*Advances in Social Science, Education and Humanities Research, volume 647*
**International Conference on Applied Science and Technology on Social Science 2021 (iCAST-SS 2021)**

# Developing Automatic English Speaking Skills Testing System Using Speech Recognition

*Aliv Faizal M.
Creative Multimedia Technology Department
Politeknik Elekronika Negeri Surabaya
Surabaya, Indonesia
aliv@pens.ac.id

Salim Nabhan
English Language Education Department
Universitas PGRI Adi Buana
Surabaya, Indonesia
salimnabhan@unipa.ac.id

Halimatus Sa'diyah
Creative Multimedia Technology Department
Politeknik Elekronika Negeri Surabaya
Surabaya, Indonesia
halimatus@pens.ac.id

Assidqi, M.H.
Creative Multimedia Technology Department
Politeknik Elekronika Negeri Surabaya
Surabaya, Indonesia
hasbi@pens.ac.id

Elizabeth Anggraeni Amalo
Department of Electronics Engineering
Politeknik Elekronika Negeri Surabaya
Surabaya, Indonesia
lisa@pens.ac.id

Imam Dui Agussalim
Department of Electronics Engineering
Politeknik Elekronika Negeri Surabaya
Surabaya, Indonesia
imam_dui@pens.ac.id

***Abstract*—English teachers have been testing students speaking skill through student presentation and other kinds of direct testing. Testing the speaking ability of a big number of students is time consuming. Hence, an automatic model of testing English speaking skill is demanded. A computer assisted testing of English speaking using the technology of speech recognition comes handy. In this paper, we propose a tool to automatically score a student's English-speaking performance. The proposed system applies the speech recognition technology with the string-matching algorithm in PHP language to process the voice input of the test candidate for scoring. The development resulted two main products namely the web-based speaking test app, and the data set for the scoring system. The initial stage focus of the development is the trained data model for the scoring system. The data was tested using Confusion Matrix, and it resulted percentage in accuracy of 80, precision of 84, recall and sensitivity of 94, and F1 Score of 88. This concludes that the app can help the researcher enrich the data and refine the score for a better automatic English-speaking testing system. Hence, once the app and the data model are perfected, it is ready for efficiency in English speaking testing.**

***Keywords—computer assisted language testing, speech recognition, machine learning, job interview simulation, computer assisted language learning***

## I. INTRODUCTION

In the settings of English as a Second Language (ESL) or English as Foreign Language (EFL), especially in Indonesia, English speaking ability has been one of the skills commonly tested. Nearly most college English lecturers, as well as school English teachers at all levels, assess their students' English speaking skill through direct presentation. Some forms of English test prompts are topic based presentation, open discussion, guided presentation, role play, debate, as well as interview [1].

One of English lessons offered in university level curriculum is job interview. The lesson of job interview consists of common interview questions used by employer to interview job applicants. These common interview questions come with common type of answers, and thus there are several common English phrases and sentences used to answer these questions. As a result, students are trained to use these commonly used phrases and sentences to answer specific interview questions. Students are invited into a job interview simulation, a form of role play in which some students become the interviewer who address the common interview questions while some the rests play as the interviewee who answer the interview questions. Another practice of job interview simulation is direct interview with the lecturer, in which a student as the

interviewee is directly asked by the lecturer as the interviewer [2].

The later form of job interview simulation, head to head between lecturer as the interviewer and the student as the interviewee, arises a problem when it comes to a big number of students [3]. This type of testing offers value to the lecturer in that the lecturer can assess the students' performance in detail, but when number of the students is big, nearly 30 students in one class; this testing can take hours. This situation demands a way to automatically score the students' performance on job interview simulation in English. It needs a medium or a tool to help English teachers or lecturers minimize the amount of time spent to test the students' performance.

The technology of Speech Recognition has been widely used for practicing and testing English pronunciation [4] [5]. There are numerous apps using speech recognition feature to test limited number of English words or phrases. Speech recognition has been used in a lot of settings, including in education. Apps like Duolingo, Talkiebuddy, Rosetta Stones, and many others use speech recognition technology to provide English pronunciation practice [6]. All these apps instruct users to read or pronounce the provided words, phrases, or even sentences. Whenever the spoken text, that is acquired through speech to text processing, is matched with the correct words, phrases, or sentences in the database, it would be marked as correct.

In relation to assessing students' responses or answers to a specific interview question, the task of matching between the student speech to text input and the answer words, phrases, or sentences in database is challenging because the students' answers may vary, unlike that used in Duolingo, Talkiebuddy, and the others, in which the text to be spoken is the same as the text in the database [7]. To get it clear, we would like to elaborate it as follows:

The interview question "Tell me about yourself" commonly requires an answer that covers about names, birth day, birth place, where he/she lives, hobbies, educational backgrounds, and so forth. We would like to narrow down the coverage to name. In order to know that the student has mentioned about his/her name is to see if there are words, phrases, or sentences indicating name mentioning. There are some possible answers that indicate about mentioning name, including "My name is ... " , or "I am ... ", or "You can call me ... ", or may be some other possibilities. The testing tool that we develop has to have this capability of detecting indicators of answer to specific question like "Tell me about yourself".

. However, to date, the development of a tool which automatically generates a score of students' English speaking performance utilizing speech recognition technology especially on a job interview is still under explored. Therefore, this research aims at developing automatic English speaking testing system using speech recognition.

## II. LITERATURE REVIEW

### *2.1 Computer Assisted Language Testing*

CALT is a combined process in which language performance is prompted and measured with the assistance of a computer. CALT embraces computer-adaptive testing (CAT), the implementation of multimedia in language test prompts, and automatic feedback. There commonly are three main motives for using technology in language testing: efficiency, equivalence, and innovation.

The use of technology in the English for Specific Purposes (ESP) has extended remarkable admiration among English as a Foreign Language (EFL) researchers. ESP teaching is goal-oriented and established on the particular necessities of students. Corpus supports to check the communicative ability and efficiency. Content, language, grammar and vocabulary knowledge is being evaluated. The evaluation of curriculum, instructional materials are continuously measured. The most significant portion of testing comprises the language practice for a specific purpose, for example in medical, business, science, law, and technology, and so forth.

### *2.2 Speech Recognition in Educational Technology*

Speech recognition has been impacting education as it facilitates the teaching and learning process with great tools and immense potentials. In language education context, speech recognition aids the students to learn pronunciation [5] as well as practice in certain situational conversation [7]. Speech recognition in education was also utilized in creating Voice Geometry Painter Application leading to easier teaching and fun learning [9]. There are several existing speech recognition tools for mobile phones including Dragon Dictate, Dragon Go, Google Now, SIRI, MeMeMeMobile, SILVIA, and Vlingo in which they have their specific features needed for education tools [10]. In addition, a study investigating two of the most prominent cloud-based speech recognition engines for language learning between Apple's Siri and Google Speech Recognition (GSR) was also conducted along with Moodle as Learning Management System (LMS) [11].

### 2.3 Web Speech API

Web speech API denotes a design of a system that is projected for speech synthesis and speech analysis. Speech Analysis refers to a method of analyzing recordings of sound by computers to collect information and be employed to augment interactions in communication. Speech Synthesis refers to a human speech ability, a computer's capacity to pronounce sounds like a human voice [11]. Mozilla is one of the web service providers, but now in its development it is starting to build a Web Speech API design system that provides audio data processing to data text or vice versa. Mozilla Developer Network Web Speech API takes input in the form of voice or audio data from users, afterward the audio data is converted to become text. Furthermore, the text data is directed back and shown on the monitor so that users can see the converted text [12].

### 2.4 PHP Regular Expression Function (Regex)

Regular expressions refer to a notation for describing sets of character strings. Whenever a specific string is in the set defined by a regular expression, we frequently state that the regular expression matches the string [13].

### 2.5 Related Works

#### 2.5.1 TalkieBuddy

TalkieBuddy is a web based English conversation practice application that uses speech recognition and speech synthesis technology for the interaction between the user and the machine [7].

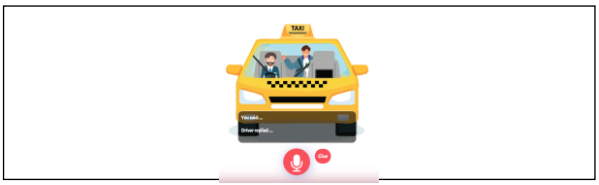

Fig. 1. TalkieBuddy

#### 2.5.2 Orientation Day

Orientation Day is the name of a visual novel game that is intended to help English learners practice conversation, pronunciation, and reading aloud [14]. The game also embraces the power of speech recognition technology for the interaction. This is unique since most other visual novel games use click-based interaction. The game was developed in Unity 3D software and was programmed in C language. The speech recognition engine used in this game is Microsoft Cortana, that is available in Windows 10. Thus, the game can only be played in a computer with Windows 10 operating system.

## III. METHOD

This development takes a specific common job interview question as the case study while it combines the technology of Speech Recognition and Strings Matching Algorithm. It began from narrowing the use case, that is a specific common job interview question, and then identifying the sub topic and indicators such as words, phrases, or sentences. The predicted correct answers, used as indicators, were stored in database. The algorithm within the application was configured to receive input through speech recognition and process the input by matching them with the stored answers in database for scoring generation. The apps were tested in terms of its precision in scoring the test takers' performance in answering the selected interview question.

### 3.1 Common Job Interview Questions

There are more than 20 common job interview questions, and some of them are:

*3.1.1 Tell me about yourself.*
*3.1.2 What attracted you to our company?*
*3.1.3 Tell me about your strengths.*
*3.1.4 What are your weaknesses?*
*3.1.5 Where do you see yourself in five years?*
*3.1.6 Have you encountered a business challenge?*
*3.1.7 What are you looking for in your next role?*
*3.1.8 etc.*

However, we do not use all of them for this initial research. We narrow them to one question only, that is "Tell me about yourself".

### 3.2 Identifiying sub-topics and indicators

After deciding that the specific job interview question selected is "Tell me about yourself", we break down the sub topic possibility and the possible indicators of words, phrases, or sentences may be used by the student as interviewee to answer that question. We come with the following initial notes:

TABLE I. INDICATORS OF SUB-TOPIC'S ANSWERS

| Name | Birth Day | Place | Hobbies | Education |
|---|---|---|---|---|
| - My name is<br>- I am<br>- You can call me<br>- My friends call me | - I was born on<br>- My birth day is | - I come from<br>- I live in<br>- I am from | - My hobbies are<br>- My hobby is<br>- I like to<br>- My favorite activities are | - I graduated from<br>- I am currently studying<br>- I am studying |

The above table is the initial prediction of what may the students respond to the questions. To enrich the list, we invited some students to respond to the question and list some other possibility into the list. The more the number of varied model of answers, the more accurate the scoring would be

To illustrate the text processing, we provide the following scenario:

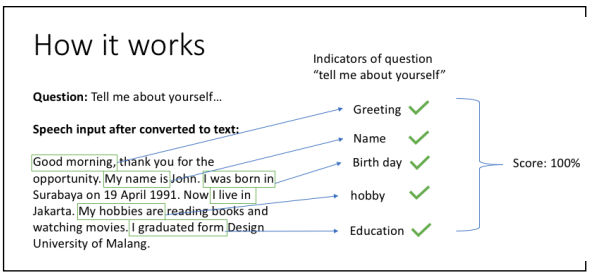


Fig. 2. Indicators extraction and matching

Fig. 2. portrays that the expected answer for the interview question "Tell me about yourself" consists at least 5 elements namely greetings, name, birth day, hobby, and education. To see if the interviewee has used greetings is by finding out if there are greetings expression used in the answer. In the illustration above, the interviewee mentioned "Good Morning" in his/her answer. That indicates that she/he has used greetings.

Furthermore, the phrase "My name" in the sentence "My name is John" is an indication that the interviewee has mentioned about his/her name.

When all of the elements in the expected answer had been met, the interviewee would score 100%.

To help organize the category/sub topic, as well as the indicators, we would also develop a back end dashboard so that the administrator finds it easy to add the indicators into the list. This would be in a form of quiz structure, in which the structure goes as the following schema:

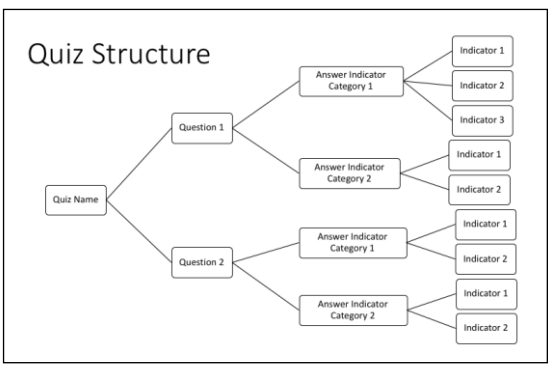


Fig. 3. How a quiz organizes questions and the answer indicators

Fig. 3. illustrates that the administrator would be given a privilege to organize the questions, like that of interview questions, into a form of quiz. For instance, the quiz name is "Job Interview Questions". The quiz would consist of several questions, like "Tell me about yourself", "Why should we hire you?", and so forth. Each question would have several elements. For example, the elements of question "Tell me about yourself" would cover elements like: greetings, name, birth day, place where he/she lives, hobby, and education. Furthermore, each of the element would be provided with some indicators, phrases likely indicating the elements. For instance, the phrases indicating greetings would be: (1) "Good Morning", (2) Good Afternoon", etc., and phrases like (1) "My name is", (2) "you can call me", (3) "I am", etc., would indicate the element of name. Each element may have more than two phrase indicators, and the number of phrases may increase along the positive variations used by the interviewees.

### *3.3 App Mock-up*

The visual representation of the above structure is initiated as the following figure 4:

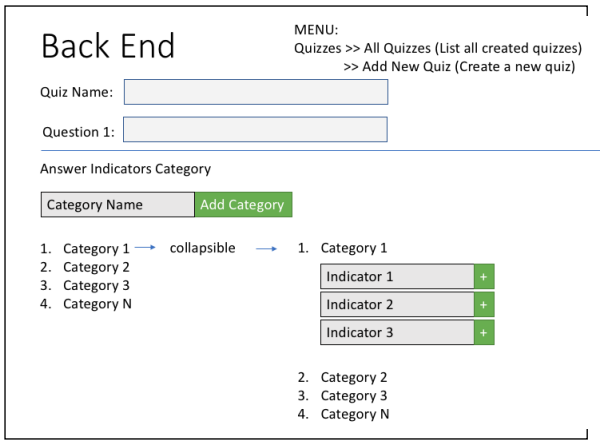


Fig. 4. How an admin could manage a quiz in the back-end

To make it easier for testing this application, we would provide a front-end representation on how the question would appear, where the speech recognition button is positioned, as well as how the score would be displayed. See the following figure 5:

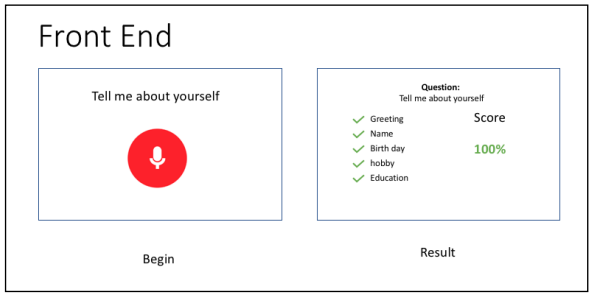


Fig. 5. Voice record button and score card in the front end

The test taker or the interviewee would press the voice record button to start answering the interview question as stated above the button. After stating the answer, the interviewee would re-press the button to stop and submit it. The app would process the answer and show the result in form of a score card.

### *3.4 App system and database*

In general, the flow of the app starts from the voice input by the test taker. The voice input is processed using Web Speech API from Mozilla.org. The converted speech to text is analyzed using string matching algorithm to match the input with the answers in the database. Furthermore, the score is shown in the score card so that the test taker can see the result.

#### *3.4.1 System*

The overall flow of the system is illustrated in the following figure 6:

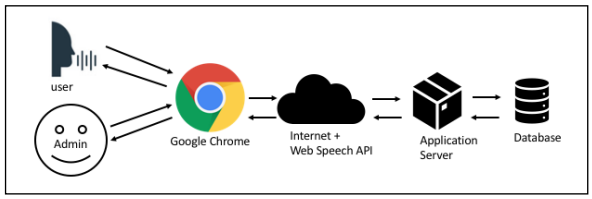


Fig. 6. System Flow

The user opens the app and uses his/her voice as input through a device using Google Chrome browser. The voice input is sent to the Web Speech API through internet. The voice/speech is converted to text. The app receives the converted input from the Web Speech API and then match the strings with the ones in the database. The result of the matching is sent back to the user in form of feedback and results.

#### *3.4.2 Database*

The database arrangement covers the process by the operator and the user. The operator sets up the master data that consists of dataset of the topic indictors and scoring. The operator has privileges to group the users as well as assign them to a specific task containing quiz and questions along with the topic indicators. The user tries to answer the question, and the answer is sent to the database for matching process. The data about the user' profile and inputs are stored.

The database arrangement is illustrated in the following figure 7:

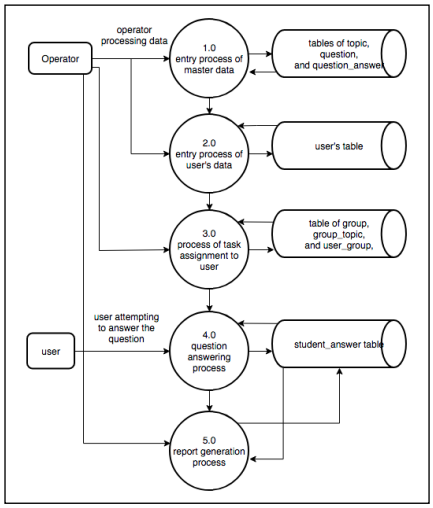


Fig. 7. Database Arrangement

### *3.5 Measurement Procedure*

The application developed in this research was tested on its accuracy, precision, recall or sensitivity, and F1 score in detecting the input by the students. The data are arranged in a confusion matrix [15]. We are trying to find out the best configuration and algorithm, and as complete indicators as possible to provide high reliability in scoring. With this small scope, we seek for confidence in the scoring so that in the next research we are ready to expand the sub topic. We tested 20 candidates as interviewees.

To measure the model, we tried to get the number of true positive as high as possible. The Accuracy formula is illustrated in (1):

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (1)$$

The Precision is measured using the following formula (2):

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

The Recall or Sensitivity is measured through the following formula (3):

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

The F1 Score is calculated under the following formula:

$$F1\ Score = \frac{(2\ x\ Recall\ x\ Precision)}{Recall + Precision} \quad (4)$$

## IV. RESULT

The result of this research and development is the app that process the automatic testing and the accuracy of the data model.

### *4.1 App*

In the current development, the app was developed mainly to support the data gathering process and it had been able to gather the data from the users so that the main data set become richer. Here are some displays representing the main features of the app.

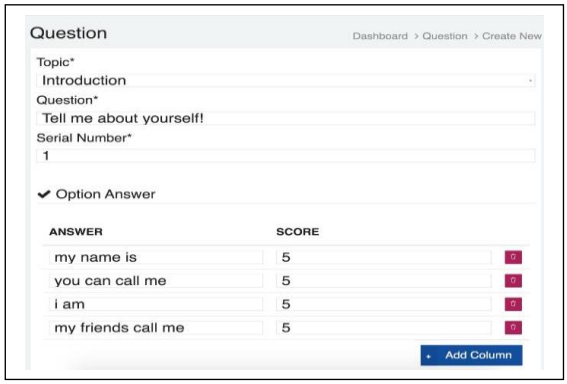


Fig. 8. Quiz editor, to manage topics, questions, and answer indicators

Previously designed in mock-up form, the quiz editor functional design and layout was implemented in the back-end for the admin to have privilege in managing the quiz components.

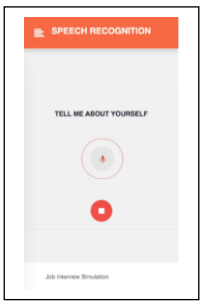


Fig. 9. How a question with the record button is presented to a user in the front-end

Fig.13. shows the layout displaying the job interview question. When the user/test taker/interviewee is ready to answer, he/she would press the record button, and an icon of recording progress would appear, indicating that the user could start to speak to answer.

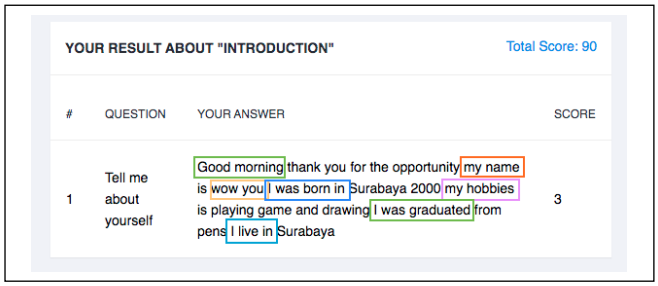


Fig. 10. A page displaying a user's result

Fig.14. portrays a user's result, displaying the converted speech to text from his/her answer for each question. The answer data were collected for analysis in order to enrich the model data for each elements of a question.

The current state of the app has been functional to help gather the data to empower the data set. However, further development and addition to the features are needed in order to have more complex analysis.

### *4.2 Data Model*

As mentioned ahead, this research would focus on one interview question, that is "Tell me about yourself". Moreover, while the answer elements for the question "Tell me about yourself" covers: Greetings, Name, Birth day, Hobby, Place of living, and Education, in this paper, we would like to focus the elaboration only one answer indicator, that is "Name". The indicator of "Name" is to ensure that the test taker has mentioned his name. That is because the element of "Name" is one of the elements of the question "Tell me about yourself". Thus, in the following table, we do not show all of the data. We only show the data under the indicator category of "Name".

The summary of the data for the indicator element of "Name" is displayed in the following Confusion Matrix table.

TABLE II. CONFUSION MATRIX

| **N=20** | **Actual Positive (1)** | **Actual Negative (0)** |
|---|---|---|
| **Predicted Positive (1)** | True Positive (TP): 16 | False Positive (FP): 3 |

| **Predicted Negative (0)** | False Negative (FN): 1 | True Negative (TN): 0 |
|---|---|---|
| | 17 | 3 |

The sample of phrases given by the test takers are shown in the following table.

TABLE III. DATA SAMPLE

| **Matrix** | **Phrases** |
|---|---|
| **TP** | My name is (11), I am(3), You can call me (1), My friends call me(1) |
| **FN** | I was given a name (1), |
| FP | My name (3) |
| TN | 0 |

The calculation results from the Confusion Matrix are presented in the following aspects: Accuracy: to describe how accurate the model is in classifying the data correctly. The accuracy calculation result for the data model is 80%. Precision: to describe the accuracy between the requested data and the prediction results provided by the model. The precision calculation result for the data model is 84%. Recall or Sensitivity: to describe the success of the model in retrieving an information. The recall calculation result for the data model is 94%. F1 Score: to illustrate the comparison of the weighted average precision and recall. We use precise accuracy as a reference for algorithm performance if our dataset has a very close (symmetric) number of False Negative and False Positive data. The F1 Score calculation result for the data model is 88%.

The performance of data model "Name" is displayed in the following figure.

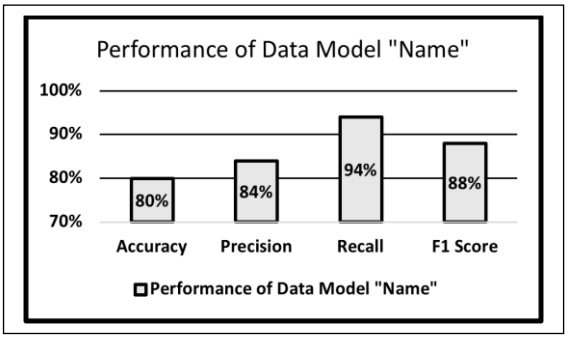


Fig. 11. Performance Chart of Data Model "Name"

Overall, the performance of data model "Name" is considered high. However, this number could be re-optimized by adding the uncovered data "I was given a name" and "my name" into the database.

One factor to consider is the grammar. Response "My name" is grammatically error because it lacks verb "is". The correct form should be "My name is", which is already in the data model. Hence, it was initially considered as False Positive (FP). On the other side, "My name" could also be seen as True Negative (TN). In the case where grammatical error is tolerated, such a data shall be seen as True Positive (TP). Thus, in such a case, the operator has privilege either to put it as TP or TN.

## V. CONCLUSION

To summarize this paper, we would like to address two main reports, namely: the app and the performance of the data model, specifically the element of "Name".

### *5.1 The App*

In this paper, while the development details of the app are not thoroughly presented, it has functioned the way it was intended and that it has allowed the researcher to gather the data in order to measure the performance of the data model. Furthermore, the app is ready for further data collection.

### *5.2 The Performance of Data Model "Name"*

The performance of data model "Name" is considered high. This data model is only one representative of other elements of the interview question "Tell me about yourself".

Grammatical error should be considered and observed more for their potential appearance along with the number of users using the app. Thus, there is a demand for determining whether to tolerate the grammatical error or to set high requirement on the grammar.

As tested using confusion matrix, the data model performed with accuracy of 80%, precision of 84%, recall and sensitivity of 94%, and F1 Score of 88%.

The future potential research topics to undergo as response to this research results are:

- Enriching the elements and data model for each interview question.
- Observing through various use cases whether to ignore grammatical error, to mind the grammatical error, or to set different value such as in form of points and percentage, or level of correctness.
- Attempting to apply the system to content area other than job interview simulation, such as speaking prompts of IELTS or the

speaking section of TOEFL IBT, other English competence tests the like.

## ACKNOWLEDGMENT

This research is funded by the Unit of Research and Community Outreach of Politeknik Elektronika Negeri Surabaya, Indonesia. We hereby thank to our fellow research group members and all research partners for the accomplishment of this research.

## REFERENCES

[1] Suwandi, "Designing Speaking Test," *Eksplanasi*, vol. 4, no. 8, pp. 183–191, 2009.

[2] Y. O. Jong and C. K. Jung, "The development of interview techniques in language studies: Facilitating the researchers' views on interactive encounters," *English Lang. Teach.*, vol. 8, no. 7, pp. 30–39, 2015, doi: 10.5539/elt.v8n7p30.

[3] S. Ma, D. Seidl, and T. McNulty, "Challenges and practices of interviewing business elites," *Strateg. Organ.*, vol. 19, no. 1, pp. 81–96, 2021, doi: 10.1177/1476127020980969.

[4] A. Elimat and A. F. AbuSeileek, "Automatic speech recognition technology as an effective means for teaching pronunciation," *JALT CALL J.*, vol. 10, no. 1, pp. 21–47, 2014, doi: 10.29140/jaltcall.v10n1.166.

[5] A. Kholis, "Elsa Speak App: Automatic Speech Recognition (ASR) for Supplementing English Pronunciation Skills," *Pedagog. J. English Lang. Teach.*, vol. 9, no. 1, p. 01, 2021, doi: 10.32332/joelt.v9i1.2723.

[6] A. F. Muhammad, D. Susanto, A. Alimudin, F. Adila, M. H. Assidiqi, and S. Nabhan, "Developing English Conversation Chatbot Using Dialogflow," *2020 Int. Electron. Symp.*, pp. 468–475, Sep. 2020, doi: 10.1109/IES50839.2020.9231659.

[7] A. F. Muhammad, D. Susanto, A. Alimudin, Z. A. Rochman, M. Hasbi Assidiqi, and S. Nabhan, "Development of English Conversation Practice App with Artificial Intelligence & Speech Recognition," *2020 Int. Electron. Symp.*, pp. 442–449, Sep. 2020, doi: 10.1109/IES50839.2020.9231570.

[8] A. C. Román-odio, B. A. Hartlaub, and C. Rom, "Classroom Assessment of Computer-Assisted Language Learning : Developing a Strategy for College Faculty Published by : American Association of Teachers of Spanish and Portuguese content in a trusted digital archive . We use information technology and tool," vol. 86, no. 3, pp. 592–607, 2013.

[9] A. Iftene, A. G. Cloud, and S. Api, "Speech recognition in education Voice Geometry Painter Application," 2017.

[10] R. Vishnupriya and T. Devi, "Speech recognition tools for mobile phone - A comparative study," *Proc. - 2014 Int. Conf. Intell. Comput. Appl. ICICA 2014*, pp. 426–430, 2014, doi: 10.1109/ICICA.2014.93.

[11] P. Daniels and K. Iwago, "The suitability of cloudbased speech recognition engines for language learning," *JALT CALL J.*, vol. 13, no. 3, pp. 229–239, 2017, doi: 10.29140/jaltcall.v13n3.220.

[12] A. F. Muhammad, D. E. Pratama, and A. Alimudin, "Development of Web Based Application with Speech Recognition As English Learning Conversation Training Media," in *IES 2019 - International Electronics Symposium: The Role of Techno-Intelligence in Creating an Open Energy System Towards Energy Democracy, Proceedings*, 2019, pp. 571–576, doi: 10.1109/ELECSYM.2019.8901594.

[13] R. Cox, "Regular Expression Matching Can Be Simple And Fast (but is slow in Java, Perl, PHP, Python, Ruby, ...)," 2007. [Online]. Available: https://swtch.com/~rsc/regexp/regexp1.html.

[14] E. A. Amalo, I. D. Agusalim, and C. D. Murdaningtyas, "Developing visual novel game with speech-recognition interactivity to enhance students' mastery on English expressions," *J. Sos. Hum.*, vol. 10, no. 2, pp. 129–130, 2017, doi: 10.12962/j24433527.v10i2.2865.

[15] E. Beauxis-aussalet and L. Hardman, "Confusion Matrix for Non-Expert Users," no. October 2014, 2018. In *IEEE Conference on Visual Analytics Science and Technology (VAST)-Poster Proceedings*.