AUTOMATIC ESTIMATION OF SINGING VOICE MUSICAL DYNAMICS

Jyoti Narang[♭*] **Nazif Can Tamer**[♭*] **Viviana de la Vega**[♯] **Xavier Serra**[♭]

[♭] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

[♯] Escuela Superior de Música de Cataluña (ESMUC), Barcelona, Spain

jyoti.narang@upf.edu, nazifcan.tamer@upf.edu,

vivianadelavega@gmail.com, xavier.serra@upf.edu

ABSTRACT

Musical dynamics form a core part of expressive singing voice performances. However, automatic analysis of musical dynamics for singing voice has received limited attention partly due to the scarcity of suitable datasets and a lack of clear evaluation frameworks. To address this challenge, we propose a methodology for dataset curation. Employing the proposed methodology, we compile a dataset comprising 509 musical dynamics annotated singing voice performances, aligned with 163 score files, leveraging state-of-the-art source separation and alignment techniques. The scores are sourced from the OpenScore Lieder corpus of romantic-era compositions, widely known for its wealth of expressive annotations. Utilizing the curated dataset, we train a multi-head attention based CNN model with varying window sizes to evaluate the effectiveness of estimating musical dynamics. We explored two distinct perceptually motivated input representations for the model training: log-Mel spectrum and bark-scale based features. For testing, we manually curate another dataset of 25 musical dynamics annotated performances in collaboration with a professional vocalist. We conclude through our experiments that bark-scale based features outperform log-Mel-features for the task of singing voice dynamics prediction. The dataset along with the code is shared publicly for further research on the topic.

1. INTRODUCTION

Musical dynamics, such as *piano* and *forte* [1], are key elements in adding expressiveness to the singing voice [2]. They enhance overall performance and facilitate the conveyance of the desired emotional impact [3]. Despite extensive research on the singing voice, the analysis of dynamics in this context has received limited attention for several reasons. Firstly, annotating dynamics is an expensive process that requires repeated listening to audio tracks to accurately identify the dynamics category. Secondly, unlike other musical features such as pitch or tempo, the categorization of dynamics is not clearly defined, and even the same annotator may interpret a piece differently on multiple listens. Finally, a significant challenge for modern deep learning applications is the lack of reliable, existing dynamics based annotated datasets that can be used for the development of automatic analysis systems [4].

Despite the challenges of dynamics-based annotations for the singing voice, investigating dynamics in singing performances is worthwhile. On one hand, dynamics are a key component of expressivity in a music performance [5, 6]. On the other hand, dynamics are also an integral part of the music writing tradition [1, 7]. The use of dynamics in Western classical music evolved significantly from the Baroque period to the Romantic era. Particularly during the Romantic era, when expressivity became prominent, the annotation of dynamics alongside the score became widespread and accepted as part of the composition process. Composers frequently utilized symbols such as *forte*, *piano*, *crescendo*, and *diminuendo* to convey their desired variations in musical dynamics, and adhering to the dynamics instructions given by the composers became an important part of a Classical music performance.

While dynamics is a musical concept, its automatic estimation for music performance analysis relies on properties derived from audio signals. The audio characteristic most similar to musical dynamics is loudness or perceptual intensity. However, the mapping of musical dynamics to audio-based features from Music Information Retrieval (MIR) technologies is still not clearly understood. Extensive research exists on dynamics and tempo as expressive dimensions for Western classical piano performances [8]. However, unlike piano, there are almost no publicly available dynamics-based annotated datasets for the singing voice, which hinders the development of such technologies for the vocal performance analysis.

In this work, we propose to take advantage of the existing OpenScore Lieder corpus to curate a dataset of vocal performances with dynamics annotations, using state-of-the-art source separation and alignment as intermediate steps [1]. Furthermore, we curate a dataset of 25 other performances of different genres annotated manually by a professional Classical vocalist to test the model. At the end,

*These authors contributed equally to this work.

[1] https://github.com/MTG/SingWithExpressions.git

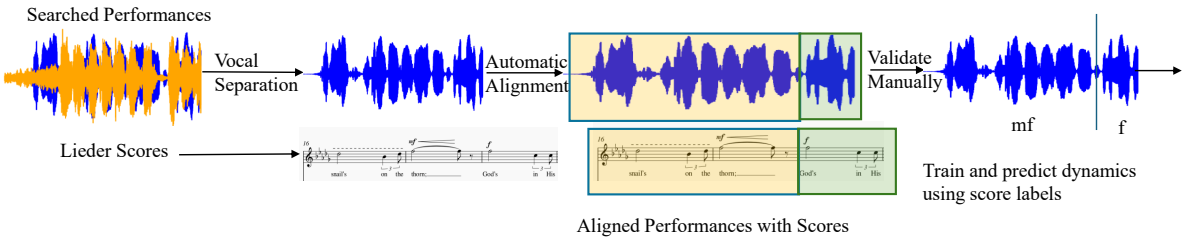


Figure 1: Data Preparation Pipeline: Corresponding to the Lieder scores from OpenScore Lieder Corpus, we apply Vocal Separation followed by Automatic Alignment. Finally, we validate the aligned score-performance data using Visualizations

we study the relationship between score based musical dynamics to perceptually motivated audio features [9] like log-Mel and bark-scale based features, testing the model with different analysis window-size, and genres of the test dataset.

Figure 1 illustrates the overall pipeline of the task. Using the meta-data information of the repository accompanying Lieder corpus, we start with searching for corresponding performances on YouTube. Further, we apply vocal separation on the performance to get vocals. Thereafter, using state-of-art alignment techniques, we align the corresponding score with the performance. At this stage, to test the accuracy of the alignment process, we develop visualization to filter out performances with mismatched aligned scores. Using the aligned score and performance data, we train a model for estimating dynamics based markings for an unknown performance.

The rest of this paper is structured as follows. In section 2, we cover the related works. Section 3 describes the dataset and the curation process. In section 4, we describe the experiments conducted with the curated data, followed by discussion and future work.

2. RELATED WORK

Although musical dynamics has been a topic of investigation in several studies [6, 7, 10, 11], especially for the case of piano [8, 12–14]there remains a notable gap in research concerning standalone musical dynamics analysis for the case of singing voice, particularly from an MIR perspective. Despite this gap, dynamics form a fundamental aspect of analysis within the interconnected fields of singing voice synthesis [15] and voice pedagogy [5].

In Singing Voice Synthesis (SVS) systems, dynamics play a crucial role in conveying expressive nuances [16]. Typically, dynamics are modelled as measures of energy in the signal [15–17] at the frame level. However, while there exists a close correlation between energy of the signal and musical dynamics, the influence of other parameters, such as pitch and timbre [10], remains largely unexplored. Understanding the relationship between pitch, timbre and dynamics could lead to more realistic representations of musical expression in SVS systems.

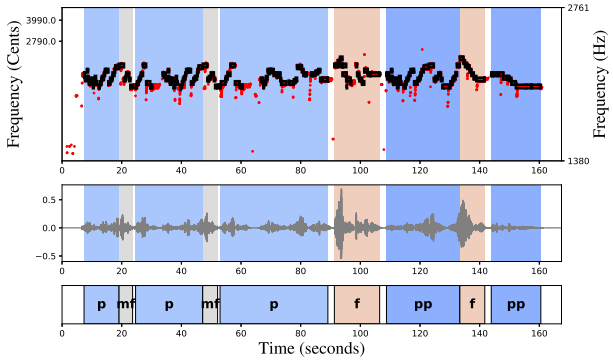
Bous and Roebel [4] explore the relationship between musical dynamics and timbral characteristics of the singing voice, employing mel-spectrogram features. Their experiment involves modifying the singing voice dynamics using a neural auto-encoder to transform voice levels. Effectiveness is assessed through evaluating perceived changes in voice level in the transformed recordings. However, a significant challenge arises as there is currently no reliable labels to determine the perceived changes in musical dynamics corresponding to "voice-level" changes as proposed in the system.

Narang et al. [18] utilize perceptually-motivated *sone* scale, comparing loudness curves of different professional renditions and student renditions for "musical dynamics" comparison following the methodology outlined by Kosta et al. [12] for comparing musical dynamics in piano. However, the study encountered limitations due to the lack of dynamics annotated datasets for evaluation.

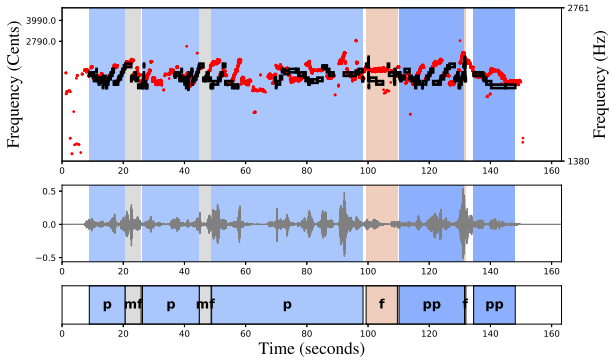
While there are some aspects of the research on Vocal Pedagogy [5] that has been utilized for the case of singing voice research from an MIR perspective, for example, Phonation mode [19] dataset or VocalSet [20] (which also contains some singing voice dynamics annotations but confined to vowel renditions), research outcomes of the vocal pedagogy remain largely unexplored by the MIR community. One direction is the role of voice source in singing voice, or how the positioning of the diaphragm affects vocal characteristics [21]. A study on vocal dynamics can help infer the voice source characteristics that can directly aid in vocal pedagogy.

3. DATASET

Dynamics are considered to be the most commonly manipulated parameter of an expressive performance and research investigations show that professionals or experts have much better control in expressive parameters in comparison to novice performers [6]. Further, songs from the 19th century Romantic era of Western classical music are widely known to be rich in expressive parameters. Drawing inspiration from this notion, we curate a dataset comprising professional renditions of 19th-century songs sourced from the OpenScore Lieder corpus [22]. Notably, composers often embed numerous dynamic markings within their scores, laying a foundational framework conducive to the analysis of dynamics.



(a) Accepted Performance Visualization



(b) Rejected Performance Visualization

Figure 2: Example visualization after automatic alignment on "The Shepherds Song" by Edward Elgar; For each sub-figure: red dots represent f0 using crepe, black dots represent note-information from the score (top), audio waveform (middle), dynamics information from the aligned score after automatic alignment (bottom)

3.1 Training Dataset Curation Process

3.1.1 Score Sources

Lieder Scores is a comprehensive collection of over 1200 19th century songs encoded over several years [22]. Within the Lieder dataset, we capitalize on two specific resources to facilitate our data curation process:

- The GitHub repository of Lieder provides MSCX files along with batch-conversion script to convert to MusicXML, enabling further processing with tools such as music21 [23]
- In the metadata section of the Lieder scores, a comprehensive compilation of composers, score names, and their respective MuseScore IDs is provided. This rich metadata serves as a valuable resource during the performance collection stage, enabling efficient querying and selection of performances.

3.1.2 Filtering Criteria for Scores

From all the batch-converted MusicXML files, we filter all scores, focusing on those with more than 3 dynamics annotations, and containing only 3 streams of score data: vocal, piano left hand and piano right hand.

3.1.3 Performance Sources

For the identified scores with greater than 3 dynamics markings, we search for multiple corresponding performances on YouTube using the query term obtained from the meta-data information of the scores. We curate multiple performances of similar pieces with the intention of extracting general dynamics based expressive patterns from professional singers. Our aim is to glean insights into varied interpretations, as there is no singular correct rendition of a performance that strictly adheres to the score. Subsequently, we carefully listen to each performance, specifically selecting those featuring vocals accompanied solely by piano. It is to be noted that not all composers have available performance data; thus, our selection process initially prioritizes renowned figures such as Schubert, Schumann, Brahms or Debussy, and ones with greater than 10 dynamics annotations. Once having exhaustively searched for performances of these well known composers, we proceed to search for lesser known composers following similar criteria. We automate the download process by utilizing YouTubeDL batch download to acquire the identified performances. The method yields a final list of 970 performances comprising identified composers, performances, and their respective MusicXML score files with dynamics based annotations.

3.1.4 Filtering Criteria for Performances

Following the filtration of scores and the manual curation of performance links, we advance to filtering performances suitable for the dynamics learning process. This process includes the following steps:

Source Separation Singing voices typically aren't presented in isolation. Even for solo performances, piano accompaniment is part of the performance. However, for our analysis, we require solo vocal renditions to accurately discern variations in performance dynamics. The initial step involves isolating the vocal component from the vocal-piano mix. This process, known as source separation, entails breaking a mixture into it's constituent components, and significant research has been dedicated to separation of vocals from the mix. We use Demucs v2 [24] to extract the vocals for the chosen songs. The robustness of using vocals resulting from source separation as an intermediate step was examined with the MusDB dataset [25].

Automatic Alignment To ensure that our curated performances can effectively serve as the basis for dynamics analysis, it's essential to achieve a basic alignment with the scores. Our approach to label creation draws inspiration from the methodology outlined by Tamer et al. [26, 27], who leverage Dynamic Time Warping (DTW) based music synchronization techniques [28] for creating pseudo labels in the realm of Violin transcription. Additionally, the concept of utilizing audio-to-score alignment as a preprocessing step for curating datasets in a semi-automatic manner for musicological endeavours was introduced in works by Weiss et al. [29], with a focus on the curation of Schubert's Winterreise dataset. While the works by Weiss et al. utilize MIDI-to-score alignment, we have chosen to

conduct the alignment using musicXML scores. This decision stems from the fact that dynamics information such as *piano, forte, crescendo, and diminuendo* can be less reliable in the process of MIDI conversion.

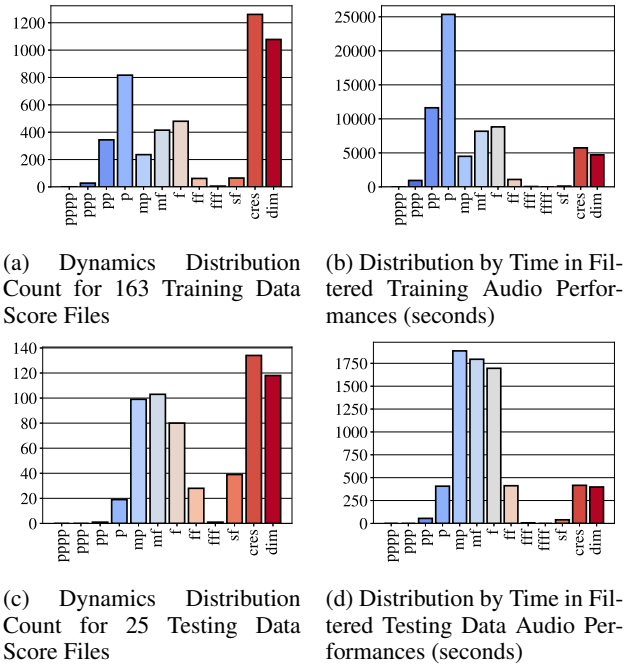
Manual Filtering using Visualizations The alignment stage yields a score with time information mapped to the corresponding performance files. Subsequently, we develop a visualization process utilizing fundamental frequency (f0) data extracted from performance files using CREPE [30] to validate the alignment between time-aligned performance and score files. Figure 2 showcases a sample visualization from the dataset. Figure 2a illustrates a performance that was accepted, and Figure 2b depicts a performance that we manually excluded during the selection process. The performance in Figure 2b was rejected because f0 curve from crepe (red dots) do not align with note-information from score (black rectangles) after automatic alignment, and hence the final labels lose reliability. The end result of this step is a comprehensive dataset of 509 performances for 163 aligned score files, which can be used to extract precise note-level expressive information from the score using tools like music21 [23].

3.1.5 Dynamics based Labels Extraction from Aligned Score Files

The aligned score files consist of all score-based information crucial for dynamics prediction. In this stage, we process the musical dynamics labels extracted using music21. Our approach adheres to the following principle: consecutive notes in the aligned audio are assumed to maintain similar dynamics unless there is a change in dynamics annotation in the score. When encountering labels like *sfz* or *sf* for a note, the value of the label of the consecutive note is assigned to be the dynamic value of the note preceding *sf* or related categories. This process results in a note-level mapping of 13 musical dynamics categories: *pppp, ppp, pp, p, mp, mf, f, ff, fff, ffff, sf, crescendo, diminuendo* directly extracted from the score. It is to be noted that we consolidate accent related categories, such as *sf*, *sfz* into a single category. Additionally, while we focus on musical dynamics for our task, the aligned score-performance data holds potential for various other Music Information Retrieval (MIR) tasks related to singing voice, including transcription, synthesis, or pedagogy.

3.2 Test Dataset Curation Process

For testing, we curated performances from a diverse selection of genres, ranging from operatic pop to theatre, R&B, or jazz, which lie outside the typical classical music domain. We collaborated with a Classical Vocalist, possessing over a decade of experience, to identify artists renowned for their wide vocal range. Once identified, we created reference scores for selected performances by these professional artists. The distribution of the genres in the selected pieces is as follows: pop(13), rock(12), jazz(3), soul(5), R&B(5), theatre(2) and other miscellaneous genres(5) including categories such as "post-disco", "acoustic" or "progressive rock", amongst others.



(a) Dynamics Distribution Count for 163 Training Data Score Files

(b) Distribution by Time in Filtered Training Audio Performances (seconds)

(c) Dynamics Distribution Count for 25 Testing Data Score Files

(d) Distribution by Time in Filtered Testing Data Audio Performances (seconds)

Figure 3: Dynamics Distribution across Train and Test Performances

3.2.1 Annotation Methodology

This section details the annotation methodology for dynamics-related markings of selected pieces as outlined by the musician: In the first listening, the piece's starting dynamic value is determined according to the dynamic markings such as: *pp, p, mp, mf, f, ff*, creating a reference point for each piece. This phase captures the most prominent features, recognizing that notation conveys more than mere amplitude. Subsequent listenings entail adding details, both in terms of dynamics and articulation of the text and musical phrases. Increased attention reveals additional layers of variation, often unnoticed during the first listening. In the third listening, decisions are made based on unification criteria. If different notations were used for the same musical effect in similar portions of the piece (e.g., different verses), the notation that best represents the musical intent is selected and unified with the rest. Rarely, genuine differences may exist between similar sections, in which case they are left distinct. In the final listening, no further notations are added. Instead, a mental musical reading of the entire work, from beginning to end, is undertaken. This involves elaborating on the interpretation following the written notations while simultaneously comparing it with the rendition produced by the artist.

3.2.2 Processing Methodology for Test Dataset

The processing methodology followed for the test dataset is similar to that of the training dataset, i.e., we apply source separation followed by automatic alignment to fetch the annotated labels using curated reference scores and performances.

Table 1: Results with Mel and Bark Features. Temporal resolution refers to the final feature rate after downsampling.

Seq Length	Temporal Resolution	Perceptual Feature	Acc	Acc($\pm$1)	Acc($\pm$2)
4096	17.4 ms	log-Mel	6.95	38.46	63.02
10000	29 ms	log-Mel	11.35	42.55	68.38
4096	16 ms	Bark	20.44	59.17	82.24
10000	30 ms	Bark	20.96	60.71	84.78

3.3 Dataset Statistics

Audio Statistics: The total duration of all the performances for the training dataset is 25.91 hours. The total duration of test files is 1.614 hours. The distribution of the labels as identified in dataset section 3 is illustrated in Figure 3. We observe that Lieder scores follow a relatively uniform distribution of dynamics with large number of dynamics annotations centered on a '*piano*'. And for the test dataset, the distribution curve is largely gaussian with majority of the distribution centered around *mp* and *mf*, which is not surprising considering the nature of pop music and mixing and mastering effects added to the final renditions.
Performance Count Per Piece: Although a single performer can deviate from the annotated score dynamics, having multiple performers per piece can help the model learn the general patterns closer to composer's intention. To leverage this effect, we collect performances with an average count of 3.12 performances per piece (std: 2.13), with a maximum of 12 performances for a piece by Robert Schumann. The average performance duration was observed to be 9.54 minutes (std: 9.36 minutes), with a maximum of 74.27 minutes for a piece by Franz Schubert and a minimum of 1.01 minutes for a piece by Peter Warlock.

4. EXPERIMENTS AND RESULTS

For the experiments outlined in this section, we utilize the curated dataset of Classical vocal performances for training and the dataset created in collaboration with the Classical vocalist for testing. We convert the note-level dynamics labels spanning from *pianississississimo (pppp)* to *fortissississimo (ffff)* into framewise labels encompassing 10 dynamics classes, and train and test our models for estimating the frame-wise dynamics. Thus, we consider dynamics estimation as a 10-class classification problem operating at the granularity of individual frames.
Input Representations: For model inputs, we consider two perceptually-motivated loudness features that are extracted after isolating the vocal tracks using DemucsV2 [24]. As our first input representation, we consider log-Mel features, which are commonly used in many audio and music processing tasks. These features are extracted using the librosa [31] library from audio sampled at 44.1 kHz using a hop size of 5.8 ms. As our second representation, we consider the specific loudness in Bark critical bands, which was previously studied in the context of piano dynamics [12] and singing voice loudness analysis [18]. The 240 dimensional Bark features are extracted using the MoSQITo library [32, 33], following the Zwicker loudness calculation method for time-varying signals [34] as specified in the ISO.532-1:2017 standard. The extraction process adheres to the default settings of a 48 kHz audio sampling rate and a 2 ms hop size.

Alongside these different input representations, we also study the effect of input sequence length and rate. To that end, we experiment with sequence lengths of 4096 and 10000. Since the original input representations have different temporal resolutions, we employ various downsampling rates to ensure that the models receive comparable feature rates during analysis. In our study with short context (4096 frames) dynamics modeling, we downsample the Bark features by 8 to operate at 16 ms, and downsample the log-Mel features by 3 to operate at 17.4 ms. For modeling dynamics detection using longer contexts (10000 frames), we downsample the log-Mel features by 5 to achieve a temporal resolution of 29 ms, and downsample the Bark features by 15 to achieve a comparable resolution of 30 ms.
Model Architecture and Training: For the frame-level estimation of dynamics, we employ a multi-scale Convolutional Neural Network (CNN) with self-attention [2] [35], originally introduced for the closely related task of framewise playing technique detection. In our implementation, the network receives input features with a fixed sequence length and outputs probabilities for 10 dynamics classes, with the class having the highest probability taken as the estimate. During training, we utilize the Adam optimizer with a learning rate of 0.002, aiming to reduce the Cross Entropy loss between the predicted dynamics classes and the aligned dynamics labels. We report our results on training the same network for different input representations, sequence lengths, and feature rates.
Metrics: One big challenge in the experimentation with musical dynamics is the subjectivity and relativity in its evaluation. For instance, one piece may span dynamics ranging from *pp* to *f*, and another piece may span dynamics ranging from *p* to *ff*. However, the measured loudness values of performances derived from both music pieces might be similar, as both sets of labels indicate a transition from relatively "soft" to "loud" dynamics. Therefore, the mapping between perceived performed dynamics and labeled musical dynamics may not be absolute. To address this challenge, we present the results in terms of exact match (Acc), relaxed accuracy 1(Acc $\pm$ 1) , and relaxed accuracy 2(Acc $\pm$ 2). Relaxed accuracy denotes that estimates are not penalized for a mismatch of 1 or 2 classes, respectively.

[2] Based on the modified version of `https://github.com/LiDCC/GuzhengTech99/blob/main/function/model.py`

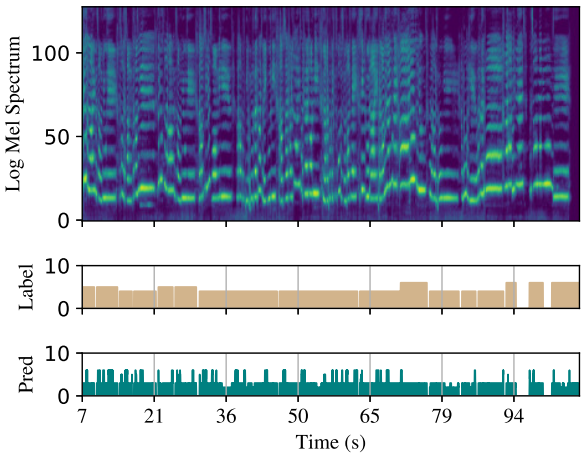


Figure 4: Model input and outputs for the log-Mel spectrum features. log-Mel-spectrogram (top), annotated labels by musician(middle), model estimates (bottom)

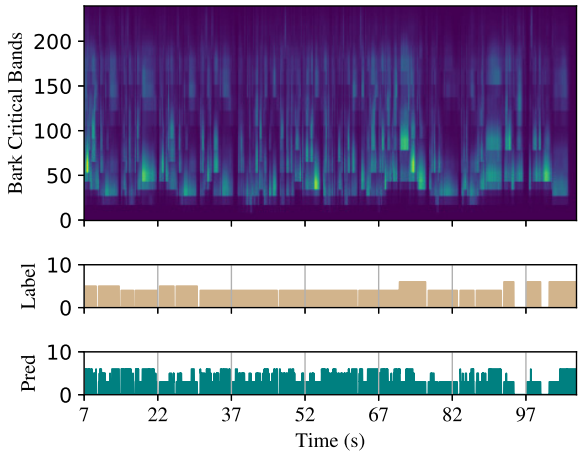


Figure 5: Model input and outputs for the bark based features: bark-critical-bands (top), annotated labels by musician (middle), model estimates (bottom)

4.1 Results

The results are summarised in Table 1. Despite the subjectivity of the task, we observe that the most confusion lie within the ± 1 or ± 2 range with significantly higher relaxed accuracies. Furthermore, we see that bark-based features outperform log-Mel features for the task. The highest relaxed accuracy ± 2 is achieved with bark-based features, indicating the models ability to differentiate between upper and lower bounds of dynamics. For example, a fortissimo is not classified to be a piano in almost 85% of the cases. An example prediction using log-Mel features and bark-based features for a theatre song "sound of music" is presented in Figures 4 and 5 respectively.

The effect of larger and smaller temporal contexts can also be seen in Table 1. Providing larger temporal contexts results in better performance for dynamics estimation. This effect is more prominent for log-Mel features compared to the Bark features. We found that the best performing model is the one with the entire song frames included in the context window i.e., the sequence length is long enough to encapsulate the whole song.

5. DISCUSSION

One of the primary challenges in predicting musical dynamics lies in the fact that performance information is available through recordings, which is a result of mixing and mastering. Consequently, the loudness information captured in recordings may diverge from performers original intentions. However, we contend that despite the influence of mixing and mastering, it is possible for musicians as well as non-musicians to infer whether a performer is singing softly, loudly or even shouting independent of raw loudness levels. Our approach leverages perceptually motivated features that encapsulate timbral characteristics, which have the potential to enhance musical dynamics estimation while remaining agnostic to variations in loudness levels.

While the labels are created semi-automatically, there are potential discrepancies due to performers not adhering strictly to the score or editors creating alternative versions of the score different from the one curated in the dataset.

Additionally, we have framed the dynamics estimation at an absolute level, with the expectation that the model will learn the variations in relative markings given a large amount of data. However, musical dynamics at any given time in a performance depend on the context rather than the absolute value of measured loudness [36]. Additionally, addressing class imbalance remains a significant challenge.

On software front, while MuseScore offers extensive annotation capabilities, some categories cannot be accurately modeled. To mitigate this, musicians often use note-level "TextExpressions" in MuseScore to add additional information. During our experimentation, we encountered terms like "sempre piano," "poco dolce," and "calando" that musicians add to the score. While we were able to mitigate challenges with some labels, achieving comprehensive coverage requires further collaboration with vocalists to refine the target labels.

6. CONCLUSION AND FUTURE WORK

We've developed a methodology for large-scale dataset curation focused on singing voice. The semi-automatically curated dataset serves as a valuable resource for tasks such as transcription, expression analysis, synthesis, and vocal pedagogy. It currently includes 509 performances aligned with 163 score files from 25 composers. Using this dataset, we trained a CNN with multi-head attention for dynamics prediction and found that bark-scale-based features outperform log-Mel features. To test the model, we curated score-performance dataset manually in collaboration with a Classical vocalist. Future work involves integrating pitch features with loudness features to enhance prediction accuracy, improving the model to address class imbalance, and expanding the dataset to include more composers.

7. ACKNOWLEDGMENTS

We would like to thank Ajay Srinivasamurthy for his invaluable feedback. IA y Música: Cátedra en Inteligencia Artificial y Música" (TSI-100929-2023-1), funded by the Secretaría de Estado de Digitalización e Inteligencia Artificial, and the European Union-Next Generation EU, under the program Cátedras ENIA 2022 para la creación de cátedras universidad-empresa en IA.

8. REFERENCES

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

[2] D. Fabian, R. Timmers, and E. Schubert, *Expressiveness in music performance: Empirical approaches across styles and cultures*. Oxford University Press, USA, 2014.

[3] R. Miller, *On the Art of Singing*. Oxford University Press, 1996. [Online]. Available: https://books.google.es/books?id=Sv1EAQAACAAJ

[4] F. Bous and A. Roebel, "Analysis and transformation of voice level in singing voice," in *Proceedings of ICASSP*. Rhodes Island, Greece: IEEE, 2023, pp. 1–5.

[5] J. Sundberg, "Perceptual aspects of singing," *Journal of voice*, vol. 8, no. 2, pp. 106–122, 1994.

[6] L. Bishop, F. Bailes, and R. T. Dean, "Performing musical dynamics: How crucial are musical imagery and auditory feedback for expert and novice musicians?" *Music Perception: An Interdisciplinary Journal*, vol. 32, no. 1, pp. 51–66, 2014.

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

[8] 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.

[9] A. Friberg, E. Schoonderwaldt, A. Hedblad, M. Fabiani, and A. Elowsson, "Using listener-based perceptual features as intermediate representations in music information retrieval," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 1951–1963, 2014.

[10] A. Elowsson and A. Friberg, "Predicting the perception of performed dynamics in music audio with ensemble learning," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 2224–2242, 2017.

[11] 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.

[12] 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, no. 5, pp. 438–461, 2018.

[13] D. Jeong and J. Nam, "Note Intensity Estimation of Piano Recordings by Score-Informed NMF," in *2017 AES International Conference on Semantic Audio*, Erlangen, Germany, Jun. 2017.

[14] L. Marinelli, A. Lykartsis, S. Weinzierl, and C. 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] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer modeling timbre and expression from natural songs," *Applied Sciences*, vol. 7, no. 12, p. 1313, 2017.

[16] M. Umbert, J. Bonada, M. Goto, T. Nakano, and J. Sundberg, "Expression control in singing voice synthesis: Features, approaches, evaluation, and challenges," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 55–73, 2015.

[17] T. Nakano and M. Goto, "Vocalistener2: A singing synthesis system able to mimic a user's singing in terms of voice timbre changes as well as pitch and dynamics," in *Proceedings of ICASSP*. Prague, Czech Republic: IEEE, 2011, pp. 453–456.

[18] J. Narang, M. Miron, A. Srinivasamurthy, and X. Serra, "Analysis of musical dynamics in vocal performances using loudness measures," in *Proceedings of the 25th International Conference on Digital Audio Effects (DAFx 2022)*. DAFx, 2022, pp. 33–39.

[19] P. Proutskova, C. Rhodes, T. Crawford, and G. Wiggins, "Breathy, resonant, pressed–automatic detection of phonation mode from audio recordings of singing," *Journal of New Music Research*, vol. 42, no. 2, pp. 171–186, 2013.

[20] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, "Vocalset: A singing voice dataset," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR 2018)*, Paris, France, Sep. 2018, pp. 468–474.

[21] J. Sundberg, "Research on the singing voice in retrospect," *TMH-QPSR*, vol. 45, no. 1, pp. 11–22, 2003.

[22] M. R. H. Gotham and P. Jonas, "The OpenScore Lieder Corpus," in *Music Encoding Conference Proceedings 2021*, S. Münnich and D. Rizo, Eds. Humanities Commons, 2022, pp. 131–136.

[23] M. S. Cuthbert and C. 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.

[24] A. Défossez, "Hybrid spectrogram and waveform source separation," in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, Nov. 2021.

[25] J. Narang, M. Miron, X. Lizarraga Seijas, and X. 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.

[26] N. C. Tamer, P. Ramoneda, and X. Serra, "Violin etudes: a comprehensive dataset for f0 estimation and performance analysis," pp. 517–524, 2022.

[27] N. C. Tamer, Y. Özer, M. Müller, and X. Serra, "High-resolution violin transcription using weak labels," in *Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR 2023)*, Milan, Italy, Nov. 2023, pp. 223–230.

[28] M. Müller, Y. Özer, M. Krause, T. Prätzlich, and J. Driedger, "Sync toolbox: A python package for efficient, robust, and accurate music synchronization," *Journal of Open Source Software*, vol. 6, no. 64, p. 3434, 2021.

[29] C. Weiß, F. Zalkow, V. Arifi-Müller, H. Grohganz, H. V. Koops, A. Volk, and M. Müller, "Schubert Winterreise dataset: A multimodal scenario for music analysis," *ACM Journal on Computing and Cultural Heritage (JOCCH)*, 2020, in press.

[30] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *Proceedings of ICASSP*, Calgary, Canada, 2018, pp. 161–165.

[31] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[32] G. F. Coop, "Mosqito," Feb. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10629475

[33] R. San Millán-Castillo, E. Latorre-Iglesias, M. Glesser, S. Wanty, D. Jiménez-Caminero, and J. M. Álvarez-Jimeno, "Mosqito: an open-source and free toolbox for sound quality metrics in the industry and education," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 263, no. 5. Institute of Noise Control Engineering, 2021, pp. 1164–1175.

[34] E. Zwicker, H. Fastl, U. Widmann, K. Kurakata, S. Kuwano, and S. Namba, "Program for calculating loudness according to din 45631 (iso 532b)," *Journal of the Acoustical Society of Japan (E)*, vol. 12, no. 1, pp. 39–42, 1991.

[35] D. Li, M. Che, W. Meng, Y. Wu, Y. Yu, F. Xia, and W. Li, "Frame-level multi-label playing technique detection using multi-scale network and self-attention mechanism," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[36] T. Nakamura, "The communication of dynamics between musicians and listeners through musical performance," *Perception & psychophysics*, vol. 41, pp. 525–533, 1987.