Lerato and Niesler *EURASIP Journal on Audio, Speech, and Music Processing* (2019) 2019:6
https://doi.org/10.1186/s13636-019-0149-9

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Feature trajectory dynamic time warping for clustering of speech segments

Lerato Lerato† and Thomas Niesler*†

Abstract

Dynamic time warping (DTW) can be used to compute the similarity between two sequences of generally differing length. We propose a modification to DTW that performs individual and independent pairwise alignment of feature trajectories. The modified technique, termed feature trajectory dynamic time warping (FTDTW), is applied as a similarity measure in the agglomerative hierarchical clustering of speech segments. Experiments using MFCC and PLP parametrisations extracted from TIMIT and from the Spoken Arabic Digit Dataset (SADD) show consistent and statistically significant improvements in the quality of the resulting clusters in terms of F-measure and normalised mutual information (NMI).

Keywords: Dynamic time warping (DTW), Feature trajectory, Speech segments, Agglomerative hierarchical clustering

1 Introduction

Dynamic time warping (DTW) is a method of optimally aligning two distinct time series of generally different length. In addition to the alignment, DTW computes a score indicating the similarity of the two sequences. This ability to quantify the similarity between time series has led to the application of DTW in automatic speech recognition (ASR) systems several decades ago [1, 2]. It has remained popular in this field, with more recent developments reported in [3] and [4].

DTW has also found application in fields related to ASR. For example, it has been used successfully in keyword spotting and information retrieval (IR) systems [5–7]. To accomplish IR, sub-sequences in a speech signal that match a template with certain degree of time warping are detected. The direct approach to keyword spotting has recently been extended by training a convolutional neural network (CNN) to emulate the template matching performed by DTW, thereby providing a substantial computational advantage [8, 9].

In the related task of acoustic pattern discovery, DTW can be allowed to consider multiple local alignments between speech signals during the overall search [10]. In this way, DTW can find similar segment pairs in speech audio, followed by a clustering step [11]. The resulting cluster labels are used to train hidden Markov models (HMMs).

In an effort to improve performance, several variations of DTW have been proposed since its inception. For example, a one-against-all index (OAI) for each time series under consideration is proposed in [4]. The OAI is subsequently used to weight the corresponding DTW alignment score in a speech recognition system.

Another modification of DTW which was reported to improve performance is the parametric derivative dynamic time warping (DDTW) that was applied to hierarchical clustering of UCR Time Series Classification Archive data [12]. Parametric DDTW combines the scores produced by DTW and by DDTW to provide a final similarity measure. A similar weighted modification of DTW has been proposed in [13].

Finally, DTW has also been applied to the direct matching of points along the best alignment for use in a signature verification system [14]. A stability function is subsequently applied, and the resulting score is used as a similarity measure.

We describe a modification of DTW and demonstrate its improved performance when used as a similarity measure to cluster speech segments. Our DTW modification exploits the asynchronous temporal structure of features extracted from speech. Related work has considered such feature trajectories by training separate hidden Markov

*Correspondence: trn@sun.ac.za
†Lerato Lerato and Thomas Niesler contributed equally to this work.
Department of Electrical and Electronic Engineering, University of Stellenbosch, Banghoek Road, Stellenbosch, South Africa

models (HMMs) for each mel frequency cepstral coefficients (MFCC) feature dimension [15]. This work reports improvements in both phoneme and word recognition. The clustering of speech segments also has several useful applications in ASR [16–18]. Recently, it has been particularly useful in the automatic discovery of sub-word units [19, 20].

Section 2 reviews the standard formulation of DTW and Section 3 describes our proposed modification. Section 4 presents the evaluation tools we employ and Section 5 describes the data we use for experimentation. Section 6 presents an experimental evaluation of the proposed method. Section 7 discusses the results and concludes the paper.

2 Classical dynamic time warping

We consider speech segments as temporal sequences of multidimensional feature vectors in the Euclidean space. Sequences are of arbitrary and generally different length, but all vectors are of equal dimension. The DTW algorithm recursively determines the best alignment between two such vector time series by minimizing a cumulative path cost that is commonly based on Euclidean distances between time-aligned vectors [2, 21].

Consider N such sequences $\mathbf{X}_i$, $i = 1, 2, \ldots, N$, each composed of T_i feature vectors, as defined in Eq. 1.

$$\mathbf{X}_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT_i}\}, \quad i = 1, 2, \ldots, N \tag{1}$$

Each feature vector $\mathbf{x}_{it}$ has m dimensions, as indicated in Eq. 2.

$$\mathbf{x}_{it} = \left\langle x_{it}^{(1)}, x_{it}^{(2)}, \ldots, x_{it}^{(m)} \right\rangle, \quad t = 1, 2, .., T \tag{2}$$

Two sequences $\mathbf{X}_i$ and $\mathbf{X}_j$ are aligned by constructing a T_i-by-T_j distance matrix $D_{ij}(p, q)$ whose entries contain the distances $d(\mathbf{x}_{ip}, \mathbf{x}_{jq})$. Typical choices for d are the Euclidean distance and the Manhattan distance. A matrix of minimum accumulated distances $\gamma_{ij}(p, q)$ is then constructed by considering all paths from $D_{ij}(1, 1)$ to $D_{ij}(p, q)$. Using the local and global path constraints, $\gamma_{ij}(p, q)$ is computed recursively according to the principle of dynamic programming, as shown in Eq. 3 [2].

$$\begin{aligned}\gamma_{ij}(p, q) = D_{ij}(\mathbf{x}_{ip}, \mathbf{x}_{iq}) \; + \min \big\{ & \gamma_{ij}(p-1, q-1), \\ & \gamma_{ij}(p-1, q), \gamma_{ij}(p, q-1) \big\}\end{aligned} \tag{3}$$

The similarity $\mathrm{DTW}(\mathbf{X}_i, \mathbf{X}_j)$ between vector sequences $\mathbf{X}_i$ and $\mathbf{X}_j$ is then given by Eq. 4. Here, K is the length of the optimal path from $D_{ij}(1, 1)$ to $D_{ij}(T_i, T_j)$ and is used to normalise the similarity value.

$$\mathrm{DTW}(\mathbf{X}_i, \mathbf{X}_j) = \frac{1}{K} \gamma_{ij}\left(T_i, T_j\right) \tag{4}$$

This standard formulation of dynamic time warping will in the remainder of the paper be referred to as *classical* DTW. Figure 1 shows the classical DTW alignment between two different sequences of 21-dimensional spectral feature vectors representing the same sound uttered by different speakers. These spectral features are obtained by straightforward binning of the short-time power spectra. To avoid clutter, the alignment of just four of the feature vectors is shown.

3 Feature trajectory DTW (FTDTW)

We define a feature trajectory $X_i^{(l)}$ as the time series obtained when considering the l-th element of each feature vector in a sequence $\mathbf{X}_i$, as shown in Eq. 5.

$$X_i^{(l)} = \left\{ x_{i1}^{(l)}, x_{i2}^{(l)}, \ldots, x_{iT_i}^{(l)} \right\}, \quad l = 1, 2, \ldots, m \tag{5}$$

Hence, $X_i^{(l)}$ is a one-dimensional time series for feature l. We now calculate the similarity of two feature vector sequences by applying classical DTW to each corresponding pair of feature trajectories, and subsequently normalise the sum, as shown in Eq. 6.

$$\mathrm{FTDTW}(\mathbf{X}_i, \mathbf{X}_j) = \frac{1}{\beta} \sum_{l=1}^{m} DTW \left\{ X_i^{(l)}, X_j^{(l)} \right\} \tag{6}$$

where $\beta = \sqrt{\sum_{l=1}^{m} K_l^2}$, K is the path length and DTW(.) is non-normalised classical DTW.

As illustrated, we repeat the alignment of the two speech segments shown in Fig. 1 with FTDTW. Figure 2a identifies seven features from each of the four feature vectors shown in Fig. 1a. Figure 2b demonstrates how each of these seven features align with the second speech segment. The features themselves are the same as those illustrated in Fig. 1. For the illustrated example, application of Eq. 6 involves 21 separate alignments, each between corresponding feature trajectories as also indicated in Fig. 2. The resulting 21 scores are summed and normalised by β. Figure 2 illustrates how, in contrast to the classical DTW, FTDTW does not require features coincident in time in one segment to align with features in the other segment also coincident in time. Finally, we note that, because each of the m DTW alignments in the summation of the right-hand side of Eq. 6 is computed independently, the FTDTW computation can be easily parallelised over m processors or cores. This provides a computational advantage over DTW, which involves the alignment of vector sequences and is not so easily parallelised.

4 Evaluation

We evaluate the effectiveness of our proposed modification to DTW by using it to compute similarities between speech segments, and then using these similarities to perform agglomerative hierarchical clustering [22, 23]. We will cluster speech segments corresponding to triphones extracted from the TIMIT corpus as well as isolated digits extracted from the Spoken Arabic Digit Dataset

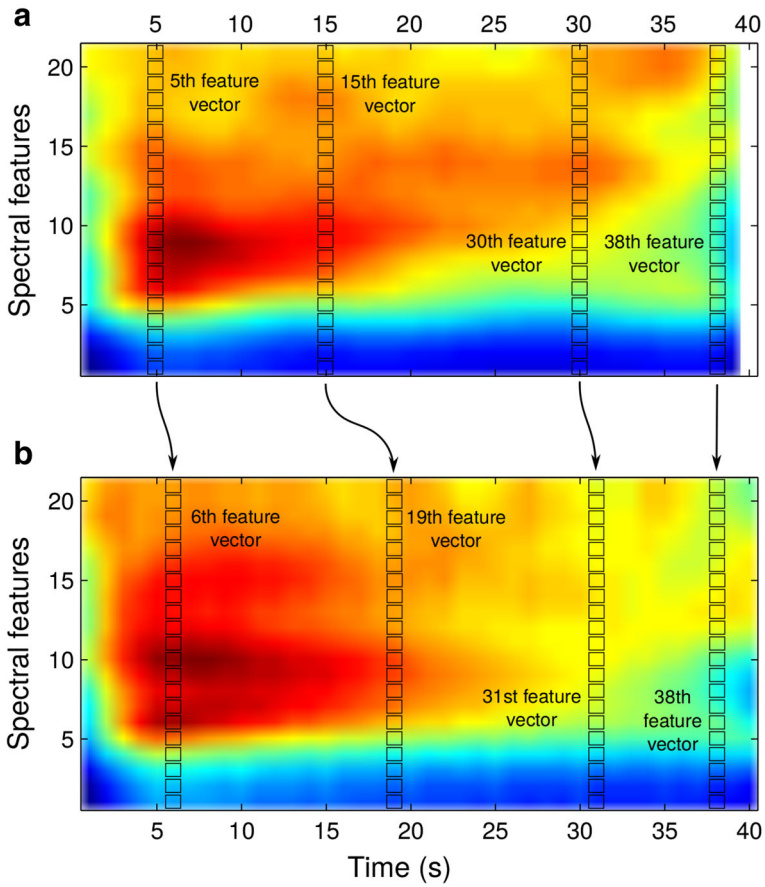

Fig. 1 Alignment by classical DTW of spectral features extracted from the triphone *b-aa+dx* as uttered by **a** male speaker mrfk0 and **b** by female speaker fdml0 in the TIMIT corpus

(SADD). Since the phonetic alignment is provided in the former and the word alignments in the latter, the ground truth is available. Hence, we can use the external metrics to quantify the quality of the resulting clusters [24]. We chose F-measure and normalised mutual information (NMI) as metrics for cluster evaluation in our experiments [25, 26]. These two metrics represent two commonly used categories of external evaluation measures called set-matching-based measures and information theoretic-based measures. The F-measure was chosen because it is a widely used set matching-based measure for the evaluation of clustering and classification systems [27]. The NMI is a popular choice among the information theoretic-based clustering evaluation measures [28].

4.1 Agglomerative hierarchical clustering

In agglomerative hierarchical clustering (AHC), the agglomeration of data objects (speech segments in the case of our experimental evaluation) is initialised by the assumption that each object is the sole occupant of its own cluster. A binary tree referred to as a dendrogram is created by successively merging the closest cluster pairs until a single cluster remains [29]. We use the popular Ward method to quantify inter-cluster similarity [30]. The input to the AHC algorithm is a symmetric $N \times N$ proximity matrix populated by the values of $\mathrm{DTW}(\cdot,\cdot)$ or $\mathrm{FTDTW}(\cdot,\cdot)$ and the output consists of the R clusters.

4.2 F-measure

The F-measure is based on the quantity precision (PR) and recall (RE). Precision indicates the degree to which a cluster is dominated by a particular class, while recall indicates the degree to which a particular class is concentrated in a specific cluster. Precision and recall are defined in Eqs. 7 and 8 respectively.

$$\mathrm{PR}(r,v) = \frac{n_{rv}}{n_r} \tag{7}$$

$$\mathrm{RE}(r,v) = \frac{n_{rv}}{n_v} \tag{8}$$

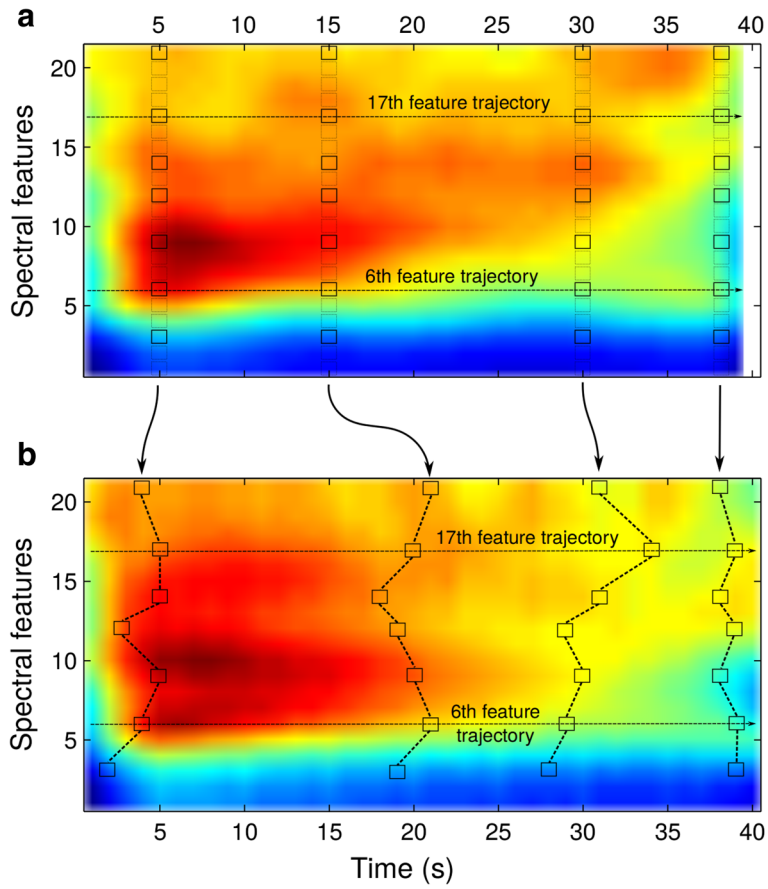

Fig. 2 Alignment by FTDTW of spectral features extracted from the triphone *b-aa+dx* as uttered by **a** male speaker mrfk0 and **b** by female speaker fdml0 in the TIMIT corpus

Here, n_{rv} indicates the number of objects of class v in cluster r; n_r and n_v indicate the number of objects in cluster r and class v respectively. The F-measure (F) is given in Eq. 9.

$$F(r,v) = \frac{2 \times \mathrm{RE}(r,v) \times \mathrm{PR}(r,v)}{\mathrm{RE}(r,v) + \mathrm{PR}(r,v)} \tag{9}$$

When the clusters are perfect, $n_{rv} = n_r = n_v$, and hence, $F(r,v) = 1$.

4.3 Normalised mutual information

Normalised mutual information (NMI) employs the following formulations:

- The set of R clusters $\mathbf{G} = \{G_1, G_2, \ldots, G_R\}$, and
- The set of V classes $\mathbf{C} = \{C_1, C_2, \ldots, C_V\}$ representing ground truth.

NMI is based on the mutual information $I(\mathbf{G},\mathbf{C})$ between classes and clusters [26, 31]. The mutual information is not sensitive to varying number of clusters, and therefore, it is normalised by a factor based on the cluster entropy $H(\mathbf{G})$ and class entropy $H(\mathbf{C})$. These entropies measure cluster and class cohesiveness respectively. The NMI criterion is given in Eq. 10.

$$\mathrm{NMI}(\mathbf{G},\mathbf{C}) = \frac{2I(\mathbf{G},\mathbf{C})}{[H(\mathbf{G}) + H(\mathbf{C})]} \tag{10}$$

The mutual information $I(\mathbf{G},\mathbf{C})$ and the entropies $H(\mathbf{G})$ and $H(\mathbf{C})$ are given in Eqs. 11, 12 and 13 respectively.

$$I(\mathbf{G},\mathbf{C}) = \sum_{r\in\mathbf{G}} \sum_{v\in\mathbf{C}} P(G_r)P(C_v) \log \frac{P(G_r \cap C_v)}{P(G_r)P(C_v)} \tag{11}$$

In Eq. 11, $P(G_r)$, $P(C_v)$, and $P(G_r \cap C_v)$ are the probabilities of a segment belonging to cluster G_r, class C_v and the intersection of G_r and C_v respectively.

$$H(\mathbf{G}) = -\sum_{r\in\mathbf{G}} P(G_r) \log P(G_r) \tag{12}$$

$$H(\mathbf{C}) = -\sum_{v\in\mathbf{C}} P(C_v) \log P(C_v) \tag{13}$$

It can be shown that $I(\mathbf{G},\mathbf{C})$ is zero when the clustering is random with respect to class membership and that it achieves a maximum of 1.0 for perfect clustering [31].

Table 1 Datasets used for experimental evaluation

Dataset	Description
1	8772 TIMIT triphones (evenly balanced).
2	8800 SADD isolated digits (evenly balanced).
3	123,182 TIMIT SI and SX triphones divided randomly into 10 subsets (not evenly balanced).

5 Data

Our first set of experiments uses speech segments taken from the TIMIT speech corpus [32]. TIMIT has been chosen because it includes accurate time-aligned phonetic transcriptions, meaning that both phonetic labels and their start/end times are known. As our desired clusters, we use triphones, which are phones in specific left and right contexts [33]. We consider triphones that occur at least 20 times and at most 25 times in the corpus. This leads to an evenly balanced set of 8772 speech segments, which also corresponds approximately to the number of segments in our second set of experiments.

For comparison and confirmation purposes, we performed a second set of experiments using the Spoken Arabic Digit Dataset (SADD) [34]. SADD consists of 8800 utterances already parametrised as 13-dimensional MFCCs. The utterances were spoken by 44 male and 44 female Arabic speakers. Each utterance in the SADD corresponds to a single Arabic digit and will therefore be considered to be a single segment in our experiments. Each digit (0 to 9) was uttered ten times by each speaker.

A third set of experiments is based on 10 independent subsets of speech segments drawn from the TIMIT SI and SX utterances, irrespective of occurrence frequency. This better represents the unbalanced distribution of triphones that may be expected in unconstrained speech. Table 1 summarises the datasets used in each of the three sets of experiments.

We considered two feature vector parametrisations popular in the field of speech processing, namely mel frequency cepstral coefficients (MFCCs) and perceptual linear prediction (PLP) coefficients [35, 36]. For the former, log frame energy was appended to the first 12 MFCCs to produce a 13-dimensional feature vector. The first and second differentials (velocity and acceleration) were subsequently added to produce the final 39-dimensional

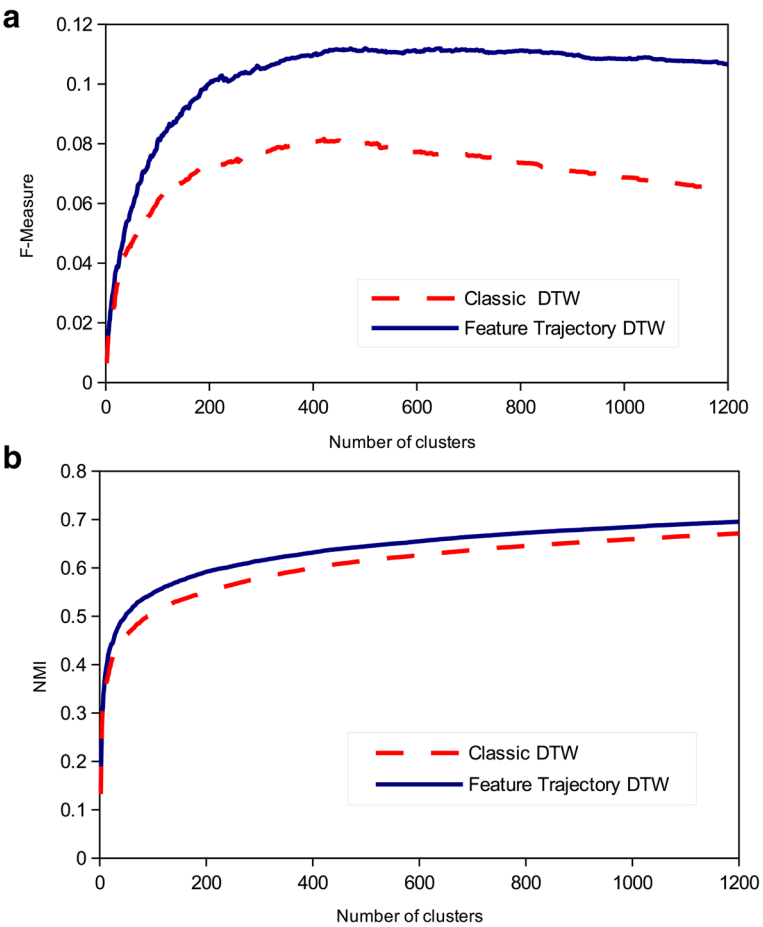

Fig. 3 Clustering performance for dataset 1 when using MFCC features in terms of **a** F-measure and **b** NMI

MFCC feature vector. For the latter, 13 PLP coefficients were considered, to which velocity and acceleration were added, again resulting in a 39-dimensional feature vector. One such feature vector was extracted for each 10 ms frame of speech, where consecutive frames overlapped by 5 ms. All TIMIT feature vectors were computed using HTK [37]. SADD provides pre-computed MFCC features, and hence, PLP features were not used in the associated experiments.

6 Experiments

To evaluate the performance of feature trajectory DTW (FTDTW) as an alternative to classical DTW as a similarity measure, we will employ it to perform AHC of the speech segments described in Section 5. The quality of the automatically determined clusters will be determined using the F-measure and in several cases also NMI.

In a first set of experiments, we cluster dataset 1 (Table 1).

Figure 3 reflects the clustering performance in terms of (a) the F-measure and (b) NMI, when using MFCCs as features. Both the F-measure and NMI are plotted as a function of the number of clusters. Note that the F-measure continues to decline as the number of clusters exceeds 1200.

Figure 3a and b show that FTDTW improves on the performance of the classical DTW in this clustering task in terms of both F-measure and NMI. Especially in terms of F-measure, this improvement is substantial.

A corresponding set of experiments using PLP features was carried out for dataset 1, and the results are shown in Fig. 4. The same trends seen for MFCCs in Fig. 4 are observed, with substantial improvements particularly in terms of F-measure.

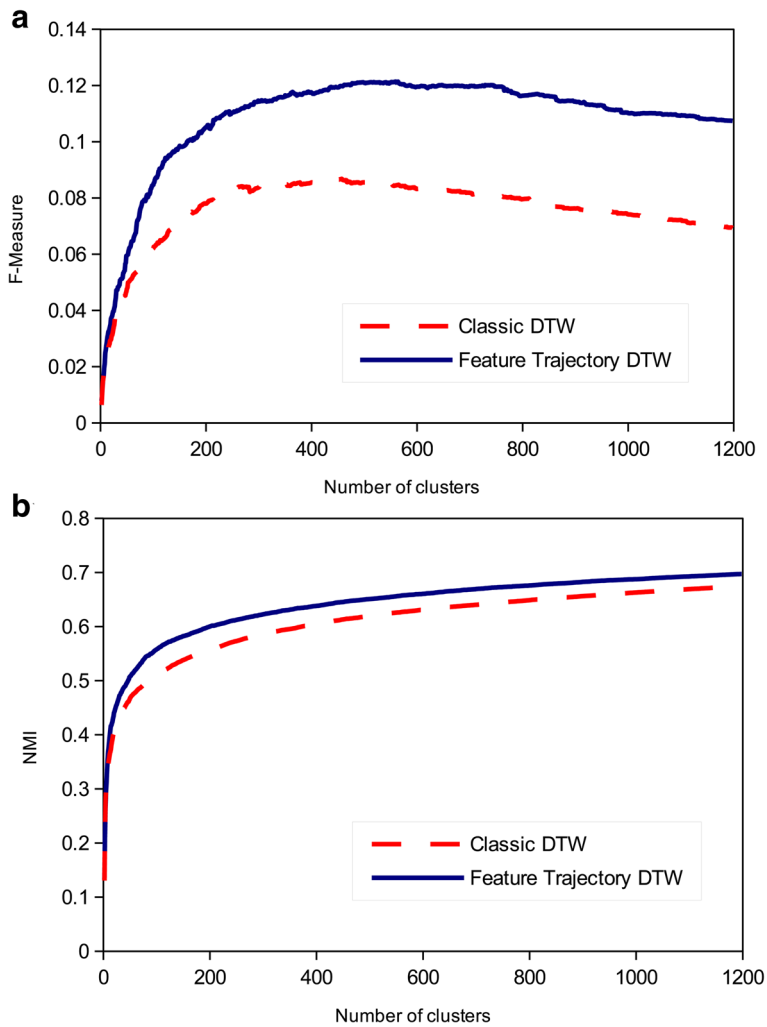

Fig. 4 Clustering performance for dataset 1 when using PLP features in terms of **a** F-measure and **b** NMI

In a second set of experiments, we clustered dataset 2 (Table 1) which consists of isolated Arabic digits. Figure 5 indicates the clustering performance, both in terms of F-measure and NMI for this dataset. Again, we observe that FTDTW outperforms the classical DTW in terms of both F-measure and NMI in practically all cases.

In a third and final set of experiments, we considered dataset 3 (Table 1). The 10 independent subsets of the TIMIT training set each contained between 12034 and 12495 triphone segments. In contrast to the experiments for dataset 1, all triphone tokens were considered irrespective of occurrence frequency. Furthermore, the number of clusters was chosen to be 2394, a figure which corresponds to the number of triphone types with more than 10 occurrences in the data. A single number of clusters, rather than a range as presented in Figs. 3, 4 and 5, has been used here in order to make the required computations practical. Figure 6 presents the clustering performance for each of the 10 subsets in terms of F-measure. We observe that FTDTW achieves an improvement over classical DTW in all cases. A paired t test indicated $p < 0.0001$, and hence, the improvements are statistically highly significant. Similar improvements were observed in terms of NMI.

7 Discussion and conclusions

The experiments in Section 6 have applied our modified DTW algorithm (FTDTW) to the clustering of speech segments. Our experiments show consistent and statistically significant improvement over the classical DTW baseline for both MFCC and PLP parametrisations and across three datasets. We conclude that FTDTW is more effective as a similarity measure for speech signals than the classical DTW.

Because the classical DTW operates on a feature-vector by feature-vector basis, it enforces absolute temporal synchrony between the feature trajectories. In contrast, FTDTW does not impose this synchrony constraint, but aligns feature trajectories independently on a pair-by-pair basis. Since FTDTW is observed to lead to better clusters

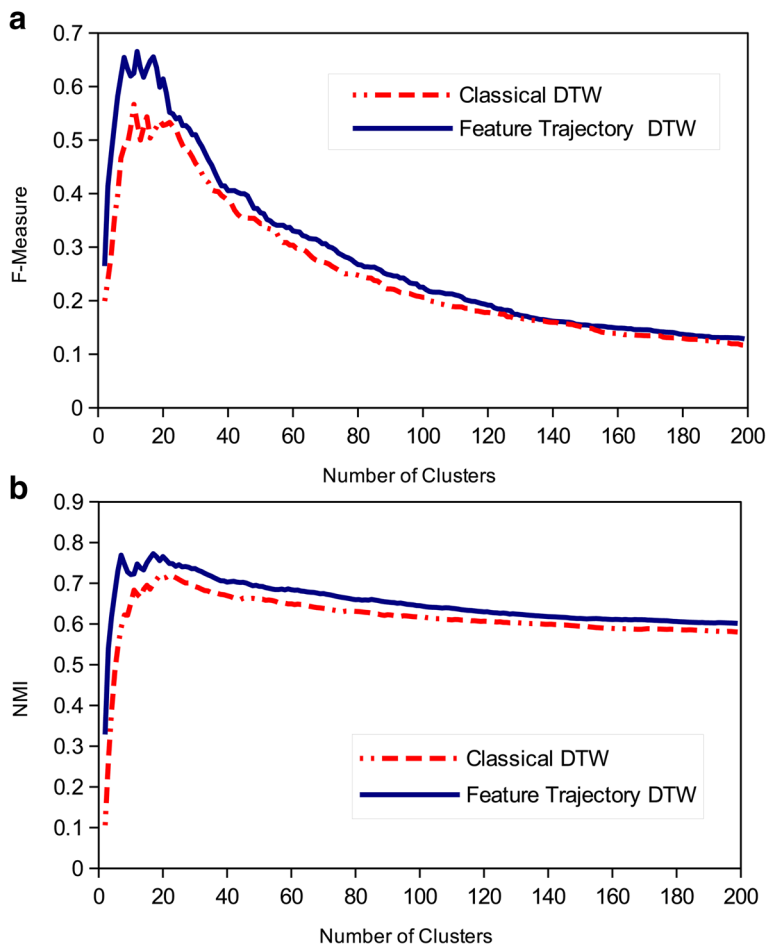

Fig. 5 Clustering performance for dataset 2 in terms of **a** F-measure and **b** NMI

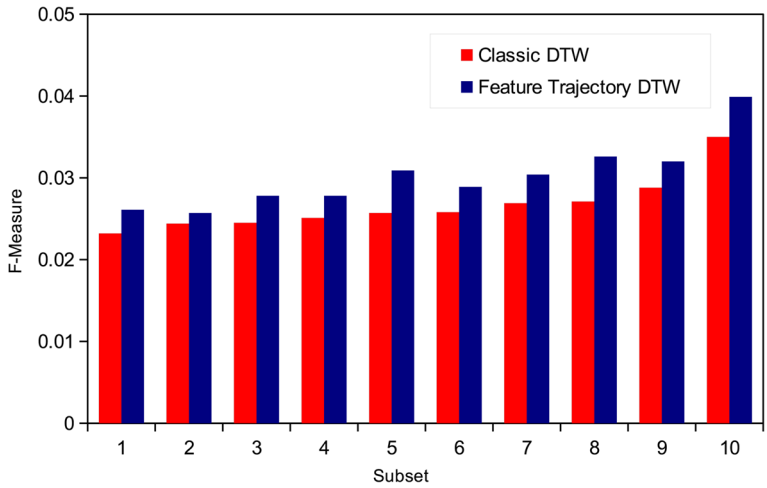

Fig. 6 Clustering performance for the 10 independent subsets of dataset 3 in terms of F-measure

in our experiments, we conclude that the strict temporal synchrony imposed by the classical DTW is counterproductive in the case of speech signals. We further speculate that segments of speech that human listeners would regard as similar also exhibit such differing time-scale warping among the feature trajectories. It remains to be seen whether this decoupling of the feature trajectories is advantageous for signals other than speech.

Finally, and noting that it is not a focus of this paper, we may consider the maxima observed in the F-measure in Figs. 3 and 4, and in both the F-measure and NMI in Fig. 5. A peak in the quality of the clusters as a function of the number of clusters may be taken to indicate the best estimate of the 'true' number of clusters in the data. For the experiments using the MFCC parametrisation of dataset 1 (Fig. 4), we see that an optimum in the F-measure is reached at 501 and 421 clusters for FTDTW and classical DTW respectively. The 'true' number of clusters corresponds to the number of triphone types in dataset 1, which is 404. Hence, both DTW formulations over-estimate the number of clusters. A similar tendency is seen for the PLP parametrisations of the same dataset, where the F-measure peaks at 439 and 559 clusters for the classical DTW and the proposed DTW respectively, and also for dataset 2 in Fig. 5.

Although the ground truth is known, the class definitions (triphones for datasets 1 and 3 and isolated digits for dataset 2) may be called into question. In particular, although all triphones correspond to acoustic segments from the same phone within the same left and right contexts, there are many other possible sources of systematic variability, such as the accent of the speaker. Hence, it may be reasonable to expect that a larger number of clusters are needed to optimally model the data. To determine whether this is the case, the clusters should be used to determine acoustic models for an ASR system. Then, the performance of varying clusterings of the data can be compared by comparing the performance of the resulting ASR systems. We intend to address this question in the ongoing work.

Acknowledgements
Not applicable.

Funding
Support received from Telkom South Africa and the Department of Arts and Culture of South Africa. The funders had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
TIMIT data are commercially available from (https://catalog.ldc.upenn.edu/LDC93S1). All interested parties can obtain the data in the same way that the authors did. A similar, publicly available alternative SADD dataset is available from: (https://archive.ics.uci.edu/ml/datasets/Spoken+Arabic+Digit).

Authors' contributions
LL and TN conceived the algorithm. LL wrote the software, executed the experiments, and drafted the document. TN directed the project. LL and TN completed the the final manuscript. Both authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 7 November 2018 Accepted: 15 March 2019
Published online: 04 April 2019

References

1. H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Signal Proc. **26**(1), 43–49 (1978)
2. C. Myers, L. R. Rabiner, A. E. Rosenberg, Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. IEEE Trans. Acoust. Speech Signal Proc. **28**(6), 623–635 (1980)
3. L. Muda, M. Begam, I. Elamvazuthi, Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. J. Comput. **2**, 138–143 (2010)

4. X. Zhang, J. Sun, Z. Luo, One-against-all weighted dynamic time warping for language-independent and speaker-dependent speech recognition in adverse conditions. PloS ONE. **9**(2), e85458 (2014)
5. Y. Zhang, J. R. Glass, in *Proc. IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*. Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams (IEEE, Merano, 2009), pp. 398–403
6. X. Anguera, in *Proc. Interspeech*. Information retrieval-based dynamic time warping (International Speech Communication Association, Lyon, 2013), pp. 1–5
7. L.-S. Lee, J. Glass, H.-Y. Lee, C.-A. Chan, Spoken content retrieval—beyond cascading speech recognition with text retrieval. Audio Speech Lang. Process. IEEE/ACM Trans. **23**(9), 1389–1420 (2015)
8. R. Menon, H. Kamper, E. Yilmaz, J. Quinn, T. Niesler, in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*. ASR-free CNN-DTW keyword spotting using multilingual bottleneck features for almost zero-resource languages, (Gurugram, 2018), pp. 20–24
9. R. Menon, H. Kamper, J. Quinn, T. Niesler, in *Interspeech*. Fast ASR-free and almost zero-resource keyword spotting using DTW and CNNs for humanitarian monitoring (ISCA, Hyderabad, 2018), pp. 2608–2612
10. A. Park, J. Glass, in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Towards unsupervised pattern discovery in speech (IEEE, San Juan, 2005), pp. 53–58
11. O. Walter, T. Korthals, R. Haeb-Umbach, B. Raj, in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. A hierarchical system for word discovery exploiting DTW-based initialization (IEEE, Olomouc, 2013), pp. 386–391
12. M. Łuczak, Hierarchical clustering of time series data with parametric derivative dynamic time warping. Expert Syst. Appl. **62**, 116–130 (2016)
13. Y.-S. Jeong, M. K. Jeong, O. A. Omitaomu, Weighted dynamic time warping for time series classification. Pattern Recog. **44**(9), 2231–2240 (2011)
14. A. P. Shanker, A. Rajagopalan, Off-line signature verification using DTW. Pattern Recogn. Lett. **28**(12), 1407–1414 (2007)
15. S. Sagayama, S. Matsuda, M. Nakai, H. Shimodaira, in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*. Asynchronous-transition HMM for acoustic modeling (IEEE, Keystone, 1999), pp. 99–102
16. T. Svendsen, K. K. Paliwal, E. Harborg, P. O. Husoy, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. An improved sub-word based speech recognizer (IEEE, Glasgow, 1989), pp. 729–732
17. K. K. Paliwal, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Lexicon-building methods for an acoustic sub-word based speech recognizer (IEEE, Albuquerque, 1990), pp. 108–111
18. H. Wang, T. Lee, C. Leung, B. Ma, H. Li, in *Proc. Interspeech*. Unsupervised mining of acoustic subword units with segment-level Gaussian posteriograms (ISCA, Lyon, 2013), pp. 2297–2301
19. K. Livescu, E. Fosler-Lussier, F. Metze, Subword modeling for automatic speech recognition: past, present, and emerging approaches. IEEE Signal Proc. Mag. **29**(6), 44–57 (2012)
20. H. Kamper, A. Jansen, S. King, S. Goldwater, in *Proc. IEEE Spoken Language Technology Workshop (SLT)*. Unsupervised lexical clustering of speech segments using fixed-dimensional acoustic embeddings (IEEE, South Lake Tahoe, 2014)
21. E. J. Keogh, M. J. Pazzani, in *Proc. SIAM International Conference on Data Mining*. Derivative Dynamic Time Warping, vol. 1 (Society for Industrial and Applied Mathematics, 2001), pp. 1–11
22. A. K. Jain, R. C. Dubes, *Algorithms for Clustering Data*. (Prentice-Hall, Inc., Upper Saddle River, 1988)
23. F. Murtagh, P. Contreras, Methods of hierarchical clustering. Comput. Res. Repository. **abs/1105.0121** (2011). http://arxiv.org/abs/1105.0121. Accessed 12 Sept 2018
24. E. Amigo, J. A. J. Gonzalo, F. Verdejo, A comparison of extrinsic clustering evaluation metrics based on formal constraints. Inf. Retr. **12**(4), 461–486 (2009)
25. B. Larsen, C. Aone, in *Proc. Fifth ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Fast and effective text mining using linear-time document clustering (ACM, New York, 1999), pp. 16–22
26. N. X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: variants, properties, normalisation and correction for chance. J. Mach. Learn. Res. **11**, 2837–2854 (2010)
27. J. Wu, H. Xiong, J. Chen, Towards understanding hierarchical clustering: a data distribution perspective. Neurocomputing. **72**(10-12), 2319–2330 (2009)
28. M. C. P. de Souto, A. L. V. Coelho, K. Faceli, T. C. Sakata, V. Bonadia, I. G. Costa, in *2012 Brazilian Symposium on Neural Networks*. A comparison of external clustering evaluation indices in the context of imbalanced data sets (IEEE Computer Society, Curitiba, 2012), pp. 49–54
29. R. Xu, D. Wunsch, Survey of clustering algorithms. IEEE Trans. Neural Netw. **16**(3), 645–678 (2005)
30. F. Murtagh, P. Legendre, Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? J. Classif. **31**(3), 274–295 (2014). https://doi.org/10.1007/s00357-014-9161-z
31. C. D. Manning, P. Raghavan, *Introduction to Information Retrieval*. (Cambridge University Press, New York, 2008)
32. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. Dahlgren, V. Zue, DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon Technical Report n. **93**, 27403 (1993)
33. B. Imperl, Z. Kacic, B. Horvat, A. Zgank, Clustering of triphones using phoneme similarity estimation for the definition of a multilingual set of triphones. Speech Comm. **39**(4), 353–366 (2003)
34. M. Lichman, UCI machine learning repository (2013). http://archive.ics.uci.edu/ml. Accessed 25 Oct 2018
35. S. B. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Sig. Process. **28**(4), 357–366 (1980)
36. H. Hermansky, Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. **87**(4), 1738–1752 (1990)
37. S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. C. Woodland, *The HTK Book, Version 3.4*. (Cambridge University Engineering Department, Cambridge, 2006)