CHAPTER IV

2STSC: SHAPE-BASED SUBSEQUENCE TIME SERIES CLUSTERING

Since Keogh and Lin proved that the clustering results of Subsequence Time Series Clustering (STSC) are meaningless (Lin et al., 2003; Keogh and Lin, 2005), many other methods, e.g., Density-based Subsequence Time Series Clustering (DSTSC) (Denton, 2005), Lag-based Subsequence Time Series Clustering (LSTSC) (Simon et al., 2006; Chen, 2007a,b), and Phrase-Analysis Subsequence Time Series Clustering (PASTSC) (Fujimaki et al., 2008), have been proposed in order to solve this meaninglessness. These previous works introduce additional parameters to discard or filter out trivial-matched subsequences. For DSTSC, a distance threshold is proposed to eliminate groups of clusters that have distances below this threshold, while LSTSC and PASTSC propose a lag value and a slide length to select only some subsequences from a large set of extracted subsequences. These parameters, however, are very sensitive to a clustering result that if inappropriate parameter values are chosen, the clustering results will still be meaningless. In addition, those works have too strict assumption that the time series sequence must be cyclic, where this assuption scarely satisfies in real-world data. Obviously, these previous work do not solve at the right point. Firstly, inappropriate parameter values may discard some useful subsequences, and secondly, distance measures used in those clustering algorithms are based on Euclidean distance that cannot capture similarity between two adjacent subsequences of trivial-matched subsequences. Lastly, a cluster representative generated from those clustering algorithms are from typical statistical values such as a mean or a median, where a mean is an averaged result of all cluster members generated from Amplitude Averaging, and a median is selected from an existing data sequence. Although a median can produce a meaningful clustering representative since it is selected from the existing sequence, the median is still not preferred to be used as a cluster representative because the median is usually sensitive to an imbalanced dataset, while the mean, on the other hand, does preserve characteristics of all data objects in the averaging.

In this chapter, Shape-based Subsequence Time Series Clustering (2STSC) is proposed to produce meaningful clustering results. Since trivial-matched subsequences are contiguous subsequences which have shifts in a time domain, an appropriate distance measure and an averaging function, i.e., Dynamic Time Warping (DTW) distance and Shape-based Averaging, are used to find the optimal alignment before distance calculation and averaging. Suppose there are three sets of trivial-matched subsequences as shown in Figure 4.1, Figure 4.2 demonstrates that Euclidean distance cannot capture the similarity of trivial-matched subsequences by identifying that subsequences from the same set of trivial-matched subsequences are different. Compared to Euclidean distance, DTW distance, on the other hand, can correctly group three sets of trivial-matched subsequences because Euclidean distance calculates a distance in one-to-one manner, while DTW distance finds an optimal alignment before distance calculation. Given the same three sets of subsequences as in Figure 4.3, the Amplitude Averaging produces an averaged result whose shapes are smoothened, while Shape-based Averaging still preserves all characteristics of the sequences, especially the peaks and valleys of the sequences.

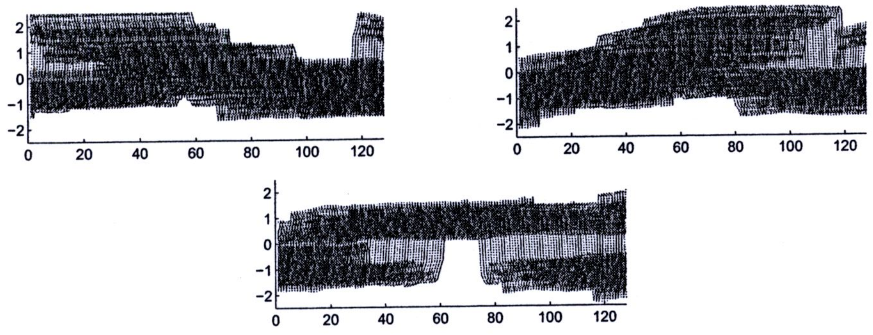


Figure 4.1: Three sets of trivial-matched subsequences.

To be more illustrative, a simple experiment demonstrates that STSC produces meaningless results. The test sequence is generated from concatenation of thirty sequences of three patterns, i.e., Cylinder, Bell, and Funnel, from the CBF dataset. The clustering results of STSC and 2STSC are shown in Figure 4.4, where the number of clusters (k) and the length of sliding window (w) are set to be 3 and 128, respectively. Clustering results of STSC are all sine waves, while 2STSC returns meaningful patterns. Note that 2STSC does not return three patterns, i.e., Cylinder, Bell, and Funnel, as expected cluster representatives because other patterns including joints between the patterns also do exist in the long time series sequence. With this proposed solution that utilizes DTW distance and Shape-based Averaging as a distance measure and an averaging function, 2STSC will demonstrate that it produces meaningful results in an experimental evaluation section.

4.1 Related Work

In this section, related works are reviewed and described to show that subsequence clustering is challenging and still an open problem. So far, no proposed work has yet efficiently solved the problem. This thesis will be the first work to introduce meaningful subsequence clustering algorithm.

Since Keogh and Lin have reported the shocking finding that the output of STSC was mean-

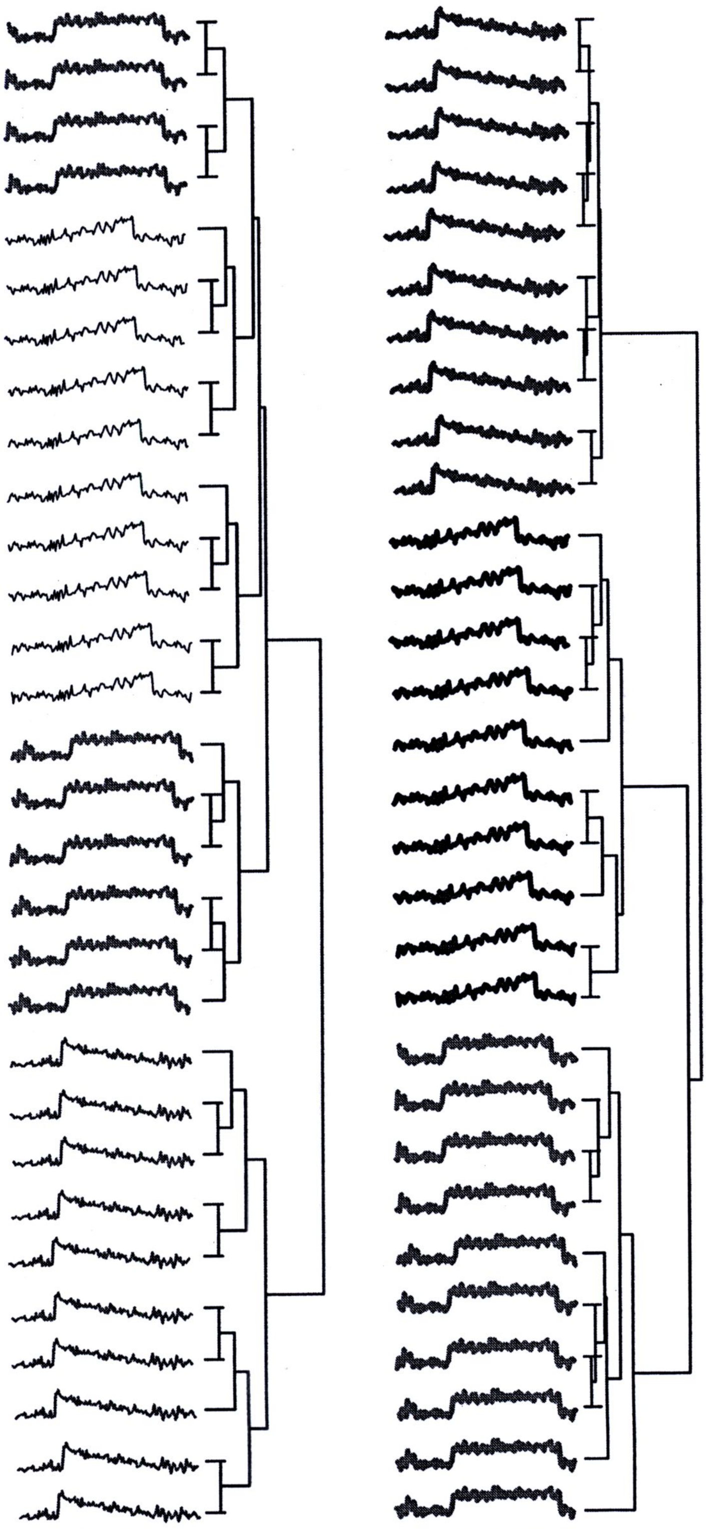
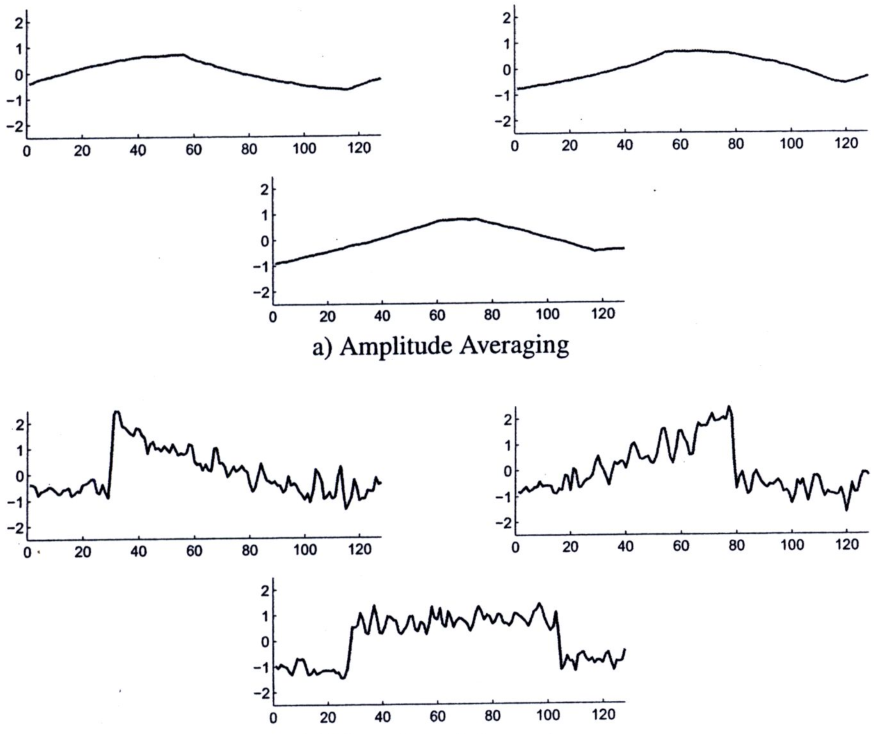


Figure 4.2: a) Euclidean cannot capture the similarity of trivial-matched subsequences, while b) DTW can.

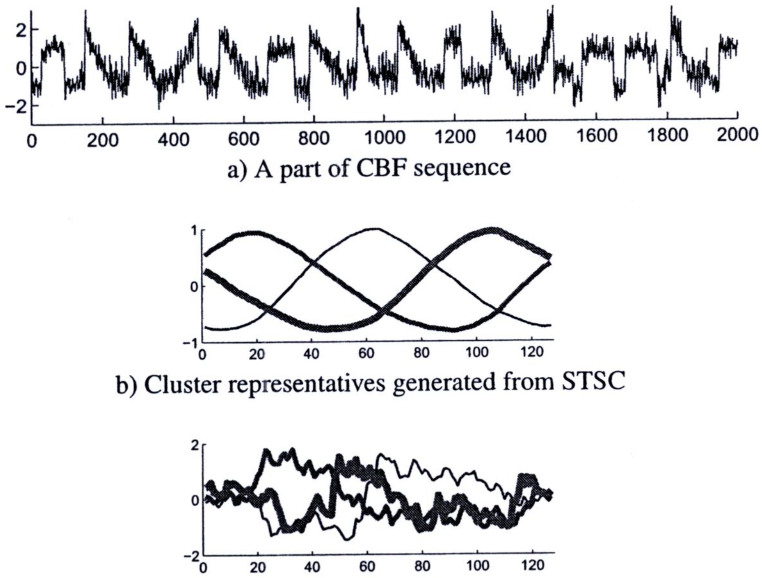
ingless (Lin et al., 2003; Keogh and Lin, 2005), hundreds of works and their successors that use STSC as a subroutine or a preprocessing step are also considered producing meaningless outputs. Keogh and Lin also proposed a tentative solution (Lin et al., 2003; Keogh and Lin, 2005) by using



a) Amplitude Averaging

b) Shape-based Averaging

Figure 4.3: a) Amplitude Averaging cannot construct meaningful representatives of trivial-matched subsequences, while b) Shape-based Averaging can.



a) A part of CBF sequence

b) Cluster representatives generated from STSC

c) Cluster representatives generated from 2STSC

Figure 4.4: a) STSC produces a meaningless clustering result, while b) 2STSC produces a meaningful clustering result.

motif discovery (Mueen et al., 2009) to remove trivial-matched subsequences, and the remaining subsequences are then clustered using k-hierarchical clustering and k-means clustering. However, the motif discovery is parameter-laden in that a real-value distance threshold must be specified in advance to define which sequences are motifs or trivial matches, and using any preprocessing steps to filter out these trivial-matched sequences may lead to an error because some important or desired subsequences are discarded.

Density-based Subsequence Time Series Clustering (DSTSC) (Denton, 2005) has then been proposed by using a kernel function to model trivial-matched subsequences as noises, and a distance threshold has been used to discover the clusters and eliminate noises. Nevertheless, the distance threshold has to be manually defined by users, and its cluster representative is selected from a median of cluster members. However, a median is an undesired cluster representative because a median is affected from imbalance distribution of cluster members that all cluster members should be averaged instead of just selecting one existing sequence. Therefore, a mean is more appropriate than other statistical values, i.e., a mode or a median, since a mean can better reflect characteristics of an interesting data collection by averaging all data sequences.

Lag-based Subsequence Time Series Clustering (LSTSC) (Simon et al., 2006; Chen, 2007b,a) is a subsequence clustering algorithm that re-samples subsequences to a specific lag value using a new distance measure (Chen, 2007b), and a cluster representative is derived from a mean (Chen, 2007b; Simon et al., 2006) or a median (Chen, 2007a). LSTSC requires a lag value by assuming that an input sequence is cyclic. However, a perfect cyclic sequence is scarcely found in real-world data; the output of LSTSC is meaningless if an improper value is chosen (Chen, 2007a). In other words, LSTSC works well when a good lag value is provided by users. To achieve a cluster representative, LSTSC uses a mean or a median of cluster members. Since resampling of subsequences using a lag value cannot be done easily, the cluster representative derived from the mean is still meaningless. In addition, using a median instead of a mean is not a good solution. Although a median selected from an existing sequence is not a sine wave, a median is still not suitable to be a cluster representative due to lack of reflection of data characteristics. Note that some papers (Chen, 2007b) utilize lag-based approach, but subsequences are not normalized before clustering; those papers are, therefore, considered meaningless as well.

Phrase-Analysis Subsequence Time Series Clustering (PASTSC) (Fujimaki et al., 2008) utilizes Discrete Fourier Transform (DFT) to convert a time series sequence to a frequency domain before clustering. For efficient transformation, PASTSC selects the phase which gives the maximum power spectrum as a parameter in DFT. After all subsequences are transformed, those data are clustered using k-means clustering or k-hierarchical clustering algorithms, and then a cluster representative for each clustering is identified in the frequency domain. After clustering is finished, a cluster representative is transformed back to the original time domain. PASTSC has an important parameter, i.e., a slide length that is a number of overlapping subsequences allowed. Since the slide length is used to eliminate trivial-matched subsequences, an inappropriate value still leads to meaninglessness as well. Although the slide length is set so that the output is not meaningless, the cluster representative is not generated from all sequences that some important subsequences are discarded by the slide length.

Perceptually Important Point (PIP) (Fu et al., 2005) has been proposed to reduce the number of dimensions of subsequences before clustering, where PIP captures peaks and valleys of subsequences. Specifically, extracted subsequences are first reduced using PIP, and then redundant subsequences which have the same PIP will be removed, where trivial-matched subsequences normally have similar PIPs. The parameter of this dimensionality reduction is the number of points that is used to represent a subsequence. Additionally, this method is suitable for noisy time series sequences but not smooth sequences since peaks and valleys are hard to be identified in the smooth sequences. However, their paper does not evaluate their clustering results with meaningfulness measurement. Similar to PIPs, many other representation techniques, e.g., Discrete Cosine Transform (DCT) (Kumar et al., 2006) and Discrete Fourier Transform (DFT) (Fujimaki et al., 2008), are also proposed to represent extracted subsequences to be an input of subsequence for clustering algorithms instead of using raw subsequences. However, data representations are not suitable since they require parameters, and precisions of clustering results are lost after these transformations.

These related works (Keogh and Lin, 2005; Denton, 2005; Chen, 2007a; Goldin et al., 2006; Fu et al., 2005; Struzik, 2003; Simon et al., 2006; Kumar et al., 2006; Fujimaki et al., 2008) do not propose the right solutions to deal with trivial-matched subsequences, i.e., new distance measures requires additional parameters and Amplitude Averaging is still used to construct a cluster representative. The distance threshold in DSTSC, the lag value in LSTSC, the slide length in PASTSC, and PIP are additional parameters that users must specify depending on characteristics of each dataset, where these values are sensitive to clustering results. With incorrect values, outputs of clustering results may be meaningless. In addition, these values are used to discard trivial-matched subsequences; therefore, some important trivial-matched subsequences are unexpectedly filtered out. For the meaningfulness measurement, all previous works used Keogh-Lin Meaningfulness Measurement (KLMM) to measure clustering output. As shown in Chapter II, KLMM turns out to be an invalid measurement since it cannot completely capture similarity of two cluster representatives, when the outputs are all sine waves with different phases and frequencies; the outputs will always be intepreted as meaningless.

In this chapter, the issues of similarity between trivial-matched subsequences and cluster representative construction are solved by using the well-known DTW distance and the proposed Shape-based Averaging instead of Euclidean distance and Amplitude Averaging, respectively.

With DTW distance and Shape-based Averaging, the proposed subsequence clustering, Shape-based Subsequence Time Series Clustering (2STSC) will be the first meaningful subsequence clustering algorithm in terms of Shape-based Meaningfulness Measurement (SMM) demonstrated in an experimental evaluation section.

4.2 Shape-based Subsequence Time Series Clustering (2STSC)

Shape-based Subsequence Time Series Clustering (2STSC), a meaningful subsequence clustering algorithm, is proposed in this thesis, where 2STSC utilizes Dynamic Time Warping (DTW) distance and Shape-based Averaging to correctly measure similarity between subsequences and average cluster members for a cluster representative. Shape-based Averaging proposed in this chapter has two variations, i.e., Cubic-Spline Dynamic Time Warping (CDTW) and Iterative Cubic-Spline Dynamic Time Warping (ICDTW) averaging functions. Both CDTW and ICDTW functions use cubic spline interpolation function (Burden et al., 1997) to re-sample x-axis of an averaged sequence, but ICDTW function is more accurate that an averaged result is guaranteed to be in the middle of two original sequences.

To solve the problem of trivial-matched subsequences, contiguous subsequences with small time shifts, 2STSC integrates DTW distance and Shape-based Averaging in k-hierarchical clustering. Specifically, like STSC, 2STSC receives a long time series sequence $S = \langle s_1, s_2, \ldots, s_n \rangle$ as an input, and then this sequence is extracted to be a set $\mathbb{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_i, \ldots, \mathcal{S}_{n-w+1}\}$ of subsequences by a sliding window of length w, where $\mathcal{S}_i = \langle s_i, s_{i+1}, \ldots, s_{i+w-1} \rangle$ and $1 \leq i \leq n - w + 1$. This set of subsequences is then normalized under z-normalization (Han and Kamber, 2000) and clustered with k-hierarchical clustering algorithm, where DTW distance and Shape-based Averaging are used as a distance measure and an averaging function in the algorithm. Finally, 2STSC returns a set $\mathbb{C} = \{C_1, C_2, \ldots, C_k\}$ of k clusters, where each cluster $C = (\mathbb{M}, R)$ contains a set $\mathbb{M} = \{\mathcal{S}_i \mid \mathcal{S}_i \in \mathbb{S}\}$ of cluster members and a cluster representative $R = \langle r_1, r_2, \ldots, r_w \rangle$ from k-hierarchical clustering. Beside an input, 2STSC requires two typical parameters which are the number of clusters (k) and the length of sliding window (w). Visually, an overview of 2STSC is illustrated in Figure 4.5.

K-hierarchical clustering used in 2STSC are agglomerative clustering which uses bottom-up strategy. Specifically, agglomerative clustering iteratively combines atomic clusters to one large cluster. K-hierarchical clustering requires an inter-cluster distance function which is used to calculate a distance between two clusters. In this thesis, 2STSC uses two inter-cluster distance functions, i.e., complete linkage and average linkage distance functions, where complete linkage and average linkage functions are maximum and mean distances, respectively, among all subse-

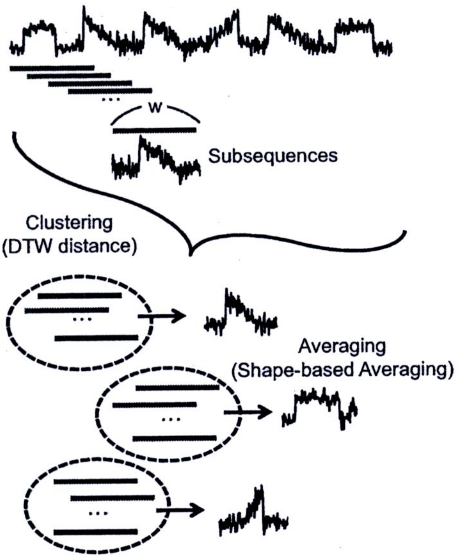


Figure 4.5: Overview of 2STSC using DTW distance and Shape-based Averaging.

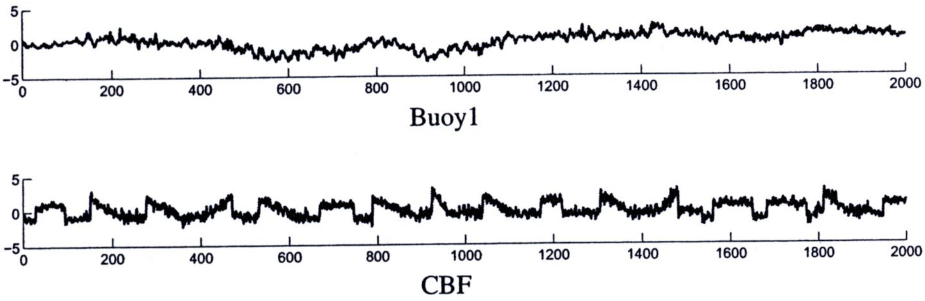
quences pairs between two cluster members. More details of agglomerative clustering algorithms are provided in Section 2.1.2. Concretely, 2STSC with the agglomerative algorithm is shown in Table 4.1. Note that 2STSC does not use the single linkage function as an inter-cluster distance function because the single linkage function cannot handle trivial-matched subsequences. Specifically, some subsequences will never be in any group if these subsequences have the largest nearest neighbor distance. Although an average distance of that subsequence is smaller than others, single linkage will only group based on the smaller nearest neighbor distance. Therefore, in this thesis, only two inter-cluster distance functions are utilized, i.e., complete linkage and average linkage distance functions.

Table 4.1: Pseudo code of Shape-based Subsequence Time Series Clustering (2STSC)

```
FUNCTION [$\mathbb{C}$] = 2STSC [$S$, $k$, $w$]
    $\mathbb{S}$ = EXTRACTSUBSEQUENCES($S$, $w$)
    $\mathbb{S}_{Norm}$ = NORMALIZESUBSEQUENCES($\mathbb{S}$)
    $\mathbb{C}$ = CLUSTERING($\mathbb{S}_{Norm}$, $k$) // with DTW distance and Shape-based Averaging
    Return $\mathbb{C}$
```

4.3 Experimental Evaluation

Shape-based Subsequence Time Series Clustering (2STSC) is evaluated by comparing with STSC in terms of meaningfulness. STSC used in this experiment is implemented on k-means clustering and k-hierarchical clustering with Euclidean distance and Amplitude Averaging, while 2STSC is implemented with k-hierarchical clustering with DTW distance and Shape-based Averaging (CDTW and ICDTW functions). Eight datasets from the Time Series Data Mining Archive (TSDMA) (Keogh and Folias, 2011) used in this experiment are normalized and shown in Appendix A, where each dataset contains 2000 data points. Two datasets, i.e., Buoy1 and CBF, used to illustrated in this experiment is shown in Figure 4.6.



Buoy1

CBF

Figure 4.6: Datasets used to evaluate meaningfulness of STSC and 2STSC

The proposed 2STSC is compared with STSC in terms of meaningfulness. However, Keogh-Lin Meaningfulness Measurement (KLMM) (Lin et al., 2003; Keogh and Lin, 2005) is invalid by the following reasons. First, the assumption of KLMM is that clustering results from the same input sequence should be similar; otherwise, the clustering results should be dissimilar. KLMM, therefore, only compares the distances between the distance of clustering results from the same input and the distance of clustering results from the different inputs. However, KLMM does not have any measurement of similarity between two inputs. Given two similar sequences, clustering reuslts from those two inputs are expected to be similar, but KLMM considers that the results are meaningless although the algorithm produces meaningful results. The second reason is that KLMM cannot capture the similarity of sine waves with different phases or frequencies since KLMM utilizes Euclidean distance to calculate distance between two cluster representatives. Therefore, these cluster results are dissimilar in terms of KLMM although the clustering results are sine waves. In Chapter II, KLMM has been shown that it is considered to be an invalid meaningfulness measurement.

In this experiment, a novel meaningfulness measurement, Shape-based Meaningfulness Measurement (SMM), is introduced to calculate meaningfulness of clustering results. The basic idea of SMM is that clustering results are meaningful if clustering results truly represent subsequences in the time series sequence. In other words, if an input sequence is not a sine wave, cluster representatives should not be sine waves, and if an input sequence is a sine wave, clustering representatives should be sine waves; otherwise, the clustering results are considered meaningless. Unlike KLMM, SMM calculates the meaningfulness between an input sequence and an output clustering result, while KLMM calculates meaningfulness between clustering results from two different datasets. Given an input sequence $S = \langle s_1, s_2, \dots, s_n \rangle$ and an output set $\mathbb{C} = \{C_1, C_2, \dots, C_k\}$ of k clusters, a set $\mathbb{S} = \{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_i, \dots, \mathcal{S}_{n-w+1}\}$ of subsequences are extracted from the input sequence S by a sliding window of length w, where each cluster

$C = (\mathbb{M}, R)$ contains a set $\mathbb{M} = \{\mathcal{S}_i \mid \mathcal{S}_i \in \mathbb{S}\}$ of cluster members and a cluster representative $R = \langle r_1, r_2, \ldots, r_w \rangle$. A set $\mathbb{R} = \{R_1, R_2, \ldots, R_k\}$ of cluster representatives are cluster representatives of all clusters. Specifically, SMM is a summation of minimum distances between each subsequence and cluster representatives. The meaningfulness value can be calculated as the following equation.

$$SMM\left(S, \mathbb{C}\right) = \frac{|\mathbb{S}| \cdot w}{\sum\limits_{i=1}^{|\mathbb{S}|} \min\left(Distance\left(\mathcal{S}_i, R_j\right)\right), \forall R_j \in \mathbb{R}} \tag{4.1}$$

where $Distance\left(\mathcal{S}_i, R_j\right)$ is a DTW distance between two sequences $\mathcal{S}_i$ and R_j.

SMM ranges from zero to positive infinity and is a relative value that SMM must be compared between two algorithms at the same set of parameters to identify that with a given dataset, which subsequence clustering algorithm produces more meaningful clustering results.

Two parameters, i.e., the length of sliding window (w) and the number of clusters (k), are varied to demonstrate the meaningfulness of clustering results of seven variations of subsequence clustering algorithms. Figures 4.7 and 4.8 show SMMs of two datasets when the number of clusters (k) is 3 and the length of sliding window (w) is varied to be 32, 64, and 128, and Figures 4.9 and 4.10 show SMMs of two datasets when the length of sliding window (w) is 64 and the number of clusters (k) is varied to be 3, 5, and 7. Figures 4.11 and 4.12 show cluster representatives of 2STSC of Buoy1 and CBF, respectively, when the number of clusters (k) is 3 and the length of sliding window (w) is 64. The results of other parameter settings and datasets are reported in Appendix D.

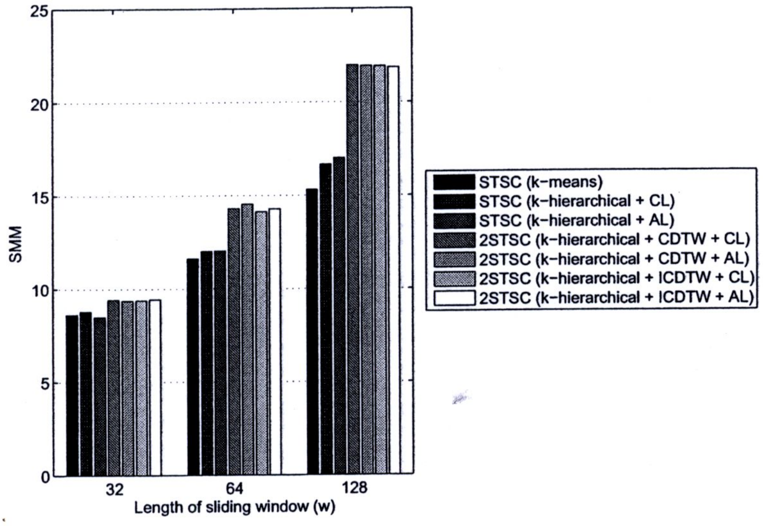


Figure 4.7: SMMs of Buoy1 when the number of clusters (k) is 3 and the length of sliding window (w) is varied.

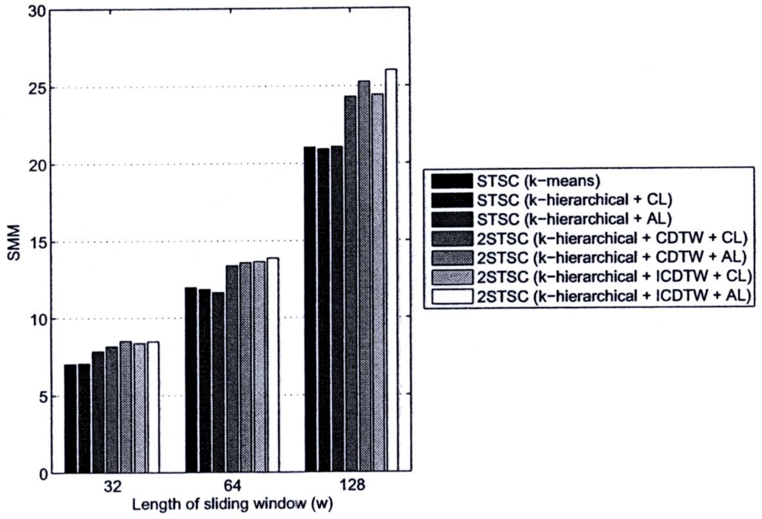


Figure 4.8: SMMs of CBF when the number of clusters (k) is 3 and the length of sliding window (w) is varied.

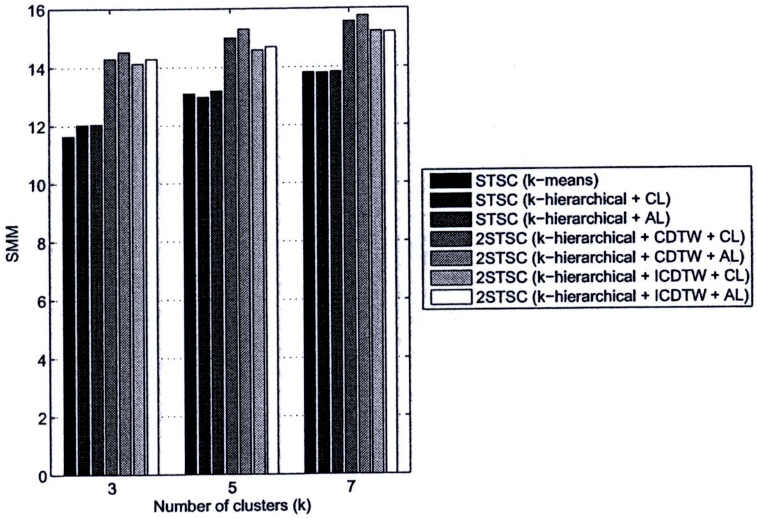


Figure 4.9: SMMs of Buoy1 when the length of sliding window (w) is 64 and the number of clusters (k) is varied.

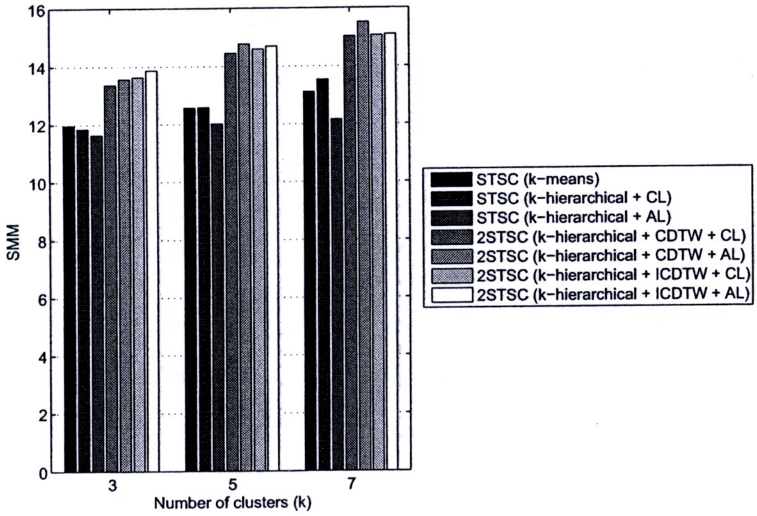


Figure 4.10: SMMs of CBF when the length of sliding window (w) is 64 and the number of clusters (k) is varied.

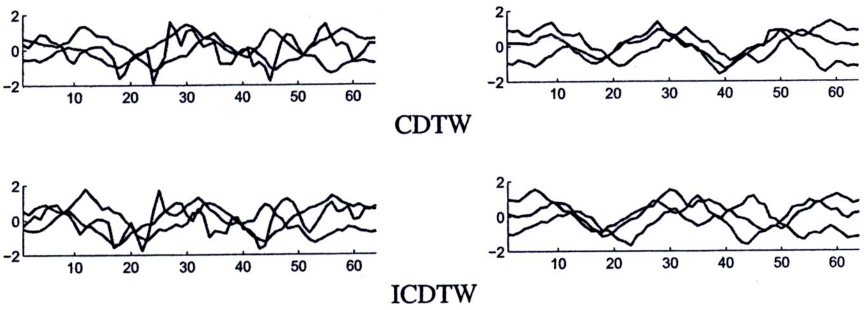


Figure 4.11: Cluster representatives generated from 2STSC of Buoy1 with complete linkage (left) and average linkage (right) when $k = 3$ and $w = 64$.

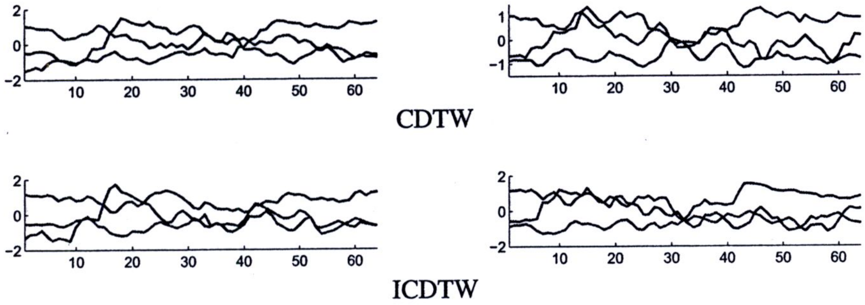


Figure 4.12: Cluster representatives generated from 2STSC of CBF with complete linkage (left) and average linkage (right) when $k = 3$ and $w = 64$.

4.4 Conclusion

DTW distance and Shape-based Averaging are proposed to be used as a distance measure and an averaging function in Shape-based Subsequence Time Series Clustering (2STSC) instead of Euclidean distance and Amplitude Averaging in Subsequence Time Series Clustering (STSC). Instead of discarding trivial-matched subsequences as in many other proposed works, 2STSC uses appropriate distance measure and averaging function that uses DTW distance to capture the similarity between a set of contiguous subsequences and Shape-based Averaging to construct a characteristic-preserved cluster representative. In addition, the cluster results from 2STSC are meaningful in terms of Shape-based Meaningfulness Measurement (SMM) which measures how well a clustering result truly represents characteristics of an input time series sequence. Cluster representatives generated from 2STSC do reflect the characteristics of input sequences, while STSC produces undesired outputs like sine waves. In addition, 2STSC requires no additional parameter like other proposed subsequence clustering algorithms, and 2STSC is extensible to support data streams in Chapter VI.