INVESTIGATING MACHINE LEARNING (AND NON-LEARNING) METHODS FOR TIME WARPING OF SPEECH

Adnan Haider

April 10, 2013

1 Goal

The purpose of this MSc project is to investigate techniques in improving the performance of the Dynamic Time Warping algorithm when applied to semi-supervised pattern recognition in speech. This approach is a bold alternative to using Hidden Markov models which has served as the standard model framework for large vocabulary speech recognition.

2 Motivation

Modern Speech recognition systems are typically built using a supervised training methodology that employs transcribed speech data to model the underlying speech process. Annotation of speech corpus are performed manually and are conducted at the level of phonemes. By a phoneme we refer to a unit of sound. Since annotation is performed at the level of phones, given the small duration of phones, it's highly likely for annotators to miss phonetic boundaries or mislabel phone intervals. Hence the entire process of annotation is not only highly time-consuming but also prone to human error. Since the acquisition of labels is very expensive, the standard state of art systems are limited on their ability to be applied to new domains. Furthermore the use of Hidden Markov models, that has served as the most popular choice of models (as they provide great flexibility:the size, type and architectural adeptness to different application domains) used to represent the underlying speech process may lead to over-fitted models when the size of the training data is substantially small[3]. This is because HMM models consist of large number of parameters that require substantial amount of training data to be trained before being applied in any application domain.

Given the relative ease with which we can create and store large quantities of speech signals, the following two questions have been often being raised in the last decade : how much can be learned from speech data alone and are there any domains where such unsupervised

methods can be applied? These two questions have been addressed in the papers[1,2] by Park and Glass. The duo in their papers propose the problem of *motif discovery* as an example of one such domain where such methods can be applied. The motivation behind motif discovery comes from the area in comparative genomics. Unlike speech recognition systems that are embedded with a lexicon dictionary, the lexicon of interesting DNA subsequences are not known ahead of time. By aligning continuous sequences with each other and identifying sub-sequences that are frequently recurrent, the discovery of biologically significant sequences can be made. Based on the observation that patterns of speech sounds are more likely to be consistent within word or phrase boundaries than across, the techniques employed to find recurrent patterns in DNA sequences can be applied to find recurrent patterns in speech that signify the presence of topic-dependent frequent words.

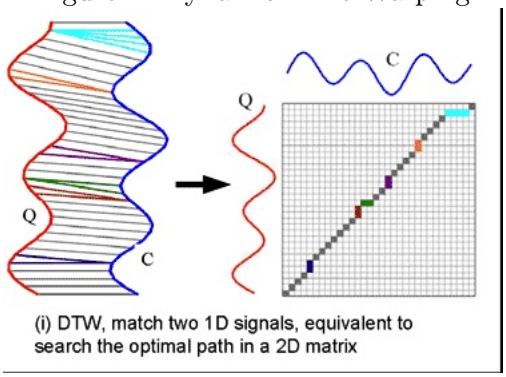
One of the most successful algorithms employed to compare the similarity between two sequences in the context of unsupervised pattern recognition is the Dynamic Time Warping algorithm. This algorithm uses the analogy of dynamic programming to search a space of mapping between the time axis of two temporal sequences in order to find the minimum distortion between them. The pairwise distance computed by DTW is domain and application independent. However in the paper[5] by Xie and Witgen, it can be seen that the alignments found by the DTW algorithm can be greatly improved if the algorithm was adapted to its ambient domain or application.

Improving the performance of the DTW is critical for detecting substantial number of motifs from an acoustic speech signal. Although the overarching goal is unsupervised technology, for this particular project, I plan to first investigate methods that improves the performance of the standard DTW algorithm in supervised context. Assuming, I have managed to successfully investigate methods that improves DTW in the supervised context, the next step (assuming I have enough time) will be investigating and evaluating the performance of the adapted DTW in semi-supervised and unsupervised context.

3 Background

The Dynamic Time Warping algorithm utilises the technique of dynamic programming to find an optimum alignment between two sequences through the computation of local distances between the points in the temporal sequence. Formally given two time series sequences R and Q, DTW finds an optimum warping path by computing and storing local distances and using the stored distance to compute the overall cumulative distance. The figure below provides a good description of how the algorithm works in practice.

Figure 1: Dynamic Time Warping



4 Previous Work

In the unsupervised context, one of the pioneering work that has employed DTW in recognising recurrent patterns in speech has been conducted by Alex Park and James Glass in their paper [1] . In this paper, the duo makes a departure from traditional models in speech recognition that classify segments of speech into classified categories using the aid of lexicon dictionary and phoneme transcriptions to a dynamic programming methodology that detects motifs by exploiting the acoustic structure of the speech signal. Basically on each pair of un-transcribed utterances, the duo conducts a variant of the DTW algorithm known as segmental DTW to extract a set of different optimum alignments distinguished by different offsets but having the same temporal rigidity. The results are then fed to a clustering algorithm that groups segments, in different regions of time, that are acoustically similar to each other. Hence resulting the discovery of motifs.

4.1 Drawbacks

The performance of Park and Glass's DTW algorithm relies heavily on the values chosen three particular parameters: L - Length constraint minimum average that controls the size of the subsequences found between two utterances,θ- the threshold used for implementing edge weights in the clustering stage and R- the offset used to find multiple alignments between two temporal sequences[1,2].The choice of values for these parameters is highly domain dependent and this inherently results in the methodology being a supervised procedure since we need to tune the parameters according to the data. The DTW algorithm employed uses a global euclidean metric to compare the similarity between a reference and a test template. In comparison, in the context of supervised settings, the HMM frameworks employ a state specific local continuous probability distribution that computes the similarity between feature vectors and the frames of the reference template. In other words, the class labels of the frames on the reference template dictate the type of distance metric that

will be used in finding the optimum alignment. This in fact results in the HMM producing better performance in the supervised setting. Further drawbacks of the above setting is the explosion of search space for continuous speech recognition tasks and poor speaker independent performance. The acoustic data used by Park and Glass constitutes of single speaker data in a consistent acoustic environment. In a real world problem however, the environment tends to be highly inconsistent and the presence of different speakers leads to greater variation in acoustic signals belonging to the same motif.

4.2 Recent developement

Some of the above problems have been tackled in the paper "Template Based Continuous Speech recognition"[3]. The group here addresses the above issues by training a gaussian kernel for instances corresponding to each class. Class labels in this context correspond to phoneme labels. Thus all feature vectors belonging to the same phoneme are assigned a particular gaussian kernel. This kernel is then employed to compute the similarity between those templates and a query template. In their setting, each reference frame is equivalent to a phone state in an HMM modelled by a single gaussian. To reduce the search space, clustering is performed in the training stage and graphical search algorithms are employed at multiple passes to reduce the number of candidate templates that are fed to the DTW algorithm which then finds the optimum alignments between the reference templates and the query template. Unlike Park and Glass, the training data used by the group comes from the Resource Management benchmark corpus which contains audio transcribed signals corresponding to multiple speaker data captured in an inconsistent environment.

5 Methods

Having outlined the pros and cons of previous work, for my project I am planning to conduct the following steps:

5.1 Stage 1

The first step of this project will be to implement the standard Dynamic Time algorithm to identify recurring patterns in an acoustic signal. The performance of the algorithm will be tested on utterances corresponding to the TIMIT corpora that consists of 6300 instances from a relatively considerable number of speakers (630). The standard DTW will serve as the baseline algorithm against which the performances of all investigated adapted versions of the Dynamic Time Warping algorithm will be tested.

5.2 Stage 2

The next stage and probably the focal point of this project will be to investigate versions of the DTW that are application and domain dependent. As I have already mentioned, my

first aim will be to improve the performance of the algorithm in the supervised context. The motivation being if we manage to identify methodologies that can boost the accuracy of the DTW in the supervised context, the identified methodologies can serve as the starting point for developing new methodologies for improving the algorithm across semi-supervised and unsupervised problem domains. One of the methodologies that I am intending to use is as follows:

- Transform each time-series sequence of MFCC feature vectors to a sequence of vectors containing both global and local information. This normally corresponds to mapping a 13 -dimensional MFFC vector to a 39 -dimensional MFCC vectors that is augmented with derivative and second derivative information of adjacent frames. This ensures that when each point i.e an MFCC vector of a sequence is compared with another point in another sequence, its position in the sequence as well as its relation to its neighbours is taken into account.
- Cluster the training data by using information on the labels
- For each cluster, adjust the weights given to the local and global features of its MFCC features by using the *In class Weighting algorithm* [5] and then for each test data compute the weighted distance using the adapted DTW (tuned for each class) with samples and each class and hence identify the K nearest neighbours for each class.
- From the exacted $K * C$ (C denotes the number of classes) samples, Identify the closest K neighbours to the test data and use a majority vote to classify the test sequence. test sequence using a majority vote of its K-nearest neighbours.

Depending on the performance given by this adapted DTW algorithm,I intend to explore machine learning methodologies that might seem viable to this problem.

5.3 Stage 3

Assuming that I still have considerable time before the write up and final submission, for the last stage of the project, I will like to address the problem of the explosion of the search space probed by DTW during continuous speech recognition tasks. In comparison to HMM, the search space explored by DTW is much bigger hence resulting in the algorithm from causing the algorithm to suffer from high time and space complexity. For this last part, I plan to investigate procedures similar to bottom-up selection mentioned in [3] that employs an approximate K nearest neighbours methodology to suggest a list of templates with high enough probability to match the segment of input beforehand when running the DTW.

6 Evaluation

To compare the performance of proposed adaptations of the DTW against the base line algorithm, I plan to use the following three evaluation criterions:

- ROC graphs. An ROC graph depicts relative trade-offs between benefits (true positives) and costs (false positives). The y axis represents perfect classification. Since random guessing represents to a curve that is the main diagonal, an ideal classifier will have a ROC curve which close to the y axis as possible.

- Precision-Recall. The task that we are concerned with is spoken term/utterance discovery. Hence, in this problem domain it is appropriate to use an evaluation in the context of information retrieval where the goal is to retrieve same word pairs from word-pair imposters. The standard precision-recall curve is a standard evaluation metric where a high recall indicated that the DTW algorithm returned most of the relevant results while a high precision implies that algorithm returned substantially more relevant results than irrelevant.

- Purity score-The Purity score[1] is a measure of how accurately the clustered algorithm is able to group together similar acoustic nodes. More specifically this corresponds to the percentage of utterances that agree with the lexical identity of the cluster. The score given to each cluster is dependent not only on the performance of the DTW algorithm but also on the clustering algorithm employed to group recurring patterns.

Table 1: Work Plan

Milestone	Estimated time of Completion
Analyse the project and do literary review	March 25^{th}
Define the problem and write research proposal	April 1^{st}
Gather and explore the TIMIT data	June 10^{th}
Implement the baseline the DTW algorithm and evaluates it's performance using the 3 evaluation metrics	June 25^{th}
Explore and compare alternative adaptations of the DTW algorithm in the supervised context against the baseline DTW	July 25^{th}
Gather results and finish writing up	August 5^{th}

.

References

[1] Alex S.Park and James Glass "Unsupervised Pattern Discovery in Speech", 2008

[2] Alex S.Park and James Glass "Toward Multi-Speaker Unsupervised Speech Pattern Discovery", 2010

[3] Mathias De Wachter,Mike Matton, Kris Demuynck and Patrick Wambacq "Template Based Continuous Speech Recognition", 2004

[4] Mathias De Wachter,Mike Matton, Kris Demuynck and Patrick Wambacq "A Locally Weighted Distance Measure For Example Based Speech Recognition", 2004

[5] Ying Xie and Bryan Wiltgen " Adaptive Feature Based Dynamic Time Warping", 2010

[6] Michael A.Carlin, Samuel Thomas, Aren Jansen, Hynek Hermansky " Rapid Evaluation of Speech Representations for Spoken Term Discovery",2011