CHAPTER II

MEANINGLESSNESS OF SUBSEQUENCE TIME SERIES CLUSTERING

Subsequence Time Series Clustering (STSC) has been proven both empirically (Peker, 2005; Chen, 2007a; Goldin et al., 2006; Denton, 2005; Keogh et al., 2003; Fujimaki et al., 2008; Kontaki et al., 2008; Chen, 2007b; Simon et al., 2006) and theoretically (Idé, 2006a,b) that its output is meaningless. Keogh and Lin (Keogh and Lin, 2005) first flagged this issue by observation that STSC always produced a set of sine waves as cluster representatives instead of expected patterns from a time series sequence. In addition, they also proposed a meaningfulness measurement, so-called Keogh-Lin Meaningfulness Measurement (KLMM). Specifically, KLMM defines that cluster representatives should be similar if the representatives are from the same input sequence, and cluster representatives should be dissimilar if the representatives are from different input sequences. However, this thesis argues that KLMM is an invalid measurement for two reasons. First, although cluster representatives from different input sequences are sine waves, these sine waves may have different phases and frequencies. Second, KLMM only measures clustering results without considering how similar input sequences are; similarity between two input sequences are not defined for KLMM. For example, clustering results from two similar sequences must be very similar, but they are considered meaningless in the view of KLMM, even a clustering algorithm does produce a meaningful result. In this chapter, the meaninglessness of clustering results of STSC will be demonstrated, and KLMM will be shown that it is an invalid meaningfulness measurement.

2.1 Background

In this section, background knowledge of Subsequence Time Series Clustering (STSC), k-hierarchical clustering, k-means clustering, Euclidean distance, and Amplitude Averaging is provided to give better understanding of STSC's the meaninglessness.

2.1.1 Subsequence Time Series Clustering (STSC)

Subsequence Time Series Clustering (STSC) has been proposed to discover patterns or to group subsequences as a part of a subroutine or a preprocessing step of various mining tasks such as rule discovery (Das et al., 1998; Fu et al., 2001; Harms et al., 2002b,a; Hetland, 2002; Jin et al.,

2002b,a; Mori and Kuni, 2001; Osaki et al., 2000; Sarker et al., 2003; Uehara and Shimada, 2002; Yairi et al., 2001), indexing (Li et al., 1998; Radhakrishnan et al., 2000), classification (Cotofrei, 2002; Cotofrei and Stoffel, 2002), prediction (Schittenkopf et al., 2000), and anomaly detection (Yairi et al., 2001). Given a time series sequences $S = \langle s_1, s_2, \ldots, s_n \rangle$ of length n, STSC first extracts a set $\mathbb{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_i, \ldots, \mathcal{S}_{n-w+1}\}$ of subsequences using a fixed-length sliding window, where a subsequence $\mathcal{S}_i = \langle s_i, s_{i+1}, \ldots, s_{i+w-1} \rangle$, $1 \leq i \leq n-w+1$, and w is the sliding window length. Then every subsquence is normalized by z-normalization (see Section 2.1.6), and subsequences are clustered by k-hierarchical clustering or k-means clustering algorithms with Euclidean distance and Amplitude Averaging as a distance measure and an averaging function. In addition, Euclidean distance is used to calculate similarity between two subsequences and Amplitude Averaging function is used to construct a cluster representative for each cluster. STSC finally returns a set of clusters returned from k-hierarchical clustering or k-means clustering. Formally, STSC receives a long time series S with two parameters, i.e., the number of clusters (k) and the length of a sliding window (w), and returns a set $\mathbb{C} = \{C_1, C_2, \ldots, C_i, \ldots, C_k\}$ of clusters, where each cluster $C_i = (\mathbb{M}, R)$ contains cluster members $\mathbb{M} = \{\mathcal{S}_i \mid \mathcal{S}_i \in \mathbb{S}\}$ and a cluster representative $R = \langle r_1, r_2, \ldots, r_w \rangle$. Pseudo code of STSC is provided in Table 2.1 and Figure 2.1 visualizes an overview of STSC.

Table 2.1: Pseudo code of Subsequence Time Series Clustering (STSC)

FUNCTION [$\mathbb{C}$] = SUBSEQUENCETIMESERIESCLUSTERING [S, k, w]
1. $\mathbb{S}$ = EXTRACTSUBSEQUENCES(S, w)
2. $\mathbb{S}_{Norm}$ = NORMALIZESUBSEQUENCES($\mathbb{S}$)
3. $\mathbb{C}$ = CLUSTERING($\mathbb{S}_{Norm}$, k) // with Euclidean distance and Amplitude Averaging
4. Return $\mathbb{C}$

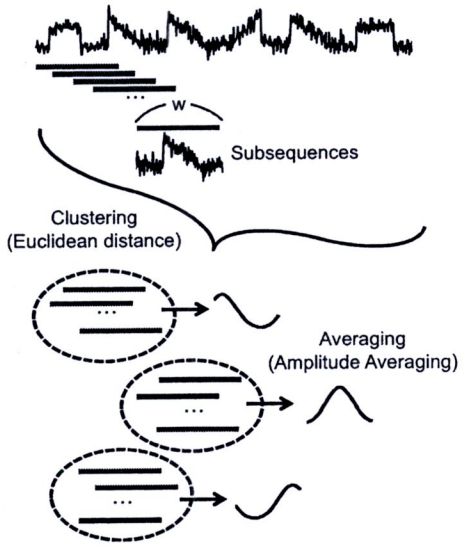


Figure 2.1: Overview of Subsequence Time Series Clustering (STSC)

2.1.2 K-Hierarchical Clustering

K-hierarchical clustering used in STSC is an agglomerative clustering algorithm. AGNES (AGglomerative NESting) is a well-known hierarchical clustering algorithms that can visualize relationships among data sequences in a hierarchical structure or a tree-based structure on distance calculations. Although many variations of hierarchical clustering algorithms have been introduced such as BIRCH (Zhang et al., 1996) (Balanced Iterative Reducing and Clustering Using Hierarchies), ROCK (Guha et al., 2000) (A Hierarchical Clustering Algorithm for Categorical Attributes), and Chameleon (Karypis et al., 1999) (A Hierarchical Clustering Algorithm Using Dynamic Modeling), AGNES is commonly used due to implementation simplicity.

Specifically, AGNES has been proposed to group data using bottom-up strategy. The method iteratively merges two atomic clusters into a larger cluster until one single cluster containing every data sequences is achieved. For each iteration, two clusters which have minimum inter-cluster distance are merged. However, grouping a dataset into one single cluster for agglomerative clustering is impractical; therefore, the number of clusters (k) is required. Concretely, pseudo codes of the agglomerative clustering algorithm which receives a set $\mathbb{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_i, \ldots, \mathcal{S}_n\}$ of time series sequences as an input and returns a set $\mathbb{C} = \{C_1, C_2, \ldots, C_i, \ldots, C_k\}$ of k clusters as an output, where each $C_i = (\mathbb{M}, R)$ contains a set $\mathbb{M} = \{\mathcal{S}_i \mid \mathcal{S}_i \in \mathbb{S}\}$ of time series sequences and a cluster representative R, are shown in Table 2.2.

Table 2.2: Agglomerative hierarchical clustering algorithm (AGNES)

FUNCTION [$\mathbb{C}$] = AGGLOMERATIVECLUSTERING [$\mathbb{S}$, k]	
1.	Initialize a set $\mathbb{C}$ of clusters which contains one sequence from $\mathbb{S}$
2.	While (the size of $\mathbb{C} > k$)
3.	$dist_{best}$ = INFINITY
4.	For each pair of C_i and C_j in $\mathbb{C}$
5.	$dist$ = INTERCLUSTERDISTANCE(C_i, C_j)
6.	if ($dist < dist_{best}$)
7.	$dist_{best} = dist$
8.	$pair_{best} = [C_i, C_j]$
9.	Endif
10.	Endfor
11.	$[C_i, C_j] = pair_{best}$
12.	C_k = MERGE(C_i, C_j)
13.	Remove C_i and C_j from $\mathbb{C}$
14.	Add C_k to $\mathbb{C}$
15.	Endwhile
16.	For each cluster C in $\mathbb{C}$
17.	$C.R$ = AVERAGE($C.\mathbb{M}$)
18.	Endfor
19.	Return $\mathbb{C}$

While many similarity functions between two clusters (called inter-cluster distances) have been proposed, three functions are typically used, i.e., single linkage, complete linkage, and average linkage inter-cluster distance functions. Single linkage function returns a minimum distance among all possible pairs between two clusters, while complete linkage function returns a maximum distance among all possible pairs between two clusters. On the other hand, average linkage function finds a mean value of all distances. Pseudo codes of single, complete, and average linkage distance functions are provided in Table 2.3, 2.4, and 2.5, respectively, and these inter-cluster distances are formalized as follows.

$$D_{single}(C_i, C_j) = \min_{\mathcal{S} \in \mathbb{M}_i, \mathcal{S}' \in \mathbb{M}_j} Distance(\mathcal{S}, \mathcal{S}') \tag{2.1}$$

$$D_{complete}(C_i, C_j) = \max_{\mathcal{S} \in \mathbb{M}_i, \mathcal{S}' \in \mathbb{M}_j} Distance(\mathcal{S}, \mathcal{S}') \tag{2.2}$$

$$D_{average}(C_i, C_j) = \frac{1}{|\mathbb{M}_i|\,|\mathbb{M}_j|} \sum_{c \in C_i} \sum_{c' \in C_j} Distance(\mathcal{S}, \mathcal{S}') \tag{2.3}$$

where D_{single}, $D_{complete}$, and $D_{average}$ are single, complete, and average linkage distance functions, respectively, C_i and C_j are any clusters, $\mathbb{M}_i$ and $\mathbb{M}_j$ are corresponding cluster members of C_i and C_j, respectively, and $\mathcal{S}$ and $\mathcal{S}'$ are sequences in $\mathbb{M}_i$ and $\mathbb{M}_j$, respectively. $Distance(\mathcal{S}, \mathcal{S}')$ is a distance function that returns a distance between two sequences $\mathcal{S}$ and $\mathcal{S}'$.

Table 2.3: Pseudo code of single linkage distance function

FUNCTION [$dist_{best}$] = SINGLELINKAGE [C_i, C_j]
1. $\mathbb{M}_i$ is a set of cluster member of C_i
2. $\mathbb{M}_j$ is a set of cluster member of C_j
3. $dist_{best}$ = INFINITY
4. For each sequence $\mathcal{S}$ in $\mathbb{M}_i$
5. For each sequence $\mathcal{S}'$ in $\mathbb{M}_j$
6. $dist$ = DISTANCE($\mathcal{S}$, $\mathcal{S}'$)
7. if ($dist < dist_{best}$)
8. $dist_{best} = dist$
9. Endif
10. Endfor
11. Endfor
12. Return $dist_{best}$

For Subsequence Time Series Clustering (STSC), Euclidean distance and Amplitude Averaging is used as a distance function and an averaging function.

Table 2.4: Pseudo code of complete linkage distance function

FUNCTION $[dist_{best}]$ = COMPLETELINKAGE $[C_i, C_j]$
1. $\mathbb{M}_i$ is a set of cluster member of C_i
2. $\mathbb{M}_j$ is a set of cluster member of C_j
3. $dist_{best}$ = INFINITY
4. For each sequence $\mathcal{S}$ in $\mathbb{M}_i$
5. For each sequence $\mathcal{S}'$ in $\mathbb{M}_j$
6. $dist$ = DISTANCE($\mathcal{S}$, $\mathcal{S}'$)
7. if ($dist > dist_{best}$)
8. $dist_{best} = dist$
9. Endif
10. Endfor
11. Endfor
12. Return $dist_{best}$

Table 2.5: Pseudo code of average linkage distance function

FUNCTION $[dist_{avg}]$ = AVERAGELINKAGE $[C_i, C_j]$
1. $\mathbb{M}_i$ is a set of cluster member of C_i
2. $\mathbb{M}_j$ is a set of cluster member of C_j
3. $dist_{avg} = 0$
4. For each sequence $\mathcal{S}$ in $\mathbb{M}_i$
5. For each sequence $\mathcal{S}'$ in $\mathbb{M}_j$
6. $dist_{avg} = dist_{avg}$ + DISTANCE($\mathcal{S}$, $\mathcal{S}'$)
7. Endfor
8. Endfor
9. $dist_{avg} = dist_{avg} / \|\mathbb{M}_i\|\|\mathbb{M}_j\|$
10. Return $dist_{avg}$

2.1.3 K-Means Clustering

K-means clustering algorithm (Lloyd, 1982; MacQueen, 1967) is a partitioning clustering that finds a group of clusters by iteratively refining members in each cluster to have the maximum objective value that minimizes summation of distances between a cluster representative and cluster members for every cluster. Beside k-means clustering, many partitioning clustering algorithms are proposed including k-medoids clustering (Kaufman and Rousseeuw, 2005) and CLARANS (Kaufman and Rousseeuw, 2005). Both k-medoids and CLARAN use a median of cluster members instead of a mean. However, a median cannot reflect all characteristics of all data sequences of a cluster because a median is selected from one of existing data sequences, while a mean is a sequence constructed by averaging all data sequences within a cluster. Therefore, k-means clustering is much more preferable than k-medoids and CLARAN.

Initially, k-means clustering first selects k centers by randomizing existing data sequences from a set $\mathbb{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_i, \ldots, \mathcal{S}_n\}$ of sequences, where $\mathcal{S}_i = \langle s_1, s_2, \ldots, s_w \rangle$ is a time series sequence of length w, and then remaining sequences are assigned to the closest cluster center,

where k is a user-defined number of clusters. After that, a new cluster center is calculated by averaging all cluster members within each cluster. The algorithm repeats assigning data sequences to the closest center and recalculating for cluster centers until the clustering result remains unchanged. When the algorithm terminates, a set $\mathbb{C} = \{C_1, C_2, \ldots, C_i, \ldots, C_k\}$ of clusters, where each cluster $C_i = (\mathbb{M}, R)$ contains a set $\mathbb{M} = \{\mathcal{S}_i \mid \mathcal{S}_i \in \mathbb{S}\}$ of cluster members and a cluster representative $R = \langle r_1, r_2, \ldots, r_w \rangle$ is returned. To be more concrete, pseudo code of k-means clustering is provided in Table 2.6.

Table 2.6: Pseudo code of k-means clustering

```
FUNCTION [ℂ] = KMEANSCLUSTERING [𝕊, k]
1.  Initialize a set ℂ of k cluster centers with existing sequence in 𝕊
2.  Do
3.    For each sequence 𝒮 in 𝕊
4.      dist_best = INFINITY
5.      For each cluster C in ℂ
6.        R = Cluster representative of C
7.        dist = DISTANCE(𝒮, R)
8.        If (dist < dist_best)
9.          dist_best = dist
10.         C_best = C
11.       Endif
12.     Endfor
13.     Assign 𝒮 to C_best
14.   Endfor
15.   For each cluster C in ℂ
16.     C.R = AVERAGE(C.𝕄)
17.   Endfor
18. While (all cluster members in ℂ change)
19. Return ℂ
```


Subsequence Time Series Clustering (STSC) with k-means clustering uses Euclidean distance and Amplitude Averaging as a distance measure and an averaging function, respectively.

2.1.4 Euclidean Distance

Euclidean distance (Keogh and Ratanamahatana, 2005) is a well-known similarity measure used in many domains including time series data. The distance is calculated in one-to-one manner shown in Figure 2.2, where the distance is a summation of difference between two data points in the same dimension. Euclidean distance between two time series sequences A and B is calculated by the following equation.

$$Euclidean(A, B) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

where $A = \langle a_1, a_2, \ldots, a_i, \ldots, a_n \rangle$ and $B = \langle b_1, b_2, \ldots, b_i, \ldots, b_n \rangle$ are two time series sequences of length n.

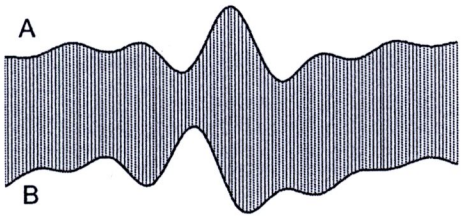
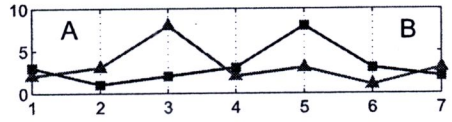


Figure 2.2: Example of Euclidean distance calculation.

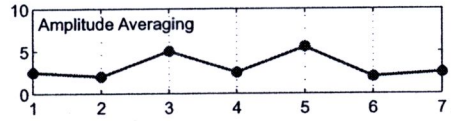
For Subsequence Time Series Clustering (STSC), Euclidean distance is used as a distance measure in k-means clustering and k-hierarchical clustering algorithms. However, at the end of this chapter (Section 2.4), Euclidean distance will be shown that it is a cause that makes a clustering result of STSC meaningless.

2.1.5 Amplitude Averaging

Amplitude Averaging function is a method to construct a mean of a set of time series sequences, where a value of each dimension of a mean is derived from averaging all values of the same dimension for all sequences. A mean $Z = \langle z_1, z_2, \ldots, z_i, \ldots, z_n \rangle$ from Amplitude Averaging of two time series sequences $A = \langle a_1, a_2, \ldots, a_i, \ldots, a_n \rangle$ and $B = \langle b_1, b_2, \ldots, b_i, \ldots, b_n \rangle$ of length n is calculated by $z_i = \frac{a_i + b_i}{2}$. The example is shown in Figure 2.3; a mean sequence is generated from two sequences $A = \langle 2, 3, 8, 2, 1, 3 \rangle$ and $B = \langle 3, 1, 2, 8, 3, 2 \rangle$ by Amplitude Averaging function.



a) Original sequences A and B



b) Averaged result generated from Amplitude Averaging

Figure 2.3: Example of Amplitude Averaging calculation.

However, if two sequences $A = \langle a_1, a_2, \ldots, a_i, \ldots, a_n \rangle$ and $B = \langle b_1, b_2, \ldots, b_i, \ldots, b_n \rangle$ have different weights, ω_A and ω_B, respectively, a mean sequence $Z = \langle z_1, z_2, \ldots, z_i, \ldots, z_n \rangle$ can be computed by $z_i = \frac{\omega_A \cdot a_i + \omega_B \cdot b_i}{\omega_A + \omega_B}$. And for averaging a set $\mathbb{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_j, \ldots, \mathcal{S}_m\}$ of

sequences, a mean sequence $Z = \langle z_1, z_2, \ldots, z_i, \ldots, z_n \rangle$ can be computed once by $z_i = \frac{\sum_{S \in \mathbb{S}} s_i}{|\mathbb{S}|}$, and the pseudo code is provided in Table 2.7.

Table 2.7: Pseudo code of Amplitude Averaging function

FUNCTION [Z] = AMPLITUDEAVERAGING [$\mathbb{S}$]
1. Initialize the sequence Z to all zeros
2. For each sequence S in $\mathbb{S}$
3. For each data point s_i in S
4. $z_i = z_i + s_i$
5. Endfor
6. Endfor
7. For each data point z_i in Z
8. $z_i = z_i / \|\mathbb{S}\|$
9. Endfor
10. Return Z

For Subsequence Time Series Clustering (STSC), Amplitude Averaging function is used as an averaging function to construct a cluster representative; however, in this section, Amplitude Averaging will be shown that it is one of the causes that makes the output of STSC meaningless.

2.1.6 Z-Normalization

Normalization is a function to rescale a sequence to a specific range. In data mining, many normalization techniques (Han and Kamber, 2000) have been proposed such as min-max normalization, sigmoid normalization, and z-normalization. For time series data, z-normalization is typically used to remove an offset and imbalanced distribution. In addition, the sequence is normalized to obtain a mean and a standard deviation of zero and one, respectively. Given a sequence $A = \langle a_1, a_2, \ldots, a_i, \ldots, a_n \rangle$ of length n, a new sequence $Z = \langle z_1, z_2, \ldots, z_i, \ldots, z_n \rangle$ is normalized according to the following equations.

$$z_i = \frac{a_i - \mu_A}{\sigma_A} \tag{2.4}$$

$$\mu_A = \frac{\sum_{i=1}^{n} a_i}{n} \tag{2.5}$$

$$\sigma_A = \sqrt{\frac{\sum_{i=1}^{n} (a_i - \mu_A)^2}{n}} \tag{2.6}$$

where μ_A and σ_A are a mean and a standard deviation of the sequence A, respectively.

Example is shown in Figure 2.4, where the original sequence is normalized to have its mean and standard deviation of zero and one, respectively.

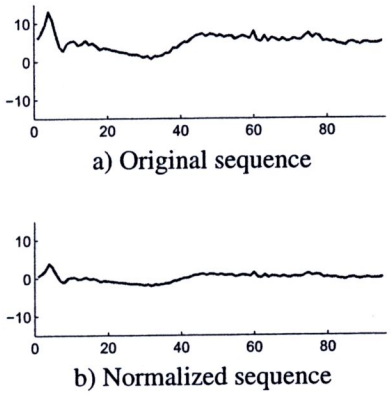


Figure 2.4: Example of z-normalization.

For Subsequence Time Series Clustering (STSC), a set of sequences extracted from a long time series sequences needs to be normalized before clustering with k-means clustering and k-hierarchical clustering algorithms. If normalization is not applied, subsequence clustering will produce undesired results since similarity between subsequences must be independent to mean and standard deviation of subsequences.

2.2 Related Work

Keogh and Lin have published a paper describing that an output of Subsequence Time Series Clustering (STSC) is a set of sine waves that is considered meaningless (Keogh and Lin, 2005). This leads to many arguments in data mining community since STSC has been implemented as a subroutine and a preprocessing step of hundreds of mining applications such as rule discovery (Das et al., 1998; Fu et al., 2001; Harms et al., 2002b,a; Hetland, 2002; Jin et al., 2002b,a; Mori and Kuni, 2001; Osaki et al., 2000; Sarker et al., 2003; Uehara and Shimada, 2002; Yairi et al., 2001), indexing (Li et al., 1998; Radhakrishnan et al., 2000), classification (Cotofrei, 2002; Cotofrei and Stoffel, 2002), prediction (Schittenkopf et al., 2000), and anomaly detection (Yairi et al., 2001). Since Keogh and Lin proved that STSC is meaningless, all the works and that successors utilized STSC are also considered invalid. Generally, STSC extracts subsequences from a long time series as an input and returns a set of clusters as an output. Keogh and Lin found that although an input changes, an output remains the same; in other words, STSC always produces the similar sine waves as cluster representatives regardless of a data input of the clustering algorithm.

Keogh and Lin claim that STSC is meaningless by the following experiment. Thirty each of three patterns, i.e., Cylinder, Bell, and Funnel (Saito, 1994), of length 128, shown in Figure 2.5, generated from the following equations are concatenated to create a long sequence in Figure 2.6.

$$c(t) = (6 + \eta) \cdot \chi[a, b](t) + \epsilon(t) \tag{2.7}$$

$$b(t) = (6 + \eta) \cdot \chi[a, b](t) \cdot (t - a) / (b - a) + \epsilon(t) \tag{2.8}$$

$$f(t) = (6 + \eta) \cdot \chi[a, b](t) \cdot (b - t) / (b - a) + \epsilon(t) \tag{2.9}$$

$$\chi[a, b] = \begin{cases} 0 & t < a \\ 1 & a \leq t \leq b \\ 0 & t > b \end{cases} \tag{2.10}$$

where η and $\epsilon(t)$ are drawn from a standard normal distribution $N(0, 1)$, a is an integer drawn uniformly from [16, 32], $b - a$ is an integer drawn uniformly from [32, 96], and t is varied from 1 to 128.

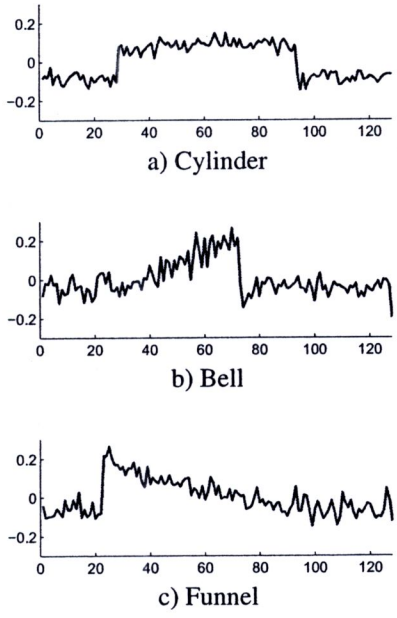


Figure 2.5: Examples of Cylinder-Bell-Funnel dataset

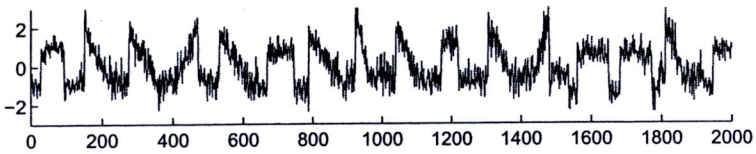


Figure 2.6: Some part of Cylinder-Bell-Funnel sequence

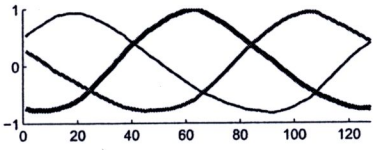


Figure 2.7: Cluster representatives generated from STSC

When this sequence is clustered by STSC, sine-wave-like cluster representatives (see Figure 2.7) are returned, while original patterns are expected to be a result. Keogh and Lin also propose a meaningfulness measurement, so-called Keogh-Lin Meaningfulness Measurement (KLMM), defining that the subsequence clustering is meaningful when the clustering algorithm returns similar cluster representatives from the same input sequence and dissimilar cluster representatives from different input sequences. Suppose $\mathbb{X} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n\}$ and $\mathbb{Y} = \{\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_n\}$ are two sets of clustering results from n different runs of two different datasets, where $\mathcal{X} = \left\{\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_k\right\}$ and $\mathcal{Y} = \left\{\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_k\right\}$ are two sets of cluster representatives, respectively. The meaningfulness of KLMM can be calculated from the following equations.

$$WithinDistance\,(\mathbb{X}) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} ClusterDistance\,(\mathcal{X}_i, \mathcal{X}_j)}{k^2} \tag{2.11}$$

$$BetweenDistance\,(\mathbb{X}, \mathbb{Y}) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} ClusterDistance\,(\mathcal{X}_i, \mathcal{Y}_j)}{k^2} \tag{2.12}$$

$$KLMM\,(\mathbb{X}, \mathbb{Y}) = \frac{WithinDistance\,(\mathbb{X})}{BetweenDistance\,(\mathbb{X}, \mathbb{Y})} \tag{2.13}$$

where $WithinDistance\,(\mathbb{X})$ is a distance between sets of cluster representatives from the same input sequence, $BetweenDistance\,(\mathbb{X}, \mathbb{Y})$ is a distance between sets of cluster representatives from different input sequences, and $ClusterDistance\,(\mathcal{A}, \mathcal{B})$ can be calculated from the summation of minimum distances between two sets of cluster representatives. The result is meaningful when KLMM returns the value close to zero since $WithinDistance\,(\mathbb{X})$ is small and $BetweenDistance\,(\mathbb{X}, \mathbb{Y})$ is very large; otherwise, the result is meaningless.

$ClusterDistance\,(\mathcal{A},\mathcal{B})$ can be formalized as the following equation.

$$ClusterDistance\,(\mathcal{A},\mathcal{B}) = \sum_{i=1}^{k} min\left[EuclideanDistance\,(A_i, B_j)\right], 1 \leq j \leq k \qquad (2.14)$$

where $\mathcal{A} = \{A_1, A_2, \ldots, A_i, \ldots, A_k\}$ and $\mathcal{B} = \{B_1, B_2, \ldots, B_j, \ldots, B_k\}$ are two sets of cluster representatives.

However, KLMM is an invalid meaningfulness measurement for two reasons. The first reason is that with the same number of clusters and the same length of sliding window, cluster representatives of two different input sequences may be sine waves with different phases and frequencies. STSC always produces sine waves regardless of an input sequence; therefore, if cluster representatives are sine waves, the clustering result would mistakenly be considered as meaningless. However, Euclidean distance utilized by KLMM cannot capture similarity between two sine waves with different phases and frequencies; therefore, KLMM considers clustering results are meaningful although results are all sine waves.

Secondly, KLMM assumes that two clustering results are meaningful if they are different. For any meaningful subsequence clustering algorithm, if two input sequences are similar, the clustering results are expected to be similar as well, and if two input sequences are different, the clustering results are expected to be different, but KLMM will always flag any two similar clustering results as meaningless regardless of similarity between two input sequences. Although a meaningful subsequence clustering algorithm exists, KLMM cannot tell how meaningful they are.

Many successor papers in finding a meaningful subsequence clustering also unawares use KLMM as a meaningfulness measurement to evaluate their algorithms; therefore, their experiments become invalid. For theoretical study, Ide (Idé, 2006b) proved that STSC always returns sine waves regardless of an input sequence. In this thesis, a new meaningfulness measure will be introduced in Chapter IV to be used as a meaningfulness measurement for Shape-based Subsequence Time Series Clustering (2STSC).

2.3 Experiments

Two following experiments will demonstrate that STSC produces meaningless clustering results and that KLMM is an invalid meaningfulness measurement. Datasets used in these exper-

iments are eight time series of length 2000 from the Time Series Data mining Archive (TSDMA) (Keogh and Folias, 2011) shown in Figure A.1. Figure 2.8 shows Buoy1 and CBF used in the experiments.

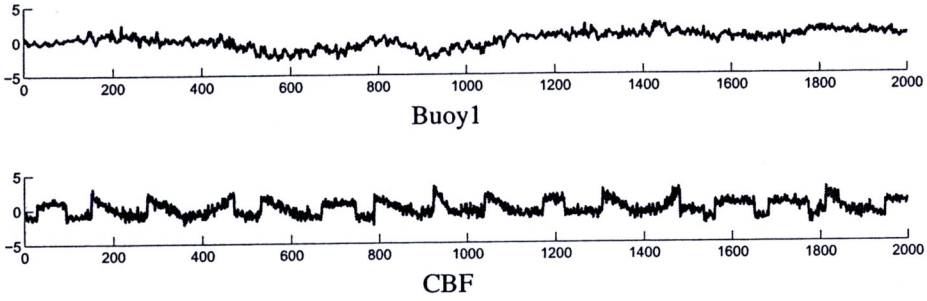


Figure 2.8: Datasets from TSDMA used in the experiments.

2.3.1 First Experiment

The first experiment demonstrates that STSC produces clustering results as sine waves regardless of an input sequence. The number of clusters (k) and the length of sliding window (w) vary. In addition, to show that cluster representatives are sine waves, perfect sine waves are constructed and compared to these cluster representatives. Generally, a sine wave can be formalized as a following equation (Hazewinkel, 2001).

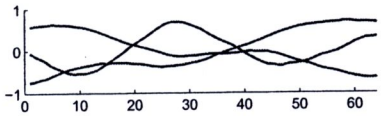
$$y(x) = A \cdot \sin(\omega x + \varphi) + \mu \tag{2.15}$$

where A is the amplitude, $\omega = 2\pi f$ is the angular frequency (in radian per second), f is the ordinary frequency (in hertz), φ is phase, and μ is an offset of the sine wave.

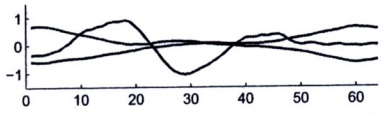
Given a set $\mathbb{R} = \{R_1, R_2, \ldots, R_k\}$ of k cluster representatives, a new set $\mathbb{R}' = \{R'_1, R'_2, \ldots, R'_k\}$ of k cluster representatives is constructed by searching for those parameters by a non-linear equation solver (Balda, 1999) implemented with Levenberg-Marquardt algorithm (Fletcher, 1971) to minimize Root Mean Square Error (RMSE).

Figures 2.9 and 2.10 show cluster representatives generated from STSC of two datasets, i.e., Buoy1 and CBF, using k-means clustering and k-hierarchical clustering (with two variations of inter-cluster distance functions) when $k = 3$ and $w = 64$. Note that single linkage distance function is not used as an inter-distance function in this experiment because k-hierarchical clustering with single linkage function cannot gracefully handle trivial-matched subsequences, where some subsequences will never in any groups if these subsequences have the largest nearest neighbor distance compared with other subsequences. In other words, single linkage group subsequences

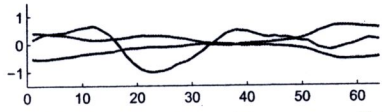
based on the smallest nearest neighbor distance. Therefore, in this study, only two inter-cluster distance functions are utilized, i.e., complete linkage and average linkage functions. The constructed sine waves from cluster representatives generated from STSC of two datasets, i.e., Buoy1 and CBF, using k-means clustering and k-hierarchical clustering are shown in Figures 2.11 and 2.12, respectively, when $k = 3$ and $w = 64$, where thick lines are constructed sine waves, and thin lines are original cluster representatives. The complete experiment results of eight datasets are provided in Appendix B, where the number of clusters (k) and the length of sliding window (w) are varied to be (3, 32), (3, 64), (5, 64), (7, 64), and (3, 128), respectively.



k-means clustering

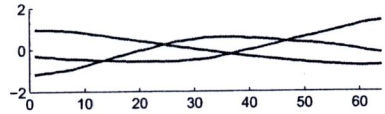


k-hiearchical clustering with complete linkage

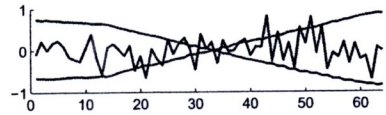


k-hiearchical clustering with average linkage

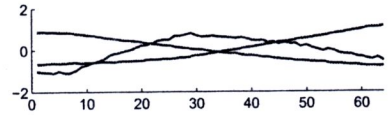
Figure 2.9: Cluster representatives generated from STSC of Buoy1 when $k = 3$ and $w = 64$.



k-means clustering

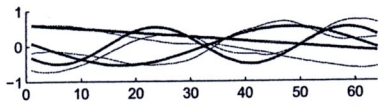


k-hiearchical clustering with complete linkage

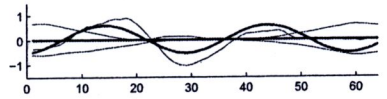


k-hiearchical clustering with average linkage

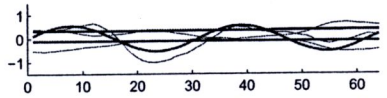
Figure 2.10: Cluster representatives generated from STSC of CBF when $k = 3$ and $w = 64$.



k-means clustering

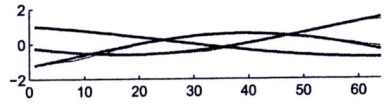


k-hiearchical clustering with complete linkage

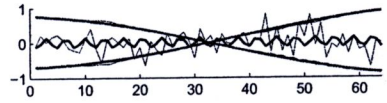


k-hiearchical clustering with average linkage

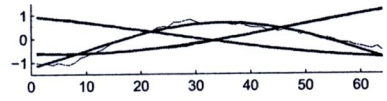
Figure 2.11: Constructed sine waves generated from STSC of Buoy1 when $k = 3$ and $w = 64$.



k-means clustering



k-hiearchical clustering with complete linkage



k-hiearchical clustering with average linkage

Figure 2.12: Constructed sine waves generated from STSC of CBF when $k = 3$ and $w = 64$.

2.3.2 Second Experiment

The second experiment demonstrates that KLMM is an invalid meaningfulness measurement. From the first experiment, clustering results of STSC are meaningless because STSC produces sine waves as cluster representatives. However, KLMM does not capture that the result is a set of sine waves, but KLMM calculates the difference between two cluster representatives using Euclidean distance. Since STSC has been proven both empirically and theoretically that it produces sine waves regardless of inputs (Idé, 2006b; Keogh and Lin, 2005), KLMM should return high values (more than one) for pairs of datasets. The following results show that KLMM is an invalid measurement since KLMM does not return high values; even though the cluster representatives are all sine waves. Figure 2.13 and Figure 2.14 show KLMM of STSC using k-means clustering and KLMM of STSC using k-hierarchical clustering by varying the number of clusters

(k) and the length of sliding window (w). From the figures, all pair comparisons of eight datasets are evaluated. The value of KLMM is represented in gray shade, where black color represents a high value of KLMM, while white color representing a low value of KLMM. From the experiments, some values are completely white, and some are gray, but not all black; however, the values are expected to be all black since STSC have been proven that it produces meaningless results.

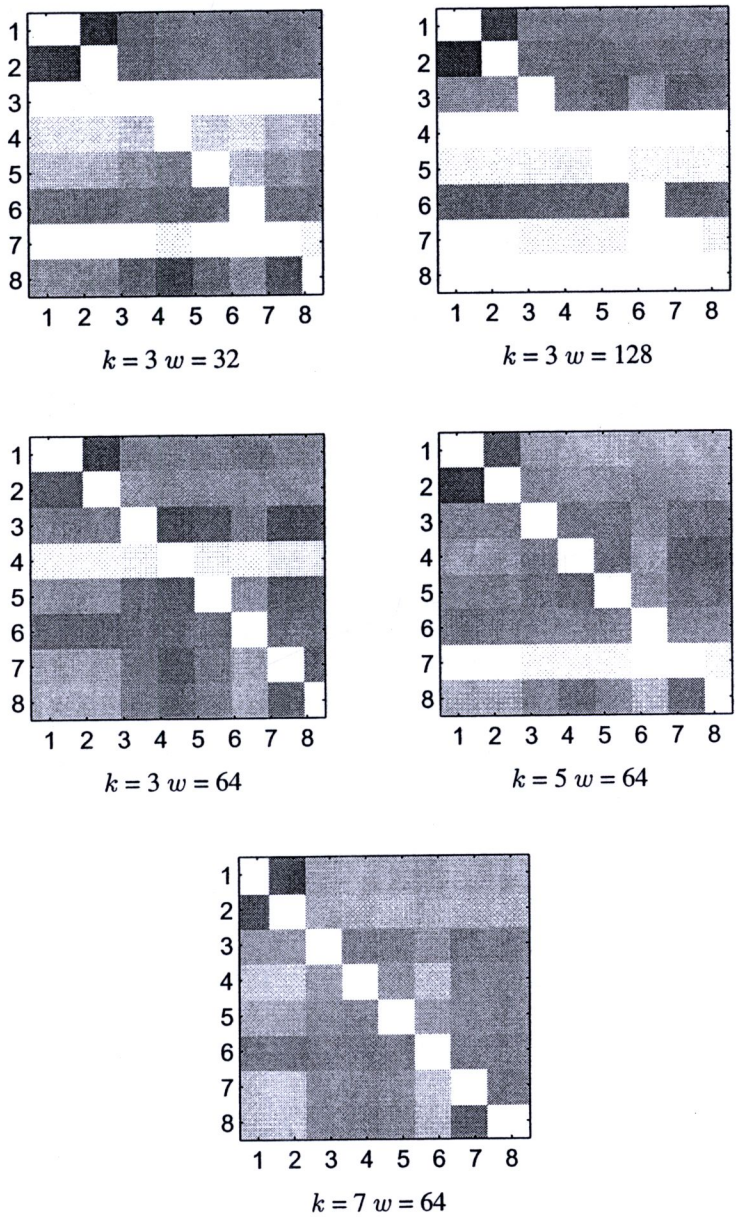


Figure 2.13: KLMMs of STSC using k-means clustering.

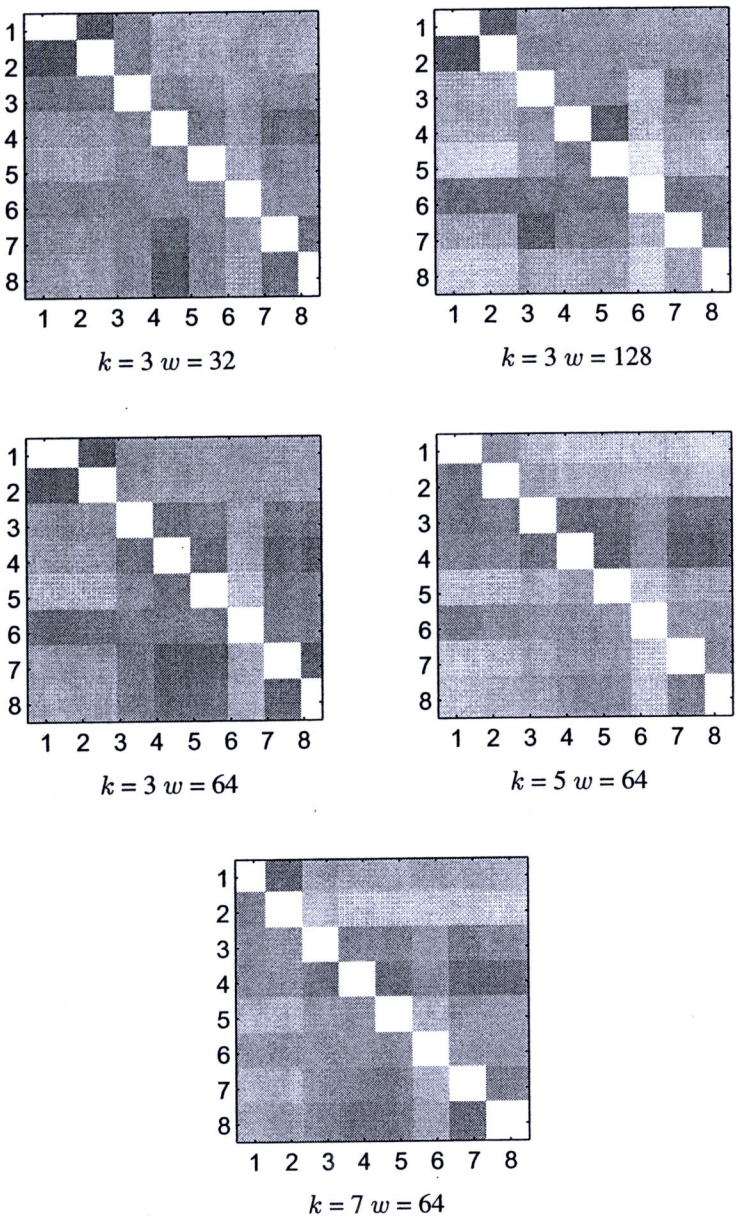
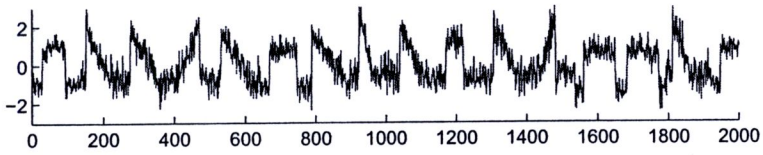


Figure 2.14: KLMMs of STSC using k-hierarchical clustering.

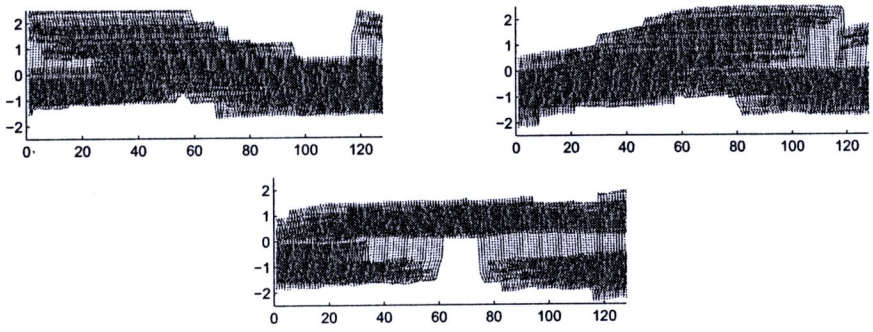
2.4 Causes of Meaninglessness

The causes of meaninglessness are inappropriate approaches to handle trivial-matched subsequences. Trivial-matched subsequences are a set of adjacent subsequences in a time series sequence, where between two adjacent subsequences, only two data points are different. Formally, given a time series sequence $S = \langle s_1, s_2, \ldots, s_n \rangle$ of length n, a set $\mathbb{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{n-w+1}\}$ of subsequences extracted from a sequence S with a fixed-length sliding window of length w, a set of trivial-matched subsequences are $\mathbb{T} = \{\mathcal{S}_i, \mathcal{S}_{i+1}, \ldots\}$, where $1 \leq i \leq n - w + 1$. Trivial-matched subsequences of CBF sequence are illustrated in Figure 2.15. In addition, inappropriate

uses of a distance measure and an averaging function to handle trivial-matched subsequences lead to an undesired clustering output. Specifically, STSC utilizes Euclidean distance and Amplitude Averaging function as a distance measure and an averaging function, respectively.



a) Some part of CBF sequence



b) Three sets of trivial-matched subsequences

Figure 2.15: Trivial-matched subsequences of CBF sequence

In Euclidean space, two adjacent subsequences may be considered as significantly different although only two data points are different, and the remaining points are the same. To be more illustrative, Euclidean distance cannot group different sets of trivial-matched subsequences shown as a dendrogram in Figure 2.16.

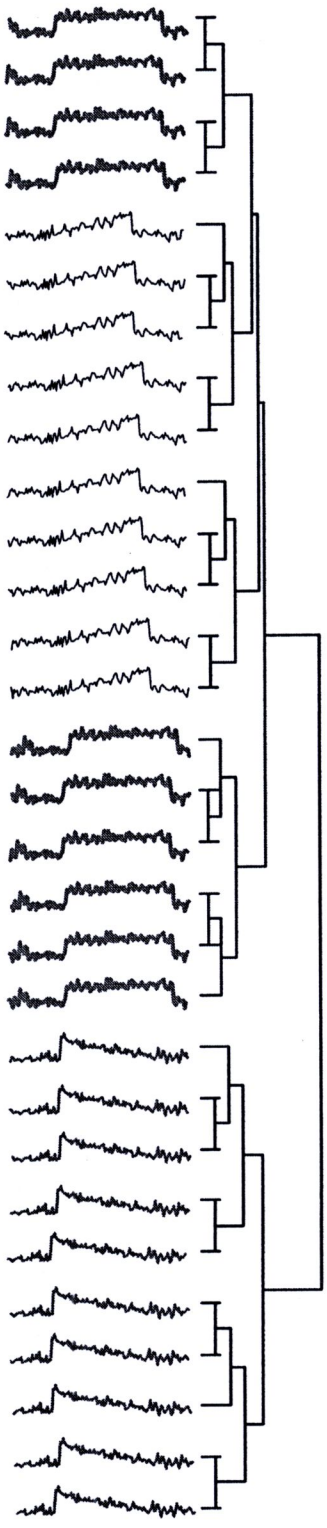


Figure 2.16: Euclidean distance cannot capture similarity between trivial-matched subsequences

To construct a cluster representative, STSC averages all subsequences within a cluster using Amplitude Averaging function, where Amplitude Averaging generates an averaged result by computing a mean of each dimension directly. In addition, Amplitude Averaging is inappropriate to be used as an averaging function of STSC since Amplitude Averaging does not align shifted data

points of adjacent subsequences. Therefore, in the end, each dimension of the result is averaged from unrelated dimensions. This leads to undesired smoothened cluster representatives. Three averaged results of trivial-matched subsequences from CBF sequence generated by Amplitude Averaging function are shown in Figure 2.17. The averaged result will be smoother and more convergent to sine waves; therefore, trivial-matched subsequence clustering can be meaningful when appropriate distance measure and average function are used instead of Euclidean distance and Amplitude Averaging function.

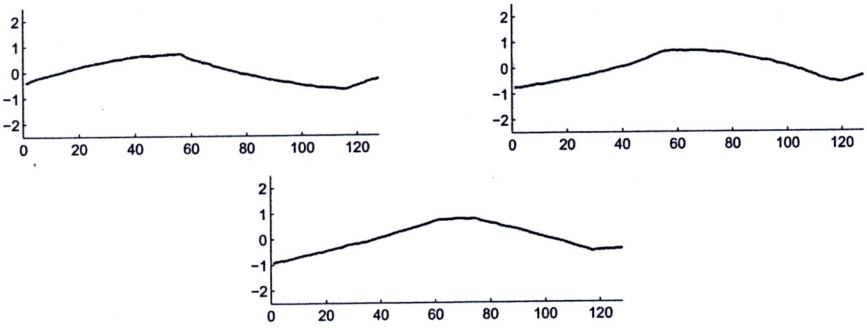


Figure 2.17: Amplitude Averaging produces an smoothened averaged result.

2.5 Conclusion

Subsequence Time Series Clustering (STSC) with both k-means and k-hierarchical clustering algorithms produces sine waves as cluster representatives regardless of an input sequence. To measure meaningfulness, Keogh and Lin have proposed a meaningfulness measurement, called KLMM, which is shown to be invalid because it returns that the result is meaningful even though cluster representatives are sine waves. The causes of meaninglessness are identified as twofold, i.e., an inappropriate distance measure and an inappropriate averaging function, where STSC utilizes Euclidean distance and Amplitude Averaging function as a distance measure and an averaging function. Therefore, the use of appropriate a distance measure and an averaging function can return a meaningful result.