2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP) | 979-8-3503-2411-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/MLSP55844.2023.10285863

# EXPLOITING MUSIC SOURCE SEPARATION FOR SINGING VOICE DETECTION

*Francesco Bonzi*[*, 1], *Michele Mancusi*[*, 2], *Simone Del Deo*[1],
*Pierfrancesco Melucci*[1], *Maria Stella Tavella*[1], *Loreto Parisi*[1], *Emanuele Rodolà*[2]

Musixmatch[1], Sapienza University of Rome[2]

## ABSTRACT

Singing voice detection (SVD) is an essential task in many music information retrieval (MIR) applications. Deep learning methods have shown promising results for SVD, but further performance improvements are desirable since it underlies many other tasks. This work proposes a novel SVD system combining a state-of-the-art music source separator (Demucs) with two downstream models: Long-term Recurrent Convolutional Network (LRCN) and a Transformer network. Our work highlights two main aspects: the impact of a music source separation model, such as Demucs, and its zero-shot capabilities for the SVD task; and the potential for deep learning to improve the system's performance further. We evaluate our approach on three datasets (Jamendo Corpus, MedleyDB, and MIR-1K) and compare the performance of the two models to a baseline root mean square (RMS) algorithm and the current state-of-the-art for the Jamendo Corpus dataset.

***Index Terms***— Singing Voice Detection, Music Source Separation, Demucs, Zero-shot Learning

## 1. INTRODUCTION

Singing voice detection (SVD) is a classification task determining whether a singing voice exists in a given audio segment. It is crucial in many Music Information Retrieval (MIR) applications, such as lyrics alignment [1], singer identification [2, 3], and lyrics transcription [4]. Traditional approaches to SVD focused on analyzing the audio mixture directly, extracting features from the raw waveform, and employing various machine-learning techniques to classify the singing voice.

In recent years, there has been a shift in the SVD research landscape due to the growing interest in utilizing music source separation (MSS) techniques. Two years ago, [3] proposed a state-of-the-art (SOTA) SVD method that leverages MSS to preprocess the input audio signal and subsequently classify the singing voice. This approach has significantly improved SVD performance, indicating the potential benefits of incorporating MSS techniques in SVD systems.

In this work, we follow this research direction and build upon the SOTA MSS method, Demucs, by integrating it with two downstream models: LRCN and a Transformer network. Our study aims to assess the effectiveness of MSS methods on SVD tasks and determine whether the two downstream models can further enhance performance.

## 2. RELATED WORKS

In recent years, there has been a growing interest in using deep learning techniques, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), such as long short-term memory (LSTM) networks, for solving the task of singing voice detection. Lee et al. [5] proposed end-to-end approaches using CNNs and LSTMs, respectively, to process the audio mixture and classify the singing voice directly.

In addition to the choice of neural network architecture, researchers have also explored different feature extraction techniques to represent the audio mixture. In 2015, Schlüter et al. [6], Lehner et al. [7], and Leglaive et al. [8] proposed three different methods based on CNNs, LSTM-RNNs, and bidirectional LSTM (Bi-LSTM) networks, respectively. These methods extract high-level audio features such as spectrograms, mel-spectrograms, MFCCs, and singing voice and percussive components from the harmonic-percussive source separation (HPSS) algorithm. Schlüter et al. [6] investigated the effectiveness of data augmentation methods on spectrograms and mel-spectrograms, while Lehner et al. [7] used 30 MFCCs and Leglaive et al. [8] combined the singing voice and percussive components from the HPSS algorithm. The best model achieved an accuracy of 0.923 [6], while the best f1-score was 0.910 [8].

In 2016 and 2017, Choi et al. [9, 10] studied the effectiveness of using pre-trained convnet features for singing voice detection and achieved similar results compared to previous works.

Recently, Zhang et al. [11] proposed a new approach based on a CNN combined with an LSTM network called the long-term recurrent convolutional network (LRCN). Un-

[*] Equal contribution,
[1] Email-addresses: {francesco.bonzi, simone.deldeo, pierfrancesco.melucci, mariastella.tavella, loreto.parisi}@musixmatch.com,
[2] Email-addresses: {mancusi, rodola}@di.uniroma1.it

979-8-3503-2411-2/23/$31.00 ©2023 IEEE

Authorized licensed use limited to: Universita degli Studi di Roma La Sapienza. Downloaded on October 27,2024 at 22:49:47 UTC from IEEE Xplore. Restrictions apply.

like previous methods, the LRCN approach used the singing voice-separated signal as input rather than the audio mixture. This approach achieved an accuracy of 0.924 and an f1-score of 0.927, outperforming previous methods. In addition to the differences in network architecture and feature extraction techniques, the LRCN approach differs from previous end-to-end approaches in using a pre-trained vocal separation algorithm to extract the singing voice signal. This approach could improve the classification quality and reduce interference from other audio sources, but it is sensitive to the separation quality.

## 3. METHOD

In 2021, a new open-source SOTA music source separator was proposed in [12]. It was developed by Facebook AI Research [1] and introduced a number of architectural changes to the previous Demucs architecture [13], considerably improving the quality of source separation for music. We used this state-of-the-art (SOTA) pre-trained model to address SVD, stacking Demucs to a downstream system, as proposed in [11]. To verify the importance of a music source separation model, we trained two networks: an LRCN and a Transformer, directly on the mixtures and the separated vocal tracks produced by Demucs on Jamendo Corpus. In addition, to assess the zero-shot capabilities of Demucs, we stack a classical signal processing root mean square (RMS) algorithm after it. Moreover, we tested the combination of Demucs and a neural network (the LRCN and the Transformer) on other datasets, such as MIR1k and MedleyDB, to assess the potential for deep learning to improve performance further.

To the best of our knowledge, we are the first to use the Transformer network for the task of SVD, motivated by the successes of this architecture in other classification tasks [14, 15].

The subsequent sections will explain these three systems in the following sequence: RMS, LRCN, and Transformer.

### 3.1. Voice Activity Detection through RMS

The root mean square (RMS) of an audio signal is defined as the square root of the mean of the squared values of the signal samples. To compute the RMS, the audio signal is first squared at each sample point, then the squared values are averaged over the entire signal, and finally, the square root of the average is taken. Mathematically, this can be expressed as:

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x[n])^2} \tag{1}$$

where $x[n]$ is the audio sample at time n, and N is the total number of samples in the signal.

[1] https://github.com/facebookresearch/demucs

The RMS metric is a reliable method for measuring the overall amplitude of an audio signal, as it captures both the strength and the duration of the signal. This metric is well-suited for identifying sections of an audio signal in which a person is singing, even in the presence of residual noise and artifacts. Applying the RMS metric to the isolated vocal source obtained from a music source separation system makes it possible to obtain a time-varying measure of the vocal amplitude throughout the song. We used the default parameters of the librosa RMS algorithm, while the threshold used to discriminate between vocals and non-vocals was determined by a grid search in the validation set of the Jamendo Corpus dataset and was set to 0.015. This method can identify the sections in which the singer is present.

### 3.2. LRCN Network

The Long-term Recurrent Convolutional Network (LRCN) architecture, first proposed in [16], is a powerful deep learning model designed for a wide range of applications, including video classification, image captioning, image classification, activity recognition, image labeling, video captioning, and singing voice detection [11]. The network has two main components: a spatial Convolutional Neural Network (CNN) and a temporal Long Short-Term Memory (LSTM) network. The CNN component of the LRCN network is responsible for extracting spatial features from the raw waveform signal. Specifically, the CNN applies a sliding window approach to convolve over the waveform, generating feature maps that capture different aspects of the signal's spatial structure. The LSTM component then processes these feature maps. It is responsible for capturing the temporal dynamics of the audio signal. It processes the feature maps the CNN component generates sequentially over time, using a set of memory cells to capture long-term dependencies between different time steps. The output of the LSTM component is a sequence of high-level features that encode the temporal dynamics of the audio signal. By combining the spatial and temporal features extracted by the CNN and LSTM components, the LRCN network can learn a rich representation of the input audio signal that is well-suited for a wide range of audio processing tasks.

### 3.3. Transformer Network

The Transformer network architecture is a highly effective deep-learning model for processing sequences data [17]. The Transformer is a type of neural network that uses self-attention mechanisms to capture the long-term dependencies in the input sequence. Specifically, the Transformer is designed to model the relationships between different feature maps by considering all pairs of feature maps simultaneously and then computing a weighted sum of the feature maps based on their relative importance. As in the previous architecture, a Convolutional Neural Network (CNN) precedes

Authorized licensed use limited to: Universita degli Studi di Roma La Sapienza. Downloaded on October 27,2024 at 22:49:47 UTC from IEEE Xplore. Restrictions apply.

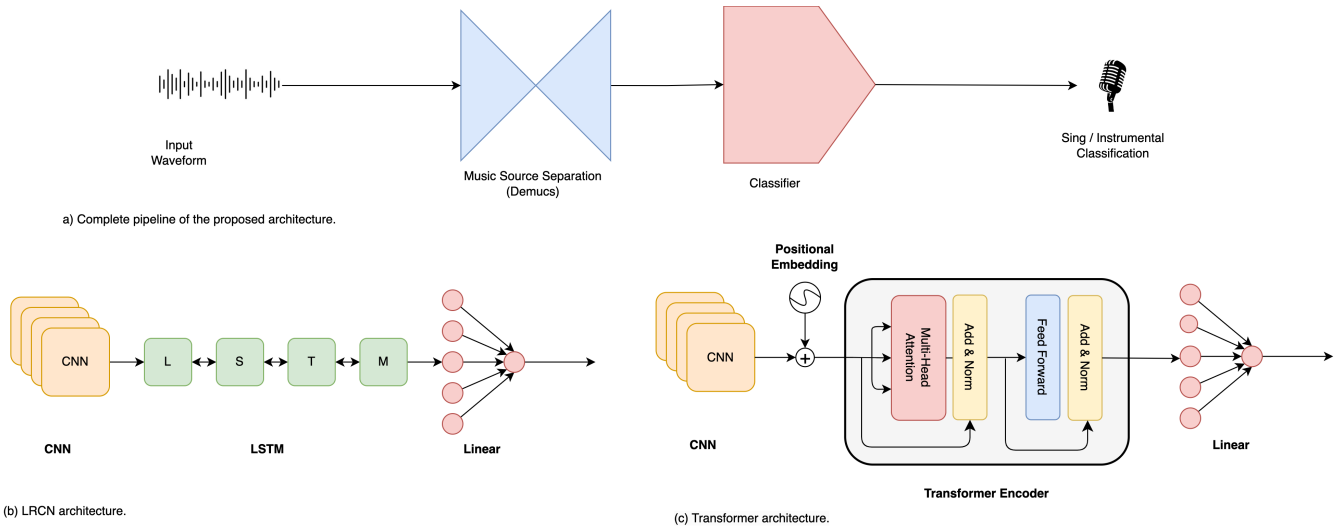


**Fig. 1**. Panel (a) shows the general pipeline, consisting of Demucs followed by a generic Classifier. Panels (b) and (c) detail the two neural Classifiers we tested that follow Demucs.

the Transformer and extracts spatial features from the raw waveform input. The resulting feature maps are then passed to the Transformer component, responsible for processing the sequence of feature maps over time to classify the singing and instrumental sections. By combining the spatial features extracted by the CNN with the self-attention mechanisms of the Transformer, the resulting model is able to learn a highly effective representation of the input audio signal that is well-suited for classification tasks.

These two approaches (LRCN and Transformer) enable the model to automatically learn to recognize singing and instrumental sections directly from the raw waveform input without needing hand-crafted feature engineering, as done in [9, 10].

## 4. EXPERIMENTS AND RESULTS

The experiments were designed to investigate two aspects: *(i)* the impact of a music source separation model, such as Demucs, and its zero-shot capabilities for the SVD task; *(ii)* the potential for deep learning to improve performance further.

To address the first aspect, we train the LRCN and the Transformer directly on mixtures on the Jamendo Corpus dataset [18] (without separating them with Demucs). Later, we kept the weights of Demucs frozen and trained the LRCN and Transformer networks on the Jamendo Corpus dataset, feeding these two models with the separated singing voice signal. Furthermore, to verify the zero-shot capabilities of Demucs, we also measure the performance of a standard signal processing metric (RMS) directly on the separated singing voice signal by Demucs, as described previously.

To address the second aspect, we further investigate the performance of the deep learning models and verify whether training on a specific dataset would enable them to outperform the RMS metric consistently. We expanded our experiments by training the LRCN and Transformer networks (keeping the weights of Demucs frozen) on the separated vocal tracks of two additional datasets: MedleyDB and MIR-1K.

Also inspired by [6], we noticed that our three augmentation techniques improved the neural networks' performance. Augmentations were applied to the separate vocal tracks. The first technique is pitch shifting, which shifts sounds up or down in the frequency spectrum without changing the tempo. The second technique is a gain adjustment, which involves multiplying the audio by a random amplitude factor to reduce or increase the volume, helping the model invariant to the input audio's overall gain. The third technique involves the addition of background noise, while the last approach is polarity inversion, which reverses the audio waveform, effectively inverting the signal phase. All these techniques aim to increase the diversity of the training data and assist the models in learning invariant features for the given task.

In the following subsection, we will describe in detail the three datasets we used to perform the experiments.

### 4.1. Datasets

The Jamendo Corpus includes 93 songs with Creative Commons licenses from Jamendo's free music-sharing website, constituting approximately 7 hours of music in total. Each file is a stereo track with a sampling rate of 44.1 kHz and is manually annotated with singing and no-singing parts by the same person to provide ground truth data. The Jamendo

Authorized licensed use limited to: Universita degli Studi di Roma La Sapienza. Downloaded on October 27,2024 at 22:49:47 UTC from IEEE Xplore. Restrictions apply.

Corpus is publicly available [2]. The official split provides 61 songs for training, 16 for validation, and 16 for testing; the total singing and no-singing frames are about 50% of the whole set for each label, so the dataset is well-balanced.

The MIR1k dataset comprises 1000 singing voice tracks with a musical background. Clips vary in duration between 4s and 13s, their sampling rate is 16kHz, and the dataset has a cumulative duration of 133 minutes. The selected snippets come from 110 karaoke tracks, and they are chosen from a pool of 5,000 Chinese pop songs and performed by MIR lab researchers (consisting of 8 women and 11 men). Annotations of pitch contours in semitone, indices, and types for unvoiced frames, lyrics, and vocal/non-vocal segments were made manually. The MIR1k dataset is publicly available [3].

The MedleyDB dataset comprises 61 audio tracks in WAV format (44.1 kHz, 16-bit) featuring vocal signals accompanied by melody annotations. Each track includes melody annotations and instrument activation data for assessing automatic instrument identification. Labels for vocal and non-vocal segments are determined by pitch values, with nonzero pitch categorized as vocal and zero as non-vocal. A semi-automated process employing monophonic pitch tracking was used for melody annotation. The dataset showcases various music genres, such as Singer/Songwriter, Classical, Rock, World/Folk, Fusion, Jazz, Pop, Musical Theatre, and Rap. The MedleyDB dataset is publicly available [4].

### 4.2. Implementation Details

After the separation step, all datasets are downsampled to 16kHz and transformed into mono samples. Furthermore, we split the MIR1k and the MedleyDB dataset into nonoverlapping training, testing, and validation sets using an 8:1:1 ratio (since the Jamendo Corpus is already split).

In order to perform augmentation, we used the open-source PyTorch-augmentations library [5]. All the training experiments were conducted on the AWS Sagemaker platform [6], with an ml.g4dn.xlarge machine equipped with 1 GPU Nvidia T4. It took around 5 hours to train a model on the original Jamendo training set for 300 epochs. We save network parameters only when the F1-score validation metric exceeds the previous score.

### 4.3. Metrics

To provide a comprehensive view of the results, as proposed in [19], model predictions were compared with the ground truth to obtain the number of false negative (FN), true negative (TN), false positive (FP), and true positive (TP). The frame-wise recall, accuracy, precision, and f1-score were computed to summarize the results.

### 4.4. Results and Discussion

#### *4.4.1. The effects of music source separation and the Demucs zero-shot capabilities in singing voice detection*

Regarding the effects of music source separation, our experiments provided strong evidence for incorporating music source separation, such as Demucs, in the singing voice detection task. As shown in Table 1, when the LRCN and Transformer models were provided with the audio mixture as input, their performance was notably inferior compared to when given the separated vocal signals as input. For example, on the Jamendo dataset, the LRCN model's accuracy increased from 0.868 when using the mixture to 0.960 when using the separated vocals, while the Transformer model's accuracy improved from 0.848 to 0.959 under the same conditions.

These significant performance improvements demonstrate that music source separation is a crucial preprocessing step for enhancing singing voice detection, as asserted by [3]. The models can concentrate on the relevant vocal information by employing Demucs to separate the vocal signals from the audio mixture, while the influence of other audio components, such as background instruments, is minimized. This enables the LRCN and Transformer models to identify and classify the singing voice more accurately and effectively, leading to state-of-the-art results.

Moreover, our investigation of zero-shot capabilities provided valuable insights into the versatility of the Demucs model in the context of singing voice detection. When used in conjunction with a standard signal processing algorithm, Demucs demonstrated competitive performance in the SVD task, as evidenced by the results presented in Table 1. For instance, on the Jamendo dataset, the combination of Demucs and the RMS yielded an accuracy of 0.949, close to the performance of the LRCN (0.960) and Transformer (0.959) models.

These promising results underscore the potential of leveraging pre-trained models, such as Demucs, for tasks beyond their original scope, such as singing voice detection. The ability of Demucs to perform well in the SVD task without any specific training or fine-tuning suggests that its inherent capacity to separate vocals from complex audio mixtures can be effectively utilized across different tasks and applications.

This finding opens up new avenues for future research. It highlights the possibility of harnessing the power of pre-trained models to achieve high-quality performance in various tasks with minimal additional training. Furthermore, it encourages the exploration of transfer learning and multi-task learning techniques to enhance further the adaptability and efficiency of models like Demucs in various audio processing tasks, including singing voice detection.

[2] https://zenodo.org/record/2585988#.YoTKaZNBxhE
[3] https://zenodo.org/record/3532216#.ZFpj8y9Bxf0
[4] https://zenodo.org/record/1715175#.XAzIzxNKjyw
[5] https://pytorch.org/audio/main/tutorials/audio_data_augmentation_tutorial
[6] https://aws.amazon.com/pm/sagemaker

Authorized licensed use limited to: Universita degli Studi di Roma La Sapienza. Downloaded on October 27,2024 at 22:49:47 UTC from IEEE Xplore. Restrictions apply.

| Input Audio | Dataset | Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|
| Mixture | Jamendo | RMS | .563 | .531 | .999 | .679 |
| | | LRCN | **.868** | **.856** | **.893** | **.864** |
| | | Transformer | .848 | .804 | .910 | .845 |
| Vocals | Jamendo | RMS | .949 | .937 | .964 | .949 |
| | | LRCN | **.960** | .945 | **.974** | .958 |
| | | Transformer | .959 | **.953** | .968 | **.960** |
| Vocals | MedleyDB | RMS | .777 | .688 | .957 | .793 |
| | | LRCN | **.854** | **.795** | .916 | **.849** |
| | | Transformer | .833 | .757 | **.936** | .833 |
| Vocals | MIR-1K | RMS | .908 | .935 | .949 | .941 |
| | | LRCN | .921 | **.946** | .955 | .949 |
| | | Transformer | **.926** | .945 | **.960** | **.952** |

**Table 1**. Results of the proposed singing voice detection systems trained and tested on the same datasets.

| Author | Input Audio | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| Schlüter et al. [6] | Mixture | .923 | - | .903 | - |
| Lehner et al. [7] | Mixture | .894 | .895 | .906 | .902 |
| Leglaive et al. [8] | Mixture | **.915** | **.895** | **.926** | **.910** |
| Ours [LRCN] | Mixture | .868 | .856 | .893 | .864 |
| Zhang et al. [11] | Vocals | .924 | .926 | .924 | .927 |
| Ours [LRCN] | Vocals | **.960** | .945 | **.974** | .958 |
| Ours [Transformer] | Vocals | .959 | **.953** | .968 | **.960** |

**Table 2**. Results of the proposed singing voice detection system compared with existing methods on the Jamendo Corpus test set.

#### *4.4.2. The potential for deep learning to further improve performance*

The results presented in Table 1 showcase the potential of deep learning models, such as LRCN and Transformer, to significantly improve performance in the singing voice detection task when compared to the baseline RMS.

In our study, we observed that the performance of our models on Jamendo and MIR1k is significantly different from the performance on MedleyDB, even if all three datasets present songs belonging to the same musical genre. This can be attributed to the fact that these datasets have been annotated differently, with the Jamendo and MIR-1K datasets having been annotated manually while MedleyDB has been annotated automatically. This discrepancy in annotation methods may have led to inconsistencies in the data, which could have, in turn, affected the overall learning process of the models. Moreover, when comparing our deep learning models with existing methods on the Jamendo Corpus test set, it becomes evident that the LRCN and Transformer models offer substantial improvements over the current state-of-the-art, as shown in Table 2. Specifically, when provided with separated vocal signals as input, our LRCN model achieves an accuracy of 0.960, while the Transformer model reaches 0.959. These results significantly overcome the previous best performance reported by Zhang et al. [11], who achieved an accuracy of 0.924. These performance improvements highlight the effectiveness of Demucs and our deep learning models in the singing voice detection task and their potential applicability to a wide range of music genres and recording conditions. In conclusion, given the results we obtained, we note that the performance of the Transformer is in line with that of the LRCN, so we believe that the Transformer has the potential to perform very well in this task and, therefore, there is a need for a more in-depth study, also supported by testing on other data.

## 5. CONCLUSION

In this paper, we presented a comprehensive study on singing voice detection, focusing on the impact of music source separation and the potential of deep learning models for improving performance in this task. Our experiments were designed to investigate two main aspects: *(i)* the impact of a music source separation model, such as Demucs, and its zero-shot capabilities for the SVD task; *(ii)* the potential for deep learning to improve performance further. Our results demonstrated that incorporating music source separation with Demucs significantly improved the performance

Authorized licensed use limited to: Universita degli Studi di Roma La Sapienza. Downloaded on October 27,2024 at 22:49:47 UTC from IEEE Xplore. Restrictions apply.

of the LRCN and Transformer models compared to using the audio mixture directly. This finding established the importance of music source separation as a crucial preprocessing step for enhancing singing voice detection. Moreover, our investigation of Demucs' zero-shot capabilities revealed its potential for leveraging pre-trained models in tasks beyond their original scope, such as singing voice detection. Lastly, our deep learning models, LRCN and Transformer, outperformed the baseline RMS and the state-of-the-art methods on the Jamendo Corpus. Based on these findings, future research efforts should address the challenges of diverse dataset annotations, refine data preprocessing techniques, and explore alternative annotation methods to improve further the models' ability to generalize across various musical contexts. Further investigation into the potential of zero-shot and transfer learning for singing voice detection could lead to more accurate and robust models.

## 6. REFERENCES


[1] Hiromasa Fujihara and Masataka Goto, "Lyrics-to-audio alignment and its application," in *Multimodal Music Processing*, 2012.

[2] Tong Zhang, "Automatic singer identification," in *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, 2003, vol. 1, pp. I–33.

[3] Xulong Zhang, Jiale Qian, Yi Yu, Yifu Sun, and Wei Li, "Singer identification using deep timbre feature learning with KNN-Net," *CoRR*, vol. abs/2102.10236, 2021.

[4] Matt McVicar, Daniel P. W. Ellis, and Masataka Goto, "Leveraging repetition for improved automatic lyric transcription in popular music," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3117–3121, 2014.

[5] Kyungyun Lee, Keunwoo Choi, and Juhan Nam, "Revisiting singing voice detection: a quantitative review and the future outlook," *CoRR*, vol. abs/1806.01180, 2018.

[6] Jan Schlüter and Thomas Grill, "Exploring data augmentation for improved singing voice detection with neural networks," in *ISMIR*, 2015.

[7] Bernhard Lehner, Gerhard Widmer, and Sebastian Bock, "A low-latency, real-time-capable singing voice detection method with LSTM recurrent neural networks," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 21–25.

[8] Simon Leglaive, Romain Hennequin, and Roland Badeau, "Singing voice detection with deep recurrent neural networks," in *40th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Ed., Brisbane, Australia, Apr. 2015, pp. 121–125.

[9] Keunwoo Choi, György Fazekas, and Mark B. Sandler, "Automatic tagging using deep convolutional neural networks," *CoRR*, vol. abs/1606.00298, 2016.

[10] Keunwoo Choi, György Fazekas, Mark B. Sandler, and Kyunghyun Cho, "Transfer learning for music classification and regression tasks," *CoRR*, vol. abs/1703.09179, 2017.

[11] Xulong Zhang, Yi Yu, Yongwei Gao, Xi Chen, and Wei Li, "Research on singing voice detection based on a long-term recurrent convolutional network with vocal separation and temporal smoothing," *Electronics*, vol. 9, no. 9, 2020.

[12] Alexandre Défossez, "Hybrid spectrogram and waveform source separation," in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, 2021.

[13] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach, "Music source separation in the waveform domain," 2019.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020.

[15] Hyunji Hayley Park, Yogarshi Vyas, and Kashif Shah, "Efficient classification of long documents using transformers," 2022.

[16] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *CoRR*, vol. abs/1506.04214, 2015.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[18] Mathieu Ramona, Gaël Richard, and Bertrand David, "Vocal detection in music with support vector machines," *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1885–1888, 2008.

[19] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, 2016.


Authorized licensed use limited to: Universita degli Studi di Roma La Sapienza. Downloaded on October 27,2024 at 22:49:47 UTC from IEEE Xplore. Restrictions apply.