Instrument Role Classification: Auto-tagging for Loop Based Music

Joann Ching[1], António Ramires[2], and Yi-Hsuan Yang[1,3]

[1] Research Center for IT Innovation, Academia Sinica, Taipei, Taiwan
[2] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[3] Taiwan AI Labs, Taipei, Taiwan
joann8512@citi.sinica.edu.tw, antonio.ramires@upf.edu, yhyang@ailabs.tw

Abstract. The proposed work introduces a new type of auto-tagging task, called "instrument role classification." We discuss why the task is necessary, and further introduce a definition regarding loop based music. We introduce a new dataset for this task, the Freesound Loop Dataset, and benchmark the performance of both neural network and non-neural network based multi-label classification models for six instrument roles.

Keywords: Music auto-tagging, instrument role classification, loop.

1 Introduction

Methods for assisting electronic music production have been emerging rapidly, and various creation interfaces, such as LogicPro, Ableton Live, Avid Pro Tools, and Bandlab, have appeared. One easy and engaging creation style is to work with "loops," which are audio excerpts, usually of short duration, that can be played repeatedly in a seamless manner (Stillar, 2005). In other words, the motive of a clip is simple and straightforward, compared to audio recordings with longer phrases, and can be summarized within seconds.

In this paper, we present a study that applies auto-tagging in a way that has rarely been done before in literature, to the best of our knowledge—to tag an audio clip by its "role." In the style of loop-based music where multiple loops are stacked together, each loop plays its "role" in the music. Specifically, we consider the following six possible roles—`Percussion`, `Bass`, `Chord`, `Melody`, `FX`, and `Voice`. See Table 1 for examples. Classifying such role provides important information for users as it eases the process of finding compatible loops that fit their needs, thereby contributing to assisting electronic music production.

We note that this task is different from general instrument prediction tasks, as different instruments can play the same role, and the same instrument can play different roles. Similar to genre or emotion labels, our labels might be susceptible to the subjective consideration of the annotators. And, our labels are non-exclusive, as a loop can fill two roles at once.

We create a new dataset for this task and prototype such a model based on a convolutional neural network that uses a data-driven harmonic filter-based

Table 1: The instrument roles and the number of associated loops in our dataset

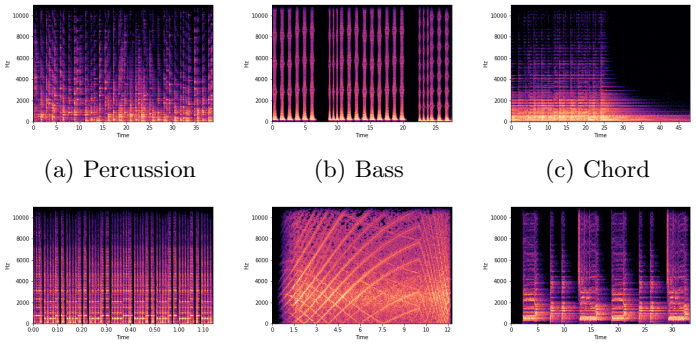
Roles	**Example instruments**	**Count**
`Percussion`	Drums, glitches, tuned percussion	1,626
`FX`	Risers, cinematic sounds, foley, scratching	845
`Melody`	Instrument playing a melody, arpeggiator	603
`Bass`	Synth bass, fingered bass	493
`Chord`	Piano chords, guitar chords, synth pads	350
`Voice`	Singing voice, spoken word, vocoder	69

front-end (HCNN) proposed lately (Won, Chun, Nieto, & Serra, 2020). With the behavior of the model capturing harmonic relations while preserving spectro-temporal locality, we show that the model can learn to distinguish the instrument roles efficiently under limited data. Both the data and code for implementing our work can be found at `https://github.com/joann8512/Loop-Classifier`.

2 Related Works

As technology evolves, music has become easily accessible by people, requiring effective music searching. As such, automatic tag prediction has been a popular task (Kim, Lee, & Nam, 2017; Chou, Jang, & Yang, 2018). Choi et al. (Choi, Fazekas, & Sandler, 2016) introduced a deep fully convolutional neural network (FCN) and showed that deep FCN with 2D convolutions can be effectively used for automatic music tagging and classification tasks. After knowing that it is possible to design efficient CNNs for modeling temporal features such as tempo and rhythm, Pons et al. (Pons & Serra, 2017; Pons, Slizovskaia, Gong, Gómez, & Serra, 2017) developed a structure that uses different filter shapes that are motivated by domain knowledge in the first layer to efficiently learn timbre representations. Among all the experiments done by Pons et al., the structure is proven to work more efficiently with twice fewer number of parameters for singing voice phoneme classification and instrument recognition.

As our task falls within the realm of loops, several new innovative works have been proposed for dealing with such type of audio material, and our task can easily be linked to these. Smith et al. (Smith, Kawasaki, & Goto, 2019) introduced an interface for extracting and remixing loops, where users are allowed to upload music, extract, remix, and mash-up loops immediately. In their user case study, expert and novice users found it easy to use. A nonnegative Tucker decomposition-based source separation model was used, and an extra factorization step with sparseness constraint improves the separation result. Along with the work of Smith et al., Chen et al. (Chen, Smith, & Yang, 2020) proposed an automated method of finding compatible loops, which presents a data generation pipeline and several negative sampling strategies for ground-truth labeling for training a machine learning model. In this work, Convolutional neural networks are shown to perform well in distinguishing between compatible loops and non-compatible loops. Fairly recently, Ramires et al. built and released the Freesound



(a) Percussion (b) Bass (c) Chord

(d) Melody (e) FX (f) Vocal

Fig. 1: Log scaled mel-spectrograms of sample loops of various instrument roles.

Loop Dataset (FSLD) (Ramires et al., 2020), a new large-scale dataset of music loops annotated by experts, in which instrument role, tempo, meter, key, and genre tags are annotated. This work is possible because of this new dataset.

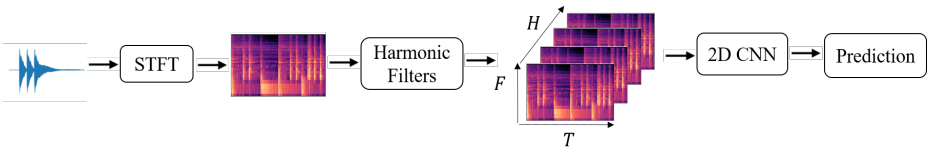
3 Dataset

We employ FSLD (Ramires et al., 2020) to train and evaluate our instrument role classification models. Differentiating from other commercial and community databases of pre-recorded loops, the source of the loops of FSLD is Freesound (Font, Roma, & Serra, 2013), a community database of audio recordings released under Creative Commons licenses, making the audios in the dataset distributable. Within all the annotations from the dataset, we take only the instrument role annotations for our task, making it 6 tags for each loop — `Percussion`, `Bass`, `Chord`, `Melody`, `FX`, and `Voice`. As a loop can be labeled with multiple tags, this makes the task a multi-label classification task.

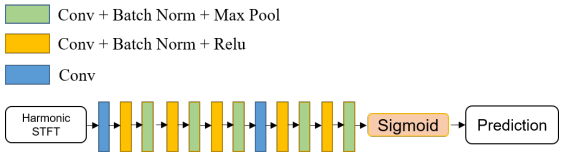
The mel-spectrograms of some example loops are shown in Figure 1. We note that `Melody` and `Bass` loops are monophonic; `Chord` loops are polyphonic; `Voice` loops can be monophonic, polyphonic, or even percussive (e.g., beatboxing), and it is the only instrument role among the six that is associated with only a certain instrument (timbre). Table 1 lists the number of loops associated with each instrument role. We see that `Percussion` is the most popular one, with 1,626 examples, whereas `Voice` is the least popular, with only 69 examples.

Although not extracted as one of the label features, the loop's styles are of multiple genres, including Bass Music, Live Sounds, Cinematic, Global, Hip Hop, Electronic, etc, which helps familiarize the model with several genres of loops.

We randomly extract one 3-second chunk from a total of 2,936 loops and split the data into training, testing, and validation sets by the ratio of 90:5:5.



(a) Overall structure



(b) Details of the '2D CNN' part in subfigure (a)

Fig. 2: Schematic plot of the HCNN model proposed in (Won et al., 2020)

4 Methodology and Results

We benchmark two neural network based model for this task. The first model, the *HCNN* (Won et al., 2020), is constructed with the basic block shown in Figure 2a. Waveforms transformed with STFT are passed through the harmonic filters to obtain a harmonic tensor representing it in six harmonics (Bittner, Mcfee, Salamon, Li, & Bello, 2017). To encourage the convolutional filters to embed harmonic information along with time and frequency, the harmonics are treated as channels to be fed. The model uses seven convolution layers and a fully connected layer. Each layer, except the last, which uses sigmoid instead, is batch normalized and ReLU-activated. The model was trained with 200 epochs using scheduled ADAM with learning rate 1e−4. We use the epoch that achieves the best result on the validation set. The second model, dubbed *non-harmonic CNN*, is an ablated version of HCNN that uses the same network structure as the HCNN but not those harmonic filters at the front-end.

We also benchmark the following non-neural network based approaches. For feature extraction, we use Essentia (Bogdanov et al., 2013) gather features from the low-level (MFCC, and pitch) and tonal (chromagram) representations. For temporal aggregation, we take the mean, standard deviation, derivative of mean, and derivative of standard deviation by frame across time, leading to a 104-dimensional feature vector for each 3-second chunk of loops. Then, we experiment with the classical methods, binary relevance (BR), label powerset (LP), and distinct Random k-Labelsets (RAkEL) (Trohidis, Tsoumakas, Kalliris, & Vlahavas, 2008), all available in the scikit-multilearn package, to convert the intended multi-label classification problem to a single-label classification problem. We then experiment with using either random forest (RF) or support vector machine (SVM) to build the classifier. For RF, we use 1,000 estimators; for SVM, we use the linear kernel and set the cost parameter C to 10.

The evaluation results are shown in Table 2. Following (Pons et al., 2018; Kim, Lee, & Nam, 2019; Won, Chun, & Serra, 2019), we present the results in

Table 2: Results of all methods. The scores are calculated respectively using roc_auc_score, average_precision_score, and f1_score from scikit learn metrics.

Methods	ROC_AUC	PR_AUC	F1
BR-RF	0.6201	0.3759	0.5538
BR-SVM	0.6248	0.3561	0.5836
LP-RF	0.6227	0.3714	0.6027
LP-SVM	0.6302	0.3405	0.5930
RAkEL-RF	0.6283	0.3787	0.5880
RAkEL-SVM	0.6325	0.3573	0.6026
non-harmonic CNN	0.8324	0.5772	0.7275
HCNN (Won et al., 2020)	0.8606	0.5860	0.7253

terms of the Area Under Receiver Operating Characteristic Curve (ROC_AUC), Area Under Precision-Recall Curve (PR_AUC), and the F1 score. We see that the neural network based models outperform the non-neural network models by a great margin. The best result is achieved by HCNN, which obtains 0.8606 ROC_AUC. The comparison between non-harmonic CNN and HCNN shows that the efficacy of using the harmonic filters.

Detailed analysis of the prediction results, the answers appear to be trending towards giving only one label, caused by the imbalance of multi-label and single-label annotations. Prediction results reflected the amount of each label count (Table 1). For the three most common tags — `Percussion`, `Melody`, and `FX`, where loops with such tags usually have high accuracy score. In contrast, as `Vocal` is least labeled, such loops are usually recognized as either `Melody` (Singing) or `Percussion` (Beatboxing). Please see the appendix for example prediction result of the HCNN model, and its confusion table.

5 Conclusion

In this paper, we have introduced a new music auto-tagging task that aims to tag each loop by its role. We have also introduced a new dataset for this task, and benchmarked a few models using the dataset. Our evaluation shows that HCNN, a neural network based model, is effective in learning useful features under limited data. For future work, we are interested in using the loop role classifiers as a building block for automatic mashup or loop-based music creation.

References

Bittner, R. M., Mcfee, B., Salamon, J., Li, P., & Bello, J. P. (2017). Deep salience representation for f0 estimation in polyphonic music. In *International society for music information retrieval (ISMIR)*.

Bogdanov, D., Wack, N., Gómez, E., Gulati1, S., Herrera1, P., Mayor, O., ... Serra, X. (2013). ESSENTIA: An audio analysis library for music infor-

mation retrieval. In *International society for music information retrieval (ISMIR).*

Chen, B.-Y., Smith, J. B. L., & Yang, Y.-H. (2020). Neural loop combiner: Neural network models for assessing the compatibility of loops. In *International society for music information retrieval (ISMIR).*

Choi, K., Fazekas, G., & Sandler, M. (2016). Automatic tagging using deep convolutional neural networks. In *International society for music information retrieval (ISMIR).*

Chou, S.-Y., Jang, J.-S. R., & Yang, Y.-H. (2018). Learning to recognize transient sound events using attentional supervision. In *IEEE international joint conference on artificial intelligence (IJCAI).*

Font, F., Roma, G., & Serra, X. (2013). Freesound technical demo. In *ACM international conference on multimedia.*

Kim, T., Lee, J., & Nam, J. (2017). Sample-level CNN architectures for music auto-tagging using raw waveforms. In *IEEE international conference on acoustics, speech and signal processing (ICASSP).*

Kim, T., Lee, J., & Nam, J. (2019). Comparison and analysis of samplecnn architectures for audio classification. In *IEEE journal of selected topics in signal processing, col. 13, no. 2.*

Pons, J., Nieto, O., Prockup, M., Schmidt, E., Ehmann, A., & Serra, X. (2018). End-to-end learning for music audio tagging at scale. In *International society for music information retrieval conference (ISMIR).*

Pons, J., & Serra, X. (2017). Designing efficient architectures for modeling temporal features with convolutional neural networks. In *IEEE international conference on acoustics, speech and signal processing (ICASSP).*

Pons, J., Slizovskaia, O., Gong, R., Gómez, E., & Serra, X. (2017). Timbre analysis of music audio signals with convolutional neural networks. In *IEEE European signal processing conference (EUSIPCO).*

Ramires, A., Font, F., Bogdanov, D., Smith, J. B. L., Yang, Y.-H., Ching, J., ... Serra, X. (2020). The freesound loop dataset and annotation tool. In *Proc. of the 21st international society for music information retrieval (ISMIR).*

Smith, J. B. L., Kawasaki, Y., & Goto, M. (2019). UNMIXER: An interface for extracting and remixing loops. In *International society for music information retrieval (ISMIR).*

Stillar, G. (2005). Loops as genre resources. *Folia Linguistica*, *39*(1-2), 197 - 212.

Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. (2008). Multi-label classification of music into emotions. In *International society for music information retrieval (ISMIR).*

Won, M., Chun, S., Nieto, O., & Serra, X. (2020). Data-driven harmonic filters for audio representation learning. In *IEEE international conference on acoustics, speech and signal processing (ICASSP).*

Won, M., Chun, S., & Serra, X. (2019, June). Toward interpretable music tagging with self-attention. *arXiv e-prints*, arXiv:1906.04972.

6 Appendix

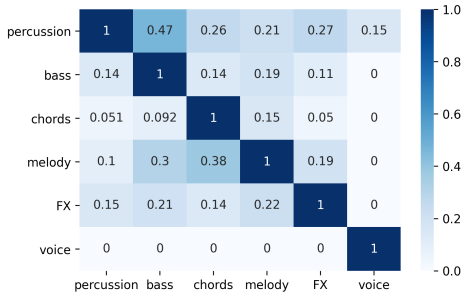
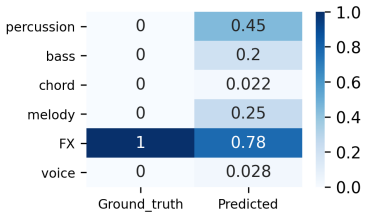
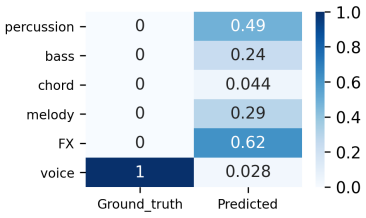


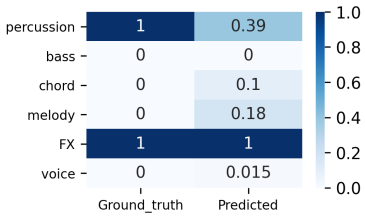
Fig. 3: Co-occurrence matrix of the six instrument role tags. The values represents normalized confidence levels of each prediction.



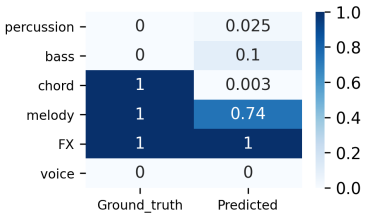
(a) FX



(b) Voice



(c) Percussion and FX



(d) Chord and melody and FX

Fig. 4: Examples of the predictions of the HCNN model we implemented. In each plot, prediction confidence levels are shown (right) next to their ground truth (left). Each "Predicted" column is the normalized average of prediction confidence of the repetitions.