# Towards Cross-Version Singing Voice Detection

Christian Dittmar, Thomas Prätzlich, Meinard Müller
*International Audio Laboratories Erlangen*
*Email: {christian.dittmar, thomas.praetzlich, meinard.mueller}@audiolabs-erlangen.de*

## Abstract

In the field of Music Information Retrieval (MIR), the automated detection of the singing voice within a given music recording constitutes a challenging and important research problem. The goal of this task is to find those segments within a given recording where one or several singers are active. In this study, we investigate the performance of state-of-the-art approaches by considering various music scenarios. First, we validate our singing voice detection system, which incorporates well-known techniques from audio signal processing and machine learning, against a public benchmark. Second, we consider a controlled yet instructive scenario using multiple versions (interpretations by different musicians) of the 24 songs of the cycle "Winterreise" by Franz Schubert. Within this cross-version scenario, which comprises various singers and pianists as well as different recording conditions, we systematically address the following research questions: Is bootstrapping a viable approach for stabilizing the singing voice detection in difficult cases? Can the results be improved by a cross-version fusion approach? Answers to these questions constitute the basis for considering more complex scenarios such as detecting the singing voices in multitimbral orchestral settings including opera recordings.

## 1 State-of-the-Art

Singing voice detection aims to determine those regions within a music recording where a singing voice is active. Although this task seems to be simple for human listeners, automatic singing voice detection poses a difficult research problem. The challenge arises from the complex characteristics of singing voice as well as the diversity of accompanying instruments. In the Music Information Retrieval (MIR) literature, it is typically assumed that the singer performs the melody and dominates over the accompaniment being played in the background [1]. Singing voices that contribute to the accompaniment (e.g., a background choir) are usually not considered as target singing. Given these preconditions, automatic singing voice detection is often approached by frame-wise classification into singing voice vs. accompaniment. Almost all procedures suggested in the literature [1, 4, 6, 7, 8, 9, 10, 12, 13] employ machine learning for this classification problem.

### 1.1 Baseline System

Our baseline system for singing voice detection closely follows the state-of-the-art approach proposed by Lehner et al. in [6, 7]. We took this procedure as starting point for our investigations, since it is described in detail and allows for a re-implementation. We only provide a summary and refer to the original publications for details about the audio features devised by the authors. Most notably, a feature referred to as Fluctogram captures pitch fluctuating signal components without the need for predominant pitch tracking. Random Forests (RF) [2] are used as classification scheme. The RF classifier generates a frame-wise decision function which we interpret as indicator for the singing voice activity. As will be described in Section 2, post-processing of decision functions is the main approach in our cross-version strategies.
We validated our re-implemented system using a subset of the publicly available JAMENDO corpus. The exact split into training and test set is given in [12]. We fixed the following parameters: The hopsize between consecutive analysis frames is 200 ms (feature rate of 5 Hz) and the analysis window size is 800 ms. In the RF classifier, we use 128 individual decision trees, each trained with a randomly selected subset of 5 features from the originally 146-dimensional feature space. The resulting decision functions are smoothed by a median filter with a kernel width of 1.4 s. The decision function threshold is set to 0.5. Using the specified 16 test songs, we achieved an averaged Accuracy of 87.3 and an averaged F-measure of 0.87, which is on par with the results reported in [7]. In Section 2, we devise two post-processing strategies to improve the baseline performance in a cross-version scenario.

### 1.2 Related Work

The negative effect of accompaniment on singing voice classification performance has been investigated in [4]. In [14], the authors tried to circumvent some of these problems by separating the singing voice from the accompaniment prior to feature extraction. A comparable approach is described in [9] with very promising results. However, the proposed signal processing chain relies on predominant pitch tracking, which bears the potential of substantial error propagation to all subsequent feature extraction and classification steps. As indicated above, singing voice detection based on machine learning faces the problem of large acoustic variance within both singing voice as well as accompaniment. An ideal classifier should be trained with an extreme range of training data covering all possible combinations of singing voices and accompaniment music. As an alternative, usage of training data taken from the target recording itself was introduced as unsupervised [10] and user-assisted [13] bootstrap strategy. Post-processing of classifier decision functions was described in [8] in the sense of a noise filtering operation.

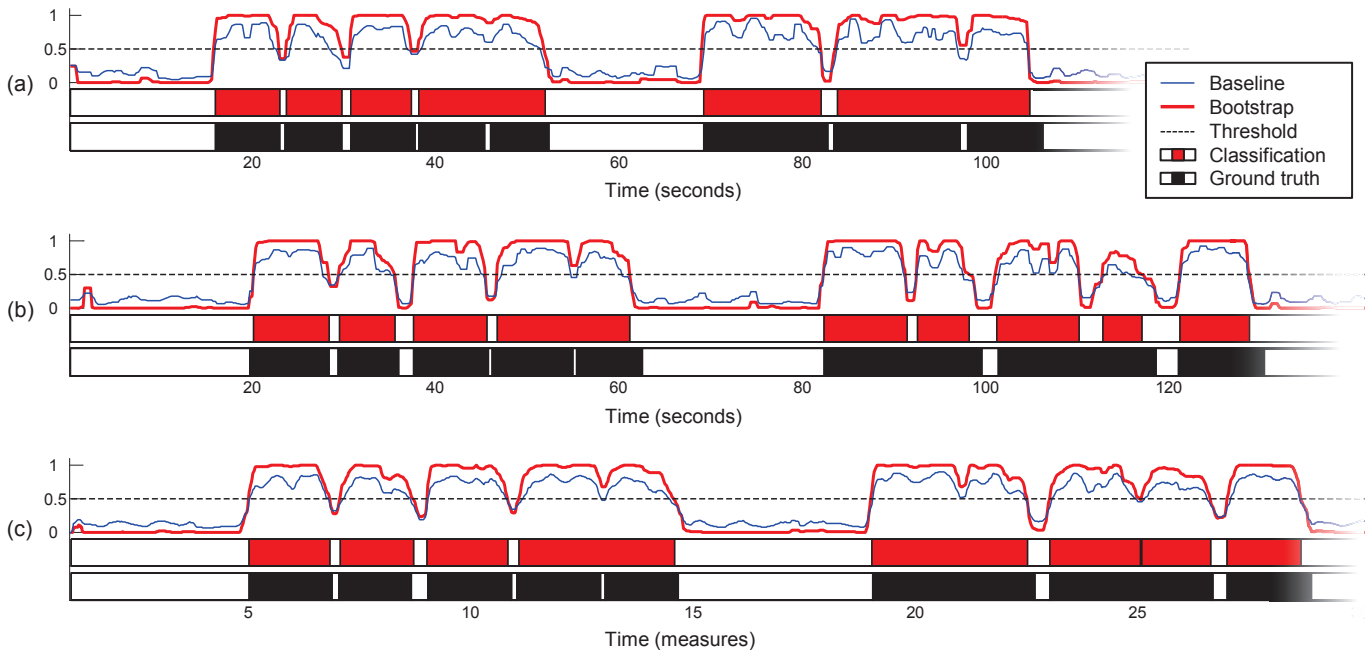


**Figure 1:** Illustration of the cross-version post-processing strategies as described in Section 2.2 and Section 2.3. The curves and annotations are based on an excerpt corresponding to the first 33 measures of the song "Wasserflut" (No. 06) from the "Winterreise" song cycle. For each case, the decision functions of the baseline (blue thin curve) and bootstrap (red bold curve) classifier are shown. The colored time-lines below the decision curves show the automatically detected singing voice activity (red segments) vs. the ground truth (black segments). **(a):** Performance by the singer Allen. **(b):** Performance by the singer Oliemans. Since this recording is performed at a slower tempo than (a), the 33 measures cover a longer time-span. **(c):** Cross-version results based on four performances (including Allen and Oliemans) after temporal alignment to a common, measure-based time axis and subsequent averaging across the individual decision functions. The improved congruence of the classification to the ground truth becomes especially evident in comparison to (b).

## 2 Case-Study in a Cross-Version Scenario

As our main contribution, we introduce two post-processing strategies that improve upon the singing voice detection capabilities of the baseline system in this section. We briefly discuss our test corpus which allows us to investigate into the peculiarities of a cross-version scenario. We will keep the explanations mostly on the conceptual level and refer to Figure 1 for an illustration of the main ideas.

### 2.1 Data

"Winterreise" (D. 911, published as Op. 89 in 1828) is a song cycle for voice and piano by the composer Franz Schubert from the Romantic era. The cycle, which is based on a setting of 24 poems (numbers) by Wilhelm Müller, was originally written for tenor voice but is frequently transposed to suit other voices as well. In our experiments, we use four different performances referenced by the respective vocalist (Allan, Oliemans, Quasthoff, Trekel). For evaluation purposes, we generated reference annotations of the singing voice activity in these pieces. This was achieved automatically by transferring singing voice activity information from a reference MIDI version of each number to the corresponding audio recording using music synchronization techniques [3]. This procedure also yields an alignment of all performances based on the measure grid of the MIDI version. In comparison to other data sets (e.g., JAMENDO), our test corpus consists of homogeneous musical material. All numbers in the four versions have instrumental piano accompaniment and male singers. Initial singing voice detection experiments yielded F-Measures around 0.95 in a leave-one-out cross-validation, i.e., taking one particular number as test item and training with the remaining songs. This upper bound can not be reached if the RF is trained with other training data (see Section 3).

### 2.2 Bootstrap Training

Inspired by the bootstrapping ideas in [10, 13], we propose to perform a second, specifically trained RF classification subsequent to the initial singing voice detection stage. The rationale is to create an automatically adapted classifier model that is trained with features taken from the current recording under analysis. In practice, no training assignment to ground truth classes is available. Thus, a central question is how to discern the extracted feature vectors into a training set for singing voice vs. a training set accompaniment? Our idea is to base this assignment on the shape of the decision function generated by the initial RF classifier. Looking at the course of this decision function, we see some extreme values in those frames where the observed features match closely to the classifier model reflecting the initial training data. However, the shape is far from being ideal, as many values reside in the middle of the range of values, where an assignment to either side is questionable. If we now select two subsets of the feature vectors, each corresponding to an upper and lower fraction (e.g., 20%) of the range of decision function values, we can use these to train a small RF classifier adapted to the feature space spanned by the recording under analysis. The new deci-

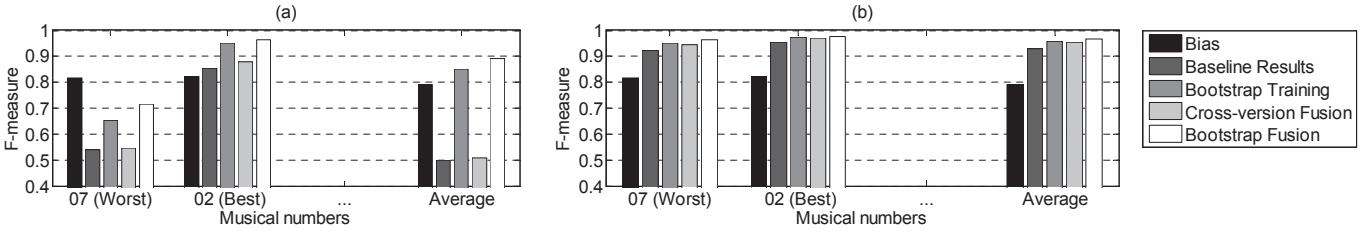


**Figure 2:** The average F-measures obtained in two different training scenarios and four different cross-version post-processing scenarios. **(a):** Results obtained by training the initial classifier with popular music recordings. **(b):** Results obtained by training the initial classifier with both popular music and classical opera recordings.

sion functions generated by classifying the current song with the adapted classifier tend to be more binary. Figure 1 illustrates this concept by overlaying the decision functions of the baseline system (blue thin curve) with the bootstrap decision functions (red bold curve). It can clearly be seen that the bootstrap decision functions behave less fuzzy.

### 2.3 Cross-Version Fusion

In [5], Konz et al. introduced the intuitive yet effective idea to exploit the availability of different recordings of the same piece of music for stabilizing automatic chord recognition results. We now pursue the same idea in order to perform a late fusion of decision functions obtained from singing voice detection in each individual version of the musical number in our test corpus. This is achieved by resampling the functions to a version-independent representation with a musical time axis given in measures (resp. sub-divisions thereof) instead of seconds. To this end, we us the measure annotations of the audio recordings (see Section 2.1). For the actual fusion, we use the most straightforward approach and just take the arithmetic mean of the decision values from the aligned decision functions. We expect to compensate for noise in the individual decision functions by averaging. Figure 1(c) illustrates the result of this operation by overlaying the fused decision function derived from baseline classification (blue thin curve) with the fused decision function derived from bootstrap training (red bold curves). It can be seen that the averaging leads to a more stable decision function. An automatic classification obtained by comparing the decision function against the decision threshold (dashed black line) shows improved agreement to the ground truth segmentation (black rectangles) in comparison to Figure 1(a) and 1(b).

## 3 Experiments

The diagrams in Figure 2 illustrate the benefits of applying bootstrap training and cross-version fusion. The bar plots show the average, frame-wise F-measures obtained under varying combinations of classifier training and post-processing strategies. The vertical axis is zoomed in to magnify the F-measure range between 0.4 and 1.0 for better visibility. **Bias** refers to the performance achievable by just assigning each frame of one test recording to the singing voice class. It can be seen that the resulting F-measures are already quite high, thus indicating that singing is the more frequent class in our test recordings. **Baseline Results** refers to the results obtained by the baseline singing voice detection system as described in Section 1.1. **Bootstrap Training** refers to the results obtained by a second classification run using an adapted RF classifier trained using the bootstrapping strategy as described in 2.2. **Cross-version Fusion** refers to the results of fusing the initial decision functions of all available versions of each test recording as described in Section 2.3. Finally, **Bootstrap Fusion** refers to the results obtained by combining both the bootstrap classification based on training with the individual test recordings and the cross-version fusion of the resulting decision functions. Besides showing the average results over all 24 songs, we also present a well-behaved example (Number 02) and an ill-behaved example (Number 07) which yield the best resp. worst F-measures in the bootstrap fusion scenario.

The results in Figure 2(a) were obtained by training the initial RF classifier with a combined data set comprising both the JAMENDO [12] and RWC [9] subsets that are annotated for singing voice. Both corpora are dominated by recordings of popular music. The resulting data set drastically differs from the music content in our test corpus, consisting of the "Winterreise" songs. This leads to substandard singing voice detection performance close to random guessing. Interestingly, even with such an unreliable initial estimate for the frames that likely contain singing voice, the strategy of bootstrap training leads to a substantial performance gain, surpassing the bias results. In contrast, cross-version fusion of the unreliable initial decision functions does not improve the result at all. The combination of both bootstrap training and cross-version fusion of the decision functions delivers the best results in this training scenario.

The results in Figure 2(b) are obtained when complementing the initial training data with recordings of classical opera. Specifically, we used all numbers from Karl Maria von Weber's Singspiel "Der Freischütz" [11] in a 1973 studio recording conducted by Carlos Kleiber. As can be seen from the F-measure of the baseline RF classifier, this additional training data gives a considerable performance boost. This is a bit surprising, since the instrumental parts of this opera are played by a symphony orchestra, whereas the instrumental parts in our test corpus are solely played by piano. However, the vibrato heavy singing style seems to be very similar in the Weber opera and the Schubert songs. The remaining

measures show that the proposed post-processing strategies seem to help again, this time to a lesser extent than in Figure 2(a). It should be noted that these results reach the upper bound of 0.95 F-measure that was obtained by leave-one-out training as described in Section 2.1. On first sight, bootstrap training could be recommended as standard post-processing in singing voice detection. Unfortunately, it has the important drawback that it may produce erroneous decision functions when no singing voice activity occurs at all throughout a recording. If these cases can not be ruled out, bootstrap training would deteriorate the results. Cross-version fusion will only be beneficial if there are no significant structural differences between the different versions and the temporal alignment is reliable enough not to introduce additional errors.

## 4 Conclusions and Future Work

In this paper, we presented two strategies to post-process automatic singing voice detection in a cross-version scenario. In our case-study involving multiple recorded versions of Franz Schubert's "Winterreise" song cycle, we showed that combining bootstrap training and cross-version fusion can lead to a substantial performance improvement. In principle, the presented strategies are applicable for singing voice detection in various music genres. However, only for classical music, it is likely to have multiple, sufficiently similar versions. Future work will be directed towards using these techniques as a preprocessing step to improve music segmentation of operas in the context of the Freischütz Digital project [11].

## 5 Acknowledgments

The authors would like to thank Bernhard Lehner for fruitful discussions and substantial support in validating the experimental results. This work has been supported by the BMBF project Freischütz Digital (Funding Code 01UG1239A to C). The International Audio Laboratories Erlangen (AudioLabs) is a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer IIS.

## References

[1] Berenzweig, A. L. & Ellis, D. P. W.: Locating Singing Voice Segments within Music Signals. In: Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA), 119–122. New Paltz, NY, USA (2001).

[2] Breiman, L.: Random forests. In: Machine learning (2001), **45**, 1: 5–32.

[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), 1869–1872. Taipei, Taiwan (2009).

[4] Gärtner, D. & Dittmar, C.: Vocal characteristics classification of audio segments: An investigation of the influence of accompaniment music on low-level features. In: Proceedings of the International Conference on Machine Learning and Applications (ICMLA). Miami, Florida, USA (2009).

[5] Konz, V., Müller, M. & Kleinertz, R.: A Cross-Version Chord Labelling Approach for Exploring Harmonic Structures—A Case Study on Beethoven's Appassionata. In: Journal of New Music Research (2013), 1–17.

[6] Lehner, B., Sonnleitner, R. & Widmer, G.: Towards Lightweight, Real-time-capable Singing Voice Detection. In: Proceedings of the International Conference on Music Information Retrieval (ISMIR). Porto, Portugal (2013).

[7] Lehner, B., Widmer, G. & Sonnleitner, R.: On the reduction of false positives in singing voice detection. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 7480–7484. Florence, Italy (2014).

[8] Lukashevich, H. & Dittmar, C.: Effective singing voice detection in popular music using ARMA filtering. In: Proceedings of the International Conference on Digital Audio Effects (DAFx). Bordeaux, France (2007).

[9] Mauch, M., Fujihara, H., Yoshii, K. & Goto, M.: Timbre and Melody Features for the Recognition of Vocal Activity and Instrumental Solos in Polyphonic Music. In: Proceedings of the International Conference on Music Information Retrieval (ISMIR), 233–238. Miami, Florida, USA (2011).

[10] Nwe, T. L. & Wang, Y.: Automatic Detection of Vocal Segments in Popular Songs. In: Proceedings of the International Conference on Music Information Retrieval (ISMIR). Barcelona, Spain (2004).

[11] Prätzlich, T. & Müller, M.: Frame-Level Audio Segmentation for Abridged Musical Works. In: Proceedings of the International Conference on Music Information Retrieval (ISMIR). Taipei, Taiwan (2014).

[12] Ramona, M., Richard, G. & David, B.: Vocal detection in music with support vector machines. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Las Vegas, Nevada, USA (2008).

[13] Tzanetakis, G.: Song-specific Bootstrapping of Singing Voice Structure. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), 2027–2030. Taipei, Taiwan (2004).

[14] Vembu, S. & Baumann, S.: Separation of vocals from polyphonic audio recordings. In: Proceedings of the International Conference on Music Information Retrieval (ISMIR). London, UK (2005).