TELKOMNIKA, Vol.16, No.6, December 2018, pp.2733~2739
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v16i6.9638

Speech Recognition Application for the Speech Impaired using the Android-based Google Cloud Speech API

Nenny Anggraini*, Angga Kurniawan, Luh Kesuma Wardhani, Nashrul Hakiem
Department of Informatics, Universitas Islam Negeri Syarif Hidayatullah Jakarta
Ir. H. Juanda St. 95 Ciputat Jakarta,15412, Indonesia
*Corresponding author, e-mail: nenny.anggraini@uinjkt.ac.id

Abstract

Those who are speech impaired (tunawicara in the Indonesian language) suffer from abnormalities in their delivery (articulation) of the language as well their voice in normal speech, resulting in difficulty in communicating verbally within their environment. Therefore, an application is required that can help and facilitate conversations for communication. In this research, the authors have developed a speech recognition application that can recognise speech of the speech impaired, and can translate into text form with input in the form of sound detected on a smartphone. By using the Google Cloud Speech Application Programming Interface (API), this allows converting audio to text, and it is also user-friendly to use such APIs. The Google Cloud Speech API integrates with Google Cloud Storage for data storage. Although research into speech recognition to text has been widely practiced, this research try to develop speech recognition, specially for speech impaired's speech, as well as perform a likelihood calculation to see the factor of tone, pronunciation, and speech speed in speech recognition. The test was conducted by mentioning the digits 1 through 10. The experimental results showed that the recognition rate for the speech impaired is about 80%, while the recognition rate for normal speech is 100%.

Keywords*: communication, google speech, speech impaired, speech recognition, speech to text*

1. Introduction

Voice communication is one of the most important basic skills that humans possess. To communicate between one individual to another individual is achieved by talking. Speaking is very important in human life, because without speaking, humans find it difficult to convey information properly with the intention that such speech can be understood by others. People should be grateful for being given the gift of the ability to speak, because there are many of our brothers and sisters who have the fate of being less fortunate and are perhaps disabled. One example of persons with a disability due to limitations in oral communication would be those who suffer from Deafness and consequently they are Tuneless. It has been said that speech impairment occurs if a person experiences abnormalities both in speech (articulation) of language and voice compared to normal talk, thus causing difficulties in communicating orally within their environment [1].

They may be able to convey messages to others through sign language or writing but not by voice. In fact, at this time many normal people who don't understand sign language and use of paper as writing media for deaf and speech impairement are very ineffective and inefficient so there will be no communication between speech impairement and normal people who dont understand sign language. One effort that can be done is to develop tools or applications that can help speech impairment with normal people to communication.

One technology that is quite well-known in the United States of America in the field of health is Medical Transcription (MT) which is available as a commercial application that uses speech recognition. To date many applications have been developed using speech recognition. Among others, in the field of health there is MT, in the military there are high-performance fighter aircraft, applications for training air traffic controllers, and tools that help people who have difficulty in using their hands, as well as the creation of computers that can be operated using user pronunciation detection. Speech recognition tools, often called speech recognisers, require actual word samples spoken by the user. The word sample will be digitised, stored in the

Received April 18, 2018; Revised September 2, 2018; Accepted October 29, 2018

computer and then used as a database in matching the next spoken word [2]. Speech to text allows a device to recognise and understand spoken words by digitising words and matching those digital signals to a specific pattern stored in a device [3].

Google Cloud Platform products one of which is Google Cloud Speech API, it allows the developers to turn audio input into text output by easily applying neural network models using the API. The API recognises more than 110 languages and variants to support a global user base. Authors can write user text dictating applications using microphones, enable voice command-and-control, or write audio files, among many other usage cases. And also, the API can recognise uploaded audio on demand, and can integrate with audio storage in Google Cloud Storage as a data storage media. A likelihood calculation can be performed to see the factor of tone, pronunciation, and speech speed in speech recognition, and the recognition rate can be observed using the Google Cloud Speech API [4].

2. Related Works

There is much previous research which is related to this study. In [5] the work intended to give an introduction to speech recognition and discuss its use in robotics. An evaluation of Google Speech using the Google speech API with regard to the word error rate and the translation speed, as well as a comparison between Google Speech with Pocketsphinx was made. Based on the overall translation times, Pocketsphinx was the better choice for implementations where low latency would be of high priority. However, Google Speech had lower Word Error Rate (WER) scores than Pocketsphinx. The results indicated that Google Speech can filter background noise more effectively than Pocketsphinx [5].

In other research [6], the activity Recognition Using a Hierarchical Hidden Markov Model (HMM) on a Smartphone with a 3D Accelerometer was conducted. In that paper, the researchers proposed a hierarchical probabilistic model-based approach to recognise the activities of a user. It was applied to acceleration data gathered from an Android smartphone. Essentially it consisted of two different kinds of probabilistic models which were continuous HMM and discrete HMM. There would appear to be many problems still to be solved in terms of mobile activity recognition such as integration of multi-modal sensor data, and modelling the variations of the user. Moreover, a comparison with other methods such as Dynamic Time Warp (DTW), and Bayesian Network (BN) is also a very important issue to be considered in future work [6].

In research into the application of pre-trained deep neural networks to large vocabulary speech recognition, the authors of that paper used a Deep Boltzmann Machine (DBN) with a pre-trained Artificial Neural Network/Hidden Markov Model (ANN/HMM) model for large vocabulary continuous speech recognition on two different datasets - 5780 hours of Voice Search and Android Voice Input data using a Compact Disc (CD) system with 7969 target states, and 1400 hours of data from YouTube using a CD system with speaker adapted features and 17552 target states [7]. Subsequently, the result obtained for an experimental setup for a Voice Search dataset showed an absolute improvement of 3.7% was observed in the WER over the baseline. For the YouTube data set, an improvement of 4.7% was observed over the baseline system after fourth epoch.

The results from the research used a Deep Boltzmann Machine (DBN) with pre-trained Artificial Neural Network/Hidden Markov Model, it indicated that ANN/HMM hybrids pre-trained with DBNs can indeed significantly outperform Gaussian Mixture Model/Hidden Markov Model (GMM/HMM) systems, even when the GMM/HMM systems was built with well-established recipes for speaker adaptive training (as was the case for the YouTube GMM/HMM baseline) and discriminative training (both GMM/HMM baselines), using much more data [7].

In further research [8], the Voice-Input Voice-Output Communication Aid (VIVOCA) was evaluated in the field trial by individuals with moderate to severe dysarthria (weak speech muscles) and confirmed that the users can make use of the device to produce intelligible speech output from disordered speech input. The VIVOCA aimed to recognises the disordered speech of the user and builds messages, which are converted into synthetic speech. System development was carried out employing user-centred design and development methods, which identified and refined key requirements for the device.

The novel methodology was applied for building small vocabulary, speaker-dependent automatic speech recognisers with reduced amounts of training data. Other research [9] employed the Google Speech API for the development of English learning media using speech

recognition technology. The learning media developed was an Android-based application. Based on previous research, the current authors intend to conduct research to build a Speech Recognition Application that can translate speech impaired dialogue and covert the audio to text by using the Google Cloud Speech API for Android.

3. System Development

3.1. Literature Review

3.1.1. Speech Impairment

According to [10], a person is speech impaired when the person is experiencing abnormalities both in the pronunciation (articulation) of language and voice in normal talk, thus causing difficulties in communicating verbally within their environment. Meanwhile, according to Frieda Mangunsong, et al. in Psychology and Education of Extraordinary Children, a speech disorder or impairment is a barrier in effective verbal communication [1].

3.1.2. Speech recognition

In a book describing artificial intelligence [2], it is mentioned that speech recognition is the process of voice identification based on the spoken word by performing a conversion of a signal, which is captured by the audio device (voice input device). Speech Recognition is also a system used to recognise the word commands of the human voice and then translate into data that can be acted upon by a computer. Sound is something that can be heard and has certain signal characteristics, while speech is a sound consisting of spoken words. Voice recognition or speech is one of the efforts required to make the sound recognisable or identifiable so that it can be utilised. Voice recognition can be divided into three approaches, namely the acoustic-phonetic oncoming, an artificial intelligence oncoming, and a pattern recognition approach. The pattern recognition approach for speech recognition can be explained with block diagram, it would be shown in Figure 1 [11].

Figure 2 shows an Automatic Speech Recognition (ASR) system architecture. It has been used in many applications [12,13], it has four components: Signal processing and feature extraction, acoustic model (AM), language model (LM), and hypothetical search. The feature processing and extraction components take an audio signal as input, improve the speech by eliminating noise and channel distortion, convert the signals from the time domain to the frequency domain, and extract vector features that stand out [14].

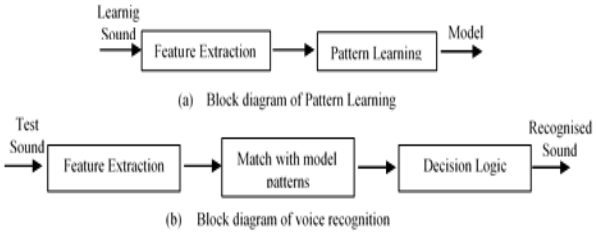

Figure 1. Speech recognition block diagram

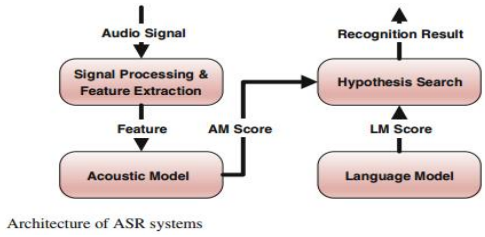

Figure 2. Architecture of an ASR System

3.1.3. Google Cloud Speech API

Machine Learning is part of the Google Cloud Platform in building applications that can hear, see, and understand the world around it. In pre-trained Machine Learning Models, the Google Translate API and Cloud Vision API, have been integrated into the Google Cloud Speech API. With such a complete API, developers can develop applications that can view, hear, and translate [9]. The Google Cloud Speech API enables the developers to turn audio into text by applying neural network models easily using the API. The API can recognises more than 110 languages and variants, to support a global user base. It is also possible to write user text by dictating using the application microphone, to enable voice command-and-control, or to write audio files, among many other usage cases by recognising the uploaded audio on demand, and integrate with the audio storage in the Google Cloud Storage [4].

3.1.4. Waterfall Model

The main stages of the waterfall model directly reflect basic development activities, According to Ian Sommerville. There are five stages in the waterfall model, which are requirements analysis and definition, system and software design, implementation and unit testing, integration and system testing, and finally, operation and maintenance [15].

3.2. System Design

3.2.1. User Interface Design

Based on the function identification at the planning stages of the requirements and modelling, the design of the speech recognition application is as follows:

a. User Interface Main Menu

As shown in Figure 3, once the user presses the "record" button then the user is asked to input the sound (voice) or to stop the input word (voice) the user presses the stop button. After the user records the voice (voice input), the application will send the voice input (data) to the Google Cloud. After processing the data on Google Cloud then the input voice will be converted to text and displayed on the device (Android), as in Figure 4.

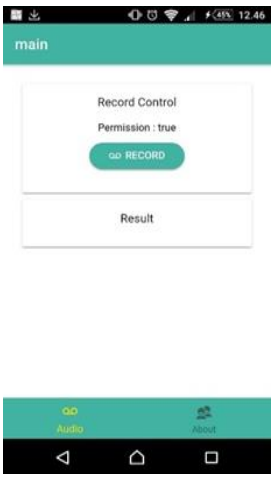

Figure 1. Main menu

Figure 4. Output view

3.2.2. Learning Pattern Recognition

This sub-section describes the sound pattern of a sentence or a word that is recorded through the open source software "Sound Meter". Computations are performed by the current authors using the method of "likelihood calculation" to calculate the pronunciation of the digits 1 to 10. The purpose of the calculation using "likelihood" is to test whether there is a relationship between the tones, pronunciation and the speaking speed. Based on the calculations contained in the "learning pattern", it is possible to analyse the likelihood values for the different tones, speech, speaking speed and the number of waves (hz) in each second.

Pronunciation of digit 1 by a normal voice which is shown in Figure 5. Table 1 represents the frequency of obsservation pronunciation of digit 1 by normal voice, while Table 2 shows the frequency of its expectation pronunciation. Likelihood value of a speech normal voice can be seen in Table 3.

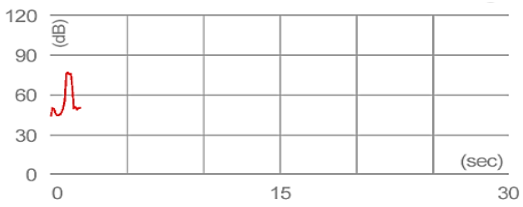

Figure 2. Histogram pronunciation digit 1 normal voice

Table 1. Frequency of Observation Pronunciation by Normal Voice

	x1	x2	x3	Total
x	1	2.8	2.9	6.7
y	45	75	50	170
Total	46	77.8	52.9	176.7

Table 2. Frequency of Expectation Pronunciation by Normal Voice

	x1	x2	x3	total
x	1.744	2.904	1.977	6.6
y	44.25	74.8	50.92	170
total	46	77.8	52.9	
	46 * 170 / 176.7 = 44.25		46 * 6.7 / 176.7 = 1.744	

Table 3. Result of Likelihood Pronunciation of Digit 1 by Normal Voice

	x1	x2	x3
x	-0.54	-0.2	1.111
y	0.725	0.204	-0.91
likelihood	0.774		
	1 * LN (1/1,744) = -0.54	Sum(table result) * 2 = 0.744	

The authors assume that s=x and db=y. Each maximal lower curve and upper maximal curve are assumed to be x1, x2,.. etc. If there is a pronunciation of the digit number 1 by a person who is speech impaired with a different speech, tone and speech speed, then Figure 6 illustrates the translation. Table 4 shows the frequency of obsservation pronunciation of digit 1 by a speech impaired voice, while Table 5 demonstrates the frequency of its expectation pronunciation. Likelihood value of a speech impaired voice can see in Table 6.

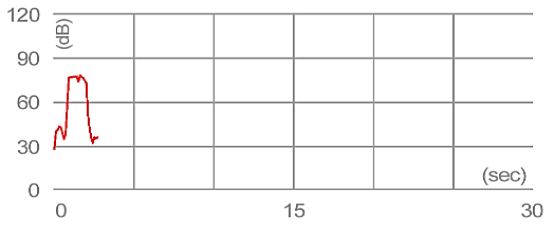

Figure 3. Histogram pronunciation of digit 1 by a speech impaired voice

Table 4. Frequency of Observation of Pronunciation of Digit 1 by a Speech Impaired Voice

	x1	x2	x3	Total
x	1	2.9	3.2	7.1
y	45	75	50	170
Total	46	77.9	53.2	177.1

Table 5. Frequency of Expectation Pronunciation of Digit 1 by a Speech Impaired Voice

	x1	x2	x3	Total
x	1.844	3.123	2.133	7.1
y	44.16	74.78	51.07	170
Total	46	77.9	53.2	

Table 6. Result of Likelihood Pronunciation of Digit 1 by a Speech Impaired Voice

	x1	x2	x3
x	-0.61	-0.21	1.298
y	0.852	0.223	-1.056
	likelihood value		0.982

It can be seen that there are different likelihood values in each table. The factors tone, pronunciation, and speech velocity all affect the likelihood value. There is a difference of likelihood value of a speech impaired voice of 0.982 as seen in Table 6, while normal voice is 0.774 which can be seen in Table 3.

3.3. Implementation

There are many steps in using the Google Cloud Speech API. The authors used Python encoding. The following are the steps:

a. Install Python and PIP. Python is used as a means of processing Google Speech and also the Speech API. PIP functions as a manageable package or python libraries.
b. Installing the modules needed by python for speech recognition. In order to connect apps with the Google Speech API Bridge, the APIs are required to bridge sending and retrieving data to the API. In addition to requiring the API Bridge the Speech Process module is also required to process Speech Recognition using Google Speech.
c. Covert audio, audio included in Google Cloud Storage in the form of extension .wav and mono type.
d. Configure the Google API Key

3.4. System Testing

Testing was conducted by speaking the numbers 1 to 10, performed by three speech impaired voices and three normal voices and the average recognition rate obtained by the use of Equation (1).

$$RR = \frac{NCorrect}{NTotal}\ x\ 100\% \quad (1)$$

where:
RR is the Recognition Rate
NCorrect is the Number of correctly recognised spoken digits
NTotal is the Total number of samples of spoken digits

For the blackbox testing, assessment is performed by knowing the function that has been determined, and the output examined to show whether the function is fully operational or otherwise. The test is performed by running the system and observing its output. Table 7 is a table of blackbox testing results that has been built for the ASR.

Table 7. Blackbox Testing

#	Desciption	Expected Result	Result
1	Run with a normal voice input by mentioning the numbers one through ten	The system can display output text	success
2	Run with a speech impaired voice input by mentioning the numbers one through ten	The system can display output text	success
3	Add audio to Google Cloud	The system can add sound recording (audio) to Google Cloud	success
4	Run recorded audio	The system can record sound and play recordings	success

4. Result

Based on the calculations contained in the "learning pattern", the likelihood values can be analysed based on the different tones, speech and speaking speed and the number of waves (Hz) in each second. Table 8 shows a comparison of patterns.

Table 8. Pattern Comparison

Number	likelihood	Hz/sec
1 (normal voice)	0.774	3
1 (speech impaired voice)	0.982	3
9 (normal voice)	2.445	5
9 (speech impaired voice)	6.851	5
2,3 and 4 (normal voice)	13.46	9
2,3 and 4 (speech impaired voice)	43.96	9

As can be seen in each row of the Table 8, there are different likelihood values. The factors of tone, pronunciation, and speech speed affect the likelihood value. After determining the number of correct and incorrect results for three normal voices and three speech impaired voices to recognise spoken numbers 1 to 10, then Equation (1) was used to obtain the

recognition rate of the application, and the results are shown in Table 9. Table 9 shows the recognition rate for speech impaired and normal voices. For the normal voices the success rate of the speech recognition application was 100% and for the speech impaired voices the recognition rate was 83.3% up to 90%. The test was performed three times to obtain a greater amount of valid data. Based on the learning pattern by calculating the likelihood value this aims to prove the tone, pronunciation, and the speed of speech affect the recognition rate result. And with these results can be one of the developments for speech recognition research, specially for google cloud speech technology.

Table 9. Recognition Rate Result

	T1	T2	T3	N1	N2	N3
Test 1	80	90	80	100	100	100
Test 2	80	80	90	100	100	100
Test 3	90	100	90	100	100	100
Mean Recognition Rate (%)	83	90	86.6	100	100	100

5. Conclusion

In this paper based on the results of the research as observed by the authors, it can be concluded that the Speech Recognition Application using Google Cloud Speech can recognise and translate the speech of the speech impaired in terms of the digits one to ten. A recognition rate of 80% was obtained for the speech impaired and 100% for normal voice speech recognition when speaking the numbers 1 to 10. It can also be concluded that the way a voice speaks into the speech system has some effect on the level of recognition and that there are three factors that affect recognition rate, namely tone, pronunciation, and speech speed.

References

[1] FMM Siahaan. Psychology and Education of Children with Special Needs First Volume (in Indonesia Psikologi dan Pendidikan Anak Berkebutuhan Khusus Jilid Kesatu). Depok, LPSP3-UI. 2009.
[2] V Amrizal, Q Aini. Artificial Intellegence (in Indonesia Kecerdasan Buatan). Jakarta Barat. Halaman Moeka Publishing. 2013. E Widiyanto, SN Endah, S Adhy. *Speech Application to Text in Bahasa using Mel Frequency Cepstral Coefficients and Hidden Markov Models (in Indonesia Aplikasi Speech To Text Berbahasa Indonesia Menggunakan Mel Frequency Cepstral Coefficients Dan Hidden Markov Model).* Prosiding Seminar Nasional Ilmu Komputer Undip. 2014: 39-44.
[3] Google Speech. Cloud Speech API. *Google Cloud Platform*. Available: https://cloud.google.com/speech/. [Accessed: 01-Jan-2017].
[4] M Stenman. Automatic speech recognition An evaluation of Google Speech. UMEA UNIVERSITY. 2015.
[5] YS Lee, SB Cho, D Muly. Activity Recognition Using Hierarchical Hidden Markov Models on a Smartphone with 3D Accelerometer. *Hybrid Artif. Intell. Syst.* 2011: 460–467.
[6] N Jaitly, P Nguyen, A Senior, V Vanhoucke. *Application of Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition*. 13rh Annu. Conf. Int. Speech Commun. Assoc. 2012: 2–5.
[7] MS Hawley, SP Cunningham, PD Green, P Enderby, R Palmer, S Sehgal, P O'Neill. A voice-input voice-output communication aid for people with severe speech impairment. *IEEE Trans. Neural Syst. Rehabil. Eng.* 2013; 21(1): 23–31.
[8] D Intan, S Saputra, SW Handani, GA Diniary. Utilization of Cloud Speech API for the Development of English Language Learning Media using Speech Recognition Technology (in Indonesia Pemanfaatan Cloud Speech API untuk Pengembangan Media Pembelajaran Bahasa Inggris Menggunakan Teknologi Speech Recognition). *TELEMATIKA.* 2017; 10(2): 92–105.
[9] H Purwanto. Ortopedagogik Umum. Yogyakarta. IKIP Yogyakarta. 1998.
[10] HA Khalilulah. Implementation of Speech Recognition on Android-based English Idiom Translator Application (in Indonesia Implementasi Speech Recognition pada Aplikasi Penerjemah Idiom Bahasa Inggris ke Bahasa Indonesia Berbasis Android). *Jurnal Teknologi dan Sistem Komputer*. 2016: 1–6.
[11] SN Endah, S Adhy, S Sutikno. Comparison of Feature Extraction Mel Frequency Cepstral Coefficients and Linear Predictive Coding in Automatic Speech Recognition for Indonesian. *TELKOMNIKA Telecommunication Computer Electronics and Control.* 2017; 15(1): 292.
[12] S Nafisah, O Wahyunggoro, LE Nugroho. An Optimum Database for Isolated Word in Speech Recognition System. *TELKOMNIKA Telecommunication Computer Electronics and Control*. 2016; 14(2): 588.
[13] D Yu, D Li. Automatic Speech Recognition, A deep learning approach. 2014; 17: 1–3.
[14] I Sommerville. Software Engineering 10th Edition. *Software Engineering.* 2015.