IPT	VSB		RW	
	num	seconds	num	seconds
vibrato	192	261.1	42	51.5
UP	488	752.0	48	94.4
DP	333	508.5	51	87.5
RP	272	327.0	94	94.7
glissando	316	595.3	42	95.9
tremolo	205	259.8	23	59.7
harmonic	318	305.8	61	42.0
plucks	204	204.0	135	99.7

Table 1. Quantity of data in the virtual sound banks (VSB) and real-world (RW) recordings. The Guzheng playing technique clips from VSB is used in the training set and that from RW is used in the test set.

- **Returning Portamento (huihuayin回滑音) (*RP* for short thereinafter):** pluck a string, press it with left hand, and then release it, creating an up-down slide (region (d) of Figure 1).

Right-hand playing technique:

- **Glissando (guazou刮奏, huazhi花指):** play several strings consecutively up or down with thumb or index finger rapidly (region (e) of Figure 1).
- **Tremolo (yaozhi摇指):** pluck a single string in rapid succession by "shaking" the index finger or thumb of the right hand back and forth across it as shown in the region (f) of Figure 1.
- **Harmonic (fanyin泛音):** tap a special location of a string with the left hand, and lift the left hand while playing the string with the right hand to produce a corresponding harmonic effect as shown in the region (g) of Figure 1.
- **Plucks (gou勾, da打, mo抹, tuo托...):** as shown in the region (h) of Figure 1, plucks are defined as simply plucking a string with a finger, including gou, da, mo, tuo [2], and so on.

3.2 Data collection and labelling

In order to ensure the diversity of the data, we collected audio clips of single playing techniques (we called these audio clips *short clips* thereinafter) from real-world recordings (RW) and 2 virtual Guzheng sound banks (VSB) which had different players and recording setups: Yellow River Sound and Kong Audio 3 KONTAKT. We sampled 2328 *short clips* covering almost all the tones in the range of Guzheng from the sound banks and annotated them into 8 classes one by one according to the sound bank manuals. Their durations range from about 0.6 to 11.5 seconds, in total 3213.4 seconds (see Table 1). The 496 real-world recordings of *short clips*, in total 625.3 seconds, were recorded and annotated by a professional Guzheng performer.

We use the *short clips* sampled in VSB as the training split and the set of *short clips* recorded in real world as the test split. This division well simulates the test situation in the real world as we usually do not have access to recordings of a testing Guzheng at training time. The dataset with detailed information and demos is available on the website [3] [14].

4. METHODS

As shown in Figure 2, our proposed model consists of two modules: the IPT detector and the onset detector. In the training phase, we train the onset detector and the IPT detector separately. Then, in the fusion phase, we apply decision fusion to the thresholded output of the onset detector and the raw output of the IPT detector to get the final result.

4.1 Input Representation

Since our goal is to generalize to actual solo recordings, we have generated series of audio sequences which simulate such recordings. Firstly, on both training and test, we concatenated our *short clips* in the dataset mentioned in Section3 one after another randomly until the length of the concatenated audio sequence is greater than 12.8 seconds. A check operation was executed to confirm that every generated audio sequence is unique. A 50 milliseconds crossfade was made in each boundary of adjacent *short clips* to ensure the audio sequence sound more realistic. We cut the audio sequence generated from the training set directly to 12.8 seconds long for training and use the unsplit audio sequence whose length is greater than 12.8 seconds generated from the test set for testing.

We also label the start time and duration of the corresponding playing techniques. Then we generate two kinds of labels (the onset lables and the IPT labels) by quantizing the continuous time labels into timestamps in the unit of 0.05-second-long frame.

The input representation is a log mel-spectrogram. Following [10], we use *librosa* [15] to compute log mel-spectrogram with 128 logarithmically-spaced frequency bins, an FFT window of 2048, a hop size of 2205, and a sample rate of 44.1kHz.

4.2 IPT Detector

Guzheng playing technique detection can be regarded as a part of the task of Guzheng music transcription. As most of the transcription task [16, 17], we treat the frame as a basic unit so we need to identify the playing techniques frame by frame. We can treat the task as a semantic segmentation task since there is no overlap of playing techniques in our data and frames in music are similar to pixels in images. Inspired by the success of FCN in semantic segmentation [18], which is also extended to audio processing research [10, 19], we applied FCN in our task.

[2] gou, da, mo, tuo are the names of the Guzheng playing techniques for simply plucking a string using the middle finger, ring finger, index finger, and thumb respectively.

[3] https://ccmusic-database.github.io/en/database/ccm.html#GZTech

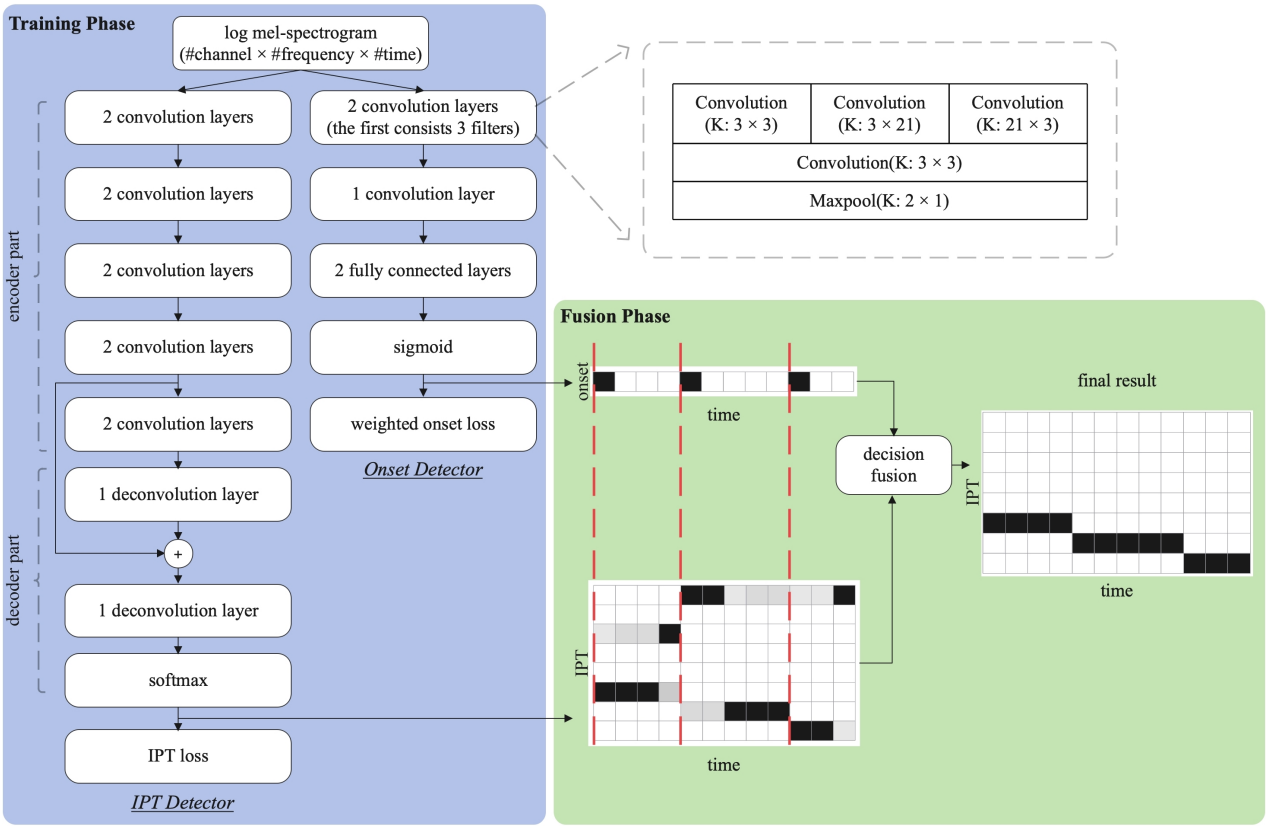

Figure 2. System architecture of our proposed model. In the training phase, the IPT detector and the onset detector were trained separately and then decision fusion was applied to their outputs in the fusion phase to get the final result. The vertical red dotted lines in the "fusion phase" stand for the starts of all onset predictions.

As shown in Figure 2, the IPT detector can be divided into the encoder and the decoder. The encoder part consists of five convolution modules. Two identical convolutional layers were stacked in each module and each of them has square (3×3)[4] filters. At each convolutional layer, zero padding was applied to make the output have the same length as the input and the output of every convolutional layer was then passed through a Rectified Linear Unit (ReLU). To apply our model to variable-length audio, a 1D max-pooling layer which only halve the length of frequency axis was implementd at the output of each convolution module followed by a dropout layer with the probability 0.25 to aid generalization. The output of the encoder part is the input of the decoder part which is made of two deconvolution layers. The decoder part is used for upsampling the feature map to the shape that we need and then make a frame-level prediction to the IPT. A skip connection of element-wise adding was used from the output of the fourth module in the encoder part to the first module output in the decoder part to help the decoder fuse the information.

4.3 Onset Detector

The onset detector is composed of two convolution modules, followed by two fully connected layers. The last layer is a fully connected sigmoid layer with 1 output for representing the probability of an onset for this frame. It has been suggested in recent research [20] that using different musically motivated filter shapes in the first convolutional layer of CNN could improve the model performance. As we discussed in Section 3, different playing techniques in Guzheng vary a lot in the temporal and spectral domain. Thus, we apply three filters of different shape (3×3, 3×21, 21×3) to the first convolutional layer and the 3×21 filter is designed to better capture the feature of long playing technique such as *portamento* or *glissando*. The following two convoliutional layers each has 3×3 filters.

We use weighted binary cross entropy (WBCE) loss as the loss function for our onset detector.

$$L_{onset} = mean(\sum_{t=0}^{T} WBCE(\beta, x_t, y_t)) \quad (1)$$

$$WBCE(\beta, x_t, y_t) = \beta y_t \log x_t + (2-\beta)(1-y_t)\log(1-x_t) \quad (2)$$

where T is the number of frames in the example, y_t is the label that is 1 when there is a ground truth onset at the frame t, x_t is the probability output by the model at frame t and β is the weight factor of the positive samples. Recent models in the onset detection research showed a lower recall than precision [16, 17] and it is largely caused by class imbalance that positive samples for onset are far less than the negative samples. So we decided to increase the weight of the positive samples to balance the performance of the model. We set β to 1.94 by coarse hyper-parameter search.

[4] 2D filters are noted: N frequency bins x M time frames.

model	Frame-level	Note-level		
	accuracy	precision	recall	F1-score
FCN+Onsets	**87.97**	**78.20**	**83.78**	**80.76**
Z.Wang's model [10]	69.44	35.18	47.45	39.53
J.-F. Ducher's model_Reproduced [8]	66.93	19.37	21.31	20.05
B.Liang's model_Reproduced [12]	53.98	41.38	44.50	42.81

Table 2. Frame-level accuracy and note-level precision, recall and F1-score on GZ_IsoTech dataset. Note-based scores calculated by the *mir_eval* library. Final metric is the mean of scores calculated per piece in the test set.

4.4 Fusion of IPT Detector and Onset Detector

In evaluation, the IPT detector and the onset detector are firstly used separately. Then we use a threshold of 0.5 to make the onset output binary. We separate the IPT output into several clips by the binary onset output. Then we selected an IPT as the final result of each clip by "voting" within it as described by Algorithm 1. We set n as the number of IPTs and t as the number of time frames. D_{onset} whose shape is [t] is the binary output of the onset detector. D_{IPT} whose shape is [n, t] is the raw output of the IPT detector, representing the probability of each IPT for each frame. D is the final one-hot result that implies the IPT at every frame. There are two intermediate variables in our algorithm. V whose shape is [n] represents the total "voting" score for each IPT in the current clip and b denotes the frame number where the current clip starts.

Algorithm 1 Decision fusion

Input: the number of IPTs n, the number of time frames t, the thresholded output of the onset detector D_{onset}, the output of the IPT detector D_{IPT}

Output: final result D

$b \leftarrow 0$
$D \leftarrow zeros(D_{IPT})$
$V[0, \ldots, n-1] \leftarrow 0$
for $i \leftarrow 0$ **to** $t-1$ **do**
  for $j \leftarrow 0$ **to** $n-1$ **do**
    $V[j] \leftarrow V[j] + D_{IPT}[j][i]$
  end for
  if $i+1 = t$ or $D_{onset}[i+1] = 1$ **then**
    $D[\text{argmax}(V)][b, \ldots, i] \leftarrow 1$
    $V[0, \ldots, n-1] \leftarrow 0$
    $b \leftarrow i+1$
  end if
end for
return D

5. EXPERIMENTS

In this section, we introduce the evaluation metrics we used, the details of the experiment and the result analyses.

5.1 Evaluation Metrics

The metrics used to evaluate a model consist of frame-level and note-level metrics. Note-level metrics include precision, recall, and F1 scores. They are calculated by the *mir_eval* library [21]. We require that both the IPT type and the onset are correct. In detail, an onset estimation is considered correct if it is within ± 0.05 seconds of the ground-truth. Frame-level accuracy is calculated by comparing the output of our model to the ground truth label frame by frame. More specifically, we calculated the ratio of the number of correct frames to the total number of frames as frame accuracy. Both note and frame scores are calculated every audio sequence and we calculated the mean of these scores as the final metric.

5.2 Results

To examine whether our proposed architecture (namely `FCN+Onsets`) can better detect IPTs in audio with continuous IPTs, we reproduced J.-F. Ducher's model [8], B.Liang's model [12] and Z.Wang's model [10] for comparison, while adapting their capacity to our data.

Because Z.Wang's model [10] cannot be tested on variable-length audio, we cut all the audio from the test set to 12.8 seconds long for the test of Z.Wang's model [10] while other models are still tested on variable-length data.

Table 2 shows the frame-level and note-level results of Guzheng playing technique detection on the GZ_IsoTech dataset. Our `FCN+Onsets` model reaches 87.97% in frame-level accuracy and 80.76% in note-level F1-score, producing the best scores in both frame-level and note-level metrics. Besides, our proposed model outperforms other models by a large margin in all metrics and results in over a 100% relative improvement in note-level F1-score compared to other models. This is mainly because our model takes better advantage of the note onset detection results which is highly interrelated to the note-level metrics. Because the test set consists of audio of real-world recordings that is different from training set in performer, instrument and recording environment, the results show that our model offers good generalization capabilities.

The visualization of the input spectrogram, the thresholded onset output, the raw IPT output, the final fusion result and the label for a recording from the test set are shown in Figure 3. Through the images in Figure 3, we can clearly see the importance of fusing IPT activations with onset predictions. The vertical red lines in the third and fourth images are the onset predictions. We can separate the audio into several notes using the onset predictions. Each note in our dataset only corresponds to one playing technique, but there may be several IPT predictions in one note as

model	Frame-level	Note-level		
	accuracy	precision	recall	F1-score
FCN+Onsets	**87.97**	78.20	**83.78**	**80.76**
CNN+Onsets	60.80	45.01	48.30	46.51
No Onset Fusion	76.48	27.74	50.55	35.47
Onset3×3	87.70	72.67	82.76	77.19
No Skip Connection	82.45	71.05	76.15	73.39
No Weighted Loss	86.26	**80.88**	79.46	80.03

Table 3. Ablation studies with frame-level accuracy and note-level precision, recall and F1-score on GZ_IsoTech dataset.

shown in the "Raw IPT Output" image. So we implement the decision fusion method to choose one IPT as the final output of a note as illustrated in Algorithm 1. The IPT results after being restricted by the onset are presented in the second-to-bottom image ("Final Fusion Result"). The "Target" image is the visualization of the truth label. As shown by Figure 3, the accuracy of the predictions in the "Final Fusion Result" image is far exceed that in the "Raw IPT Output" image.

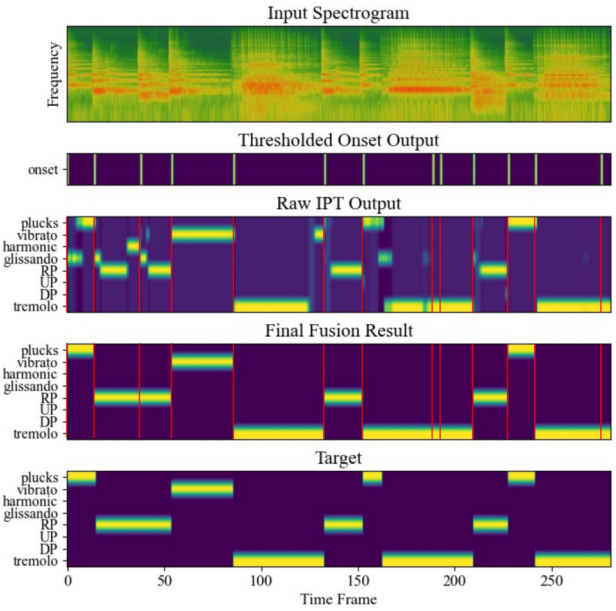

Figure 3. Inference on an audio in the test set. The log mel-spectrogram input, the thresholded onset output, the raw IPT output, the final fusion result and the target are shown from top to bottom. The vertical red lines in the third and fourth images are the onset predictions.

5.3 Ablation Studies

We conducted an ablation study where the performance of `FCN+Onsets` is compared with the following five models (namely `CNN+Onsets`, `No Onset Fusion`, `Onset3×3`, `No Skip Connection`, `No Weighted Loss`). In `CNN+Onsets`, we used the same CNN architecture for the IPT detector as presented in the onset detector instead of FCN. The result of `No Onset Fusion` is the output of the IPT detector without the decision fusion with the onset detector. In `Onset3×3`, we replaced the whole first convolutional layer of the onset detector by 3×3 filters. In `No Skip Connection`, we remove the skip connection part of the `FCN+Onsets` model. `No Weighted Loss` is the model that uses ordinary binary cross entropy (BCE) Loss in the onset detector.

The results show the importance of the onset information - `No Onset Fusion` results in a significant 11.49% decrease in the frame-level accuracy and a 45.29% decrease in the note-level F1-score. Our model can locate the IPT boundaries precisely and rectify some mispredicted frames for IPTs through fusing the IPT results with the note onset information. `CNN+Onsets` results in decreases in all metrics. It shows the FCN can better distinguish different IPTs than CNN. `No Skip Connection` shows a 5.52% decrease in frame-level accuracy and a 7.37% decrease in note-level F1-score. This proves the effectiveness of our skip-connection operation. `No Weighted Loss` has a little rise in note precision but decrease in all other metrics that shows weighted loss can better balance the precision and recall.

6. CONCLUSION

In this paper, we create a new dataset consisting of abundant Guzheng playing technique clips from multiple sound banks and real-world recordings. We propose an end-to-end Guzheng playing technique detection system using FCN that can be tested on variable-length audio. We propose a new decision fusion method where an onset detector is trained to divide an audio into several notes and the predicted IPTs are added frame by frame inside each note during fusion. The final output of each note is the IPT class with the highest probability within the note. Our method outperforms existing works by a large margin, which proves that our model is effective and offers good generalization capabilities that transfer well between different training and test sets. Moreover, by visualising the results, we find that the onset information is crucial for the Guzheng playing technique detection task.

In different genres of Guzheng music, there are slight differences in the same playing technique. In the future, we will continue to expand our dataset and add real-world Guzheng music pieces from different genres. Moreover, Guzheng is polyphonic and even one note may have more than one playing techniques. We will investigate how to detect more complicated playing techniques.

7. ACKNOWLEDGEMENT

This work was supported by National Key R&D Program of China (2019YFC1711800), NSFC (62171138). Wei Li, Fan Xia and Yi Yu are corresponding authors of this paper.

8. REFERENCES

[1] S. Zhuo, *The Chinese zheng zither: contemporary transformations*. Routledge, 2016.

[2] X. Zhao, "From 'complementing tone with yun' to 'expressing with sound': On the changes of left-hand playing techniques in zheng playing and the thinking aroused from them (In Chinese)," *Yellow River of the Song*, no. 19, pp. 33–37, 2009.

[3] V. Lostanlen, J. Andén, and M. Lagrange, "Extended playing techniques: the next milestone in musical instrument recognition," in *Proceedings of the 5th International Conference on Digital Libraries for Musicology*, 2018, pp. 1–10.

[4] L. Su, L.-F. Yu, and Y.-H. Yang, "Sparse cepstral, phase codes for guitar playing technique classification." in *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*, 2014, pp. 9–14.

[5] 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 *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR*, 2015, pp. 708–714.

[6] T.-W. Su, Y.-P. Chen, L. Su, and Y.-H. Yang, "Tent: Technique-embedded note tracking for real-world guitar solo recordings," *Transactions of the International Society for Music Information Retrieval*, vol. 2, no. 1, pp. 15–28, 2019.

[7] A. Livshin and X. Rodex, "The importance of cross database evaluation in sound classification," in *Proceedings of the 4th International Society for Music Information Retrieval Conference, ISMIR*, 2003, pp. 1–1.

[8] J.-F. Ducher and P. Esling, "Folded cqt rcnn for real-time recognition of instrument playing techniques," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR*, 2019, pp. 708–714.

[9] Y.-F. Huang, J.-I. Liang, I.-C. Wei, and L. Su, "Joint analysis of mode and playing technique in guqin performance with machine learning," in *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR*, 2020, pp. 85–92.

[10] Z. Wang, J. Li, X. Chen, Z. Li, S. Zhang, B. Han, and D. Yang, "Musical instrument playing technique detection based on fcn: Using chinese bowed-stringed instrument as an example," *arXiv preprint arXiv:1910.09021*, 2019.

[11] J. Abeßer, H. Lukashevich, and G. Schuller, "Feature-based extraction of plucking and expression styles of the electric bass guitar," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*. IEEE, 2010, pp. 2290–2293.

[12] B. Liang, G. Fazekas, and M. Sandler, "Piano sustain-pedal detection using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*. IEEE, 2019, pp. 241–245.

[13] J. Qiu, "On the classification and evolution of zheng playing techniques (In Chinese)," *Chinese Music*, no. 4, pp. 215–224, 2004.

[14] Z. Li, S. Yu, C. Xiao, Y. Geng, W. Qian, Y. Gao, and W. Li, "Ccmusic database: Construction of chinese music database for mir reaserch (In Chinese)," *Journal of Fudan University (Natural Science)*, vol. 58, no. 3, pp. 351–357, 2019.

[15] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.

[16] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR*, 2018, pp. 50–57.

[17] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, "High-resolution piano transcription with pedals by regressing onset and offset times," *IEEE/ACM Transactions on Audio, Speech, and Language Processing, TASLP*, vol. 29, pp. 3707–3717, 2021.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition, CVPR*, 2015, pp. 3431–3440.

[19] L. Ardaillon and A. Roebel, "Fully-convolutional network for pitch estimation of speech signals," in *20th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 2005–2009.

[20] J. Pons and X. Serra, "Designing efficient architectures for modeling temporal features with convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*. IEEE, 2017, pp. 2472–2476.

[21] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. Ellis, "mir_eval: A transparent implementation of common mir metrics," in *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*, 2014, pp. 367–372.

KDC: AN OPEN CORPUS FOR COMPUTATIONAL RESEARCH OF *DASTGĀHI* MUSIC

Babak Nikzat
Kunstuniversität Graz
b.nikzat@kug.ac.at

Rafael Caro Repetto
Kunstuniversität Graz
rafael.caro-repetto@kug.ac.at

ABSTRACT

Iranian *dastgāhi* music is considered as the classical repertory of contemporary Iran. In the 19th century, the melodic modes that developed during its long history were grouped in categories, each of them known as *dastgāh*. The *dastgāhi* system presents unique features, that have been object of musicological study since its inception. However, computational methods for its research are still scarce, due in good part to the lack of open, well curated corpora. The aim of the KUG *Dastgāhi* Corpus (KDC) is to contribute to the development of computational corpus driven research for this tradition. KDC is created following the FAIR principles, and in close collaboration with performers and scholars, who contribute to it with annotations and qualitative evaluations. Besides presenting the first version of KDC, in this paper we explore the possibilities that Iranian *dastgāhi* music offers to computational research. In order to test the performance of state-of-the-art technologies applied to this music tradition, we present preliminary results for several analytical tasks, and discuss their opportunities and limitations learnt in the process.

1. INTRODUCTION

The potential of computational methods for supporting musicological research tasks is increasingly being recognized and adopted by musicologists. Due to the high degree of specialization required both from the technical and musicological sides of this kind of research, collaboration between specialists from both disciplines is one of the most fruitful approaches for computational musicology. Developed at the Institute for Ethnomusicology of the Kunstuniversität Graz (KUG), the KUG *Dastgāhi* Corpus (KDC) aims to offer a growing collection of well curated and annotated data of Iranian *dastgāhi* music around which engineers, musicologists and researchers from other disciplines interested in this music tradition can collaborate. In this first section we describe the main features of the *dastgāhi* musical system and offer an overview of the state-of-the-art of its computational research. Then, we introduce and describe in detail the first version of KDC. In section 3 we discuss the potential of Iranian *dastgāhi* music for computational research. In the following sections, preliminary analyses carried out to test such potential are presented and discussed. The paper closes with our plan for expanding KDC and future work.

1.1 Iranian *dastgāhi* music

Iranian *dastgāhi* music is modal. The basic modal entities in this tradition are known as *gušе*s and they are defined by a model melody built on a specific intervallic structure, a series of characteristic melodic patterns, and a modal centre known as *šāhed* [1]. Related *gušе*s are grouped in specific collections, and in a specific order. These collections are known either as *dastgāh* or *āvāz*, each of them with its own specific name. A *dastgāh* is larger than an *āvāz* in terms of number of *gušе*s, and an *āvāz* is considered to be a subcategory of a specific *dastgāh*. Nowadays, the canonical repertoire consists of 7 *dastgāh*s and 5 *āvāz*es (see Table 1). For simplicity, in this paper we are using *dastgāh* to refer to both of them collectively. The *gušе*s in a particular *dastgāh* are organized in a fixed order, and even though the performance of a *dastgāh* does not required all its *gušе*s to be performed, the order has to be maintained. The first *guše* in all *dastgāh*s is called *darāmad*, and it has a very significant meaning since it introduces the *dastgāh*. Most performances start and end with *darāmad*.

Along the history of Iranian *dastgāhi* music, great masters (*ostād*s) of this tradition defined their own canonical repertory of the 7 *dastgāh*s and 5 *āvāz*es, fixing the number of *gušе*s in each of them, their particular order, their specific melodic contour and their rhythmic rendition. These canonical repertoires are known as *radif*s, always associated with their corresponding *ostād*. These fixed *radif*s are nowadays only used for learning the tradition. Contemporary musicians perform their own interpretation of the repertory, but through the rules they experientially extracted from the *radif* they learnt. As a more encompassing concept, different performance schools are recognized, each of them with idiosyncratic performance styles. The most extended ones today are Tabriz, Tehran and Isfahan.

Iranian *dastgāhi* music uses a microtonal tuning system, resulting in specific intervallic structures defined for each *guše*. Three basic interval categories are considered, namely tone, semitone and an intermediate one between the previous two, whose specific width is not fixed but flexible [1], depending on *guše* or even performer, and that is the result of altering one natural tone less than a semitone

up (*sori*) or down (*koron*). These intervallic structures form specific scales, in which particular tones receive predominant roles, such as *šāhed* and *ist*, which are fundamental for defining the modal features of a *guše*.

Specific techniques for vocal and instrumental performance, including an extensive repertory of ornamentation, is another essential aspect of this music, showing a great diversity depending on school, instrument, and personal style. The most common instrumental ornaments include: *ešāreh*, *tekiyeh*, *qalt*, *dorrāb*, *riz*, and *šalāl*. *Tahrir* is the most significant singing ornament, which requires a specific technique in which the singer switches between chest and head voice (see Fig. 5).

*Guše*s can be performed in free rhythm or using rhythmic cycles. Even though this is fixed per each *guše* in the *radif*s, it is decided by each musician in contemporary performance. When a *guše* is sung in free rhythm, the prosodic rules, known as *aruz*, of Persian classical poetry, to which most of the sung lyrics in Iranian *dastgāhi* music belong, informs the rhythmic rendition of the singing. This results in a characteristic rhythmic style that is also adopted when a *guše* is perfomed only instrumentally [2, 3]. *Aruz* not only informs musical rhythm through the sequence of long and short syllables, but also through the qualitative features of the vowels (*hejā*) [4, 5]. For metred performance, musicians can draw on rhythmic cycles known as *adwār-e iqāi*, the most common ones being *kerešme*, *čāhār-pāreh*, *zanguleh* and *gereyli* [4]. Each *dowr-e iqāi* is defined by a number of pulses, known as *naqareh*, and a structure of accents.

A typical ensemble of Iranian *dastgāhi* music is formed by *santur* (hammered zither), *tār* (long-neck lute), *setār* (long-neck lute), *ney* (bamboo flute), *kamanče* (spike fiddle), *tonbak* (goblet drum) and a singer. The music, however, is also commonly performed by a solo singer or instrumentalist.

Dastgāhi music is an oral tradition and conventionally the students learn it by memorizing a specific *radif* taught in private lessons by an *ostād*. The learning process is based on observation and imitation. In the contemporary context some teachers use transcriptions in staff notation as teaching material. However the notation functions just as memory aid, since many details cannot be represented in the staff notation system.

1.2 Computational analysis of Iranian *dastgāhi* music

Considering the magnitude of Iranian *dastgāhi* music, in terms of history, wide performance realm, complex theorization and extensive musicological research, this music tradition has received comparatively little attention from computational approaches. Pioneering work has been carried out by the researcher and *santur* performer Peyman Heydarian. Since his master thesis [6], Heydarian is working on the computational analysis of this instrument, especially its acoustic features [7–9], as well as on *dastgāhi* music performed in this instrument, with a special focus on *dastgāh* classification [10–12]. In fact, this second task, automatic *dastgāh* identification and classification, has been the most common one by other researchers who approached this tradition from computational methods [13–16].

Computationally aided musicological research on *dastgāhi* music is still very rare. Shafiei used score to audio alignment for the analysis of *tahrir* [17], while Sanati analysed intervals from *ney* recordings [18]. There has also been some work on *dastgāhi* music generation using deep neural networks [19].

We argue that one of the reasons that explains that the computational analysis of *dastgāhi* music is not yet been fully established in an unified trend, instead of occasional isolated works, is the lack of a comprehensive, publicly accessible datasets upon which continuous research can be built. All the publications referenced in the previous paragraphs gathered their own research datasets, but none of them offer any indication about possibilities for accessing them. Only Heydarian has been able to sustain a decades long research on this tradition based on his own dataset [20], but it is not public, and it only includes recordings of *santur*. Hence, the goal of KDC precisely is to contribute to the establishment of a coherent line of computational research on Iranian *dastgāhi* music by providing an open, well curated corpus of high quality recordings, upon which this research can be built.

2. DESCRIPTION OF KDC

KDC is conceived as an ever growing corpus, so in this paper we present its first version. Although far from its ideal size in terms of coverage, it already offers a coherent and comprehensive representation of the *dastgāhi system*. Currently, KCD contains 213 solo recordings by four professional musicians, 65 of which consist of complete *guše* performances and hence form the core of the corpus.

In order to analyse the corpus, we draw on the five criteria defined by Serra [21]:

Purpose: As described in section 1, KDC is created for the analysis of Iranian *dastgāhi* music using corpus driven computational methods. Due to the lack of a notational system originally devised for this music tradition, and the difficulties of staff notation for capturing key elements of this music, especially regarding tuning, intonation and ornamentation, KDC primarily aims at gathering audio recordings. However, the inclusion of other data types, such as scores, lyrics, videos, images, etc., that might contribute to the general purpose of the corpus, is not excluded. A key point in the creation of KDC is the close collaboration with musicians. They do not only contribute with recordings and annotations, but also with their knowledge, evaluation, and research ideas. Musicians are thus considered as active collaborators of KDC.

Coverage: According with KDC's purpose, we focus on voice and melodic instruments. In this first version of the corpus, being one of our goals to test state-of-the-art technologies for the analysis of this music tradition, we have focused on solo performances, and consciously excluded ensemble performances, that would present a greater challenge for computational analysis. In order to