frontiers | Frontiers in Signal Processing

ORIGINAL RESEARCH
published: 05 April 2022
doi: 10.3389/frsip.2022.808594

A Deep-Learning Based Framework for Source Separation, Analysis, and Synthesis of Choral Ensembles

Pritish Chandna[1,2]*, *Helena Cuesta*[1], *Darius Petermann*[3] *and Emilia Gómez*[1,4]

[1]*Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain,* [2]*Voicemod S.L., Valencia, Spain,* [3]*Intelligent Systems Engineering Department, Indiana University Bloomington, Bloomington, IN, United States,* [4]*Joint Research Centre, European Commission, Seville, Spain*

Choral singing in the soprano, alto, tenor and bass (SATB) format is a widely practiced and studied art form with significant cultural importance. Despite the popularity of the choral setting, it has received little attention in the field of Music Information Retrieval. However, the recent publication of high-quality choral singing datasets as well as recent developments in deep learning based methodologies applied to the field of music and speech processing, have opened new avenues for research in this field. In this paper, we use some of the publicly available choral singing datasets to train and evaluate state-of-the-art source separation algorithms from the speech and music domains for the case of choral singing. Furthermore, we evaluate existing monophonic F0 estimators on the separated unison stems and propose an approximation of the perceived F0 of a unison signal. Additionally, we present a set of applications combining the proposed methodologies, including synthesizing a single singer voice from the unison, and transposing and remixing the separated stems into a synthetic multi-singer choral signal. We finally conduct a set of listening tests to perform a perceptual evaluation of the results we obtain with the proposed methodologies.

Keywords: audio signal processing, deep learning, choral singing, source separation, unison, singing synthesis

OPEN ACCESS

Edited by:
Wenwu Wang,
University of Surrey, United Kingdom

Reviewed by:
Akinori Ito,
Tohoku University, Japan
Emmanouil Benetos,
Queen Mary University of London, United Kingdom

***Correspondence:**
Pritish Chandna
pritish.chandna@voicemod.net

Specialty section:
This article was submitted to Signal Processing Theory, a section of the journal Frontiers in Signal Processing

Received: *03 November 2021*
Accepted: *07 March 2022*
Published: *05 April 2022*

Citation:
Chandna P, Cuesta H, Petermann D and Gómez E (2022) A Deep-Learning Based Framework for Source Separation, Analysis, and Synthesis of Choral Ensembles. Front. Sig. Proc. 2:808594. doi: 10.3389/frsip.2022.808594

1 INTRODUCTION

Choral singing is one of the most widespread types of ensemble singing (Sundberg, 1987). It is a central part of musical cultures across the world and it has been an important activity for humans to express ideas and beliefs, as well as for social entertainment and mental wellbeing (Clift et al., 2010; Livesey et al., 2012). *Vocal ensemble* is the term we commonly use to describe a set of multiple singers singing simultaneously. These singers are often divided into different sections based on their vocal range, and an ensemble comprising multiple sections is often referred to as a *choir*. Singers within a section typically sing the same melodic line, referred to as a *unison*, which is complementary to the melodic lines sung by other sections. A commonly used configuration for Western choral singing, which is the focus of this paper, is the Soprano, Alto, Tenor and Bass (SATB), comprising the aforementioned four sections. Combined together, these sections produce a harmonious effect, known as the SATB choir sound. This form of singing is widely practised and studied in Western culture.

Music Information Retrieval (MIR) is the field of research combining techniques from the fields of musicology, signal processing, and machine learning among others, to computationally analyze music so as to assist practitioners, learners, and enthusiasts. In this field, we find some early literature

about the acoustic properties of choral singing (Rossing et al., 1986; Ternström, 1991; Ternström, 2002) which attempted to understand the "choral sound." Over the last few years, the MIR community has shown an increasing interest on the topic, particularly with the emergence of data-driven machine learning methodologies, as well as novel datasets. Studies on singers' intonation and interaction in vocal ensembles (Devaney et al., 2012; Dai and Dixon, 2017; Cuesta et al., 2018; Dai and Dixon, 2019; Weiss et al., 2019), the analysis and synthesis of vocal unisons (Cuesta et al., 2018; Cuesta et al., 2019; Chandna et al., 2020), multi-pitch estimation of polyphonic vocal performances (Su et al., 2016; Schramm and Benetos, 2017; Cuesta et al., 2020), automatic transcription (McLeod et al., 2017), or the separation of voices (Gover and Depalle, 2020; Petermann et al., 2020; Sarkar et al., 2020) have been published recently.

Most of these studies focus on vocal ensembles with one singer per voice part. More specifically, the majority of them examine vocal quartets, i.e., four singers of different vocal ranges. However, we find a large amount of amateur and professional vocal ensembles with multiple singers per part singing in unison, i.e., choirs. While quartet and choral singing are conceptually very similar, they differ in the amount of singers within one section. According to Ternström (1991), while a single performer produces tones of well-defined properties, e.g., pitch, loudness, and timbre, an ensemble of performers, i.e., multiple singers per part, will generate sounds characterized by statistical distributions of these properties. Hence, we need to take this characteristic into consideration when working with choir recordings that include unisons instead of single singers.

In this study we evaluate source separation algorithms for separating the individual parts which have unison singing. In particular, we study various types of recently proposed deep learning based source separation algorithms while keeping an eye on the limited data available for training and evaluation. We also extend the work done in (Chandna et al., 2020), which proposes a model to synthesize a single singer signal from a unison input and vice-versa. The model proposed extracts linguistic content from an input unison signal and uses this to along with the F0 to synthesize a voice signal. To facilitate this synthesis from a unison track separated from a choral mixture, we study the pitch characteristics of a unison signal extracted by the source separation process. Using this study, we propose an application framework that uses the separated unison signals to synthesize a single singer with the unison linguistic and melodic content. Then, this solo singer signal can be pitch-transposed, converted back to a unison, and remixed with other voice parts to generate a synthetic multi-singer choral recording. Finally, we propose a framework for separation, analysis, and synthesis of SATB ensembles, depicted in **Figure 7**.

The rest of the paper is structured as follows: **Section 2** reviews the most relevant literature related to source separation, analysis, and synthesis of four-part vocal ensembles. **Section 3** presents the proposed methodology for this paper while **Section 4** discusses the experiments we carried out to evaluate the various parts of the methodology, including a perceptual evaluation in **Section 4.6**. Proposed applications of the analysis are discussed in **Section 5**. Finally, conclusions drawn from our study are discussed in **Section 6**.

2 RELATED WORK

2.1 Four-Part Singing Ensembles

A musical ensemble of singers singing simultaneously is commonly known as a choir. Choral music is a tradition that has been practiced throughout society from the medieval ages to modern times, involving diverse groups of singers of various capabilities and with different vocal ranges. As such, choral singing is a social activity that can be performed in various arrangements with or without instrumental accompaniment, the latter referred to as *a cappella* singing. The earliest form of ensemble choral singing can be traced back to the Gregorian chants of the 4th century, which involved multiple singers singing the same content simultaneously, in unison.

Such compositions are widely practiced today in dedicated conservatories across the world. Being a social activity, one of the most popular formats of choral singing makes use of the distinct male and female vocal range, with female singers capable of singing high pitches arranged in choir parts known as *soprano* and *alto*, while male singers are consigned to *tenor* and *bass* parts. The soprano part is typically for singers comfortable in the 260–880 Hz vocal range. For the alto section, the associated range is 190–660 Hz. Singers comfortable in the lower ranges, 145Hz–440 Hz and 90–290 Hz, are generally assigned the tenor and bass voices, respectively (Scirea and Brown, 2015). An SATB choir may also have just four singers, one singing each of the parts, resulting in a quartet arrangement. It is also common to have multiple singers singing in unison in each of the parts, resulting in even more pronounced *choral* effect.

We denote the voice signal of a singer in the soprano part as x_S^j, where $j = 1, \ldots, J$, with J being the number of singers singing in unison in the soprano voice. The signal for the unison of sopranos, x_S^U, is a linear mixture of the individual singers:

$$x_S^U = \sum_{j=1}^{J} a_S^j x_S^j \tag{1}$$

where a_S^j represents the gain of the individual singers in the unison signal. This gain depends on the position of the individual singer with respect to the microphone used for recording and on the voice's loudness. Similarly, the individual voice signals of the singers in the alto, tenor and bass voices are denoted as x_A^j, x_T^j and x_B^j, respectively. The unison signals for the respective parts are denoted by x_A^U, x_T^U and x_B^U. The sum of the unison signal gives us the choral mixture signal,

$$x_C = b_S^U x_S^U + b_A^U x_A^U + b_T^U x_T^U + b_B^U x_B^U \tag{2}$$

where b_S^U, b_A^U, b_T^U and b_B^U represent the gains of the soprano, alto, tenor and bass unison signals in the mixture signal. In Western choral music, voices are commonly written to harmonize with each other, which combined with voices having relatively similar timbres, leads to a high number of overlapping harmonics in the resulting mixture.

Focusing on Western choral singing, we aim at separating the four voices from an audio recording of a choir mixture. Our goal is to separate four unisons, and not each voice in the ensemble:

TABLE 1 | Summary of the multitrack datasets of ensemble singing we use in this project. The reported durations refer to the total recording duration, not considering multiple stems per recording.

Dataset	Voices and nb. of singers	Multitrack	Duration (hh:mm:ss)	Music material	Annotations
Choral Singing Dataset	4S-4A-4T-4B	Yes	00:07:14	3 songs	F0, notes, MIDI
Dagstuhl ChoirSet	2S-2A-4T-5B	Yes	00:55:30	2 songs exercises	F0 (automatic),score, beats
ESMUC Choral Dataset	5S-3A-3T-2B	Yes	00:21:08	3 songs exercises	F0, notes
Bach Chorales	SATB	Yes	00:58:20	26 songs	F0 (automatic),MIDI

$$x_C \rightarrow \hat{x}_S^U, \hat{x}_A^U, \hat{x}_T^U, \hat{x}_B^U \tag{3}$$

where $\hat{x}_v^U, v \in \{S, A, T, B\}$ denotes each separated unison stem. The unison separation process enables several subsequent applications such as the analysis of the unison signal, or the synthesis of a single voice with pitch and lyrical content resembling the one in the original unison, among others. We explore some of these further applications in **Section 5**.

2.2 Vocal Ensembles Datasets

The data-driven models for source separation that we aim to adapt in this study require large amounts of data for learning. *MUSDB18* (Rafii et al., 2017) dataset, commonly used for training and evaluating source separation algorithms for the case of musical source separation, contains a collection of 150 multitrack recordings with their isolated drums, bass, vocals, and "others" stems. The dataset contains individual stems for vocals, drums, bass, and other musical instruments for each of the tracks. These stems were recorded and mixed in professional studios with each instrument tracked individually.

Recording such a database of choral singing is a technically challenging task since a typical choir setting requires multiple singers to sing simultaneously, and recording an individual singer within the ensemble requires highly directional microphones which reduce leakage from other singers. Consequently, there has been a scarcity of datasets for research on vocal music. However, in the last years, great effort was brought towards releasing curated multitrack datasets of ensemble singing. In this work, we select some of these multitrack datasets to train and evaluate source separation methods and for F0 analysis in polyphonic vocal music. **Table 1** provides a summary of the main features of the datasets we consider, and we describe some additional details about them such as recording conditions and how they affect the inter-microphone bleeding in the following.

Choral Singing Dataset (CSD) (Cuesta et al., 2018) is a publicly available multitrack dataset of Western choral music. It comprises recordings of three songs in the SATB format, each in a different language (Catalan, Spanish, and Latin). The songs are all performed by a choir of 16 singers, organized in four per section. Each section of the choir was recorded independently, using microphones to isolate the voice of each singer. F0 trajectories for each recording as well as section-wise MIDI notes are available for each of the songs. The total audio duration is around 7 min, which makes it a small dataset in comparison to MUSDB18 dataset. Recordings from CSD contain some leakage from contiguous singers of the same section, which is less problematic than leakage from other sections in the context of this work.

Similarly, Dagstuhl ChoirSet (DCS) (Rosenzweig et al., 2020) is a dataset with ensemble singing recordings of two songs in Latin and Bulgarian languages. The dataset also contains a set of vocal exercises consisting of scales, cadences, and intonation exercises. Combinations of handheld dynamic microphones, headset microphones, and throat microphones, as well as a stereo pair microphones were used to record 13 singers, grouped into uneven SATB sections. All singers were recorded simultaneously, leading to slightly higher leakage in the individual tracks than the CSD, including some inter-section leakage. The dataset contains annotations for beats, synchronized score representations, and automatically extracted F0 contours. The total audio duration is around 55 min.

ESMUC Choir Dataset (ECD)[1] is a multitrack data collection that comprises three songs, in German and Latin, performed by a choir of 12 singers, unevenly distributed into SATB choir sections. The singers were recorded simultaneously using handheld dynamic microphones, and with a stereo pair microphone capturing the overall choir sound. Individual recordings from ECD contain high inter-singer and inter-section leakage, significantly higher than CSD and DCS. The dataset contains 20 min of audio, with manually corrected F0, and note annotations for each of the tracks.

Bach Chorales Dataset (BCD) is a commercial multitrack dataset used in the experiments in (McLeod et al., 2017; Schramm and Benetos, 2017). It consists of 26 songs performed by a SATB quartet, i.e., one singer per part. The total amount of audio of the BCD is around 58 min. Each singer in the quartet was recorded individually in a professional setup and there is not inter-singer leakage in the recordings. BCD contains each individual audio track and the mixture of the four voices. Besides, it provides MIDI files and automatically extracted F0 trajectories. However, due to the original commercial source of the recordings, this dataset is not publicly available for research purposes.

All audio files used in this project are resampled to a common sampling rate of 22 050 Hz.

2.3 Source Separation

Source separation, a task which consists in separating a mixture of signals into the constituent signals, has been well researched

[1]https://zenodo.org/record/5848990

across many different fields, including finance, medicine, geology, and audio and video signals. For audio signals, the task entails breaking a mixture composed of multiple signals (i.e., sources), down into its individual components. A significant amount of research in the domain has been dedicated to separating the voice signal from an audio mixture.

Over the last decade, research in source separation has shifted to data-driven machine learning techniques, particularly with the advent of deep learning. Several deep learning based models have been proposed for the related but distinct tasks of music source separation and asynchronous speech separation. The first of these tasks entails separating temporally and harmonically correlated sources in terms of distinct musical instruments, including the singing voice. The sources to be separated in this case have distinct spectral structures, e.g., the singing voice has a distinct formant structure, which is internally modeled by the source separation algorithms. Speech source separation refers to separating two asynchronous speech signals from distinct speakers. The distinction modeled by source separation algorithms pertains to temporal cues and the distinctive timbre of the speakers involved. Both of these tasks are closely related to our study, which consists of separating four sources with similar spectral structures, which can primarily be distinguished by their fundamental frequency (F0).

Source separation for synthetic choral data has been studied employing score-informed non-negative matrix factorization (NMF) and the Wave-U-Net architecture by Gover and Depalle (2020). The authors synthesized 371 Bach chorales using a commercial MIDI synthesizer named *FluidSynth*. This allowed for synthesized choral mixes and stems aligned with score information. Then, the Wave-U-Net (Stoller et al., 2018) architecture was adapted to accept temporal conditioning. The conditioning was applied both at the input and output layers, as well as the downsampled bottleneck layer. It was shown that the Wave-U-Net architecture outperformed the NMF-based baseline, even without the conditioning.

In the context of real-world SATB choir recordings, voice separation has been studied using transfer learning (Bugler et al., 2020) with a ChimeraNet model (Luo et al., 2017) pre-trained on the MUSDB and Slakh datasets (Manilow et al., 2019). This model was then fine tuned to segregate the male and female voices in SATB recordings in the DCS.

Our previous research on source separation for SATB recordings (Petermann et al., 2020) investigated the performance of the deep learning architectures mentioned above specifically on voice segregation for choir recordings. The first part of this work involved the assessment of state-of-the-art (SOTA) models given two use-cases: 1) using mixtures with exactly one singer per singing group, in a *quartet*, and 2) using up to four singers per SATB group, for a total of up to 16 singers. We also proposed an adapted version of the Conditioned U-Net by Meseguer-Brocal and Peeters (2019), leveraging the varied frequency ranges of the constituent parts of an SATB choir to segregate the voices. We conditioned the U-Net on the oracle F0 of the individual parts using a feature-wise linear modulation (*FiLM*) layer (Perez et al., 2018). This led to an increase in performance over the vanilla U-Net model.

2.4 Unison Analysis and Synthesis

We recently proposed a system for unison analysis and synthesis (Chandna et al., 2020). This system extracts the linguistic content from an input unison signal using a network trained *via* a student-teacher schema. A language independent linguistic content representation is extracted by using the intermediate layers of a SOTA voice conversion model (Qian et al., 2019). This representation is used to generate the harmonic and in-harmonic parts of the WORLD vocoder (Morise et al., 2016), which are used along with the F0 to synthesise the waveform pertaining to a single singer singing the unison signal. While our previously proposed model has proven to effectively model the linguistic content of the unison signal, extraction of the F0 from the unison remains a challenging task. To model this F0, we must analyze the pitch of the unison.

Previous studies have shown that listeners perceive unison performances to have a single pitch, even though this pitch is produced by multiple singers (Ternström, 1991). We would require large perceptual studies with enormous amounts of unison recordings to study in depth which is the pitch that a listener perceives when they hear a unison performance. In an early study by Ternström (1991), perceptual experiments were conducted with expert listeners to investigate, among other aspects, the preferred level of *pitch scatter* in unison vocal performances. Pitch scatter is defined in the original paper as the standard deviation over voices in the mean F0—the average F0 computed over the duration of each note of a song. The authors used synthesized stimuli with different levels of scatter, and found that while listeners tolerate up to 14 cents of scatter, the preferred level of pitch scatter for a unison ranges between 0 and 5 cents. These findings suggest that while slight deviations in pitch between singers are preferred, they should be small enough so that the overall sound is still perceived as a unique pitch. To extract this unique pitch from an input unison signal, we use monophonic F0 estimation, introduced in the following.

2.5 Monophonic F0 Estimation

Although unison performances are commonly considered monophonic signals, depending on the magnitude of pitch and timing deviations they contain, they can be more challenging for monophonic F0 trackers than single singer recordings. In this work, we assess the performance of two SOTA methods for monophonic F0 tracking on unison performances to see which one performs better. We hypothesize that a reliable F0 contour extracted from a unison signal approximates its melodic content, which can be used for further analysis and synthesis applications, as we detail it in the next sections. We consider a knowledge-based F0 tracker, pYIN, and a data-driven F0 tracker, CREPE.

pYIN (Mauch and Dixon, 2014) is a knowledge-based F0 tracking method based on the well-known YIN algorithm (de Cheveigné and Kawahara, 2002). YIN is a time-domain monophonic F0 tracker for speech and music signals based on the auto-correlation function (ACF). In this method, the authors propose a modified difference function based on the ACF, where they locate the dip corresponding to the period of the signal. pYIN was later introduced as a probabilistic version of YIN,

where multiple F0 candidates are computed together with their probabilities, and then one F0 value per analysis frame is selected through hidden Markov models (HMM) and Viterbi decoding.

CREPE (Convolutional Representation for Pitch Estimation) (Kim et al., 2018) is a data-driven F0 tracking algorithm based on a deep convolutional neural network that operates on the time-domain waveform. The output of the network is a 360-dimensional output vector, $\hat{y}$, and each of the 360 nodes is associated to a specific F0 value in *cents*, covering six octaves with a 20 cents resolution. The output F0 estimate is calculated as the average of the associated F0 values weighted by their corresponding values in the output $\hat{y}$. One F0 estimate is obtained for each analysis window of the input signal.

3 PROPOSED METHODOLOGY

This section introduces the methodology we follow. In the proposed framework, we experiment with three main tasks: first, source separation models to separate the individual unison signals from a choral mixture; second, approximating the perceived pitch of the unisons; and third, the analysis-synthesis framework to regenerate individual singing voice signals which can be remixed together.

3.1 Source Separation

We assess the performance of a set of SOTA models for both music and speech source separation applied to our target case (SATB choirs). We first select the Open-Unmix model (Stöter et al., 2019), which provided SOTA results on the musical source separation task over the 2018 Signal Separation Evaluation Campaign (Stöter et al., 2018) (SISEC). For asynchronous speech separation, we select the Conv-TasNet (Luo and Mesgarani, 2019), which has been shown to outperform the ideal time-frequency (TF) mask for the case of synchronous source separation. We note that while this algorithm has been adapted to the task of music source separation (Défossez et al., 2019; Samuel et al., 2020), we use the original variant specifically proposed for speech source separation for our study. This is done because and we want to compare models proposed for musical source separation with speech source separation when adapted to the task at hand. Choral singing is a mixture of multiple voices, which is the case addressed in speech source separation, but it also has a musical structure for which musical source separation might be more apt. By comparing state-of-the-art models in both domains, we can assess which models might be more suitable for adaptation to choral source separation.

Deep learning based algorithms have been proposed for both tasks over the last few years. While many of the models are based on a time-frequency representation like the spectrogram, recently proposed models have explored end-to-end waveform based separation (Lluís et al., 2019). It has been postulated that waveform based source separation algorithms require a larger amount of data for training than spectrogram based models. Given the limited availability of data in our case, we assess whether waveform based source separation models can perform as well as spectrogram based models for segregating the parts of an SATB choir. For this we compare the U-Net (Jansson et al., 2017) model for source separation with its waveform-based counterpart, operating directly on the waveform, Wave-U-Net (Stoller et al., 2018).

3.2 Modeling the Pitch of a Unison

Several aspects influence the pitch we perceive from a unison. Some of them are the relative loudness of each singer with respect to the others, the listener's position, or the frequency range. However, in this paper we follow Occam's Razor and use the hypothesis that the simplest possible function of the individual F0 of each singer, the mean, is thus the simplest possible representation of the perceived pitch in a unison performance. Therefore, to obtain a reference F0 contour to characterize the melodic content of the unison, we approximate the perceived pitch of a unison as the average along the F0s produced by each singer of the group. However, to account for potential differences in the energy of each contributing source, we consider a weighted average where each source has an associated weight (as we mentioned in **Section 2.1**, a_S^j denotes the weight of soprano singer *j*.) In practice, in our experiments we compute the weights as the normalized root mean square (RMS) of each source, which we can roughly associate to their volume. For the rest of this paper we refer to this weighted average as the approximated perceived pitch of the unison. In **Section 4.6**, we validate this approximation using perceptual experiments. The multitrack nature of our working dataset allows the calculation of the RMS for each recording individually, for which we use the RMS algorithm from `librosa` (McFee et al., 2015).

We compute the F0 weighted average on a frame basis considering only the active voices per frame. Due to timing deviations between singers, note transitions have a set of frames where not all voices are in sync, i.e., one singer starts or ends a note slightly before or after another singer. To account for such passages, at each frame of analysis, we only use the F0s from active singers for the average calculation.

3.3 Synthesis and Remixing

Once we have extracted the individual unison stems from a choral mixture and modeled the perceived pitch of the unison signal, we can use the Unison-to-Solo (UTS) methodology we proposed in (Chandna et al., 2020) to synthesize a single singing voice pertaining to the melodic and linguistic content of the signal. The framework for the UTS methodology is illustrated in **Figure 1**: it extracts the linguistic content, the melody, and the timbre from an input unison signal to generate vocoder features for synthesizing a single singing voice.

From this synthetic single singing voice representation of the individual parts of the choral mixture, we employ the Solo-to-Unison (STU) methodology, also proposed in (Chandna et al., 2020), to generate a unison signal. As illustrated in **Figure 2**, the STU methodology uses the analysis-synthesis framework that enables the addition of pitch and timing deviations, along with singer timbre variations using voice conversion models. Since we are synthesising the unison signal, the methodology can be used to apply transformations such as pitch shifting, time stretching and other score based transformations including increasing the

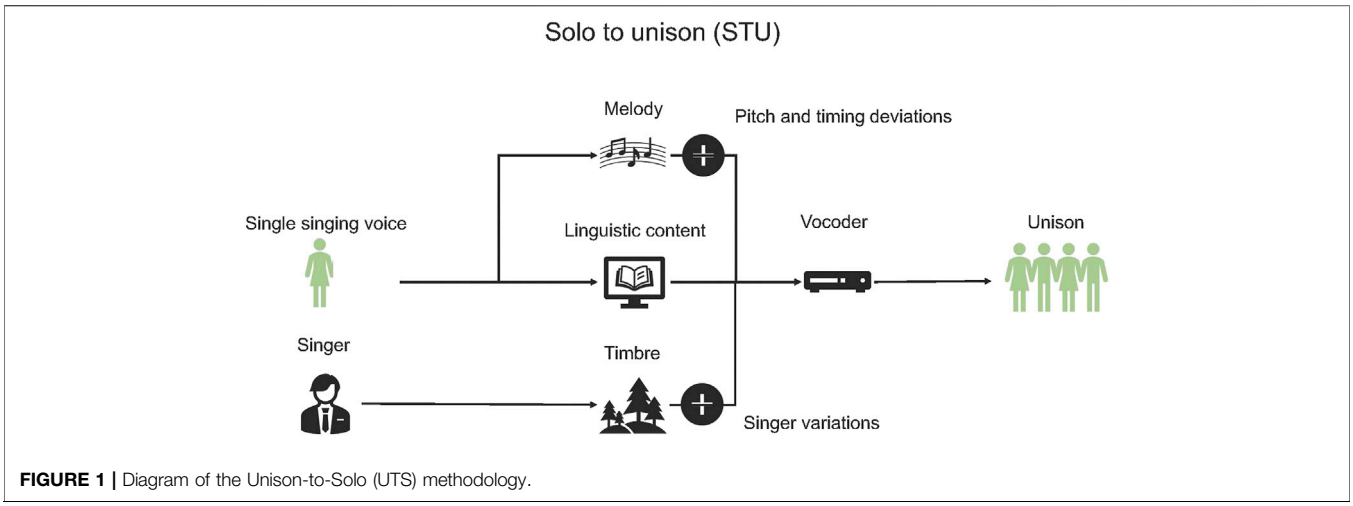

FIGURE 1 | Diagram of the Unison-to-Solo (UTS) methodology.

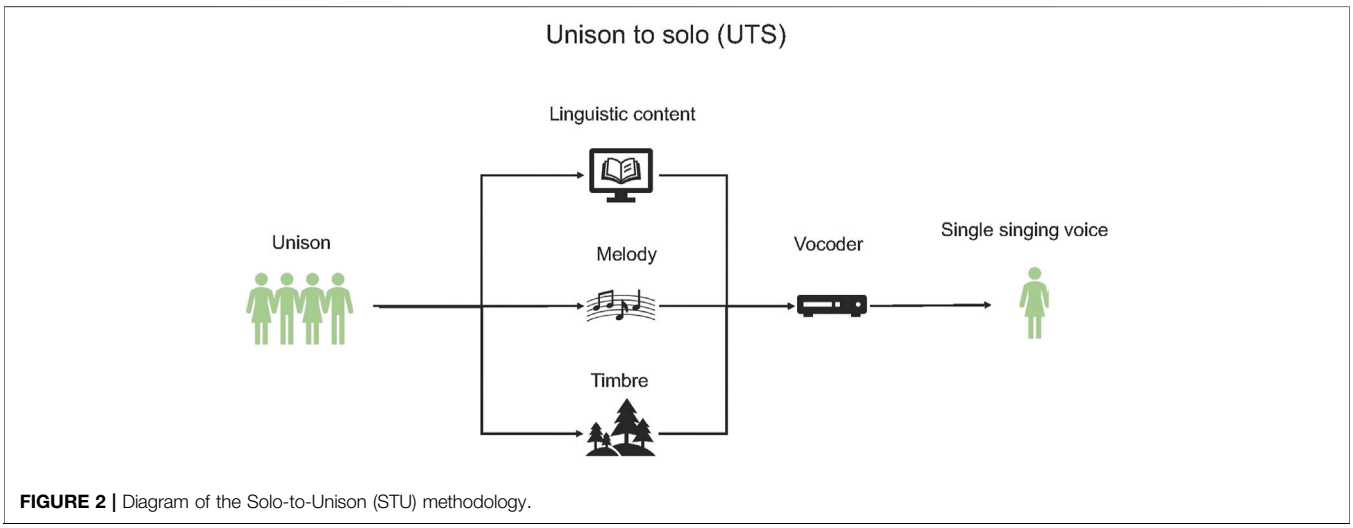

FIGURE 2 | Diagram of the Solo-to-Unison (STU) methodology.

number of singers singing the in unison. These synthesised unison signals can then be remixed using the framework illustrated in **Figure 7** to support the editing of choral recordings for more accessible and targeted singing practice.

4 EXPERIMENTS

In this work, we perform two main experiments to assess the type of source separation model and the data required for the task of segregating the parts of an SATB choir. We train and evaluate four models for the task, as listed in (**Table 2**). We compare the spectrogram-based U-Net model (*UNet*) (Jansson et al., 2017) with its waveform-based counterpart, Wave-U-Net (*WaveUNet*) (Stoller et al., 2018). We note that while the original U-Net proposed for singing voice separation uses two separate networks to predict TF masks for the vocal and accompaniment stems, we use a single network to predict four masks to be applied for each of the four parts. This allows us to assess if waveform-based models can perform as good as spectrogram based models for the task, given the limited data used for training. We also compare the Open-Unmix (*UMX*) model with the Conv-TasNet (*ConvTasNet*) model. While the former represents the SOTA for music source separation, the later represents the SOTA for speech source separation. *UMX* uses four sub-networks, one for each of the parts.

As discussed in **Section 2.2**, there is a limited amount of data available for training and evaluation of the models. While CSD, DCS, and ECD contain individual tracks for multiple singers per part of an SATB choir arrangement, there is significant overlap in the songs present in CSD and DCS, while ECD has significant inter-singer leakage in the tracks. BCD has the cleanest data amongst the datasets, but only has quartet recordings with a single singer per part. We see that, while it is easier to obtain clean data for quartets, real-world choir recordings often have multiple singers per part. As such, we need to assess if quartet based data is sufficient for training a deep learning based source separation model for the task of segregating parts with multiple singers.

TABLE 2 | The models we adapt for voice segregation in SATB choirs.

Model	Input	Originally proposed for
U-Net Jansson et al. (2017)	Spectrogram	Music source separation
Wave-U-Net Stoller et al. (2018)	Waveform	Music source separation
Open-Unmix Stöter et al. (2019)	Spectrogram	Music source separation
ConvTasNet Luo and Mesgarani, (2018)	Waveform	Speech separation

TABLE 3 | Summary of the models trained for each experiment.

Models	Name	Trained on	Evaluated on	Experiment
UMX UNet WaveUNet ConvTasNet	$modelname_C$	CSD	ECD_{clean}	Incremental training
UMX UNet WaveUNet ConvTasNet	$modelname_{CB}$	CSD, BCD	ECD_{clean}	Incremental training
UMX UNet WaveUNet ConvTasNet	$modelname_{CBDE}$	CSD*, BCD, ECD_{clean}**, DCS	Die Himmel, CSD_{Q1}	Full training

ECD_{clean} refers to ECD after the leakage removal. CSD denotes CSD excluding the quartet formed by the combination of the first singer of each section (CSD_{Q1}). ECD_{clean}** represents all clean ECD excluding the three takes of the song* Die Himmel.

Table 3 summarizes the datasets and data partitions we consider for each experiment and model, and the experiments we conduct are described as follows.

4.1 Experiment 1: Incremental Training

In this first experiment, we first train the four models on CSD, with all possible combinations of singers within a song constrained by two cases: *Quartet* case, wherein the number of singers per part of the input is limited to one (quartet input), and *Choir* case, wherein we allow all possible combinations of singers from one to four singers per part of a song. We denote the models trained with the CSD as $modelname_C$, where *modelname* in {*UNet*, *WaveUNet*, *UMX* and *ConvTasNet*}. We then augment the data from CSD with quartet data from BCD. As the training data is incrementally increased using the quartets from BCD, we term this experiment "Incremental training." These models are termed as $modelname_{CB}$, using a nomenclature similar to the one previously mentioned.

4.2 Experiment 2: ESMUC Choir Dataset Leakage Removal

To evaluate the models from the incremental training experiment, we need to consider a dataset which is not part of the training set to avoid overlap between the songs and singers used for training and evaluation. While ECD fits this criteria, there is a significant amount of inter-section leakage, i.e., presence of alto, tenor, and bass tracks in the soprano track within the tracks of the dataset. To alleviate this problem, we use a cleaning procedure using the models trained in the incremental training to clean the individual tracks of ECD before mixing them together to create the evaluation set.

Heuristically, in the incremental training experiment we find UMX_{CB} model to perform the best amongst those trained. Hence, we select this model to clean the individual stems of ECD by passing them through the corresponding sub-network of UMX_{CB}. We pass the soprano stem of each song from ECD through the trained Open-Unmix model for the soprano part to clean the soprano stem. This procedure is repeated for alto, tenor, and bass stems. These cleaned stems are then mixed together to form the evaluation set for the first experiment. We conduct a perceptual evaluation of this cleaning procedure, presented in **Section 4.6**. This allows filtering out interference from parts of the choir that do not belong to the target part. Then, we consider the cleaned ECD, denoted as ECD_{clean}, for evaluation of the models from the incremental training experiment.

4.3 Experiment 3: Full Training

For the full training experiment, we augment the training set with ECD and DCS and train the models on the both cases of data. These models are termed $modelname_{CBDE}$ and we evaluate them on three takes of the song *Die Himmel* from ECD, which were excluded from the training set (see **Table 3**). Excluding these recordings allows for the evaluation of the models on an unseen song. To further evaluate the models for unseen singers, we exclude the first singer from each of the parts from CSD and use the quartet of singer one (composed of the first singer of each section and denoted CSD_{Q1}) songs from the CSD for evaluation.

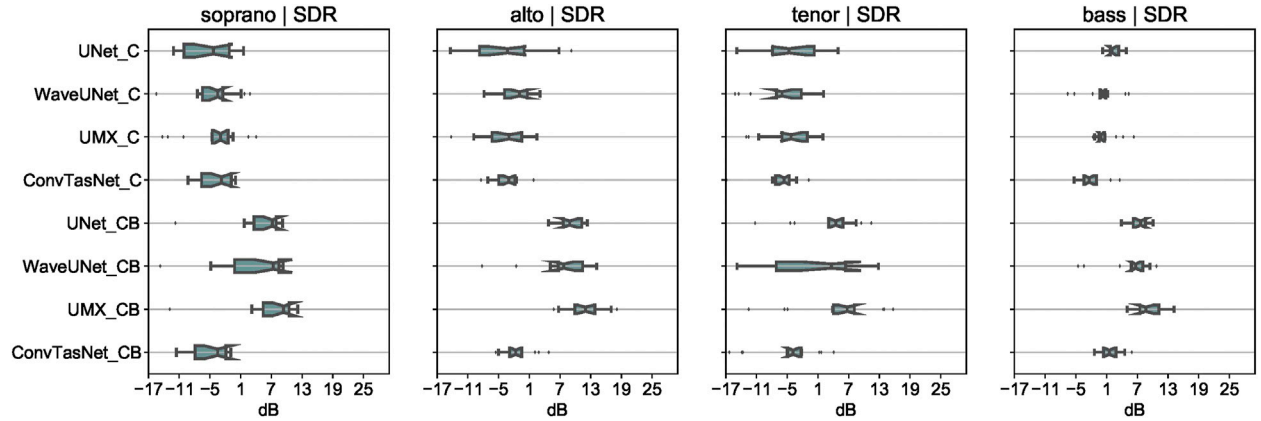

FIGURE 3 | Boxplot SDR results for the *Choir* case models trained on CSD ($modelname_C$) and CSD + BCD ($modelname_{CB}$).

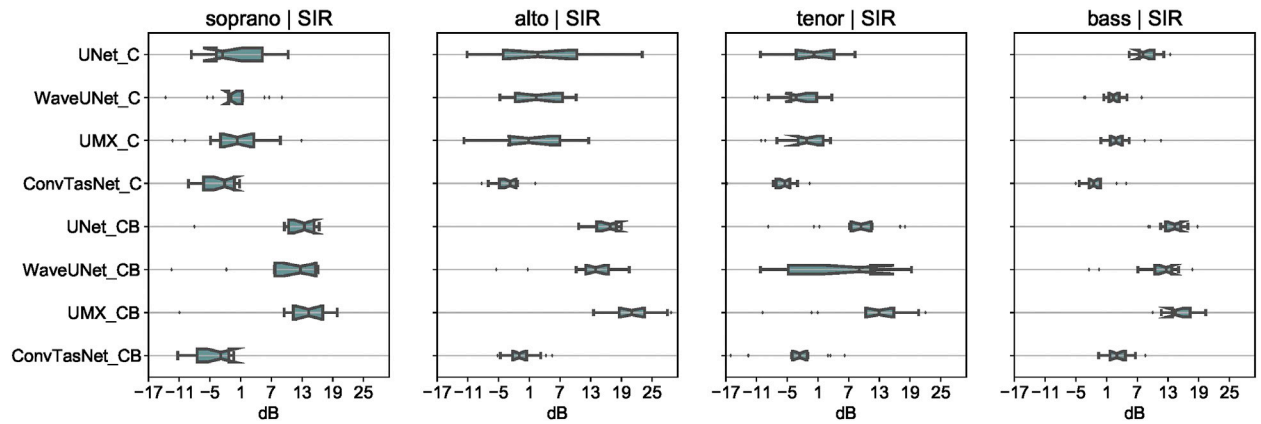

FIGURE 4 | Boxplot SIR results for the *Choir* case models trained on CSD ($modelname_C$) and CSD + BCD ($modelname_{CB}$).

4.4 Evaluation Results

This section presents the evaluation metrics we consider to measure the performance of our models, followed by the results we obtain for the described experiments.

4.4.1 Evaluation Metrics

We evaluate our models with the *BSS eval set* of objective metrics (Vincent et al., 2006). In particular, we consider the *bss_eval_sources* set of metrics, pertaining to single-channel source signals. The three metrics we select are the Source-to-Interferences Ratio (SIR), which measures the amount of interference in an estimated source from the other sources in the mixture, the Sources-to-Artifacts Ratio (SAR), which measures the artifacts introduced by the source separation process, and the Source-to-Distortion Ratio (SDR), which provides an estimate of the overall quality of the separation, compared to the ground truth.

4.4.2 Results Incremental Training

The results of the incremental training experiment are depicted in **Figures 3**–**5**. The evaluation shown is for models trained using *Choir* case data, with all possible combinations of singers from CSD (except the first singer of each section), augmented with quartet data from BCD. The evaluation set contains all the songs from ECD_{clean} dataset with all singers in the mixture. **Figure 3** shows the SDR metrics for the four models trained with CSD data, denoted as $modelname_C$, and with CSD data augmented with quartet data from the BCD, denoted as $modelname_{CB}$. **Figures 4**, **5** show the SIR and SAR metrics for the same models, respectively.

We note that the performance on all three evaluation metrics of the Wave-U-Net model is comparable to that of the U-Net model. As the two models are similar in architecture, this allows us to conclude that waveform-based models for source separation are just as effective as their spectrogram-based counterparts for the task of segregating the different parts of an SATB choir recording. We also note that the Open-Unmix model outperforms the other models evaluated in the study while the Conv-TasNet model under-performs the rest. This suggests that music source separation algorithms are more suited for the choral singing domain than models proposed for asynchronous speech

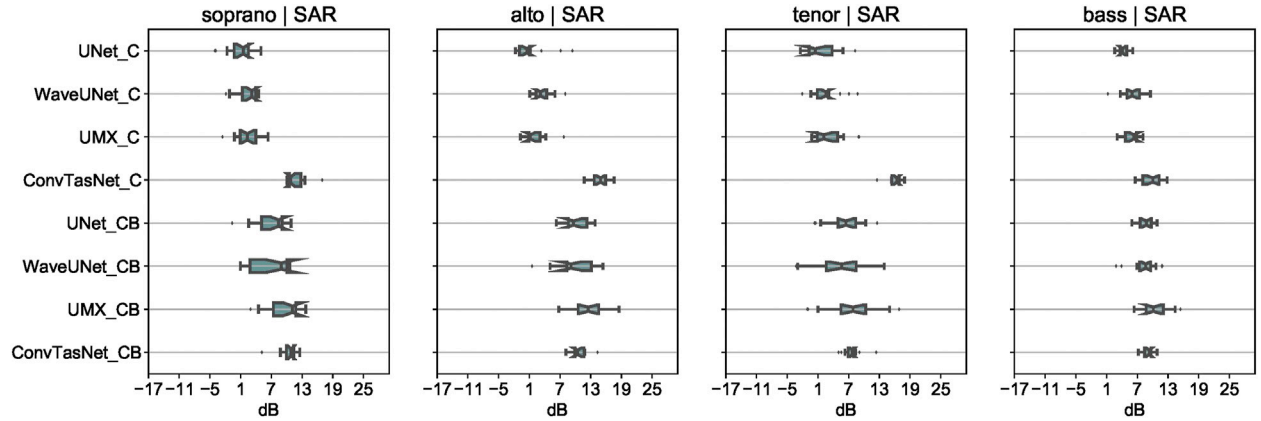

FIGURE 5 | Boxplot SAR results for the *Choir* case models trained on CSD (*modelname*$_C$) and CSD + BCD (*modelname*$_{CB}$).

TABLE 4 | SDR results from *modelname_CBDE* evaluated on ECD songs (13 singers with all the singers in the mix), showing the training with data from the quartet case (left) and choir case (right).

Model	Quartet case: SDR (dB)					Choir case: SDR (dB)				
	Soprano	Alto	Tenor	Bass	Avg	Soprano	Alto	Tenor	Bass	Avg
Wave-U-Net	−3.92 ± 4.69	8.91 ± 1.37	−11.43 ± 2.10	5.46 ± 0.98	−0.24	−4.45 ± 4.61	8.50 ± 1.14	−12.81 ± 1.41	4.61 ± 0.94	−1.05
U-Net	1.67 ± 3.78	9.80 ± 0.96	−7.13 ± 2.07	7.20 ± 0.91	2.88	2.06 ± 3.37	10.40 ± 1.19	−4.79 ± 2.04	7.60 ± 0.92	3.81
Open-Unmix	−1.21 ± 5.88	10.70 ± 3.59	−8.20 ± 4.96	7.42 ± 2.01	2.17	−0.14 ± 7.06	10.82 ± 3.89	−7.09 ± 4.18	8.02 ± 2.03	2.90
ConvTasNet	−7.29 ± 2.18	1.05 ± 1.54	−15.78 ± 1.03	4.98 ± 0.29	−4.26	−7.14 ± 2.28	−0.57 ± 2.03	−16.45 ± 1.52	3.70 ± 0.74	−5.11

separation. However, we observe that the Conv-TasNet model shows better results on the SAR metric than the other models, especially when trained with CSD (*ConvTasNet*$_C$). The Open-Unmix model uses 4 separate networks while the U-Net and Wave-U-Net models use a single network with 4 output masks. We observe that while the Open-Unmix model outperforms the other two models, this improvement can be attributed to a more complex network architecture, rather than the use of separate networks for each of the sources.

We further note that augmenting the multiple data combinations of CSD with the quartet data of BCD leads to a significant improvement of results across all the models and parts. This suggests that we can use quartet based data in the future to further train the source separation models for the choral singing case. We note that pitch shifting can also be used for data augmentation, as in (Cuesta et al., 2020) for multi-pitch estimation. However, we did not experiment with this technique in this study. Finally, we also note that the performance of all the models worsens a little for the tenor part, as compared to the other parts, whereas the bass part is easily separated. We assume this difference is because of the overlap in the melody range for the tenor and alto part. As the F0 of the parts is the major distinguishing feature between them, we believe this overlap in range leads to confusion between the two parts. We further investigate this in **Section 4.5.3**.

4.4.3 Results Full Training

Table 4 shows the SDR metric for models trained on data from all 4 datasets. The results are shown for models trained with both cases, i.e., *Quartet* case, which is restricted to one singer part, and *Choir* case, which uses combinations of multiple singers for each part.

For this evaluation, we consider 3 takes of a single song from ECD$_{clean}$, which were excluded from the training set, and the quartets formed by mixing the stems of the singers excluded from training from CSD, CSD$_{Q1}$ (see **Table 3**). This allows us to test the performance of the models for unseen songs and unseen singers. We observe that there is an improvement in performance in the U-Net and Open-Unmix models, when trained with *Choir* case data, while the Wave-U-Net and Conv-TasNet are slightly worse when multiple singers are used for training the models.

Table 5 shows the SIR results for the same models, while **Table 6** shows the SAR results. We again observe that the performance of all models is lower for the tenor part than it is for the other parts. We believe this is due to the confusion between the overlapping pitch ranges of the soprano and alto parts, and the tenor and bass parts. We also note that the performance of the U-Net and Open-Unmix models improves when considering *Choir* case data for training, but does not consistently improve for Wave-U-Net and Conv-TasNet. However, from the incremental training experiment, we observe that the model training can be improved by

TABLE 5 | SIR results from *modelname_CBDE* evaluated on ECD songs (13 singers with all the singers in the mix), showing the training with data from the quartet case (left) and choir case (right).

Model	Quartet case: SIR (dB)					Choir case: SIR (dB)				
	Soprano	Alto	Tenor	Bass	Avg	Soprano	Alto	Tenor	Bass	Avg
Wave-U-Net	−0.05 ± 5.99	15.07 ± 1.54	−7.54 ± 2.77	13.05 ± 1.34	5.13	−0.68 ± 5.88	15.19 ± 1.49	−9.45 ± 1.83	11.63 ± 1.06	4.17
U-Net	4.82 ± 4.45	17.52 ± 1.38	−3.62 ± 2.34	13.44 ± 1.16	8.04	9.46 ± 4.56	17.57 ± 1.59	−0.99 ± 2.45	13.47 ± 0.78	9.87
Open-Unmix	5.54 ± 7.19	18.91 ± 2.36	−2.34 ± 4.80	11.23 ± 3.09	8.33	7.03 ± 8.72	19.03 ± 3.21	−1.75 ± 4.75	13.21 ± 2.21	9.38
ConvTasNet	−6.54 ± 2.43	2.46 ± 1.26	−15.39 ± 1.03	8.56 ± 0.50	−2.72	−6.17 ± 2.66	1.13 ± 1.68	−15.56 ± 1.59	8.17 ± 0.58	−3.10

TABLE 6 | SAR results from *modelname_CBDE* evaluated on ECD songs (13 singers with all the singers in the mix), showing the training with data from the quartet case (left) and choir case (right).

Model	Quartet case: SAR (dB)					Choir case: SAR (dB)				
	Soprano	Alto	Tenor	Bass	Avg	Soprano	Alto	Tenor	Bass	Avg
Wave-U-Net	2.63 ± 0.28	10.26 ± 1.26	−0.75 ± 0.82	6.51 ± 0.83	4.66	2.45 ± 0.32	9.69 ± 1.00	−0.13 ± 0.70	5.88 ± 0.96	4.47
U-Net	1.17 ± 2.16	10.69 ± 0.87	0.76 ± 0.14	8.58 ± 0.79	5.30	3.73 ± 2.35	11.41 ± 1.06	1.27 ± 0.32	9.10 ± 0.95	6.37
Open-Unmix	2.12 ± 2.94	11.52 ± 3.78	−1.99 ± 1.69	10.34 ± 0.77	5.49	3.18 ± 3.70	11.61 ± 3.94	−1.03 ± 0.73	9.82 ± 1.79	5.89
ConvTasNet	8.55 ± 1.55	8.76 ± 2.06	10.96 ± 2.43	8.05 ± 0.06	9.08	7.55 ± 1.96	7.02 ± 2.32	6.88 ± 1.85	6.25 ± 0.79	6.92

augmenting the training data with multiple singers per part along with quartets.

4.5 F0 Modeling Experiments

In this section, we present the evaluation of the monophonic F0 trackers (see **Section 2.5**) in two different scenarios. We first consider a reference unison created as the mixture of individual singers from the same vocal part, and evaluate the prediction of the F0 trackers on this performance against the approximated reference melodic content (cf. **Section 3.2**).

With the results of the first part, we select the F0 tracker that obtains the best performance on the reference unisons for the second evaluation. In this case, we replace the reference unisons with the unisons extracted by the source separation methods described above. Such process enables a fully automated extraction of the melodic content of one choir section given a choir mixture as input.

4.5.1 Datasets

For the F0 modeling experiments we select recordings from CSD and recordings from ECD. As mentioned in **Section 2.2**, these datasets have manually corrected annotations for the F0 of each stem for all the songs. While this manual correction leads to an F0, that is, closer to the approximated perceived pitch, we acknowledge that the annotator's perception might introduce some bias in the F0 labels.

The number of singers per section is different in both datasets: four singers per part in CSD and two to five in ECD. For each song in these datasets, we create the unison mixture for each SATB section. We evaluate a total of 14 unison performances.

4.5.2 Results

This section presents the results of the F0 modeling experiment. **Table 7** summarizes the results of the evaluation with two different monophonic F0 trackers to estimate the pitch of unison performances. While the Raw Pitch Accuracy (RPA) only measures the proportion of voiced frames for which the F0 estimation is correct (within half semitone), the Overall Accuracy (OA) metric also considers the voicing detection: it represents the proportion of frames which are correct both in terms of pitch and voicing. We need to consider both aspects (pitch and voicing) for our task. Hence, we select CREPE for our further experiments, as it scores a higher OA on average. CREPE outputs a voicing confidence value along with the F0 predictions. Consequently, we apply a threshold on the confidence to decide whether a frame is voiced or unvoiced. We calculate this threshold as the average of the thresholds that maximize the OA on an external set of recordings. In particular, we select four monophonic recordings from Cantoría dataset (Cuesta, 2022)[2] and 8 monophonic recordings from DCS. We have manually-corrected F0 annotations for all 12 recordings, which we consider as reference in the evaluation. With this process, we obtain an optimal threshold of 0.7 that we employ in the following experiments.

TABLE 7 | Results of the evaluation of monophonic F0 trackers on the ground truth unison mixtures. We use the reference melodic content as described in **Section 3.2**. The results are averaged across songs and voice parts. Standard deviation is shown in parentheses.

F0 tracker	Raw pitch accuracy (%)	Overall accuracy (%)
pYIN	76.8 (18.3)	68.7 (16.9)
CREPE	70.0 (18.3)	71.7 (14.7)

Figure 6 displays the results of the second experiment, i.e., the evaluation of CREPE applied to the unison signals extracted by

[2]Cantoría dataset comprises multi-track recordings of an SATB quartet and is available here: https://zenodo.org/record/5851069

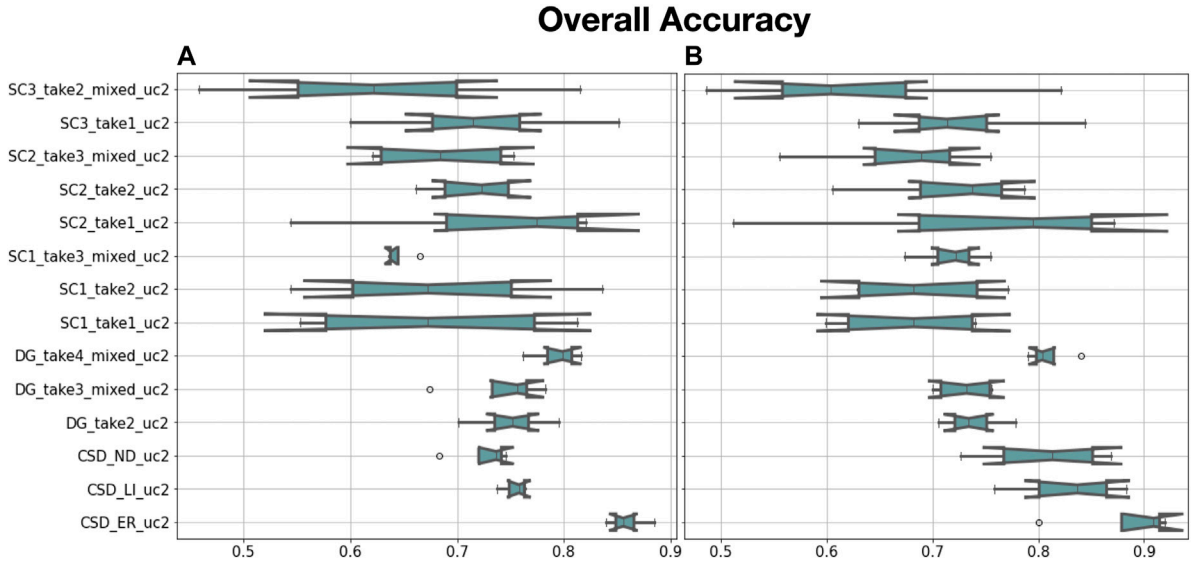

FIGURE 6 | Overall accuracy obtained with CREPE on the output of Open-Unmix. **(A)** Results for *Quartet* case (model trained with one singer per part). **(B)** Results for *Choir* case (model trained with multiple singers per part). Song labels beginning with SC and SG refer to songs from ECD while song labels beginning with CSD refer to songs from CSD.

Open-Unmix, grouped by song. We present the results for *Quartet* case (**Figure 6A**, model trained with one singer per part), and for *Choir* case (**Figure 6B**, model trained with multiple singers per part). Let us recall that the inputs are recordings of a choir, i.e., multiple singers per part.

We obtain a higher OA on CSD songs that were used for training the source separation models than on ECD songs that were excluded from training. These results show that CREPE can be used to effectively approximate the perceived pitch of the unison signal extracted from the source separation models, even for songs that are not used for training the source separation models.

4.5.3 Cross-Section Evaluation

A qualitative inspection of a small subset of the source separation results suggested some confusion between contiguous voice parts. To assess this phenomenon quantitatively, we conduct a brief cross-section evaluation, i.e., we repeat the F0 estimation evaluation using the F0 from another voice section as reference. These evaluations reveal some confusion especially between alto and tenor (up to a RPA of 56%), and soprano and alto (up to 58%). In both scenarios, the confusion happens from the higher pitch voice to the lower one, i.e., the algorithm extracts the alto voice instead of the soprano, and not vice-versa. Furthermore, we found that all confusions with an RPA above 40% belong to songs from ECD collection and not CSD. These results confirm the limitations we detect in the source separation numerical evaluation, where alto and tenor voices obtain worse performances.

4.6 Perceptual Evaluation

We use subjective listening tests to conduct a perceptual evaluation of the results of the proposed methods. We evaluate three outcomes of this work: the source separation, the solo singing synthesis, and the cleaning process of ECD. For the source separation, we focus on the following criteria: audio quality, melodic content similarity, and separation quality, i.e., level of bleeding from other voices. In terms of the solo singing synthesis, we aim to validate our approximation to the F0 of a unison as the weighted average of the individual F0 contributions. Hence, we focus on the melodic similarity between the synthesis and the original unison. Finally, we evaluate the process of cleaning the stems from ECD, in an attempt to remove bleed from other singers.

4.6.1 Perceptual Evaluation Methodology

For the subjective listening tests related to separation, the participants were provided three examples from each of the soprano, alto, tenor, and bass parts separated using the Open-Unmix model trained with the three variants of the datasets. The participants were also provided the mixture as a reference and the ground truth part as an anchor and were asked to rate each of the examples in terms of isolation of the target part from the rest of the parts on a scale of [0–5]. For the questions related to quality, similar examples were provided to the participants, but they were asked to rate these examples in terms of the audio quality of the output, taking into account the artefacts and other distortions that may have been added during the separation process. For these listening tests, we used the separated outputs from Open-Unmix model, i.e., UMX_C, UMX_{CB}, and UMX_{CBDE}.

To evaluate the adherence to the melody of the weighted average F0 or the approximated perceived used to model the perceived single pitch of the unison, participants were asked to rate the similarity of the melody of a single singing voice synthesized with the weighted average of the individual voices

TABLE 8 | Normalized MOS results of the subjective listening tests pertaining to the separation and quality criteria.

	Soprano		Alto	
	Quality	**Separation**	**Quality**	**Separation**
UMX_C	1.01 ± 0.81	3.09 ± 1.42	2.38 ± 0.65	4.32 ± 1.19
UMX_CB	0.74 ± 0.85	2.53 ± 1.59	2.34 ± 0.76	4.02 ± 1.63
UMX_CBDE	0.78 ± 0.50	2.58 ± 1.36	2.36 ± 0.83	4.48 ± 1.12
	Tenor		**Bass**	
	Quality	**Separation**	**Quality**	**Separation**
UMX_C	2.44 ± 0.91	4.73 ± 2.25	2.36 ± 1.15	3.28 ± 0.76
UMX_CB	1.50 ± 0.88	4.89 ± 2.73	2.29 ± 1.59	2.20 ± 1.17
UMX_CBDE	1.97 ± 0.93	4.92 ± 4.49	2.05 ± 0.97	2.61 ± 0.78

in the unison given a reference of a unison recording. This synthesis was done using the UTS system proposed in (Chandna et al., 2020). In addition, participants were asked to rate the unison part separated from a mixture for the same criteria on a scale of [0–5]. One question was provided for each of the SATB parts. Finally, to evaluate the cleaning of ECD individual stems, we asked participants to rate the original and cleaned versions of the parts given an example of the same part from the clean CSD dataset.

4.6.2 Perceptual Evaluation Results

The results of the subjective listening tests are shown in **Tables 8**, **9**. There were 12 participants in our evaluation, most of whom had prior musical training. **Table 8** shows the results of the listening tests on the quality and separation criteria. The rating for each of the questions was normalized by dividing by the rating given for the reference part in the question. It can be observed that the separation for the alto and tenor parts was rated higher than the soprano and bass parts, contrary to the observations in the objective evaluation. It can also be observed that the rating for separation for the model trained with just CSD was higher than the models trained with data augmentation. This is also contrary to the results of the objective evaluation. A similar trend is seen for the quality criteria, with the soprano part being rated lower than the alto, tenor and bass parts.

The results for listening test questions pertaining to the adherence to melody of the unison criteria and the cleaning process of the ECD dataset are shown in **Table 9**. We observe that the cleaned signal was rated higher than the original signal for the soprano and alto parts while the opposite was true for the tenor and bass parts. We believe that the lower ratings given to the cleaned signal for the later parts was due to the artefacts introduced to the signal during the separation process. We also observe that single singing voice signal synthesized using the weighted mean of the individual singers in the unison was rated higher than the separated part in terms of adherence to melody for the unison signal. This supports our hypothesis that the weighted mean can be used as an effective representation of the single pitch perceived while listening to a unison recording.

TABLE 9 | Normalized MOS results of the subjective listening tests pertaining to the cleaning of the ECD dataset and the adherance to the melody of the unison synthesis.

	Soprano	**Alto**	**Tenor**	**Bass**
Cleaned	4.23 ± 2.42	2.93 ± 1.28	3.73 ± 2.59	1.76 ± 0.89
Original	3.97 ± 1.66	2.20 ± 2.06	3.92 ± 2.70	2.43 ± 1.31
Separated	3.87 ± 1.23	2.07 ± 1.75	0.91 ± 0.88	1.53 ± 1.33
Weighted	4.68 ± 1.42	3.02 ± 2.65	2.98 ± 0.87	4.37 ± 4.43

5 APPLICATIONS

The analyses presented in the previous sections can be combined for several applications related to SATB choir analysis and synthesis. Once separated, the unison stems can be remixed to emphasise a particular part of the choir, or to remove one for individual choral practice, e.g., one singer may want to remove their part from the mixture and sing along the other parts. Similarly, the analysis of the F0 of the separated unison signals can be used along with the Unison-to-Solo (UTS) and Solo-to-Unison (STU) networks presented in our previous work (Chandna et al., 2020) to generate material for individual or collective choir practice. As illustrated in **Figure 7B**, the UTS model extracts linguistic content, $\hat{z}_v$, for each of the unison stems, $v \in \{S, A, T, B\}$, and uses this information along with the extracted F0, $\widehat{F0}_v$, to synthesize a single singing voice signal, y_v, representing the perceived content of the unison signal. Synthesizing such a signal can be interesting in itself, especially for transcription purposes, using an automatic lyrics transcription system (Demirel et al., 2020). Before the synthesis step, further audio transformations such as pitch-shifting and time-stretching can be applied to the estimated content, so that the user can obtain a modified signal for their practice.

We also proposed an STU model to generate a unison mixture from a single *a cappella* singing voice signal (**Figure 7C**). This model uses voice conversion to generate various clones of the input signal and adds timing and pitch deviations to create a unison effect. Such a model can be used to create modified remixes of an SATB choir recording, y_c, as shown in **Figure 7D**. An additional application of this part is the generation of a full choir recording with multiple singers per voice, given a quartet recording as input.

For this study, we test the effectiveness of our model for transposition and remixing of a full choir recording. As shown in **Figure 7**, we first use one of the source separation models from **Section 3.1**, e.g., UMX_{CBDE}, to separate the soprano, alto, tenor, and bass unison parts from a full choir mixture. This is followed by content extraction and re-synthesis to generate individual single voice signals for each part. We apply transpositions of ±1 semitone for the parts in this step, which corresponds to the UTS model. We then employ the STU model to create unison signals for each of the transformed parts, which are then remixed together.

Furthermore, we present some audio examples on the accompanying website.

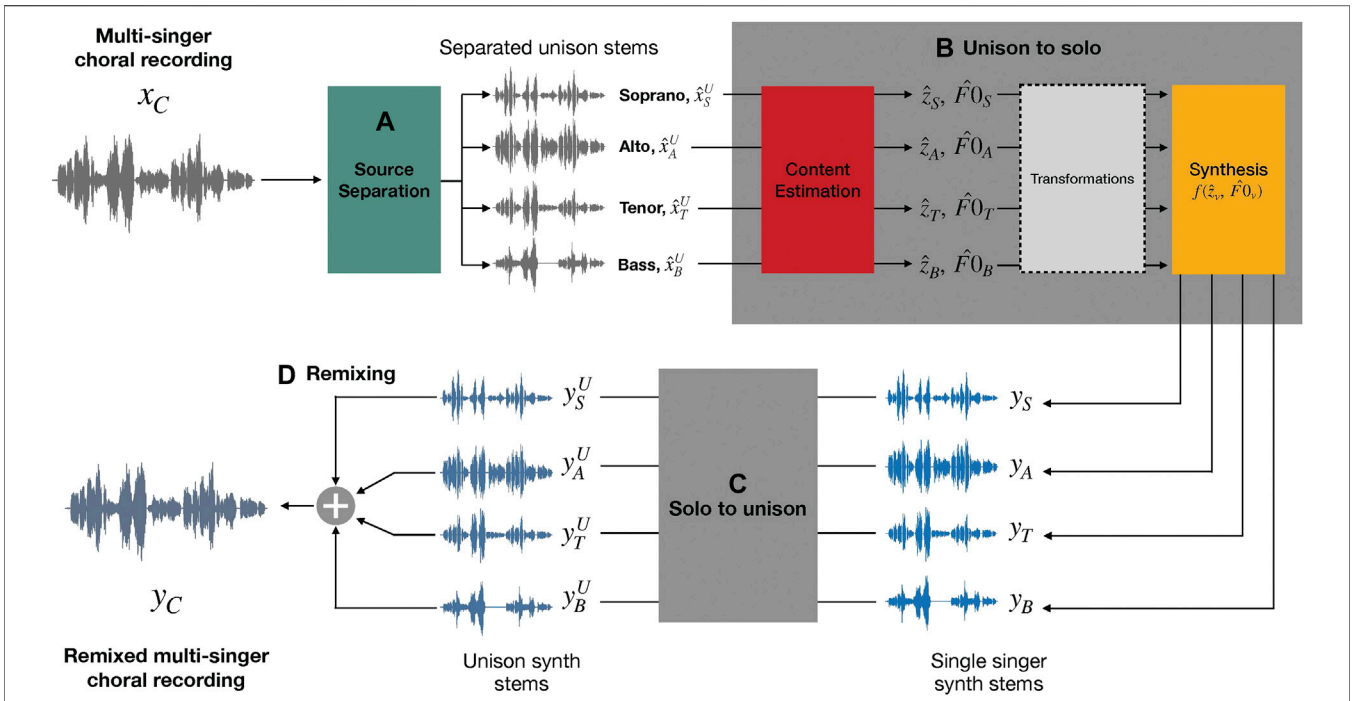

FIGURE 7 | Diagram of the full proposed framework. Given an input recording of a multi-singer choral recording, x_C, the first part of the paper proposes the use of **(A)** source separation techniques to obtain each underlying unison stem, $\hat{x}_v^U$, where $v \in \{S, A, T, B\}$. The second part of this work focuses on the analysis and synthesis of unisons, which we present as a set of applications that include **(B)** unison to solo, with optional audio transformations like pitch-shifting, **(C)** solo to unison, and **(D)** remixing.

6 CONCLUSION

We have conducted an initial foray into source separation for SATB choral recordings. We adapt four recently proposed data-driven source separation models for music and speech to the task of separating the soprano, alto, tenor, and bass voices from a choral mixture. For the experiments, we consider some of the recently published datasets of choral music to train and evaluate source separation models and find that the models proposed for music source separation are more suited to this task than those proposed for speech source separation. We also find that waveform-based models are just as effective as models using intermediate representations such as spectrograms, and that quartet based data can be effectively employed for training models for separating multiple voices in unison per part. These findings provide the foundation for future work in this domain as deep-learning based source separation models move towards end-to-end waveform separation. Due to the difficulty in recording individual singers in a choir section, it is likely that future datasets for SATB choir singing will be in the quartet format and these can be used for augmentation of data for training source separation models. We note that further augmentations such as pitch shifting, as used in Cuesta et al. (2020) for multi-pitch estimation, can also be used for data augmentation, but were not considered in this study. We also show that models trained with fewer but clean data can be used for cleaning the inter-singer leakages that may be present in SATB recordings.

Further, we analyse unison singing within the SATB voices and through perceptual listening tests, and show that the perceived melody of a vocal unison performance can be approximated by the weighted average of the F0s of the individual singers in a unison. This weighted average can be estimated by monophonic F0 tracking algorithms, both data-driven and knowledge-based. Finally, we propose a separation and remixing system which can be used for modifying choral recordings for practice and teaching purposes. The system leverages on research presented in this paper as well as models developed by us earlier to re-synthesise individual singing voice signals from separated parts of the SATB choir. The approximated perceived pitch is used for this synthesis which also allows for modifications to the signal. We provide examples of such modifications on the **Supplementary Material**. We note that the quality of the synthesis can be improved through improvements in each of the constituent components of the framework, including source separation, monophonic F0 estimation, linguistic modeling, and synthesis techniques. We also believe that increased data as well as augmentations such as pitch shifting can improve the performance of the components. We hope that through this study we can lay the foundations for future work in this field.

DATA AVAILABILITY STATEMENT

They can be found here: Choral Singing Dataset: {https://zenodo.org/record/1286485}, Dagstuhl ChoirSet: {https://zenodo.org/

record/3897181}, ESMUC Choir Dataset: {https://zenodo.org/record/5848989}, and Cantoría Dataset: {https://zenodo.org/record/5878677}.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Committee for Ethical Review of Projects (CIREP)—Universitat Pompeu Fabra (UPF). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors contributed to conception and design of the study. PC and DP conducted the source separation experiments, while HC organized the databases and performed the analysis of the unison recordings. DP carried out the data cleaning of one of the datasets. PC and HC designed the listening test, and PC was in charge of creating it. All authors were involved in the search for participants of the listening test. PC, HC, EG, and DP wrote the first versions of the manuscript, and all authors contributed to manuscript revision, as well as approved the submitted version.

FUNDING

This work is partially supported by the European Commission under the TROMPA project (H2020 770376), and the project Musical AI (PID 2019-111403GB-I00/AEI/10.13039/501100011033) funded by the Spanish Ministerio de Ciencia, Innovación y Universidades (MCIU) and the Agencia Estatal de Investigación (AEI).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frsip.2022.808594/full#supplementary-material

REFERENCES

Bugler, A., Pardo, B., and Seetharaman, P. (2020). A Study of Transfer Learning in Music Source Separation. *arXiv preprint arXiv:2010.12650*

Chandna, P., Cuesta, H., and Gómez, E. (2020). "A Deep Learning Based Analysis-Synthesis Framework for Unison Singing," in Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR), 598–604.

Clift, S., Hancox, G., Morrison, I., Hess, B., Kreutz, G., and Stewart, D. (2010). Choral Singing and Psychological Wellbeing: Quantitative and Qualitative Findings from English Choirs in a Cross-National Survey. *J. Appl. arts Health* 1, 19–34. doi:10.1386/jaah.1.1.19/1

Cuesta, H. (2022). *Data-Driven Pitch Content Description of Choral Singing Recordings*. Barcelona, Spain: Ph.D. thesis, Universitat Pompeu FabraTo appear.

Cuesta, H., Gómez, E., and Chandna, P. (2019). "A Framework for Multi-F_0 Modeling in Satb Choir Recordings," in Proceedings of the Sound and Music Computing Conference (Málaga, Spain): SMC), 447–453.

Cuesta, H., Gómez, E., Martorell, A., and Loáiciga, F. (2018). "Analysis of Intonation in Unison Choir Singing," in Proceedings of the International Conference of Music Perception and Cognition (ICMPC) (Graz, Austria), 125–130.

Cuesta, H., McFee, B., and Gómez, E. (2020). "Multiple F0 Estimation in Vocal Ensembles Using Convolutional Neural Networks," in Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR), 302–309.

Défossez, A., Usunier, N., Bottou, L., and Bach, F. (2019). Music Source Separation in the Waveform Domain. *arXiv preprint arXiv:1911.13254*

Dai, J., and Dixon, S. (2017). "Analysis of Interactive Intonation in Unaccompanied SATB Ensembles," in Proceedings of the International Society for Music Information Retrieval Conference (ISMIR) (Suzhou, China), 599–605.

Dai, J., and Dixon, S. (2019). Singing Together: Pitch Accuracy and Interaction in Unaccompanied Unison and Duet Singing. *The J. Acoust. Soc. America* 145, 663–675. doi:10.1121/1.5087817

de Cheveigné, A., and Kawahara, H. (2002). Yin, a Fundamental Frequency Estimator for Speech and Music. *J. Acoust. Soc. America* 111, 1917–1930. doi:10.1121/1.1458024

Demirel, E., Ahlbäck, S., and Dixon, S. (2020). "Automatic Lyrics Transcription Using Dilated Convolutional Neural Networks with Self-Attention," in 2020 International Joint Conference on Neural Networks (IJCNN) (IEEE). doi:10.1109/ijcnn48605.2020.9207052

Devaney, J., Mandel, M. I., and Fujinaga, I. (2012). "A Study of Intonation in Three-Part Singing Using the Automatic Music Performance Analysis and Comparison Toolkit (AMPACT)," in Proceedings of the International Society for Music Information Retrieval Conference (ISMIR) (Porto, Portugal), 511–516.

Gover, M., and Depalle, P. (2020). "Score-informed Source Separation of Choral Music," in Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR), 231–239.

Jansson, A., Humphrey, E. J., Montecchio, N., Bittner, R. M., Kumar, A., and Weyde, T. (2017). "Singing Voice Separation with Deep U-Net Convolutional Networks," in Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR).

Kim, J. W., Salamon, J., Li, P., and Bello, J. P. (2018). "Crepe: A Convolutional Representation for Pitch Estimation," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Calgary, Canada), 161–165. doi:10.1109/ICASSP.2018.8461329

Livesey, L., Morrison, I., Clift, S., and Camic, P. (2012). Benefits of Choral Singing for Social and Mental Wellbeing: Qualitative Findings from a Cross-National Survey of Choir Members. *J. Public Ment. Health* 11, 10–26. doi:10.1108/17465721211207275

Lluís, F., Pons, J., and Serra, X. (2019). End-to-end Music Source Separation: Is it Possible in the Waveform Domain? *ProcInterspeech*, 4619–4623. doi:10.21437/interspeech.2019-1177

Luo, Y., Chen, Z., Hershey, J. R., Le Roux, J., and Mesgarani, N. (2017). "Deep Clustering and Conventional Networks for Music Separation: Stronger Together," in 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (IEEE), 61–65. doi:10.1109/icassp.2017.7952118

Luo, Y., and Mesgarani, N. (2018). "Tasnet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE), 696–700. doi:10.1109/icassp.2018.8462116

Luo, Y., and Mesgarani, N. (2019). Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. *Ieee/acm Trans. Audio Speech Lang. Process.* 27, 1256–1266. doi:10.1109/taslp.2019.2915167

Manilow, E., Wichern, G., Seetharaman, P., and Le Roux, J. (2019). "Cutting Music Source Separation Some Slakh: A Dataset to Study the Impact of Training Data Quality and Quantity," in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (IEEE). doi:10.1109/waspaa.2019.8937170

Mauch, M., and Dixon, S. (2014). "pYIN: A Fundamental Frequency Estimator Using Probabilistic Threshold Distributions," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Florence, Italy), 659–663. doi:10.1109/icassp.2014.6853678

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., et al. (2015). "Librosa: Audio and Music Signal Analysis in python," in Proceedings the Python Science Conference (Texas, USA: Austin), 18–25. doi:10.25080/majora-7b98e3ed-003

McLeod, A., Schramm, R., Steedman, M., and Benetos, E. (2017). Automatic Transcription of Polyphonic Vocal Music. *Appl. Sci.* 7, 1285. doi:10.3390/app7121285

Meseguer-Brocal, G., and Peeters, G. (2019). Conditioned-u-net: "Introducing a Control Mechanism in the U-Net for Multiple Source Separations," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 159–165. doi:10.5281/zenodo.3527766

Morise, M., Yokomori, F., and Ozawa, K. (2016). World: a Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Trans. Inf. Syst.* E99.D, 1877–1884. doi:10.1587/transinf.2015edp7457

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. C. (2018). "FiLM: Visual Reasoning with a General Conditioning Layer," in Proceedings of the 32nd AAAI Conference on Artificial Intelligence.

Petermann, D., Chandna, P., Cuesta, H., Bonada, J., and Gómez, E. (2020). "Deep Learning Based Source Separation Applied to Choir Ensembles," in Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR), 733–739.

Qian, K., Zhang, Y., Chang, S., Yang, X., and Hasegawa-Johnson, M. (2019). "Autovc: Zero-Shot Voice Style Transfer with Only Autoencoder Loss," in International Conference on Machine Learning, 5210–5219.

Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., and Bittner, R. (2017). The {MUSDB18} Corpus for Music Separation. *Res. Data*. doi:10.5281/zenodo.1117372

Rosenzweig, S., Cuesta, H., Weiss, C., Scherbaum, F., Gómez, E., and Müller, M. (2020). Dagstuhl ChoirSet: A Multitrack Dataset for MIR Research on Choral Singing. *Trans. Int. Soc. Music Inf. Retrieval (Tismir)* 3, 98–110. doi:10.5334/tismir.48

Rossing, T. D., Sundberg, J., and Ternström, S. (1986). Acoustic Comparison of Voice Use in Solo and Choir Singing. *J. Acoust. Soc. America* 79, 1975–1981. doi:10.1121/1.393205

Samuel, D., Ganeshan, A., and Naradowsky, J. (2020). "Meta-learning Extractors for Music Source Separation," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE), 816–820. doi:10.1109/icassp40776.2020.9053513

Sarkar, S., Benetos, E., and Sandler, M. (2020). "Choral Music Separation Using Time-Domain Neural Networks," in Proceedings of the DMRN+15: Digital Music Research Network Workshop (Centre for Digital Music, QMUL), 7–8.

Schramm, R., and Benetos, E. (2017). "Automatic Transcription of a Cappella Recordings from Multiple Singers," in AES International Conference on Semantic Audio (Audio Engineering Society).

Scirea, M., and Brown, J. A. (2015). "Evolving Four Part harmony Using a Multiple Worlds Model," in Proceedings of the 7th International Joint Conference on Computational Intelligence (IJCCI) (IEEE), 220–227. doi:10.5220/0005595202200227 1

Stoller, D., Ewert, S., and Dixon, S. (2018). "Wave-u-net: A Multi-Scale Neural Network for End-To-End Audio Source Separation," in Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR), 334–340.

Stöter, F.-R., Liutkus, A., and Ito, N. (2018). "The 2018 Signal Separation Evaluation Campaign," in International Conference on Latent Variable Analysis and Signal Separation (Springer), 293–305. doi:10.1007/978-3-319-93764-9_28

Stöter, F.-R., Uhlich, S., Liutkus, A., and Mitsufuji, Y. (2019). Open-unmix - a Reference Implementation for Music Source Separation. *Joss* 4, 1667. doi:10.21105/joss.01667

Su, L., Chuang, T.-Y., and Yang, Y.-H. (2016). "Exploiting Frequency, Periodicity and Harmonicity Using Advanced Time-Frequency Concentration Techniques for Multipitch Estimation of Choir and Symphony," in Proceedings of the International Conference on Music Information Retrieval (ISMIR) New York City, USA, 393–399.

Sundberg, J. (1987). *The Science of the Singing Voice* (DeKalb, Illinois (USA): Northern Illinois University Press).

Ternström, S. (1991). Perceptual Evaluations of Voice Scatter in Unison Choir Sounds. *STL-Quarterly Prog. Status Rep.* 32, 041–049.

Ternström, S. (2002). Choir Acoustics – an Overview of Scientific Research Published to Date. *Speech, Music Hearing Q. Prog. Status Rep.* 43, 001–008.

Vincent, E., Gribonval, R., and Fevotte, C. (2006). Performance Measurement in Blind Audio Source Separation. *IEEE Trans. Audio Speech Lang. Process.* 14, 1462–1469. doi:10.1109/tsa.2005.858005

Weiss, C., Schelcht, S. J., Rosenzweig, S., and Müller, M. (2019). "Towards Measuring Intonation Quality of Choir Recordings: A Case Study on Bruckner's Locus Iste," in Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR) (Delft, Netherlands), 276–283.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.