ABSTRACT

One of the major parameters in music is the overall speed of a musical performance. Speed is often associated with tempo, but other factors such as note density (onsets per second) seem to be important as well. In this study, a computational model of speed in music audio has been developed using a custom set of rhythmic features. The original audio is first separated into a harmonic part and a percussive part and onsets are extracted separately from the different layers. The characteristics of each onset are determined based on frequency content as well as perceptual salience using a clustering approach. Using these separated onsets a set of eight features including a tempo estimation are defined which are specifically designed for modelling perceived speed. In a previous study 20 listeners rated the speed of 100 ringtones consisting mainly of popular songs, which had been converted from MIDI to audio. The ratings were used in linear regression and PLS regression in order to evaluate the validity of the model as well as to find appropriate features. The computed audio features were able to explain about 90% of the variability in listener ratings.

1. INTRODUCTION

This study is focused on one of the major parameters in music, the overall speed of a musical performance. From a music theoretic background we are used to associate speed with the tempo of the music. However, as suggested earlier, the perceived speed is related to the tempo but may also be dependent on other aspects like the note density (number of onsets per second) [1]. An indirect indication of this was provided in [2] where it was found that the note density (and not the tempo) was constant for a certain emotional expression across different music examples. Madison & Paulin [3] asked listeners to rate the speed for 50 music examples spanning a variety of musical styles and rhythms. They found that speed correlated with tempo but also indicated that there must be other aspects involved in the perceptual judgment of speed. In Figure 1, three examples with different tempos and onset densities are shown. As outlined in Table 1, example A has a slow tempo but the hi-hat plays on 16th notes. As a result, the number of onsets coming from percussive instruments is high. Example B has a high number of onsets from harmonic instruments (e.g. vocals, piano, etc.) but a moderate tempo. Finally, example C has the highest tempo but the lowest overall note density. How do these different aspects affect the perceived speed? In this study we will model the perception of speed by extracting specifically developed features (such as tempo and onset densities) from music audio. An important idea is that the model should exploit the characteristics of the onsets to better understand the music.

![Figure 1](https://example.com/figure1.png)

Figure 1. Several factors that can influence the perceived speed of a piece of music. The tempo is one important factor but onset density is relevant as well.

<table>
<thead>
<tr>
<th>Example</th>
<th>Tempo</th>
<th>Drum-Ons</th>
<th>Harm-Ons</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Slow</td>
<td>Fast</td>
<td>Mid</td>
</tr>
<tr>
<td>B</td>
<td>Mid</td>
<td>Mid</td>
<td>Fast</td>
</tr>
<tr>
<td>C</td>
<td>Fast</td>
<td>Slow</td>
<td>Slow</td>
</tr>
</tbody>
</table>

Table 1. Different characteristics of the music which are related to speed. A song may have a slow tempo but many onsets that increases the perceived speed.
The current work is part of an ongoing study about perceptually determined features in music information retrieval. In a previous study it was shown that speed could be modeled by a combination of tempo and different note densities of the instruments using symbolic data [4]. The explained variation was about 90 % using linear regression. This indicates that a similar result could in theory be obtained using audio data provided that the appropriate low-level audio features could be extracted.

A flowchart of the processes used in the model is shown in Figure 2. As a first step, source separation (Section 3) was used to separate harmonic content and percussive content in the audio as well as to cluster onsets into different groups. Features were computed from both the percussive and the harmonic part as well as from the original audio as described in Section 4. To find appropriate features as well to evaluate the validity of the model, regression was used, in which the audio features were mapped against ground truth data consisting of listener ratings of speed. This is described in Section 5.

2. SPEED DATA AND AUDIO EXAMPLES

The speed estimations were perceptually determined in a previous experiment in which 20 listeners rated speed for each music example on a quasi-continuous scale marked slow-fast with the range 1-9. The music examples were a set of 100 ringtones consisting mainly of popular songs, originally in MIDI format and converted to audio [5, 6].

3. SOURCE SEPARATION AND ONSET DETECTION

The intermediate processing steps between audio and feature extraction (green boxes in Figure 2) are described in this section.

3.1 HP-Separation

Source separation was used to separate harmonic and percussive content. Source separation has been used in the past in computational models related to rhythm [7]. The method proposed by FitzGerald [8] was used as the first step of the separation. The basic idea of the method is that percussive sounds are broadband noise signals with short duration and that harmonic sounds are narrow band signals with longer duration. To be able to separate these different sounds, the audio is transformed to the spectral domain by using a short-time Fourier transform (STFT). By applying a median filter across each frame in the frequency direction, harmonic sounds are suppressed. By applying a median filter across each frequency bin in the time direction percussive sounds are suppressed. After median filtering, the signal is transformed back to the time domain again using the inverse STFT.

To further suppress harmonic content in the percussive waveform a second separation stage incorporates a constant-Q transform (CQT) [9]. The CQT can be understood as an STFT with logarithmically spaced frequency bins, accomplished by varying the length of the analysis window. The implication relevant to this study is that a high frequency resolution can be achieved also in the low frequencies, at the expense of a poor time resolution.

The frequency resolution of the CQT was set to 60 bins per octave and each frame was median filtered across the frequency direction with a window size of 40 bins. After filtering, the percussive signal was transformed back to the time domain using an inverse CQT.

By transforming back to the time-domain, the underlying phase information is retained. The phase can be regarded as a mapping that connects a frequency bin to a certain point in time. This is especially useful in the CQT-stage as the filtering can be performed at a low

![Figure 2](image-url)
time-resolution (with window lengths up to a second); but subsequent onset detection algorithms can be computed at a higher time-resolution. The resulting percussive and harmonic waveforms are shown in Figure 3.

Figure 3. The result of the HP-separation. The original waveform is separated into a percussive and a harmonic waveform. The example is a 3-second section of the song Candy Shop, by 50 cent, which will be used to visualize the feature extraction throughout this paper.

3.2 Onset Detection
Audio features were computed from all three waveforms (original, harmonic and percussive) by the scheme shown in Figure 2. The first step, independent of feature and waveform, was to compute a spectral flux (SF) [10], where spectral fluctuations along the time-domain are detected. The SF was computed several times in different ways. Some shared steps will be described here, with unique steps described in Sections 4.1-4.8. The power spectrum was computed with a CQT or an STFT and converted to sound level. A range of 30 dB was used. Thus, the maximum sound level of each band is set to 0 dB and sound levels below -30 dB are set to -30 dB. Let \( L(n, i) \) represent the sound level at the \( i \)th frequency bin/band of the \( n \)th frame. The SF is given by

\[
SF(n) = \sum_{i=1}^{b} H \left( L(n, i) - (L(n-s, i) \right)
\]

where \( b \) is the number of bins/bands. The variable \( s \) is the step size and \( H \) is a half-wave rectifier function, or for the percussive SF:

\[
H(x) = \begin{cases} 
  x & \text{if } x > 0 \\
  0.2x & \text{if } x \leq 0 
\end{cases}
\]

The implication of Eq. 2 is that negative spectral fluctuations have a slight influence on the onset detection function. Onsets were detected by peak picking on a low-pass filtered curve of the spectral flux (see Figure 4).

3.3 Clustering
To better exploit the characteristics of the percussive onsets they were clustered into groups. The clustering was based on sound level in 8 frequency bands, spaced approximately an octave apart, as well as the RMS sound level. As the appropriate number of clusters is unknown beforehand, three k-means clusterings [11, 12] were carried out, with the number of clusters \( k \), set to 2, 3 and 4. The fit of each clustering attempt was defined by the smallest Euclidian distance between any two clusters, where a large smallest distance gave a higher fit. When choosing which clustering attempt to use, a higher number of groups \( k \) were premiered over a lower if their fit was similar. The result is shown in Figure 5.

Figure 4. The onset detection functions that discover onsets by finding peaks in the SF. In this example harmonic onsets are tracked.

Figure 5. The clustering of percussive onsets. In example A the drums are clustered into three different clusters. In example B three clusters are initially discovered, but the onsets in Cluster 1 are assigned to Cluster 2 & 3.
When the clustering is completed the onsets have been divided into 2, 3 or 4 clusters. At this point the clusters are further analyzed to find out if the sound of the onsets in two of the clusters can be combined to form the sound of the onsets in a third cluster. This happens in example B of Figure 5. The k-means clustering has divided the onsets into three different clusters, corresponding to the sound of the kick and the hihat combined, as well as both played separately. The algorithm then compares the different clusters and discovers that Cluster 2 (the kick) and Cluster 3 (the hihat) can be combined to form the sound of Cluster 1 (the kick and the hihat). To account for this, each onsets belonging to Cluster 1 will instead be set as belonging to both Cluster 2 and Cluster 3, and Cluster 1 will cease to exist. This does not happen in example A of Figure 5 where 3 unique clusters have been identified.

4. FEATURE EXTRACTION

A total of 8 audio features were computed, 2 from the original waveform, 5 from the percussive waveform and 1 from the harmonic waveform. The audio features are shown as the end result in the flowchart in Figure 2. The 8 features are explained in Sections 4.1-4.8, with one subsection for each feature.

4.1 Onset Density – Harmonic

Onsets in harmonic instruments were tracked from the original waveform, with the SF of a CQT. The bins of the CQT were not combined into broader bands before the SF. This facilitates the detection of harmonic onsets, as a pitch shift of a semitone in an instrument will result in an increase in energy in the half wave rectified SF.

To avoid false onset detections at pitch glides from vibratos, shifts of a peak by 20 cents (one bin), without an increase in sound level, were restricted from affecting the SF. This was accomplished by subtracting the sound level of each bin of the new frame, by the maximum sound level of the adjacent bins in the old frame. The onset detection function for harmonic onsets is shown in Figure 4.

4.2 Onset Density – Bass

Onsets in the low register (frequencies between 40 Hz and 210 Hz) were tracked with an SF of the lower bins of an STFT. The frequency bins were summed to a single band before the SF.

4.3 Onset Density – Perceptual weighting

Percussive onsets were tracked with an SF of an STFT on the percussive waveform. The bins of the frequency domain representation were divided into 13 non-overlapping frequency bands (half-octave spacing). Sub-band processing for onset detection has been described in [13], and can be motivated by its similarity to human hearing [14]. The strength of each detected onset was calculated based on the average sound level of the first 50 ms from the onset position, where lower frequencies were given a higher relevance.

To further determine the perceived strength of the onsets, each onset was compared to the surrounding onsets within 1.5 seconds. This time span (3 seconds in total) was defined as the perceptual present of the particular onset. By comparing it with the strongest onset within the perceptual present its strength could be altered to represent its perceptual impact. The onset was given a higher strength if there were no significantly stronger onsets within the perceptual present. If there were onsets that were significantly stronger, its strength was lowered. The height of the cluster-bars in Figure 6 represents the perceptual strength of each onset. To derive at a measure of onsets density, the sum of the perceptual strength of the onsets was used.

4.4 Onsets Density – Strong

The strongest clusters of the clustering process were used to compute two features. The first feature was simply the number of onsets, belonging to a strong cluster, per second. This feature was only computed for periods of strong onsets within 1.5 seconds of each other.

4.5 Strong Cluster IOI

The second feature derived from the strong clusters was developed to catch the assumed perception of a slow speed, when the interonset intervals (IOIs) of onsets belonging to the same strong cluster are long. As an example, a song with equally spaced drum onsets consisting of “Kick, Snare, Kick, Snare,..” was assumed to

![Figure 6. An overview of the processes involved in extracting 5 features (described in Section 4.3-4.7) from the percussive waveform. Onsets are detected and clustered into different components to gain an understanding of how the music will be perceived. The perceptual weighting of the onsets is represented by the height of the bars. In this particular song, Tempo is derived from the IOI between kick and handclaps and only the kick belonged to a strong cluster. The percussiveness feature is related to the height of the peaks in the onset detection function, as visualized by the dotted line.](image-url)
have a higher perceived speed than a song where the drums instead plays “Kick, Kick, Snare, Kick, Kick, Snare, Kick, etc..”. This is accounted for in the Tempo feature as well, because the tempo in the second example would be half the tempo of the first example.

In Figure 6, this feature is derived from the IOI between onsets belonging to Cluster 1. Common IOIs are detected by peak picking in a low pass filtered histogram of cluster IOIs. Each found peak contributes to the feature based on its relative height as well as the cluster strength.

4.6 Tempo S-Curve

The tempo detection algorithm is part of an ongoing project, and a detailed description is in preparation. All distances between onsets within 5 seconds from each other are used to detect the tempo.

4.6.1 Period Length

First, the period length of the percussive waveform is detected. The period length corresponds to the length of the most prominent pattern of repeated rhythmic sounds in the music. A histogram over onset distances is generated, where the contribution of each onset-pair increases with increasing similarity in spectrum as well as increasing onset strength. The leftmost peak in the low pass filtered histogram, within 92 % of the highest peak, is chosen as the period length.

4.6.2 Tempo

Secondly, the tempo (beat length) is detected. A histogram over onset distances is once again generated, where the contribution of each onset-pair increases with increasing dissimilarity in spectrum as well as increasing onset strength. The final probability distribution for tempo is the Hadamard product of the histogram and several filters. One filter is based on the determined period length. The idea is that the beat will be a simple ratio of the period length, so Hanning windows are produced at positions

$$P_{\text{low}} \times \left( \frac{1}{2} \right)^n, \quad P_{\text{low}} \times \left( \frac{1}{2} \right)^{n+1} \times \left( \frac{1}{3} \right) \quad n = 0,1,2,... \quad (3)$$

Another filter is based on IOIs within strong clusters as described in Section 4.5. The general distribution of tempos in popular music is taken into account in one filter and several filters are connected to the onset density of the particular song. The highest peak in the final probability distribution was chosen as the tempo.

4.6.3 S-Curve

In compliance with the findings in [3], an S-Curve (Figure 7) was applied to the tempo value, giving differences in tempo a higher impact between 60 and 160 BPM.

4.7 Percussiveness

One feature was based on the percussiveness of the onsets. This estimate is derived from the height $h$ of the peaks in the SF of the percussive waveform, as shown in Figure 6.

$$\text{Percussiveness} = \frac{\sum_{i=1}^{n} h(i)^{1+p}}{\sum_{i=1}^{n} h(i)^p} \quad (4)$$

Equation 4 gives the mean peak height when $p$ is 0, an estimate closer to the lowest peaks when $p$ is negative, and an estimate closer to the highest peaks when $p$ is positive. In this study $p$ was set to 0.4.

4.8 SF CQT

When extracting information from the harmonic waveform the integral of the SF was used; indicated as the colored area in Figure 8. The use of an onset detection function was avoided as the HP-separation had removed all transients from the harmonic waveform. The use of a CQT was motivated by the harmonic nature of the processed audio. Spectral changes in high frequencies were used for this feature.

5. PREDICTING SPEED FROM THE FEATURES

Two regression techniques were used to analyze the mapping between the computed audio features and the listener ratings of speed. First, a multiple linear regression was used, justified by a predictor-to-case ratio higher
than 1:10. Secondly, PLS regression was used [15]. PLS regression carries out data reduction, whilst maximizing covariance between features and predicted data [16].

The multiple linear regression between listener ratings and computed audio features is presented in Table 2. As shown, a linear combination of the computed audio features was able to explain about 90 % of the variability. In comparison, the agreement among the listeners estimated by the mean intersubject correlation was 0.71 and Cronbach’s alpha 0.98 [4].

Table 2. The prediction of the perceptual feature speed from computed audio features. The variable $sr^2$ is the squared semi-partial correlation coefficient.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Beta</th>
<th>$sr^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>On Dens. - Harmonic</td>
<td>0.205</td>
<td>0.033</td>
<td>0.000***</td>
</tr>
<tr>
<td>On Dens. - Bass</td>
<td>0.130</td>
<td>0.007</td>
<td>0.016*</td>
</tr>
<tr>
<td>On Dens. - Perceptual</td>
<td>0.302</td>
<td>0.018</td>
<td>0.000***</td>
</tr>
<tr>
<td>On Dens. - Strong</td>
<td>-0.155</td>
<td>0.010</td>
<td>0.004**</td>
</tr>
<tr>
<td>Strong Cluster IOI</td>
<td>0.127</td>
<td>0.006</td>
<td>0.021*</td>
</tr>
<tr>
<td>Tempo S-Curve</td>
<td>0.430</td>
<td>0.056</td>
<td>0.000***</td>
</tr>
<tr>
<td>Percussiveness</td>
<td>-0.095</td>
<td>0.005</td>
<td>0.041*</td>
</tr>
<tr>
<td>SF CQT</td>
<td>0.107</td>
<td>0.004</td>
<td>0.053</td>
</tr>
</tbody>
</table>

The most important feature was Tempo S-Curve, followed by Onset Density - Harmonic, Onset Density - Perceptual and Onset Density - Strong (negative contribution). The independent contribution in terms of the squared semi-partial correlation coefficient $sr^2$ indicates that Onset Density - Bass, Strong Cluster IOI, Percussiveness and SF CQT each increased the explained variance with less than 1 %. The negative contribution of Percussiveness could be explained as a higher perceived speed when the percussive onsets are less clear.

A partial least square regression (PLS) of the same features is shown in Table 3. With 3 components, the cross-validated adjusted $R^2$ indicates that just below 90 % of the variability could be explained. Note also that the cross-validation procedure only lowers the result marginally, supporting the validity of the features.

Table 3. The prediction of the perceptual feature speed from computed audio features. The squared correlation coefficient $R^2$ was derived using Partial Least-square Regression (PLS), with 10-fold cross validation. In the lower part, $R^2$ as a function of the number of components is shown. Components 4-8 did not contribute and are not shown.

<table>
<thead>
<tr>
<th>PLS Regression - Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Components Used = 3</td>
</tr>
<tr>
<td>$R^2 = 0.907$</td>
</tr>
<tr>
<td>$R^2$ cv = 0.883</td>
</tr>
<tr>
<td>Component</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

The fitted values of the linear regression from Table 2 are shown in Figure 9 below. As seen in the figure, the deviations from the target are rather evenly distributed across the range and with a maximal deviation of about one unit.

6. CONCLUSIONS AND DISCUSSION

The computed audio features were able to explain about 90 % of the variability in listener ratings. The most important features were tempo together with onset densities for different layers of the music. The validity of the features was supported by a cross-validation, and fitted values were relatively close to target values.

The results show that it was possible to reach the same high explained variance on audio data as on MIDI data using similar features [4]. This indicates that the appropriate low-level audio features have been extracted, which is reassuring for the ongoing study.

Since good results were achieved only after we applied source separation, both in terms of clustering and HP-separation, the segmentation of data seems to be a promising path forward. From an ecological point of view it seems reasonable to assume that the interaction between onsets of the same source is relevant; especially if the sound of this source is one of the most prominent ones. By clustering onsets we can detect onsets belonging to the same source and thus use the rhythmic pattern of this source in the model. By using several onset detection functions on separate parts of the audio, different aspects of the music can be captured. The CQT seems to be suitable for detecting onsets in harmonic instruments, while the better time-resolution of the STFT in lower frequencies facilitates the detection of percussive instruments. A drawback with the proposed system is that the computation of several STFTs and CQTS is relatively time-consuming.
7. ACKNOWLEDGEMENT
This work was supported by the Swedish Research Council, Grant Nr. 2009-4285 and 2012-4685.

8. REFERENCES