5-2024

DLVS4Audio2Sheet: Deep learning-based vocal separation for audio into music sheet conversion

Nicole TEO

Zhaoxia WANG
Singapore Management University, zxwang@smu.edu.sg

Ezekiel GHE

Yee Sen TAN

Kevan OKTAVIO

See next page for additional authors

Citation

TEO, Nicole; WANG, Zhaoxia; GHE, Ezekiel; TAN, Yee Sen; OKTAVIO, Kevan; LEWI, Alexander Vincent; ZHANG, Allyne; and HO, Seng-Beng. DLVS4Audio2Sheet: Deep learning-based vocal separation for audio into music sheet conversion. (2024). *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2024 Workshops, RAFDA and IWTA, Taipei, May 7-10: Proceedings*. 14658, 95-107.
Available at: https://ink.library.smu.edu.sg/sis_research/9160

DLVS4Audio2Sheet: Deep Learning-based Vocal Separation for Audio into Music Sheet Conversion

Zhaoxia WANG[1 [0000−0001−7674−5488]], Nicole TEO[1 [0009−0006−1497−2785]], Ezekiel GHE[1 [0009−0005−9606−2867]], Yee Sen TAN[1 [0009−0007−4373−9824]], Kevan OKTAVIO[1 [0009−0007−4648−7394]], Alexander Vincent LEWI[1 [0009−0006−0455−8441]], Allyne ZHANG[1 [0009−0008−3631−653X]], and Seng-Beng HO[2 [0000−0003−4839−1509]]

[1] Singapore Management University, 80 Stamford Rd, Singapore 178902, Singapore
zxwang@smu.edu.sg;nicolet.2023@engd.smu.edu.sg
[2] Institute of High Performance Computing, A*STAR , 1 Fusionopolis Way, Singapore 138632, Singapore
hosb@ihpc.a-star.edu.sg;hosengbeng@gmail.com

Abstract. While manual transcription tools exist, music enthusiasts, including amateur singers, still encounter challenges when transcribing performances into sheet music. This paper addresses the complex task of translating music audio into music sheets, particularly challenging in the intricate field of choral arrangements where multiple voices intertwine. We propose DLVS4Audio2Sheet, a novel method leveraging advanced deep learning models, Open-Unmix and Band-Split Recurrent Neural Networks (BSRNN), for vocal separation. DLVS4Audio2Sheet segments choral audio into individual vocal sections and selects the optimal model for further processing, aiming towards audio into music sheet conversion. We evaluate DLVS4Audio2Sheet's performance using these deep learning algorithms and assess its effectiveness in producing isolated vocals suitable for notated scoring music conversion. By ensuring superior vocal separation quality through model selection, DLVS4Audio2Sheet enhances audio into music sheet conversion. This research contributes to the advancement of music technology by thoroughly exploring state-of-the-art models, methodologies, and techniques for converting choral audio into music sheets. Code and datasets are available at: https://github.com/DevGoliath/DLVS4Audio2Sheet

Keywords: Music · Choral audio · Music sheet · Vocal separation · Audio-to-Sheet · Deep learning · Open-Unmix · Band-Split Recurrent Neural Networks (BSRNN)

1 Introduction

In the rapidly evolving domain of music technology, a significant focus lies on the intricate process of translating complex audio into precise music sheets [2]. This

challenge, often termed automatic music transcription, presents a formidable task at the intersection of signal processing and artificial intelligence [1]. Automatic music transcription involves the development of computational methodologies to convert acoustic music data into various forms of music notation [23]. Typically, such systems take an audio waveform as input, analyze its time-frequency characteristics, and generate either a typeset music sheets, or a representation of pitches over time [12, 22].

The complexity of automatic music transcription is particularly pronounced in choral music, where multiple voices harmonize to create a rich auditory tapestry [2]. Unlike popular music recordings, which often isolate instruments and vocals, choral compositions inherently involve multiple voices blending together to form intricate harmonie [12]. This inherent complexity poses unique challenges for transcription and audio separation, distinct from the predominant focus of existing literature on standard music source separation in popular songs [11].

Choral music, known for its polyphonic nature, often involves intricate layering of multiple voices singing in various pitches and timbres [2, 12]. This complexity makes it challenging to precisely differentiate between the audio source components such as Soprano, Alto, Tenor, and Bass (SATB) sections [2].These challenges underscore the need for further exploration and development of techniques tailored to choral music source separation.

Music source separation, particularly in vocal separation, has a rich history predating the emergence of deep learning. Early traditional algorithms, such as Independent Component Analysis (ICA) and Hidden Markov Models (HMMs), established the foundation of this field [9].

Recent developments in music research have expanded beyond traditional methodologies to incorporate deep learning models [2, 10, 22]. While adoption of deep learning-based approaches is not yet widespread, these methods show promise for handling intricate tasks such as singing vocal separation.

In this paper, we present DLVS4Audio2Sheet, a novel approach that utilizes deep learning models, including Open-Unmix [18, 19] and Band-Split Recurrent Neural Networks (BSRNN) [10], to accurately segment choral music into distinct audio sections. DLVS4Audio2Sheet is specifically designed to overcome challenges associated with converting audio into music sheets.

The following is a summary of our contributions:

1. This research propose a new method, DLVS4Audio2Sheet, which leverages and evaluates two advanced models, Open-Unmix and Band-Split Recurrent Neural Networks (BSRNN), to tackle the complex task of transcribing choral music audio into notated music sheets.
2. This research offers practical insights into the difficulties that arise while managing the complexity of choral arrangements. Creative ways are also suggested to get beyond these obstacles, offering innovative techniques for professionals and scholars involved in source separation of choral music.
3. This research contributes to a thorough comprehension of the advantages and disadvantages of the evaluated deep learning models in the particular setting

of source separation for choral music. This can be an important resource for the larger community of music technology scholars and practitioners.

2 Related Work

Successful automatic music transcription systems have the potential to revolutionize various interactions between individuals and music, spanning music education, creation, production, search, and musicology [1, 14]. This technology serves as a catalyst for societal and economic impacts. Additionally, tasks such as music source separation, which involve estimating and inferring source signals from mixed data, are closely intertwined with automatic music transcription, highlighting its broader relevance and applications [23].

Furthermore, the music source separation tasks presents a substantial opportunity for enhancing automatic music transcription systems [16]. By effectively separating individual sources within complex audio recordings, such as vocals, and background noise, these systems can more accurately transcribe the underlying musical notes [11]. This not only improves the fidelity of the transcription process but also opens up new avenues for applications in music education, production, and analysis [10].

The field of music source separation has a long history, having its beginnings in the pre-deep learning era with traditional techniques like Independent Component Analysis (ICA) and Hidden Markov Models (HMMs) [9]. In addition, music source separation has undergone a revolution thanks to the development of deep learning algorithms [7]. More recently, Transformer models, Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs) have demonstrated remarkable achievements in accurately identifying sources, allowing for more subtle separation even in complex audio mixes [10, 11, 19]. Although these methods were groundbreaking at the time, they have trouble managing overlapping and complex audio sources, which is a major obstacle in the separation of choral singing [8].

Petermann proposed domain-specific adjustments based on the basic frequency contour of each singing group, following an evaluation of several approaches for music source separation [13]. Additionally, Choi investigated various neural transformation methods and compared their performance [3]. Chandna demonstrated the potential of deep learning approaches in choir ensemble separation by training and evaluating state-of-the-art source separation algorithms using publicly available choral singing datasets [2].

Open-Unmix and Band-Split RNN stand out among other deep learning models for their superior performance in music transcription domains [18, 2, 10]. Open-Unmix, an open-source model, excels in isolating vocals from background noise, thanks to its specialized architecture tailored for subtle separations [18]. Conversely, Band-Split RNNs are particularly adept at discerning sounds in similar frequency ranges, a common challenge in choral music scenarios [10]. Despite their remarkable achievements, these methods encounter difficulties in handling overlapping and complex audio sources, especially evident in choral singing sep-

aration [8]. This research aims to leverage the strengths of advanced models to address challenges and enhance the effectiveness of music transcription processes.

3 Methodology

We develop DLVS4Audio2Sheet, a novel method that leverages two deep learning models, Open-Unmix and BSRNN, to convert choral music audio into notated music sheets. DLVS4Audio2Sheet undergoes evaluation through two distinct modules, allowing for a comprehensive assessment of each individual model's advantages and disadvantages in the context of source separation for choral music. Based on the outcome of this analysis, the most effective models are selected as components of the DLVS4Audio2Sheet method for further processing.

3.1 Overall Design of the DLVS4Audio2Sheet Method

As depicted in Fig. 1, our proposed DLVS4Audio2Sheet method consists of two sub-modules: (A) Vocal Separation and (B) Audio to Music Sheet. In sub-module (A) Vocal Separation, two distinct deep learning models, Open-Unmix and BSRNN, are employed, each tailored for processing choral music inputs. Notably, separate training is conducted for each section of the choir, resulting in distinct trained models for individual sections. Following training, the performance of each model is rigorously tested and evaluated for every choir section. Subsequently, the model with the highest evaluation score is selected for further processing.

It is essential to highlight that the output of the selected deep learning models in the Vocal Separation sub-module (A) results in four distinct audio files corresponding to specific choir sections: Soprano, Alto, Tenor, and Bass. These audio files serve as inputs for the Audio to Music Sheet sub-module (B), where they undergo a conversion procedure to transform them into music score sheets. The audio file with the highest evaluation score is chosen for the next step in the process. This systematic approach ensures the accurate transcription of choral music into notated form, facilitating comprehensive analysis and interpretation.

3.2 Vocal Separation Leveraging the Two Deep learning Models: Open-Unmix and BSRNN

In the overall methodology design of the DLVS4Audio2Sheet method (Fig. 1), vocal separation is identified as one of the core modules. Two individual deep learning-based models are selected to be components of the DLVS4Audio2Sheet method after undergoing thorough evaluation.

Open-Unmix model: In the open-unmix model, choral music is fed as input at the preprocessing stage. Next, the audio signal's frequency range is cropped to concentrate on the pertinent frequencies. Individual frequencies can then be

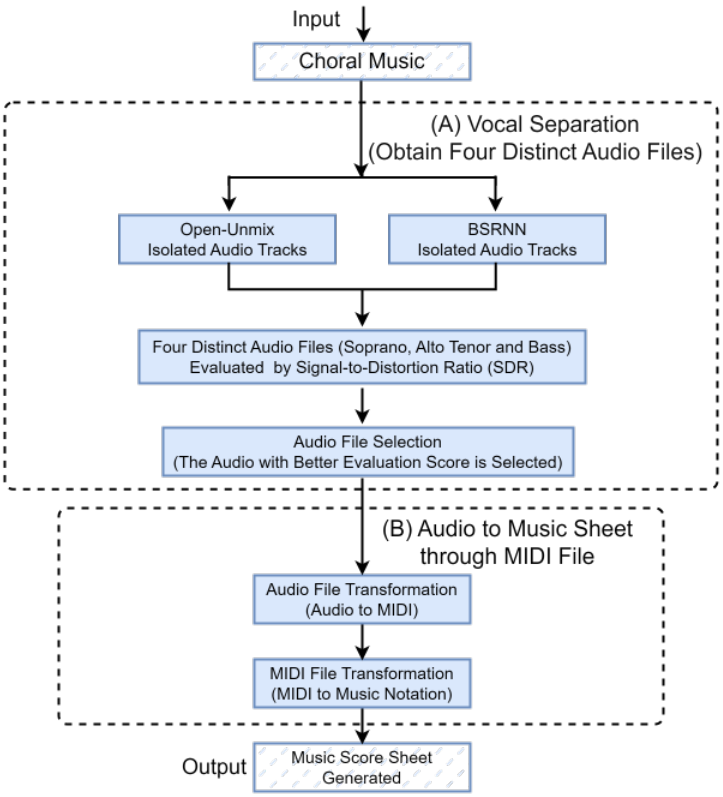

Fig. 1. Overall Design of the DLVS4Audio2Sheet Method

altered by applying a Short-Time Fourier Transform (STFT) on the clipped frequencies, which transforms the time-domain signal into the frequency domain. Following the STFT, the data is subjected to normalization, which modifies the audio signal's amplitude to keep it within a constant range. This procedure is essential for the stability and effectiveness of the neural network processing that follows.

Neural network processing starts with spectral feature extraction. In this stage, the model locates and extracts particular features from the frequency spectrum that are essential for differentiating between various audio sources in the mix. Following feature extraction, source estimation occurs. This is the process by which the model forecasts the distinct sources or elements (such various voices or instruments) in the audio stream. The next step is masking, which creates filters that can isolate particular audio components by suppressing the non-targeted elements using the collected features and source estimations.

Next is the postprocessing phase, where the frequency-domain signal is transformed back into the time domain using an inverse Short-Time Fourier Transform

(ISTFT), which essentially reconstructs the audio waveform from its spectral components. Phase reconstruction is the last stage of the procedure, and is essential in order to guarantee that the separated audio tracks have the proper temporal alignment and coherence. This is because STFT and ISTFT have the potential to alter the signal's phase information. From the original choral music input, this process generates isolated audio tracks, and the isolated tracks are then retrieved [18].

Band-Split RNN (BSRNN): BSRNN [10] introduces a novel approach to music separation using the "band-split" strategy, where the model intricately splits the input audio spectrograms into various sub-bands across different frequency levels. Due to the nature of choral music, the frequencies allotted to each SATB choir section serve as the primary means of differentiation. Compared to previous musical source separation models that focuses more on other audio variables like timbre and reverberation, BSRNN's emphasis on sub-frequencies makes it appropriate for our goal.

First, choral music is introduced into the process and is subjected to frequency decomposition. This decomposition is a part of the audio signal's breakdown into its component frequencies during the pre-processing phase. Next, the broken-down frequencies undergo processing via Fully-Connected layer band-specific RNNs, signifying that distinct RNNs are assigned to distinct audio signal frequency bands. This stage is essential for adjusting the RNN's processing and learning to the properties of various frequency ranges, which can enhance the precision and caliber of the audio separation. The model then executes feature extraction, where it discerns and extracts noteworthy features from the frequency bands that hold significance for the ensuing mask generation phase. After the masks are created, the model performs an inverse frequency transformation after combining the bands. This conversion, which is the opposite of the original frequency decomposition, essentially combines the frequency bands that were processed into a single audio signal again. Isolated audio recordings are this technique's end product, indicating that the RNN model has effectively distinguished the various elements of the choral music.

With this design, BSRNN can precisely adjust its approach for every choir part by using a loss function that combines Mean Absolute Error (MAE) in the frequency domain and in the time domain. By using a dual loss function, BSRNN is guaranteed to preserve the integrity and quality of the time-domain signals in addition to capturing the subtle spectral details of the choir sounds. This works especially well for choir music as it provides clear and accurate vocal separation by emphasizing the complex interactions between temporal and frequency elements in choir compositions.

3.3 Audio to Score Sheet through MIDI File

As depicted in the overall methodology design of the DLVS4Audio2Sheet method (Fig. 1), the Audio to Music Sheet module (Module (B)) is another core com-

ponent. The output of the Vocal Separation module (Module (A)) serves as the input for the Audio to Music Sheet module (Module (B)).

To convert vocal tracks obtained from deep learning models into musical notes, we utilize Python libraries or existing methods such as Librosa, Aubio, SciPy, versatile music21 [6] and AnthemScore software [17]. These methods employ probabilistic modeling techniques to infer the most likely sequence of musical states, resulting in an intermediate piano-roll representation detailing note onsets, offsets, pitches, and names. Subsequently, a MIDI file is generated, incorporating tempo information. In our research, we employ music21 and AnthemScore software to convert segmented audio files into score sheets, simplifying access for musicians and choral groups.

4 Dataset

4.1 Scarcity of Datasets

One obstacle in investigating machine learning methods for multiple music voice separation is the absence of a suitable and large annotated dataset. In order to get around this problem, we created our own multi-track dataset for training by combining different combinations from multi-track datasets that already exist. This is a common data augmentation method used in various multiple singing voice separation papers [5, 15] to address the lack of data. In this section, we will cover the datasets used, and explain the data augmentation process.

4.2 Raw Datasets

Table 1 shows the datasets used in this study. They are all freely accessible through public sources or have been published as parts of other academic works [2, 4].

Table 1. Existing datasets for choral music separation [2, 4]

Dataset	***No. of songs***	***Duration (minutes)***
Choral Singing Dataset	3 songs	7
ESMUC Choir Dataset	3 songs	31
Cantoria Dataset	11 songs	20

All three datasets comprise full-length choir songs consisting of vocal sections including Soprano, Alto, Tenor, and Bass sections. Each individual performer was recorded using a close-up microphone, while distant microphones simultaneously captured the entire choir's sound. Consequently, the separate audio recordings were processed for each solo singer as well as recordings of the full choir singing in unison ('mixture') for every song. When utilizing the complete choir recording as input, the individual recordings serve as the ground truth or reference for the model's output.

While the ESMUC Choir Dataset was sampled at 22 KHz using mono channels, the Choral Singing Dataset and Cantoria Dataset were sampled at 44 KHz using stereo audio channels. Of the above datasets, we chose Choral Singing Dataset and ESMUC Choir Dataset for training, validation and testing, while Cantoria Dataset was used for demonstration purposes when presenting our findings. This is because the former two datasets had 16 and 12 singers respectively, distributed between SATB sections, which allowed for data augmentation to be carried out.

4.3 Data Augmentation

Before being employed in the corresponding models, some measures had to be taken to assure consistency because the two datasets used for training had some minor differences. For example, the recordings for the ESMUC Choir Dataset had to be transformed from the standard 22KHz sample rate to 44KHz sample rate [2, 4]. In addition, the other models needed to be arranged according to song, which meant that each of the four voice stems for a certain song needed to go into the same folder as the mixture that included all of the voice recordings.

To address the lack of a comprehensively annotated dataset, data augmentation was applied to the three training datasets [2, 4]. This was possible due to the multi-track nature of the datasets, along with there being multiple singers for each vocal stem. Each song in the dataset is divided into four vocal stems, each containing three to four distinct recordings by different vocalists. Consequently, artificial mixed recordings for each song are generated by combining one singer per vocal SATB sections, resulting in a total of 256 distinct mixes for each segment, given that there are four singers involved.

To maintain consistency across various models trained locally or in the cloud, these procedures are automated using Python scripts executed on the raw dataset. To generate undiscovered mixes for the testing dataset, we excluded one voice stem from each dataset. This reduction significantly decreased the total number of possible combinations. The resulting dataset comprises 687 unique mixes, which were divided into a roughly 80-10-10 split for training, validation, and testing purposes.

5 Experimentation and Results

5.1 Vocal Separation Results and Comparisons

To begin separating vocals in our choir dataset, we first conducted an 80-10-10 split to allocate data for training, validation, and testing purposes. We evaluated two models, Open-Unmix and BSRNN, for dividing mixed choir vocals into different SATB sections.

The evaluation was based on the Signal-to-Distortion Ratio (SDR), a widely used benchmark for source separation competitions [2, 20, 21]. SDR, measured in decibels (dB), indicates the quality of separation, with a higher SDR indicating clearer distinction of the desired voice from other vocalists or background noise.

At the core of how SDR functions is the assumption that the estimate of a signal source comprises four separate components:

$$\hat{s_i} = s_{target} + e_{interf} + e_{noise} + e_{artif} \tag{1}$$

where s_{target} is the true source; e_{interf}, e_{noise}, and e_{artif} are error terms for interference, noise and artifacts respectively [21].

In our context, the estimate of the source is the mixture track, while the true source refers to the individual recordings of each vocal stem. Interference, or unwanted signals from other sources, would mainly be the voices of other signals, while noise and artifacts are random signals added during measurement and processing. These components are found in the formula for SDR [2, 20, 21]:

$$SDR := 10log_{10}(\frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2}) \tag{2}$$

It can be inferred that SDR is related to the ratio of the target source to the unwanted signals, and can take negative and positive values. A positive value would mean that in the output, the signal of the true source is greater than that of the unwanted signals, while the opposite is true for a negative value. Hence, we aim to maximise the SDR value, which implies that the voice of the desired SATB singer is much more prominent than the other singers and background noise.

We compared our findings with those of Petermann's study [13], which employed State-of-the-Art (SOTA) models for the same choir source separation challenge. For comparison purposes, we selected the top and bottom performing SOTA models based on specific evaluation criteria.

Our model underwent evaluation using the Choral Singing Dataset, which aligns with the dataset utilized in Petermann's study [13]. This dataset choice ensures consistency and enables a direct comparison between our results and those reported by Petermann.

By leveraging the same dataset and evaluating against both the best and the worse performing SOTA models, we aimed to provide a reliable assessment of our findings in relation to existing research in the field of choir source separation.

The Results Leveraging Open-Unmix: The Open-Unmix model was specially modified for choir recordings, with a focus on training separate models for the SATB choir voices. A few notable changes were the lengthening of the training epoch count from 100 to 200, the augmentation of the batch size from 16 to 64, and the rise in sequence time from 6 to 8 seconds. In addition, the arrangement of the audio channels was changed from stereo to mono in order to enhance the choir recordings.

The model outperformed State-of-the-Art (SOTA) models in Soprano separation, displaying a noteworthy proficiency with an SDR of 4.68 while achieving an average SDR of 2.92. This achievement is due to the sequence window's 8-second duration increase, which successfully captured distinct voices. However,

lower frequency problems plagued the model, especially when it came to differentiating Bass voices. This highlights the need for more improvement in the management of lower frequency bands.

The Results Leveraging BSRNN: Since BSRNN is a frequency-domain model, band-split configurations were used to improve its performance when used with SATB choir music. The model underwent modifications involving a decrease in batch size from 8 to 4, as well as a reduction in the epoch range from 100 to 500 to 30 to 50. Resource and computational limitations are the cause of these modifications.

In spite of this, the model came close to the best SOTA model with an average SDR of 2.84. For example, BSRNN outperformed Open-Unmix in the category of voice separation, with an SDR of 2.86, especially in the tenor voice domain. This performance can be attributed to its band-split arrangement, which successfully distinguished choir sections according to variations in frequency.

Table 2. Performance Comparison

Model	SDR				
	Average	***Soprano***	***Alto***	***Tenor***	***Bass***
Open-Unmix	**2.92**	**4.68**	3.12	2.13	1.74
BSRNN	2.84	2.16	3.12	**2.86**	3.21
SOTA (Worst) [2]	-4.26	-7.29	1.05	-15.78	4.98
SOTA (Best) [2]	2.88	1.67	**10.70**	-7.13	**7.42**

Comparison and Further Discussion As outlined in the methodology, Open-Unmix and BSRNN were selected based on their distinct capabilities, with the evaluation focusing on their efficacy in separating choir vocals. In terms of the Bass audio component, neither Open-Unmix nor BSRNN match the performance of previous methods, as illustrated in Table 2.

Open-Unmix demonstrates promise in this regard, achieving an average SDR of 2.92, the highest among the evaluated methods. This result is consistent with previous work, which found Open-Unmix to outperform other models on the same dataset [2]. Interestingly, the highest SDR obtained for Alto and Bass components was also achieved by the Open-Unmix model in the previous study [2].

Even though Open-Unmix exhibits better performance in separating the Soprano component compared to other models, it demonstrates weaker performance in separating the Tenor component compared to BSRNN, which emerges as the superior choice for this particular component.

5.2 Converting Audio to Score Sheet through MIDI File

In the final step of the proposed method, the separated audio files (e.g., vocal tracks) are converted into MIDI format and then into sheet music. For this task,

we utilized music21 and AnthemScore software, comparing their performance to select the superior option for this case study. Our findings revealed that Music21 outperformed AnthemScore for this case. Music21, a Python library tailored for computer-aided musicology, offers a plethora of tools for working with musical data, analysis, and notation. It facilitates the processing of MIDI files, their conversion into music21 objects, and the display of music notation. Leveraging Music21, we transform separated vocal tracks into raw MIDI data, subsequently rendering them into human-readable sheet music.

6 Conclusion, Limitations and Future Works

6.1 Conclusion

In conclusion, this paper proposed DLVS4Audio2Sheet, a novel method designed to address the challenges of transcribing choral music into notated music sheets. DLVS4Audio2Sheet demonstrates promising results in segmenting choral audio into individual vocal sections by leveraging advanced deep learning models such as Open-Unmix and Band-Split Recurrent Neural Networks (BSRNN) for vocal separation. Through rigorous evaluation using these deep learning algorithms, we have assessed the effectiveness of DLVS4Audio2Sheet in producing isolated vocals suitable for notated scoring music conversion. The method's ability to select optimal models for further processing enhances the quality of vocal separation, consequently improving the overall audio-to-music sheet conversion process. This research significantly contributes to the advancement of music technology by thoroughly exploring state-of-the-art models, methodologies, and techniques for converting choral audio into music sheets. Moving forward, DLVS4Audio2Sheet holds promise for facilitating more efficient and accurate transcription of choral music, benefiting music enthusiasts, performers, and composers alike.

6.2 Limitations and Future Works

While DLVS4Audio2Sheet presents a promising approach for converting choral audio into music sheets, there are certain limitations to be considered. Firstly, the effectiveness of the method heavily relies on the quality of the input audio data, including factors such as recording quality and background noise levels. In scenarios where the input audio contains significant noise or overlapping vocal sections, DLVS4Audio2Sheet may encounter challenges in accurately segmenting and isolating individual vocal parts.

Secondly, another limitation of DLVS4Audio2Sheet is the scarcity of training data available for deep learning models. Choral music datasets suitable for training and evaluating vocal separation algorithms are limited. This shortage of data may hinder DLVS4Audio2Sheet's ability to generalize effectively across diverse choral music styles and contexts. Furthermore, the lack of diversity in training data could result in overfitting or biases in the learned representations, thereby limiting the method's robustness and performance on unseen or challenging inputs.

Additionally, while DLVS4Audio2Sheet enhances the audio-to-music sheet conversion process by improving vocal separation quality, it may not fully address all challenges associated with transcribing choral music. Factors such as nuances in musical interpretation, tempo variations, and non-standard notation conventions may still require manual intervention or additional post-processing steps.

Lastly, the performance of DLVS4Audio2Sheet is contingent upon the capabilities and limitations of the deep learning models, Open-Unmix and Band-Split Recurrent Neural Networks (BSRNN), utilized for vocal separation. While these models have demonstrated effectiveness in various applications, they may still struggle with certain complexities inherent in choral music arrangements, such as dynamic vocal interactions with noises. Considering such limitations, future research could also look into exploring other deep learning techniques. For example, applying various large language models (LLMs) to this research domains will be an interest topic.

In summary, while DLVS4Audio2Sheet represents a significant advancement in the field of music transcription, it is essential to acknowledge its limitations and continue exploring avenues for improvement, particularly in ensuring scalability and robustness in real-world applications.

Acknowledgment

The authors express their sincere appreciation to the following SMU students for their keen interest and valuable contributions to the code and report development of this music analysis related research: DARRYL SOH SOON YONG, TAY WAN LIN, NG WEI HAN NORMAN, TOG ZHEN MING, JOEL JOHN TAN, SIM YAN YI, ENQI CHAN, ERIC LI TONG, LEE ZHEXIAN THADDEUS, and TEO WEI LUN.

References

1. Benetos, E., Dixon, S., Duan, Z., Ewert, S.: Automatic music transcription: An overview. IEEE Signal Processing Magazine **36**(1), 20–30 (2018)
2. Chandna, P., Cuesta, H., Petermann, D., Gómez, E.: A deep-learning based framework for source separation, analysis, and synthesis of choral ensembles. Frontiers in Signal Processing **2**, 808594 (2022)
3. Choi, W., Kim, M., Chung, J., Jung, D.: Investigating deep neural transformations for spectrogram-based musical source separation. arXiv preprint arXiv:1912.02591 (2019)
4. Cuesta, H., Gómez Gutiérrez, E., Martorell Domínguez, A., Loáiciga, F.: Analysis of intonation in unison choir singing. In: Proceedings of the 15th International Conference on Music Perception and Cognition / 10th Triennial Conference of the European Society for the Cognitive Sciences of Music. Graz (Austria). pp. 125–130 (2018)
5. Cuesta, H., M.B., Gómez, E.: Multiple f0 estimation in vocal ensembles using convolutional neural networks. arXiv preprint arXiv:2009.04172 (2020)

6. Cuthbert, M.S., Ariza, C.: music21: A toolkit for computer-aided musicology and symbolic music data (2010)
7. Grais, E.M., Sen, M.U., Erdogan, H.: Deep neural networks for single channel source separation. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3734–3738. IEEE (2014)
8. Hershey, J., C.M.: Audio-visual sound separation via hidden markov models. In: Advances in Neural Information Processing Systems. vol. 14 (2001)
9. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural networks **13**(4-5), 411–430 (2000)
10. Luo, Y., Yu, J.: Music source separation with band-split rnn. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023)
11. Mitsufuji, Y., Fabbro, G., Uhlich, S., Stöter, F.R., Défossez, A., Kim, M., Choi, W., Yu, C.Y., Cheuk, K.W.: Music demixing challenge 2021. Frontiers in Signal Processing **1**, 808395 (2022)
12. Nikolsky, A., Alekseyev, E., Alekseev, I., Dyakonova, V.: The overlooked tradition of "personal music" and its place in the evolution of music. Frontiers in Psychology **10**, 3051 (2020)
13. Petermann, D., Chandna, P., Cuesta, H., Bonada, J., Gómez, E.: Deep learning based source separation applied to choir ensembles. arXiv preprint arXiv:2008.07645 (2020)
14. Román, M.A., Pertusa, A., Calvo-Zaragoza, J.: Data representations for audio-to-score monophonic music transcription. Expert Systems with Applications **162**, 113769 (2020)
15. Rosenzweig, S., Cuesta, H., Weiß, C., Scherbaum, F., Gómez, E., Müller, M.: Dagstuhl choirset: A multitrack dataset for mir research on choral singing. Transactions of the International Society for Music Information Retrieval **3**(1) (2020)
16. Schedl, M., Gómez, E., Urbano, J., et al.: Music information retrieval: Recent developments and applications, vol. 8. Foundations and Trends® in Information Retrieval (2014)
17. Sofronievski, B., G.B.: Scorpiano–a system for automatic music transcription for monophonic piano music. arXiv preprint arXiv:2108.10689 (2021)
18. Stöter, F.R., Uhlich, S., Liutkus, A., Mitsufuji, Y.: Open-unmix-a reference implementation for music source separation. Journal of Open Source Software **4**(41), 1667 (2019)
19. Thakur, K.K., Choudhury, S., Ghosh, S., Dash, S., Chabbra, T.S., Ali, I., Shankarappa, R.T., Tiwari, S., Goyal, S.: Speech enhancement using open-unmix music source separation architecture. In: 2022 IEEE Delhi Section Conference (DELCON). pp. 1–6. IEEE (2022)
20. Tzinis, E., Wisdom, S., Hershey, J.R., Jansen, A., Ellis, D.P.: Improving universal sound separation using sound classification. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 96–100. IEEE (2020)
21. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE transactions on audio, speech, and language processing **14**(4), 1462–1469 (2006)
22. Wen, Y.W., Ting, C.K.: Recent advances of computational intelligence techniques for composing music. IEEE Transactions on Emerging Topics in Computational Intelligence **7**(2), 578–597 (2022)
23. Wu, Y.T., Chen, B., Su, L.: Multi-instrument automatic music transcription with self-attention-based instance segmentation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 2796–2809 (2020)