Recognition of Isolated Musical Patterns using Context Dependent Dynamic Time Warping

Aggelos Pikrakis[1] , *Sergios Theodoridis*[1], *Dimitris Kamarotos*[2]
[1]Dept. of Informatics and Telecommunications, University of Athens
Panepistimioupolis, Ilisia 15784, Athens, Greece
[2]IPSA Institute of the Aristotle University of Thessaloniki
e-mail: pikrakis@di.uoa.gr, stheodor@di.uoa.gr, dimik@otenet.gr

ABSTRACT

This paper presents an efficient method for recognizing isolated musical patterns in a monophonic environment, using a novel extension of Dynamic Time Warping, which we call Context Dependent Dynamic Time Warping. Each pattern is converted into a sequence of frequency jumps by means of a fundamental frequency tracking algorithm, followed by a quantizer. The resulting sequence of frequency jumps is presented to the input of the recognizer which employs Context Dependent Dynamic Time Warping. The main characteristic of Context Dependent Dynamic Time Warping is that it exploits the correlation exhibited among adjacent frequency jumps of the feature sequence. The methodology has been tested in the context of Greek Traditional Music, which exhibits certain characteristics that make the classification task harder, when compared with Western musical tradition. A recognition rate higher than 95% was achieved.

1 INTRODUCTION

This paper proposes a scheme for the recognition of predefined musical patterns in a monophonic environment in the context of Greek Traditional Music. The patterns to be recognized have been isolated from their context by means of a manual segmentation process, thus the term "isolated musical patterns". The term monophonic refers to a *single non-polyphonic instrument, the clarinet, recorded under laboratory conditions with an ambient noise of less than* $5dB$.

In the first stage of the recognition scheme a feature generation algorithm converts the unknown musical pattern into a sequence of frequency jumps (multiples of one quarter-tone). At the heart of this stage lies a fundamental frequency tracking algorithm.

In the second stage, Context Dependent Dynamic Time Warping (CDDTW) is employed in order to match the previously extracted feature sequence to a set of twelve reference sequences (one reference sequence per musical type). The unknown pattern is determined based on the lowest matching cost. We propose CDDTW as a novel extension of the standard Dynamic Time Warping (DTW) methodology [1], [2], [3], [4], [5], [6]. Standard DTW schemes assume that each feature in the resulting sequence is uncorrelated with its neighboring ones (i.e., its context). In contrast, CDDTW permits flexible grouping of neighboring features (i.e., forming feature segments) in order to exploit possible underlying mutual dependencies.

To our knowledge, this is the first time that the CDDTW methodology is proposed and applied to the recognition of isolated musical patterns in the time domain. Previous work by the authors employed standard DTW schemes [7] which present certain limitations. A standard DTW scheme was also used in [8], but for signals available in MIDI format, which has severe limitations for the majority of real world signals and in particular for the case of Greek Traditional musical patterns. Most previously published literature related to music sound recognition has focused on MIDI representation, e.g. [9], [10].

The reported research focuses on Greek Traditional music. The musical system of Greek Traditional music and the techniques of instrument players give the resulting sound material a radically different structure when compared with that of the western equal-tempered intervalic system (system of musical scales).

Section 2 presents the aforementioned feature generation procedure. Section 3 presents CDDTW and Section 4 gives details of the application of our method to patterns from the Greek Traditional music. Conclusions and future work are presented in Section 5.

2 FEATURE GENERATION

At first, a sequence of fundamental frequencies are extracted from the musical pattern to be recognized. We experimented with the following frequency domain and time domain methods:

a) Frequency-domain approaches: Schroeder's histogram [11], Schroeder's Harmonic Product Spectrum [11], Piszscalski's method [12] and Brown's pattern recognition method based on the properties of a constant-Q transform [13].

b) Time-domain methods: Cooper and Kia's method [14] and Brown's narrowed autocorrelation method [15]. Also Tolonen's method was used [16]. In addition, we developed a new frequency-domain algorithm [7, 17] that can be considered as a modification of Schroeder's histogram.

After extensive experimentation with all the above algorithms, we concluded that Brown's narrowed autocorrelation method [15] and Tolonen's multipitch analysis model [16] gave the best results with respect to accuracy and frequency doubling, provided all required parameters were rightly tuned. The new algorithm that we employed is more efficient from a computational point of view, yet fails to cope with the problem of absent fundamentals. Since this is not a crucial issue for the signals of our study this new algorithm can be used alternatively. However, any fundamental frequency tracking algorithm can be used.

Let $\mathbf{f} = \{f_i, i = 1 \ldots M\}$, be the generated sequence of fundamental frequencies corresponding to M successive frames of a pattern. At first, each f_i is mapped to a positive number, say k, equal to the distance (measured in quarter-tone units) of f_i from f_s, where f_s is the lowest frequency that the instrument used in the experiments can produce (for the signals that we studied $f_s = 146.8Hz$). Therefore,

$$k = round(24 \log_2 \frac{f_i}{f_s})$$

where $round(.)$ denotes the roundoff operation. As a result, the sequence of frequencies is mapped to a sequence of positive numbers, $\mathbf{L} = \{l_i, i = 1 \ldots M\}$. The goal of this step is to imitate some aspects of the human auditory system, which is known to analyse an input pattern using a logarithmic frequency axis.

In order to deal with the fact that instances of the same musical type may have different starting frequencies, (i.e. may appear at different frequency bands), a sequence of frequency jumps is extracted from the symbol sequence $\mathbf{L}$. This is achieved by calculating the difference $\mathbf{D}$ of $\mathbf{L}$, i.e.,

$$\mathbf{D} = \{d_{i-1} = l_i - l_{i-1}, i = 2 \ldots M\}$$

Since most of the time, l_i is equal to l_{i-1}, $d_i = 0$ for most of the frames (i's). $\mathbf{D}$ turns out to be a sequence of frequency jumps falling in the range of $-G$ to G, where G corresponds to the maximum allowed frequency jump ($G = 60$ quarter-tones, i.e., 15 tones for the signals that we studied).

3 Context Dependent Dynamic Time Warping

In the sequel, the resulting (from the unknown pattern) sequence $\mathbf{D} = \{d_i, i = 1 \ldots M-1\}$ is matched against a set of twelve reference patterns (one reference pattern per musical type) using CDDTW. The choice of reference patterns is based on the fact that all musical patterns of a specific type can be considered as variations of a theoretically established model. Such models are the result of musicological research in the context of Greek Traditional music [18] and describe the ideal structure that should be present in all patterns of a specific type. Each model is translated to a reference sequence $R_l = \{0, S_1, 0, S_2, 0, \ldots, S_{R_l}, 0\}$, $l = 1, \ldots, 12$, where $\{S_1, \ldots, S_{R_l}\}$ are positive or negative frequency jumps, multiples of one quarter-tone. There is only one zero separating successive S_i's because, as we will soon discuss, successive zeros do not contribute to the cost. For example, the reference pattern for musical type II is $R_2 = \{0, -2, 0, -4, 0, -4, 0, 4, 0\}$. The representative reference patterns for the twelve musical types that we studied can be accessed at http://www.di.uoa.gr/pikrakis/eus2002.

Feature sequences corresponding to patterns of the same musical type, should possess the following structure

$$\{\mathbf{0}_{z_1}, S_1, \mathbf{0}_{z_2}, S_2, \mathbf{0}_{z_3}, \ldots, \mathbf{0}_{z_{R_l}}, S_{R_l}, \mathbf{0}_{z_{R_l+1}}\}$$

where $\mathbf{0}_{z_k}$ stands for z_k successive zeros. In other words, due to the phenomenon of time elasticity, such feature sequences should, ideally, differ only in the number of successive zero-valued d_i's, separating any two S_i's. However in practice, the following deviations from this ideal situation are often encountered:

(a) Some S_i's can be one quarter-tone higher or lower than what one would expect. This is due to variations among instrument players and/or to errors during the feature generation stage (see Figure 1).

(b) Negative or positive jumps, equal to one quarter-tone, *usually encountered in pairs*, are likely to appear in the feature sequence due to errors in the feature generation stage. Such pairs manifest themselves as subsequences of d_i's of the form $\{-1, \mathbf{0}_{k_1}, 1, \mathbf{0}_{k_2}\}$ or of the form $\{1, \mathbf{0}_{k_1}, -1, \mathbf{0}_{k_2}\}$ in place of the expected sequence $\mathbf{0}_{k_1+k_2+2}$ of $k_1 + k_2 + 2$ zeros. (see Figure 1)

(c) Large pitch estimation errors, generated by the fundamental frequency tracker, (pitch doubling or pitch halving errors spanning more than one consecutive frames), are also likely to appear. Such errors usually manifest themselves as a large negative (positive) frequency jump P_1 followed by a number of zeros and a large positive (negative) jump Q_1 followed by a number of zeros.

(d) In some cases, certain S_i's are "broken" into two successive jumps whose sum is equal to the original S_i (see Figure 1).

It must be emphasized that, with the exception of variations of type (a) and (d), all these phenomena are due to errors in the feature generation process and have no relation whatsoever with what the ear perceives.

3.1 Description of the CDDTW algorithm

At a first step, we define the *"context of length N of a symbol d_i in the feature sequence"* to be the set of symbols $\{d_{i-N+1}, d_{i-N+2}, \ldots, d_i\}$. At a second step, we assume that node (j, i) of the cost grid can be reached from nodes $\{(j, i-1), (j-1, i-1), (j, i-2), (j-1, i-2), \ldots, (j, i-$

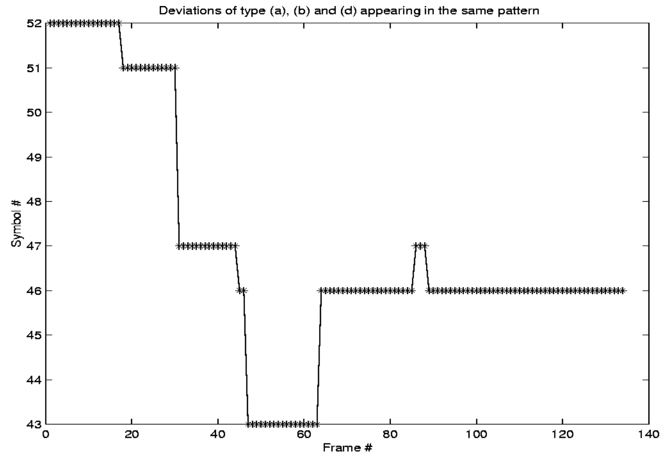

Figure 1: *Symbol sequence L of a musical pattern of type II, prior to calculating differences. Deviations of type (a), (b) and (d) can be observed. If differences are calculated, the resulting feature sequence is* $\mathbf{D} = \{\mathbf{0}_{z_1}, -1, \mathbf{0}_{z_2}, -4, \mathbf{0}_{z_3}, -1, \mathbf{0}_{z_4}, -3, \mathbf{0}_{z_5}, 3, \mathbf{0}_{z_6}, 1, \mathbf{0}_{z_7}, -1, \mathbf{0}_{z_8}\}$

$N), (j-1, i-N)\}$. In other words, the set of allowed predecessors of (j,i) is extended to include nodes ranging up to N columns on the left of (j,i) in the cost grid, excluding vertical paths, i.e excluding node $(j-1,i)$. In order to define the transition costs, let us first start with an example. Figure 2 shows a possible transition $(4,3) \rightarrow (4,6)$. The cost depends on the symbols d_4, d_5 and d_6 of the feature sequence, which form the context of length 3 of d_6. In order to calculate this cost, one has, at first, to sum these symbols, thus generating a new symbol $S = d_4 + d_5 + d_6$. In the sequel, the Euclidean distance between S and $R_l(4)$ is computed and this is defined as the cost associated with the specific transition.

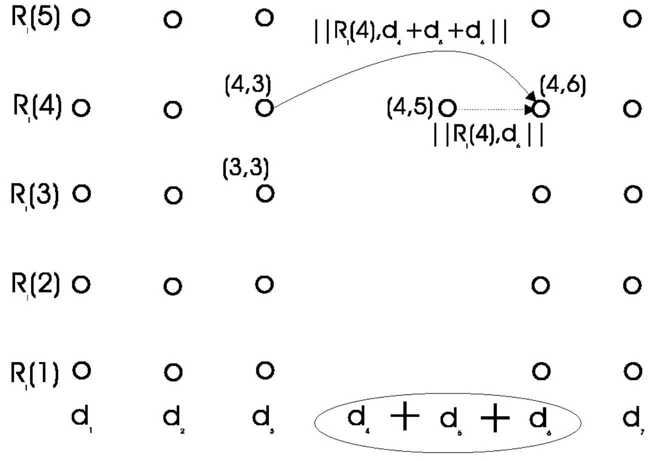

Figure 2: *The Euclidean distance below the dotted line is the cost assigned to transition* $(4,5) \rightarrow (4,6)$ *via a standard DTW scheme, whereas the distance above the long solid line is the cost assigned to the long transition* $(4,3) \rightarrow (4,6)$

Summing symbols is an attempt to cancel out (from sequence $\mathbf{D}$), the deviations described in Section 3. The longer the transition, the more complicated the deviations that are canceled out. For simple deviations, such as those of type (a), (b), and (d), short transitions are sufficient. However, when the complex version of deviations of type (c) is encountered, or when deviations are combined to generate complex phenomena, long transitions, involving up to nine symbols, are necessary. The transition marked with a solid line in Figure 2 is a relatively short one, involving the symbols d_4, d_5 and d_6. Transitions of this type are expected to cancel out simple deviations of type (b).

In the general case, the cost of a transition $(j, i-k) \rightarrow (j,i)$ or $(j-1, i-k) \rightarrow (j,i)$ is equal to $||R_l(j), \sum_{m=i-k+1}^{i} d_m||$. The cost $D_{min}(j,i)$ of the best path reaching node (j,i) is therefore equal to the minimum cost generated by the paths reaching (j,i) and is computed according to the following equation

$$\begin{aligned} D_{min}(j,i) = \min\{&D_{min}(j,i-1) + ||R_l(j), d_i||, \\ &D_{min}(j-1,i-1) + ||R_l(j), d_i||, \\ &D_{min}(j,i-2) + ||R_l(j), d_i + d_{i-1}||, \\ &D_{min}(j-1,i-2) + ||R_l(j), d_i + d_{i-1}||, \ldots \\ &\ldots, D_{min}(j,i-N) + ||R_l(j), \sum_{m=i-N+1}^{i} d_m||, \\ &D_{min}(j-1,i-N) + ||R_l(j), \sum_{m=i-N+1}^{i} d_m||\} \qquad (1) \end{aligned}$$

It is possible to reduce further the computational complexity of CDDTW if the following observation is taken into account: in the feature sequence

$$\mathbf{D} = \{0, -2, 0, -3, 0, -1, 0, -4, 0, 4, 0, 1, 0, -1, 0\}$$

each pair of non-zero symbols is separated by a single zero. Due to the way CDDTW works a zero does not alter the cumulative jumps that are calculated. Therefore, it is possible to omit zeros entirely, both from the features sequence of the unknown pattern and the reference pattern. In our example, $\mathbf{D}$ becomes $\{-2, -3, -1, -4, 4, 1, -1\}$ and R_2 becomes $\{-2, -4, -4, 4\}$. This suggests that it suffices to keep the non-zero d_i's from the original feature sequence.

4 APPLICATION OF THE METHOD

The Greek Traditional clarinet is an instrument that closely resembles the western-type clarinet. The lowest possible fundamental that the instrument can produce depends on its tuning. For the purpose of our study, this was measured to be equal to D3=146.8Hz.

A set of 1200 musical patterns were generated by four professional Greek Traditional Clarinet players in a monophonic environment, involving all the aforementioned twelve types of musical patterns. For the feature generation stage, the new fundamental frequency tracking algorithm along with the narrowed autocorrelation method and Tolonen's multipitch analysis model were extensively tested.

For the quantization step we used an alphabet of 121 discrete symbols, with each symbol being equal to a frequency jump in the range of $-60 \ldots +60$ quarter-tones, i.e. $G = 60$ (Section 2).

Two sets of experiments were carried out. One using the standard DTW technique employing Itakura and Sakoe-Chiba constraints. The latter proved to be more robust for the signals of our interest since these constraints allow for long horizontal and vertical paths in the cost grid. The success rate obtained was of the order 93%, with little variations depending on the pitch extraction algorithm used.

The other set of experiments employed the new CDDTW scheme and the success rate was significantly improved to above 95%. It must be stated that this method was basically immune to variations of the pitch tracking method used. The context length for the CDDTW scheme was set equal to 9 symbols.

All experiments were carried out using the MATLAB environment.

5 CONCLUSIONS AND FUTURE RESEARCH

In this paper an efficient scheme for the recognition of isolated musical patterns was presented. The scheme is based on CDDTW, a novel extension of standard DTW schemes. The feature generation stage of the scheme employs a new fundamental frequency tracking algorithm. The methodology was applied with success in the context of Greek Traditional Music. A reason for this choice is that it provides a musically homogeneous material, generated by the traditional mode of playing the instrument, and at the same time presents many constraints (like the unequal musical intervals and the change of the spectral content of the sound depending on the playing mode). Future research will focus on applying this new recognition scheme in the context of Classic Western Music, with other instruments besides clarinet and with multidimensional feature vectors.

References

[1] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 26, Feb. 1978.

[2] H. Sakoe, "Two-level DP matching: A dynamic programming based pattern recognition algorithm for connected word recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 27, Dec. 1979.

[3] F. Itakura, "Minimum prediction residual principle applied to speech recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 23, Feb. 1975

[4] H.F. Silverman and D.P. Morgan, "The Application of Dynamic Programming to Connected Speech Recognition", *IEEE ASSP Magazine*, July, 1990.

[5] J.Deller and J.Proakis and J.Hansen, *Discrete-Time Processing of Speech Signals*, McMillan, 1993.

[6] S.Theodoridis and K.Koutroumbas, *Pattern Recognition*, Academic Press", 1998.

[7] A. Pikrakis, S. Theodoridis, D. Kamarotos, "Recognition of Isolated Musical Patterns in the context of Greek Traditional Music using Dynamic Time Warping Techniques", *Proceedings of the International Computer Music Conference (ICMC), 1997*, Thessaloniki, Greece.

[8] D.R. Stammen and B. Pennycook, "Real-time Recognition of Melodic Fragments using the Dynamic Time Warping Algorithm", *Proceedings of the International Computer Music Conference (ICMC)*, 1993.

[9] B. Kostek, "Computer-Based Recognition of Musical Phrases Using the Rough-Set Approach", *Information Sciences 104*, Elsevier Science, 1998.

[10] B. Kostek and M. Szczerba, "Parametric Representation of Musical Phrases", *101st Convention of the AES*, Nov. 1996.

[11] M.R Schroeder, "Period Histogram and Product Spectrum: New Methods for Fundamental-Frequency Measurement", *Journal of the Acoustical Society of America*, Vol. 43, no. 4, 1968.

[12] M. Piszczalski and B. Galler, "Predicting musical pitch from component frequency ratios", *Journal of the Acoustical Society of America*, Vol. 66, no. 3, 1979.

[13] J.C. Brown, "Musical fundamental frequency tracking using a pattern recognition method", *Journal of the Acoustical Society of America*, Vol. 92, no. 3, 1992.

[14] D. Cooper, K.C. Ng, "A Monophonic Pitch-Tracking Algorithm based on Waveform Periodicity Determinations using Landmark Points" *Computer Music Journal*, Vol. 20, no. 3, Fall 1996.

[15] J.C. Brown and B. Zhang, "Musical frequency tracking using the methods of conventional and narrowed autocorrelation", *Journal of the Acoustical Society of America*, Vol. 89, no. 5, 1991.

[16] T. Tolonen and M. Karjalainen, "A Computationally Efficient Multipitch Analysis Model", *IEEE Transactions on Speech and Audio Processing*, Vol. 8, no. 6, Nov. 2000.

[17] A. Pikrakis, S. Theodoridis, D. Kamarotos, "Recognition of Isolated Musical Patterns using Discrete Observation Hidden Markov Models", *Proceedings of the European Signal Processing Conference (EUSIPCO), 1998*, Rhodos, Greece.

[18] S. Karas, "Theoritikon - Methodos , on Greek Traditional Music (in Greek)". Athens, 1982.