International Journal of Advanced Smart Convergence Vol.10 No.2 79-85 (2021)
http://dx.doi.org/10.7236/IJASC.2021.10.2.79

IJASC 21-2-12

Comparative Analysis of Speech Recognition Open API Error Rate

Juyoung Kim*, Dai Yeol Yun **, Oh Seok Kwon ***, Seok-Jae Moon**** and Chi-gon Hwang****

**Graduate Student, Graduate School of Smart Convergence Kwangwoon University, Seoul, Korea*
*** Professor, Department of Plasma Bioscience and Display, KwangWoon University, Seoul 01897, Korea.*
**** Professor, Department of Plasma Bioscience and Display, KwangWoon University Graduate School, Seoul 01897, Korea*
***** Professor, Department of Computer Science, Kwangwoon University, Seoul, Korea*
***** Professor,Department of Computer Engineering, Institute of Information Technology, Kwangwoon University, Seoul, 01897, Korea*
E-mail { kjyjx, hibig10, ohskwon, msj80386, duck1052 }@kw.ac.kr

Speech recognition technology refers to a technology in which a computer interprets the speech language spoken by a person and converts the contents into text data. This technology has recently been combined with artificial intelligence and has been used in various fields such as smartphones, set-top boxes, and smart TVs. Examples include Google Assistant, Google Home, Samsung's Bixby, Apple's Siri and SK's NUGU. Google and Daum Kakao offer free open APIs for speech recognition technologies. This paper selects three APIs that are free to use by ordinary users, and compares each recognition rate according to the three types. First, the recognition rate of "numbers" and secondly, the recognition rate of "Ga Na Da Hangul" are conducted, and finally, the experiment is conducted with the complete sentence that the author uses the most. All experiments use real voice as input through a computer microphone. Through the three experiments and results, we hope that the general public will be able to identify differences in recognition rates according to the applications currently available, helping to select APIs suitable for specific application purposes.

Keywords: *Open API, Speech Recognition Technology, Recognition Rate, Artificial Intelligence*

1. Introduction

Speech recognition technology is a technology that converts sound voice signals obtained by computers through sound sensors such as microphones into words or sentences. [1] To briefly explain the technical principle, when a sound comes in through an input device, it converts the voice signal into a frequency spectrum and recognizes the incoming voice by classifying it into phonemes. And the process of probabilistic calculation of the relationship between words to produce sentences that one can understand. Speech recognition technology is already being serviced in many parts of our lives such as Smartphones, cars, and call centers. Currently, speech recognition technology translates what is said in language into Korean and vice versa. A translator that converts language as if talking to a native speaker. The ultimate goal can be to develop intelligent secretaries who know what to say and manage schedules and contacts.

In the past, the function of speech recognition was limited to texting sounds, but now it includes natural language processing (NLP) [2] technology that enables human-machine conversation. Recently, speech

Manuscript Received: April. 23, 2021 / Revised: May. 1, 2021 / Accepted: May. 4, 2021
Corresponding Author: hibig10@kw.ac.kr
Tel: +82-10-2742-6084, Fax: +82-2-940-5289
Professor, Department of Plasma Bioscience and Display, KwangWoon University, Seoul 01897, Korea.

recognition technologies that show innovative performance of deep learning technologies are also becoming common. Major deep learning techniques used in speech recognition recently include Deep Neural Networks (DNN) [3] , Convolutional Neural Network (CNN) [4] , and Long Short-Term Memory (LSTM-RNN) [5] . Another important background in the development of voice recognition and artificial intelligence technologies is the activation of open-source-based ecosystems. examples include Kaldi (Jones Hopkins University), Caffe (University of Berkeley), Tensorflow (Google), Theano (University of Montreal, Canada), Torch (Facebook), CNTK/DMTK (Microsoft), and cuDNN (Nvidia) [6]. Currently, language intelligence technology, which includes speech recognition, is the foundation technology commonly applied to products and services in various industries, and is one of human-computer interaction technologies. In this paper, we check the recognition rate of free Korean speech recognition open APIs. It uses three open APIs: Kakao, Google, and Microsoft. Each api three experiments to see recognition in comparison. In the first experiment, "number" recognition, in the second experiment, "Ga Na Da Hangul" recognition, and in the third experiment, the author's most frequently used "five-word sentence" were used as input values. [7] The inputs used in the experiment were generated using a computer microphone, and each experiment was conducted 10 times. Through these experiments, we expect to be able to check the recognition rate of free APIs and help select speech recognition APIs.

2. Related Works

2.1 Speech recognition technology

Speech recognition technology is a technology that allows computers to interpret human speech languages and convert their contents into text data, which is also described as Speech to Text (STT), Voice Recognition, and Artificial Hearing. Speech recognition system can be divided into offline processing module and online processing module. The offline processing module is a learning step in generating speech recognition models from speech data, and the online processing module is a navigation step in recognizing user-vocalized speech.

2.1.1 Models, methods, and algorithms

2.1.1.1 Hidden Markov models

Typical general-purpose voice recognition systems are based on the Hidden Markov Model (HMM) [8]. HMM is a statistical model that produces a series of outputs. Continuous speech input signals can be considered partial stop signals or short interval stop signals, so they are used for speech recognition. Voice can be considered as HMM for statistical purposes. HMM models are frequently used because they can be trained automatically and are simple to use and computationally feasible.

2.1.1.2 Dynamic time warping (DTW)-based speech recognition

Dynamic time warping (DTW) is an algorithm for measuring the similarity between two sequences that can vary with time or speed [9]. DTW is applied to video, audio, and graphics, allowing you to analyze any data that can be switched to linear representation. Automatic speech recognition is a representative application that deals with various speaking speeds. It is a method that allows us to find optimal matches between specific constraints and given two sequences (e.g., time series) data. That is, sequence data are nonlinearly "warped" to match each other.

2.1.1.3 Neural networks

Neural networks are often used in phoneme classification, phoneme classification through multi-objective evolutionary algorithms, isolated word recognition, audiovisual speech recognition, and audiovisual speaker recognition. Recently, LSTM and related recurrent neural networks (RNN) [10] and time-delayed neural networks (TDNN) [11] wn good performance. Deep feedforward and recurrent networks are artificial neural networks(DNN) with multiple device layers hidden between the input and output layers. DNN model complex nonlinear relationships. It shows good performance on large-scale lexical speech recognition.

2.1.1.4 End-to-end Automatic Speech Recognition

Existing speech-based (HMM-based models) approaches require separate components and training for pronunciation, acoustic, and language models. This recognition model is learned by combining all components of speech recognizers. This simplifies the training process and the distribution process.

2.1.2 Applications

2.1.2.1 In-car systems

You can use simple voice commands to make calls, select a radio station, or play music on a compatible smartphone, MP3 player, or flash drive with music. [12]

2.1.2.2 Health care

- **Medical documentation**

In the medical field, speech recognition is implementable at the beginning and the end of medical document processing. Speech recognition is directed by the provider to the speech recognition engine, displayed as the recognized words are used, and the author is responsible for editing and signing documents. Delayed speech recognition is now widely used in the industry.

- **Therapeutic use**

Long-term use of speech recognition software with word processors has shown advantages for short-term memory enhancement in patients with cerebral Arteriovenous Malformation (AVM) who underwent resection therapy.

2.1.2.3 Military

- **High-performance fighter aircraft**

The Advanced Fighter Technology Integration (AFTI)/F-16 Vista (F-16 Vista) voice recognition program in the United States, the Mirage program in France, and other programs covering a variety of aircraft platforms in the United Kingdom. In these programs, speech recognizers are successfully operating within fighter jets through applications such as radio frequency setting, autopilot system command, steering point coordinates and weapon emission parameter setting, and flight display control.

- **Helicopters**

The problem of achieving high recognition accuracy in stress and noise is strongly required not only for helicopter environments but also for jet fighter environments. Speech recognition applications include communication radio control, navigation system setup, and automatic target handover system control.

2.1.2.4 Telephony and other domains

ASR coming commonplace in telephony and is increasingly prevalent in computer games and simulations. In telephony systems, ASR is currently used primarily in service centers by integrating with Interactive Voice Response (IVR) systems. [12]

2.2 Speech Recognition Offline Processing Module

The offline processing module is a model generation process for speech recognition, which consists in detail of acoustic (pronunciation) modeling, language (grammar) modeling, and pronunciation pre-construction. Acoustic modeling is the process of modeling the properties of phonological environment-specific pronunciation through frequency analysis, and language modeling is the process of statistically modeling sentence-unit syntactic structures specific to the language of the country. [13]

2.3 Speech Recognition Online Processing Module

Online processing modules can be divided into voice end inspection, pre-processing, exploration, and post-processing technologies. The speech recognition engine performs a speech end check on consecutive input speech utterances to find the starting and ending points of speech, while the pre-processing module performs frequency analysis from speech signals to extract feature vectors reflecting acoustic features, and signal processing for noise processing. In the exploration phase, the input feature vectors are compared to the model using acoustic models, language models, and pronunciation dictionaries, which are the results of the learning phase, and the word is finally determined through scoring. In the case of Korean, the process of reconstructing the recognition result into syllables in the post-processing module is called post-processing because the pseudo-morphosis is used as a recognition unit, and the process of converting it to numbers or English is usually called post-processing. [13]

3. Experiment and evaluation

The first experiment performs a speech recognition rate experiment on Arabic numerals, from 0 to 19. It is a microphone attached to a computer that receives and stores voice sounds from 0 to 19. We store the same voice as data by repeating it 10 times.

We experiment with the recognition rate using three speech recognition free APIs for such stored data.

Table 1. Google Speech Error Rate – Number

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	sum	Error
1	4	5	5	5	3	8	3	4	7	45	22.5%

Table 1 shows the number of incorrect answers obtained using the Google Speech API and the overall incorrect answer rate. The overall wrong answer rate was 22.5 per cent. Here, we confirm that the largest number of recognition errors occur in numbers 6 and 16.

Table 2. Kakao Error Rate – Number

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	sum	Error
0	3	4	3	1	4	5	7	5	8	40	20%

Table 2 shows the number of incorrect answers obtained using the Kakao API and the overall incorrect answer rate. The overall wrong answer rate was 20.0 per cent. We confirm that the largest number of recognition errors occur in numbers 1, 2, 6, and 16.

Table 3. MS Error Rate – Number

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	sum	Error
3	3	4	3	6	9	9	7	6	6	56	14%

Table 3 shows the number of incorrect answers obtained using the MS Azure API and the overall incorrect answer rate. The overall incorrect answer rate was 14.0%, the best of the three free API recognition rates.

There were many cases in which 11 was recognised as 10.

Table 4. Google Speech Error Rate – Hangul

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	sum	Error
5	5	8	2	3	5	6	5	5	5	49	35%

Table 4 shows that the error count for character recognition using the Google Speech API is 35%. It showed a tendency not to recognize "ka", "ta", and "pa".

Table 5. Kakao Error Rate – Hangul

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	sum	Error
11	13	14	11	9	10	12	11	11	12	114	81%

Table 5 shows the number of errors in character recognition using the Kakao API, with an overall incorrect answer rate of 81%. It could be confirmed that the recognition rate has fallen significantly. Although only one character was allowed to be recognized, it appears to have shown such a low recognition rate because it was created for search or chat, given that something was added to it.

Table 6. MS Error Rate – Hangul

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	sum	Error
3	1	1	5	3	2	2	4	3	2	26	18.6%

Table 6 shows that the overall error rate is 18.6% for character recognition using the MS Azure API. The Google Speech API showed good recognition rates on "ka", "ta", and "pa", which had difficulty recognizing. So the overall recognition rate showed the best performance.

Finally, the sentence experiment confirmed the recognition rate of the sentence. The sentences used here are two words: "Text mom I'll call her later" and "Set the alarm for 9 a.m. tomorrow." The data to be used in the experiment were created the same as the methods to be generated in the previous experiments.

Table 7. Google Speech Error Rate – Sentence

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	sum	Error
0	1	2	0	1	0	1	0	1	0	6	3%

Table 7 uses the Google Speech API, showing a 3% overall incorrect answer rate. There were no more than three incorrect answers. It showed better recognition rates than numbers and letters.

Table 8. Kakao Error Rate – Sentence

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	sum	Error
0	3	1	0	2	1	2	1	5	1	16	8%

Table 8 uses the Kakao API, showing a total incorrect answer rate of 8%. It showed better sentence recognition than number or letter recognition. Finally, Table 9 shows experimental results on sentences using the Azure API.

Table 9. MS Error Rate – Sentence

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	sum	Error
1	1	2	4	1	2	3	1	5	4	24	12%

There was not enough data on Hangul yet, so the incorrect answer rate was higher than other APIs, and

above all, the resulting sentences did not fit the spelling. Given that the spacing part of Hangul is also wrong, it seems to be falling a lot about Korean.

Figure 1 illustrates the results of the above three experiments as a graph of the recognition rate. Perception of "number" has increased in order of Google, Kakao and Microsoft. The recognition rate of "letters" is in the order of MS, Google, and Kakao, while "sentences" are in the order of Google, Kakao, and MS.\

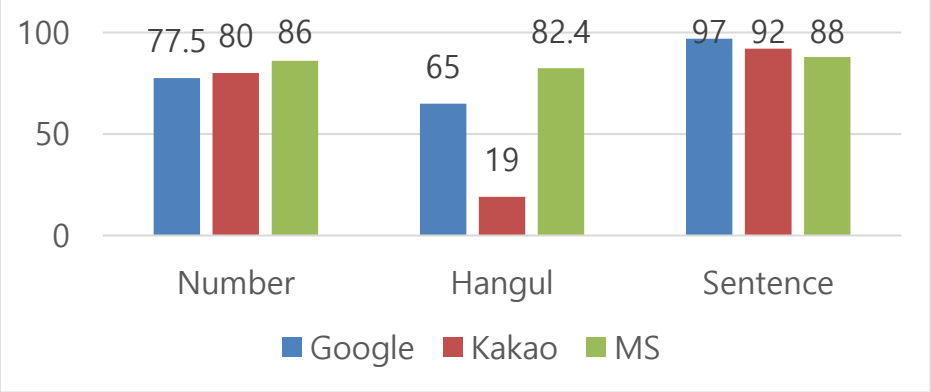

Figure 1. Recognition Rate Comparison – Number, Hangul, Sentence

Figure 2 shows a graph of the three experimental results averaged. If listed in terms of average performance, it can be listed in the order of MS, Google, and Kakao. Kakao's recognition rate seems to fall more than the other two, but this cannot be said to be directly related to the overall recognition rate because Kakao automatically changes to a word when it enters just one letter.

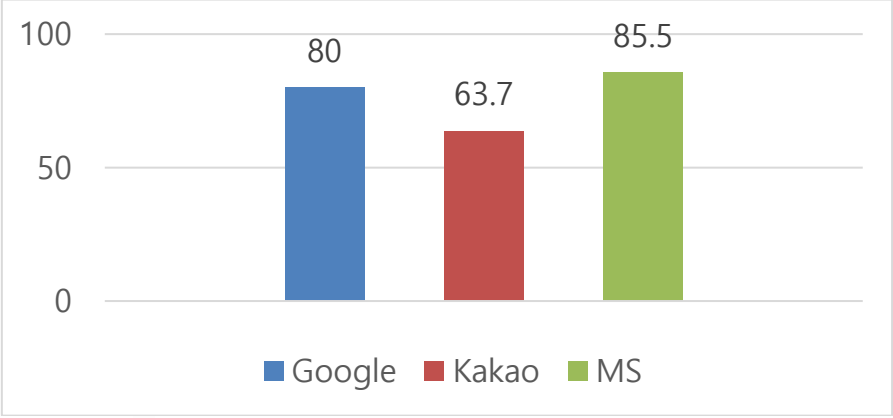

Figure 2. Recognition Rate Mean – Number, Hangul, Sentence

4. Conclusion

Speech recognition technology is a technology that converts sound voice signals obtained by computers through sound sensors such as microphones into words or sentences. I compared and analyzed the recognition rate using voice recognition API that can be easily used and supports Hangul. Based on the average recognition rate, Kakao's performance was much lower, and Google and Microsoft were similar. However, each API has its own strong and weak parts. For example, Microsoft's Azure had the best recognition rate when it needed to recognize numbers, but Google Speech had the best recognition rate when it needed to recognize sentences. In the case of Kakao, the recognition rate was not good when speaking in letters, but on the contrary, if spoken in words, the recognition rate would be high and the calibration performance was higher than that of other companies' APIs. Kakao is not that big of a company compared to Google or MS, but as it is the only Korean company, it has a better understanding of Korean language than other companies and has more data, so the

Korean language correction performance seems to be higher than other companies.

References

[1] Park, Hyeon-Sin, et al. "Trend of state-of-the-art machine learning-based speech recognition technology." *The Magazine of the IEIE* 41.3 (2014): 18-27.

[2] Chang Soo Ko. "The Prospects of Natural Language Process." Urimal, 31.0 (2012): 5-22.

[3] Qian, Yao, et al. "On the training aspects of deep neural network (DNN) for parametric TTS synthesis." *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014.

[4] Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom. "A convolutional neural network for modelling sentences." *arXiv preprint arXiv:1404.2188* (2014).

[5] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.

[6] Chetlur, Sharan, et al. "cudnn: Efficient primitives for deep learning." *arXiv preprint arXiv:1410.0759* (2014).

[7] Seung Joo Choi, and Jong-Bae Kim. "Comparison Analysis of Speech Recognition Open APIs' Accuracy." Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology 7.8 (2017): 411-418. B. Sklar, Digital Communications, Prentice Hall, pp. 187, 1998

[8] Young, Steve J., and Sj Young. "The HTK hidden Markov model toolkit: Design and philosophy." (1993): 69.

[9] Müller, Meinard. "Dynamic time warping." *Information retrieval for music and motion* (2007): 69-84.

[10] Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." *IEEE transactions on Signal Processing* 45.11 (1997): 2673-2681.

[11] Liang, Jinling, et al. "Robust synchronization of an array of coupled stochastic discrete-time delayed neural networks." *IEEE Transactions on Neural Networks* 19.11 (2008): 1910-1921.

[12] Corkrey, Ross, and Lynne Parkinson. "Interactive voice response: review of studies 1989–2000." *Behavior Research Methods, Instruments, & Computers* 34.3 (2002): 342-353.

[13] Higuchi, Takuya, et al. "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise." *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.