TUNING FREQUENCY DEPENDENCY IN MUSIC CLASSIFICATION

Yi Qin

School of Information Science and Technology
ShanghaiTech University
qinyi1@shanghaitech.edu.cn

Alexander Lerch

Center for Music Technology
Georgia Institute of Technology
alexander.lerch@gatech.edu

ABSTRACT

Deep architectures have become ubiquitous in Music Information Retrieval (MIR) tasks, however, concurrent studies still lack a deep understanding of the input properties being evaluated by the networks. In this study, we show by the example of a Music Genre Classification system the potential dependency on the tuning frequency, an irrelevant and confounding variable. We generate adversarial samples through pitch-shifting the audio data and investigate the classification accuracy of the output depending on the pitch shift. We find the accuracy to be periodic with a period of one semitone, indicating that the system is utilizing tuning information. We show that proper data augmentation including pitch-shifts smaller than one semitone helps minimizing this problem and point out the need for carefully designed augmentation procedures in related MIR tasks.

Index Terms— tuning frequency, music genre classification, model evaluation, convolutional recurrent neural networks

1. INTRODUCTION

Music has become ubiquitous in our daily lives, and with an increasing amount of data online comes an increasing demand for automated analysis and categorization of this data. Music Information Retrieval (MIR) is "a multidisciplinary research endeavor that strives to develop innovative content-based searching schemes, novel interfaces, and evolving networked delivery mechanisms in an effort to make the world's vast store of music accessible to all" [1]. While early MIR systems were often based on expert-designed signal processing systems (e.g., [2]), most concurrent systems are built on data-driven machine learning approaches [3]. More specifically, various forms of deep learning approaches are nowadays considered state-of-the-art for the majority if not all of MIR tasks.

Deep Neural Networks (DNNs), however, often cannot be interpreted easily and lack intuitive approaches to understanding *why* the network makes specific decisions. This opacity has triggered recent work on reverse engineering and visualizing states of the network to allow for insights into the learned model [4, 5].

As an insightful analysis of internal and intermediate results remains hard despite these efforts, we analyze what a DNN has learned by investigating the output of the network in this study. The system is therefore treated as a black box instead of a program that can be modified to gain access to intermediate or internal results. More specifically, we test a classification system with input files that contain perceptually identical content, i.e., humans would categorize them in the same class and to evaluate how the system reacts to these imperceptibly modified inputs. For this purpose, we choose Music Genre Classification (MGC) as an example task. It is a well-researched topic that has been one of the early and probably most popular MIR tasks with a large number of relevant publications spanning the past two decades. Genre classification is also a fitting prototype classification task because of its high similarity to other MIR tasks such as Mood Recognition, Music Tagging, and Artist Identification, as well as its relation to tasks such as Vocal Activity Detection and Instrument Identification. We modify the pitch of the test signals of a MGC system under the assumption that a small pitch variation is irrelevant and will not impact the genre of a piece of music. That means that we assume that, for example, that a piece played in the musical key of D-Major can also be played in D#-Major (or starting from a root note tuned between D and D#) and still be considered the same musical genre.

The remainder of this paper is structured as follows: Sect. 2 introduces related work on genre classification, model evaluation, and the role of tuning frequency in MIR. Section 3 introduces the experimental setup and the systems used, and Sect. 4 presents and discusses the results. Finally, we draw our conclusions in Sect. 5.

2. RELATED WORK

2.1. Genre Classification

The definition of MGC poses some fundamental problems (e.g., inconsistency and non-orthogonality of genre labels, compare [6]) that even lead researchers to suggest abandoning all attempts and focusing on other tasks [7]. Nevertheless, it

remains a popular research topic in the field of MIR, initially set in motion by Tzanetakis and Cook's seminal publication in 2002 [8], proposing a standard machine learning pipeline with feature extraction and aggregation followed by a classifier. Their system triggered research which usually focused either on improving the dominant low-level features describing timbral content, such as Mel-Frequency Cepstral Coefficients (e.g., [9, 10]), or on using more sophisticated classifier models [11]. An overview of feature-driven approaches to MGC can be found in [12].

These traditional machine learning methods were quickly replaced by the often more successful deep learning methods [13, 14]. In MGC specifically, a common deep learning architecture is a Convolutional Neural Network (CNN) with a mel-spectrogram input [15]. Recurrent Neural Networks (RNN) are, in contrast to CNNs which look at one snippet of data at a time, able to learn patterns in a sequence [16]. Convolutional Recurrent Neural Networks (CRNNs) combine these two approaches in a hybrid model by replacing the last convolutional layers with a RNN [17].

2.2. Tuning Frequency in MIR Systems

The concert pitch A4 is the pitch commonly used for tuning one or more musical instruments. Its frequency, the tuning frequency, is standardized internationally to 440 Hz [18], but the exact frequency used by musicians can vary due to a variety of reasons. These reasons can include, for example, the usage of historic instruments, timbre preferences, and even low or high room temperatures while recording. These deviations can easily span a quarter-tone or more [19]. Therefore, the estimation of tuning frequency [20, 21] is considered essential in pitch-focused MIR systems (key and chord detection, pitch transcription). For other MIR systems, such as MGC, tuning frequency is ignored as it is considered an irrelevant property for the task. For example, listeners will categorize genre identically if the same song is played at a slightly higher or lower pitch.

2.3. Model Evaluation

In image classification, researchers have worked on understanding network behavior by generating adversarial test inputs resulting in misclassification even though the input varies only imperceptibly from a correctly classified input [22, 23]. The results imply that it might be the linear behavior in high-dimensional spaces that causes such behavior.

Other ways of evaluating the functionality of neural networks investigate the inner state of the network by, e.g., visualizing (projections of) network layer activations [22, 4, 5].

In the context of MGC, Sturm addresses adversarial examples by pointing out that not only the classification accuracy on a specific dataset should matter in the evaluation of an MGC system, but also — among others — its robustness, meaning that the system should be invariant to aspects inconsequential for the human identification of genre [24]. In a different study, the same author shows that systems which appear capable of the task of MGC might actually be utilizing irrelevant characteristics or confounds [25].

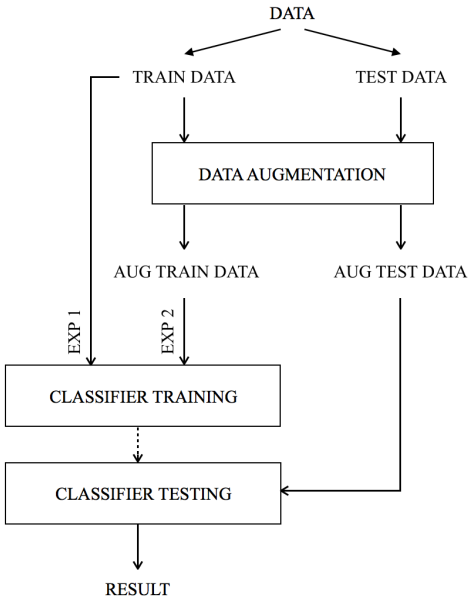

Fig. 1. Experimental flowchart.

3. EXPERIMENTAL SETUP

Building on Sturm's findings [25], our experimental setup is based on the hypothesis that state-of-the-art DNNs for MGC are, while quite capable of achieving high accuracy on common datasets, impacted by irrelevant characteristics. In our study, we are investigating the impact of the tuning frequency on the classification accuracy of a Music Genre Classifier. As humans will not classify a piece of music differently just because it is slightly shifted in pitch, pitch shifting by small amounts should not impact classification accuracy.

Our general methodology is shown in Fig. 1. We use a state-of-the-art MGC architecture and two established datasets. The test data is augmented by pitch shifting up and down in small increments. We then measure the accuracy depending on the pitch shift factor. The details are outlined in the subsections below.

3.1. Classification Model

The CRNN model presented by Choi [17] uses a 2-layer RNN to summarize temporal patterns on top of the 2-dimensional 4-layer CNNs. Based on Choi's model which is available online (trained for Auto-Tagging with the Million Song Dataset), we use transfer learning [26, 27] to adapt the model to MGC with

our data (see Sect. 3.2). The input is the log-amplitude mel-spectrogram of the down-sampled (12 kHz) audio as generated by the library librosa [1] (96 mel-bins, hop-size 256 samples). The resulting input dimensions are 96×1366 (mel-frequency bands $\times$ time frames). The output layer is replaced with a new layer with sigmoid activation functions and with as many output nodes as the number of classes. We use ADAM as optimizer [28] and categorical cross-entropy as a loss function.

3.2. Data

The experiment is based on two datasets, the GTZAN dataset [8] with 1000 audio clips covering 10 genres and the FMA-small dataset [29] with 8000 audio clips and 8 genres. Both datasets are balanced and all audio clips have a length of 30 s.

We perform 5-fold cross-validation to examine the performance of each model. We randomly split the dataset into five equal-sized subsets, resulting in a training/testing ratio of $4:1$. We then treat the average accuracy as the test accuracy of each of the experimental models.

3.3. Data Augmentation

That data is augmented by changing the pitch using a state-of-the-art commercial pitch-shifting engine which is industry standard.[2] This pitch shifting engine is integrated in the majority of professional music production software and should provide high quality pitch shifting with a minimum of artifacts. A code review and discussions with the developers ensured that the pitch shifting engine is not optimized for specific pitch factors. The tempo of the songs keeps unchanged as tempo has been shown to be a feature relevant for MGC [30]. Corresponding to a pitch-shifting factor range of 0.75–1.34 with 0.01 per step, the data is pitch-shifted in the range of $\pm 5\,\mathrm{semitones}$ with 4–6 steps per semitone.

3.4. Experiments

The following experiments are being presented:

- **Exp. 1**: Train the network (initialized with the original tagging-weights) with both training sets and report the results on both augmented test sets.
- **Exp. 2**: Train the network with both augmented training sets and report the results on both augmented test sets.

In both scenarios, the classification accuracy is reported.

4. RESULTS

4.1. Experiment 1

The results for Exp. 1 are shown in Fig. 2. As expected, the test accuracy of the unpitched songs (pitch shift = 0) is the absolute maximum for both datasets. The maximum accuracy for GTZAN and FMA is about 88.0% and 61.9%, respectively. This is roughly in line with our expectations based on previous results [12], but note that we concern ourselves not too much about the absolute values of our results as our experiments are focused on evaluating the relative changes. It can be observed that for both datasets the accuracy drops with increasing pitch shift. There is a variability of this decrease of up to approx. 10%.

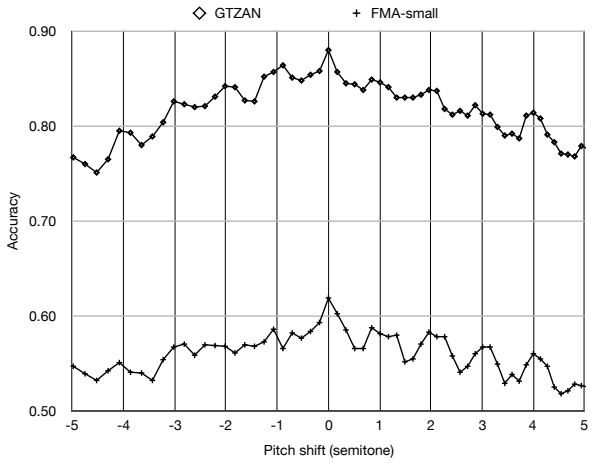

Fig. 2. Classification accuracy on both datasets in dependence of the pitch-shift

The general decrease in accuracy with increasing pitch shift is expected. On the one hand, instrument and general timbre characteristics of the audio might be shifted out of the generally expected range and thus impact classification accuracy; on the other hand, the pitch-shifting is expected to start producing noticeable artifacts for extreme pitch shift factors which could in turn impact the classification accuracy.

The surprising and noteworthy result, however, is the periodicity of the result. We observe local maxima at (or close to) pitch shift factors at integer multiples of one semitone and local minima in-between. This means that the classification result of one song and the same song pitch-shifted by half a semitone might be different. In other words, the classifier is not invariant to irrelevant pitch changes within a semitone range. This indicates that the network actually takes into account the tuning frequency of a song, an irrelevant variable. Obviously, this is an undesired effect with implications on the performance of a multitude of music analysis and synthesis systems.

To further investigate this phenomenon, we track the accuracy of individual genres in the GTZAN dataset. Figure 3 shows four prototypical genres: strong periodicity plus general decrease with increasing pitch shift (Pop), slight periodicity and general decrease (Rock), fluctuation with general decrease (Blues), and general robustness to pitch shifting (Jazz). From these results, we see two trends: first, the more acoustic in-

[1] https://doi.org/10.5281/zenodo.1342708

[2] zplane ELASTIQUE: www.time-stretching.com, last access: Oct 19, 2018

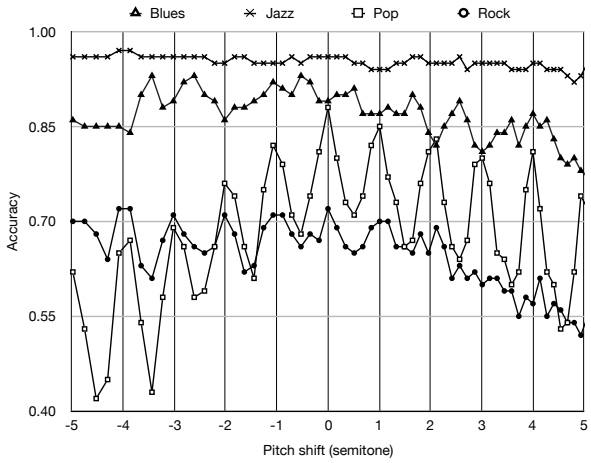

Fig. 3. Classification accuracy on selected GTZAN genres

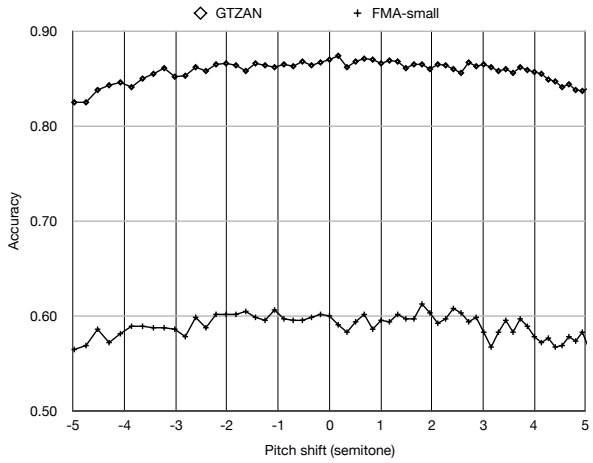

Fig. 4. Classification accuracy in dependence of the pitch-shift when trained with augmented data

struments are used, the less pronounced the effect appears as the likelihood of a diverse training set increases (synthesizers and electronic music equipment usually defaults to 440 Hz while the tuning of acoustic instruments is in the hands of the musicians). Second, the impact on genres with very high classification accuracy is minimal, indicating that the classification is so robust that a pitch shift cannot influence the result.

4.2. Experiment 2

When the model is trained with the augmented data, however, the periodicity of the results disappears as shown in Fig. 4. As expected, we can still identify a tendency of generally declining accuracy at large pitch-shifts, but otherwise we receive a roughly constant pitch-shift-independent classification accuracy, albeit with noticeable variability. The absolute maximum is not found anymore at pitch shift 0, which speaks towards the general noisiness of the results.

As expected, the semitone periodicity of the result can be removed by training with augmented data. The tuning frequency dependency without data augmentation, however, is confounding and should be generally tested for in neural audio analysis and synthesis systems. Although data augmentation is increasingly used in the training of neural networks with audio, the usual augmentation approaches usually pitch-shift by integer multiples of semitones [31, 32]. This could lead to the same tuning frequency dependency that we observe here, and is clearly not desirable for the majority of classification tasks. This problem applies specifically to neural networks; traditional timbre-feature-based classification systems use features (e.g., Mel Frequency Cepstral Coefficients) that are designed to be pitch independent.

A side effect of augmenting the training data is that the general decrease at large pitch shifts disappears or at least is not strong. It indicates that the system is learning to deal with larger pitch shifts, but whether this is desired or not is a different discussion.

5. CONCLUSION

We presented an experiment examining the robustness of a neural network against irrelevant tuning frequency changes. The results show that a state-of-the-art network, trained for the task of genre classification, is confounded by modifications of the tuning frequency other than shifts by semitones, as the classification accuracy drops for pitch shifts of 0.5, 1.5, ... semitones. This is a noteworthy result as it might indicate the vulnerability of general music classification systems even if they are trained with data augmented by pitch-shifting. We call for a careful reevaluation of training and data augmentation practices in MIR systems to ensure that tuning frequency may not become a generally confounding variable for these systems.

Future work will aim at evaluating this dependency for other tasks such as mood classification or instrument detection as well as conducting experiments with other architectures. We also plan to look into other transformations that are irrelevant for specific tasks; these might include time-stretching, amplification, filtering, etc.

6. REFERENCES

[1] JS Downie, "The scientific evaluation of music information retrieval systems: Foundations and future," *CMJ*, vol. 28, no. 2, pp. 12 – 23, 2004.

[2] M Goto and Y Muraoka, "Music Understanding At The Beat Level – Real-time Beat Tracking For Audio Signals," in *IJCAI*, 1995.

[3] EJ Humphrey, JP Bello, and Y LeCun, "Feature learning and deep architectures: new directions for music informatics," *JIIS*, vol. 41, no. 3, pp. 461–481, 2013.

[4] MD Zeiler and R Fergus, "Visualizing and Understanding Convolutional Networks," in *ECCV*, D Fleet, T Pajdla, B Schiele, and T Tuytelaars, Eds. 2014, Lecture Notes in Computer Science, pp. 818–833, Springer.

[5] S Mishra, BL Sturm, and S Dixon, "Understanding a Deep Machine Listening Model through Feature Inversion," in *ISMIR*, Paris, 2018.

[6] A Lerch, *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*, Wiley-IEEE Press, Hoboken, 2012.

[7] C McKay and I Fujinaga, "Musical genre classification: Is it worth pursuing and how can it be improved?," in *ISMIR*, Victoria, 2006.

[8] G Tzanetakis and P Cook, "Musical genre classification of audio signals," *Trans. on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[9] JJ Burred and A Lerch, "Hierarchical Automatic Audio Signal Classification," *JAES*, vol. 52, no. 7/8, pp. 724–739, 2004.

[10] G Tzanetakis, LG Martins, K McNally, and R Jones, "Stereo Panning Information for Music Information Retrieval Tasks," *JAES*, vol. 58, no. 5, pp. 409–417, 2010.

[11] S Sharma, P Fulzele, and I Sreedevi, "Novel hybrid model for music genre classification based on support vector machine," in *ISCAIE*, 2018.

[12] Z Fu, G Lu, KM Ting, and D Zhang, "A Survey of Audio-Based Music Classification and Annotation," *Trans. on Multimedia*, vol. 13, no. 2, pp. 303–319, 2011.

[13] K Choi, "Automatic Tagging Using Deep Convolutional Neural Networks," in *ISMIR*, New York, 2016.

[14] S Gururani, C Summers, and A Lerch, "Instrument Activity Detection in Polyphonic Music using Deep Neural Networks," in *ISMIR*, Paris, 2018.

[15] S Sigtia and S Dixon, "Improved music feature learning with deep neural networks," in *ICASSP*, 2014.

[16] J Dai, S Liang, W Xue, C Ni, and W Liu, "Long short-term memory recurrent neural network based segment features for music genre classification," in *ISCSLP*, 2016.

[17] K Choi, G Fazekas, M Sandler, and K Cho, "Convolutional recurrent neural networks for music classification," in *ICASSP*, 2017.

[18] ISO 16:1975, "Acoustics – Standard tuning frequency (Standard musical pitch)," Standard, ISO, 1975.

[19] A Lerch, "On the Requirement of Automatic Tuning Frequency Estimation," in *ISMIR*, Victoria, 2006.

[20] K Dressler and S Streich, "Tuning Frequency Estimation using Circular Statistics," in *ISMIR*, Wien, 2007.

[21] A Degani, M Dalai, R Leonardi, and P Migliorati, "Comparison of tuning frequency estimation methods," *Multimedia Tools and Applications*, 2014.

[22] C Szegedy, W Zaremba, I Sutskever, J Bruna, D Erhan, IJ Goodfellow, and R Fergus, "Intriguing Properties of Neural Networks," in *ICLR*, Banff, 2014.

[23] IJ Goodfellow, J Shlens, and C Szegedy, "Explaining and Harnessing Adversarial Examples," in *ICLR*, San Diego, 2015.

[24] BL Sturm, "Classification accuracy is not enough: On the evaluation of music genre recognition systems," *JIIS*, vol. 41, no. 3, pp. 371–406, 2013.

[25] BL Sturm, "A simple method to determine if a music information retrieval system is a "horse"," *Transòn Multimedia*, vol. 16, no. 6, pp. 1636–1644, 2014.

[26] Y Bengio, "Deep learning of representations for unsupervised and transfer learning," in *JMLR W&CP*, 2012, vol. 27, pp. 17–36.

[27] J Yosinski, J Clune, Y Bengio, and H Lipson, "How transferable are features in deep neural networks?," in *NIPS*, Montreal, 2014.

[28] DP Kingma and J Ba, "Adam: A method for stochastic optimization," in *ICLR*, San Diego, 2015.

[29] K Benzi, M Defferrard, P Vandergheynst, and X Bresson, "FMA: A dataset for music analysis," in *ISMIR*, Suzhou, 2017.

[30] F Gouyon, S Dixon, E Pampalk, and G Widmer, "Evaluating Rhythmic descriptors for Musical Genre Classification," in *Int. Conf. on Metadata for Audio*. 2004, AES.

[31] B McFee, EJ Humphrey, and JP Bello, "A Software Framework for Musical Data Augmentation," in *ISMIR*, Malaga, 2015.

[32] S Kum, C Oh, and J Nam, "Melody Extraction on Vocal Segments Using Multi-Column Deep Neural Networks," in *ISMIR*, New York, 2016.