

Response to Eamonn Keogh’s Criticism of my ICDM’05/KAIS Work

I have written this document in response to the document titled “JasonMeaningless.pdf” on Dr Eamonn Keogh’s homepage. In that document Dr Keogh attacks the work published in [2, 3] Specifically, the criticism revolves around the fact that work in [2, 3] used a CBF time series with fixed feature onset and duration, rather than the variable feature onset and duration version he used in his original paper [1]. Ordinarily, I don’t believe in posting non peer reviewed assertions on the web, however I have been contacted a number of times by people looking for clarification on Dr Keogh’s comments. Hence the following response.

Dr Keogh uses the word “deceptive” in his document which gives the flavour that I engaged in deceptive behaviour over this issue. This is not true. As soon as this problem became clear, I modified the version of the paper on my website and contacted the 10 or so authors of papers who had published papers to that point on this topic. Standardisation of data sets in the data mining community would seem to be an accepted problem, and this was broadly reflected in the response to my email, e.g. “ I know how tricky it is to compare with other people’s data and/or algorithms, so I don’t think it’s surprising at all that you wouldn’t have used exactly what Eamonn did. What’s rather exceptional is that you take the effort of informing people who are interested in the work! ... “

Now a response to the technical issues raised by Dr Keogh. The main assertion by Dr Keogh is that work in [2, 3] is flawed because fixed onset and duration features are used, rather than the variable onset and duration ones used in his paper. However, this really goes against what Dr Keogh says in his original paper, “ that no matter what the data set, clustering method, etc, leads to meaningless results”. Lets take three instantiations of the CBF data set, (a) one with noise and variable feature onset and duration, (b) one with noise and fixed feature onset and duration, and (c) one without noise and fixed feature onset and duration (see Figure 1 ¹). If we STS cluster each of these time series, we end up in all cases with smoothed sine-type wave centroids (see Figure 2), i.e. in Dr Keogh’s terminology, meaningless results. This would suggest to me that the problem we are trying to solve in finding a valid time series clustering method has nothing to do with whether the CBF time series has features with fixed or variable onset and duration, or even whether there is (a reasonable amount of - see below) noise present. However, this is my assertion. Let me now do an experiment which shows it is the case. Dr Keogh claimed that the fixed feature onset and duration version of the CBF time series leads to a much easier clustering problem to the variable CBF instantiation, and this is why my work in [2, 3] was able to produce meaningful clustering outcomes. In recent work [4], which is based on but generalises the work in [2, 3] to non-cyclic time series, I show that

¹these time series have 20 cylinder, bell and funnel features; only the first 12 are shown

the method I'm proposing does cluster the variable feature onset and duration CBF time series correctly. Figure 3 shows the resulting clustering of this time series (i.e time series (a) in Figure 1 above). For more details on this experiment, see [4].

Dr Keogh also makes the assertion that the version of CBF he used has more noise than mine. This maybe true (it looks visually true) however this again is a “red herring” issue. As we saw above, even when the CBF time series has no noise, we still get the smoothed cluster centroid problem. Regarding the method of time series clustering I propose in [4], we saw above that it returns the correct outcome in the presence of noise. If I keep adding noise to the basic underlying CBF signal, then of course there will come a point when it will fail to return the 3 base features. However, is this then really a time series of C, B and F features? At what level of signal to noise ratio does the underlying signal become sufficiently swamped with noise that the signal in the time series becomes fundamentally different. Maybe the approach here is to try to clean the signal of noise first, e.g. by applying a smoothing window over the data first, however this is really a separate topic of research. In my work I have focused on trying to solve the fundamental problems that Dr Keogh's work in [1] raised. The results of this work suggest that the feature onset and duration in the CBF time series, and the level of noise, are not fundamental to the problem with STS-clustering. For my opinion on what are the fundamental problems, see my recent paper (available on my website), “Useful Clustering Outcomes from Meaningful Time Series Clustering” [4].

Jason Chen

24th September 2007

References

- [1] E.Keogh, J.Lin, and W.Truppel. Clustering of time series subsequences is meaningless: Implications for previous and future research. In *Proceedings of the International Conference of Data Mining*, 2003.
- [2] J.R.Chen. Making subsequence time series clustering meaningful. In *Proceedings of IEEE International Conference on Data Mining*, pages 114–121, Houston, USA, November 2005.
- [3] J.R.Chen. Making clustering in delay vector space meaningful. *Knowledge and Information Systems, an International Journal*, 11(3):369–385, April 2007.
- [4] J.R.Chen. Useful clustering outcomes from meaningful time series clustering. In *Proceedings of the Australasian Data Mining Conference*, Gold Coast, Australia, December 2007.

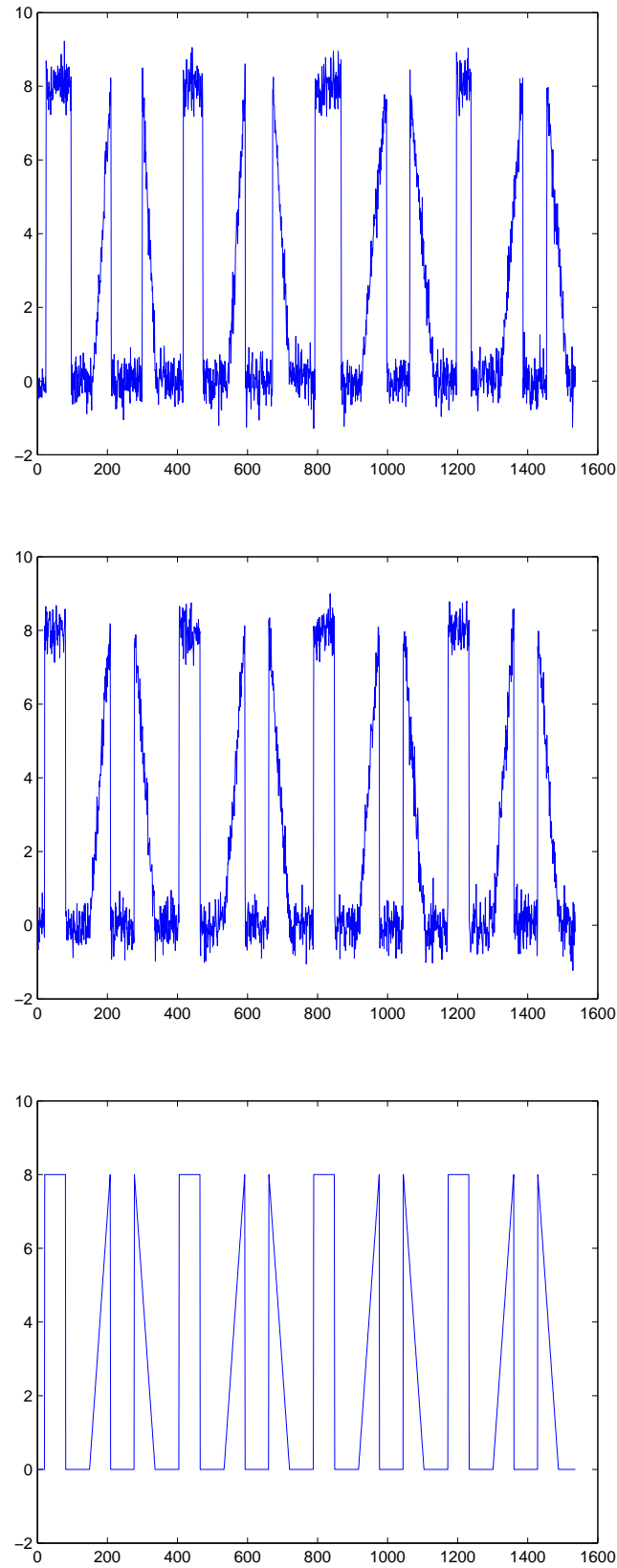


Figure 1: Three instantiations of the CBF time series: variable feature onset and duration with noise (top), fixed feature onset and duration with noise (middle), and fixed feature onset and duration without noise (bottom)

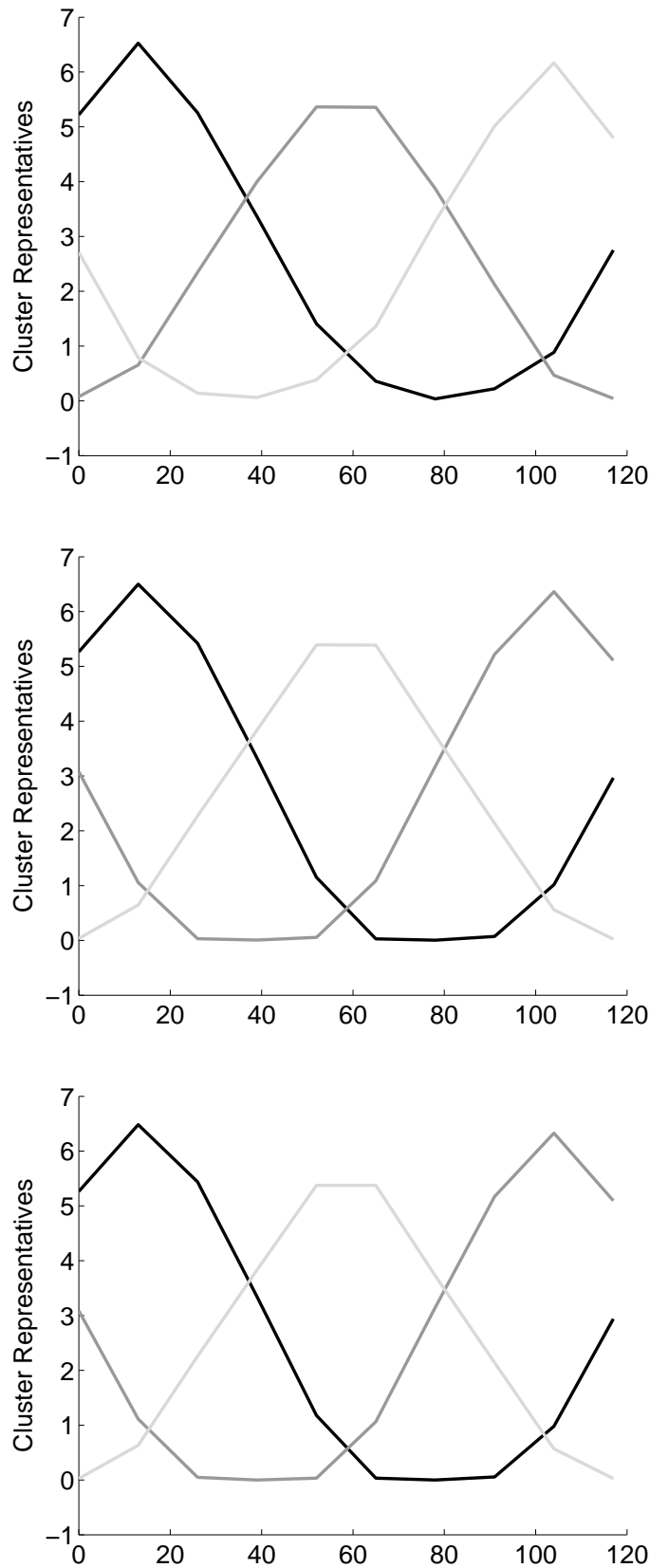


Figure 2: The STS-clustering results for the three instantiations of the CBF time series shown previously, i.e. for the : variable feature onset and duration with noise (top), fixed feature onset and duration with noise (middle), and fixed feature onset and duration without noise (bottom) CBF time series

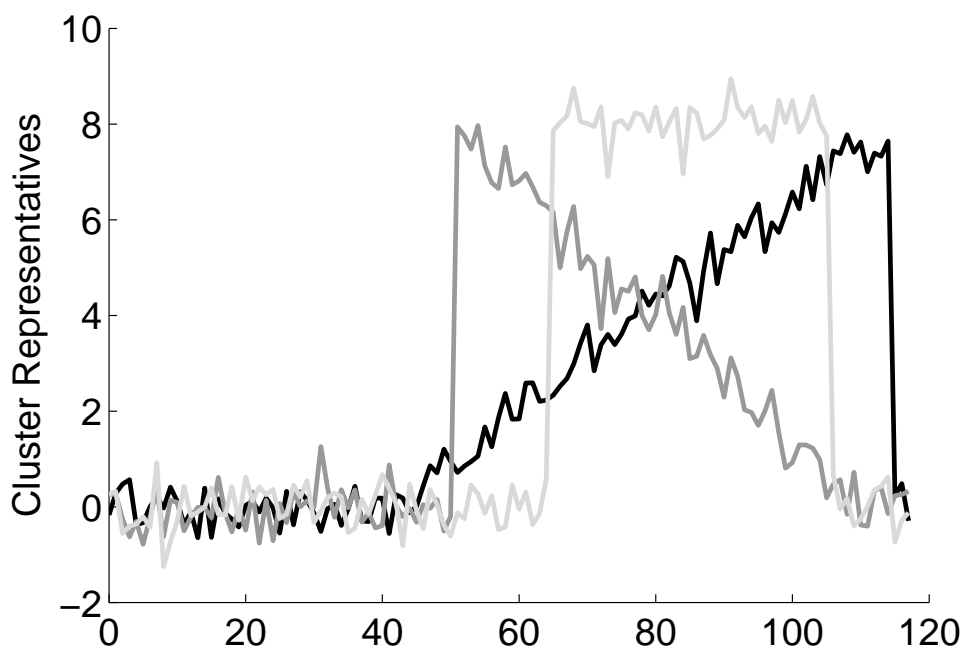


Figure 3: Clustering the CBF time series with variable feature onset and duration using the method proposed in [4]. The result is the correct clustering outcome, in contrast to that shown in Figure 2