Converting Vocal Performances into Sheet Music Leveraging Large Language Models

Jinjing JIANG[2], Nicole TEO[1], Haibo PEN[3], Seng-Beng HO[4], Zhaoxia WANG[1*]
Research and Development, Mastercard, Singapore, Singapore[2]
School of Computing and Information Systems, Singapore Management University, Singapore, Singapore[1]
School of Electrical and Information Engineering, Tianjin University, China, China[3]
AI Institute Global, 79 Ayer Rajah Crescent, 139955 Singapore[4]
jiangjinjing19@gmail.com, nicolet.2023@engd.smu.edu.sg, penhaibo@tju.edu.cn,
sengbeng.ho@aiiglobal.ai, zxwang@smu.edu.sg*

***Abstract*—Advanced natural language processing (NLP) models are increasingly applied in music composition and performance, particularly in generating vocal melodies and simulating singing voices. While NLP techniques have been successfully used to analyze vocal performance data, providing insights into performance quality and style, the automatic transcription of vocal performances into sheet music remains a complex challenge. Existing tools for manual transcription are often insufficient due to the intricate dynamics of vocal expression. This study addresses the challenge of automating the conversion of vocal performances into sheet music using a combination of novel techniques, including large language models (LLMs). We present a method that successfully translates vocal audio input into display-ready sheet music. Our findings highlight the strengths and limitations of various approaches, particularly in the transcription of a cappella performances into notes and lyrics. This research contributes to the expanding field of NLP-driven music analysis and showcases the potential of these models to revolutionize vocal transcription in the future.**

***Index Terms*—Natural Language Processing, Vocal Performance, Automatic Music Transcription (AMT), Large Language Models, Machine Learning, A Cappella, Lyric Transcription, Sheet Music**

I. Introduction

Advancements in machine learning (ML) and natural language processing (NLP), particularly with large language models (LLMs), are increasingly permeating a wide range of research and application domains [4], [5], [16], [23], [50], [51]. Among these applications, aligning lyrics with musical notes during transcription has emerged as a significant task [22], [30], [36]. This process involves accurately mapping the timing and content of lyrics to corresponding musical sections, a task that demands sophisticated NLP models capable of handling variations in pronunciation and tempo. Additionally, NLP techniques are being applied to analyze the sentiment and emotional tone of lyrics, providing deeper insights into the emotional dimensions of vocal performances [39]. The application of deep learning to analyze and transcribe vocal performances into accurate sheet music is thus becoming a focal point of research [46].

The convergence of advanced NLP and music analysis has sparked considerable interest due to the ability of modern models to both interpret and generate musical content [11], [30], [36]. Recent progress has enabled the development of sophisticated systems that can generate vocal melodies and simulate singing voices, offering new creative opportunities for music composition and performance. These technologies have begun to unlock novel creative possibilities for artists and music producers alike.

However, despite notable advancements in utilizing NLP for analyzing vocal performances—yielding insights into performance quality, style, and emotional expression—the automated conversion of vocal performances into sheet music remains a complex and unsolved challenge. Existing software tools provide some assistance for manual transcription, but the nuanced and dynamic nature of vocal performances often diminishes their efficacy [6]. As a result, automating the transcription of vocal music into standardized notation is an ongoing and pressing issue in this field.

This study aims to address this challenge by exploring various automated methods for converting vocal performances into sheet music, including LLMs. By evaluating the strengths and limitations of various techniques, particularly in the transcription of a cappella music, this research contributes to the expanding body of literature on NLP and LLM applications in music analysis. Ultimately, this work demonstrates the limitations and transformative potential of LLMs to enhance the interpretation of vocal performances and facilitate their transformation into accessible musical notation, thereby enriching the music creation process.

The main contributions of this paper are as follows:

- This research proposes a novel method for converting vocal performances, including a cappella music, into sheet music, addressing a significant challenge in music transcription.
- This research conducts a comprehensive comparative analysis of various vocal isolation techniques crucial for accurate transcription. This analysis outlines the strengths and weaknesses of these methods, particularly in their application to a cappella music transcription.

Jinjing JIANG and Nicole TEO contributed equally to this research and are both recognized as first authors.

2375-9259/24/$31.00 ©2024 IEEE
DOI 10.1109/ICDMW65004.2024.00063

- This study presents a framework for automating vocal transcription into sheet music by incorporating advanced techniques, including LLMs, to tackle the inherent complexities of vocal performance transcription.
- Leveraging NLP techniques, this paper provides valuable insights into the quality and stylistic elements of vocal performances, offering a deeper understanding of their impact on transcription accuracy.
- This research contributes to the expanding literature on NLP and LLM applications in music analysis, highlighting both the limitations and potential of these technologies in automating vocal performance transcription, and identifying new avenues for future research.

II. Existing Works

The integration of NLP techniques, such as machine learning and LLMs, with music technology has garnered significant attention in recent years, with numerous studies exploring the potential applications of NLP techniques in music analysis, composition, and performance [11], [14]. Early research focused primarily on using NLP to analyze lyrical content, examining themes, sentiment, and stylistic elements within song lyrics. These efforts laid the groundwork for more complex interactions between language and music, particularly as the capabilities of NLP models have evolved [29], [33].

Recent advancements in machine learning, especially with the advent of deep learning and transformer-based models, have expanded the scope of NLP applications in music [10], [45]. Researchers have leveraged these models to generate vocal melodies and simulate singing voices, demonstrating the potential for creating innovative musical compositions without human intervention. Notable projects include the use of generative models to produce original music pieces, as well as systems designed to enhance live performances through real-time vocal synthesis.

Despite these promising developments, the task of converting vocal performances into sheet music remains a formidable challenge [41]. Various software tools have been developed to assist with manual transcription; however, the intricate nuances of vocal expression—such as pitch variations, timing, and phrasing—often evade accurate representation using conventional methods. Studies indicate that while existing automated transcription systems can be effective in certain contexts, they struggle with the complexities of vocal performances, particularly in genres that emphasize improvisation and a cappella arrangements.

Recent research has started to explore the potential of LLMs for music transcription [6], [11], [22]. These models have shown promise in analyzing structured musical data and generating corresponding notations. However, the application of LLMs to transcribe vocal performances into sheet music remains largely underexplored, presenting an opportunity to advance the field by evaluating the strengths and limitations of LLMs compared to traditional transcription methods. In the domain of audio signal processing, key objectives such as audio source separation and speech enhancement play a critical role in isolating specific signals from complex recordings, a prerequisite for accurate vocal transcription.

For the intricate task of separating an a cappella song into six distinct vocal tracks (Lead Vocal, Soprano, Alto, Tenor, Bass, and Vocal Percussion), our initial approach involved using Non-negative Matrix Factorization (NMF), which has demonstrated its efficacy in capturing complex audio structures [29]. However, our focus expanded to evaluate several alternative methods, including Independent Low Ranking Matrix Analysis (ILRMA), Convolutional Recurrent Neural Network with Attention (CRNN-A), and Spleeter, each of which brings unique strengths and capabilities, as detailed in Section III-A. We aim to employ the most promising method from this evaluation. This innovative approach, integrating multiple techniques, effectively addresses the complexities inherent in a cappella source separation [9].

Moreover, accurately assessing the fundamental frequency (F0) during the conversion of audio WAV files to MIDI remains a considerable challenge. [13]. Hidden Markov Models (HMMs) and the Viterbi algorithm are popular in handling sequence-to-sequence tasks in NLP domains. Such techniques can effectively address complexities in extracting pitch, timing, and duration information from audio signals. While fundamental frequency-based methods struggle with F0 inaccuracies, HMMs provide a probabilistic framework adept at modelling temporal dependencies within audio data [33]. Their distinct advantage in handling intricate audio recordings contributes to a more faithful MIDI representation [33]. Playing a pivotal role in the decoding process within HMMs, the Viterbi algorithm ensures precise alignment between observed audio data and MIDI note sequences [2].

Similarly, speech-to-text conversion is a critical task within the domain of NLP. Recent advancements in speech recognition, such as OpenAI's Whisper, have achieved notable success, with a reported transcription accuracy of 98.5% for spoken English [25]. This technology facilitates the conversion of audio lyrics into text. The proposed workflow involves rigorous text processing to correct transcription inaccuracies, followed by the synchronization of lyrics with audio through forced alignment techniques, such as the afaligner Python package. This approach capitalizes on established technologies to ensure precise and synchronized transcription of audio content.

The subsequent conversion of MIDI to sheet music entails the extraction of essential musical components from the MIDI file, including notes, chords, time signatures, and key signatures. The music21 Python package from MIT plays an integral role in this process, facilitating the import of MIDI files and the associated transcribed lyrics, which are then systematically transformed into sheet music notation [9].

III. Proposed Methodology

As illustrated in Fig.1, the first proposed method consists of a series of sequential processing steps (see Fig.1 (a)). The process begins with audio preprocessing aimed at isolating distinct vocal tracks. Subsequently, the isolated vocal track

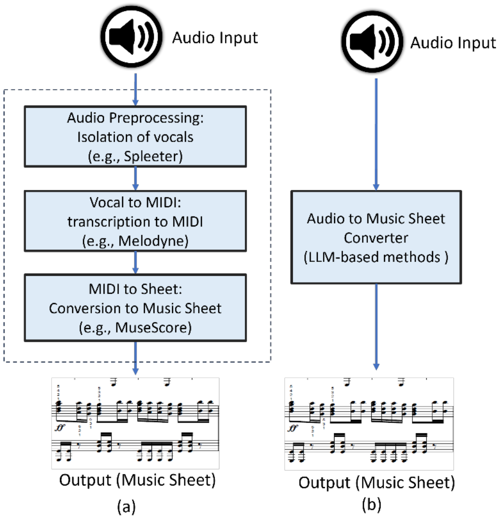

Audio Input
Audio Preprocessing: Isolation of vocals (e.g., Spleeter)
Vocal to MIDI: transcription to MIDI (e.g., Melodyne)
MIDI to Sheet: Conversion to Music Sheet (e.g., MuseScore)
Output (Music Sheet)
(a)
Audio Input
Audio to Music Sheet Converter (LLM-based methods)
Output (Music Sheet)
(b)

Fig. 1. Overview of the Proposed Methods: (a) A multi-step music transcription solution involving vocal isolation, vocal to MIDI and MIDI to Sheet conversions; (b) An integrated solution for music transcription with LLMs

is converted from WAV format to MIDI. Also, the lyrics are extracted specifically from the lead vocal track. In the final stage, MIDI files are translated into musical notation, and the extracted lyrics are aligned with their corresponding musical notes to produce a complete sheet music representation. This approach provides a structured and efficient methodology for transcribing audio inputs, such as a cappella vocal recordings, into accurate and comprehensive sheet music. For this study, we employ the Jacappella dataset, which contains 35 Japanese a cappella sheet music alongside corresponding vocal audio recordings, including six distinct individual vocal parts' recordings [48].

Fig. 1 (b) shows the LLM-based approach from the user's point of view. Research has been done on several LLM-based methods and frameworks, typically involving general-purpose Transformer models, that can assist in converting vocal performances into sheet music. This approach envisions an integrated solution for music transcription with fewer steps.

A. Vocal Isolation

The first step in Fig.1 (a) involves separating the vocal tracks from a cappella songs. In this study, we attempted four typical sound decomposition methods– Non-negative Matrix Factorization (NMF) [24], Independent Low Ranking Matrix Analysis (ILRMA) [20], [21], CRNN-A model [8], [53], and Spleeter [15]. NMF effectively decomposes mixed sounds into individual sources by capturing the non-negative and additive characteristics of audio signals. ILRMA is recognized for its high-quality blind source separation by assuming sources are less dependent on time-frequency [19]. If we attempt to split the tracks by speech, the CRNN-A model would be a promising approach, as it combines a recurrent and a convolutional neural network (CNN) for voice separation. [45]. Spleeter can be used to learn various vocal features as it uses time-frequency (TF) masking, which filters the mixture of Short-time Fourier transform (STFT) frequencies in a piece of music, allowing for the selective isolation of desired elements [15]. At the end of this whole process. the resulting isolated audio tracks for a cappella song are obtained.

B. Vocal to MIDI

The second step of the proposed method (a) is to convert each separated vocal track obtained from Vocal Separation (Section III-A) into MIDI. In other words, we transformed the WAV audio into MIDI format files, encapsulating the musical notes and their timings into a digital representation. Various methods are experimented with to convert using Python libraries such as Librosa [28], Aubio [1] and SciPy [49]. Each attempt builds upon the shortcomings of the previous one, incorporating more accurate pitch estimation methods, considering note duration and introducing probabilistic models.

The results were produced in an existing work that leverages several signal processing techniques and probabilistic modelling [47], including HMM, Viterbi Algorithm and PYIN. The HMM transition matrix defines the probabilities of transitioning between different states (silence, onset, sustain) for each note, while the prior probabilities are estimated from the input audio signal using the PYIN pitch estimator [7], [27], which considers pitch, voicing, and onset information. The Viterbi algorithm is then applied, utilising the transition matrix and prior probabilities, along with the insights from PYIN, to infer the most likely sequence of musical states. An intermediate piano-roll representation is constructed from this sequence, detailing note onsets, offsets, pitches, and note names. A MIDI file is then generated, incorporating tempo information. This approach combines probabilistic modelling with detailed note representation, enhancing the accuracy of the MIDI conversion process.

C. MIDI to Sheet

The final step in Fig.1 (a) is to convert the MIDI format into a piece of sheet music by utilising the versatile music21 Python library for music notation handling and analysis. The MIDI files were first converted to MuseScore [52] format, a widely used open-source music notation software, which allows for the precise rendering of sheet music. Then, we add the lyrics obtained by OpenAI's Whisper automatic speech recognition model. As a Transformer sequence-to-sequence model, Whisper undergoes training on various speech-processing tasks concurrently, such as multilingual speech recognition and translation, spoken language identification, and voice activity detection [37]. For transcription, Whisper segments the input audio into 30-second intervals, converting each segment into a log-Mel spectrogram, employing its decoder to predict the text transcript word by word, utilising the informative visual representation of audio signals. Eventually, by leveraging music21, we can transform raw MIDI data into human-readable sheet music in various formats, integrating lyrics and ensuring an accurate representation of a cappella performances [9].

D. Vocal Performances into Sheet Music using LLMs

In Fig. 1 (b), we present an integrated solution for music transcription leveraging LLMs. The Multi-Task Multitrack Music Transcription (MT3) [17] is a multi-instrument automatic music transcription model based on the T5X framework [40], a general-purpose Transformer model. While the original framework was developed to handle a wide range of instrumental combinations, our work specifically explores its application in transcribing a cappella performances. Additionally, we conducted several experiments with custom music transcription models available through OpenAI's ChatGPT [34], utilizing GPT-4o [35] to evaluate their efficacy in this domain.

IV. Experiment Results and Discussion

A. Dataset Specifcations

The Ja Cappella dataset [31] is used to develop and evaluate the proposed automatic music transcription method. It contains 35 a cappella vocal ensemble songs and their corresponding individual voice part recordings. This dataset is also segmented into seven genres, jazz, punk rock, bossa nova, popular, reggae, enka, and neutral. These subsets contain songs that showcase the characteristic traits of their respective genres, which might help classify audio based on their features, such as chordal harmony and key the vocals are sung in.

B. Algorithm Performance and Result Analysis

(1) Vocal separation algorithm parameter settings and results

The experiment feature extractions heavily rely on time, amplitude, and frequency as the primary dimensions for our model training. For evaluation, we used the Signal-to-Distortion Ratio (SDR), the most commonly used foundational yardstick for various source separation competitions, which is measured in decibels(dB) [43]. SDR essentially quantifies the ratio of the desired source signal to unwanted components, including interference, noise, and artifacts.

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

The higher the SDR, the more distinctly separated the vocal tracks are, suggesting the more accurate the separation of the a cappella.

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

where, s_{target}, e_{interf}, e_{noise}, and e_{artif}. s_{target} represents the true source, and e_{interf}, e_{noise}, and e_{artif} represent error terms for interference, noise, and added artefacts respectively [26].

We compared the results with existing works that employed various deep learning methods [46] and found that the Spleeter model performed particularly well in vocal separation. Our analysis included four models from the study conducted on the Ja Cappella dataset: the NMF model [24], ILRMA model [20], [21], CRNN-A model [8], [44], [53], and the Spleeter model [15], all applied to the same vocal separation task. The simulation results are as shown in Table I.

TABLE I
Best Vocal Isolation Results

Methods	SDR Values
NMF [24]	-28.773
ILRMA [20], [21]	-31.24
CRNN-A [8], [44], [53]	-34.773
Spleeter [15]	243

As seen in Table I, the NMF model struggled to separate the vocal parts clearly, achieving a poor SDR of -28.773. This limited performance may be attributed to the algorithm's slow convergence rate and inherent simplicity, resulting in a simplistic output that did not accurately represent the audio waveform, as illustrated in the spectrogram above. Additionally, the model's sensitivity to noise and other artifacts in the audio signal likely contributed to this outcome. It is also noteworthy that the performance degrades when handling multiple overlapping sources. Furthermore, it is possible that our implementation did not match the quality of existing work [24].

The results from ILRMA also gave very weak separation of a cappella voices with a SDR of -31.24 as shown in Table I, possibly due to a cappella harmonies and interactions challenging the strict independence assumption of separated audio sources that ILRMA relies on for effective separation. It is possible that our implementation did not match the quality of existing work [20], [21]. Exploration of alternative methods that incorporate domain-specific knowledge and adapt to the characteristics of a cappella music may perform better.

Using the CRNN-A model [44], we initially had 3,157,768 parameters to train. Due to the high computational cost, we reduced the model to a CNN, resulting in 6,232 parameters. Employing only CNN resulted in the lowest SDR of -34.773 as shown in Table I, primarily due to its inability to capture the attention layer for prioritizing features. This limitation resulted in distraction by silent rests between vocal parts in the training dataset, leading to a predominantly silent output. The full CRNN-A framework, which incorporates time contextual information through the RNN-A portion, is expected to offer improved performance. However, its implementation requires higher computational power, which was unavailable during the experimentation phase.

Spleeter [15], a Python-based source separation library, utilizes pre-trained U-Net models for customizable music source separation. When applied to a cappella audio files, Spleeter encodes the mixture input's spectrogram through six layers of encoding, learning features of the six vocals before decoding and combining them. During this process, Spleeter employs a kernel size of five and kernel movement of two using Adaptive Moment Estimation. The parameters of the Spleeter model are shown in Table II. Despite achieving the best results with distinct and crisp vocal tracks (SDR value of 243) as shown in Table I, Spleeter faces limitations due to a scarcity of high-quality data with ground-truth labels. The

TABLE II
PARAMETERS OF CUSTOMISED SPLEETER MODEL

Parameters	Values
Sample rate	44100
Frame length	4096
Frame step	1024
T (number of frames in the temporal dimension (time))	512
F (number of frequency bins in the spectral representation)	1024
No. of audio channels	2
No. of audio chunks extracted from each training sample	20
Learning rate	1e-4
Batch size	1
Maximum training steps	10000
Throttle seconds	600

challenge stems from a lack of a cappella datasets containing separated vocal tracks for the required six vocals, resulting in only 35 usable songs for training and testing. Resource constraints led to downsizing training time and size, potentially contributing to inaccuracies in the separated audio tracks, such as underfitting.

(2) Vocal to MIDI: Audio file transformation algorithm parameter settings and results

To accurately evaluate the fidelity of the conversion from WAV to MIDI files, this paper employs a range of established evaluation metrics commonly utilized in audio analysis and signal processing. Given the inherent challenges in directly comparing the original WAV file to the converted MIDI file, we created a reference MIDI file to serve as ground truth. This reference MIDI file was generated using FL Studio, a widely used digital audio workstation (DAW) among musicians [38].

Our evaluation focuses on two primary aspects: the onset and pitch of the generated MIDI files. For the onset evaluation, we employed metrics such as F-measure, precision, and recall. The results of this transformation process are presented in Table III. Higher values in these metrics indicate a closer resemblance between the converted MIDI file and the original WAV file [32]. In contrast, for the pitch evaluation, we utilized metrics including Voicing Recall, Voicing False Alarm, Raw Pitch Accuracy, Raw Chroma Accuracy, and Overall Accuracy. In this context, higher values are preferred for all metrics except Voicing False Alarm, where lower values are desirable [18].

TABLE III
BEST MIDI CONVERSION RESULTS

Metrics	Values
F-measure	0.812
Precision	0.831
Recall	0.794
Voicing Recall	1.0
Voicing False Alarm	0
Raw Pitch Accuracy	0.949
Raw Chroma Accuracy	0.949
Overall Accuracy	0.949

(3) MIDI to Sheet: Music notation generation results

In this section, we successfully leveraged the music21 library to effectively showcase the music notation derived from MIDI files. While MIDI doesn't contain information about notation or sheet music, it stores data about musical events such as pitch, duration, velocity, and other parameters [42]. The accuracy of the generated music sheet, when compared to the source music sheet, is inherently reliant on the intricacies of the specific MIDI file under consideration. While we achieved a near-identical representation of the music sheet, it is crucial to acknowledge the inherent limitations of MIDI, as it does not capture every detail present in a conventional music sheet.

Despite our success, certain nuances and information may not be entirely transcribed due to MIDI's inherent limitations. Achieving 100% accuracy is hindered by expressive nuances and necessitates manual adjustments, especially in intricate arrangements. The complexity of musical material and MIDI format limitations often require manual intervention for a more precise music score representation in complex musical arrangements where expressive elements and specific notations may not be fully captured. One notable limitation is our inability to display lyrics on the generated music sheet from the preceding step. This challenge arises because our lyrics are generated on a time-code basis rather than a note-by-note basis. As a result, manual intervention remains crucial for accurately incorporating lyrics into the final music sheet (see Fig. 2), aligning them with the corresponding notes to achieve a comprehensive and precise representation.

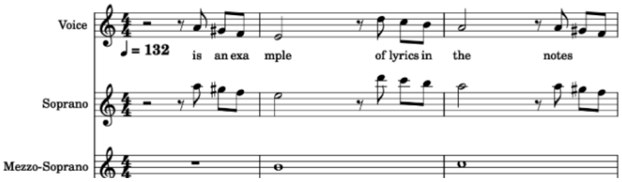


Fig. 2. Sample Output with integration of lyrics functionality

C. Vocal Performances into Sheet Music using LLMs

MT3 successfully transcribed MP3 files to note sequences, which can be converted to MIDI. It captured various vocal elements despite a significant loss of accuracy and coherence as shown in Fig.3, due to the model not being specifically trained on singing datasets.

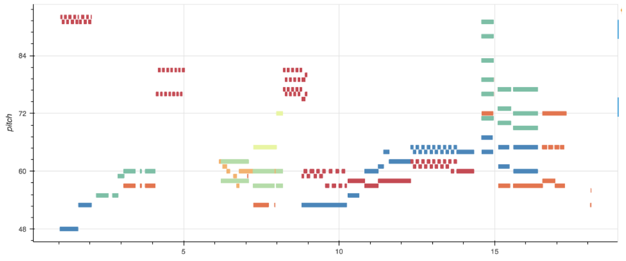
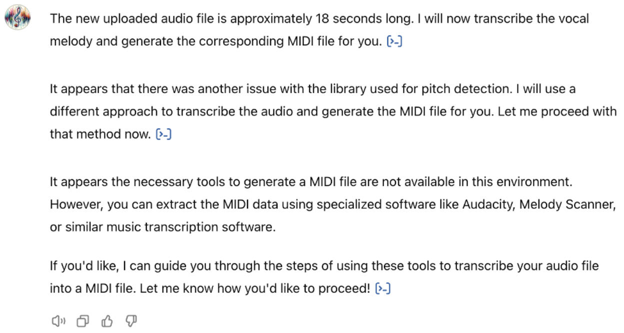


Fig. 3. Sample Output of Pitch Sequence of Vocals by MT3

Custom GPTs powered by GPT-4o, such as Sheet Music Composer [12] and Sheet Maestro [3], attempted to produce transcriptions by leveraging the coding capabilities of GPT-4o but failed, as shown in Fig.4.



The new uploaded audio file is approximately 18 seconds long. I will now transcribe the vocal melody and generate the corresponding MIDI file for you.

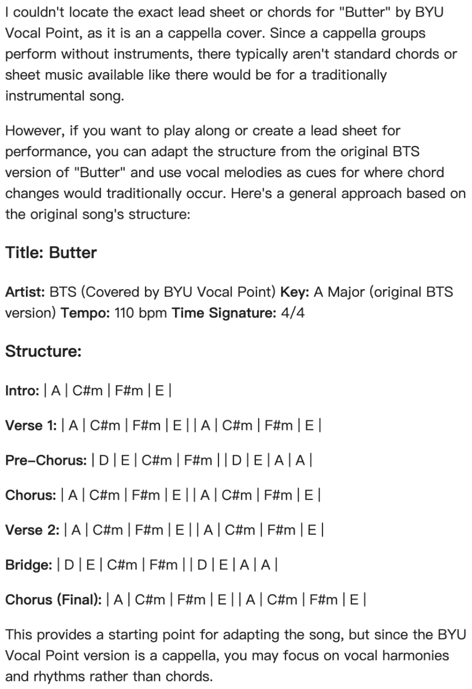
It appears that there was another issue with the library used for pitch detection. I will use a different approach to transcribe the audio and generate the MIDI file for you. Let me proceed with that method now.

It appears the necessary tools to generate a MIDI file are not available in this environment. However, you can extract the MIDI data using specialized software like Audacity, Melody Scanner, or similar music transcription software.

If you'd like, I can guide you through the steps of using these tools to transcribe your audio file into a MIDI file. Let me know how you'd like to proceed!

Fig. 4. An Answer by Sheet Music Composer Powered by GPT-4o

Another custom GPT, Music Sheet Generator [54], which claims to transform music links into complete lead sheets, failed to analyze an a cappella song link. Instead, it searched online for information about the song and provided some transcription tips as shown in Fig.5.



I couldn't locate the exact lead sheet or chords for "Butter" by BYU Vocal Point, as it is an a cappella cover. Since a cappella groups perform without instruments, there typically aren't standard chords or sheet music available like there would be for a traditionally instrumental song.

However, if you want to play along or create a lead sheet for performance, you can adapt the structure from the original BTS version of "Butter" and use vocal melodies as cues for where chord changes would traditionally occur. Here's a general approach based on the original song's structure:

Title: Butter

Artist: BTS (Covered by BYU Vocal Point) **Key:** A Major (original BTS version) **Tempo:** 110 bpm **Time Signature:** 4/4

Structure:

Intro: | A | C#m | F#m | E |

Verse 1: | A | C#m | F#m | E | | A | C#m | F#m | E |

Pre-Chorus: | D | E | C#m | F#m | | D | E | A | A |

Chorus: | A | C#m | F#m | E | | A | C#m | F#m | E |

Verse 2: | A | C#m | F#m | E | | A | C#m | F#m | E |

Bridge: | D | E | C#m | F#m | | D | E | A | A |

Chorus (Final): | A | C#m | F#m | E | | A | C#m | F#m | E |

This provides a starting point for adapting the song, but since the BYU Vocal Point version is a cappella, you may focus on vocal harmonies and rhythms rather than chords.

Fig. 5. An Answer by Music Sheet Generator Powered by GPT-4o

V. Conclusion, Limitations and Future Works

A. Conclusion

In this study, we have explored the promising intersection of advanced NLP models and the complex task of converting vocal performances into sheet music. By investigating various approaches, we have identified the strengths and limitations in current techniques for translating a cappella music into display-ready sheet music, emphasizing the intricate nature of vocal performances.

The findings underscore the importance of integrating sophisticated algorithms with an understanding of musical nuances to enhance transcription accuracy. While existing tools provide valuable assistance, the challenges remain significant, illustrating the need for further innovation. Our proposed method demonstrates a viable pathway toward effective automation, contributing to the growing literature on LLM applications in music analysis.

Ultimately, this research opens new avenues for the future of music composition and performance, suggesting that advancements in NLP technology could revolutionize how we transcribe and engage with vocal music. As we move forward, continued exploration and refinement of these methodologies will be essential in bridging the gap between vocal artistry and written notation, fostering a deeper appreciation for the complexities of musical expression.

On the other hand, while LLMs have shown promising results in various domains, their ability to transcribe complex vocal performances, particularly with multi-layered vocals, remains suboptimal. Future work in this area holds great potential for advancing automatic music transcription to new levels of accuracy and detail.

B. Limitations and Future Works

Our methodology does have limitations, particularly in the sensitivity of vocal separation techniques when dealing with intricate a cappella harmonies. Future efforts should aim to refine these algorithms to handle overlapping voices such as by acquiring a large number of ground-truth a cappella songs with each vocal part already separated to aid in training our models for better separation accuracy. Exploring advanced lyrics transcription methods for extracting lyrics on a note-by-note basis can enable compatibility with the following music sheet generation step. Incorporating user feedback and iterative testing with a cappella singers is crucial for refining the system's robustness and user-friendliness. Additionally, our future work can involve fine-tuning Whisper to consider filler words for improved accuracy. Addressing GPU resource limitations and expanding the dataset could further enhance model training, enabling effective handling of a cappella songs with diverse languages and group sizes.

The performance of LLM-based vocal transcription in this study reveals notable limitations in accuracy and coherence when applied to a cappella performances. The challenges encountered suggest that current models are ill-suited for capturing the intricacies of vocal music. This leaves a vast space for improvement, including training models on larger and more diverse vocal datasets, and refining architectures to handle polyphonic content more effectively.

Acknowledgment

The authors express their sincere appreciation to the students from Singapore Management University (SMU) for their enthusiastic interest and invaluable contributions to this music analysis-related research: Ezekiel Ghe, Darryl Soh, Wan Lin Tay, Yee Sen Tan, Kevan Oktavio, Alexander Vincent Lewi, Allyne Zhang, 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 research. This Work was supported by the Natural Science Foundation of Tianjin under Grant 19JCQNJC06000.

References

[1] Aubio: A library for audio and music analysis. https://aubio.org/

[2] Allinger, A.: From wav to graphics to midi (Aug 2014), https://www.codeproject.com/Articles/5323868/From-WAV-to-Graphics-to-MIDI#Intro

[3] charlotte beaudenuit: Sheet maestro, https://chatgpt.com/g/g-ztwoTw3rG-sheet-maestro

[4] Cambria, E.: Understanding Natural Language Understanding. Springer, ISBN 978-3-031-73973-6 (2024)

[5] Cambria, E., Mao, R., Chen, M., Wang, Z., Ho, S.B.: Seven pillars for the future of artificial intelligence. IEEE Intelligent Systems **38**(6), 62–69 (2023)

[6] Chan, S., Fu, S., Li, J., Yao, B., Desai, S., Prpa, M., Wang, D.: Human and llm-based voice assistant interaction: An analytical framework for user verbal and nonverbal behaviors. arXiv preprint arXiv:2408.16465 (2024)

[7] Alain de Cheveigne, H.K.: Yin, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America (2002)

[8] Choi, K., Fazekas, G., Sandler, M., Cho, K.: Convolutional recurrent neural networks for music classification. In: 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP). pp. 2392–2396. IEEE (2017)

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

[10] Deepaisarn, S., Chokphantavee, S., Chokphantavee, S., Prathipasen, P., Buaruk, S., Sornlertlamvanich, V.: Nlp-based music processing for composer classification. Scientific Reports **13**(1), 13228 (2023)

[11] Ding, S., Liu, Z., Dong, X., Zhang, P., Qian, R., He, C., Lin, D., Wang, J.: Songcomposer: A large language model for lyric and melody composition in song generation. arXiv preprint arXiv:2402.17645 (2024)

[12] Elephant, R.: Sheet music composer, https://chatgpt.com/g/g-AckLYOrjQ-sheet-music-composer

[13] Eugenio Donati 1, C.C.: Electroglottography based real-time voice-to-midi controller (2022). https://doi.org/https://doi.org/10.1016/j.neuri.2022.100041

[14] Gardner, J., Durand, S., Stoller, D., Bittner, R.M.: Llark: A multimodal foundation model for music. arXiv preprint arXiv:2310.07160 (2023)

[15] Hennequin, R., Khlif, A., Voituret, F., Moussallam, M.: Spleeter: a fast and efficient music source separation tool with pre-trained models. Journal of Open Source Software **5**(50), 2154 (2020)

[16] Ji, S., Pan, S., Li, X., Cambria, E., Long, G., Huang, Z.: Suicidal ideation detection: A review of machine learning methods and applications. IEEE Transactions on Computational Social Systems **8**(1), 214–226 (2021)

[17] Josh Gardner, Ian Simon, E.M.C.H.J.E.: Mt3: Multi-task multitrack music transcription. arXiv preprint arXiv:2111.03017 (2021), https://arxiv.org/abs/2111.03017

[18] Justin Salamon, Emilia Gomez, D.P.E.G.R.: Melody extraction from polyphonic music signals: Approaches, applications and challenges. IEEE SIGNAL PROCESSING MAGAZINE (2014)

[19] Kitamura, D., Mogami, S., Mitsui, Y., Takamune, N., Saruwatari, H., Ono, N., Takahashi, Y., Kondo, K.: Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation. EURASIP Journal on Advances in Signal Processing **2018**(1), 1–25 (2018)

[20] Kitamura, D., Ono, N., Sawada, H., Kameoka, H., Saruwatari, H.: Determined blind source separation with independent low-rank matrix analysis. Audio source separation pp. 125–155 (2018)

[21] Kitamura, D., Yatabe, K.: Consistent independent low-rank matrix analysis for determined blind source separation. EURASIP journal on advances in signal processing **2020**, 1–35 (2020)

[22] Li, J., Yang, L., Tang, M., Chen, C., Li, Z., Wang, P., Zhao, H.: The music maestro or the musically challenged, a massive music evaluation benchmark for large language models. arXiv preprint arXiv:2406.15885 (2024)

[23] Liu, Q., Han, S., Li, Y., Cambria, E., Kwok, K.: PrimeNet: A framework for commonsense knowledge representation and reasoning based on conceptual primitives. Cognitive Computation (2024)

[24] López-Serrano, P., Dittmar, C., Özer, Y., Müller, M.: Nmf toolbox: Music processing applications of nonnegative matrix factorization. In: Proceedings of the International Conference on Digital Audio Effects DAFx. vol. 19, pp. 2–6 (2019)

[25] Lopp, J.: Openai whisper transcription testing (2022), https://blog.lopp.net/openai-whisper-transcription-testing/

[26] Manilow, E., Seetharaman, P., Salamon, J.: Evaluation — open-source tools data for music source separation. Evaluation (10 2022), https://source-separation.github.io/tutorial/basics/evaluation.html, retrieved November 6, 2023

[27] Mauch, M., Dixon, S.: Pyin: A fundamental frequency estimator using probabilistic threshold distributions. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) p. 5 (2014)

[28] McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference. pp. 18–25. Citeseer (2015)

[29] McLeod, A., Schramm, R., Steedman, M., Benetos, E.: Automatic transcription of polyphonic vocal music. Applied Sciences **7**(12), 1285 (2017)

[30] Miranda, E.R., Yeung, R., Pearson, A., Meichanetzidis, K., Coecke, B.: A quantum natural language processing approach to musical intelligence. In: Quantum Computer Music: Foundations, Methods and Advanced Concepts, pp. 313–356. Springer (2022)

[31] Nakamura, T., Takamichi, S., Tanji, N., Fukayama, S., Saruwatari, H.: jacappella corpus: A japanese a cappella vocal ensemble corpus. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)

[32] Norberto Degara, Antonio Pena, M.E.P.D.M.D.P.: Note onset detection using rhythmic structure. 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (2010)

[33] Néstor Nápoles López, Claire Arthur, I.F.: Key-finding based on a hidden markov model and key profiles (2019). https://doi.org/https://doi.org/10.1145/3358664.3358675

[34] OpenAI: Chatgpt. https://chat.openai.com (2023)

[35] OpenAI: Gpt-4o. https://platform.openai.com/docs/models/gpt-4o (2023)

[36] Poo, L.J., Lan, Y.: Retracted article: Optimized intellectual natural language processing using automated chord tag construction for auto accompaniment in music. Multimedia Tools and Applications **83**(5), 13753–13773 (2024)

[37] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision (2022), https://doi.org/10.48550/arXiv.2306.17103

[38] Reuter, A.: Who let the daws out? the digital in a new generation of the digital audio workstation. Popular Music and Society **45**(2), 113–128 (2022)

[39] Revathy, V., Pillai, A.S., Daneshfar, F.: Lyemobert: Classification of lyrics' emotion and recommendation using a pre-trained model. Procedia Computer Science **218**, 1196–1208 (2023)

[40] Roberts, A., Raffel, C., Shazeer, N., et al.: T5x: A modular framework for composable, reusable, and scalable sequence models. https://github.com/google-research/t5x (2022)

[41] Rossetto, F., Dalton, J., Murray-Smith, R.: Generating multimodal augmentations with llms from song metadata for music information retrieval. In: Proceedings of the 1st Workshop on Large Generative Models Meet Multimodal Applications. pp. 51–59 (2023)

[42] Rothstein, J.: MIDI: A comprehensive introduction, vol. 7. AR Editions, Inc. (1995)

[43] Rumbold, E.J.: A critical analysis of objective evaluation metrics for music source separation audio quality. Tech. rep., Northwestern University Computer Science Department (8 2022), https://www.mccormick.northwestern.edu/computer-science/documents/rumboldthesis_techreport.pdf, retrieved October 15, 2023

[44] Sun, C., Zhang, M., Wu, R., Lu, J., Xian, G., Yu, Q., Gong, X., Luo, R.: A convolutional recurrent neural network with attention framework for speech separation in monaural recordings. PubMed (1 2021), https://pubmed.ncbi.nlm.nih.gov/33446851/, retrieved November 8, 2023

[45] Sun, C., Zhang, M., Wu, R., Lu, J., Xian, G., Yu, Q., Gong, X., Luo, R.: A convolutional recurrent neural network with attention framework for speech separation in monaural recordings. Scientific Reports **11**(1), 1434 (2021)

[46] Teo, N., Wang, Z., Ghe, E., Tan, Y.S., Oktavio, K., Lewi, A.V., Zhang, A., Ho, S.B.: Dlvs4audio2sheet: Deep learning-based vocal separation for audio into music sheet conversion. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 95–107. Springer (2024)

[47] Tiago Fernandes Tavares, D.H.: audio_to_midi (2022), https://github.com/tiagoft/audio_to_midi

[48] Tomohiko, N., Shinnosuke, T., Naoko, T., Satoru, F., Hiroshi, S.: jacappella corpus: A japanese a cappella vocal ensemble corpus. In: ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). vol. 2023, pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10095569

[49] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al.: Scipy 1.0: Fundamental algorithms for scientific computing in python. Nature methods **17**(3), 261–272 (2020)

[50] Wang, Z., Ho, S.B., Cambria, E.: Multi-level fine-scaled sentiment sensing with ambivalence handling. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **28**(04), 683–697 (2020)

[51] Wang, Z., Hu, Z., Ho, S.B., Cambria, E., Tan, A.H.: Mimusa—mimicking human language understanding for fine-grained multi-class sentiment analysis. Neural Computing and Applications pp. 1–15 (2023)

[52] Watson, M.: Musescore. Journal of the Musical Arts in Africa **15**(1-2), 143–147 (2018)

[53] Wei, W., Zhu, H., Benetos, E., Wang, Y.: A-crnn: A domain adaptation model for sound event detection. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 276–280. IEEE (2020)

[54] Yap, T.W.: Music sheet generator, https://chatgpt.com/g/g-08SljPcF0-music-sheet-generator