A fully differentiable model for unsupervised singing voice separation

Gael Richard, Pierre Chouteau, Bernardo Torres

A FULLY DIFFERENTIABLE MODEL FOR UNSUPERVISED SINGING VOICE SEPARATION

Gaël Richard *Pierre Chouteau* *Bernardo Torres*

LTCI, Télécom Paris, Institut polytechnique de Paris, France

ABSTRACT

A novel model was recently proposed by Schulze-Forster et al. in [1] for unsupervised music source separation. This model allows to tackle some of the major shortcomings of existing source separation frameworks. Specifically, it eliminates the need for isolated sources during training, performs efficiently with limited data, and can handle homogeneous sources (such as singing voice). But, this model relies on an external multipitch estimator and incorporates an Ad hoc voice assignment procedure. In this paper, we propose to extend this framework and to build a fully differentiable model by integrating a multipitch estimator and a novel differentiable assignment module within the core model. We show the merits of our approach through a set of experiments, and we highlight in particular its potential for processing diverse and unseen data.

***Index Terms*—** Unsupervised source separation, multiple singing voices, differentiable models, deep learning

1. INTRODUCTION

Music Source Separation (MSS) [2], the task of estimating the individual music signals when only a mixture is available, has become an indispensable tool in various applications, ranging from remixing tracks and audio transcription to music recommendation and melody extraction for beginner musicians. The field's state-of-the-art now relies on leveraging deep neural networks (DNNs) for this task [3–5]. DNNs offer the advantage of working directly on raw audio signals, bypassing the need for hand-crafted features. However, these methods have their own shortcomings. Firstly, they mainly rely on supervised training, meaning that isolated sources must be accessible and available for training. Secondly, these models are often excessively complex, requiring large amounts of data and computational resources. Finally, although these models perform well on specific datasets, they can run into difficulties when the input data is homogeneous, such as when dealing with sets of choruses of similar voices.

Given these limitations, methods that can operate effectively without access to isolated sources have emerged. Techniques like Non-negative Matrix Factorization (NMF) [6] have shown promise, but often rely on prior information [7] or heavy assumptions about the sources [8], and might have low flexibility due to a pre-defined number of spectral templates. Unsupervised deep-learning-based approaches are promising but only few works address homogenous or correlated sources.

This paper builds upon the work of Schulze-Forster et al. [1], which proposes an unsupervised DNN model that has shown state-of-the-art performance in separating choral singing. We expand their work by integrating the multi-F0 estimation and voice assignment modules as differentiable blocks within the model and by proposing a novel method for differentiable F0 contour extraction. We then obtain an end-to-end, fully differentiable model for unsupervised source separation. We also conduct an extensive experimental validation to demonstrate the efficacy of our methods.

The paper is organised as follows: we recall our baseline unsupervised source separation method in Section 2 before presenting in Section 3 our new fully differentiable approach. The experiments and results are respectively presented in Sections 4 and 5. Finally, we suggest some conclusions in Section 6.

2. UNSUPERVISED MUSIC SOURCE SEPARATION

The original model proposed in [1], referred to as UMSS, was shown to be efficient for complex source separation problems in which individual sources are not available for training, the sources are homogeneous (only singing voices), or only a limited amount of mixture recording data is obtainable. The approach is inspired by the recent hybrid deep learning paradigm, which integrates signal processing models in DNNs to incorporate domain knowledge [9, 10]. In UMSS, each source is represented with a differentiable parametric source-filter model. During training, the task of the DNN is to re-synthesize the observed mixture as a sum of the sources by estimating the source parameters given their fundamental frequencies. The source estimates can be obtained either directly as the synthesized sources or by filtering the initial mixture with soft masks obtained from the synthesized source signals. The latter strategy obtains the best overall results and is then selected in this work.

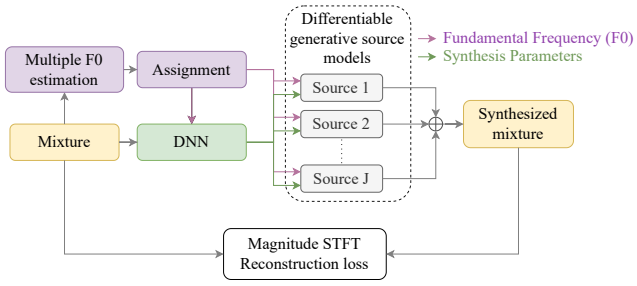

Fig. 1: Overview of the unsupervised music source separation approach proposed in [1].

The model used in [1] to obtain source fundamental frequencies was given in [11] and performs multi-F0 extraction by first processing a spectral representation through a DNN, and then converting the output multi-frequency salience map to F0 contours. In [1], a voice assignment heuristic based on temporal pitch continuity

This work was funded by the European Union (ERC, HI-Audio, 101052978). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

was further added to enable training of the source separation model. A closer look at this pipeline reveals that both salience-map-to-F0 and voice assignment operations are not differentiable. Addressing these issues, we introduce differentiable extensions to these non-differentiable blocks, achieving a fully differentiable architecture.

3. A FULLY DIFFERENTIABLE MODEL

3.1. The complete model

The complete model is shown in Fig. 2. It is based on the unsupervised model described above [1] but with the integration of the multi-F0 estimation and vocal assignment as three differentiable blocks (displayed in purple on Fig. 2). The resulting architecture is then fully differentiable and can be trained end-to-end.

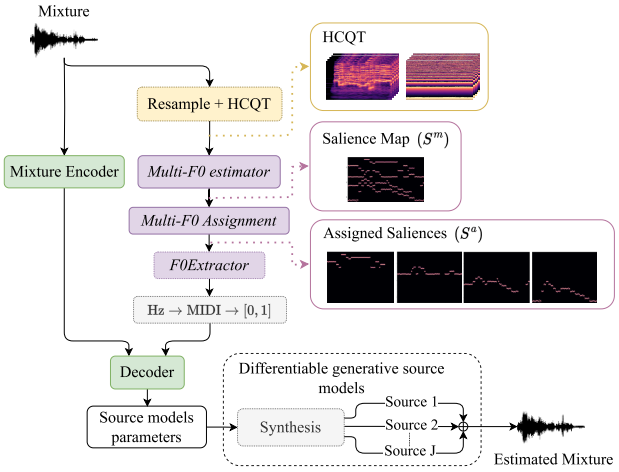

Fig. 2: Fully differentiable architecture for joint learning. The blocks in purple are those that have been added to the initial model.

More precisely, the proposed architecture takes as input a 4-second audio mixture which is processed in parallel in two branches: 1) the first branch is based on the encoder of [1], which extracts the main characteristics of the observed audio. 2) The second branch is dedicated to the estimation of multiple fundamental frequencies. It consists of three blocks: the first estimates a multi-frequency salience map from the observed audio (*Multi-F0 estimator* from [11]); then, this multi-frequency salience map is converted to assigned salience maps, corresponding to the time-frequency representation of the fundamental frequencies of each voice (*Multi-F0 Assignment*, from [12]); finally, our new contour extraction method is applied to each assigned salience map to obtain the F0 sequences of each source over time (*F0Extractor*).

The full loss used to optimize the training is given by:

$$\mathcal{L}_{\text{full}} = \mathcal{L}_{\text{rec}} + \alpha\mathcal{L}_1 + \beta\mathcal{L}_2 + \gamma\mathcal{L}_3, \tag{1}$$

where $\mathcal{L}_{rec}$ is the commonly adopted Multi-Scale Spectral Loss [10] computed between the magnitude STFT's of the input mixture x and estimated mixture $\hat{x}$ (the sum of the estimated sources):

$$\mathcal{L}_{\text{rec}}(x, \hat{x}) = \sum_{n \in \mathcal{N}} \mathcal{L}^n_{\text{lin}} + \mathcal{L}^n_{\text{log}}, \tag{2}$$

with $\mathcal{L}^n_{\text{lin}} = \|\text{STFT}_n(x) - \text{STFT}_n(\hat{x})\|_1$ for scale (window size) n and $\mathcal{L}^n_{\text{log}} = \big(\|\log(\text{STFT}_n(x)) - \log(\text{STFT}_n(\hat{x}))\|_1\big)$.

$\mathcal{L}_1$ is a loss between the sum of the individual assigned salience maps $S^a_j \in \mathbb{R}^{L\times M}$ and the unassigned multi-frequency salience map $S^m \in \mathbb{R}^{L\times M}$, to constrain the sum of assignments to equal the input multi-frequency salience map S^m:

$$\mathcal{L}_1 = \text{MSE}(\sum_{j=1}^{J} S^a_j, S^m), \tag{3}$$

where MSE denotes the *Mean Squared Error*.

$\mathcal{L}_2$ is used to force the assignment model to return frequencies only within a predefined range for each voice (Soprano [260Hz-880Hz]; Alto [190Hz-660Hz]; Tenor [145Hz-440Hz]; Bass [90Hz-290Hz]) [13]. This is achieved by multiplying each assigned salience map S^a_j by a mask corresponding to the desired interval and computing the MSE between the two salience maps, before and after masking.

$$\mathcal{L}_2 = \sum_{j=1}^{J} \text{MSE}(S^a_j, S^{\text{mask}}_j) \tag{4}$$

$\mathcal{L}_3$ is used to force the assigned salience maps to return only one voice, by means of the MSE between the assigned salience maps S^a_j and their binary reconstructions S^b_j.

$$\mathcal{L}_3 = \sum_{j=1}^{J} \text{MSE}(S^a_j, S^b_j) \tag{5}$$

3.2. Differentiable voice assignment

We describe herein the new differentiable voice assignment process in more detail. As depicted in Fig. 3, the overall extraction of the sequence of F0s values for each source includes a number of steps. Starting from an estimated multi-frequency salience map, the *multi-F0 Assignment* module assigns it to each voice in the form of an assigned salience map. For this purpose, we exploit the model proposed by [12] which is, to the best of our knowledge, the only DNN-based model suitable for multiple singing voices. Then, it is necessary to convert these assigned salience maps to sequences of F0 values for each voice. This is classically done by peak-picking followed by thresholding, which are non-differentiable operations.

To cope with this problem, we follow a similar strategy than used in VQ-VAE for gradient propagation of a non-differentiable function [14]. First, from the assigned salience maps S^a, we reconstruct proxy binary salience maps $S^b = \{S^b_j\}$, where the binary map for source j, $S^b_j \in \mathbb{R}^{L\times M}$, has the same dimension as the corresponding S^a_i. S^b contains ones in the frequency bin selected after peak-picking and zeros on all other bins. The F0 contour is then computed as $F0_j = \sum_{i=0}^{M-1} f_i \cdot (S^b_j)_i$, where f is a frequency vector containing the corresponding bin frequencies in Hz. During the *forward* process, the estimation of the sources proceeds normally. In the *backward* pass, however, gradients from S^b are copied to S^a.

3.3. Implementation details and training parameters

The whole architecture is designed for signals sampled at 16 kHz except for the multi-F0 estimation model [11] which works at 22 kHz. To cope with this, the input to the multi-F0 estimation module is upsampled to 22 kHz. Harmonic Constant-Q transform (HCQT) [15] is computed on-the-fly using the *nnAudio* [16] library. For training, we used the ADAM optimizer with a learning rate of 1×10^{-4}

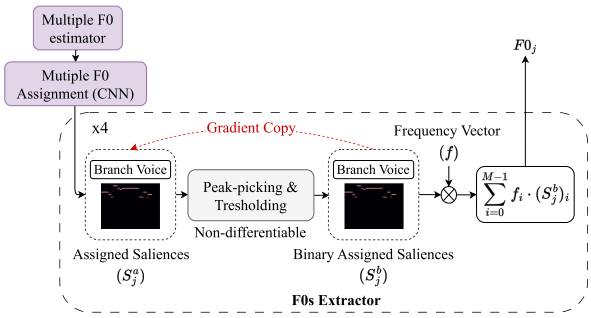

Fig. 3: Extraction of F0 sequences from assigned salience maps. The gradient is copied from S^b to S^a during the *backward* pass. The weighted sum of the binary salience maps is performed on the frequency axis, resulting in frame-level F0 estimations.

and a batch size of 15. For the most part, models are trained with an early stopping set at 200 epochs. Depending on the training options, regularization terms are added with a multiplicative factor (α, β, γ) so that each term has the same weight as $\mathcal{L}_{rec}$. The input to the mixture encoder is an STFT with a window size of 512 and a hop size of 256 samples, following [1]. HCQT representations used span 6 octaves, with 60 channels per octave, resulting in 20 cents per frequency bin. The minimum frequency is set to 32.7 Hz to ensure compatibility with the multi-F0 model in [11]. For the Multi-Scale Spectral loss ($\mathcal{L}_{rec}$), we used window sizes $\mathcal{N} = \{2048, 1024, 512, 256, 128, 64\}$.

4. DATASETS AND EXPERIMENTS

4.1. Datasets

For our experiments, we use the different databases exploited in [1], namely: *BC1Song*, *BCBSQ* and *Choral Singing Dataset*[1]. A fourth database, *Cantoría* is also used to test the generalization ability of our approaches. These four databases are briefly described below:

- *Bach Chorales-Barbershop Quartet (BCBSQ)* is built from the databases Bach Chorales [17] and Barbershop Quartet [18]. It is a commercial database containing 26+22 songs, performed by a quartet of singers (Soprano, Alto, Tenor and Bass (SATB) voices for the BC songs and tenor, lead, baritone and bass voices for BSQ). It contains recordings of all the individual voices, as well as recordings of the choir. This database is used for training with a total duration of 91min 20s and 9min 10s for validation.
- *BC1Song* is a reduced database that takes into account a single Bach Chorales song with a total duration of 2min 40s for training and 2min 20s for validation.
- *Choral Singing Dataset (CSD)* [19] is a public database for choral singing. It consists of recordings of 3 songs performed by an SATB choir, with four singers per section (4 Soprano, 4 Alto, 4 Tenor and 4 Bass). The recordings of the 16 singers are separate, obtained from a spot mic and in a context where each section is recorded at the same time. Residues of the other 3 singers in the section are then present in the different recordings. The database is only used for testing
- *Cantoría* [20] is a choral database consisting of 11 songs performed by a professional SATB choir. It contains recordings of all the voices as well as recordings containing the entire choir (Total duration of 36min 10s). The dataset is only used for testing.

[1]All datasets are resampled at 16 kHz; all audio mixes used are four seconds long allowing direct and consistent comparisons with the results of [1]

4.2. Experiments

Our experiments aim at assessing the global performance of our model and how it compares to selected baselines under different training strategies. We only describe herein experiments where the multi-F0 estimation and voice assignment models are initialized with pre-trained weights since a training of the complete architecture from scratch did not lead to satisfying results. We evaluate several training strategies as described below. When not specified otherwise, $\mathcal{L}_{\text{full}}$ is used for training (Eq. 1).

- $S_F S_F$ where the models *SalienceExtractor* and *SalienceAssignment* are initialized with the pre-trained weights (from [11] and [20]) and fixed during training (only $\mathcal{L}_{\text{rec}}$ is used);
- $S_{FT} S_{FT}$ similar as $S_F S_F$ but after initialisation, the models are fine-tuned jointly with the training of the encoder/decoder;
- $S_F S_{FT}$ where all submodules are pre-trained for initialisation and only the model *SalienceExtractor* is fixed during training;
- W_{UP}: in this setup, the assignment model is fine-tuned (with $\alpha\mathcal{L}_1 + \beta\mathcal{L}_2 + \gamma\mathcal{L}_3$) for 50 epochs. Next, the set of models for F0 estimation is frozen, and the separation/synthesis model is trained for a further 50 epochs ($\mathcal{L}_{\text{rec}}$ only). Finally, all modules are unfrozen and training continues using $\mathcal{L}_{\text{full}}$.

We compare our results to the reference baseline UMSS [1] and the U-Net model from [21].

4.3. Evaluation

The main metric used for evaluation is the Signal-to-Distortion Ratio (SI-SDR) [22], which measures the quality of the separated sources in relation to the original signals.

$$\text{SI-SDR} = 10\log_{10}\left(\frac{\|\eta s^2\|}{\|\eta s - \hat{s}^2\|}\right), \tag{6}$$

where s denotes the target source, $\hat{s}$ the estimated source, and $\eta = \text{argmin}_\eta |\eta s - \hat{s}|^2$. The F0 estimation accuracy is evaluated using three metrics. The Raw Pitch Accuracy (RPA) measures the percentage of voiced frames in which the estimated pitch is within a 0.5 semitone range of the ground-truth [23]. The Raw Chroma Accuracy (RCA) measures the same quantity as RPA but allows for octave errors. Finally, the Overall Accuracy (OA) takes into account all frames, including unvoiced cases.

5. RESULTS

5.1. Discussion

Tables 1a and 1b show the experimental results for models trained on BC1Song and BCBSQ (larger database), respectively. Out of the trained models, $S_F S_F$ has the worst performance. The addition of regularization in models $S_{FT} S_{FT}$, $S_F S_{FT}$, and W_{UP} significantly improves performance across all metrics. Our best-performing model, W_{UP}, achieves 85% RPA (and 85% RCA) on BC1Song and 87% RPA (88% RCA) on BCBSQ. There is, however, a significant drop of approximately 8% in the OA metric, indicating a tendency to incorrectly predict periodicity in unvoiced frames.

Our models consistently outperform Petermann et al.'s [21] U-Net model in the SI-SDR metric. The best results were achieved

Model	SI_SDR [dB]		OA [%]		RPA [%]		RCA [%]	
	μ	Md	μ	Md	μ	Md	μ	Md
UMSS [1]	**6.65**	**7.56**	-	-	-	-	-	-
U-Net [21]	1.5	2.72	-	-	-	-	-	-
$S_F S_F$	2.93	3.59	66	68	72	75	73	77
$S_{FT} S_{FT}$	5.03	6.2	76	81	83	89	84	89
$S_F S_{FT}$	4.84	6.1	76	81	83	89	84	90
W_{UP}	5.22	6.34	77	82	85	90	85	91

(a) BC1Song

Model	SI_SDR [dB]		OA [%]		RPA [%]		RCA [%]	
	μ	Md	μ	Md	μ	Md	μ	Md
UMSS [1]	**6.91**	**7.60**	-	-	-	-	-	-
U-Net [21]	4.44	5.71	-	-	-	-	-	-
$S_F S_F$	2.93	3.59	66	68	72	75	73	77
$S_{FT} S_{FT}$	4.81	6.07	73	79	80	87	82	88
$S_F S_{FT}$	5.77	6.46	78	82	85	90	85	89
W_{UP}	6.20	6.91	79	84	87	91	88	92

(b) BCBSQ

Table 1: Evaluation of proposed approaches on CSD (Test dataset) using source separation and pitch accuracy metrics, for the BC1Song (a) and BCBSQ (b) training datasets. μ stands for the mean and Md for the median. The models are trained and evaluated on mixtures with 4 sources. For all metrics, higher is better.

by W_{UP}, scoring 5.52 dB for BC1Song and 6.20 dB for BCBSQ. However, they do not meet the performance benchmarks of our baseline [1] (6.65 and 6.91 dB), which uses pre-extracted F0 and manual source assignment. This trend holds across both datasets. We believe that this slight performance decrease may be due to the differentiation strategy adopted for the voice assignment module with gradient copy. Models trained on the larger BCBSQ dataset have better performance, as shown in Table 1b. Unlike reported in [1] we do not observe improved relative performance on smaller datasets. We hypothesize that larger datasets are helpful for F0 estimation generalization, as evidenced by the increase in RPA and RCA metrics on BCBSQ compared to BC1Song.

Integrating multi-F0 estimation during training adds complexity and impacts performance. Models like W_{UP} and $S_F S_{FT}$, which use frozen salience extraction, yield better results, with W_{UP} being the most effective. We observed that errors in F0 estimation have a cascading effect on the training and performance of the source separation model, which is consistent with findings in [1] and results in the seminal DDSP paper [10], where even monophonic F0 joint estimation was reported to be a major challenge.

5.2. Ablation

We discuss here in an ablation experiment the effectiveness of the proposed differentiable voice assignment module in the whole training process. This analysis is limited to the two best-performing approaches, W_{UP} and $S_F S_{FT}$. For the warm-up approach (W_{UP}), we give the results obtained for each main stage of training: warm-up of the assignment model (*F0*), training of the synthesis model (*Synth*) and further training of the complete model which leads to (W_{UP}).

Model	SI_SDR [dB]		OA [%]		RPA [%]		RCA [%]	
	μ	Md	μ	Md	μ	Md	μ	Md
F0	1.99	2.67	77	82	84	90	85	90
Synth	5.44	6.23	77	82	84	90	85	90
W_{UP}	**6.2**	**6.91**	**79**	**84**	**87**	**91**	**88**	**92**
$S_{FT} S_{FT}$	5.77	6.46	78	82	85	90	86	90
$S_F S_{FT} F$	**5.95**	**6.76**	**79**	**83**	**87**	**91**	**88**	**92**

Table 2: Evaluation of training stages for *W-Up* and $S_F S_{FT}$ on BCBSQ

For the $S_F S_{FT}$ approach, we continue training with the entire trainable architecture, which is referred to $S_F S_{FT} F$ in Table 2.

It can be seen that for both approaches, the training of the complete architecture is beneficial which confirms the effectiveness of our proposed contour extraction process. It also shows that the model used to extract the multi-frequency saliency maps (blocked for the $S_F S_{FT}$ method) becomes more efficient thanks to information derived from the Multi-Scale Spectral loss.

5.3. Generalization capabilities on a new dataset: Cantoría

We study herein the ability of our approach to generalize on another database (Cantoría). The results presented in Table 3 correspond to the performance on Cantoría of the models trained on BC1song and BCBSQ. All scores are clearly lower, underlining the difficulty of the task, but our approach (W_{UP}) is clearly most robust in this context. We relate this to the generalisation capabilities of our voice assignment module. In the original model, the heuristical assignment model was very efficient, but apparently overfitted on the characteristics of the training databases BC1song and BCBSQ.

Model	BC1Song		BCBSQ	
	μ	Md	μ	Md
UMSS [1]	0.31	0.73	0.86	1.38
U-Net [21]	-2.31	-2.07	0.97	1.47
W_{UP}	**1.93**	**2.61**	**3.29**	**3.79**

Table 3: SI-SDR [dB] on Cantoría for the models trained on BC1song and BCBSQ

To support reproducibility, we open source our code and provide a demo page with sound examples[2].

6. CONCLUSION

We have proposed a fully differentiable architecture for unsupervised music source separation which can be trained end-to-end. Our results demonstrate the merits of this new architecture and in particular they show that our model has better generalization capabilities when applied to more diverse data. Future work will be dedicated to the design of an architecture that could be efficiently trained from scratch, extending the current model which relies on a thoroughly designed warm-up procedure with pre-trained sub-modules.

[2]Audio demo at: https://pierrechouteau.github.io/audio ; Code at https://github.com/PierreChouteau/umss_icassp

7. REFERENCES

[1] K. Schulze-Forster, G. Richard, L. Kelley, C. S. J. Doire, and R. Badeau, "Unsupervised Music Source Separation Using Differentiable Parametric Source Models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–14, 2023.

[2] E. Cano, D. FitzGerald, A. Liutkus, M. D. Plumbley, and F.-R. Stöter, "Musical source separation: An introduction," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 31–40, 2018.

[3] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Demucs: Deep extractor for music sources with extra unlabeled data remixed," *arXiv preprint arXiv:1909.01174*, 2019.

[4] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix-a reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019.

[5] N. Takahashi and Y. Mitsufuji, "D3net: Densely connected multidilated densenet for music source separation," *arXiv preprint arXiv:2010.01733*, 2020.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[7] S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 129–132.

[8] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1180–1191, 2011.

[9] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-based deep learning," *Proceedings of the IEEE*, vol. 111, no. 5, pp. 465–499, 2023.

[10] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable Digital Signal Processing," in *International Conference on Learning Representations*, Aug. 2020.

[11] H. Cuesta, B. McFee, and E. Gomez, "Multiple F0 Estimation in Vocal Ensembles Using Convolutional Neural Networks," in *Proceedings of the 21th International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, Canada, Oct. 2020.

[12] H. Cuesta and E. Gómez, "Voice Assignment in Vocal Quartets Using Deep Learning Models Based on Pitch Salience," *Transactions of the International Society for Music Information Retrieval*, vol. 5, no. 1, pp. 99–112, May 2022.

[13] M. Scirea and J. A. Brown, "Evolving Four Part Harmony using a Multiple Worlds Model," in *Proceedings of the 7th International Joint Conference on Computational Intelligence (IJCCI)*, 2015, pp. 220–227.

[14] A. van den Oord, O. Vinyals, and k. kavukcuoglu, "Neural Discrete Representation Learning," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[15] R. M. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello, "Deep salience representations for F0 estimation in polyphonic music," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017.

[16] K. W. Cheuk, H. Anderson, K. Agres, and D. Herremans, "nnAudio: An on-the-Fly GPU Audio to Spectrogram Conversion Toolbox Using 1D Convolutional Neural Networks," *IEEE Access*, vol. 8, pp. 161 981–162 003, 2020.

[17] "PG Music - The Bach Chorales." [Online]. Available: https://www.pgmusic.com/bachchorales.htm

[18] "PG Music - The Barbershop Quartet." [Online]. Available: https://www.pgmusic.com/barbershopquartet.htm

[19] H. Cuesta, E. Gómez, A. Martorell, and F. Loáiciga, "Analysis of intonation in unison choir singing," in *Proceedings of the 15th International Conference on Music Perception and Cognition*, Jul. 2018.

[20] H. Cuesta and E. Gómez, "Cantoría Dataset," Jan. 2022. [Online]. Available: https://zenodo.org/record/5851070

[21] D. Petermann, P. Chandna, H. Cuesta, J. Bonada, and E. Gomez, "Deep Learning Based Source Separation Applied to Choir Ensembles," in *Proceedings of the 21th International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, Canada, Oct. 2020.

[22] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 626–630.

[23] G. E. Poliner, D. P. W. Ellis, A. F. Ehmann, E. Gómez, S. Streich, and B. Ong, "Melody transcription from music audio: Approaches and evaluation," *IEEE Trans. Speech Audio Process.*, vol. 15, no. 4, pp. 1247–1256, 2007.