Speech Recognition with Hidden Markov Model and Multisensory Methods in Learning Applications Intended to Help Dyslexic Children

B Siregar, AJ Tarigan, S Nasution, U Andayani, F Fahmi

Dept. of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara
Dept. of Electrical Engineering, Faculty of Engineering, Universitas Sumatera Utara

E-mail: baihaqi@usu.ac.id

Abstract. Dyslexia disorders commonly occur in children. Dyslexia is a difficulty in recognizing letters and numbers. Helpful tools are needed to detect and treat dyslexic children so that early anticipation can be taken. So that later this disorder will no longer be suffered by the child when he is an adult. For this reason, information technology-based learning media was created in this study with the aim of helping therapy for children with dyslexia. This study utilizes the hidden Markov method for speech recognition and multisensory methods for learning media. In this study learning is divided into three levels, namely learning to recognize numbers, learning to recognize letters, and learning to understand words. Each level uses a pattern that is generally difficult to accept children with dyslexia disorders. During the testing process, level one and level three efficiency has reached 80%, while for level two it is still worth 60%.

1. Introduction
At present, we are no longer familiar with the term dyslexia disorder. This disorder affects many children. Although they have an IQ above the average normal child, dyslexics have difficulty recognizing letters and numbers. In the schooling process, some teachers think that dyslexic children are stupid, low achievers, lazy, lack of concentration, or sick children. This assumption arises because teachers do not understand about this disorder, so that the efforts made by the teacher are less optimal or not by the needs and abilities of children [1].

Therefore we need an application to help their learning, one of them with learning applications using speech recognition or detecting the sound of children who have dyslexia. Requests will be made as attractive as possible to produce children who have dyslexia want to learn and easy to remember. The algorithm used for speech recognition is the Hidden Markov Model, while for learning is a multisensory method.

Hidden Markov Models are located in almost all modern voice recognition systems, and although the basic framework has not changed significantly in the last few decades or more, detailed modeling techniques developed in this framework have evolved to a state of sophistication. With results that have been stable and significant [2]. Multisensory learning methods are based on the assumption that children will learn better if the subject matter is presented in various modalities. Modalities that are often involved are visual (vision), auditory (hearing), kinesthetic (movement), and tactile (touching) [3].
In general, the method used for learning materials for dyslexic children is done manually with written media or print media. For this reason, the author tries to make learning aid into the computer application system in the form of learning media for children with dyslexia disorder based on information technology. The research that has been done is the implementation of speech recognition on Android-based English learning applications using the hidden Markov model method. In this application questions in the form of questions about English are faced with users. And users answer questions using sound. If there is an error in answering, it will be faced with learning to find out the answer to the English question. At the conclusion of the study, using the hidden Markov model of practice and implementation of the system has gone well.

2. Material & Method
Materials and methods used in this study are Microsoft's speech application programming interface with the Hidden Markov Model and multisensory method.

2.1. Speech Recognition
Speech recognition, known as Automatic Speech Recognition (ASR) or computer voice recognition, is the process of converting sound signals into a sequence of words, through an algorithm that is implemented as a computer program. Speech recognition is also a pattern recognition, where there are two phases in supervised pattern recognition, namely, training and testing. The feature extraction process is relevant for general classification in both stages. During the transmission phase, the parameters of the classification model are estimated using a large number of class examples (training data) [4]. During the testing phase or introduction phase, the test pattern feature (data speech test) is matched with the model trained from every class. The test pattern is then stated in a model that has the best test pattern. Four techniques can be seen in speech recognition, namely [5]:

2.1.1. Speech Analysis Technique. Speech data contains various types of information that shows the identity of the speaker. The stages of speech analysis relate to the frame size suitable for voice signal segmentation in further study and extraction.

2.1.2. Feature Extraction Technique. Speech feature extraction in problem categorization is about reducing the dimensions of the input vector when maintaining the difference in signal strength. As we know from the formation of primary speaker identification and verification systems that the number of training and test vectors is needed for growing classification problems with the input dimensions provided so that we need the extraction feature of the sound signal.

2.1.3. Modeling Technique. The purpose of the modeling technique is to produce speaker models that use unique speaker vector features. Modeling speakers are divided into two classifications, namely speaker recognition and speaker identification. Speaker identification technique automatically identifies who speaks based on individual information integrated into the sound signal. Speaker recognition is also divided into two parts: dependent speakers and independent speakers. In speaker mode standalone of speech recognition, the computer must ignore the distinctive characteristics of the speaker from the sound signal and extract the intended message. In the case of the speaker, recognition machine must remove the components of the speaker in the acoustic signal. The primary purpose of speaker identification is to compare speech signals from unknown speakers to public speaker databases. This system can recognize speakers, who have been trained with some speakers. Speaker recognition can also be divided into two methods, text-dependent, and text independent. In the text dependent method, the speaker says a keyword or sentence that has the same text to test the training and introduction. While the independent version does not depend on specific passages that are spoken.

2.1.4. Matching Technique. The speech recognition engine matches a word detected with the phrase known to one of the following techniques: Whole word matching, which is a machine comparing digital-
audio signals that come to the word recording template. This technique requires less processing time than sub-word pairing but needs that the user (or someone) record every word that will be recognized, sometimes several hundred thousand words. The entire word template also requires large storage memory (between 50 and 512 bytes per word) and is only practical if the vocabulary recognition is known when the application is developed. Subword matching, which is a machine looking for sub-words, usually phonemes and then doing advanced pattern recognition. This technique requires more processing than whole word matching but requires less storage (between 5 and 20 bytes per word). Besides, word pronunciation can be guessed from the English text without requiring the user to speak the previous word.

2.2. Speech Application Programming Interface
Speech Application Programming Interface (SAPI) is an API developed by Microsoft that is used as a voice identifier in the Windows application programming environment. The SAPI programming architecture can be seen as a middleware located between the application and the speech engine [5]. The main components in SAPI are as follows:

- **Voice Command.** A high-level object for commands and controls using voice recognition.
- **Voice Dictation.** A high-level object for continuous dictation speech recognition.
- **Voice Talk.** A high-level object for speech synthesis.
- **Voice Telephony.** An object for writing telephone applications based on voice recognition.
- **Direct Speech Recognition.** An object as a machine to control voice recognition (direct control of recognition engine)
- **Straightforward Text to Speech.** An object like a machine that controls synthesis.
- **Audio Object.** An object to read from an audio device or an audio file.

2.3. Hidden Markov Model
The foundation of the modern Hidden Markov Model (HMM) based on continuous speech recognition technology was established in the 1970s by groups at Carnegie-Mellon and IBM who introduced the use of HMM and then at Bell Labs where the HMM was introduced. HMM provides a practical and straightforward framework for modeling time variations in vector spectral sequences [6]. As a consequence, almost all of today's large vocabulary in continuous speech recognition systems is based on HMM. The practical application of HMMs in modern systems involves considerable sophistication in presenting the core architecture of HMM-based continuous speech recognition systems. HMMs are located almost in all contemporary speech recognition systems, and although the basic framework has not changed significantly in the past decade or so, detailed modeling techniques developed within this framework have evolved to a state of sufficient sophistication. The results have been stable and significant [7].

HMM is possible to be used by each speech model. Even if a lousy speech unit is selected, HMM can absorb suboptimal characteristics in the parameter model, this of course limits system performance. Words seem to be the most natural unit for modeling because what you want to recognize and the model in language also uses words as basic units. Indeed, recognizers who use the word-level model appear quite reasonable. Part of this success is since they can capture the effects of the coarticulation phoneme in words that is indicated by the more massive the unit, the better the recognizer will be. However, because there are many unique words, the training data needed for each of these words, makes this kind of system not easily expanded. So for an extensive vocabulary of natural speech recognition, the word unit is not an option. But for small, well-defined vocabulary, for example, a set of commands, they are very suitable. Usually, the left to right topology model is used in which the number of states depends on the number of phonemes in the word. One part per grammar is a good rule of thumb [5].

The main components of the large continuous speech recognizer vocabulary are illustrated in Figure 1. The audio wave input from the microphone is converted to a fixed size sequence of acoustic vector
\(y_1, y_2, ..., y_t\) in a process called feature extraction. The decoder then tries to find the series of words \(w_1, w_2, ..., w_t\) which is most likely to have produced \(Y\).

\[
\hat{w} = \text{arg} \max \{P(w|Y)\}
\]

However, because this equation cannot be calculated directly because the number of observation sequences that may never run out, the Bayes 1 rule is used to change into:

\[
\hat{w} = \text{arg} \max \{p(Y|w)P(w)\}
\]

where

- \(Y\) is the sound signal as a source of observation.
- \(w\) is a word order that has the highest probability spoken.
- \(P(w)\) is the probability that the string word \(w\) will be pronounced. Also called the language model.
- \(P(Y|w)\) is the probability that when the phrase string \(w\) is pronounced, acoustic \(Y\) will be observed. Also called an acoustic model.

The possibility of \(P(Y|w)\) is determined by the acoustic model, and \(P(w)\) was previously identified by the language model. The basic unit of sound is shown by Gales & Young’s telephone acoustic model, 2008. As a result, a speech recognizer consists of three components, namely the preprocessing section that translates the sound signal into an order of observation symbols, a language model that tells us how likely a string of certain words will occur and an acoustic model that tells us how string words might be spoken.

**Figure 1. Hidden Markov Model Architecture.**
2.4. Method

The methodology used in this research can be seen in Figure 2.

![System Flowchart](image)

**Figure 2.** System Flowchart

In Figure 2 the flow diagram of the system process shows that the incoming sound will be converted into digital form after first extracting. The signal that has been changed will be matched into the existing database. Pairing with the highest probability will be chosen as a language model.

3. Results & Discussion

Testing was carried out directly on dyslexic respondents to get the feasibility value of the application system that was built. Tests are carried out in stages on three levels of difficulty in the form of recognition of number, letter, and word recognition patterns.
3.1. Application Start Page Display
The primary page display when the application is first to run is shown in Figure 3.

![Figure 3. Application Main Page View.](image)

3.2. Display Learning Level
The level page view is shown in Figure 4. Level 1 is number recognition, level 2 is letter recognition, and level 3 is word recognition.

![Figure 4. Display Learning Level Pages.](image)
3.3. Display of Patterns of Numbers, Letters, and Words

Figure 5(a) is a display of a number pattern, Figure 5(b) display of letter patterns, and Figure 5(c) display of word patterns.

![Figure 5](image)

**Figure 5.** Display Pages of Numbers, Letters, and Word Patterns.

3.4. System Testing

In this study, the respondents who are positive children have dyslexia. The experiment was carried out in stages. I know the number pattern five times, then the letter pattern five times and the word pattern five times.

![Figure 6](image)

**Figure 6.** System Testing on Respondent
3.5. System Testing Results
For the test results in the category on the numbers conducted on the child, respondents can be seen in Table 1.

Table 1. Value of the Number Pattern Testing

| Testing Number | Number of Questions | Number of Correct Answers | Number of Wrong Answers | Value |
|----------------|---------------------|---------------------------|-------------------------|-------|
| 1.             | 5                   | 3                         | 2                       | 60    |
| 2.             | 5                   | 5                         | 0                       | 100   |
| 3.             | 5                   | 4                         | 1                       | 80    |
| 4.             | 5                   | 5                         | 0                       | 100   |
| 5.             | 5                   | 5                         | 0                       | 100   |

From the Table 1 it appears that the first experiment, the respondent has not been able to answer correctly all the questions that arise. But in the second, third, fourth and fifth trials there is an increase in the ability to recognize numbers. For the test results in the category of letter patterns conducted on respondents can be seen in Table 2.

Table 2. Value of the Letter Pattern Testing

| Testing Number | Number of Questions | Number of Correct Answers | Number of Wrong Answers | Value |
|----------------|---------------------|---------------------------|-------------------------|-------|
| 1.             | 5                   | 2                         | 3                       | 40    |
| 2.             | 5                   | 3                         | 2                       | 60    |
| 3.             | 5                   | 4                         | 1                       | 80    |
| 4.             | 5                   | 4                         | 1                       | 80    |
| 5.             | 5                   | 5                         | 0                       | 100   |

From Table 2 it appears that after several attempts at letter recognition, there was an increase in the child's ability. For the test results in the category of word patterns conducted on respondents can be seen in Table 3.

Table 3. Value of the Word Pattern Testing

| Testing Number | Number of Questions | Number of Correct Answers | Number of Wrong Answers | Value |
|----------------|---------------------|---------------------------|-------------------------|-------|
| 1.             | 5                   | 1                         | 4                       | 20    |
| 2.             | 5                   | 3                         | 2                       | 60    |
From Table 3 it appears that the first experimental pattern of respondents has difficulty in recognizing words formed from similar letters. Even though in the next stage it has increased, it needs more intensive training.

4. Conclusion
From the results of the analysis and testing carried out on the application, it can be concluded that:
- Based on a direct trial of dyslexic respondents, using this system, therapy for children with dyslexia is more comfortable, more exciting and enjoyable.
- The use of a desktop microphone on a computer running this application is less useful because it is affected by noise in the surrounding environment or sound, so a device is needed in the form of a headset microphone.
- The effectiveness of the system at level 2 is not the same as level 1 and level 3. The system is still not sensitive to the letters of the Indonesian alphabet because the Speech API does not fully support the Indonesian version.

References
[1] M. A. Anusuya and S. K. Katti, “Speech Recognition by Machine, A Review,” Jan. 2010.
[2] L. A. Sessiani, “Pengaruh Metode Multisensori dalam Meningkatkan Kemampuan Membaca Permulaan Pada Anak Taman Kanak-kanak,” pp. 1–113, 2007.
[3] M. Gales and S. Young, “The Application of Hidden Markov Models in Speech Recognition,” Found. Trends® Signal Process., vol. 1, no. 3, pp. 195–304, 2007.
[4] F. R. Vellutino, J. M. Fletcher, M. J. Snowling, and D. M. Scanlon, “Specific reading disability (dyslexia): what have we learned in the past four decades?,” J. Child Psychol. Psychiatry, vol. 45, no. 1, pp. 2–40, Jan. 2004.
[5] S. K. Gaikwad, B. W. Gawali, and P. Yannawar, “A Review on Speech Recognition Technique,” Int. J. Comput. Appl., vol. 10, no. 3, pp. 16–24, Nov. 2010.
[6] Sukarso and A. Syarif, “Aplikasi Pengenalan Suara Menggunakan Microsoft Sapi sebagai Pengendali Peralatan Elektronik,” 2007.
[7] U. Goswami, “Reading, dyslexia and the brain,” Educ. Res., vol. 50, no. 2, pp. 135–148, Jun. 2008.