INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Vocal Harmony Separation using Time-domain Neural Networks

Saurjya Sarkar, Emmanouil Benetos, Mark Sandler

Centre for Digital Music, Queen Mary University of London, United Kingdom
{saurjya.sarkar,emmanouil.benetos,mark.sandler}@qmul.ac.uk

Abstract

Polyphonic vocal recordings are an inherently challenging source separation task due to the melodic structure of the vocal parts and unique timbre of its constituents. In this work we utilise a time-domain neural network architecture re-purposed from speech separation research and modify it to separate *a capella* mixtures at a high sampling rate. We use four-part (soprano, alto, tenor and bass) *a capella* recordings of Bach Chorales and Barbershop Quartets for our experiments. Unlike current deep learning based choral separation models where the training objective is to separate constituent sources based on their class, we train our model using a permutation invariant objective. Using this we achieve state-of-the-art results for choral music separation. We introduce a novel method to estimate harmonic overlap between sung musical notes as a measure of task complexity. We also present an analysis of the impact of randomised mixing, input lengths and filterbank lengths for our task. Our results show a moderate negative correlation between the harmonic overlap of the target sources and source separation performance. We report that training our models with randomly mixed musically-incoherent mixtures drastically reduces the performance of vocal harmony separation as it decreases the average harmonic overlap presented during training.

Index Terms: source separation, polyphonic vocal music, end-to-end, singing analysis.

1. Introduction

Choral music consists of a group of singers typically singing the same lyrics but in different vocal styles and notes creating a polyphonic harmony. These different vocalists are usually categorised into 4 parts by their singing style and vocal registers [1]. These classes are often also used to identify parts of other musical ensembles such as brass sections. Such musical ensembles consisting of sources with similar timbres can be defined as monotimbral ensembles. The task of separating sources in a monotimbral ensemble is significantly different than the much better researched problem of class-based Music Source Separation [2] where there are distinct instrument types to be separated. Unlike the vocal vs. accompaniment (drums/bass/others) problem, the sources in a choral mixture are very similar to each other and highly synchronised in both time and harmony. In contrast, drums are highly percussive sources with distinctive temporal structure and the bass occupies a completely different frequency range as compared to the vocals. This makes vocal harmony separation a significantly more challenging task than the vocal vs. drums and bass separation problem.

Music source separation has been an actively researched field in the last decade. Significant advances have been made in this field since the advent of deep learning based models. More recently the SiSEC challenge [3] has provided a common baseline, a large publicly available dataset and an evaluation framework for the most popular music separation task of decomposing pop music mixtures into as many as 4 parts (vocals, drums, bass and other remaining instruments). A reduced form of the same problem is to separate the mixture into vocals and accompaniment only, which enables commercial applications to generate automatic karaoke [2]. This task has attracted a lot of interest from the research community with vast improvements in recent years [4, 5, 6, 7]. Very little research has been seen in more specific and challenging music separation tasks like monotimbral separation which can be a powerful music production tool as ensemble recordings typically have some bleed[1].

Popular approaches to perform music source separation have typically relied on spectrogram masking based methods [8, 9], where the spectrogram of the mixture is provided as the input to the model which subsequently predicts a mask which is applied on the mixture spectrogram to suppress all non-target sources. Although spectrogram based methods have consistently reported state-of-the-art results in music separation [7], the lack of accurate phase estimation still proves to be the Achilles heel of this task. Without accurate phase estimation, the best achievable of performance for such models is limited by the Ideal Ratio Mask performance [10].

Time-domain source separation models have surpassed this threshold in the domain of speech separation, due to their ability to encapsulate phase information in the learnt filterbanks which removes the requirement of phase reconstruction[11]. Since Conv-TasNet [10], further developments on time-domain source separation [12, 13, 14] methods have pushed speech separation performance far beyond soft-masking based approaches for single-channel 2 and 3 speaker separation tasks. Although music separation has seen some success with time-domain approaches [6, 15], these approaches introduce time-domain processing in a direct regression fashion to model musical sources, whereas popular speech separation models work with a different encoder-masker-decoder philosophy with a significantly smaller number of model parameters.

In this work, we adapt the encoder-masker-decoder type TasNet [16] based architectures for vocal harmony separation. We find the task of Vocal Harmony Separation to fit in a unique space between music and speech separation, where we find challenging aspects of both tasks to merge. Sources present in choral mixtures are often very similar with weak distinction between them thus allowing the possibility of training them using permutation invariant training [17] like speech separation models. Meanwhile, unlike speech separation, the sources present in these mixtures are highly correlated and synchronised to each other as they sing the same words with different harmonisations in synchronisation. This poses a unique problem where there is very little timbral distinction between the sources and high temporal synchronisation and frequency overlap due to their musical structure. We explore the implications of these constraints on training methods of these models. We present a new method

[1] Sound picked up by a microphone from a source other than that which is intended.

http://dx.doi.org/10.21437/Interspeech.2021-1531

to estimate the separation difficulty of a mixture with a music theory based harmonic overlap score. We find that the measured harmonic overlap does present moderate negative correlation with separation performance for vocal harmony mixtures. This implies that the musical structure/complexity of a mixture does impact the separation performance of a mixture, as compared to a mixture of random speech with similar number of sources. We also find that models trained on randomly mixed (musically incoherent) data show higher variability in performance for separation tasks with higher harmonic overlap.

The remaining of the paper is structured as follows: in Section 2 we present other works in the literature that tackle the problem of vocal harmony separation. In Section 3 and 3.1 we present the details of the models we use for our task and the modifications required to work at higher sampling rates. Section 3.2 introduces the proposed metric to calculate the musical complexity of a vocal harmony mixture. We present the details of the data used in our experiments in Section 3.3 and training methods in Section 3.4. We then compare our results with the state-of-the-art in Section 4 and present our findings related to random mixing and harmonic overlap. We discuss the implications of our findings and future work in Section 5.

2. Related Work

While music source separation is a well researched topic especially since the publication of the MUSDB dataset [3], the majority of the work has been limited to the separation tasks enabled by the above dataset i.e. vocals, bass, drums and others. A few papers have discussed separation tasks for other instrument/mix types but they rely on either datasets with limited size [18] or large synthesised datasets [19]. The topic of choral music separation faces the same problem, where the datasets with real recordings are limited in size and scope [20], and alternatively there are large synthesised datasets [21].

Two recent works [22, 21] explore score-informed choral separation utilising conditioned U-Net [5] and Wave-U-Net [6] architectures. While both models show reasonable success, it is difficult to compare the performance of the two since [22] is trained and evaluated on real data with bleed, and [21] utilises synthesised vocal choirs. While both these methods present poor baseline scores for non-informed separation, in our work we present a non-informed application of time-domain source separation models that outperforms the non-informed separation baselines presented in [22, 21]. Our model performs comparably to the score-informed models presented in [22, 21].

3. Method

We used modified versions of the Conv-TasNet [10] and Dual-path Transformer (DPTNet) [13] based time-domain separation architectures from [23] for our experiments using permutation invariant training (PIT) [17]. We modify the network parameters[2] to accommodate for the higher sampling rate data in our usecase within the available GPU resources. Higher sampling rates significantly impact the GPU memory consumption of such time-domain models. TasNet based models [16, 10, 12, 13] tend to have extremely short encoder filterbanks (2-20 samples) as compared to spectrogram based models. This generates feature representations at a much higher time resolution requiring significantly higher GPU memory for backpropagation. Moreover, [11] show that the performance gains observed in these models can be strongly attributed to this high time resolution. Given our GPU memory constraints, we have to find a balance between the filter lengths, input segment length and batch size to find optimal performance at higher sampling rates.

3.1. Accommodating higher sampling rates

Speech separation models typically operate at 8 kHz while commercial music is consumed at much higher sampling rates (44.1kHz and higher). In our dataset the original vocals were bandlimited to 11kHz, thus we trained our models at 22.05 kHz.

For Conv-TasNet we increased the filter length to 20 samples and hop size to 10 samples and also increased the number of dilated layers to 9 to achieve a similar receptive field of ≈1.5 second at 22.05 kHz as the original implementation at 8 kHz.

For DPTNet, we could not accommodate 5 second input segments with a filter length of 2 samples and 6 repeat units on our GPU memory. Thus, we tried different combinations of filter lengths, input durations and repeat units (R) in our experiments to utilise all available memory and show the impact of the different parameters. Depending on the input audio segment duration and filter length, we had to modify the chunk size K as per the eq. $K = \sqrt{2L}$ as mentioned in [12]. L is the length of the latent representation generated by the encoder determined by eq. 1 where T is the duration of the training samples, F_s is sampling rate and L_f is the length of the filters in the encoder.

$$L = \frac{2 \times T \times F_{\mathrm{s}}}{L_{\mathrm{f}}} \tag{1}$$

3.2. Harmonic Overlap Score

We present a novel measure for calculating the harmonic overlap for any two given monophonic sources based on calculating the number of coinciding partials observed in the first 16 overtones for a given pair of F0s being played by two sources. This also correlates well to the perceived resonance for any given interval (pair of notes), where the strongest resonances are seen for octave intervals, followed by perfect fifths, perfect fourths, major thirds and so on. This measure is particularly apt for monophonic sources in an ensemble where such instruments perform together with the intention of blending well with each other to create a coherent sonic texture. We design our harmonic overlap metric to give us a measure of coherence for such ensembles by calculating pair-wise harmonic overlaps normalised by their duration of activity.

Given a set of N sources x_i for $i \in \{1, 2, ..., N\}$, we utilise the pYIN pitch detection algorithm [24] to estimate their pitches F_i^0 and convert them to a 20 cent log-frequency scale for each of the sources resulting in P_i^0. We then compute the first 16 overtones P_i^j for $j \in \{1, 2, ..., 16\}$ as per eq. 2. We convert the obtained set of harmonic pitches to a binary vector $B_{\mathrm{i,k}}$ as per eq. 3 for $k \in \{1, 2, 3, ...\}$. We subsequently get the harmonic overlap for a given time frame by counting the total number of overlaps per frame for each pair of sources in a mixture as per eq. 4. We then aggregate the pair-wise scores over the entire input segment and normalise the score by dividing by the overall pair wise activity duration, i.e. for each pair we calculate the total number of frames where both sources were active and divide the aggregated score by that value.

$$P_{\mathrm{i}}^{\mathrm{j}} = \left\lceil 60 \times log_2(\frac{j \times F_{\mathrm{i}}^{0}}{440}) \right\rceil + 345 \tag{2}$$

[2] Complete model parameters and audio examples for each model can be found at: http://c4dm.eecs.qmul.ac.uk/ChoralSep/.

$$B_{\text{i,k}} = \begin{cases} 1, & \text{for k} \in P_{\text{i}}^{\text{j}} \\ 0, & \text{for k} \notin P_{\text{i}}^{\text{j}} \end{cases} \quad j \in \{1, 2, ..., 16\} \tag{3}$$

$$\textit{Harmonic Overlap} := \sum_{i \neq j}^{N} B_{\text{i,k}} \cdot B_{\text{j,k}}{}^{\text{T}} \tag{4}$$

3.3. Dataset

There are very few clean datasets available for choral music, where isolated ground truth for each source is present. This is especially challenging as compared to other ensembles, since choral singers typically perform together and are rarely recorded in isolation [25]. It is known that choral singers tend to perform much better when the entire choir performs together in a physical space [26], i.e. each singer can monitor themselves and the rest of the choir with every participant making minor adjustments during performance [27]. This makes it very difficult to record each individual singer without any bleed from the other sources. There is one publicly available dataset that consists of 3 choral pieces performed by 16 singers [20], but the recordings are not clean as all the sources are recorded simultaneously. This causes the non-target sources to bleed into each of the recordings, resulting in a noisy ground truth.

In this work, we use two datasets of *a capella* recordings without bleed from [28] for our experiments, 26 songs from Bach Chorales (BC) and 22 songs from Barbershop Quartets (BQ). This gives us a total of 104 minutes of 4 parts: Soprano, Alto, Tenor and Bass (SATB) recordings, where BC contain 2 male (tenor and bass) and 2 female (Soprano and alto) vocalists, and BQ contain all 4 male vocalists. We split the songs present in the dataset into 3 groups for training, cross validation and testing roughly in ratio 8:1:1 (since song lengths vary), making sure that the test and cross-validation sets consist of songs (and not just segments) that are unseen in the training set.

3.4. Training

We train both variants of models for 200 epochs with early stopping given a patience of 30 epochs. We use the SI-SNR [29] loss function as shown in eq. 5 where $\bar{x}$ is the predicted source and x is the target source. We use SI-SNR in a class-agnostic/permutation invariant [17] fashion where we compute the pair-wise SI-SNR for each predicted source w.r.t. each target source, and then consider the prediction-target assignments for the lowest cumulative SI-SNR.

$$\begin{aligned} s_{\text{target}} &= \frac{\langle \bar{x}, x \rangle x}{\|x\|^2} \\ e_{\text{noise}} &= \bar{x} - s_{\text{target}} \\ SI - SNR &:= 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2} \end{aligned} \tag{5}$$

For our Conv-TasNet based model, we initialise the learning rate to $5e^{\text{-3}}$ with a scheduler that halves the learning rate if the validation loss (cross-validation set of 2 unseen songs) does not improve for 3 consecutive epochs. For the DPTNet based model, we use the linear warmup followed by exponential decay scheduler as presented in the original paper [13].

4. Results

We evaluate the performance of our models on 2 unseen songs each from the Bach Chorales and Barbershop Quartet dataset (9 minutes in total). We find that although the training and test sets are similar (due to similar singing style and limited variety of vocalists), our models perform better than other non-informed separation models based on U-Net and Wave-U-Net [22] which were trained on a dataset of similar duration and diversity [20]. Our non-informed model performs at par with the state-of-the-art score-informed separation models Conditioned U-Net [22] and Conditioned Wave-U-Net [21]. It must be noted that results from both [22, 21] were reported on different datasets than us, thus it is difficult to make conclusive performance comparisons on the reported results.

Table 1: *Results for 4-source Choral Music Separation w.r.t. other works in literature. It must be noted that both [22, 21] use different datasets to train and evaluate their models thus are not directly comparable.*

Model	**SIR**	**SAR**	**SDR**
ConvTasNet	+12.23 dB	+9.27 dB	+7.52 dB
DPTNet	**+14.42 dB**	**+10.25 dB**	**+8.61 dB**
U-Net[22]	+9.30 dB	+5.69 dB	-
Wave-U-Net[22]	+7.07 dB	+5.54 dB	-
Wave-U-Net[21]	-	-	+5.4 dB
C-U-Net[22]	+12.08 dB	+7.21 dB	-
C-Wave-U-Net[21]	-	-	+8.1 dB

Table 2: *Performance comparison for DPTNet models trained with various filter sizes, repeat units and input segment lengths.*

Model	**L_f**	**T**	**R**	**SDR**
ConvTasNet	20	5 sec	6	+7.52 dB
DPTNet	16	5 sec	8	**+8.61 dB**
DPTNet	8	5 sec	4	+7.9 dB
DPTNet	4	2 sec	5	+7.56 dB
DPTNet	2	2 sec	3	+7.16 dB

4.1. Model Variation

We compare the performance of Conv-TasNet and DPTNet for our task at 22.05 kHz with our best performing Conv-TasNet model and various filter lengths and repeat units for DPTNet. While [11] concluded that a large part of the performance gains observed in TasNet based solutions with respect to spectrogram based separation architecture could be largely attributed to the small filter lengths, we do not observe the same improvement in our experiments. This is likely due to us having to use shorter input segments of 2 seconds and reduced repeat units to accommodate models with shorter filter lengths of 2 and 4 samples on DPTNet. We instead find that using larger filterbanks of 16 samples and increasing the repeat units to 8 gives us our best performance with DPTNet.

4.2. Random Mixing

Random mixing is a commonly used data augmentation method which was introduced for Music Source Separation in [4] and subsequently also utilised in [14, 15] where they find that mixing segments from different musical pieces to generate new training examples for training improves separation performance. Since finding clean multi-tracks for highly polyphonic ensembles like choral music is difficult, random mixing would enable us to generate data with any amount of polyphony by mixing monophonic singing tracks from various songs.

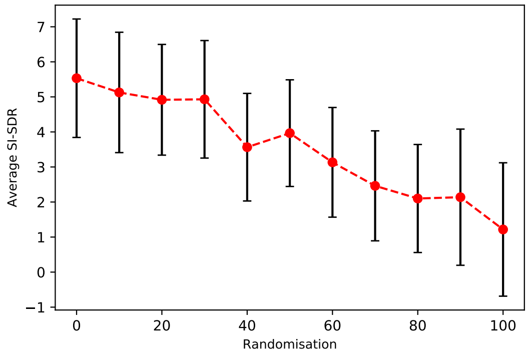

Figure 1: *Average output SI-SDR achieved by different ConvTasNet based models trained with varying balance of randomised (musically incoherent) and synchronised mixtures.*

We systematically study the impact on model performance of randomly mixing vocal parts from different songs during training. We randomly choose a number of data samples from the training set and shuffle their constituent parts to generate a training set with a desired percentage of randomisation. In Figure 1 we see that the model performance monotonically decreases as the amount of randomised mixtures in the training data is increased. We see a 4.32 dB decrease in average SI-SDR improvement between a model trained on synchronised mixtures vs. randomised mixtures. It is noteworthy that our randomisation process preserves the SATB choral structure.

We also carry out experiments where we increase the overall dataset size by adding new mixtures of randomised samples without replacement. We observe that increasing the dataset size does not improve separation performance. Models trained on expanded datasets with $10 - 100\%$ additional training samples of randomised mixtures show an average performance difference of $+0.07 \pm 0.32$ dB ΔSI-SDR w.r.t. our baseline model.

4.3. Harmonic Overlap Analysis

We use the Harmonic Overlap metric as introduced in section 3.2 and find a moderate negative correlation (Pearson correlation coefficient: -0.334) between the harmonic complexity of the audio mixture and the separation performance achieved, as shown in Figure 2. This shows that mixtures with stronger resonant intervals are more difficult to separate, thus musical structure does impact separation negatively. This agrees with our perceptual ability of distinguishing harmonies being sung as the Harmonic Overlap score ranks resonant intervals much higher than dissonant intervals. We observe that models trained on randomised data show higher performance degradation for mixtures with higher Harmonic Overlap score (Pearson correlation coefficient: -0.215) and higher performance variance. This may also explain why randomisation of mixtures during training affects the overall model performance as randomising mixtures significantly reduces the average harmonic overlap in the mixtures presented during training. Hence, models need to be presented with synchronised, musically coherent training data to be able to separate monotimbral ensembles.

5. Conclusion

We show that TasNet based separation models with PIT are effective for separating vocal harmonies. The main challenges to achieve a good level of performance and generalisability is limited by two main factors: lack of sizeable clean multi-track data, memory and resource consumption for training models at high sample rates. While we present that adapting DPRNN and DPTNet like architectures with larger filter lengths does alleviate some of the memory limitations, the lack of sufficient data does not allow us to make conclusive statements regarding the impact of larger filter lengths on performance. Also, our TasNet based models show good performance on similar datasets but cross-dataset performance is really poor. In such scenarios, we see a significant performance advantage of the score-informed model presented in [22].

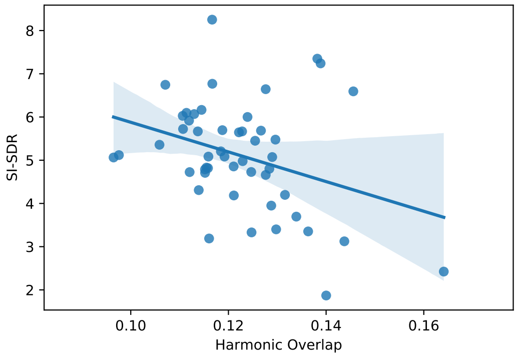

Figure 2: *Linear fit with 95% confidence interval of Harmonic Overlap score for test audio mixtures vs. output SI-SDR achieved with ConvTasNet model.*

While training with randomised mixtures is a common approach in source separation, we report a significant negative impact of randomisation of training mixtures for vocal harmony separation. This behaviour is not seen in other music and speech separation tasks and we suspect that this is due to the combination of musical structure and timbral similarity between sources which is unique to our task. This is especially problematic as available data for vocal harmonised mixtures with more than 2 sources is very limited, so while randomisation could be used to generate higher polyphony training examples from isolated vocal tracks from [18], models trained on such data did not succeed in our experiments.

A potential solution for this problem can be based on recursive separation using one-and-rest permutation invariant training (OR-PIT) [30]. Having shown that PIT based objective functions are successful at vocal harmony separation, OR-PIT provides a unique opportunity to utilise data with varying polyphony for training. While using such a method may not provide significant performance improvement for the task of speech separation, we believe that being able to utilise musical mixtures with varying degrees of polyphony may allow us to create sizeable datasets for this task, without compromising on musical structure and relevance of the training set. We intend to explore such approaches in the future.

6. Acknowledgements

S. Sarkar is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by UK Research and Innovation [grant number EP/S022694/1] and Queen Mary University of London. This work was supported by a Turing Fellowship for E. Benetos under the EPSRC grant EP/N510129/1.

7. References

[1] R. Shewan, "Voice classification: An examination of methodology," *The NATS Bulletin*, vol. 35, no. 3, pp. 17–25, 1979.

[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. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2017, pp. 323–332.

[4] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 261–265.

[5] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *Proc. of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017, pp. 23–27.

[6] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proc. of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 334–340.

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

[8] 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.

[9] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 2020.

[10] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[11] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying tasnet: A dissecting approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6359–6363.

[12] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[13] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *Proc. Interspeech 2020*, pp. 2642–2646, 2020.

[14] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *arXiv preprint arXiv:2002.08933*, 2020.

[15] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019.

[16] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[17] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[18] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research." in *Proc. of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, vol. 14, 2014, pp. 155–160.

[19] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some slakh: A dataset to study the impact of training data quality and quantity," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 45–49.

[20] H. Cuesta, E. Gómez Gutiérrez, A. Martorell Domínguez, and F. Loáiciga, "Analysis of intonation in unison choir singing," in *Proc. of the 15th International Conference on Music Perception and Cognition (ICMPC)*, 2018.

[21] M. Gover and P. Depalle, "Score-informed source separation of choral music," Master's thesis, McGill University, 2019.

[22] D. Petermann, P. Chandna, H. Cuesta, J. Bonada, and E. Gómez Gutiérrez, "Deep learning based source separation applied to choir ensembles," in *Proc. of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[23] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas *et al.*, "Asteroid: the pytorch-based audio source separation toolkit for researchers," *arXiv preprint arXiv:2005.04132*, 2020.

[24] M. Mauch and S. Dixon, "pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 659–663.

[25] K. Ihalainen, "Methods of choir recording for an audio engineer," Ph.D. dissertation, Tampere Polytechnic, 2008.

[26] T. Fischinger, K. Frieler, and J. Louhivuori, "Influence of virtual room acoustics on choir singing." *Psychomusicology: Music, Mind, and Brain*, vol. 25, no. 3, p. 208, 2015.

[27] J. Dai and S. Dixon, "Analysis of interactive intonation in unaccompanied satb ensembles." in *Proc. of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[28] R. Schramm and E. Benetos, "Automatic transcription of a cappella recordings from multiple singers," in *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*. Audio Engineering Society, 2017.

[29] 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)*. IEEE, 2019, pp. 626–630.

[30] N. Takahashi, S. Parthasaarathy, N. Goswami, and Y. Mitsufuji, "Recursive speech separation for unknown number of speakers," *Proc. Interspeech 2019*, pp. 1348–1352, 2019.