Polynomial Dynamic Time Warping Kernel Support Vector Machines for Dysarthric Speech Recognition with Sparse Training Data

Vincent Wan, James Carmichael

Department of Computer Science
University of Sheffield, Sheffield, UK
{v.wan, j.carmichael}@dcs.shef.ac.uk

Abstract

This paper describes a new formulation of a polynomial sequence kernel based on dynamic time warping (DTW) for support vector machine (SVM) classification of isolated words given very sparse training data. The words are uttered by dysarthric speakers who suffer from debilitating neurological conditions that make the collection of speech samples a time-consuming and low-yield process. Data for building dysarthric speech recognition engines are therefore limited. Simulations show that the SVM based approach is significantly better than standard DTW and hidden Markov model (HMM) approaches when given sparse training data. In conditions where the models were constructed from three examples of each word, the SVM approach recorded a 45% lower error rate (relative) than the DTW approach and a 35% lower error rate than the HMM approach.

1. Introduction

Hidden Markov model (HMM)-based speech recognition systems require a reasonable quantity of training data in order to perform well in real conditions. It is not always possible, however, to collect sufficient data to achieve the desired accuracy. Classification of dysarthric speech is an example of such a task where training corpora are usually severely limited.

Debilitating neurological conditions which often accompany dysarthria can make the collection of speech samples from patients a time-consuming and low-yield process [1]. It is often not feasible, therefore, to provide ample corpora for training statistical speech recognition engines. Off-the-shelf HMM speech recognition systems typically employ pretrained models that are, to some extent, speaker-adaptable. Such systems, however, usually record unacceptably high error rates [2] when decoding speech resulting from certain types of dysarthria, particularly the progressive varieties where the patient’s pronunciation becomes increasingly dissimilar from the population norm as articulator control worsens. The usual solution to accommodate variability — training on larger corpora — is not applicable in this instance.

This paper presents an isolated word recognition system based on the support vector machine (SVM) with a polynomial dynamic time warping (DTW) kernel (DTW-SVM) for the classification of dysarthric speech. This work is an advancement on the efforts of Green et al. [1]. It has been shown that support vector machines (SVMs) [3] generalise well on sparse training data. Unlike some other machine learning algorithms, including Gaussian mixture models, they can be made to classify variable length sequences directly and without the aid of a finite state machine (such as HMMs) to capture the temporal information. In this paper, sequence classification is achieved via the well known DTW algorithm, described in section 2. That section also discusses the concept of sequence kernels and introduces the notation used.

The formulation of the novel polynomial DTW kernel is outlined in section 3. The name comparison of the DTW-SVM to a standard DTW system and to a statistical HMM system are described in section 4 and the results presented in 5. Section 6 concludes the paper.

2. Background

2.1. Dynamic Time Warping

In symmetric DTW [4] a global distance \( D(X, Y) \) between two sequences of vectors \( X = \{x_1, \ldots, x_N\} \) and \( Y = \{y_1, \ldots, y_N\} \) is determined. In speech recognition tasks, \( x \) and \( y \) are vectors of cepstral coefficients and \( N \) is the number of vectors in each sequence. Let the local distance be the Euclidean distance between two cepstral vectors, \( d(x_i, y_j) = ||x_i - y_j|| \). Since the two sequences may not be of equal length, an alignment must be defined so that each vector in \( X \) is matched with a corresponding vector in \( Y \). This alignment is denoted by \( \phi_x(k) \) and \( \phi_y(k) \) for \( k = 1, \ldots, K \), where \( K \) is the length of the alignment. Using this notation, \( x_{\phi_x(k)} \) is matched with \( y_{\phi_y(k)} \). For the purpose of this work, the alignment maintains the temporal sequence within \( X \) and \( Y \) and no vectors may be omitted. The DTW alignment distance is determined for a given alignment by summing the local distances between matched vectors,

\[
D_{\text{alignment}}(X, Y) = \frac{1}{K} \sum_{k=1}^{K} d(x_{\phi_x(k)}, y_{\phi_y(k)})
\]

The global distance is the alignment distance along the Viterbi path, which is defined by

\[
D(X, Y) = \min D_{\text{alignment}}(X, Y).
\]

A dynamic programming beam search is used to compute the global distance efficiently.

2.2. Support Vector Machines

A support vector machine, as described in [5], is a general data classification technique for binary decision problems. The objective is to define a linear boundary that is “midway” between...
two given data groups: the concept is known as margin maximisation. The data points closest to the decision boundary are known as the support vectors since they alone among all the data points define the hyperplane’s exact position. The maximum margin criterion makes the SVM particularly suitable for recognition tasks with limited training data.

Although the SVM is based upon a linear discriminator, it is not restricted to making linear hypotheses. Non-linear decisions are made possible by a non-linear mapping of the data to a higher dimensional space. The phenomenon is analogous to folding a flat sheet of paper into any three-dimensional shape and then cutting it into two halves, the resultant non-linear boundary in the two-dimensional space is revealed by unfolding the pieces. The SVM’s non-parametric mathematical formulation allows these transformations to be applied efficiently and implicitly: the SVM’s objective is a function of the dot product between pairs of vectors; the substitution of the original dot products with those computed in another space eliminates the need to transform the original data points explicitly to the higher space. The computation of dot products between vectors without explicitly mapping to another space is performed by a kernel function, examples of such include the radial basis function kernel,

\[ K_{RBF}(x, y) = \exp \left( -\frac{1}{2\sigma^2} (x - y)^2 \right) \] (3)

and the polynomial kernel,

\[ K_{poly}(x, y) = (x \cdot y)^n \] (4)

where \( \sigma \) is the width of the radial basis functions and \( n \) is the order of the polynomial.

2.3. Sequence kernels

A sequence kernel is a function that operates on two complete sequences. They enable SVMs to perform discriminative classification of sequences of unequal length directly at the sequence-level. This contrasts with HMM-based techniques where sequences are divided into frames and discrimination is then applied at the frame-level, potentially resulting in the loss of information useful for sequence-level classification. Examples of sequence kernels include the Fisher kernel [6], the pair HMM kernel [7] and the generalised linear discriminant kernel [8].

The well known DTW approach provides a way of computing a ‘distance’ between two sequences that can be exploited to derive sequence kernels. In the approach by Shimodaira et al. [9], the local distance in the DTW algorithm was substituted with a similarity measure defined by the right-hand side of (3) and the global distance redefined to maximise the similarity. However, the resulting alignment path found by this method suffers significantly from that obtained when using the Euclidean local distance metric, a factor which can negatively influence recognition accuracy.

An alternative approach by Bahlmann et al. [10] kept the Euclidean local distance and replaced the \((x - y)\) term in (3) with the DTW global distance thereby creating a ‘radial basis function DTW kernel’.

3. Polynomial DTW kernel

The method employed herein for converting the DTW global distance to a dot product uses a technique called spherical normalisation [11, 12] and the polynomial kernel (4). Figure 1 illustrates the process of derivation. Let \( H \) be the space in which the sampled vectors \( x \) and \( y \) normally exist, represented in the figure as a straight line. Let \( S \) be a unit hypersphere embedded in a space that has one dimension more than \( H \). Consider two points in \( H \) represented by vectors \( x \) and \( y \). The spherical normalisation procedure projects each point along the dotted lines onto the surface of \( S \). Each projected point is represented by unit vectors \( \hat{x} \) and \( \hat{y} \) that originate from the centre of \( S \). The formula for the transformation from \( x \) to \( \hat{x} \) is

\[ \hat{x} = \frac{1}{\sqrt{x^2 + \alpha^2}} x \] (5)

where \( \alpha \) is the perpendicular distance from the centre of \( S \) to \( H \). Since the transformation is reversible, there is no loss of information. A detailed description of the properties of spherical normalisation can be found in [11].

Let \( X = \{x_1, \ldots, x_N\} \) denote the sequence \( X \) after it is mapped onto \( S \) and let \( Y \) be similarly defined. It is in this new space that distances can be converted readily to dot products. By definition, the length, \( d_S(\hat{x}, \hat{y}) \), of the geodesic on a unit hypersphere between two points represented by vectors \( \hat{x} \) and \( \hat{y} \), equals the angle between the two vectors, \( \theta_{\hat{x}, \hat{y}} \). Instead of a Euclidean distance, therefore, let the local DTW distance be the shortest distance between two points on the surface of the hypersphere,

\[ d_S(\hat{x}, \hat{y}) = \theta_{\hat{x}, \hat{y}} \] (6)

Using the relation \( \hat{x} \cdot \hat{y} = \cos \theta_{\hat{x}, \hat{y}} \), the new local DTW distance may be expressed in terms of a dot product,

\[ d_S(\hat{x}, \hat{y}) = \arccos(\hat{x} \cdot \hat{y}) \] (7)

The same principle can be used to reconvert the DTW global distance measured in \( S \) to a dot product,

\[ K_{\text{linear DTW}}(\hat{X}, \hat{Y}) = \cos D_S(\hat{X}, \hat{Y}) \] (8)

where \( D_S \) is the global distance computed using the local distance defined by (7).

It is important to note that (8) can only be a valid kernel if the symmetric form of DTW is used, i.e. \( D_S(\hat{X}, \hat{Y}) = D_S(\hat{Y}, \hat{X}) \). In addition to symmetry, the kernel function must satisfy Mercer’s condition [13] or be proven by alternative means to be a valid dot product. By definition, the cosine of the angle between two unit vectors is guaranteed to yield a dot product. From equations (1) and (6), the global distance is the mean of the angles between the pairs of vectors from the two sequences. However, it may be disputed whether the cosine of such a mean is still a dot product. Our simulations have not
encountered any of the expected convergence anomalies if the kernel were invalid.

Assuming that equation (8) is a valid dot product, it can be substituted into equation (4), replacing the term \((x \cdot y)\) to obtain higher order non-linear solutions. Thus the polynomial DTW kernel takes the form,

\[
K_{\text{poly DTW}}(\hat{x}, \hat{y}) = \cos^n D_{\text{g}}(\hat{x}, \hat{y})
\]  

(9)

The procedure for computing the polynomial DTW kernel is summarised as follows:

1. Map the vectors of each sequence to a hypersphere using equation (5).
2. Perform DTW to get the global distance as defined by equation (2) using the local distance defined by equation (7).
3. Take the cosine of the global distance and raise to the \(n\)th power according to equation (9).

4. Experiments

Isolated word speech recognition simulations were performed using speech collected from both normal and dysarthric speakers [1]. This is necessarily a speaker dependent task because dysarthric speech can differ enormously between speakers depending upon the severity of their disability. There were 7 dysarthrics and 7 non-dysarthrics, each uttering between 30 and 40 examples of each of the words: alarm, lamp, radio, TV, channel, volume, up, down, on and off.

Given the limited quantity of data, \(N\)-fold cross-validation was used to increase the number of tests performed so that statistically significant results could be obtained. Cross-validation was performed using training set sizes ranging from 3 to 15 examples per word as follows: for each specified training set size, that number of utterances were randomly selected for model building and the remaining utterances were used for testing; this process was repeated until all of the utterances were selected once for training. This yielded over 17000 individual tests on dysarthric speech when training sets were at the 3-sample minimum and approximately 1700 when training corpus size was maximised to 15 utterances per words.

The speech was parameterised using 13 Mel-frequency cepstral coefficients (MFCCs) computed using a 32 millisecond window with a 10 millisecond frame shift, augmented with their first order derivatives and normalised to zero mean and unit diagonal covariance resulting in vectors of 26 dimensions. A Gaussian mixture silence detector, employing smoothing techniques, was used to remove all frames representing silence or containing short bursts of non-speech noise.

Three word modelling techniques were compared. The first was a standard DTW system. It used all the utterances in the training selection as its templates and classified a test utterance according to the single closest matching template. The training utterances were manually end-pointed to contain only the relevant speech. During testing, this system required the use of a detector to remove frames of silence.

The second technique was the HMM system, Stardust\(^1\), implemented in [15], employing a silence-word-silence grammar, using a continuous density ‘straight-through’ model topology.

\(^1\)This system was built around the HTK toolkit [14] using a multi-pass re-estimation strategy that was bootstrapped from manually end-pointed speech.

The third was the polynomial DTW-SVM described in section 3. SVMs are binary classifiers so the directed acyclic graphs approach [16] was used to achieve multi-class classification. That approach considers the classes in pairs: an SVM trained to discriminate between those two classes is used to eliminate one of them. The process is repeated until one class remains. Like the standard DTW system, the training utterances were manually end-pointed and the test utterances were stripped of silence frames using an automatic detector.

5. Results

Trials on the DTW-SVM system showed that good accuracy could be achieved using a third order polynomial, where \(n = 3\) in (9) and \(\alpha = 1\) in (5).

In these simulations, the SVMs were set up to classify separable data (i.e., the regularisation parameter, or upper bound on the Lagrange multipliers, was set to infinity). The average Lagrange multiplier value was of the order 0.1 with maximum values in the range 1 to 5. Due to the paucity of training data, the SVM solutions – save a few exceptions – used all of the training data for support vectors.

A summary of the results is shown in figure 2. The standard DTW system returned the highest error rate across all conditions. Results on normal speech showed no significant difference between the DTW-SVM and HMM techniques except when training sets were reduced to 3 examples per word; when so constrained, the SVMs reduced the error rate to 4.0% compared to the HMM’s 7.4%, a 46% relative reduction.

On dysarthric speech, when trained on 3 examples per word, the DTW-SVM system had a 19.7% error rate and the HMM 30.1%, a relative reduction of 35%. When training corpus size ranged from 6 to 15 samples per model, the SVM continued to outperform the HMM, with the difference in respective

![Figure 2: The error rates of DTW-SVMs, HMMs and standard DTW systems on normal and dysarthric speakers as a function of the size of the training set.](image-url)
error rates remaining roughly constant at about 2% absolute.

6. Conclusion
A novel polynomial DTW kernel SVM was presented in this paper. The new kernel was obtained by converting the DTW distance measure between two sequences into a kernel function appropriate for an SVM.

Simulations compared the DTW-SVM with standard DTW and HMM techniques. The new approach was consistently more accurate than both standard DTW and HMM systems when presented with the sparse training data from dysarthric speakers. In tests on normal speech, there was no observable difference in accuracy between the HMM and DTW-SVM techniques except in one condition when the training data was most sparse. In that single condition, the DTW-SVM yielded its greatest relative (46%) reduction in the error rate compared to the HMM.

This improvement is of more than just academic interest for this research team: the next phase of investigation will involve the production of a speech-activated environmental control system (SPECS) which, unlike its predecessor, will progress beyond proof-of-concept stage and become a commercially available product. The DTW-SVM’s performance qualifies this technology as a candidate for the SPECS speech recognition engine, a possibility which would represent a refreshing alternative to HMM statistical modelling which seems to predominate the research effort in this domain.

7. Acknowledgements
The authours would like to thank Roger Moore for his guidance and contributions.

8. References