# ANALYSIS OF MUSICAL DYNAMICS IN VOCAL PERFORMANCES USING LOUDNESS MEASURES

*Jyoti Narang, Marius Miron, Ajay Srinivasamurthy and Xavier Serra*

Universitat Pompeu Fabra
Barcelona, Spain
{jyoti.narang,marius.miron,ajays.murthy,xavier.serra}@upf.edu

## ABSTRACT

In addition to tone, pitch and rhythm, dynamics is one of the expressive dimensions of the performance of a music piece that has received limited attention. While the usage of dynamics may vary from artist to artist, and also from performance to performance, a systematic methodology to automatically identify the dynamics of a performance in terms of musically meaningful terms like *forte*, *piano* may offer valuable feedback in the context of music education and in particular in singing. To this end, we have manually annotated the dynamic markings of commercial recordings of popular rock and pop songs from the Smule Vocal Balanced (SVB) dataset which will be used as reference data. Then as a first step for our research goal, we propose a method to derive and compare singing voice loudness curves in polyphonic mixtures. Towards measuring the similarity and variation of dynamics, we compare the dynamics curves of the SVB renditions with the one derived from the original songs. We perform the same comparison using professionally produced renditions from a karaoke website. We relate high values of Spearman correlation coefficient found in some select student renditions and the professional renditions with accurate dynamics.

## 1. INTRODUCTION

Dynamics are used to convey expressiveness of a musical performance. In musical terms, the term *dynamics* often refers to the intended or perceived "sound strength", while in technical terms, the musical dynamics are usually mapped to loudness of the resulting audio [1]. The classification of dynamic markings for performances into categories like - *pp* (very soft), *p* (soft), *mp* (moderately soft), *mf* (moderately loud), *f* (loud), *ff* (very loud) remains widely accepted [2], and several studies have been conducted analyzing the relationship between the dynamic markings in the score to the observed values of loudness in audio [3, 4], particularly for the case of Western Classical piano performances [4–6]. However, not many studies have been conducted analyzing the role of dynamics in vocal performances [7].

The task of automatic transcription [8] of dynamics from audio is useful in scenarios where the availability of scores is limited or the primary source of learning is via oral means, for example in traditions like pop and jazz. In such oral traditions, learning entails not only following the original performance in terms of rhythmic [9] and pitch accuracy [10], but also implicitly reproducing the expressive techniques employed by the original artist. With automatically transcribed dynamic markings, it is possible for a vocal practitioner or learner to understand the interpretation of a given piece of music as intended by the artist, and reproduce them in the same way. This can be particularly useful in vocal music learning and assessment applications [11], or singing along with karaoke tracks. In addition, a system that yields the dynamic range of a song based on audio analysis may help the students to choose songs within a certain dynamic range, corresponding to their own. However, the lack of annotations and data make the evaluation of this task particularly challenging.

The variation of musical dynamics is usually instrument dependent [2], and several approaches exist to model musical dynamics [12], approaching it as a classification problem i.e. categorizing the label (p, m, f etc) based on the observed loudness levels, or prediction problem, where loudness levels corresponding to the dynamics markings are predicted, using machine learning approaches like decision trees, Support Vector Machine (SVM), etc [3]. Jeong et al. [13] predict the note level intensity for the case of piano using non-negative matrix factorization (NMF) based techniques taking the aligned score and performance audio as input. Marinelli et al. [14] use convolutional neural network (CNN) with modulation power spectra as an input feature for dynamics classification into categories *pp* and *ff*. However, existing work on computational modelling of vocal dynamics is rather limited with almost no annotation availability to the best of our knowledge.

In our previous work [7], we devised a methodology to extract musical dynamics from audio via loudness features either from a mix or monophonic vocal audio recordings. To validate our approach, we conducted a case study where we asked a music teacher to provide feedback on the musical dynamics employed by the artist ‘Norah Jones’ in her rendition of the song ‘Don’t know why’, in reference to a karaoke version. We found in our analysis that the musical dynamics markings by the teacher were closely correlated with the loudness feature extracted from the audio. However, the methodology was conducted for a small number of renditions, extracted from professionally produced karaoke songs, and the case study was carried out for one song.

In the current work, we extend a similar analysis for a large number of songs, part of publicly available Smule Vocal Balanced dataset [15]focusing on measuring and comparing the dynamics in vocal rock and pop performances using audio recordings. We first annotate the score in the form of musical notes corresponding to the singing voice for five popular songs. We then collaborate with a music teacher to annotate the dynamic markings in these songs. Only five songs have been chosen for analysis due to either source separation artifacts, overall song length mismatch of karaoke/professional renditions with original songs, or challenges with creating a score that is aligned with the original renditions. We compare the loudness contours computed on the original recordings with the associated renditions from the SVB dataset [15] using sone scale [16], which is based on psychoacoustic model inspired by the human ear, for comparing the mapping of dynamics markings indicated by the teacher to the loudness values obtained from the audio signals. Further, we propose metrics for comparing dynamics based on Spearman correlation of loudness curves (1) for entire performances and (2) around the annotated dynamic markings. The rest of the paper is structured as follows. In section 2, we introduce the dataset we use in our analysis, with Section 3 describing the methodology used in the process. Section 4 contains the details of the conducted experiments along with the challenges and limitations of the proposed methodology. Finally, we conclude with a discussion section and possible directions of future work.

## 2. DATASETS

Because the goal of our research is to compare singing voice dynamics between original songs, professionally produced renditions, and student renditions, we derive a dataset from three data sources:

(i) Five commercial pop rock songs listed in Table 1, for which we obtained the original audio from YouTube.

(ii) For the five songs we obtain professionally produced renditions and the associated audio stems from a karaoke website[1]. The choice of karaoke tracks of the same songs in our dataset helps us validate if the reproduction of a song by a professional singer involves reproducing the dynamics of the original artist.

(iii) From the Smule Vocal Balanced (SVB) dataset [15] we select the audio renditions corresponding to the five songs. The SVB dataset comprises student recordings of 24874 solo singing performances from 5429 singers singing a collection of 14 songs. The recordings comprise different levels of singing training and recording quality. In addition, some of the performances may be duets and some are incomplete.

**Score creation for the original audio tracks.** In order to carry out evaluation, we need scores with precise dynamics as intended or perceived in the original recordings. Hence, we collaborated with a music teacher to identify the dynamics markings at the note level, and eventually at the phrase level. We create a score for select reference recordings in musicXML format from the SVB dataset and via collaboration with a music teacher, we annotate the score with dynamics markings into 8 categories *ppp*, *pp*, *p*, *mp*, *mf*, *f*, *ff* and *fff*. The annotation process is defined as follows. The teacher listens to the complete source separated version of the reference recording, and thereafter annotates the song sections with corresponding dynamics markings at the note level. The note level annotations of each rendition is created by the author by listening to the source separated reference recordings. MusicXML is chosen as an intermediate tool for score transcription in order to extract information like note pitch, note start time, note end time, measure number, beat number etc from the score, such that timing information can be mapped between the score and audio. The end result of this step is a score that can be parsed automatically using tools like music21 [17] for extracting the dynamics from the score. Each annotation for our task took close to 6 hours. The annotations were also validated by the author in the process.

## 3. METHODOLOGY

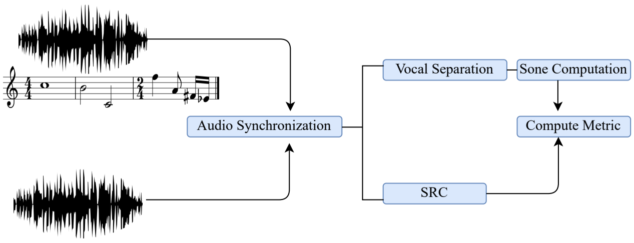

Figure 1: Methodology for extracting and comparing Dynamics

The methodology for extracting loudness is presented in Figure 1. There are 3 parts to the process 1) Audio Synchronization 2) Data pre-processing and feature computation 3) Metric Computation

### 3.1. Audio Synchronization

Since the score is annotated on the original recording, we assume that the score is coarsely aligned with the audio signal of this recording. However, renditions may not be perfectly aligned with the original recordings. We assume that any of the renditions are performed with a backing track. In some cases we have to account for an initial offset, before the backing track starts. To align the renditions with reference recordings we compute cross correlation of the pitch contour of the two audio signals using melodia [18] implementation of essentia [19] using a hop size of 0.003 ms. The offset in frames is then mapped to the corresponding frame position in the extracted loudness curve.

### 3.2. Data pre-processing and feature computation

#### *3.2.1. Source Separation*

The recent progress in the field of audio source separation, especially for contemporary rock and pop genre of music facilitated us to use it as an intermediate step. We validated the efficacy of this step in our previous work [7] with the MusDB dataset [20], where the correlation coefficient between the loudness curves of source separated vocals with the loudness curve of the vocal stem was very high, in most cases, being greater than 0.9.

#### *3.2.2. Loudness Extraction from Audio*

With isolated vocal tracks from the mix or monophonic recordings from renditions of professional/amateur singers, the next step is to extract loudness curves from each of the sources to compare them. We use the *sone* scale for this purpose. Sone scale is inspired by psychoacoustic concept of equal loudness curves, with the measurement being linear i.e doubling of perceived loudness doubles the sone value [16]. The phon scale is closely associated with the dB scale, where 1 phon is equivalent to 1 deciBel at 1000 Hz (1 kHz). The sone scale is based on the observation that a 10 phon increase in sound level is perceived as doubling of loudness. A phon value of 40 translates to 1 sone, and the relationship between phons and sones can be modelled with the equation:

[1] https://www.karaoke-version.com/

$$S = \begin{cases} 2^{(L-40)/10}, & \text{if } P >= 40. \\ (L/40)^{2.642}, & P < 40. \end{cases} \tag{1}$$

The sone scale computation along with the consecutive smoothing operation is carried out in the same way as proposed by Kosta et al [3] in their analysis. Each of the curves are normalized by dividing by the max value to compare the relative values, and not the absolute ones.

### 3.3. Metric Computation

Vocal dynamics may fluctuate throughout the song. However, only certain parts of the song may include dynamic changes and may be deemed more important in judging the expressiveness of dynamics. To account for that we compute comparisons between renditions and original recordings at the song level (global) and at change points we annotated (local).

#### *3.3.1. Global Loudness Comparison*

To compare the loudness curves of the rendition and the original song, we did not want to make any assumptions about the underlying data distribution [21], and hence decided to compute the non-parametric Spearman Correlation Coefficient ($\rho$) of the smoothed curves of the aligned renditions.

$$r_s = \rho = 1 - \frac{6 \sum {d_i}^2}{n(n^2 - 1)} \tag{2}$$

where

- $d_i$ = $R(X_i)$ - $R(Y_i)$ is the difference between the two ranks of each observation, where $X_i$ and $Y_i$ represent raw scores
- $n$ is the number of observations

#### *3.3.2. Local Loudness Comparison at the Change Points*

**Change points** refer to the points in time where the dynamic changes occur, for example from *mf* to *f* and so on. The local changes are measured by computing the Spearman correlation, $\rho$ of the smoothened loudness curves at the change point window, where a change point window is estimated from the score using music21 [17] and thereafter mapped to the corresponding position in the audio. In order to do so, we first compute the beats using madmom's [22] *DBNBeattracker*, and then manually check for the initial beat until the score is well aligned with the reference recording. To make sure alignment is in order, we use synthesized audio from the score using *fluidsynth* [23] library and play the synthesized version with the reference recording, making sure all the change points are mapped correctly. We use a total of 4 beats as part of the change point window (2 beats before the change point, and 2 beats after the change point) in order to carry out evaluation at the change point window.

## 4. EXPERIMENTS

### 4.1. Experimental Setup

| Song Name | Artist | Filtered Student Count |
|---|---|---:|
| All of me | John Legend | 416 |
| Chandelier | Sia | 4 |
| Love Yourself | Justin Beiber | 149 |
| Say you won't let go | James Arthur | 362 |
| When I was your man | Bruno Mars | 7 |

Table 1: Student Data Statistics

For the commercial popular recordings, we solely have access to the mixed tracks for the subsets (i) and (iii) of our dataset, while for subset (ii) the karaoke versions, we have access to all the stems. To that extent, for the subsets (i) and (iii) we extract the vocal stem using the Spleeter implementation of the U-Net singing voice source separation [24]. Thereafter, the loudness curve is extracted from the separated vocal track or vocal stem using sone scale. The methodology for extracting loudness in terms of the sone scale is described above, and we use similar set of parameters to Kosta et al. [4]. We use a block size of 512 samples or 11 ms with a Hanning window, and a hop size of 256 samples or 5.5 ms. We use the ma_sone function in Elias Pampalk's Music Analysis toolbox [25] in Matlab. Further, we apply smoothing operation using "loess" with *smooth* function in matlab (based on locally weighted non-parametric regression fitting using a 2nd order polynomial). We experimentally determine the time span for the loess method to 5%. The loudness curves are normalized by dividing by the max value to carry out a relative comparison between renditions of different amplitude levels. We also experimented with using dynamic range for normalization, however, the minimum sone value when comparing the entire rendition is always 0, and source separation artifacts sometimes interfere with minimum value selection using peak-picking leading to a smaller dynamic range. For evaluating the correlation solely around the change points, we take a change point window around each change point indicated by the score. We take a time interval of 2 beats before and 2 beats after the change point.

#### *4.1.1. Sone scale loudness curve comparison with other scales*

We also experimented with Loudness Unit Full Scale (LUFS) loudness extraction using the Essentia implementation of EBUR128 [26], comparing the smoothed loudness curves of karaoke/professional renditions with original renditions. We found the results to be quite similar with a variation of around 6 to 7% for all the songs. We use the momentary loudness with 400 ms block-size, 5.5 ms hop-size and 5% time-span of the "loess" function. The variability of the results was primarily due to tuning of time-span parameter of the "loess" function. Further, we also compared the sone scale values to RMS values of the signal in our initial experiments, and found the RMS output to be quite noisy. We continued our experimentation with the sone scale considering the robustness of the resulting values with parameter tuning in reference to other scales.

### 4.2. Pre-processing Student Recordings

The Smule dataset consists of 24874 monophonic recordings of 14 commercial songs, of which we select five songs as mentioned above. Many of the renditions are sung in duets. In order to carry out a dynamics comparison for the entire song, we manually filter out duet recordings from complete renditions in our analysis. We make use of percentage voice in the audio track for doing so, thresholding by different numbers based on the input song, us-

Table 2: Observed mean, median and standard deviation for aligned student recordings and $\rho$ for Karaoke Recordings

| | Student Recordings | | | | | | Karaoke Recordings | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Global | | | Local(Change Points) | | | Global | Change Points | | |
| Song | Mean | Median | SD | Mean | Median | SD | $\rho$ | Mean | Median | SD |
| All of me | 0.72 | 0.71 | 0.06 | 0.59 | 0.96 | 0.61 | 0.88 | 0.62 | 0.99 | 0.64 |
| Chandelier | 0.77 | 0.76 | 0.05 | 0.47 | 0.90 | 0.77 | 0.92 | 1.0 | 1.0 | 0.0 |
| Love yourself | 0.71 | 0.72 | 0.06 | 0.71 | 0.99 | 0.58 | 0.89 | 0.99 | 0.99 | 0.001 |
| Say you wont let go | 0.71 | 0.70 | 0.07 | 0.35 | 0.67 | 0.73 | 0.90 | 0.67 | 0.86 | 0.52 |
| When I was your man | 0.85 | 0.87 | 0.06 | 0.73 | 0.94 | 0.43 | 0.88 | 0.67 | 0.87 | 0.46 |

Table 3: Difference mean, median of Student Recordings and Karaoke Recordings

| | Global | | Local(Change Points) | |
|---|---|---|---|---|
| Song | Mean Difference | Median Difference | Mean Difference | Median Difference |
| All of me | 0.16 | 0.17 | 0.03 | 0.03 |
| Chandelier | 0.15 | 0.16 | 0.53 | 0.10 |
| Love yourself | 0.18 | 0.17 | 0.28 | 0.0 |
| Say you wont let go | 0.19 | 0.20 | 0.32 | 0.19 |
| When I was your man | 0.03 | 0.02 | -0.06 | -0.07 |

ing *split* function present in the librosa library [27] for this preprocessing step. Moreover, not all renditions are sampled at the same rate, hence, we carry our resampling operation over the entire filtered set, keeping the sampling rate at 44100 Hz. Further, we filter out all recordings shorter than the length of the corresponding reference tracks. We also filter out all recordings where the global Spearman Correlation Coefficient of the loudness curves is less than 0.6. This threshold was chosen by first holding out 10% of the recordings for each song, and then manually listening to 5 songs for each song from the held out data. We found that songs with values smaller than $\rho$ of 0.6 either had background noise that could not be filtered using vocal activity detection, or were not complete renditions. Table 1 presents the song count of student renditions after pre-processing and filtering the remaining 90% data with the threshold. As evident from the table, there aren't many recordings available for the songs 'Chandelier' and 'When I was your man', suggesting preference of certain songs over others. This further leads to class imbalance, however, since we carry out evaluation at the song level, the class imbalance does not impact the evaluation.

### 4.3. Results

We compute and average Spearman correlation values globally and locally for all student renditions compared to the original songs. In addition, we report similar metrics when comparing the karaoke /professional renditions with the original songs.

Table 2 presents the mean, median, and standard deviation (SD) for the entire set of student renditions and at change points. It is to be noted that for student renditions, global mean, median and SD's refer to the mean of Spearman Correlation or $\rho$ values for the entire set of student renditions for any given piece of music, while for karaoke/professional renditions, we report the $\rho$ for the entire song as a global metric, and mean, median and SD's of the change points as a local metric. Following are some primary conclusions from our investigation:

- The median values at change points are generally higher as compared to mean values
- Professional/karaoke renditions have higher values as compared to student/SVB renditions
- The scale used for loudness measurement i.e. EBUR or sone does not impact the results much with correct parameter tuning.

The median values being higher than mean values could stem from the fact that correlation values are sensitive to silence originating from aligned smoothed curves, leading to higher values at change points where dynamics variation coincides with aligned renditions. On the other hand, a mismatch at any of the change point pushes down the mean values at change points. For example, Figure 2 presents the sone values, along with smoothed loudness curves and detected change points for the song 'Say you won't let go'. The $\rho$ value at change points 7 and 8 in this example turn out to be negative pushing down the mean values for the entire rendition. The median for all professional renditions is greater than 0.8 validating our hypothesis that professional singers are able to reproduce the dynamics in most cases.

Table 3 presents the difference of mean and median values of student renditions from karaoke renditions at change points as well as globally. Most values in the difference table are positive, suggesting validation of the hypothesis that professional/karaoke singers follow the dynamics of the original/reference rendition better than the students on average. For the case of the song 'When I was your man', the mean and median difference is negative. We analysed the recordings by listening to them for the song, and found them to be following the notes as well as the dynamics relatively well.

There is only one change point for the case of the song 'Love yourself', and a high mean and median value across student renditions indicates that most students follow that particular change point.

### 4.4. Challenges and Limitations

Despite the encouraging results that we find in our investigations, the proposed system works only for specific conditions at the mo-

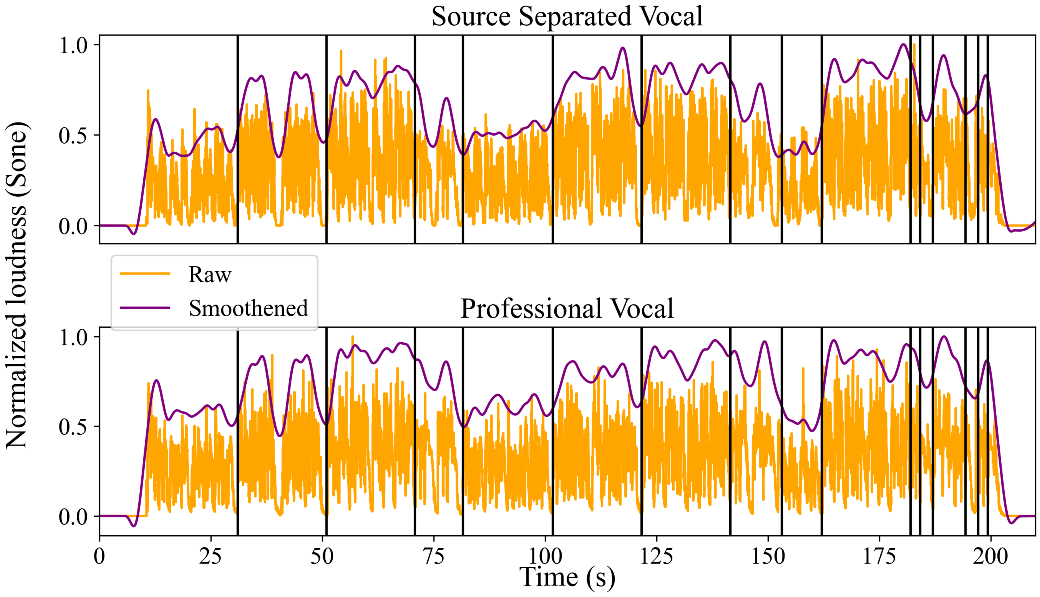

Figure 2: Comparison between source separated vocals of the rendition and the professional vocal stem

ment. The analyzed recordings should be monophonic, without any noise or backing track interference, and approximately the same length as the original/reference recording. Moreover, we filter out all renditions where the singers sing only the partial or one part of the duet track. In this investigation, we have also filtered out renditions with Spearman Correlation values less than 0.60. This threshold was found by perceptually testing student renditions with lower values, which either had a backing track or noise interference or were renditions where the students stopped singing/repeated certain sections amongst other challenges.

The loudness curves are also dependent on the robustness of the source separation algorithm applied prior to loudness extraction. We discarded some original recordings and the associated renditions in the SVB dataset because the source separation output had interference from the instrumental portions as well, that were affecting the dynamics of the source separated version of the track.

Apart from score creation, we need to make sure that the score is completely aligned with the reference track to fetch the change point window correctly. However, for some tracks, the BPM value is not static and changes over the course of the song that makes it challenging to use this approach for analysis. In the current songs chosen for analysis, the time signature is 4/4, and the BPM value remains constant through out the performance.

## 5. DISCUSSION

Work on dynamics extraction and measurement is a challenging task for several reasons. The first being lack of sufficiently annotated data for singing voice. While we take some preliminary steps to address this gap by creating some scores with dynamics markings, it is challenging to scale this approach to include any given piece of music. Moreover, while creating annotations for this work, many a times, the music teacher would discretize a given dynamics category to further levels, for example *p+ or p-*, which we could not address due to the limitations of tools like Musescore and music21.

Another important factor that plays a role in dynamics analysis is the compression factor applied to professionally produced music. While the artists may have a range of dynamics that they employ in a performance, many a times, the producers chose to limit or compress the vocal range within a specific limit. However, we simplify the problem statement with an assumption that mastering is done in such a way that musical dynamics are retained in any given performance and can be easily perceived and distinguished by a music teacher.

In the current investigation, we have simplified the problem statement with an assumption that students imitate the performance of the teacher, including the dynamics of the original performance without addition of a subjective interpretation. The annotations are created by a teacher with expertise in the Rock and Pop genre of music, and since the annotations are created by the same teacher for all the songs, the analysis is consistent and coherent for the entire dataset. There is also a possibility that the annotations may vary from one expert to another, however, we do not take teacher variation into consideration for our analysis. Further, the students always perform with an accompaniment track, which are similar across renditions of the same piece leading to a similar musical context.

The mapping between absolute value of loudness measured from audio signals to specific musical dynamics category will also depend on the genre of music being evaluated. For Rock and Pop genre of music, the expected dynamic range is generally much narrower than genres like orchestra or opera.

Although the results are promising, the analysis is dependent on the parameters related to smoothing and sones computing, for which we determine values experimentally. Moreover, the usage of dynamics in a performance is very much artist and song dependent, that adds to the difficulty of a piece of music. For example, the music teacher annotated ‘Love yourself’ to be a song with easy dynamics structure in terms of difficulty, and 'All of me' to be a song with much more dynamics variation. Hence, we would need to explore and use a combination of metrics to cater to song variations.

## 6. CONCLUSION AND FUTURE WORK

We propose a system for dynamics analysis, particularly testing it for the case of vocal music education. Our system proposes comparing original recordings and renditions using Spearman correlation of loudness curves globally and locally, at change points. The evaluation we perform on a dataset we derive from the SVB dataset shows that professional produced recordings have higher correlation than amateur renditions. In addition, the local comparison is more sensitive to outliers and it is more discriminative. For the current investigation, we have limited the analysis to one annotator, keeping the analysis consistent and coherent with one teacher with expertise in Rock and Pop genre of music. The annotations were also reviewed by the author of the paper. Several directions can be explored going forward, the primary one being addition of more evaluation metrics testing it for varying levels of difficulty of a music piece, and also carrying out subjective evaluation with the help of the same teacher, testing whether the objective metrics are in line with the subjective evaluation. We also intend to extend the analysis with addition of closely correlated features, especially the relationship of loudness with timbre. Further, we intend to apply machine learning approaches to predict the dynamics of a music piece, taking advantage of the annotations that are created as part of this work.

**Acknowledgement** We would like to thank Divakar Nambiath for his invaluable contributions. Ajay Srinivasamurthy is currently with Amazon Alexa, India and contributed to the ideas and work described in the paper before joining Amazon. Part of this research is funded by Project Musical AI - PID2019-111403GB-I00/AEI/10.13039/501100011033 funded by the Spanish Ministerio de Ciencia, Innovación y Universidades (MCIU) and the Agencia Estatal de Investigación (AEI).

## 7. REFERENCES

[1] Stefan Weinzierl, Steffen Lepa, Frank Schultz, Erik Detzner, Henrik von Coler, and Gottfried Behler, "Sound power and timbre as cues for the dynamic strength of orchestral instruments," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1347–1355, 2018.

[2] Blake Patterson, "Musical dynamics," *Scientific American*, vol. 231, no. 5, pp. 78–95, 1974.

[3] Katerina Kosta, Rafael Ramírez, Oscar F Bandtlow, and Elaine Chew, "Mapping between dynamic markings and performed loudness: a machine learning approach," *Journal of Mathematics and Music*, vol. 10, no. 2, pp. 149–172, 2016.

[4] Katerina Kosta, Oscar F Bandtlow, and Elaine Chew, "Dynamics and relativity: practical implications of dynamic markings in the score," *Journal of New Music Research*, vol. 47, no. 5, pp. 438–461, 2018.

[5] Neil P McAngus Todd, "The dynamics of dynamics: A model of musical expression," *The Journal of the Acoustical Society of America*, vol. 91, no. 6, pp. 3540–3550, 1992.

[6] Gerhard Widmer and Werner Goebl, "Computational models of expressive music performance: The state of the art," *Journal of new music research*, vol. 33, no. 3, pp. 203–216, 2004.

[7] Jyoti Narang, Marius Miron, Xavier Lizarraga Seijas, and Xavier Serra, "Analysis of musical dynamics in vocal performances," in *Proceedings of the 15th International Symposium on Computer Music Multidisciplinary Research (CMMR 2021)*, Tokyo, Japan, Nov. 2021, pp. 99–108.

[8] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri, "Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.

[9] Fabien Gouyon and Simon Dixon, "A review of automatic rhythm description systems," *Computer music journal*, vol. 29, no. 1, pp. 34–54, 2005.

[10] David Gerhard et al., *Pitch extraction and fundamental frequency: History and current techniques*, Department of Computer Science, University of Regina Regina, SK, Canada, Dec. 2003.

[11] Vsevolod Eremenko, Alia Morsi, Jyoti Narang, and Xavier Serra, "Performance assessment technologies for the support of musical instrument learning," 01 2020, pp. 629–640.

[12] Axel Berndt and Tilo Hähnel, "Modelling musical dynamics," in *Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound*, 2010, pp. 1–8.

[13] Dasaem Jeong and Juhan Nam, "Note intensity estimation of piano recordings by score-informed nmf," in *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*. Audio Engineering Society, 2017.

[14] Luca Marinelli, Athanasios Lykartsis, Stefan Weinzierl, and Charalampos Saitis, "Musical dynamics classification with CNN and modulation spectra," in *Proceedings of the 17th Sound and Music Computing Conference*, Torino, Italy, 2020, pp. 193–199.

[15] Inc. Smule, "DAMP-VPB: Digital Archive of Mobile Performances - Smule Vocal Performances Balanced," Nov. 2017.

[16] Jacob Beck and William A Shaw, "Ratio-estimations of loudness-intervals," *The American journal of psychology*, vol. 80, no. 1, pp. 59–65, 1967.

[17] Michael Scott Cuthbert and Christopher Ariza, "music21: A toolkit for computer-aided musicology and symbolic music data," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, Utrecht, Netherlands, Aug. 2010, pp. 637–642.

[18] Justin Salamon, Emilia Gómez, Daniel PW Ellis, and Gaël Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 2, pp. 118–134, 2014.

[19] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José Zapata, and Xavier Serra, "Essentia: an open-source library for sound and music analysis," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 855–858.

[20] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner, "Musdb18-a corpus for music separation," 2017.

[21] Emery Schubert, "Correlation analysis of continuous emotional response to music: Correcting for the effects of serial correlation," *Musicae scientiae*, vol. 5, no. 1_suppl, pp. 213–236, 2001.

[22] Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer, "Madmom: A new python audio and music signal processing library," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 1174–1178.

[23] Jan Newmarch, "Fluidsynth," in *Linux Sound Programming*, pp. 351–353. Springer, 2017.

[24] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde, "Singing voice separation with deep u-net convolutional networks," 2017.

[25] Elias Pampalk, "A matlab toolbox to compute music similarity from audio.," in *Proceedings of the 5th International Society for Music Information Retrieval Conference (ISMIR 2004)*, Barcelona, Spain, Oct. 2004.

[26] R EBU-Recommendation, "Loudness normalisation and permitted maximum level of audio signals," 2011.

[27] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*. Citeseer, 2015, vol. 8, pp. 18–25.

[28] Dimitrios Giannoulis, Michael Massberg, and Joshua D Reiss, "Digital dynamic range compressor design—a tutorial and analysis," *Journal of the Audio Engineering Society*, vol. 60, no. 6, pp. 399–408, 2012.

[29] Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer, "Madmom: A new python audio and music signal processing library," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 1174–1178.