Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Sciences 176 (2006) 1629–1655
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

Identifying the classical music composition of an unknown performance with wavelet dispersion vector and neural nets ☆

Stephan Rein [a,1], Martin Reisslein [b,*]

[a] *Communications Systems Group, Institut für Telekommunkationssysteme, Technical University Berlin, Sekr. EN 1, Einsteinufer 17, D-10587 Berlin, Germany*
[b] *Department of Electrical Engineering, Arizona State University, Goldwater Center, MC 5706, Tempe, AZ 85287-5706, United States*

Received 3 January 2005; received in revised form 1 June 2005; accepted 3 June 2005

Abstract

As the internet search evolves toward multimedia content based search and information retrieval, audio content identification and retrieval will likely become one of the key components of next generation internet search machines. In this paper we consider the specific problem of identifying the classical music composition of an unknown performance of the composition. We develop and evaluate a wavelet based methodology for this problem. Our methodology combines a novel music information (audio content) descriptor, the *wavelet dispersion vector*, with neural net assessment of the similarity between unknown query vectors and known (example set) vectors. We define the wavelet dispersion vector as the histogram of the rank orders obtained by the wavelet

☆ A shorter preliminary version of this paper appears in the *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), Montreal, Canada, May 2004, pp. iv-341–iv-344.

* Corresponding author. Tel.: +1 480 965 8593; fax: +1 480 965 8325.
E-mail addresses: stephan.rein@tu-berlin.de (S. Rein), reisslein@asu.edu (M. Reisslein).
URL: http://www.fulton.asu.edu/~mre (M. Reisslein).

[1] Performed this work while visiting Arizona State University, Tempe.

0020-0255/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2005.06.002

coefficients of a given wavelet scale among all the coefficients (of all scales at a given time instant). We demonstrate that the wavelet dispersion vector precisely characterizes the audio content of a performance of a classical music composition while achieving good generalization across different performances of the composition. We examine the identification performance of a combination of 39 different wavelets and three different types of neural nets. We find that our wavelet dispersion vector calculated with a biorthogonal wavelet in conjunction with a probabilistic radial basis neural net trained by only three independent example performances correctly identifies approximately 78% of the unknown performances.
© 2005 Elsevier Inc. All rights reserved.

Keywords: Audio content description; Classical music; Generalization; Identification; Internet search machine; Music information retrieval; Neural net; Wavelet

1. Introduction

Due to the immense and growing amount of audiovisual data that is available on the world wide web (WWW), techniques for multimedia content classification and retrieval are becoming increasingly important. Next generation internet search machines are expected to be able to understand and process multimedia information (content). More precisely, a user query can be a mixture of multimedia data including text, audio, picture, and video content. The search machine should give a reasonable answer providing content that is highly related to the query and of relevance to the user. A music information (audio content) description and retrieval methodology for implementation in internet search machines should allow for a very compact content representation since there is an immense volume of audio data on the WWW. In addition, the methodology should allow for an efficient computation of these descriptors.

In this paper we focus on the problem of music information retrieval for classical music and in particular on the problem of identifying the classical music composition of an unknown query performance (as well as the retrieval of other performances of the composition), as described in more detail in Section 2. Our main contributions to this problem domain is to develop and evaluate a wavelet transform based audio content description and retrieval methodology that involves (i) a novel wavelet dispersion vector for describing the characteristic audio content (classical music composition) features, and (ii) neural net processing of these vectors to assess the similarities between the vector describing a classical music performance entered as query and the vectors describing the performances in the search machine's knowledge base. Our methodology is based on the insight that the audio content of a performance of a classical music composition corresponds to characteristic patterns in the wavelet coefficients. Our definition of the wavelet dispersion vector provides a compact

representation of these characteristic patterns by directly combining wavelet coefficients across scales and time instants (audio signal samples). More specifically, the calculation of the wavelet dispersion vector from the wavelet coefficients (for a set of scales and time instants) of an audio piece proceeds in the following main steps. We first determine the rank order of a coefficient for a given wavelet scale and time instant among the coefficients from all wavelet scales for the considered time instant. We then determine the histogram of the number of times the coefficients of a given scale attain a given rank order across all the time samples of the music piece. We finally obtain the wavelet dispersion vector from this histogram as detailed in Section 4. We find that the so obtained wavelet dispersion vector efficiently describes the wavelet patterns corresponding to the classical music composition, i.e., the dispersion vector describes how the wavelet coefficients are scattered (dispersed) to form the characteristic pattern. We also find that the wavelet dispersion vectors can be processed in a computationally efficient manner by a neural net to assess the similarities between a vector entered as query and the vectors of known performances of the composition. We demonstrate that the proposed methodology of combining wavelet dispersion vectors and neural net processing has good generalization properties as it identifies performances that are not part of the search machine's knowledge base (example set) with a high success rate.

We examine the performance of our methodology for combinations of 39 different wavelets with three different types of neural nets. In our performance evaluation, the search machine is provided with a performance of a classical music movement (piece of a composition) and the task is to find the same movement in a different performance/recording, whereby the performances differ in time, frequency, sound environments, and recording quality. We consider four different performances/recordings of the same 32 movements in our evaluation. By combining the biorthogonal wavelet with the order numbers 3 (for reconstruction) and 9 (for decomposition) with the scales $1, 3, 5, \ldots, 47$ with a probabilistic radial network trained with three different performances, our methodology achieves a mean success rate of 78% for identifying the movements of a performance that is not in the search system's knowledge base. The identification success rate for a performance known to the system is approximately 100%.

This paper is organized as follows. In the following subsection we review related work. In Section 2 we describe in detail the problem setting. We also present the classical music recordings, which we have used as the sample music content throughout this study. In Section 3 we report our observation that the audio content corresponds to characteristic patterns of the wavelet coefficients, which is the basis for our identification methodology. We also outline the development path of our methodology on which we have considered wavelet envelope descriptors and elementary summary statistics of wavelet coefficients which in turn have led us to the novel wavelet dispersion vector

for solving the considered problem. In Section 4, we present our novel wavelet dispersion vector. We demonstrate that this vector efficiently describes the wavelet patterns discovered in Section 3. In Section 5, we describe how the wavelet dispersion data can be processed in a neural net to identify the classical music composition of an unknown performance in a computationally effective manner. We examine the performance of our identification methodology employing the wavelet dispersion vector in conjunction with a neural net for different wavelet families and wavelet scales. In Section 6, we summarize our findings.

1.1. Related work

Audio content description, sound classification, and audio retrieval have been studied extensively, see for instance [1–6]. Related to our research are the lines of work on audio classification/indexing and audio fingerprinting/retrieval in this literature. The existing body of literature on audio classification/indexing considers systems that are trained by a number of example sounds for classification of novel sound segments into elementary content based classes or genres, see for instance [7–15]. The system developed in [14], for instance, classifies sports audio data into one of the six sound classes applause, ball-hit, cheering, music, speech, and speech with music. The systems developed in [8,10] classify audio sounds into 16 sound classes, including the sounds animal, bells, female, and telephone. There exist also systems for artist detection [16] and music type detection [17–19].

While the goal of audio classification is to categorize audio pieces into a relatively small number of classes/genres, the goal of audio fingerprinting and retrieval is to identify a particular audio piece and/or audio pieces that are very similar to a given piece. (Each audio piece or set of similar pieces may thus be thought of as an individual class.) The existing approaches for audio fingerprinting (which is also referred to as audio hashing) and retrieval can be categorized according to their design goals into two main groups. One group aims to identify and retrieve the "same" piece as the query piece, whereby the query piece is typically "known", i.e., the query piece is contained in the audio database (or the query piece is a somewhat distorted or noisy version of the known piece in the database). The other group aims to identify and retrieve pieces that are "similar" to the query piece, whereby the query piece may be unknown, i.e., the query piece may not be contained in the audio data base. The first area is relatively more mature and several approaches have been developed employing a wide variety of audio signal transforms, feature (fingerprint) extraction, and matching methods, see for instance [20–33]. The second area of identifying similar audio pieces is relatively less mature and relatively few feature extraction and matching methods have been explored. A texture score representation which is based on Mel Cepstrum coefficients

and a hidden Markov model is developed in [34]. A polyphonic binary feature vector which is obtained through a band pass filter bank and beat tracking is developed in [35]. The zero-crossing rates are extracted with an octave-band filter bank in [36] and are used to characterize the audio content through the dominant frequencies in each subband. The signal power and spectrogram are used in [37,38] to develop characteristic sequences. The psycho-acoustic perception of rhythm patterns is used in [39] to develop self-organizing maps of the audio pieces. In [40], audio features are extracted from audio compressed by the MPEG audio compression algorithm, which is based on a psycho-acoustic model, and the audio features are then processed by a fuzzy logic based clustering algorithm to identify the similar audio pieces. A characteristic signature based on a sequence of fundamental frequencies, which are based on the psycho-acoustic perception of the audio is developed in [41]. While these initial works have significantly advanced the fundamental understanding of identifying and retrieving audio, they have only explored a part of the spectrum of approaches for the transformation and feature extraction from the audio signal. In particular, the existing studies have focused primarily of the feature extraction using various forms of the Fourier transform and spectral filtering. In contrast to the existing studies, we explore the use of the *wavelet transform*, which has many attractive properties for feature extraction, and develop a wavelet based methodology for the identification of the classical music composition of an *unknown* query performance.

We note that the use of the wavelet transform for classification as well as for the identification of a known query audio piece has been studied in a few works. The statistical properties of the wavelet coefficients, such as mean, standard deviation, and zero-crossing rate, are exploited by the schemes developed in [17,29]. The scheme developed in [42] forms feature vectors from the wavelet approximation coefficients and a subset of the wavelet detail coefficients. The scheme developed in [43] forms feature vectors from the covariances between the original audio signal and the various wavelet detail signals. Our proposed wavelet dispersion vector differs from these approaches in that it does not extract the statistical properties of the wavelet coefficients, but rather captures the pattern formed by the wavelet coefficients. The scheme proposed in [44] is similar in spirit to ours in that it captures a part of the relationship between the wavelet coefficients across the wavelet scale dimension and across the time dimension in the fingerprints. The main difference between the fingerprints in [44] and our wavelet dispersion vector is that the fingerprints in [44] include information for the relationships across the time dimension for each individual time sample of the audio signal, resulting in fingerprints with a size on the order of the number of time samples in an audio segment. In contrast, our wavelet dispersion vector aggregates the wavelet coefficient relationships across the time dimension into a histogram with a bin for each wavelet scale. The number of wavelet scales is typically over two or more orders of magnitude

smaller than the number of samples in an audio segment, resulting in a correspondingly more compact audio content characterization with our wavelet dispersion vector, which as we demonstrate allows for highly precise audio piece identification. Overall, our work complements the existing literature on wavelet transform based techniques for audio identification in that we examine wavelet transform based techniques specifically for the identification of similar unknown classical music pieces. We demonstrate that the classical music compositions are represented by characteristic patterns in the wavelet domain, which allow for accurate content description and at the same time good generalization to unknown pieces.

We note for completeness that a hardware implementation of a wavelet based classifier which categorizes audio files into either a voice class or a music class has been studied in [45]. Wavelets have also been employed to reconstruct audio recordings [46], to observe the timing and frequency characteristics of cardiac cycles [47], and for audio transcription [48]. In addition, several studies have employed wavelets for classifying the texture of images, see for instance [49,50] and the shot structure of video, see for instance [51].

2. Problem setting and audio data base

In this section we describe the specific identification problem considered in this paper as well as the example classical music pieces used in this study. We begin by reviewing the generally desirable properties of content description for next generation internet search machines. First, due to the immense amount of audio data available on the world wide web, the descriptors should have a very compact representation. Secondly, the methodology should provide an efficient computation scheme for the construction of these descriptors. In addition, the methodology should provide an efficient procedure for determining—based on the descriptors—the similarities between query input and the knowledge base of the search machine to allow for a fast user-oriented search and retrieval service.

2.1. Identification problem

The problem setting that we consider in this paper is illustrated in Fig. 1. Suppose the user has a performance of unknown classical music composition and would like to find the title of the composition and to find other performances of the composition. In our example the unknown composition is the movement iv of Sonata No. 1 of Bach's *Sonatas and Partitas* performed by N. Milstein (which the user does not know). The user feeds the audio piece into the next generation internet search machine. We suppose the search machine "knows" the performance (recording) by Y. Menuhin of Bach's *Sonatas and*

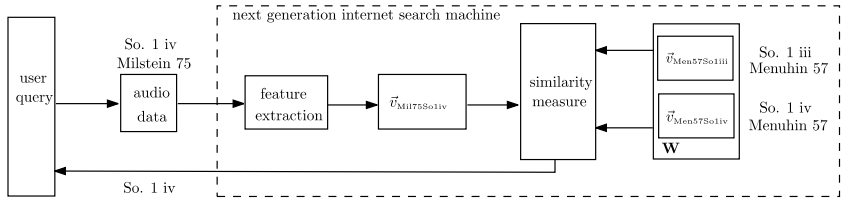


Fig. 1. Next generation audio identification and retrieval scenario. The search machine identifies an unknown performance of a classical music composition by generalizing the descriptor of a known performance of the composition.

Partitas, i.e., it has the content descriptors of the individual pieces of the Menuhin performance in its knowledge base (example set). However, we also suppose that the search machine does not know the Milstein performance. The problem for the search machine now is to identify the audio piece entered by the user as movement iv of Sonata No. 1 of Bach's *Sonatas and Partitas* and provide the user with a link to the performance by Menuhin of this piece.

To solve this problem, the feature extraction component of the audio identification and retrieval methodology has to solve a demanding problem. On one hand, a very precise content description has to be extracted because we need to distinguish different pieces within the classical music genre. On the other hand, the extracted feature descriptors should allow for a generalization. That is, the feature descriptors should not describe the content too precisely, because if the description is too precise, then it would not be possible to identify a performance that is not part of the example set of the system.

2.2. Example classical music performances

Generally, a data base of example audio pieces is required for the experimental evaluation and study of audio content description schemes. Some studies use audio data bases that contain examples of different elementary sounds, including sounds of birds, telephone, or laughter (e.g., see http://www.musclefish.com). Other studies use audio data bases constructed from popular music charts.

Our goal is to identify classical music compositions. We have chosen the *Sonatas and Partitas* composed by Johann Sebastian Bach for the Solo Violin, Bachwerkeverzeichnis (BWV) 1001–1006, as example pieces for our experiments. The *Sonatas and Partitas* composition consists of three sonatas and three partitas, each containing between four to seven distinct movements. The entire composition consists of 32 distinct movements. We denote Pa1i for the movement i of Partita 1, and denote the other movements analogously. Especially the sonatas have a very similar structure, thus posing a particular challenge. We consider four different performances of these 32 movements;

namely the performances Menuhin 1934–6 (Men36), Menuhin 1957 (Men57), Heifetz 1952 (Hei52), and Milstein 1973 (Mil75). We denote Men36Pa1i for the Men36 recording of Pa1i and denote the other movement recordings analogously.

Our music data base meets the requirements for a good test data base in that it contains music of consistent relevance, manuscripts are available for the music, and the pieces represent a wide range of quality, ranging from the Men36 performance recorded with the studio technique of 1934 to the Mil75 recording which can be considered as up-to-date audio quality. Importantly, the audio pieces in the data base should contain polyphonic and not separable phenomena. Bach's Sonatas and Partitas demand the player to concurrently use different cords. Although there is only one solo violin, a sound comparable to the performance of several violin players is present. This is a particular challenge for the compactness of the descriptors. In addition, the recordings used in this study are available on the publicly available music CDs specified in [52].

For our studies we down sampled the recordings to 8 kHz using the software cooledit 2.3 (see http://www.syntrilium.com).

3. Characteristic wavelet coefficient pattern for classical music identification

In this section we report on the characteristic patterns in the wavelet transform coefficients which correspond to the audio content in the performances of a classical music composition and explore compact representations of the characteristic patterns. The wavelet transform decomposes a signal into a weighted sum of wavelet functions. The weights are called wavelet coefficients. A wavelet coefficient is calculated for a scale s and a position τ. The scale s describes how the mother wavelet function is scaled. It can either be dilated or compressed. The position τ describes the shift of the wavelet function. The wavelet coefficients are calculated as

$$C(s,\tau) = \int_{-\infty}^{\infty} s(t)\frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right)\mathrm{d}t. \tag{1}$$

When performing a wavelet decomposition, the s-scaled mother wavelet function is slid along the entire signal $s(t)$. For each shift τ, a wavelet coefficient is calculated. This procedure is repeated for each scale. The higher the scale, the more dilated is the mother function. Similarly, the lower the scale the more compressed is the mother function. Therefore, a high scale refers to a low frequency, whereas a low scale refers to a high frequency. A wavelet transform that only uses scales and shifts of powers of two is called a dyadic wavelet transform.

For illustration of the characteristic pattern in the wavelet transform coefficients, we plot in Fig. 2 the Meyer wavelet transform of the first 13 seconds of

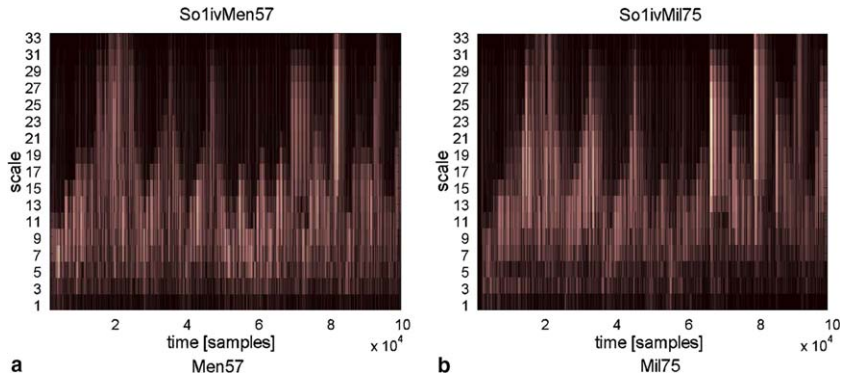


Fig. 2. Meyer wavelet transform of two different performances of So1iv. The sharply delimited patterns indicate the wavelet's ability to describe highly complex audio signals In addition, the two performances have similar wavelet patterns, indicating the wavelet's generalization ability.

the performances Men57 and Mil75 of So1iv. We observe sharp and clearly delimited patterns in the plots, which indicates that the wavelet coefficients describe very specific details of the audio signals. Wavelets can reveal very small discontinuities which would be very difficult to describe by sinusoids. Each individual musical event is resolved by very sharp and bounded patterns. The remaining challenge is to find a compact description for these patterns to solve the problem posed in Section 2. Importantly, the patterns for the two different performances are very similar, despite the differences between the individual interpretations by Menuhin and Milstein of Bach's manuscript. The similarities of the patterns in Fig. 2 indicate that a wavelet transform based characterization allows for good generalization across different performances (recordings) and thus the identification of unknown classical music performances.

From now on we consider a non-dyadic wavelet transform. Recall that a dyadic wavelet transform employs powers of two for the shifts and scales. A dyadic wavelet transform results in a more compact representation of the content, however, the extracted features are of lower precision. For our continued development of a technique for identifying classical music compositions we initially need all the details that can be resolved by the wavelet technique.

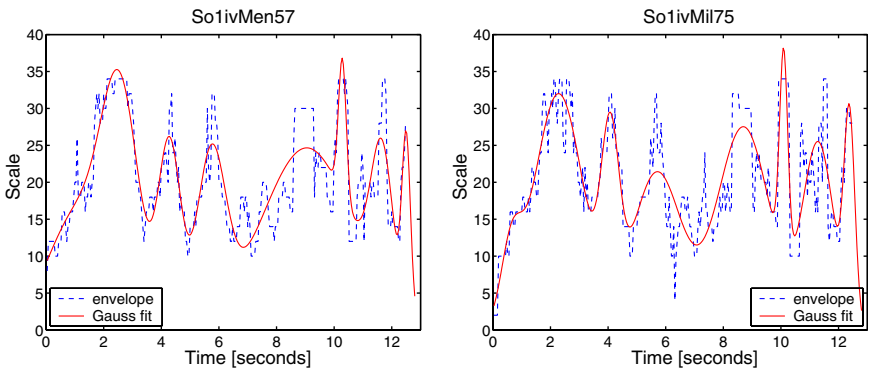
3.1. Challenges of compact content description with wavelets

Motivated by the insights reported in the preceding section, we proceed to develop audio content descriptors from the wavelet coefficients. The wavelet coefficients very closely characterize the audio content, but these coefficients are a very verbose characterization. The challenge is to extract a very compact

characterization that is practical for efficient internet search and information retrieval. During our explorations leading to the development of our wavelet dispersion vector characterization, which is our main result and is presented in Section 4, we have investigated a number of different feature extraction techniques. In this section we summarize the investigations of two techniques to overcome the challenge of wavelet data extraction and summarization; we refer the interested reader to [52] for details. We include these outlines here as they lend valuable insights that have eventually led to the development of the wavelet dispersion vector technique and may be of independent interest for other identification and information retrieval problems.

3.1.1. Gaussian wavelet envelope descriptor

As we have found in Section 3, the audio content of the classical music performances is represented by very specific patterns in the wavelet domain, which look very similar, even for different performances of the same movement (composition). In this section we outline a wavelet envelope descriptor, which describes the shape of the wavelet patterns. To obtain a *numerical* function describing the shape of the patterns shown in Fig. 2, we first estimated the average energy of the coefficients for each scale. We then set all energy values lower than a threshold, which represents the intensity of barely visible coefficients, to zero and determine the envelope of the wavelet energy patterns as the first nonzero value from the top of each column of the energy matrix. We plot the resulting envelope functions in Fig. 3 (ignore the smooth "Gauss fit" curves for now). Recall from the query example illustrated in Fig. 1 that we want the internet search machine to identify the unknown query performance Mil75So1iv by measuring similarities with the known performance Men57So1iv. In



(a) Men57So1iv (b) Mil75So1iv

Fig. 3. Numerical wavelet envelope function and corresponding analytical Gaussian fit.

order to do so based on the envelope function characterization, the search machine has to measure a good correlation between these two envelope functions illustrated in Fig. 3. Although the two envelope functions look very similar, a correlation measure would not provide a high correlation due to the varying time shift. The time shift could be compensated by a synchronization algorithm at the expense of significant added complexity as each piece in the search machine's knowledge base would need to be synchronized to the query piece before computing the correlation. We chose not to pursue this approach and explored instead an analytical description of the envelope functions.

To obtain an *analytical* representation of the envelope function, we consider a Gaussian curve fit employing functions of the form

$$y(x) = \sum_{i=1}^{N} a_i \mathrm{e}^{-\left(\frac{x-p_i}{w_i}\right)^2}. \tag{2}$$

We obtained the coefficients of the curve fit using multiple linear regression models. As illustrated in Fig. 3, the Gaussian fit smoothes the envelopes and may allow for a good generalization, because the obtained curves look very similar. We have indeed confirmed that the locations of the peaks referring to the two different recordings are highly correlated. However, we found an insignificant correlation for the peak width, see [52] for details.

The Gaussian wavelet envelope descriptor describes the shape of the wavelet patterns. However, it does not meet the requirements detailed in Section 2. This descriptor would still need a parametrization technique for the estimation of the number of peaks that should be approximated. If the number of real peaks is larger than the number of Gaussian peaks, a less precise approximation is obtained. Furthermore, the correlation or distance measure to indicate the similarity of the functions is very sensitive to single, not correctly resolved peaks. We therefore expect that this approach would require significant additional complexity to achieve sufficient generalization for precision audio content description. We next explore statistical summarization tools to solve the challenge of parsimoniously characterizing the wavelet coefficients.

3.1.2. Statistical wavelet analysis for content description

In this section we outline our investigations of descriptive statistical summarization tools to find similarities between the wavelet coefficients. In this investigation we consider two sets of representative audio data to evaluate the statistical tools' abilities to reveal similarities. The considered set I contains three different recordings (Men57, Hei52, and Mil73) of the first movement of Partita 3. Set II contains the movements ii, iii, and iv of Partita 3, each one recorded by a different player; in particular the movement recordings Pa3iiMen57, Pa3iiiHei52, and Pa3ivMil73. We consider the first 4 seconds of every audio file. Each of the two sets contains three subsets, whereby we refer

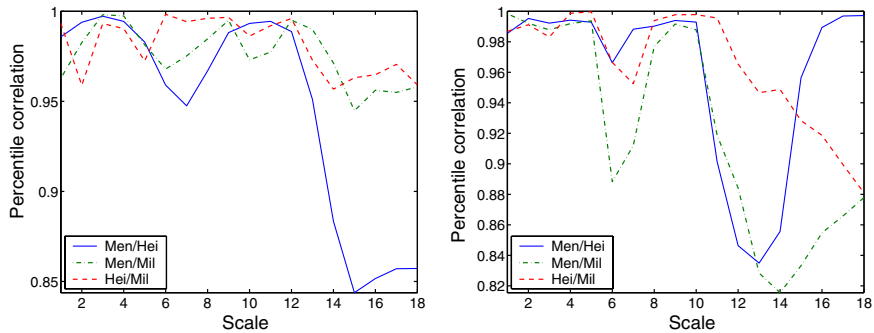
to a subset as the combination of two of the pieces in a given set. Ideally, the investigated statistical tools should measure similarities for the subsets of set I and dissimilarities for the subsets of set II.

We have considered three sets of descriptive tools, namely (i) statistical variability summarization tools, (ii) scale frequency measure (derived from our wavelet envelope descriptor), and (iii) percentile correlation plots.

Statistical data summarization tools. For comparing the distribution of the wavelet coefficients, we consider the following summarization tools: (1) arithmetic mean, (2) geometric mean, (3) harmonic mean, (4) standard deviation, (5) variation, (6) mean absolute deviation, (7) median, (8) interquartile range, (9) range, and (10) skewness, see [52] for detailed definitions. We summarize our investigation methodology as follows: For each audio piece we perform a continuous Meyer wavelet transform for the scales $1, \ldots, 18$ (see [53, p. 115 ff]). For each scale, the wavelet coefficients are summarized using one of the summarization tools. Thus, using one of the 10 summarization tools, an audio piece is represented by 18 values, which we scaled to a range of $-1, \ldots, 1$. For each audio subset we evaluate the correlation between the so obtained values. In summary, we found that the geometric mean, the standard deviation, the mean absolute deviation, and the interquartile range give consistently high correlation for all subsets of set I. However, we also found that these summarization tools give high correlations for the subsets in set II. We obtained the largest difference in correlation between the subsets in sets I and II for the skewness indicator, which gave a correlation difference (between the mean skewness correlation of set I and the mean skewness correlation of set II) of 0.47.

Scale frequency measure. We define the scale frequency measure as a histogram giving the frequency with which the numerical wavelet envelope of Section 3.1.1 attains a given scale. We found that the correlations of the scale frequency measure for the subsets within a set are not consistent and that the correlations between the subsets in sets I and II are not significantly different.

Percentile correlations. We employ percentile correlation plots to measure similarities between the distributions of the wavelet coefficients. We derive these plots from percentile plots, which we obtain for each subset and wavelet scale. We calculate one correlation index for each percentile plot. We plot these correlation indices as a function of the scale to obtain the percentile correlation plots for sets I and II, which we present in Fig. 4. We observe from Fig. 4(a) that the correlations for the subsets Men/Mil and Hei/Mil are between 0.95 and 1, whereas the correlations for subset Men/Hei drop down to 0.85 for the higher wavelet scales. On the other hand, we observe from Fig. 4(b) that the correlations for all subsets in set II drop below 0.9 for some of the higher wavelet scales. Thus, we could correctly identify the similarities and



(a) Audio set I (b) Audio set II

Fig. 4. Percentile correlation plots. A high correlation indicates that the two distributions tend to be similar.

dissimilarities for the Men/Mil and Hei/Mil subsets, but not for the Men/Hei subset.

3.2. Conclusions from explorations of compact wavelet descriptors

In summary, our investigations outlined in this section manifest the general challenges in feature extraction for precision classifiers: The wavelet envelope descriptor is precise since it describes individual events. However, the description appears to be too precise to allow for a reasonable generalization. The statistical tools, on the other hand, allow for a generalization; however, their description appears to be too broad to distinguish similar, but different audio pieces. To successfully solve our problem posed in Section 2, it would be desirable to find a descriptor between these two approaches. This descriptor should be very precise, but still allow for a generalization.

4. Audio piece content description with wavelet dispersion vector

In this section we present the wavelet dispersion vector, which extracts the characteristic features of a classical music performance and permits the identification of the corresponding composition. The intuition behind this vector is as follows. The dispersion measures discussed in the preceding section first extracted features of each scale and then concatenated these features to indicate similarities. This approach leads to a distinction between scale and time. Our novel wavelet dispersion vector, on the other hand, directly combines time and scale, as detailed in this section, and thereby describes the structure of the wavelet patterns. In order to explain the wavelet dispersion vector, we first

give a general mathematical description and then present an illustrative example.

4.1. Formal description of wavelet dispersion vector

Suppose an audio piece with T samples is given. In our numerical work we consider the first 5 seconds of each performance sampled at 8 kHz, for a total of $T = 40{,}000$ samples for a given piece. Suppose a wavelet transform with S scales is performed on the audio samples. We examine the performance of the different wavelet families, order numbers, and scales for audio identification in Section 5. Let

$$\mathbf{C} = (c_{s,t}), \quad s = 1, \ldots, S; \ t = 1, \ldots, T, \tag{3}$$

denote the matrix of obtained wavelet coefficients. Note that this matrix has S rows and T columns.

The wavelet coefficient matrix $\mathbf{C}$ is now processed as follows to obtain the wavelet dispersion vector. First, we obtain a *rank order matrix*

$$\mathbf{R} = (r_{s,t}), \quad s = 1, \ldots, S; \ t = 1, \ldots, T, \tag{4}$$

from the matrix $\mathbf{C}$. The elements $r_{s,t}$, $s = 1, \ldots, S$, of a given column t in the rank order matrix $\mathbf{R}$ are the ranks (ordered positions) of the corresponding elements $c_{s,t}$, $s = 1, \ldots, S$, of the wavelet coefficient matrix $\mathbf{C}$, i.e.,

$$r_{s,t} = \text{order}_{1 \leqslant s \leqslant S}[c_{s,t}], \tag{5}$$

whereby $\text{order}_{1 \leqslant i \leqslant I}[x_i]$ gives the position of x_i when the x_i, $i = 1, \ldots, I$, are sorted in decreasing order. In particular, the largest wavelet coefficient of a given column, i.e., $\max_{1 \leqslant s \leqslant S} c_{s,t}$ is assigned the rank 1, i.e.,

$$s_{\max} = \arg \max_{1 \leqslant s \leqslant S} c_{s,t} \tag{6}$$

$$\Rightarrow r_{s_{\max},t} = 1. \tag{7}$$

Similarly, the smallest wavelet coefficient is assigned the rank S, i.e.,

$$s_{\min} = \arg \min_{1 \leqslant s \leqslant S} c_{s,t} \tag{8}$$

$$\Rightarrow r_{s_{\min},t} = S. \tag{9}$$

Next, we calculate a *rank histogram (wavelet dispersion) matrix* $\mathbf{D}$ by counting how often a given scale (row) in the rank order matrix $\mathbf{R}$ attained a given rank ρ, i.e.,

$$\mathbf{D} = (d_{s,\rho}), \quad s = 1, \ldots, S; \ \rho = 1, \ldots, S, \tag{10}$$

with

$$d_{s,\rho} = \sum_{t=1}^{T} 1_{(r_{s,t}=\rho)}, \tag{11}$$

whereby $1_{()}$ denotes the indicator function, i.e., $1_{(A)} = 1$ if A is true and $1_{(A)} = 0$ otherwise. Note that the rank histogram matrix elements satisfy $1 \leqslant d_{s,\rho} \leqslant T$ and $\sum_{\rho=1}^{S} d_{s,\rho} = T \ \forall s = 1, \ldots, S$.

Finally, we arrange the wavelet dispersion matrix elements into the *wavelet distortion vector* $\vec{v}$ by reading the matrix elements row-by-row, i.e.,

$$\vec{v} = [d_{1,1}, d_{1,2}, \ldots, d_{1,S}, d_{2,1}, d_{2,2}, \ldots, d_{2,S}, \ldots, d_{S,1}, d_{S,2}, \ldots, d_{S,S}]. \tag{12}$$

Note that the wavelet dispersion vector has a slight relation to the percentile plots considered in Section 3.1.2. Intuitively, both for obtaining the wavelet dispersion vector and the percentile plot, the wavelet coefficients are sorted and histograms are constructed. For the wavelet dispersion vector the coefficients are sorted across the scales for a given time instant, and the histograms (one for each scale) are then constructed across time; the data summarized in a histogram, however, captures the ordering pattern of the coefficients across different scales. For the percentile plot, on the other hand, the coefficients are sorted across time for a given scale and a histogram is constructed from the sorted coefficients; the data thus summarized in a histogram does not capture the ordering pattern of the coefficients across scales.

4.2. Illustrative example of wavelet dispersion vector

To illustrate the calculation of the wavelet dispersion vector we consider an example with $T = 5$ audio samples and $S = 3$ wavelet scales and the wavelet coefficient matrix

$$\mathbf{C} = \begin{array}{ccccc|c} 1 & 2 & 3 & 4 & 5 & \\ \hline 0.43 & 0.22 & 0.14 & 0.76 & 0.33 & 1 \\ 0.10 & 0.32 & 0.11 & 0.28 & 0.90 & 2 \\ 0.54 & 0.49 & 0.34 & 0.18 & 0.91 & 3 \end{array} \tag{13}$$

(which contains only positive values for ease of illustration). The row and column indices do not belong to $\mathbf{C}$ and are only included for clarity. We now determine the rank order for audio sample $t = 1$.

$$\mathbf{C} = \begin{array}{ccccc|c} 1 & 2 & 3 & 4 & 5 & \\ \hline 0.43(2) & 0.22 & 0.14 & 0.76 & 0.33 & 1 \\ 0.10(3) & 0.32 & 0.11 & 0.28 & 0.90 & 2 \\ 0.54(1) & 0.49 & 0.34 & 0.18 & 0.91 & 3 \end{array}. \tag{14}$$

The number in brackets represents the rank order within the column. This process is repeated for all audio samples:

$$\mathbf{C} = \begin{array}{ccccc|c} 1 & 2 & 3 & 4 & 5 & \\ \hline 0.43\ (2) & 0.22\ (3) & 0.14\ (2) & 0.76\ (1) & 0.33\ (3) & 1 \\ 0.10\ (3) & 0.32\ (2) & 0.11\ (3) & 0.28\ (2) & 0.90\ (2) & 2 \\ 0.54\ (1) & 0.49\ (1) & 0.34\ (1) & 0.18\ (3) & 0.91\ (1) & 3 \end{array}. \tag{15}$$

To form the rank order matrix we only retain the ranks, i.e.,

$$\mathbf{R} = \left[\begin{array}{ccccc|c} 1 & 2 & 3 & 4 & 5 & \\ \hline 2 & 3 & 2 & 1 & 3 & 1 \\ 3 & 2 & 3 & 2 & 2 & 2 \\ 1 & 1 & 1 & 3 & 1 & 3 \end{array}\right]. \tag{16}$$

Next, we calculate the wavelet dispersion matrix by counting the ranks within a scale (row)

$$\mathbf{D} = \left[\begin{array}{ccc|c} 1 & 2 & 3 & \\ \hline 1 & 2 & 2 & 1 \\ 0 & 3 & 2 & 2 \\ 4 & 0 & 1 & 3 \end{array}\right]. \tag{17}$$

The first row, for instance, obtained the first rank once, the second rank twice, and the third rank twice. The wavelet dispersion data is then stringed to form the wavelet dispersion vector.

$$\vec{v} = [1 \quad 2 \quad 2 \quad 0 \quad 3 \quad 2 \quad 4 \quad 0 \quad 1]^T. \tag{18}$$

We further illustrate this procedure for the first five seconds ($T = 40{,}000$) of the Men36So1 recording. We employ the Meyer wavelet with $S = 18$ scales. In Fig. 5 we show the S rank histograms corresponding to the wavelet dispersion matrix $\mathbf{D}$.

4.3. Summarizing search machine knowledge in wavelet classifier matrix

Generally, for every audio piece (file) p, the wavelet dispersion vector can be constructed to represent the audio characteristics as detailed in the preceding sections. The wavelet dispersion vectors $\vec{v}_p$ of all audio pieces known to the search machine p, $p = 1, \ldots, P$, can then be combined in a *wavelet classifier matrix* $\mathbf{W}$, whereby each dispersion vector forms a column of the matrix, i.e.,

$$\mathbf{W} = (\vec{v}_p), p = 1, \ldots, P. \tag{19}$$

Our audio data base contains $P = 128$ different audio files. Thus we calculate 128 different wavelet dispersion vectors, which form the 128 columns of the wavelet classifier matrix $\mathbf{W}$. Note that this assumes that all 128 pieces are known to the search machine. If only a subset of the pieces is known to the search machine, then the number of columns P is correspondingly smaller. The number of rows of $\mathbf{W}$ depends on the number of employed wavelet scales and the dimension reduction technique, as detailed in the next section.

4.4. Dimension reduction of wavelet classifier matrix

In this section we discuss the dimension reduction of the wavelet classifier matrix $\mathbf{W}$, which is desirable to obtain a more compact representation of the

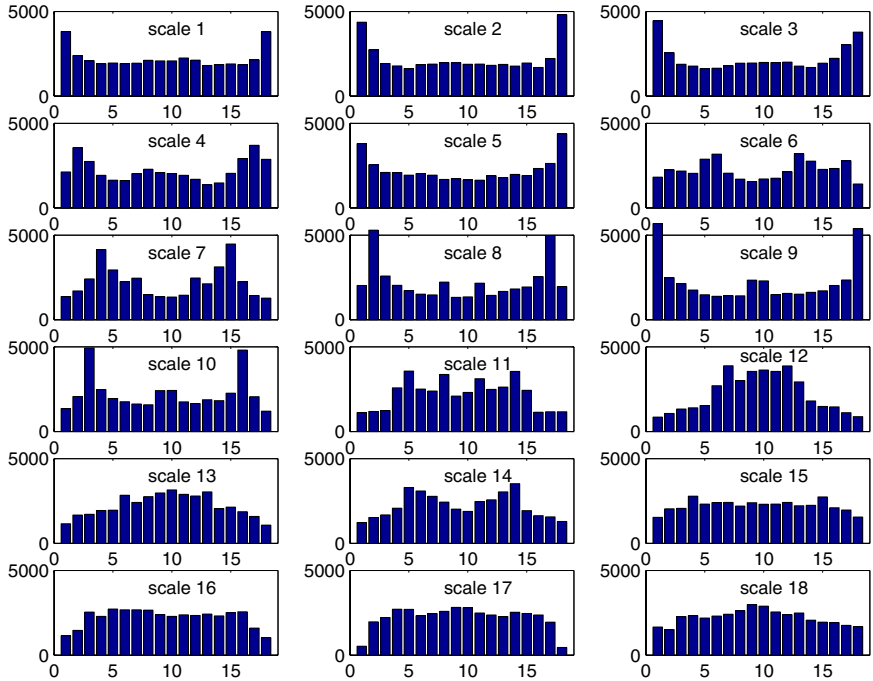


Fig. 5. Rank histograms corresponding to wavelet dispersion matrix **D** for So1Men36. The histogram values are stringed to form the wavelet dispersion vector $\vec{v}$.

audio pieces known to the search machine. We first note that there are redundancies in the rank order matrix **R** and dispersion matrix **D** calculated for a given piece p. Specifically, one row of **R** and one column of **D** are redundant and could be eliminated, which would provide a typically minor dimension reduction.

Instead of pursuing this minor dimension reduction, we propose a simple yet effective dimension reduction technique that eliminates some of the histogram columns of the dispersion matrix **D** before forming the wavelet dispersion vector. More specifically, we eliminate a certain number of the columns $d_{s,\rho}$, $\rho = 1, \ldots, S$, starting from the lowest rank $\rho = 1$ and highest rank $\rho = S$. We study this approach quantitatively in Section 5. This dimension reduction approach is motivated by the results of Section 3.1.1 where the wavelet envelope descriptor was obtained by deleting very small wavelet coefficients. These values were barely visible in the wavelet domain pictures and were not part of the specific and bounded wavelet patterns. Similarly, we discard the lowest ranks of the wavelet dispersion histograms as they represent the small wavelet coefficients. We also discard the highest ranks because they represent a set of outlying wavelet coefficients.

We also conducted experiments for dimension reduction employing the principal component analysis technique [54]. We found that the principal components of the classifier data do no longer contain the characteristic information to allow for a reasonable identification.

We also note that although the considered audio pieces contain multiple voices, a dimension reduction through sub-space estimation techniques is not possible, because the different voices do not fulfill the statistical requirements for this technique. Furthermore, such a separation is difficult because the different voices do not occur at fixed frequency bands.

4.5. Preliminary performance evaluation of wavelet dispersion vector content description

In this section we conduct a preliminary evaluation of the identification and generalization performance of the classical music description using the wavelet dispersion vector. Recall that each audio piece (file) p is described by a wavelet dispersion vector $\vec{v}_p$. In abstract terms these vectors lie in a dispersion vector space. The vectors describing performances of different compositions should be "distant" from each other in order to allow for a correct identification. On the other hand, the vectors describing different performances of the same composition (piece) should be "close" (similar) to each other in order to allow for a generalization, and thus identification of performances unknown to the search machine.

In this preliminary evaluation we measure the similarities between the wavelet dispersion vectors using the correlation between the vectors. (In the following section we employ neural nets to assess the similarity.) We consider user query scenarios in which the user queries the search machine with each of the 32 movements of a given recording and the search machine only knows the movements of one of the other performances, i.e., the wavelet classifier matrix of the search machine only contains the $P = 32$ wavelet dispersion vectors of one of the other performances of the 32 movements. The four different performances in our audio data base allow for 12 different combinations of "query performance" and "known performance". For each query the search machine identifies the closest movement among the "known performances" as the movement that attains the largest correlation with the wavelet dispersion vector of the queried movement. In Fig. 6 we show the results for the three combinations with Men36 as the query performance and each of the other three performances as the known performances, i.e., each of the 32 wavelet dispersion vectors of the Men36 recording is entered as user query and the 32 wavelet dispersion vectors from one of the other three performances are employed by the search machine's wavelet classifier matrix $\mathbf{W}$. Each individual Men36 movement query (column on x-axis) is assigned an answer (y-axis) from each of three other performances. Therefore, there are 3 points in each column on

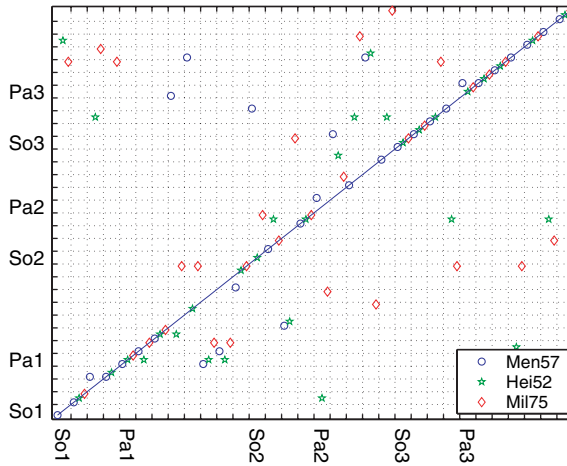


Fig. 6. Preliminary performance evaluation for wavelet dispersion vector. The user queries each of the movements of a given performance (Men36 in the shown plot). The search machine only knows the movements of one other performance (Men57, Hei52, or Mil75 in the shown plot) and identifies the movement attaining the largest correlation of the wavelet dispersion vector. Over 50% of the queried movements are correctly identified.

the plot. The three points in a given column represent from left to right the matched query results in the Men57, Hei52, and Mil75 recordings, respectively.

For example, observe in Fig. 6 how a user query containing the second movement of Partita 1 of the Men 36 recording (Men36Pa1ii) is answered. When employing the 32 column classifier matrix of Hei52, the search machine would identify this piece as the first movement of Partita 1. This identification is obtained by the maximum of 32 different correlations between the Men36-Pa1ii classifier and the 32 Hei52 classifiers. The maximum correlations between Men36So1ii and the 32 Men57 and Mil75 classifiers give the correct answers. If all points are on the line in Fig. 6, then all pieces were correctly identified. We observe from the results plotted here (and the experiments with the other three query performances which we can not include here due to space constraints and for which we refer to [52]) that approximately 60% of the movements are correctly identified. Noting that the probability of correctly identifying a movement by random choice is only 1/32, this indicates a good generalization ability.

Each point in the plots in Fig. 6 represents an audio retrieval of only one recording. In the related literature, search and retrieval systems are proposed that are trained by many audio files describing the same content. For example, in [7], more than 48 sound clips of laughter are employed to construct a laughter classifier. In this work, we are interested in employing a minimum number of audio files to construct a classifier that already achieves reasonable results.

Therefore, our data base contains only 4 different recordings of the same composition. Thus a piece unknown to the search and retrieval system can be identified by a classifier that has been constructed by 3 different recordings (training pieces). It remains to combine the different training pieces to improve the identification performance; we examine the use of low-complexity neural nets for this task in the next section.

5. Neural nets for wavelet dispersion vector classification

In this section we study the use of neural nets for assessing the similarity of a query wavelet dispersion vector to the vectors in the wavelet classifier matrix of the search machine. In particular, we examine a wide variety of combinations of different wavelets (for extracting the wavelet dispersion vector) and neural nets (for classifying the query vector).

5.1. Overview of examined neural nets

In signal communications and information sciences neural nets have been proposed to solve a variety of different problems, including pattern recognition and vector classification, especially when large amounts of novel data are to be processed with limited computational effort. As detailed in Section 4, we store the audio content description data in a wavelet dispersion vector $\vec{v}$. The search machine has to classify a vector entered as query using its wavelet classifier matrix $\mathbf{W}$, which contains all the wavelet dispersion vectors known to the search machine. Toward this end, we train a neural net with the wavelet dispersion vectors in the search machine's wavelet classifier matrix. We note that a trained neural net is defined by a set of parameters. Thus, after completing the training, the training data (wavelet classifier matrix) is no longer needed for answering queries.

There exist many different types of neural nets. Each of these types can be employed with different algorithms (see for instance [55,56]). Since there exists no generally valid recipe for choosing the most suitable neural net configuration for processing novel data, we examined different configurations to determine a reasonable methodology. In particular, we considered single-layer perceptron neural networks, backpropagation neural networks, and probabilistic radial basis neural networks.

The neural nets were trained with a minimum number of epoches (where one epoch corresponds to one traverse through all of the training wavelet dispersion vectors in $\mathbf{W}$) and neurons to allow an identification of known vectors with a success rate of approximatively 100%. In particular, the perceptron network was trained with a learning rate of 0.001 until a small performance error or a maximum of 50 epoches was reached. For the backpropagation network

we employed two tan-sigmoid function layers, each with 80 neurons, and a single neuron linear output layer. We trained the network for a maximum of 70 epoches. For backpropagation training, we employed the Fletcher-Reeves conjugate gradient algorithm. The probabilistic network is characterized by a bias parameter, which we set to 2×0.8326 (for $x = 0.836$ the radial basis function $f_{\mathrm{rdb}}(x)$ is 0.5). (A detailed description of the employed neural nets along with matlab code is provided in [52].)

5.2. Overview of examined wavelets

We examine 39 different wavelets, namely the Meyer wavelet, the Mexican hat wavelet, the Morlet wavelet, 7 types of symlets (modified Daubechies wavelets), 5 types of coiflets, 14 types of biorthogonal wavelets, and 10 types of Daubechies wavelets. In our performance plots we use abbreviated names to denote the wavelets. The number following the wavelet name denotes the wavelet order number. The biorthogonal wavelets have two order numbers, because they use a different wavelet mother function for reconstruction (first number) and decomposition (second number). Although, we only employ the wavelet decomposition, each of these so denoted wavelets describe different functions even if they are of the same order, because the hi-pass decomposition filters employ different sets of filter coefficients.

5.3. Performance evaluation set-up

For each of the 39 different considered wavelets we conduct the following evaluation. We perform a wavelet decomposition (wavelet scales 1, 3, 5, ..., 47) and construct the wavelet dispersion vectors as detailed in Section 4 for the first five seconds of each of the 128 (4 performances $\times$ 32 movements) audio pieces in our data base.

We consider a query scenario, where the user queries each of the 32 movements of one performance and the search machine knows the other three performances of these movements. In other words, the wavelet classifier matrix of the search machine contains the wavelet dispersion vectors of the other 3×32 movement performances. We discard the first two and the last two bars of each histogram (first two and last two elements of each row of wavelet dispersion matrix $\mathbf{D}$ of each known piece). Further, we discard the last two histograms of the higher scales. (Instead, the first two histograms of the lower scales can be discarded with no significant performance difference.) This process reduces the dimension of the wavelet dispersion matrix from 576×96 to 440×96. According to our experiments, a more extensive reduction resulted into a significantly lower performance. Therefore, we here do not report measurements with other reduction parameters. For processing this matrix with the neural nets, each vector of this matrix is normalized to zero mean and unit standard deviation.

5.4. Identification and retrieval performance results

Note that our audio data base of four performances of the same 32 movements allows for four combinations of one "query performance" and three "known performances". We denote each such combination with the query performance, e.g., the performance results for Men36 refer to the combination where each of the 32 movements of the Men36 performance is entered as query and the search machine knows the performances Men57, Hei52, and Mil75 of these movements. In Fig. 7, we report the performance as the percentage of correctly identified movements for the probabilistic radial neural net. (The backpropagation net gives relatively poor performance of 10–20% of correct identification, while the perceptron net gives somewhat better performance with 25–50% of correct identification, see [52] for details.) The points *nov* (novel) give the mean success percentage rate (piece identification) across the four query combinations, i.e., they give the mean value of the 4 points Men36, Men57, Hei, and Mil. In addition, the points *fgp* (fingerprint) give the success rates for scenarios where the queried performance is known to the search machine, i.e., has been used for the training.

We observe from Fig. 7 that the probabilistic radial neural net identifies known wavelet dispersion vectors (label *fgp*) with mean success rates of 100% over the entire range of the 39 different wavelets. This indicates that the neural net perfectly learned to identify the known wavelet dispersion vectors. The success rates for unknown pieces range from 62% to 76% for the Morlet wavelet function. We also observe that the performance trends are

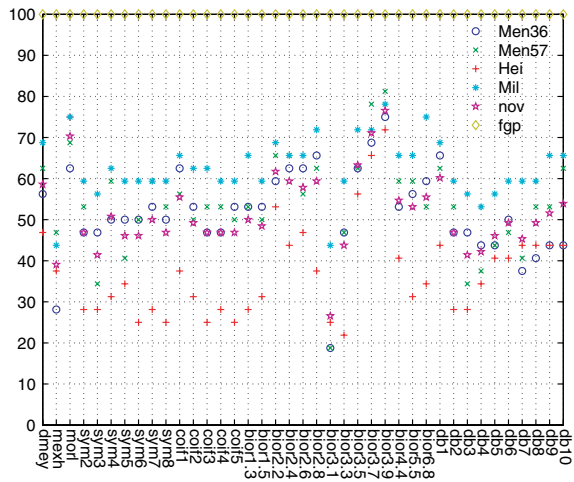


Fig. 7. Percentages of correctly identified pieces with probabilistic radial network for different wavelets. The morlet and bior3.9 wavelet achieve good generalization.

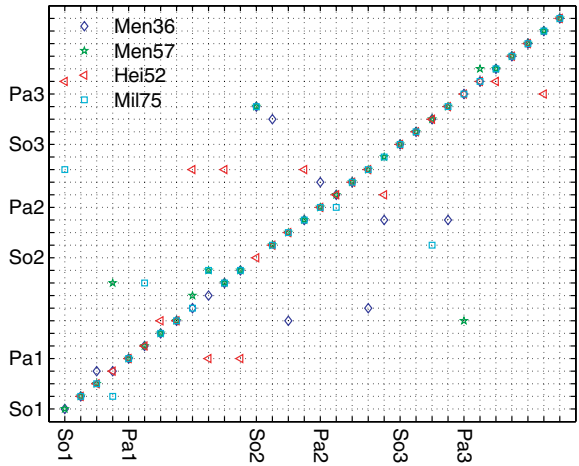


Fig. 8. Detailed unknown piece identification results for bior3.9 wavelet. A user query movement from one performance (x-axis) is assigned an answer (y-axis) by a search machine knowing the other three performances. 78% of the identifications are correct.

similar for all four performances. We furthermore observe that there is a second maximum in performance for the biorthogonal 3.9 wavelet, which achieves a mean identification success rate of 78%. (The plotted results are obtained with a wavelet decomposition using the wavelet scales 1, 3, 5, ..., 48. In more extensive evaluations [52] we found that with a smaller scale bandwidth using the wavelet scales 1, 3, 5, ..., 24, the percentages of correct identification are generally in the range from 20% to 50%, with the Morlet and biorthogonal 3.9 wavelets reaching identification performances in the 60–80% range.)

We provide a more detailed view of the identification results for the probabilistic net with the bior3.9 wavelet (using the wavelet scales 1, 3, 5, ..., 48) in Fig. 8. For user queries on each of the 32 movements from each of the four performances (Men36, Men57, Hei52, and Mil75), the figure gives the identification result. Note that each of the identification results in Fig. 8 is obtained by the probabilistic net with a wavelet classifier matrix containing the other three performances, whereas in Fig. 6, the identification is obtained by a maximum correlation measure with only one performance in the search machine's wavelet classifier matrix.

5.5. Summary of identification and retrieval methodology

Based on our results, we summarize our proposed methodology for the considered audio identification and retrieval problem as follows. First, a Morlet or biorthogonal (order 3.9) wavelet decomposition with the scales $s = 1, 3, 5, \ldots, 47$ is performed to obtain the wavelet coefficients. The coefficients are then

summarized in the wavelet dispersion vector $\vec{v}$. The dimension of this vector is reduced as detailed in Section 4.4 and the wavelet dispersion matrix **W** is formed, which represents the knowledge base (set of example vectors) of the search machine. Finally, a probabilistic radial neural net is trained with the set of example vectors and subsequently employed for answering queries.

6. Conclusion

In this paper we have considered a problem from the domain of music information retrieval, namely the problem of identifying the classical music composition of an unknown performance of the composition. We have developed and evaluated a wavelet transform based methodology for the identification of classical music compositions and the retrieval of other performances of the composition, which we expect to become an important service of next generation internet search machines. Our methodology consists of two main components: the characterization of the audio in the wavelet dispersion vector, and the assessment of the similarity of these vector by a neural net. The wavelet dispersion vector extracts the characteristic features of the audio content into a compact descriptor, which is accurate yet has good generalization properties.

The proposed methodology has been evaluated with four different performances of 32 different classical music movements. Thus, one class (for a given movement) in the search machine's knowledge base is constructed from only three example wavelet dispersion vectors. The system achieves a mean success rate of 78% for correctly identifying an unknown performance of a movement. This performance is quite promising, given that the different performances of the 32 movements differ significantly in time, frequency, audio environments, and audio recording quality. The historical recording of Y. Menuhin from 1936 with a correspondingly low recording quality is identified with a success rate of 75%.

Wavelet transform based approaches have to date received relatively little attention in the domain of music information retrieval, yet as demonstrated with the solution of the specific problem considered in this paper hold significant potential for the domain of music information retrieval. With the present study we provide ground work for the development of wavelet transform based techniques for other problems in the domain of music information retrieval, which are an exciting area for future work.

Acknowledgment

We are grateful to Prof. Thomas Sikora, head of the Communications Systems Group at the Technical University Berlin, Germany, for his support of this work.

References

[1] P. Cano, E. Batle, T. Kalker, J. Haitsma, A review of algorithms for audio fingerprinting, in: Proceedings of the IEEE International Workshop on Multimedia Signal Processing, St. Thomas, US Virgin Islands, December 2002, pp. 169–173.
[2] J.S. Downie, SIGIR 2003 workshop reports: Report on the panels and workshops of the music information retrieval (MIR) and music digital library (MDL) evaluation frameworks project, ACM SIGIR Forum 37 (2) (2003) 32.
[3] J. Foote, An overview of audio information retrieval, ACM/Springer Multimedia Systems Journal 7 (1) (1999) 2–10.
[4] T. Kalker, Applications and challenges for audio fingerprinting, in: Proceedings of 111th AES Convention, Watermarking versus Fingerprinting Workshop, December 2001.
[5] S.R. Subramanya, R. Simba, B. Narahari, A. Youssef, Transform-based indexing of audio data for multimedia databases, in: Proceedings of the IEEE International Conference on Multimedia Computing and Systems, October 1997, pp. 211–219.
[6] E. Wold, T. Blum, D. Keislar, J. Wheaton, Content-based classification, search, and retrieval of audio, IEEE Multimedia 3 (3) (1996) 27–36.
[7] M. Casey, MPEG-7 sound recognition tools, IEEE Transactions on Circuits and Systems for Video Technology 11 (6) (2001) 737–747.
[8] G. Guo, S. Li, Content-based audio classification and retrieval by support vector machines, IEEE Transactions on Neural Networks 14 (1) (2003) 209–215.
[9] H. Harb, L. Chen, A query by example music retrieval algorithm, in: Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '03), April 2003.
[10] S.Z. Li, Content-based audio classification and retrieval using the nearest feature line method, IEEE Transactions on Speech and Audio Processing 8 (5) (2000) 619–625.
[11] L. Lu, H.-J. Zhang, H. Jiang, Content analysis for audio classification and segmentation, IEEE Transactions on Speech and Audio Processing 10 (7) (2002) 504–516.
[12] G. Peeters, X. Rodet, Automatically selecting signal descriptors for sound classification, in: Proceedings of the International Music Computer Conference (ICMC '02), vol. 1, Goteborg, Sweden, 2002.
[13] G. Tzanetakis, P. Cook, Musical genre classification of audio signals, IEEE Transactions on Speech and Audio Processing 10 (5) (2002) 293–302.
[14] Z. Xiong, R. Radhakrishnan, A. Divakaran, T. Huang, Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03), vol. 5, Hong Kong, April 2003, pp. 628–631.
[15] T. Zhang, C.-C. Kuo, Hierarchical classification of audio data for archiving and retrieving, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '99, vol. 6, Phoenix, AZ, March 1999, pp. 3001–3004.
[16] B. Whitman, G. Flake, S. Lawrence, Artist detection in music with minnowmatch, in: Proceedings of the 2001 IEEE Workshop on Neural Networks for Signal Processing, Falmouth, MA, September 2001, pp. 559–568.
[17] T. Lambrou, P. Kudumakis, R. Speller, M. Sandler, A. Linney, Classification of audio signals using statistical features on time and wavelet transform domains, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '98, vol. 6, Seattle, WA, May 1998, pp. 3621–3624.
[18] H. Soltau, T. Schultz, M. Westphal, A. Waibel, Recognition of music types, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98), vol. 2, Seattle, WA, May 1998, pp. 1137–1140.

[19] K. Umapathy, S. Krishnan, S. Jimaa, Audio signal classification using time-frequency parameters, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '02), vol. 2, Lausanne, Switzerland, August 2002, pp. 249–252.
[20] E. Allamanche, J. Herre, O. Hellmuth, B. Froba, T. Kasten, M. Cremer, Content-based identification of audio material using MPEG-7 low level description, in: Proceedings of the International Symposium of Music Information Retrieval, October 2002.
[21] C. Burges, J. Platt, S. Jana, Extracting noise-robust features from audio data, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '02, vol. 1, Orlando, FL, May 2002, pp. 1021–1024.
[22] C. Burges, J. Platt, S. Jana, Distortion discriminant analysis for audio fingerprinting, IEEE Transactions on Speech and Audio Processing 11 (3) (2003) 165–174.
[23] J. Haitsma, T. Kalker, A highly robust audio fingerprinting system, in: Proceedings of 3rd International Conference on Music Information Retrieval, Paris, France, October 2002.
[24] J. Haitsma, T. Kalker, J. Oostveen, Robust audio hashing for content identification, in: Proceedings of 2nd International Workshop on Content Based Multimedia and Indexing, September 2001.
[25] J. Herre, E. Allamanche, O. Hellmuth, Robust matching of audio signals using spectral flatness features, in: Proceedings of IEEE Applications of Signal Processing to Audio and Acoustics Workshop, October 2001, pp. 127–130.
[26] A. Kimura, K. Kashino, T. Kurozumi, H. Murase, Very quick audio searching: introducing global pruning to the time-series active search, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, May 2001, pp. 1429–1432.
[27] F. Kurth, A ranking technique for fast audio identification, in: Proceedings of IEEE International Workshop on Multimedia Signal Processing, St. Thomas, US Virgin Islands, December 2002, pp. 186–189.
[28] F. Kurth, A. Ribbrock, M. Clausen, Identification of highly distorted audio material for querying large scale databases, in: Proceedings of AES 112th Convention, May 2002.
[29] G. Li, A.A. Khokhar, Content-based indexing and retrieval of audio data using wavelets, in: Proceedings of ICME, August 2000, pp. 885–888.
[30] M.K. Mihcak, R. Venkatesan, A perceptual audio hashing algorithm: a tool for robust audio identification and information hiding, in: Proceedings of the 4th Information Hiding Workshop, 2001, pp. 51–65.
[31] G. Richly, R. Kozma, F. Kovacs, G. Hosszu, Optimized soundprint selection for identification in audio streams, IEE Communications 148 (5) (2001) 287–289.
[32] S. Sukittanon, L.E. Atlas, J.W. Pitton, Modulation-scale analysis for content identification, IEEE Transactions on Signal Processing 52 (10) (2004) 3023–3035.
[33] A. Wang, An industrial strength audio search algorithm, in: Proceedings of International Symposium on Music Information Retrieval, Baltimore, MD, October 2003.
[34] J.-J. Aucouturier, M. Sandler, Using long-term structure to retrieve music: representation and matching, in: Proceedings of the International Symposium on Music Information Retrieval, October 2001.
[35] H. Nagano, K. Kashino, H. Murase, Fast music retrieval using polyphonic binary feature vectors, in: Proceedings of IEEE International Conference on Multimedia and Expo, August 2002, pp. 101–104.
[36] Y. Shiu, C.-H. Yeh, C.-C. J. Kuo, Audio fingerprint extraction for content identification, in: Proceedings of SPIE Internet Multimedia Management Systems IV, vol. 5242, November 2003, pp. 55–64.
[37] J. Foote, ARTHUR: retrieving orchestral music by long-term structure, in: Proceedings of International Symposium on Music Information Retrieval, 2000.
[38] C. Yang, Content-based music retrieval on acoustic data, Ph.D. Dissertation, Stanford University, August 2003.

[39] E. Pampalk, A. Rauber, D. Merkl, Content-based organization and visualization of music archives, in: Proceedings of ACM Multimedia, December 2002.
[40] X. Zhao, Y. Zhuang, J. Liu, F. Wu, Audio retrieval with fast relevance feedback based on constrained fuzzy clustering and stored index table, in: Proceedings of Advances in Multimedia Information Processing—PCM 2002: Third IEEE Pacific Rim Conference on Multimedia, Lecture Notes in Computer Science, vol. 2532, December 2002.
[41] S. Pfeiffer, S. Fischer, W. Effelsberg, Automatic audio content analysis, in: Proceedings of the fourth ACM International Conference on Multimedia, 1997, pp. 21–30.
[42] S.R. Subramanya, A. Youssef, Wavelet-based indexing of audio data in audio/multimedia databases, in: Proceedings of Multi-Media Database Management Systems, August 1998, pp. 46–53.
[43] M.G. Albanesi, M. Ferretti, A. Giancane, Time-frequency decomposition for analysis and retrieval of 1-D signals, in: Proceedings of IEEE International Conference on Multimedia Computing and Systems, June 1999, pp. 974–978.
[44] C.-S. Lu, Audio fingerprinting based on analyzing time-frequency localization of signals, in: Proceedings of IEEE International Workshop on Multimedia Signal Processing, St. Thomas, US Virgin Islands, December 2002, pp. 174–177.
[45] J. Hughes, R. Gaborski, K. Hsu, A. Titus, An auditory classifier employing a wavelet neural network implemented in a digital design, in: Proceedings of 14th Annual IEEE International ASIC/SOC Conference, September 2001, pp. 8–12.
[46] A. Czyzewski, Some methods for detection and interpolation of impulsive distortions in old audio recordings, in: Proceedings of IEEE Applications of Signal Processing to Audio and Acoustics Workshop, October 1995, pp. 139–142.
[47] B. Tovar-Corona, M. D. Hind, J. N. Torry, R. Vincent, Effects of respiration on heart sounds using time-frequency analysis, in: Proceedings of Computers in Cardiology, September 2001, pp. 457–460.
[48] Y.-R. Chien, S.-K. Jeng, An automatic transcription system with octave detection, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, May 2002, pp. 1865–1868.
[49] A. Busch, W.W. Boles, Texture classification using wavelet scale relationships, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, May 2002, pp. 3584–3587.
[50] M.N. Do, M. Vetterli, Texture similarity measurement using Kullback-Leibler distance on wavelet subbands, in: Proceedings of IEEE International Conference on Image Processing, September 2000, pp. 730–733.
[51] J. Nam, A.E. Cetin, A.H. Tewfik, Speaker identification and video analysis for hierarchical video shot classification, in: Proceedings of IEEE International Conference on Image Processing, September 1997, pp. 550–553.
[52] S. Rein, M. Reisslein, Identifying the classical music composition of an unknown performance with wavelet dispersion vector and neural nets, Technical Report, Arizona State University, Department of Electrical Engineering, December 2004. Available from: <http://www.fulton.asu.edu/~mre>.
[53] I. Daubechies, Ten Lectures on Wavelets, Society for Industrial & Applied Mathematics, 1992.
[54] R. Jain, The Art of Computer Systems Performance Analysis, Wiley & Sons, 1991.
[55] M. Misiti, Y. Misiti, G. Oppenheim, J.-M. Poggi, Wavelet Toolbox User's Guide, second ed., MathWorks Inc., Natick, MA, 2002. Available from: <http://www.mathworks.com>.
[56] H. Demuth, M. Beale, Neural Network Toolbox User's Guide, fourth ed., MathWorks Inc., Natick, MA, 2002. Available from: <http://www.mathworks.com>.