Performance analysis method of dynamic time warping and k-nearest neighbor in sound based presence system

Herbet Simangungsong*, Herman Mawengkang, Maya Silvi Lidya

Department of Informatic, University of North Sumatra, Medan, Indonesia

*herbetandus0@gmail.com

Abstract. Verification and identification of a person using biometrics have been widely used as the retina of the eye, the face and voice. In this experiment, such as voice biometrics to identify a person who is used to the system presence. Voice recognition is done by pattern matching between training data and test data. In this study used methods Dynamic Time Warping (DTW), K-Nearest neigbors (KNN) and Fast Frequency Transform (FFT) for voice recognition. DTW is used as a method of pattern recognition, while KNN is used for sound classification. Before testing conducted prior extraction using FFT method. This study uses 100 votes out of 10 people with the amount of each 10 people. Presentations were used as training data by 70% and 30% of test data. Results obtained by dividing the recognized voice to the overall sound. From the results 83.33 % voice recognition.

1. Introduction

Verification and identification using biometrics have been widely used in various instances. Biometric is characteristic of human biology that is divided into two parts, which is based on physical characteristics and behavioral characteristics. Based on the physical characteristics are such as fingerprints (finger print), the retina of the eyes, hands, and face while behavioral characteristics are body temperature and sound. The process of identifying a person's fingerprint (finger print) is often used for presence, but still there are problems in the process of fingerprint identification caused by conditions such as abnormal fingers like wet, dirty and chipped.Chaudhari & Wagh (2014) in his study explaining fingerprint identification methods Crossing Number (CN) obtained results of up to 62.20% of the data from which the National Institute of Standards and Technology (NIST). In addition to identification by fingerprint researchers tried researching voice recognition using Dynamic Time Warping (DTW), K-Nearest Neighbor (KNN) and Fast Fourier Transform (FFT).

2. Basis Theory

Pattern recognition is a data classification to determine the degree of similarity between the testing and training data. These data may include images, text, video or voice (Murty and Devi, 2011). Sound is one biometric characteristic of human beings that are useful for communication. Everyone has a different sound so that the soundis very often used for security or secrecy and validity of the data. Voice recognition is done through two processes called training and testing. Training process is the stage to get a voice model that will be used as samples, while testing process is the phase matching with the voice models or samples that already exist.

2.1. Feature Extraction

Feature extraction is a process for separating useful sound of noise. This study uses FFT

as feature extraction. The stages of the FFT method is normalization, cutting, blocking until windowing frame. Normalization is used to calculate the average sound samples that will be used as a divider of signal sample values. Cutting used to cut frequencies to provide freuency limit value. Frame blocking is used to divide the signal into several parts while windowing is to convert the time domain into the frequency domain.

2.2. Clasiffication

Sound classification is used to classify the sound at a distance closer. In this experiment, K-Nearest neigbor method (KNN), the method is used to classify data based on distance. Phase KNN method of determining the number of k, calculates the distance of testing and training data, sortedfrom the smallest value to the largest value.

2.3. Pattern recognition

This study uses Dynamic Time Warping (DTW) for pattern recognition. DTW method is used to calculate the distance between the two frequencies. The closest value indicates the target on pattern recognition.

3. Research methods

This research will use 100 voices in the acquisition of 10 people with 10 votes each. Sounds used 70% or 70 voice as training data and 30% or 30 vote as testing data. The first voice in the extraction prior is to classify or to identify. Here the architecture of these stages.

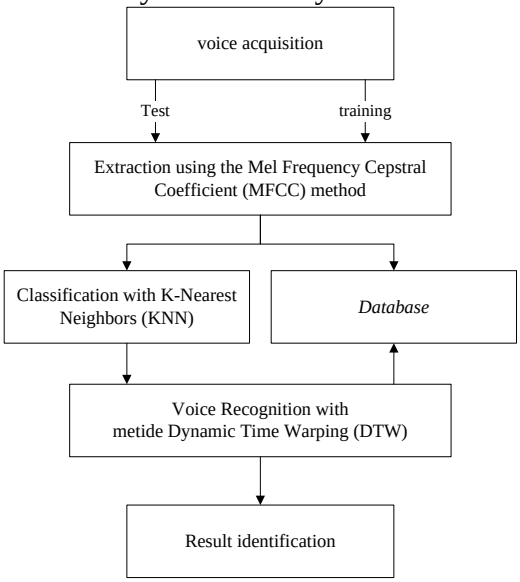


Figure 1. Architecture voice recognition

3.1. acquisition of Sound

Acquisition phase noise uses a voice recorder with the word "present" during 1ms with a frequency of 8000 Hz.

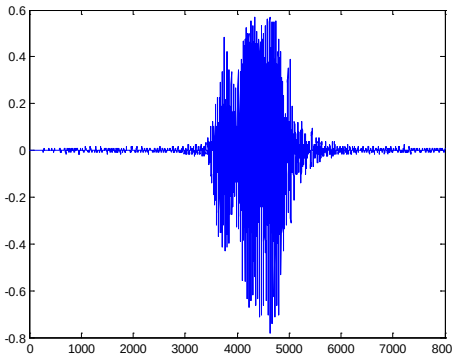


Figure 2. Frequency sound acquisition results

3.2. extraction Vote

Sound extraction stage using FFT method is normalization, cutting, frame blocking, windowing.

Step 1) *normalization*

The process of calculating the value of normalization inthe frequency of seeking maximum value at a frequency is used as a divider to the other data. The maximum value of the frequency is 0.570 so the first data on the frequency is -0.16.

$$y[n] = \frac{x}{\max(x)}$$

$$\bar{x} = \frac{-0{,}09}{0{,}570}$$

$$\bar{x} = \textit{-0.16}$$

$y[n]$ =normalization result
x =score of signal sample
$\max(x)$ =maximum score from signal sample

Step 2) *cutting*

Cut the frequency values with the upper limit of 0.3 and below -0.3.

$$ba[n] = 0{,}3 < limit < -0{,}3$$

$$cu[n] = ba * 0.25$$

ba [n] = maximum and minimum limits
cu [n] = result of cutting frequency
0:25 = ¼ of frequency

Step 3) *frame Blocking*

With a frequency of 8000Hz with 1s used time frame 128.

Where :
Ts =Time of taking the voice (ms)
M =frame length (ms)
Ts(m)= frame accumulation

Step 4) *windowing*

windowing used to reduce the gap between the early and late signals.

$$x(n) = x_i(n)w(n)$$

$x[n]$ =signal sample of windowing result

$x_i[n]$ =signal sample score from signal frame to i
$w[n]$ =window function

Step 5) *Fast Fourier Transform* (FFT)
*Fast Fourier Transform*is used to convert the time domain into the frequency domain. FFT formula is:

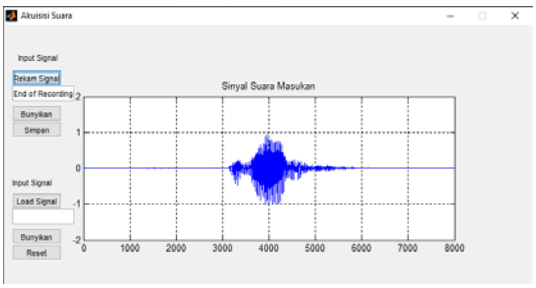
$$f(n) = \sum_{k=0}^{n-1} y_k e^{-2\pi jkn/N}, n = 0,1,2,3 \dots N-1$$

Where :
$f[n]$ = frequency
N =sample of each frame
k =window function
j =imajiner

4. Results and Discussion

Testing of testing data to trainingis a process to determine the suitability of the sound that determines whether the voice is recognized or not. The results of this study using K-Nearest neigbors (KNN), Fast Fourier Transform (FFT) and Dynamic Time Warping (DTW).

4.1. acquisition sound



4.2. training data

In this study, the frequency is divided into multiple frames, and researchers used as training data at a frame 128. The following data 1 after extractions takes the value as much as 128.

To-person voice	1	2	3	4	5	6	:::::>	127	128	Target
1	0.272	.390	2.402	6.123	2.832	8.490	:::::>	2.402	.390	*hhhh*
	0,003	0.045	0.122	0.061	0.099	0,042		0.122	0.045	*hhhh*
	0.088	.202	1,715	4.369	2,300	6.093		1,715	.202	*hhhh*
	0,027	0.198	1,620	4.256	2.125	4.359		1,620	0.198	*hhhh*
	0.273	.255	1,657	4.687	3.203	1,739		1,657	.255	*hhhh*
	0,085	0.077	2.342	4,290	1,182	6.709		2.342	0.077	*hhhh*
	0.126	.130	2,470	5.592	2,344	7.005		2,470	.130	*hhhh*

4.3. The test data

In this study, the frequency is divided into multiple frames, and researchers used as training data at a frame 128. The following data 1 after extraction takes the value as much as 128.

To-person voice	1	2	3	4	5	6	:::::>	127	128	Target
1	0.149	0,071	0.968	3.104	1,482	2,274	:::::>	0.968	0,071	*hhhh*
	0.134	0.306	2,817	2.908	2,072	2.727		2,817	0.306	*hhhh*
	0,287	0.163	.748	3.788	2.575	1,182		.748	0.163	*hhhh*

4.4. classification of sound

Sound classification of test data to training data is done using K-Nearest Neighbor (KNN) to obtain the shortest distance. KNN The results will be used as training data for testing data using the Dynamic Time Warping. From the sound of testing data to 1 in the training data is to test as many as 70 voices. The test results are:

28.886 25.398 13.597 12.072 27.647 13.369 19.759 25.347 25.268 25.245 25.364 26.247 30.698 23.375 23.443 25.531 25.391 25.457 25.403 25.311 21.488 18.268 15.037 25.345 25.364 23.166 21.496 22.877 18.626 22.685 33.381 28.830 25.479 23.242 25.897 24.591 25.149 25.456 24.869 24.607 40.320 22.456 25.471 24.598 25.431 20.879 17.502 18.996 26.356 25.415 25.461 28.679 23.622 21.187 25.428 25.292 18.865 17.380 23.822 27.516 22.775 23.497 24.992 25.356 26.283 24.978 25.207 24.314 29.136 27.951,

then from KNN method results sorted from the smallest value to the greatest value of nine samples.

to-person	*To-Test Data*	Distance Value	Target	Unrecognized / Unrecognized
1	1	12.07	*hhhh*	*unrecognizable*
	2	8.90	*hhhh*	*unrecognizable*
	3	10.83	*hhhh*	*unrecognizable*
2	1	-	*bbbb*	*Unrecognized*
	2	9.70	*bbbb*	*unrecognizable*
	3	16,90	*bbbb*	*unrecognizable*
3	1	14.38	*JJJJ*	*unrecognizable*
	2	14.53	*JJJJ*	*unrecognizable*
	3	22.39	*JJJJ*	*unrecognizable*
4	1	12.32	*pppp*	*unrecognizable*
	2	13,02	*pppp*	*unrecognizable*
	3	20.59	*pppp*	*unrecognizable*
5	1	-	*aaaa*	*Unrecognized*
	2	9.90	*aaaa*	*unrecognizable*
	3	15.57	*aaaa*	*unrecognizable*
6	1	-	*cccc*	*Unrecognized*
	2	11.76	*cccc*	*unrecognizable*
	3	8.90	*cccc*	*unrecognizable*
7	1	12.89	*dddd*	*unrecognizable*
	2	13,02	*dddd*	*unrecognizable*
	3	-	*dddd*	*Unrecognized*
8	1	22.39	*oooo*	*unrecognizable*
	2	12.34	*oooo*	*unrecognizable*
	3	15,50	*oooo*	*unrecognizable*
9	1	17.20	*nnnn*	*unrecognizable*
	2	20.12	*nnnn*	*unrecognizable*
	3	13.35	*nnnn*	*unrecognizable*
10	1	14.46	*yyyy*	*unrecognizable*

	2	9.80	*yyyy*	*unrecognizable*
	3	-	*yyyy*	*Unrecognized*

$$Accuracy\ \frac{amount\ of\ data\ recognition}{total\ data}\ x\ 100\%$$
$$Accuracy = \frac{25}{30}\ x\ 100\%$$
$$Accuracy = 83{,}33\ \%$$

5. Conclusion

After analizing by adding KNN method between FFT and DTW obtain the level of accuracy of 83.33%. With known voice as much as 25 voices and sounds are not known as much as 5 voice.

6. References

[1] Andhini, RB, Irawan, B., & Vitello, I. 2015. Analysis and Implementation of Application Being Text Voice Recognition Method Using Backpropagation Neural Network. e-Proceedings of Engineering: Vol.2, No.2: 3526-3532

[2] Bansal, P., Imam, SA and Bharti, R. 2015. Speaker Recognition using MFCC, shifted MFCC with Vector Quantization and Fuzzy. International Conference on Soft Computing Techniques and Implementations- (ICSCTI) Department of ECE, FET, MRIU, Faridabad, India: 41-44

[3] Chaudhari, GN, & Wagh, RB 2014. A Novel Approach for Latent to Rolled Fingerprint Matching. International Conference on Contemporary Computing and Informatics (IC3I). 387-391

[4] Dinata, C., Puspitaningrum, D. & Ernawati. 2017. Implementation Techniques Dynamic TimeWarping (DTW) in the Text To Speech Applications. Journal of Information Engineering Vol.10, No.1: 1-5.

[5] Fajrin, T. & Nurina, AF Analysis Presence system with fingerprint students of SMK Negeri 2 Karangayar. Journal of speed - Enginering Center for Research and Education Vol. 3: 78-83

[6] Fitrilina, Kelly, R., & Aulia, S. 2013. Speech Recognition MFCC-HMM method for Motion commands Robot Car Tracker Color Identification. National Journal of Electrical Engineering: 31-40