Scale Transform in Rhythmic Similarity of Music

André Holzapfel and Yannis Stylianou, *Member, IEEE*

***Abstract*—As a special case of the Mellin transform, the scale transform has been applied in various signal processing areas, in order to get a signal description that is invariant to scale changes. In this paper, the scale transform is applied to autocorrelation sequences derived from music signals. It is shown that two such sequences, when derived from similar rhythms with different tempo, differ mainly by a scaling factor. By using the scale transform, the proposed descriptors are robust to tempo changes, and are specially suited for the comparison of pieces with different tempi but similar rhythm. As music with such characteristics is widely encountered in traditional forms of music, the performance of the descriptors in a classification task of Greek traditional dances and Turkish traditional songs is evaluated. On these datasets accuracies compared to non-tempo robust approaches are improved by more than 20%, while on a dataset of Western music the achieved accuracy improves compared to previously presented results.**

***Index Terms*—Computational ethnomusicology, music information retrieval (MIR), rhythmic similarity, scale transform.**

I. Introduction

TWO time sequences can be compared by measuring similarities of various kinds, depending on what is the task at hand. Looking at a speech signal, for example, one can ask if in two sequences the same vowel is contained. A suitable similarity measure for this is based on the similarity of the spectral envelopes of the signals. When the question is the language of the recording, one might focus on the different temporal development of utterances, because languages typically differ in their syllable rate. A similar situation is found in music information retrieval (MIR): the most appropriate cues depend on the kind of similarity that is to be determined. In case the task is to find if a piece of music is more similar to classic music or to folk music, usually characteristics derived from the spectral content are sufficient [1], [2]. When the task is to classify into a genre of dance music, such as tango or waltz, then temporal characteristics have to be taken into consideration [3]–[8]. In [3], a self similarity measure is used to derive beat spectra, that are compared by using a cosine distance. This measure is shown to work well within a narrow range of tempo variation only. The approaches in [4] and [5] do work in presence of different tempi, but for this either the tempo or meter characteristics have to be estimated. As indicated in [9], these type of estimation is not very reliable for music signals without strong percussive content or with complex rhythmic structure, such as Folk or Jazz. The findings in [10] indicate that these type of estimation is difficult on traditional forms of music. Furthermore, state of the art meter tracking approaches have not been applied yet to music forms with time signatures unusual in Western popular music. In [6]–[8], some features are presented that do not need any tempo estimation, such as periodicity histograms, inter-onset interval histograms, or temporal modulation patterns. The common shortcoming of these descriptors is that they cannot be directly compared in presence of tempo differences, and for that reason characteristics of the descriptors such as their flatness or energy have to be used.

In this paper, a novel method for the measurement of rhythmic similarity in music is presented. In Western music, tempo changes appear within certain boundaries, as observed in [4] in the example of dance music. In traditional dances the tempo of the performance usually varies between different performances but also within the duration of the piece [11], [12]. Thus, in order to compare dance music that accompanies the same dance but is performed in different tempo, a similarity measure robust to these changes is necessary. Apart from traditional dances, other forms of traditional music are also characterized by wide tempo changes. An example is classic Ottoman music, where compositions are categorized by their melodic scheme, the *makam*, and their rhythmic scheme, the *usul*. As these rhythmic categories are not in general connected to a certain form of dance, they can vary widely in tempo. Furthermore, the *usul* can have complex or compound time signatures. For these types of music signals, a rhythmic similarity measure was recently proposed in [13] and it was based on the scale transform [14]. The scale transform is scale invariant, or equivalent in music, is not sensitive to tempo changes. In [13], it was shown that it can be applied in rhythmic similarity of music without previous tempo or meter estimations. Until now, the scale transform has been applied in various fields of signal processing in order to compare signals that have been changed by a scale factor. For example, in [15] the scale transform is applied to vowel recognition in speech. The usage of the scale transform is motivated by the fact, that between two speakers uttering the same vowel, there is a scaling in frequency domain due to the different vocal tract lengths (VTLs). Similar observations can be found in [16], where the scaling of the impulse response of the vocal tract due to different VTLs is shown to disappear when applying a Mellin transform. In [17], the scale transform was applied in order to estimate the speed gaps between mechanical systems, which are assumed to cause the related signals to be different by a scale factor. To the best of our knowledge, the scale transform has been applied to music signals only for audio effects [18]. However, two studies have

Manuscript received June 10, 2009; revised February 09, 2010; accepted February 11, 2010. Date of publication March 15, 2010; date of current version October 01, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dan Ellis.

A. Holzapfel is with the Technological Education Institute, Heraklion GR 710 04, Crete, Greece (e-mail: hannover@csd.uoc.gr).

Y. Stylianou is with Institute of Computer Science, FORTH, Heraklion GR-711 10, Crete, Greece, and also with the Multimedia Informatics Lab, Computer Science Department, University of Crete, Heraklion 71409, Crete, Greece (e-mail: yannis@csd.uoc.gr).

Digital Object Identifier 10.1109/TASL.2010.2045782

observed improvements when including a scale invariance into their approaches. In [19], scale invariance helped to investigate multiple fundamental frequencies with common harmonic structure. In terms of rhythm, the authors of [20] presented a method to compensate for tempo changes between two pieces of music by applying a logarithmic scale, which is closely related to the relation between the scale transform and the Fourier transform as will be denoted in Section II-A. The authors of the present paper introduced scale transform for the analysis of music signals in [13], where autocorrelation sequences are used as descriptor for the rhythmic content of a piece of dance music. When the same piece of music is performed at a different tempo, its autocorrelation is scaled in time. Thus, the scale transform magnitudes of the autocorrelations remain essentially the same and can be compared in a straightforward way. In this paper, this method will be detailed and extended so that it can be used for different types of signals. Here the focus lies on using signals different in a musicological perspective as well as under a technical perspective. This is achieved by examining a dataset of Turkish traditional music which is available in a symbolic description format (MIDI). The influence of critical system parameters will be analyzed in detail and insights into the characteristics of the obtained scale transform descriptors will be given.

This paper is structured as follows. Section II introduces the proposed method, by giving a general overview in Section II-A. The methods for computing the scale invariant rhythm descriptors for audio signals and for MIDI signals will be presented in Sections II-B and II-C, respectively. In order to facilitate a better understanding of the proposed scale domain descriptors, in Section II-D some of their characteristics are detailed. In Section III-A, the music collections will be described. The characteristics of these datasets will be outlined, and their different demands to a rhythmic similarity measure will be described. Section III-B describes previously proposed measures that will serve as a baseline for comparison, and the evaluation method is detailed in Section III-C. The experimental results are discussed in Section IV and the paper is concluded in Section V.

II. Suggested Rhythm Descriptors

In this section, we provide the necessary background of scale transform for supporting our suggestions. Then, we describe the suggested method of measuring rhythmic similarities in music by distinguishing the cases of music representation by an audio waveform and by the MIDI format. More specifically, the necessary background will be provided in Section II-A, and thereafter in Sections II-B and II-C the different demands of the waveform and the MIDI data will be addressed. Section II-D gives further information about characteristics of the proposed features.

A. Scale Invariant Rhythm Descriptor

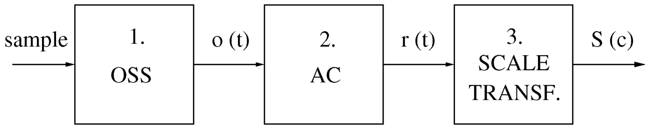


Fig. 1. Computational steps of scale-invariant rhythm descriptors.

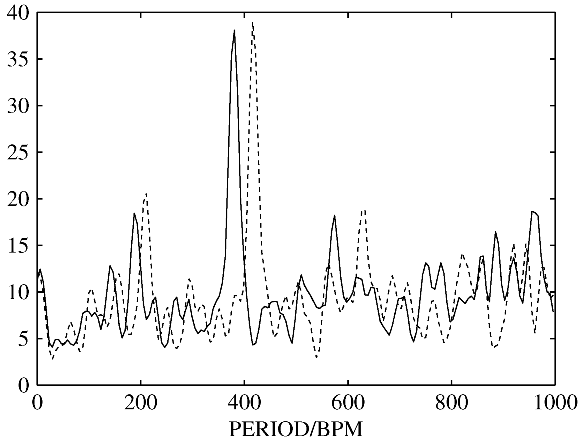


Fig. 2. Periodicity spectra of original (bold) and time-scaled (dashed) Cretan dance sample, Time scale factor: $a = 1.1$.

In Fig. 1, the three steps in the computation of scale invariant rhythm descriptors are shown. As a preprocessing step towards a scale invariant description of rhythm, onset strength signals (OSSs, denoted as $o(t)$) at a sampling frequency of 50 Hz are computed. This sampling period ensures that only frequencies related to the perception of rhythm are contained. OSSs have salient peaks at the instants where a musical instrument starts playing a note. For example, in [21] OSSs have been computed from audio signals by using a method based on spectral magnitude differences, and in [22] a method to compute OSS from a MIDI file was proposed. From the computed OSS, salient periodicities that are characteristic of the rhythm of the sample have to be found. In [23], STFTs of the onset strength signals were computed, referred to as periodicity spectra. If $X(f)$ is the Fourier transform of $x(t)$, then it is well known that

$$\sqrt{a}x(at) \leftrightarrows \frac{1}{a}X(f/a) \quad (1)$$

In Fig. 2, a periodicity spectrum of a Cretan dance sample of the class *Siganos* is shown in bold lines, while the periodicity spectrum of its time scaled version is depicted in dotted lines. The scaled version was obtained using the *audacity*[1] software, by applying the included plug-in for changing tempo of an audio file with a scale factor of $a = 1.1$. The scaling in the frequency domain representation can be recognized in Fig. 2. The immediate computation of a pointwise distance between the depicted periodicity spectra is affected by the time scaling caused by the different tempi.

In this paper, the use of the scale transform is suggested to overcome the differences in the tempo between similar, in terms of their rhythm, music pieces. The scale transform is a special case of the Mellin Transform, defined as [14]

$$X(c) = \frac{1}{2\pi}\int_0^\infty x(t)e^{(-jc-1/2)\ln t}dt \quad (2)$$

and it can be shown to be scale-invariant, which means that the magnitude distributions of the scale transforms of signals $x(t)$ and $\sqrt{a}x(at)$ are equal [14]. Although the scale transform is scale invariant, it is not shift invariant. This means that $x(t)$ and $x(t-a)$ have different magnitude scale transform. Instead of

[1] [Online]. Available: http://audacity.sourceforge.net/.

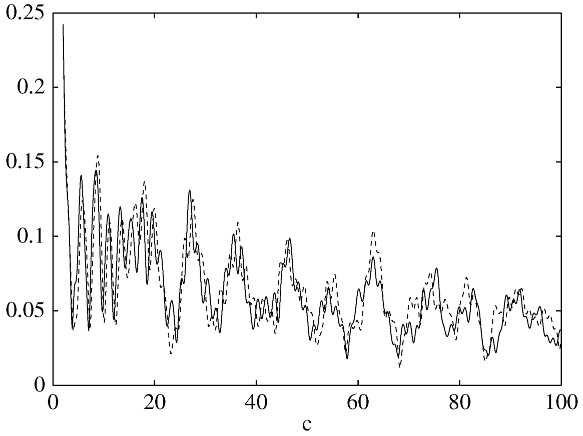


Fig. 3. Mean scale transform magnitudes of original (bold) and time-scaled (dashed) Cretan dance sample, time scale factor: $a = 1.1$.

using OSS, as usually suggested in this context (i.e., [23] and references there in), and motivated by the approach described in [17], we suggest to use the autocorrelation function $r(t)$ of OSS as a descriptor for the rhythm. It is worth noting that the autocorrelation function of a scaled signal is equal to the scaled (by the same scale factor) version of the autocorrelation of the original signal. By using the autocorrelation function of OSS we overcome the shift-variant property of the scale transform. Therefore, the suggested approach is scale (or tempo) and shift invariant. Throughout this paper, the computed autocorrelations were normalized, so that their value at the zero lag equals to one. In Fig. 3, the scale magnitudes for the same examples used in Fig. 2 are depicted. It is evident that their scale magnitudes are essentially the same and they can be compared by a point to point distance measure in a straightforward way, avoiding the dynamic programming procedure proposed in [23].

The computation of the scale transform can be performed efficiently by using its relation to the Fourier transform [24]

$$R(c) = \int_0^\infty r(e^t)e^{1/2t}e^{-jct}dt \tag{3}$$

which is the Fourier transform of the exponentially warped signal weighted by an exponential window. Since the autocorrelation computed from OSS is a real signal, this relation to the Fourier transform clarifies that negative scale values need not to be considered since the magnitude spectrum is an even function of frequency. While in [13] the implementation of the scale transform based on (3) was used, in this paper the algorithm for computing the direct scale transform (DST) as presented in [25] was applied. DST is derived from (2), by approximating the integral in (2) as follows:

$$R(c) \approx \frac{\sum_{k=1}^{\infty}\left[r(kT_s - T_s) - r(kT_s)\right](kT_s)^{1/2-jc}}{(1/2 - jc)\sqrt{2\pi}} \tag{4}$$

where T_s denotes the minimum lag size of $r(t)$, which is equal to the sampling period of $o(t)$. Compared to the implementation presented in [24], the way of computation depicted in (4) avoids the interpolation that is necessary to get exponentially spaced samples from signal $r(t)$. The transform was obtained by precomputing the base function matrix $(kT_s)^{1/2-jc}$, multiplying it with the difference vector $r(kT_s - T_s) - r(kT_s)$ and normalizing using the denominator in (4). The highest scale value C computed in (4) will be determined in the experiments shown in Section IV-A. The scale resolution Δc, which defines at which scale values the scale transform in (4) is computed, was not found critical. In [17], a value of $\Delta c = 1$ was referred to be sufficient for their application. In general, Δc is related to the time domain as

$$\Delta c = \frac{\pi}{\ln \frac{T_{\rm up}+T_s}{T_s}} \tag{5}$$

where $T_{\rm up}$ is the maximum retained lag time of the used autocorrelation [17]. For example, if $T_{\rm up} = 8$ s and $T_s = 0.02$ s then a value of $\Delta c = 0.52$ is obtained, which means that the nth scale coefficient is computed for $c = n\Delta c$. In this paper, we will apply (5) for the computation of Δc.

B. Computation From Audio Signals

On waveform data, OSSs are computed using the method proposed in [21]. Then, the sample autocorrelation r_a is computed from the OSS $o(t)$ as

$$r_a(t,w) = \sum_{n=0}^{T_{\rm win}-t-1} o(n+t+wH)o(n+wH) \tag{6}$$

where $T_{\rm win}$ denotes the length of the rectangular analysis window in seconds, w denotes the index of the analysis frame, and H the analysis hop size, which was set to 0.5 s. The maximum lag $T_{\rm up}$ of the autocorrelation was set equal to $T_{\rm win}$. For each analysis frame w, the sample autocorrelation is transformed into scale domain by applying the DST as denoted in (4), and only the magnitude values for scales $c < C$ are kept. This way, slight tempo changes within the piece are compensated, because they cause a scaling between autocorrelations computed in different analysis windows, which does not affect the scale transform magnitudes. To get a single description vector for a song i, the mean of the scale transform magnitudes is computed, which will be denoted by S_i^C. In Fig. 3, the mean scale transform magnitudes (STM) computed using the described method are depicted.

C. Computation From MIDI Data

For MIDI data, there are mainly two differences in computing the STM:

1) First, the onset times and the note durations are exactly known as they can be read from a MIDI file. For that reason, tools from the miditoolbox [26] could be used to derive the sample autocorrelations. The onset times in milliseconds were read from the MIDI files for the MIDI channels that contain the song melody (channels 1 and 2). Using these onset times, onset vectors are generated at the same sampling period of $T_s = 20$ ms as for the audio signals. The amplitudes at the onset times are determined regarding the duration annotated in the MIDI file, as suggested in [22].
2) The second difference is that the windowed computation of the autocorrelation as defined in (6) has been found to cause problems. This is related to two facts: OSSs derived

from MIDI data are much more sparse than OSSs derived from waveform data, as the onsets are discrete impulses of varying height. Furthermore, the tempo of pieces in MIDI format remains absolutely constant. No noise is induced by the way humans play musical instruments, which can cause the peaks in OSS to deviate from the position determined by the meter. Because of that, one sample autocorrelation is obtained using the whole onset strength signal as input. The autocorrelation is then transformed into scale space by using (4), resulting in the STM descriptor for a MIDI signal.

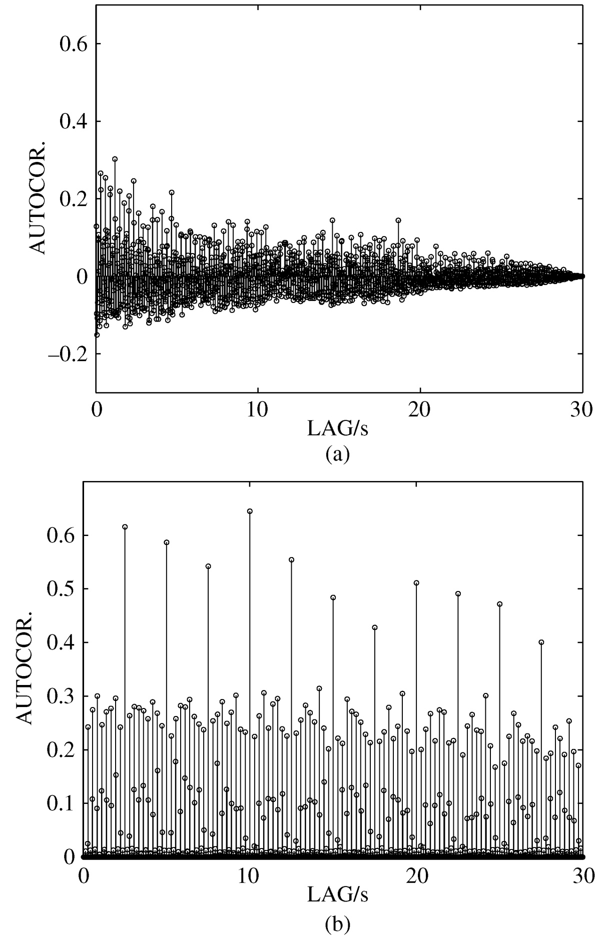
D. *Some Properties of STM*

In order to enable better understanding of the features in the scale domain, some more details about the scale transform will be provided in this section. Two autocorrelation sequences of OSS computed over audio (a) and MIDI data (b) are depicted in Fig. 4. Note that both autocorrelations show a periodicity that is related to the tatum, i.e., the smallest metrical unit in the piece [9]. Especially, the autocorrelation sequence computed from MIDI data shows a similarity with a pulse train of the tatum period. Considering a pulse train $\sum_{n=1}^{\infty} \delta(t-pn)$ with period $p > 0$, the scale transform pair of this pulse train is given by [27]

$$\sum_{n=1}^{\infty} \delta(t-pn) \Longleftrightarrow p^{-jc-0.5}\zeta(jc+0.5) \tag{7}$$

where $\zeta(s)$ denotes the Riemann Zeta function [28]. In panel (a) of Fig. 5, the magnitude of the Riemann Zeta function $\zeta(jc+0.5)$ is depicted. In panel (b) of Fig. 5, two STM derived from autocorrelations of samples from two traditional Turkish songs represented in MIDI format are shown. It is apparent that these STM have similarities with the envelope of the Riemann Zeta function. Note that for the STM computed on the autocorrelation sequences obtained from audio waveforms (see an example in Fig. 3) depicted in Fig. 3, this similarity is not so distinct. This is because, as it was shown in Fig. 4, the autocorrelation sequences derived from waveform data are less spiky than the corresponding sequences computed from MIDI data. Note that the overall shape of the Riemann Zeta magnitude does not depend on period p, and thus leads to a similar shape of the STM envelope for pieces with different tempi. In practice, one more problem we have to face is the energy compensation between scaled signals. In theory, because of the energy normalization factor $\sqrt{a}$ the scale transform magnitude remains the same for scaled signals. However, in our case, the autocorrelation functions cannot easily be normalized since they are derived from different signals, with unknown scale relation. This infeasibility of correct normalization in the time domain would lead to a constant factor change in scale magnitude. For that reason a Euclidean distance measure between STM is not applicable. As the appearance of p in the scale transform of a pulse train constitutes a constant factor in magnitude, instead of measuring Euclidean distance we suggest to measure the angle between two STM.

It is worth to clarify the effect of choosing some range of scale coefficients $c < C$ at this point. As mentioned above, autocorrelation sequences derived from musical signals are typically characterized by the period defined by the tatum of the piece.



(a)

(b)

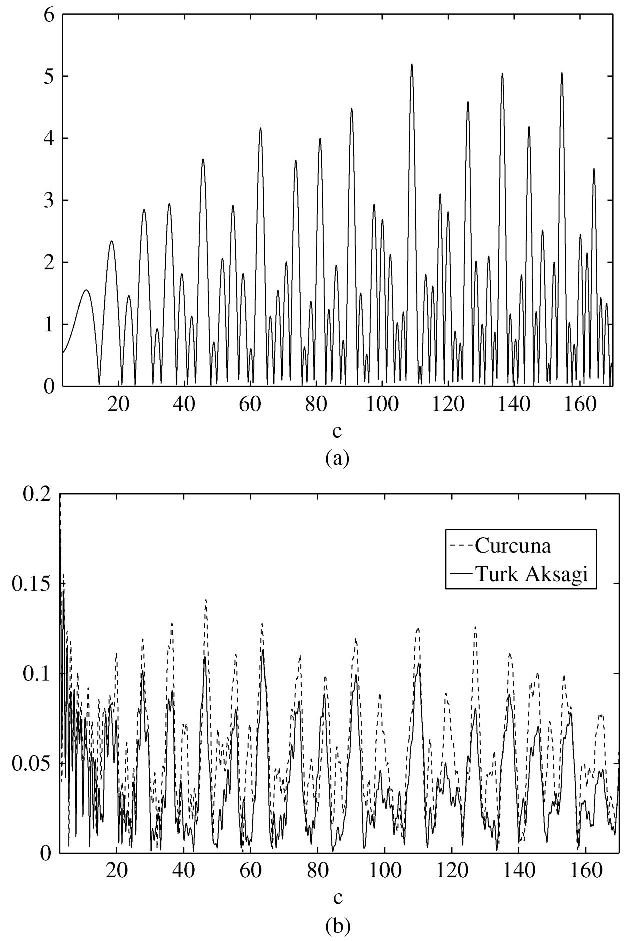
Fig. 4. Two examples of autocorrelation vectors for (a) waveform and (b) MIDI data.

In Fig. 6, three pulse trains, as a simplified model for such type of autocorrelation sequence, are reconstructed using the complex scale coefficients smaller than $C = \{50, 100, 200\}$. The pulse train has a length of 5 s and a period length of 100 ms, and it was sampled at a sampling period of $T_s = 20$ ms. It can be seen that by using more scale coefficients for the reconstruction, the approximation of samples at large time values gets improved. This is caused by the type of the base function applied in the scale transform as denoted in (2): functions $e^{(-jc-1/2)\ln t}$ are chirp functions for which the period is increasing as time increases. This increment is realized faster for small scale values. Thus, the base functions of c_1 will match the period of the pulse train earlier in time than the base function of c_2, if $c_1 < c_2$. This leads to an interesting interpretation: fixing the maximum lag T_{up} of the autocorrelation results in a vector of a given length, and increasing the number C in the STM descriptors equals to giving more weight to higher lag values within this vector.

III. Experimental Setup

A. *Evaluation Data*

In this paper, three different datasets are used: the first dataset, which will be referred to as D1, is a set of ballroom dances that



(a)

(b)

Fig. 5. Comparison of the Riemann Zeta function in (a) and two STM computed from two autocorrelations of MIDI samples in (b).

was used in the rhythm classification contest in the ISMIR conference 2004 [29]. It has been used for the evaluation of dance music classification for example in [4] and [6]. In [4], it was found that a classification accuracy of 78% can be achieved given the true tempo of the pieces as the only input to the classifier. Because there is a small overlap in the tempo distribution of the classes, this dataset can be considered as simple and it was chosen in order to prove the general validity of the approach presented in this paper. The second dataset, D2, is a dataset of traditional dances encountered in the island of Crete in Greece, and the third dataset, D3, consists of samples of traditional Turkish music. The latter two datasets were compiled by the first author of the paper. The distribution of tempi per dataset is provided in Table I.

Dataset D2 was used previously in [13] and contains samples of the following six dances: *Kalamatianos*, *Siganos*, *Maleviziotis*, *Pentozalis*, *Sousta*, and *Kritikos Syrtos*. Each class contains 30 instrumental song excerpts of about 10-s length. As shown in [13], there are large overlaps between their tempo distributions. In the case of tempo-halving and doubling errors in a tempo estimation preprocessing step, these overlaps would become even larger. Thus, a similarity measure that does not rely on tempo information is necessary to achieve a good classification in that dataset. Regarding their rhythmic properties, all traditional dances from the Greek islands share the property of having a 2/4 time signature [30, p. 32]. Only the dance class *Kalamatianos* in D2 has a 7/8 time signature.

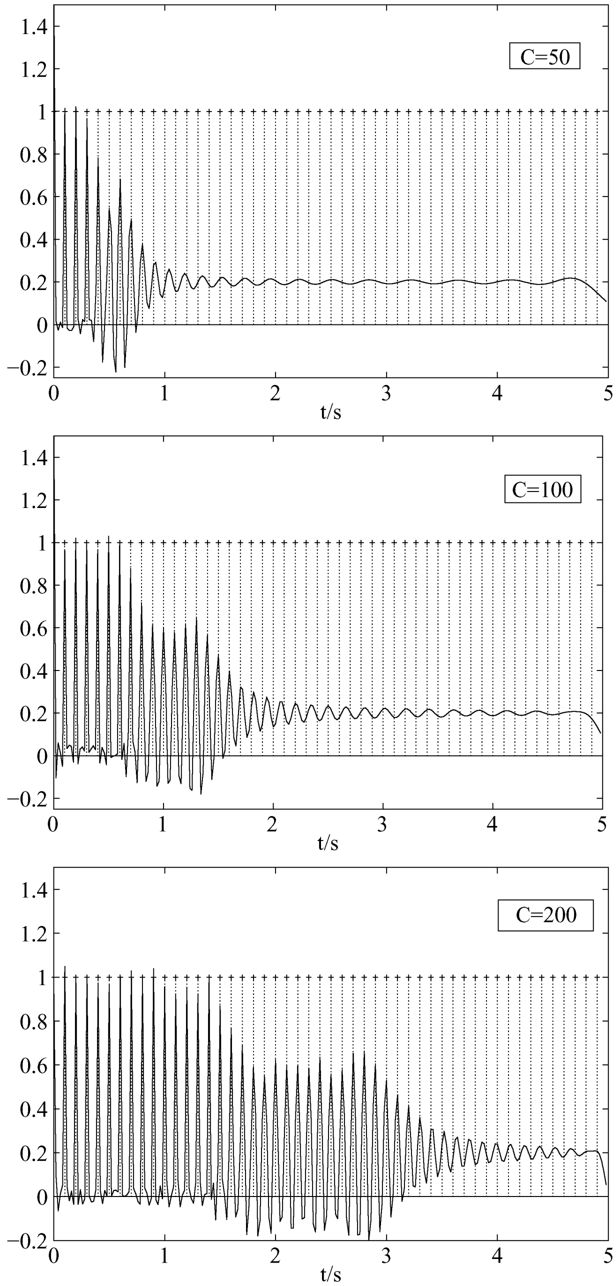


Fig. 6. Reconstruction of an impulse train by filtering in scale domain.

The dataset of Turkish music, D3, consists of six different classes of rhythm, but unlike the other two datasets, the classes are not related to specific dances. The musicological term used for the different types of rhythm in this music is *usul*. Each *usul* specifies a rhythmic pattern that defines the temporal grid for the composition. These patterns can be of various lengths from 2 up to 124 beats. The six *usul* in D3 have lengths from 3 up to 10: *Aksak* (9/8), *Curcuna* (10/8), *Düyek* (8/8), *Semai* (3/4), *Sofyan* (4/4), and *Türkaksagi* (5/8). These short *usuls* were chosen, because no sufficient number of songs with longer *usuls* were available to the authors. According to Table I, the tempo variances within each class are much bigger than in D1 and D2. This is because samples in D2 are connected to specific dance

TABLE I
STATISTICS OF THE TEMPO DISTRIBUTIONS

	D1							
CLASS	CHA	JIV	QUI	RUM	SAM	TAN	VW	WAL
MEAN	122	166	201	100	102	127	178	86
STD	5.6	14.5	11.5	11.2	18.0	4.0	2.2	4.4
N_{Songs}	111	60	82	98	86	86	65	110

	D2					
CLASS	KAL	SIG	MAL	PENT	SOUS	SYRT
MEAN	128	98	147	145	123	68
STD	8.7	4.5	8.8	10.8	8.7	5.9
N_{Songs}	30	30	30	30	30	30

	D3					
CLASS	AKS	CURC	DUY	SEM	SOF	TURK
MEAN	99	98	71	132	81	73
STD	27.9	13.5	12.6	26.3	16.7	22.3
N_{Songs}	64	57	47	22	60	38

movements which puts a natural constraint to the range of tempo variations. Most of the samples in D3 are not dance music and as such, their tempo can vary in a much wider range. Thus, features for the description of the rhythmic content have to be robust to these changes. In order to acquire the samples, the teaching software *Mus2okur* [31] was used, resulting in a collection of 288 songs, distributed among the six *usul* as shown in the last row of Table I. The software gives a list of songs for a chosen *usul*, which are then exported to a MIDI file. Thus, the data in D3 is available in form of symbolic descriptions, which means that their onset times can be read from the description. The MIDI files contain the description of the melody lines, usually played by only one or two instruments in unison, and the rhythmic accompaniment by a percussive instrument. As this content is separated into different voices, the rhythmic accompaniment can be excluded. This enables to focus on the relation between the melody of the composition and the underlying *usul*. To the best of our knowledge, such a study on *usul* has not been conducted before.

B. Similarity Measures

Because of the scale invariance property of STM, a simple point wise distance can be applied to get a (dis)similarity measure between two STM. As shown in [3] and [23], the cosine distance outperforms the Euclidean distance. Furthermore, as described in the previous section, measuring the angle between two STMs is to be preferred from using Euclidean distance due to the unknown normalization factor. Because of that, the rhythmic dissimilarity between songs i and j can be measured by computing the cosine distance between their mean STMs S_i^C and S_j^C

$$d_{sc}(i,j) = 1 - \frac{S_i^C \cdot S_j^C}{|S_i^C|\,|S_j^C|}. \tag{8}$$

In order to confirm the superiority of the cosine distance compared to the Euclidean distance, the Euclidean distance between two mean STM $d_{\mathrm{eucl}}(i,j)$ will also be used. For reasons of comparison, some previously proposed measures of rhythmic similarity will be used as well. As shown in [3] and [23], the cosine distance denoted in (8) is a good measure for rhythmic similarity directly applied to periodicity spectra if the tempi do not differ widely between the pieces that are compared. Because of that, such measures can be expected to perform well on D1 with its small tempo variations while it should decrease in performance on the other datasets. The cosine measure will be denoted as $d_{\cos}(P)$ when directly applied to periodicity spectra, and $d_{\cos}(R)$ when directly applied to the autocorrelation sequences derived from OSS.

In [23], a dissimilarity measure based on a warping strategy was introduced: periodicity spectra as shown in Fig. 2 are computed from OSS, and then the periodicity spectrum of one song is warped in order to be aligned with the periodicity spectrum of another song, a process referred to as dynamic periodicity warping (DPW). The linearity of the warping path derived in DPW serves as a measure of rhythmic similarity: the more linear the warping path, the more similar the two pieces are considered. This dissimilarity measure will be denoted as d_{DPW}.

C. Evaluation Procedure

For a given dataset, all pairwise dissimilarities between songs are computed using the measures described in Section III-B. This results in dissimilarity matrices, having values close to zero whenever two pieces are found to be similar. In order to determine the accuracy of the proposed rhythmic similarity measure, the accuracies of a k-Nearest Neighbor (kNN) classification will be determined. For this, each single song will be used as a query for which a classification into one of the available classes is desired, i.e., a leave-one-out cross validation is performed using the computed dissimilarity matrix as an input. The value k that determines the number of neighbors is varied in the interval $[2\ldots 30]$, and the best accuracy achieved by varying k is then reported. In order to determine if these accuracies are over optimistic, the kNN accuracies will be compared with results achieved using a Fisher LDA classifier and a pairwise SVM classification using a linear kernel. For SVM, the implementation included in the WEKA software [32] has been used without any parameter changes. Both LDA and SVM classifiers are evaluated using leave-one-out cross-validations.

In Section IV-A, the accuracy of the proposed STM features for the discrimination of different rhythms will be discovered. Therefore, it is necessary to evaluate the optimum set of scale coefficients for each dataset. In the first experiments, the accuracy depending on the choice of the highest included scale coefficient will be determined. In Section IV-D it is evaluated if a maximum relevance feature selection as proposed in [33] can provide us with a consistent way to derive a compact set of features that is optimal for the classification task. For this, the relevance to the target class of each feature in a training set is computed by determining their mutual information

$$I(x_i,c) = \iint p(x_i,c)\log\left(\frac{p(x_i,c)}{p(x_i)p(c)}\right)dx_i dc. \tag{9}$$

In practice, the integration in (9) is problematic for continuous valued features as the scale coefficients in our case. For that reason, each feature has been discretized by using an adaptive quantization as proposed in [33], using $b = 5$ bins. In order to select a set of relevant features, all mutual information values between the single scale coefficients and the target class have been computed. Then, a threshold has been applied to the computed mutual information, which for a value of 100% chooses all features and for a value of 0% only the one feature with the maximum relevance for the training set. Changing this threshold continuously from 0% to 100% leads to choosing a subset of features regarding their individual relevance for the classification. The influence of varying this threshold will be determined in Section IV-A.

IV. Experiments

For the proposed similarity measure d_{sc}, there are mainly two critical parameters: the length of the maximum lag T_{up} considered in the autocorrelation and the numbers of coefficients C of STM in (8). The influence of these parameters will be explored by computing the accuracies in a grid search of these two parameters. For each dataset the optimum number for the maximum lag will be determined, and the effect of varying the number of scale coefficients will be explored.

A. Optimum Upper Scale and Maximum Lag

On both waveform datasets D1 and D2, the optimum maximum lag T_{up} found in the grid search was 8 s. The accuracies for D3 improved until a maximum lag of 14 s is reached. It was observed that further increase does not lead to a decrease in accuracy on this dataset, as it is the case on the waveform data in D1 and D2. In Fig. 7, the accuracies of kNN classifiers are depicted when changing the number of scale coefficients C. The optimum maximum lag was used for each dataset. It can be seen that the accuracy of the classification depends on the number of chosen scale parameters in a different way for each dataset. The highest classification accuracy was achieved for D1, which confirms the hypothesis of this dataset being simple due to small overlaps of tempo distributions and small tempo variances in comparison to D3. More specifically, the classification accuracy increases up to 88.1% at $c = 170$. In general, an area of almost constant accuracy is reached for $C > 80$, as can be seen from Fig. 7. A similar behavior can be observed for D3, where the best accuracy using kNN is achieved at $C = 140$ (78.1%). On D2, a maximum is reached at $c = 30$ with an accuracy of 76.1%. Unlike for D1 and D3, when further increasing C on D2 the accuracy decreases. As mentioned in Section III-C, the shown kNN accuracies are the maximum values achieved by varying k, and thus the values might be overoptimistic. However, similar results are obtained using the SVM and LDA classifiers, as can be seen in Table II. For SVM, on D1 and D3 a saturation is reached while for D2 this does not hold, just like for the kNN results depicted in Fig. 7. The LDA classification could not be evaluated for very large values of C, as the increasing dimensionality causes numerical problems. In Table II, the best accuracies for all three classifiers using the proposed features are depicted along with the value of C at which this accuracy is reached. It seems that for higher scale values on D2, the STM contain more noise than for the other two datasets. As shown in Section II-D, higher scale values lead to a more accurate reconstruction at larger autocorrelation lags. Thus, regarding Fig. 6, for D2 a stronger weighting for lags smaller than one second is optimal, while for D1 this weighting is extended to about two seconds. This behavior will be further explored in Section IV-D.

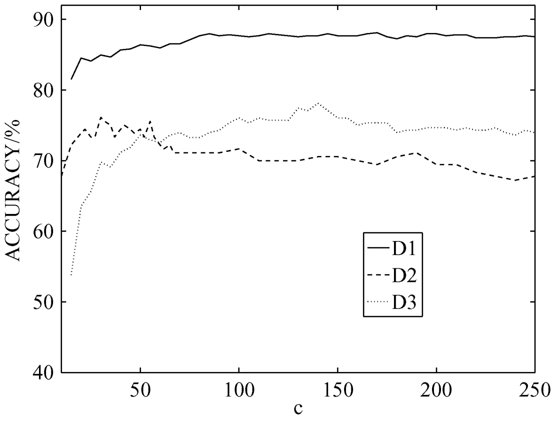


Fig. 7. Accuracies on the three datasets for varying number of scale parameters, using kNN.

TABLE II
Classification Accuracies at C Using STM Features

	kNN	SVM	LDA
D1	88.1 ($C = 170$)	91.7 ($C = 160$)	89.5 ($C = 120$)
D2	76.1 ($C = 30$)	76.1 ($C = 35$)	77.8 ($C = 25$)
D3	78.1 ($C = 140$)	82.3 ($C = 140$)	77.1 ($C = 40$)

TABLE III
kNN-Classification Accuracies

	$d_{cos}(P)$	$d_{cos}(R)$	d_{DPW}	d_{eucl}	d_{sc}
D1	86.1	86.0	83.5	86.1	**88.1**
D2	54.3	44.7	60.9	73.9	**76.1**
$D3_{mel}$	53.1	56.2	50.5	75.7	**78.1**
$D3_{all}$	63.5	66.7	71.0	83.7	**86.0**

B. Comparison of Distance Measures

Table III shows the classification accuracies on the datasets, using the measures as described in Section III-B and kNN classification. Similar to the results presented in [23], the direct cosine measures between the periodicity spectra, $d_{\cos}(P)$, and between the autocorrelation sequences, $d_{\cos}(R)$, work well on D1. The proposed scale method d_{sc} achieves a slightly improved accuracy of 88.1%. However, this improvement is not significant regarding the confidence interval, which is 2.4% (level of confidence $= 95\%$). Comparing these results to the highest accuracy, without the usage of the tempo annotations, of 85.7% as presented in [34] on the same dataset D1, the accuracy presented here using d_{sc} appears to be a satisfying proof of concept. The improvements in comparison to [23] and [13] must be assigned to the changed sample rate of the OSS (50 Hz instead of 160 Hz) which in general improved results throughout the experiments, and to the different computation of the scale transform.

For D2, Table III shows a considerable advantage for the proposed scale distance measure d_{sc}, which achieves an accuracy of 76.1% with a confidence interval of 6.2%: on this dataset it outperforms the cosine measures by 21.8/31.4 percentage points.

This clear improvement can be assigned to the robustness to tempo changes of the scale transform.

The accuracies for the dataset of Turkish MIDI files are listed in the third and fourth row of Table III. The third row gives the accuracies when using the melody lines only for the onset computation as described in Section II-C. Using the dissimilarity measure d_{sc} proposed in this paper leads to the best results: an optimum accuracy of 78.1% is reached at $C = 140$, with a confidence interval of 4.8%. Direct comparisons of either periodicity spectra or autocorrelation sequences are clearly inferior due to the large changes in tempo for each *usul*. The DPW approach presented in [23] does not lead to good results on D3. This must be assigned to the large standard deviation of the tempi in one class since DPW assumes that there are no differences in tempo larger than 20% between two songs. When tempo differences exceed this threshold, the whole procedure is becoming unreliable [23].

The fourth row of Table III (i.e., for D_{3all}) shows the accuracies that can be achieved when the tracks containing percussive accompaniment are also included in the computation of OSS. The accuracies are then in general improved, since the percussive accompaniment is typically the same for one specific *usul*. The relatively high values in the third row $D3_{mel}$ clarify the information about the *usul* that is contained solely in the melody line of the composition. As the difference between the best accuracy in the third row and the best in the fourth row is only 7.9 percentage points, it can be concluded that this relation between the melody and the *usul* is very strong.

Comparing the measures based on the scale transform (i.e., d_{eucl} using Euclidean distance and d_{sc} using cosine distance) we see that d_{sc} indeed outperforms d_{euc}. This was expected, because of the normalization factor in (1) (i.e., a) is unknown, and this affects the magnitude of vectors being compared, but not the angle between them. Compared to d_{DPW}, the distance derived using dynamic periodicity warping [23], the advantage of d_{sc} regards accuracy as well as computational: while in DPW there is the need to compute a warping path using dynamic programming, the most time consuming operation in the scale distance measure is the scale transform which is performed using a matrix multiplication.

C. Further Exploring MIDI

Two more experiments have been conducted to evaluate the robustness of the proposed method. For these experiments the SVM classification that resulted in the best accuracy of 82.3% on the MIDI data has been used, which means that all scale coefficients until $c = 140$ have been used in the STM (see Table II). Again, only the melody lines have been included in the OSS computations, while the percussive instruments have been left out.

The first experiments explores the influence of tempo deviations within the classes. Since for the MIDI files the tempo information is given, experiments could be conducted with the tempo of the pieces changed in a deterministic way. For this, from the data in D3 the global tempo mean value has been computed. Then, all pieces have been assigned this tempo mean plus a uniformly distributed noise. This noise has been varied in steps of 5% from 0% up to 85%. For 0% noise all pieces share the same tempo, and no scaling effects the autocorrelations. At 85% noise level noise level the global mean of about 87 bpm results in a possible tempo range from 13 to 161 bpm. In order to compensate for the noise introduced by the randomly changed tempo for each noise level the experiment has been rerun ten times, and the mean accuracies of the ten runs are reported. Computing the mean SVM-accuracy for the noise free case leads to an accuracy of 82.9%. The small difference to the accuracy of 82.3% (as shown in Table II) in presence of the original tempo variance of the data proves the robustness of the proposed method to this variance. Increasing the noise level leads to an almost linear decrease in classification accuracy. However, at the largest tempo noise level of 85% the accuracy is still 73.2%. This confirms that the theoretical properties of the scale transform make the features robust to large tempo changes in practice as well.

The second experiment explores the way accuracy might get affected when dealing with real audio signals of Turkish music instead of the MIDI signals as contained in D3. For that purpose, the functionality of the *MIDI toolbox* [35] for the synthesis of an audio file from a MIDI has been used. The synthesis locates Shepard tones [36] of constant intensity wherever an onset is listed in the MIDI file. Thus, computing an OSS from the signals synthesized in this way results in almost constant onset strengths amplitudes at the locations of the note onsets. The accuracy clearly decreased to 63.5% (from 82.3%), again using SVM on STM features at $C = 140$. It was investigated if this decrease is caused by the flat characteristic of the OSS that does not allow the differentiation between strong and weak onsets. For this, audio files were synthesized using the *timidity*[2] software, which uses the velocity information contained in MIDI file, which means that onsets have varying strength. A standard piano sound has been used for synthesis. In the same experimental setup, using SVM on STM features at $C = 140$, an accuracy of 77.4% was obtained. In another experiment, the durational accent type used in the OSS computation from the MIDI files was replaced by flat accents. This means that impulses of constant height were positioned at the location of all note onsets. Indeed, removing the information about the intensity of the onset leads to the accuracy of 68.7%, and it can be concluded that the weighting of an onset according to its strength is a crucial information. Thus, it can be assumed that this method will work comparably well when applied to real world audio, which contain the full range of dynamics that characterizes human performance.

D. Mutual Information Based Feature Selection

In order to find a way to obtain an optimal set of features for classification independent of the dataset, various criteria based on the coefficient energies or the scale bandwidth as introduced in [14] have been evaluated without success. We decided then to compute the mutual information, MI, between each scale coefficient and the class label as this was described in Section III-C in order to select the best features for our task from a given STM based on information theorem criteria. This was further motivated by the fact that for D1 and D3 classification accuracies improve, when low scale coefficients are left out. Thus, for each dataset different scale coefficients appear to be relevant for classification. It was decided to use the SVM classifier, which

[2] [Online]. Available: http://timidity.sourceforge.net/.

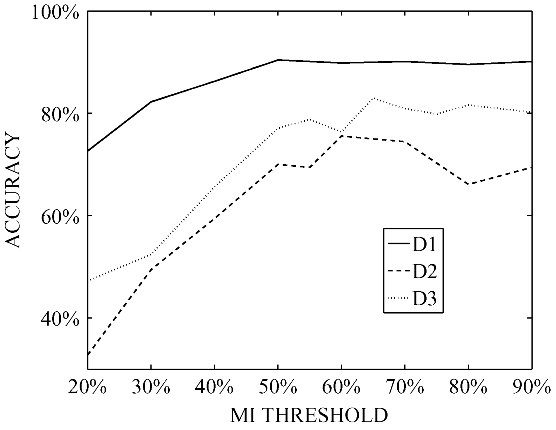


Fig. 8. SVM classification accuracies on the three datasets for varying mutual information threshold.

TABLE IV
Compression Values for Mutual Information Threshold of 60%

	D1	D2	D3
N_{feat}	249	27	98
Compression	34.7%	92.9%	76.5%

achieved the highest accuracies in Table II, and to vary the mutual information threshold as described in Section III-C on the set of features obtained for $C = 200$ for all datasets. The classification accuracies are depicted in Fig. 8. It can be seen that from an MI threshold value of about 60% upwards for all three datasets a saturation effect is reached. These saturation levels are about the same as the best classification accuracies depicted in Table II. Thus, it can be concluded that using mutual information criteria a common way to get to an optimal feature set can be defined. From Fig. 8 it is clear that the number of samples in a dataset affects the way the accuracy changes when increasing the threshold. Increasing the threshold leads to an increasing dimensionality of the feature vector, which leads to problems especially on the smallest dataset, D2. It is interesting to compare the compression achieved using mutual information thresholds for the three datasets. Table IV shows the number of coefficients corresponding to an MI threshold value of 60%. It can be seen that for D2, a much higher compression is achieved than for D1. It was observed that for D2 scale coefficients for low scales $(c < 50)$ are the most relevant, while for D1 the relevant scales were found among the whole scale range. This phenomenon is not related to the size of the datasets, but only to the different musical characteristics of the contained data. We recall from Fig. 6 that the scale coefficients until $c = 50$ allow for a reconstruction of the autocorrelation for lags up to one second. This means that small lags are more important for this type of music than the others.

E. Listening Test

In order to evaluate the relation between the proposed distance measure and the way human subjects perceive rhythmic similarity on the used data, a listening test was conducted. For this test, 11 subjects were asked to judge the similarity measurements performed on D2 which lead to the optimum classification performance for this dataset in Section IV-A ($C = 35$ for LDA). Each subject was asked to decide which of two comparison samples was rhythmically closer to a reference sample. A total amount of 25 reference samples were randomly chosen from D2 and presented to each subject. One of the comparison samples was the closest to the reference according to the proposed rhythm similarity measurement, while the other was the sample which was positioned in the middle of the ranked list of samples produced by the suggesting method as being similar to the reference sample. The subjects could decide for one of the two samples being closer, or they had the possibility to state that both comparison samples are equally close to the reference. All subjects had practical experience in all style of dances present in the dataset (Cretan dances). They were informed that all music will be traditional Cretan dances, but not exactly which type of dances. Furthermore, they were asked not to restrict their judgement on the recognition of the class, but to concentrate on judging rhythmical similarity, independently of the class affiliation. The result is shown in Table V, and it can be seen that in 64% of the cases the proposed measurement agrees with the listeners' judgements. In only 16% of the cases, the proposed measurement contradicted the listeners' opinion. No difference regarding the similarity of the two comparison samples was perceived in 20% of the cases. These results prove that apart from the objective verification of the proposed method in the classification task, the method is characterized by a high correlation of the way subjects perceive rhythmic similarity.

V. Conclusion

A description of the rhythmic content of a piece of music based on the scale transform was proposed. This description is robust to large tempo variations that appear within a specific class and to large tempo overlaps between classes. Using simple distance measure and classifier techniques, the descriptor vectors can be used to classify the samples with high accuracies. The approach is computationally simple and has no need of any tempo or meter estimation which might be desirable for certain kinds of music signals. Based on mutual information criteria, a method was proposed for choosing a feature set that is optimal for the classification task. The relation between autocorrelations sequences and the Riemann Zeta function in scale domain was explored, while a discussion of the signal reconstruction by applying inverse transform enabled to gain valuable insight into the relation between variables in scale and in time domain. The inclusion of the traditional Turkish dataset provided us with a potential starting point for a detailed study of rhythmic characteristics of Turkish traditional music. The suggested measure provides a simple and efficient tool for the description and comparison of rhythm content, especially applicable to music with little or no percussive content and strong tempo variations.

References

[1] A. Holzapfel and Y. Stylianou, "Musical genre classification using non-negative matrix factorization based features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 424–434, Feb. 2008.

[2] T. Li and M. Ogihara, "Toward intelligent music information retrieval," *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 564–574, Jun. 2006.

[3] J. Foote, M. D. Cooper, and U. Nam, "Audio retrieval by rhythmic similarity," in *Proc. ISMIR—Int. Conf. Music Inf. Retrieval*, 2002, pp. 265–266.

TABLE V
RESULTS OF LISTENING TEST FOR D2

CONTRADICTION	NEUTRAL	CONSENSUS
16%	20%	64%

[4] G. Peeters, “Rhythm classification using spectral rhythm patterns,” in *Proc. ISMIR—Int. Conf. Music Inf. Retrieval*, 2005, pp. 644–647.

[5] J. Paulus and A. Klapuri, “Measuring the similarity of rhythmic patterns,” in *Proc. ISMIR—Int. Conf. Music Inf. Retrieval*, 2002, pp. 150–156.

[6] F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer, “Evaluating rhythmic descriptors for musical genre clasification,” in *Proc. AES 25th Int. Conf.*, 2004.

[7] E. Pampalk, “Computational models of music similarity and their application to music information retrieval,” Ph.D. dissertation, Vienna Univ. of Technol., Vienna, Austria, 2006.

[8] T. Lidy, “Evaluation of new audio features and their utilization in novel music retrieval applications,” M.S. thesis, Vienna Univ. of Technol., Vienna, Austria, 2006.

[9] A. P. Klapuri, A. J. Eronen, and J. T. Astola, “Analysis of the meter of acoustic musical signals,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 14, no. 1, pp. 342–355, Jan. 2006.

[10] A. Holzapfel and Y. Stylianou, “Beat tracking using group delay based onset detection,” in *Proc. ISMIR—Int. Conf. Music Inf. Retrieval*, 2008, pp. 653–658.

[11] I. Loutzaki, “Audio report: Greek folk dance music,” *Yearbook for Traditional Music*, vol. 26, pp. 168–179, 1994.

[12] B. Aning, “Tempo change: Dance music interactions in some Ghanaian traditions,” *Inst. of African Studies: Res. Rev.*, vol. 8, no. 2, pp. 41–43, 1972.

[13] A. Holzapfel and Y. Stylianou, “A scale transform based method for rhythmic similarity of music,” in *Proceedings of the IEEE Int. Conf. Acoustics, Speech, and Signal Processing. ICASSP*, 2009, pp. 317–320.

[14] L. Cohen, “The scale representation,” *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3275–3292, Dec. 1993.

[15] S. Umesh, L. Cohen, N. Marinovic, and D. J. Nelson, “Scale transform in speech analysis,” *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, pp. 40–46, Jan. 1999.

[16] T. Irino and R. D. Patterson, “Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilized wavelet-Mellin transform,” *Speech Commun.*, vol. 36, no. 3, pp. 181–203, 2002.

[17] F. Combet, P. Jaussaud, and N. Martin, “Estimation of slight speed gaps between signals via the scale transform,” *Mech. Syst. Signal Process.*, vol. 19, pp. 239–257, 2005.

[18] A. D. Sena and D. Rocchesso, “A fast Mellin transform with applications in DAFx,” in *Proc. 7th Int. Conf. Audio Effects (DAFx’04)*, 2004, pp. 65–69.

[19] S. Saito, H. Kameoka, K. Takahashi, T. Nishimoto, and S. Sagayama, “Specmurt analysis of polyphonic music signals,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 639–650, Mar. 2008.

[20] J. H. Jensen, M. G. Christensen, D. P. W. Ellis, and S. H. Jensen, “A tempo-insensitive distance measure for cover song identification based on chroma features,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. ICASSP*, 2008, pp. 2209–2212.

[21] D. P. W. Ellis, “Beat tracking by dynamic programming,” *J. New Music Res.*, vol. 36, no. 1, pp. 51–60, 2007.

[22] R. Parncutt, “A perceptual model of pulse salience and metrical accent in musical rhythms,” *Music Percept.*, vol. 11, no. 4, pp. 409–464, 1994.

[23] A. Holzapfel and Y. Stylianou, “Rhythmic similarity of music based on dynamic periodicity warping,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 2217–2220.

[24] A. D. Sena and D. Rocchesso, “A fast Mellin and scale transform,” *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, pp. 75–84, 2007.

[25] W. Williams and E. Zalubas, “Helicopter transmission fault detection via time-frequency, scale and spectral methods,” *Mech. Syst. Signal Process.*, vol. 14, no. 4, pp. 545–559, Jul. 2000.

[26] T. Eerola and P. Toiviainen, “MIDI Toolbox: MATLAB tools for music research,” Univ. of Jyväskylä [Online]. Available: www.jyu.fi/musica/miditoolbox/. Jyväskylä, Finland, 2004

[27] A. D. Poularikas, *The Handbook of Formulas and Tables for Signal Processing*. Boca Raton, FL: CRC, 1999.

[28] G. F. B. Riemann, “Ueber die anzahl der primzahlen unter einer gegebenen groesse,” *Monatsber. Koenigl. Preuss. Akad. Wiss. Berlin*, pp. 671–680, Nov. 1859.

[29] “Audio description contest—Rhythm classification, 5th Int. Conf. Music Inf. Retrieval (ISMIR),” in *Proc. ISMIR2004* [Online]. Available: http://mtg.upf.edu/ismir2004/contest/rhythmContest/

[30] S. Baud-Bovy, *An Essay on the Greek Folk Song*, (in Greek) Laographic Inst. of Peleponese, 1984.

[31] M. K. Karaosmanoğlu, S. M. Yılmaz, O. Tören, S. Ceran, U. Uzmen, G. Cihan, and E. Başaran, :Mus2okur,” Data-Soft Ltd. [Online]. Available: http://www.musiki.org/, Turkey, 2008

[32] I. H. Witten and E. Frank*, Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005.

[33] M. Markaki and Y. Stylianou, “Dimensionality reduction of modulation frequency features for speech discrimination,” in *Proc. InterSpeech*, 2008.

[34] S. Dixon, F. Gouyon, and G. Widmer, “Towards characterization of music via rhythmic patterns,” in *Proc. ISMIR—Int. Conf. Music Inf. Retrieval*, 2004.

[35] T. Eerola and P. Toiviainen*, MIDI Toolbox: MATLAB Tools for Music Research*. Jyväskylä, Finland: Univ. of Jyväskylä, 2004.

[36] R. N. Shepard, “Circularity in judgements of relative pitch,” *J. Acoust. Soc. Amer.*, vol. 36, pp. 2346–2353, 1964.

André Holzapfel received the graduate engineer degree in media technology from the University of Applied Sciences, Duesseldorf, Germany, and the M.Sc. and Ph.D. degrees in computer science from University of Crete, Heraklion, Greece.

He is currently a Lecturer at the Music Technology and Acoustics Department, Technological Educational Institute of Crete, Greece. His research interests are in the field of speech processing, music information retrieval, and ethnomusicology.

Yannis Stylianou (M’95) received the Diploma of Electrical Engineering degree from the National Technical University of Athens (NTUA), Athens, Greece, in 1991 and the M.Sc. and Ph.D. degrees in signal processing from the Ecole National Superieure des Telecommunications, ENST, Paris, France, in 1992 and 1996, respectively.

He is currently an Associate Professor at the Department of Computer Science, University of Crete, Heraklion, Greece, and an Associate Researcher in the Networks and Telecommunications Laboratory, Institute of Computer Science, FORTH. From 1996 until 2001, he was with AT&T Labs Research, Murray Hill and Florham Park, NJ, as a Senior Technical Staff Member. In 2001, he joined Bell-Labs Lucent Technologies, Murray Hill, NJ, (now Alcatel-Lucent). Since 2002, he has been with the Computer Science Department, University of Crete, Heraklion, Greece, and the Institute of Computer Science at FORTH, Heraklion. He holds nine patents. He enjoys working with speech and voice signals, music, and sounds produced by marine mammals. His current research focuses on speech signal processing algorithms for speech analysis, statistical signal processing (detection and estimation), and time-series analysis/modeling.

He is on the Board of the International Speech Communication Association (ISCA), member of the IEEE Speech and Language Technical Committee and of the IEEE Multimedia Communications Technical Committee, on the Editorial Board of the *Journal of Electrical and Computer Engineering*, Associate Editor of the *EURASIP Journal on Speech, Audio, and Music Processing*, and Associate Editor of the *EURASIP Research Letters in Signal Processing*. He is Vice-Chairman of the Cost Action 2103: “Advanced Voice Function Assessment,” VOICE. He was Associate Editor for the IEEE SIGNAL PROCESSING LETTERS and on the Management Committee for the COST Action 277: “Nonlinear Speech Processing.” Among other projects in FP6, he participated in the SIMILAR Network of Excellence coordinating the task on the fusion of speech and handwriting modalities. He is a member of ISCA and of the Technical Chamber of Greece, TEE.