ISSN 2347 - 3983
Volume 8. No. 7, July 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter25872020.pdf
https://doi.org/10.30534/ijeter/2020/25872020

Comparison of Hearing Support System Core Technologies

WooLim Kim[1], Sang Boem Lim[2, *]
[1]Department of Smart ICT Convergence, Konkuk University, Seoul, Korea, kwl930521@konkuk.ac.kr
[2, *]Department of Smart ICT Convergence, Konkuk University, Seoul, Korea, sblim@konkuk.ac.kr
([*]corresponding author)

ABSTRACT

Worldwide, 34 million children under the age of 15 suffer from hearing problems, and about 100 million young people are exposed to the risk of hearing major damage as they use unsafe personal audio devices such as smart phones for long periods of time or regularly attend noisy environments such as clubs, sporting events and concerts. In this paper, we discuss Speech-to-Text (STT) and Text-to-Speech (TTS) services which are core technology of our hearing support system. We also present architecture of hearing support system.

Key words : Hearing Support System, Speech-to-Text (STT), Text-to-Speech (TTS), Voice Service Platform.

1. INTRODUCTION

According to the World Health Organization, the world's population with hearing problems, such as hearing loss or hearing loss, accounted for about 460 million people, more than 6 percent of the total population in 2019, and one in three people aged 65 or older has hearing problems such as hearing loss or hearing loss. In addition, the problem of auditory health is on the rise not only in the elderly but also in the lower age or the younger generation.[1] The half of the world's 12 to 35 years of age, or 1.1 billion people, at risk of hearing. In addition, the WHO estimates that the number of hearing-impaired people will reach about 900 million by 2050[2].

Table 1 shows the progress of Korea's disabled registration. In 2018, Korea's disabled population stood at 2,585,876 out of 51,826,059 people, 5.0 percent of the total population. The total population is the number of people registered as owners and the disabled population represents the number of registered disabled people. The number of people with disabilities increased 1.6 percent in 2018 following a sharp increase in 2017. The 2013 to 2018 trend shows that the number of people with the largest number of people with disabilities is declining noticeably, while the number of people with hearing disabilities is on the rise until 2018.

[1] This paper is based on the master thesis of WooLim Kim who is the first author of this paper at Konkuk University.

As shown in Table 1, hearing and hearing impairment are on the rise worldwide, but compared with the number of hearing impairments, the production and distribution rates of auditory medical devices, such as hearing aids, are below the global demand for hearing impairments. In addition, Korea, which has a hearing aid penetration rate of less than 17 percent, has very low hearing-related medical devices and services compared to people who suffer from hearing problems as well as hearing problems. [3]

Table 1: Registered disabled transition [3]

Year	Disabled population	Ear failure rate (%)
2013	2,501,112	10.2
2014	2,494,460	10.1
2015	2,490,406	10.1
2016	2,511,051	10.8
2017	2,545,637	11.9
2018	2,585,876	13.2 (342,582)

In this paper, we are comparing various Speech to Text (SST) and Text to Speech (TTS) technologies to provide communication mechanisms with visitors to hearing-impaired person. We also present system architecture of hearing support system.

2. RELATED WORK

There are also some papers that are using Internet of Things (IoT) devices to monitor person's health information and support better daily life. [4][5] With regard to the smart doorbell species proposed in this paper, we have looked at a total of two related studies: how to recognize the sound of the doorbell of the currently hearing-impaired, smart doorbell products sold on the market, and papers related to the doorbell.

There are two typical ways that hearing-impaired people currently recognize the sound of a doorbell. The first is hearing dogs. Just as blind guide dogs watch what they cannot see and guide them through the streets, hearing dogs listen and tell them what they cannot hear. When the doorbell rings, the

hearing dog touches the owner and sends a signal to guide the hearing-impaired to the sound, and the hearing dog leads to the sound. [6]

The second is a doorbell lamp (Figure 1). It is composed of relatively low prices for hearing-impaired Multiple lamps make it easy to install in living rooms, kitchens, etc. and show that the doorbell was pressed with vision, not sound, by repeatedly blinking through the light. The product can be used not only for hearing impaired people, but also for generations with babies, senior citizens, and test takers. [7]

Figure 1: Doorbell lamp [5]

There is a paper that implements a wearable bracelet system that uses sound detection sensors, ultrasonic sensors, and vibration sensors to identify the surrounding environment by sound detection and ultrasonic waves and then informs them through vibration (Figure 2). Noise above 40 dB for home use and more than 70 dB for external use is transmitted to mobile phone applications. Ultrasonic sensors notify users when objects are detected within 1.5M for hazards that cannot be solved by sound detection sensors alone, and when values received from each sensor enter the device using vibration motor module, they notify users by vibration. [8]

Figure 2: Sound Detection Safety Bracelet [6]

3. SPEECH-TO-TEXT AND TEXT-TO-SPEECH SERVICE COMPARISON

We compared the voice services available with Speech-to-Text (STT) and Text-to-Speech (TTS) functions to communicate with users of smart doorbell and with external visitors.

First, STT stands for Speech-to-Text, also known as Speech Recognition, which translates a person's voice language from a computer to Text data. It is mainly used when controlling a device or retrieving information by voice, and STT's algorithms typically model the voices produced by multiple people using the Hidden Markov Model (HRM) to construct an acoustic model and a language model.

TTS stands for Text-to-Speech, also known as speech synthetics, and is an engine that converts text that is written by a computer into a voice language. It is a technology that produces sound waves of voice by a machine, recording a person's voice, dividing it by a certain number of voice units, and typing it into a synthesizer with a sign, and then combining the required voice units according to the instructions to create a voice with a person. TTS is used as a screen-read software for people who are blind or cannot read text messages, such as children and foreigners.

3.1 Voice Service Platform

The platforms that provide voice services include the Newtone provided by Kakao, Android speech and cloud API provided by Google through Android studios, Clova speech provided by Naver, Amazon speech provided by Amazon, Bing speech provided by Microsoft, SwiftScribe from Baidu, and Watson from IBM.

Kakao's Newtone [9] is available for free without time or character restrictions and supports multiple languages including Korean, English and Spanish. In addition, the web search word mode, which increases the recognition rate based on words and sentences that are frequently searched in the web portal search window, supports the specialized continuous language mode that can be recognized continuously as people say, and has a recognition radius of about 0.5m. However, it is only available in the mobile application development environment.

Google's Android Speech [10] is a way to run the Google Speech Recognition app using Android studios to return the result value. Although it is available for free, voice recognition events should be set up in order to implement them in the way that developers want. It supports Korean and English and supports map mode optimized for recognizing regional names such as address, name and location so that it can be known only by its Web search word mode, continuous word mode, name, and name, and location, and an isolated language mode that selects the most similar words among the list of recognized voices and preregistered words. The recognition radius is about 0.4 meters. Also, it is only available in mobile application environments, such as Kakao's Newtone.

Google Cloud API [11] is one of the Google Cloud-based features that provides the ability to convert human voice into text and the ability to convert text into human voice. It supports a total of 110 languages and customization by providing a list of recognizable words. It can also work in both batch mode and real-time mode and filter inappropriate words in some languages. The system was created using Deep

Neural Network (DNN). Up to 60 minutes is free for all users and 0.006 USD per 15 seconds is required to proceed for more than 60 minutes. The total monthly capacity is limited to 1 million audio minutes.

Naver's Clova Speech [12] is specialized in Korean recognition and supports English, Japanese and Chinese in addition to Korean. In addition, it is an API that can be continuously learned using machine learning technology, and it can set daily and monthly limit of usage in web-based consoles and handle various management tasks such as checking usage statistics. REST API method and Android and iOS SDK are provided, which are available on various platforms such as servers as well as mobile devices. The recognizable time is recognizable for 60 seconds and the service charge must be paid 4 won per 15 seconds.

Amazon's Transcribe [13] has recently become available for Korean language as well as English and Spanish. It can add punctuation and text formats and the system can be developed to add timestamps for each word in the body, matching each word in the text to the appropriate location in the audio file. Users will be able to add names of products or other specific words It is available free for up to 60 minutes a month for 12 months after registration and requires $0.0004 per second after a period.

Microsoft's Bing Speech [14] is real-time processing capable, supports user-defined, text formatting, verbal filtering, and text normalization. It also supports a variety of speaking situations, such as conversational situations and dictation. In addition, services that convert text from text to voice can adjust a variety of voice parameters, such as gender, volume, pronunciation, speaking speed, and rhythm. 5,000 per month will be provided free of charge, and if used in addition, user will have to pay $4 per 1,000.

Watson [15] from IBM supports only English, Spanish, and Japanese and provides keyword search capabilities. Keyword search is a function that senses a user-defined string directly from the voice. In addition, it provides word substitution, timestamp, and filtering When converting text into speech, there is a function to detect the tone of a sentence and to specify the intonation, gender of the voice. However, these functions are only available in English. Word timing features allow voice synchronization with text streaming. The fee is $1,000 per month for free voice, $0.03 per minute for excess, and $1 million a month for text and $0.02 per 1,000 characters.

3.2 Voice Service Test

We would like to know the accuracy of the recognition of each platform. Before conducting the test, I looked for the most used word when using the doorbell to record a voice. Reference was made to a list of vocabulary words for learning Korean [16] based on the frequency ranking of the frequency survey provided by the National Institute of Korean Language.

Table 2 brings a list of Korean vocabulary words. A total of 5,965 words are recorded, with 982 for the first stage, 2,111 for the second stage and 2,872 for the third stage. The first stage was A, second stage was B, third stage was C, and then the words were selected. The homonym of the promoted word is based on the standard Korean dictionary and the homonym not in the standard Korean dictionary is marked as '80'. In addition to pronouns, adjectives, verbs, adverbs, nouns and exclamations, the part of speech is distinguished by proper nouns, police detectives, auxiliary usage, inability to analyze, investigation, and dependence nouns. The explanation represents a simple pool, a native word, or an example that can tell the exact meaning of the word.

Table 2: Survey on the frequency of use of modern Korean

Rank	Word	Part of speech	Rating
8	I	pronouns	A
16	No	adjectives	A
35	Come	verbs	A
36	Know	verbs	A
81	Again 01	adverbs	A
112	Next 01	nouns	A
114	Who	pronouns	A
155	Where 01	pronouns	A
196	Go outside	verbs	A
293	There 01	pronouns	A
357	Yes 03	exclamations	A
415	Put	verbs	A
722	Hold on a second	adverbs	A
1008	When	pronouns	A
1018	Later	nouns	A
1484	Sorry	adjectives	A
2607	Wait a minute	adverbs	A
4966	Appreciation	verbs	A

Table 3 is a table in which all the phrases are combined into sentences, using the words selected in Table 2 to change to phrases used in everyday life. For example, the words 'later', 'again', and 'come' changed to the words 'later', 'again', and 'come back' and combined them to 'come back again later'. In addition to the combined sentences, the most used sentences for visitors were 'That's OK' and 'Delivery'.

Voice service tests were recorded with 18 sentences shown in Table 3 and a total of 50 real-time voice service accuracy tests were conducted using recorded files. Before conducting the test, Swift Scribe, which does not support Hangul itself,

Amazon Transcribe, which does not support Hangul in real-time service, and Clova Speech, which does not have a free usage period, were excluded from the list of voice service tests.

Table 3: Test sentence

Rank	Word	Syntactic word	Sentence
8	I	It's me	It's me
16	No	No	No
35	Come	Come over	Come again next time
36	Understand	I understand	Ok, i understand
81	Again 01	Again	Come back again later
112	Next 01	Next time	Come again next time
114	Who	Who is it	Who is it
155	Where 01	from where	Where are you from
196	Go out	Go outside	I'm going out
293	There 01	There	Please leave it there
357	Yes 03	Yes, sir	Yes, sir
415	Put	Put	Please put it there
722	on a second	Hold on a second	Hold on a second
1008	When	When	When will i come
1018	Later	later	Come back again later
1484	Sorry	I'm sorry	I'm sorry
2607	moment	Wait a moment	Wait a moment
4966	Thank	Thank you	Thank you
			Delivery
			That's OK

Figure 3 shows a ratio of the number of wrong answers in the tested voice service. Since 18 sentences were tested 50 times, the total number of wrong answers was considered 900. Wrong answers were treated as wrong answers if they did not make sense in a single sentence. The ratio was calculated by translating 900 sentences and the number of wrong answers.

The voice test divided 18 sentences into one long sentence and one short sentence by dividing them into one short sentence, and drew the results together, excluding repetitive pattern sentences.

With a long sentence perceived as incorrect, Android speech showed 264 wrong answers compared to 900 sentences in total. Android speech showed accurate results for most sentences, including space writing, but it was possible to confirm that several sentences were missing without proper recognition in a series of recorded files.

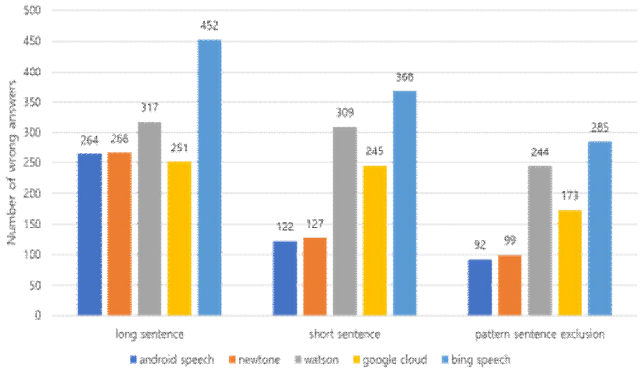


Figure 3: Comparison of wrong answers

For Newtone, the number of wrong answers was 266, showing the top line of the results around the most obvious sentences, and 10 other uncertain but likely sentences together. However, some missed sentences were printed, which were completely different from the original meaning.

Watson had 317 wrong answers and the Google Cloud API had 251. Watson also had accurate sentences, including periods and spaces, but most failed to print the first sentence correctly and repeatedly presented a wrong sentence as another sentence.

The Google Cloud API printed the least number of wrong answers among the six services tested. The Google Cloud API, like Watson, showed most accurate sentences as a result, including spacing, except for a few sentences that are not specifically recognized.

For Bing speed, the largest number of incorrect answers was 452. Bing Speech did not recognize most sentences, did not even recognize the voice, and the reaction slowed down to a halt.

Recognizing the short sentences, Android speech and Newtone displayed a significantly lower number of 122 wrong answers and 127 wrong answers compared to long sentences. Watson and the Google Cloud API, on the other hand, showed 309 and 245 incorrect answers, with no significant difference from long sentences. Bing speed dropped slightly to 368, but still showed the highest number of wrong answers compared to other voice services.

All the tests were conducted using recorded files, but because of the use of real-time voice recognition services, we printed slightly different sentences depending on the differences in the surrounding environment and some were not recognized by the pronunciation or sound quality of the test voice. To pinpoint this, out of 900 sentences, each voice service treated sentences that were not commonly recognized or misprinted as pattern words and identified the wrong answers except pattern words.

The result was a significantly lower number of wrong answers for all voice services. Android speech and Newtone had a very low number of wrong answers, 92 and 99. Watson had 244 wrong answers and Bing speech 285 wrong answers, less than 300 wrong answers from more than 300 wrong answers, while the Google Cloud API had 173, with less than 200 wrong.

However, Android speech and Newtone, which significantly lower the number of wrong answers in short sentences compared to long sentences, may vary depending on the module of the microphone that accepts the voice compared to other voice services.

For the reason, voice recognition is also possible only through mobile microphone because it is only available on mobile. In addition, unlike other voice services that can be executed by the end of recognition with a time limit, the voice service was shut down based on the point where the voice was cut off, so many long sentences were incorrect due to the failure to accept in long sentences, or problems that caused the loss of recognition, while the short sentences ended in one sentence were significantly lower.

While Android speech and Newtone showed the lowest number of incorrect answers, one of the main functions of the project proposed in this paper, Speed to Text (STT) was not appropriate because external visitors receive voice through raspberry pi and users send text through smart phone.

In order to deliver more accurate sentences to users in all of these environments, we would like to use Google's Cloud API, which is available in Raspberry Pi and has a low number of incorrect answers in both long and short sentences, as a smart doorbell STT, Text to Speech (TTS) service.

4. SYSTEM STRUCTURAL

4.1 System Architecture

All the smart doorbell system allows users to see visitors in real-time through a camera sensor that can receive alarms through a button sensor that allows them to recognize visitors from the raspberry pi and view visitors. In addition, through a microphone sensor that listens to visitors' visit purpose, the user receives it from mobile to text and converts the user's answer from text to voice to respond using a speaker sensor that informs the visitor.

Figure 4 shows the architecture of a smart doorbell species. The structure of the smart doorbell is divided into two parts: the raspberry pi and the mobile application. Raspberry pi serves as the body of a doorbell species and mobile applications are the actual part of the user's view and use. Raspberry Pi and mobile applications are connected via Wi-Fi communication.

The Raspberry Pi will show that visitors have come once it receives a button signal and will launch a real-time streaming service through a camera connected to the Raspberry Pi. In addition, streaming service images are stored with startup. When It receive a voice signal, user can chat through the voice so that user can talk to the other person, and the voice send will be saved just like the streaming service. In addition, conversations sent by the other party can be heard by voice through raspberry pi, the previously stored streaming service video and voice files can be stored in the database, and the database can be accessed through data management.

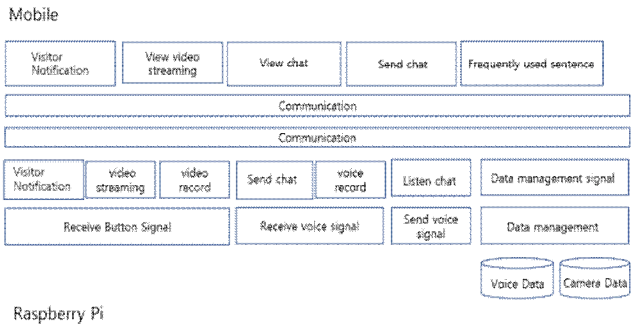


Figure 4: System architecture

The application lets users know that visitors are currently outside and watch real-time streaming videos launched through Raspberry Pi's camera. It can also send a conversation through text for identification or conversation and receive a text message from an external visitor. Users can communicate with external visitors more quickly by managing frequently used phrases in advance.

The detailed features and architecture of the raspberry pi and mobile are described in more detail in Figure 5 and Figure 6 below.

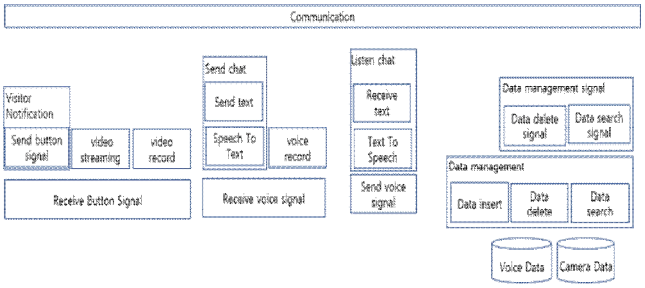


Figure 5: Raspberry pi architecture

Figure 5 illustrates the architecture of the raspberry pi in more detail. raspberry pi accepts sensor data connected to raspberry pi through servers. Sensors connected to the raspberry pi have buttons, cameras, microphones, and speakers.

The button sensor signals the raspberry pi that the button sensor was pressed at the same time as it was pressed and sends signals to camera sensors and mobile applications. The camera sensor, which received a signal from a button sensor, provides a real-time video streaming service so that users can see visitors outside through the camera and at the same time record. Recorded images are stored in the database.

Users who receive signals through mobile applications that they have visitors on the outside use a text-to-speech (TTS) library installed in the raspberry pi to read the text sent for identification and dialogue to visitors by voice through a speaker sensor, which sends the purpose of their visit through a microphone sensor, converts the text into text and sends the converted text to a text using a speech-to-to-text library installed on the raspberry pi.

At this time, the visitor's voice file is recorded and stored in the database. Images and voice files of unfamiliar external visitors stored in the database are newly registered, searched and deleted to manage data and accessible by users directly through deletion and search signals. Using this, it can be shared with others other than the police in case of a possible future situation.

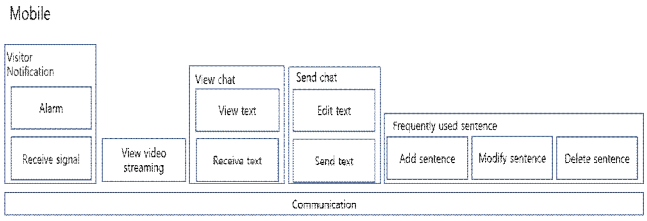


Figure 6: Mobile Architecture

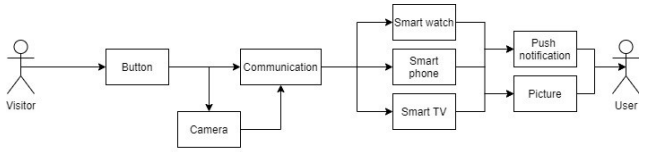
The data from each sensor shown in Figure 5 is passed to the mobile application through the communication service. Figure 6 is an architecture of mobile applications viewed by users who use smart doorbell inside the home.

With the signal received through a button sensor, the mobile phone will notify the user through a push notification, along with vibrations, so that the user can recognize that a visitor is currently outside without looking at the smart phone. In addition, video footage of visitors obtained through real-time streaming services can be viewed in real-time, such as what the current visitor looks like and what he/she does. Users who have checked the visitor's appearance can write down what they want and send it via text to ask the visitor's identity and receive the visitor's response in text. The system allows continuous dialogue with visitors without continuously opening doors.

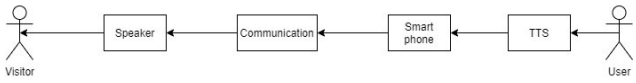
In addition to writing down the desired text in real time, users can add, modify, and delete new lists of commercial districts by pre-registering frequently used sentences. This allows users to send the desired text faster to external visitors.

4.2 System Flow Chart

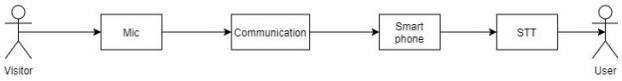
Figure 7 shows the flow of functions according to the sensor shown in Figure 5, the overall service flow diagram of the smart interracial system proposed in this paper.



(a)



(b)



(c)

Figure 7: Service flow chart

Figure 7-(a) shows that when a visitor first arrives, the visitor presses a button on the doorbell, which sends signals to the smart phone through communication, and in addition to smart phones, smart watches and smart TVs can be able to send signals. The smart phone, which received the signal, informs the user that the current visitor has come through push notification along with vibrations so that users can recognize that visitors have come without having to look at their smart phones, while the user can check the real-time images of visitors outside.

Figure 7-(b) recognizes that there is currently a visitor outside and sends text asking for the identity or purpose of the visitor from applications such as smart phones, smart watches, smart TVs, etc. using TTS (Text to Speech) installed on the Raspberry Pi to convert the text transmitted by the user to voice and send it to the visitor via speakers. Text sent by the user can be easily sent by clicking through frequent registration of a commercial phrase, without having to write text in real time.

Figure 7-c is a speaker voice that allows visitors to know if there is currently a person in the house and uses the microphone to respond to the user's questions and to voice the purpose of the visit. With the Speed to Text (STT) installed in the raspberry pi, visitors' voices can be converted to text, then transmitted to the application and users can read outsiders' responses from their smart phones to text.

Repeatedly using the flow of such services, it allows external visitors and hearing-impaired people who have difficulty communicating through voice to communicate smoothly with non-disabled visitors. It also allows communication difficulties to be improved not only for hearing impaired people but also for those who have difficulty communicating with strangers.

5. CONCLUSION

In this paper, we select most used words in daily life and combine into sentences to create to phrases used in everyday life and perform comparison tests with various Speech to Text (SST) and Text to Speech (TTS) technologies. As a test result, we choose to use Google's Cloud API as a smart doorbell STT and TTS service.

All the smart doorbell system allows users to see visitors in real-time through a camera sensor that can receive alarms through a button sensor that allows them to recognize visitors from the raspberry pi and view visitors. In addition, through a microphone sensor that listens to visitors' visit purpose, the user receives it from mobile to text and converts the user's answer from text to voice to respond using a speaker sensor that informs the visitor. Based on these requirements, we also present architecture and service flow chart of our system.

REFERENCES

1. WHO (World Health Organization), **Deafness and hearing loss**, https://www.who.int/ (Accessed on Jun. 10, 2020)
2. WHO (World Health Organization), **World Hearing Day 2019**, Mar. 2019, https://www.who.int/deafness/world-hearing-day/2019/en/ (Accessed on Jun. 12, 2020)
3. Survey Statistics Team, **2018 Survey of on the Economic Activity of Persons with Disabilities**, *Statistics report, Korea Employment Development Institute*, Nov. 2018.
4. K. U. K. Reddy and S. Shabbiha, **Design of High Security Smart Health Care Monitoring System using IoT**, *International Journal of Emerging Trends in Engineering Research*, Vol. 8, no. 6, pp. 2259 - 2265, Jun. 2020. https://doi.org/10.30534/ijeter/2020/09862020
5. W. Kim and S. B. Lim, **Smart Chair Cover for Posture Correction**, *International Journal of Emerging Trends in Engineering Research*, Vol. 7, no. 8, pp. 191-196, Aug. 2019. https://doi.org/10.30534/ijeter/2019/14782019
6. Korea Assistance Dog Association, **What guide dog?**, http://www.helpdog.org/ (Accessed on Jun. 7, 2020)
7. Rehabilitation International Korea, “Big ideas for everyday life for poem/audience disabled people 'doorbell lamp' and 'Stick of Love'”, http://www.freeget.net/main.php
8. S. Lee, S. Kim, E. Kim and T. Kang, **Sound Detection Safety Bracelet for Hearing Impaired Person**, in *Proc. The Korean Institute of Electrical Engineers. Information and control symposium*, 2018, pp246-248.
9. H. Mun and Y. Lee, **Accelerating Smart Speaker Service with Content Prefetching and Local Control**, in *Proc. 17th IEEE Annual Consumer Communications and Networking Conference, CCNC 2020*, Las Vegas, 2020, Article number 9045455. https://doi.org/10.1109/CCNC46108.2020.9045455
10. Google developers, **SpeechRecognizer**, https://developer.android.com/reference/android/speech/SpeechRecognizer/ (Accessed on Jun. 12, 2020)
11. M. Dahiya and Sonal, **Integration of gender verification mechanism to authenticate voice for google cloud speech to text**, *International Journal of Advanced Science and Technology*, Vol. 29, no. 4, pp. 496-510, 2020.
12. Naver Developers, **Clova Speech Recognition (CSR)**, https://www.ncloud.com/product/aiService/csr/ (Accessed on Jun. 10, 2020)
13. AWS, **Amazon Transcrib**, https://aws.amazon.com/ko/transcribe/ (Accessed on Jun. 8, 2020)
14. Microsoft Azure, **Cognitive Services**, https://azure.microsoft.com/ko-kr/services/cognitive-services/speech-to-text/ (Accessed on Jun. 8, 2020)
15. IBM, **Watson Speech to Text**, https://www.ibm.com/watson/kr-ko/developercloud/speech-to-text.html (Accessed on Jun. 8, 2020)
16. Korean Language Promotion Division, **Vocabulary List for Learning Korean**, *National Institute of Korean Language.*