SINGING PITCH ESTIMATION IN JINGJU MUSIC

David Bedoya
Universitat Pompeu Fabra, Barcelona
josedavid.bedoya01@estudiant.upf.edu

ABSTRACT

Leveraging the increasing attention from MIR researchers in jingju (also known as Beijing or Peking opera) music, this paper provides a comparison of the performance of a state-of-the-art deep learning architecture against alternative approaches for the singing pitch estimation task in jingju. For this, the singing pitch estimations are first computed using three algorithms in subsets of jingju audio collections, then the accuracies of these estimations are evaluated, and base on these evaluations, the performance of the algorithms is compared. The results show that the neural network performed better in most of the cases addressed here.

1. INTRODUCTION

Singing pitch estimation refers to the task of extracting the fundamental frequency to a monophonic melody, i.e. a single singer, or the predominant melody from polyphonic music signals, i.e. a singer with musical accompaniment. The main goal of this paper is to compare the performance of a state-of-the-art deep convolutional neural network in this task against alternative approaches, using Jingju music as source material.

1.1 Jingju Music Datasets

Jingju is one of the genres of Chinese traditional theatre arts, arguably the most widespread and acclaimed one. Since its core component is singing, the accurate estimation of the fundamental frequency of the voice is of utmost importance in the research of this tradition. Here, two jingju music datasets are used for the estimations:

1.1.1 Monophonic

For monophonic melody extraction, here are used the a cappella dataset used in [7]. Its audio recordings, hereinafter called the a cappella recordings, comprise 41 interpretations of 33 unique arias sung by 13 jingju singers, and are available upon request [1] . The dataset also contains the ground truth pitch tracks.

[1] https://zenodo.org/record/832736

1.1.2 Predominant

For predominant melody extraction, here are used the audio recordings used in [6], hereinafter called the commercial recordings, which are a subset of the Jingju Audio Recordings Collection (JARC) [2] . The first author of the just cited paper manually annotated the pitch tracks of this recordings using Sonic Visualiser [3] .

The remainder of the paper is organized as follows. The following section presents the methodology for estimating the jingju singing fundamental frequency in the datasets. Section 3 explains how these estimations are evaluated. Accuracy results of these evaluations are presented and discussed in Section 4. In the last section conclusions and future work are pointed out.

2. METHODOLOGY

To compute the singing pitch estimations, each of the two types of recordings (a cappella and commercial) are processed with two algorithms, one using a pre-trained deep learning architecture and other that follows a hand-engineering approach. The code is openly available [4] and has precise explanations for each step of the process.

2.1 A cappella recordings

It would be desirable that the recordings had the metadata and annotations in terms of the jingju musical system. As a further matter, this would be useful for studying jingju singing pitch estimation in terms of its musical system elements [5] (not addressed here). The a cappella recordings used in [7] do not have these metadata and annotations, but the Jingju a Cappella Recordings Collection (JaCRC) [6] , which include these recordings and many more (albeit with a different name), does. Then, as a way of better identifying the audio samples, the a cappella recordings are matched with the appropriate JaCRC recordings using the cosine of the angle between audio vectors as a similarity function. After this, the pitch estimations from these audio samples are computed using the following algorithms:

[2] https://zenodo.org/record/1475846
[3] http://www.sonicvisualiser.org/
[4] https://github.com/jdavibedoya/f0-jingju
[5] For more information about the music elements in jingju music, please refer to [8, 9, 12]
[6] https://zenodo.org/record/3251761

2.1.1 pYIN

This algorithm [5] is a modification of the well-known YIN algorithm for fundamental frequency ($F0$) estimation. Here, the Essentia implementation [7] is used. With pYIN, three frame size steps $(512, 1024, 2048)$ and three hop size steps $(256, 441, 512)$ are explored for hyper-parameter tuning.

2.1.2 CREPE

This algorithm [4] is a monophonic pitch tracker based on a deep convolutional neural network operating directly on the time-domain waveform input. With CREPE, two time steps $(5, 10)$ for hyper-parameter tuning are explored here.

2.2 Commercial recordings

As already stated, the commercial recordings used here are the same ones used in [6]. In that paper, the recordings are identified with a MusicBrainz ID. With the metadata available on MusicBrainz [8], the audio files can be easily found in the JARC. Once the audio samples are found, their pitch estimations are computed using the following algorithms:

2.2.1 MELODIA

This algorithm [10] estimates the fundamental frequency of the predominant melody from polyphonic music signals. Here, the Essentia implementation [9] is used. With MELODIA, four frame size steps $(512, 1024, 2048, 4096)$ and five hop size steps $(128, 256, 441, 512, 1024)$ are explored for hyper-parameter tuning.

2.2.2 CREPE

Although CREPE is not specifically designed to estimate the predominant melody from polyphonic music signals, here it is used for this purpose in the commercial recordings in order to provide performance results that can be compared to those of the MELODIA algorithm. The same approach is followed in [2], albeit with a different dataset (iKala [11]). With CREPE, this time eight time steps $(5, 10, 15, 20, 25, 30, 35, 40)$ are explored for hyper-parameter tuning.

3. EVALUATION

The mir_eval [1] library is used to evaluate the accuracy of the algorithm's predictions. For hyper-parameter tuning, the raw pitch accuracy (RPA) metric at a tolerance of $1/2$ semitone is used. The raw pitch accuracy computes the proportion of melody frames in the reference for which the frequency is considered correct. Subsequent to the extractions of the pitch estimations per recording, the performance of each algorithm is computed in the respective dataset where it estimated the pitch tracks, as the proportion of melody frames in the dataset ground truths for which the estimated frequency is considered correct, i.e. the RPA for the respective dataset. For this, three pitch tolerances ($10, 25$, and 50 cents) are used here.

[7] https://essentia.upf.edu/reference/std_PitchYinProbabilistic.html
[8] https://musicbrainz.org/
[9] https://essentia.upf.edu/reference/std_PredominantPitchMelodia.html

4. RESULTS AND DISCUSSION

Tables 1 and 2 expose the average raw pitch accuracies and their standard deviations for the a cappella and commercial recordings respectively, and at the three pitch evaluation tolerances.

Cents	CREPE	pYIN
10	0.804 ± 0.060	0.705 ± 0.088
25	0.942 ± 0.023	0.912 ± 0.050
50	0.977 ± 0.012	0.971 ± 0.017

Table 1. RPA results for the monophonic melody extraction task performed on the cappella recordings.

Cents	CREPE	MELODIA
10	0.603 ± 0.084	0.695 ± 0.084
25	0.802 ± 0.075	0.775 ± 0.107
50	0.859 ± 0.066	0.794 ± 0.115

Table 2. RPA results for the predominant melody extraction task performed on the commercial recordings.

From these results, it can be seen that CREPE outperforms the pYIN algorithm in all the cases addressed here. At a tolerance of $1/2$ semitone, the performance of the pYIN algorithm is less than that of CREPE by just 0.6%. Nonetheless, as pitch tolerance becomes more restrictive, the difference between the performance of these two algorithms increases in favor of the neural network by 3% and 9.9% at tolerances of 25 and 10 cents, respectively. Despite the fact that the CREPE algorithm has not been specifically designed for the predominant melody estimation task, it performed better than MELODIA by 6.5% at a tolerance of $1/2$ semitone. However, as pitch tolerance becomes more restrictive, the difference between the performance of these two algorithms decreases, and the roles are swapped at a tolerance of 10 cents, that is, the performance of the MELODIA algorithm is greater than that of CREPE by 9.2% at the most restrictive tolerance evaluated here. It can also be observed that the results for monophonic melody extraction exhibit greater variability than those for predominant melody extraction.

5. CONCLUSIONS AND FUTURE WORK

This paper has presented a comparison of the performance of a pre-trained deep learning architecture (CREPE) against alternative approaches (pYIN and MELODIA) for the singing pitch estimation task using jingju music as source material. The neural network has obtained the best results in all cases for the estimation of the monophonic melody, and in most cases for the estimation of the predominant melody in the polyphonic commercial recordings, this is notable considering the fact that this deep

learning algorithm has not been specifically designed for the latter purpose.

In future work, the performance of the CREPE network for singing pitch estimation in jingju music could be improved by carrying out transfer learning from its pretrained model using the jingju audio collections. Correspondingly, to improve the performance of predominant pitch estimation with an alternative approach, vocal timbral models specifically trained for jingju music could be used to select pitch contours or even to generate them from the salience function within the MELODIA algorithm, this latter approach is followed in [3] for Carnatic music. It would also be relevant to compute and store the singing pitch estimations per melodic unit [10]. With this, it would be possible to compare the RPA results in terms of the elements of the jingju musical system.

6. ACKNOWLEDGEMENTS

The author wishes to express his gratitude to Prof. Rafael Caro for helping to solve some questions about the datasets and sharing the manually annotated pitch tracks of the commercial recordings used here. The methodology followed in this paper has been significantly inspired by the analysis performed in Section 2.1 of [2].

7. REFERENCES

[1] E. J. Humphrey-J. Salamon O. Nieto D. Liang C. Raffel, B. McFee and D. P. W. Ellis. mir_eval: A transparent implementation of common mir metrics. In *Proceedings of the 15th International Conference on Music Information Retrieval, 2014.*, pages 367–372, 2014.

[2] J. Bonada P. Chandna E. Gómez, M. Blaauw and H. Cuesta. Deep learning for singing processing: Achievements, challenges and impact on singers and listeners. arXiv 1807.03046, 2018.

[3] V. Ishwar. Pitch estimation of the predominant vocal melody from heterophonic music audio recordings. Master's thesis, Universitat Pompeu Fabra, Barcelona, 2014.

[4] P. Li J. W. Kim, J. Salamon and J. P. Bello. Crepe: A convolutional representation for pitch estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2018)*, pages 161–165, 2018.

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

[6] N. Kroher R. Caro Repetto, R. Gong and X. Serra. Comparison of the singing style of two jingju schools. In *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015*, pages 507–513, 2015.

[7] Y. Yan R. Gong and X. Serra. Pitch contour segmentation for computer-aided jingju singing training. In *SMC 2016 - 13th Sound and Music Computing Conference, Proceedings*, pages 172–178, 2019.

[8] R. Caro Repetto and X. Serra. Creating a corpus of jingju (beijing opera) music and possibilities for melodic analysis. In *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014*, pages 313–318, 2014.

[9] R. Caro Repetto and X. Serra. A collection of music scores for corpus based jingju singing research. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017*, pages 46–52, 2017.

[10] J. Salamon and E. Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, 2012.

[11] Z.-C. Fan H.-W. Chen L. Su Y.-H. Yang T.-S. Chan, T.-C. Yeh and R. Jang. Vocal activity informed singing voice separation with the ikala dataset. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015)*, pages 718–722, 2015.

[12] E. Wichmann. *Listening to Theatre: The Aural Dimension of Beijing Opera*. University of Hawaii Press, Honolulu, 1991.

[10] The melodic unit in jingju corresponds to the line of lyrics