HIERARCHICAL CLASSIFICATION OF SINGING ACTIVITY, GENDER, AND TYPE IN COMPLEX MUSIC RECORDINGS

Michael Krause and Meinard Müller

International Audio Laboratories Erlangen

ABSTRACT

Traditionally, work on singing voice detection has focused on identifying singing activity in music recordings. In this work, our aim is to extend this task towards simultaneously detecting the presence of singing voice as well as determining singer gender and voice type. We describe and compare four strategies for exploiting the hierarchical relationships between these levels. In particular, we introduce a novel loss term that promotes consistency across hierarchy levels. We evaluate the strategies on a dataset containing over 200 hours of complex opera recordings with various singers of different genders and voice types, with a particular focus on hierarchical consistency. Our experiments show that by adding our loss term, a joint classification strategy using a single neural network achieves slightly improved evaluation scores and significantly more consistent results.

Index Terms— hierarchical classification, singing voice detection, opera, music processing, music information retrieval

1. INTRODUCTION

Singing voice detection (SVD) has been a long standing task in the field of music information retrieval (MIR) [1]. Prior work has focused on increasing the accuracy for detecting singing activity in music recordings [2, 3, 4]. Aside from mere activity, however, singing voice can be classified with regard to a multitude of aspects related to singing styles and techniques. Western opera, for example, is performed by singers with certain voice types (e. g., baritone, tenor, soprano) who may sing individually or simultaneously, creating a complex sound mixture of singing and orchestral music. In this context, we propose to simultaneously detect singing activity, singer gender, and voice type in music recordings. A system for detecting gender and voice type may be useful for annotating recorded and live performances in order to enhance audience experience or to aid in navigating music collections. It may also be useful as a pre-processing step for tasks such as singer identification.

In our scenario, the classes under consideration form hierarchical relationships, as illustrated in Figure 1. As a main contribution of this paper, we describe four different strategies for classification of singing voice that incorporate hierarchical information in different ways. In particular, we compare these strategies with regard to consistency of their predictions across hierarchy levels and discuss a joint classification strategy requiring only a single neural network which uses a novel loss to promote consistency. The strategies are comprehensively evaluated using a large dataset of over 200 hours of audio recordings from Richard Wagner's cycle of operas *Der Ring des Nibelungen*.

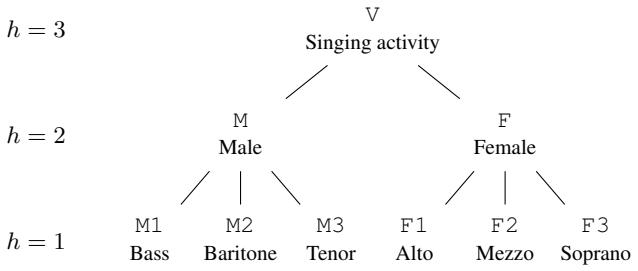

Fig. 1: Class hierarchy as considered in this paper (number of hierarchy levels $H = 3$).

We make the following contributions: first, we introduce a novel hierarchical classification problem by extending singing voice detection towards singer gender and voice types. Second, motivated by this problem, we formalize an appropriate hierarchical class model and provide evaluation and hierarchical consistency measures (Section 3). Third, we describe four strategies to approach this problem that utilize the hierarchical relationships between classes in different ways (Section 4). In particular, we propose a novel loss term that promotes consistent predictions. Finally, we evaluate these strategies in the context of a dataset containing over 200 hours of complex opera music (Section 5). We show that the joint strategy using our additional loss achieves strong and consistent results.

2. RELATED WORK

Historically, SVD has received a lot of attention from the MIR community and numerous approaches have been proposed over the years. Earlier systems have usually relied on signal processing and classical machine learning techniques [2, 5] while more recent approaches typically rely on deep-learning [6, 7]. We refer to [4] for a review. Few works have considered finer grained classes, such as the classification of singing gender [8]. For applications beyond music processing, some researchers have considered the automatic classification of singers according to gender or voice type [9, 10], but this is usually constrained to short, isolated excerpts of singing. Here, we want to detect and classify singing activity across entire music recordings.

Several works have investigated singing voice detection for opera [11, 12, 13], without considering finer grained classes. Other works have focused on identifying emotion [14] or melody [15] from opera recordings.

This work was supported by the German Research Foundation (DFG MU 2686/7-2). The authors are with the International Audio Laboratories Erlangen, a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS. The authors gratefully acknowledge the compute resources and support provided by the Erlangen Regional Computing Center (RRZE). Moreover, the authors thank Vlora Arifi-Müller, Christof Weiß, Cäcilia Marxer, and all student assistants involved in the preparation of data and annotations.

Hierarchical classification has been explored for tasks such as bird call classification [16], singing transcription [17] or general sound event detection [18, 19]. It has also been considered in the wider machine learning literature [20, 21]. Often, however, these works rule out simultaneous class activity or do not evaluate results with regard to hierarchical inconsistencies. We refer to [22] for a comprehensive overview of pre-deep-learning hierarchical classification approaches.

3. HIERARCHICAL CLASS MODEL

We begin by formalizing our class hierarchy and detection task. Let $\mathbf{C} = \{\texttt{V}, \texttt{M}, \texttt{F}, \texttt{M1}, \texttt{M2}, \texttt{M3}, \texttt{F1}, \texttt{F2}, \texttt{F3}\}$ be the set of all classes in our scenario. These classes are organized in a hierarchy, where H is the total number of hierarchy levels and $\mathbf{C}^h$ are the classes at hierarchy level $h \in [1:H]$. The lowest level $h = 1$ corresponds to the most specific classes, whereas the highest level $h = H$ corresponds to the most general one. We assume that $(\mathbf{C}^h)_{h\in[1:H]}$ forms a partition of $\mathbf{C}$. In our case, $H = 3$ and the hierarchy levels correspond to voice type ($h = 1$), singer gender ($h = 2$), and singing activity ($h = 3$), respectively. In particular, we have $\mathbf{C}^1 = \{\texttt{M1}, \texttt{M2}, \texttt{M3}, \texttt{F1}, \texttt{F2}, \texttt{F3}\}$, $\mathbf{C}^2 = \{\texttt{M}, \texttt{F}\}$, and $\mathbf{C}^3 = \{\texttt{V}\}$.

For a class $c \in \mathbf{C}$ we write $c\uparrow$ for the immediate parent of c in the class hierarchy (e. g., $\texttt{F2}\uparrow = \texttt{F}$). Additionally, we write $c\downarrow$ for the set of immediate children of c (e. g., $\texttt{M}\downarrow = \{\texttt{M1}, \texttt{M2}, \texttt{M3}\}$).

We formulate our singing detection task as a frame-wise, multi-label classification problem. Formally, let $\mathcal{I}$ be the set of items under consideration. In our case, the elements of $\mathcal{I}$ are audio frames. We describe our reference annotations as well as predictions made by some detection model as families $\left(\mathcal{I}_c\right)_{c\in\mathbf{C}}$ of subsets $\mathcal{I}_c \subseteq \mathcal{I}$. For $i \in \mathcal{I}_c$ we say that class c is active for frame i. Note that the sets $\mathcal{I}_c$ for different c need not be disjoint (i. e., multiple singers with different genders and voice types may be active for the same audio frame), and there may be items $i \in \mathcal{I}$ that are not contained in any set $\mathcal{I}_c$ (i. e., there may be audio frames without any singing). In this way, we account for our multi-label scenario.[1]

Generally, we would like the $\mathcal{I}_c$ to be in some sense consistent with the hierarchical structure of $\mathbf{C}$. For example, an item $i \in \mathcal{I}_c$ should also be an element of $\mathcal{I}_{c\uparrow}$. We will refer to this requirement as bottom-up consistency. Likewise, if $i \in \mathcal{I}_c$, then there should be some child $c' \in c\downarrow$ such that $i \in \mathcal{I}_{c'}$. We will call that top-down consistency. We now introduce three measures that capture the degree of bottom-up consistency ($\gamma_c^{\uparrow}$), top-down consistency ($\gamma_c^{\downarrow}$), or both (γ_c) for the set $\mathcal{I}_c$.

First, for a subset $\mathbf{C}' \subseteq \mathbf{C}$, we introduce the notation $\mathcal{I}_{\mathbf{C}'} = \bigcup_{c\in\mathbf{C}'} \mathcal{I}_c$. Now, for any $h > 1$ and $c \in \mathbf{C}^h$, we define the following consistency measures with values in the range $[0, 1]$:

$$\gamma_c = \frac{|\mathcal{I}_c \cap \mathcal{I}_{c\downarrow}|}{|\mathcal{I}_c \cup \mathcal{I}_{c\downarrow}|}, \gamma_c^{\downarrow} = \frac{|\mathcal{I}_c \cap \mathcal{I}_{c\downarrow}|}{|\mathcal{I}_c|}, \gamma_c^{\uparrow} = \frac{|\mathcal{I}_c \cap \mathcal{I}_{c\downarrow}|}{|\mathcal{I}_{c\downarrow}|}. \quad (1)$$

Intuitively, these measures capture the amount of agreement between $\mathcal{I}_c$ and $\mathcal{I}_{c\downarrow}$. If all $i \in \mathcal{I}_c$ are also contained in $\mathcal{I}_{c'}$ for some $c' \in c\downarrow$, then $\gamma_c^{\downarrow} = 1$. If for all $i \in \mathcal{I}_{c\downarrow}$ it holds that $i \in \mathcal{I}_c$, then $\gamma_c^{\uparrow} = 1$. Finally, $\gamma_c = 1$ if and only if $\mathcal{I}_c = \mathcal{I}_{c\downarrow}$. γ_c is also called intersection-over-union or Jaccard index.

[1] In the terminology adopted by [22], we are dealing with a hierarchically multi-label problem on a tree with full depth labeling.

4. HIERARCHICAL SINGING DETECTION

In this section, we introduce various strategies for hierarchical singing detection. For now, we assume a given classification model $\mathcal{M}$ that can be trained on subsets of items $\mathcal{I}' \subseteq \mathcal{I}$ to yield predictions for subsets of classes $\mathbf{C}' \subseteq \mathbf{C}$. After training, we can use such a model to get probabilities p_c per class $c \in \mathbf{C}'$ for unseen inputs $i \in \mathcal{I} \setminus \mathcal{I}'$. For the classification, we threshold these probabilities at 0.5 to obtain $\mathcal{I}_c^{\text{Est}}$, the set of items that have been predicted for a certain class $c \in \mathbf{C}$. Details on $\mathcal{M}$ are given in Section 5. Next, we describe our detection strategies.[2]

Strategy A: Independent Decisions. In this first strategy, we train one independent model for each hierarchy level $h \in [1 : H]$. Thus, predictions are made separately at each hierarchy level and no consistency between predictions is enforced.

Strategy B: Bottom-Up Aggregation. Here, we train only one model for hierarchy level $h = 1$. We then obtain predictions for higher levels by iteratively aggregating results from lower levels, setting $\mathcal{I}_c^{\text{Est}} = \mathcal{I}_{c\downarrow}^{\text{Est}}$ first for all $c \in \mathbf{C}^2$ and then for all $c \in \mathbf{C}^3$. By design, this ensures that $\gamma_c = \gamma_c^{\downarrow} = \gamma_c^{\uparrow} = 1$ for all c. However, classification errors made at lower levels are propagated upwards.

Strategy C: Top-Down Divide-and-Conquer. Here, we begin with a model for hierarchy level $h = H$ and divide items into subsets $\mathcal{I}_c$ for $c \in \mathbf{C}^h$ according to the classification results. We then iterate that process, proceeding with separate classification models for the subsets $\mathcal{I}_c$, where for $\mathcal{I}_c$ one considers the classes in $c\downarrow \subseteq \mathbf{C}^{h-1}$. In other words, only one model is trained on the entire dataset and operates at the highest hierarchy level. Subsequent models differentiate between more specific classes and are trained and evaluated only on frames for which the parent class is active. By design, this ensures that $\gamma_c^{\uparrow} = 1$ for all c. However, since each model considers a multi-label classification problem, there may be frames for which some class is predicted at a higher hierarchy level, but subsequent models predict none of its child classes as active, leading to $\gamma_c^{\downarrow} < 1$. Furthermore, errors made at higher levels are propagated downwards and many separate models need to be trained.

Strategy D: Joint Classification. Finally, we consider a strategy with a single model for all classes $\mathbf{C}$. To this end, we utilize a multi-task model that performs singing activity detection, gender recognition, and voice type classification at the same time. This model makes predictions jointly at all hierarchy levels (as opposed to the independent decisions in Strategy A), but may violate consistency properties. To address this, we introduce two losses that encourage consistent predictions in a soft way. Intuitively, in order to promote bottom-up consistency and improve $\gamma_c^{\uparrow}$, a loss should encourage the predictions for a parent class c to be at least as high as the prediction for any of its child classes c'. Writing p_c for the probability output by the model for class c, this is realized by the loss

$$\mathcal{L}_{\uparrow} = \frac{1}{|\mathbf{C} \setminus \mathbf{C}^H|} \sum_{h=2}^{H} \sum_{c\in\mathbf{C}^h} \sum_{c'\in c\downarrow} \max\{0, p_{c'} - p_c\}^2, \quad (2)$$

which contains penalties for every $p_{c'} > p_c$. This loss formulation has been proposed in [20]. The normalization factor ensures that the loss is in the range $[0, 1]$.

Similarly, to promote top-down consistency and improve $\gamma_c^{\downarrow}$, a loss should penalize predictions for a parent class c that are above the

[2] In the terminology adopted by [22], Strategy D would be considered a "global" approach, whereas B corresponds to a "flat" approach. Strategies A and C are both "local", with A consisting of local classifiers per layer (LCL) and Strategy C employing local classifiers per parent node (LCPN).

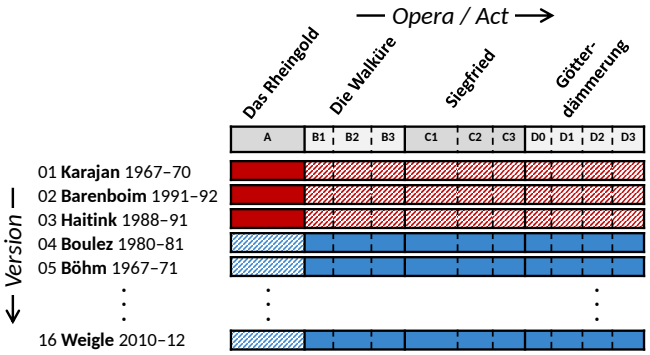

Fig. 2: Structure of Richard Wagner's *Ring* cycle and overview of 16 recorded versions, see [23] for details. Versions 04–16 (blue) are used for training, while versions 01–03 (red) are used for testing. The solid cells indicate one run of cross-validation, where a certain act is removed from training and used for testing.

highest probability predicted for a child class c'. Thus, we propose a novel loss by defining

$$\mathcal{L}_{\downarrow} = \frac{1}{|\mathbf{C} \setminus \mathbf{C}^1|} \sum_{h=2}^{H} \sum_{c \in \mathbf{C}^h} \max\{0, p_c - \max_{c' \in c\downarrow} p_{c'}\}^2. \quad (3)$$

The final loss for the model is obtained as

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \alpha\mathcal{L}_{\downarrow} + \beta\mathcal{L}_{\uparrow},$$

where $\alpha, \beta \in \mathbb{R}$ are weights associated with each consistency loss and $\mathcal{L}_{\text{BCE}}$ is the standard binary cross-entropy loss applied at each output of the network. We will refer to the variants of Strategy D without or including additional consistency losses as strategies $\text{D}^{0,0}$ and $\text{D}^{\alpha,\beta}$, respectively.

5. EXPERIMENTS

In this section, we introduce the dataset and evaluation measures used for comparing our hierarchical detection strategies and describe the specific classification model used. Finally, we discuss results.

5.1. Dataset

To compare the effectiveness of the strategies outlined in Section 4, we make use of a dataset containing a multitude of singers with different genders and voice types, as well as complex orchestral accompaniment. Specifically, we consider 16 recordings (versions) of *Der Ring des Nibelungen* by Richard Wagner, a cycle of four operas that each last between two and five hours. In total, our dataset consists of over 200 hours of opera music. An overview of the dataset and the operas' structure is given in Figure 2. Reference annotations have been obtained with a semi-automatic procedure using score-to-audio synchronization, see [11, 24] for details.

We reserve three versions for testing and take the rest for training, in line with [11]. The reference annotations for the training and test sets can be represented as families $\left(\mathcal{I}_c^{\text{Ref}}\right)_{c \in \mathbf{C}}$, where $\mathcal{I}_c^{\text{Ref}} \subseteq \mathcal{I}$ contains items labeled as class c. These families naturally fulfill all consistency properties, i. e., $\gamma_c = \gamma_c^{\downarrow} = \gamma_c^{\uparrow} = 1$ for all c.

Writing $\delta_c = |\mathcal{I}_c|/|\mathcal{I}| \in [0, 1]$ for the fraction of items where class c is active, we make several observations on the distribution of classes in $\mathcal{I}^{\text{Ref}}$ for our test set: around half of all audio frames in our dataset contain singing ($\delta_{\text{V}} = 0.55$) and male voices are more common ($\delta_{\text{M}} = 0.36$) than female voices ($\delta_{\text{F}} = 0.196$). Some voice types occur more often ($\delta_{\text{M3}} = 0.175$) than others ($\delta_{\text{F1}} = 0.033$). Around 2% of frames contain activity for more than one voice type.

In addition to splitting our dataset across versions, we perform cross-validation over opera acts in the test set. Figure 2 illustrates one run of cross-validation (solid cells). As a consequence, our approaches need to generalize both to unseen versions (containing different singers and acoustic conditions) and unseen acts (i. e., different musical compositions).[3]

5.2. Evaluation Measures

As described in Section 3, we formulate our detection task as frame-wise classification on full recordings. The detection strategies described in Section 4 yield predictions for all classes at all hierarchy levels for each frame. The performance of the detection strategies can be evaluated using standard measures from information retrieval such as class-wise[4] precision, recall, and F-measure. Formally:

$$P_c = \frac{|\mathcal{I}_c^{\text{Ref}} \cap \mathcal{I}_c^{\text{Est}}|}{|\mathcal{I}_c^{\text{Est}}|}, R_c = \frac{|\mathcal{I}_c^{\text{Ref}} \cap \mathcal{I}_c^{\text{Est}}|}{|\mathcal{I}_c^{\text{Ref}}|}, F_c = \frac{2 \cdot P_c \cdot R_c}{P_c + R_c}. \quad (4)$$

5.3. Model

The detection strategies described in Section 4 depend on a classification model $\mathcal{M}$ that can be trained to classify audio frames. In our experiments, we use a state-of-the-art model for singing activity detection, introduced in [6, 3]. This system is a VGGNet-inspired convolutional neural network applied to log-mel spectrograms patches (of length 1.64 s) with a single sigmoid output. The network is trained to predict singing activity for the center frame of the input patch. For details on the architecture and the specific reimplementation used, we refer to [6, 11]. We only slightly modify this system by increasing the number of sigmoid outputs at the final layer depending on the number of classes we wish to predict (for example, the network used in Strategy B has $|\mathbf{C}^1| = 6$ outputs, while the network for Strategy $\text{D}^{\alpha,\beta}$ has $|\mathbf{C}| = 9$). For Strategy $\text{D}^{\alpha,\beta}$, we use the additional losses described in Section 4. In our experiments we set $\alpha = \beta = 0.1$. We determined these values empirically such that all terms in $\mathcal{L}$ have a similar magnitude. As some classes in our dataset occur less frequently than others, we resample the training set (with replacement) such that each class occurs the same number of times. For post-processing, we follow [11] by applying a median filter of length 1.4 seconds and then downsampling the predictions to a frame rate of 5Hz.

5.4. Results

Figure 3 shows results for the four detection strategies on our test set. The results for Strategy A demonstrate that, using models trained independently per hierarchy level, one can achieve high evaluation scores for the upper two levels ($F_{\text{V}} = 0.94, F_{\text{M}} = F_{\text{F}} = 0.93$) and lower results for the finest level (e. g., $F_{\text{M1}} = 0.40$ or $F_{\text{M3}} = 0.77$). Bottom-up consistency is high ($\gamma_c^{\uparrow} = 0.99$ for all c), even though the strategy does not enforce this. Therefore, models at higher levels can

[3] In [11], this is referred to as a "neither split".

[4] Many works on hierarchical classification use evaluation measures that aggregate across classes (see [25] for an overview). In contrast, we use class-wise measures to analyze the behavior of our systems with regard to the specific musical challenges associated with different genders and voice-types.

identify all frames that are predicted as active at lower levels. The opposite does not hold, as evident in the low top-down consistency values. For example, $\gamma_{\text{F}}^{\downarrow} = 0.74$ implies that some frames for which female singing is being predicted at level $h = 2$ were misclassified as nonactive or as a male voice type at level $h = 1$.

Strategy B involves the same model for level $h = 1$ as Strategy A, but results for higher levels are obtained through bottom-up aggregation. As such, no inconsistencies arise for Strategy B, but evaluation results are much worse on higher levels (e. g., $F_{\text{V}} = 0.85$ as opposed to $F_{\text{V}} = 0.94$ for Strategy A) owing to frames incorrectly classified as non-singing (see e. g. $R_{\text{V}} = 0.76$).

For Strategy C, we observe the same F-measures as for Strategy A at levels $h = 3$ (by design) and $h = 2$. In addition, predictions are mostly top-down consistent, even though the strategy does not enforce this. On the finest level, there are large improvements for some classes (e. g., $F_{\text{F3}} = 0.74$ as opposed to $F_{\text{F3}} = 0.66$ for strategies A and B) but also degradations for some results (e. g., $F_{\text{F1}} = 0.47$ as opposed to $F_{\text{F1}} = 0.52$ for strategies A and B).

Employing the joint classification strategy without additional loss terms $\text{D}^{0,0}$, we obtain less accurate predictions for most of the classes on the lowest level (e. g., $F_{\text{F3}} = 0.58$ as opposed to $F_{\text{F3}} = 0.66$ for strategies A and B) and also a large amount of top-down inconsistencies (e. g., $\gamma_{\text{F}}^{\downarrow} = 0.49$). We are able to improve this using the additional loss terms of Strategy $\text{D}^{0.1,0.1}$. In particular, our proposed loss term $\mathcal{L}_{\downarrow}$ leads to high top-down consistency scores (e. g., $\gamma_{\text{F}}^{\downarrow} = 0.94$ compared to $\gamma_{\text{F}}^{\downarrow} = 0.49$ for $\text{D}^{0,0}$ and $\gamma_{\text{F}}^{\downarrow} = 0.74$ for Strategy A). Strategy $\text{D}^{0.1,0.1}$ also improves F-measures for some classes on the finest level (notably $F_{\text{F3}} = 0.75$). As another advantage, unlike strategies A and C, this joint strategy requires only a single model.

From a musical point of view, our results indicate that singing activity and singer gender can be identified reliably (as evident by the high evaluation results for these classes across all strategies, except Strategy B). This is particularly important for Strategy C, where mistakes made at higher level are propagated downwards. Thus, Strategy C may perform worse in settings where classes at higher levels are more difficult to separate. Our evaluation results also indicate that, in contrast to higher hierarchy levels, differentiating between different voice types is much more challenging. In addition, by looking more closely at the results for Strategy $\text{D}^{0.1,0.1}$, we found that most false negative predictions for a certain voice type co-occur with a false positive prediction for another voice type from the same gender. Confusions often occur between baritone and bass as well as between soprano and other female voice types. For example, around half of all frames annotated as mezzo are incorrectly predicted to contain soprano instead of mezzo singing.

6. CONCLUSION

In this paper, we have formalized a hierarchical extension of singing voice detection towards singer gender and voice type. We evaluated four possible strategies for solving our task in the context of a large dataset of opera recordings. We compared these strategies with regard to evaluation scores and consistency of predictions, and, in particular, proposed a novel loss term for promoting consistent predictions across hierarchy levels. We showed that a joint classification strategy with our additional loss achieves high results and consistency. The singing scenario considered in this paper indicates the potential of our hierarchical modeling and loss term. These may also be helpful in other use cases. For example, future work may focus on more complex class hierarchies such as those encountered in musical instrument recognition.

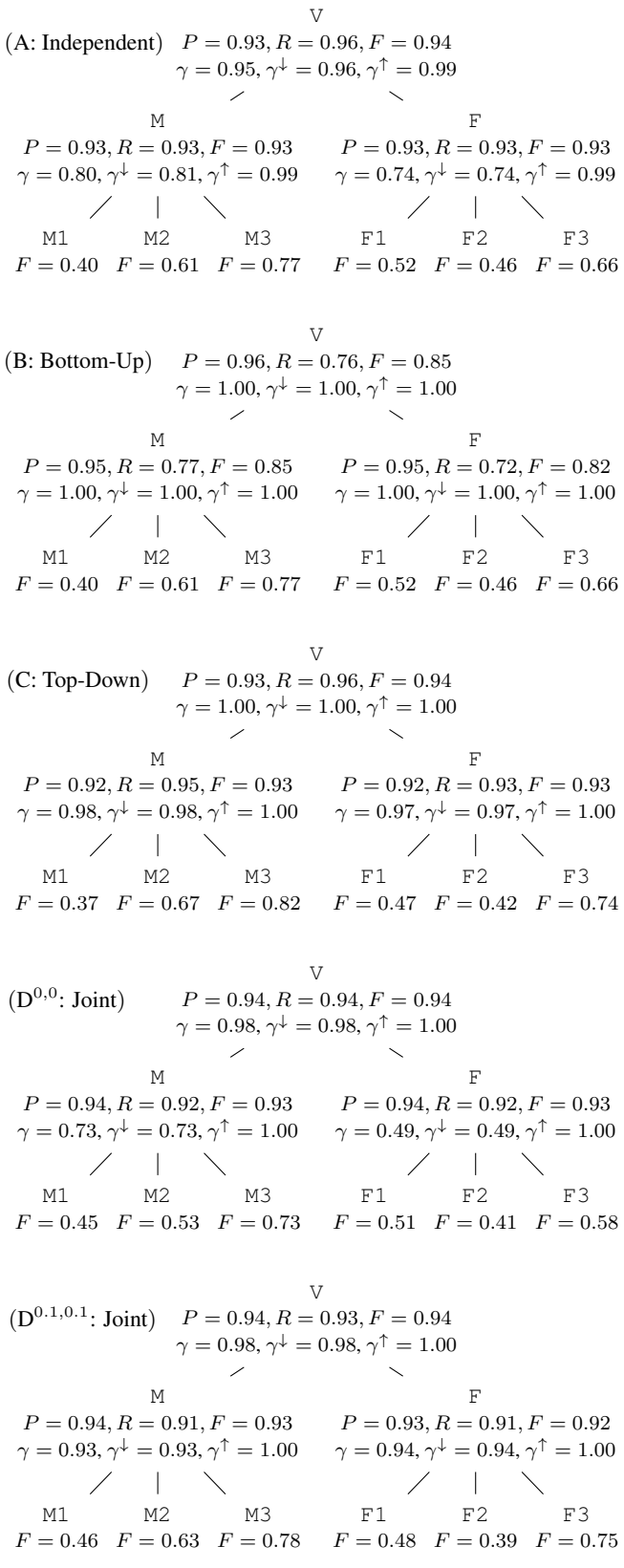

Fig. 3: Results for our detection strategies on the full test set. Subscripts (such as in P_{M2}) are omitted for readability.

7. REFERENCES

[1] Eric J. Humphrey et al., "An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 82–94, 2019.

[2] Bernhard Lehner, Gerhard Widmer, and Reinhard Sonnleitner, "On the reduction of false positives in singing voice detection," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 7480–7484.

[3] Jan Schlüter and Thomas Grill, "Exploring data augmentation for improved singing voice detection with neural networks," in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, Málaga, Spain, 2015, pp. 121–126.

[4] Kyungyun Lee, Keunwoo Choi, and Juhan Nam, "Revisiting singing voice detection: A quantitative review and the future outlook," in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, Paris, France, 2018, pp. 506–513.

[5] Mathieu Ramona, Gäel Richard, and Bertrand David, "Vocal detection in music with support vector machines," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, 2008, pp. 1885–1888.

[6] Jan Schlüter and Bernhard Lehner, "Zero-mean convolutions for level-invariant singing voice detection," in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, Paris, France, 2018, pp. 321–326.

[7] Simon Leglaive, Romain Hennequin, and Roland Badeau, "Singing voice detection with deep recurrent neural networks," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 121–125.

[8] Felix Weninger, Jean-Louis Durrieu, Florian Eyben, Gaël Richard, and Björn W. Schuller, "Combining monaural source separation with long short-term memory for increased robustness in vocalist gender recognition," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 2196–2199.

[9] Matthias Müller, Thilo Schulz, Tatiana Ermakova, and Philipp P. Caffier, "Lyric or dramatic - vibrato analysis for voice type classification in professional opera singers," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 943–955, 2021.

[10] Edward Polrolniczak and Michal Kramarczyk, "Estimation of singing voice types based on voice parameters analysis," in *Signal Processing: Algorithms, Architectures, Arrangements, and Applications, SPA*, Poznan, Poland, 2017, pp. 63–68.

[11] Michael Krause, Meinard Müller, and Christof Weiß, "Singing voice detection in opera recordings: A case study on robustness and generalization," *Electronics*, vol. 10, no. 10, pp. 1214:1–14, 2021.

[12] Stylianos I. Mimilakis, Christof Weiß, Vlora Arifi-Müller, Jakob Abeßer, and Meinard Müller, "Cross-version singing voice detection in opera recordings: Challenges for supervised learning," in *Proc. of the Int. Workshops of ECML PKDD 2019, Part II*, pp. 429–436.

[13] Christian Dittmar, Bernhard Lehner, Thomas Prätzlich, Meinard Müller, and Gerhard Widmer, "Cross-version singing voice detection in classical opera recordings," in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, Málaga, Spain, October 2015, pp. 618–624.

[14] Emilia Parada-Cabaleiro, Maximilian Schmitt, Anton Batliner, Simone Hantke, Giovanni Costantini, Klaus R. Scherer, and Björn W. Schuller, "Identifying emotions in opera singing: Implications of adverse acoustic conditions," in *Proc. of the Int. Conf. on Digital Libraries for Musicology*, Paris, France, 2018, pp. 376–382.

[15] Zheng Tang and Dawn A. A. Black, "Melody extraction from polyphonic audio of Western opera: A method based on detection of the singer's formant," in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, Taipei, Taiwan, October 2014, pp. 161–166.

[16] Jason Cramer, Vincent Lostanlen, Andrew Farnsworth, Justin Salamon, and Juan Pablo Bello, "Chirping up the right tree: Incorporating biological taxonomies into deep bioacoustic classifiers," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing ICASSP*, Barcelona, Spain, 2020, pp. 901–905.

[17] Fu Zih-Sing and Li Su, "Hierarchical classification networks for singing voice segmentation and transcription," in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, Delft, The Netherlands, 2019, pp. 900–907.

[18] Arindam Jati, Naveen Kumar, Ruxin Chen, and Panayiotis G. Georgiou, "Hierarchy-aware loss function on a tree structured label space for audio event detection," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing ICASSP*, Brighton, United Kingdom, 2019, pp. 6–10.

[19] Yong Xu, Qiang Huang, Wenwu Wang, and Mark D. Plumbley, "Hierachical learning for dnn-based acoustic scene classification," Tech. Rep., DCASE2016 Challenge, 2018.

[20] Jonatas Wehrmann, Ricardo Cerri, and Rodrigo C. Barros, "Hierarchical multi-label classification networks," in *Proc. of the Int. Conf. on Machine Learning (ICML)*, Stockholm, Sweden, 2018, pp. 5225–5234.

[21] Luca Bertinetto, Romain Müller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord, "Making better mistakes: Leveraging class hierarchies with deep networks," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition CVPR*, Seattle, WA, 2020, pp. 12503–12512, Computer Vision Foundation / IEEE.

[22] Carlos Nascimento Silla Jr. and Alex Alves Freitas, "A survey of hierarchical classification across different application domains," *Data Mining and Knowledge Discovery*, vol. 22, no. 1-2, pp. 31–72, 2011.

[23] Frank Zalkow, Christof Weiß, and Meinard Müller, "Exploring tonal-dramatic relationships in Richard Wagner's Ring cycle," in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, Suzhou, China, 2017, pp. 642–648.

[24] Frank Zalkow, Christof Weiß, Thomas Prätzlich, Vlora Arifi-Müller, and Meinard Müller, "A multi-version approach for transferring measure annotations between music recordings," in *Proc. of the AES Int. Conf. on Semantic Audio*, Erlangen, Germany, 2017, pp. 148–155.

[25] Aris Kosmopoulos, Ioannis Partalas, Éric Gaussier, Georgios Paliouras, and Ion Androutsopoulos, "Evaluation measures for hierarchical classification: a unified view and novel approaches," *Data Mining and Knowledge Discovery*, vol. 29, no. 3, pp. 820–865, 2015.