TOWARDS AN INVERTIBLE RHYTHM REPRESENTATION

AggelosGkiokas and Vassilis Katsouros

Institute for Language and Speech,
Processing / R.C. Athena
Athens, Greece
{agkiokas,vsk}@ilsp.gr

Stefan Lattner and Arthur Flexer

The Austrian Research Institute for Artificial
Intelligence,
Vienna, Austria
{stefan.lattner,arthur.flexer}@ofai.at

George Carayannis

National Technical University of Athens
Athens, Greece
carayan@softlab.ece.ntua.gr

ABSTRACT

This paper investigates the development of a rhythm representation of music audio signals, that (i) is able to tackle rhythm related tasks and, (ii) is invertible, i.e. is suitable to reconstruct audio from it with the corresponding rhythm content being preserved. A conventional front-end processing schema is applied to the audio signal to extract time varying characteristics (accent features) of the signal. Next, a periodicity analysis method is proposed that is capable of reconstructing the accent features. Afterwards, a network consisting of Restricted Boltzmann Machines is applied to the periodicity function to learn a latent representation. This latent representation is finally used to tackle two distinct rhythm tasks, namely dance style classification and meter estimation. The results are promising for both input signal reconstruction and rhythm classification performance. Moreover, the proposed method is extended to generate random samples from the corresponding classes.

1. INTRODUCTION

Invertible signal transformations play an essential role in the music processing field. From everyday use of music, like adjusting the equalizer of a stereo system up to the process of sound production and mixing, the transformation of audio is a crucial step for most of these applications. Most of them usually rely on the well-studied Short Time Fourier Transform (STFT) (a.k.a. Gabor Transform), which offers a very simple and intuitive way to edit audio signals. However, the main limitation of STFT is the linear spacing of frequency bins, which can be a drawback for music analysis systems, since in this case the energy concentrates on frequencies in a logarithmic scaling. The Constant Q Transform (CQT), firstly introduced by Brown [1], although it has been an alternative to STFT for almost three decades, it didn't get as much attention; not only due to its computational cost, but mostly because it is irreversible. In [2] authors presented a computational framework for a CQT with almost perfect reconstruction, and just one year later, a perfect reconstruction was achieved [3].

However, most of the signal analysis and transformation applications involve transformations based solely in time slices, i.e. observing the spectral content on a segment basis. Transformations in the other dimension (i.e. across time for each frequency bin) have not yet been studied. Such transformations would have an impact on the temporal organization of the input signal, which in the case of music is very closely related to rhythm.

The aim of this paper is to introduce a rhythm representation that can be exploited to tackle rhythm related tasks, and it is invertible, so that a rhythmically relevant signal can be reconstructed from it. Such a representation is useful, since it can provide a better insight of rhythm related features.

The rest of the paper is organized as follows. Section 2 will present current rhythm analysis systems, and will highlight the limitations of such methods towards developing an invertible rhythm representation. Section 3 will present an overview of the proposed method, while algorithmic details will be described in Section 4. An evaluation of the model's reconstruction ability and its rhythm classification performance will be presented in Section 5. Section 6 concludes this paper with a discussion and considerations for future work.

2. BACKGROUND

Most of the rhythm analysis methods involve a two step-process framework. Firstly, from a Time-Frequency representation (either STFT or CQT) of the input signal, the extraction of spectral characteristics through time (as for example frequency band energies), hereafter referred to as *accent features*, takes place. Accent features are then processed in a periodicity analysis step, such as the autocorrelation function [4-6], convolution with a bank of resonators [7-10], considering Inter-Onset-Interval histograms [11] or just by taking the Discrete Fourier Transform (DFT) of each feature [12]. The result of such an analysis is usually referred to as *spectral rhythm patterns* or *periodicity function* (PF).

Regarding the beat-tracking methods, most of them include an additional step, which combines spectral (periodicity function) with temporal (accent features) information to infer beat positions. A notable exception is the beat-tracking method presented in [5], where tempo is induced from the beat activation function.

Although the accent feature extraction step is a *lossy* transformation, it can be considered as being *lossless* in the context of

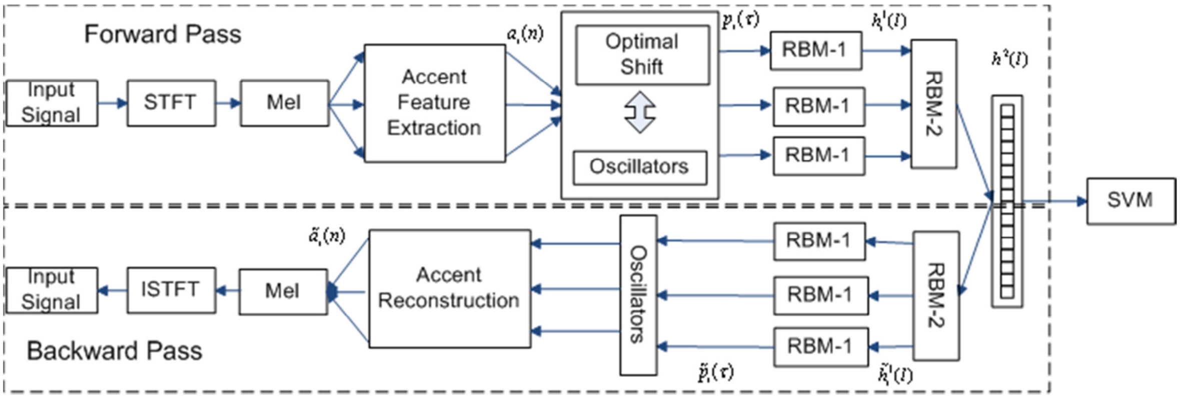

Figure 1. *Overview of the proposed method.*

rhythm analysis, as it is possible to reconstruct audio that preserves the rhythmic information of the original signal. Such a result was demonstrated in [7], where the envelopes of sub-band energies were used to modulate a noise signal. The derived signal had a rhythmic characteristic very similar to that of the original audio.

However, this is not the case for the periodicity analysis step methods. While most of them preserve spectral information of the accent features, phase or timing information gets lost, which makes it impossible to get any meaningful reconstruction of the accent features and consequently block the way to an invertible rhythm transformation.

Regarding beat-tracking methods, although beat-inference can be considered as an inverse process from the spectral to the time domain, it cannot be viewed as an inverse transformation. An exception can be found in [13], where a beat-tracking method was proposed based on the Non-Stationary Gabor Transform (NSGT) [14]. NSGT was applied to the accent features and the derived complex valued periodicity function was further processed by resampling and by peak-selection. The inverse NSGT was consequently applied to the processed periodicity function to get an estimate of the beat positions. Although this approach focuses on inferring the beats, it provides a framework for reconstructing rhythmically meaningful components of the accent features.

In this paper, we present a method for a lossy invertible rhythm transformation, which consists of two key features. Firstly, a periodicity analysis step (i.e. computing the PF) which allows for an imperfect but sufficient reconstruction of the accent features takes place. Second, we deploy a network of Restricted Boltzmann Machines (RBM) [15] on the periodicity function. RBMs are energy based, stochastic Neural Networks with a visible input layer and a hidden layer, which are able to learn the probability distribution over the training data. RBMs are generative models, able to reconstruct input data given the states of the hidden units, to sample from the learned distribution and to extract meaningful features from the data. The hidden layer provides a latent representation of the input data and can be used to tackle rhythm related tasks.

3. METHOD OVERVIEW

Figure 1 presents an overview of the proposed method. The input signal is decomposed into I frequency bands and the corresponding accent features $a_i[n]$ are extracted from each band. Next, a periodicity analysis takes place for each accent feature to extract a periodicity function $p_i[\tau]$ for each band. The periodicity analysis design has focused on preserving both amplitude and phase information of the target periodicities, such that it is possible to reconstruct the accent features from it. The periodicity functions derived from all accent features are then fed to train a single RBM (RBM-1). Actually, the RBM-1 learns a distribution over all periodicity functions of the training dataset, irrespectively from which frequency band of the signal the accent feature was computed. Afterwards, the outputs of the RBM-1, denoted by $h_i^1[l]$, are concatenated to a single vector, which is used as an input to the 2nd single-layer RBM (RMB-2). The motivation behind this architecture is that RBM-1 will learn a distribution over the individual PFs, while the RBM-2 will learn a distribution over combinations of PFs from the I frequency bands. The output of the RBM-2, denoted by $h^2[l]$, can be used as an input to a discriminative method such as a Support Vector Machine (SVM) to tackle the rhythmic task under consideration. The whole network is capable of reconstructing the accent features starting from the top-level RBM output $h^2[l]$, and then calculating sequentially $\tilde{h}_i^1[l]$, $\tilde{p}_i[\tau]$ and $\tilde{a}_i[n]$, as shown in the bottom part of Figure 1. Finally, from the reconstructed accent features $\tilde{a}_i[n]$, it is possible to derive an audio signal that preserves the most dominant rhythmic characteristics of the original audio signal.

4. METHOD DETAILS

4.1. Extracting Accent Features

The input signal is downsampled to 22.05 kHz and the STFT is computed with a sliding window of 1024 samples and half overlap between successive windows. From the amplitude spectrum $\mathbf{X}$, I band energies $e_i[n]$, $i = 1..I$ are computed with equally spaced triangular filters and half overlap in the mel-scale. Formally we can write

$$\mathbf{E} = \mathbf{X} \cdot \mathbf{M} \tag{1}$$

where $\mathbf{E} = [\mathbf{e}_1|\mathbf{e}_2|..\mathbf{e}_I]$ and $\mathbf{M}$ is the filter matrix. Then, the logarithm of the filter energies are differentiated to extract the accent feature sequence $a_i[n]$, $i = 1..I$, where n denotes the frame index. The use of logarithms followed by differentiation to extract the $a_i[n]$ is in line with [16], as it indicates relative changes w.r.t. the features' level. Each $a_i[n]$ is segmented by a sliding square window of N frames length with half hop size, resulting in approximately 12s of audio. Finally, the segments $a_i^s[n]$ are normalized w.r.t. their mean value and standard deviation.

To reconstruct an audio signal from the accent features $\tilde{a}_i^s[n]$ $\tilde{a}_i^s[n]$ the inverse pipeline is applied to $\tilde{a}_i^s[n]$, which can be summarized in the following equation:

$$\tilde{e}_i^s[n] = \exp\left(\sum\nolimits_{m=1}^{n}(\tilde{a}_i^s[m]\cdot\sigma_{a_i^s} + \mu_{a_i^s})\right) \tag{2}$$

where $\mu_{a_i^s}$, $\sigma_{a_i^s}$ denote the mean and standard deviation of $\mathbf{a}_i^s$. Afterwards, the residual spectrogram $\tilde{\mathbf{X}}$ from is reconstructed from $\tilde{\mathbf{E}}$ as

$$\tilde{\mathbf{X}} = \tilde{\mathbf{E}}\mathbf{M}^{\mathbf{T}} \circ [e^{j\angle\mathbf{X}}] \tag{3}$$

where $\circ$ denotes the Hadamard product and $\angle\mathbf{X}$ denotes the phases of the initial spectrogram $\mathbf{X}$. It should be noted that if the number of bands $I \ll M/2$ where M is the FFT size, most of the harmonic content is truncated.

4.2. Periodicity Analysis

A periodicity function or a periodicity vector (PF) is an essential rhythmic representation of the accent features. Its domain is frequency of beats per minute or Hertz with typical values ranging from 0.5 Hz (30 b.p.m.) up to 5 Hz (300 b.p.m.) [17] and its value represents the salience of these periodicities. One of the most important contributions of the proposed method is to derive a periodicity function that (a) is able to reconstruct the accent features and (b) is an efficient representation which can be used to learn higher level rhythm features with Restricted Boltzmann Machines.

A typical family of periodicity analysis methods may comprise of the convolution of each accent feature sequence with a bank of resonators $\mathbf{o}_\tau$. Each $\mathbf{o}_\tau$ has an inherent oscillation frequency that corresponds to tempo τ. The maximum value of the convolution within a certain window for each accent-oscillator pair represents the salience of tempo τin this window for the accent feature $\mathbf{a}$, i.e. $p_a(\tau) = \max(\mathbf{o}_\tau * \boldsymbol{a})$, where $p_a(\tau)$ denotes the periodicity vector.

If $\mathbf{o}_\tau$ and $\mathbf{a}$ have the same length, an alternative calculation of a periodicity function $p_a(\tau)$ is given by

$$p_a(\tau) = \max_k(\mathbf{a}^{\mathbf{T}}\mathbf{o}_\tau^k) \tag{4}$$

where $\mathbf{o}_\tau^k$ denotes the circular shift of $\mathbf{o}_\tau$ by k samples. In other words, $p_a(\tau)$ corresponds to the value of the "best fit" between $\mathbf{o}_\tau$ and $\mathbf{a}$. The objective is to derive an invertible periodicity analysis step, i.e. it should be possible to reconstruct $\mathbf{a}$ *solely* by $p_a(\tau)$. For the oscillators we consider the derivative of the resonators proposed in [18], as follows

$$o_\tau(n) = d_L(n) * (1 + \tanh(\gamma\cdot(\cos(2\pi\omega_\tau n) - 1))) \tag{5}$$

where d_L denotes a non-causal differential filter of order L, ω_τ is the frequency corresponding to tempo τ and γ is called the output gain.

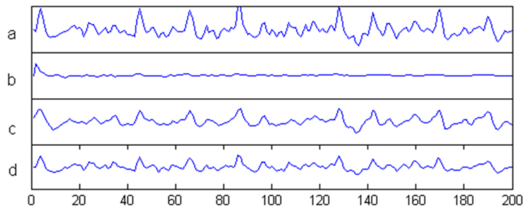

Figure 2. *Reconstruction of accent features from the periodicity vector.*

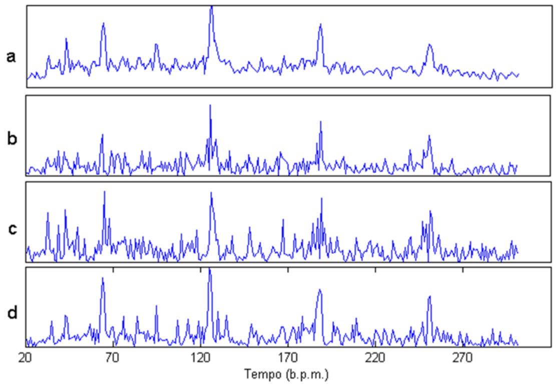

Figure 3. *The "ideal" PF (a) along with the derived PF with two random shifts k (b) and (c), and the PF computed with the proposed method (d).*

Let us denote $\mathbf{O}^{\mathbf{k}} = [\mathbf{o}_{\tau_1}^{k_1}|\mathbf{o}_{\tau_2}^{2}|..|\mathbf{o}_{\tau_M}^{k_M}]$ where $\mathbf{k}$ is a vector containing the shifts of the oscillators of the target tempi and $\mathbf{O}^k = [\mathbf{o}_{\tau_1}^k|\mathbf{o}_{\tau_2}^k|..|\mathbf{o}_{\tau_M}^k]$ where k is a scalar, i.e. considering a constant shift for all $\mathbf{o}_\tau$. A naive choice to reconstruct the accent feature *solely* from $\mathbf{p}_a$ could be

$$\tilde{\mathbf{a}} = \mathbf{O}^0\mathbf{p}_a^{\mathbf{T}} \tag{6}$$

that is to consider zero shifts for all oscillators. However, this approach has proven to result in a poor quality of accent features, because the phase, i.e. the time offsets k_τ which maximize $\mathbf{a}^{\mathbf{T}}\mathbf{o}_\tau^k$ ($k_\tau = \mathrm{argmax}_k(\mathbf{a}^{\mathbf{T}}\mathbf{o}_\tau^k)$), are needed. To demonstrate this, Figures 2 (a) and (b) show the original and the reconstructed accent features $\mathbf{a}, \tilde{\mathbf{a}}$ respectively and it is clear that the temporal information of $\mathbf{a}$ has been lost. Figure 2(c), shows the reconstruction of the accent features, when the time offsets k_τ of each oscillator are considered in the calculation of Eq. 6.

On the other hand, if we compute the PF as $\bar{p}_a(\tau) = \mathbf{a}^{\mathbf{T}}\mathbf{o}_\tau$ instead as in Eq. 6, i.e. by setting k=0 for all oscillators, the reconstructed accent signal derived from the inverse operation (i.e. $\bar{\mathbf{a}} = \mathbf{O}\bar{\mathbf{p}}_a^{\mathrm{T}}$), is much closer to the original one, and the phase information is preserved, as shown in Figure 2 (d). The same result would hold for any constant shift k. In other words, the oscillator bank $\mathbf{O}^k$ is able to provide a rough reconstruction of $\mathbf{a}$ if all the individual shifts k_τ are constant. In other words, $\mathbf{O}^k$ can be considered as an approximate basis of the accent features. Note that although $\bar{p}_a(\tau)$ is used for the reconstruction of $\mathbf{a}$, the actual periodicity function representing the rhythm saliences is its absolute value $|\bar{p}_a(\tau)|$.

However, $\bar{p}_a(\tau)$ proves to be a poor periodicity estimator of the accent signal $\mathbf{a}$ if it is computed for an arbitrary shift k. Figure 3 (a) shows the periodicity function derived from Eq. 4 and Figure 3 (b) and (c) show $|\bar{p}_a(\tau)|$ for different time shifts k of the oscillators. It is clear that although $p_a(\tau)$ and $|\bar{p}_a(\tau)|$ exhibit peaks at the same positions, the amplitude of these peaks are very different. Moreover, the derived PF is sensitive to time shifts of the oscillators, or equivalently to time shifts of the accent signal **a**. To sum up, we can deduce that there is a trade-off between a good periodicity function and a good reconstruction of the accent signal.

To overcome this limitation and have an efficient PF while at the same time keeping the reconstruction capabilities almost unaffected, we propose the following method. Let $a_i[n]$ denote the accent features of the $i = 1..I$ band energies. Firstly, an "ideal" periodicity function is computed as in Eq. 4, which is then averaged for all frequency bands I

$$\hat{p}_a(\tau) = \sum_i \left| p_{a_i}(\tau) \right| = \sum_i \left| \max_k (\mathbf{a}_i^{\mathrm{T}} \mathbf{o}_\tau^k) \right| \tag{7}$$

Next, the *actual* PF $p_a^k(\tau)$ is computed for a number of time shifts k as

$$p_a^k(\tau) = \sum_i \mathbf{a}_i^{\mathrm{T}} \mathbf{o}_\tau^k, \;\; \mathbf{p}_a^k = \sum_i \mathbf{a}_i^{\mathrm{T}} \mathbf{O}_\tau^k, \tag{8}$$

Note that k is the same for all the oscillators and for all accent bands i. Finally, the time shift k_0 that exhibits the higher cosine similarity between $\mathbf{p}_a^k$ and $\widehat{\mathbf{p}}_a$ is chosen:

$$k_0 = \arg\max_k \left(\frac{|\mathbf{p}_a^k|^{\mathrm{T}} \hat{\mathbf{p}}_a}{\left\| \mathbf{p}_a^k \right\| \left\| \hat{\mathbf{p}}_a \right\|} \right) \tag{9}$$

The corresponding PF $p_a^{k_0}(\tau)$ can be considered as being the best approximation of $\hat{p}_a(\tau)$. The resulting periodicity function for each $\mathbf{a}_i$ is then computed by setting $k = k_0$ for all oscillators:

$$\mathbf{p}_{a_i} = \mathbf{a}_i^{\mathrm{T}} \mathbf{O}_\tau^{k_0} \tag{10}$$

Finally, the accent features are reconstructed as

$$\hat{\mathbf{a}}_i^{\mathrm{T}} = \mathbf{p}_{a_i} \left(\mathbf{O}_\tau^{k_0} \right)^{\mathrm{T}} \tag{11}$$

The choice of the same k on the calculation of Eq. 8 ensures a good balance between accent feature reconstruction and periodicity function approximation. Since the time shift k_0 is the same for all oscillators, the reconstruction accents will be close to $\mathbf{a}_i$ (as in Fig. 2d). In the case where only the periodicity functions $\mathbf{p}_{a_i}$ without the time shift k_0 are known, then the reconstructed accent features will differ by a time-shift value of k_0. Moreover, by choosing the same shift when computing the PF for the different bands i, there is no need to keep the phase information of the oscillators in order to represent the PF for all energy bands, and at the same time the derived PF is as similar as possible to the "ideal" PF. With this method, both frequency (rhythm analysis) capability and temporal (reconstruction) capability are preserved. Periodicity analysis is shift-invariant to the accent signal, while the residual signal is shifted by a constant value k_0 which is already known. Figure 3 (d) shows the PF derived with this method, which is closer to the "ideal" PF. Quantitative results of this analysis will be given in Section 5.

4.3. Learning Features with Restricted Boltzmann Machines

In a previous work [19] a Restricted Boltzmann Machine was trained on the periodicity function. The features learned by the RBM were used to successfully tackle a variety of rhythm analysis tasks. In this paper, we deploy a similar approach to exploit the periodicity function described in the previous Sections. The main difference is that instead of the absolute values of the PF (as mentioned in the previous Section), we consider the actual values of the PF. This choice relies on the reconstruction prerequisite of the proposed method. We expect that the salience of periodicities corresponding to large negative values of the PF will be preserved in the features learned by the RBM network.

The motivation of using RBMs instead of other methods such as the Auto-Encoder is firstly that RBMs are proved to derive better features, and secondly they are generative models, a property that is exploited in this paper as will be described later in Section 5.5.

If the reconstruction error of the RBM network is small enough, both periodicity and timing information are preserved. In the first step, the PF extracted for all excerpts denoted by $p_i^m[\tau]$, where m, i indicate the instance and the band respectively are all grouped into a single training dataset D_1 irrespectively of the band index i. $D_1$is then used to train the first RBM. After RBM-1 is trained, for a target excerpt with periodicity functions $p_i^m[\tau]$, the corresponding RBM-1 outputs $\mathbf{h}_{i,m}^1$ of all bands i=1...I are concatenated to a single vector to form the training dataset D_2. If we denote the dimension of the hidden layer by N_1 the feature dimension of D_2 is $I \cdot N_1$. D_2 is subsequently used to train RBM-2, with $I \cdot N_1$ visible and N_2 hidden units. While RBM-1 is dedicated to learn the distribution of the individual accent features, RBM-2 learns an overall distribution across all I band energies. We denote the output of the RBM-2 as $\mathbf{h}_m^2$.

4.4. Rhythm Classification

To demonstrate the discriminative potential of the extracted features, we have deployed an SVM classifier with a Radial Basis Kernel for tackling meter estimation and dance style classification tasks. The LIBSVM [20] implementation was used. Given a dataset $\{\mathbf{a}_i^m\}_{m=1..M}$with a predefined set of classes C, the corresponding RBM outputs denoted by $H = \{\mathbf{h}_m^2\}_{m=1..M}$ are computed and used as the feature space on which we employ SVM classifiers using the one-to-one multiclass approach. All datasets were split to 10 folds such that the distribution of the classes is the same for all folds and a 10-fold cross validation approach was used. 8 folds were used for training, one fold for testing and one for validation. SVM was trained on a segment basis and the classification decision was taken with a majority vote over each test excerpt.

5. EVALUATION

5.1. Evaluation Setup

The proposed method was evaluated on five datasets for two distinct tasks. The first task is related to Meter estimation, where experiments were conducted on the Essen Folk Song database [21] and Finish Folk Collections database [22] comprising of 6207and 7735 melodies in MIDI format respectively. MIDI files were synthesized to 22 kHz audio and ground-truth time signature information was extracted from the MIDI files. The second task is Dance Style Classification performed on the ballroom dataset [23], which consists of audio samples 30s long of 8 dance rhythm classes.

Apart from recognition performance, in order to demonstrate the inversion capabilities of the proposed model, the Speedo [24] and GTZAN [25] datasets which mainly consist of popular music audio, were used.

5.2. Network Training

The RBM network was trained on a subset of the Million Song Dataset (MSD) [26] consisting of 130.000 excerpts. This results in approximately 4.5 million instances for training the RBM-1 and 900K instances for training the RBM-2. The training instances for the RBM-1 were normalized to zero mean and unit standard deviation for each dimension and we used Gaussian Visible and Noisy Rectified Linear Hidden Units (NReLU) [27]. For the RBM-2, both layers consist of NReLUs. The number of accent bands was chosen $I = 5$, and the periodicity function was calculated for $\tau_{\min} = 20$, $\tau_{\max} = 300$ with a $\delta\tau = 1$step. The number of hidden units for RBM-1 and RBM-2 were chosen $N_1 = 300$ and $N_2 = 500$ respectively, resulting in an (281x300) architecture for RBM-1 and in an (1500x500) for RBM-2. Both RBMs were trained using Contrastive Divergence-1 [15].

5.3. Reconstructing Accent Signals from the Periodicity Vector

In the inverse pipeline presented in Figure 1 for reconstructing $\tilde{a}_i[n]$ from $\mathbf{h}^2$, the errors for each step are accumulated yielding the final reconstruction error. This Section will present a deeper insight into the reconstruction errors of the accent features, produced by the periodicity analysis step and the details of the method presented in Section 4.2 will be established experimentally. Moreover, PF approximation (Eq. 10) along with RBM and overall network reconstruction errors will be reported.

Let $\hat{a}_i[n]$ denote the reconstruction of accent feature $a_i[n]$ solely from the periodicity function as computed by Eq. 8, and let $\tilde{a}_i[n]$ denote the accent feature reconstructed from the whole network. The differences between $(\hat{a}_i[n], a_i[n])$, $(\tilde{a}_i[n], \hat{a}_i[n])$ and $(\tilde{a}_i[n], a_i[n])$ correspond to the reconstruction errors of the accent features introduced by the periodicity analysis step (RBMs are ignored), the RBM influence on accent signal reconstruction and the whole architecture respectively. Regarding periodicity function approximation, the difference of the "ideal" PF $\hat{p}_a(\tau)$ (Eq. 7) and $p_a(\tau) = \sum_i |p_{a_i}(\tau)|$ (Eq. 10) corresponds to the error introduced by the periodicity analysis step. If $\tilde{p}_{a_i}(\tau)$ denotes the *pf* derived from the RBM reconstruction, then the difference between $\tilde{p}_{a_i}(\tau)$ and $p_{a_i}(\tau)$ corresponds to the reconstruction error of the PF introduced by the RBM network, while the difference between $\hat{p}_a(\tau)$ and $\tilde{p}_a(\tau) = \sum_i \tilde{p}_{a_i}(\tau)$ can be viewed as a measure of the overall approximation of $\hat{p}_a(\tau)$ by the whole network.

To quantify the performance of the proposed method w.r.t. accent feature reconstruction and PF approximation we consider the cosine similarity $R_{\mathbf{x}\tilde{\mathbf{x}}} = \mathbf{x}^T\tilde{\mathbf{x}}/\|\mathbf{x}\|\|\tilde{\mathbf{x}}\|$ as an evaluation measure. The choice of the cosine measure instead of other conventional measures such as the L^2 norm relies on the fact that cosine measure is more intuitive with respect to the proposed context. For example, the cosine similarity between $a_i[n]$ and an amplified version of it $A \cdot a_i[n]$ will be 1, which is not the case for any norm. The same holds for comparing the PFs as well.

Dataset	Periodicity		RBM Error		Overall Error	
	$R_{\hat{a}a}$	$R_{\hat{p}p}$	$R_{\hat{a}\tilde{a}}$	$R_{\tilde{p}p}$	$R_{\tilde{a}a}$	$R_{\tilde{p}\hat{p}}$
Essen	74.0	89.7	94.7	91.2	63.6	87.0
F-Folk	72.0	90.5	92.5	89.9	59.4	87.8
Genres	78.7	90.4	94.9	92.6	69.3	87.9
Speedo	79.5	90.4	95.1	92.8	70.4	87.9
Ballroom	78.6	89.8	95.2	92.9	69.9	87.6

Table 1. *Detailed reconstruction approximation results of the proposed method. Values correspond to* $100 \cdot R_{\mathbf{x}\tilde{\mathbf{x}}}$

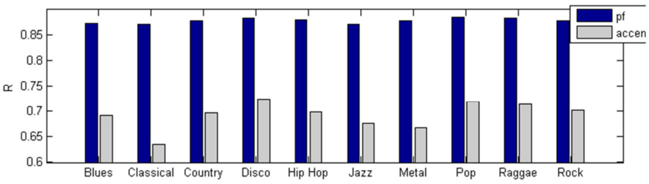

Figure 4. *Reconstruction approximation of accent features and periodicity vector across all genres on the TGZAN dataset.*

Table 1 summarizes the results for the accent feature and PF reconstruction for the periodicity analysis step, the RBM network and the whole architecture for all evaluation datasets described in previous section. Values correspond to mean values for each dataset. Regarding the approximation of the "ideal" periodicity function ($R_{\hat{p}p}$value) by finding a single optimal time shift for all oscillators as described in Eq. 7 and Eq. 10, a value around 0.9 for cosine similarity was achieved for all datasets. RBM's reconstruction of $\tilde{p}_{a_i}(\tau)$ is around 0.92 for all datasets while the similarity of $\tilde{p}_{a_i}(\tau)$ with the $\hat{p}_a(\tau)$ is above 0.87 for all datasets. In other words, the network learns a latent representation of the periodicities $\mathbf{h}^2$ from which we can reconstruct a PF that is very close to the "ideal PF". It is also noteworthy that the reconstruction rates of the PFs are almost the same for every dataset, even for those which stem from midi files.

Regarding *accent feature* reconstruction rates of the proposed method, the 1st column of Table 1 indicates a larger reconstruction error of the accent features due to the inverse PF analysis step (Eq. 11). Although the accent feature reconstruction error introduced by the RBMs, i.e. the difference between accent features reconstructed from $p_{a_i}(\tau)$ and $\tilde{p}_{a_i}(\tau)$, the relatively small $R_{\hat{a}a}$ has an impact on the final reconstruction capabilities of the whole network $R_{\tilde{a}a}$ which is around 0.7 for audio and 0.6 for the midi datasets.

To get a better insight into the reconstruction errors, Figure 4 shows $R_{\tilde{a}a}$ and $R_{\tilde{p}\hat{p}}$ for each genre of the GTZAN dataset. As expected, the reconstruction error of the accent features is larger for some genres, such as classical, metal and jazz, while it is better for some genres with more steady rhythm, such as disco and pop. On the other hand, it is noteworthy that the periodicity function approximation value $R_{\tilde{p}\hat{p}}$ is almost the same for all genres. Audio examples of the reconstructions for all excerpts of the GTZAN can be found in[1]. For a direct comparison to the original tracks, the audio files reconstructed from the initial accent features (Eq. 3) are also provided.

5.4. Classifying to Rhythm Classes

5.4.1. Meter Estimation

To report comparable results with other methods [28] for the meter estimation task, 9 meter classes were considered, meters 2/4, 3/2, 3/4, 3/8, 4/1, 4/2, 4/4, 6/4, 6/8 were chosen for the Essen Collection and meters 2/4, 3/2, 3/4, 3/8, 4/4, 5/2, 5/4, 6/4, 6/8 for the Finish Folk Collection. The proposed method achieved a classification accuracy of 80.7% and 75% for the Essen and Finish Folk song collections on a track basis. For a better insight into

[1] http://mir.ilsp.gr/invertible_rhythm.html

	Predictions								
	2/4	3/2	3/4	3/8	4/1	4/2	4/4	6/4	6/8
2/4	**83**	0	4	1	0	0	12	0	1
3/2	0	**42**	9	0	0	31	18	0	0
3/4	7	0	**83**	0	0	0	10	1	0
3/8	36	0	12	**19**	0	0	2	0	31
4/1	0	0	0	0	**86**	14	0	0	0
4/2	0	1	0	0	2	**92**	4	0	0
4/4	5	0	3	0	0	1	**91**	0	0
6/4	0	0	61	0	0	0	13	**26**	0
6/8	5	0	6	2	0	0	1	0	**86**

Table 2. *Confusion matrix for the meter estimation task on the Essen Folk Song Collection. Values are percentages (%). Rows correspond to ground truth and columns to predictions.*

	Essen		Finish Folk	
Method	9 class	2 class	9 class	2 class
Proposed	80.7	89.2	75.0	93.8
Toivianen[28]	83.2	95.3	68	96.4
Eck [6]	-	90	-	93

Table 3. *Comparison with other reference methods for the 2-class and 9-class meter estimation.*

the classification performance, the confusion matrix of the classification results for the Essen dataset is presented in Table 2. Most of the meter classification errors are for similar meters, as for example in the case of 3/8 examples, where 43% of the cases were classified either as 6/8 or ¾.

If we consider only two broad meter categories, i.e. *duple* (e.g. 2/4, 4/4, 4/8 etc. meters) and *triple/compound* (e.g. 3/8, 6/8, 9/8, 3/2 etc. meters) the classification accuracy for the Essen dataset and the Finish Folk dataset become 89.2% and 93.8% respectively. Table 3 presents comparative results to other existing methods for the 2-class and 9-class classification problem. The proposed method achieves comparative results to state-of-the-art methods.

5.4.2. Dance Style Classification

Table 4 presents the confusion matrix of the classification results on the *Ballroom* dataset. The most correctly classified genre is the Quickstep (QS), with a classification accuracy of over 97%, while Tango (TA), ChaCha (CH) and Waltz (W) were correctly recognized with around 90% accuracy. On the other hand, many instances of Jive (JI) and Viennese Waltz (VW) were confused with Waltz. Table 5 presents the overall classification results of the proposed method (81.95%), compared to other methods. It is noteworthy that as in the case of meter estimation, without any prior knowledge about the task, the features learned from the network achieved a performance close to the state-of-the-art methods. Note that for the reference methods, the values in parenthesis correspond to results that were achieved when the ground-truth tempo was given. Consequently, they are not comparable to the proposed method.

	Predictions							
	CH	JI	QS	RU	SA	TA	VW	W
CH	**88.3**	0	2.7	3.6	0.9	4.5	0	0
JI	3.3	**60**	0	3.3	3.3	1.7	5	23.3
QS	0	0	**97.6**	2.4	0	0	0	0
RU	1	4.1	1	**77.6**	0	4.1	5.1	7.1
SA	1.2	1.2	4.7	8.1	**79.1**	1.2	1.2	3.5
TA	5.8	0	0	0	0	**89.5**	0	4.7
VW	0	1.5	0	1.5	0	0	**58.5**	38.5
W	0	2.7	0	1.8	0	0.9	4.5	**90**

Table 4. *Confusion matrix for the dance style classification task on the Ballroom dataset. Values are percentages (%). Rows correspond to ground truth and columns to predictions.*

Proposed	Gouyon[29]	Peeters[30]	Dixon [31]
81.95	79.6 (90.1)	81 (90.4)	84 (96)

Table 5. *Classification accuracies (%) on the Ballroom dataset of the proposed method along with three reference methods.*

5.5. Sampling from Rhythm Classes

To understand the features learned by the RBM network as well as the reconstruction capabilities of the proposed method, we provide audio examples *sampled* from the overall architecture. To do so, we followed Hinton's approach in [32] to sample digit images from the ten digit classes of MNIST, as shown in Figure 5. A binary class vector **c** consisting of softmax units was concatenated to the network output vector $\mathbf{h}^2$ to form the visible vector of an additional RBM with *N*=3000 hidden binary units in order to learn a joint distribution of **c** and $\mathbf{h}^2$. The RBM was trained using Persistent Contrastive Divergence [33]. After training, samples for a given class where drawn by clamping **c** to a constant value corresponding to this class and running a Gibbs chain with $\mathbf{h}^2$ being randomly initialized. During Gibbs sampling, audio examples were reconstructed from $\mathbf{h}^2$.

We applied this method to draw samples from the eight *Ballroom* classes. In order to provide a quantitative measurement of the quality of the samples, we followed the following procedure. For each class, after 2000 initial Gibbs steps, one sample was drawn every 250 Gibbs steps. Afterwards the samples were filtered out, such that the value *c* of the class vector **c** on the negative phase of the Gibbs sampling for the corresponding class was over 0.7. With this procedure, 50 samples were finally drawn from each class. Afterwards, for each sample the 10 most similar training instances were retrieved. Similarity was computed as the cosine similarity measure of the corresponding periodicity functions. The retrieval matrix obtained with the above procedure is presented in Table 6. Rows correspond to samples drawn from the model and columns correspond to training instances retrieved from each class. The last row shows the mean value of *c* for each class computed over all Gibbs steps. Interestingly, for some classes, retrieval rates are very high, as for example for ChaCha and Samba. As expected, samples from Waltz and Viennese Waltz are similar and as a consequence many Waltz samples are closer to Viennese Waltz training instances and vice-versa. Rumba is confused with Samba, and similarly to Table 4, many Jive samples are close to Waltz instances. An interesting effect which should be further investigated is the case of Quickstep and Tango samples, which were closer to the Samba and the ChaCha training instances respectively. However it should be noted that this experiment corresponds only to a random snapshot; several trials indicated a high sensitivity of these results to both training epochs of the RBM and Gibbs sampling steps. Such an effect, should be investigated in the future. Nevertheless, the overall results of Table 6 indicate that in most of the cases the samples drawn from the model are closer to the actual training instances

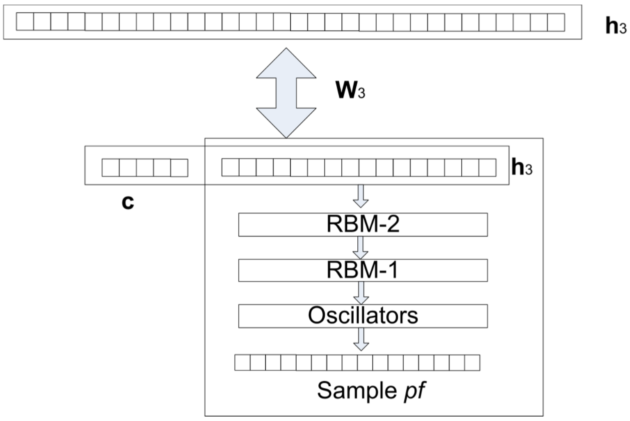

Figure 5. *Class sampling method overview.*

of the corresponding class, than to those of all other classes. Audio examples of the samples can be found in[1]

6. CONCLUSION AND FURTHER WORK

In this paper we presented a method for constructing a rhythm representation that is capable of reconstructing an audio signal that resembles the rhythmic properties of the original. Besides the reconstruction efficiency, the derived features proved to be successful for tackling two important rhythm analysis tasks, namely dance style classification and meter estimation. Without any prior knowledge of the tasks, the performance of the proposed method is comparable to state-of-the-art methods.

The accent feature reconstruction error introduced by the periodicity analysis step, should be investigated further in future work, as for example using other oscillator types. Another possible solution could be to increase the tempo analysis range. However, a balance between reducing this error and keeping the size of the network on a relative small size should be preserved.

The building block of the proposed network, the Restricted Boltzmann Machine, apart from being exploited as a generative model for reconstructing the input and as feature detector to tackle rhythm related tasks, has another important property. It provides a framework for generating samples from classes. Such an extension in the audio domain is a powerful tool since it will provide a deeper insight of what rhythmic features and classes are learnt.

Convolution plays an important role in rhythm processing, since many rhythm analysis methods involve a convolution step with a bank of template filters that correspond to certain frequencies, in order to compute a periodicity function. Instead of jointly learning rhythmic and timing information of the accent features by the periodicity analysis step, a Convolutional Restricted Boltzmann Machine acting directly on the time-domain accent features and representing rhythmic information could be explored as an alternative approach.

7. ACKNOWLEDGEMENTS

This research is funded by the Vienna Science and Technology Fund (WWTF) through project MA14-018, by the Future and Emerging Technologies (FET) Programme Lrn2Cr8 within the Seventh Framework Programme for Research of the European Commission, under FET grant number 610859 and by the LangTERRA project(FP7- REGPOT-2011-1)

	CH	JI	QS	RU	SA	TA	VW	W
CH	**100**	0	0	0	0	0	0	0
JI	0.8	**39.6**	0	3.3	0	0.2	11.8	44.3
QS	2	2	**40.2**	9.4	44.7	0.8	0.2	0.8
RU	0.8	2.9	4.5	**50**	38.4	0	0.6	2.7
SA	0	0	0.8	12.9	**86.3**	0	0	0
TA	33.1	5.7	0.2	1.2	0.2	**55.9**	1.6	2.2
VW	1	6.9	0	3.3	3.5	1	**55.9**	28.4
W	1.4	32.4	0.8	3.3	1.4	2.4	14.3	**44.1**
c	98.3	68.2	77.2	55.9	91.1	78.7	58.4	64.4

Table 6. *Samples to training instances retrieval precision for each class pair. Rows correspond to class samples and columns to class training instances. Values are in percentage.*

8. REFERENCES

[1] Brown, J. C., “Calculation of a constant Q spectral transform,” *The Journal of the Acoustical Society of America*, *89*(1), pp. 425-434, 1991

[2] Schörkhuber, C., and Klapuri, A. ,”Constant-Q transform toolbox for music processing,” in *7th Sound and Music Computing Conference,* Barcelona, Spain ,2010

[3] Velasco, G. A., Holighaus, N., Dörfler, M., & Grill, T. (2011). “Constructing an invertible constant-Q transform with non-stationary Gabor frames,” in *Proceedings of DAFX11, Paris*.

[4] G. Percival and G. Tzanetakis. "Streamlined tempo estimation based on autocorrelation and cross-correlation with with pulses." *IEEE/ACMTrans. Speech Audio Process.* 22(12), pp. 1765-1776, Dec. 2014.

[5] S. Böck and M. Schedl. "Enhanced beat tracking with context-aware neural networks." In *Proc. DAFX,* 2011.

[6] D. Eck, and N. Casagrande, "Finding meter in music using an autocorrelation phase matrix and shannon entropy," in *Proc. ISMIR*, 2005, pp. 504-509.

[7] E. D. Scheirer, “Tempo and beat analysis of acoustic musical signals,” *J. Acoust. Soc. Amer.*, 103(1), pp. 588–601, 1998.

[8] Alonso, M. A., Richard, G., & David, B.”TempoAnd Beat Estimation Of Musical Signals. In *Proc. ISMIR*, 2005

[9] A. Klapuri, A. Eronen, and J. Astola, “Analysis of the meter of acoustic musical signals*,” IEEE Trans. Audio, Speech, Lang. Process*., 14(1), pp. 342–355, Jan. 2006.

[10] A. Gkiokas, V. Katsouros, G. Carayannis & T. Stafylakis, "Music tempo estimation and beat tracking by applying source separation and metrical relations," *in Proc. ICASSP*, pp. 421-424, 2012

[11] S. Dixon, “Automatic extraction of tempo and beat from expressive performances,” *J. New Music Res*., 30(1), pp. 39–58, 2001.

[12] G. Peeters and J. Flocon-Cholet, “Perceptual Tempo Estimation using GMM-Regression,” in *Proc. MIRIUM*, Nara, Japan, 2012

[13] A. Holzapfel, G.A. Velasco, N. Holighaus, M. Dörfler & A. Flexer, “Advantages of nonstationary gabor transforms in beat tacking,” in *Proc. of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pp. 45-50, 2011.

[14] Holighaus N., Dörfler M., Velasco G. A. and Grill T.: “A framework for invertible, real-time constant-Q transforms,” *IEEE Transactions on Audio, Speech and Language Processing*, 21(4), pp. 775-785, 2013.

[15] G. Hinton, "Training products of experts by minimizing contrastive divergence." *Neural computation,* 14(8), pp. 1771-1800, 2002.

[16] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. ICASSP*, 1999, pp. 3089-3092.

[17] L. van Noordenan D. Moelants, "Resonance in the perception of musical pulse," *Journal of New Music Research*, 28(1), 43-66, 1999

[18] Large E. and Kolen J., “Resonance and the Perception of Musical Meter”, *Connection Science 6(1), pp 177-208,* 1994

[19] A. Gkiokas, V. Katsouros and G. Carayannis, “Towards Universal Spectral Rhythm Features: An Application to Dance Style, Meter and Tempo Estimation,” Submitted to *IEEE/ACMTrans. Speech Audio Process.,* Feb. 2015, under review.

[20] C.C. Chang, and C.J. Lin. "LIBSVM: a library for support vector machines." *ACM Trans. on Intel. Systems and Tech.*, 2(3), pp. 27:1-27, 2011.

[21] H. Schaffrath and D. Huron. "The Essen folksong collection in the humdrum kern format," Menlo Park, CA: Center for Computer Assisted Research in the Humanities, 1995.

[22] T. Eerola and P. Toiviainen. "Digital archive of Finnish folk tunes." Computer Database, University of Jyvaskyla, 2004.

[23] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, “An experimental comparison of audio tempo induction algorithms,” *IEEE Trans. Audio, Speech, Lang. Process.*, 14(5), pp. 1832–1844, Sep. 2006

[24] M. Levy, “Improving perceptual tempo estimation with crowd-sourced annotations,” in *Proc. ISMIR*, 2011, pp. 317–322.

[25] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” I*EEE Trans. Speech Audio Process.*, 10(5), pp. 293–302, Jul. 2002.

[26] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman and P. Lamere. "The million song dataset," in *Proc. ISMIR*, 2011, pp. 591-596.

[27] V. Nair and G. Hinton. "Rectified linear units improve restricted boltzmann machines," in *Proc. ICML*, 2010, pp. 807-814.

[28] P. Toiviainen and T. Eerola. "Autocorrelation in meter induction: The role of accent structure," *J. Acoust. Soc. Amer*., 119(2), pp. 1164-1170, 2006.

[29] F. Gouyon, S. Dixon, E. Pampalk and G. Widmer. "Evaluating rhythmic descriptors for musical genre classification," in *Proc. AES* 2004, pp. 196-204.

[30] G. Peeters, “Rhythm Classification Using Spectral Rhythm Patterns,” in *Proc.ISMIR*, 2005 pp. 644-647.

[31] S. Dixon, F. Gouyon and G. Widmer, “Towards Characterisation of Music via Rhythmic Patterns,” in *Proc.ISMIR*, 2004.

[32] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, 18, pp. 1527-1554, 2006.

[33] T. Tieleman, "Training restricted Boltzmann machines using approximations to the likelihood gradient." In *Proceedings of the 25th international conference on Machine learning*, pp. 1064-1071. ACM, 2008.