BRIEF REPORT
published: 13th October 2020
doi: 10.21070/jees.v5i2.867

Automatic speech recognition in computer-assisted language learning for individual learning in speaking

Esti Junining[1]*, *Sony Alif*[2], *Nuria Setiarini*[1]

[1]*Language and Literature Department, Universitas Brawijaya, Indonesia,* [2]*Language Education Department, Universitas Brawijaya, Indonesia*

This study is intended to help English as a Foreign Language (EFL) learners in Indonesia to reduce their anxiety level while speaking in front of other people. This study helps to develop an atmosphere that encourages students to practice speaking independently. The interesting atmosphere can be obtained by using Automatic Speech Recognition (ASR) where every student can practice speaking individually without feeling anxious or pressurized, because he/she can practice independently in front of a computer or a gadget. This study used research and development design as it tried to develop a product which can create an atmosphere that encourages students to practice their speaking. The instrument used is a questionnaire which is used to analyze the students' need of learning English. This study developed a product which utilized ASR technology using C# programming language. This study revealed that the product developed using ASR can make students practice speaking individually without feeling anxious and pressurized.

Keywords: Automatic Speech Recognition, Computer-Asssisted Language Learning, Speaking

INTRODUCTION

The rapid development in technology during the past decades opens the probability to improve many aspects of human life, including on the education field. It also has improved the demand for foreign language education. Muslichatun (2013) says that foreign language students consider speaking ability as their main goals of study, either because they would get some personal satisfactions from being able to speak a foreign language or because they think it would be useful in pursuing other interests or career goals. Moreover, people are considered mastering a language if they can speak the target language well. However, mastering the speaking itself is rather difficult to accomplish. There are many problems that language learners' face while they are mastering speaking. One of the factors that make speaking difficult are stated by Muslichatun (2013) as follows: "Some of the students do not want to practice speaking outside the class because they are not confident due to their incapability of speaking English".

On the contrary, Brown and Douglas (2007) claims that lacks willingness to communicate (self-efficacy and risk-taking) can make speaking more difficult to master than it already is. Therefore, to overcome this problem language student needs an atmosphere that encourages them to try out language Brown and Douglas (2007).

The recent advances of computer hardware and software have provided Computer-Assisted Language Learning (CALL) with limitless resources for foreign language learning. The most up-to-date language teaching and learning using CALL system is state-of-the-art Automatic Speech Recognition (henceforth, ASR) technology which mostly emphasizes on pronunciation. Instead

OPEN ACCESS
ISSN 2503 3492 (online)

Correspondence:
Esti Junining
esti@ub.ac.id

Received: *31st August 2020*
Accepted: *30th September 2020*
Published: *13th October 2020*

Citation:
Junining E, Alif S and Setiarini N (2020) Automatic speech recognition in computer-assisted language learning for individual learning in speaking. J. Eng. Educ. Society. 5:2. doi: 10.21070/jees.v5i2.867

of practicing speaking in front of instructor or other people, students can practice speaking in front of computer. This situation can reduce students' anxiety in speaking and encourages students to try language and to give a response. In other words, by ASR language students can overcome one of the difficulties in learning to teach, which lacks willingness to communicate.

Computer-Assisted Language Learning (CALL) is described as the search for and study of applications of computer and information technology (ICT) in language teaching and learning. The main purpose of CALL is to find ways for using computers for teaching and learning language. Specifically, CALL is the use of computer that involve educational learning, including word processing, presentation, guided drill and practice, tutor, simulation, and problem solving using various computer-based technologies like games, multimedia, and internet applications such as e-mail, chat and the World Wide Web (WWW) for language learning purposes.

Automatic Speech Recognition (ASR) is a technology which allows students to experiment with the language in a safe, private setting. Levis and Suvorov (2012) describes ASR as "an independent, machine-based process of decoding and transcribing oral speech. A typical ASR system receives acoustic input from the speaker through a microphone, analyzes it using some pattern, model or algorithm, and produces an output, usually in the form of a text". Therefore, ASR is a technology that shows great potential for individual pronunciation work. Stone (2013), Tell Me More Auralog (2013), and English (2014) are examples of a language learning software that based on ASR Mccrocklin (2016).

Individual leaning is a learning where students become self-motivated and self-directed. They are able to set goals, take actions, make decisions, and make use of available resource, both human and non-human, towards their own learning based on their interests, needs and capabilities. Learning independently is not necessarily means learning alone where the student must do everything by themselves, but it is reducing student's exposure of continuous direction and supervision from teachers, parents or others whose acts as guides, motivators, and resources and make them able to identify their own needs and capabilities and know when they needed help (Delong, 2009; White, 2008).

Speaking is one of the important productive skills in learning English. It is unquestionable that speaking is used in most of daily activities such as socializing and working. Writing is also used a lot, but when we compare the frequent uses and the efficiency of the two skills, speaking is more dominant than writing. Moreover, people often show their idea or opinion through speaking instead of writing. The term of speaking itself has several meanings. Speaking is an interactive process of constructing meaning by producing, receiving and processing information.

Some students experience some level of speech anxiety when they have to speak in front of a class. In fact, speaking in public might be most people's greatest fear. Speech anxiety can come in different level; from a slight feeling of worries to a nearly devastating fear. Some of the most common symptoms of speech anxiety are shaking, sweating, dry mouth, rapid heartbeat, and squeaky voice.

Woodrow (2006) assessed the major "stressor" that caused student to have speaking anxiety, those were performing in English in front of classmates, giving an oral presentation, speaking in English to native speakers, speaking in English in classroom activities, speaking in English to strangers, not being able to understand when spoken to, talking about an unfamiliar topic, talking to someone of higher status, and speaking in test situations.

Some previous studies are "*Speech Recognition Bahasa Indonesia untuk Android*" conducted by Wijaya et al. (2013). The result of the study states that the applications are efficient and usable. It made the users feel comfortable to learn Bahasa Indonesia. In addition, the application is very beneficial and useful for the users. The next study is "*Implementasi Speech Recognition pada Aplikasi Pembelajaran dalam Bentuk Permainan Menebak Kata baku Bahasa Indonesia*" conducted by Bahri (2019). The result shows that the application is effective to increase Indonesian vocabularies.

In this study, the writer aimed to compile an application using ASR technology which allows students to practice speaking individually and to find the benefits and limitations of ASR for individual learning in speaking.

METHODS

This study used instructional design (ID) models contain the core element of ADDIE. ADDIE is an acronym for analyze, design, develop, implement, and evaluate and it is based on a systematic product development concept Branch (2009). The researcher used ADDIE model because ADDIE remains one of today's most effective and systematic ID, step by step framework used by interaction designers and training developers is needed to make sure that CALL development occurs in a controlled and structured phase.

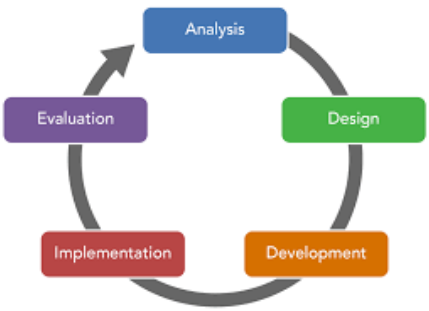

FIGURE 1 | **ADDIE model** Branch (2009)

There were five steps in ADDIE model. The steps were; (1) analyze. It involves identifying and clarifying the problem, the gap, or desired outcomes. Several key components are to be utilized to make sure analysis is thorough; however, in

this study; the researcher collected the course document, syllabic, and questionnaire as primary data collection in December 2019. (2) design. The design stage defines specific learning objectives, subject matter analysis, exercise, lesson planning, assessment instruments used and media selection to fulfill instructional goals. The researcher also determines tools to be used to gauge performance, tests, subject matter analysis, planning, and resources. Here, pinpointing the main idea of the ASR application is determined. (3) develop. The development phase may involve creating and testing the content. Using the data and information collected from previous stages, it allows the researcher to use this ample source to create ASR application. (4) implement. The implementation stage reflects the development of required materials, associated applications or websites, and preparing participants to use any required tools or technology. Here, the researcher tested the ASR application to find out whether the application is working according to the initial objective or not. The researcher tested the product two times using two different ways. (5) evaluate. Evaluation is an essential stage of the ADDIE model as it aims to answer whether the problems are solved or the goals are met. Every stage of ADDIE model deal with continual or formative feedback. The researcher analyzed the implementation process and the participants' questionnaire to come up with the necessary evaluation that has to be made for the product to work better.

RESULTS AND DISCUSSION

The result of this study is divided into Product Analysis, Product Design. Product Development, Product Implementation, and Product Evaluation.

Product Analysis

The initial step of design-based research is need analysis which analyzed the EFL learners' need in speaking. As previously mentioned in the Methods, questionnaires which elaborate the learners' problems in speaking, aspects of difficulties in speaking, the reasons of having the problems and what tools they need in speaking are distributed to the EFL learners. The result of the questionnaires show that they need a particular tool to measure the feedback in speaking. By knowing the need, the instrument for measuring the speaking skills in the form of Automatic Speech Recognition is designed.

Product Design

The design of Automatic Speech Recognition (ASR) product is in the form of application to recognize speech. ASR is a machine to detect voice which is identified by fluctuating symbols of voice. The voice recorded by the machine memory will be selected then identified as correct or not in the computer. The correct voice will be indicated by signals which appeared in the computer. From this machine design, the correct or incorrect voice on the basis of English pronunciation will be easily detected.

Product Development

In accordance with the objective and research method of this study a product in the form of an application was designed, developed and implemented to equip the students with good English-speaking practice. This application was developed using certain hardware and software. In developing this application, the researcher used hardware with the specifications: 1) Intel® Pentium® CPU B980 @ 2.40GHz 2.40 GHz Processor; 2) 4.00 GB (2.60 GB usable) RAM; and 3) Built-in Microphone. While the software used for developing this application were: 1) Windows 7 Professional with 32-bit Operating System; and 2) Microsoft Visual Studio 2012 Ultimate.

Product Implementation

Before the product was implemented to the participants, the writer tested the product to find out whether the application is working according to the initial objective or not. The product was tested two times using two different ways. The first test was conducted in a proper way, which means properly followed the instruction given by the application. While in the second test, it was not conducted in a proper way in which the application was not said anything at all and taken the practice over the time expected.

The result of the first test showed that the application worked properly based on its objective. The application gave 7 (seven) as a score based on speaker speech. The feedback were "All vowels and consonants are produced in a manner that is easily understood ..." and "Speech maybe uneven ...". While in the second test the application gave 0 (zero) as a score based on speaker speech. It also gave feedback such as "Pronunciation seems completely characteristic of another language ..." and "Speech is slow ..." which is understandable considering the speaker stayed silent for the whole practice process.

After the product was tested and confirmed to work properly, the product then was implemented to the participants. The product could give different scores and feedback to the participants according to the speaker's speech input. However, there was a case where the product cannot properly recognized speech input. Thus, it gave inaccurate score and feedback to the speaker's speech.

Participant 4 received 0 (zero) score on accuracy while participant 5 got 0 (zero) both on accuracy and fluency. This problem occurred because the product was used in crowded place. As a consequence, instead of receiving relevant speech input from the speaker it also received other noises from the crowd which made the automatic speech recognition cannot accurately give score and feedback to the speaker's speech.

Participant 1, 2, and 3 received their score and feedback without any problem at all. In this case, the participants used

the product in quiet room with only their speech to be recognized by the product. To summarize, the product only works well while it is used in calm and quiet places without any disturbing noises otherwise it will not recognize the speaker speech because there are many noises to be recognized.

Product Evaluation

The implementation process and the participants' questionnaires were analyzed to evaluate the product implementation. The result of the questionnaire analysis revealed the quality of the visual graphic. It was said that the visual graphic was good yet it is too simple. However, the score and feedback given by the product were helpful for the students' speaking. Four out of the five participants think it reflects their speaking ability accurately. Therefore, there is no evaluation needed regarding the product visual graphic, the scoring rubric and the feedback.

In spite of all the good feedback given by the participants, a problem occurred when implementing the product in crowded place. The microphone did not work properly because there were lots of other people noises in the crowd. The built-in microphone used by the researcher as an input device not only took the speaker speech as an input but also those noises. As a result, the automatic speech recognition engine was unable to recognize the speech given by the speaker. Thus, the product kept giving 0 (zero) as a score.

Considering above problem can really affect how the product work, the researcher decided to change the built-in microphone configuration. At first, the microphone was configured to 100% volume and 0.0 dB boost. After the evaluation, the microphone volume was set to 100% and the microphone boost was enhanced to +36.0 dB boost. In addition, the researcher also used noise suppression sound effect to reduce the background noise with the purpose of improving the microphone performance.

DISCUSSION

The purpose of this study is to describe ASR application which allows EFL learners to practice speaking individually and to find the benefits and limitations of ASR for individual learning in speaking. The analysis of needs explores the learners' problems in speaking, aspects of difficulties in speaking, the reasons of having the problems and what tools they need in speaking are distributed to the EFL learners.

The learners' problems in speaking include feeling afraid of making mistakes, being ashamed of starting to talk and not wanting to start a communication. These problems commonly happened to beginners as what Goodrich and Namkung (2019) claimed that the teachers' challenge in teaching speaking is the learners' low motivation in speaking.

The aspects of difficulties are failure to employ speaking skills in real life communication, linguistic and psychological barriers and insufficient exposure in target language. These aspects of difficulties are similar to the case of learning speaking difficulties in Libya Diaab (2016). These aspects of difficulties can be overcome by individual learning with tools.

One of the tools EFL learners' need in speaking is developing ASR. The result of questionnaires indicated that EFL learners are happy learning speaking using ASR, and they do not have high anxiety anymore in speaking. From the above discussion, it can be concluded that ASR is effective enough in helping the EFL learners to practice teaching

However, the previous studies have also revealed the research gaps in implementing Automatic Speech Recognition (ASR). Some people said that it was not useful to increase speaking scores. While others state that ASR is beneficial to increase speaking skills because it can reduce the anxiety level of the EFL learners in speaking practice. However, the result of the study shows that ASR indicates positive feedback to increase speaking skills.

In addition, Benk et al. (2019) state that ASR provides a link between humans and machine, the most important is computer, however, machine alone still cannot fulfill the need of human to produce better speaking skills. It is supported by Cook et al. (2014) who emphasized using ASR can help people learn English pronunciation well. However, selecting words or phrases in English needs challenging attempts to be easily understood by the machine. This chance needs hard working skills to implement.

The researcher applied this phenomenon and developed it to become a key branch in human communication with the machine where the sound has helped to facilitate the use of the machine with the user and to make a natural communication. Automatic speech recognition has greatly contributed to the development of artificial intelligence, which seeks to create between very flexible methods of handling the machine, this allows the user to communicate and exchange information without using known input/output modules such as the keyboard. Voice-based input/output techniques are very useful in several areas, such as the care of disabled people, the use of cars, in particular hand driving, distress calls, and many more. Further, Cook et al. (2014) argue that this ASR is commonly used to pronounce words uttered by an individual voice, the voice is typed in the monitor, as a result of the voice uttered. This invention is helpful to learn pronunciation for beginners, to identify the spelling being uttered.

Similar to Cook et al. (2014), Yousem (2008) underlines the other type of ASR as voice recognition dictation. He explains that the role of this product working is by giving voice transcription. The voice is transcribed just like the process of voice dictation.

Kikel (2020) elaborates the differences between speech recognition and voice recognition. Speech recognition and voice recognition are innovations that have rapidly developed over the centuries. With their level of growth, voice and speech recognition have multiple purposes that can enhance usability, boost defense, support law enforcement efforts. Furthermore, learn the difference between voice recognition and speech

recognition as stated in the following.

The Primary Difference

Voice and speech recognition differ in its primary function. Speech recognition can not only detect words that are being uttered, while voice one cannot only be used by language but other common voices. It is not only from the language being uttered but also any sound being heard. From this definition, it is clearly seen that both terms are different in each function.

Speech Recognition

Speech Recognition emphasizes on the use of machine as a tool to identify any speech uttered by human when uttering words or languages being spoken. It has connection with dialect accent to pronounce words or languages. The voice, then, is transformed into graphical symbols indicating the tone of the voice whether it is strong or weak. Two main things to consider in speech recognition are namely accuracy and speed.

Voice Recognition

Voice recognition is a bit different from speech recognition in the way that voice recognition focuses on the voice difference among humans. Besides, it is used to identify the individual voice, it also can be used as a speaker identity.

CONCLUSION

From the result of this study, it can be concluded that the application of ASR is comfortable and can be used for individual learning. The participants can easily navigate from each page to another page. They also did not have any problem in understanding the given instructions by the application. The feedback provided by the application are also very helpful in finding the aspect of speaking the participants need to improve. On the other hand, the application is unfit to be used in crowded places since extraneous voice might come through the microphone. In addition, there is a complaint from the participants that sometimes they had to speak loudly in front of the microphone so that it can be listened well by the device. The trouble of the microphone could reflect to the scores, and sometimes it did not reflect the speakers' speaking ability.

ACKNOWLEDGMENTS

Acknowledgment was given to the Dean of the Faculty of Cultural Studies for all ease and support in conducting this research.

REFERENCES

Auralog (2013). Tell Me More. https://www.crunchbase.com/organization/tell-me-more.

Bahri, S. (2019). Implementasi Speech Recognition Pada Aplikasi Belajar Bentuk Permainan Menebak Kata Baku Bahasa Indonesia. *Ubiquitous: Computers and Its Applications Journal* 2, 93–98.

Benk, S., Elmir, Y., and Dennai, A. (2019). A Study on Automatic Speech Recognition. *Journal of Information and TechnologyReview* 10.

Branch, R. M. (2009). Instructional design: The ADDIE approach. vol. 722 (Springer Science & Business Media).

Brown, H. and Douglas (2007). *Teaching by Principles an Interactive Approach to Language Pedagogy* (New York: Pearson Education).

Cook, A. M., Polgar, J. M., and M, J. (2014). Assistive technologies-e-book: principles and practice (Elsevier Health Sciences).

Delong, S. (2009). Teaching Methods to Encourage Independent Learning and Thinking. http://www.usma.edu/cfe/Literature/DeLongS_09.pdf.

Diaab, S. (2016). Role of faulty instructional methods in Libyan EFL learners' speaking difficulties. vol. 232, In *Procedia - Social and Behavioral Sciences* (Elsevier BV), 338–345. doi: 10.1016/j.sbspro.2016.10.032.

English, B. (2014). https://www.burlingtonenglish.com/.

Goodrich, J. M. and Namkung, J. M. (2019). Correlates of reading comprehension and word-problem solving skills of Spanish-speaking dual language learners. *Early Childhood Research Quarterly* 48, 256–266. doi: 10.1016/j.ecresq.2019.04.006.

Kikel, C. (2020). Difference Between Voice Recognition and Speech Recognition. Total Voice. https://www.totalvoicetech.com/difference-between-voice-recognition-and-speech-recognition/.

Levis, J. and Suvorov, R. (2012). Automatic speech recognition. *The encyclopedia of applied linguistics*.

Mccrocklin, S. M. (2016). Pronunciation learner autonomy: The potential of Automatic Speech Recognition. *Pronunciation learner autonomy: The potential of Automatic* 57, 25–42.

Muslichatun, I. (2013). Improving the Students' Speaking Practice in Describing People by Using Contextualized Card Game. *Journal of Language and Literature* 8.

Stone, R. (2013). Official Rosetta Stone: Language learning, Learn - a Language. https://www.rosettastone.com/.

White, C. (2008). *Language Learning Strategies in Independent Language Learning: An Overview*, Hurd, T. W. L. M. S. (ed.) (Clevedon, England: Multilingual Matters), 3–24.

Wijaya, T., Santoso, S., and Salman, A. G. (2013). Speech Recognition Bahasa Indonesia Untuk Android (Jakarta: Universitas Bina Nusantara.).

Woodrow, L. (2006). Anxiety and Speaking English as a Second Language. *RELC Journal* 37, 308–328. doi: 10.1177/0033688206071315.

Yousem, D. M. (2008). Voice Recognition Dictation. In Radiology Business Practice: How to Succeed (Elsevier Inc), 231–245.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.