DLVS4audio2sheet: Deep Learning-based Vocal Separation for Audio into Music Sheet Conversion

Nicole Teo, Zhaoxia Wang, Ezekiel Ghe, Yee See Tan, Kevan Oktavio, Alexander Vincent Lewi, Allyne Zhang, and Seng-Beng Ho

Presenter: Ezekiel Ghe, Nicole Teo

Email: ezekiel.ghe.2020@scis.smu.edu.sg, nicolet.2023@engd.smu.edu.sg

Outline

- ❑ Introduction
- ❑ Related Work
- ❑ Proposed Method
- ❑ Results & Discussion
- ❑ Conclusion

Leverages Open-Unmix and BSRNN to convert choral music audio into notated music sheets

Introduction/Background

- Choral music involves intricate layering of multiple voices singing in various pitches and timbres
- Hard to precisely differentiate between Soprano, Alto, Tenor, and Bass (SATB)
- Adoption of deep learning-based approaches remains somewhat constrained

Question: Does deep learning models enhance audio into music sheet conversion?

- Luo, Y., Yu, J.: Music source separation with band-split rnn. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023)

Contributions

➢ Using Deep learning methods

- Open-Unmix and BSRNN
- **Tackle the complex task of transcribing choral music audio into notated music sheets**

➢ Manage the complexity of choral arrangements

- Offer innovative techniques for professionals and scholars involved in source separation of choral music

➢ Thorough comprehension of advantages and disadvantages of deep learning models

- Particular setting of source separation for choral music
- Important resource for the community of music technology scholars and practitioners

Related Work

- Automatic music transcription systems (Benetos et al., 2018)
  - Serves as a catalyst for societal and economic impacts
- Music source separation (Luo et al., 2023)
  - Vocal separation
  - Open-Unmix and Band-Split Recurrent Neural Networks (BSRNN)

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

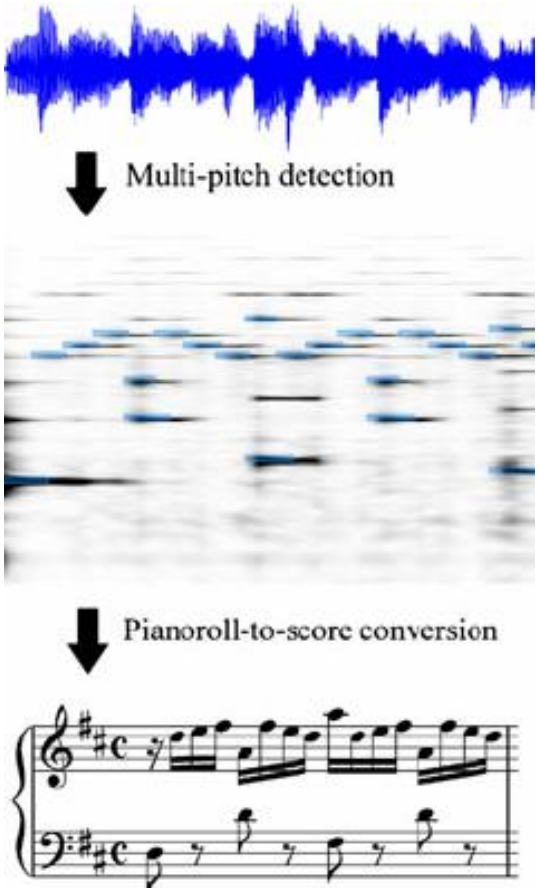
- Benetos, E., Dixon, S., Duan, Z., Ewert, S.: Automatic music transcription: An overview. *IEEE Signal Processing Magazine* 36(1), 20–30 (2018)
- Luo, Y., Yu, J.: Music source separation with band-split rnn. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023)

Related Work

1. Automatic music transcription systems

- Potential to revolutionize various interactions between individuals and music, spanning music education, creation, production, search, and musicology

- Music source separation are closely intertwined with automatic music transcription, highlighting its broader relevance and applications



• Román, M.A., Pertusa, A., Calvo-Zaragoza, J.: Data representations for audio-to-score monophonic music transcription. *Expert Systems with Applications* 162, 113769 (2020)

Picture credit: https://link.springer.com/article/10.1007/s10844-013-0258-3

Related Work

2. Music source separation

- By effectively separating individual sources within complex audio recordings, these systems can more accurately transcribe the underlying musical notes
  - Improves the fidelity of the transcription process and opens up new avenues for applications in music education, production and analysis
- Transformer models (RNNs and CNNs) have demonstrated achievements in accurately identifying sources, allowing for more subtle separation even in complex audio mixes
  - However, these methods have trouble managing overlapping and complex audio sources, which is a major obstacle in the separation of choral singing
- Open-Unmix excel in isolating vocals and Band-Split RNNs at discerning similar frequency ranges
  - Despite their advancements, overlapping and complex audio sources, particularly in choral singing, remain problematic

- Tzinis, E., Wang, Z., Smaragdis, P.: Sudo rm-rf: Efficient networks for universal audio source separation. In: *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP).* pp. 1–6. IEEE (2020)
- 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)

DLVS4Audio2Sheet: A method that leverages two deep learning models

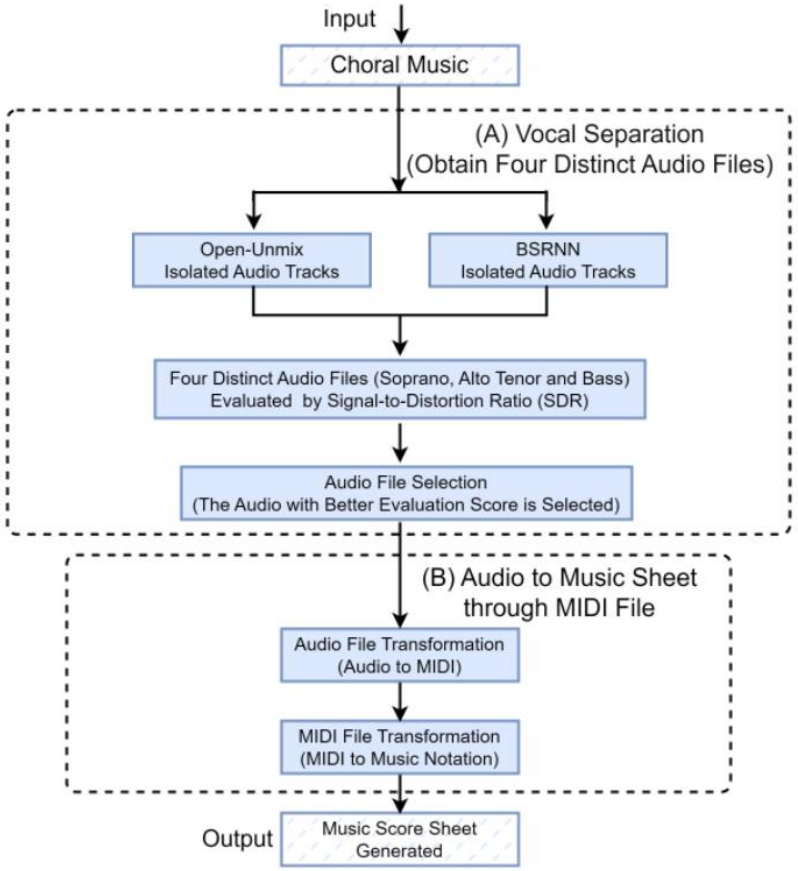


Fig. 1. Overall Design of the DLVS4Audio2Sheet Method

(A) Vocal Separation: Open-Unmix and BSRNN are employed, each tailored for processing choral music inputs

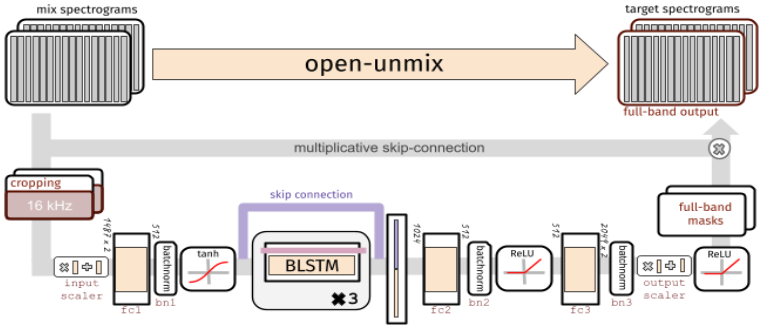
Separate training is conducted for each section of the choir, resulting in distinct trained models for individual sections

(B) Audio to Music Sheet: Audio files undergo a conversion procedure to transform them into music score sheets

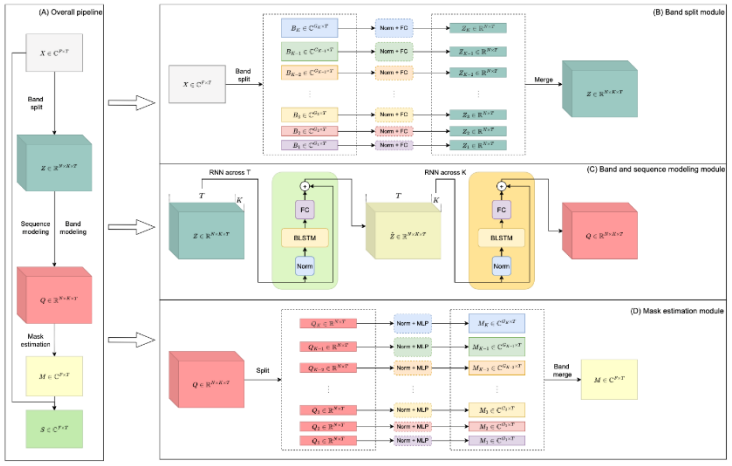
Audio file with the highest evaluation score is chosen for the next step. This ensures the accurate transcription of choral music into notated form, facilitating comprehensive analysis and interpretation

Two important components in the proposed DLVS4Audio2Sheet method

Open-Unmix



Picture credit: https://github.com/sigsep/open-unmix-pytorch



BSRNN (Luo and Yu 2023)

Open-Unmix: Preprocessing → STFT to convert the signal into the frequency domain → Feature extraction → Source Estimation → Postprocessing using ISTFT

Phase reconstruction ensures proper temporal alignment and coherence of separated audio tracks

BSRNN: Preprocessing → Fully-connected layer band-specific RNNs → Feature extraction → Inverse frequency transformation

Dual loss function ensures preservation of time-domain signal integrity and capturing subtle spectral details

- 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)
- Luo, Y., Yu, J.: Music source separation with band-split rnn. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023)

Audio to Score Sheet through MIDI File

- Output of the Vocal Separation module (Module (A)) serves as the input for the Audio to Music Sheet module (Module (B))
- We used Python libraries or existing methods
- These methods employ probabilistic modelling 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

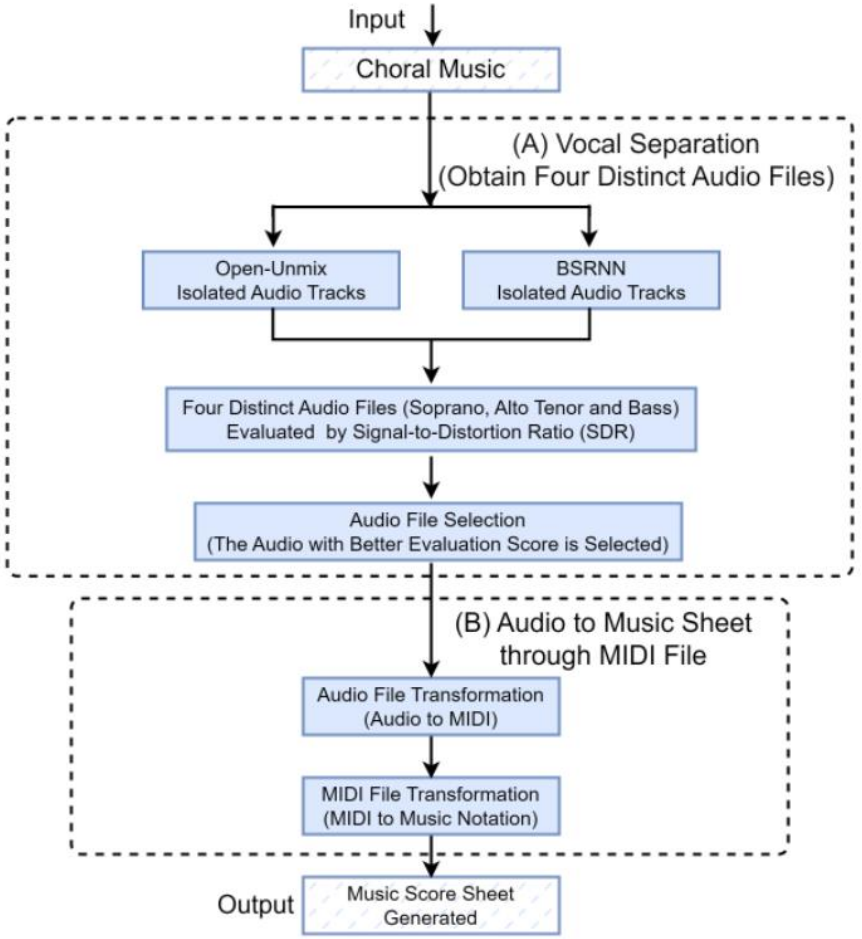


Fig. 1. Overall Design of the DLVS4Audio2Sheet Method

Dataset

- Choral Singing Dataset
  - 3 songs, 7 mins
  - Sampled at 44 KHz using stereo audio channels
  - Chosen for training, validation, and testing
- ESMUC Choir Dataset
  - 3 songs, 31 mins
  - Sampled at 22 KHz using mono channels
  - Chosen for training, validation, and testing
- Cantoria Dataset
  - 11 songs, 20 mins
  - Sampled at 44 KHz using stereo audio channels
  - Used for demonstration purposes when presenting findings

Results and Discussion
Vocal Separation Results and Comparisons

(1) Four separate components of SDR

$$\hat{s}_i = s_{target} + e_{interf} + e_{noise} + e_{artif}$$

(2) SDR formula

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

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**

Results and Discussion

Converting Audio to Score Sheet through MIDI File

- In the final step, the separated audio files are converted to MIDI format and then into sheet music
- We used music21 and AnthemScore software, comparing their performance to select the superior option
- Findings reveal that Music21 outperformed AnthemScore
- Leveraging Music21, we transform separated vocal tracks into raw MIDI data, subsequently rendering them into human-readable sheet music

Conclusion

- **DLVS4Audio2Sheet**——a novel method designed to address the challenges of transcribing choral music into notated music sheets
- It demonstrates promising results by leveraging advanced deep learning models (Open-Unmix and BSRNN) for vocal separation
- Facilitate more efficient and accurate transcription of choral music, benefitting music enthusiasts, performers, and composers

Limitations and Future Work

- Effectiveness of this method relies heavily on quality of the input audio data
- Scarcity of training data available for deep learning models
- Lack of diversity in training data could result in overfitting or biases in the learned representations
- May not fully address all challenges associated with transcribing choral music
- Future work could explore other deep learning techniques (e.g. applying various LLMs)

Acknowledgements

The authors express their sincere appreciation to the following SMU students for their enthusiastic interest and invaluable contributions to this music analysis-related research: Darryl Soh, Wan Lin Tay, Norman Ng, Zhen Ming Tog, Joel Tan, Yan Yi Sim, Enqi Chan, Eric Li Tong, Thaddeus Lee, and Wei Lun Teo. Their dedication has significantly enriched our work.

References

Benetos, E., Dixon, S., Duan, Z., Ewert, S.: Automatic music transcription: An overview. IEEE Signal Processing Magazine 36(1), 20–30 (2018)

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)

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)

Cuesta, H., M.B., Gómez, E.: Multiple f0 estimation in vocal ensembles using convolutional neural networks. arXiv preprint arXiv:2009.04172 (2020)

Cuthbert, M.S., Ariza, C.: music21: A toolkit for computer-aided musicology and symbolic music data (2010)

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)

Hershey, J., C.M.: Audio-visual sound separation via hidden markov models. In: Advances in Neural Information Processing Systems. vol. 14 (2001)

Hu, Z., Wang, Z., Ho, S.B., Tan, A.H.: Stock market trend forecasting based on multiple textual features: a deep learning method. In: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI). pp. 1002–1007. IEEE (2021)

Hu, Z., Wang, Z., Wang, Y., Tan, A.H.: Msrl-net: A multi-level semantic relation-enhanced learning network for aspect-based sentiment analysis. Expert Systems with Applications 217, 119492 (2023)

Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural networks 13(4-5), 411–430 (2000)

Luo, Y., Yu, J.: Music source separation with band-split rnn. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023)

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)

Ni, J., Young, T., Pandelea, V., Xue, F., Cambria, E.: Recent advances in deep learning based dialogue systems: A systematic survey. Artificial intelligence review 56(4), 3055–3155 (2023)

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)

Parth, Y., Wang, Z.: Extreme learning machine for intent classification of web data. In: Proceedings of ELM-2016. pp. 53–60. Springer (2018)

Román, M.A., Pertusa, A., Calvo-Zaragoza, J.: Data representations for audio-to-score monophonic music transcription. Expert Systems with Applications 162, 113769 (2020)

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)

Schedl, M., Gómez, E., Urbano, J., et al.: Music information retrieval: Recent developments and applications, vol. 8. Foundations and Trends® in Information Retrieval (2014)

Tzinis, E., Wang, Z., Smaragdis, P.: Sudo rm-rf: Efficient networks for universal audio source separation. In: 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. IEEE (2020)

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)

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)

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)

…

Picture credit: http://www.marriageenvy.com/new-blog/