Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Speech recognition using Dynamic Time Warping (DTW)

To cite this article: Yurika Permanasari *et al* 2019 *J. Phys.: Conf. Ser.* **1366** 012091

Speech recognition using Dynamic Time Warping (DTW)

[1]**Yurika Permanasari,** [2]**Erwin H. Harahap,** [3]**Erwin Prayoga Ali**

*[1,2,3]*Departement of Mathematic, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Islam Bandung, Jl. Tamansari No.1 Banung 40116 Indonesia

email: [1]Yurikakoe@gmail.com, [1]Erwin2h@gmail.com, [3]Erganear0808@gmail.com

Abstract. Sound is one of the most common communication medias used by humans. Every human has different sound characteristics. To recognize the compatibility of a sound, a special algorithm is needed, which is Dynamic Time Warping (DTW). DTW is a method to measure the similarity of a pattern with different time zones. The smaller the distance produced, the more similar between the two sound patterns. Both sound patterns are similar, thus the two voices are said to be the same. The initial data on the speech recognition process is transformed into frequency waves. Pronounce volume, pronunciation time, and noise from the sound around the recording takes place affecting the distance generated. The smaller the effect, the smaller the distance that will be generated.

Keywords: Speech Recognition, Sound Pattern, Dynamic Time Warping (DTW).

1. Introduction

Sound is one of the most frequently used communication medias by humans. Humans can produce sound easily without requiring great energy. In addition, every human being has different forms of sound so that they can be distinguished according to the character of each voice [1]. Human voice is one of the biometric technologies that can be used to identify someone (person identification) [2]. Biometrics technology is the introduction of special characteristics in an individual such as iris, face, fingerprint, and others. The biometrics system has characteristics that cannot be forgotten and are not easily lost.

As technology develops, sound can be translated into digital data that is understood by computers. So that words spoken by humans can be used as a system so that computers can recognize commands from human voices. The system is called speech recognition. This system identifies the sound based on the spoken word by converting an acoustic signal captured by the sound input device. Algorithm is needed for matching between the initial words used as speech recognition and inputted words, while it is impossible for us to be able to equate the pronunciation time. Dynamic Time Warping (DTW) is a method to measure the similarity of a pattern with a different time zone [3].

2. Speech recognition

Speech recognition is the process of identifying sounds based on words spoken by converting an acoustic signal, which is captured by an audio device. Speech recognition is also a system used to recognize word commands from human voices and then translate into data that is understood by computers. This technology allows a device to recognize and understand the words spoken by digitizing words and matching the digital signals with a certain pattern stored in a device [4].

The words spoken are transformed into digital signals by converting sound waves into a set of numbers which are then adjusted to certain codes to identify those words. The results of the identification of spoken words can be displayed in written form or can be read by technology devices as a command to do a job, for example a search command on Google by saying "ok google".

3. Dinamic time warping

Dynamic time warping is an algorithm that calculates the optimal warping path between two data from sound so that the output is the path warping values and the distance between the two data. Warping path is the distance between a comparison of two patterns, the smaller the warping path that is produced, the two patterns can be said to be the same.

Two words from the same word by the same user can have different times. For example, two can be pronounced with two or twoo. DTW solves this problem by aligning words correctly and calculating the minimum distance between two words [5].

Different timing of speech alignment is a core problem for distance measurement in speech recognition. Small shifts result in incorrect identification. DTW is an efficient method for solving time alignment problems. Therefore this algorithm is more realistic to be used in measuring the similarity of a pattern (pattern / template matching) [6]. The processed data is always in the time zone, so that the sequence of data we have is considered to vary with time. The illustration of matching with the DTW method is shown in the figure [7]

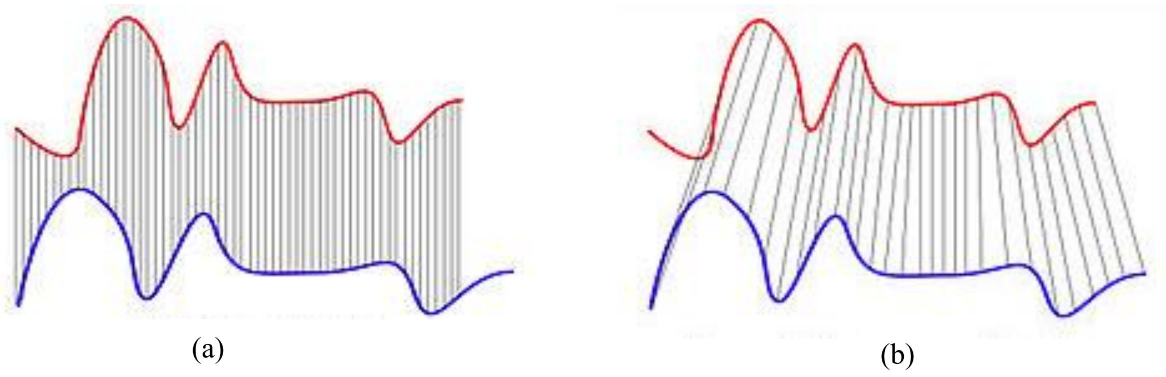


Figure 1. (a) The original alignment of two sequences (b) alligments with DTW

The DTW algorithm is intended to align two vector sequences by turning the time axis repeatedly until the optimal match between the two sequences is found. This algorithm performs as a linear mapping of the axis to align the two signals. Suppose there are two vector sequences in n-dimensional space:

$$x = [\quad x_1 \quad x_2 \quad \dots \quad x_n \quad] \text{ and } y = [\quad y_1 \quad y_2 \quad \dots \quad y_n \quad]$$

The two sequences are aligned on the sides of the box, with one above and the other on the left side. Both sequences start at the bottom left of the grid.

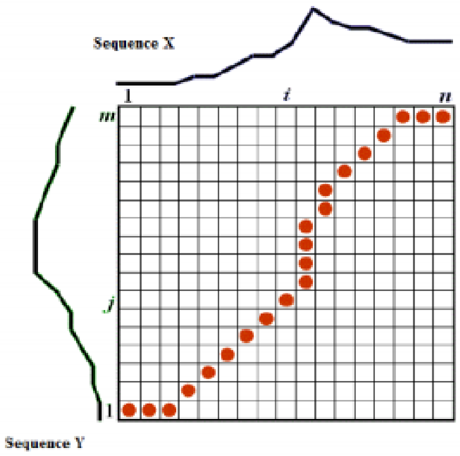


Figure 2. Scatter diangram of Matching sequence

In each cell, a measure of distance is placed, comparing the corresponding elements of two sequences. The distance between two points is calculated through Euclidean distance.

$$Dist\ (x,y) = |\ x - y\ | = [\ (x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + \ (x_n - y_n)^2]^{\frac{1}{2}} \tag{1}$$

The best match or alignment between these two sequences is the path through the grid, which minimizes the total distance between the two, which is called global distance. The entire distance (global distance) is calculated by finding and going through all possible routes through the grid, each calculating the overall distance.

The minimum global distance from the number of distances (Euclidean distance) between individual elements on the road divided by the number of weighting functions. For each sequence long enough the number of possible paths through the grid will be very large. The optimal value function is defined as *D (i, j)* as the DTW distance between *t (i: m)* and *r (j: n)* with the mapping path from *(i, j)* to *(m, n)*

$$D(i,j) = |t(i) - r(j)| + min\begin{Bmatrix} D(i+1,j) \\ D(i+1,j+1) \\ D(i,j+1) \end{Bmatrix} \tag{2}$$

with initial conditions $D(m,n) = |t(m) - r(n)|$

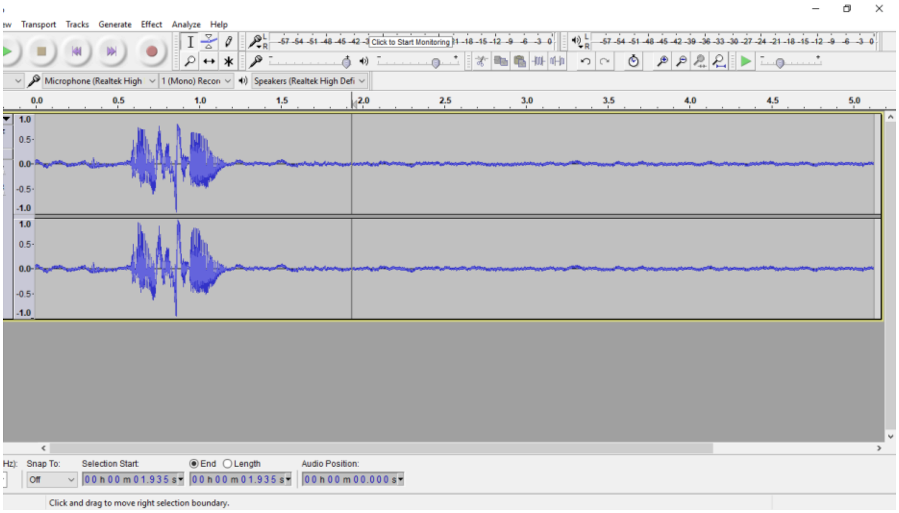
4. Experiments and results

In the initial process carried out for speech recognition is the process of storing training data. The training data process is used to store user voice references. In the process of training the first user to do the sound recording process using a microphone, after which the sound traits obtained from the training data were stored in the database database.

After making reference data for training data, then the test data (testing) are then carried out. The test data has stages that are almost the same as the process in the training data. In the process of this test data, the user returns to recording the sound in * .wav. After obtaining sound characteristics from the test data, matching was done using the DTW method. By calculating the distance from the two vectors between the data tested with the data that is in the reference data.

4.1 Sound data

Data in the form of sound signals are recorded using a microphone connected to a computer. Voice recording is done with the help of the Audacity program, with a sampling frequency of 16000 Hz, a stereo channel.



Snap To: Off
Selection Start: 00 h 00 m 01.935 s
● End ○ Length: 00 h 00 m 01.935 s
Audio Position: 00 h 00 m 00.000 s

Click and drag to move right selection boundary.

Figure 3. Sound signal results from Audacity

Then the sound is saved as an ".wav" file. then the voice data is entered into the Matlab program.

Tabel 1. The voice data details used

Data Name	Time Pronouncuation (s)	Sound Volume (Db)
User1.1	0.35	0
User1.2	0.410	0
User1.3	0.850	0
User1.4	1.00	0
User1.5	0.35	+1
User1.6	0.35	-1
User1.7	0.35	+4
User1.8	0.35	0
User1.9	0.35	0
User1.10	0.35	0
User2.1	0.35	0
User2.2	0.420	0
Mail3	0.35	+4

4.2 Warping process

Three things that are considered in the compatibility test as follows:

(i) The same time of pronunciation and the same volume

(ii) The same pronunciation time and different volume

(iii) Different pronunciation times and the same volume in the test match there are two users who make voice input namely user 1 and user 2. User 1 has ten samples and two samples for User 2, with training data used is User 1.1.

The results are :

(i) The same time of pronunciation and the same volume

In this trial the pronunciation time starts at seconds to 0.35 with a sound volume of 0 db (normal sound is not amplified or reduced). The results as follows:

Data Name	**Output DTW**
User 1.8	15.2730
User 1.9	12.5054
User 1.10	14.7427
User 2.1	223.5862

(ii) Same pronunciation time and different sound volume

This trial time of speech starts at seconds to 0.35 with volumes including + 1db (sound is hardened by 1db), -1db (sound is reduced by 1db), + 4db (sound is loud at 4db), the results are :

Data Name	**Output DTW**
User 1.5	14.8578
User 1.6	14.5985
User 1.7	61.1329
User 2.3	248.0940

(iii) Different pronunciation times and the same volume

In this trial the pronunciation time is done randomly with the volume used is 0db (the sound is not hardened or reduced). The results obtained from User1 as training data and the rest are test data :

Data Name	**Output DTW**
User 1.2	206.1125
User 1.3	200.7430
User 1.4	199.3750
User 2.2	513.3904

5. Conclusion

From the results of the trial, with 1 training data and 11 test data, the smallest distance is obtained when the volume is the same and the pronunciation time is 12.5. Acquiring distance value of User1 which is the same person is a minimum of 12.5 maximum 15.3 While the value of distance from User2 (different people) the value of the distance is quite large, namely 223. The difference in the value of a considerable distance indicates that Dynamic Time Warping (DTW) is quite good for voice recognition.

The smaller the distance that is generated then the sound will be more similar to the initial sound input, in other words the sound can be recognized. The pronunciation time and volume affect the distance that is generated. The closer the pronunciation distance of the test data to the training data the smaller the distance. The higher the volume of the test data, the greater the distance generated.

The trial was conducted with 3 conditions, namely the same pronunciation time and the same volume, different pronunciation times with the same volume, and the same pronunciation time and different volume. Of the three conditions, the minimum distance value is obtained at the same pronunciation time and the same volume with the minimum distance produced 12.5 compared with different pronunciation time periods with the same volume of 199.4 and minimum distance at the time of pronunciation the same and the volume is different at 14.6. Thus it can be concluded that the pronunciation time is more influential than the volume. So that the more similar the pronunciation time between test data and training data, the distance generated will also be smaller.

References

[1] A. R. Darma Putra, "Verifikasi Biometrika Suara Menggunakan Metode MFCC dan DTW," *Lontar Komputer UNUD,* vol. 2 No.1, 2011.

[2] J. Riyanto, "Perangkat Lunak Pengenalan Suara (Voice Recognition) untuk Absensi Karyawan dengan Menggunakan Metode DTW," Fakultas Teknik dan Ilmu Komputer, UNIKOM, Bandung, 2011.

[3] R. Varathrajan, "Wearable Sensor Devices for Early Detection of Alzheimer Disease Using DTW Algorithm," *Cluster Computing,* vol. 21, no. 1, pp. 681-690, 2018.

[4] R. W. S. Lawrence Rabiner, Theory and Application of Digital Speech Processing, Upper Saddle River New York: Prentice Hall, 2011.

[5] K. Paliwal, "A Modification over Sakoe and Chiba's Dynamic Time Warping Algorithm for Isolated Word Recognition," *Signal Processing,* vol. 4, pp. 329-333, 1982.

[6] R. J. Kate, "Using DTW Distance As Features for Improved Time Series Classification," *Data Mining and Knowledge Discovery,* vol. 30, no. 2, pp. 283-312, 2015.

[7] Y. I. Nurhasanah, "Aplikasi Pendeteksi Emosi Manusia Menggunakan Metode MFCC dan DTW," in *Seminar Nasional Teknologi Informasi ITENAS*, Bandung, 2016.