Template-Based Vibrato Analysis in Music Signals

Driedger, J; Balke, S; Ewert, S; Müller, M

TEMPLATE-BASED VIBRATO ANALYSIS OF MUSIC SIGNALS

Jonathan Driedger[1], **Stefan Balke**[1], **Sebastian Ewert**[2], **Meinard Müller**[1]
[1]International Audio Laboratories Erlangen, Germany
[2]Queen Mary University of London
{jonathan.driedger,stefan.balke,meinard.mueller}@audiolabs-erlangen.de

ABSTRACT

The automated analysis of vibrato in complex music signals is a highly challenging task. A common strategy is to proceed in a two-step fashion. First, a fundamental frequency (F0) trajectory for the musical voice that is likely to exhibit vibrato is estimated. In a second step, the trajectory is then analyzed with respect to periodic frequency modulations. As a major drawback, however, such a method cannot recover from errors made in the inherently difficult first step, which severely limits the performance during the second step. In this work, we present a novel vibrato analysis approach that avoids the first error-prone F0-estimation step. Our core idea is to perform the analysis directly on a signal's spectrogram representation where vibrato is evident in the form of characteristic spectro-temporal patterns. We detect and parameterize these patterns by locally comparing the spectrogram with a predefined set of vibrato templates. Our systematic experiments indicate that this approach is more robust than F0-based strategies.

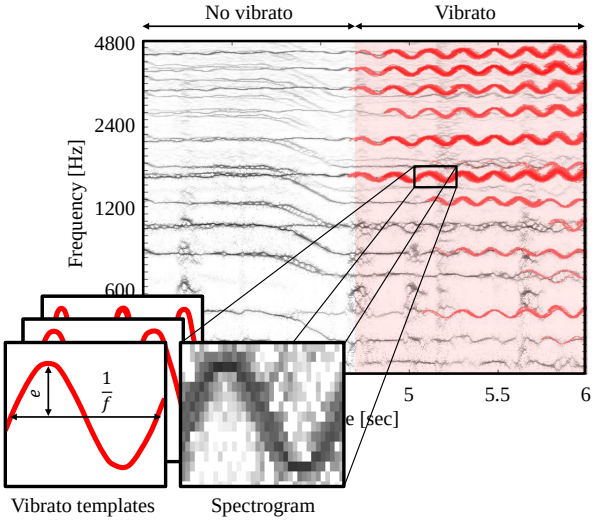

Figure 1. Template-based vibrato analysis. A matching vibrato template lets us infer the rate f and extent e of vibrato present in the music signal.

1. INTRODUCTION

The human voice and other instruments often reveal characteristic spectro-temporal patterns that are the result of specific articulation techniques. For example, vibrato is a musical effect that is frequently used by musicians to make their performance more expressive. Although a clear definition of vibrato does not exist [20], it can broadly be described as a musical voice's "periodic oscillation in pitch" [16]. It is commonly parameterized by its *rate* (the modulation frequency given in Hertz) and its *extent* (the modulation's amplitude given in cents [1]). These parameters have been studied extensively from musicological and psychological perspectives, often in a cumbersome process of manually annotating spectral representations of monophonic music signals, see for example [5, 10, 18, 20, 22].

To approach the topic from a computational perspective, the signal processing community has put considerable research efforts into developing automated vibrato analysis methods for monophonic, as well as for more complex music signals with multiple sound sources. While some applications implicitly exploit spectro-temporal characteristics of vibrato to approach higher-level tasks such as harmonic-percussive decomposition [9], singing voice detection [6], or singing voice separation [21], there also exist methods for explicitly detecting and parameterizing vibrato components in a given music signal. A common approach is to perform the vibrato analysis in two consecutive steps. In the first step, a fundamental frequency trajectory (F0-trajectory) is estimated for the musical voice that is most likely to exhibit vibrato. This trajectory is then analyzed in the second step to detect and parameterize periodic modulation patterns, see for example [4, 8, 12–14, 23]. However, computing F0-trajectories for complex signals with multiple instruments is a highly non-trivial and error-prone task by itself [15]. Therefore, a trajectory estimated in the first step may not appropriately reflect the relevant modulation patterns. This in turn renders the vibrato detection and parametrization in the second step problematic, if not impossible.

To avoid the error-prone F0-estimation step, in this work we propose a novel approach for automatically analyzing vibrato components in complex music signals. Our core idea is to detect spectro-temporal vibrato patterns di-

[1] A *cent* is a logarithmic frequency unit. A musical semitone is subdivided into 100 cents.

rectly in a music signal's spectrogram by locally comparing this representation with a set of predefined vibrato templates [2] that reflect different vibrato rates and extents. The measured similarity yields a novel mid-level feature representation—a *vibrato salience spectrogram*—in which spectro-temporal vibrato patterns are enhanced while other structures are suppressed. Figure 1 illustrates this idea, showing three different vibrato templates as well as a spectrogram representation of a choir with a lead singer who starts to sing with strong vibrato in the excerpt's second half. Time-frequency bins where one of the templates is locally similar to the spectrogram, thus yielding a high vibrato salience, are indicated in red. As we can see, these time-frequency bins temporally coincide with the annotated vibrato passage at the top of Figure 1. Additionally, a high vibrato salience does not only indicate the presence of vibrato in the music signal, but also reveals the vibrato's rate and extent encoded in the similarity maximizing template.

The remainder of this paper is structured as follows. In Section 2 we describe our template-based vibrato analysis approach in detail. In Section 3, we evaluate the performance of our proposed method, both by means of a quantitative evaluation on a novel dataset as well as by discussing illustrative examples. Finally, in Section 4, we conclude with an indication of possible future research directions. Note that this paper has an accompanying website at [2] where one can find all audio examples and annotations used in this paper.

2. TEMPLATE-BASED VIBRATO ANALYSIS

In this section, we describe our proposed template-based vibrato analysis approach. We discuss relevant spectrogram representations (Section 2.1) and describe how the vibrato templates are modeled (Section 2.2). Both our choice of spectrogram representation and the vibrato template's design are motivated by the correlation-like similarity measure that we use to locally compare the templates with the spectrogram. We then introduce the derivation of the vibrato salience spectrogram (Section 2.3) and comment on our approach's computational complexity (Section 2.4). As a running example, we use the choir signal from Figure 1.

2.1 Spectral Representation

Given a discrete music signal $x : \mathbb{Z} \to \mathbb{R}$, we first compute the *short-time Fourier transform* (STFT) $X : \mathbb{Z} \times \mathbb{Z} \to \mathbb{C}$ of x by

$$X(m,k) = \sum_{r \in \mathbb{Z}} w(r) \cdot x(r+mH) \cdot \exp(-2\pi i k r/N) \ , \quad (1)$$

where m is the frame index, k is the frequency index, N is the frame length, w is a window function, and H is the hopsize (w.l.o.g. we assume $m, k \in \mathbb{Z}$). Figure 2a shows an excerpt of our example signal's *magnitude spectrogram* $|X|$ where one can clearly see wave-like vibrato patterns in the lead singer's fundamental frequency and its overtones. However, due to the STFT's linear frequency sampling, the vibrato patterns' amplitudes increase with higher overtones.

[2] Note that this approach is conceptually similar to the Hough transform [3], a mathematical tool known from image processing for the detection of parameterized shapes in binary images. However, the Hough transform is known to be very sensitive to noise and therefore not suitable for detecting vibrato patterns in spectrograms that are commonly rather noisy.

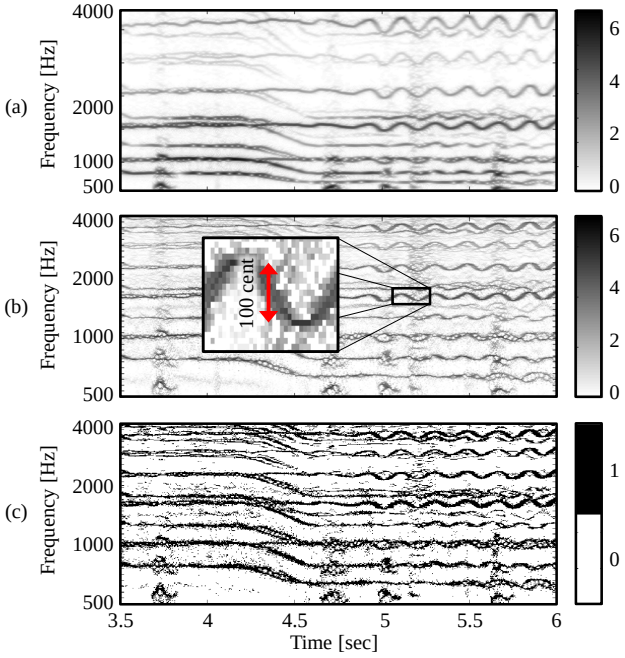

Figure 2. Spectrogram representations of the input signal x. **(a):** Magnitude spectrogram. **(b):** Log-frequency spectrogram. **(c):** Binarized log-frequency spectrogram Y.

In the context of our template-based analysis it is desirable that vibrato patterns stemming from the same frequency modulated tone have the same amplitude that directly reflects the vibrato's extent. We therefore compute a *log-frequency spectrogram* from the STFT X, using a phase vocoder-based reassignment approach as discussed in [7, Chapter 8] or [14]. In this representation, which can be seen in Figure 2b, frequency bands are spaced logarithmically and have a constant logarithmic bandwidth specified in cents. This ensures the desired property in this spectrogram representation.

In a last step, we normalize the spectrogram in order to achieve two goals. First, we aim to make the representation independent of the signal's volume such that we can also detect vibrato in quiet signal passages. Second, when locally comparing our vibrato templates with the representation, the resulting similarity measure should yield values in a fixed range. A method that showed to be simple and effective to achieve both goals is spectrogram binarization, where we set the ten percent highest values of each frame in the log-frequency spectrogram to one and all remaining values to zero. This yields a *binarized log-frequency spectrogram* $Y : \mathbb{Z} \times \mathbb{Z} \to \{0, 1\}$, see Figure 2c. In our experiments, we choose parameters such that Y has a time res-

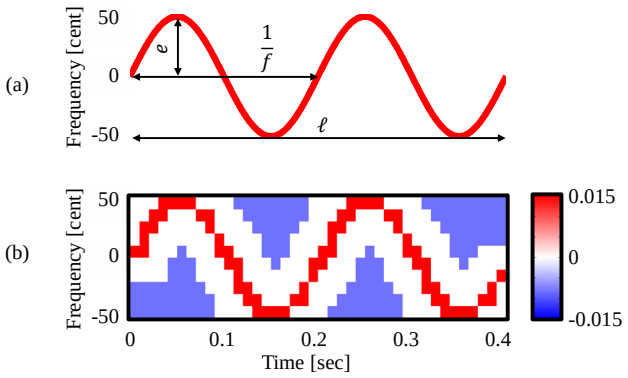

Figure 3. Generation of a vibrato template T with a vibrato rate $f = 5$ Hertz, extent $e = 50$ cent, and a duration of $\ell = 0.4$ seconds. **(a):** Sinusoidal vibrato trajectory s. **(b):** Vibrato template T.

olution of roughly 150 frames per second and a frequency resolution of ten bands per semitone.

2.2 Vibrato Templates

Next, we introduce a set $\mathcal{T}$ of templates that reflect spectro-temporal vibrato patterns as expected in Y. Let us model such a template $T \in \mathcal{T}$ for vibrato having a rate of f Hertz, an extent of e cents, and a duration of at least ℓ seconds. When locally comparing the template T with Y, one should obtain high similarity values when T is aligned with a matching spectro-temporal vibrato pattern in Y and low values otherwise. The idea is therefore to have a positive portion in T that reflects the spectro-temporal vibrato pattern as well as a negative portion that prevents the template from correlating well with regions in Y that are homogeneously equal to one.

Assuming a sinusoidal vibrato, we can describe the vibrato's trajectory (up to phase) by

$$s(t) = e \, \sin(2\pi f t) \, , \tag{2}$$

$t \in [0, \ell]$. Figure 3a shows such a trajectory for $f = 5$ Hertz, $e = 50$ cent, and $\ell = 0.4$ seconds. The trajectory is then discretized such that its time- and frequency resolution matches the binarized log-frequency spectrogram. Time-frequency bins that are close to s are assigned with positive values, while bins having a certain distance from s get negative values. To allow for some tolerance of the width of vibrato patterns in Y, the remaining time-frequency bins are defined to be zero. Finally, positive and negative entries in T are normalized to sum up to one and minus one, respectively, see Figure 3b.

2.3 Vibrato Salience

In order to locate and parameterize vibrato structures in the binarized log-frequency spectrogram Y, we aim to compute a *vibrato salience spectrogram* S—a kind of mid-level feature representation—in which vibrato structures are enhanced while other kinds of structures are suppressed. To this end, we define the vibrato salience spectrogram S_T for a single vibrato template $T : [0 : A-1] \times [0 : B-1] \to \mathbb{R}$, $A, B \in \mathbb{N}$. The computation process is illustrated in Figure 4a. Let $\mathcal{I}$ be the set of all index pairs $(a, b) \in [0 : A-1] \times [0 : B-1]$ such that $T(a, b)$ is positive (the indices of all red entries in Figure 3b). Furthermore, let

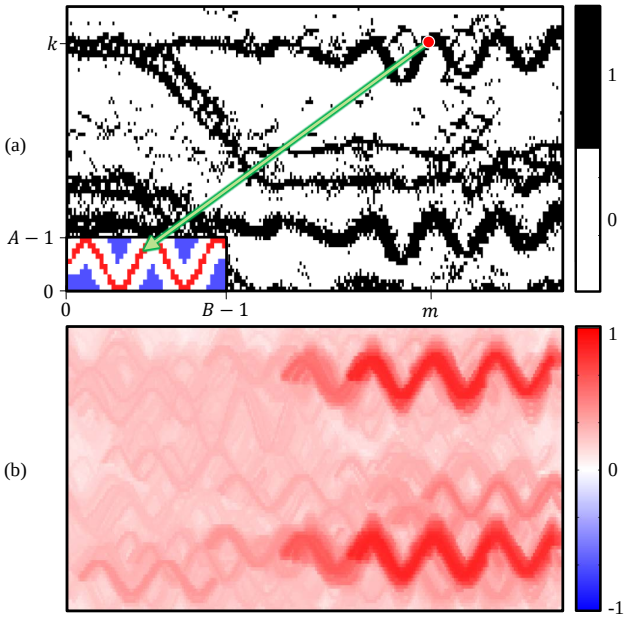

Figure 4. Vibrato salience spectrogram computation. **(a):** Process to compute S_T. The similarity-maximizing shift (μ, κ) that maps (m, k) onto an index pair in $\mathcal{I}$ is indicated by a green arrow. **(b):** Vibrato salience spectrogram S.

$$Y^{(\mu,\kappa)}(m, k) = Y(m - \mu, k - \kappa) \, , \tag{3}$$

$\mu, \kappa \in \mathbb{Z}$, denote a version of Y that is shifted by μ and κ indices in time- and frequency direction, respectively. Intuitively, the vibrato salience $S_T(m, k)$ should be high if $Y(m, k)$ is part of a spectro-temporal vibrato pattern as reflected by T. To this end, we verify if there is a shift (μ, κ) that aligns $Y(m, k)$ (red dot in Figure 4a) with one of the positive entries in the vibrato template T such that T and $Y^{(\mu,\kappa)}$ are similar (the optimal shift for our example in Figure 4a is indicated by a green arrow). To compute $S_T(m, k)$, we therefore maximize the correlation-like similarity measure

$$c(T, Y) = \sum_{a=0}^{A-1} \sum_{b=0}^{B-1} T(a, b) Y(a, b) \tag{4}$$

over all shifts (μ, κ) that map (m, k) onto one of the index pairs in $\mathcal{I}$:

$$S_T(m, k) = \max_{\{(\mu,\kappa):(m,k)-(\mu,\kappa)\in\mathcal{I}\}} c(T, Y^{(\mu,\kappa)}) \, . \tag{5}$$

The full vibrato salience spectrogram can then be computed by maximizing over all vibrato templates $T \in \mathcal{T}$:

$$S(m, k) = \max_{T\in\mathcal{T}} S_T(m, k) \, . \tag{6}$$

Item name	L_{x}	L_{vib}	-0 dB TB-A	-0 dB F0-M	-5 dB TB-A	-5 dB F0-M	-10 dB TB-A	-10 dB F0-M	BL
Sound On Sound Demo—Mystery	9.79	1.78	0.83	0.93	0.84	0.86	0.73	0.30	0.31
Giselle—You	5.12	2.99	0.91	0.94	0.91	0.88	0.86	0.53	0.73
Leaf—Full	5.36	1.64	0.84	0.86	0.74	0.29	0.82	0.00	0.46
Phre The Eon—Everybody is Falling Apart	2.47	0.47	0.98	0.97	0.96	0.97	0.95	0.00	0.32
Secretariat—Borderline	7.69	1.98	0.79	0.69	0.73	0.76	0.79	0.00	0.41
Sunshine Garcia Band—For I Am The Moon	12.54	3.36	0.63	0.73	0.67	0.62	0.74	0.44	0.42
Angela Thomas Wade—Milk Cow Blues	4.50	2.10	0.44	0.82	0.32	0.63	0.32	0.00	0.63
Triviul—Dorothy	5.22	0.85	0.77	0.88	0.73	0.85	0.65	0.00	0.28
Funny Valentines—Sleigh Ride	7.18	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00
∅	6.65	1.69	0.80	0.87	0.77	0.76	0.76	0.25	0.39

Table 1. Quantitative evaluation (F-measure), comparing our proposed template-based detection approach TB-A, F0-based vibrato detection F0-M (manual vibrato selection in F0-trajectories), and a baseline BL. Lengths of the signals (L_x) and accumulated lengths of ground truth vibrato passages (L_{vib}) are given in seconds.

Figure 4b shows the vibrato salience spectrogram S resulting from the binarized log-frequency spectrogram Y shown in Figure 4a. Note that by the vibrato template's design and $Y(m,k) \in \{0,1\}$, one obtains $S(m,k) \in [-1,1]$ for all $m,k \in \mathbb{Z}$. While the vibrato structures present in Y are also clearly visible in S, the horizontal structures as well as the glissando at the excerpt's beginning do not correlate well with the vibrato templates. They are therefore, as intended, suppressed in S.

2.4 Computational Complexity

The vibrato salience spectrogram's derivation as defined in the previous section is a computationally expensive process. When implemented naively, it is necessary to use a quadruply nested loop to iterate over all combinations of time-frequency bins (m,k) in Y, vibrato templates $T \in \mathcal{T}$, index shifts $\{(\mu,\kappa) : (m,k) - (\mu,\kappa) \in \mathcal{I}\}$, and index pairs (a,b) in T. However, note that many computations are redundant and that it is therefore possible to optimize the calculation process, for example by exploiting two-dimensional convolutions. Furthermore, one can speed up the derivation by considering only a limited frequency range in Y as well as by applying further heuristics such as only taking into account vibrato salience values above a threshold $\tau \in [-1,1]$. Although still being computationally demanding, the derivation therefore becomes feasible enough to be used in practice. For example, deriving S for a music signal with a duration of 60 seconds takes our MATLAB implementation roughly 40 seconds on a standard computer.

3. EXPERIMENTS

In this section, we present our experimental results. In Section 3.1, we quantitatively evaluate our proposed approach in the context of a vibrato detection task. Then, in Section 3.2 we demonstrate the method's potential for automatically analyzing vibrato rate and extent. Finally, in Section 3.3, we indicate open challenges and potential solutions.

3.1 Evaluation: Vibrato Detection

In a first experiment, we considered the task of temporally identifying vibrato passages in a music signal. We therefore compiled a dataset of nine items (see Table 1), which are excerpts of music signals from the "Mixing Secrets" multitrack dataset [17]. Each item consists of a monophonic vocal signal x_{voc} and a polyphonic accompaniment signal x_{acc}. Annotations of vibrato passages in the vocal signals were created manually to serve as ground truth for the subsequent evaluation (none of the accompaniment signals x_{acc} has vibrato). To vary the difficulty of the vibrato detection task, we created three different mixes for each of the items—one were x_{voc} and x_{acc} were mixed without modification (-0 dB), one were x_{voc} was attenuated by -5 dB prior to mixing the signals, and a third mix with x_{voc} being attenuated by -10 dB.

To construct an automated vibrato detection procedure based on our proposed template-based analysis approach, we first computed vibrato salience spectrograms S for all of the resulting 27 mix signals. Since only high vibrato salience values in S are likely to indicate the presence of spectro-temporal vibrato patterns, we then chose a threshold $\tau \in [-1,1]$. Time instances where the maximal vibrato salience in a frame exceeded τ were then labeled as having vibrato while all other time instances were labeled as having no vibrato. For this experiment we used a set $\mathcal{T}$ of 30 templates, reflecting vibrato rates from five to seven Hertz in steps of 0.5 Hertz, as well as extents from 50 to 100 cents in steps of 10 cents. These parameters were chosen particularly to detect the vibrato in singing voice as these are typical vibrato rates and extents for human singing, see [10, 11]. All templates had a length corresponding to $\ell = 0.4$ seconds. The threshold τ was experimentally set to $\tau = 0.55$, yielding good vibrato detection results for all items in the dataset.

One of this experiment's main objectives was to compare our template-based method's performance with F0-based strategies as discussed in Section 1. To emulate such an approach, we used MELODIA [14]—a state-of-the-art algorithm for estimating F0-trajectories of predominant musical voices in complex music signals—to esti-

mate trajectories for all mix signals. Instead of automatically analyzing the extracted trajectories in a second step, we then manually inspected them for passages that reflect vibrato. This was done to obtain an upper bound on the performance an automated procedure could achieve in this second step when detecting vibrato solely based on the estimated F0-trajectory.

We then computed precision (P), recall (R), and F-measure (F) for the detection results of our automated template-based procedure (`TB-A`), for the procedure based on the manually inspected F0-trajectory (`F0-M`), as well as for a baseline approach that simply labels every time instance as having vibrato (`BL`):

$$\mathrm{P} = \frac{\mathrm{TP}+\epsilon}{\mathrm{TP}+\mathrm{FP}+\epsilon}, \mathrm{R} = \frac{\mathrm{TP}+\epsilon}{\mathrm{TP}+\mathrm{FN}+\epsilon}, \mathrm{F} = \frac{2\mathrm{PR}}{\mathrm{P}+\mathrm{R}} \ . \quad (7)$$

Here, TP is the number of true positives, FP the number of false positives, FN the number of false negatives, and $\epsilon > 0 \in \mathbb{R}$ is some small number to prevent division by zero. Note that all music signals and annotations used in the experiment can be found at this paper's accompanying website [2].

The evaluation's results are summarized in Table 1 which shows for each item its name, the music signal's length, the accumulated duration of vibrato in this signal, as well as the F-measures of `TB-A` and `F0-M` for the three different mixes (-0 dB, -5 dB, and -10 dB). The F-measure for the baseline `BL` is indicated in the last column and the table's last row indicates mean values. Here we can observe a clear trend. For mixes where x_{voc} was not attenuated (-0 dB), both `TB-A` and `F0-M` yield average F-measures ($\mathrm{F} = 0.80$ and $\mathrm{F} = 0.87$) clearly above the baseline `BL` ($\mathrm{F} = 0.39$). For this mixing condition, `F0-M` outperforms our template-based approach. However, recall that `F0-M` constitutes an upper bound on the performance of F0-based vibrato detection approaches. Automating the vibrato detection step may therefore result in lower scores.

For mixes where x_{voc} was attenuated by -5 dB, the average F-measure of `TB-A` only slightly decreases to $\mathrm{F} = 0.77$, while the performance of `F0-M` drops to $\mathrm{F} = 0.76$. This tendency becomes even more extreme when considering vocal signals attenuated by -10 dB where `TB-A`'s performance stays almost constant ($\mathrm{F} = 0.76$) while `F0-M`'s average F-measure goes down to $\mathrm{F} = 0.25$, many of the individual items scoring F-measures of zero.

The reason for this trend becomes obvious when investigating individual items. Figure 5 depicts the vibrato detection results of both `TB-A` and `F0-M` in all mixing conditions for the item *Leaf—Full*. In the condition -0 dB, the results of `TB-A` (Figure 5a) and `F0-M` (Figure 5b) coincide well with the ground truth (Figure 5c), leading to high F-measures ($\mathrm{F} = 0.84$ and $\mathrm{F} = 0.86$). Here, our template-based analysis approach detects most of the spectro-temporal vibrato patterns in the signal's spectrogram (time-frequency bins where the vibrato salience exceeds the threshold τ are indicated in red in Figure 5a). `F0-M` also achieves a good result since the F0-trajectory extracted by MELODIA (indicated in blue in Figure 5b) captures the singing voice's fundamental frequency well

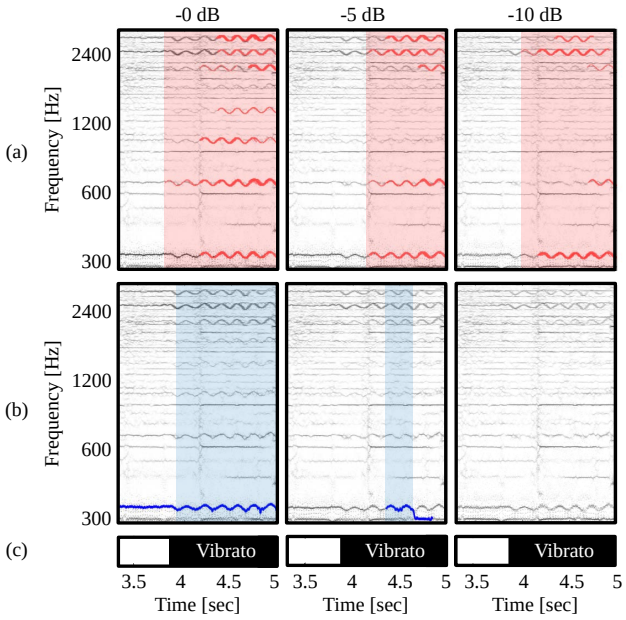

Figure 5. Comparison of `TB-A` and `F0-M` for the item *Leaf—Full*. **(a):** `TB-A`. Automatically derived vibrato passages are indicated in red. **(b):** `F0-M`. Manually annotated vibrato passages in the trajectory are indicated in blue. **(c):** Ground truth annotation.

in this mix. However, this changes when attenuating the vocal signal by -5 dB. While `TB-A` still identifies many vibrato patterns, therefore detecting the vibrato present in the mix ($\mathrm{F} = 0.74$), the F0-estimation becomes problematic and MELODIA retrieves only a small segment of the singing voice's F0-trajectory correctly, leading to a poor vibrato detection ($\mathrm{F} = 0.29$). When attenuating x_{voc} by -10 dB, the F0-trajectory's estimation fails completely ($\mathrm{F} = 0.00$) since MELODIA's assumption of a predominant melodic voice is violated. On the other hand, our proposed detection procedure is capable of detecting the vibrato in the mix.

As a final remark, note that our proposed approach also succeeds to recognize that the item *Funny Valentines—Sleigh Ride* does not contain any vibrato at all.

3.2 Evaluation: Vibrato Analysis

As we have seen in the previous section, the vibrato salience spectrogram S can be used to determine *when* vibrato is present in a music signal. Additionally, when computing S, we also implicitly obtain information about the vibrato's parameters. The rate and extent of vibrato present in the music signal are encoded by the similarity maximizing vibrato templates T in Equation (6). In Figure 6a, we see the log-frequency spectrogram of a mixture of piano music (no vibrato) and three consecutive artificial vibrato tones. The tones have vibrato rates of seven, five, and ten Hertz and extents of 40, 200, and 70 cents, respectively. Time-frequency bins where the vibrato salience exceeds τ are indicated in red. Note that for this experiment we used a much larger template set $\mathcal{T}$, consisting of 285 tem-

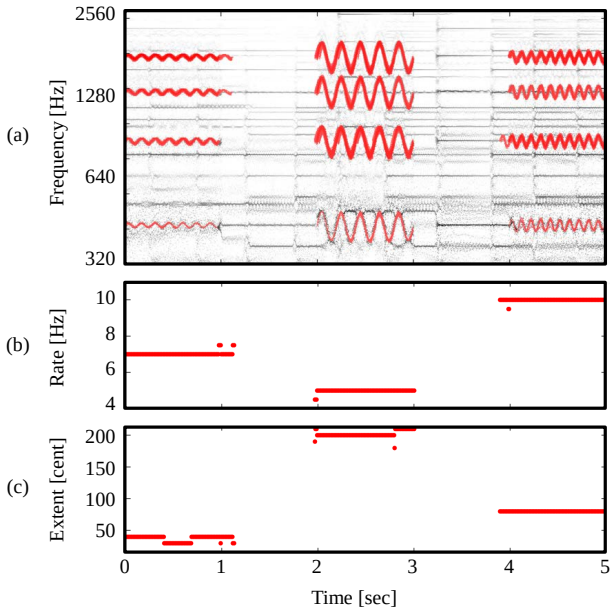

Figure 6. Vibrato rate and extent analysis. **(a):** Log-frequency spectrogram. Time-frequency bins (m, k) with $S(m, k) > \tau$ are indicated in red. **(b)/(c):** Vibrato rate and extent of the template T with the highest vibrato salience per frame.

plates that reflected vibrato rates from four to eleven Hertz in steps of 0.5 Hertz, as well as extents from 30 to 210 cents in steps of 10 cents. Figures 6b/c indicate the vibrato rate and extent of the vibrato template T that maximized the vibrato salience per frame. The two plots correctly reflect the tones' vibrato rates and extents, while showing only a few outliers. Note that values in the plots are quantized since our approach can only give estimates for rates and extents as they are reflected by one of the templates in $\mathcal{T}$. This kind of vibrato analysis could be helpful in scenarios like informed instrument identification when it is known that different instruments in a music signal perform with different vibrato rates or extents.

3.3 Challenges

In general, our proposed procedure yields useful analysis results for the music examples discussed in the previous sections. We now want to discuss a few difficult examples.

One potential source for incorrect analysis results are false positives as visualized in Figure 7a, which shows a log-frequency spectrogram excerpt of *Sunshine Garcia Band—For I Am The Moon* from our dataset. In this excerpt, one of our vibrato templates T is similar enough (with respect to our similarity measure) to a non-vibrato spectro-temporal pattern to yield vibrato salience values above the threshold τ. This could cause incorrect vibrato detection results or meaningless vibrato parametrizations. However, we experienced such spurious template matches to often occur in an isolated fashion. Here, one could exploit additional cues such as multiple template matches at the same time instance due to overtone structures of instruments to reinforce the vibrato analysis' results.

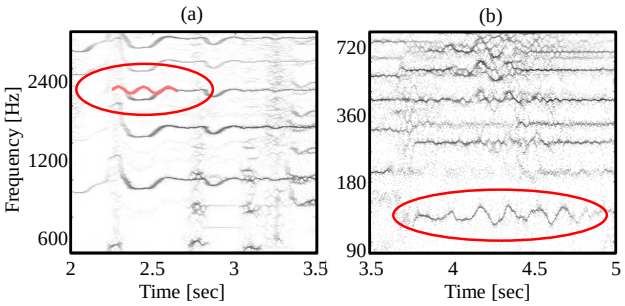

Figure 7. Error sources for our template-based vibrato analysis. **(a):** Spurious template matches. **(b):** Vibrato does not have a sinusoidal form.

The opposite situation is visualized in Figure 7b. It shows a log-frequency spectrogram excerpt of "Gute Nacht", a song from Schubert's "Winterreise" for piano and tenor. In this excerpt, the singer sings a long note with strong vibrato. However, although there is a template reflecting an appropriate vibrato rate and extent in our template set $\mathcal{T}$, the vibrato is not detected by our procedure. This is the case since by our vibrato template's design—as described in Section 2.2—we generally assumed vibrato to have a sinusoidal spectro-temporal structure. This assumption is violated in the shown vibrato pattern. However, our approach is conceptually not limited to sinusoidal vibrato templates and one could further improve the templates' design in order to also capture these kind of vibrato patterns.

4. CONCLUSIONS AND FUTURE WORK

In this paper we presented a novel approach for analyzing vibrato in complex music signals. By locally comparing a signal's spectrogram with a set of predefined vibrato templates, we derived a vibrato salience spectrogram—a kind of mid-level feature representation—in order to locate and parameterize spectro-temporal vibrato patterns. Our approach has the advantage that the analysis does not rely on the estimation of a (possibly erroneous) F0-trajectory. Experiments indicated that our proposed procedure allows for a more robust vibrato detection than F0-based approaches, in particular for complex music signals.

In future work we would like to further explore the use of vibrato templates in various application scenarios. For example, deriving spectral masks from the vibrato salience spectrogram S could open up novel ways of decomposing a music signal into vibrato and non-vibrato components. Furthermore, we believe that the use of vibrato templates could be beneficial for tasks like F0-tracking [14, 19] or performance analysis [1].

Acknowledgments:
This work has been supported by the German Research Foundation (DFG MU 2686/6-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institut für Integrierte Schaltungen IIS. Sebastian Ewert is funded by the EPSRC (EP/L019981/1).

5. REFERENCES

[1] Jakob Abeßer, Hanna M. Lukashevich, and Gerald Schuller. Feature-based extraction of plucking and expression styles of the electric bass guitar. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2290–2293, Dallas, Texas, USA, March 2010.

[2] Jonathan Driedger, Stefan Balke, Sebastian Ewert, and Meinard Müller. Accompanying website: Towards template-based vibrato analysis in complex music signals. `http://www.audiolabs-erlangen.de/resources/MIR/2016-ISMIR-Vibrato/`.

[3] K. Glossop, P. J. G. Lisboa, P. C. Russell, A. Siddans, and G. R. Jones. An implementation of the hough transformation for the identification and labelling of fixed period sinusoidal curves. *Computer Vision and Image Understanding*, 74(1):96–100, 1999.

[4] Perfecto Herrera and Jordi Bonada. Vibrato extraction and parameterization in the spectral modeling synthesis framework. In *Digital Audio Effects Workshop (DAFX98)*, Barcelona, Spain, November 1998.

[5] Yoshiyuki Horii. Acoustic analysis of vocal vibrato: A theoretical interpretation of data. *Journal of Voice*, 3(1):36–43, 1989.

[6] Bernhard Lehner, Gerhard Widmer, and Reinhard Sonnleitner. On the reduction of false positives in singing voice detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7480–7484, Florence, Italy, 2014.

[7] Meinard Müller. *Fundamentals of Music Processing*. Springer Verlag, 2015.

[8] Hee-Suk Pang. On the use of the maximum likelihood estimation for analysis of vibrato tones. *Applied Acoustics*, 65(1):101–107, 2004.

[9] Jeongsoo Park and Kyogu Lee. Harmonic-percussive source separation using harmonicity and sparsity constraints. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 148–154, Málaga, Spain, 2015.

[10] Eric Prame. Measurements of the vibrato rate of ten singers. *The Journal of the Acoustical Society of America (JASA)*, 96(4):1979–1984, 1994.

[11] Eric Prame. Vibrato extent and intonation in professional western lyric singing. *The Journal of the Acoustical Society of America (JASA)*, 102(1):616–621, 1997.

[12] Lise Regnier and Geoffroy Peeters. Singing voice detection in music tracks using direct voice vibrato detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1685–1688, Taipei, Taiwan, 2009.

[13] S. Rossignol, P. Depalle, J. Soumagne, X. Rodet, and J.-L. Collette. Vibrato: detection, estimation, extraction, modification. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Trondheim, Norway, December 1999.

[14] Justin Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, 2012.

[15] Justin Salamon, Emilia Gómez, Daniel P. W. Ellis, and Gaël Richard. Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine*, 31(2):118–134, 2014.

[16] Carl E. Seashore. The natural history of the vibrato. 17(12):623–626, 1931.

[17] Mike Senior. Mixing secrets for the small studio—additional resources. `www.cambridge-mt.com/ms-mtk.htm`. Web resource, last consulted in January 2016.

[18] T. Shipp, R. Leanderson, and J. Sundberg. Some acoustic characteristics of vocal vibrato. *Journal of Research in Singing*, IV(1):18–25, 1980.

[19] Fabian-Robert Stöter, Nils Werner, Stefan Bayer, and Bernd Edler. Refining fundamental frequency estimates using time warping. In *Proceedings of EUSIPCO 2015*, Nice, France, September 2015.

[20] Johan Sundberg. Acoustic and psychoacoustic aspects of vocal vibrato. *STL-QPSR*, 35(2-3):45–68, 1994.

[21] Hideyuki Tachibana, Nobutaka Ono, and Shigeki Sagayama. Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):228–237, January 2014.

[22] Renee Timmers and Peter Desain. Vibrato: Questions and answers from musicians and science. In *Proceedings of the International Conference on Music Perception and Cognition (ICMPC)*, volume 2, 2000.

[23] Luwei Yang, Elaine Chew, and Khalid Z. Rajab. Vibrato performance style: A case study comparing erhu and violin. In *Proceedings of the International Conference on Computer Music Modeling and Retrieval (CMMR)*, 2013.