Adaptive Time-frequency Scaterring for Periodic Modulation Recognition in Music Signals

Changhong Wang, Emmanouil Benetos, Vincent Lostanlen, Elaine Chew

ADAPTIVE TIME–FREQUENCY SCATTERING FOR PERIODIC MODULATION RECOGNITION IN MUSIC SIGNALS

Changhong Wang[1], Emmanouil Benetos[1], Vincent Lostanlen[2], Elaine Chew[3]
[1]Centre for Digital Music, Queen Mary University of London, UK
[2]Music and Audio Research Laboratory, New York University, NY, USA
[3]CNRS-UMR9912/STMS IRCAM, Paris, France
{changhong.wang,emmanouil.benetos}@qmul.ac.uk;
vincent.lostanlen@nyu.edu; elaine.chew@ircam.fr

ABSTRACT

Vibratos, tremolos, trills, and flutter-tongue are techniques frequently found in vocal and instrumental music. A common feature of these techniques is the periodic modulation in the time–frequency domain. We propose a representation based on time–frequency scattering to model the inter-class variability for fine discrimination of these periodic modulations. Time–frequency scattering is an instance of the scattering transform, an approach for building invariant, stable, and informative signal representations. The proposed representation is calculated around the wavelet subband of maximal acoustic energy, rather than over all the wavelet bands. To demonstrate the feasibility of this approach, we build a system that computes the representation as input to a machine learning classifier. Whereas previously published datasets for playing technique analysis focus primarily on techniques recorded in isolation, for ecological validity, we create a new dataset to evaluate the system. The dataset, named CBF-periDB, contains full-length expert performances on the Chinese bamboo flute that have been thoroughly annotated by the players themselves. We report F-measures of 99% for flutter-tongue, 82% for trill, 69% for vibrato, and 51% for tremolo detection, and provide explanatory visualisations of scattering coefficients for each of these techniques.

1. INTRODUCTION

Expressive performances of instrumental music or singing voice often abound with vibratos, tremolos, trills, and flutter-tongue. A common feature of these four playing techniques is that they all result in some periodic modulation in the time–frequency domain. However, from a musical standpoint, these techniques convey distinct stylistic effects. Discriminating between these spectrotemporal patterns requires a compact and informative representation that remains stable to time shifts, time warps, and frequency transpositions. Time–frequency scattering [2] provides such mathematical guarantees. Besides the local invariance to translation and stability to deformation provided by the scattering transform [1, 10], time–frequency scattering goes further by applying frequential scattering along the log-frequency axis. This operation provides invariance to frequency transposition and captures regularity along log-frequency dimension.

Prior work in the representation of vibrato and tremolo can be divided into three broad categories: F_0-based representations [4, 14, 20], template-based techniques [5], and modulation spectra [13, 15, 17]. The error-prone stage of fundamental frequency estimation hinders the performance of F_0-based methods. Template-based methods may work for vibratos with a large modulation extent (frequency variation), while for subtly-modulated vibratos, both the definition of templates and the matching between templates and test segments are problematic. The modulation spectra is another well-known representation for modulated sounds, which is averaged on the audio clip level [17]. It may work well for long-term music information retrieval tasks such as genre classification or instrument recognition, but struggling with providing temporal positions for short-duration playing technique recognition. To our knowledge, there is not yet any computational research that compares and discriminates between these periodic modulations in real-world music pieces.

Besides the question of coming up with an adequate signal representation, there is a critical need for human-annotated playing techniques in audio recordings. Up to now, most of the available research literature has focused on playing techniques that have been recorded in highly controlled environments [8, 16, 19]. Yet, recent findings demonstrate that, in the context of a music piece, playing techniques exhibit considerable variations as compared to when they are played in isolation [18]. For periodic modulations, these variations are more evident in folk music, which highly depends on the interpretation of the performer. Such inter-performer variability in folk music performance necessitates data collection with full pieces.

This paper includes three contributions: representation, application, and dataset. We propose a representation based on time–frequency scattering to model the inter-

class variability for fine discrimination of vibrato, tremolo, trill, and flutter-tongue. Rather than decomposing all the wavelet bands as the scattering transform, we calculate a time–frequency scattering around the wavelet subband of maximal acoustic energy, i.e. the transform is calculated adaptively on the predominant frequency. On the application side, to our knowledge this is the first attempt at creating a system for detecting and classifying periodic modulations in music signals. To evaluate our methodology, we create a dedicated dataset of the Chinese bamboo flute, also known as the *dizi* or *zhudi*, and thereafter abbreviated as CBF. This dataset, named CBF-periDB, contains full-length solo performances recorded by professional CBF players and has been thoroughly annotated by the players themselves.

The rest of this paper is organised as follows. The characteristics of each periodic modulation and how this information can be represented by an adaptive time–frequency scattering are described in Section 2. Section 3 shows details of the feature extraction process and the proposed recognition system. The dataset, evaluation methodology, and results are discussed in Section 4. Section 5 presents our conclusions and directions for future research.

2. SCATTERING REPRESENTATION OF PERIODIC MODULATIONS

Prior to discriminating between the four periodic modulations, we analyse characteristics of each modulation in Section 2.1. A short introduction of the scattering transform is provided in Section 2.2. Section 2.3 describes the proposed representation for modelling periodic modulations.

2.1 Characteristic Statistics of Periodic Modulations

The characteristic statistics of each modulation in discussion are shown in Table 1. As can be seen, flutter tonguing has a much higher modulation rate as compared to the other three modulations; thus, the modulation rate can be used as a main feature to distinguish it from others. For the other three techniques with similar modulation rate, the discriminative information lies in the modulation extent and shape of the modulation unit. The *modulation unit* refers to the unit pattern that repeats periodically within the modulation. It can be one-dimensional, either amplitude modulation (AM) or frequency modulation (FM), or two-dimensional as a spectro-temporal modulation. This can be intuitively observed from the partially enlarged spectrograms given in Fig. 1. Trills are note-level modulations, for which the frequency variations are larger than one semitone. This extent of modulation is much larger than vibratos and tremolos. The shape of the modulation unit for trill is more square-like rather than sinusoidal ones which vibratos and tremolos exhibit. The difference between vibrato and tremolo is that vibratos are FMs, while tremolos are AMs. We show later how this discriminative information is encoded into the proposed representation in Section 2.3.

Type	**Rate** (Hz)	**Extent**	**Shape**
Flutter-tongue	25-50	< 1 semitone	Sawtooth-like
Vibrato	3-10	< 1 semitone	Sinusoidal (FM)
Tremolo	3-8	$\approx$ 0 semitone	Sinusoidal (AM)
Trill	3-10	Note level	Square-like

Table 1. Characteristic statistics of four periodic modulations in music signals.

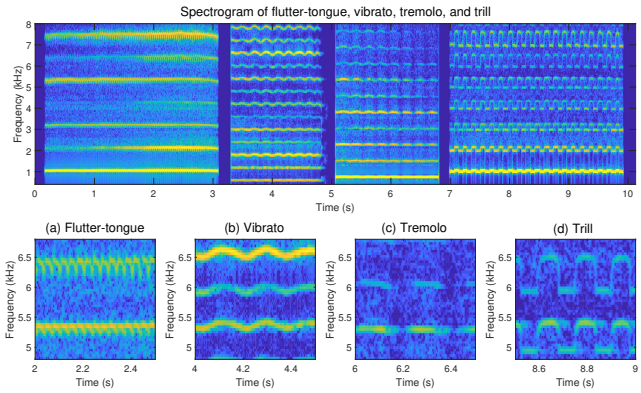

Figure 1. Visual comparison of four periodic modulations. Top: Spectrogram of flutter-tongue, vibrato, tremolo, and trill; bottom: partially enlarged spectrogram of each modulation for detailed comparison.

2.2 Scattering Transform

Proposed by [10], the scattering transform is a cascade of wavelet transforms and nonlinearities. Its structure is similar to a deep convolutional network. The difference is that its weights are not learnt but can be hand-crafted to encode prior knowledge about the task at hand. Since little energy is captured by the scattering transform with orders higher than two [2], we focus in this paper to the second order.

Let ψ_λ denote the wavelet filter bank obtained from a mother wavelet ψ, where λ is the centre frequency of each wavelet in the filter bank. Likewise, ψ_{λ_m} refers to the wavelet filter bank of the m^{th}-order scattering transform, for $m \geq 1$. The second-order temporal scattering transform of a time-domain signal x is defined as:

$$\left| \left| x \overset{\text{t}}{*} \psi_{\lambda_1} \right| \overset{\text{t}}{*} \psi_{\lambda_2} \right| \overset{\text{t}}{*} \phi_T, \tag{1}$$

where $\overset{\text{t}}{*}$ is the wavelet convolution along time. $|x \overset{\text{t}}{*} \psi_{\lambda_m}|$ is the modulus of the m^{th}-order wavelet transform, which hereafter we refer to as the m^{th}-order wavelet modulus transform. The temporal scattering coefficients are obtained by an averaging at a time scale T by means of a low-pass filter ϕ_T.

In addition to the invariance to time-shifts and time-warps provided by the scattering transform, time–frequency scattering provides frequency transposition invariance by adding a wavelet transform along log-frequency axis [7]. The specific time–frequency scattering we apply is a *separable* scattering proposed in [3]. Here,

separable refers to separate steps of temporal and frequential operations of wavelet scattering, arranged in a cascade. The separable scattering representation comprises a second-order temporal scattering and a first-order frequential scattering. The latter is calculated by another wavelet transform along the log-frequency dimension on top of the second-order temporal scattering:

$$\left|\left|\left|x \overset{t}{*} \psi_{\lambda_1}\right| \overset{t}{*} \psi_{\lambda_2}\right| \overset{t}{*} \phi_T \overset{fr}{*} \psi_{\gamma_1}\right| \overset{fr}{*} \phi_F. \quad (2)$$

where $\overset{fr}{*}$ is the wavelet convolution along log-frequency. ψ_{γ_1} is the wavelet filter bank applied in the first-order frequential scattering. The frequential scattering coefficients are obtained by an averaging of the frequential wavelet modulus transform with transposition invariance of F (in octave unit) using a low-pass filter ϕ_F. All scattering coefficients in this paper are normalised and have their logarithm calculated, to capture only the temporal structure and to motivate auditory perception [2]. Hereafter, we use Morlet wavelets throughout the whole scattering network for wavelet convolutions. This is because Morlet wavelets have an exactly null average while reaching a quasi-optimal tradeoff in time–frequency localisation [9]. Our source code is based on the ScatNet toolbox [1] .

2.3 Scattering Representation of Periodic Modulations

Periodic modulation recognition, as suggested by the analysis above, is a pitch invariant task. The core discriminative information is based on the modulation itself, which is indicated by its modulation rate, extent, and shape. Fig. 2 shows respectively (a) the spectrogram, (b) the second-order temporal scattering representation, and (c) the first-order frequential scattering representation of a series of periodic modulation examples in CBF-periDB. The spectrogram used here is only for illustration purposes. The first four examples are regular cases (modulations based on stable pitch or with constant parameters): vibrato, tremolo, trill, and flutter-tongue. The last three are cases with time-varying parameters (modulations based on time-varying pitch or with time-varying rate, extent, or shape): rate-changing trill, rate- and extent-changing trill, and flutter-tongue with time-varying pitch. We use these examples to show how the characteristic information of each pattern is captured and discriminated by a separable scattering transform, which consists of a second-order temporal scattering and a first-order frequential scattering transform.

2.3.1 Second-order temporal scattering

Different from a standard two-order temporal scattering transform, we do not decompose all the frequency bands (exchangeable with wavelet bands) in the first-order wavelet modulus transform. As can be observed from Fig. 2 (a), the patterns of these modulations, either in the regular case or time-varying case are similar for each harmonic partial. This indicates that the decomposition of one partial is sufficient to capture modulation information. Fig. 2 (b) shows the second-order temporal scattering representation decomposed only from the frequency band with the highest energy. Flutter-tongue is the most discriminable one with the highest modulation rate. For the other three patterns with close modulation rate value, other characteristic information is considered. By dominant band decomposition, the trill can also be discriminated because of its large modulation extent. This can be interpreted by filters with bandwidth larger than one semitone, which blurs other subtle modulations.

To specifically detect vibrato or tremolo, frequency bands less than one semitone should be obtained. We then make use of their modulation shape information by introducing a *band-expanding* technique. Assume we have frequency bands of 1/16 octave bandwidth in the first-order wavelet modulus transform. Ideally for tremolo, the modulation information is contained only in the dominant frequency band since it is an AM. This is verified by the second example in Fig. 2 (b), which has almost only the fundamental modulation rate with no upper harmonics in the second-order temporal scattering representation. However, vibratos are FMs, which means the modulation information spreads over neighbouring frequency bands. Decomposing neighbouring frequency bands above or below the dominant band provides additional information to distinguish vibrato from tremolo. All this discriminative information can be visualised from the fundamental modulation rate and the richness of the harmonics in the second-order temporal scattering representation in Fig. 2 (b).

2.3.2 First-order frequential scattering

The temporal scattering transform is sensitive to attacks and amplitude modulations, which results in the high en-

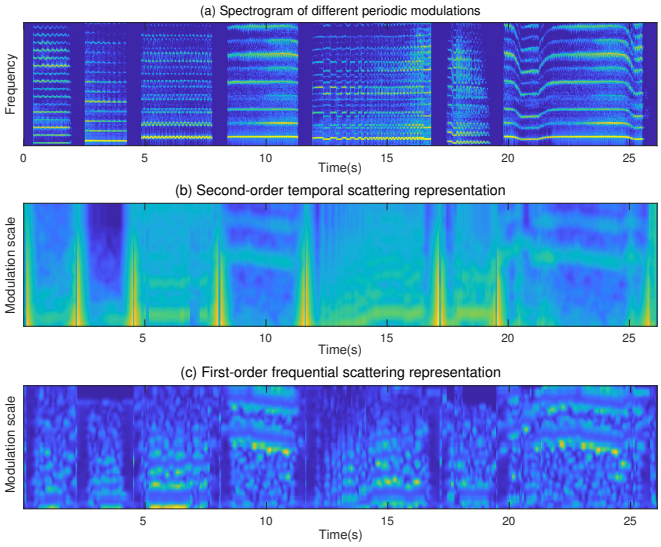

Figure 2. Separable scattering representation of different periodic modulations. From left to right, the first four are regular cases: vibrato, tremolo, trill, flutter-tongue based on stable pitches; the last three are time-varying cases: rate-changing trill, rate- and extent-changing trill, flutter-tongue with time-varying pitch.

[1] https://www.di.ens.fr/data/software/scatnet/

ergy part of the boundaries as clearly observed in Fig. 2 (b). To suppress this noisy information while retaining the frequential structure offered by the second-order temporal scattering, we use frequential scattering along the log-frequency axis on top of Fig. 2 (b). Frequential scattering has a similar framework as temporal scattering while the former captures regularity along log-frequency. As shown in Fig. 2 (c), we obtain a clearer representation without reducing the discriminative information necessary for the task. Although the last example is flutter-tongue bounded to time-varying pitch, its modulation rate is relatively stable. This verifies our method of using just the dominant frequency band or expanded frequency bands from the first-order wavelet modulus transform. The rate-changing and rate-extent changing cases show that the time-varying modulation rates are also captured.

3. PERIODIC MODULATION RECOGNITION

With the proposed representation prepared, we build a recognition system consisting of four binary classification schemes that each predicts one modulation type. Section 3.1 describes the feature extraction process. The calculated features are then fed to a machine learning classifier illustrated in Section 3.2.

3.1 Feature Input

As described in Section 2.3, we adapt the scattering transform by decomposing the dominant and its neighboring frequency bands in the first-order wavelet modulus transform. The feature extraction process of dominant band decomposition is shown in Fig. 3. Using a waveform as input, we first obtain the first-order wavelet modulus transform, where the frequency band with the highest energy for each time frame is localised. Decomposing these bands, we obtain a second-order temporal scattering. A first-order frequential scattering is then conducted on top of the second-order temporal scattering coefficients. Concatenating the two representations in a frame-wise manner, we obtain the feature input to the classifiers. For expanded band decomposition, additional features are calculated similarly by decomposing the neighbouring frequency bands around the dominant band.

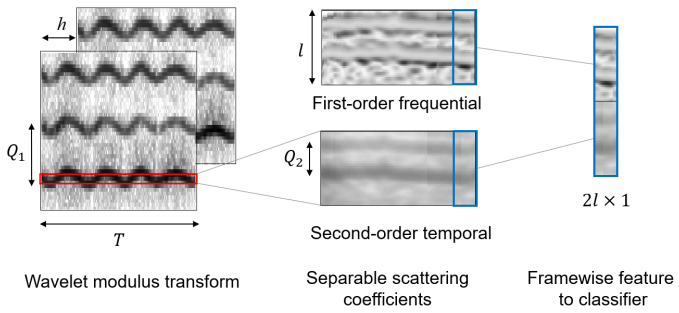

Figure 3. Feature extraction process.

Table 2 gives the parameters which encode the core discriminative information for the recognition. T is the averaging scale for the temporal scattering coefficients. This parameter is useful for discriminating modulations with large differences on modulation rate, for example, on distinguishing flutter-tongue from other low-rate periodic modulations. Averaging scales covering at least four unit patterns are recommended for reliable estimation of the modulation rate. Q_1 are the filters per octave in the first-order temporal scattering transform. Since the modulations discussed here are all oscillatory patterns, setting Q_1 should ensure that each of the modulations are not blurred in the first-order wavelet modulus transform. Here, we use $Q_1 > 12$ for the first-order temporal scattering to support subtly-modulated vibratos and tremolos, of which the modulation extent is less that one semitone. N is the number of neighbouring frequency bands besides the dominant band decomposed from the first-order wavelet modulus transform. $N = 0$ refers to dominant band decomposition only while $N > 0$ means expanded band decomposition. This is a key parameter to encode the unit shape information of subtle modulations. However, if the task at hand is only to detect modulations with high modulation rate or with large extent, this parameter is not necessary.

Parameter	Notation	Main information encoded
Averaging scale	T	Modulation rate
Filters per octave	Q_1	Modulation extent
Expanded bands	N	$= 0$, temporal shape
		> 0, spectro-temporal shape

Table 2. Parameters encoding modulation information in the adaptive time–frequency scattering framework.

Other parameters involved in the feature calculation include frame-size h, filters per octave Q_2 in the second-order scattering decomposition, frequency bands l extracted from the second-order temporal scattering. For frequential scattering, we apply a single scale wavelet transform. The frame size is inversely log-proportional to the oversampling parameter α by $h = T/2^{\alpha}$ (samples), which is designed to compensate for the low temporal resolution resulting from the large averaging scales. Since these parameters carry little discriminative information, we set them consistently for all classification schemes, with $\alpha = 4$, $Q_2 = 8$, and $l = 2Q_2$ based on experimental results. A general example with $T = 2^{15}$ corresponds to frame size of $h = T/2^{\alpha} = 2048$ samples (46ms, assuming the sampling rate is 44.1kHz). The dimensionality of the final representation at each time frame equals to $(N+1) \times 2l = 4(N+1)Q_2$.

3.2 Recognition System

Due to the existence of combined playing techniques, such as the combination of flutter-tongue and vibrato, and the combination of tremolo with trill, one frame of the input may have multiple labels. Multi-label classification is considered beyond the scope of this paper and is regarded as future work. Here, we conduct binary classifications for

each modulation, which enables us to explicitly encode the characteristic information specifically for the corresponding pattern. Four binary classifiers are constructed using support vector machines (SVMs) with Gaussian kernels. The model parameters to be optimized in the training process are the error penalty parameter and the width of the Gaussian kernel [6]. The best parameters selected in the validation stage are used for testing. The input feature to the classifiers is the proposed adaptive time–frequency transform of the current time frame.

Taking flutter-tongue as an example, its relatively high modulation rate (25-50Hz) can be emphasized by setting $T = 8192$ (sampling rate is 44.1kHz). The modulation extent which is less than one semitone is interpreted by setting $Q_1 = 16$. Fig. 4 shows the detection result of flutter-tongue from a piece in CBF-periDB using dominant band decomposition ($N = 0$), which can be clearly observed from the harmonic structure in Fig. 4 (a). This is then enforced by removing the noisy attacks using a frequential scattering transform as shown in Fig. 4 (b). Concatenating the two representations, we form frame-wise separable scattering feature vectors with $(N + 1) \times 2l = 32$ dimensions and frame size of $h = T/2^{\alpha} = 512$ samples (12ms). Fig. 4 (c) visualises the binary classification result of flutter-tongue compared with the ground truth for an example excerpt. Overall it can be seen that the proposed approach is successful at detecting flutter-tongue, even in short segments, although occasionally the output is over-fragmented. Similarly, binary classifiers can be implemented for detecting vibratos, tremolos, and trills, with parameters fine-tuned to the corresponding modulations.

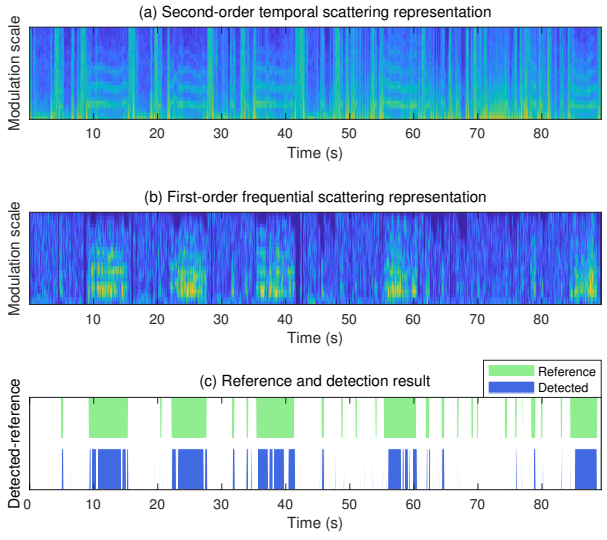

Figure 4. Binary classification result of flutter-tongue in an example excerpt of the CBF-periDB.

4. EVALUATION

4.1 Dataset

To verify the proposed system, we focus on folk music recordings which have more inter-performer variations than Western music. The proposed periodic modulation analysis dataset, CBF-periDB, comprises monophonic performances recorded by ten professional CBF players from the China Conservatory of Music. All data is recorded in a professional recording studio using a Zoom H6 recorder at 44.1kHz/24-bits. Each of the ten players performs both isolated periodic modulations covering all notes on the CBF and two full-length pieces selected from *Busy Delivering Harvest* «扬鞭催马运粮忙», *Jolly Meeting* «喜相逢», *Morning* «早晨», and *Flying Partridge* «鹧鸪飞». Players are grouped by flute type (C and G, the most representative types for Southern and Northern styles, respectively) and each player uses their own flute. This dataset is an extension of the CBF-glissDB dataset in [18], with ten pieces containing periodic modulations added. The playing techniques are thoroughly annotated by the players themselves. Details of both isolated techniques and full-piece (performed) recordings are shown in Table 3. The dataset and annotations can be downloaded from `c4dm.eecs.qmul.ac.uk/CBFdataset.html`.

Isolated		**Performed**	
Type	Length	Piece, number	Length
Flutter-tongue	4.9	Mo, 3	16.0
Vibrato	7.3	BH, 7	28.0
Tremolo	5.0	JM, 4	12.4
Trill	12.3	FP, 6	51.9

Table 3. Length of both isolated techniques and full-piece recordings in CBF-periDB (Mo=*Morning*; JM=*Jolly Meeting*; BH=*Busy Delivering Harvest*; FP=*Flying Partridge*; all numbers for length are measured in minutes).

In the recognition implementation process, the dataset is split into a 6:2:2 ratio according to players (players are randomly initialised) and a 5-fold cross-validation is conducted. This way of data splitting ensures a non-overlap between players in the train, validation, and test sets. Since each player uses their own flute during recording, there is no overlap across flutes. Due to the limited piece types, non-overlap of pieces is not feasible at the current stage. We regard this as future work with a dataset expansion plan to address piece diversity.

4.2 Metrics

Due to the short duration and periodic nature of vibratos, tremolos, trills, and flutter-tongue, feature dependencies over sequential frames are slightly required. The precision $\mathcal{P} = \frac{TP}{TP+FP}$, recall $\mathcal{R} = \frac{TP}{TP+FN}$, and F-measure $\mathcal{F} = \frac{2\mathcal{P}\mathcal{R}}{\mathcal{P}+\mathcal{R}}$ are used here for frame-based evaluation, where TP, FP, FN are true positives, false positives, and false negatives respectively [12]. Assigned labels by SVMs are then compared to the ground truth annotations in a frame-wise manner.

4.3 Baseline

According to our knowledge, there is not yet any previous work on discriminating between these four periodic mod-

Type	Dominant band						Expanded band					
	Temporal scattering			Separable scattering			Temporal scattering			Separable scattering		
	$\mathcal{P}(\%)$	$\mathcal{R}(\%)$	$\mathcal{F}(\%)$	$\mathcal{P}(\%)$	$\mathcal{R}(\%)$	$\mathcal{F}(\%)$	$\mathcal{P}(\%)$	$\mathcal{R}(\%)$	$\mathcal{F}(\%)$	$\mathcal{P}(\%)$	$\mathcal{R}(\%)$	$\mathcal{F}(\%)$
Flutter-tongue	96.1	99.8	97.9	96.3	99.8	98.0	96.7	99.6	98.1	97.8	99.5	**98.7**
Trills	87.1	66.7	75.1	87.4	68.2	76.2	89.5	73.3	80.4	89.8	76.3	**82.3**
Vibrato	75.2	17.5	26.4	72.2	33.1	45.3	75.9	59.4	66.5	75.1	64.7	**69.3**
Tremolo	92.5	1.21	2.2	80.9	6.7	10.6	70.8	38.5	49.1	67.6	41.4	**50.7**

Table 4. Performance comparison of binary classification for flutter-tongue, vibratos, tremolos, and trills in CBF-periDB using separable scattering and temporal scattering representations based on the dominant frequency band and expanded frequency band decomposition ($\mathcal{P}$=precision; $\mathcal{R}$=recall; $\mathcal{F}$=F-measure).

ulations; thus we compare the proposed systems against a state-of-art detection method for vibrato. The filter diagonalisation method (FDM), which efficiently extracts high resolution spectral information for short time signals, was first applied to vibrato detection in erhu performance [20]. Based on the high similarity of the music style between erhu and CBF, both being traditional Chinese instruments, we use FDM as a baseline method for vibrato detection. Using automatically estimated fundamental frequency by pYIN [11] as input for FDM, we try different parameter ranges for vibrato rate and extent based on the vibrato characteristics of the CBF. The best result we obtain based on a 256ms-frame-wise evaluation is $\mathcal{P}$=36.5%, $\mathcal{R}$=58.7%, $\mathcal{F}$=45.0%. The rate range and extent range we use are 3-10 Hz and 5-20 cent, respectively.

4.4 Results

In order to explicitly show information captured in each classification task, a small set of parameter settings for T, Q_1, and N is used. Besides the different meta-parameter settings specifically designed for each modulation classification, we run further experiments by using the same parameters for all four binary classification processes: $T = 2^{15}$, $Q_1 = 16$, and $N = 6$ frequency bands symmetrically expanded around the dominant band. This corresponds to frame size of 46ms and feature dimension of 224. Note that for specific modulation detection, meta parameters of the scattering transform can be fine-tuned to have a much lower feature dimension, as the flutter-tongue detection example demonstrated in Section 3.2. The binary classification results for each pattern are given in both the dominant band decomposition and expanded band decomposition, as shown in Table 4. The detection results using the second-order temporal scattering coefficients only as feature are also provided. The comparison between temporal scattering and separable scattering verifies our analysis in Section 2.3 that for periodic modulation recognition, frequential scattering in the separable scattering representations provides additional discriminative information by removing noisy information. Better performance for flutter-tongue and trill detection shows that for modulations with high modulation rate and large extent, decomposition of the dominant band is sufficient. For discriminating between temporal modulation and spectro-temporal modulation, expanded band decomposition works much better.

Generally, detection performance on vibrato and tremolo is worse than that for flutter-tongue and trill detection. Identifying the errors in the original audio, we find that in most cases these are combined techniques, i.e. subtle frequency variations are accompanied with amplitude modulations or vice versa. In the case of CBF, such combinations are common because of the instrumental gestures of vibrato and tremolo. Vibrato can be generated by fingering or tonguing, while tremolos are commonly produced by breathing variations. Performers are also expressively intended to add tremolo effects on top of other playing techniques.

5. CONCLUSIONS AND FUTURE WORK

Periodic modulation recognition is a pitch invariant task and should only capture modulation information. This is realised by calculating a time–frequency scattering around the wavelet subband of the maximum acoustic energy. We found that the proposed representation decomposed from the dominant band is sufficient to detect modulations with high modulation rate (flutter-tongue) and large modulation extent (trill), while expanded band decomposition captures subtle frequency modulations (vibrato and tremolo). We introduce the ecologically valid dataset, CBF-periDB, to evalate the recognition system. Results show that the proposed representation captures discriminative information between vibratos, tremolos, trills, and flutter-tongue in real-world pieces.

The current work only considers continuous periodic modulations; other periodic patterns, such as tonguing, will be considered as future work due to their noncontinuous nature and more complicated parameter variations. Inspired by research on polyphonic music transcription, current work may be expanded to polyphonic periodic modulation detection. Additional comparison of the current result with other equivalent representations such as the modulation spectra, will be conducted. We conclude that the scattering transform presents a versatile and compact representation for analysing periodic modulations in performed music and opens up new avenues for computational research on playing techniques.

6. ACKNOWLEDGEMENTS

Changhong Wang is funded by the China Scholarship Council (CSC). Emmanouil Benetos is supported by a RAEng Research Fellowship RF/128. Elaine Chew is supported by the European Union's Horizon 2020 research and innovation program (grant agreement No 788960) under the European Research Council (ERC) Advanced Grant (ADG) project COSMOS.

7. REFERENCES

[1] J. Andén and S. Mallat. Scattering representation of modulated sounds. *15th International Conference on Digital Audio Effects (DAFx)*, 2012.

[2] J. Andén and S. Mallat. Deep scattering spectrum. *IEEE Transactions on Signal Processing*, 62(16):4114–4128, 2014.

[3] C. Baugé, M. Lagrange, J. Andén, and S. Mallat. Representing environmental sounds using the separable scattering transform. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8667–8671, 2013.

[4] Y. P. Chen, L. Su, and Y. H. Yang. Electric guitar playing technique detection in real-world recording based on F0 sequence pattern recognition. In *Proc. Conf. International Society for Music Information Retrieval (ISMIR)*, pages 708–714, 2015.

[5] J. Driedger, S. Balke, S. Ewert, and M. Müller. Template-based vibrato analysis in music signals. In *Proc. Conf. International Society for Music Information Retrieval (ISMIR)*, pages 239–245, 2016.

[6] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning: data mining, inference, and prediction, 2009.

[7] V. Lostanlen. *Convolutional operators in the time-frequency domain*. PhD thesis, PSL Research University, 2017.

[8] V. Lostanlen, J. Andén, and M. Lagrange. Extended playing techniques: the next milestone in musical instrument recognition. In *5th International Conference on Digital Libraries for Musicology (DLfM)*, 2018.

[9] S. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way.* Academic Press, 2008.

[10] S. Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

[11] M. Mauch and S. Dixon. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663, 2014.

[12] M. Müller. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer, 2015.

[13] Y. Panagakis, C. Kotropoulos, and G. R. Arce. Music genre classification via sparse representations of auditory temporal modulations. In *17th IEEE European Signal Processing Conference (Eusipco)*, pages 1–5, 2009.

[14] R. Panda, R. Malheiro, and R. Paiva. Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing*, (1):1–1, 2018.

[15] B. L. Sturm and P. Noorzad. On automatic music genre recognition by sparse representation classification using auditory temporal modulations. *Proc. Computer Music Modeling and Retrieval (CMMR)*, 2012.

[16] L. Su, H. M. Lin, and Y. H. Yang. Sparse modeling of magnitude and phase-derived spectra for playing technique classification. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 22(12):2122–2132, 2014.

[17] S. Sukittanon, L. E. Atlas, and J. W. Pitton. Modulation-scale analysis for content identification. *IEEE Transactions on Signal Processing*, 52(10):3023–3035, 2004.

[18] C. Wang, E. Benetos, X. Meng, and E. Chew. Hmm-based glissando detection for recordings of chinese bamboo flute. In *Proc. Conf. Sound and Music Computing (SMC)*, pages 545–550, May 2019.

[19] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo. Vocalset: A singing voice dataset. In *Proc. Conf. International Society for Music Information Retrieval (ISMIR)*, pages 468–474, 2018.

[20] L. Yang, K. Rajab, and E. Chew. The filter diagonalisation method for music signal analysis: frame-wise vibrato detection and estimation. *Journal of Mathematics and Music*, 11(1):42–60, 2017.