Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology
Vol.7, No.8, August (2017), pp. 411-418
http://dx.doi.org/10.14257/ajmahs.2017.08.76

음성 인식 오픈 API의 음성 인식 정확도 비교 분석

최승주[1], 김종배[2]

Comparison Analysis of Speech Recognition Open APIs' Accuracy

Seung Joo Choi[1], Jong-Bae Kim[2]

요 약

음성인식기술은 마이크와 같은 소리 센서를 통해 얻은 음향학적 신호를 단어나 문장으로 변환시키는 기술을 말한다. 이 기술과 인공지능을 결합한 음성 대화 시스템은 차세대 인터페이스로 주목받고 있으며, 스마트폰, 스마트TV, 자동차 등 다양한 분야에서 사용되고 있다. 최근에는 삼성전자에서 인공지능과 음성인식을 결합한 '빅스비'를 출시하였으며, Google, Naver 등 다양한 기업들은 음성인식기술을 오픈 API로 제공하고 있다. 본 논문에서는 대표적인 음성 인식 오픈 API 3개를 선택하여 각 특징을 비교 분석한다. 또한 한 3번의 실험을 통해 모바일 환경에서 각 음성인식 API별 인식률을 비교하였다. 첫 번째로 숫자 인식을 실험하였고, 두 번째로는 가나다 한글 인식을 실험하였다. 세 번째 실험에서는 모바일 음성인식 프로그램에서 쓰이는 대표적인 명령 문장을 입력하여 문장 인식 실험을 진행하였다. 이러한 비교실험을 통해 한국어를 지원하는 음성인식 오픈 API의 선택 기준을 제시하여 상황별로 적절한 API를 사용하는 데에 도움을 줄 수 있을 것으로 기대한다.

핵심어 : 음성 인식, 음성 대화 시스템, 음성 이해, 오픈소스, 오픈 API

Abstract

Speech recognition technology is transformation skill using sound sensor such as microphone to transfer the acoustical signal to words or sentence. Speech conversation system using this technology and artificial intelligence is receiving attention as next generation of interface, and it is used in variable areas like smartphone, smart TV, car and so on. Recently, Samsung released 'Bixby' which is speech conversation program with artificial intelligence, and a lot of company such as Google and Naver are providing speech recognition open API. In this paper, we select three typical APIs and do comparison analysis of APIs' features. In addition to that, we do three experiment in mobile for analysis of APIs' speech recognition accuracy. First, we test number recognition. In second test, we test Korean word recognition. Lastly, we test sentence recognition with mobile instruction sentence. With result, we expect developers can select appropriate speech recognition open API in each situation.

Keywords : speech recognition, voice recognition, speech understanding, open source, open API

Received (June 2, 2017), Review Result (June 19, 2017)

Accepted (June 26, 2017), Published (August 31, 2017)

[1]06978 Graduate School of Software, Soonsil University, Seoul, Korea
email: csj5183@gmail.com
[2](Corresponding Author) 06978 Graduate School of Software, Soongsil University, Seoul, Korea
email: kjb123@ssu.ac.kr

ISSN: 2383-5281 AJMAHS

1. 서론

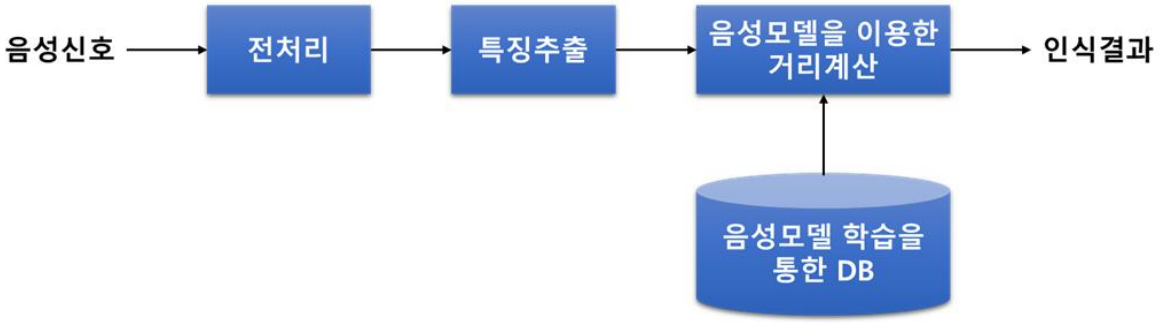
음성인식기술은 마이크와 같은 소리 센서를 통해 얻은 음향학적 신호를 단어나 문장으로 변환시키는 기술을 말한다[1]. 이 기술과 인공지능을 결합한 음성 대화 시스템은 차세대 인터페이스로 주목받고 있으며, 스마트폰, 스마트TV, 자동차 등 다양한 분야에서 사용되고 있다[2]. 특히 최근 Google은 인공지능 음성인식 비서인 'Google Assistant'를 출시했으며, 자동차, 가전 제품 등에 활용 가능한 음성인식 기능들을 오픈 API로 제공하고 있다. 또한 국내에서도 Naver 2017년 5월 12일 음성인식 인공지능인 Naver Clova가 출시되었고, 삼성전자에서는 빅스비를 출시하는 등 음성인식에 대한 개발이 활발하게 이루어지고 있다[3].

본 논문에서는 한국어가 지원되는 대표적인 음성 인식 오픈 API들의 특징들을 분석하여 각 오픈 API의 장단점을 비교하였다. 또한 3번의 실험을 통해 모바일 환경에서 각 음성인식 API별 인식률을 비교하였다. 첫 번째 실험에서는 숫자 음성 인식 실험을, 두 번째 실험에서는 한글 음성 인식을 실험하였다. 세 번째 실험에서는 모바일에서 많이 쓰이는 명령어 문장을 입력하여 문장 인식 실험을 진행하였다. 이러한 음성인식 API의 특징 비교와 인식률 실험을 통해 한국어를 지원하는 음성인식 오픈 API의 선택 기준을 제시하여 상황별로 적절한 API를 사용하는 데에 도움을 줄 수 있을 것으로 기대한다.

2. 관련 연구

2.1 음성인식

음성인식은 자동적 수단에 의하여 음성으로부터 언어적 의미 내용을 식별하는 것으로 구체적으로 음성 파형을 입력하여 단어나 단어열을 식별하고 의미를 추출하는 처리 과정이며, 크게 음성분석, 음소 인식, 단어 인식, 문장 해석, 의미 추출의 5가지로 분류된다. 좁은 의미로는 음성 분석에서 단어 인식까지를 말하는 경우가 많다. 음성 인식의 궁극적인 목표는 자연스러운 발성에 의한 음성을 인식하여 실행 명령어로서 받아들이거나 자료로서 문서에 입력하는 완전한 음성/텍스트 변환(full speech-to-text conversion)의 실현이다. 단지 단어를 인식할 뿐 아니라 구문 정보(문법), 의미 정보, 작업에 관련된 정보와 지식 등을 이용하여 연속 음성 또는 문장의 의미 내용을 정확하게 추출하는 음성 이해 시스템(speech understanding system)을 개발하는 것이다[4][5][6]. 음성 인식 기술의 원리는 [그림 1]과 같다.



[그림 1] 음성인식 기술의 원리[7]

[Fig. 1] Principal of speech recognition technology[7]

2.2 응용 프로그램 인터페이스(Application Program Interface, API)

운영 체계(OS)에서 응용 프로그램을 만들 수 있도록 제공하는 소프트웨어이다. 응용 프로그램은 API를 사용하여 OS 따위가 가지고 있는 다양한 기능을 이용할 수 있다. 초기의 개인용 컴퓨터(PC)에서는 응용 프로그램이 하드웨어의 기능을 직접 조작하는 경우가 많았다. 그러나 파일 관리나 정보의 화면 표시 기능처럼 모든 기능을 응용 프로그램 내에 두면 프로그램 개발 효율이 떨어지고, 복수의 응용 프로그램을 번갈아 사용하였을 때 문제가 발생하기 쉽다. 따라서 많은 응용 프로그램이 공통으로 이용할 수 있는 기능은 OS 따위에 두는 것이 일반이다. 응용 프로그램 작성자가 프로그램 중에 함수를 기술하기만 하면 함수 호출에 따라 다양한 기능을 이용할 수 있게 된다. 이 함수의 집합이 API이며, 종류가 다른 OS 사이에 API의 공통 형식이 규정되면 이기종의 컴퓨터 사이에 응용 프로그램의 이식성(portability)이 확보된다[4].

2.3 음성 인식 오픈 API

오픈(공개) API란 누구나 사용할 수 있도록 공개된 API를 말한다. 임의의 응용 프로그램을 쉽게 만들 수 있도록 준비된 프로토콜, 도구 같은 집합으로 프로그램 개발자는 운영 체제의 상세한 기능은 몰라도 공개된 몇 개의 API만으로도 쉽게 응용 프로그램을 개발할 수 있다[4]. 세계적인 음성 인식 오픈 API로는 Google의 Cloud Speech API가 있으며, 국내에서는 대표적으로 카카오의 뉴톤(Newtone)과 네이버의 Clova Speech API가 제공되고 있다.

3. Open API의 비교

본 논문에서는 한국어 음성 인식이 지원되는 Google의 Cloud Speech API와 카카오의 뉴톤(Newtone), 네이버의 음성인식 API Clova Speech Recognition의 특징들을 비교하였다. 비교 항목은 지원하는 자연언어와 컴퓨터 언어, 무료 서비스 범위, 활용 분야 등이 있다. [표 1]은 비교 분석표이다.

[표 1] 음성인식 오픈 API 비교 분석 표[8]
[Table 1] Comparison analysis table of speech recognition open API

	Google Cloud Speech API[9]	카카오 뉴톤(Newtone)[10]	Naver Clova Speech Recognition[11]
지원 자연 언어	한국어 포함 80여개 언어 지원	한국어	한국어, 영어, 일어, 중국어(간체)
지원 컴퓨터 언어	C#, GO, JAVA, NODE.JS, PHP, PYTHON, RUBY	C#, JAVA	C#, JAVA
무료 서비스 제한 범위	한 달에 음성인식 1시간 무료	없음	하루 3,600초
활용 분야	스카트폰, PC, 테블릿, IoT 기기(자동차, TV, 스피커 등)	스마트폰	스마트폰, 웹
기술 지원 범위	SDK 제공, API 문서 제공, 설치 및 설정 가이드, 설정 최적화 예시, 음성인식 개념 설명 제공, 샘플 어플리케이션 제공	SDK 제공, API 문서 제공, 설치 및 설정 가이드, 설정 최적화 예시, SDK 및 API에 대한 일반적인 질문 등	SDK 제공, API 문서 제공, 설치 및 설정 가이드, 설정 최적화 예시, 샘플 APK 제공
커뮤니케이션 채널	공식 커뮤니티 제공, 외부 오픈 소스 커뮤니티 링크 제공	공식 카페 제공	외부 오픈 소스 커뮤니티 링크 제공, 공식 개발자 포럼 제공
인공지능 지원	지원	미지원	지원

자연 언어와 컴퓨터 언어를 가장 다양하게 지원하는 API는 Google Cloud Speech API였으며, 이에 따라 활용 분야도 가장 다양하였다. 무료 서비스의 제한이 없는 API는 카카오의 뉴톤이 유일했다. 셋 API 모두 SDK와 API 문서, 설치 및 설정 가이드를 제공하였으며, 또한 구글과 네이버는 샘플 APK를 지원하였다. 세 API 모두 공식 커뮤니티를 지원하여 개발자간의 커뮤니케이션을 적극 지원하였다. 구글과 네이버의 두 API는 인공지능을 함께 지원하였다.

4. 실험 방법

본 논문에서는 한국어 음성 인식이 지원되는 Google의 Cloud Speech API와 카카오의 뉴톤, 네이버의 Clova Speech Recognition의 Sample APK를 설치하거나 샘플코드를 Android Studio를 이용하여 모바일 어플리케이션으로 컴파일한 후 음성을 입력하여 음성 인식 정확도를 비교하였다. 실험은 입력장치 종류별 음성 인식률을 비교한 논문을 참고하였다[12]. 카카오의 뉴톤이 모바일 어플리케이션 개발 환경만 지원하기 때문에 입력 장치는 스마트폰을 사용하여 실험을 진행하였다.

실험은 3단계로 진행하였고, 각 실험은 잡음이 적은 같은 공간에서 API 종류별로 각각 10회씩 발성하였다. 첫 번째 실험은 0부터 19까지의 한국어 숫자 20단어를 입력하였다. 두 번째 실험에서는 가나다 14개의 음절을 입력하였다. 세 번째 실험에서는 모바일에서 많이 쓰이는 명령 문장을 입력하여 문장 인식률을 실험하였다. 입력한 명령 문장은 다음과 같다.

1. 최근 찍은 사진 친구한테 보내줘
2. 화요일 오후 3시 30분에 알람해줘

각각의 문장은 5개의 단어로 이루어져 있고, 총 10개의 단어를 한 번의 실험 횟수로 정하였다.

5. 실험 결과

Naver의 Clova Speech Recognition은 음성 인식 후 최대 5개의 음성인식 결과를 보여주었으며, 카카오의 Newtone은 10개의 음성인식 결과를 보여주었다. Google의 Cloud Speech API는 한 개의 결과만을 보여주었다. 음성이 입력된 값이 출력된 여러 결과 값 중에 일치하는 것이 있으면 음성인식이 실패하지 않은 것으로 간주하였다.

숫자 20단어를 입력한 첫 번째 실험 결과는 [표 2]의 표와 같다. 인공지능을 통해 연속으로 입력되는 값이 숫자로 판단되는 것을 방지하기 위해 입력 시 한 입력 시간에 숫자 1개만을 입력하여 숫자 20개가 한 문장으로 입력되지 않게 하였다.

[표 2] 음성인식 API 종류별 숫자 음성인식 결과
[Table. 2] number speech recognition result of speech recognition APIs

음성인식 API 종류	각 회별 틀린 개수 (각각 전체: 20)										계
	1	2	3	4	5	6	7	8	9	10	(전체:200)
Google	6	9	6	6	5	8	6	9	5	8	68(34%)
Naver	2	2	5	2	5	3	4	2	3	6	34(17%)
카카오	0	0	1	0	0	1	1	0	0	0	3(1.5%)

숫자 음성 인식에서 Google Speech의 오인식률은 34%로 가장 높았다. Google Speech는 특히 '가'와 '다', 또는 '나', '라', '마', 또는 '바', '파'의 발음을 구별하지 못했다. 그 다음으로 Naver의 Clova Speech Recognition이 17%, 카카오의 Newtone은 1.5%의 오인식률을 보였다.

[표 3]의 표는 한글 가나다 14개의 음절을 입력한 결과표이다. 이 또한 첫 번째 실험과 마찬가지로 한 입력 시간당 한 개의 음만을 입력하여 한글이 한 문장으로 입력되지 않게 하여 인공지능을 통해 연속되는 한글로 판단되는 것을 방지하였다.

[표 3] 음성인식 API 종류별 14개 한글 음성인식 결과
[Table 3] speech recognition result of Korean 14 words.

음성인식 API 종류	각 회별 틀린 개수 (각각 전체: 20)										계
	1	2	3	4	5	6	7	8	9	10	(전체:140)
Google	6	5	9	8	7	9	7	9	6	7	73(52.14%)
Naver	5	3	3	3	6	4	3	1	2	4	34(24.28%)
카카오	4	4	6	4	5	4	5	4	4	6	46(32.85%)

한글 음성 인식에서 Google Speech의 오인식률은 73%로 가장 높았다. 그 다음으로 카카오의 Newtone이 32.85% Naver의 Clova Speech Recognition은 24.28%의 오인식률을 보였다.

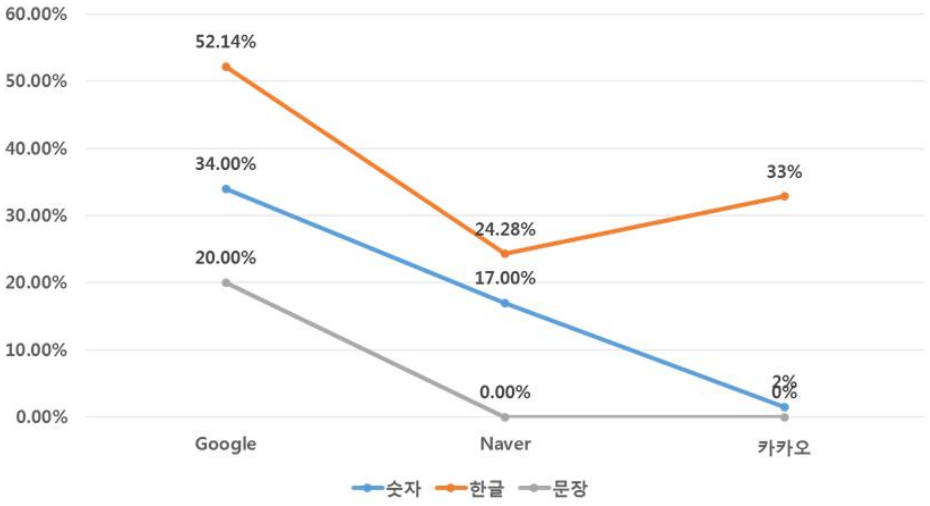
[표 4]의 표는 문장을 입력한 결과표이다. 한 단어의 입력과 결과 값이 일치하지 않으면 오인식된 것으로 간주하였다.

[표 4] 음성인식 API 종류별 문장 음성인식 결과
[Table 4] speech recognition result of sentences.

음성인식 API 종류	각 회별 틀린 개수 (각각 전체: 20)										계
	1	2	3	4	5	6	7	8	9	10	(전체:140)
Google	0	1	0	1	0	0	0	0	0	0	2(20%)
Naver	0	0	0	0	0	0	0	0	0	0	0(0%)
카카오	0	0	0	0	0	0	0	0	0	0	0(0%)

문장 인식에서는 앞의 실험들과 다른 결과를 보였다. Google 같은 경우 두 번째 문장의 첫 시작 부분인 '화요일'을 제대로 인식하지 못한 경우을 제외하면 모든 문장의 단어를 인식했다. Naver와 카카오의 오인식률을 0%였다.

[그림 2]는 세 실험을 종합한 결과 그래프이다.



[그림 2] 음성인식 API 실험 결과
[Fig. 2] Test result of speech recognition API

6. 결론

본 논문에서는 각 오픈 API의 특징들을 비교하였다. Google은 약 80여개의 언어를 지원하는 반면, 네이버는 4개 언어, 카카오는 한국어만을 지원하였다. Google은 또한 다양한 컴퓨터 언어를 지원하여 스마트폰 뿐만 아니라 다양한 IoT 기기에도 API를 활용할 수 있다. 기술의 지원 범위는 전체적으로 비슷하였지만, 음성인식의 개념 설명과 부가적인 설명을 제공한 곳은 Google이 유일했다. Google과 Naver는 인공지능 기술을 지원하기도 하였다.

또한 본 논문에서는 세 번의 실험을 통해 각 음성인식 API별 숫자, 한글, 문장에 대한 인식률을 측정하였다. 숫자, 한글, 문장 모두에서 구글의 오인식률이 높았지만 문장 인식에 있어서는 그 차이가 미미했다.

본 논문에서는 오픈 API들의 특징과 각각의 음성 인식 정확도를 비교하여 음성인식 오픈 API 선택 기준을 제시하고자 했다. 이를 통해 개발자들이 상황에 적절한 오픈 API를 선택할 수 있을 것으로 기대한다. 본 연구에서는 음성 데이터를 제공하는 실험자가 한 명이였다. 따라서 향후 연구로는 각 음성 API의 음성 인식률을 다양한 음성 데이터를 통한 실험이 필요하다.

References

[1] Jong-Hun Kim, Chang-Woo Song, Ju-Hyun Kim, Kyung-Yong Chung, Kee-Wook Rim, Jung-Hyun Lee, Smart Home Personalization Service based on Context Information using Speech Recognition. Journal of the Korea Contents Association. **(2009),** Vol.9, No.11, pp.80-89.

[2] http://www.segye.com/newsView/20170724003054**(2017)**

[3] http://news.hankyung.com/article/2017072469031**(2017)**

[4] Telecommunications Technology Association(http://www.tta.or.kr/)

[5] L. Rabiner and B. Juang, Fundamentals of Speech Recognition, 1st ed. Englewood Cliffs, NJ: Prentice Hall, **(1993)**

[6] G. Hinton et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition," The IEEE Signal Processing Magazine, **(2012)**, Vol. 29, No. 6, pp. 82-97

[7] Hyun Shin Park, Sung Woong Kim, Min-Ho Jin, Chan Dong Yoo, Current trend of speech recognition based machine learning. The Magazine of the IEIE. **(2014),** pp.18-27.

[8] Seung Joo Choi, Jong-Bae Kim, Comparison Analysis of Speech Recognition Open API, Convergence Research Letter. HSST, **(2017),** Vol.3, No.2, pp.1201-1204.

[9] https://cloud.google.com/speech/

[10]http://developers.daum.net/services/apis/newtone

[11]https://developers.naver.com/docs/labs/vrecog/

[12]Sam-Joo Doh, Myoung-Wan Koo, A Comparison of the Speech Recognition According to the Input Device, Journal of the Communications of the Korean Institute of Information Scientists and Engineers(HCI) **(1995),** pp.171-176.