International Journal of Computer Applications Technology and Research
Volume 9–Issue 02, 76-82, 2020, ISSN : -2319-8656

The Verification of Voice Recognition Using Cmusphinx and DTW

Gusti Made Arya Sasmita
Department of Information Technology
Faculty of Engineering
Udayana University
Badung, Bali, Indonesia

I Putu Arya Dharmaadi
Department of Information Technology
Faculty of Engineering
Udayana University
Badung, Bali, Indonesia

Henrico Aldy Ferdian
Department of Information Technology
Faculty of Engineering
Udayana University
Badung, Bali, Indonesia

Abstract: The development of smartphones increasingly affects the human life, such as data security. The use of smartphones to maintain data security is still low, which generally only uses conventional method as security systems available on smartphone devices, such as word combinations, PINs or patterns. Various weaknesses of conventional methods cause the development of biometrics systems. Therefore, a technology is created to replace the conventional security system that is biometric security system using natural human characteristics, one of them using voice. Data security uses the android based sound biometric app using CMUSphinx as the word library. CMUSphinx does not require an internet connection to run it. MFCC (Mel Frequency Cepstrum Coefficients) is as a characteristic extraction method on a digital sound signal. Matching process with saved data uses DTW (Dynamic Time Warping) method. This study can help improve the data security of smartphone; therefore, it cannot be sabotaged or accessed by other parties that are not desired.

Keywords: Voice Recognition, Android, CMUSphinx, MFCC, DTW

1. INTRODUCTION

The development of the internet is also followed by the development of an increasingly large smartphone, required the development of an adequate security system as well. Meanwhile, awareness on the data security of user is still low, which generally only use the conventional methods available on smartphones, such as word combinations, PIN or pattern. The use of PIN and password causes some problems, such as forgotten, can be used together, and can be cracked with various algorithms.

Various weaknesses of conventional methods above cause of the development of biometrics system. Users try to apply the new science of biometrics as a medium for personal recognition. Biometrics systems use body parts or behavior, which weaknesses in conventional methods can be reduced. The advantages of using biometrics are difficult to duplicate body parts, cannot be used together, and cannot be forgotten. This technology fulfills two important functions, they are identification and verification. The identification system aims to solve one's identity. Meanwhile the verification system aims to refuse or accept identity claimed by someone. The things that encourage the use of biometric identification and verification are universal (in everyone's case), unique (each has its own characteristics), and is not easily falsified[1].

Voice communication is one of the most rapid and appropriate communication media in humans in conveying information. The distinctiveness of the people' voice are in the loud or weak voice when people speak in normal circumstances, the way of pronunciation, intonation, rhythm of speech, accent etc. The use of sound is important to be analyzed in several processes related to voice processing which are divided into two types, namely speech recognition and speaker recognition. In contrast to speaker recognition which is the recognition of identity claimed by someone from his voice (special characteristic can be intonation of voice, depth of This research is applied in the Android platform because Android is one of the popular operating system used by the community. The used word library is CMUSphinx with the MFCC method as the feature extraction process on the voice level, etc.), speech recognition is a process done by computer to recognize words spoken by someone regardless of the related person identity[2].

This research is applied in the Android platform because Android is one of the popular operating system used by the community. The used word library is CMUSphinx with the MFCC method as the feature extraction process on the inputted digital sound pattern and the DTW method as the matching process. Library CMUSphinx is used because the security system designed does not require an internet connection to run it; therefore, users do not have to worry about the limitations of internet connection. The sound pattern feature matching process that has been stored in the database using the DTW method allows a device to recognize speech by digitizing words and matching those digital signals to a certain pattern stored in the device. This technology can be a good alternative for smartphones not easily sabotaged or accessed by other parties that are not desired.

The preparation of this paper is as follows: Section 2 presents some earlier work on speech recognition. Section 3 describes the proposed voice recognition system. The experimental results are discussed in Section 4 and conclusion in Section 5.

2. LITERATURE REVIEW

Research on voice recognition with several methods has been discussed several times. In 2011, Darma Putra and Adi Resmawan discuss how to design and create a software that can verify a speaker using MFCC method as feature extraction and DTW for matching process[3].

In 2013, B. Raghavendhar Reddy and E. Mahender discuss a system of acquiring speech signals that run through the microphone and processing sample speech to recognize spoken text. Recognized text can be stored in a file. The development is on the android platform using Eclipse Workbench. The speech-to-text system directly acquires and converts voice to text. It gives users different options for data entry. Furthermore, the speech-to-text system can improve the accessibility of the system by providing data entry options for blind, deaf, or physically disabled users[4].

In 2014, Bhadragiri Jagan Mohan and Ramesh Babu. N conducts research that explains the continuous speech recognition can be used in the security system to verify the keywords spoken by the user. The speech recognition system processes the spoken word using the MFCC algorithm and through feature matching stages using the DTW method. The whole system is implemented using matlab where the input of speech samples is recorded by sound card on windows[5].

In 2015, Mansour, et al. conducts a research on voice recognition using MFCC and DTW. This study focuses on developing a system for speech recognition using dynamic time warping (DTW) algorithms by comparing sound speaker signals with sound signals already stored in the database, and extracting the main features of speaker sound signals using MFCC[6].

3. RESEARCH METHOD

This study uses the Android platform with analog voice signal processing using CMUSphinx as the word library. The analog voice signal is converted to a digital voice signal; furthermore it is processed using the MFCC method as feature extraction. Sample data already registered, will be tested for data matching using DTW method. This allows the system to know which users match the data that has been registered or not. Figure 1 presents the block diagram of the created application.

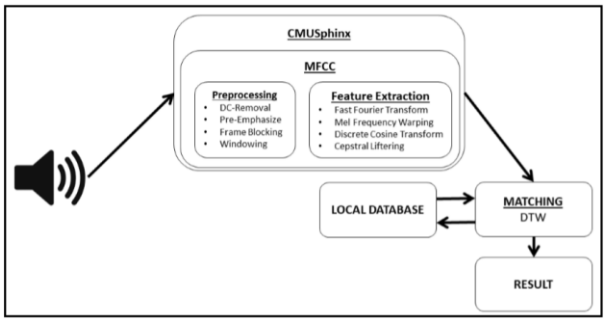

Figure 1: Block Diagram

4. CONCEPTS AND THEORIES

Concepts and theories contain explanations of supporting theories that will be used in this study. These theories include Voice Recognition, CMUSphinx, MFCC, DTW. The theory will be discussed as follows.

4.1 Voice Recognition

Sound is a combination of various signals; however, the pure sound can theoretically be explained by the velocity of the oscillation or frequency measured in Hertz (Hz) and the amplitude or loudness of the sound by measurement in decibels (dB). Voice recognition first appeared in 1952 and consists of a device for the introduction of a single digit of spoken word. Then in 1964, there is IBM Shoebox, one of the most well-known technology in the United States in the field of health is Medical Transcriptionist (MT) is a commercial application that uses speech recognition. Voice recognition is divided into two types, namely speech recognition and speaker recognition. Speech recognition is the process of voice identification based on the spoken word. The comparable parameter is the level of voice suppression which will then be matched to the available database templates. Meanwhile the voice recognition system based on people who speak called speaker recognition. User recognition can be classified into three stages: identification, detection, and verification. User identification is a process for determining the identity of a user through spoken voice, meanwhile user detection is the process of discovery of the user's voice from a bunch of votes in a database, and user verification is a process to verify the user's voice conformance to the identity claimed by the user. User recognition focuses more on user voice recognition and not on user speech recognition[7].

4.2 CMUSphinx

In the early 1920s the first machine to recognize a commercially significant level of speech called Radio Rex was personalized in the 1920s, the effort to design speech recognition systems was automatically made in the 1950s. During the 1950s, most speech recognition systems investigated the resonance spectral vowel regions of each speech extracted from the analog filter output signal and the logic circuit[8].

In 2000, the Sphinx group at Carnegie Mellon is committed to opening some voice recognition components, including Sphinx 2 and later. Sphinx 3, Sphinx 4, PocketSphinx for mobile devices. PocketSphinx as an open source library allows developers to add new languages. However, it takes an acoustic model and a language model. The use of PocketSphinx technology in 2014 by P.Vijai Bhaskar and Dr. S. Rama Mohana Rao came with the Telugue welcome recognition system and there was little effort made in translating the Tamil language into English by voice. Sphinx is an opensourcetoolkit for speech recognition developed by Carneige Mellon University (CMU) located in the United States. In order to recognize and respect the creator, the Sphinx is often referred to as CMUSphinx. CMUSphinx uses the HMM method and the n-gram statistical language model to build an Automatic Speech Recognition (ASR) system. CMUSphinx was first developed by Kai-Fu Lee[9].

4.3 Mel Frequency Cepstrum Coefficients (MFCC)

Mel Frequency Cepstrum Coefficients (MFCC) is one of the most widely used methods in the field of speech processing, both speech recognition and speaker recognition used to perform feature extraction. This method adopts the workings of the human auditory organ, so as to capture the very important sound characteristics which are used to perform parameter extraction, a process that converts voice signals into several parameters. The parameter extraction steps using the MFCC method is as shown in Figure 2[3].

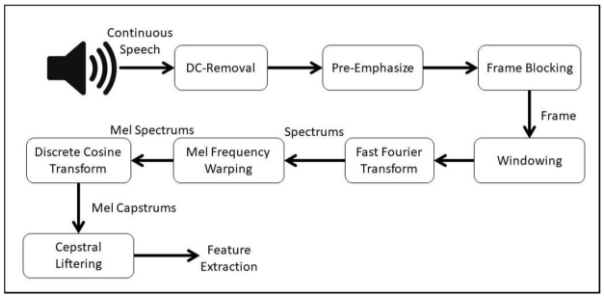

Figure 2: Block MFCC

4.3.1 Conversion of Analog Signals Into Digital Signals

Natural signals in general such as voice signal is a continuous signal which has an unlimited value. On the computer, all the signals that can be processed by the computer is just a discrete signal or often known as the term digital signal. Changing the signal can be done through 3 processes, such as data sampling process, quantization process and coding process[3].

4.3.2 DC-Removal

DC Removal aims to calculate the mean of the sound sample data, and subtract the value of each sound sample by the average value. The goal is to get the normalization of the input voice data.

$$y[n] = x[n] \cdot \bar{X} \ , 0 \leq n \leq N\text{-}1 \qquad (1)$$

Where:
y[n] : The signal sample of DC removal process results.
x[n] : The original signal sample.
$\bar{X}$: The average value of the original signal sample.
N : The signal length.

4.3.3 Pre-Emphasize

Pre-emphasize is one type of filter that maintains high frequencies on a spectrum, which is generally eliminated during the sound production process. The x (n) speech signal is sent to the high-pass filter[10]:

$$y(n) = x(n) - a * x(n-1) \qquad (2)$$

Where y (n) is the output signal and the value of a is usually between 0.9 and 1.0. The Z transform of this equation is given by:

$$H(z) = 1 - a * z^{-1} \qquad (3)$$

The purpose of pre-emphasize is to offset the high frequency parts suppressed during the human voice production mechanism. Furthermore, it can reduce the noise ratio of the signal in order to improve signal quality.

4.3.4 Frame Blocking

Frame blocking is a process in which the voice signal is divided into several pieces that can later facilitate the calculation and analysis of sound. Each piece of the sound signal is called a frame. Frames in frame blocking generally have a length of 10-30ms. One frame consists of several samples depending on each second the sound will be sampled, how big the sampling frequency and overlapping. Overlapping is done to avoid loss of feature or voice characteristics on the boundary of intersection of each frame. The length of the overlap area commonly used is approximately 30% -50% of the frame length.

4.3.5 Windowing

Framing process can cause spectral or aliasing leakage. This can happen because of the low number of sample rate, or due to the frame blocking process which causes the signal to be discontinue. In order to reduce the possibility of spectral leakage, the result of the farming process must pass through the windowing process. Each frame must be multiplied by a hamming window to keep the first and last point continuity in the frame called Hamming Window.

4.3.6 Fast Fourier Transform (FFT)

The core of the fourier transform is to decipher the signal into the component of the sine shape of different frequencies. Spectral analysis shows that the time difference in speech signal is related to different energy distributions over frequency. Therefore, FFT is done to get the frequency response of each frame. When FFT is done on a frame, it is assumed that the signal in the frame is periodic and continuous when it is enclosed. If this does not happen, FFT can still be performed; however, discontinuities on the first frame and last point will likely introduce undesirable effects on the frequency response. In order to solve this problem, the user multiplies each frame by windowing to increase its continuity on the first and last points. FFT is a fast algorithm for the implementation of Discrete Fourier Transform (DFT) operated on a discrete time signal consisting of N sample as follows:

$$f(n)=\sum_{K=0}^{N-1} y_k e^{-2\pi jkn/N}, n=0,1,2,\ldots,N\text{-}1 \qquad (4)$$

4.3.7 Mel Frequency Warping

Psychophysical studies have shown that the human perception of sound frequencies for speech signals does not follow a linear scale. Therefore, for any tone with real frequency f, in Hz, a pattern is measured on a scale called 'mel'. The 'mel frequency' scale is a linear frequency scale below 1000 Hz and a logarithmic scale above 1000 Hz. This scale is defined by Stanley Smith, John Volkman and Edwin Newman as[11]:

$$mel(f)=2595*\log_{10}(1+\frac{f}{700}) \quad (5)$$

An approach for spectrum simulation on a mel scale is to use a filterbank placed uniformly on a mel scale. Filterbank is one of the forms of the filter performed in order to know the energy size of a particular band frequency in the sound signal. At MFCC, filterbank is applied in the frequency domain.

4.3.8 Discrete Cosine Transform (DCT)

DCT is the last step of the main process MFCC feature extraction. Basically the DCT concept is the same as the inverse fourier transform. However, the results of DCT close to PCA (Principle Component Analysis). PCA is a classic static method that is widely used in data analysis and compression. Therefore, DCT often replaces the inverse fourier transform in the MFCC feature extraction process. In order to get the value of MFCC cepstrum, the mel frequency must be transformed back into time domain using Discrete Cosien Transform (DCT) method. DCT is applied to the output of a triangular N bandpass filter in order to obtain a coefficient of cepstral L mel-scale. The formula for DCT is,

$$C(n) = \Sigma Ek * \cos(n * (k - 0.5) * \pi / 40)) \quad (6)$$

Where n = 0,1, .. to N

Where N is the number of triangle bandpass filter, L is the number of cepstral mel-scale coefficients. There are N = 40 and L = 13. Since it has done FFT, DCT converts the frequency domain into a domain like time called domain quefrency. The features obtained are similar to cepstrum, which is called the coefficient of cepstral mel-scale, or MFCC.

4.3.9 Cepstral Liftering

The results of the DCT process have some disadvantages. Low order from cepstral coefficients is very sensitive to spectral slope; meanwhile the high order part is very sensitive to noise. Therefore, cepstral liftering becomes one of the standard techniques applied to minimize the sensitivity. This process can be done by implementing the window function to the cepstral features.

$$W[n] - \{1+\frac{L}{2}\sin(\frac{n\pi}{L}) \; n - 1,2,\ldots,L \quad (7)$$

Where:
L = number of cepstral coefficients.
N = index of cepstral coefficients.

4.4 Dynamic Time Warping (DTW)

The problems in voice recognition are quite numerous; one of them is the recording process that is often different in duration, even if the spoken word or phrase is the same. Although for the same syllable or vowel, the recording process often occurs in different durations. Consequently the matching process between test signals and reference signals (templates) often does not yield optimal values. DTW (Dynamic Time Warping) is an algorithm that focuses on matching two feature vector sequences by repeatedly shrinking or extending the time axis until an exact match is obtained between two sets. It is used in order to check the similarity between two voice signals or non-linear curved time series. The DTW distance between two vectors is calculated from the optimal bending path of the two vectors.

The DTW algorithm is more realistic to use in measuring pattern matching than simply using linear measurement algorithms such as Euclidean Distance, Manhattan, Canberra, Mexican Hat and others[12]. The principle provides a range of 'steps' in space (a time frame in the sample, time frames in the template) and is used to match paths that show the largest local match (similarity) between straight time frames. The total similarity cost obtained with this algorithm is an indication of how well these samples and templates have in common which will be selected as the bestmatching templates.

5. EXPERIMENT AND RESULT

5.1 Experiment

Figure 3 is an app view that has been created for testing. The sample used is 50 different person voice sample data by performing tests on three conditions around a silent environment, noise with low level, and noise with high level.
The application created will be tested by looking for an error ratio that states the probability of a matching error in the system. There are two types of ratios, namely the False Accepted Rate (FAR) and False Rejected Rate (FRR) ratios. The testing phase in this study was carried out by a total of 50 people who were asked to pronounce a word from a given word list where everyone has one word to register and test. Furthermore, the user is asked to say the words given to the conditions surrounding the environment of registration and testing sounds silent, noise with low level and high level and matching results recorded.

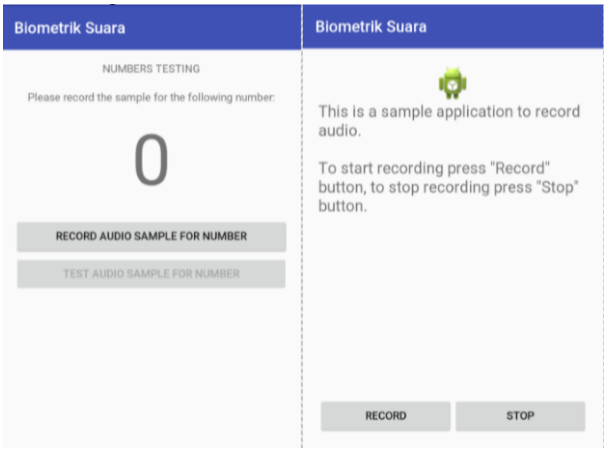

Figure 3: Application Interface

5.2 Result

5.2.1 Low Noise Condition Testing

Low noise condition testing is the result of matching when the environment of noise conditions is heard with low levels, such as the sound of small children's cries from a distance, people chatting from a distance, etc. At registration and testing as well as recording results, out of 50 users are divided into 5 groups in which each group has 10 members. In speech recognition, the test is repeated 3 times with the same word of each user and the matching results are recorded.

Test results on conditions around the test environment have low audible noise, in each group resulting in average variation of recognition as well as FRR values and FAR values obtained.

Group 1 gets an average introduction of 100% with an FRR value and a FAR value of 0%. Group 2 gets an average introduction of 60% with a FRR value of 40% and a FAR value of 20%. Group 3 gets an average introduction of 80% with a FRR value of 20% and a FAR value of 10%.

Group 4 gets an average introduction of 90% with FRR value and a FAR value of 10%. Meanwhile group 5 gets an average introduction of 60% with a FRR value of 40% and a FAR value of 30%. The graph of accuracy of each group on the environment condition sounds low-level noise is shown in Figure 4.

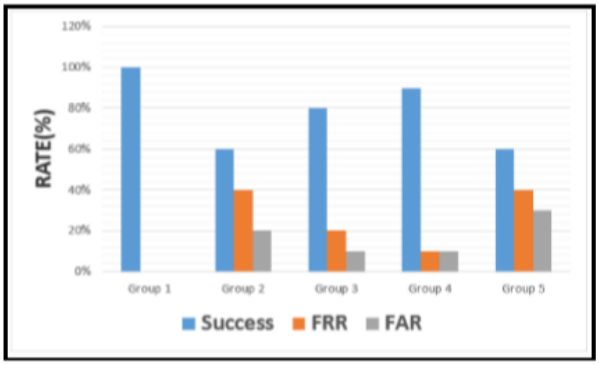

Figure 4: The Graph of Accuracy in Each Group

Figure 4 is a graph of accuracy in each group, in which the success rate of the system in the test can be said to be successful in voice recognition because the average introduction in each group can be said evenly, that is 60% to 100% which there is no result too low

5.2.2 High Noise Condition Testing

High noise condition testing is a test where conditions around the test is noisy and noise is caused as a child is playing closely, neighbors are turning on the sound system too loud, etc. The first test, under high environmental noise condition, is carried out on 50 users with one test. The second test, from 10 users taken in sequence to sample data entered in the database cleaned and removed from the noise and tested using noise obtained previously. Noise removal and cleaning is done using the Adobe Audition CS6 application. The following test scores for each user are shown in Table 1.

Table 1: Test Score 1 High Noise Condition.

No.	Kata	Klasifikasi Benar	Klasifikasi Salah	Rata-rata Pengenalan
1	Satu	1	0	100%
2	Dua	0	1	0%
3	Tiga	1	0	100%
4	Empat	0	1	0%
5	Lima	1	0	100%
6	Enam	1	0	100%
7	Tujuh	0	1	0%
8	Delapan	0	1	0%
9	Sembilan	1	0	100%
10	Sepuluh	0	1	0%
11	Sebelas	1	0	100%
12	Dua Belas	0	1	0%
13	Tiga Belas	0	1	0%
14	Gajah	0	1	0%
15	Lima Belas	0	1	0%
16	Enam Belas	0	1	0%
17	Tujuh Belas	0	1	0%
18	Delapan Belas	1	0	100%
19	Sembilan Belas	0	1	0%
20	Dua Puluh	0	1	0%
21	Dua Puluh Satu	0	1	0%
22	Dua Puluh Dua	1	0	100%

23	Dua Puluh Tiga	1	0	100%
24	Dua Puluh Empat	1	0	100%
25	Dua Puluh Lima	0	1	0%
26	Dua Puluh Enam	1	0	100%
27	Lima Puluh Enam	1	0	100%
28	Tujuh Puluh Delapan	0	1	0%
29	Dua Puluh Sembilan	1	0	100%
30	Tiga Puluh	1	0	100%
31	Domba	1	0	100%
32	Tiga Puluh Dua	1	0	100%
33	Tiga Puluh Tiga	0	1	0%
34	Tiga Puluh Empat	0	1	0%
35	Tiga Puluh Lima	0	1	0%
36	Tiga Puluh Enam	1	0	100%
37	Sembilan Puluh	0	1	0%
38	Dua Ratus Lima Puluh	1	0	100%
39	Seratus	1	0	100%
40	Delapan Puluh Tujuh	0	1	0%
41	Empat Puluh Satu	0	1	0%
42	Enam Puluh Tujuh	1	0	100%
43	Lima Puluh Lima	1	0	100%
44	Empat Puluh Empat	1	0	100%
45	Empat Puluh Lima	1	0	100%
46	Delapan Puluh Sembilan	0	1	0%
47	Enam Puluh Tiga	1	0	100%
48	Enam Puluh Sembilan	0	1	0%
49	Empat Puluh Sembilan	1	0	100%
50	Lima Puluh	1	0	100%
TOTAL		**26**	**24**	**52%**
Nilai FRR				**48%**
Nilai FAR				**32%**

Table 1 is the first test of high noise conditions where 50 user data are tested once test and high noise results result in less accurate matching with an average acquired 52%. The result is a FRR value of 48%, meanwhile for FAR value of 32%.

Table 2: Test Score 2 High Noise Condition.

No.	Kata	Klasifikasi Benar	Klasifikasi Salah	Rata-rata Pengenalan
1	Satu	3	0	100%
2	Dua	3	0	100%
3	Tiga	-	-	-
4	Empat	0	3	0%
5	Lima	0	3	0%
6	Enam	0	3	0%
7	Tujuh	3	0	100%
8	Delapan	3	0	100%
9	Sembilan	-	-	-
10	Sepuluh	3	0	100%
TOTAL		**15**	**9**	**62,5%**
Nilai FRR				**30%**
Nilai FAR				**30%**

Table 2 is a test of 10 sample data where 10 sampled data entered in the database are cleared and removed from noise and tested using data containing noise. The average introduction is 62.5% with FRR value and 30% FAR value. The result is not good because the effect of noise, user 3 and user 9 looks empty because when tested with high noise, the application does not find the matching data in the database entered. However, when the test data is done noise reduction of 100% using Adobe Audition CS6 application and re-tested; therefore, the data is not detected initially or wrong will rematch with existing data in the database.

5.2.3 The Silent Condition Testing

Testing data is carried out when conditions is quiet, the absence of noise interference from near or far, such as sound system, screams small children etc. Test is performed on 30 users by means of every 10 users who have registered and tested, then reenrolled 10 users until 30 users are enrolled and tested. The test graphs in silent environment conditions are described in Figure 5.

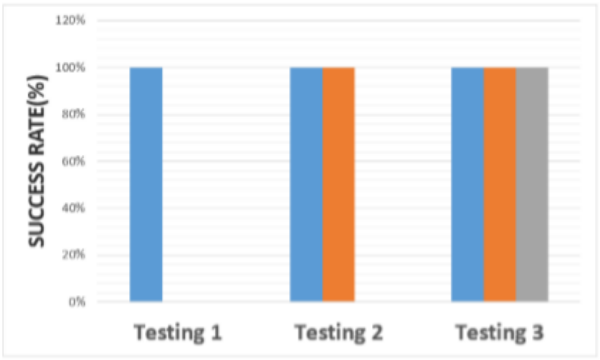

Figure 5: Graph of Accuracy Gradual Testing on Silent Condition

Figure 5 is a gradual testing graph in a quiet state. The results obtained are said to be successful because of all the tests at the time of silence the average introduction of 100% with the value of FRR and FAR value of 0%. 5.2.4 The accuracy comparison results of any environmental condition The accuracy comparison results of any environmental condition at registration or test are presented in the form of an accuracy comparison graph described in Figure 6.

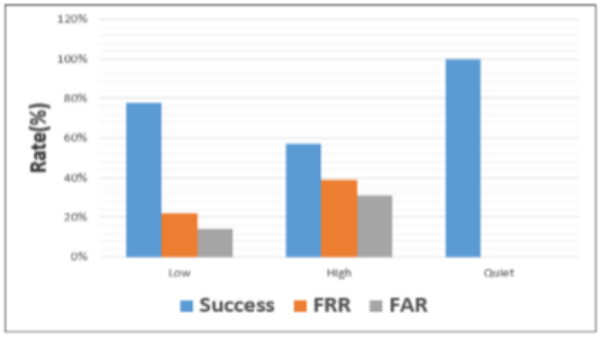

Figure 6: Graph of Comparison of Accuracy of Each Condition

Figure 6 is the average accuracy of the system during environmental conditions around the test with low noise conditions with a low level of 78%, 57.25% of high audible noise conditions and a 100% silent sound condition. Through the comparison graph on each condition, it can be concluded that noise on the environment is one of the factors that have an important effect on the quality of the system in the results of matching test data. Other factors can affect the system in the introduction of testing:

a. Environmental conditions.
b. Conditions and intonation of the user's voice.
c. Location of Microphone.
d. Ways of recording a voice signal.
e. Equipment condition.

6. CONCLUSION

Voice recognition is one type of the biometrics introduction. The use of the CMUSphinx library is as a part of an Androidbased Biometrics Application system that runs without an internet connection and it has a good performance in converting the analog voice signal into a digital sound signal using the MFCC method. Matching using the DTW method is a flexible mathematical method, giving high accuracy results. The best result is obtained when condition is quiet around the test environment, with all the tests matching the data stored in the database and the FRR value and the FAR value of 0%. Testing at a time when noise condition is heard with low levels can be said to be successful because almost all the tests detected match the data stored. Meanwhile, the bad result is when testing at highaudible noise conditions with a FRR value of 48% and a FAR value of 32%.

REFRENCES:

[1] K. Agustini, “Biometrik Suara dengan Transformasi Wavelet Berbasis Orthogonal Daubenchies,” Gematek J. Tek. Komput., vol. 9, no. 1, pp. 49–57, 2007.

[2] G. Melissa, “Pencocokan Pola Suara (Speech Recognition) dengan Algoritma FFT dan Divide and Conquer,” Makal. If2251 Strateg. Algoritm., 2008.

[3] D. Putra and A. Resmawan, “Verifikasi Biometrika Suara Menggunakan Metode MFCC Dan DTW,” Lontar Komputer, vol. 2, no. 1, pp. 8–21, 2011.

[4] B. R. Reddy and E. Mahender, “Speech to Text Conversion using Android Platform,” Int. J. Eng. Res. Appl., vol. 3, no. 1, pp. 253–258, 2013.

[5] B. J. Mohan and N. Ramesh Babu, “Speech recognition using MFCC and DTW,” 2014.

[6] A. H. Mansour, G. Z. A. Salh, and K. A. Mohammed, “Voice Recognition using Dynamic Time Warping and MelFrequency Cepstral Coefficients Algorithms,” Int. J. Comput. Appl., vol. 116, no. 2, pp. 34–41, 2015.

[7] C. Ho, Speaker Recognition System. California: California Institut of Technology, 1998.

[8] L. R, A. S, S. S, B. C, and G. Fernando, “Android Speech-to-speech Translation System for Sinhala,” Int. J. Sci. Eng. Res., vol. 6, no. 10, pp. 1660–1664, 2015.

[9] I. K. Suryadharma, G. Budiman, and B. Irawan, “Perancangan Aplikasi Speech To Text Bahasa Inggris ke Bahasa Bali Menggunakan Pocketsphinx Berbasis Android (Design Application Speech To Text English To Balinese Language Using Pocketsphinx Base On Android),” Bandung Univ. TELKOM, pp. 1–10, 2014.

[10] K. Chakraborty, A. Talele, and P. S. Upadhya, “Voice Recognition Using MFCC Algorithm,” Int. J. Innov. Res. Adv. Eng., vol. 1, no. 10, pp. 2349–2163, 2014.

[11] A. Setiawan, A. Hidayatno, and R. R. Isnanto, “Aplikasi Pengenalan Ucapan dengan Ekstraksi Mel-Frequency Cepstrum Coefficients (MFCC) Melalui Jaringan Syaraf Tiruan (JST) Learning Vector Quantization (LVQ) untuk Mengoperasikan Kursor Komputer,” TRANSMISI, vol. 13, no. 3, pp. 82–86, 2011.

[12] A. Muhammad, “Penggunaan Jarak Dynamic Time Warping (DTW) pada Analisis Cluster Data Deret Waktu (Studi Kasus pada Dana Pihak Ketiga Provinsi Seindonesia),” pp. 277–280, 2005.