A NOVEL METHOD TO DETECT INSTRUMENTAL MUSIC IN A LARGE SCALE MUSIC CATALOG

Wo Jae Lee, Emanuele Coviello

Amazon Music

ABSTRACT

A large scale music catalog contains diverse types of sound recordings. In this paper, we present a methodology to identify instrumental music with high precision and high recall. Our method starts by separating a recording into a vocals track and a background track. Then, we process the vocals track with a singing voice detection model, to estimate the amount of singing voice in a song. Finally, we analyze the background track using a neural network trained to classify between instrumental and non-instrumental sounds. We demonstrate the effectiveness of our approach on instrumental music detection through a comparative evaluation against several state-of-the-art models.

Index Terms— singing voice detection, instrumental music detection, music signal processing

1. INTRODUCTION

Music streaming services cater to a variety of intents of a global user bases. Discriminating between instrumental music vs. music with vocals is critical to many use cases. Examples range from identifying instrumental or utility music (for instance to create playlists for focus or relaxation), or as a preliminary steps to singing language classification which is particularly important in multi-lingual markets. In this paper we tackle the problem of automatic instrumental music detection.

There is ample literature investigating scalable content-based methods for automatic music tagging [1, 2, 3], where a common practice is to train a supervised multi-class multi-label model on low-level content-features, covering audio or multiple data modalities. This has registered success in many applications such as prediciton of genre, mood, instrumentation, or language [1, 4, 5]. However, our investigation suggests that plain adoption of this approach results in low recall at high level of precision for the detection of instrumental music. We hence propose a novel multi-stage method, that first uses a source separation model to split vocals and accompaniment (background), then quantifies the amount of singing voice on the vocals signal, and finally, if that's below a threshold, it uses a binary classifier to detect whether the track is instrumental. Our experimental results demonstrate that our method is superior to state-of-the-art pretrained music tagging models.

2. PRIOR WORK

With the aim of accurately detecting the presence of voice within audio signals, there has been extensive research on voice activity detection (VAD). The task of VAD involves predicting whether segments of input audio contain speech or background noise, and is typically framed as a binary classification problem. Early VAD methods relied on high-level acoustic features [6, 7], which were effective in scenarios with low levels of background noise [8]. More recent work has switched to deep-learning models for VAD, investigating recurrent neural networks [8], convolutional neural networks [9], or convolutional long short-term memory deep neural networks (CLDNNs) [10]. Typically, these models predict the presence of speech in individual segments of audio using acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coefficients (LPCs), and Perceptual Linear Predictive Coefficients (PLPs) [9, 11]. There are several state-of-the-art VAD models, such as SpeechBrain [12] and Marblenet [9], publicly available for inference. However, models trained on conversational speech data tend to fail when directly applied to music, due to the different intonation and phonation between singing and speaking, and the difference in acoustic background [13, 14].

To identify singing voices in music signals, a source separation model is also used to isolate vocals from a source track [14]. Hsu et al. [15] applied Harmonic/Percussive Source Separation (HPSS) [16] to source signals to attenuates the energy from music accompaniment. Leglaive et al. [17] also utilized HPSS in a pre-processing phase to decompose input signals into harmonic and percussive components. To enhance the model performance, Schlüter et al. [18] introduced a data augmentation technique (e.g., pitch shifting). Also, song (music with voice) and instrumental music (music without voice) are distinguished based on the existence of vocal components within audio signals [19, 20]. Earlier work [19] directly applied a classification model to hand crafted features to detect the presence of singing voice. Ghosal et al. [19] proposed a model categorizing music into two distinct types, instrumental and song. The proposed model was motivated by spectrogram analysis, where discriminative features were identified in the frequency peaks of instrumental and song signals – instrumental signals are characterized by more stable frequency peaks, whereas the presence of voice eliminates such stability.

3. PROPOSED METHOD

In this study, we present a novel approach for detecting instrumental music. We define an instrumental track as a recording that excludes any form of vocalization. Therefore, the presence of vocals is a key determinant in identifying a song as instrumental if the accompanied music is the sound of an instrument. Figure 1 illustrates our approach for instrumental music detection. Our methodology starts by separating an audio file into a vocals track (i.e., singing voice and other vocal sounds) and a background track (i.e., accompaniment), using a pretrained neural network [21]. Subsequently, we split the vocals track into short clips, and classify each clip for the presence of voice to then compute the overall prominence of vocals. If this is below a given threshold, we classify the background track with a model to identify whether the track is instrumental music. We further elaborate each step in the following sections.

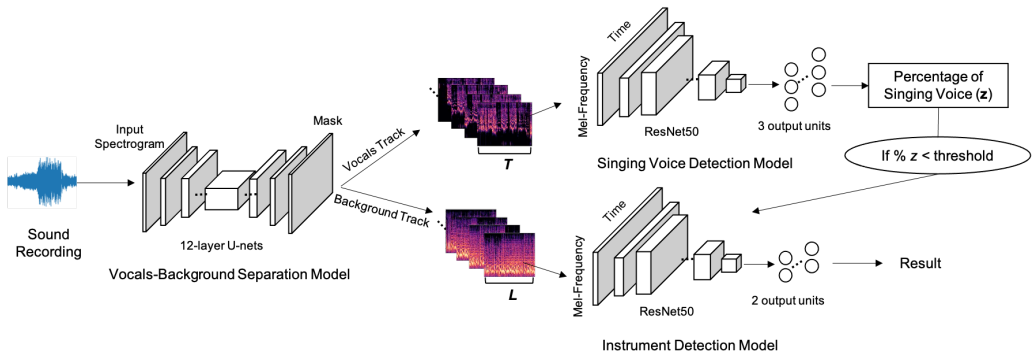

Figure 1: A proposed method to detect instrumental music using vocals-background separation (VBS) model, singing voice detection (singing-VD) model, and instrument detection (ID) model

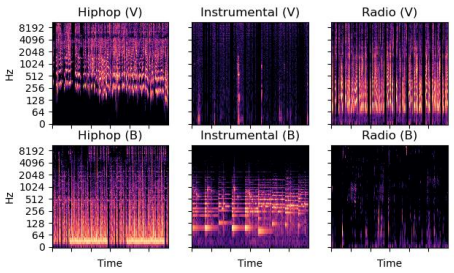

Figure 2: Mel-Spectrogram of vocals track (top) and background track (bottom) separated from different types of sound recordings where V and B denote vocals and background, respectively.

3.1. Vocals-background Separation

We rely on a pretrained music source separation model, Spleeter [21] to separate an audio file into a vocals part and background part. The Spleeter model is based on a U-net convolutional neural network (CNN) architecture, which had demonstrated to achieve state-of-the-art performance on a music source separation task on the musdb18 dataset [22]. Furthermore, its robustness has been confirmed across a diverse range of music genres, thus making it a popular choice in music source separation tasks [23, 24].

To show Spleeter at work, in Fig. 2 we plot the Mel spectrogram of the vocals track (V – top row) and background track (B – bottom row) extracted from a hip hop recording, an instrumental music recording, and a radio/spoken recording. Consistently with our intuition, both the vocals and background tracks extracted from the hip-hop recording have comparatively large Mel coefficients. Instead, for the instrumental music example we see lower Mel coefficients in the vocals tracks and larger Mel coefficients for the background track, and vice-versa for the radio play example. Still we observe an amount of leakage (that could harm classification performance) that our method will address.

Source separation provides signals with background-removed and vocals-removed from the original music signals. These separated signals may be more reliable for detecting instrumental music, which is a recording that does not include any vocals but does contain sound from musical instruments. In Section 4, we compare our proposed method to baseline models that analyze the original music signals directly.

3.2. Singing Voice Detection (singing-VD) Model

In the next step, we split the vocals track into non-overlapping segments of duration T, and train a ResNet50 [25] to classify each segment as silence (including background noise), instrumental sound, or vocals (including singing or speech). The ResNet50 has 50 layers with residual connection, and it has proven successful in several ML tasks across different domains. More generally, CNNs have been a common choice for vocal activity detection [9, 18].

To compile a ground truth dataset for this sub-task, we first collect audio recordings that cover silence/noise, instrumental sounds, and vocals/speech. We then randomly sample segments of duration T from these recordings, and propagate the recording-level labels to each of its clips. We further elaborate on our training data collection strategy in Section 4.

In contrast to the conventional approach widely adopted in VAD that employs a binary label, i.e., vocals and non-vocal, our model estimates a posterior probability of silence, instrumental sounds, and singing voice (or spoken words) using a softmax function. The rationale for this decision is rooted in the observation that, despite using source separation, we can observe both periods of silence as well as leakage of dampened instrumental sounds in the vocals track (see Fig. 2). This approach can enhance the robustness of our model to inaccuracies or imperfect vocals/background separation. At inference time, the model makes a prediction on each segment, and the percentage of voice, $\mathbf{z}$, is calculated as follows:

$$\mathbf{z} = \frac{\text{number of segments with voice}}{\text{number of total segments}} \tag{1}$$

In this study, we classify a segment as a voice segment if it has the highest observed probability on the voice class. We use a threshold that is computed using a validation set to filter tracks with no or low vocals prominence for the next instrument detection step. More details are discussed in 4.

3.3. Instrument Detection (ID) Model

A zero or low percentage singing of voice is an indication that the given recording falls within the categories of instrumental music

or other non-voice non-music categories (e.g., white noise, natural/environmental sounds, etc). To distinguishes between instrumental and non-instrumental sounds, we train a binary classifier based on the ResNet50 [25] architecture. Similarly as for the model described in Section 3.2, we compile a training dataset by collecting multiple non-overlapping segments of L seconds from instrumental music and non-instrumental music recordings. We train the model at the segment level, and at inference time we use average pooling to aggregate segment-level predictions into a song-level prediction. Finally, the song-level score for the instrumental class is compared against a threshold tuned on a validation set, to decide whether a source recording is instrumental music.

4. EXPERIMENTS

4.1. Datasets and Experimental Setup

We compiled three datasets, $\mathcal{D}_{svd}$, $\mathcal{D}_{id}$, $\mathcal{D}_{imd}$ from an internal music catalog for training and evaluation of the proposed models.

First, to train the singing-VD model with instance-level training, we need segment-level labels. Directly labelling a dataset at the segment-level is a costly and non scalable process. Access to a strongly-labeled datasets can have its advantages (e.g., for accurate evaluation of segment-level tasks), but in practice such datasets have limited size [26]. Since in our end-to-end application we are ultimately interested in recording-level annotations, we can adopt a more scalable approach to compile a larger dataset that is *weakly* annotated at the segment level. Specifically, we sample 39,282 recordings with associated human labels corresponding to *voiced* and *non-voiced* recordings (including songs from 530 distinct genres and micro-genres as well as radio plays). We apply VAS (as described in Section 3.1) to extract the vocals and background tracks, and then samples segments of T (from vocals) or L seconds (from background), and propagate the recording labels (voiced and non-voice) to the corresponding segments. In addition, we apply a *silence* label to segments with energy below a given level.[1] The resulting dataset contains 18,790, 22,585, and 68,917 segments of silence, instrument, and voice, respectively. To prevent potential information leakage between train, validation, and test sets, the dataset is split at the artist-level into train/validation/test sets with a ratio of 0.7/0.15/0.15. This dataset is referred to as $\mathcal{D}_{svd}$.

Second, a diverse range of sound recordings, including instrumental music and non-instrumental music, were collected from a catalog to train and evaluate an ID model. Non-instrumental music encompassed white noise, nature sounds, animal sounds, etc. A total of 88,631 and 151,040 segments of instrument and non-instrument sounds were collected from 31,661 and 39,153 sound recordings, respectively. Following the methodology used in $\mathcal{D}_{svd}$, silence segments were removed from the collected samples and the dataset was split into train, validation, and test sets at the artist level. The dataset is referred to as $\mathcal{D}_{id}$.

Lastly, the end-to-end system shown in Fig. 1 is evaluated on 37,711 tracks not included in $\mathcal{D}_{svd}$ and $\mathcal{D}_{id}$ for the task of instrumental music detection. The dataset, denoted as $\mathcal{D}_{imc}$, contains 15,040 instrumental tracks and 22,671 non-instrumental tracks, and the latter of which have been gathered from a diverse selection of sound recordings including acoustic, acapella, and utility music.

[1]We treat a segment below a certain dB threshold as silence in vocals and background tracks. We adjusted this threshold in a preliminary experiment based on a qualitative assessment.

4.1.1. *Evaluation metrics and Baseline models*

The evaluation of the singing-VD and ID models is conducted using two metrics, F1 score and mean average precision (mAP). We compare several models including (1) models based on various architectures using Mel spectral features, (2) models based on the ResNet50 architecture using Mel spectral features or MFCC features, and (3) a state-of-the-art VAD model based on a convolutional recurrent deep neural network (CRDNN) [27].

To evaluate the end-to-end system in Fig. 1, we report precision, recall, and F1-score using song-level predictions, and compare the proposed approach to several state-of-the-art tagging models including MTT-MUSICNN [28, 1], MTT-MUSICNN [29, 1], and YAMNet [2] (here, relevant classes associated with instrumental music are selected for the evaluation from a vocabulary of tags, including tags such as "no singer," "no singing," "instrumental," and "musical instrument"). For MTT-MUSICNN and MSD-MUSICNN, the models return the top five tags as predictions for each song. For YAMNet, we use the validation set to calculate a threshold optimizing a F1 score. This threshold is then used to evaluate the performance on the test set.

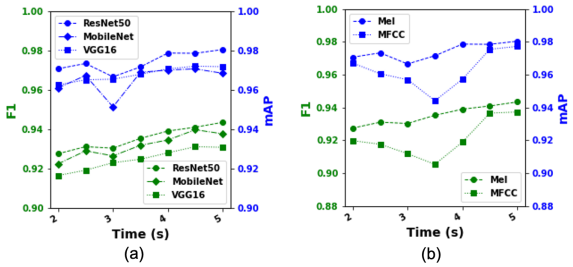

Figure 3: Overall performance of the proposed singing-VD model measured in F1 and mAP at different T and comparison to the baseline models: (a) different architectures using Mel feature (left) and (b) ResNet50 using Mel or MFCC features (right).

4.2. Results

In this section we present results for singing voice detection, instrumental sound detection, and the end-to-end system.

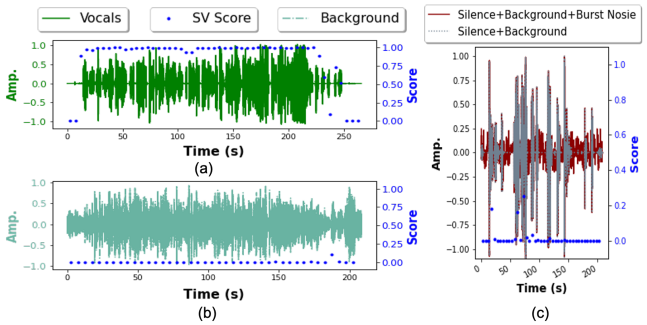

Figure 4: Visualization of singing voice (SV) scores on (a) vocals track (upper left), (b) background track (lower left), and (c) silence + background + burst noise (right).

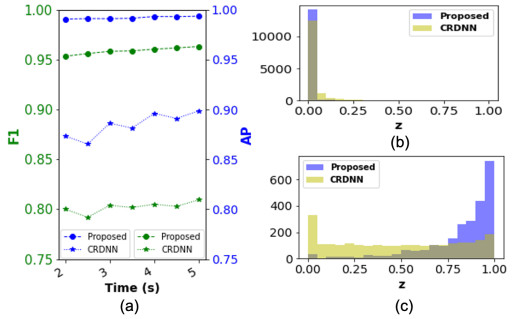

Figure 5: Comparison between proposed method and CRDNN [27]: (a) overall performance measured in F1 and AP at different T, (b) histogram of $\mathbf{z}$ on instrumental recordings (upper right), and (c) histogram of $\mathbf{z}$ on acapella recordings (lower right).

4.2.1. *Singing Voice Detection*

We start by computing low-level audio features (Mel Spectrogram, MFCCs) with n_fft=2048, n_mels=90, win_length=2048, and hop length=1024. We then group features into segments of T seconds, and train the singing voice detection model on these segments. We compare different T values, as well as different neural network architectures and audio features. The evaluation in this section is at the segment level.

The results of evaluation of the trained singing-VD models using various architectures and features at different values of T are presented in Fig. 3.[2] In the left and right panel of the figure, we compare different architectures (ResNet50, MobileNet [31] and VGG16 [32]) and different low-level features (Mel spectrograms and MFCCs), for different values of T. ResNet with Mel features is the best performing setting, both in terms of F1 and mAP (at 0.9434 and 0.9804, respectively). Additionally, we see that performance increase as T increases, and levels off at around $5s$. We hence set $T = 5$. Since we're sampling non-overlapping segments, the choice of T does not impact run-time performance significantly.

To illustrate the best model at work, in Fig 4, we visualize the singing voice (SV) score (i.e., posterior probabilities at the frame level) for (a) a vocals track (extracted singing voice), (b) a background track, and (c) a track combining silence, background track, and burst noise. We see that the model assigns high scores to the segments with vocals, and nearly all zero scores to the background segments. To illustrate the model's robustness, in Fig. 4(c) we added zero-mean Gaussian noise with a standard deviation of 0.1 (to a random frame every 0.1s, i.e., burst noise) in the background track. We can see that the model still assigns low scores to frames.

Finally, we compare the proposed singing-VD model with CRDNN [27]. The results in Fig. 5(a) indicate that the proposed model consistently outperforms CRDNN, as evidenced by higher F1 and average precision (AP), with an average improvement of 0.1567 and 0.1074, respectively. Additionally, in Fig. 5(b) we compare the percentage of singing voice (i.e., $\mathbf{z}$) computed with the proposed model and CRDNN on instrumental (top) and acapella recordings (bottom). We see that, compared to CRDNN, the proposed model results in a more discriminative score, which support our earlier conclusions in terms of F1 score and average precision.

[2]We conducted a hyper-parameter search based on a validation set. For the ResNet50 model on Mel spectral features, we selected a mini-batch size of 32, 20 epochs, an Adam optimizer [30] with an initial learning rate of 0.001 that is reduce by monitoring the validation loss over the course of the training epochs, L2 regularization with a weight decay parameter of 0.001.

Table 1: Comparison between proposed method and baseline models on the task of instrumental music detection.

Model	Precision	Recall	F1
MTT_MUSICNN	0.7692	0.2133	0.3341
MSD_MUSICNN	0.6812	0.4772	0.5612
YAMNet	0.5486	0.6960	0.6136
Ours	***0.9050***	***0.8722***	***0.8883***

4.2.2. *Instrumental Sound Detection*

We adopt a similar strategy as for the singing-VD model to train and evaluate the ID model using $\mathcal{D}_{id}$. The ID model examines the backgorund track to classify between instrumental and non-instrumental sounds. Specifically, we compared (1) models using Mel feature with a variety of architectures, including ReseNet50, MobileNet, and VGG16, (2) models based on the ResNet50 architecture that incorporate either Mel or MFCC features, and (3) different values of L. For the task of instrumental sound detection, we found that a ResNet50 model with Mel as input features and L=4.5 was better than the alternatives, with F1 and mAP values measuring 0.9511 and 0.9862, respectively.

4.2.3. *End-to-end system*

We tune a threshold value for $\mathbf{z}$ using song-level prediction to apply the "instrumental music" label by optimizing F1 score on a validation set (we set the threshold at 0.05). In Table 1, using $\mathcal{D}_{imd}$, we compute precision, recall, and F1 score for the proposed method and three state-of-the-art baseline tagging models. Our model differs from the baseline models in that it operates on separated signals, whereas the baseline models make inference directly on the original sound recordings. In comparison to MTT_MUSICNN and MSD_MUSICNN, which return the top five tags (we assume a prediction to be valid if the associated tag described in Section 4.1.1 is included in the top five tags), our model's F1 score is higher by 0.5542 and 0.3271, respectively. YAMNet shows the highest performance in terms of F1 score compared to the baseline models. Compared to the proposed model, F1 score of the proposed model is higher by 0.2747.

5. CONCLUSIONS

We proposed a method which involves the separation of audio signals into distinct components of vocals and background tracks, with each source evaluated through the singing-VD and ID models to detect the presence of singing voice and instrumental sounds, respectively. The experiment results clearly demonstrate the effectiveness of the proposed approach for the task of instrumental music detection with a notable improvement in precision and recall through a comparison with state-of-the-art tagging models. Although our evaluation was only on 37k tracks (this was limited by the availability of ground truth data), the method is amenable to large scale catalog. In future work, we will explore how to make inference more efficient, for example by developing a simpler single end-to-end architecture, potentially using the current approach in a teacher/student setting.

6. REFERENCES

[1] J. Pons and X. Serra, "musicnn: Pre-trained convolutional neural networks for music audio tagging," *arXiv preprint arXiv:1909.06654*, 2019.

[2] "Yamnet." [Online]. Available: https://github.com/tensorflow/models/tree/master/research/audioset/yamnet.

[3] H.-H. Wu, C.-C. Kao, Q. Tang, M. Sun, B. McFee, J. P. Bello, and C. Wang, "Multi-task self-supervised pre-training for music classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 556–560.

[4] J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," *arXiv preprint arXiv:1711.02520*, 2017.

[5] K. Choi and Y. Wang, "Listen, read, and identify: multimodal singing language identification of music," *arXiv preprint arXiv:2103.01893*, 2021.

[6] K.-H. Woo, T.-Y. Yang, K.-J. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.

[7] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3-4, pp. 271–287, 2004.

[8] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 483–487.

[9] F. Jia, S. Majumdar, and B. Ginsburg, "Marblenet: Deep 1d time-channel separable convolutional neural network for voice activity detection," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6818–6822.

[10] R. Zazo, T. N. Sainath, G. Simko, and C. Parada, "Feature Learning with Raw-Waveform CLDNNs for Voice Activity Detection," in *Proc. Interspeech 2016*, 2016, pp. 3668–3672.

[11] B. Lehner, J. Schlüter, and G. Widmer, "Online, loudness-invariant vocal detection in mixed music signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1369–1380, 2018.

[12] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[13] Y. Ohishi, M. Goto, K. Itou, and K. Takeda, "Discrimination between singing and speaking voices," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[14] R. Monir, D. Kostrzewa, and D. Mrozek, "Singing voice detection: a survey," *Entropy*, vol. 24, no. 1, p. 114, 2022.

[15] C.-L. Hsu, D. Wang, J.-S. R. Jang, and K. Hu, "A tandem algorithm for singing pitch extraction and voice separation from music accompaniment," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 5, pp. 1482–1491, 2012.

[16] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *2008 16th European Signal Processing Conference*. IEEE, 2008, pp. 1–4.

[17] S. Leglaive, R. Hennequin, and R. Badeau, "Singing voice detection with deep recurrent neural networks," in *2015 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 121–125.

[18] J. Schlüter and T. Grill, "Exploring data augmentation for improved singing voice detection with neural networks." in *ISMIR*, 2015, pp. 121–126.

[19] A. Ghosal, R. Chakraborty, B. C. Dhara, and S. K. Saha, "Song/instrumental classification using spectrogram based contextual features," in *Proceedings of the CUBE International Information Technology Conference*, 2012, pp. 21–25.

[20] ——, "A hierarchical approach for speech-instrumental-song classification," *SpringerPlus*, vol. 2, no. 1, pp. 1–11, 2013.

[21] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 2020, deezer Research. [Online]. Available: https://doi.org/10.21105/joss.02154

[22] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The musdb18 corpus for music separation," 2017.

[23] R. Henning, A. Choudhry, and M. Ma, "Deep learning based music source separation," *SCSU Journal of Student Scholarship*, vol. 1, no. 2, p. 3, 2021.

[24] S. Park and B. S. Chon, "Gsep: A robust vocal and accompaniment separation system using gated cbhg module and loudness normalization," *arXiv preprint arXiv:2010.12139*, 2020.

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

[26] Y. Hou, F. K. Soong, J. Luan, and S. Li, "Transfer learning for improving singing-voice detection in polyphonic instrumental music," *arXiv preprint arXiv:2008.04658*, 2020.

[27] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[28] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging." in *ISMIR*. Citeseer, 2009, pp. 387–392.

[29] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *ISMIR*, 2011, pp. 591–596.

[30] K. Diederik and B. Jimmy, "Adam: A method for stochastic optimization. arxiv prepr. int," *arXiv preprint arXiv:1412.6980*, 2014.

[31] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.