# Study on 人間国宝 山口五郎 尺八の神髄 尺八本曲

DRAFT VERSION: 9/6/2024

There are 4500+ phrases in the entire body works of the 琴古流 repertoire of 本曲 pieces for the 尺八. Over years 山口五郎 — designated as 重要無形文化財保持者 (人間国宝) in 1992 — recorded this repertoire, which was then compiled and released as a 12 Compact Disc (CD) box set by Victor Japan in 1999 with support from 日本伝統文化振興財団. The box set is titled 人間国宝 山口五郎 尺八の神髄 尺八本曲 and contains a total of 36 tracks.

Described herein is a computational analysis on the entire set of recordings from all 12 of the CDs. The most general description of this analysis is to measure similarity/difference among all or a subset of phrases in the repertoire. As a quick teaser, the mean phrase duration and standard deviation across the entire corpus of phrases is 7.691 seconds and 3.220 seconds, respectively. And, the phrase representing what could be the “most quintessential” across the entire corpus starts at 13'22.76" into disc 5 track 1, 吉野鈴慕, and ends 3.179 seconds later. [Based on mean DTW distance of 1206.82. The read buffer size is 1024 frames, the start buffer index is 34572, the end buffer index is 34721, the duration between the start and end buffer indexes is 3.460 seconds, and the duration of the pitch data is 3.179 seconds.]

After extracting each track from each CD, discarding the right channel, and saving the audio as monaural, 16-bit, 44.1kHz WAV files, the pYIN pitch tracking algorithm is used on each WAV file for phrase boundary identification (i.e., segmentation).

The segmentation method works as follows: The first non-zero pitch value marks the start of a phrase. If the number of zero pitch values (e.g., 5) within a window of pitch values (e.g., 10) equals/exceeds a set number (i.e., a percentage of the window size) the end of the phrase is marked. (The number of zero pitch values and window size ["gate_hold": 5, "gate_release": 10] can be reset before each full analysis.) Upon marking the end of a phrase, all pitch values in the last window are removed. The pitch data is then saved in its own file with a filename that includes a unique tag for that phrase. [NOTE: this paragraph requires double-checking.]

The same unique tag is included in the filename of another file which contains the WAV file filename and the start and end buffer indexes into the sound file. This allows for the pitch data to be matched with its corresponding audio region from the CD track.

The read buffer size used for the WAV files is 1024 frames, roughly corresponding to a duration of 23.2 milliseconds. Each pitch data point is based on 4096 frames of audio with a hop size of 1024 frames (or 4x overlap), roughly corresponding to 23.2 milliseconds or 43.1 pitch data points per second.

The duration of each phrase is also saved in its own file. The filename also includes the phrase's unique tag. The duration value is calculated from the pitch data set, which is not identical to the duration calculated from the start and end buffer indexes. This is because the latter includes the buffers representing the trailing zero pitch values at the end of the phrase.

At this point we have three types of files for each phrase, each including the same unique tag in the filename. The files are pitch data, audio track and region data, and duration data. After a full analysis, 4500+ of each of those three types of files are generated.

What might one do with all of that data?

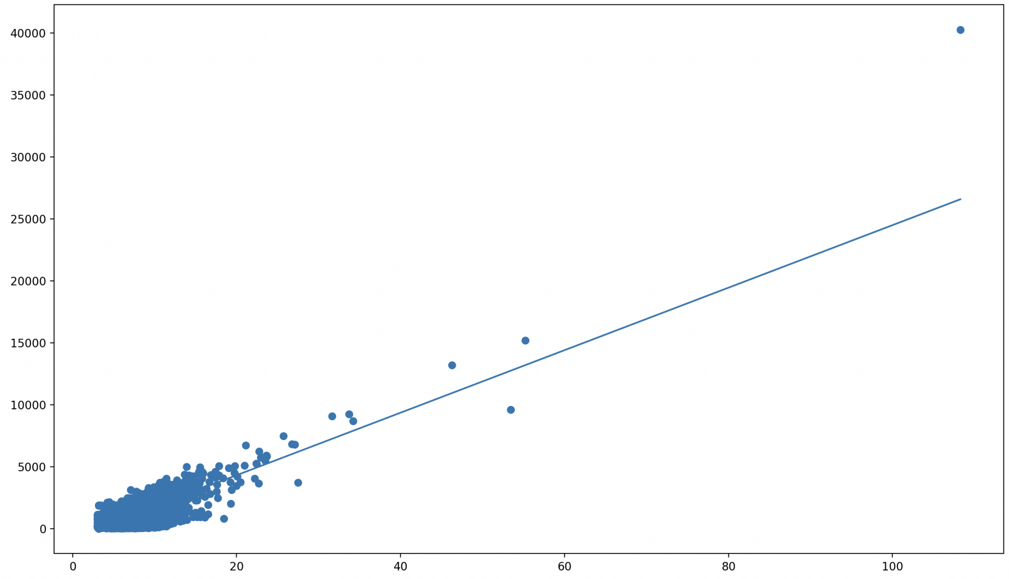
Enter Dynamic Time Warping (DTW), a decades old method often used for calculating a distance between two time series data sets. Each pitch data set associated with a phrase is a time series. Thus, a distance measurement can be calculated between any two phrases in the corpus. The distance value between a pitch data set and itself — an absolute perfect match — is 0. All other distance values are positive. The larger the value, the greater the distance.

Taking one phrase and calculating the distance between it and each and every other phrase in the corpus generates a vector of distance measurements. In the case of the 吉野鈴慕 phrase mentioned earlier, the 6 phrases with the lowest distance values in the vector (including itself) is summarized in the following table:

| WAV File | Start Time | Duration | Distance Value |
|---|---|---|---|
| 5-01 Yoshiya Reibo.wav | 13:22.760 | 3.179 | 0.000 |
| 9-01 Meguro Jishi.wav | 14:06.901 | 4.8 | 29.874 |
| 9-01 Meguro Jishi.wav | 4:50.180 | 5.867 | 31.999 |
| 8-01 Reibo Nagashi.wav | 14:31.607 | 6.144 | 32.992 |
| 9-01 Meguro Jishi.wav | 12:32.837 | 5.099 | 33.267 |
| 6-01 Sakae Jishi.wav | 51:8.006 | 4.629 | 38.181 |

One notable characteristic of DTW is that smaller data set sizes will tend to produce smaller distance measurements. The following figure plots distance against duration. It also includes a linear regression line derived from all data points. The distances are between the one 吉野鈴慕 phrase mentioned earlier and each and every other phrase in the corpus.

```
*** plot costs against durations for one phrase ***
```



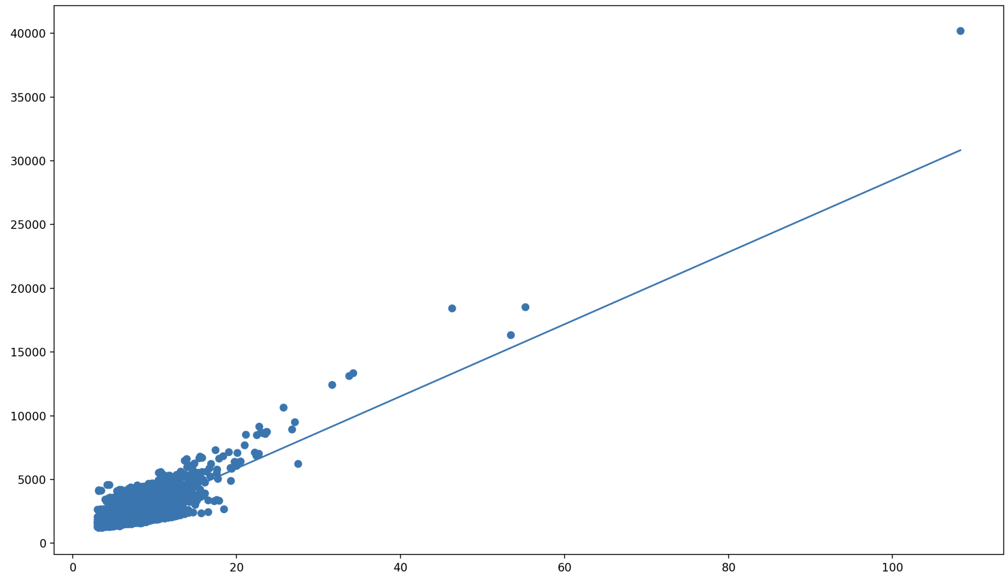
This plot suggests in general that the longer the duration of the phrase the larger the distance value.

While a single vector of distances associated with one phrase can provide insight into the repertoire, one might also consider looking at the vectors for all phrases in the repertoire. [The full set of vectors forms an NxN matrix, where N is the total number of phrases in the corpus.]

Summing all distance values contained in a single vector (or taking its mean value) generates an aggregate distance value. One simple conclusion across all vectors, then, might be that the phrase with the lowest aggregate distance value represents the one that is on the average most closest to all phrases or "most quintessential" in the corpus.

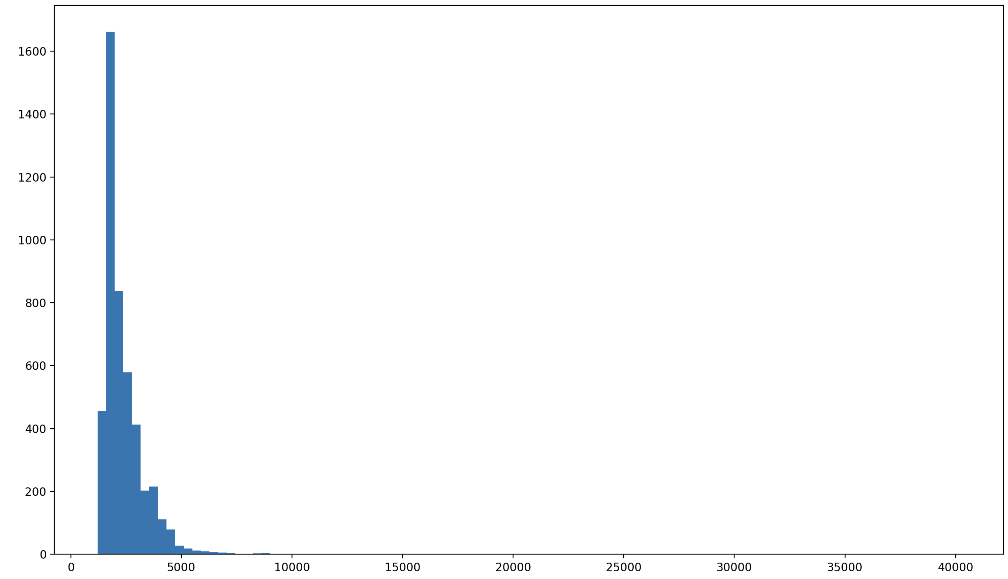
A kind of meta- or corpus-level vector can be created by generating aggregate distance values for each phrase in the corpus. Each element of the vector is associated with a unique phrase and its properties, such as duration. Pairing phrase duration with aggregate distance, the following figure plots aggregate distance against duration for each unique phrase. Again, a linear regression line is also provided.

```
*** plot costs against durations for all phrases ***
```



Here, too, one observes a data set size bias, one that favors phrases with shorter durations.

```
*** plot histogram of costs for all phrases ***
```



```
*** play lowest costs for 20230715161509845710 ***
```

```
5-01 Yoshiya Reibo.wav 1024 34572 34721 802.7602721088435 806.2200453514739
13:22.760272108843537 3.179
['20230715161541463881', 29.874374389648438]
9-01 Meguro Jishi.wav 1024 36473 36698 846.9014058956916 852.12589569161
14:6.901405895691596 4.8
['20230715161539614208', 31.999221801757812]
9-01 Meguro Jishi.wav 1024 12497 12772 290.1797732426304 296.56526077097504
4:50.179773242630404 5.867
['20230715161558694845', 32.992027282714844]
8-01 Reibo Nagashi.wav 1024 37537 37825 871.6074376417233 878.2947845804989
14:31.607437641723322 6.144
['20230715161541163783', 33.26720428466797]
9-01 Meguro Jishi.wav 1024 32422 32661 752.8373696145125 758.3869387755102
12:32.837369614512454 5.099
['20230715161653369089', 38.18104553222656]
6-01 Sakae Jishi.wav 1024 132128 132345 3068.0061678004536 3073.044897959184
51:8.006167800453568 4.629
```