CHARACTERIZING MUSICAL VOCAL DYNAMICS IN PERFORMANCE

Jyoti Narang **Marius Miron** **Ajay Srinivasamurthy*** **Xavier Serra**
Universitat Pompeu Fabra, Spain
{jyoti.narang,marius.miron,ajays.murthy,xavier.serra}@upf.edu

ABSTRACT

Dynamics play a fundamental role in varying the expressivity of any performance. While the usage of this tool can vary from artist to artist, and also from performance to performance, a systematic methodology to derive dynamics in terms of musically meaningful terms like *piano, forte* etc can offer valuable feedback in the context of vocal music education. To this end, we make use of commercial recordings of some popular rock and pop songs from the Smule vocal balanced dataset and transcribe it with dynamic markings with the help of a music teacher. Further, we compare the dynamics of the source separated original recordings with the aligned karaoke versions to find the variations in dynamics. We compare and present the differences using statistical analysis, with a goal to provide the dynamic markings as guiding tools for students to learn and adapt with a specific interpretation of a piece of music.

1. INTRODUCTION

Several tools are employed by musicians to intensify the expressivity of any performance, one of them being dynamics or loudness variation. 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 [1], and several studies have been conducted analyzing the relationship between the dynamic markings in the score to the observed values of loudness in audio [2], particularly for the case of Western Classical piano performances [2–4]. However, not many studies have been conducted analyzing the role of dynamics in vocal performances [5].

The task of automatic transcription [6] of dynamics from audio can be particularly 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 [7] and pitch accuracy [8], 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 [9], or singing with karaoke tracks. Also, a system that can produce the dynamic range of a song based on audio analysis can facilitate the learners in song search and selection where they prefer to chose songs or artists based on their own dynamic range. However, lack of annotations and data for corresponding evaluation make the task particularly challenging.

In the current work, we focus on deriving the dynamic markings for vocal rock and pop performances from audio recordings. We first collaborate with a music teacher to annotate the dynamics of some select recordings which are part of the the Smule Vocal balanced dataset [10], and compare the markings with loudness contours extracted from audio recordings. For our analysis, we make use of the sone scale [11], which is based on a psychoacoustic model inspired by the human ear, and compare our results to RMS values computed from the signals directly.

In a previous work on similar lines [5], a methodology was devised to extract dynamics from audio via loudness features either from a mix or monophonic vocal audio recordings. To validate the approach, a case study was conducted where collaboration was carried out with a music teacher, asking him to transcribe dynamic markings from audio. It was found in the analysis that the markings by the teacher were in line with the changes in loudness features extracted from the audio. However, this was conducted for a small excerpt of a song. In the current work, the plan is to extend a similar analysis for the entire length of the song and also for multiple songs, which are part of the publicly available Smule vocal balanced dataset, with a goal to eventually test the approach with the student recordings available in the same dataset.

2. METHODOLOGY

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

2.1 Audio Synchronization

2.1.1 Audio to Score Synchronization

We first obtain the score of a given piece of music in XML format and transcribe it with dynamic markings with the

 *Ajay Srinivasamurthy is currently with Amazon Alexa, India.

help of a music teacher. Since the score is created using the original recording, we assume that the score is coarsely aligned with the audio. The end result of this step is an XML file that contains the pitch and dynamics information of the official recordings.

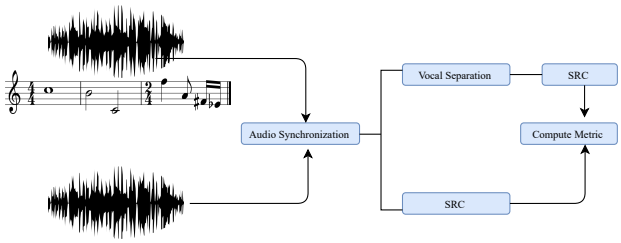

Figure 1. Methodology for extracting and comparing Dynamics

2.1.2 Audio to Audio Synchronization

Before extracting any of the features, we need to make sure that the rendition is aligned with the original audio recording. We assume that any of the renditions are performed with a backing track, and have the same length as the original track. However, there might be slight adjustments required to the offset position. The offset mostly depends on the song, and is verified individually from song to song.

2.2 Data preprocessing and feature computation

Each of the songs for our experiments are chosen from vocal learning and analysis perspective. We first collect the audio corresponding to the original songs from Youtube, viewing it as ground truth for our analysis. Further, we collect the karaoke versions of the same tracks performed by professional singers, with included stems. The choice of karaoke tracks of the same songs in our dataset helps us test if the reproduction of a track involves reproducing the dynamics of the original artist. Finally, once the dynamics of the original artist are obtained, we use the canonical score with dynamic markings for analysis of student recordings or amateur singers. However, we need access to monophonic vocal tracks in order to compute the loudness contours. Since the original recordings are available in the form of a mix, we use source separation as a preprocessing step to get isolated vocal tracks.

2.2.1 Source Separation

The recent progress in the field of audio source separation, especially for contemporary rock and pop music facilitated us to use it as an intermediate step. We validated the efficacy of this step in our previous work [5] with MusDB dataset [12], 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.

2.2.2 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 and RMS values of the signal for this step. The sone scale computation along with the consecutive smoothening operation is carried out in the same way as proposed by Kosta et al [13] in their analysis. Each of the curves are normalized by dividing by the max value to compare the relative values. Finally, we apply peak picking to get the overall dynamic range of different renditions.

Song Name	**Artist**	**PCC**
All of me	John Legend	0.95
Chandelier	Sia	0.92
Lost Boy	Ruth B.	0.48
Love Yourself	Justin Beiber	0.87
More than Words	Extreme	0.77
Say you won't let go	James Arthur	0.95
When I was your man	Bruno Mars	0.91

Table 1. Chosen Songs and Pearson Correlation Coefficent of loudness curves using Sone Scale

3. PRELIMINARY RESULTS

Metric Computation between different renditions is still a work in progress, but we present some preliminary results based on correlation values between the loudness curves.

To compare the structural similarity of the loudness curves between renditions, we compute the Pearson Correlation Coefficient (PCC) between the aligned smoothened curves of islolated vocal track of the corresponding YouTube recording, and vocal stem of the karaoke recording obtained from the website [1] . Table 3 shows the chosen songs from the smule dataset with the corresponding values.

With a deeper analysis of the corresponding values, we find that for audio tracks where source separation resulted in clean vocal stems, the correlation values were greater than 0.9. For the case of the song 'Lost Boy' by 'Ruth B.', it was found that the karaoke version was time stretched leading to a value lower than 0.50. The song 'More than words by Extreme' was labelled to be a very difficult song in terms of dynamics by the music teacher. A value of 0.77 for this song by the professional artist suggests perhaps that the karaoke artist used limited number of dynamics as compared to the original artist.

4. CONCLUSION

Work on dynamics transcription is a challenging task, primarily because of lack of annotated data for singing voice. Through our work, we intend to bridge this gap by providing some annotations on vocal dynamics by collaborating with a music teacher and devising a baseline methodology for extracting this expressive parameter from audio signals.

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

5. REFERENCES

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

[2] K. Kosta, O. F. Bandtlow, and E. Chew, "Dynamics and relativity: practical implications of dynamic markings in the score," *Journal of New Music Research*, vol. 47, pp. 438–461, 2018.

[3] N. 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.

[4] G. Widmer and W. 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.

[5] A.Anonymous, "Analysis of musical dynamics in vocal performances," *International Symposium on Copmuter Music Multidisciplinary Research*.

[6] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.

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

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

[9] V. Eremenko, A. Morsi, J. Narang, and X. Serra, "Performance assessment technologies for the support of musical instrument learning," 2020.

[10] I. Smule, "DAMP-VPB: Digital Archive of Mobile Performances - Smule Vocal Performances Balanced," Nov. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.2616690

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

[12] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "Musdb18-a corpus for music separation," 2017.

[13] K. Kosta, R. Ramírez, O. F. Bandtlow, and E. 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.