Identification of Upscaled Lossless Audio based on Spectrogram Image Features

Yufei Du [1] **Feng Xu** [1]

Abstract

Music listening brings joy and reduces mental stress. In this project, we investigated a novel problem of classifying whether a music file is genuinely hi-resolution. Briefly, lossless music files were first converted into 30-second FLAC or MP3 audio snippets. Then, we extracted spectrogram features from each of the audio snippets. A decision tree classifier was trained on these features to perform the binary classification of whether a audio snippet is loseless or lossy. We compared our approach with the openSMILE audio feature extractor. The results showed that our approach achieves higher in F-1 score (95.8%) than the openSMILE approach (49.5%), indicating that our approach is a more accurate and reliable way to identify upscaled lossless audio.

1. Introduction

The way we experience music has been substantially changed thanks to the development of audio technology. Arising from 30 years ago, hi-resolution (hi-res) audio was introduced to improve the listening experience (Melchior, 2019). Hi-res audio may be defined as high sampling frequency (more than 44.1 kHz) and large number of bits per sample (more than 16 bits) (Reiss, 2016). It requires to be stored in a lossless compression format so that it could provide the full range of audio from the original recordings. However, it requires larger storage space in the meantime. Nevertheless, the hi-res audio is more accessible and is gaining popularity as the storage becomes cheaper in recent days.

Compared with high-cut audio (below 20 kHz), studies have showed that the brain and nervous system are activated more by high-resolution audio (Kuribayashi et al., 2014; Ito et al., 2016a), potentially leading to a more enjoyable listening experience. Further studies showed that the hi-res audio increases α wave on EEG and reduces sweat when participants suffering from mental stress (Harada et al., 2010; Ito et al., 2016b; Harada et al., 2017). Moreover, the digitalized hi-res audio also reshaped how the professional seek and explore the music (Im & Kim, 2016).

However, artificially upsampled hi-res audio may negatively impact those benefits of hi-res audio. Due to the lack of available hi-res audio in the market, one can obtain artificial hi-res audio via upsampling high-cut audio. Although the upsampled audio matches the sampling frequency of the hi-res audio, the high-frequency components in original sounds cannot be restored.

To this end, we proposed an efficient framework to classify whether an audio is genuinely hi-res or was upscaled from high-cut audio file.

2. Related Works

Music information retrieval (MIR) is a well studied task within the past 20 years (Fu et al., 2010). For example, music classification has been performed on genre classification (Tzanetakis & Cook, 2002; Li et al., 2003; Song & Zhang, 2007), artist classification (Kim & Whitman, 2002), and instrument recognition (Marques & Moreno, 1999; Chakraborty & Parekh, 2018). Audio features, such as zero crossing rate and spectral centroid have been proposed in these music classification tasks. Additionally, classifiers such as k-nearest neighbors (Fix & Hodges, 1989), Gaussian mixture models (Duda et al., 1973), and support vector machine (Boser et al., 1992) have been implemented for music classification tasks (Pampalk et al., 2003), (Tzanetakis & Cook, 2002), and (Pampalk et al., 2002), respectively. However, the classification of artificial hi-res audio remains blank.

In recent years, deep learning has gathered considerable interest in many fields such as computer vision (Krizhevsky et al., 2012; He et al., 2016; Liu et al., 2021) and natural language processing (Xu et al., 2015; Devlin et al., 2018). Not only because of its outstanding performance but also the practical property of learning feature representations. At first glance, recurrent neural networks (RNNs) (Rumelhart et al., 1986) are promising on this task because their lateral propagation structure, which allows them to exhibit temporal information. However, Schluter et al. show that (Schlüter

[1]University of North Carolina at Chapel Hill. Correspondence to: Yufei Du <yufeidu@cs.unc.edu>, Feng Xu <fengxu@unc.edu>.

COMP 562 Course Project

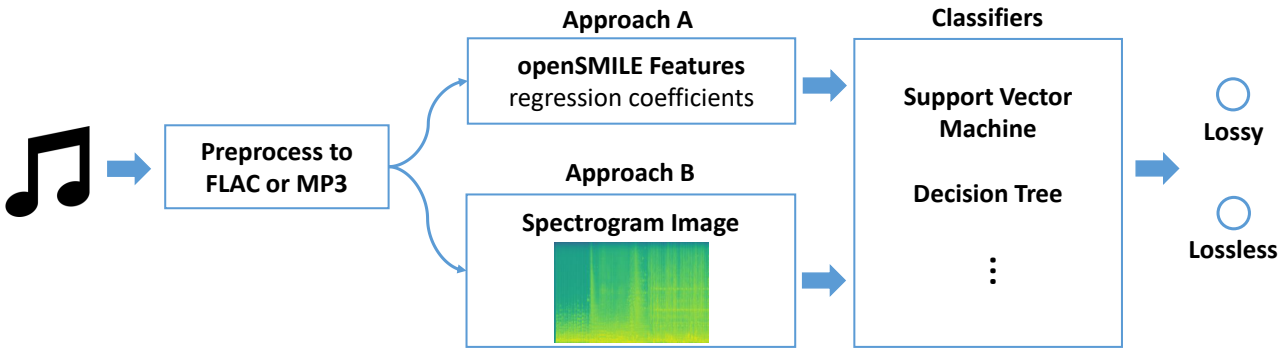

Figure 1. An overview of the two approaches.

& Böck, 2014) convolutional neural networks (CNNs) could achieve better results than RNNs in a musical onset detection task, potentially due to CNNs' ability to capture the temporal-frequency relationship in the audio features from draft. The features are in the form of 2D feature images and extracted from the spectrogram of the audio signals. Moreover, compared with RNNs' lateral structure, convolution operation can be fully parallelized and leverage the GPU hardware (Gehring et al., 2017). The capability of CNNs in music classification was also demonstrated via transfer learning on multiple MIR tasks (Choi et al., 2017).

3. Method

We implemented and tested two different approaches to attack this problem [1].

3.1. Approach A

First, we searched for common feature extractor for audio processing and decided to use openSMILE (Eyben et al., 2010), a state-of-the-art audio feature extractor that still receives regular updates. To ensure that the tool extracts sufficient amount of features, we used the `ComParE_2016` feature set of openSMILE, which generates 6,373 features. We use the "functional" feature level of openSMILE, which applies statistical, polynomial regression coefficients, and transformations to the low-level features, because openSMILE suggests that this feature level is common for music information retrieval.

Figure 1 shows the structure of approach A. First, our preprocessor script takes all the audio files, extracts features using openSMILE, and then save the features into a CSV file. Then, our classifier that utilizes the Scikit-learn module (Pedregosa et al., 2011) loads the CSV file and trains the linear Support Vector Machine (SVM) classifier. As suggested by the documentation of Scikit-learn, we added a standard scaler pass to the SVM classifier.

[1]Code available at https://github.com/yufeidu/COMP562Project

3.2. Approach B

While we were exploring methods to extract audio features, we also considered an alternative approach: could we simply transform the audio files into images and then treat each pixel of an image as a feature, like the MNIST handwritten digit classification problem (LeCun, 1998). To verify our assumption, we built another approach that uses the each pixel of the spectrogram of the audio file as a feature.

Figure 1 shows the structure of approach B. First, our preprocessor script takes all the audio files, converts them into spectrograms using Matplotlib (Hunter, 2007), and saves the spectrograms as TIFF (Adobe Developers Association et al., 1992) images. We disabled axis and borders, so each spectrogram would only include the graph itself. Then, our classifier loads all the images and trains the decision tree classifier using Scikit-learn (Pedregosa et al., 2011). We decided to use decision tree for this approach because of the overwhelming amount of features: the resolution of a spectrogram image is 334 by 217 pixels, and each pixel has four channels (red, green, blue, and alpha), generating a total of 289,912 features, which is impractical for SVM due to memory consumption.

4. Experiments

4.1. Dataset

Our dataset contains 21 music albums by 10 different groups of artists with a total of 1004 audio files. We collected the audio files by either ripping audio CD that we legally purchased or buying digital albums from lossless music distribution websites (Sony Music Solutions Inc.; OTOTOY). Our collection contains both voice pieces and instrumental pieces, as well as both live recordings and "virtual" instruments generated using a synthesizer.

For all the audio files, we converted them into the

Approach	F-1	Precision	Recall	Accuracy
A	28.4%	28.6%	28.2%	29.0%
B	95.8%	95.8%	95.7%	95.8%
A with decision tree	49.6%	49.5%	49.6%	49.5%

Table 1. Experiment results for approach A, B and a version of approach A with decision tree instead of SVM.

FLAC (Xiph.Org Foundation) format, one of the most popular file formats for lossless music, with the sample rate set to 44100Hz, the same as the sample rate of audio CD. We needed to manually unify the sample rate because some digital lossless albums use higher sample rate such as 96000Hz. This set of audio files is used to generate the lossless part of our dataset.

Next, we converted the lossless audio files into the lossy MP3 (Hoffman et al., 1998) and converted the MP3 files back to FLAC to create the lossy (i.e., the fake lossless audio) part of our dataset. For conversion to MP3, we used the "Variable (VBR)" bitrate mode, since the converter tool (Portet) claims that this mode has the best quality, and we set the quality to "insanely high".

Finally, to unify the length of the audio files and to increase the amount of both lossless and lossy audio data, we divide each audio file into sections of 30 seconds, generating 6,877 lossless audio files and 6,877 lossy audio files in total, each contains 30 seconds of audio.

4.2. Experimental Setup

We use the same dataset and the same setup to evaluate approach A and approach B. We utilized Scikit-learn's (Pedregosa et al., 2011) cross-validation library function to perform a 5-fold cross-validation and took the average of the results of the five runs. We measure the F-1 score, precision, recall, and accuracy of each approach.

4.3. Results

Table 1 lists the results of approach A and approach B. The results shows that surprisingly, approach A with audio features and SVM performs significantly worse than approach B with spectrogram as features and decision tree. We expected that the SVM as a more robust and complex learning model would perform better than a simple shallow learning model like the decision tree. We believe that the huge difference in the results between approach A and approach B is mostly because of the different features. While the audio features extracted by openSMILE (Eyben et al., 2010) would fit tasks such as music genre identification or song emotion prediction, these audio features may not fit our task of identifying audio compression.

To verify our assumption about the cause of the differences in results, we built another version of approach A with the decision tree classifier instead of SVM. The results of this version is also included in Table 1. While the results of this version is still significantly worse than the results of approach B, it shows a noticeable increase from the results of approach A with SVM classifier. This likely indicates that the linear model in SVM is not suitable for this task, and if we were to continue improving approach A, we likely need to utilize some kernel tricks.

The results of approach B show that spectrograms are sufficient to show artifacts from lossy audio compression. This matches our expectation: lossy compression would likely change or remove some audio frequencies in order to save space, and spectrograms can directly show the frequencies.

5. Conclusion

In this report, we presented two machine learning approaches for upscaled lossless audio identification. Our first approach uses audio features extracted from openSMILE (Eyben et al., 2010), a state-of-the-art feature extractor for audio, and SVM classifier; our second approach uses spectrograms generated from the audio files as features and decision tree classifier. We evaluated both approaches using a dataset of 6,877 real lossless audio files from real-world commercial music and 6,877 upscaled audio files from converting the real ones. The results show that the second approach, with spectrogram features and decision tree classifier, could achieve highly accurate results, with an F-1 score of 95.8%.

References

Adobe Developers Association et al. Tiff revision 6.0. *Internet publication: http://www. adobe. com/Support/TechNotes. html*, 1992.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, 1992.

Chakraborty, S. S. and Parekh, R. Improved musical instrument classification using cepstral coefficients and neural networks. In *Methodologies and Application Issues of Contemporary Computing Framework*, pp. 123–138. Springer, 2018.

Choi, K., Fazekas, G., Sandler, M., and Cho, K. Trans-

fer learning for music classification and regression tasks. *arXiv preprint arXiv:1703.09179*, 2017.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Duda, R. O., Hart, P. E., et al. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973.

Eyben, F., Wöllmer, M., and Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

Fix, E. and Hodges, J. L. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3):238–247, 1989.

Fu, Z., Lu, G., Ting, K. M., and Zhang, D. A survey of audio-based music classification and annotation. *IEEE transactions on multimedia*, 13(2):303–319, 2010.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pp. 1243–1252. PMLR, 2017.

Harada, T., Ishizaki, F., Hamada, M., Horie, N., Nitta, Y., Nitta, K., Katsuoka, H., and Nakamura, S. Circadian rhythm of heart-rate variability and autonomic cardiovascular regulation in parkinson's disease. *Autonomic Neuroscience: Basic and Clinical*, 158(1):133, 2010.

Harada, T., Kurai, R., Ito, S., Nitta, Y., Aoi, S., Ikeda, H., Iida, T., Miyazaki, H., Umei, N., Chikamura, C., et al. Effect of joyful and anxiety-provoking music on autonomic nervous system function. *International Medical Journal*, 24(2):211–213, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hoffman, D., Fernando, G., Goyal, V., and Civanlar, M. Rfc2250: Rtp payload format for mpeg1/mpeg2 video, 1998.

Hunter, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

Im, H. and Kim, N. W. Three personas of potential high-resolution music users. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2851–2856, 2016.

Ito, S., Harada, T., Ishizaki, F., Yamamoto, R., Niyada, K., Miyazaki, H., Nitta, Y., Chikamura, C., Suehiro, K., and Nitta, K. Effect of high-resolution audio on function of autonomic nervous system. *International Medical Journal*, 23(4):1–3, 2016a.

Ito, S., Harada, T., Miyaguchi, M., Ishizaki, F., Chikamura, C., Kodama, Y., Niyada, K., Yamamoto, R., Nitta, Y., Shiromoto, O., et al. Effect of high-resolution audio music box sound on eeg. *Int. Med. J*, 23:1–3, 2016b.

Kim, Y. E. and Whitman, B. Singer identification in popular music recordings using voice coding features. In *Proceedings of the 3rd international conference on music information retrieval*, volume 13, pp. 17. Citeseer, 2002.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105, 2012.

Kuribayashi, R., Yamamoto, R., and Nittono, H. High-resolution music with inaudible high-frequency components produces a lagged effect on human electroencephalographic activities. *NeuroReport*, 25(9):651–655, 2014.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Li, T., Ogihara, M., and Li, Q. A comparative study on content-based music genre classification. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 282–289, 2003.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

Marques, J. and Moreno, P. J. A study of musical instrument classification using gaussian mixture models and support vector machines. *Cambridge Research Laboratory Technical Report Series CRL*, 4:143, 1999.

Melchior, V. R. High-resolution audio: a history and perspective. *Journal of the Audio Engineering Society*, 67 (5):246–257, 2019.

OTOTOY. Ototoy. URL https://ototoy.jp.

Pampalk, E., Rauber, A., and Merkl, D. Content-based organization and visualization of music archives. In *Proceedings of the tenth ACM international conference on Multimedia*, pp. 570–579, 2002.

Pampalk, E., Dixon, S., and Widmer, G. On the evaluation of perceptual similarity measures for music. In *of: Proceedings of the sixth international conference on digital audio effects (DAFx-03)*, pp. 7–12. Citeseer, 2003.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Portet, G. Soundconverter - gnome sound conversion. URL https://soundconverter.org.

Reiss, J. D. A meta-analysis of high resolution audio perceptual evaluation. *Journal of the Audio Engineering Society*, 2016.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Schlüter, J. and Böck, S. Improved musical onset detection with convolutional neural networks. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 6979–6983. IEEE, 2014.

Song, Y. and Zhang, C. Content-based information fusion for semi-supervised music genre classification. *IEEE Transactions on Multimedia*, 10(1):145–152, 2007.

Sony Music Solutions Inc. Mora. URL https://mora.jp/.

Tzanetakis, G. and Cook, P. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.

Xiph.Org Foundation. Flac - free lossless audio codec. URL https://xiph.org/flac.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057. PMLR, 2015.