Dynamic Time Warping with Path Control and Non-local Cost

Yoram Stettiner[1], David Malah[2] and Dan Chazan[3]

[1,2] Dept. of Electrical Engineering, Technion-Israel Institute for Technology, Haifa 32000 , Israel

[1] Also with Nexus Telecommunication Systems Ltd., 6 Tfutzot Israel, Givatayim 53583, Israel

[3] IBM Science and Technology center, Matam, Haifa, Israel

ABSTRACT

Dynamic Time Warping (DTW) is a Dynamic Programming technique widely used for solving time-alignment problems. The classical DTW constrains only the first derivative of the warping function, hence allowing no direct control over the warping function curvature. Moreover, it implicitly assumes- inappropriately for some applications- that the noise is white. We propose a multi-dimensional Dynamic-Programming technique which can efficiently solve time-warping optimization problems involving colored noise, and allows control over the warping function curvature. The technique is demonstrated for the co-channel speech separation problem. Applications employing DTW can benefit from the new technique, which offers improved accuracy and robustness in the presence of colored noise and competing speech.

1: Introduction

Dynamic Time Warping (DTW) is a Dynamic Programming (DP) technique widely used for solving time alignment problems in diverse speech processing applications [4]. Recently, DTW was used in estimating the parameters of a long-term model for voiced speech, aimed primarily at the co-channel speech separation problem [1][2]. According to the model, each voiced phoneme is the outcome of two transformations applied to a strictly periodic signal, plus additive Gaussian noise. The first transformation is a non-linear time-warp, resulting in a quasi-periodic signal. The second transformation is a linear time-varying filtering operation. The derivative of the time-warping function represents the instantaneous pitch.

As part of a ML estimation procedure proposed in [1], it is required to find a warping function that optimizes a quadratic cost function. Unfortunately, the cost function value at any given time depends not only on the warping function value at that time, but also on near ***past and future*** values. Consequently, the cost function is not local (according to the classical DTW terminology), hence the conventional DTW technique does not apply. Nevertheless, lacking a better alternative at the time, a conventional DTW [4] was used in [1] as a sub-optimal solution.

When applying the long-term model to the co-channel speech separation problem, the model parameters of both speakers - including the warping functions - need to be estimated simultaneously. For robust estimation, the warping functions must be tightly constrained, so that each can track the pitch variations of one, and only one, speaker.

The classical DTW technique constrains only the first derivative of the warping function, hence allowing no direct control over the warping function curvature. One possible improvement is to construct a cost function that depends also on the low-order derivatives of the warping function. Such a cost function could provide path control by penalizing deviations of the warping function from a smooth curve, and can be regarded as a "soft" constraint.

We will show that in cases where the non-local cost function has a bounded support, and the "soft" constraint involves a small number of low-order derivatives, the above two problems may be jointly formulated in a such a way that the Optimality Criterion is still met and Dynamic Programming is applicable. This is done by augmenting the dimension of the DTW formulation. Once the multi-dimensional warping function is estimated, it is projected onto the original dimension to provide the sought after warping function. The number of augmented dimensions is equal to the maximum between the length of the non-local cost function support, and the number of derivatives used for the "soft" constraint.

Applications where DTW is currently used may benefit from the new technique, which allows improved accuracy and robustness, in the presence of colored additive noise or competing speech.

The paper is organized as follows: Section 2 reviews the conventional DTW technique. Section 3 defines the time-warping problem with non-local cost functions. Section 4 discusses path control by non-local soft constraints. Section 5 describes the proposed multi-dimensional DTW (MD-DTW) solution. In section 6, the application to the co-channel speech separation

1051-4651/94 $04.00 © 1994 IEEE

problem is outlined, , and section 7 concludes the paper.

2: Principle of optimality and DTW

In the context of DTW problems [4], Bellman's Principle of Optimality [3] implies that given an optimal path ϕ from A to B, and a point C lying somewhere on ϕ, the path segments AC and CB constitute optimal paths from A to C and from C to B, respectively. See [4] for various DTW schemes. A DTW problem is solved using a grid. A *local cost* is associated with each grid point (node) and/or with each transition from one node to another. The optimal path is constrained by the *local path constraint*. The optimal path is found by accumulating the local costs, finding the optimal endpoint, and finally back-tracking along the optimal path. We will next consider a time-warping problem involving non-local costs and path constraints.

3: Non-local cost

Consider the following cost function

$$J_1 = \int_{t_0}^{t_f} \left| \int_{-\infty}^{\infty} r(t-\tau)\big(x(\tau) - s(\phi(\tau))\big) d\tau \right|^2 dt \qquad (1)$$

where $x(t)$ and $s(t)$ are given signals, $r(t)$ is a given kernel, and ϕ is a time warping function satisfying some constraints. Clearly the integrand - named a Lagrangian in the terminology of optimal control - is a function of t alone if $r(t) = \delta(t)$, hence the Optimality Principle applies. Otherwise, the Optimality Principle no longer applies since the Lagrangian at time t depends also on some past and perhaps future values of $\phi(t)$. Consequently, an optimal decision on the present value of $\phi(t)$ is impossible since this decision also affects future values of the Lagrangian. However, assuming $r(t)$ is causal and of finite duration d, the integration limits of the inner integral become 0 to d.

Cost functions of the form (1) can be encountered in diverse fields. In the context of the co-channel speech separation problem, we wish to estimate a time warping function ϕ in the ML sense. Assuming a Gaussian noise, not necessarily white, one arrives at the weighted least squares (WLS) problem stated in (1), where the kernel $r(t)$ is the impulse response of a noise whitening filter. Rewriting (1) in discrete-time notation, one obtains

$$J_2 = \big(x - s(\phi)\big)^T Q \big(x - s(\phi)\big) \qquad (2)$$

where x, s and ϕ denote the discrete-time versions of the signals x , s and ϕ, respectively, and $Q = R^T R$, where R is a convolution matrix such that Rx is the discrete equivalent of $r(t)*x(t)$. Clearly, Q is a symmetric positive semi-definite matrix, and is actually the inverse of the covariance matrix of the Gaussian noise. Furthermore, assuming as before that $r(t)$ decays fast, Q becomes a diagonally-dominant sparse matrix.

4: Path control

In DTW, the local slope of the optimal path is constrained by the *local path constraint*. Yet in the co-channel speech separation problem-as mentioned in the introduction-it is beneficial to control some higher derivatives (e.g. curvature) of the optimal path as well. To this end, we add to the cost function (1) the following term:

$$J_3(\phi) = \int_{t_0}^{t_f} \sum_{p=1}^{P} \left| b_p\big(\phi^{(p)}(t)\big) \right|^2 dt \qquad (3)$$

where $b_n(\cdot)$ are cost functions of the corresponding n-th order derivative $\phi^{(n)}$ of ϕ. Combining (2) and (3) we write in discrete-time notation

$$J_4 = \big(x - s(\phi)\big)^T Q \big(x - s(\phi)\big) + \sum_{p=1}^{P} \big[b_p(D_p\phi)\big]^T \big[b_p(D_p\phi)\big] \qquad (4)$$

where D_p is the matrix representing the p-th order discrete differentiation operator. For small enough values of p, D_p is also a band-diagonal matrix.

5: Multi-dimensional DTW (MD-DTW)

We wish to solve the time-warping problem where both the cost and the constraints are non-local, i.e., we wish to minimize J_4 with respect to the warping function ϕ. To that end, we propose a multi-dimensional DTW (MD-DTW) algorithm with a dimension equal to one plus the largest expected number of non-zero sub-diagonals in either Q or D_p.

Since the simple local constraints commonly used in DTW schemes [4] are too coarse for our application, we upsample the reference vector by a factor M, and look for the deviation of the warping function from a pure time delay. Consequently, the DP problem is solved for a new function $\eta(t)$ related to the original warping function $\phi(t)$ by $\eta(t) = M \cdot (\phi(t) - t)$. Hence, using discrete notation and a discrete grid, a pure delay corresponds to a straight horizontal line.

To keep the dimension of the problem small, we assume: a. the non-local cost kernel Q has only P_2

non-zero sub-diagonals, and b. the non-local constraints involve only the first P_1 derivatives of $\eta^{(n)}$. Hence $P=max(P_1,P_2+1)$ is the problem's dimension. We further assume the existence of hard constraints - each of the first P_1 discrete derivatives may take its values only from a corresponding finite set of integers.

From the standpoint of discrete optimal control, the state vector consists of the warping function and its discrete derivatives of orders 1 through (P-1), while the P-th derivative serves as the control variable. The cost function is given in (4).

To demonstrate the MD-DTW, let us consider the simple the case where Q is diagonal, and only the first two derivatives are of interest, i.e., $P_1=2$, $P_2=1$. The dimension of the problem is $P=max(P_1,P_2)=2$. Let the first and second derivatives assume values from the sets $Z_1=\{2, 1, 0, -1, -2\}$, and $Z_2=\{1, 0, -1\}$, respectively. Fig. 1 below depicts all the transitions that constitute feasible (i.e., with finite cost), left-to-right, length 2, path segments, ending at some specific node E. The numbers are the ordinates of the shown grid nodes. Likewise, Fig. 2 below shows all 13 feasible path segments of length 2 leading to node E. In fact, Fig. 1 may be viewed as a projection of Fig. 2 on the (t, η) plane.

Referring to fig. 2, we note that a final decision on the optimal path leading to node E cannot be taken at t=2, since this would explicitly select some path segment from t=1 to t=2, which shall in turn affect the cost of arriving at future nodes at t=3 (not shown in the figures). There are 5 possibilities to go from t=1 to t=2. Since the final decision must be delayed until t=3, the solution is to compute 5 hypothesized accumulated costs at node E, each conditioned on one of the 5 possibilities. By induction, all nodes in the grid must have these 5 conditional accumulated costs.

Let us denote each node by its t and η time coordinates. Node E is thus denoted by the ordered pair (2,0). Since only 3 values of the second order derivative are allowed, we need to consider only 3 possibilities for the nodes {(1,-1), (1,0), (1,1)}. Furthermore, because the first derivative cannot exceed 2 in absolute value, there are only 2 possible preceding nodes for the nodes {(1,-2), (1,2)}. Each node (t,η) has 5 conditional costs, denoted $C_n(t,\eta)$, n=1,...5, associated with it. Likewise, let us denote a path segment by listing all the nodes it travels through. Accordingly, all path segments originating from nodes at t=0 and leading to node E (Fig. 2) are denoted by groups of the form {(0,i),(1,j),(2,0)}, where i, j and 0 are the η coordinates at t=0 (origin), t=1 (intermediate) and t=2 (destination), respectively. We need to decide, for each hypothesized intermediate node (1,j) j=1,...,5, on the optimal source node (0,i) $i\in\{j-1,j,j+1\}$ of a path segment going through (1,j) and arriving at the destination (2,0). To that end, we need to minimize

$$C_j(2,0)=\underset{i}{Min}\left\{C_{i-j}(1,j)+L\left((0,i),(1,j),(2,0)\right)\right\} \quad (5)$$

with respect to i. C_{i-j}(1,j) is the accumulated cost at node (1,j) assuming that the optimal path previously passed through (0,i), and $L(\{(0,i), (1,j), (2,0)\})$ is the incremental cost (i.e., the Lagrangian value) due to the path segment {(0,i), (1,j), (2,0)}. Eq. (5) is the basic recursion of the MD- DTW for this simple case.

Unlike conventional DTW, here the accumulated cost function is multi-valued, but DP may still be employed.

Consider next the same example from an optimal control standpoint. Let the state variables be η and $\dot{\eta}$, or η_1 and η_2 in discrete notation. The discrete state equations are

$$\begin{aligned}\eta_1(n+1)&=\eta_1(n)+\eta_2(n)\\ \eta_2(n+1)&=\eta_2(n)+u(n)\qquad u(n)\in\{-1,0,1\}\end{aligned} \quad (6)$$

where $u(n)$ is the control variable. Fig. 3 below depicts a 3 dimensional grid. The horizontal coordinate is time, the vertical coordinate is η_1 - the warping function - and the depth axis is η_2 - the first derivative of the warping function. This constitutes a two dimensional DP problem. In our example, η_2 can take only 5 possible values. Each of the nodes of Fig. 1 described above, with its 5 conditional costs, is replaced by 5 separate nodes, spread along the $\dot{\eta}$ axis. Each such node corresponds to one of the 5 possible values of η_2, and takes the corresponding conditional cost described above. The MD-DTW problem can now be solved as an optimal control problem by a two-dimensional DP procedure. The local constraint of conventional DTW is replaced by the dynamics of the state equations and the allowed range of control. Fig. 3 depicts an example of an optimal path (middle curve). Its projections on the (t,η) plane, the $(t,\dot{\eta})$ plane and the $(\eta,\dot{\eta})$ plane, are the optimal warping function, its derivative, and the state-space trajectory, respectively.

The above procedure is readily extendible to problems having either cost functions with larger supports, or higher order constraints. This is done simply by augmenting dimensions and conditional costs to the state-space formulation, solving the multi-dimensional DP problem, and projecting the optimal path on the (t, η) plane. The same is true for adding higher order constraints.

Although the computational complexity of the above technique remains O(N) - N being the utterance length in samples - as in conventional DTW techniques, the

number of operations is also proportional to an exponential function of the dimension P. Hence the technique is practically limited to small values of P.

6: Application to the co-channel speech separation problem

In the co-channel speech separation problem, the goal is to decompose the co-channel speech signal into several signal components, one for each speaker, mainly for intelligibility improvement. Recently, a long-term model for voiced speech, was applied to the co-channel speech separation problem [1][2], where as part of the estimation procedure , it is required to find a warping function that optimizes a quadratic cost function of the form (2), where Q is a band-diagonal matrix. In addition, each warping function must be tightly constrained so that each can track the pitch variations of one, and only one, speaker. The MD-DTW was incorporated into an EM scheme [1][2] for co-channel speech separation, where it was required to minimize (4) with respect to the warping function ϕ. Due primarily to its tight higher-order constraints, it managed to robustly estimate the warping functions in most of the cases where the previously installed conventional DTW scheme failed.

7: Discussion and summary

We propose a multi-dimensional DTW technique which can efficiently solve optimizations problems of the form (2) or (4), involving non-local cost and constraints. Applications where DTW is currently used may benefit from the new technique, which allows a more accurate and robust time-warping in the presence of colored additive noise or competing speech.

References

[1] Stettiner Y., Malah D. and Chazan D., "Estimation of the parameters of a long term model for accurate representation of voiced speech", Proc. IEEE Conf. Acoustics Speech and Signal Processing, ICASSP-93, pp. 534-537, 1993

[2] Chazan D., Stettiner Y. and Malah D., "Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation", Proc. IEEE Conf. Acoustics Speech and Signal Processing, ICASSP-93, vol 2, pp. 728-731, 1993

[3] Bellman R. and Dreyfus S., "Applied dynamic programming", Princeton University Press, Princeton, NJ, 1962

[4] Deller, Proakis and Hansen, "Discrete-time processing of speech signals", Chap. 11, MacMillan, 1993

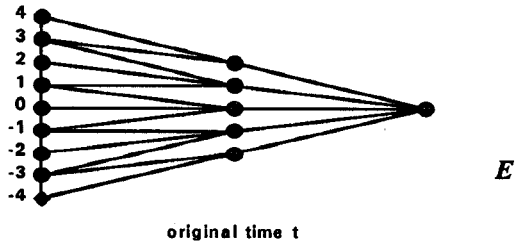

Figure 1 - All feasible path segments leading (left to right) to node E (rightmost node). The horizontal axis is the original time axis t. The vertical axis is $\eta(t) = M(\Phi(t) - t)$ The numbers on the vertical axis are the ordinates of the shown grid nodes.

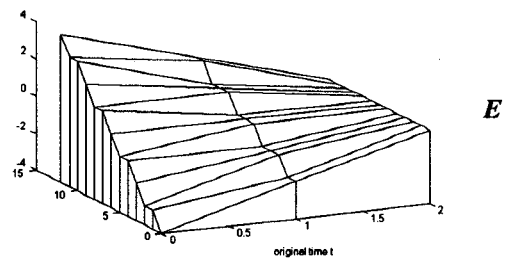

Figure 2 - All 13 feasible path segments leading (left to right) to node E are shown stacked one after the other. The horizontal axis is the time axis t. The vertical axis is $\eta(t) = M(\Phi(t) - t)$. The numbers on the vertical axis are the ordinates of the shown grid nodes. The depth axis enumerates the feasible path segments.

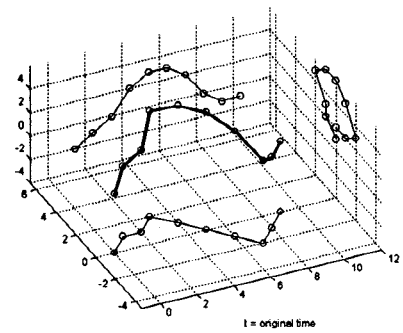

Figure 3 - MD-DTW example. The optimal path is the middle curve. Its projection on the (t, η) plane is the optimal warping function. The projection on the $(t, \dot{\eta})$ plane is the 1st order difference of the optimal warping function. The projection on the $(\eta, \dot{\eta})$ plane is the state-space trajectory.