EXPLORING DÍZI PERFORMANCE PARAMETERS WITH MACHINE LEARNING

Kai Ren Teo
Kairen_teo@mymail.sutd.edu.sg
Singapore University of Technology & Design

Balamurali B T
balamurali_bt@sutd.edu.sg
Singapore University of Technology & Design

Teck Seng Ng
tonexplore@gmail.com
Asian Music Technology (Singapore)

Jer-Ming Chen
Jerming_chen@sutd.edu.sg
Singapore University of Technology & Design

ABSTRACT

In musical performance, three questions are often asked: what is the contribution of the instrument? What is the contribution of the player? Which musical exercise was performed? Here, we attempt to objectively quantify and compare the effects of the player, instrument and performed exercise by analyzing audio samples collected from a series of musical exercises performed on the traditional and modernized Chinese transverse flute, Dízi, for 5 expert players. Feature extraction is employed on the samples and a machine learning classifier algorithm is applied to the same dataset to confidently identify distinct populations: firstly, separating traditional vs modernized Dízi, secondly, separating player identity, and finally identifying the musical exercise performed all based purely on the Dízi performance's acoustic output alone.

1. INTRODUCTION

The Chinese transverse flute, Dízi (笛子; pronounced [tǐt͡si]) has a long history in traditional Chinese classical and folk music since it was first documented in the Han Dynasty (206 BC-220 AD), playing essential musical roles both as solo instrument and also in chamber music. Despite its pedigree (or perhaps because of it), there has been minimal innovation in its design relating particularly to its acoustic and playing response, but in the last two decades, a Singaporean master Dízi-maker (Teck Seng Ng, one of co-authors of this paper) has systematically modernized the Dízi, informed by acoustic and performance considerations (See Figure 1) (Balamurali BT 2018).

These modern instruments maintain the characteristic shape, timbre and fingering of the traditional Dízi (see Figure 1) but are described by players as having improved 'control', 'uniformity', 'responsiveness' and 'playability', when compared to standard (traditional) Dízi. Consequently, these modern Dízi are now highly sought-after by professional players internationally. Of course, many of these perceptual qualities are subjective and its salience may vary from player to player.

Therefore, in this exploratory study, we are interested in objectively quantifying and comparing the playability response of the standard (traditional) Dízi and Ng's (modern) Dízi to test if there is anything qualitatively different or identifiable between them. To investigate, we asked players to perform a series of musical exercises on both sets of instruments, performed feature extraction on the audio samples collected and applied a machine learning classifier algorithm to determine if (a) we could identify two distinct populations of Dízi simply by comparing the Dízi's acoustic output when played by expert players. Further, using the same acoustic dataset we investigate whether it can be (b) possible to distinguish between players performing the exercise and (c) the exercise itself. We note (to our knowledge) there is no similar previous work applying machine learning classification to woodwind instrument performance, particularly in these classification categories.

The remainder of this paper is organized as follows. Details about data collection can be found in Section 2. Experimental methodology is described in Section 3. This section overviews extracted features and the classification process. Section 4 contains results produced as part of this investigation and finally conclusions in Section 5.

Figure 1. Modern (top) and standard (bottom) Dízi in the key of C, referenced alongside the embouchure hole (first hole on the left). Note the superficial similarities of hole placements and external dimensions.

2. DATA COLLECTION

5 expert Dízi players were tasked to perform musical exercises on three Dízi sizes (*): 'Low G', 'C', 'High G' for both standard (traditional) and modern (Ng) Dízi. All 6 instruments used are professional/concert-grade, and tuned to a reference pitch of A4 = 442 Hz. (*Note: the Dízi is a transposing instrument. Consequently, "*movable doh*" solfège is used here to refer to note name-fingerings on the Dízi, because this is how they are referred to in Dízi practice and tradition.)

The seven musical exercises performed were:

1. Diatonic scale from the lowest to the highest note possible, performed at 4 dynamic levels:
    1.1. constant *pp*, soft volume
    1.2. constant *mf*, moderate volume
    1.3. constant *ff*, loud volume
    1.4. *Messa di Voce*, increasing volume gradually from *pp* to *ff* back to pp while sustaining a constant pitch over 6-8 seconds with no vibrato and a stable embouchure
2. Overblowing: 'overblow' while holding the first four fingerings (xxx xxx; xxx xxo; xxx xoo; xxx ooo, nominally "*sol*", "*la*", "*ti*" and "*do*") and sound the first 4-5 overtones produced, held for 2-3 seconds.
3. Pitch bending: bend the sounded notes when playing the "*sol*", "*do*" and "*mi*" fingerings, as low as possible, and as high as possible
4. Octave break: play smoothly (*legato*) over the octave break: e.g. *sol-sol'-sol*; *sol'-sol-sol'* (where the apostrophe indicates the upper-octave note) for the fingerings producing "*sol*", "*do*" and "*mi*".

The audio recordings (near-field and far-field) were made in a room specially treated for excellent noise isolation (<20 dB re 20 μPa) and low reverberation time (<0.23 seconds), with condenser microphones (Rode NT3) placed level at 0.5 m, 1.2 m and 2.0 m away from the player's mouth, along the player's main forward axis. Each microphone signal was rendered into mono PCM WAV (44,100 Hz, 16 bits) channels and pyAudio library was used to extract audio features for analysis.

3. EXERIMENTAL METHODOLOGY

Figure 2 shows the experimental methodology. A classifier system using ensemble learning techniques was trained to classify the Dízi-type, player identity and the musical exercise performed. Ensemble learning perturbs-and-combines a number of machine learning techniques together. Random forest, one of the powerful ensemble learning algorithms, was trained and tested as part of this investigation. The chosen classifier for this investigation contains a multitude of decision trees and has been trained using a variety of audio features (Giannakopoulos and Pikrakis 2014).

The features were labelled with the instrument (traditional/modern) type used, player identity and exercise played. In the first investigation, a trained classifier was used to predict if a given audio sample was produced by the traditional or modern Dízi. In the second investigation, a newly trained classifier was used to determine which player performed the audio sample in question. And in the third investigation, a classifier was trained to identify the musical exercise performed

3.1. Audio Feature Extraction

Dízi audio recordings were firstly divided into frames of 50 ms size and for every 50 ms audio frame a number of 'short term' features (i.e., features extracted from a short audio frame) were extracted. A 50% frame overlap was also incorporated. These features include mel-frequency cepstral coefficients (MFCCs), chroma vectors, zero crossing rate (ZCR), energy, energy entropy, spectral entropy, spectral flux, spectral roll off, spectral spread, spectral centroid and chroma deviation (Giannakopoulos and Pikrakis 2014, Giannakopoulos 2016). Altogether, 34 features were extracted for every audio frame. Data normalization was performed on the extracted features to ensure the contributions of each feature is proportional and this further avoid bias such as recording volume between recording sessions. Short explanation of these features can be found in the following section.

3.1.1. Mel-frequency cepstral coefficients (MFCCs)

MFCCs focus on the perceptually relevant aspects of the audio spectrum and are arguably the most commonly used audio features in the speech/speaker recognition arena (Muda, Begam, and Elamvazuthi 2010). Human perception of musical pitch is logarithmic in nature, and is therefore most sensitive to low and mid-range frequency sounds but loses its ability to distinguish adjacent high frequency sound. MFCCs imitate this perceptual characteristic by estimating energy in various region of audio spectrum over a set of overlapped non-linear mel-filter bank. Mel-banks are narrow at low frequency and get wider at higher frequencies. In short, the whole extraction process of MFCCs can be summarized as: cosine transform of log power audio spectrum on a non-linear mel frequency scale (Rabiner and Schafer 2011, Rabiner and Juang 1993). In this investigation, we have used the first 13 MFCCs.

3.1.2. Chroma Vectors

Chroma vectors are one of the widely used features in music-related application and are often used to capture

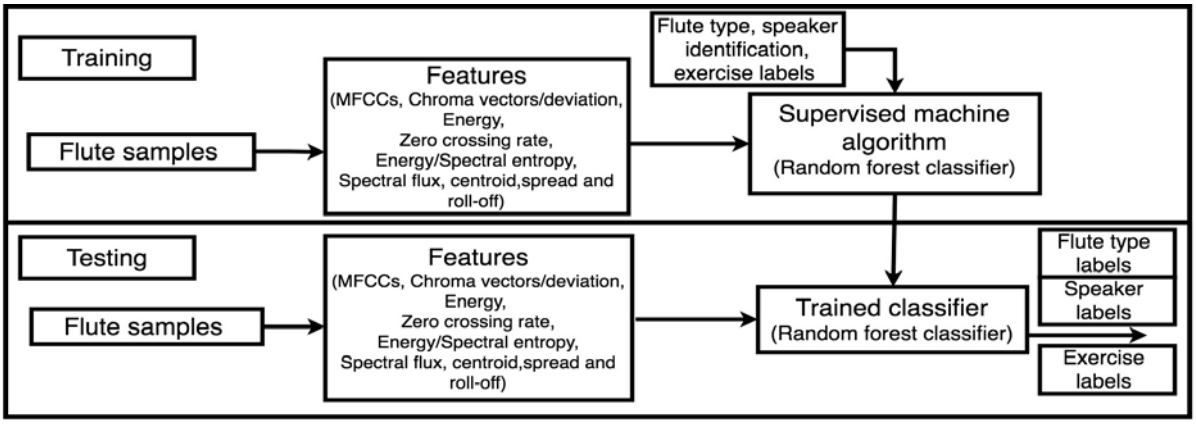

Figure 2 Experimental methodology

harmonic and melodic characteristics (Peeters 2006). They are related to twelve equal-tempered pitch classes of western music. For extracting this feature, firstly, the discrete Fourier transform (DFT) coefficients calculated for a short term windowed signal is grouped into twelve bins. Now, chroma feature for a particular bin is estimated by taking the average of the respective DFT coefficients in that specific bin. In this investigation, we have used chroma vector of 12 elements long. Along with this chroma vector, chroma deviation was also extracted. Chroma deviation represents standard deviation of 12 chroma coefficients.

3.1.3. Average energy

The average energy of a short frame sequence $x_i(n),\ n = 1, 2, \ldots, W_l$ is calculated as

$$Average\ Energy = \frac{1}{W_l}\sum_{n=1}^{W_l}|x_i(n)|^2 \qquad (1)$$

where W_l is the number of samples in a given frame (i.e., frame length). Presence of audio from silence can be identified using this short-term feature.

3.1.4. Zero crossing rate (ZCR)

The rate at which sign-changes occur in an audio signal is given by ZCR. It is a measure of frequency content of a signal. DC offset in the signal, if exists, should be removed prior to determining ZCR.

3.1.5. Energy Entropy

Energy entropy can be interpreted as a measure of abrupt change in audio signal energy. In order to calculate energy entropy, an audio frame is subdivided into *K* sub-frames of fixed duration. Then energy content of the audio frame and every sub-frame are calculated by the procedure mentioned in Equation (1). Now, entropy is given by

$$Energy\ Entropy = \left(-\sum_{j=1}^{K} e_j.log_2(e_j)\right) \qquad (2)$$

where e_j is the ratio of energy of the j^{th} sub-frame to total energy of the audio frame. The entropy value will be low if there are sudden changes in the audio signal amplitude and vice-versa (Giannakopoulos et al. 2006).

3.1.6. Spectral flux

The spectral change between two successive frames is measured as spectral flux.

3.1.7. Spectral roll-off

Spectral roll-off measures the frequency below which 90% of the spectral magnitude is concentrated.

3.1.8. Spectral centroid and Spectral spread

The center of gravity and the second central moment of spectrum constitute spectral centroid and spread respectively.

3.1.9. Spectral entropy

Spectral entropy is calculated in a similar way to energy entropy. However, here spectrum of a short-term audio is used instead of the original sample values. The spectrum of a particular audio frame is sub-divided into *K* bins (sub-bands) and energy of every sub-band is estimated. This sub-band energy (for e.g. j^{th} sub-band) is then normalized by dividing the total spectral energy of the frame to produce n_j (Shen, Hung, and Lee 1998, Giannakopoulos and Pikrakis 2014). Finally, spectral entropy is calculated as in Equation (3).

$$Spectral\ Entropy = \left(-\sum_{j=1}^{K} n_j.log_2(n_j)\right) \qquad (3)$$

3.2. Random forest classifier

Random forest uses a multitude of decision trees. Based on the significance of input feature values, the decision trees split the given samples into many homogenous sets. The top-most node in a decision tree contains samples from the entire population and is highly non-

homogenous. However, homogeneity of samples in sub-nodes increases with tree splitting. In the case of random forest, decisions from many trees are considered as opposed to single decision tree. The final prediction will be that class label which most of the trees voted for (Seni and Elder 2010, Liaw and Wiener 2002). In this investigation, we used Scikit-learn random forest implementation (Pedregosa et al. 2011). Ten decision tree estimators were considered for the random forest in this investigation.

3.3. Training vs Testing Data

In order to separate traditional vs modernized Dízi-type (first investigation), musical exercises performed by players are identified from the entire recordings. Each exercise was then grouped to traditional and modernized Dízi-type. A similar procedure has been followed to investigate if an audio sample was performed by a specific player (second investigation) and for a particular exercise (third investigation). However, in this investigation, once the musical exercises performed by players were identified, these exercises were grouped to that of specific players who created them. In the third investigation, the musical exercises performed replaced the player groups to identify a particular exercise performed from the audio sample. Four out of five players were considered for this second investigation. One of the players was left out because the corresponding raw recording contains a lot of extraneous artefacts and thus required a lot of effort in labelling a particular exercise.

The available audio data for every investigation is split into two sets, a training set and a testing set. Very often, the proportion chosen for splitting is 70% for the training set and 30% for the test set and we have followed the same protocol in this investigation (This split has resulted in approximately 17500 samples for training and 6500 samples for testing the models for most of the investigations). The classification model is trained using the former and its performance is evaluated using the latter. In order to evaluate the realistic performance of the trained model, the test set chosen has never been seen by the model during training. Therefore, the resulting performance can be considered a good guide to what can be expected when the model is applied to unseen data. However, performance may be affected if the testing recordings happen to arise from different recording conditions or from non-contemporaneous sessions. This latter assumption has not been tested in this investigation.

4. RESULTS

4.1. Separating traditional vs modernized Dízi

From the investigation, it is clear that the chosen classifier was able to predict if an audio sample was produced by a traditional or modern Dízi to a high degree of accuracy (Table 1), ranging from 81-93%. The model was able to predict the type of Dízi used with varying degrees of accuracy based on the exercises performed, achieving the highest accuracy (≥90%) with all four scale exercises. Nevertheless, the performance of other exercises was neither that inferior (≥81%).

The relative importance of each feature attribute was also computed as part of this investigation. These 'importance values' can be used to enable a feature selection process and give some idea about why some exercises perform better than the rest. The top four features contributing to Dízi-type separation are listed in Table 2, with Energy being the most important. This is very much intuitive given the fact the scale exercise, especially *ff,* is the best among the exercises that separate a traditional Dízi from a modern one.

Exercise	Accuracy (%)
Scale *pp*	90
Scale *mf*	91
Scale *ff*	93
Scale *mdv*	90
Overblow	81
Pitch Bend	87
Octave Break	86

Table 1. Accuracy of Dízi-type prediction by exercise.

Top four features for Dízi-type prediction	
1	Energy
2	Chroma Deviation
3	Zero Crossing Rate
4	Spectral Entropy

Table 2. Top 4 features for Dízi-type prediction.

Exercise	Accuracy (%)
Scale *pp*	94
Scale *mf*	93
Scale *ff*	93
Scale *mdv*	94
Overblow	94
Pitch Bend	94
Octave Break	90

Table 3. Accuracy of player prediction by exercise.

4.2. Separating player identity

The chosen classifier (newly trained) was able to predict which player generated the audio sample in question with a high degree of accuracy across the different exercises (Table 3) at ≥ 90% in all cases, with the majority at the 93-94% mark.

The high accuracy might be related to the fact that the number of players considered for this investigation was relatively small. The players were asked to perform using two different types of Dízi (3 traditional; 3 modern) and this eventually models their idiosyncrasies while performing various exercises. Such idiosyncrasy vividly explains 'intra-player' and 'inter-player' variations (i.e., variation within a player and variation between players) that eventually contributed toward player identification.

The most significant features contributing to player prediction are listed in Table 4, with entropy of energy being the most important here. Since the energy entropy value directly reflects the sudden changes in the audio signal amplitude, one would expect its importance in separating the players is related to how a particular player performs transitions from one part of the play to another, while performing a particular exercise.

Another important result, one can infer from this investigation, is that the top 4 features (see Table 2 and 4) for both the separation problem are somewhat same. This means that the features that contain cues about the played instruments further contain cues about player-specific information.

Top four features for player prediction	
1	Energy Entropy
2	Energy
3	Zero Crossing Rate
4	Chroma Deviation

Table 4. Top 4 features for player prediction.

4.3. Separating exercises performed

Using the chosen classifier, the seven musical exercises performed were also identifiable from the audio sample with a high degree of accuracy across the two types of Dízi (see Table 5) at ≥ 80%. An intuitive observation from this result is that, the accuracy of the exercise prediction is more or less the same irrespective of the chosen Dízi-type, indicating that they are both of comparable playability.

Type of Dízi	Accuracy (%)
Traditional Dízi	82
Modernized Dízi	83

Table 5. Accuracy of exercise prediction by Dízi-type.

Several exercises had similar transitions between notes sounded and were mostly distinguished by their volume and spectral characteristics (there is typically less variation within-exercise but large variation between-exercise for volume and spectrum). Having said that, for exercises such as *Messa di Voce* there would necessarily be variation in volume within-exercise, however, energy or volume envelope here has a distinct signature. Since the values for each feature in the audio samples were normalized, signal volume ranges observed for each exercise would then solely be related to the transitions from one part of the exercise to another and is thus significant in characterizing the exercise performed. Looking at the top four features that contributed to exercise prediction (as listed in Table 6), our intuition was indeed right.

Top four features for exercise prediction	
1	Energy
2	Zero Crossing Rate
3	Spectral Roll-off
4	Chroma Deviation

Table 6. Top 4 features for exercise prediction.

5. CONCLUSIONS

As is reported in Section 4, the machine learning algorithm used and approach outlined in Section 3 is effective in both instrument/player separation as well as in identifying the musical exercise performed. This acknowledges the fact that both player and instruments play significant and differentiating roles in musical performance – the contributions of both player and instrument cannot be ignored. It would then give a basis to further investigate links between perceptual characteristics of the Dízi played, acoustic features associated with the output sound of the Dízi, and how expert players interact accordingly. Further, based on the accuracy performance of all three investigations, it is evident that in fact – for this sample set at least – players are marginally more easily discriminated than the instrument used or musical exercise performed. This offers a first-hand qualitative indicator of the relative roles and the interaction between player and instrument which is critical for advanced musical performance.

6. REFERENCES

Balamurali BT, Da Yang Tan, Teck-Seng Ng, Jer-Ming Chen. 2018. "Fabrication of Chinese Transverse Flute, Dízi: an Acoustic Impedance Analysis." *Proceedings of WESPAC2018, Delhi, India*.

Giannakopoulos, Theodoros. 2016. "pyAudioAnalysis: Feature Extraction in pyAudio library", https://github.com/ tyiannak/pyAudioAnalysis/ wiki/3.-Feature-Extraction, accessed Jan 2018.

Giannakopoulos, Theodoros, Dimitrios Kosmopoulos, Andreas Aristidou, and Sergios Theodoridis. 2006. "Violence content classification using audio features." Hellenic Conference on Artificial Intelligence.

Giannakopoulos, Theodoros, and Aggelos Pikrakis. 2014. *Introduction to audio analysis: a MATLAB® approach*: Academic Press.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and regression by randomForest." *R news* 2 (3):18-22.

Muda, Lindasalwa, Mumtaj Begam, and Irraivan Elamvazuthi. 2010. "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques." *arXiv preprint arXiv:1003.4083*.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. 2011. "Scikit-learn: Machine learning in Python." *Journal of machine learning research* 12 (Oct):2825-2830.

Peeters, Geoffroy. 2006. "Musical key estimation of audio signal based on hidden Markov modeling of chroma vectors." Proceedings of the International Conference on Digital Audio Effects (DAFx).

Rabiner, Lawrence R, and Biing-Hwang Juang. 1993. "Fundamentals of speech recognition."

Rabiner, Lawrence R, and Ronald W Schafer. 2011. *Theory and applications of digital speech processing*. Vol. 64: Pearson Upper Saddle River, NJ.

Seni, Giovanni, and John F Elder. 2010. "Ensemble methods in data mining: improving accuracy through combining predictions." *Synthesis Lectures on Data Mining and Knowledge Discovery* 2 (1):1-126.

Shen, Jia-lin, Jeih-weih Hung, and Lin-shan Lee. 1998. "Robust entropy-based endpoint detection for speech recognition in noisy environments." Fifth international conference on spoken language processing.