Translation Invariant Approach for Measuring Similarity of Signals

A.Darvishi
Department of Computer and Electrical Engineering
Babol Noushirvani University of Technology,
P.O.Box 47144, Babol, Iran
E-mail: alidarvishi.nit@gmail.com

Abstract

In many signal processing applications, an appropriate measure to compare two signals plays a fundamental role in both implementing the algorithm and evaluating its performance. Several techniques have been introduced in literature as similarity measures. However, the existing measures are often either impractical for some applications or they have unsatisfactory results in some other applications. This problem becomes more evident when signals involve translations in amplitude and time. This paper presents a new one dimensional similarity measure to compare two signals. The proposed measure accepts transformations like time-shift, amplitude-scale, amplitude-shift, and phase delay in measuring the similarity. The results in this paper indicate that the proposed approach overcomes exiting techniques in measuring similarity among different signals.

Keywords: *Similarity measure, Time shift, Amplitude shift, Amplitude scale, Phase delay.*

1. Introduction

Similarity measure is a vital tool in many applications, such as in decision making, pattern classification, and data mining. Depending on the application, a k-dimensional (k>0) similarity measure may be needed for measuring the similarity. This paper considers a one-dimensional similarity measure which is an important tool in many signal processing applications.

Quantifying similarity by developing a suitable similarity measure is often a difficult task. Currently, there are many measures that can be used for quantifying the similarity [1,2,7]. These measures are classified into three different groups in this paper: distance measures, proximity measures, and binary measures. In distance measures individual samples of a signal are correspondingly compared with the individual samples of the other signal in quantifying the similarity. In other words, in this group only the raw data of the two sequences are used in measuring the similarity. In proximity measures, some extra information about the signals, like the average value and standard deviation, may also be used in quantifying the similarity apart from the raw data used for the computation in the distance measures. The binary measures are used to quantify the similarity among binary data.

In this paper, two signals are considered as similar if difference of the two signals or their translations (like time-shift, amplitude shift, amplitude scale) is considerably low.

Although much research has been performed on similarity measures, the combination of different similarity measure have rarely been considered [4]. Majority of existing similarity measures fail to search proximities in the signals when dealing with some situations like time-shifting, amplitude-scaling, amplitude-shifting (DC-offset) and phase-delay. Hence, it may be desirable to propose a similarity measure algorithm that considers the overall characteristics of the signals, not the raw values of the individual samples in the two signals under consideration.

This paper reviews the existing techniques of similarity measure, and then proposes a new similarity measure. The results in this paper indicate that the proposed approach outperforms exiting similarity measuring methods.

2. Reviewing existing similarity measures

There are a number of similarity measures in the literature, each of which may have a different view on the signal. These similarity measures can be categorized into three groups described in the following subsections.

2.1. Distance Measures

Measures in this group consider samples of the two sequences one-by-one. In this group, the similarity measure d(x,y), considers two sequences x and y, should satisfy the following criteria[3]:

- Positivity: for any x and y, d(x, y) is a real number greater than or equal to zero. The d is zero if and only if x=y.
- Symmetry: d(x, y) = d(y, x).
- Triangle inequality: d(x, z) ≤ d(x, y) + d(y, z)

2.1.1. Euclidian Distance

The distance between two sequences, x and y, is the square root of the sum of the squared differences between the values of the sequences:

$$Euclid\,(x\,,y\,)=\sqrt{\Sigma_i\,(x_{\,i}\,-y_{\,i}\,)^2} \qquad (1)$$

1.1.2. Chebychev

The distance between two sequences is the maximum absolute difference between the values of the sequences:

$$Chebychev\,(x\,,y\,)=\max_i\left|x_{\,i}\,-y_{\,i}\,\right| \qquad (2)$$

2.1.3. Block (Hamming)

The distance between two sequences is the sum of the absolute differences between the values of the sequences:

$$\text{Block}(x, y) = \sum_i \left| x_i - y_i \right| \tag{3}$$

2.1.4. Minkowski

The distance between two sequences is the p th root of the sum of the absolute differences to the power of p between the values of sequences:

$$\textit{Minkowski}(x, y) = \left(\sum_i \left| x_i - y_i \right|^p \right)^{1/p} \tag{4}$$

Indeed, the Minkowski distance can represent some other similarity measures from this group. For example, by setting the p to 1 and 2 it represents the Hamming and Euclidian measures, respectively.

2.2. Proximity Measures

The measures categorized in this group consider the general characteristics of the two sequences. These measure are often applicable to specific applications [1,2,3,7]. These measures may have more intuitive results compared to the measures in the previous group.

2.2.1. Cosine

This is an angle-based similarity measure that expresses the similarity between the normalized versions of the two sequences. In other words, only directions of the unit vectors representing the two sequences are considered in this measure:

$$\text{Cosine}(x, y) = \frac{\sum_i (x_i y_i)}{\sqrt{\left(\left(\sum_i x_i^2 \right) \left(\sum_i y_i^2 \right) \right)}} \tag{5}$$

This measure accepts the amplitude scale.

2.2.2. Correlation

The Correlation measure can be considered as a measure representing correlation of the two sequences x and y under comparison:

$$Correlation(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (6)$$

where $\bar{x}=\frac{1}{n}\sum_{i}^{n}x_i$ and $\bar{y}=\frac{1}{n}\sum_{i}^{n}y_i$ are the mean of x and y .

This measure accepts the amplitude shift.

2.2.3. Chi-Square

This similarity measure depends on the total frequencies of the two sequences or variables whose proximity is computed. Expected values are from the model of independence of cases (or variables), x and y.

$$Chisq(x,y)=\sqrt{\sum_{i}\frac{\left(x_i-E\left(x_i\right)\right)^2}{E\left(x_i\right)}+\sum_{i}\frac{\left(y_i-E\left(y_i\right)\right)^2}{E\left(y_i\right)}} \quad (7)$$

This measure accepts the phase delay.

2.2.4. Jensen

The output of the Jensen function is between 0 and 1. If two sequences P and Q are exactly the same, the output of the Jensen function will be equal to zero [6]:

$$Jensen\ (P,Q)=$$
$$\frac{1}{2}\sum_{i=1}^{k}\{p_i'\log_2 p_i'+q_i'\log_2 q_i'-(p_i'+q_i')\log_2((p_i'+q_i')/2)\} \quad (8)$$
$$Where \quad p_i'=p_i/\sum_{i=1}^{k}p_i \quad , \quad q_i'=q_i/\sum_{i=1}^{k}q_i$$

This measure accepts the amplitude scale. It should be noted that this function can be used for comparing sequences with a positive value for each samples.

2.3. Binary Measures

In some applications, we may need to compare sequences with a binary value. The similarity measures described in the two previous groups may not be suitable for comparing binary sequences. To quantify the similarity between these sequences, four summation variables a, b, c and d are used which contains the number of positive matches (1-1), left mismatches (0-1), right mismatches (1-0), and the number of negative matches (0-0) between the individual corresponding samples of the two

sequences [5]. Binary measures can be extended to non-binary data by modification of the summation variables as described in [7].

2.3.1. Jaccard

This is one of the most popular measures that also known as the similarity ratio:

$$Jaccard(x, y) = \frac{a}{a+b+c} \tag{9}$$

2.3.2. Ochiai Similarity Measure

This is the binary form of the Cosine measure. The value of this measure varies between 0 and 1:

$$Ochiai(x, y) = \sqrt{\left(\frac{a}{a+b}\right)\left(\frac{a}{a+c}\right)} \tag{10}$$

2.3.3. Hamann

This measure gives the probability that a characteristic has the same state in both sequences minus the probability that a characteristic has different states in the two sequences. The value of this measure varies between –1 to +1:

$$Hamman(x, y) = \frac{(a+d)-(b+c)}{a+b+c+d} \tag{11}$$

2.3.4. Goodman and Kruskal Lambda

This measure assesses the predictability of the state of a characteristic in one sequence given the state in the other sequence. Specifically, this measure quantifies the proportional reduction in error using one sequence to predict the other, when the directions of prediction are of equal importance

$$\begin{aligned} t_1 &= \max(a,b) + \max(c,d) + \max(a,c) + \max(b,d) \\ t_2 &= \max(a+c, b+d) + \max(a+b, c+d) \\ Lambda(x, y) &= \frac{t_1 - t_2}{2(a+b+c+d) - t_2} \end{aligned} \tag{12}$$

Lambda has a range of 0 to 1.

3. Novel similarity measure

In this paper a new similarity measure is introduced. In this measure we integrate the strengths of both Cosine and Correlation measures that are invariant to amplitude-

scaling and amplitude-shifting respectively. We show that this new measure has additional features; it is invariant to time-shifting and phase-delay. The algorithm of this novel similarity measure is described below.

Given two sequences $X = (x_1,x_2,\ldots,x_n)$ and $Y= (y_1,y_2,\ldots,y_m)$, where generally $n \neq m$, by assuming $m \leq n$ the proposed algorithm takes the following steps to measure the similarity:

I. $$m_Y = \frac{1}{m}\sum_{i=1}^{m} Y_i$$

II. $$X_k = Circshift\ (X,(-k+1)) \quad , \quad m_{X_k} = \frac{1}{m}\sum_{i=1}^{m} X_{k_i}$$
$$k = 1,2,\ldots,n$$

III. $$S_k = \frac{\sum_{i=1}^{m}(X_{k_i} - m_{X_k})(Y_i - m_Y)}{\sqrt{\sum_{i=1}^{m}(X_{k_i} - m_{X_k})^2 \sum_{i=1}^{m}(Y_i - m_Y)^2}}$$

IV. $$Sim(x,y) = \underset{1 \leq i \leq n}{Max}\left[S(i)\right]$$

In this algorithm, circshift (S, shiftsize) is a circular shift operator which circularly shifts S by shiftsize samples. The shiftsize is an integer scalar. If the value of shiftsize is positive, S is shifted to the right, otherwise shifted to the left. The basic idea of this algorithm is to consider the sequences under comparison as circular. Then the similarity of the two sequences is examined at different starting points. The maximum similarity obtained at any starting point is considered as the similarity of the two sequences. For the case that the two sequences have unequal lengths, the longer sequence is circularly shifted by k samples (k varying from 1 to n). Then m sub-samples of the longer sequence are compared with the shorter sequence to quantify the similarity.

In Table 1, some properties of the proposed approach have been compared with the existing similarity measures.

As the table shows none of the Distance measures has any of the properties mentioned in this table. How ever, among this group, the normalized Euclidian accept amplitude-shift and amplitude-scale. These properties are evaluated in the next section.

Table 1. *Comparing properties of the proposed measure with the existing similarity measures.*

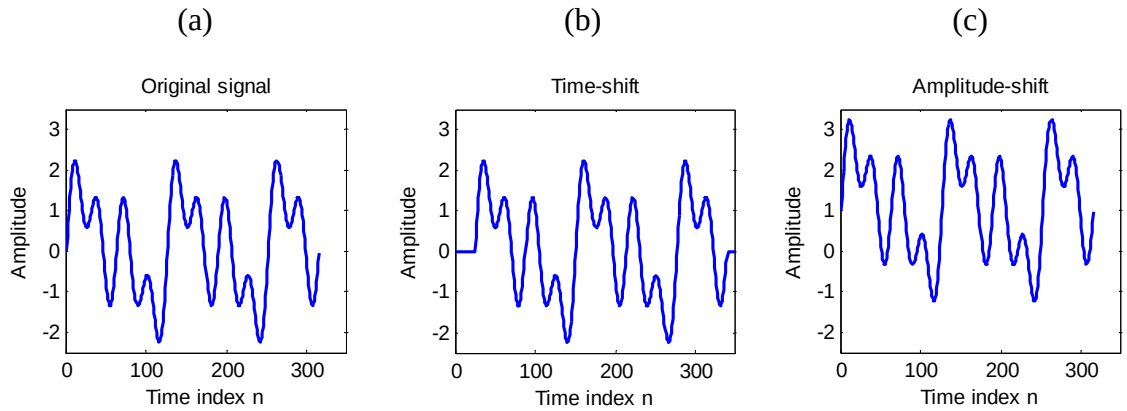
Measures	Properties				
	Time-shift	**Amplitude-shift**	**Time-scale**	**Amplitude-scale**	**Phase-delay**
Distance measures	**Í**	**Í**	**Í**	**Í**	**Í**
Normalized Euclidian	**Í**	**P**	**Í**	**P**	**Í**
Chi-Square	**Í**	**Í**	**Í**	**Í**	**P**
Jensen	**Í**	**Í**	**Í**	**P**	**Í**
Cosine	**Í**	**Í**	**Í**	**P**	**Í**
Correlation	**Í**	**P**	**Í**	**P**	**Í**
Proposed measure	**P**	**P**	**Í**	**P**	**P**

4. Evaluation

In order to assess the performance of the proposed approach and to compare its performance with the existing similarity measures a multi components signal is used as an original signal and the following transformations are considered which map a point (t,x) from original signal into a point ($\bar{t}$, $\bar{x}$):

- Time shift: T1: $\bar{t} = t + e, \quad \bar{x} = x$
- Amplitude shift: T2: $\bar{t} = t, \quad \bar{x} = x + e$
- Time scale: T3: $\bar{t} = (1+e)t, \quad \bar{x} = x$
- Amplitude scale: T4: $\bar{t} = t, \quad \bar{x} = (1+e)x$
- Phase delay: T5: $Y = \text{Phase\_Delay}(X, q)$

Figure 1 shows the original signal and any of the above mentioned transformation on it. It is expected that the similarity measure can consider any of these signals as similar to the original signal.



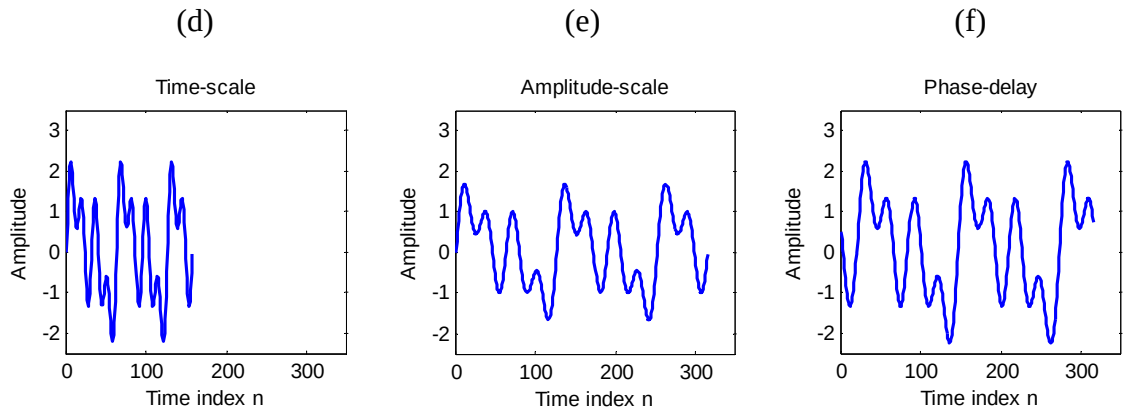


Figure 1. *Test signals used to assess performance of different measures: (a) Original signal, output of transformations: (b Time-Shift, (c) Amplitude-Shift, (d) Time-Scale, (e) Amplitude-Scale, (f) Phase-Delay.*

Among different similarity measures introduced in Section 2, we have only chosen one similarity measure from the Distance measures (the normalized Euclidian measure), and one from the Proximity measures (the Cosine measure). Table 2 shows the results of these similarity measures on the signals shown in Figure 1.

Table 2. *Outputs of different similarity measures in comparing the signals in Figure 1 with the original signal (Figure 1-a).*

Similarity Measure	Normalized Euclidian	Cosine	Proposed measure
Output Range	[0 1]	[-1 1]	[0 1]
Time-Shift	0.62635	-0.0336	1
Amplitude-Shift	1	0.7363	1
Time-Scale	0.75329	0.48786	0.69814
Amplitude-Scale	1	1	1
Phase-Delay	0.58195	-0.16454	1
Original Signal	1	1	1

To further evaluate the proposed approach with the existing similarity measures, we have added a white Gaussian noise ($m = 0$ and $s = 7$) to the signal in Figure 1. This experiment has been repeated 100 times and the averaged results are shown in Table 3. As this table shows, the proposed approach outperforms the exiting similarity measures.

Table 3. *Outputs of different similarity measures in comparing the noise corrupted signals in Figure 1 with the original signal (Figure 1-a). (the averaged results on 100 realizations)*

Similarity Measure	Normalized Euclidian	Cosine	Proposed measure
Output Range	[0 1]	[-1 1]	[0 1]
Time-Shift	0.6541	-0.0337	0.9610
Amplitude-Shift	0.9134	0.7235	0.9611
Time-Scale	0.7522	0.4675	0.6688
Amplitude-Scale	0.9677	0.9954	0.9955
Phase-Delay	0.6169	-0.1574	0.9611
Original Signal	0.9127	0.9616	0.9613

5. Conclusions

In this paper, a novel measuring technique based on the Cosine distance is proposed to evaluate signals similarity by allowing many-to-many matching between sub-samples. Results in this paper indicate that the proposed measure can overcome the limitations of the existing measures that only one-to-one matching is allowed between sequences. The proposed approach provides a more robust measure invariant to different forms of translations which might be involved in signals, such as amplitude-scale, amplitude-shift, time shift, and phase delay. Performance comparison results in this paper show that the proposed similarity measure outperforms existing approaches in comparing one-dimensional sequences.

References

[1] Schalkoff, R.J, Pattern Recognition: Statistical, Syntactic and Neural Approaches, John Wiley and Sons, 1992.

[2] Witold Pedrycz, Knowledge-based clustering: from data to information granules, John Wiley and Sons, 2005.

[3] Vijay K.Madisetti and Douglas B.Williams;(1999); ***"The Digital Signal Processing Handbook"***; Chapman&Hall,CRC Press LLC,1999.

[4] Todd K. Moon, ***"Similarity Methods in Signal Processing"***, IEEE Transactions on Signal Processing, vol. 44, No. 4, pages 827-833 , April 1996.

[5] S. Bayram, I. Avcibas, B. Sankur, N. Memon. ***"Image manipulation detection with binary similarity measures"***, European Signal Processing Conference, Turkey, September 2005.

[6] H. Hassanpour and M. Mesbah; ***"Neonatal EEG seizure detection using spike signatures in the time-frequency domain"***, IEEE Int. Sympo. On Sig. Proc. And Its Appl. (ISSPA), vol.2, pp.41-44, Paris, France, July 2003.

[7] K. Rieck, P. Laskov, K.-R. M¨uller, ***"Efficient Algorithms for Similarity Measures over Sequential Data: A Look Beyond Kernels"***, DAGM 2006, LNCS 4174, pp. 374–383, 2006.