Cross-Version Singing Voice Detection in Opera Recordings: Challenges for Supervised Learning

Stylianos I. Mimilakis[1⋆], Christof Weiß[2⋆], Vlora Arifi-Müller[2], Jakob Abeßer[1], and Meinard Müller[2]

[1] Fraunhofer IDMT, Ilmenau, Germany
[2] International Audio Laboratories Erlangen, Germany
mis@idmt.fraunhofer.de

Abstract. In this paper, we approach the problem of detecting segments of singing voice activity in opera recordings. We consider three state-of-the-art methods for singing voice detection based on supervised deep learning. We train and test these models on a novel dataset comprising three annotated performances (versions) of Richard Wagner's opera "Die Walküre." The results of our cross-version experiments indicate that the models do not sufficiently generalize across versions even in the case when another version of the same musical work is available for training. By further analyzing the systems' predictions, we highlight certain correlations between prediction errors and the presence of specific singers, instrument families, and dynamic aspects of the performance. With these findings, our case study provides a first step towards tackling singing voice detection with deep learning in challenging scenarios such as Wagner's operas.

Keywords: opera · singing voice detection · supervised deep learning.

1 Introduction

The automatic identification of vocal segments in music recordings—known as singing voice detection (SVD)—is a central problem in music information retrieval (MIR) research [1]. In relevant literature, most SVD approaches are tailored to popular music [5–7,11,12,14,15]. However, Scholz et al. [18] showed that SVD quality considerably depends on the music genre, and that systems do often not generalize well across genres. Partly, this is due to the genre-specific usage of instruments and singing styles. A particular case is Western opera, where singing is often embedded in a rich orchestral accompaniment and instruments often imitate singing techniques such as vibrato [19]. Dittmar et al. [2] studied SVD within an opera scenario involving several versions of Weber's "Der Freischütz." Using carefully selected audio features and random forest classifiers, they showed that bootstrap training [11,20] helps to leverage the genre-dependency problem. They further demonstrated the benefit of a cross-version scenario by performing late fusion of the individual versions' results. We are not aware of any studies dealing with Wagner's operas, which constitute a challenging scenario due to their large and complex orchestration and highly expressive singing styles.

⋆ Equally contributing authors

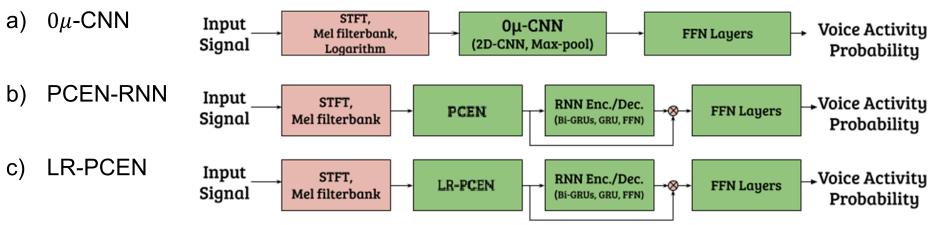

Fig. 1: The three examined DL models. Red modules denote non-trainable, predefined functions. Green modules denote parameterized functions subject to optimization. The cross symbol denotes the element-wise multiplication from [10].

As the system used in [2], early approaches to SVD [13,14] typically consist of two parts—the extraction of audio features and the supervised training of classifiers such as random forests. Recently, SVD based on deep learning (DL) has become popular [6–8, 16]. As the contributions of this paper, we apply several state-of-the-art (SOTA) approaches [10, 17, 21]—proposed for SVD in popular music—to the opera scenario. We systematically assess their efficacy on a limited dataset comprising three semi-automatically annotated versions of Richard Wagner's opera "Die Walküre" (first act). Our experiments demonstrate that the models do not sufficiently generalize across versions even when the training data contains other versions of the same musical work. Finally, we highlight specific challenges in Wagner's operas, pointing out interesting correlations between errors and the voices' registers as well as the activity of specific instruments.

2 Deep-Learning Methods

In this paper, we examine three SVD approaches based on supervised DL (Fig. 1).[3] Lee et al. [5] give an overview and a quantitative analysis of DL-based SVD systems. Our first model (Fig. 1a) is based on Convolutional Neural Networks (CNNs) followed by a classifier module. CNNs have been widely used for SVD [15–17]. To achieve sound-level-invariant SVD, Schlüter et al. [17] introduce zero-mean convolutions—an update rule that constrains the CNN kernels to have zero mean. For our first model (denoted as 0μ-CNN), we use the specific architecture presented in [17] with zero-mean update rules. As an alternative approach to sound-level-invariant SVD, Schlüter et al. [17] suggest per-channel energy normalization (PCEN) [21]. For our second model (Fig. 1b, denoted as PCEN-RNN), we consider this technique as front-end followed by recurrent layers and the classifier, realized by feed-forward layers. Recurrent neural networks (RNNs) have been used for SVD in [6], among others. As our third model, we examine a straightforward extension to PCEN involving a low-rank autoencoder (Fig. 1c, denoted as LR-PCEN). For both RNN-based models (PCEN-RNN and LR-PCEN), we include skip-filtering connections [10], which turns out to be useful for "pin-pointing" relevant parts of spectrograms [9].

[3] Due to limited space, we only provide an overview. For details, we refer to the relevant literature [6,10,17,21] and our source code: https://github.com/Js-Mim/wagner_vad.

For pre-processing, we partition the monaural recording into non-overlapping segments of length 3 seconds. Inspired by previous approaches [5,6,17], we compute a 250-band mel-spectrogram for each segment. As input to the 0μ-CNN model [16], we use the logarithm of the mel-spectrogram. For the PCEN-RNN model, we use the mel-spectrogram as input to the trainable PCEN front-end [21] followed by a bi-directional encoder with gated recurrent units (GRUs) and residual connections [10]. The decoder predicts a mask (of the original input size) for filtering the output of the PCEN. For the LR-PCEN, we replace the first-order recursion [21, Eq.(2)] with a low-rank (here: rank one) autoencoder that shares weights across mel-bands. The output of the autoencoder is used alongside residual connections with the input mel-spectrogram. We randomly initialize the parameters and jointly optimize these using stochastic gradient descent with the Adam [4] solver, binary cross-entropy loss, and a batch size of 64. We set the initial learning rate to 10^{-4} and the exponential decay rates for the first- and second-order moments to 0.9. We optimize for 100 iterations throughout the training data and adapt the learning rate depending on the validation error. We perform early stopping after 10 non-improving iterations.

3 Dataset

We evaluate the systems on a novel dataset comprising three versions of Wagner's opera "Die Walküre" (first act) conducted by Barenboim 1992 (`Bar`), Haitink 1988 (`Hai`), and Karajan 1966 (`Kar`), each comprising 1523 measures and roughly 70 minutes. Starting with the libretto's phrase segments, we manually annotate the phrase boundaries as given by the score (in musical measures/beats). To transfer the singing voice segments to the individual versions, we rely on manually generated measure annotations [22]. Using the measure positions as anchor points, we perform score-to-audio synchronization [3] for generating beat and tatum positions, which we use to transfer the segmentation from the *musical time* of the libretto to the *physical time* of the performances.

Since alignment errors and imprecise singer performance may lead to offsets between the transferred segment boundaries and the actual singing, we manually refined our semi-automatic annotations for the `Kar` recording, which we use as test version in our experiments. While the majority of phrase boundaries was adjusted, these adjustments were on a very subtle level, affecting only 1% of the frames. However, due to our annotation strategy, there is another issue. Since we start from the libretto with its phrase-level segments, the annotations do not account for smaller musical rests within phrases—an issue that is also common for SVD annotations in pop music. To estimate the impact of these gaps within phrases (labeled as "singing"), we compute the overlap between the phrase-level singing regions from the libretto (`Kar`) and note-level annotation derived from an aligned score. The two annotations match for only 94% of all frames. This suggests that in the opera scenario, phrase-level annotations may not be precise enough for high-quality SVD evaluated on a frame level. We therefore regard an accuracy of 94% as a kind of upper bound for our experiments.

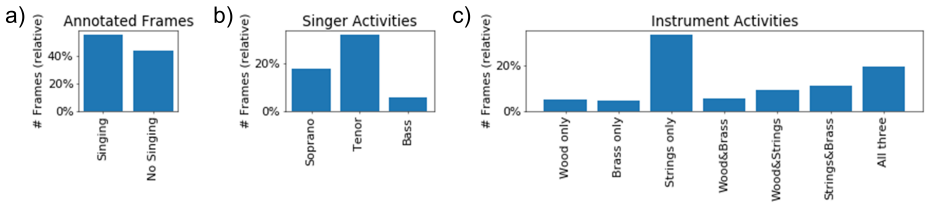

Fig. 2: Percentage of frames (`Kar` version) with (a) annotated singing voice, (b) activity of individual singers, (c) activity of instrument sections and their combination.

Table 1: Data splits used for the experiments.

Data Split	DS-1	DS-2	DS-3
Training	`Bar`, `Hai`, `Kar`	`Bar`, `Hai`	`Bar`
Validation	`Kar`	`Kar`	`Hai`
Test	`Kar`	`Kar`	`Kar`

Our dataset is quite balanced regarding singing frames (Fig. 2a). Among the three singers performing in the piece, the tenor dominates, followed by soprano and bass (Fig. 2b), while they never sing simultaneously. Regarding instrumentation, the string section alone plays most often, followed by all sections together, and other constellations (Fig. 2c). For systematically testing generalization to unseen versions, we create three data splits (Table 1). In DS-1, the test version (`Kar`) is available during training and validation. DS-2 only sees the test version at validation. DS-3 is the most realistic and restrictive split.

4 Experiments

For our results, Table 2 reports precision, recall, and F-measure with singing as the relevant class. Let us look at the results of the scenario DS-1 where the `Kar` version is used both for training and testing. While all models perform well, 0μ-CNN exhibits a higher tendency to overfit to the `Kar` version, going even beyond the upper bound of 94% discussed above. For the more realistic scenario DS-2, where `Kar` is only available for validation, the F-measure of 0μ-CNN decreases, being worse than the PCEN models that decrease only slightly. 0μ-CNN and PCEN-RNN tend towards more false negatives (precision $>$ recall). Looking at the scenario DS-3, where the `Kar` version is only used for testing, the models further decrease. All models have a tendency towards false negatives (most prominently 0μ-CNN). This points to detection problems in presence of the orchestra, which become particularly relevant when generalizing to unseen versions with different timbral characteristics and acoustic conditions.

We want to study such hypotheses in more detail for the realistic split DS-3. Regarding individual singers, the 0μ-CNN model obtains higher recall for the bass (76% of frames detected) than for tenor and soprano (each 70%). Interestingly, both PCEN models behave the opposite way, obtaining low recall ($<$50%)

Table 2: SVD results for all models (0μ-CNN, PCEN-RNN, LR-PCEN) and data splits.

Data Split	DS-1			DS-2			DS-3		
Models	0μ-CNN	PCEN-RNN	LR-PCEN	0μ-CNN	PCEN-RNN	LR-PCEN	0μ-CNN	PCEN-RNN	LR-PCEN
Precision	**0.98**	0.91	0.92	0.94	**0.98**	0.90	**0.95**	0.85	0.88
Recall	**0.96**	0.95	0.94	0.84	0.90	**0.91**	0.70	**0.77**	**0.77**
F-Measure	**0.97**	0.93	0.93	0.89	**0.90**	**0.90**	0.80	0.81	**0.82**

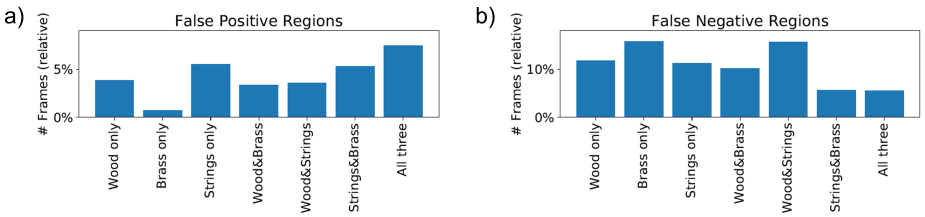

Fig. 3: (a) False positive and (b) false negative frames as detected by the 0μ-CNN model (Kar version). We plot the percentage of errors for regions with certain instrument sections or constellations playing, in relation to these regions' total duration.

for the bass and high recall (around 80%) for the others. We might conclude that the 0μ-CNN is less affected by the imbalance of singer in the training data. Since segments typically imply a certain length, we conduct a further experiment using median filtering for post-processing (not shown in the table). As observed in [2], F-measures improve by 2–4% for all models using a filter of length 2 sec.

Finally, we look at the correlation between errors and specific instrument activities (Fig. 3). Most instrument combinations do not show a preference for producing false positives or negatives, with two interesting exceptions. When only brass instruments are playing without singing, the 0μ-CNN practically never produces false positive predictions. However, when brass only occurs with singing, we observe an increase of false negatives. The highest frequency of false positives occurs for all three sections playing. Listening to false-positive regions, we often find expressive strings-only passages. False-negative regions often correspond to soft and gentle singing. Examining this in more detail, we observed a slight loudness-dependency for all models. As reported for pop music [17], singing frames are usually louder leading to more "loud" false positives and "soft" false negatives. This indicates that, despite the models' level invariance, confounding factors such as timbre or vibrato might affect SVD quality.

Our experiments and analyses only provide a first step towards understanding the challenges of SVD in complex opera recordings. From the results, we conclude that the systems do not sufficiently generalize across versions due to their different acoustic characteristis—even if the specific musical work is part of the training set. While all models are capable of fitting the data to a reasonable degree (given the reliability and precision of our annotations), generalization becomes problematic as soon as the test version is not seen during training or validation. Even if loudness dependencies are eliminated, our results suggest that more work has to be done to impose further invariances and constraints.

Acknowledgements

This work was supported by the German Research Foundation (AB 675/2-1, MU 2686/11-1, MU 2686/7-2). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institut für Integrierte Schaltungen IIS. We thank Cäcilia Marxer and all students who helped preparing the data and annotations.

References

1. Berenzweig, A.L., Ellis, D.P.W.: Locating singing voice segments within music signals. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). pp. 119–122 (2001).
2. Dittmar, C., Lehner, B., Prätzlich, T., Müller, M., Widmer, G.: Cross-version singing voice detection in classical opera recordings. In: Proc. Int. Soc. for Music Information Retrieval Conf. (ISMIR). pp. 618–624 (2015).
3. Ewert, S., Müller, M., Grosche, P.: High resolution audio synchronization using chroma onset features. In: Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). pp. 1869–1872 (2009).
4. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proc. Int. Conf. for Learning Representations (ICLR) (2015).
5. Lee, K., Choi, K., Nam, J.: Revisiting singing voice detection: A quantitative review and the future outlook. In: Proc. Int. Soc. for Music Information Retrieval Conf. (ISMIR). pp. 506–513 (2018).
6. Leglaive, S., Hennequin, R., Badeau, R.: Singing voice detection with deep recurrent neural networks. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). pp. 121–125 (2015).
7. Lehner, B., Schlüter, J., Widmer, G.: Online, loudness-invariant vocal detection in mixed music signals. IEEE/ACM Trans. on Audio, Speech & Language Processing **26**(8), 1369–1380 (2018)
8. Lehner, B., Widmer, G., Böck, S.: A low-latency, real-time-capable singing voice detection method with LSTM recurrent neural networks. In: Proc. European Signal Processing Conf. (EUSIPCO). pp. 21–25 (2015).
9. Mimilakis, S.I., Drossos, K., Cano, E., Schuller, G.: Examining the mapping functions of denoising autoencoders in music source separation. CoRR **abs/1904.06157** (2017), https://arxiv.org/abs/1904.06157
10. Mimilakis, S.I., Drossos, K., Virtanen, T., Schuller, G.: A recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation. In: Proc. IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6 (2017).
11. Nwe, T.L., Wang, Y.: Automatic detection of vocal segments in popular songs. In: Proc. Int. Soc. for Music Information Retrieval Conf. (ISMIR). pp. 138–144 (2004).
12. Ramona, M., Peeters, G.: Audioprint: An efficient audio fingerprint system based on a novel cost-less synchronization scheme. In: Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). pp. 818–822 (2013).
13. Ramona, M., Richard, G., David, B.: Vocal detection in music with support vector machines. In: Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). pp. 1885–1888 (2008).
14. Regnier, L., Peeters, G.: Singing voice detection in music tracks using direct voice vibrato detection. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). pp. 1685–1688 (2009).
15. Schlüter, J.: Learning to pinpoint singing voice from weakly labeled examples. In: Proc. Int. Soc. for Music Information Retrieval Conf. (ISMIR). pp. 44–50 (2016).
16. Schlüter, J., Grill, T.: Exploring data augmentation for improved singing voice detection with neural networks. In: Proc. Int. Soc. for Music Information Retrieval Conf. pp. 121–126 (2015).
17. Schlüter, J., Lehner, B.: Zero-mean convolutions for level-invariant singing voice detection. In: Proc. Int. Soc. for Music Information Retrieval Conf. (ISMIR). pp. 321–326 (2018).
18. Scholz, F., Vatolkin, I., Rudolph, G.: Singing voice detection across different music genres. In: Proc. AES Int. Conf. on Semantic Audio. pp. 140–147 (2017).
19. Seashore, C.E.: The natural history of the vibrato. Proc. National Academy of Sciences **17**(12), 623–626 (1931)
20. Tzanetakis, G.: Song-specific bootstrapping of singing voice structure. In: Proc. IEEE Int. Conf. on Multimedia and Expo (ICME). vol. 3, pp. 2027–2030 (2004).
21. Wang, Y., Getreuer, P., Hughes, T., Lyon, R.F., Saurous, R.A.: Trainable frontend for robust and far-field keyword spotting. In: Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). pp. 5670–5674 (2017).
22. Weiß, C., Arifi-Müller, V., Prätzlich, T., Kleinertz, R., Müller, M.: Analyzing measure annotations for Western classical music recordings. In: Proc. Int. Conf. on Music Information Retrieval (ISMIR). pp. 517–523 (2016).