Web-Based automation speech-to-text application using audio recording for meeting speech

Muhammad Yasir¹, Marlince NK Nababan¹, Yonata Laia¹, Windania Purba¹, Robin², Asaziduhu Gea³

¹Faculty of Technology and Computer Science, Universitas Prima Indonesia, Indonesia
²STMIK Mikroskil, Medan, Indonesia
³Department of Information System, Faculty of Technology and Computer Sciences, Universitas Methodist Indonesia, Indonesia

*marlince@unprimdn.ac.id

Abstract- The use of technology is very important nowadays. Most of everyday activities are documented and stored in digital form. One of those activity is recording. Activity of an organization is very important, therefore recording of meeting material is one of those important activities. Usually the meeting material is documented by writing them into papers or typing them and stored them into computer. Sometimes, the meeting information is written or typed incorrectly so a speech-to-text application is required to solve this problem. In this study, the solution offered is to implement a web-based automation speed-to-text application which can record the voice of meeting participants then converted them into text automatically, so the results of the recording process of meeting materials are more effective and efficient. By using voice recognition feature its called web kit Speech Recognition, this system can be implemented. After successfully implementing this application which is so called Speech Meeting Web kit system, the average value for duration system for Indonesia and English language is 96.63 % and 82.78 %.

1. Introduction

Today the importance of using technology for every activity we do it will be documented. In some agencies or organizations every activity carried out must be documented and must be stored properly. As is the case at the University of Indonesia Prima Medan, which is still carrying out the process of documentation of the meeting material manually by using a series of meeting notes. So that errors often occur in the delivery of information to meeting participants who cannot attend the meeting.

Computer-based information systems (SI) are the needs of every organization both private and government in supporting the processes that occur within the organization. SI that can produce information quickly and accurately is the need of each organization. So that many activities in the work are very much related to SI both in schools, universities, supermarkets, companies, and even government offices. SI can help the process of archiving the recording of letters [1].

Speech is the easiest way to communicate with each other. Speech-to-Text Conversation (STT) system that has many benefits for the deaf. The purpose of this system is to change the speech signal to produce text output in the field of education by using Mel Frequency Cepstral Coefficients (MFCC) from spoken speech sound signals isolated. With used method Hidden Markov Model (HMM) [1], End Point Detection a algorithm that can solve disturbing noise noises [2]. Recognition of speech, an
approach that is useful for feature extraction of sound signals with various patterns of approach. Voice recognition systems are motivated for ASR systems with vocabulary in different languages [3][6].

This research develops the conversion of text into sound, the Text-To-Speech Concatenative system. The development of this research has been carried out in the last four years [7]. The use of Text-To-Speech has been applied in training focusing on learning English words [8][11]. With the design and implementation of the system which aims to encourage interest in learning English by using a personal account [12][13]. The development of the Slovak language text-to-speech system applying speech techniques directly using the Markov model was trained using a new speech correspondent which revealed that the sound produced was very clear and very good [15]. Signal Processing Laboratory provides personalized speech synthesizers for people with speech disorders, and also for those who have no voice [15][16].

2. Methodology

Recording is the term for a trip record of an activity, whether a meeting, seminar, discussion, or trial from the beginning to the end of the event and written by a notepad. Minutes can contain the main points discussed in the activity, objectives of the activity, place of activity, time of activity, proposals and decisions in activities and participants who attended the activity.

The Web Speech API allows users to record audio from the microphone, which is then sent via a HTTP post request to the speech recognition web service. The Web Speech API aims to enable web developers to provide features speech-to-text. In a web browser we can build a system with speech-input and will be text-output features from using a Web Speech API. The system using a JavaScript Web Speech API there can be implemented for a speech recognition feature. The speech recognition interface lives on the browser’s window object as Speech Recognition in Firefox and as web kit Speech Recognition in Chrome. The browser should support the speech recognition API if the web kit Speech Recognition object exists.

For the main code to make a speech recognition speech-to-text it's using an object which provides the speech interface, and set some of its attributes and event handlers. The recognition continuous function its set for default value is false, that meaning when the user stops talking, the speech recognition will be ended. But in this system the function will be set for true value so that recognition will continue even if the user pauses while speaking. The spoken language for the speech recognizer its set for two language there's en-US for English-United States and id-ID for Indonesian. The function used for put the language its recognition language. Because two language already put on this system the users can speak English and Indonesian simultaneously. After setting the language has done, the recognition start function will be called to activate the speech recognizer. When the capturing audio it’s begin, it will call the on start event handler, and then for each new set of results, it calls the result event handler.

In this research a system designed using only a few codes of the general function of the web speech feature as a way to use the available database from Google web speech API. Furthermore, in this research the code was modified by providing additional functions for minutes of web speech, design models, and various other features that have been made based on meeting needs. This is the process of system for speech nothicl webKit in general. In the early stages of the speech meeting system webKit speech is in active state by opening the website directly. The system will be controlled by the admin at the meeting. For simulations of the speech-to-text meetings system, 6 people are placed in a meeting room. Text used as meeting material has 243 words for Indonesian language and 266 words for English language. Two people were chosen as speakers to present the meeting topic. For the other participants, they will give some instructions when the main speaker delivers the meeting material so that a discussion will be formed, but their voices can’t be recorded and detected by the system. Because the system is only able to record one main voice, in this case the voice from material speaker. This is becomes a limitation of the system created. At the start of the meeting, the speaker will convey the topic through the microphone that has been provided. The system only allows the speaker to discuss material one by one, in other words when the meeting simulation takes place if
there is a message or voice from other meeting participants, the system cannot record the sound. Repetition of meeting simulation is done 5 times per language to get a constant result. The speaker's voice will be transmitted into the speech system of the speech meeting webKit. The voice will be processed using the features of the web speech API from Google which has been incorporated into the Notulen Meeting Speech Website system to convert the speech into text form. Because for the speech recognition events system need to take database from Google it will requires an internet connection to connect the system to Google web server. After that the result will be stored into database of the Notulen Meeting Speech Website by the admin. During the simulation, the minutes will be divided into two, the first minutes will process the recording of the meeting topics from the two speakers manually, the second minutes will observe the system and check to determine whether there are sentences or words that are not recorded through the system. So that in the next stage the results of recording the topic of the meeting using the system can be assessed. The system have a form specifically for the admin, where the conversion results will be directly submitted to the form convert speech to text. From the convert result the admin can get a conclusion from the result of the conversation in the meeting, so that can be presented in the form of meeting conclusions. The admin can manage meeting schedule and the attendance list for meeting participants. This research simulation can be illustrated as in Figure 1.

Figure 1. Process of the Recording Speech Meeting Web Kit System

For the method to find the system accuracy as a form of system feasibility test, in this research using a multi-variance statistical analysis method. To test the system again and provide maximum results, in this research the calculation of the speed of conversion from the sound form to the text was also carried out by using a random words with the number of characters as a parameters.

3. Results and Discussion
In this research the web-based system speech-to-text for meeting activities has been successfully created and implemented. The system is called Notulen Meeting Speech Website. Some meetings process can identify by the documents that have been collected during the research period. Among them are meeting schedule documents, news documents, presenter documents and decision papers. Similarly, the flow chart of the document processing on the current system is running.

Meeting schedule process is done by determining what type of meeting will be done in one year. Then the administrative staff list the meeting schedule according to the type of meeting, day, date, place and participant involved in the meeting. Once the list a created then confirmed by the caretaker
and disseminated to the participants of the meeting. The security of the process will be made into the web. Where have separate parts and functionality of each existing menu.

The results of the recording of meeting conversations from simulation has been obtained that have been automatically converted into the form of text systems as shown in Table 1. Simulation is done 5 times for each language in order to obtain constant results. The variables needed for the process of testing the system are the exact number of words detected by the system and according to the spoken word (word detected correctly) and total words form testing. Only original words are counted as total words, the number of spaces in a sentence is not counted.

As seen in Table 1 and Table 2, the average value of accuracy system for Indonesia and English language is 96.63 % and 82.78 %. During the simulation there are several words that are spoken at the meeting but in the system the words that are converted are not the same. As seen in Figure 3 in the fourth row, the word "Florida" is a kind of system error when detecting words spoken by the speaker. The word should be "For" and "The". Some words that cannot be detected too. This is happened because the articulation and pronunciation of the meeting speakers are not clear and sometimes he talks too fast. The system still cannot recognize the words, so that sometimes system provides incorrect feedback and even the word cannot be detected. The result from meetings simulation can be seen at a Recording Web Kit form Figure 2 (Indonesian) and 3 (English) for displaying the results of converting sound from meeting speakers into text that is done automatically. At the end of the process, conversion result will be input by the minutes into report meetings form to display the result for the main page website so it can be read by all users.

In this research also do a testing for speed of voice conversion into text as seen in Table 3. The test is done by using three random words which is has a different length of character. The words used it tries, evaluate, and experimentation. In this testing conversion speed for the experimental word has been obtained. The average value for try words its 0.54 s, for the evaluate word the average value its 1.03 s and for the experimentation word the average value its 1.87 s. This result states that more character of a word the system will take more time to convert the word into text file.

![Figure 2. Result convert speech to text with Indonesian language at recording webkit form](image-url)
Figure 3. Result convert speech to text with English language at recording webkit form

Table 1. The testing results for simulation Indonesian language meetings using Speech Meeting Web kit System for the average accuracy

| No | Simulation (Indonesian Language) | Manually | System | Accuracy (%) |
|----|---------------------------------|----------|--------|--------------|
|    |                                 | Total Words | Detected Words | Undetected Words | Word Detected Incorrectly | Word Detected Correctly |               |
| 1  | Simulation 1                    | 243       | 237     | 6             | 3                         | 234                     | 96.30 %       |
| 2  | Simulation 2                    | 243       | 239     | 4             | 1                         | 238                     | 97.94 %       |
| 3  | Simulation 3                    | 243       | 234     | 9             | 3                         | 231                     | 95.06 %       |
| 4  | Simulation 4                    | 243       | 241     | 3             | 2                         | 239                     | 98.35 %       |
| 5  | Simulation 5                    | 243       | 236     | 7             | 4                         | 232                     | 95.47 %       |
| 6  | Total                           | 1215      | 1187    | 29            | 13                        | 1174                    | 96.63 %       |

Table 2. The testing results for simulation English language meetings using Speech Meeting Web kit System for the average accuracy

| No | Simulation (English Language) | Manually | System | Accuracy (%) |
|----|--------------------------------|----------|--------|--------------|
|    |                                 | Total Words | Detected Words | Undetected Words | Word Detected Incorrectly | Word Detected Correctly |               |
| 1  | Simulation 1                    | 266       | 262     | 4             | 43                        | 219                     | 82.33 %       |
| 2  | Simulation 2                    | 266       | 249     | 17            | 50                        | 199                     | 74.81 %       |
| 3  | Simulation 3                    | 266       | 256     | 10            | 42                        | 214                     | 80.45 %       |
| 4  | Simulation 4                    | 266       | 262     | 4             | 24                        | 238                     | 89.47 %       |
| 5  | Simulation 5                    | 266       | 260     | 6             | 29                        | 231                     | 86.84 %       |
| 6  | Total                           | 1330      | 1289    | 41            | 188                       | 1101                    | 82.78 %       |
Table 3. There is a results for experiment of the conversion speed of the sound into text form in second (s)

| No | Experiment | Try         | Evaluate   | Experimentation |
|----|------------|-------------|------------|-----------------|
| 1  | 1st Experiment | 00:00:56    | 00:01:02   | 00:01:94        |
| 2  | 2nd Experiment | 00:00:53    | 00:01:02   | 00:01:96        |
| 3  | 3rd Experiment | 00:00:52    | 00:01:06   | 00:01:79        |
| 4  | 4th Experiment | 00:00:58    | 00:00:92   | 00:01:77        |
| 5  | 5th Experiment | 00:00:53    | 00:01:07   | 00:01:96        |
| 6  | 6th Experiment | 00:00:52    | 00:01:09   | 00:01:77        |
|    | Average Value | 00:00:54    | 00:01:03   | 00:01:87        |

4. Conclusion
Conversion of sound into text form depends on the length of the character of a spoken word. The average value for try words its 0.54 s, for the evaluate word the average value its 1.03 s and for the experimentation word the average value its 1.87 s. The longer words mentioned will be have more time needed to convert sound into text. Another factor that affects conversion speed is the internet connection used. If the internet connection is stable the conversion process will run normally if the internet connection is not stable then there is a delay during the conversion process. The average value of accuracy system for Indonesia and English language is 96.63% and 82.78 %. The system need a good and clear articulation or pronunciation from meeting speakers. Or system cannot recognize the words so that sometimes system provides incorrect feedback and even the word cannot be detected.

References
[1]. S.Myat Mon, Hla Myo Tun, “Speech-To-Text Conversion (STT) System Using Hidden Markov model (hmm)”, international journal of scientific & technology research volume 4, issue 06, june 2015.
[2]. N. Allawadi, “Speech-to-Text System for Phonebook Automation”, Computer Science And Engineering Department Thapar University,June 2012.
[3]. Erro, D.—Hernaez, I.—Alonso, A.—Garc´ìa-Lorenzo, D.—Navas, E.—Ye, J.—Arzelus, H.—Jauk, I.—Hy, N. Q.—Magari´nos, C.—Pere´rez-Ram´on, R.—Sul´ir, M.—Tian, X.—Wang, X.: Personalized Synthetic Voices for Speaking Impaired: Website and App. Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH 2015), 2015, pp. 1251–1254.
[4]. Miss Prachi Khilari, Prof. Bhope V. P “A Review On Speech To Text Conversion Methods”, International Journal of Advanced Research in Computer Engineering & Technology Volume 4, Issue 7, July 2015.
[5]. Shaheena Sultana, M. A. H. Akhand, Prodipt Kumar Das, M. M. Hafizur Rahman, “Bangla Speech-to-Text Conversion using SAPI”, International Conference on Computer and Communication Engineering(ICCCE 2012), July 2012.
[6]. Dhanush Kumar S, Lavanya S, Madhumita G, Mercy Rajaselvi V, “Journal of Speech to Text Conversion”, International Journal of Advance Research, Ideas and Innovations in Technology Volume 4, Issue 2, 2018
[7]. Suman K. Saksamudre, P.P. Shrishrima, R.R. Deshmukh, “A Review on Different Approaches for Speech Recognition System”, International Journal of Computer Applications (0975 – 8887) Volume 115 – No. 22, April 2015.
[8]. Nuzhat Atiqua Nafis, Md. Safaet Hossain, “Speech to Text Conversion in Real-time”, International Journal of Innovation and Scientific Research Volume 17, August 2015, pp. 271-277.
[9]. B. Raghavendhar Reddy, E. Mahender, “Speech to Text Conversion using Android Platform”, International Journal of Engineering Research and Applications Volume 3, Issue 1, February 2013.

[10]. Paul Daniels, “Using Web Speech Technology With Language Learning Applications”, The Jalt Call Journal Volume 11, No.2, pp. 177-187, 2015.

[11]. Prerana Das, Kakali Acharjee, Pranab Das, Vijay Prasad, “Voice Recognition System: Speech-To-Text”, Journal of Applied and Fundamental Sciences Volume 1, November 2015.

[12]. G. Swathi, C. Kiran Mai, B. Raveendra Babu “Speech Synthesise System for Telugu Language”, International Journal of Computer Applications (0975 – 8887) Volume 81 – No5, November 2013.

[13]. Bong-Hyun Kim, Ki-Chan Kim, Sang-Young Oh, Sung-Eon Hong, “Implementation of English word learning system using Smartphone TTS”, Multimed Tools Appl (2016) 75:13179–13192.

[14]. Lee DW (2015) An analysis of the early English education in some countries, A master degree paper of Kongju University.

[15]. Lee HR (2011) Development and Application of Web based English Writing System through Cooperative Learning. J Korean Inf Educ Soc 15(1):137–146.

[16]. Martin Sul’ir, Jozef Juh’ar, “Development Of The Slovak Hmm-Based Tts System And Evaluation Of Voices In Respect To The Used Vocoding Techniques”, Computing and Informatics, Vol. 35, 2016, 1467–1490.