Intonation: a Dataset of Quality Vocal Performances Refined by Spectral Clustering on Pitch Congruence

Sanna Wager[1], George Tzanetakis[2,3], Stefan Sullivan[3], Cheng-i Wang[3], John Shimmin[3], Minje Kim[1], Perry Cook[3,4] scwager@indiana.edu, gtzan@cs.uvic.ca, minje@Indiana.edu

[1] Indiana University, School of Informatics, Computing, and Engineering, USA
[2] University of Victoria, Department of Computer Science, Canada
[3] Smule, Inc, San Francisco, USA
[4] Princeton University, Departments of Computer Science and Music, Princeton, NJ, USA

SAIGE
INDIANA UNIVERSITY
SIGNALS & ARTIFICIAL INTELLIGENCE GROUP IN ENGINEERING
http://saige.sice.indiana.edu

INTRODUCTION

- We introduce the "Intonation" dataset of amateur vocal performances
- It contains **public performances collected from Smule, Inc.**
- They are selected from a large database for **tendency for good intonation**
- We describe the semi-supervised approach for choosing these performances
- This approach generalizes to other datasets
- We compare the intonation distributions of the selected performances versus the remaining ones in the large collection

CUSTOM DATASETS

- **Scenario:** A research topic in audio or music information retrieval is uncommon
- Data is hard to find
- A **subset** of a **huge dataset** for another task is suitable
- Desired features are not labeled and can be hard to model
- Manual filtering is labor intensive
- How can we automate the process?
- One approach involves **feature extraction** and **clustering**
- **Semi-automatic** process
- Reduces manual component to a manageable size

MUSICAL INTONATION

- We wish to select "in-tune" performances from tens or thousands of performances
- **"In tune" is subjective**
- We can measure pitch patterns across performances that we consider "in tune"
- Directly defining a model is difficult
- Intonation studies [1, 2] show frequent, deliberate deviations from the equal-tempered scale
- Pitch also varies due to pitch bending, vibrato, natural characteristics of the voice, and harmonization

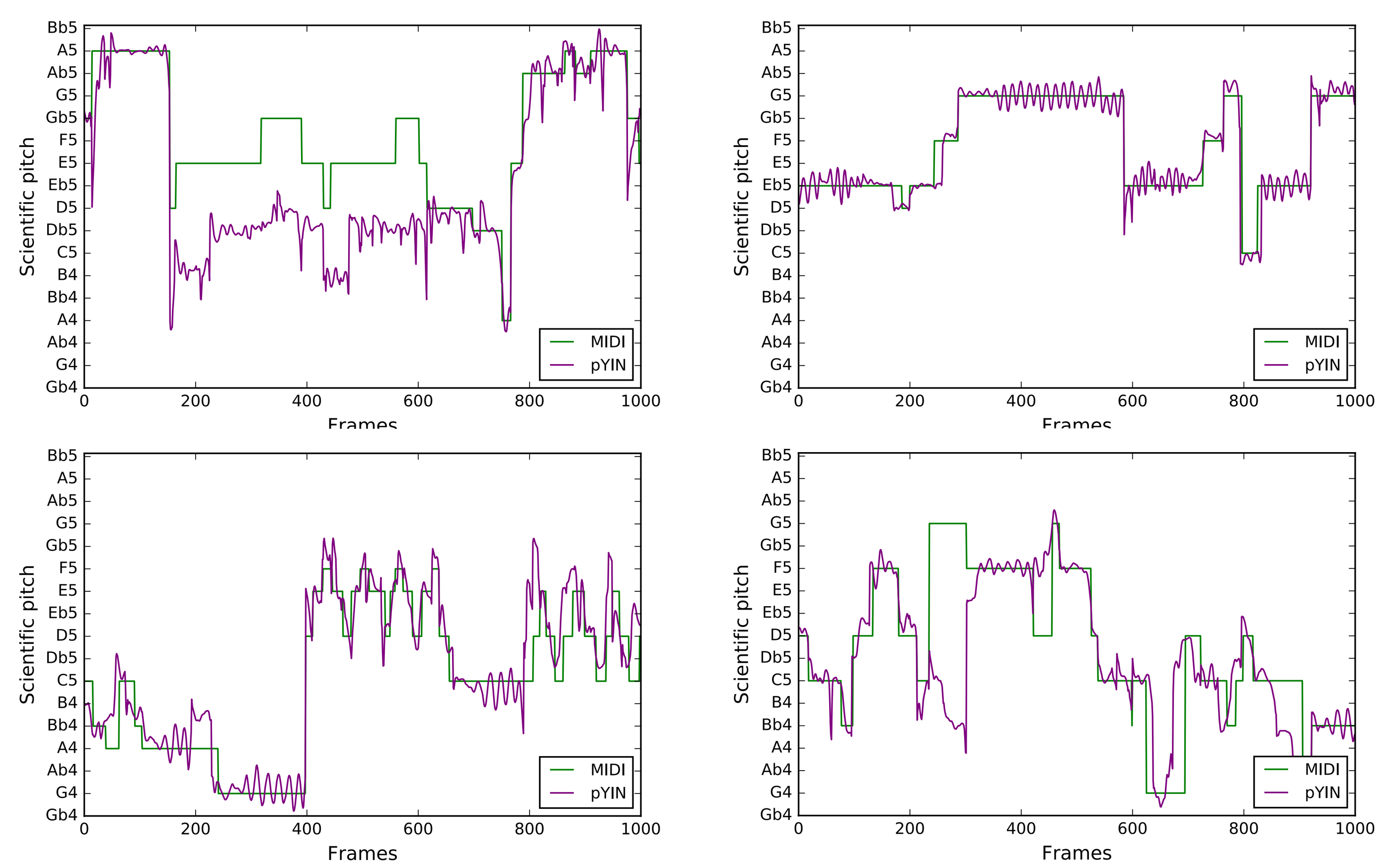

Figure 1. Singing pitch analysis (pYIN algorithm [3]) and aligned MIDI score in four performances. Which performances are "in tune"? Our analysis resulted in choosing the top two but not the bottom two.

- Avoid creating an explicit definition by using a semi-supervised approach

RELATED WORK

- Nichols *et al*. predict singer talent on YouTube based on features extracted from the audio [4] using a **pitch deviation histogram** from the short-Time Fourier Transform amplitude peaks
- Our feature extraction task is different: **We have access to the musical scores** and the **audio sources are separate**
- Lim *et al*. compare performance pitch and musical score in the context of a tool for musical performance visualization [5]

FEATURE EXTRACTION

- Pre-filter tens of thousands of performances for basic score alignment
- Summarize intonation patterns using low-dimensional set of features
- Compare frame-wise pitch analysis (pYIN) to MIDI score with 11ms. resolution
- **Deviations in Cents**: $1200 * \log_2 \frac{f_1 + \epsilon}{f_2 + \epsilon}$

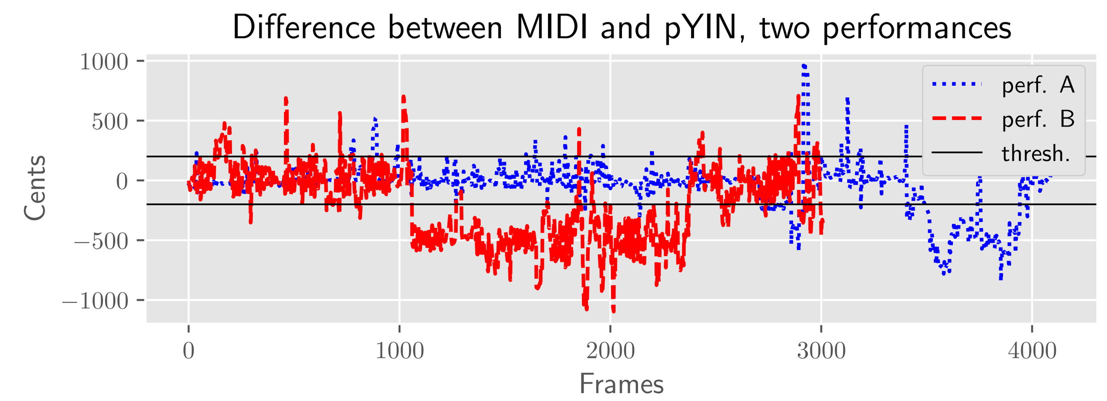

Figure 2. frame-wise deviations in cents

- **Treat each performance's deviations as a distribution**
- Compute 31 evenly spaced **quantiles**

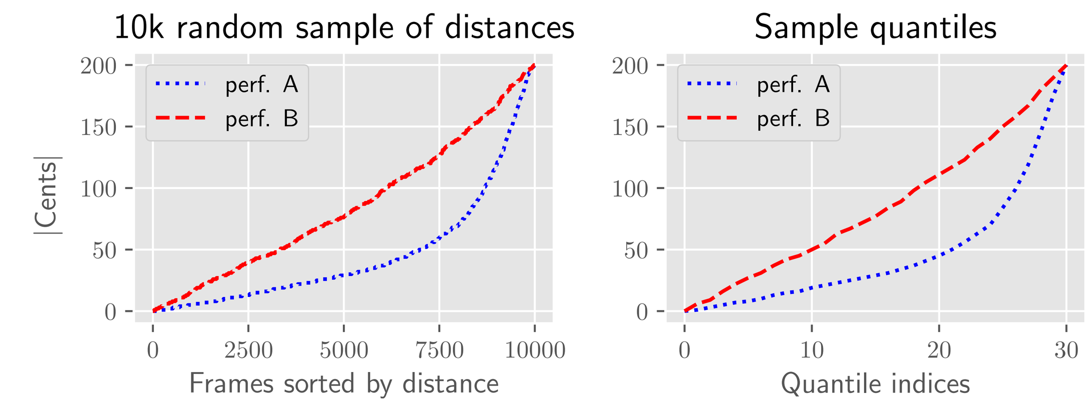

Figure 3. Sorted quantiles for two performances. The red one is included in the "Intonation" dataset.

SPECTRAL CLUSTERING

- Cluster the quantized performances (Speclus algorithm [6])
- Use **signless Laplacian** matrix as the adjacency graph (50 nearest neighbors)

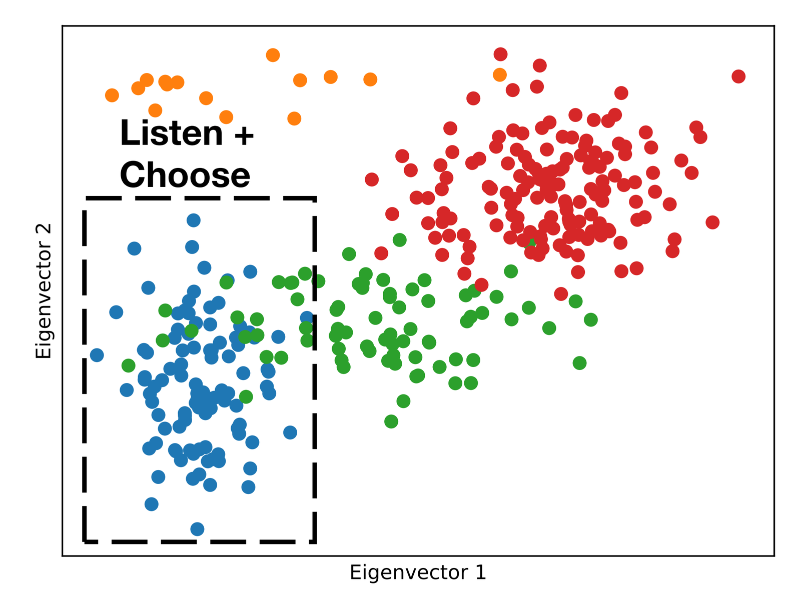

Figure 4. Cluster 5000 songs at a time into 3 or 4 clusters, depending on which value produced better Newman modularity.

- Listen to 50 samples from every cluster and subjectively determine the intonation of every performance, evaluating it as "in tune", "neutral", or "out of tune".
- Consistently, one cluster produced significantly better results

DATASET CONTENTS

- **4703 vocal performances** of 474 unique arrangements by 3556 singers
- Metadata
- Frame-wise pitch analysis of vocals (probabilistic YIN algorithm)
- Backing track features for 30-90 second range (computed using Librosa)
  - Constant-Q transform
  - Chroma
  - Mel-frequency cepstrum coefficients
  - Root mean square error
  - Onset

Figure 5. Dataset and detailed description available on Stanford DAMP page.

APPLICATIONS

Applications include:

- Singing style analysis
- Informed source separation
- Query by humming

Notes on the dataset:

- Not every selected performance is in tune and not every other one is out of tune
- Good enough for many machine-learning applications
- Intonation dataset represents **majority genre**
- Less common genres like Blues and Country performances got left out
- Excellent performances in these genres have a different pitch behavior (flatter)
- They are in a different cluster

INTONATION ANALYSIS

- Compare distributions of performances from selected clusters versus the others
- Same analysis as before, but keeping everything
  - No absolute value
  - No threshold at 200 cents
- Dynamic time warping to align the MIDI and singing pitch

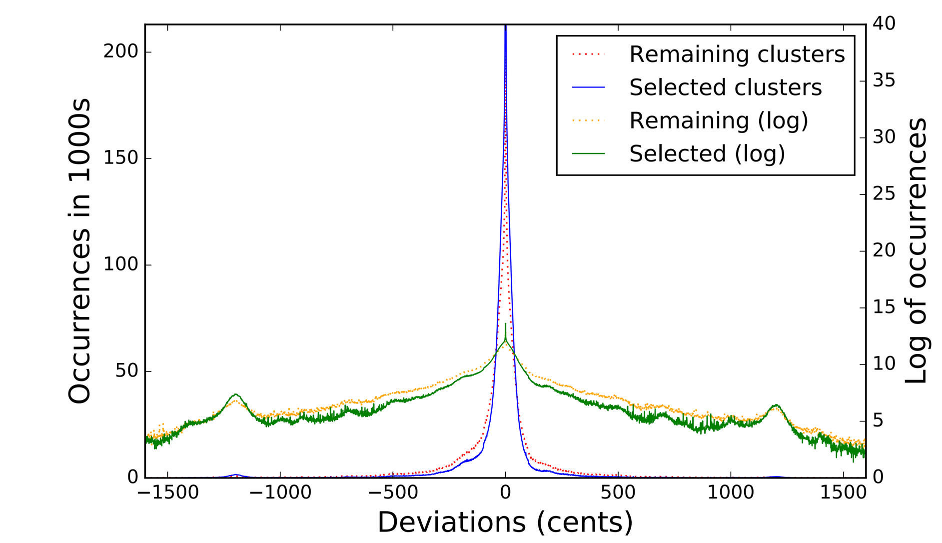

Figure 6. Global histograms of singing pitch deviations from equal-tempered MIDI

- Analyze distribution of positive versus negative deviations from the score
- Unexpected higher concentration on the negative side

Results from "Intonation" dataset (4702 performances)		
Cents range	Negative/positive deviation ratio	Var
1 to 2	0.500	0.001
2 to 16	0.506	0.001
1 to 100	0.532	0.002
100 to 300	0.727	0.002
Results from other performances (9701 performances)		
Cents range	Negative/positive deviation ratio	Var
1 to 2	0.500	0.001
2 to 16	0.509	0.001
1 to 100	0.541	0.002
100 to 300	0.700	0.002

Figure 7. Positive and negative deviation counts for cents ranging from 1 to 100
Table 1. Probability estimates of positive versus negative deviations, computed using bootstrapping [7]

REFERENCES

[1] R. Parncutt and G. Hair, "A psychocultural theory of musical interval: Bye bye Pythagoras," *Music Perception: An Interdisciplinary Journal*, vol. 35, no. 4, 2018.
[2] J. Devaney, J. Wild, and I. Fujinaga, "Intonation in solo vocal performance: A study of semitone and whole tone tuning in undergraduate and professional sopranos," ISPS, 2011.
[3] M. Mauch and S. Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," ICASSP, 2014.
[4] E. Nichols, C. DuHadway, H. Aradhye, and R.F. Lyon, "Automatically discovering talented musicians with acoustic analysis of YouTube videos," ICDM, 2012.
[5] K.A. Lim and C. Raphael, "Intune: A system to support an instrumentalist's visualization of intonation," Computer Music Journal, vol. 34, no. 3, 2010.
[6] M. Lucinska and S.T. Wierzchon, "Spectral clustering based on k-nearest neighbor graph," IFIP, 2012.
[7] B. Efron and R. J. Tibshirani, An introduction to the bootstrap, CRC press, 1994.