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
mis@idmt.fraunhofer.de
[2] International Audio Laboratories Erlangen, Erlangen, Germany

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 that 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 [6–8,12,13,15,16]. However, Scholz et al. [19] 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 [20]. 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 [12,22] helps to leverage the genre-dependency problem.

S. I. Mimilakis and C. Weiß—Equally contributing authors

P. Cellier and K. Driessens (Eds.): ECML PKDD 2019 Workshops, CCIS 1168, pp. 429–436, 2020.
https://doi.org/10.1007/978-3-030-43887-6_35

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 SVD for Wagner's operas, which constitute a challenging scenario due to their large and complex orchestration and highly expressive singing styles.

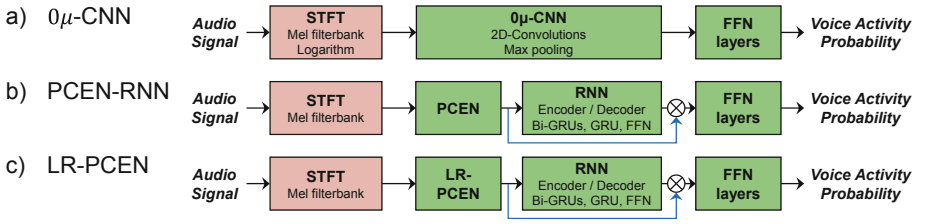

Fig. 1. The three examined DL models. Red modules denote non-trainable, predefined functions. Green modules denote parameterized functions subject to optimization. The cross symbol ⊗ denotes the element-wise multiplication introduced in [11]. (Color figure online)

As the system used in [2], early approaches to SVD [14,15] 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 [7–9,17]. As for the contributions of this paper, we apply several state-of-the-art (SOTA) approaches [11,18,23]—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).[1] Lee et al. [6] give an overview and a quantitative analysis of DL-based SVD systems. Our first model (Fig. 1a) is based on a convolutional neural network (CNN) followed by a classifier module. CNNs have been widely used for SVD [16–18]. To achieve sound-level-invariant SVD, Schlüter et al. [18] introduce zero-mean convolutions—an update rule that constrains the CNN kernels to have zero mean. We use this zero-mean update rule within the specific architecture presented in [18] for our first model (denoted as 0μ-CNN). As an alternative

[1] Due to limited space, we only provide an overview of the models. For details, we refer to the relevant literature [7,11,18,23] and our source code: https://github.com/Js-Mim/wagner_vad.

approach to sound-level-invariant SVD, Schlüter et al. [18] suggest per-channel energy normalization (PCEN) [23]. 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 network (FFN) layers. Recurrent neural networks (RNNs) have been used for SVD in [7], 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 [11], which turns out to be useful for "pin-pointing" relevant parts of spectrograms [10].

For pre-processing, we partition the monaural recording into non-overlapping segments of length 3 s. Inspired by previous approaches [6,7,18], we compute a 250-band mel-spectrogram for each segment. As input to the 0μ-CNN model [17], 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 [23] followed by a bi-directional encoder with gated recurrent units (GRUs) and residual connections [11]. 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 [23, 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 binary cross-entropy loss and the Adam [5] solver setting the initial learning rate to 10^{-4} and the exponential decay rates for the first- and second-order moments to 0.9. We optimize over the training data for 100 iterations and adapt the learning rate depending on the validation error. Moreover, 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 min of music. 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 [24]. 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. Almost every phrase boundaries was adjusted, thus affecting rouhgly 4% of all frames in total. Due to our annotation strategy, there might be another issue. Since we start from the libretto with its

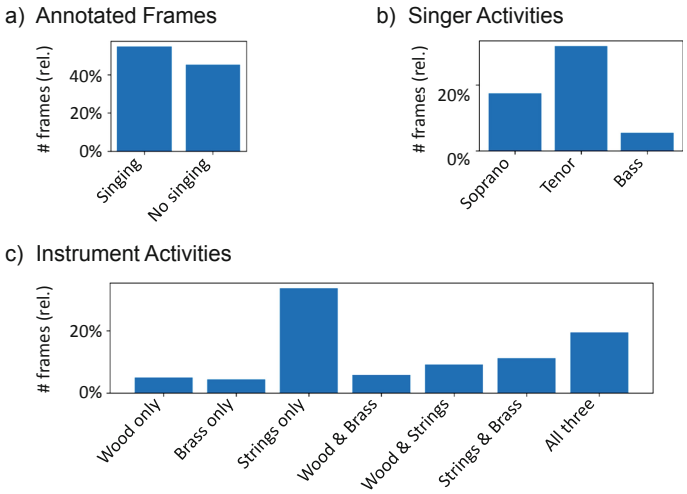

Fig. 2. Percentage of frames (`Kar` version) with (a) annotated singing voice, (b) activity of individual singers, (c) activity of instrument sections and their combination.

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`

phrase-level segments, the annotations do not account for smaller musical rests within textual phrases—an issue that is also common for SVD annotations in popular 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 as well as automatic alignment strategies may not be precise enough for high-quality SVD evaluated on the frame level. We therefore regard an accuracy or F-measure of 94% as a kind of upper bound for our experiments.

In our dataset, singing and non-singing frames are quite balanced (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. All models perform well here and almost reach the upper bound of 94% discussed above. For the more realistic scenario DS-2, where the test version (`Kar`) is only available for validation, the F-measures of all models decrease. Furthermore, the models tend towards more false negatives (precision > recall). Both effects are particularly prominent for 0μ-CNN. In the scenario DS-3, where the `Kar` version is only used for testing, the results further deteriorate. Again, all models show a clear 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.

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	**0.95**	0.93	0.94	**0.96**	0.91	0.92	**0.97**	0.87	0.90
Recall	0.91	**0.92**	0.90	0.81	**0.88**	**0.88**	0.69	**0.76**	0.74
F-Measure	**0.93**	**0.93**	0.92	0.88	0.89	**0.90**	0.80	0.81	**0.82**

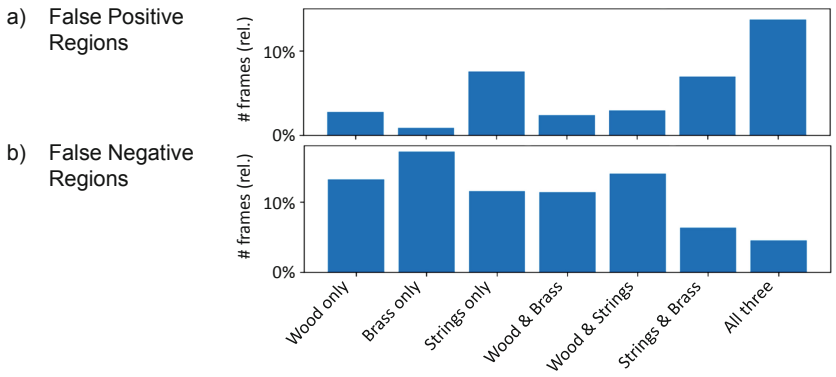

Fig. 3. (a) False positive and (b) false negative frames as detected by the 0μ-CNN model (`Kar` version). We plot the percentage of errors for regions with certain instrument sections or constellations playing, in relation to these regions' total duration.

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 (74% of frames detected) than for tenor and soprano (each 68%). Interestingly, both PCEN models behave the opposite way, obtaining low recall (<50%)

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

Finally, we want to investigate correlations between errors and specific instrument activities for the LR-PCEN model's results (Fig. 3). For most instrument combinations, we cannot observe any strong preference for producing false positives or negatives, with two interesting exceptions. When only brass instruments are playing without singing, the LR-PCEN practically never produces false positive predictions. In contrast, when brass only occurs together with singing, we observe a strong increase of false negatives. The highest frequency of false positives occurs for tutti passages (all three sections playing). When listening to false-positive regions, we often find expressive strings-only passages. In contrast, 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 popular music [18], 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 characteristics—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. A nice example is given in [21] where the generalization performance of the models is optimized. Furthermore, considering techniques such as data augmentation or unsupervised domain adaptation [4] might be useful to achieve robust SVD systems for the opera scenario.

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: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 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: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Málaga, Spain, pp. 618–624 (2015)
3. Ewert, S., Müller, M., Grosche, P.: High resolution audio synchronization using chroma onset features. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Taipei, Taiwan, pp. 1869–1872 (2009)
4. Gharib, S., Drossos, K., Çakir, E., Serdyuk, D., Virtanen, T.: Unsupervised adversarial domain adaptation for acoustic scene classification. Computing Research Repository (CoRR) abs/1808.05777 (2018)
5. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the International Conference for Learning Representations (ICLR), San Diego, California, USA (2015)
6. Lee, K., Choi, K., Nam, J.: Revisiting singing voice detection: A quantitative review and the future outlook. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Paris, France, pp. 506–513 (2018)
7. Leglaive, S., Hennequin, R., Badeau, R.: Singing voice detection with deep recurrent neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Brisbane, Australia, pp. 121–125 (2015)
8. Lehner, B., Schlüter, J., Widmer, G.: Online, loudness-invariant vocal detection in mixed music signals. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(8), 1369–1380 (2018). https://doi.org/10.1109/TASLP.2018.2825108
9. Lehner, B., Widmer, G., Böck, S.: A low-latency, real-time-capable singing voice detection method with LSTM recurrent neural networks. In: Proceedings of the European Signal Processing Conference (EUSIPCO), Nice, France, pp. 21–25 (2015)
10. Mimilakis, S.I., Drossos, K., Cano, E., Schuller, G.: Examining the mapping functions of denoising autoencoders in singing voice separation. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 266–278 (2020)
11. Mimilakis, S.I., Drossos, K., Virtanen, T., Schuller, G.: A recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation. In: Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Tokyo, Japan, pp. 1–6 (2017)
12. Nwe, T.L., Wang, Y.: Automatic detection of vocal segments in popular songs. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Barcelona, Spain, pp. 138–144 (2004)
13. Ramona, M., Peeters, G.: Audioprint: An efficient audio fingerprint system based on a novel cost-less synchronization scheme. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada, May 2013, pp. 818–822. https://doi.org/10.1109/ICASSP.2013.6637762
14. Ramona, M., Richard, G., David, B.: Vocal detection in music with support vector machines. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, Nevada, USA, pp. 1885–1888 (2008)
15. Regnier, L., Peeters, G.: Singing voice detection in music tracks using direct voice vibrato detection. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Taipei, Taiwan, pp. 1685–1688 (2009)

16. Schlüter, J.: Learning to pinpoint singing voice from weakly labeled examples. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), New York City, USA, pp. 44–50 (2016)
17. Schlüter, J., Grill, T.: Exploring data augmentation for improved singing voice detection with neural networks. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Málaga, Spain, pp. 121–126 (2015)
18. Schlüter, J., Lehner, B.: Zero-mean convolutions for level-invariant singing voice detection. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Paris, France, pp. 321–326 (2018)
19. Scholz, F., Vatolkin, I., Rudolph, G.: Singing voice detection across different music genres. In: Proceedings of the AES International Conference on Semantic Audio, Erlangen, Germany, pp. 140–147 (2017)
20. Seashore, C.E.: The natural history of the vibrato. Proc. Nat. Acad. Sci. USA **17**(12), 623–626 (1931)
21. Tolstikhin, I., Bousquet, O., Gelly, S., Schoelkopf, B.: Wasserstein auto-encoders. In: Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, Canada, pp. 1–16 (2018)
22. Tzanetakis, G.: Song-specific bootstrapping of singing voice structure. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, vol. 3, pp. 2027–2030 (2004)
23. Wang, Y., Getreuer, P., Hughes, T., Lyon, R.F., Saurous, R.A.: Trainable frontend for robust and far-field keyword spotting. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), New Orleans, USA, pp. 5670–5674 (2017)
24. Weiß, C., Arifi-Müller, V., Prätzlich, T., Kleinertz, R., Müller, M.: Analyzing measure annotations for Western classical music recordings. In: Proceedings of the International Conference on Music Information Retrieval (ISMIR), New York, USA, pp. 517–523 (2016)