Analysis of Musical Dynamics in Vocal Performances

Jyoti Narang, Marius Miron, Xavier Lizarraga, and Xavier Serra

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
{jyoti.narang, marius.miron, xavier.lizarraga,
xavier.serra}@upf.edu

Abstract. Dynamics are one of the fundamental tools of expressivity in a performance. While the usage of this tool is highly subjective, a systematic methodology to derive loudness markings based on a performance can be highly beneficial. With this goal in mind, this paper is a first step towards developing a methodology to automatically transcribe dynamic markings from vocal rock and pop performances. To this end, we make use of commercial recordings of some popular songs followed by source separation and compare them to the karaoke versions of the same songs. The dynamic variations in the original commercial recordings are found to be structurally very similar to the aligned karaoke/multi-track versions of the same tracks. We compare and show the differences between tracks using statistical analysis, with an eventual goal to use the transcribed markings as guiding tools, to help students adapt with a specific interpretation of a given piece of music. We perform a qualitative analysis of the proposed methodology with the teachers in terms of informativeness and accuracy.

Keywords: Vocal Performance Assessment, Music Education, Loudness Measurement, Dynamics Transcription

1 Introduction

Musical expression is an integral part of any performance. The subjective nature of this term makes it difficult to identify "whether the expressive deviations measured are due to deliberate expressive strategies, musical structure, motor noise, imprecision of the performer, or even measurement errors" [1]. While the choice of expressions used may vary from performer to performer and also from performance to performance, deriving the expressions used in a specific interpretation of a performance can offer significant advances in the realm of music education. Not only can it help students learn from a specific musical piece, insights about the variations in expressions can add to possible set of choices that one can employ during a performance.

With the advent of online practice tools like music minus one, audio accompaniments, users have a wide variety of mediums to chose to practice with [2]. However, most of these tools are limited to pitch and rhythm correctness, offering little or no insight about the expressive variations of the performance. In this work, we focus on deriving the dynamic variations of vocal rock and pop performances via loudness feature extracted from the audio recordings. The goal of this paper is to develop a methodology to extract and compare the dynamic variations of similar pieces of vocal performances that can lay the foundation of transcribing dynamic markings of vocal performances.

This overall idea can be broken down into a set of 2 questions that we intend to address through our work.

(i) Given a mix, is it possible to transcribe dynamics using the source separated voice signal with the same accuracy as would be achieved when the vocal stem of the mix is available?

(ii) Can we analyze the similarities and differences between two loudness curves in order to provide feedback on dynamics?

In order to address the first question, we use state of the art source separation algorithms to extract vocal tracks from mixes followed by loudness computation, and compare them to the loudness curves of the vocal stems available for the same mix. To address the second question, we have conducted a preliminary experiment comparing the loudness curves of the source separated commercial mixes with multi-track karaoke versions with vocal stems. Overall the structure of the paper is as follows. Section 2 presents some fundamental information about the kind of loudness scales and the study of dynamics in music information retrieval. In section 3, we describe a methodology of the proposed approach followed by preliminary investigation of the comparison of loudness curves in section 4. The influence of vocal source separation on loudness computation is also presented in section 4.

In section 5, we conduct a case study where the dynamic variations of the two versions (karaoke and commercial) have been analyzed by a teacher to give feedback followed by section 6 with conclusions and future work.

2 Background and Related Work

Significant work has been done to model performance dynamics by measuring the loudness variations [3] with a conclusion that the variations in dynamics are not linear. Several measurement techniques have been defined to measure the loudness of signals.

2.1 Loudness Measurement Scales

Of the scales available for loudness measurement, some are inspired by the subjective psychoacoustic phenomenon of human ear, while others are objective in terms of measurement. The most commonly used measurement is the dBFS scale, or loudness unit full scale. The more recently adopted industry standard is the EBUR scale [9]. For our analysis, we make use of the sone scale, which is based on psychoacoustic model, and compare our results to RMS values computed from the signals directly.

Sone Scale This scale is inspired by the psychoacoustic concept of equal loudness curves, with the measurement being linear i.e. doubling of the perceived loudness doubles the sone value [10]. While the phon scale is more closely associated with dB scale, a phon value of 40 translates to 1 sone. The relationship between phons and sons can be modelled using the equation:

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

RMS RMS or root mean square is the square root of the mean square of the amplitude of the signal.

$$RMS = sqrt((x_1^2 + x_2^2....x_n^2)/N) \tag{2}$$

2.2 Dynamics in Music Information Retrieval

Work on measurement of dynamics has been typically centered around Western Classical piano performances, incorporating dynamics as an expressive performance parameter that can vary across performers/performances [4]. Kosta et al. [5] used change-point detection algorithm to measure dynamic variations from audio performances and compared them to the markings in the score. Further, they applied machine learning approaches like decision trees, support vector machines (SVM), artificial neural networks [6] to predict loudness levels corresponding to the dynamic markings in the score. They found that the loudness values can be predicted relatively well when trained across recordings of similar pieces, while failing when trained across pianists' other performances.

Another approach to model dynamics is using linear basis functions to encode structural information from the score [8]. Each of the "basis function" stand for one score marking like *stacatto*, *crescendo*, the active state being a representation of the expressive marking present in the score and vice-versa. Chacón et al. [7] carry out a large scale evaluation of expressive dynamics on piano and orchestral music using linear and non-linear models.

3 Methodology

A diagram of the proposed methodology is presented in Figure 1. In case solely the mix is available, the input audio mix is passed to a source separation algorithm, U-Net [16] to get the separated vocal track. Thereafter, we extract the loudness from the separated vocal track or vocal stem using the sone scale and RMS as described earlier. The loudness extraction for the sone scale is carried out in the same way as proposed by Kosta et al [5] in their analysis. Each of the loudness curves are normalized by dividing with the max value for the rendition in order to carry out a fair relative comparison between different renditions. This step makes sure that only the relative values are compared and not the absolute ones. Finally, we apply peak picking operation to get a range of overall dynamics that can be further processed to map to specific dynamics based on musicological knowledge. It is to be noted that we limit the current set of experiments to comparison of loudness curves, leaving the actual mapping of loudness values to musically meaningful values as future work.

4 Experiments

4.1 Data Curation

We have primarily used three sources of data for our analysis:

(i) Commercial official recordings of rock and pop songs

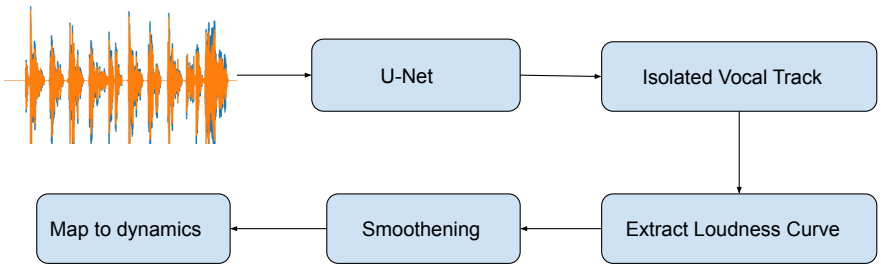

Fig. 1: Methodology for extracting loudness from a mix.

(ii) Custom karaoke tracks from the site[1] exactly replicating the official tracks
(iii) Musdb dataset to validate the efficacy of source separation algorithm

To evaluate the impact of singing voice source separation we use the musdb dataset containing 150 multi-track songs. For the commercial recordings, we conducted a preliminary investigation with 7 popular tracks shown in Table 1.

For the commercial popular recordings, only the mixes are available while for the karaoke versions, we have access to all the stems. This leads to 3 sources of data for the analysis of the same tracks - source separated vocals from the commercial mix (CSS), source separated vocals from the karaoke mix (KSS), vocal stems from the karaoke stems (KSV).

4.2 Experimental Setup

As mentioned above in the methodology, we first apply source separation using the spleeter implementation of UNet [13] to separate the mix into two stems - vocal track and the accompaniment. This step is skipped in case vocal stems are available for analysis. 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 follow the same block and hop size for the sone scale as well as RMS values. For loudness extraction using the sone scale, we use ma_sone function in Elias Pampalk's Music Analysis toolbox [11] in Matlab. The RMS values are extracted using the essentia library [15]. We further apply smoothening operation using two methods - "loess" with *smooth* function in matlab (based on locally weighted non-parametric regression fitting using a 2nd order polynomial) and exponential moving average [19][EMA]. Based on experimental testing, we use a span of 5% for the loess method. With the exponential moving average smoothening, we use an attack of 2 ms and release time of 20 ms. In the current set of experiments, the RMS smoothening is carried out using EMA methodology, and sone scale is smoothened using loess method. This operation was followed by peak picking operation to get a sense of overall dynamics followed. The peak picking parameters were experimentally set to a threshold of 0.1, and a peak distance of 1.2 seconds. We used the madmom library [14] for peak picking operation with RMS, and findPeaks function in maltab with sone scale loudness extraction. Figure 2 and Figure 3 show an example of computation of loudness

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

value using Sone scale and RMS respectively, followed by smoothening operation and detected peaks for the song ‘Don’t know why’ by Norah Jones.

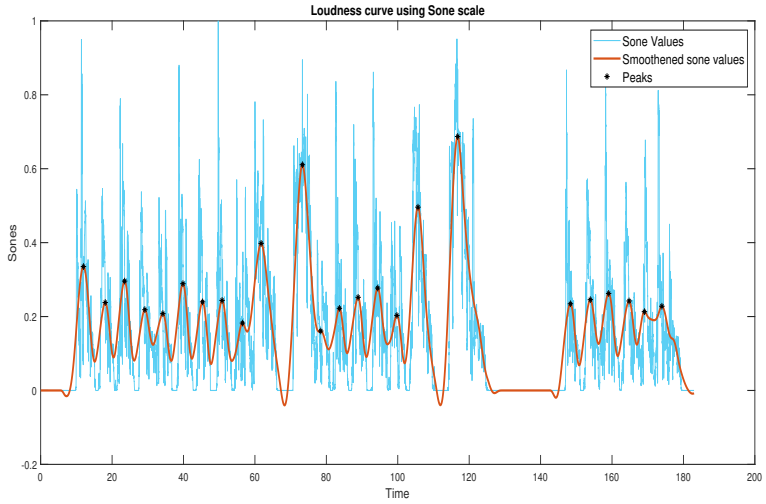

Fig. 2: Loudness using sone scale for Don’t Know Why by Norah Jones

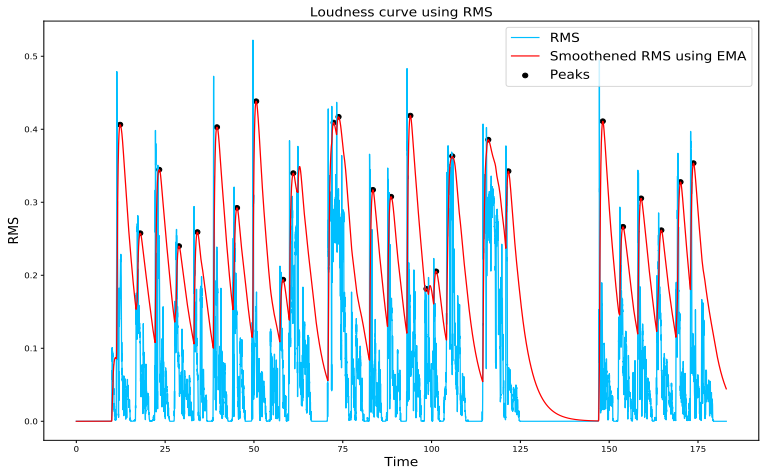

Fig. 3: Loudness using RMS values for Don’t Know Why by Norah Jones

4.3 Results

Overall Loudness Comparison between Renditions In order to compare the structure similarity of the loudness curves, we computed Pearson Correlation Coefficient of the smoothened curves extracted from the audio signals. Table 1 shows the values observed for each of the 7 songs. As evident from the table, most values are greater than 0.8, and in the case of comparing source separated version with the clean karaoke version, most values are greater than 0.9 indicating the robustness of the methodology with the pre-processing step of applying source separation.

Local dynamics To account for local dynamic changes, we compute the differences between consecutive peaks and derive a histogram from all the local differences. Further, the computed peak differences for each song are combined together for all songs from the same source i.e. commercial source separated, karaoke source separated and karaoke stem vocal. Thereafter, we use the non-parametric Kolmogorov-Smirnov 2 sample test which fits the properties of our data. This test is computed between each pair of the 3 histograms corresponding to the 3 sources. We find that for each of the comparisons, the p-value was 0.99 indicating no statistically significant differences between the histogram plots. These results are in line with our initial claim that the overall structure of the local dynamics changes as reflected in the loudness curves. These analysis results were the same for the histograms obtained using RMS values and sone values.

Table 1: Chosen songs and Pearson Correlation Coefficients for smoothened loudness sone curves

Song Name	Artist	CSS, KSV	KSS, KSV	CSS, KSS
Skyfall	Adele	0.867	0.994	0.931
Torn	Natalie Imbruglia	0.701	0.946	0.800
Fade into you	Mazzy Star	0.943	0.887	0.897
Imagine	John Lennon	0.889	0.981	0.440
Say you won't let go	James Arthur	0.955	0.835	0.800
Don't know why	Norah Jones	0.866	0.997	0.870
Son of a preacher man	Dusty Springfield	0.701	0.957	0.669

Global Dynamic Range The global dynamic range of each of the songs is computed using difference in max peak and min peak extracted from the smoothened loudness curve. As indicated in Table 2, the observed global dynamic range based on peak values are mostly similar in the case of karaoke source separated version and the karaoke vocal stem version with the exception of the song 'Son of a preacher man' with RMS values, and 'Fade into you' with sone values.

Outlier Analysis With a deeper analysis of the song 'fade into you', we find that there is a guitar section in the original song that becomes an artifact in the source separation output. This leads to a peak being wrongly detected increasing the overall dynamic

Table 2: Observed dynamic range with RMS and sone values

	RMS			Sone		
Song Name	**CSS**	**KSS**	**KSV**	**CSS**	**KSS**	**KSV**
Skyfall	0.460	0.156	0.176	0.503	0.477	0.489
Torn	0.092	0.138	0.206	0.355	0.199	0.213
Fade into you	0.144	0.195	0.167	0.306	0.354	0.182
Imagine	0.172	0.149	0.171	0.320	0.287	0.271
Say you won't let go	0.187	0.138	0.142	0.272	0.190	0.199
Don't know why	0.256	0.222	0.217	0.526	0.489	0.462
Son of a preacher man	0.150	0.227	0.371	0.275	0.339	0.295

range for both CSS and KSS resulting from peak detection. A high value of Pearson Correlation Coefficient between CSS and KSS as compared to KSS and KSV reflects from the fact that both of them have source separation as a pre-processing step, and both the versions contain similar artifacts.

4.4 Influence of voice source separation on loudness computation

In order to validate the efficacy of the source separation algorithm prior to using it for evaluating dynamics, we computed the Pearson Correlation of the smoothened loudness curves extracted from the mix with the smoothened loudness curves of the vocal stem tracks available with the musdb dataset [17].

As evident from the histogram in Figure 4, 138 values of the 149 songs evaluated are greater than 0.90. There are 6 songs with values between 0.80 and 0.90, and only 1 song with a value less than 0.50. The mean of the values is 0.960 and the standard deviation is 0.081. These results look promising to be able to use source separation as a prior step for dynamics analysis.

Outliers The song with the lowest value of correlation coefficient "PR-Happy Daze" contains a lot of instrumental music without much vocal component. Hence, the output of source separation algorithm is mostly artifacts. The song "Skelpolu - Resurrection" with a correlation coefficient of 0.58 has similar challenges.

5 Discussion

Work on transcription of dynamics is a challenging task for several reasons. One of the primary reasons being lack of sufficiently annotated data for singing voice to validate the efficacy of these algorithms.

Hence, in order to validate our approach, we conducted a case study with the song 'Don't know why by Norah Jones' where we asked a teacher with 6 years of Western singing teaching experience to compare the two tracks and provide feedback on the dynamic changes. Following is the feedback that we received from the teacher for some phrases of both tracks.

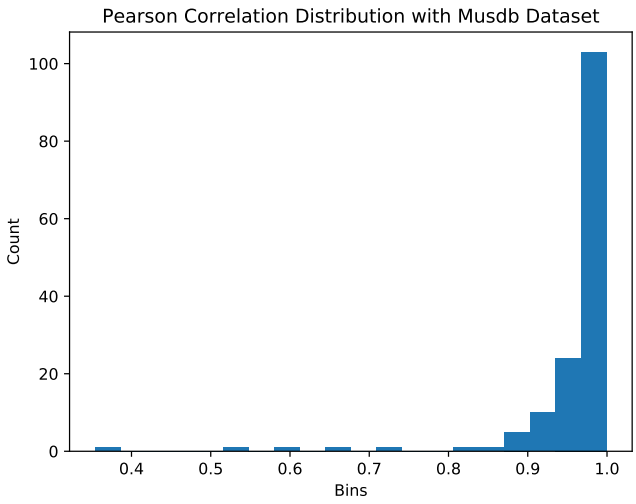

Fig. 4: Distribution of Pearson Correlation Coefficient applied to smoothened loudness curves of musdb dataset

I waited 'til I saw the sun

For Norah's Version:"Norah's dynamics change over the line. "I've" is 'mp'. "Waited till" starts as 'mf', which gradually drops down to 'mp' as she ends the line, can be seen as a diminuendo." For the Backing Track Version: "Dynamically, the singer is 'mf' throughout. This sounds like the kind of vocal take where the original vocals have been compressed one too many times."

I don't know why I didn't come

For Norah's Version: "Dynamically between an 'mp' and 'mf'". For the Backing Track Version: "Once again at an 'mf'. Vocals have definitely been compressed to sound at the same level consistently".

Case Study Results As evident from the first phrase, the teacher claimed that Norah Jones used a wider range of dynamics in her performance as compared to the cover version. Figure 5 shows the loudness curve of the cover version along with Norah Jones version using the sone scale. The classified dynamic markings for the two renditions are shown in the same plot. As compared to Norah's version of the same song, there is definitely a relatively very low difference between consecutive initial peaks in the cover version. The global dynamic range observed in the results section for this song is also in line with this observation. Similar results can be seen with RMS computation.

Challenges Despite having noisy artefacts and interferences from other instruments, state of the art source separation may be adequate for music analysis, when extracting dynamics. However, the peak detection method may not be robust enough to different

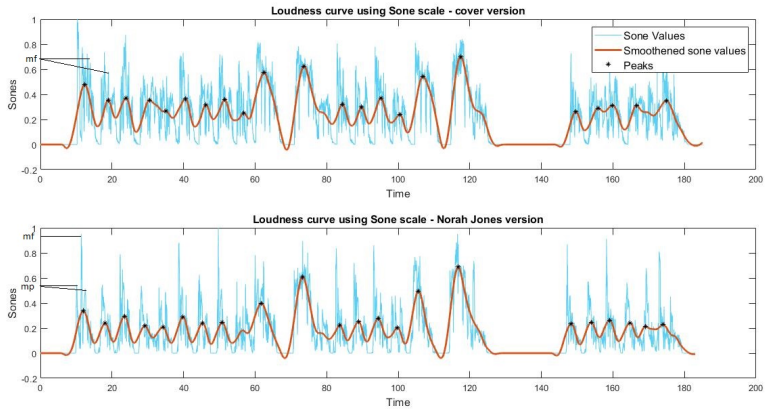

Fig. 5: Loudness using sone scale for Don't Know Why

performances and require calibration. Smoothing should be done w.r.t the tempo of the song.

While our initial case study showed some promising results, scaling such a system is still a very cumbersome task. Apart from the limitations with data and annotations, we are constrained by the knowledge that can help us realize the right granularity of transcription. For example, expressive markings like crescendo and diminuendo are associated with phrase boundaries [18], but the reverse might not be true. We would need collaborative efforts from multiple fronts in order to take advantage of the recent advances in the field of audio signal processing.

6 Conclusion and Future Work

We presented a methodology to extract dynamics from a performance using loudness as a feature. In the current investigation, we found that it is possible to use these loudness metrics to reach a level of relative changes that can in turn be mapped to dynamics. In future, we intend to discretise these relative values to map them to musically meaningful terms that can be used for providing the right feedback to students. Apart from that, in order to realize the overall goal of transcription, we intend to continue annotations of popular songs and further apply data driven approaches of machine learning to automatically derive the dynamic markings.

We also intend to apply the current methodology to student recordings to validate the efficacy of the system, and if the approach can be used to provide feedback on dynamics to students.

Acknowledgements We would like to thank Ajay Srinivasmurthy and Divakar Nambiath for their invaluable contributions to this work. Part of this research is funded by the projects 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)) and NextCore (RTC2019-007248-7 funded by the

Spanish Ministerio de Ciencia, Innovación y Universidades (MCIU) and the Agencia Estatal de Investigación (AEI)).

References

1. Langner, Jörg, and Werner Goebl. "Visualizing Expressive Performance in Tempo—Loudness Space." Computer Music Journal 27.4 (2003): 69-83.
2. Eremenko, Vsevolod, et al. "Performance assessment technologies for the support of musical instrument learning." (2020).
3. Berndt, Axel, and Tilo Hähnel. "Modelling musical dynamics." Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound. 2010.
4. Widmer, Gerhard, and Werner Goebl. Computational models of expressive music performance: The state of the art." Journal of New Music Research 33.3 (2004): 203-216.
5. Kosta, Katerina, Oscar F. Bandtlow, and Elaine Chew. "Dynamics and relativity: Practical implications of dynamic markings in the score." Journal of New Music Research 47.5 (2018): 438-461.
6. Kosta, Katerina, et al. Mapping between dynamic markings and performed loudness: a machine learning approach." Journal of Mathematics and Music 10.2 (2016): 149-172.
7. Cancino-Chacón, Carlos Eduardo, et al. "An evaluation of linear and non-linear models of expressive dynamics in classical piano and symphonic music." Machine Learning 106.6 (2017): 887-909.
8. Grachten, Maarten, and Gerhard Widmer. "Linear basis models for prediction and analysis of musical expression." Journal of New Music Research 41.4 (2012): 311-322.
9. EBU–Recommendation, R. "Loudness normalisation and permitted maximum level of audio signals." (2011).
10. Beck, Jacob, and William A. Shaw. "Ratio-estimations of loudness-intervals." The American journal of psychology 80.1 (1967): 59-65.
11. Pampalk, Elias. "A Matlab Toolbox to Compute Music Similarity from Audio." ISMIR. 2004.
12. Kosta, Katerina. Computational Modelling and Quantitative Analysis of Dynamics in Performed Music. Diss. Queen Mary University of London, 2017.
13. Hennequin, Romain, et al. "Spleeter: a fast and efficient music source separation tool with pre-trained models." Journal of Open Source Software 5.50 (2020): 2154.
14. Böck, Sebastian, et al. "Madmom: A new python audio and music signal processing library." Proceedings of the 24th ACM international conference on Multimedia. 2016.
15. Bogdanov, Dmitry, et al. "Essentia: An audio analysis library for music information retrieval." Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4-8; Curitiba, Brazil.[place unknown]: ISMIR; 2013. p. 493-8.. International Society for Music Information Retrieval (ISMIR), 2013.
16. Jansson, Andreas, et al. "Singing voice separation with deep u-net convolutional networks." (2017).
17. Rafii, Zafar, et al. "MUSDB18-a corpus for music separation." (2017).
18. Smith, Jeffrey C. Correlation analyses of encoded music performance. Stanford University, 2013.
19. Giannoulis, D., Massberg, M., & Reiss, J. D. (2012). Digital dynamic range compressor design—A tutorial and analysis. Journal of the Audio Engineering Society, 60(6), 399-408.