ID	Environment		Number of subjects assigned to each velocity level				
	Instrument	Acoustic environment	$v_{\text{ref}} = 32$	$v_{\text{ref}} = 44$	$v_{\text{ref}} = 60$	$v_{\text{ref}} = 80$	Total
Env. I	Grand Disklavier	Anechoic chamber	6	6	6	5	23
Env. II	Upright Disklavier	Non-anechoic chamber	7	7	7	7	28

Table 1: Experiment settings and subjects assignment in two environments.

We address the loudness modeling on the other pitches in section 4.

We set the number of RMLSP test round to be 32. In the first two test rounds, the subject always listens to two piano tones of constant variable velocities $v_{\text{var}} = 90$ and $v_{\text{var}} = 30$, respectively. After that, the model yields the next variable velocity uniformly sampled from a $[-6..+6]$ interval centered at the current-step estimation of the PSE. The order of the piano tone pairs between the reference stimulus and the variable stimulus is random at each test round.

The experiment is conducted on two player pianos located in different acoustic environments (as shown in Table 1). Environment I contains a grand YAMAHA Disklavier piano in an anechoic chamber, and Environment II contains an upright YAMAHA Disklavier piano in a non-anechoic chamber. Subjects are seated half a meter in front of the pianos where the pianists usually sit and use a laptop to respond.

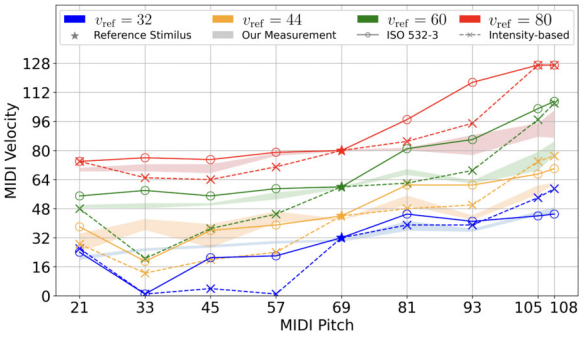
The subjects are 15 adults (18–35 years), 10 male and 5 female. All subjects have normal hearing sensitivity. 11 subjects participate in the experiment in Environment I and 9 subjects participate in the experiment in Environment II. Each subject is tested in a separate section.

The subject is tested at two or three of the four velocity levels randomly assigned (as shown in Table 1). For Environment I, 23 measurement results were collected. Each velocity level has a sample number of 6 except for velocity 80 with a sample number of 5. For Environment II, 28 measurement results were collected. Each velocity level has a sample number of 7.

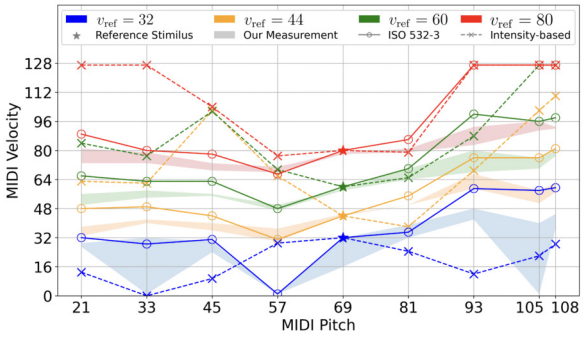
3.3 Result

Figure 1 shows our experiment results in Environment I and Environment II, respectively. Since MIDI velocity is an ordinal variable, it is improper to average the subjects' PSEs and show the mean equal-loudness curves. Instead, we present *equal-loudness ribbons*, where we replace the mean with the range of subjects' PSEs between the first and third quartiles among all data. Here, we use four different colors to indicate four reference velocities, and the diamond markers indicate the four reference stimuli on each graph. Note that in the lowest ($v_{\text{ref}} = 32$) ribbons, some of the variable pitches have PSEs equal to one, meaning that the variable pitches at the least audible velocity are still louder than or equal to the reference tone.

We see the patterns of equal-loudness ribbons are very different across the two environments. In Environment I, we see a gradual growth trend in all velocity levels, meaning higher pitches of the same velocity tend to be softer.



(a) Environment I.



(b) Environment II.

Figure 1: Measurement of equal-loudness ribbons of piano tone in two environments compared with ISO 532-3 [24,25] and intensity-based loudness computation [14,26].

In Environment II, the trend is less evident, and we see a valley at A3 (i.e., $p_{\text{var}} = 57$) at all reference velocity levels. Moreover, the range of equal-loudness ribbons varies, indicating the just noticeable difference in velocity change is uneven on different pianos, pitches and velocities.

3.4 Comparison with Existing Methods

We use the experiment result to validate two common loudness computation methods. The first method is ISO 532-3 by Moore *et al.* [24,25], the state-of-the-art loudness model based on the modeling of human auditory systems. The loudness value is computed as the maximum value of the predicted long-term loudness curve. The other method is used in [14,26] by naively computing the average intensity of the first 10 ms after the peak. For the two methods, we use the recordings of all the piano tones (discussed in section 4.3) and compute their corresponding loudness. The induced equal-loudness contours are shown in Figure 1 in solid lines with circles and dashed lines with crosses, respectively. Other popular methods such as applying A-

weighting is not considered since such methods are proved not generalizable to natural tone [32].

As shown in Figure 1(a) and Figure 1(b), in both environments, the results induced from ISO 532-3 are in general consistent with the equal-loudness measurement except for the pitch in the higher registers. The results induced from the intensity-based method fluctuate around the equal-loudness ribbons, failing to describe the perceptual effect.

4. LOUDNESS MODEL

In the previous section, we measure the equal-loudness contours (ELC) in two environments on 9 variable pitches and 4 reference velocities. In this section, we extend our findings to learn loudness models to account for loudness estimation of the remaining tones via a non-parametric approach as baseline (discussed in section 4.1) and a parametric approach (discussed in section 4.2). The parametric model is also capable of predicting loudness given the waveform of an arbitrary piano tone. We compare the prediction result with ISO 532-3 and the intensity-based method in section 4.3.

4.1 Non-parametric Method

We first propose a naive approach to predict the loudness of arbitrary piano tones in each environment via linear interpolation of the measured ELCs in section 3. First, we assign the loudness in *sone* to the tones of the reference pitch A4 under all the velocities using the calibration method discussed in section 4.2.3. Then, we linearly interpolate the ELCs measured on 9 variable pitches to all 88 pitches, and assign the same loudness level along each contour. Finally, for each pitch, we linearly interpolate between the ELCs. In section 5, we see such a simple method already yields satisfactory performance in downstream applications.

4.2 Parametric Method

Moreover, a parametric model is proposed to estimate the loudness using machine learning.

4.2.1 Model

We use $x = (p, v)$ to denote a piano tone, where p is the pitch and v is the velocity levels. We learn a parametric loudness model $\ell = f_\theta(x)$ to compute the loudness of a given piano tone x.

The loudness model is learned as a supervised classification problem (as shown in Figure 2). The input is a pair of piano tones (x_1, x_2), where $x_1 = (p_1, v_1)$ and $x_2 = (p_2, v_2)$. The target is the ground-truth label y, which is the indicator of whether x_1 is the louder one inferred from the ELCs. The difference between the estimated loudness ℓ_1 and ℓ_2 is used to predict the y. Specifically, the loss function is:

$$\mathcal{L}(\theta; x_1, x_2, y) = \text{BCE}(f_\theta(x_1) - f_\theta(x_2), y), \quad (1)$$

where $\text{BCE}(\cdot)$ is the binary cross entropy function.

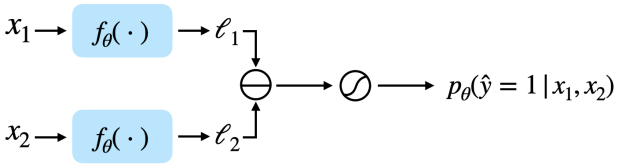


Figure 2: Illustration of the objective function.

The pairs (x_1, x_2) are sampled from all the tone pairs in each environment whose loudness comparisons are derivable based on the human subject results. Specifically, a pair must satisfy either of the two conditions:

(**C.1**) The two tones have the same pitch and thus can be compared based on velocity. The one with the larger velocity is assumed to be louder.

(**C.2**) The two tones have different variable pitches and the loudness comparison can be inferred from the equal-loudness ribbons in Figure 1.

4.2.2 Implementation

We define our loudness model $\ell = f_\theta(x)$ as the linear combination of the values on the mel-spectrogram of x (without an intercept term). The mel-spectrogram has 8 mel-frequency bins, a 2048 window size, and a 512 hop size under the sample rate of 22050 Hz. We select the 5 time frames (approx. 0.1s) right after the onset of the tone detected by [33]. In all, we have 80 input features of a pair of piano tones and the model can be optimized using logistic regression.

The small mel-frequency bin number in the current model can be understood as a constant convolution kernel to achieve better generalization on other non-variable pitches. In the future, the model can also be extended to neural networks given a larger amount of human subject results.

4.2.3 Calibration

The parametric method learns an unstandardized estimation of piano tone loudness based on classification, which needs to be further calibrated to the standard loudness unit in sone. We address the problem by matching the piano tones under reference pitch A4 to the corresponding equal-loudness pure tones and compute the corresponding pure tone loudness using ISO 532-3. The matching procedure follows from the *method of adjustment* [32].

4.3 Evaluation

4.3.1 Data Preparation

We record the piano tones of two Disklavier pianos in both environments for all pitches and velocities. The recorder is placed 0.5 meters in front of the piano (similar to the subjects' ear position) and is kept still during the recording procedure. For all piano tones, the MIDI duration is 0.3 seconds and each recording clip lasts 1.3 seconds.

To train the parametric model, we select all the piano tone pairs that satisfy the two conditions (defined in 4.2.1),

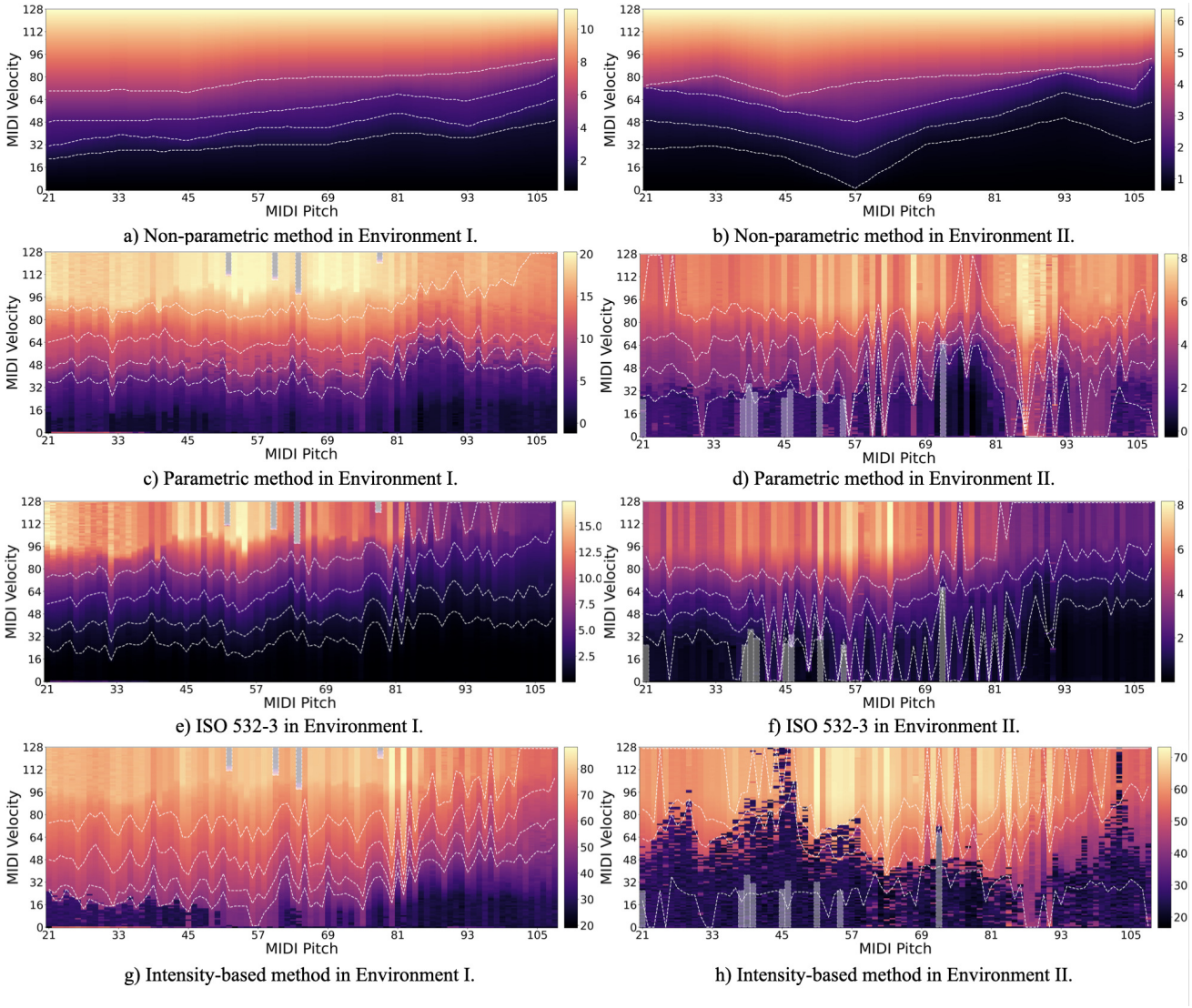


Figure 3: Visualization of calculated loudness value in Environment I and Environment II using our proposed loudness models, ISO 532-3 [24,25], and intensity-based method [14,26]. Points with a rectangular mask indicate mechanical failure of the player piano.

including 673,035 pairs satisfying the **C.1** and 427,633 pairs satisfying the **C.2** in Environment I, and 695,860 pairs satisfying the **C.1** and 422,386 pairs satisfying the **C.2** in Environment II. The dataset is randomly split into train set (80%) and test set (20%).

4.3.2 Results

Figure 3 shows the heatmaps of the predicted loudness of all piano tones in both environments by four methods: 1) our non-parametric method, 2) our parametric method, 3) ISO 532-3, and 4) intensity-based method. We show the ELCs derived from the heatmaps in dashed lines for better readability.

We see the non-parametric method presents a smooth interpolation of the measured ELCs. The result of the parametric method demonstrates a similar trend as ISO 532-3 except in the higher pitch range, which is more consistent with human subject measurement in Figure 1. Moreover, both parametric method and ISO 532-3 show more discontinuity than Figure 1, suggesting the MIDI velocity is not assigned in a uniform manner for different pitches. Finally, the result of intensity-based method are noisy and inconsistent with the human perception.

Table 2 shows the model accuracy on the test set of two conditions defined in 4.2.1 separately. The results demonstrate that our proposed method outperforms the two baselines in both environments, especially in the accuracy of **C.2**. We show the generalization ability in two ways. First, we apply the model trained in one environment to the other environment and achieves satisfactory accuracy (as indicated by the underlined numbers). Second, we combine the training data in both environments and train a hybrid model. The model outperforms the baselines and even achieves the best performance in environment II in both conditions. The evaluation result shows our proposed model capture certain features in the piano-tone mel-spectogram contributing to the perceptual loudness.

Methods	Acc. in Env. I		Acc. in Env. II	
	C.1	C.2	C.1	C.2
ISO 532-3	**0.9689**	0.9631	0.9037	0.9455
Intensity-based	0.9658	0.9290	0.8377	0.8555
PM - Env. I	**0.9689**	**0.9893**	0.8976	0.9614
PM - Env. II	0.9528	0.9607	0.9121	0.9793
Hybrid PM	0.9401	0.9785	**0.9370**	**0.9881**

Table 2: Evaluation of loudness comparison accuracy in Environment I and Environment II. The methods to compare are: 1) ISO 532-3: [24,25], 2) Intensity-based [14,26], 3) PM - Env. I: Parametric model trained on the data recorded in Environment I, 4) PM - Env. II: Parametric model trained on the data recorded in Environment II, and 5) Hybrid PM: Parametric model trained on the data recorded in both environments. Underlined data are tested by the parametric model training in the other environment.

5. PERFORMANCE CONTROL TRANSFER

In this section, we apply our theory of piano tone loudness to the downstream application of *performance control transfer*, in which we adjust the MIDI velocities on two different player pianos to achieve the same perceptual effect. An example scenario of the application is to reproduce a pianist's performance in the concert hall to one's living room.

5.1 Modification

Performance control transfer is originally proposed in [14], in which piano tone loudness is treated as a fundamental invariant property between the original performance and the transferred version. However, the study mistakenly uses the physical intensity as the loudness measure. Our approach replace the loudness estimator by the proposed loudness models (discussed in section 4) and keep the remaining algorithm the same. We also use ISO 532-3 as the loudness estimator as a baseline method.

5.2 Listening Test

We conduct a listening test to invite participants to evaluate the performance transferred by different algorithms, similar to the one conducted in [14]. Specifically, we prepare four pieces, including two monophonic pieces and two polyphonic pieces. The pieces are selected to cover classical and popular genres.

We invite two pianists to play the four pieces in Environment I and record both pianists' MIDI control and the performance audio. Then, we transfer the MIDI files recorded in Environment I to Environment II using different performance transfer methods, including 1) our non-parametric method, 2) our parametric method, 3) ISO 532-3, 4) intensity-based model (i.e., the original transfer method [14]), and 5) a raw transfer without any MIDI velocity editing. We play the transferred MIDI in Environment II and record them from the same microphone position discussed in section 4.3.

We invite people to subjectively rate the transfer quality through a double-blind online survey. During the survey, the subjects listen to four groups of samples. In each group, the original performance in Environment I is played, followed by the five transferred versions. Both the order of groups and the sample order within each group are randomized. After listening to each sample, the subjects rate them based on a 5-point scale from 1 (very low) to 5 (very high) according to three criteria: *faithfulness* (to the original performance), *naturalness* and *overall musicality*.

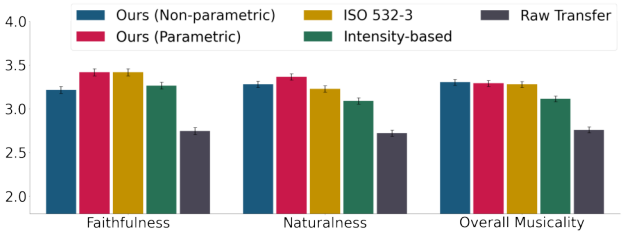


Figure 4: The subjective evaluation results of the five transfer methods.

5.3 Results

A total of 21 participants (14 male, 7 female) with different musical backgrounds have completed the survey. Figure 4 shows the result where the heights of the bars represent the means of the ratings and the error bars represent the confidence intervals computed via within-subject ANOVA [34].

The results show that all the loudness-based methods significantly outperforms the original intensity-based implementation and raw transfer (with p-value < 0.005). Moreover, our parametric method is marginally better than ISO 532-3 in terms of naturalness and overall musicality, and comparable with ISO 532-3 in faithfulness. Both our parametric and non-parametric methods demonstrate with only sparse human subject measurement data, we can achieve promising performance transfer results.

6. CONCLUSION

In this paper, we have contributed the first study to model the perceptual loudness of piano tones in a hybrid approach combining psychoacoustic experiments, and data-driven methods. Our theory include: 1) measurements of piano-tone equal-loudness contours on two player pianos, and 2) data-driven loudness models to estimate piano-tone loudness based on the measured contours. Experiments show our model provides more satisfactory estimation than existing loudness theories. Based on our findings, we validate the state-of-the-art loudness model ISO 532-3 on piano-specific tasks and argue that existing methods can be greatly improved if we have a more detailed human subject measurement and a deeper understanding of the timbral features of piano tone.

On the application side, we improve the existing performance control transfer method with the proposed loudness estimator. We believe the other downstream applications can also benefit from our study to propose a clearer problem definition and more accurate feature extraction.

7. REFERENCES

[1] H. Takeshima, Y. Suzuki, H. Fujii, M. Kumagai, K. Ashihara, T. Fujimori, and T. Sone, "Equal-loudness contours measured by the randomized maximum likelihood sequential procedure," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 389–399, 2001.

[2] Y. Suzuki and H. Takeshima, "Equal-loudness-level contours for pure tones," *The Journal of the Acoustical Society of America*, vol. 116, no. 2, pp. 918–933, 2004. [Online]. Available: https://doi.org/10.1121/1.1763601

[3] B. C. Moore, B. R. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *Journal of the Audio Engineering Society*, vol. 45, no. 4, pp. 224–240, 1997.

[4] E. Zwicker, G. Flottorp, and S. S. Stevens, "Critical band width in loudness summation," *The Journal of the Acoustical Society of America*, vol. 29, no. 5, pp. 548–557, 1957. [Online]. Available: https://doi.org/10.1121/1.1908963

[5] B. R. Glasberg and B. C. Moore, "A model of loudness applicable to time-varying sounds," *Journal of the Audio Engineering Society*, vol. 50, no. 5, pp. 331–342, 2002.

[6] H. Fletcher, "Loudness, pitch and the timbre of musical tones and their relation to the intensity, the frequency and the overtone structure," *The Journal of the Acoustical Society of America*, vol. 6, no. 2, pp. 59–69, 1934.

[7] D. E. Hall and A. Askenfelt, "Piano string excitation v: Spectra for real hammers and strings," *The Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1627–1638, 1988.

[8] B. Bank, F. Avanzini, G. Borin, G. De Poli, F. Fontana, and D. Rocchesso, "Physically informed signal processing methods for piano sound synthesis: a research overview," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 10, pp. 1–12, 2003.

[9] B. Bank and L. Sujbert, "Generation of longitudinal vibrations in piano strings: From physics to sound synthesis," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2268–2278, 2005.

[10] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, "High-resolution piano transcription with pedals by regressing onset and offset times," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3707–3717, 2021.

[11] D. Jeong, T. Kwon, and J. Nam, "A timbre-based approach to estimate key velocity from polyphonic piano recordings." in *ISMIR*, 2018, pp. 120–127.

[12] 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.

[13] A. Maezawa, K. Yamamoto, and T. Fujishima, "Rendering music performance with interpretation variations using conditional variational RNN." in *ISMIR*, 2019, pp. 855–861.

[14] M. Xu, Z. Wang, and G. G. Xia, "Transferring piano performance control across environments," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 221–225.

[15] B. Kingsbury, "A direct comparison of the loudness of pure tones," *Physical Review*, vol. 29, no. 4, p. 588, 1927.

[16] H. Takeshima, Y. Suzuki, K. Ashihara, and T. Fujimori, "Equal-loudness contours between 1 khz and 12.5 khz for 60 and 80 phons," *Acoustical Science and Technology*, vol. 23, no. 2, pp. 106–109, 2002.

[17] K. Betke, "New measurements of equal-loudness level contours," in *Proc. Inter-noise*, vol. 89, 1989, pp. 793–796.

[18] J. Hall, "Maximum-likelihood sequential procedure for estimation of psychometric functions," *The Journal of the Acoustical Society of America*, vol. 44, no. 1, pp. 370–370, 1968.

[19] H. Møller and J. Andresen, "Loudness of pure tones at low and infrasonic frequencies," *Journal of Low Frequency Noise, Vibration and Active Control*, vol. 3, no. 2, pp. 78–87, 1984.

[20] H. Fasti, A. Jaroszewski, E. Schorer, and E. Zwicker, "Equal loudness contours between 100 and 1000 hz for 30, 50, and 70 phon," *Acta Acustica united with Acustica*, vol. 70, no. 3, pp. 197–201, 1990.

[21] L. L. Beranek, J. Marshall, A. Cudworth, and A. P. G. Peterson, "Calculation and measurement of the loudness of sounds," *The Journal of the Acoustical Society of America*, vol. 23, no. 3, pp. 261–269, 1951.

[22] B. Roß, C. Borgmann, R. Draganova, L. E. Roberts, and C. Pantev, "A high-precision magnetoencephalographic study of human auditory steady-state responses to amplitude-modulated tones," *The Journal of the Acoustical Society of America*, vol. 108, no. 2, pp. 679–691, 2000.

[23] S. Kuwada, R. Batra, and V. L. Maher, "Scalp potentials of normal and hearing-impaired subjects in response to sinusoidally amplitude-modulated tones," *Hearing research*, vol. 21, no. 2, pp. 179–192, 1986.

[24] "Acoustics — Methods for calculating loudness — Part 3: Moore-Glasberg-Schlittenlacher method," International Organization for Standardization, Geneva, CH, Standard, 2017.

[25] B. C. J. Moore and B. R. Glasberg, "Modeling binaural loudness," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. 1604–1612, 2007. [Online]. Available: https://doi.org/10.1121/1.2431331

[26] R. B. Dannenberg, "The interpretation of midi velocity," in *ICMC*, 2006.

[27] A. Adli, Z. Nakao, and Y. Nagata, "Calculating the expected sound intensity level of solo piano sound in midi file," vol. 2006, 2006, pp. 731–736.

[28] M. Keane, "Statistical analysis of classical piano recordings in midi format," in *Proceedings New Zealand Acoustical Society Conference*, 2004.

[29] B. Bank and J. Chabassier, "Model-based digital pianos: from physics to sound synthesis," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 103–114, 2018.

[30] S. Ewert and M. Müller, "Estimating note intensities in music recordings," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 385–388.

[31] O. Ortmann, *The Physical basis of piano touch and tone: An experimental investigation of the effect of the player's touch upon the tone of the piano*. K. Paul, Trench, Trubner & Company Limited, 1925.

[32] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and models*. Springer Science & Business Media, 2013, vol. 22.

[33] B. McFee, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto, D. Ellis, J. Mason, E. Battenberg, S. Seyfarth, R. Yamamoto, viktorandreevichmorozov, K. Choi, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, A. Weiss, D. Hereñú, F.-R. Stöter, P. Friesch, M. Vollrath, T. Kim, and Thassilo, "librosa/librosa: 0.9.1," Feb. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6097378

[34] H. Scheffe, *The analysis of variance*. John Wiley & Sons, 1999, vol. 72.

ON THE IMPACT AND INTERPLAY OF INPUT REPRESENTATIONS AND NETWORK ARCHITECTURES FOR AUTOMATIC MUSIC TAGGING

Maximilian Damböck,[1] **Richard Vogl,**[1] **Peter Knees**[1,2]
[1] Faculty of Informatics, TU Wien, Austria
[2] School of Music, Georgia Institute of Technology, USA
maxi.damboeck@gmail.com, richard.vogl@tuwien.ac.at, peter.knees@tuwien.ac.at

ABSTRACT

Automatic music tagging systems have once more gained relevance over the last years, not least through their use in applications such as music recommender systems. State-of-the-art systems are based on a variant of convolutional neural networks (CNNs) and use some type of time-frequency audio representation as input, in a fitting combination, to predict semantic tags available through expert or crowd-based annotation. In this work we systematically compare five widely used audio input representations (STFT, CQT, Mel spectrograms, MFCCs, and raw audio waveform) using five established convolutional neural network architectures (musicnn, VGG-16, ResNet, a squeeze and excitation network (SeNet), as well as a newly proposed musicnn variant using dilated convolutions) for the task of music tag prediction. Performance of all factor combinations are measured on two distinct tagging datasets, namely MagnaTagATune and MTG Jamendo. A two-way ANOVA shows that both input representation and model architecture significantly impact the classification results. Despite differently sized input representations and practical impact on model training, we find that using STFT as input representations provides the best results overall. Furthermore, the proposed dilated convolutional architecture shows significant performance improvements for all input representations except raw waveform.

1. INTRODUCTION

Music tagging is a multi-label classification task to predict semantic high-level tags like "rock", "piano" or "fast" for music pieces. As for many music information retrieval (MIR) tasks, deep learning has become the state-of-the-art for music tagging [1]. Specifically, CNNs have provided the best results for this task, using either types of audio spectrograms or raw waveform as network input [2, 3].

The selection of audio input representations for neural networks is often decided heuristically or by falling back to standard spectrograms and usually not evaluated [4]. However, the choice of input representation can have a significant impact on performance, as shown in other tasks like natural language processing (NLP), justifying also an evaluation of different representations [5].

While existing work compares different aspects of neural network architectures and input representations (see section 2), an evaluation of their interplay is missing. In this work, as our first contribution, we systematically compare five different architectures and five different input representations in a full-factorial design. The baseline architectures are: *(i)* VGG-16 [6], *(ii)* ResNet [7], *(iii)* SENet [8], *(iv)* musicnn [9], and *(v)* a musicnn architecture with dilated convolutional layers. Five different input representations are tested on each of those architectures: *(i)* raw waveform, *(ii)* short time Fourier transformation (STFT), *(iii)* Mel spectrogram, *(iv)* Mel frequency cepstral coefficients (MFCCs), and *(v)* constant-Q transformation (CQT) [2, 3, 10]. All experiments are repeated on two different datasets and a statistical analysis is used to test if differences of results are significant.

Through this extensive evaluation, we can state that our second contribution is the introduction of dilated convolutional layers in a music tagging network architecture (musicnn). This observed improvement can be related to the fact that the dilated layers allow for a large receptive field when only using a limited amount of layers, cf. [10].

Next, in section 2, we review related work, followed by an outline of the five input representations, five neural network architectures, and datasets used in section 3. Section 4 describes the experimental setup, training, and evaluation strategy and presents the statistical evaluation and significance analysis of the results. Finally, conclusions on the results and findings are drawn in section 5.

2. RELATED WORK

In [2], the authors evaluate state-of-the-art CNN models for music auto-tagging either by using the audio waveform or Mel spectrogram as input representations. MagnaTagATune (MTAT), Million Song dataset (MSD), and MTG-Jamendo (MTGJ) serve as datasets for training and evaluation. Among others, the study considers the following model architectures: A fully convolutional network, *musicnn* (cf. section 3.2.4), a Sample-Level CNN to process raw waveforms using squeeze- and excitation blocks (cf. section 3.2.3), a convolutional recurrent neural network, as well as a self-attention based approach using an adaptation of the transformer architecture. Furthermore, the work fo-

Attribution: M. Damböck, R. Vogl, and P. Knees, "On the Impact and Interplay of Input Representations and Network Architectures for Automatic Music Tagging", in *Proc. of the 23rd Int. Society for Music Information Retrieval Conf.,* Bengaluru, India, 2022.

cuses on effects of data augmentation like pitch shifting or time stretching to improve generalization abilities. Note that while we follow a similar strategy wrt. evaluated architectures in our work, we focus on the impact of different input representations on these different architectures and evaluate their interplay.

More recently, transformer networks have been applied for music tagging [11,12]. While these models improve the state-of-the-art performance, they are significantly more complex than plain convolutional models. Therefore, we consider them not a good match for the evaluation in this work. Other deep learning techniques like transfer-learning, data augmentation, and model aggregation which have been used in music tagging [13], are mainly model agnostic and can be applied on top of any input representation and model combination.

In [14], the authors tested the effect of training data size for music tagging. They trained models with dataset sizes ranging from 100k to 1.2 million training instances using the private 1.2M-Songs dataset. Two CNN architectures are used, one for raw waveforms and a second one for Mel spectrograms inputs. The Mel spectrogram-based model outperformed the raw waveform-based model in both datasets on all tested configurations, which encouraged the conclusion that domain knowledge can be beneficial for designing model architectures.

Using dilated convolutions has shown promising results in image segmentation- and classification over the last few years [15–17]. A similar trend can be observed for the audio domain [18–21]. Dilation can be used to increase the receptive field using only a few layers, which can be beneficial for tasks focusing on global properties of the input.

In [22], the authors use dilated convolutions for environmental sound classification. This task (similarly to music tagging) requires an architecture with a large receptive field. Previously this was achieved by using very deep CNNs. The evaluation shows performance improvements of up to 10% on different datasets compared to existing deep CNNs while reducing model size.

In [4], the authors compare the effects of several audio preprocessing methods for music auto-tagging when using a CNN. For this, the same neural network was trained on the MSD using either STFT or Mel spectrogram as input representations. The evaluation shows no significant differences in performance for the two input representations. However, using log-compressed magnitude spectrograms showed a significant performance increase for all tested configurations.

A more comprehensive comparison of spectrogram representations for audio was conducted in [10] in the context of environmental sound classification. Different configurations for STFT, Mel spectrogram, CQT, continuous wavelet transform, and MFCCs were evaluated in the experiments, using different CNN configurations for training. The evaluation shows that in nearly all configurations, shallower CNN models outperformed deeper ones. The authors suspect that overfitting of the deeper model is the cause of this. The results also show that Mel spectrograms performed consistently well, while STFT and CQT did not, while MFCCs performed worst.

The effects of the reduction of frequency and time resolution of Mel spectrograms for CNN architectures in the context of music auto-tagging were investigated in [23]. Two different state-of-the-art CNN architectures served as a reference for the comparison of Mel spectrograms with different frequency and time resolutions. The results show that while reducing the frequency bands from 128 to 48, the performance loss was negligible.

3. METHOD

This section covers the description of the input representations as well as the neural network architectures used in the evaluation. The experiments are implemented in Python, using librosa [1] for audio processing and TensorFlow as deep learning framework. The source code is available online, refer to the source code for all implementational details [2], there are also additional figures on the accompanying website [3].

3.1 Input Representations

In the following section, the used input representations are discussed and parameter settings are provided. In this work, standard parameter settings established in the literature are used. Figure 1 visualizes the 2D input representations using a 10s audio snippet.

3.1.1 Raw Waveform

In the context of this work, the audio material is resampled to 16kHz mono. This is used consistently also for the source of all spectrograms. As waveform input representation, 16kHz mono 32bit float pulse-code modulation (PCM) data is used.

3.1.2 STFT

A STFT is calculated by repeatedly applying a discrete Fourier transformation (DFT) to small frames of the audio signal. As a baseline spectrogram, in this work, the librosa STFT with default parameters is used: frame size of 2 048 at 16kHz, a hop length of 512, and a Hann window function.

3.1.3 Mel Spectrogram

A usual modification to plain spectrograms is to use psychoacoustically-motivated frequency scales. A common choice is the Mel scale, which has a logarithmic characteristic. In this work, 96-frequency-bin Mel spectrograms are used. A minimum frequency of 32Hz and 12 bins per octave are used. These or similar values are consistently used throughout the literature [23].

3.1.4 MFCC

MFCCs are a compact representation reducing the frequency dimension of the Mel spectrogram using the discrete cosine transform (DCT). MFCCs have been successfully applied in speech recognition and to some extend

[1] https://librosa.org
[2] https://gitlab.com/MaxDamb/dl-autotagging
[3] https://tinyurl.com/ismir22-317