2199	2199	2199	2199	2199	2199
틀린 어절수	511	191	639	192	308	132	214
Accuracy	76.76%	91.31%	70.94%	91.27%	85.99%	94.00%	90.27%

표 5. 분야별 음성인식 Open API 정확도 비교
Table 5. Comparison of speech recognition Open API accuracy by Field

구분	Google	Amazon	IBM	MS	Naver	Kakao	ETRI	계
문화	82.14	88.84	74.11	92.86	87.50	98.21	87.95	87.37
경제	75.00	91.67	85.09	92.98	92.11	89.47	92.98	88.47
부동산	69.57	94.35	72.61	93.48	93.48	95.65	91.74	87.27
의료	85.91	85.91	68.18	92.27	82.27	90.45	86.82	84.55
국방	75.90	94.38	76.31	93.57	88.35	96.79	90.36	87.95
정치	78.01	94.61	78.84	93.78	90.04	98.34	94.61	89.75
과학	75.23	90.83	61.01	87.61	82.57	89.45	88.53	82.18
사회	80.34	93.59	69.66	90.17	78.21	92.74	89.74	84.92
스포츠	69.31	86.24	49.21	83.60	75.66	94.18	88.36	78.08
날씨	74.70	90.96	68.67	90.36	87.95	93.98	90.96	85.37
합계	76.76	91.31	70.94	91.27	85.99	94.00	90.27	85.79

이 전체 음성인식 정확도는 카카오가 94.00%로 가장 좋았고, IBM이 70.94%로 가장 낮았다. 분야별로는 Table 5에서 와 같이 문화 분야는 카카오가 98.21%, 경제 분야는 ETRI, Microsoft가 92.98%, 부동산 분야는 카카오가 95.65%, 의료 분야는 Microsoft가 92.27%, 국방 분야는 카카오가 96.79%, 정치 분야는 카카오가 98.34%, 과학 분야는 Amazon이 90.83%, 사회 분야는 Amazon이 93.59%, 스포츠 분야는 카카오가 94.18%, 날씨 분야는 카카오가 93.98%로 좋은 성능을 보였다. 그리고 Table 6에서 분야별 상위 순위 정확도 비교표를 보면, 1위는 카카오가 6개 분야, ETRI는 1개 분야, Microsoft는 2개 분야, Amazon은 2개 분야에서 좋은 성능을 보였다.

Ⅳ. 결 론

본 논문에서는 클라우드 기반의 음성인식 Open API의 분야별 한국어 연속음성인식 정확도를 비교하였다. 실험 결과는 그림 3에서와 같이 분야별 구분 없이 전체 오류율은 Kakao가 가장 좋은 성능을 보였으며, 그 뒤를 Amazon이 좋은 성능을 보였다. IBM은 가장 낮은 성능을 보였다. 그림 5에서는 보면 분야별 성능이 경제, 의료, 과학, 사회 분야를 제외하고는 Kakao가 모든 분야에서 좋은 성능을 보였다. ETRI는 경제 분야에서, Microsoft는 경제, 의료 분야에서, Amazon은 과학 분야와 사회 분야에서 좋은 성능을 보였다. Kakao가 분야별 통합 정확도에서 가장 우수

표 6. 분야별 상위 순위 음성인식 Open API 비교
Table 6. Comparison of top-ranking speech recognition Open API by field

분 야	1위		2위		평 균
	기업	Accuracy	기업	Accuracy	
문화	Kakao	98.21	MS	92.86	87.37
경제	ETRI,MS	92.98			88.47
부동산	Kakao	95.65	Amazon	94.35	87.27
의료	MS	92.27	Kakao	90.45	84.55
국방	Kakao	96.79	Amazon	94.38	87.95
정치	Kakao	98.34	Amazon	94.61	89.75
과학	Amazon	90.83	Kakao	89.45	82.18
사회	Amazon	93.59	Kakao	92.74	84.92
스포츠	Kakao	94.18	ETRI	88.36	78.08
날씨	Kakao	93.98	Amazon	90.96	85.37
분야	Kakao	94.00	Amazon	91.31	85.79

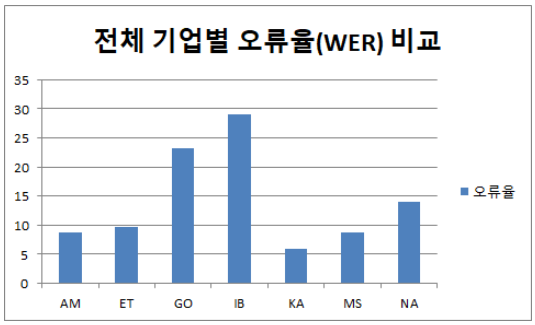


그림 3. 전체 기업별 오류율 비교
Fig. 3. Comparison of WER by company

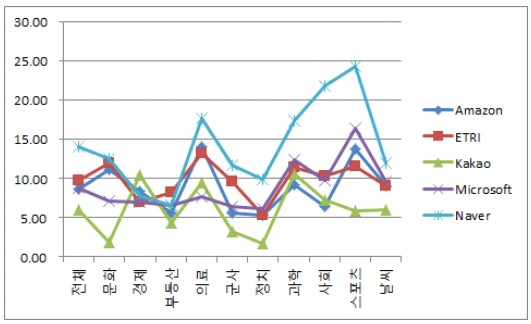


그림 4. 상위 순위 기업 분야별 오류율 비교
Fig. 4. Comparison of WER by top-ranked companies

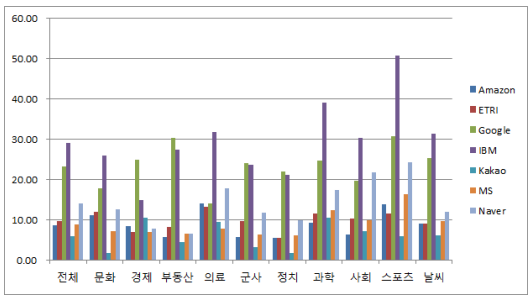


그림 5. 기업 분야별 오류율 비교
Fig. 5. Comparison of WER by company sector

하였지만 분야별로 보면 전 분야에서 우수한 성능을 보이지는 않았다. 그림 4에서 보면 분야별 음성인식의 정확도 순위는 분야 통합 전체의 순위와 특정 분야에서의 순위가 같지 않음을 확인할 수 있었다. 본 연구를 통해서 클라우드 기반의 음성인식 Open API를 지원하는 기업의 음성인식 엔진은 분야별로 정확도의 차이를 보이는 특성이 있음을 확인하였다. 따라서 본 연구는 클라우드 기반의 음성인식 Open API를 지원하는 기업들의 음성인식 엔진의 분야별 성능 개선에 기여하기를 바라고, 음성인식 개발자에게는 응용 음성인식 시스템을 개발하는데 해당 응용 분야에 가장 적합한 음성인식 Open API를 선택하는데 도움이 되기를 기대한다.

향후 과제로는 음성인식 Open API별 음성인식 결과 문장에 대한 특성을 분석하는 것이다. 본 연구에서 음성인식 Open API별 음성인식 결과는 의미는 같지만 표현 방법이 다른 특성을 보여 분석에 앞서 교정 작업을 하였다. 이러한 특성들에 대한 분석이 없이는 응용 음성인식 시스템 개발에 직접 적용하는 것은 어렵다. 따라서 향후 과제는 응용 음성인식 시스템의 개발에 직접 적용할 수 있는 음성인식 Open API별 음성인식 결과 문장에 대한 특성을 분석하고 그에 따른 후처리 방법을 제시하는 것에 있다.

References

[1] Telecommunications Technology Association (TTA), *Information and Communication Glossary*, http://terms.tta.or.kr/dictionary/dictionaryView.do?word_seq=097616-1

[2] G. Hinton, et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Process. Mag.*, pp. 15-17, Jun. 2012.

[3] D. Lee, M. Lim, H. Park, and J.-H. Kim, “LSTM RNN-based Korean speech recognition system using CTC,” *J. Digital Contents Soc.*, vol. 18, no. 1, pp. 93-99, Feb. 2017.

[4] H. Park, D. Lee, M. Lim, Y. Kang, J. Oh, S. Seo, D. Rim, and J.-H. Kim, “Hybrid CTC-Attention based end-to-end speech recognition using korean grapheme unit,” *Conf. Human & Cognitive Lang. Technol.*, 2018.

[5] National Institute of the Korean Language, *Standard Korean Dictionary*, https://stdict.korean.go.kr/main/main.do

[6] O.-W. Kwon, “Research trends in WFST-Based speech recognition,” *J. KIISE*, vol. 21, no. 2, pp. 1-4, Oct. 2013.

[7] Korea Creative Content Agency, “*Trends and Prospects of Voice Recognition Technology*,” Cultural Technology (CT) In-depth Report, Nov. 2011.

[8‘] National IT Industry Promotion

Agency(NIPA), "*Weekly Technology Trend*," Oct. 2011.

[9] https://ko.wikipedia.org/wiki/%ED%81%B4%EB %9D%BC%EC%9A%B0%EB%93%9C_%EC%BB%B4%ED%93%A8%ED%8C%85

[10] https://www.ncloud.com/product/aiService/csr

[11] https://developers.kakao.com/docs/latest/ko/voice/common

[12] http://www.aihub.or.kr/ai_software/370#group00

[13] https://cloud.kt.com/portal/ktcloudportal.epc.productintro.aivoice.html#

[14] https://developers-doc.nugu.co.kr/nugu-sdk

[15] https://cloud.google.com/speech-to-text/

[16] https://www.ibm.com/kr-ko/cloud/watson-speech-to-text

[17] https://azure.microsoft.com/ko-kr/services/cognitive-services/speech-to-text/

[18] https://aws.amazon.com/ko/transcribe

[19] V. Këpuska and G. Bohouta, "Comparing speech recognition systems (Microsoft API, Google API And CMU Sphinx)," *J. Eng. Res. and Appl.*, ISSN : 2248-9622, vol. 7, no. 3, Part. 2, pp. 20-24, Mar. 2017.

[20] S. J. Choi and J.-B. Kim, "Comparison analysis of speech recognition open APIs' accuracy," *Asia-pacific J. Multimedia Services Convergent with Art, Humanities, and Sociology*, vol. 7, no. 8, pp. 411-418, Aug. 2017.

[21] A. L. Herchonvicz, C. R. Franco, and M. G. Jasinski, "A comparison of cloud-based speech recognition engines," *Computer on the Beach*, Oct. 2017.

[22] H.-K. Roh and K.-H. Lee, "A basic performance evaluation of the speech recognition APP of standard language and dialect using Google, Naver, and Daum KAKAO APIs," *Asia-pacific J. Multimedia Services Convergent with Art, Humanities, and Sociology*, vol. 7, no. 12, pp. 819-829, Dec. 2017.

[23] Igor Bobriakov, *Comparison of the Top Speech Processing APIs*, https://activewizards.com/blog/comparison-of-the-top-speech-processing-apis, Dec. 2018.

[24] H. Oh, K.-N. Lee, and D. Yook, "Performance comparison of open APIS for speech recognition," *The Acoustical Soc. Korea*, 2019.

[25] Co-production of Artificial Intelligence Lab of Pusan National University and Narain Four Tech Co. Ltd., http://164.125.7.61/speller/

[26] W. Chan, "End-to-End speech recognition models," BASc Comput. Eng., Univ. of Waterloo MS Electrical and Comput. Eng., Carnegie Mellon Univ. Carnegie Mellon Univ., Pittsburgh, PA, Oct. 2016.

[27] H. Choi, J. Park, and D. Park, "Language modeling in speech recognition," *Commun. of the KIISE*, vol. 16, no. 2, pp. 17-22, Feb. 1998.

[28] J. Jeong, "*Current status and challenges of cloud computing*," National Assembly Research Service Pending Report, vol. 313, Dec. 2017.

[29] C.-B. Lee, "Legal tasks for safe use and revitalization of cloud computing," *J. Info. Secur. Assoc.*, vol. 20, no. 2, Apr. 2010.

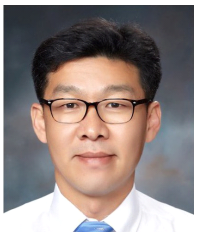
유 현 재 (Hyun-Jae Yoo)

12018년 8월 : 서강대학교 정보통신대학원 소프트웨어공학 석사

2019년 3월~현재 : 숭실대학교 IT정책경영학과 박사과정

<관심분야> 음성인식, 음성합성, 자연어 처리, AI, 보이스봇, 빅데이터

박 상 길 (Sang-Kil Park)



2020년 8월 : 숭실대학교 IT정책경영학과 박사

2007년 11월~현재 : (주)웨어비즈 대표이사

<관심분야> 스마트시티, 클라우드, 네트워크 보안, 빅데이터, 자연어 처리학, 광통신 공학

김 명 화 (Myung-Hwa Kim)

2005년 2월 : 서울대학교 환경대학원 석사

2019년 3월~현재 : 숭실대학교 IT정책경영학과 박사과정

<관심분야> AI, 딥러닝, 빅데이터, 소프트웨어 개발방법론

김 광 용 (Gwang-Yong Kim)

1984년 : 고려대학교 공학사 졸업

1991년 : 조지아 주립대학 보험수리학 석사

1995년 : 미국 조지아 주립대학 박사

1999년~현재 : 숭실대학교 경영학부 교수

<관심분야> 데이터사이언스, 디지털트랜스포메이션, 인공지능, 빅데이터, 블록체인, 클라우드, IOT, 전자정부, 핀테크, 비즈니스 모델링(디자인싱킹, TRIZ, 캔버스모델 등) 등