Journal of Computer Science

Original Research Paper

Speech to Text in Indonesian Personal Assistant

[1]Intan Sari Areni, [2]Ayu Dhiya Mufidah, [2]Indrabayu, [2]Sri Wahyuni, [2]Ingrid Nurtanio, [1]Ida Rachmaniar Sahali and [2]Anugrayani Bustamin

[1]Department of Electrical Engineering, Universitas Hasanuddin, Indonesia
[2]Department of Informatics, Universitas Hasanuddin, Indonesia

Article history
Received: 20-03-2019
Revised: 22-05-2019
Accepted: 18-06-2019

Corresponding Author:
Intan Sari Areni
Department of Electrical Engineering, Universitas Hasanuddin, Indonesia
Email: intan@unhas.ac.id

Abstract: Short Message Service (SMS) is one of the most often used features on smartphones. Delivery of SMS messages while driving can interrupt driver's concentrations that may lead to an accident. Hence, Speech Recognition system in SMS message activity is required. In this research, the Speech Recognition system is able to convert speech as an Indonesian query in making messages, entering contacts without searching the phone list and equipped with push button for sending a message using Google Speech Recognition Application Programming Interface (API) System created using Java programming language with Android Studio Editor. The input data consist of training and testing data. Training data used is 20 voice data samples on STT message that consist of 10 different male voice and 10 different female voice samples for 7 similar words. While for testing data, 10 voice data samples are used that consist of 5 different voice samples for male and female. System performance based on Result Training Data (RTD), the Result Random Data (RRD) and Grade Success System (GSS). The results for send message show that RTD, RRD and GGS reach 100%, 96.74% and 98.37%, respectively. For add contact, the performance system obtained 100% for all parameters.

Keywords: Personal Assistant, Speech Recognition, Short Message Service, ASR, Indonesian Language

Introduction

The rapid development of telecommunication technology over the last decade and the technology of Short Message Service (SMS) is one of the most frequently used features on a smartphone. However, writing and reading a message while driving can cause traffic accident.

Therefore, Text to Speech (TTS) and Speech to Text (STT) technologies for SMS application is developed for various languages. In 2017, authors' have made TTS application to read message in Indonesian and also added an abbreviation reader feature that frequently used on SMS. The user can add more abbreviation into the application's database (Areni *et al*., 2017a). To complete the previously made TTS system, authors' added STT feature to send a message and to read contact's name in the smartphone. This selected feature is based on a questionnaire from 120 respondents where 80.25% experiencing difficulties in typing SMS message and searching name in phone contact while driving. By implementing Indonesian speech recognition, the difficulties can be overcome.

Speech Recognition is a technique to recognize command words from human voice and translate them into understandable data by system. The benefit of Speech Recognition can be observed from its speed and ease of use. Those words are converted into digital signal by converting voice waves into several numbers then adjusting it with certain codes and comparing it with a stored pattern. The compared parameter is voice suppression that will be paired with the available template database (Areni *et al*., 2017a).

Research about STT in various languages has been conducted in earlier years by utilizing Google Application Programming Interface (API). Iizuka *et al.* (2012) made system enhancement and service enhancement in Speech Recognition "VOICE IT!" technology by applying language processing to the sound signal processing (Iizuka *et al*., 2012).

Gauthier *et al.* (2016) developed an ASR application for speech services in various sectors with Africans. This research compares the application of vowel length contrast on voice recognition. The human voice has complex characteristics. Some factors that might affect those characteristics are sex, age, health, etc. Kwon *et al.* (2015) have done early stage research of preprocessing optimization on elderly voice recognition by using a smart device. Some voice recognition researches up to this day can effectively process adult voice but not elderly. Their voice pattern is much slower and the level of clarity in articulation is very low. Based on that research, there is an increasing of elderly voice recognition process for up to 12% (Kwon *et al*., 2015). Speech recognition research for Indonesian language has been performed by Bustamin *et al*. (2016; Areni *et al*., 2017b) using the Mel Frequency Cepstral Coefficient (MFCC) method for feature extraction on the word homophone. Cavus (2016) has done a research regarding to intelligent mobile application for learning English (pronunciation) by using voice recognition. The features presented in the application are succeeded in increasing student's motivation and enthusiastic in learning process compared to the traditional way (Cavus, 2016). Furthermore, the voice recognition app for disabled was also developed by Abdallah and Fayyoumi (2016). This Assistive Technology could facilitate the communication process between disabled with sign language recognition that is integrated with Arabic voice recognition. The testing process of the system is done with 3 steps, i.e. checking SDK functionality performance, installing the apps on various types of smartphones and filling out the questionnaire form from 10 deaf and 15 normal participants with percentage of user satisfaction by 95% (Abdallah and Fayyoumi, 2016).

In the development of interface Speech Recognition technology, there are several challenges that must be addressed to develop a device that is comfortable and useful for the human communication process. Based on Mittal and Navdeep (2016) some of these challenges include processing power, memory usage, an accuracy of speech signal recognition, use of time and energy consumption (Mittal and Navdeep, 2016). Furthermore, similar technology was developed by Yousaf *et al.* (2018) that is called Vocalizer to Mute (V2M). The novelty presented in this research is the integration between speech recognition and 3D avatar animation visualization to support deafmute communication. The used method is Mel Frequency Cepstral Coefficient (MFCC) for feature extraction of training, Hidden Markov Model (HMM) Toolkit for recognition process and also 3D avatar to create sign language visualization. The result obtained is 97.9% for 15 participants from deaf-mute children social foundation (Yousaf *et al*., 2018).

Research conducted in this paper is the development of TTS application on (Areni *et al*., 2017b) by adding a Speech Recognition system with an Indonesian query to create messages and enter contact names. The application is created using the Google Speech API with Java programming language using Android Studio platform.

Materials and Methods

Personal Assistant (PA) has been developed by our research group since 2017 by developing TTS application on receiving SMS with Indonesian abbreviation reading feature. This TTS system can also be updated with abbreviation that is not included in the default database (Areni *et al*., 2017a). The activity diagram of the TTS system is shown in Fig. 1. In this research, the PA is added Speech to Text (STT) technology for sending SMS and searching for phone contact features.

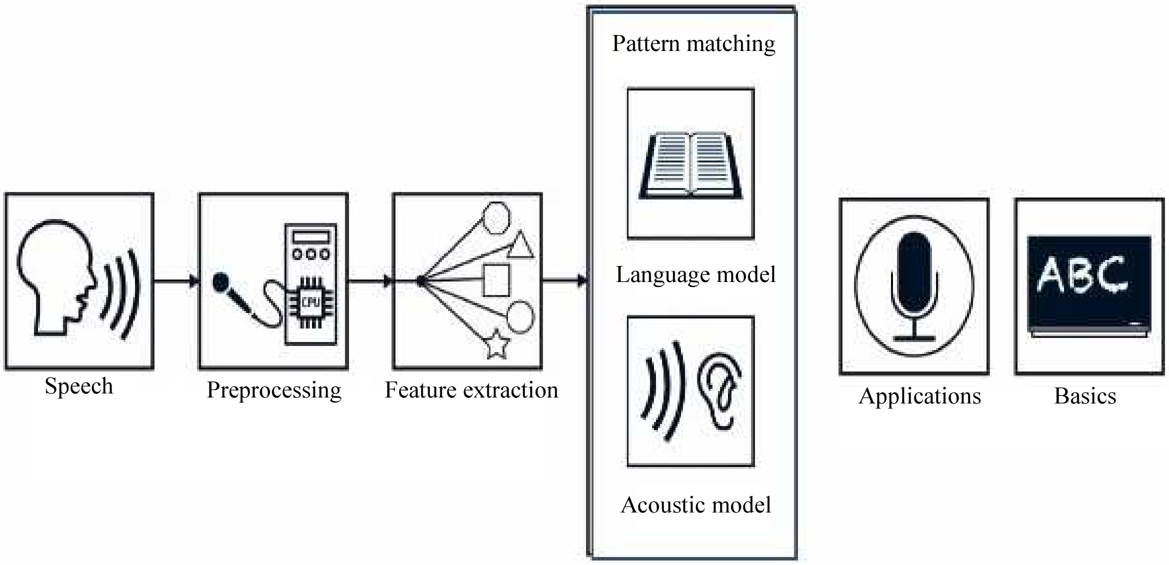

Fig. 1: Speech recognition scheme

Speech Recognition is the process of recognizing words from human speech that is converted into text. Human voice waves will be converted into a set of numbers based on certain codes and adapted to a pattern stored in a device. For each different utterance, different characteristic patterns will be generated. There are two modes on the speech recognition system, namely: Dictation Mode and Command and Control Mode. In dictation mode, users can say a word or phrase that will be recognized by the machine and converted into text. The possible number of recognized words is limited depending on the words in the database. The introduction of this mode depends on the speaker's sound pattern and accent. In Command & Control mode, the user speaks a pre-defined word/phrase in the database that will be used to execute certain commands on the application. The number of recognized commands depends on the application that has been defined first in the database. This mode is an independent speaker because the number of recognized words is usually very limited. There is a possibility that the speaker does not need any training system. The scheme of speech recognition is shown in Fig. 1 (Chelba *et al*., 2012).

The using of Google Speech Recognition API allows developers to convert speech into text. This service can be processed offline and online. However, in the offline processing, supported languages are limited depending on the language of each device/smartphone. In addition, this feature cannot be operated on some versions of Android. In contrast to online processing that allows complete language support and can be operated almost in all versions of Android. However, the using of this service requires an internet connection because the Speech Recognition process takes place on Google's servers.

The Google Speech API is a framework developed by Google to recognize sounds, convert them into strings (text) and put them into Google search pages so that search results will appear based on voice input. Speech recognition is performed on Google servers using the Hidden Markov Model (HMM) algorithm and developing it using large *n*-gram language models (Reddy *et al*., 2013).

The steps of the speech recognition system in this study are described as follows.

Input Data

The data used in this system consist of 2 types of voice data, i.e., training and testing data:

- Training Data. 20 voice data samples on STT message are used as training data that consist of 10 different male voice and 10 different female voice samples that speak out 7 similar words. In total, there are 70 voice samples used as training data
- Testing data. Testing data is the data that has not been trained into the system so the system is expected to recognize the words that will be inputted. 10 voice data samples are used as testing data that consist of 5 different male voice and 5 different female voice samples

Speech to Text (Message)

To compose an SMS by using STT android based, an architecture model is designed by internet connection to Speech Recognition Google server database through Application Programming Interface (API) Speech Recognition inside an Android smartphone. Google has Speech Recognition program. The Google's data come from voice recording and written query search to predict words that are possible to be spoken by people. Google uses two HTTP connections, the first one is a request to upload a voice signal to Google server and the second is a request to access the recognition result. Then, Google data center will use a certain statistical modeling to determine the contents of the words spoken.

Google Speech API is called by using ACTION_RECOGNIZE_SPEECH framework. This framework is implemented using onActivityResult() method. Next, this library will get the voice through the smartphone's microphone.

The process of voice recognition to text in composing an SMS message using Google Speech API is described as follows:

a. Inside Google Speech Recognition database, the user firstly needs to hold mic icon to start recording and input their voice in Indonesian to the android smartphone in SMS service that supports Speech Recognition API. Then the voice signal will be transferred into Speech Recognition server. The user's voice signal will be treated by several stages of the voice recognition process to text from the user's smartphone to the connection of Google's Speech Recognition database server. The voice recognition process is carried out on Google servers using the Hidden Markov Model (HMM) algorithm
b. Google Database Speech Recognition server will send an output result in a form of text to the user by a transmission medium. After the result is found, then the type of text based on Indonesian that is previously inputted will be sent back to the user as an output with the desired Indonesian text. If the text output does not match with the desired results, the user can perform the voice recognition process on Android smartphone devices like the previous step

Diagram Activity Text to Speech for Abbreviation in SMS Text is shown on Fig. 2 and STT process to send SMS is shown by an activity diagram that can be seen on Fig. 3.

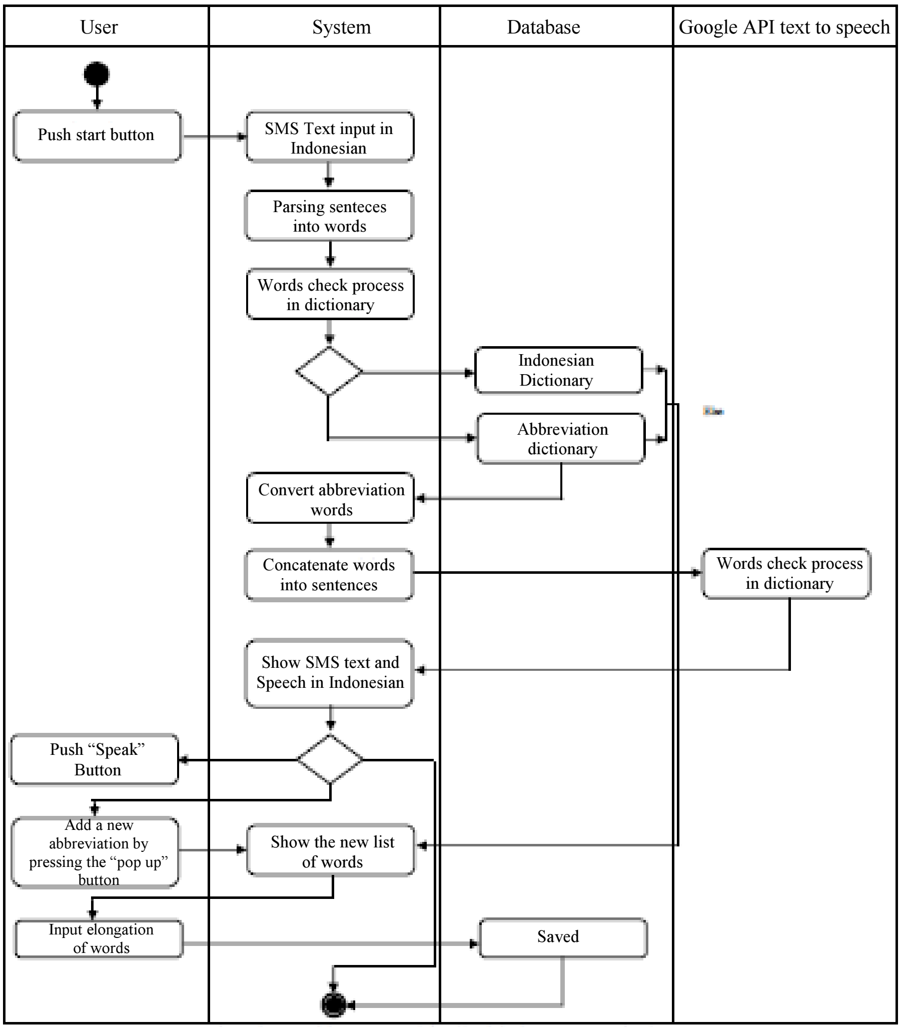
User
System
Database
Google API text to speech
Push start button
SMS Text input in Indonesian
Parsing senteces into words
Words check process in dictionary
Indonesian Dictionary
Abbreviation dictionary
Convert abbreviation words
Concatenate words into sentences
Words check process in dictionary
Show SMS text and Speech in Indonesian
Push "Speak" Button
Add a new abbreviation by pressing the "pop up" button
Show the new list of words
Input elongation of words
Saved

Fig. 2: Diagram activity text to speech for abbreviation in SMS text (Areni *et al*., 2017a)

Google Speech API Server Call

A server system is responsible for voice processing cycle. Voice recognition is done on the server side. Google is able to translate hundreds of words including Indonesian. To translate words into Indonesian nicely, the smartphone language needs to be set to Indonesian first. Control in Indonesian language is done by EXTRA_LANGUAGE_MODEL library.

In the process, to convert voice into text, there are several steps that need to follow:

a. User voice as analog signal converted by the device into a discrete one then it is changed into binary and sent to the server real-time for further conversion.
b. After user stops talking (stop recording or end detect) then server receives all conversation data in digital/binary form and does the conversion.
c. After the conversion is completed, the server will send the result in the string form to the device

In order to use Speech Recognition API, importing android.speech.RecognizerIntent class is needed Basically, Intent initialization from RecognizerIntent class is used to display "Speak now/Coba Ucapkan Sesuatu" dialog box to recognize input voice. Next, by calling startActivityForResult (REQUEST_OK) method, the system will send the capture result real-time and wait for the serve to respond. Analog to Digital Converter (ADC) is included in this process. If the respond is "REQUEST_OK", the system will run getStringArrayListExtra() method to obtain the conversion result in the form of ArrayList. The Intent initialization from RecognizerIntent class EXTRA_LANGUAGE_MODEL is needed as an extra value. On this application, LANGUAGE_MODEL_FREE_FORM is used as an extra value, meanwhile optional extra value used on EXTRA_PROMPT and EXTRA_LANGUAGE parameters is obtained native language selection. EXTRA_PROMPT parameter value that will be used is an Indonesian string and EXTRA_LANGUAGE parameter value is id. When Intent initialization is running on class then system will start the Indonesian voice recognition.

Speech to Text for Add Contact

In this phase, user needs to input contact name using Speech to Add Contact button by saying the contact name in accordance to the saved contact in the smartphone. If system succeeds in searching the contact name that user request, then the name and the phone number requested will be displayed in accordance with the saved contact in the smartphone. In order to read the phone contact on the smartphone android.permission. READ_CONTACTS is needs to be added on AndroidManifest.xml using CONTACT_PICKER_RESULT framework to create smartphone contact number.

The Activity diagram for the Speech to Add Contact process is shown in Fig. 4. When the user has written a short message and it will immediately pass the process of sending a message, it will do speech recognition to recognize the name of the recipient message.

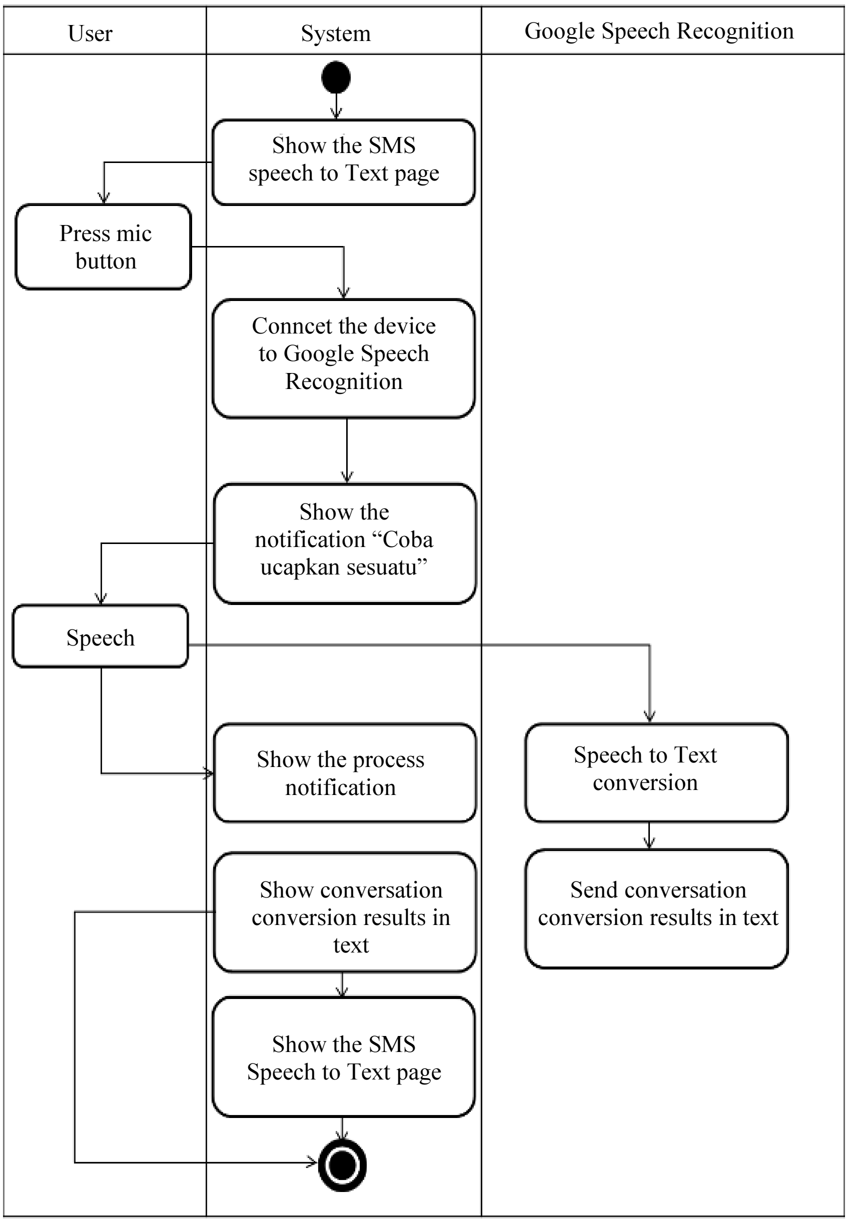
User
System
Google Speech Recognition
Show the SMS speech to Text page
Press mic button
Conncet the device to Google Speech Recognition
Show the notification “Coba ucapkan sesuatu”
Speech
Show the process notification
Speech to Text conversion
Show conversation conversion results in text
Send conversation conversion results in text
Show the SMS Speech to Text page

Fig. 3: Activity diagram of the STT to write SMS

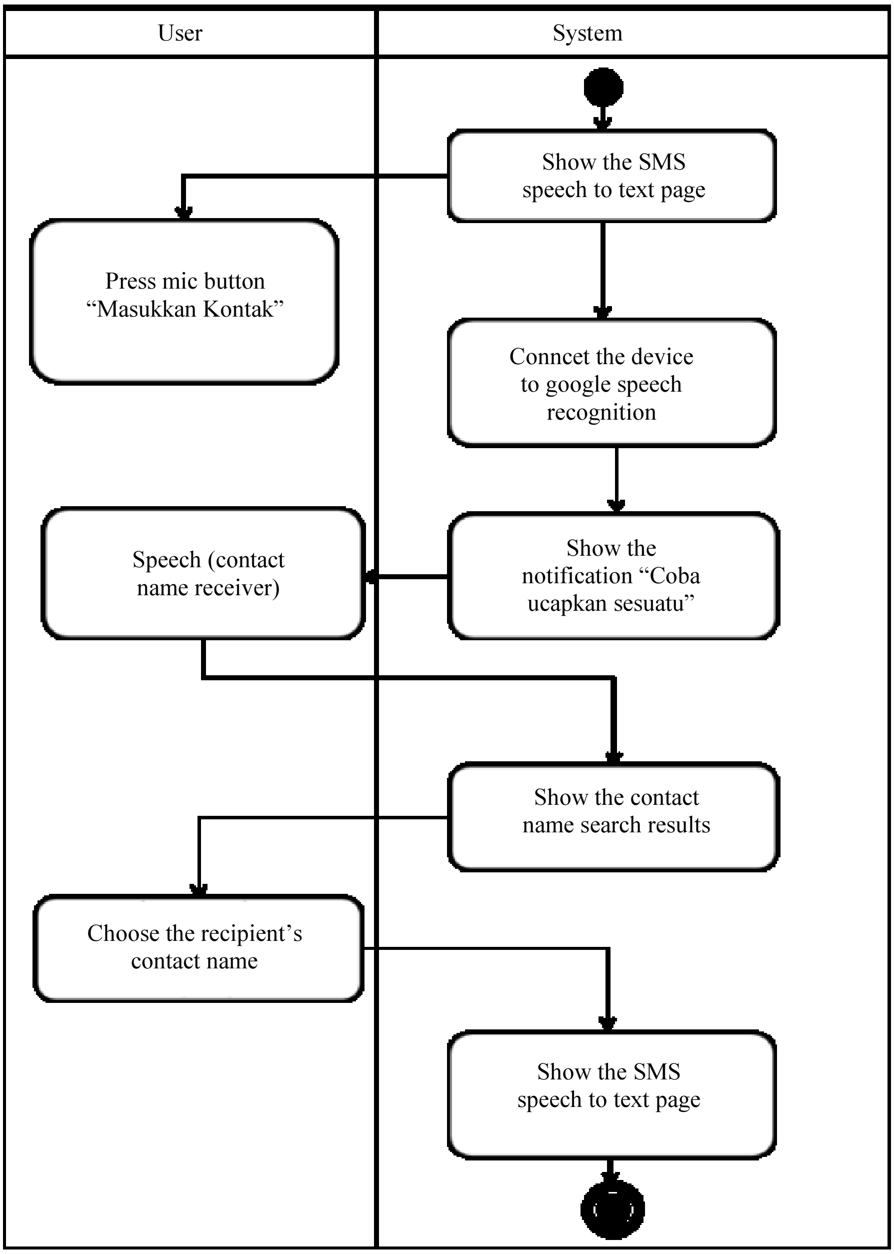
User
System
Show the SMS speech to text page
Press mic button "Masukkan Kontak"
Conncet the device to google speech recognition
Show the notification "Coba ucapkan sesuatu"
Speech (contact name receiver)
Show the contact name search results
Choose the recipient's contact name
Show the SMS speech to text page

Fig. 4: Activity diagram for speech to add contact

Database Design

Database design in this system is used to support data processing facilities. SQLiteOpenHelper is the used tool on the database.

Send Message

In this phase, sending SMS message can be done in two ways, by using speech command and by pushing send button. The first way is by using user voice command, by saying Indonesian query "kirim" then the system will send the message automatically. This process works by activating Recognition Service which is a component running behind the application. This service runs in the main thread of the application.

On the send button using SMS manager, there's a special class on Android OS that specifies to access SMS features in Andorid. This class is used to send SMS to the destination number. When using SMS manager, the SMS will be sent from the apps without using other apps. It needs to call send Text Message() method and then input the parameter in the form of destination phone number, sender phone number, SMS content, executed intent when the SMS is delivered or failed to send and executed intent when the SMS is delivered. Sending SMS will need some access to the feature by declaring SEND_SMS permisson on and roidManifest.

Validation

System testing is performed by using training and testing data taken randomly. The validation process will be calculated on a percentage scale represented by Result Training Data (RTD), Result Random Data (RRD) and Grade Success System (GSS), each equation shown in Eqs. (1-3):

$$RTD = \frac{TDsuccess}{nTD} \times 100\% \quad (1)$$

$$RRD = \frac{RDsuccess}{nRD} \times 100\% \quad (2)$$

$$GSS = \frac{RTD + RRD}{2} \quad (3)$$

Where:
Tdsuccess = The number of training data successfully classified
nTD = Total number of training data
Rdsuccess = The number of random test data successfully classified
nRD = The total number of random test data

RTD is the result of a previously trained data test, whether the system can recognize correctly, not recognize correctly, or not recognize it at all. While Result Random Data is the result of data test that is not yet recognized by the system so that the system is expected to recognize the data well in the input, although learning about the pattern of input data has never been done previously. The results obtained from the *RTD* and RRD will be averaged to obtain the overall *GSS*.

Results and Discussion

The software used is Android Studio 2.3.3.version android Software Development Kit (SDK) Manager and Java (jdk-8u121-windows-x64). The application is tested using Android version 5.1.1 (Lollipop).

System performance is based on the success of generating contact numbers and typing messages with Indonesian speech input. Trials were conducted by 10 men and 10 women. The results show that the system accuracy is 100% if the pronunciation and articulation match with Google Speech API database.

Performance analysis system is the main parameter to know the success rate of system design. These parameters are RTD, RRD and GSS. Table 1 shows the process of validating speech to text systems for typing messages and speech to add contacts using testing data. The results of Table 2 based on Eq. (3).

Table 1: Validation system with RTD & RRD parameters

Validation system	Message	Add contact
Result Training Data (RTD)	100%	100%
Result Random Data (RRD)	96.74%	100%

Table 2: Time execution process of the system

No	Character	Duration (sec)
1	165	29
2	163	26
3	160	24
4	158	19
5	151	16

In addition, testing the conversion time of speech to text is also completed based on the character length of the spoken message. The test results are shown in Table 2 with a conversion time range of 16-29 sec where the number of characters is 151-165.

Conclusion

Speech Recognition application on SMS delivery with Indonesian query designed and built using Android Studio Java programming language has been performed in this research. This application has been able to convert speech into text in writing SMS, adding phone contacts and can send messages through Speech Command. The success rate of the system is 100% for trained sound test and 98.37% for the untrained randomized sound test. While the system success rate is based on the test conducted on test spech test to enter the name of the SMS recipient contact is 100%. The test results show that this application is able to recognize the spoken voice and successfully send SMS. The system success rate for speech recognition reaches 99% with the Grade Success System parameter. Unrecognized words are caused by the intonation of sound, pronunciation and poor articulation during speech data retrieval.

For the development of the Personal Assistant system with a query in the Indonesian language, a database of Indonesian corpus which in its development also needs the best method will be made. This application is also expected to be useful for smartphone users who experience physical disability, so that the command extension on the application is required. In addition, this application is expected to cover all the commands on the smartphone, so it can help human interaction with computers becomes easier.

Acknowledgment

This research is supported by PDUPT-RISTEKDIKTI with contract number of 1569/UN4.21/PL.00.00/2018, LPPM UNHAS and AIMP Research Group.

Author's Contributions

All authors equally contributed to this work.

Ethics

This article is the original contribution of the authors and is not published elsewhere. There is no ethical issue involved in this article.

References

Abdallah, E. and E. Fayyoumi, 2016. Assistive Technology for deaf people based on android platform. Procedia Comput. Sci., 94: 295-301. DOI: 10.1016/j.procs.2016.08.044

Bustamin. A., I.I.S. Areni and N.N. Mokobombang, 2016. Speech to text for Indonesian homophone phrase with mel frequency cepstral coefficient. Proceedings of the International Conference on Computational Intelligence and Cybernetics, Nov. 16, IEEE Xplore Press, pp: 29-32. DOI: 10.1109/CyberneticsCom.2016.7892562

Areni, I.S., S. Wahyuni, Indrabayu and Anugrahyani, 2017. Solution to abbreviated words in text messaging for personal assistant application. Proceedings of the International Seminar on Application for Technology of Information and Communication (iSemantic), Nov. 17, IEEE Xplore Press, pp: 238-241. DOI: 10.1109/ISEMANTIC.2017.8251876

Areni. I.S., Indrabayu and A. Bustami., 2017. Improvement in speech to text for bahasa Indonesia through homophone impairment training. J. Comput. (Taiwan), 28: 1-12. DOI: 10.3966/199115992017102805001

Chelba, D., M. Bikel, P.S. Nguyen and S. Kumar, 2012. Large scale language modelling in automatic speech recognition. Computat. Language. arXiv preprint arXiv: 1210.8440

Gauthier, E., L. Besacier and S. Voisin, 2016. Automatic speech recognition for african languages with vowel length contrast. Procedia Comput. Sci., 81: 136-143. DOI: 10.1016/j.procs.2016.04.041

Yousaf, K., M. Zahid, S. Tanzila, R. Amjad and R. Muhammad *et al*., 2018. A novel technique for speech recognition and visualization based mobile application to support two-way communication between deaf-mute and normal peoples. Wireless Communicat. Mobile Comput., 1: 1-12. DOI: 10.1155/2018/1013234

Cavus, N., 2016. Development of an intellegent mobile application for teaching english pronunciation. Procedia Comput. Sci., 102: 365-369. DOI: 10.1016/j.procs.2016.09.413

Mittal, P. and S. Navdeep, 2016. Speech based command and control system for mobile phones: Issues and challenges. Proceedings of the 2nd International Conference on Computational Intelligence and Communcation Technology, Feb. 12-13, IEEE Xplore Press, Ghaziabad, India, pp: 729-732. DOI: 10.1109/CICT.2016.150

Reddy, D.B.R.E., 2013. Speech to text conversion using android platform. Int. J. Eng. Res. Appl., 3: 253-258. DOI: 10.1.1.415.4162

Iizuka, S., T. Kosuke, O. Shin and F. Hirotaka, 2012. Speech recognition technology and applications for improving terminal functionality and service usability. NTT Docomo Technical. J., 13: 79-84. DOI: 10.1016/j.proeng.2014.03.054

Kwon, S., K. Sung-Jae and C.J. Yeon, 2015. Preprocessing for elderly speech recognition of smart device. Comput. Speech Language, 36: 110-121. DOI: 10.1016/j.csl.2015.09.002