Master thesis on Sound and Music Computing

Universitat Pompeu Fabra

Automatic Assessment of Singing

Ninad Vijay Puranik

Supervisor: Prof. Baris Bozkurt

Co-Supervisor: Prof. Xavier Serra

August 2019

Master thesis on Sound and Music Computing

Universitat Pompeu Fabra

Automatic Assessment of Singing

Ninad Vijay Puranik

Supervisor: Prof. Baris Bozkurt

Co-Supervisor: Prof. Xavier Serra

August 2019

Contents

Acknowledgement

I feel extremely fortunate to have you, Xavier Serra, as my co-supervisor for this thesis project. Thank you for your online course which introduced the amazing SMC world to me. Thank you for your trust and confidence in my ability to work on this project for MusicCritic. Thank you for showing me the simpler broader perspectives whenever I was caught up with the complicated details.

I would like to thank my supervisor Baris Bozkurt for helping me learn to frame my curiosities into research problems, develop plans to solve them and put the positive and negative results into context. Thank you for always being available when I needed support, advice and encouragement.

Special thanks to Vishwajit Godbole, Girija Erande and all the students from Sargam Sangeet Vidyalaya, Pune without whom it would have been impossible to build a dataset for this project. Also, a great thanks to all MTG members and friends who contributed to this project with their singing.

Thank you Nitin Amin for all the valuable musical discussions and your suggestions for the project. Big thanks to my ASPlab colleagues Krishna, Rafael, Blazej, Seva, Miguel and Alia for the fruitful collaborations, discussions, help, and encouragement.

Thank you, Sohan Kale, for being a great friend right from the Master's application to my last day in Barcelona.

Thanks to my parents for being a constant moral support.

Finally, thanks to my little boy Anvay and wife Pranjali who inspired and encouraged me to take up and complete this study.

Abstract

This thesis aims to develop an automatic singing evaluation system specially suited to evaluate notes singing exercises. We build Hindustani singing datasets with a combined collection of audio samples of 349 reference and performance pairs. The samples are annotated with an overall grade based on pitch accuracy. For this purpose, we develop a GUI grading tool which provides a visual feedback for the performance. This tool helps reduce human biases in the grade annotations. The existing baseline system (BMCS) for singing assessment developed using Turkish Conservatory dataset (MAST dataset) is extensively studied to identify audio alignment as one of the possible areas of improvement. A methodology and appropriate metrics are devised to test the audio alignment performance. Using this methodology, different features are tested to demonstrate that an improved audio-to-audio alignment system can be achieved using a 120-dimensional HPCP feature.

The second part of this study is focused on finding suitable features to assess a singing performance, given a good alignment of reference and student audio. A novel 'pitch-histogram cosine distace' feature is devised to measure note-level accuracy of singing. The effectiveness of these features with respect to the baseline features is shown by linear regression models trained and tested using the Hindustani and MAST datasets. The effectiveness of 'pitch-histogram cosine distance' is indicated by the low mean absolute errors and the interpretability of the linear models developed. These features are also used to provide a note-level accuracy visualization of student performance.

Chapter 1

Introduction

1.1 Context

The TECSOME (TEChnologies for Supporting Online Music Education) project is an European Research Council funded project (ERC grant agreement 768530) being developed at the Audio Signal Processing Lab (ASPLab) of the Music Technology Group (MTG) at Universitat Pompeu Fabra (UPF) in Barcelona. As part of the project an automatic music performance assessment system, named Music Critic, is being developed with a target to support music performance courses and help them scale up to MOOC (Massive Online Open Course) level.

The MTG, in collaboration with Ragasphere [1] and on the Kadenze on-line education platform, offers a set of courses[2] on North Indian Classical Music. The assignments for these courses require the students to listen to a reference melody sung by a teacher and repeat it as accurately as possible. The student rendition is recorded and submitted through the web-browser based interface provided by MusicCritic. The MusicCritic framework supports the analysis of the student submissions using Music Information Retrieval (MIR) technologies to provide a qualitative and quantitative feedback of the student performance to the course instructors. A final grade can

[1] `https://www.ragasphere.com/`,

[2] `https://www.kadenze.com/courses/north-indian-classical-music-i-fundamental-elements/`, last accessed on 15 August, 2019

be given to the student based on the MusicCritic feedback and the instructor's judgement after listening to the performance. This thesis presents the work done in this context for the development of a system for the automatic evaluation of singing for Hindustani (North Indian Classical) Music exercises.

1.2 Motivation

Online education technologies help to overcome practical difficulties of availability of classroom space, physical distances between places, adjusting the time schedule and customizing the pace of lectures for every individual. It has made good teachers and courses accessible to a wide audience at the international level. Online degree programs and MOOCs have been especially successful in fields like computer science, mathematics and programming. A major reason for this is the deterministic nature of these fields. Exercises and Assignments in such topics can be easily constructed to have a limited set of correct answers which can be received in machine readable format by a software program to perform an automatic evaluation on the basis of a well defined rubric. A grade can thus be assigned to every student submission without the intervention of a human expert. This allows the online course to scale massively.

Evaluation of musical performance is a subjective task. Each person may have a different opinion about the same performance. A person's opinion of the same performance may differ with time or mood. We have traditionally relied on human experts to judge a music performance on the basis of technical aspects like correctness of pitch and tempo and aesthetic aspects like dynamics, emotion and voice quality. For an absolute evaluation, i.e. when a performance is not compared with another, we could be satisfied with a qualitative assessment. However, a quantitative assessment is desired when we have to rate a performance relative to a set of performances or to a standard. For example we would need a quantitative assessment to select/reject a performance during a music competition or audition, or when we need to give numerical or letter grades to students enrolled for a music performance course.

North-Indian Classical Music, popularly known as Hindustani music is one of the oldest form of music existing today. While the study of European Art music was popularized all over the world during the colonial era, Hindustani Music retained much of its indigenous characteristic due to its transmission through the generations by 'Guru-Shishya Parampara'. As per the 'Guru-Shishya' tradition of Hindustani Music, the guru (teacher) demonstrates the accurate way of singing a melodic phrase in a Raga, the basic melodic framework of Hindustani music, which the shishya (student) is expected to repeat after. This process is repeated till the Guru is satisfied with the shishya's rendition after which they move to the next phrase in the Raga.[3] The student is expected to memorize these melodic patterns by repeating them several times and create a mental map of the collection of melodic phrases to develop a tacit understanding of a Raga. A subconscious learning of the Raga is also achieved by observing how the Guru improvises within the Raga framework during public and private concerts. This system of learning has preserved the peculiar characteristics of this music surprisingly well while still allowing a lot of scope for individualistic expression.

Hindustani music continues to be one of the few popular Classical music forms which is still transmitted by the oral tradition. Systems for transcription of Hindustani music were developed by Bhatkhande and Paluskar [1] in the 20th century. These are now popular with amateurs, and the experts use them at times as a memory aid. However, these systems are at best prescriptive and to fully understand the rendition of a transcription, it is considered necessary to listen to an audio performance.

The focus on oral-aural learning without the use of a Western Music style transcription, the inherent improvisational nature and micro-tonality make supporting an online course on Hindustani music a challenging Music Information Retrieval (MIR) task. The growing popularity of Hindustani music, especially in the Western countries, coupled with the scarcity of expert Gurus in the Western world makes it a relevant task in today's online music education market.

[3] Audio recordings of guru Pt. Gajanbuva Joshi teaching his illustrious disciple Pt. Ulhas Kashalkar highlight this style of learning. `http://www.gajananbuwajoshi.com/guru_gajananbuwa_joshi#taleem`, last accessed on 15 August, 2019

1.3 Objectives

Main objectives of this study are:

- curate a dataset of reference-student pairs of recordings for Hindustani Music exercises.
- improve the methods of previous systems for automatic assessment of singing in order to achieve a higher correlation of automatic grade with human annotators.
- provide a visual feedback of the student performance for students and teachers to use.
- a practical requirement was that the solutions developed should be easily integrable into the MusicCritic framework to support the online course on Kadenze platform.

1.4 Structure of the Report

We introduce the previous models developed by researchers at MTG and other relevant works in the area of musical performance assessment in chapter 2.

In the same chapter, we also discuss the performance of the baseline model developed for the assessment of Turkish Makam exercises when it was re-implemented for the Hindustani exercises data. In this preliminary study, we identify that f0-pitch-estimation, audio-to-audio alignment, dataset quality and the perceptual aspects of human grading are the issues that need to be addressed while developing a system for the assessment of Hindustani exercises.

In chapter 3, we describe the approaches used to gather a sufficiently large and consistently annotated dataset.

In chapter 4, the experiments done to identify the best possible automatic f0-pitch extraction method are described and the results presented. We conclude that al-

though many pitch detection algorithms give promising results, a single best pitch estimation scheme may not be realizable at the moment.

In chapter 5, the problem of audio-to-audio alignment using DTW algorithm is discussed. A variety of low-level features and distance metrics are tested with custom-built metrics and a manually annotated dataset to identify that a higher dimensional (120 dimensional) Harmonic Pitch Class Profile (HPCP) is the best feature of a DTW based alignment for our dataset.

The chapter 6 presents the requirements for hand-crafted features to test pitch accuracy of student singing. A novel feature 'Pitch histogram cosine distance '(PHCD) is introduced and explained. The performance of PHCD is evaluated on the basis of the mean absolute error obtained by training and testing a linear regression model with it. Similar models are trained with the features used by the baseline system and the performance is compared.

Finally, we summarize and discuss our findings in chapter 7 and present the conlusions and future work in chapter 8

Chapter 2

Background

2.1 Foundations

The germ of this thesis lies in the baseline system for singing voice assessment developed by Bozkurt et.al [2] and its successor implemented within the state-of-the-art technological framework of MusicCritic,[1] which is currently in development at the MTG-UPF. [3]. These works would be briefly discussed here as they form the foundation on which this thesis is built.

2.1.1 Baseline System

The problem of automatic assessment of singing is tackled in this baseline system by Bozkurt et. al. in following steps:

- Data Collection done through video recordings of the student auditions for acceptance at Turkish Music Conservatory at Istanbul Technical University
- Dataset preparation: Low-level representation of the data using f0-series extracted with the state-of-the-art pitch detection algorithm
- Creating pairs of reference and performance pairs of f0-series along with a 4 level grade label from multiple annotators.

[1] `https://musiccritic.upf.edu/demo/`

- Compute a histogram of the distance obtained by subtracting the reference feature from its student counterpart obtained by a DTW (Dynamic Time Warping) matching.

- Use the histogram with the dtw-cost and dtw-length-change as features to train a machine learning model to do a binary pass/fail classification.

2.1.2 Music-Critic

Music-Critic is a web-based framework to develop exercises for online courses involving music performance. It is fully compliant with the LTI standard [4] which allows it to be seamlessly integrated into all major MOOC platforms like Coursera, Edx, Kadenze etc. It provides interfaces for music teachers to create exercises, collect student performance submissions, grade the submitted performances and provide a feedback to the students. It provides interfaces for researchers to collect anonymous submissions and related data to train models for automatic evaluation of the exercises. The framework supports the grading of student performances to be done in a semi-automatic fashion wherein the automatic grade assigned by the model is verified by the concerned instructors. A qualitative feedback of the performances can be provided to the students in the form of graphical visualizations or teacher/expert comments in addition to the numerical grade. As stated previously, the framework has already been deployed to support a MOOC for Hindustani Music on Kadenze platform.

It was proposed by Bozkurt et. al. [3], that machine learning models can be trained on small subsets of performances assessed by the instructors which could then be used to do a semi or fully automatic assessment of a large number of performances.

To demonstrate this, a subset of the dataset used in the baseline system (henceforth referred as 'MAST dataset') was evaluated by 6 different human annotators and also by a machine learning approach (using 10-fold cross-validation). The mean absolute error (MAE) between the ratings of an individual annotator or the automated system and the mean rating by all annotators was used to determine the extent of inter-

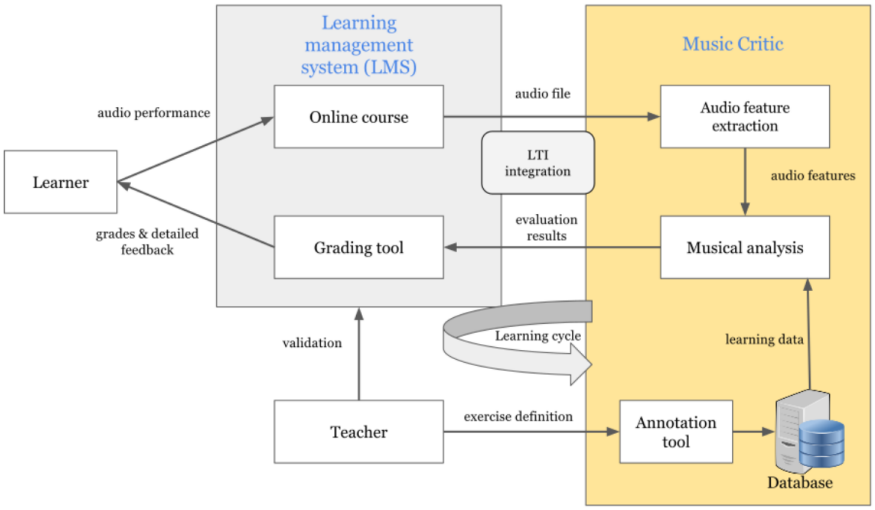

Figure 1: Music-Critic Framework Workflow

annotator agreement and understand the effectiveness of the automated system. The authors report an MAE of 0.45 for the automated system and an inter-grader MAE of 0.38.

This model of automatic singing assessment, henceforth referred as Baseline Music-Critic System (BMCS), served as the baseline for this thesis study.

2.2 State of the Art

There have been other approaches for the task of automatic evaluation of singing in different contexts. Nakano et. al. [5] was one of the early works in singing skill evaluation. This work used pitch interval accuracy and vibrato as features to train an SVM-classifier to obtain a binary (good/poor) classification with average accuracy of 83.5%. Importantly, this approach relied on singing skill evaluation without any reference or score information of what was sung.

Abeßer et. al [6] did a notable work of controlled collection of singing samples from pupils from German schools. The samples were graded based on a rubric by multiple musical experts with significant agreement. Features were extracted from the recorded samples using an automated polyphonic melody extraction algorithm. Finally SVM based classification was done to achieve average accuracy of 55.7%.

This score may appear to be on the lower side. However, this paper is the first to present a robust process pipeline of systematic data collection, annotation based on rubric, feature extraction and machine learning based classification.

Recent work by Pati K , Gururani S, et. al [7] has employed deep learning strategies for the task of music performance assessment. They use pitch contour and Mel Spectrograms of the audio recordings provided by Florida Bandmasters' Association to train fully convolutional neural networks (FCN) and convolutional recurrent neural networks (CRNN). The performance of these models is evaluated on the basis of pearson correlation (r) and determination (R^2). The authors claim that the deep learning based models show significant improvement when compared with models developed using hand-crafted features and SVMs.

Gupta C et. al. [8] have approached the automatic grading problem for singing without a reference. They hypothesize that good singing generally has only a set of dominant notes. Their grading approach relies on the finding the 'spikiness' of pitch-histograms. A spikier histogram is considered to be better singing. Their model is compared with a best-worst scaling score of human evaluations and they report an average correlation of 0.716.

Rong Gong [9] has focused on automatic singing syllable and phoneme segmentaion and detection of mispronunciation in Jingju singing.

Lerch A et. al. [10] have recently released the preprint of a detailed literature survey on Music Performance Assessment which presents a broader overview of this task. They have opined that in spite of several attempts across varied performance parameters using different methods, the important features for assessing music performances remain unclear. Many systems work well only for a selected dataset or have low accuracies for the systems to be deployed in real world scenarios. They have also highlighted that there are difficulties and costs associated with obtaining large amount of expert annotated data.

2.3 Preliminary Studies and Definition of Research Problems

Since the baseline BMCS model had been proved to be effective for exercises from Turkish Music, we started with an assumption that the same strategy could be useful for Hindustani Music exercises. To serve as our dataset, we had collected 111 human-graded singing samples from the student assignment submissions for the online Hindustani Course on Kadenze platform. The human grades indicated the overall performance accuracy on a scale of 1 to 4, with 4 denoting a flawless performance. We used the same feature extraction and the standard 10-fold cross-validation scheme used by the baseline model to train and test a linear regression model for this dataset.

The results for this process can be visualized in figures 2 and 3.

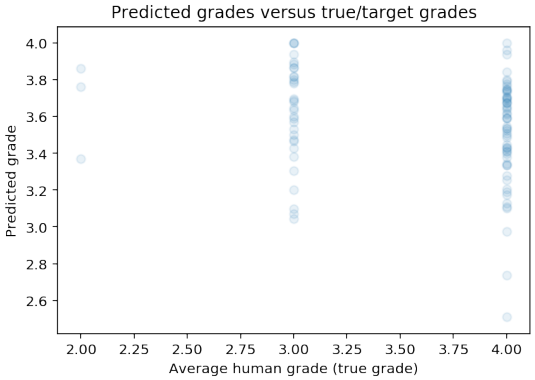

Figure 2: Predicted vs True grade

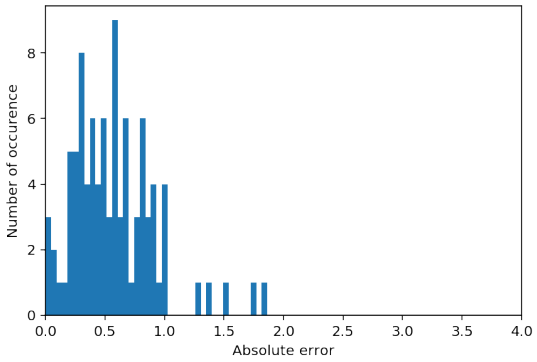

Figure 3: Grading error

2.3.1 Grading Bias

We can observe two issues in these figures:

- Human annotated grades (true grades) for almost all the samples is 3 or 4. Clearly there was a grading bias in not assigning low grades to the students. I confirmed with the course instructor Mr. Nitin Amin, that higher grades were given to encourage students to continue learning as this was an introductory course.

- The model gives an error between 0 to 1 with a mode near 0.5. The mean absolute error for this model was found to be 0.5 with a standard deviation of 0.33. This suggests that the extracted features do not correlate with the grades due to inconsistent grading and/or deficiencies in the feature extraction process for this data.

These observations suggest that we need to define more precise rubrics and address the issues related to grading biases.

2.3.2 Dataset Size

At the time of training the model, we had 111 audio files to process of which the feature extraction process could not extract the features for 9 files. These 9 files have noisy audio and it seems that the feature extraction process was not robust enough to deal with these cases. The effective dataset size of 102 samples is very small in itself. The online course expected a student to submit 3 takes for each performance. We observed that in a significant number of cases, all the three takes of a student appeared to be identical perceptually. This could mean that even in the 102 samples of our data, we potentially had some duplicates. Due to reasons beyond the scope of this study, the online course was not attracting sufficient students which could potentially provide us with a large enough dataset. Thus we found that collecting a large enough dataset for our model training would be a necessary part of solving this problem.

2.3.3 F0-Pitch Detection Errors

While the issues of grading bias and dataset size were obvious, there was also a possibility of deficiencies in the feature extraction process. At the time of training the model, noisy files could not be processed. This led me to believe that there was a need to do a complete review of the processes in the baseline method at a micro-level. The baseline method worked basically on the pitch-contour or an automatic note transcription of the student audio. However, on visualizing the pitch-contours extracted using the popular pitch detection algorithms (including the method used

in the BMCS model) for randomly selected student performances, I observed that the pitch contours were very noisy. They varied significantly from one algorithm to another. In addition, they also depended significantly on the parameters such as window size, hop size and other parameters specific to the algorithms.

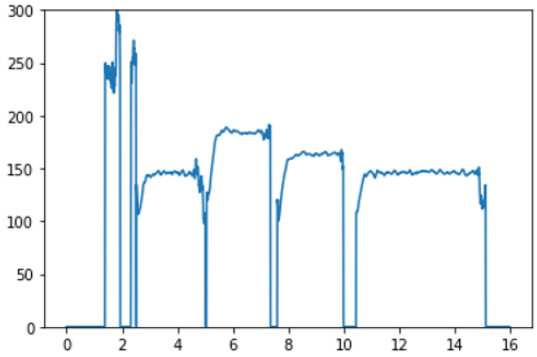

Figure 4: PredominantMelodiaMakam

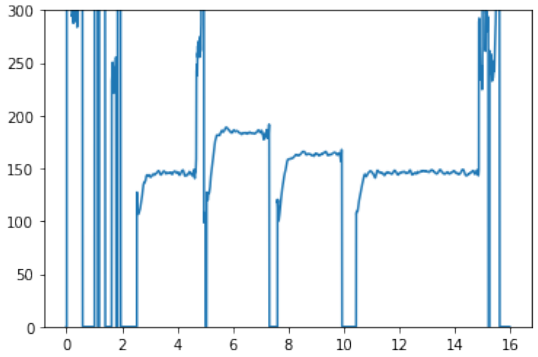

Figure 6: PredominantPitchMelodia

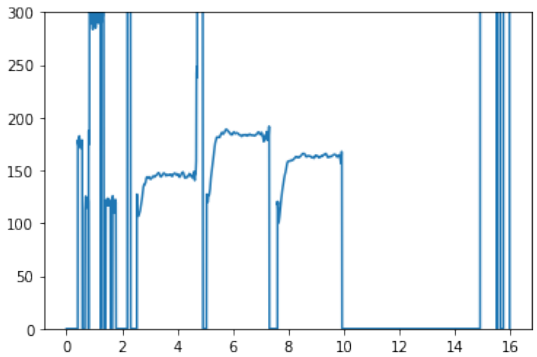

Figure 5: PitchMelodia

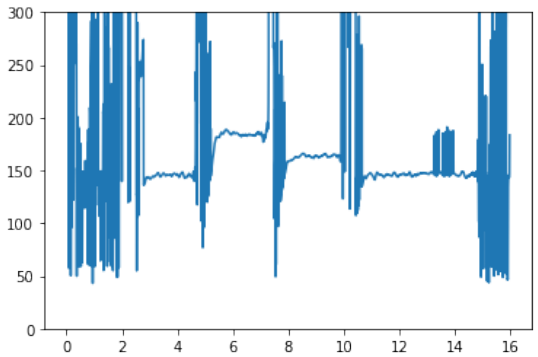

Figure 7: ProbabilisticYIN

As a representative case of this issue, we can see the different pitch-contour representations of the same audio extracted using different algorithms in figures 4, 5, 6 and 7. The audio in consideration was a fairly good performance of student singing a simple exercise having 4 notes. While PitchMelodia fails to detect the 4th note, all the other algorithms mostly agree with each other at the flat regions in the pitch-contours where we would expect to find the notes. However for the non-sung regions where we would perceive 'silence' we see that the pitch detection algorithms have output some noisy data. In the 4th note (around 14 sec), we see that ProbabilisticYIN has made some errors.

The next steps in the BMCS model, namely the audio-to-audio alignment and fea-

ture extraction were both dependent on the extracted pitch contours. This made me believe that identifying a suitable pitch detection method which worked suitably for the dataset in question was an important issue to be solved.

2.3.4 Audio-to-Audio alignment

The next step in the BMCS model is the audio-to-audio alignment using Dynamic Time Warping (DTW) algorithm with the reference and student pitch-contours as the time series to be aligned. The time aligned audio is then segmented to notes to extract note-level features. Audio alignment errors are to be expected if we use the erroneous pitch contours as the input time series for the DTW. This was confirmed by visualization of the audio alignments for randomly selected performances.

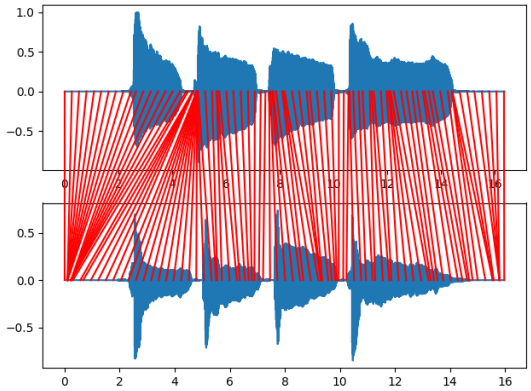

Figure 8: Alignment error due to erroneous pitch contours

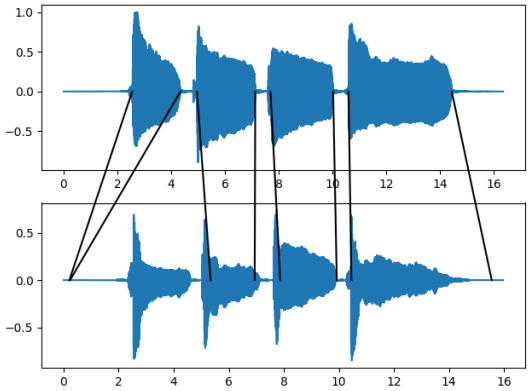

Figure 9: Alignment Error leads to incorrect note-segmentaion

Figure 8 is an example of erroneous pitch contour leading to alignment errors. The red lines show the mapping between the reference and student performance. Clearly the first note in the reference has got aligneed to the initial silence in the student version. This has led to incorrect note-segmentation as can be seen in figure 9. Erroneous note segmentation would lead to errors while doing note-level comparison in the extracted features.

If we are able to reliably and accurately obtain the pitch-contours, the audio alignment problem would be solved automatically. However, the student audio files could be noisy. As mentioned in the previous section, the pitch estimation from the state

of the art algorithms is sensitive to the input parameters. It is thus quite likely that a fool-proof pitch detection algorithm may not be a realized at the moment. So we decided to treat the audio-alignment as a separate problem independent from the pitch detection problem.

Thus in the preliminary investigation of the baseline method we found the four issues namely grading bias, dataset size, pitch detection errors and audio-to-audio alignment as the principle research problems to work on. We will now discuss the approaches we selected to tackle these issues in the next chapters. A short state-of-the-art review is done for each sub problem that we discuss at the start of the chapters.

Chapter 3

Datasets and Grade Annotations

3.1 Dataset

For the reasons stated in the previous chapter, we needed to obtain a large enough dataset to train our model. Some freely available singing datasets like MAST-melody_dataset [2] and Kara1k dataset [11] have extracted features of reference-student or origin-cover respectively. However, these do not contain raw audio data. From their description it appears that the recordings were done with good quality microphones. There being no precedent to an automatic singing assessment system developed to grade singing exercises in a MOOC for Hindustani music, we were skeptical of the generalization potential of any model developed using an unrelated dataset. Hence, it was deemed necessary to collect an 'a capella' dataset of reference and student performances, preferably recorded through an online portal using readily available devices such as smartphones or laptops to represent the real use case scenario.

For this purpose, within the MusicCritic framework, we created a web application [1] to allow creation of courses which allowed anonymous submissions of exercises from anyone. A subset of five exercises present in the Kadenze course was ported to this application and publicly presented as a website link. The five exercises presented

[1] `https://staging.musiccritic.upf.edu/training/hindustani_training/contribute`

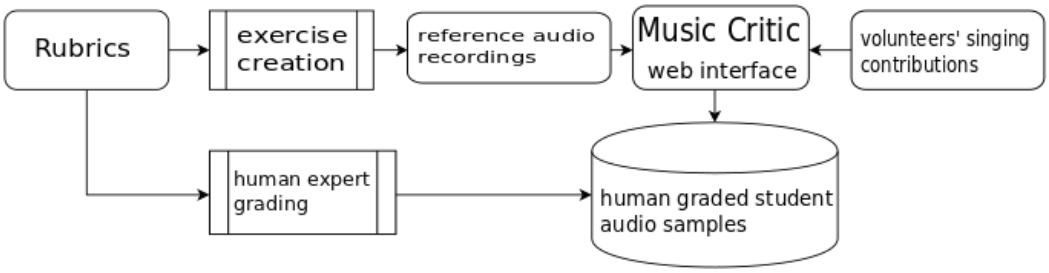

Figure 10: Dataset creation process

were arranged in the increasing order of difficulty. Each exercise had instructions based on the rubrics described in the next section. The contributors were asked to listen to a reference melody with a backing track of tanpura and tempo track.

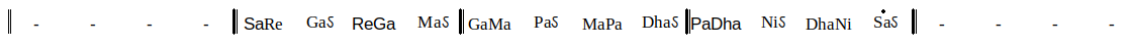

Figure 11: prescriptive transcription of an exercise

The Kadenze course did not provide any prescriptive transcription for the reference melodies in the exercises as the melodies were musical phrases taught in the course videos and practice sessions and were not new to the students. Also, we believe that the students who enrolled for the course were motivated enough to listen to the reference multiple times, if required, to memorize the melody. This would not be the case with the volunteer contributors whom we expected to approach. Hence, a prescriptive transcription of the reference melody was provided to help the contributors memorize the melody.

3.2 Grade Annotations

Grading of musical exercises is a subjective task. Inconsistencies in grading arise due to two main reasons. First, the lack of a well defined rubric to enable an objective grading of the student performances. And second, the perceptual limitations of a human grader. Our work to reduce the grading inconsistencies due to both these aspects is summarized here.

3.2.1 Rubrics definition

Effectively constructed rubrics can enhance the experience of learning music while helping the music teachers and students know where they stand. DeLuca, C et. al. [12] present three types of music performance rubrics namely discrete-component rubric, integrated-component rubric, and self-reflection rubric. Discrete component rubrics measure one individual component of music performance by specifying a descriptive criteria while integrated-control rubrics attempt to evaluate the holistic effect of the individual components. The authors claim that self-reflection is an important part of learning process and self reflection rubrics are designed to evaluate how well a student is able to self-reflect on her performance. Wesolowski, B. C.[13] claims that analytic rubrics which try to evaluate complex concepts by breaking them into simple components can provide valuable feedback to the student for improvement than holistic rubrics which evaluate the overall effect of a performance. He recommends the usage of a four level scale with 1: beginning, 2: developing, 3: accomplished, 4: exemplary.

We took inspiration from these and used pitch and timing as our parameters for the rubrics. These two parameters are independent of each other and objective enough for human evaluators across musical cultures to arrive at a uniform definition of correct and incorrect. We use the 4 point scale as 1: Large errors or fail 2: Major errors 3: Minor errors 4: Excellent. This scale is used as it is simple for human annotators to understand and hence expected to produce a large inter-evaluator agreement. Also this choice of scale does not allow a 'middle' grade thus forcing the evaluators to assign a grade as above average or below average which would be beneficial to train a model to understand as what leads to a high or low score. For the purpose of this study, however, we focused only on the pitch accuracy rubric.

3.2.2 Grader interface

With an experience of learning and performing Hindustani Classical music since more than 10 years, I feel that I am well qualified to assign grade labels to all the

student submissions that we collected. Initially, I attempted to grade the entire dataset purely by listening to the audio. My experience of grading can be summarized as follows:

- It is easy to identify both a flawless performance(grade 4) or a performance with large errors (grade 1).It is however difficult to judge the intermediate performances.
- At times a large error in a single note or a small percentage of notes made me penalize the performance more than necessary. An objective judgement based on a rough estimate of percentage of the notes correctly sung and the nature of errors in the incorrect notes was difficult to obtain.
- An error made at the beginning of a performance tends to get penalized more than an error done at the end.
- The emotional state of mind at the time of grading may influence the grade.
- At times it was difficult to maintain a rubric level focus. For example, we may not perceive small pitch errors if the voice timbre is pleasant. Hence the grading of pitch rubric is influenced by an unrelated voice timbre rubric.

These experiences highlighted the perceptual limitations of a human grader in giving consistent grading. Hence, a need for having a visual feedback as an aid for grading was felt. I developed a GUI grader tool as shown in figure 12 which provided a visual feedback of note-level pitch accuracy of student performances. The accurate note-segments are indicated by green color while inaccuracies are in grey. Partially correct notes will have a linearly varying color between green to grey based on the note accuracy score. The note accuracy score is a rule based score dependent on the pitch histogram cosine distance feature which we will discuss further.

There can be a critical view that possible inaccuracies in the visual feedback may influence the grader's perception of the audio and thus bias the grades in its own way. In my capacity as the grader, I can confirm that all effort was taken to ensure

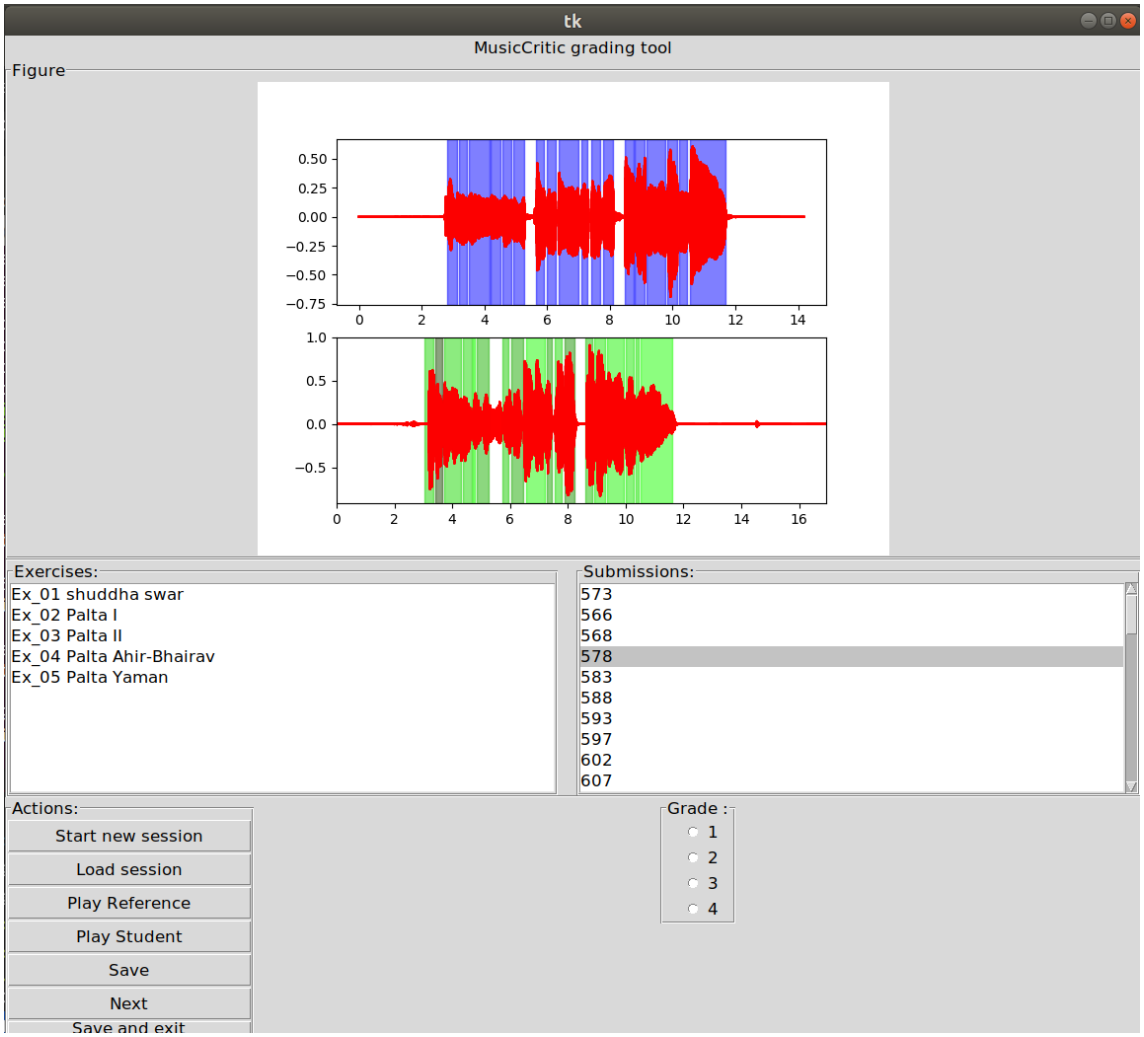

Figure 12: Grader interface

that the visual feedback was used purely as an aid to identify and quantify the singing errors. If there seemed like a disagreement between the visual feedback and the audio, grade was assigned purely based on listening the audio. Using the grader tool, a grade report was generated as a csv file containing the unique sound id of the student performance and the grade.

3.3 Datasets description

As an outcome of the dataset collection activity conducted with the help of MusicCritic framework, we collected two datasets which we will refer as HT_MTG and SSV datasets. HT_MTG dataset contains 130 valid student submissions with

Grade	HT_MTG	SSV
1	12	55
2	27	50
3	33	55
4	59	49

Table 1: Dataset grade distribution

major contributions from MTG-UPF researchers. The SSV dataset contains 210 valid submissions with all the contributions from Sargam Sangeet Vidyalay, a music school at Pune, India. The distribution of the grade classes from 1 to 4 for both the datasets can be seen in table 1.

It can be seen that SSV is almost a balanced dataset while HT_MTG has a larger proportion of good performances over poor ones. Additionally, the MAST dataset used to train the BMCS was also used to perform tests to evaluate the generalisability of the model developed.

Chapter 4

Automatic F0-Pitch detection

Automatic F0-pitch detection has been an actively researched topic in Music Information Retrieval and there are some approaches that achieve good performances with both monophonic and poly-phonic signals. Prominent approaches for pitch detection can be classified into three different classes: time-domain, frequency-domain and data-driven approaches. The time-domain approaches use an auto-correlation function to determine the periodicity of the audio signal which is used to infer the fundamental frequency. The YIN algorithm, developed by De CheveignÃĺ et al. [14] is one of the oldest time domain approach for pitch detection. pYIN is a probabilistic version of YIN presented by Mauch and Dixon [15]. In the frequency domain, PredominantPitchMelodia developed by Salamon and Gomez [16] works on polyphonic audio to determine the pitch contours of the most prominent melody. PredominantMelodyMAKAM by Atli, H. S. et. al. [17] is based on adaptations and post-processing approaches done on PredominantPitchMelodia to suit Turkish MAKAM music. Among data-driven approaches, CREPE developed by Kim JW , et. al [18] is a deep Convolutional Neural Network (CNN) based algorithm which operates on the time domain waveform.

Pitch being a prominent grading rubric for singing evaluation, accurate pitch detection is fundamental for an automatic singing assessment system. This is even more true for a system such as BMCS which uses the pitch-contours for audio alignment

and segmentation in addition to feature extraction. The focus of our work was not to find the best pitch detection, but to identify the method and parameters which work best for our use case which is the singing human voice. A fundamental issue for comparison of different pitch detection algorithms is the availability of a dataset with ground truth pitch labels to compare with the pitch estimates. To test which algorithms work best for the singing human voice, we tested the several prominent ones using the Ikala dataset [19] which contains 252 audio samples of 'a capella' singing each of 30 sec duration and their annotated pitch values. We compared the pitch es-

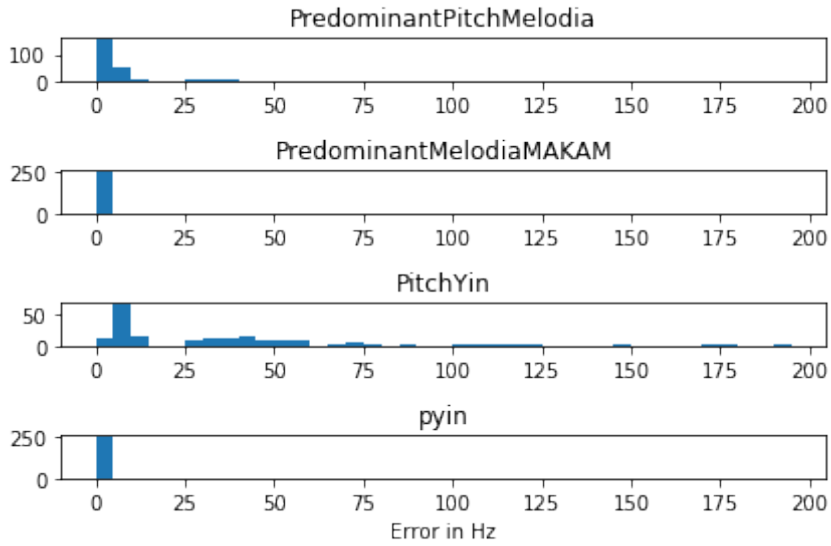

Figure 13: Comparison of prominent pitch detection algorithms

timates by PredominantPitchMelodia, PredominantMelodyMAKAM, PitchYin and pYIN on the basis of the Mean Absolute Error (MAE) for each audio file. CREPE was not included in the tests as we observed that it required very long execution times. For the pYIN extraction we used the vamp plugin implementation [20] while for the other algorithms, we used their Essentia [21] implementations with window size of 2048 and a hop size of 128 samples as the parameters. The results can be visualized in the histogram plots of the MAEs in figure 13.

These results suggest that, while all algorithms typically have low errors, Pitch Makam and pYIN are the better performers for this task. It should however be noted that the Ikala dataset has good quality recordings, also the annotations have been done where the pitch is mostly stable. The student submissions in our dataset

are recorded using poor quality devices and acoustic environments. Hence, it cannot be assumed that the same results will be true for our datasets. Also, the MAE metric may mask small number of large errors. But these particular errors may lead to improper audio alignment if the pitch contours are used for the DTW based alignment as in the BMCS. While these results help to identify a suitable pitch detection algorithm, they do not completely solve the issues we discussed in the previous chapter, especially the erroneous audio alignment as a result of the pitch detection errors.

The conclusion I drew from these experiments is to accept that automatic pitch detection may not be perfect and we need to find a way such that any reasonably accurate pitch detection will help us to develop the singing evaluation system, thus making the choice of a pitch detection algorithm practically redundant.

Nevertheless, the vamp plugin implementation of 'pYIN smoothedpitchtrack' was used for all pitch contour extractions in the further study, since it was one of the best performers in the tests conducted on Ikala dataset and also was empirically found to have reasonably high accuracy on the datasets we collected.

Chapter 5

Audio to Audio alignment

Since the evaluation rubrics assess pitch and tempo accuracy independently, it is important to develop a model that does not penalize one for the inaccuracies in the other. Very often music students can be accurate in pitch but make errors in timing or duration of notes. This problem can be overcome by aligning the corresponding notes from reference to the student version. The BMCS model uses Dynamic Time Warping (DTW) algorithm to align the pitch contours of reference and student versions. We inspected the audio alignment performance of the baseline for a small subset of our dataset and found that there could be major errors in alignment. While DTW has been used successfully for a variety of time series alignment tasks, a good alignment can be only obtained if there exists some similarity in the two series to be aligned. The main reason for the audio alignment to fail with the baseline system was identified to be the inaccuracy in the pitch contours obtained by the automatic pitch detection algorithm. We decided to improve the audio alignment by using a three-pronged strategy, 1) Investigate different pitch detection algorithms and pitch contour post-processing strategies to obtain 'accurate' pitch contours. 2) Investigate the usage of other low level features such as MFCC, HPCP, energy envelope, etc. instead of pitch contours to obtain a better alignment performance. 3) Understand the DTW algorithm and its variants to find the appropriate metric to determine the 'similarity' between the reference and student performances.

We have discussed the issues related to pitch detection in the previous chapter and concluded that a solution which does not rely on pitch detection for audio alignment may be necessary. We will discuss the DTW algorithm and the Investigation of features for alignment in the following sections.

5.1 Dynamic Time Warping (DTW)

Dynamic Time Warping [22] is an algorithm used to optimally find a non-linear alignment between two discrete time sequences. Let us assume two sequences, $X = [x_1, x_2, ...x_n]$ and $Y = [y_1, y_2, ..., y_m]$, where every point x_i and y_j in X and Y could be a scalar value (eg. f0-pitch) or a multi-dimensional feature vector (eg. HPCP, MFCC). In general the two sequences are expected to be similar but may be misaligned in time. In our topic of research this could mean that the reference and student versions of the singing performance may have the same notes sung in a different tempo or a variable tempo. Our goal then is to find a mapping between the points which are similar to each other. Intuitively this means that at any pair of time instants, x_i and y_j, we have to make a choice whether to move forward in X, Y or both. A suitable distance metric is chosen to evaluate point-wise distance between any two points. Commonly used metrics are the euclidean and cosine distances. The distance metric should typically evaluate small values for 'closer' or similar points and large values for dissimilar ones. Using the chosen distance metric, a cumulative cost-matrix is constructed to reflect the smallest cost to arrive at any point starting from $[x_1, y_1]$. A backtracking algorithm then traverses the optimal path from $[x_n, y_m]$ to $[x_1, y_1]$ to give us the alignment path which is our desired output.

Another important aspect of the DTW is that its time as well as space complexity is $O(N^2)$. So using the DTW to align long sequences would lead to very long processing times and large memory requirement. Raw audio data for example, typically has 44100 samples per second of audio. To align two 1 second long raw audio sample data with DTW, we would require our cost matrix of 44100×44100 dimension. The memory requirement to store this matrix would be of the order of gigabytes. Hence, a compressed representation of the audio with a suitably large hop-size would be

necessary for using the DTW on a readily available personal computer. The other, arguably more important reason, for not using raw audio is that we want to align the two sequences on the basis of perceptual similarity. So a pre-processing of the audio to features which preserve the similarity and get rid of the noise in raw audio data would improve the alignment performance.

It is commonly found that optimal alignment paths rarely deviate far away from the principal diagonal of the cost matrix. Taking advantage of this fact, methods such as Sakoe-Chuba band [23] and Itakura Parallelogram [24] attempt to reduce the execution time of the DTW by limiting the number of cells that are evaluated in the cost matrix. Salvador and Chan [25] attempt use approximation techniques to implement the DTW with linear completely (O(N)). All these methods however are approximations of the DTW and there is a possibility of a sub-optimal alignment path when they are used. The reference and student performances that we intend to align are typically of short durations (less than 15 sec). Using suitable features and hopsize, it was easily possible to execute the DTW in less than a second for a reference-student pair. Hence, we chose to use the standard DTW implementation for our experiments and models.

5.2 Investigation of low-level features for audio alignment

While understanding the DTW process, we realize that there are two main decisions to be made:

- Which features should be used for audio to audio alignment?
- What distance metric should be used to evaluate the cost matrix?

Muller [22] suggests use of chroma features in combination with cosine distance for alignment in tonal similarity based alignment tasks. An exploration of different features for audio alignment has been done by Kirchoff and Lerch [26]. Their

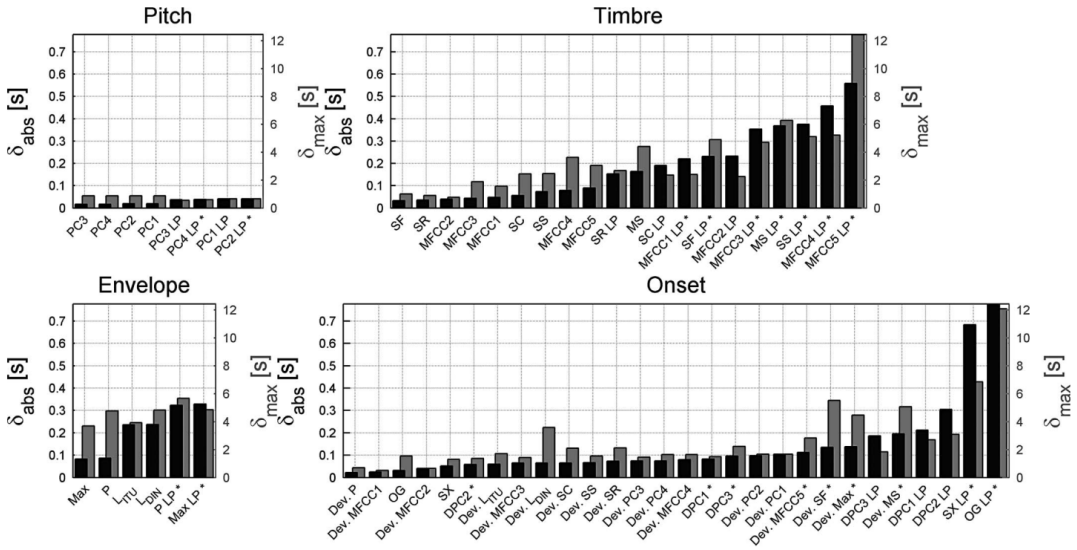

Fig. 5. Synchronization results of individual features: the left ordinate displays the mean absolute deviation of each feature (black bars), the right ordinate displays the maximum deviation (grey bars).

Figure 14: Exploration of features for audio alignment by Kirchoff and Lerch [26]

results summarized in figure 14 also suggest that pitch chroma based features, abbreviated as PC in the figure, show the least deviation from ground truth and are the most suitable features for tonal similarity based alignment. Their method used synthesized samples generated using a variety of time stretching and noise additions applied to the original audio. Thus it is exploring the best features for aligning the two inherently identical audio samples, where one is a synthetically modified copy of another. A benefit of using synthetic transformations is that the ideal DTW alignment path is known beforehand, which serves as the ground-truth. The metrics that they use to evaluate the audio alignment performance are based on the deviation from the ideal path.

Our use case involves aligning two inherently distinct audio samples which have been sung by different people but have some musical similarity. This presents a different challenge to define the metrics and methods for audio alignment. Additionally, the audio samples we work with are not commercial quality recordings and frequently consist of artifacts like ambient noises, reverberations, etc. Hence, we proposed and performed our own experiments to explore different feature and distance metric combinations for audio alignment.

5.2.1 Audio alignment experiments

Dataset:

For these experiments, I manually annotated note onset and offsets for a subset of 60 student samples from the SSV dataset. Similar annotations were done for the reference recordings as well. These 60 reference-student audio pairs along with their manually annotated note onset-offsets served as the dataset for our audio alignment experiments.

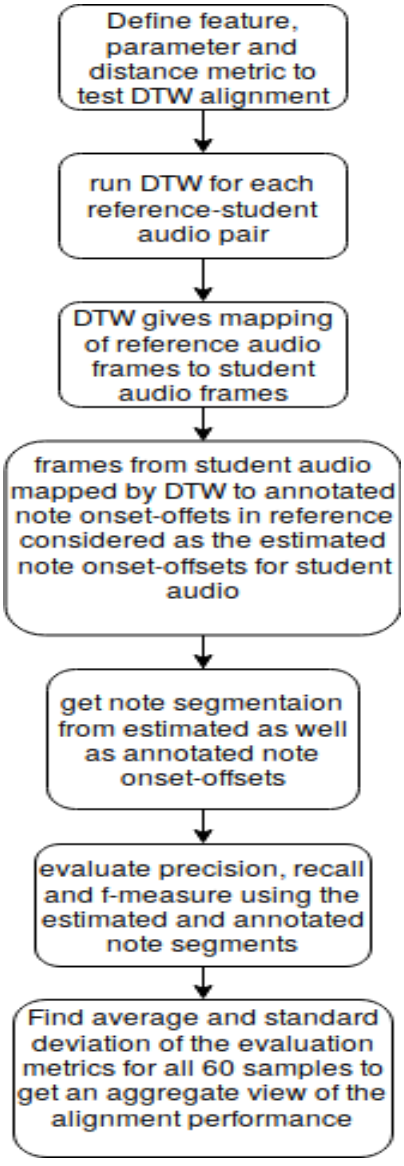

Figure 15: Audio alignment experiment methodology

Methodology: Multiple feature-parameters and distance metrics were tested for their alignment performance on this dataset with the steps explained below.

To perform the experiment, we first define the feature, distance metric and parameters and run the DTW for each reference-student audio pair. DTW gives a mapping between reference and student audio frames. The frames from the student audio that are mapped by the DTW to the annotated note onset-offsets from the reference audio are considered as the estimated note onset-offsets. The region between a note onset and offset was considered as a note segment. Note segments were obtained from the annotated and estimated onset-offsets for each of the 60 samples. Note segments from the annotated onset-offsets served as the ground truth with which those obtained from the estimated onset-offsets were compared. We defined 'overlap' as the length of the region common between the estimated and annotated note segments.

For the purpose of evaluation the following metrics were proposed :

- precision = length of overlap / length of estimated segment
- recall = length of overlap / length of annotated segment
- f-measure = Harmonic mean of precision and recall

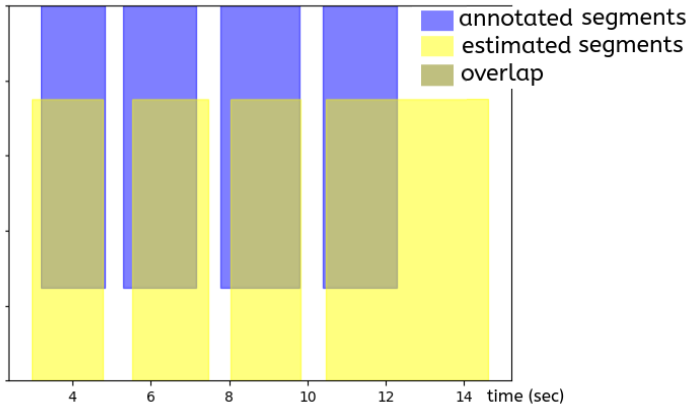

Figure 16: Alignment evaluation metrics

Parameters:

A 'hanning' window with size = 4096 samples and hop size = 512 samples was used for all analysis tasks. Additionally, special features were defined as:

'short-time-energy' : the total energy for a windowed frame.

'energy_mask ' : '1' for a frame with short-time-energy greater than 10% of the mean short-time-energy of the entire audio, otherwise '0'.

For the feature extraction of standard features (eg. MFCC, HPCP [27]) I used the essentia [21] implementations with the specified parameters.

F0-pitch and nnls-chroma extraction was done using their vamp plugin implementations namely pyin:smoothedpitchtrack [15] and nnls-chroma:chroma [28]. All special features defined previously are extracted as per their definitions.

The results for each set of feature, parameters and distance metric tested are summarized with their mean and standard deviation and presented in Table 2.

Feature_params dist_metric	Avg precision	Avg recall	Avg f-measure	Stdev precision	Stdev recall	Stdev f-measure
energy_euclidean	0.3205	0.3108	0.3078	0.3391	0.3168	0.3287
energy_mask_euclidean	0.2508	0.2984	0.2661	0.2694	0.2697	0.2695
hpcp_12_cosine	0.8241	0.6284	0.7065	0.1972	0.1883	0.1800
hpcp_12_euclidean	0.8435	0.7396	0.7843	0.1569	0.1651	0.1524
hpcp_120_cosine	0.8288	0.7154	0.7631	0.1811	0.1902	0.1805
hpcp_120_euclidean	0.8137	0.8155	0.8106	0.1223	0.1237	0.1091
hpcp_240_euclidean	0.8138	0.8152	0.8106	0.1223	0.1234	0.1091
hpcp_48_cosine	0.8266	0.7123	0.7605	0.1808	0.1878	0.1792
hpcp_48_euclidean	0.8131	0.8138	0.8095	0.1175	0.1196	0.1040
mfcc_cosine	0.5122	0.5082	0.4950	0.3173	0.2740	0.3019
mfcc_euclidean	0.7323	0.7477	0.7322	0.2414	0.2108	0.2226
nnls_chroma_cosine	0.7328	0.6570	0.6873	0.2549	0.2365	0.2418
nnls_chroma_euclidean	0.7945	0.7395	0.7638	0.2439	0.2232	0.2320
pitch_euclidean	0.6035	0.5332	0.5560	0.3055	0.2332	0.2658

Table 2: Alignment result: Individual features

Results: A representative case highlighting the importance of feature used in DTW is shown in Figs. 17 and 18, where we have an inaccurate alignment with pitch feature, while an accurate alignment with 120-dim HPCP feature used for the DTW, respectively. The red lines show the mapping of reference to student audio while the black lines show the mapping of onset and offsets in reference to the student audio, which we use to determine the estimated note-segments.

From Table 2, we observe that chroma based features e.g. nnls_chroma and hpcp generally perform better than pitch and mfcc for this alignment task. Cosine and euclidean metrics have comparable performance for precision, but cosine typically has a lower recall score. Thus, cosine distance metric tends to precisely detect a small portion of the whole note sung by the student, however it leaves out a significant portion of the human annotated note. This is due to the tendency of cosine distance to cause pathological warping. Pathological warping is a common problem observed with DTW, where large number of frames from one series are mapped to the same frame in the second series. An example of this problem is the first note in figure 17.

Our context involves beginner students' performances which may not be sung accurate in pitch. The alignment-feature and distance metric should, therefore, look

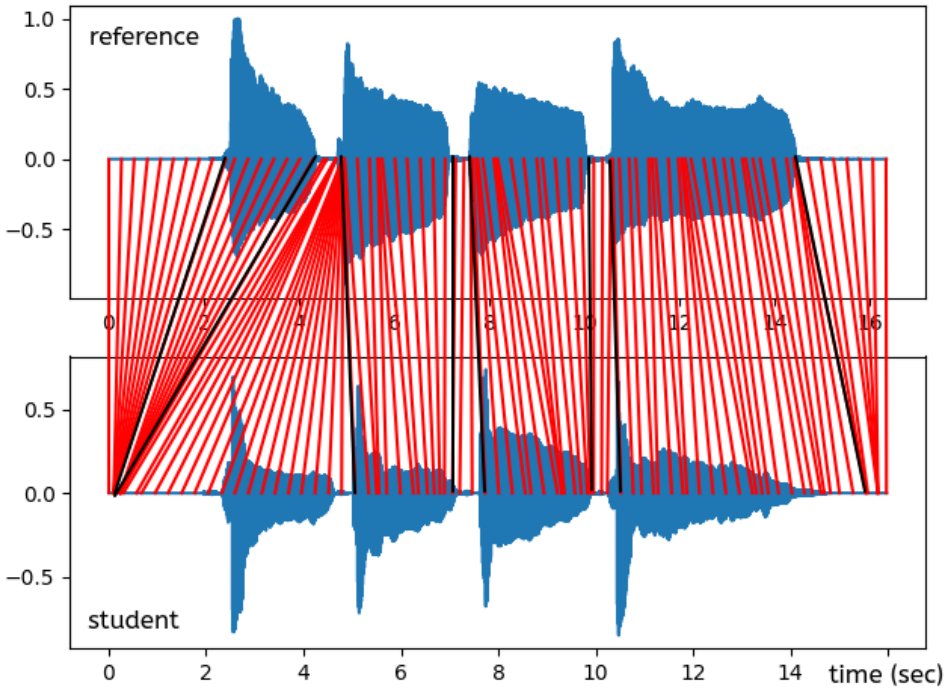

Figure 17: Incorrect alignment and segmentation with pitch contours

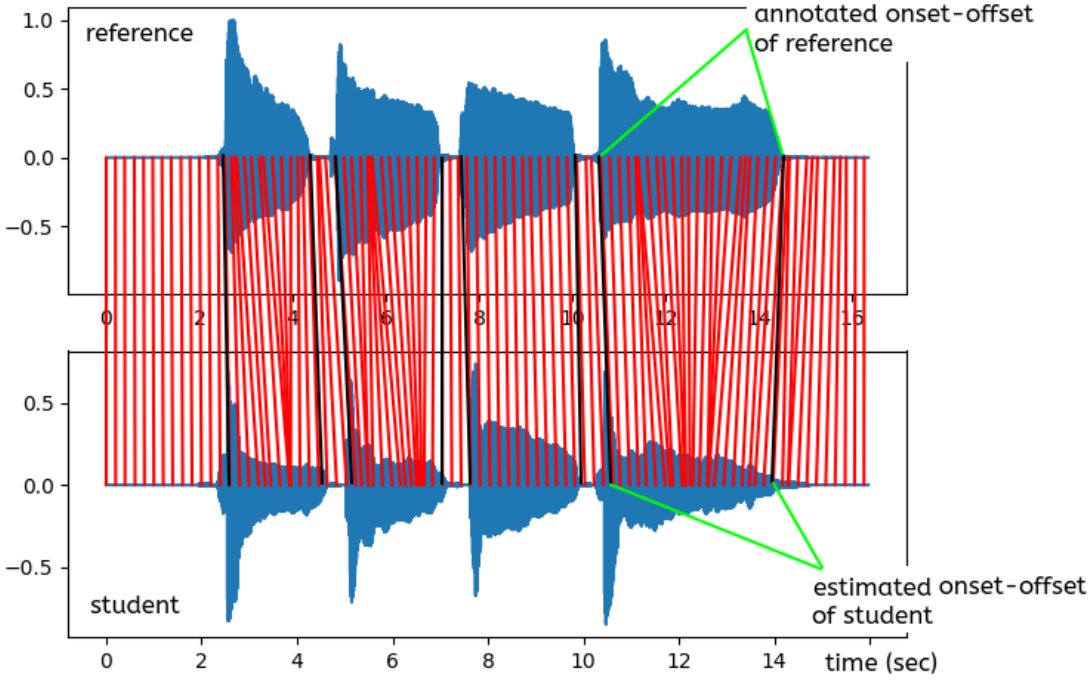

Figure 18: Improved alignment and segmentation with 120-dim HPCP

for a 'coarse' similarity between the audio frames. Euclidean distance appears to be doing that better than cosine.

An important revelation is that using a higher dimensional HPCP, upto a limit, improves the alignment performance. It can be seen in Table 2, that when we use 12, 48, 120 and 240 dimensional HPCPs, the average f-measures are 0.7843, 0.8095, 0.8106 and 0.8106 respectively. Most chroma features represent the audio in 12 chroma bins, with the resolution of one bin per semitone. Higher dimensional HPCPs, however, have more bins per semitone. So a 120 dimensional HPCP will have 120 bins with a bin resolution of $\frac{1}{10}$ semitone.

5.2.2 Multi-feature DTW Alignment

A single feature used to represent the audio may lead to over-compression or noisy representation of the audio which may lead to improper alignment. A novel approach was tried out to combine information from multiple features to generate the DTW path. To implement this, a modular DTW implementation was done which could evaluate multiple cost-matrices using different features and then a linear combination of the cost-matrices would be used to determine the path.

Typically, it was observed by visual inspection of figures like Fig. 18, that silence from one audio was being aligned with voiced region from another in certain cases. To rectify these kind of errors, we tested our multi-feature alignment approach by combining the best performing individual feature i.e. 120-dimensional hpcp in combination with a varying percentage (5, 10 or 15 percent) of energy_mask feature. The idea was to have an additional cost if silence is mapped with voice, thus deterring their alignment with each other. The results for these are presented in Table 3.

When compared with pure HPCP based alignment, we see that a 5% weight to the energy mask improves alignment performance (f-measure) by about 1%, and decreases standard deviation by about 2%, thus improving the reliability of the alignment. This improvement, however, was mainly due to the improved alignment in one particular sample where f-measure increased from 0.32 to 0.84. Nevertheless,

Features_combined dist_metric percent_energy_mask	**avg precision**	**avg recall**	**avg f-measure**	**stdev precision**	**stdev recall**	**stdev f-measure**
120-dim hpcp + energy_mask euclidean 5%	0.8259	0.8258	0.8218	0.1033	0.1042	0.0867
120-dim hpcp + energy_mask euclidean 10%	0.8189	0.8148	0.8131	0.1151	0.1159	0.1017
120-dim hpcp + energy_mask euclidean 15%	0.8132	0.8068	0.8065	0.1273	0.1231	0.1141

Table 3: Alignment Results: Multi-Feature DTW

we observed small improvements (upto 4% increase in f-measure) in alignment in all other cases except one sample where alignment f-measure decreased from 0.85 to 0.84.

Chapter 6

Feature Extraction and Model Training

Once the a good alignment between the two audio signals is accomplished, a good set of features is required to depict the reality of the student performance with respect to the reference performance. The rubrics chosen for the performance assessment are pitch and timing accuracy. This study is however focused on evaluating pitch accuracy of student singing.

For estimating the pitch accuracy, the BMCS system uses the absolute error in pitch between the aligned reference and student note segments to obtain statistical aggregate high-level features such as mean absolute error, mean error and centre of mass of error histogram. We will refer to these features as the 'pitch_error_stats' features. These statistical features are used to train a linear model to predict the student grade. A noteworthy point is that any feature based on error between the reference and student pitch should ideally be negatively correlated with the predicted grade since a larger pitch error should correspond to a lower grade. For a linear regression model, this means that the successful features would be those with significant negative weights.

Vidwans A. et al. [29] present a survey of various features and their effectiveness to evaluate music performance with respect to specific rubrics. They have presented

note steadiness, average pitch accuracy and percentage of in-tune notes as the suitable pitch based evaluation features. Similar to the BMCS features, these features are also statistical aggregation of the pitch error between the student performance and an ideal.

Two factors are important while extracting pitch accuracy related features:

- **Perceptual Factors:** What we detect using pitch detection algorithms is the f0-fundamental frequency, while pitch is a perceptual phenomenon. We need to be aware of the perceptual aspects which will influence the human judgement of sung pitch, while we create features for estimating pitch accuracy. For example:
  - For an accurate estimation of the pitch by the human brain, the note has to be stable for some duration [30]. Put in other words, our pitch estimate is an 'aggregate' estimate for a small time duration rather than an estimate for a time instant. This duration may be more or less depending on the musical training and abilities of a person. So, if we were to listen to a note with vibrato, it would be difficult to estimate the amount of vibrato although we may identify the base note. On the perceptual level, the absolute difference between the base note and instantaneous f0-pitch may be insignificant.
  - While judging the pitch, we also tend to put more emphasis on the sustained note than the attack and decay where perceptual pitch is ill-defined. So numerical errors in f0-pitch in the attack-decay regions may be a noise while determining the note accuracy.
  - The just noticeable difference (JND) for pitch is typically found to be between 5-10 cents [31]. Numerical f0-differences within the JND is simply noise at the perceptual level.
  - While judging pitch accuracy, all errors beyond a threshold would be perceived as large errors and the note would be classified as grossly incorrect. The magnitude of the error would be insignificant here. Using absolute

error in f0-pitch would lead to differential treatment of large errors when perceptually they are treated same.

- **Sensitivity to Algorithmic Errors:**The pitch detection algorithm may sometimes output noisy values due to various reasons including but not limited to noise in the audio, inherent algorithm limitations or the pitch itself being ill-defined during attack-decay phases. Any frame-wise errors calculated for noisy pitch data would essentially lead to noise in the extracted features. Errors in audio alignment may lead to a portion of one note to get aligned to its adjacent note. Frame-wise errors calculated for misaligned parts would also lead to noise in the features.

Considering these factors, I designed a new feature, 'pitch-histogram cosine similarity', to extract note level pitch accuracy information. This feature and its usefulness is explained in the next section.

6.1 Pitch-Histogram Cosine Distance (PHCD)

The idea behind this feature is to assume the entire note-segment happens in a single unit of time, the logic being that as humans, we would be able to perceive if a whole note was accurately sung or not, with less regard to its finer temporal pitch variations. This is more true for short note segments with almost flat pitch contours. To calculate the pitch-histogram cosine distance (PHCD) for a note, we first evaluate the histogram of reference and student pitch values in cents. Imagining the reference and student note histograms as multi-dimensional vectors, we evaluate the cosine distance between them which gives us a measure of how aligned the histograms are. We use the histograms with bin sizes of 100, 50, 20 and 10 cents to determine the pitch accuracy with varying tolerances. The idea is that two notes may be close to each other within semi-tone level accuracy, but they may have inaccuracies at the half-semitone level. This would be reflected by a low distance for bin size = 100 cents but a high value for bin size = 50 cents. The note-level features can be aggregated for the whole performance to assign an overall grade for the performance.

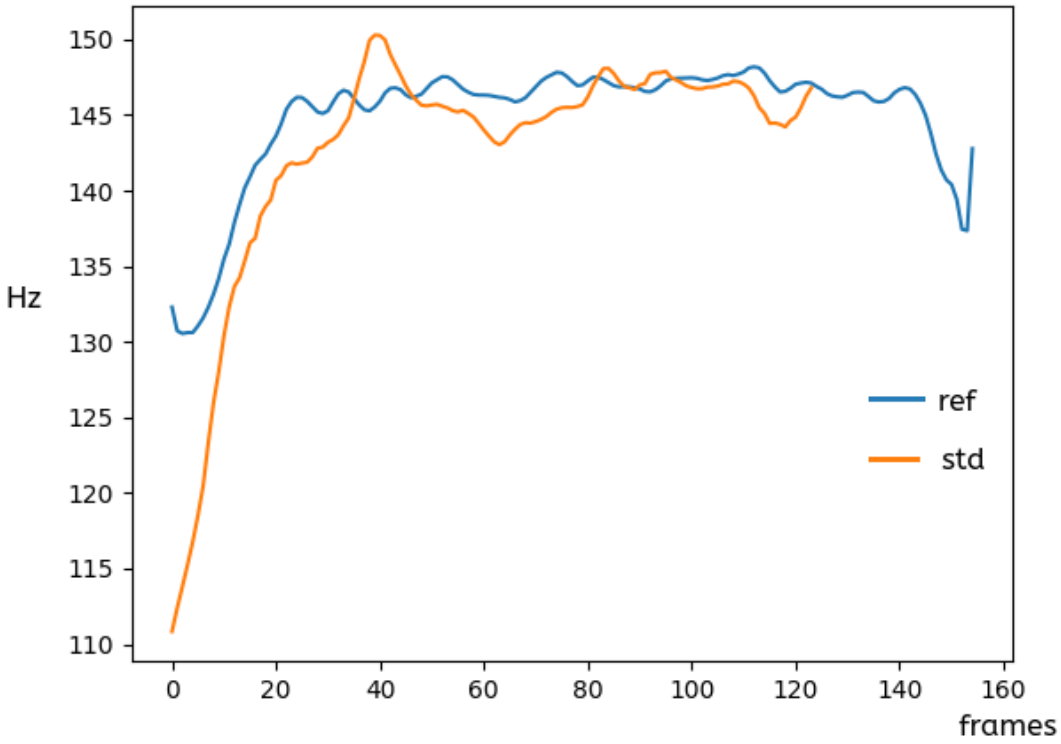

Figure 19: Pitch contours for a note segment

For our implementation, we have used a weighted sum with weights proportional to the length of the reference note-segments as the aggregation policy, the idea being that longer notes should get more weight in the final score. Note level PHCD by definition would lie between 0 and 1 since the pitch histogram vectors have all non-negative components. The weighted sum approach for aggregation of the note-level PHCD to obtain the performance level PHCD would preserve this property. Like other error measures, PHCD is also an error measure between the reference and student pitch. Hence, it is expected to have negative correlation with performance accuracy or student grade.

To understand the working of the PHCD feature, let's consider the pitch contour in figure 19 as an example. It can be observed that reference and student pitch contours are of unequal lengths. This may happen due to the student making a duration error or at times due to audio alignment method detecting a shorter or a longer note segment. To evaluate absolute pitch error, the two pitch contours have

to be converted to equal lengths by interpolation. This may introduce additional noise in the pitch contours. An advantage of PHCD is that no such interpolation is necessary for its evaluation.

The note in figure 19 has been sung fairly accurately, however, there is a large pitch error at the attack. This is more likely an error due to inaccurate pitch estimation than an actual error in singing. But the error will have a large contribution to pitch error based features. The pitch histograms with bin sizes of 100 and 50 cents can be visualized in figures 20 and 21 respectively. At the 100 cents resolution, the student and reference pitch histograms are almost identical. This will lead to a low, close to zero, cosine distance if we consider the pitch histogram bin values as the components of a vector. At the 50 cent resolution however, a significant portion of the student pitch values are classified into an adjacent bin. So the cosine distance between them at the 50 cent resolution will be higher.

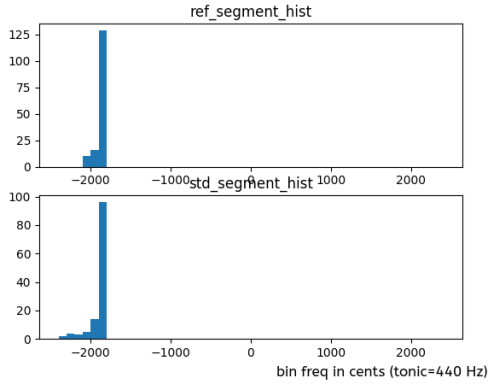

Figure 20: pitch histogram, binsize = 100 cents

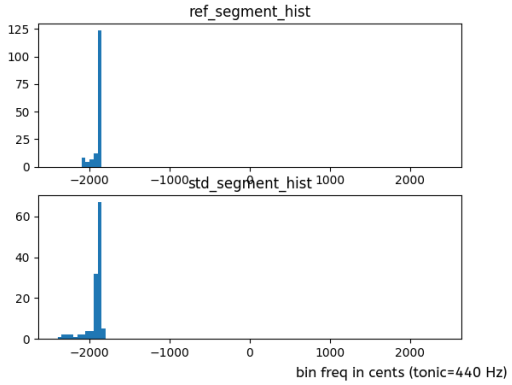

Figure 21: pitch histogram, binsize = 50 cents

6.2 Performance visualization with PHCD

While a numerical grade would be useful for ranking singing performances or for making a select/reject decision for a singing audition, a qualitative feedback about the errors in singing could be more useful for a learner. Since the PHCD is a robust feature, less sensitive to noise, it can be used to develop rule based note accuracy scores. These scores can by used to develop a visualization of the student performances.

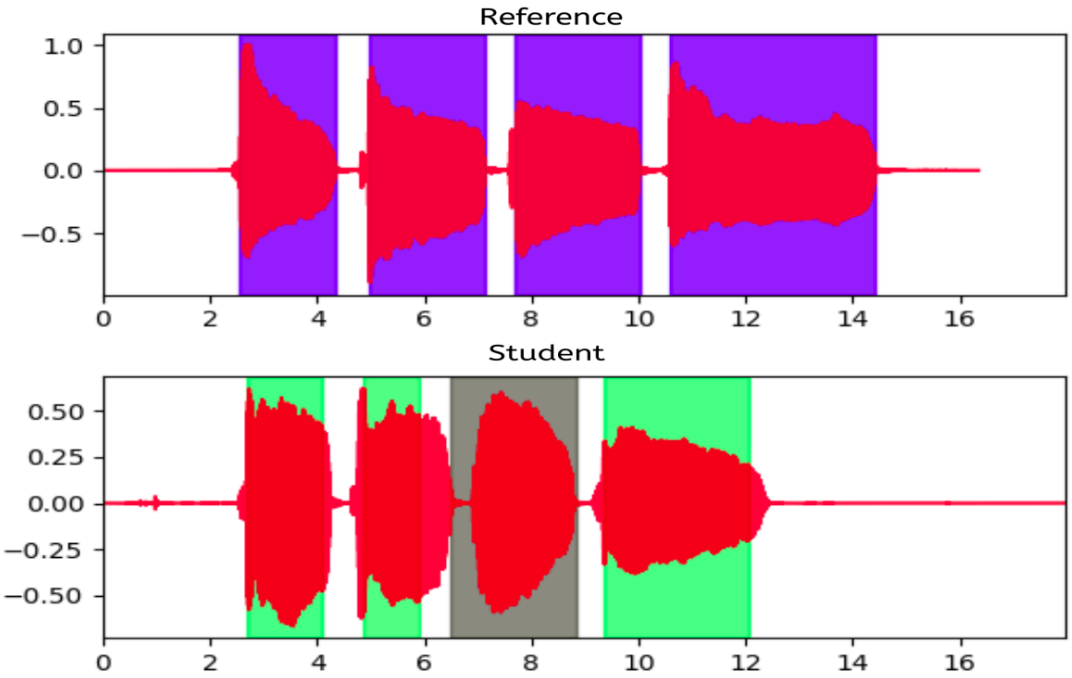

Figure 22: Performance visualization with PHCD

Such a rule based visualization was developed and can be seen in figure 22. The score formula used is based on the weighted average of all PHCD values with the weights as the binsizes.

$$score = 1 - \frac{100*(PHCD\_100)+50*(PHCD\_50)+20*(PHCD\_20)+10*(PHCD\_10)}{100+50+20+10}$$

The score thus determined will always lie between 0 and 1 with 0 denoting large error while 1 denoting virtually no errors. It is used to assign a colour to each estimated note segment in the student audio as a linear interpolation between green and grey, with green denoting accurate while grey denoting inaccurate note. It can be observed in figure 22 that the third note is inaccurately sung by the student. Note that small inaccuracy in the audio alignment has resulted in an incomplete detection of the second note. This, however, does not affect the evaluation of the note indicating the robustness of the PHCD feature. This visualization was used in the grader tool discussed previously as an aid to identify and quantify the errors by students.

6.3 Model Training and Testing

The aim of this study was to independently review and improve the baseline system in the context of Hindustani singing exercises. We treated this problem with a two step approach, 1) Audio-to-audio alignment of reference and student performances and 2) Feature extraction for the aligned audio pair.

In the chapter on Audio alignment, we have described the methods we used to show that for this dataset and application, the 120 dimensional HPCP is the best feature for audio alignment using DTW. To evaluate the performance of feature-extraction we described in the previous section, we compare the features we developed with the features used by the baseline system. For this purpose, we trained and tested linear regression models using a standard cross-validation scheme for machine learning systems: a group (%90) of the recordings are used for training the system and then the remaining (%10) of the recordings are used for testing (guaranteeing no overlap exists between training and test samples). The tests are repeated for ten times, each time using another group for testing. This is the same strategy used by the baseline model.

We trained and tested the linear regression models using the following strategies:

- Train and test SSV and HT_MTG datasets with PCHD features.
- Train and test SSV and HT_MTG datasets with pitch_error_stats features.
- Train and test on MAST dataset with PHCD features.
- Train and test on MAST dataset with pitch_error_stats features
- Train with SSV dataset using PHCD feature and test on the MAST dataset.

Note that in all these tests, we have used 120-dim HPCP for audio alignment while the BMCS model used pitch contour based audio alignment.

The results of the linear regression model trained with different combinations of features and dataset are shown in the following tables

PHCD feature	absolute error	
Dataset	Mean	Std dev
HT_MTG	0.3318	0.2668
MAST	0.5219	0.3746
SSV	0.3554	0.2972
Train: SSV, Test: MAST	0.5449	0.3804

Table 4: pitch_error_stats feature performance with linear regression models

pitch_error_stats features	absolute error	
Dataset	Mean	Std dev
HT_MTG	0.3478	0.3259
MAST	0.3990	0.3448
SSV	0.4567	0.4550

Table 5: pitch_error_stats feature performance with linear regression models

The BMCS involved analysis of cross-annotator grading consistency. They report the mean and std_dev of absolute error for the individual graders as 0.375487 and 0.368389 respectively. These numbers indicate the extent of the inter-annotator agreement and provide a base to interpret MAE for the BMCS. The grading tool visualization that was used to grade the HT_MTG and SSV datasets may improve inter-annotator agreement. However, since a cross-annotator grading analysis was not done for these datasets, we continue to use the BMCS values as a reference to interpret the results in tables 4 and 5

The Baseline MusicCritic System used MAST dataset with pitch_error_stats and reports an MAE of 0.45 with std dev of 0.36. We used the same dataset and features, albeit with the audio alignment using 120-dim HPCP instead of pitch contours, and obtained an improved result with MAE of 0.3990 and std dev of 0.3448. The improvement has to be attributed to the audio-alignment strategy since all other factors are identical.

The test results for both the Hindustani exercise datasets, HT_MTG and SSV, using PHCD features are comparable with MAE's of 0.33 and 0.35 respectively. However, for the same datasets using pitch_error_stats leads to a significant difference in the MAE and std dev. The potential for generalization for PHCD based model is tested

coefficient	HT_MTG	SSV	MAST
pitch_hist_cos_10	0.11995621	-1.43517695	0.44042155
pitch_hist_cos_20	-0.39142623	1.31682886	0.03144872
pitch_hist_cos_50	-0.83011561	-1.38104312	0.18924696
pitch_hist_cos_100	-2.7255353	-2.21555304	-3.9557599
constant	4.03347993437322	4.14117312751159	3.25244835462104

Table 6: PHCD feature: model interpretability

coefficient	HT_MTG	SSV	MAST
mean_diff	0.00430415776	0.00223936104	-0.00618255466
'std_diff'	0.00207092872	-0.00137574247	0.00441928193
'com_diff_dist'	0.178110342	0.00152854376	0.435407016
'diff_0_20'	3.48430199	2.75163046	0.824060921
'diff_20_50'	2.35107012	3.24348545	0.567932734
'diff_50_100'	-1.57870463	-3.09987344	0.0326336797
'diff_100_Inf'	-4.25666748	-2.89524247	-1.42462734
'mean_diff_ext'	-0.00347276405	-0.000587615555	0.00518758848
'diff_ext'	0.000102588453	-0.000166726554	0.00139088855
'com_diff_dist_ext'	0.216059937	0.213082845	-0.952116204
const	-0.908191779633695	0.14101252155054	7.18583225642118

Table 7: pitch_error_stats feature: model interpretability

by training using SSV dataset and testing on MAST dataset. The MAE and std of 0.5449 and 0.3804 respectively is comparable to the MAE and std of 0.5219 and 0.3746 for the model trained and tested with the MAST dataset.

In addition to the performance evaluation of the models on the basis of MAE and std dev, it is also important to discuss the interpretability of the learned models. Both the PHCD and pitch_error_stats are features based on error in pitch. If accuracy in pitch is what is desired from a student, any error based feature should have a negative correlation with the grade. This will be indicated by a negative coefficient for the feature in the regression equation.

The feature coefficients for the models using PHCD and pitch_error_stats are shown in tables 6 and 7 respectively. It can be seen that for the Hindutani datasets, the PHCD models appear very intuitive. The constant term is close to 4 which should be the score in absence of any error. Error at the 100 cent level is penalized the most followed by error at 50 cent.

The case of 10 and 20 cent level accuracy is peculiar. We should note that a large number of contributions are from amateur students. Also it is difficult to perceive pitch accuracy at 10 and 20 cent level of accuracy for the grader. We would expect these features to play a significant role only when the dataset contains singing by expert singers and is annotated by expert graders. For the current datasets, however, these coefficients may be merely artefacts of curve fitting. It is still important that in all the three models, their values are either low (as in HT_MTG and MAST) or mostly cancelling each other (as in SSV). The ground truth for MAST dataset is the average grade by six annotators who have graded without the help of a visual feedback of the performance. This may have resulted in the linear model only having the 100 cent level pitch accuracy as a significant feature.

The coefficients for pitch_error_stats based models on the contrary do not reflect such clear interpretability. The 'diff_100_Inf' feature which denotes the fraction of pitch error greater than 100 cents has a large negative coefficient in all the three models. But any other feature does not show such a clear trend. The 'mean_diff, std_diff and com_diff features' are not normalized. Hence small coefficients for these features do not necessarily mean low correlation. A very important factor is the variation in the constant term across the models. If the reference itself is considered as the student performance, all error based features would evaluate to zero and the grade predicted is the constant term. While the model with MAST data gives a very high grade of 7.18, the models with SSV gives a grade close to zero and HT_MTG gives a negative grade.

Chapter 7

Summary and Discussion

7.1 Contributions

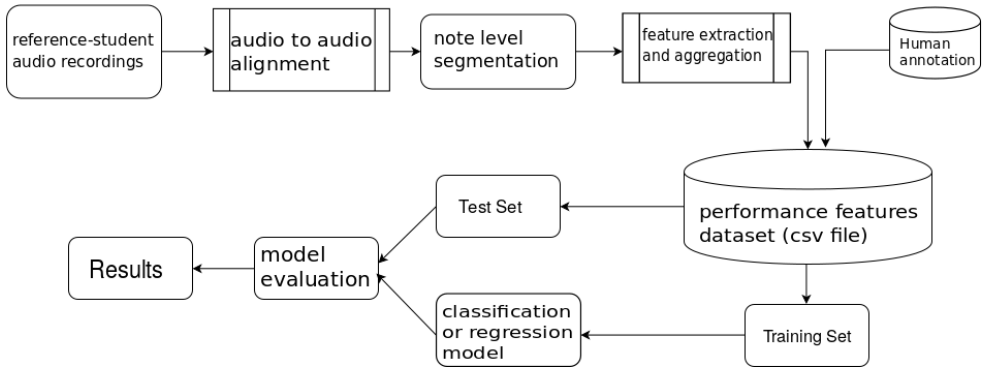

Figure 23: Notes singing assessment model Workflow

The principle contributions of this thesis are as follows:

- Two new datasets consisting of singing samples for Hindustani exercises are created. The first dataset (SSV dataset) contains recordings from students at Sargam Sangeet Vidyalay, a music school at Pune (India), who are familiar with Hindustani Music. The other dataset (HT_MTG dataset) has major contributions from MTG-UPF researchers, who may or may not be familiar with the music style.

- The problem of audio-to-audio alignment was extensively studied using a

methodology to evaluate alignment performance with custom-built metrics and a manually annotated subset of the dataset as the ground truth. Our experiments identified that higher dimensional (120 dimensions) Harmonic Pitch Class Profile (HPCP) as the best feature for a DTW based audio alignment for our dataset. To my knowledge, this is the first work which demonstrates the usefulness of higher dimensional HPCPs for the audio-alignment tasks.

- A novel Multi-feature DTW based alignment was tested with some success.
- A novel note level 'pitch histogram similarity feature' is introduced to evaluate the pitch accuracy of sung notes. This is shown to be more effective at developing interpretable linear regression models. The feature also helps us to provide a rule based score and a visualization for note-level accuracy of the student.
- Linear regression models are trained using these features and the first prototype for assessment of notes singing exercises in Hindustani Music is developed.

7.2 Critical Analysis

DTW based audio alignment would work as expected if there is an underlying similarity between the time series being aligned. In general for student performances with large errors, the audio alignment performance using HPCP is poor. Usually this is not a problem as a poor audio alignment still leads to high PHCD thus assigning poor grade to the student. We will discuss two cases where improper alignment leads to inappropriate student feedback.

Note : Green sections in the visualizations indicate accurate pitch as estimated by PHCD based score, while grey suggests inaccuracy in the pitch.

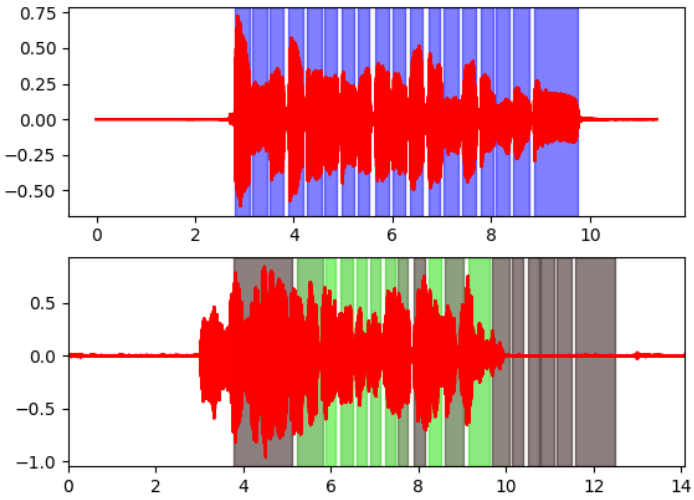

Figure 24: performance visualization of sound id 646

In the case of sound id 646, the student has sung all the notes inaccurate in pitch. HPCP based DTW alignment however finds tonal similarity between unrelated notes. This falsely predicts a high pitch accuracy in the performance as indicated in the figure 24.

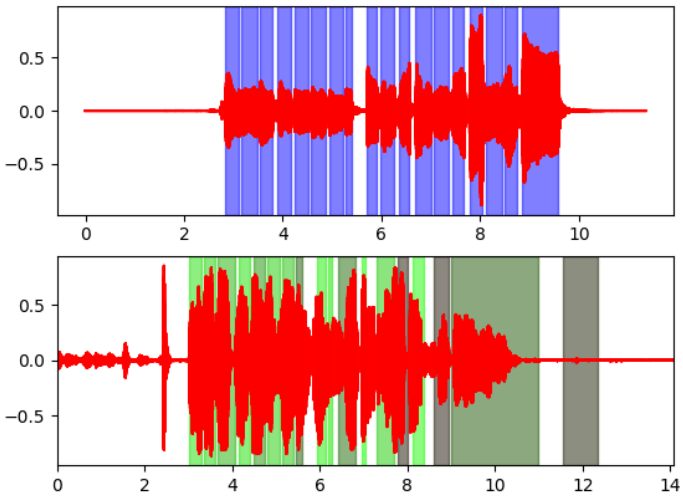

Figure 25: performance visualization of sound id 584

In the case of sound id 584 (figure 25), the student's pitch is lower than the reference in the notes between 5 seconds to 9 seconds, and the last few notes are mostly

accurate in pitch. The alignment error makes it appear as if the performance is mostly accurate till 8 seconds with errors at the end of the performance. This happens because the inaccurately sung notes are actually in tune with adjacent notes in the reference. The alignment error has resulted in an improper note-level visual feedback. However, this alignment error may not significantly impact the final grade estimate.

The next two examples discussed will demonstrate the robustness of the system to poor quality audio.

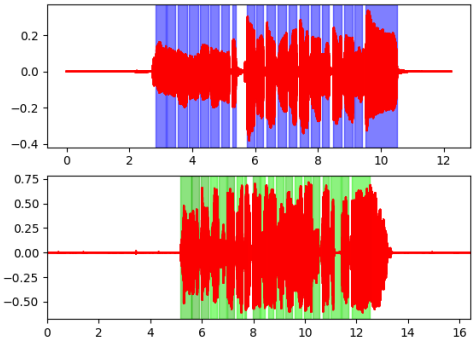

Figure 26: performance visualization of sound id 114

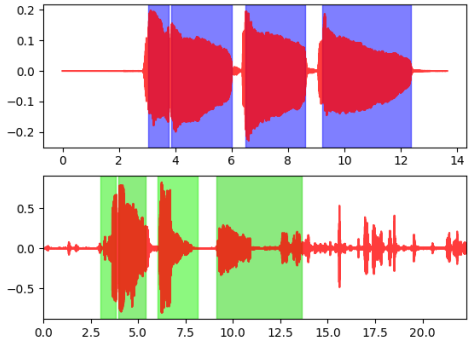

Figure 27: performance visualization of sound id 659

Sound id no 114 (figure 26) is accurately sung in pitch but has very poor quality audio due to clipping. The alignment or pitch accuracy estimate is unaffected by the audio quality.

Sound id no 659 (figure 27) has some background noise. The third note is barely audible. However, the alignment and pitch accuracy estimate is largely accurate.

The audio files of the examples presented will be available on the GitHub repository for this thesis.

7.3 Reproducibility

This work is a part of the MusicCritic project which is in active development. The HT_MTG and SSV audio datasets have not been published yet. However, representative samples of the data and the codes which implement the methods described in the thesis will be made available in the the following GitHub repository:

`https://github.com/PNinad/automatic-singing-assessment`

Chapter 8

Conclusions and Future Work

8.1 Conclusions

In the scope of this thesis, I have participated in the collection of Hindustani singing datasets with a collection of 349 audio samples of reference and student pairs. I have annotated the audio samples with an overall grade based on pitch accuracy. For this purpose, a grading tool with GUI which provided a visual feedback of the performance was developed. This tool helped to reduce grading biases in the annotations. The existing baseline system (BMCS) for singing assessment developed using Turkish Conservatory dataset was extensively studied to identify audio alignment as one of the possible areas of improvement. A formal approach and metrics were devised to test the audio alignment performance using different features. Using this methodology, I demonstrated that an improved audio-to-audio alignment system can be achieved using a 120-dimensional HPCP feature.

The second part of this study was focused on finding suitable features to assess a singing performance, given a good alignment of reference and student audio. I devised the 'pitch-histogram cosine distance' feature to measure note-level accuracy of singing. The effectiveness of these features with respect to the baseline features was shown by linear regression models trained and tested using the Hindustani and MAST datasets. The pitch-histogram cosine distance based models gave lower mean

absolute errors and lower variances in general, except for the case when the training-testing was done on MAST dataset. The pitch-histogram cosine distances as well as the baseline features are indicators of error in the student performance. Any linear regression model with error based features should ideally have a high positive constant term and the feature weights should be negative. The effectiveness of 'pitch-histogram cosine distance' is also indicated by the fact that the linear models developed with these features follow this trend. These features were also used to provide a note-level accuracy visualization of student performance. A rule based score was used to generate the visualizations. In the absence of a dataset to train a model, this score or its suitable adaptation can be used as a rule based grade.

A fully functional singing assessment system developed as an outcome of this work is deployed in a demo exercise which can be accessed at `https://musiccritic.upf.edu/demo/#demo`[1]

8.2 Future Work

This study primarily followed the same basic strategy as the baseline system. In this sense, it was not a completely independent approach to solve the singing assessment problem. The audio alignment, for example, was done using DTW since it was the approach followed by the baseline model. The task can also be done using a Gaussian Mixture Model based approach. The DTW alignment obtained from the current system could serve as a prior estimate for this approach as done by Devaney J. et. al. [32]. The Multi-feature DTW which showed some promising results should be explored further. The timing differences between the reference and student notes (as detected by the DTW based alignment) can be used to extract timing related features and provide a feedback of timing accuracy of the student. Additionally, the music performance can be judged on the basis of slightly subjective parameters such as dynamics, pronunciation and voice timbre. Work needs to be done to improve our system to include these rubrics. The exercises that I have used to collect the dataset are all mostly flat note exercises. Hindustani Music in its full grandeur would have

[1] last accessed 31 Aug 2019

legato style phrases. It needs to be tested how the existing system performs with these kind of exercises. This thesis was focused more on the data collection and signal processing to extract interpretable features from the student performance. The machine learning model used, i.e linear regression, gives a proof of concept that such a machine learning system can be trained using the features we extract. A rigorous study should be done to extract more features and train-test different machine learning systems to improve the assessment accuracy. Recent studies [7] in music performance assessment have employed feature learning instead of feature extraction using deep learning architectures. Deep learning models would require a larger dataset and more computational power to train the model. It also has a drawback that the models developed may not interpretable. However, purely on the criterion of accuracy, it needs to be seen how our model performs in comparison with deep learning approaches.

List of Figures

List of Tables

Bibliography

[1] Misra, C., Chakraborty, T., Basu, A. & Bhattacharya, B. Swaralipi: A framework for transcribing and rendering indic music sheet (2016).

[2] Bozkurt, B., Baysal, O. & Yuret, D. A dataset and baseline system for singing voice assessment. In *Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR), Matosinhos, Portugal*, 25–28 (2017).

[3] Bozkurt, B., Gulati, S., Romani, O. & Serra, X. Musiccritic: A technological framework to support online music teaching for large audiences. In *33rd World Conference of International Society for Music Education (ISME)* (Baku, Azerbaijan, 2018). URL `https://doi.org/10.5281/zenodo.1211450`.

[4] Severance, C., Hanss, T. & Hardin, J. Ims learning tools interoperability: Enabling a mash-up approach to teaching and learning tools. *Technology, Instruction, Cognition and Learning* **7**, 245–262 (2010).

[5] Nakano, T., Goto, M. & Hiraga, Y. An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. In *Ninth International Conference on Spoken Language Processing* (2006).

[6] Abeßer, J., Hasselhorn, J., Dittmar, C., Lehmann, A. & Grollmisch, S. Automatic quality assessment of vocal and instrumental performances of ninth-grade and tenth-grade pupils. In *International Symposium on Computer Music Multidisciplinary Research* (2013).

[7] Pati, K., Gururani, S. & Lerch, A. Assessment of student music performances using deep neural networks. *Applied Sciences* **8**, 507 (2018).

[8] Gupta, C., Li, H. & Wang, Y. Automatic evaluation of singing quality without a reference. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 990–997 (IEEE, 2018).

[9] Gong, R. *Automatic assessment of singing voice pronunciation: A case study with jingju music.* Ph.D. thesis, Universitat Pompeu Fabra (2018).

[10] Lerch, A., Arthur, C., Pati, A. & Gururani, S. Music performance analysis: A survey. *arXiv preprint arXiv:1907.00178* (2019).

[11] Bayle, Y. *et al.* Kara1k: A karaoke dataset for cover song identification and singing voice analysis. In *2017 IEEE International Symposium on Multimedia (ISM)*, 177–184 (2017).

[12] DeLuca, C. & Bolden, B. Music performance assessment: Exploring three approaches for quality rubric construction. *Music Educators Journal* **101**, 70–76 (2014).

[13] Wesolowski, B. C. Understanding and developing rubrics for music performance assessment. *Music Educators Journal* **98**, 36–42 (2012).

[14] De Cheveigné, A. & Kawahara, H. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America* **111**, 1917–1930 (2002).

[15] Mauch, M. & Dixon, S. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 659–663 (IEEE, 2014).

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

[17] Atli, H. S., Uyar, B., Sentürk, S., Bozkurt, B. & Serra, X. Audio feature extraction for exploring turkish makam music. In *Proceedings of 3rd International Conference on Audio Technologies for Music and Media, Ankara, Turkey*, 142–153 (2014).

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

[19] Chan, T.-S. *et al.* Vocal activity informed singing voice separation with the ikala dataset. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 718–722 (IEEE, 2015).

[20] Cannam, C. *et al.* Mirex 2013 entry: Vamp plugins from the centre for digital music. *Proceedings of the Music Information Retrieval Evaluation Exchange.[Cited in pages 102 and 104.]* (2013).

[21] Bogdanov, D. *et al.* Essentia: An audio analysis library for music information retrieval. In *Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4-8; Curitiba, Brazil.: ISMIR; 2013. p. 493-8.* (International Society for Music Information Retrieval (ISMIR), 2013).

[22] Müller, M. Dynamic time warping. *Information retrieval for music and motion* 69–84 (2007).

[23] Sakoe, H. & Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing* **26**, 43–49 (1978).

[24] Itakura, F. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **23**, 67–72 (1975).

[25] Salvador, S. & Chan, P. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis* **11**, 561–580 (2007).

[26] Kirchhoff, H. & Lerch, A. Evaluation of features for audio-to-audio alignment. *Journal of New Music Research* **40**, 27–41 (2011).

[27] Gómez, E. Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing* **18**, 294–304 (2006).

[28] Mauch, M. & Dixon, S. Approximate note transcription for the improved identification of difficult chords. In *ISMIR*, 135–140 (2010).

[29] Vidwans, A. *et al.* Objective descriptors for the assessment of student music performances. In *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio* (Audio Engineering Society, 2017).

[30] Balaguer-Ballester, E., Clark, N. R., Coath, M., Krumbholz, K. & Denham, S. L. Understanding pitch perception as a hierarchical process with top-down modulation. *PLoS computational biology* **5**, e1000301 (2009).

[31] Kollmeier, B., Brand, T. & Meyer, B. Perception of speech and sound. In *Springer handbook of speech processing*, 61–82 (Springer, 2008).

[32] Devaney, J., Mandel, M. I. & Ellis, D. P. Improving midi-audio alignment with acoustic features. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 45–48 (IEEE, 2009).