PLAYING TECHNIQUE RECOGNITION BY JOINT TIME–FREQUENCY SCATTERING

Changhong Wang[1], Vincent Lostanlen[2], Emmanouil Benetos[1], 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.

ABSTRACT

Playing techniques are important expressive elements in music signals. In this paper, we propose a recognition system based on the joint time–frequency scattering transform (jTFST) for pitch evolution-based playing techniques (PETs), a group of playing techniques with monotonic pitch changes over time. The jTFST represents spectro-temporal patterns in the time–frequency domain, capturing discriminative information of PETs. As a case study, we analyse three commonly used PETs of the Chinese bamboo flute: acciacatura, portamento, and glissando, and encode their characteristics using the jTFST. To verify the proposed approach, we create a new dataset, the CBF-petsDB, containing PETs played in isolation as well as in the context of whole pieces performed and annotated by professional players. Feeding the jTFST to a machine learning classifier, we obtain F-measures of 71% for acciacatura, 59% for portamento, and 83% for glissando detection, and provide explanatory visualisations of scattering coefficients for each technique.

Index Terms— Music signal analysis, scattering transform, performance analysis, playing technique recognition

1. INTRODUCTION

Performance analysis is an important task in music information retrieval, which focuses on the different ways we play a music piece. A typical example of expressive music playing is the application of playing techniques, such as vibratos and tremolos. The modeling and detection of playing techniques benefits research in automatic transcription of musical ornaments [1, 2], realistic music generation [3], computer-aided music pedagogy [4], instrument classification [5, 6], and performance analysis [7].

When displaying playing techniques in the time–frequency domain, we observe that each has a distinctive spectro-temporal pattern. Fig. 1 shows seven commonly used playing techniques in music signals. The four playing techniques in Fig. 1 (a), vibrato, tremolo, trill, and flutter-tongue, are periodic modulations that elaborate on stable notes and are temporally symmetric. The modulation patterns for their harmonic partials are parallel. Conducting periodicity analysis on one harmonic partial is sufficient for fine discrimination of these playing techniques [8]. We refer to these playing techniques as *pitch modulation-based playing techniques* (PMTs).

CW is funded by the China Scholarship Council (CSC). EB is supported by RAEng Research Fellowship RF/128. EC is supported by the European Union's Horizon 2020 research and innovation program (788960) under the European Research Council (ERC) Advanced Grant (ADG) project COSMOS. We acknowledge Meinard Müller for his valuable advice.

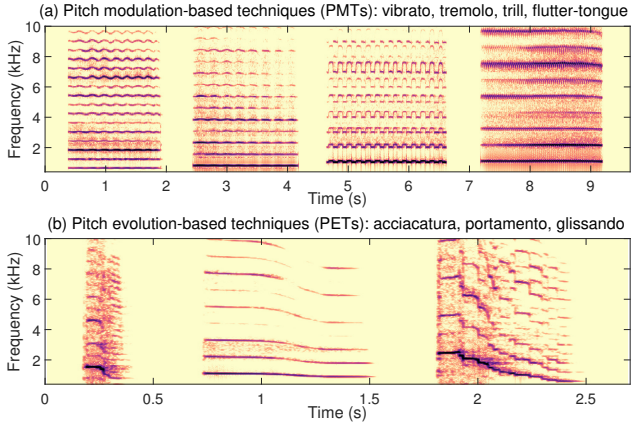

Fig. 1. Spectrograms of (a) PMTs and (b) PETs. Whereas modulations of PMTs were captured by separable scattering [8], PETs are represented by joint scattering due to their evolutionary nature.

However, as regards playing techniques containing monotonic pitch changes, finding a representation with clear music information encoding is a challenging task. Fig. 1 (b) shows three examples from this group of playing techniques: acciacatura, portamento, and glissando. In the case of the Chinese bamboo flute (CBF), these playing techniques are known as 垛音 (duoyin), 滑音 (huayin), and 历音 (liyin), respectively. Acciacatura is played with a sharp attack and lots of air on the first note followed by a rapid transition to the second note, and is a characteristic CBF playing technique. Portamento is a continuous slide between two notes. Glissando is a slide across a series of discrete tones. We call this group of playing techniques *pitch evolution-based playing techniques* (PETs), which contain monotonic pitch evolution over time and are temporally asymmetric. Unlike PMTs with regularity along time, PETs exhibit variations both along time and along frequency, which are thereafter referred to as temporal and spectral variations. The interaction of the two types of variations plays an important role for discriminating PETs.

Prior research on recognising PETs typically focused on only one kind of playing technique, without generalisation across to other playing techniques. Hidden Markov models (HMMs) were used in [9] for detecting glissando and in [7] to recognise portamento. Rule-based features introduced in [10] were specifically for glissando detection. Patterns of regularity across playing techniques motivate us to build a generic model for music playing technique recognition. We find that the scattering transform, an approach for building invariant, stable, and informative signal representations [11], offers such flexibility by its different operators, such as separable scat-

tering and joint scattering. Whereas our previous work for PMT recognition [8] relied on separable time–frequency scattering. This current paper applies the joint time–frequency scattering transform (jTFST) [12] for PET recognition.

This paper includes three contributions: (1) A supervised learning system for detecting and classifying pitch evolution-based playing techniques. This system, based on joint time–frequency scattering, is robust to frequency transposition, variations in instruments, performer identity, and regional musical and playing styles. (2) A new dataset, named CBF-petsDB, comprising full-length Chinese traditional musical recordings and expert playing technique annotations, suitable for computational performance analysis evaluation. (3) A formal interpretation of the role of each component in the joint time–frequency scattering feature extractor, confirmed by explanatory visualizations of real-world acoustic data.

2. JOINT TIME–FREQUENCY SCATTERING FOR REPRESENTING PLAYING TECHNIQUES

2.1. Pitch Evolution-based Playing Techniques (PETs)

Prior to discriminating between acciacatura, portamento, and glissando, we analyse characteristics of each playing technique and calculate statistical information from the CBF-petsDB, as shown in Table 1. Each of these playing techniques has a specific duration range: 0.1–0.4s for acciacatura, 0.2–1.2s for portamento, and 0.2–1.1s for glissando. For temporal variations, although all three playing techniques contain monotonic pitch changes over time, portamento exhibits smooth pitch changes while the pitch changes within acciacatura and glissando are both changes at the note level. Acciacatura contains only one note change, while glissando spans a series of note changes. For spectral variations, acciacatura has a noisy attack while glissando and portamento exhibit clear harmonic structures. The possible directions of their pitch changes are different: acciacatura in CBF playing only occurs downwards, while the other two playing techniques can exhibit both upward and downward directions.

Characteristics	**Acciacatura**	**Portamento**	**Glissando**
Duration (s)	0.1-0.4	0.2-1.2	0.2-1.1
Temporal variation	One note change	Continuous pitch changes	Consecutive note changes
Spectral variation	Noisy attack	Harmonic	Harmonic
Pitch direction	$\searrow$	$\nearrow$ or $\searrow$	$\nearrow$ or $\searrow$

Table 1. Characteristic information of PETs.

2.2. Joint Time–Frequency Scattering Transform

Assume we have a two-dimensional (2D) time–frequency image $\mathbf{X}(t,\lambda)$ obtained, for example, the representations in Fig. 1 (Spectrograms are used throughout the paper for clear visualisation of the spectro-temporal patterns). $t \in \mathbb{R}_+$ is the time variable and $\lambda = \log_2(\mathbb{R}_{\geqslant 1})$ denotes log-frequency. Similar to convolutional neural networks (CNNs) with horizontal and vertical filters [13], the jTFST consists of both temporal (along the time axis) and spectral wavelets (along the frequency axis) [12]. The interaction of the two types of wavelet decompositions captures the spectro-temporal patterns in the time–frequency domain.

Motivated by the recognition task for PETs, we interpret the definition of the jTFST in [12] from a new perspective. Rather than formulating a 2D mother wavelet, we consider the temporal and spectral wavelet convolutions in a sequential manner. This is more precise in terms of what computations perform and provide explicit information of what has been captured at each step.

Let $\boldsymbol{\psi}^{(t)}(t)$ and $\boldsymbol{\psi}^{(f)}(\lambda)$ denote the mother wavelets in the time and frequency domains, respectively. To obtain temporal and spectral wavelet filterbanks, we dilate $\boldsymbol{\psi}^{(t)}(t)$ along t by 2^{-v_t} and $\boldsymbol{\psi}^{(f)}(\lambda)$ along f by 2^{-v_f}. The scaling factors $v_t \in \mathbb{R}_+$ and $v_f \in \mathbb{R}_+$ measure the variations along time and along log-frequency. An orientation variable $\theta = \pm 1$ is introduced to reflect the oscillation direction (up or down) of the spectro-temporal pattern. $\theta = -1$ flips the center frequency of wavelet $\boldsymbol{\psi}^{(f)}(\lambda)$ from λ to $-\lambda$. The resulting temporal and spectral wavelet banks are respectively:

$$\boldsymbol{\psi}^{(t)}_{v_t}(t) = 2^{v_t}\boldsymbol{\psi}^{(t)}(2^{v_t}t) \quad \text{and} \tag{1}$$

$$\boldsymbol{\psi}^{(f)}_{v_f,\theta}(\lambda) = 2^{v_f}\boldsymbol{\psi}^{(f)}(\theta 2^{v_f}\lambda). \tag{2}$$

The joint wavelet transform of $\mathbf{X}(t,\lambda)$ computes convolutions, $\mathbf{X} * \boldsymbol{\psi}^{(t)}_{v_t}(t) * \boldsymbol{\psi}^{(f)}_{v_f,\theta}(\lambda)$. It captures the joint variability of $\mathbf{X}(t,\lambda)$ localised at (t,λ), measured by the temporal variability, v_t, spectral variability, v_f and orientation, θ. For a specific recognition task at hand, we normally focus on a spectro-temporal pattern smaller than a "time–frequency box" restricted by some time scale T and frequency scale F. To ensure time-shift invariance, time-warping stability, frequency-transposition invariance, and frequency-warping stability, we take the modulus of $\mathbf{X} * \boldsymbol{\psi}^{(t)}_{v_t}(t) * \boldsymbol{\psi}^{(f)}_{v_f,\theta}(\lambda)$ and average it by a 2D low-pass filter $\boldsymbol{\phi}_{T,F}$. The joint time–frequency scattering coefficients of $\mathbf{X}(t,\lambda)$ are defined as

$$\mathbf{S}_2\boldsymbol{x}(t,\lambda,v_t,v_f,\theta) = \left|\mathbf{X} * \boldsymbol{\psi}^{(t)}_{v_t}(t) * \boldsymbol{\psi}^{(f)}_{v_f,\theta}(\lambda)\right| * \boldsymbol{\phi}_{T,F}. \tag{3}$$

Notated as $\mathbf{S}_2$ is to form a consistent framework (explained below).

The above analysis is based on the assumption that the time–frequency image $\mathbf{X}(t,\lambda)$ is given. In fact, we can also obtain $\mathbf{X}(t,\lambda)$ from audio waveforms $\boldsymbol{x}(t)$ within the time–frequency scattering framework. In the scattering framework, we refer to $\mathbf{X}(t,\lambda)$ as the scalogram, which is calculated by convolving $\boldsymbol{x}(t)$ with a temporal wavelet bank, $\boldsymbol{\psi}_\lambda(t)$, and calculating the modulus: $\mathbf{X}(t,\lambda) = \left|\boldsymbol{x} * \boldsymbol{\psi}_\lambda(t)\right|$. Averaging $\mathbf{X}(t,\lambda)$ temporally by a low-pass filter $\boldsymbol{\phi}_T$, we obtain the first-order temporal scattering transform, $\mathbf{S}_1\boldsymbol{x}(t,\lambda) = \mathbf{X}(t,\lambda) * \boldsymbol{\phi}_T$. Thus the jTFST obtained on top of $\mathbf{X}(t,\lambda)$ is regarded as the second-order time–frequency scattering coefficients $\mathbf{S}_2\boldsymbol{x}$. To capture only the temporal variation regardless of the absolute energy of the audio waveform, we normalise $\mathbf{S}_2\boldsymbol{x}$ over $\mathbf{S}_1\boldsymbol{x}$. Motivated by auditory perception [14], the logarithm is applied to the normalised coefficients. The log-normalised jTFST is expressed as

$$\widetilde{\mathbf{S}}_2\boldsymbol{x}(t,\lambda,v_t,v_f,\theta) = \log\left(\frac{\mathbf{S}_2\boldsymbol{x}}{\mathbf{S}_1\boldsymbol{x}+\varepsilon}\right), \tag{4}$$

where ε is a regularization parameter that zeros out negligible scattering coefficients.

A diagram of the jTFST is shown in Fig. 2. Convolving (a) the scalogram $\mathbf{X}(t,\lambda)$ with a temporal wavelet bank, $\boldsymbol{\psi}^{(t)}_{v_t}(t)$, we obtain (b) the temporal wavelet transform. (b) mainly captures the temporal variations of each frequency band, without taking the interaction across frequency bands into account. To capture correlations

across temporal wavelet bands, we apply a wavelet convolution with $\psi_{v_f}^{(f)}(\lambda)$ along the log-frequency axis and obtain (c) the jTFST. According to Eq. (4), for each "time–frequency" box around (t, λ), we obtain a three-dimensional tensor (v_t, v_f, θ), which captures the joint activation of temporal and spectral variations, and its direction, as shown in Fig. 2 (d). 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 [11]. Our source code is based on the ScatNet toolbox[1].

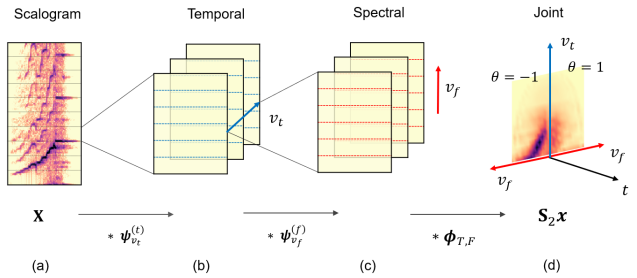

Fig. 2. Diagram of the joint time–frequency scattering transform.

2.3. Joint Time–Frequency Scattering for Playing Techniques

For recognising PETs, we explicitly encode their characteristic information into the jTFST by setting appropriate transform parameters (see Table 2). The averaging scale T (in samples) carries duration information via setting T equivalent to the maximum duration of each type of playing technique. According to the duration of PETs in Table 1, we use $T = 2^{12}$ (93 ms at a sampling rate of 44.1kHz) for acciacatura, $T = 2^{15}$ (743 ms) for portamento, and $T = 2^{14}$ (372 ms) for glissando. $J_1^{(t)} = Q_1^{(t)} \log_2(T)$ is the maximum scale in the first-order wavelet bank, which means that $\lambda = 1, ..., J_1^{(t)}$. Convolving the audio waveform $\boldsymbol{x}(t)$ with the temporal wavelet bank $\boldsymbol{\psi}_\lambda(t)$ using $Q_1^{(t)} \in \mathbb{N}$ filters per octave, we obtain the scalogram $\mathbf{X}(t, \lambda)$. $Q_1^{(t)}$ is useful for distinguishing note changes from smooth pitch changes. For acciacatura and glissando, we set $Q_1^{(t)} = 12$ due to their note-change property. To capture the smooth pitch evolution within portamento, $Q_1^{(t)} > 12$ is required, and we set $Q_1^{(t)} = 16$ for portamento. For the temporal wavelet bank in the second-order jTFST, we have the maximum scale $J_2^{(t)} = Q_2^{(t)} \log_2(T)$. We use $Q_2^{(t)} = 2$ due to their less oscillatory nature, resulting in the wavelet bank $\boldsymbol{\psi}_{v_t}^{(t)}(t)$, with $v_t = 1, ..., J_2^{(t)}$.

One may observe from Fig. 1 (b) the different harmonic structures between the selected PETs. This timbral information can be captured by applying a spectral wavelet bank $\boldsymbol{\psi}_{v_f}^{(f)}(\lambda)$ with $Q_1^{(f)}$ filters per octave. Similar to the temporal wavelet convolution, we have the maximum scale $J_1^{(f)} = Q_1^{(f)} \log_2(Q_1^{(t)} F)$, resulting in the spectral wavelet bank $\boldsymbol{\psi}_{v_f}^{(f)}(\lambda)$ with $v_f = 1, ..., J_1^{(f)}$. The averaging scale, F (in octaves), depends on the frequency transposition invariance requirement of the task. Here we use $Q_1^{(f)} = 2$ filters per octave and $F = 2$ octaves. We then obtain the log-normalised jTFST of PETs for each time frame according to Eq. (3) and Eq. (4).

Fig. 3 shows the jTFST of acciacatura, portamento, and glissando: (a) is the spectrogram; (b), (c), and (d) are the 2D joint activations for each type of PET. As observed, although both acciacatura and glissando have high energy regions in the jTFST, their energy distributions along the variation scales are different. From (b) and (d), noisy attacks show as diffused energy in the jTFST, and the time and frequency regularity of glissando results in clear slopes. Due to the uni-directional nature of acciacatura, we calculate the jTFST only for the downward direction. For portamento and glissando, we calculate the jTFST for both directions and select the one with the maximum energy to form a direction-invariant representation.

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

Characteristics	Parameter for encoding	Notation
Duration	Averaging scale	T
Pitch change	Temporal filters per octave	$Q_1^{(t)}, Q_2^{(t)}$
Harmonicity	Spectral filters per octave	$Q_1^{(f)}$
Pitch direction	Orientation variable	θ

Table 2. Joint scattering parameters which encode discriminative information for PETs.

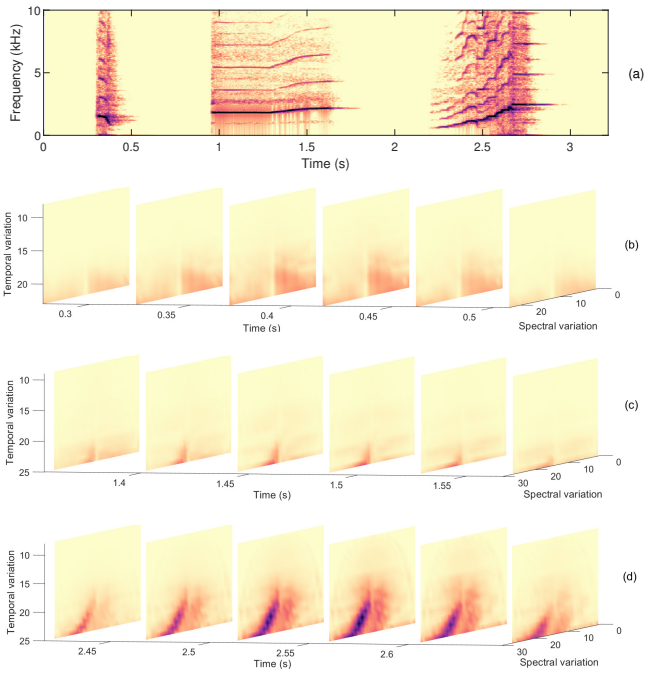

Fig. 3. Joint activation of temporal and spectral variations for PETs. (a) Spectrogram of acciacatura, portamento, and glissando examples. (b), (c), and (d) are the corresponding jTFST for each case.

3. PLAYING TECHNIQUE RECOGNITION

3.1. Dataset

To verify the proposed system, we focus on folk music recordings which often exhibit more inter-performer variations than in Western music. The proposed analysis dataset, CBF-petsDB, comprises monophonic performances recorded by ten professional CBF players from the China Conservatory of Music. All performances were acquired in a studio using a Zoom H6 recorder at 44.1kHz/24-bits. Each of the ten players performed both isolated PETs 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 were grouped by flute type (C and G, the most representative types for Southern and Northern styles, respectively) and each player used their own flute. The dataset and annotations can be downloaded from `c4dm.eecs.qmul.ac.uk/CBFdataset.html`. The number of PET instances in CBF-petsDB is shown in Table 3.

Dataset	Type	Acciacatura	Portamento	Glissando
CBF-petsDB	Isolated	310	495	334
	Performed	119	347	153

Table 3. Number of PET instances in CBF-petsDB.

3.2. Recognition System

Unlike PMTs, which recur across adjacent time frames, the PETs exhibit non-stationary patterns. In this case, long-term context carries essential information for discriminating between PETs. Here, we use N frames centered at the current frame to represent contextual information. To extract coefficients which contain most of the modulation energy, we use only coefficients from one-third of each modulation scale. Concatenating the extracted jTFST, we obtain a long vector with dimensions of 150, 249, and 253 for acciacatura, portamento and glissando at each frame. The mean, standard deviation, and first-order difference of the $N = 5$ frame-long vectors together form the input feature for classification. The frame size h (in samples) is inversely log-proportional to an oversampling parameter α, whereby $h = T/2^{\alpha}$. This is designed to compensate for the low temporal resolution resulting from the large averaging scale T. We use $\alpha = 2$ for all the experiments. Thus, for acciacatura, portamento, and glissando, $T = 2^{12}, 2^{15}$, and 2^{14}, which means their frame sizes h equal 23 ms, 186 ms, and 93 ms, respectively.

With the extracted representation, we build a recognition system consisting of three binary classifiers, one for each type of PET. Our classifier is a support vector machine (SVM) with a Gaussian kernel. The model parameters for optimization in the training process are the error penalty and the width of the Gaussian kernel [15]. In the recognition 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. The best model parameters selected in the validation stage are used for testing.

3.3. Metrics and Baseline

We use precision $\mathcal{P} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$, recall $\mathcal{R} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$, and F-measure $\mathcal{F} = \frac{2\mathcal{P}\mathcal{R}}{\mathcal{P}+\mathcal{R}}$ as the evaluation metrics, where $\mathrm{TP}, \mathrm{FP}, \mathrm{FN}$ are true positives, false positives, and false negatives, respectively [16]. Labels assigned by the SVMs are then compared to the ground truth annotations in a frame-wise manner.

There is not yet any previous work on acciacatura detection and discrimination between these three PETs. Thus we compare the proposed systems against state-of-art detection methods for portamento [7] and glissando [9], respectively, both based on HMMs. In each case, the input is the frame-wise fundamental frequency estimated by pYIN [17]. Both systems are evaluated on a framesize of 20ms. The best F-measures obtained for portamento and glissando are 38% and 49%, respectively, as shown in Table 4.

3.4. Results

Table 4 compares the binary classification results for acciacatura, portamento, and glissando in the CBF-petsDB based on the jTFST and the baseline methods. Better performance for acciacatura and glissando may be attributed to their to more distinctive characteristics. Cross checking the detection errors with the original audio, we find that the false negatives in portamento detection can often be attributed to the players combining playing techniques in a form of co-articulation. Fig. 4 shows the portamento detection result from an example piece compared to the ground truth. The false negative at 126s is an instance of portamento and flutter-tongue co-articulation. In such cases, portamento is no longer smooth but modulated with small ripples, making them hard to detect, even with 16 filters per octave in the first-order wavelet transform.

Dataset	PETs	Joint scattering			Baseline
		$\mathcal{P}(\%)$	$\mathcal{R}(\%)$	$\mathcal{F}(\%)$	$\mathcal{F}(\%)$
CBF-petsDB	Acciacatura	84.2	66.9	71.3	N/A
	Portamento	70.0	51.1	58.6	37.9
	Glissando	83.9	83.6	83.3	48.8

Table 4. Performance comparison of binary classification for acciacatura, portamento, and glissando in the CBF-petsDB using joint scattering and baselines. ($\mathcal{P}$=precision; $\mathcal{R}$=recall; $\mathcal{F}$=F-measure).

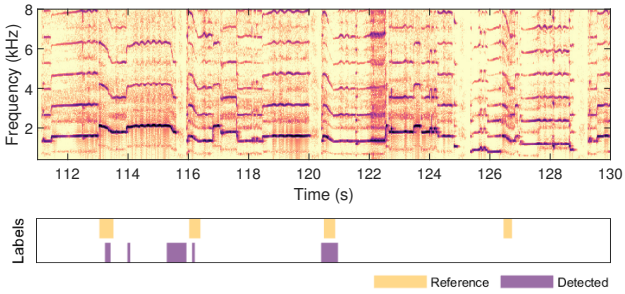

Fig. 4. Portamento detection result for an excerpt in the piece *Busy Delivering Harvest* by Player 8.

4. CONCLUSION

In this paper, we have proposed a recognition system for pitch evolution-based playing techniques (PETs) based on the joint time–frequency scattering transform (jTFST). The characteristics of each type of PET are explicitly encoded into the jTFST, which forms the input to classifiers. For ecological validity, we have created a new dataset with real-world folk music recordings on which to evaluate the system. Frame-based F-measures of 71% for acciacatura, 59% for portamento, and 83% for glissando are obtained, which confirms the feasibility of building a generic model for playing technique recognition. Analysing the results, we find that portamento false negatives are often attributed to playing technique co-articulation.

Future work will further verify the approach on other datasets with playing techniques, such as Studio-Online [6] and ConTimbre [18]. We will also compare the jTFST with other equivalent time–frequency representations, such as the two-dimensional Fourier transform and the modulation spectra [19].

5. REFERENCES

[1] Mikel Gainza and Eugene Coyle, "Automating ornamentation transcription," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2007, vol. 1, pp. I–69.

[2] Marc Leman, Luc Nijs, and Nicola Di Stefano, "On the role of the hand in the expression of music," in *The Hand*, pp. 175–192. Springer, 2017.

[3] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," in *arXiv:1609.03499*, 2016.

[4] Yoonchang Han and Kyogu Lee, "Hierarchical approach to detect common mistakes of beginner flute players," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 77–82.

[5] Glenn Eric Hall, Hassan Ezzaidi, Mohammed Bahoura, and Christophe Volat, "Classification of pizzicato and sustained articulations," in *European Signal Processing Conference (EUSIPCO)*, 2013, pp. 1–5.

[6] Vincent Lostanlen, Joakim Andén, and Mathieu Lagrange, "Extended playing techniques: the next milestone in musical instrument recognition," in *5th International Conference on Digital Libraries for Musicology (DLfM)*, 2018.

[7] Luwei Yang, *Computational Modelling and Analysis of Vibrato and Portamento in Expressive Music Performance*, Ph.D. thesis, Queen Mary University of London, 2017.

[8] Changhong Wang, Emmanouil Benetos, Vincent Lostanlen, and Elaine Chew, "Adaptive Time–frequency Scattering for Periodic Modulation Recognition in Music Signals," in *International Society for Music Information Retrieval Conference (ISMIR)*, Nov 2019.

[9] Changhong Wang, Emmanouil Benetos, Xiaojie Meng, and Elaine Chew, "HMM-based glissando detection for recordings of Chinese bamboo flute," in *Sound and Music Computing Conference (SMC)*, May 2019.

[10] R. Panda, R. Malheiro, and R. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, 2018.

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

[12] Joakim Andén, Vincent Lostanlen, and Stéphane Mallat, "Joint time–frequency scattering," *IEEE Transactions on Signal Processing*, vol. 67, no. 14, pp. 3704–3718, July 2019.

[13] Jordi Pons, Thomas Lidy, and Xavier Serra, "Experimenting with musically motivated convolutional neural networks," in *14th IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2016, pp. 1–6.

[14] Joakim Andén and Stéphane Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, 2014.

[15] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, "The elements of statistical learning: data mining, inference, and prediction," 2009.

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

[17] Matthias Mauch and Simon Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 659–663.

[18] Thomas Hummel, "Algorithmic orchestration with contimbre," *Journées d'Informatique Musicales-JIM2014*, pp. 139–140, 2014.

[19] Etienne Thoret, Philippe Depalle, and Stephen McAdams, "Perceptually salient regions of the modulation power spectrum for musical instrument identification," *Frontiers in psychology*, vol. 8, no. 587, 2017.