Lyrics Transcription in Western Classical Music with Whisper: A Case Study on Schubert's Winterreise

Hans-Ulrich Berendes and **Simon Schwär** and **Meinard Müller**
International Audio Laboratories Erlangen

Abstract

Automatic Lyrics Transcription (ALT) aims to transcribe sung words from music recordings and is closely related to Automatic Speech Recognition (ASR). Although not specifically designed for lyrics transcription, the state-of-the-art ASR model Whisper has recently proven effective for ALT and various related tasks in music information retrieval (MIR). This paper investigates Whisper's performance on Western classical music, using the "Schubert Winterreise Dataset." In particular, we found that the average Word Error Rate (WER) with the unmodified Whisper model is 0.56 for this dataset, while the performance varies greatly across songs and versions. In contrast, spoken versions of the song lyrics, which we recorded, are transcribed with a WER of 0.14. Further systematic experiments with source separation and time-scale modification techniques indicate that Whisper's accuracy in lyrics transcription is less affected by the musical accompaniment and more by the singing style.

1 Introduction

Lyrics, the words of a song, are vital to vocal music. They contain important information for listeners and bridge the gap between music and language. Automatic Lyrics Transcription (ALT) extracts these words, often from a mix of instruments and vocals (Tsai et al., 2018). Automatic Speech Recognition (ASR) performs a similar task for normal speech (Malik et al., 2021). While both involve processing the human voice, speech, and singing differ in pitch fluctuations, pronunciation, speed, time variations, and vocabulary (Humphrey et al., 2019). Musical accompaniment can further complicate ALT, as it superimposes the singing voice, often with high temporal and spectral correlations (Gupta et al., 2020). Due to these differences, ASR and ALT have long been considered separate tasks (Kruspe, 2024).

Recent ASR advances rely on large, diverse datasets and often use weakly-supervised or self-supervised training (Baevski et al., 2020; Peng et al., 2024). One state-of-the-art model, Whisper, is trained on a total of 5 million hours of data (Radford et al., 2023). Trained on such extensive data, Whisper shows promising capabilities for ALT as well. It can either be used without modifications (Cífka et al., 2023), in combination with a Large Language Model (LLM) for transcript post-processing (Zhuo et al., 2023) or be fine-tuned on specific music genres (Wang et al., 2023). Understanding large pre-trained models is crucial, as these models can be useful for tasks with limited data like ALT, in particular for underrepresented languages or genres (Latif et al., 2023; Wang et al., 2024).

This paper aims to better understand Whisper's ALT performance and the challenges of transcribing singing compared to speech. Different from the other works mentioned above, we focus on Western classical music. In particular, we use the "Schubert Winterreise Dataset" (SWD) (Weiß et al., 2021) as a case study, which comprises nine complete recordings of the 24-song cycle "Winterreise" by Franz Schubert. The Winterreise is composed for solo voice with piano accompaniment, based on German poems from the early 19th century.

Our contributions are twofold: an in-detail analysis of Whisper's ALT performance on the SWD, and a comparison of speech and singing transcription through experiments with spoken versions of the lyrics, source separation, and time-scale modification.

2 Experimental Setup

2.1 Whisper

The multilingual ASR model Whisper, introduced by Radford et al. (2023), is based on a transformer architecture and available in various sizes. In this work, we use the largest and latest pre-trained version, `large-v3` [1]. For simplicity, we refer to this model as Whisper. It has been trained on 4 million hours of unlabeled data and 1 million hours of weakly-supervised data, both not publicly available. Despite being tailored for ASR, there are indications that music is included to some extent in the training data (Zhuo et al., 2023). Although there has been work on improving Whisper for ALT (Zhuo et al., 2023; Wang et al., 2023, 2024), we use the model in its original state to better understand its behavior and potentially evaluate differences between speech and singing.

2.2 Evaluation Dataset

The SWD (Weiß et al., 2021), contains nine commercial recordings of all 24 songs of the Winterreise. These versions feature different male singers, pianos, acoustic conditions, and audio quality. The total number of words per version in the lyrics is 2644. In the following, we denote the songs using their respective number ranging from `SWD-01` to `SWD-24`. Following the dataset paper, we denote the versions with a two-letter identifier alongside the recording year, e.g., `AL98`. For more details on the versions, see Weiß et al. (2021). Since Whisper's training data is not public, we cannot ensure that there is no overlap with the publicly available SWD. We use this dataset because we consider the classical singing style together with the accompaniment to be a challenging scenario for an ASR system. Additionally, the SWD enables cross-version analysis by offering multiple performances of each piece.

2.3 Evaluation Metrics

The most commonly used metric to measure the accuracy of ASR and ALT is the Word Error Rate (WER) (Malik et al., 2021). Given a reference text and a transcript, it is defined as

$$\mathrm{WER} = \frac{D + I + S}{R}, \tag{1}$$

where D is the number of deletions, I the number of insertions, S the number of substitutions, and R the number of words in the reference text. The WER can exceed 1 when a transcript has more words than its reference. While our focus lies on the WER, we additionally report the Character Error Rate (CER) for a more fine-grained analysis. It is defined similarly to the WER but on a character level, rather than a word level. To ensure consistency, we standardize both the reference and transcript texts by removing all punctuation and capitalization before calculating the metrics. Considering the stochastic decoding in the Whisper model, we average the metrics over five independent trials to ensure result stability, as done in Cífka et al. (2023). We will briefly discuss the impacts of this in Section 3.

3 Lyrics Transcription Results

In this section, we evaluate the transcription performance of Whisper for singing with accompaniment. Figure 1 shows the WER of the Whisper transcription of SWD for each song and each version, along with the respective averages. The overall mean WER is $\mu = 0.56$ but we can see considerable differences, both across songs and versions with an overall standard deviation of $\sigma = 0.234$.

3.1 Results across Versions

We first investigate the differences between versions. The average WER varies from $\mu_{\texttt{FI66}} = 0.49$ to $\mu_{\texttt{AL98}} = 0.64$, an absolute difference of up to 0.15 for the same songs and lyrics. The standard deviation is $\sigma_{\text{version}} = 0.044$. Notably, the oldest recording `HU33` (with the worst audio quality) has a mean WER of $\mu_{\texttt{HU33}} = 0.54$, just below the average, indicating Whisper's robustness against poor audio quality (Radford et al., 2023). No version consistently gives better or worse results. For example, `FI66` has the lowest average WER but shows the highest WER of 0.46 for `SWD-02` and the lowest WER of 0.24 for `SWD-05`.

3.2 Results across Songs

Next, we examine WER variations across songs. The mean WER (across versions) ranges from $\mu_{\texttt{SWD-02}} = 0.29$ to $\mu_{\texttt{SWD-21}} = 0.98$, an absolute difference of 0.69. The standard deviation of per-song averages is $\sigma_{\text{song}} = 0.148$, larger than $\sigma_{\text{version}} = 0.044$ mentioned above.

For deeper insight, we examine songs `SWD-02` and `SWD-21`. Musically, `SWD-02` features a fast tempo with subtle piano accompaniment, mainly supporting the voice. Figure 2 shows the lyrics of the first two stanzas of `SWD-02` alongside the corresponding transcript. Many errors are substitutions, e.g., “Wetterfahne” becomes “Wetterfalle”. Whisper also struggles with compound words, e.g.,

[1] Available at https://github.com/openai/whisper/

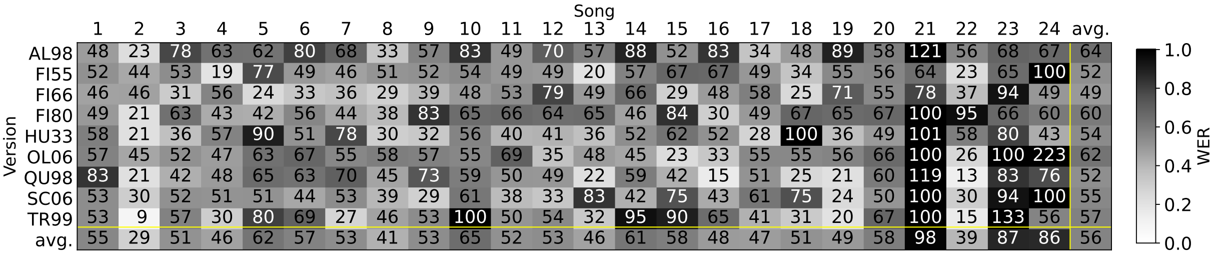

Figure 1: WER of each song and version in SWD, sorted by averages over songs and versions. For better visibility, the numbers are given in 100·WER.

"Liebchens Haus" is written as "Liebchenshaus", and "nimmer" is split into "nie mehr", semantically equivalent in German.

SWD-21 has the highest average WER, with seven out of nine versions showing a WER of 1.0 or higher. In these instances, Whisper often fails to produce meaningful transcriptions. The song has a slow tempo with long piano-only sections. Transcripts frequently contain irrelevant text, such as music descriptors ("Piano Music") or unrelated phrases ("Thank you for listening"), an issue already previously documented (Cífka et al., 2023; Zhuo et al., 2023).

Reference	Transcript
Der Wind spielt mit der Wetterfahne auf meines schönen Liebchens Haus. Da dacht ich schon in meinem Wahne, sie pfiff' den armen Flüchtling aus.	Der Wind spielt mit der Wetterfalle auf meine schöne Liebchenshaus. Verdacht ich schon in meinem Wale, sie pfiff den armen Flüchtling aus.
Er hätt es eher bemerken sollen, des Hauses aufgestecktes Schild, so hätt er nimmer suchen wollen im Haus ein treues Frauenbild.	Er hätte sicher bemerken sollen, des Hauses aufgestellte Schild. So hätte er nie mehr suchen wollen, im Haus ein treues Frauenbild.

Figure 2: Comparison between the first two stanzas of SWD-02 (reference text on the left) and a Whisper-generated transcript (on the right), with errors highlighted in red. The WER of this excerpt is 0.33.

3.3 Discussion

The average WER of 0.56 on the SWD is considerably higher compared to speech benchmark datasets for long-form transcription, which are in the range of 0.04 to 0.2 (Radford et al., 2023). Cífka et al. (2023) utilized Whisper for ALT with a variety of modern genres, including rock and pop music, and reported a WER of 0.36, which is still significantly lower than our results. This suggests that the music in the SWD presents a more challenging task compared to rock and pop music. Although it is difficult to reason about errors of black box systems like Whisper, we hypothesize that some errors, e.g., seen in Figure 2, can be attributed to the poetic style and old language.

Whisper's stochastic decoding introduces noise, leading to some uncertainty in our results. Averaging over five trials, the average standard deviation is 0.13, with a confidence of 0.06 for the mean WER of a single track. We argue that this is sufficiently small to maintain the validity of our observations.

4 Comparative Analysis of Speech and Singing Transcription

In the previous section, we have seen that the ALT performance of Whisper for the classical music dataset SWD is low compared to speech benchmarks. In this section, we further explore this difference, by investigating possible factors that may deteriorate the transcription performance from speech to classical singing.

4.1 Influence of Musical Accompaniment

One major difference between ASR and ALT is the musical accompaniment, which acts as a correlated "background noise" when transcribing singing. The varying accompaniment could potentially account for WER differences between ALT and ASR datasets, as well as differences between the songs in the SWD. To test this hypothesis, we employ Musical Source Separation (MSS) to extract vocal tracks of the SWD, which we denote with V-MSS. For source separation, we use the commercial system provided by the company *AudioShake*, further denoted by V-MSS$_{\mathrm{AS}}$, as well as the open-source model hybrid Demucs introduced in Défossez (2021), denoted by V-MSS$_{\mathrm{HDMC}}$. In Table 1 we report the respective WERs and CERs. The MSS pre-processing does not improve the results significantly, which aligns with previous findings by Cífka et al. (2023). This indicates that the musical accompaniment is not the primary source of errors. However, small artifacts introduced by the MSS algorithms could be detrimental to the transcription performance and more work on clean multi-track data could give more insight into this

	MIX	V-MSS$_{AS}$	V-MSS$_{HDMC}$	V-SP
WER [%]	56.1	54.1	55.6	14.6
CER [%]	44.3	42.4	43.9	9.4

Table 1: WER and CER for the three signal types: unprocessed polyphonic input (MIX), vocals extracted with MSS (V-MSS) with further indication of the used MSS system, and the spoken version of the lyrics (V-SP).

aspect. The robustness against musical accompaniment might not hold for models other than Whisper, since previous work has shown that jointly training an MSS system with a lyrics transcriber model can be beneficial (Gao et al., 2023).

4.2 Sung vs. Spoken Lyrics

There is a plethora of work, comparing the acoustic differences between speech and singing, e.g., List (1963); Patel et al. (2006); Gao et al. (2018); Vanden Bosch der Nederlanden et al. (2023). We want to directly compare these two domains in terms of Whisper's respective transcription accuracy. To this end, we recorded spoken versions of the song lyrics for the SWD, which we denote with V-SP. Our recordings feature two speakers, male and female, both native German speakers.

Table 1 shows the WER and CER for the spoken lyrics (V-SP). The WER for V-SP is 0.146 and the CER is 0.094 and therefore considerably lower, compared to the original SWD. Therefore we can rule out the distinct vocabulary of the SWD as the single source of errors. Since MSS pre-processing did not improve the results, we hypothesize that singing itself, and particularly the classical singing style in the SWD, poses a challenge for Whisper. One distinct difference between speech and singing is the duration of individual phonemes (Kruspe, 2024). To investigate the influence of this, we apply time-scale modification to the spoken lyrics V-SP, using the `libtsm` Python package[2], based on Driedger and Müller (2014). Note that this introduces artifacts, which grow more noticeable with stronger modification, however, the pitch is not changed. Each time-scale modified signal is characterized by a single time stretch factor, where a value smaller than 1 denotes a higher speed compared to the original signal. Figure 3 shows the average WER across various time stretch factors.

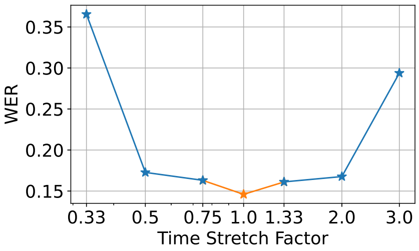

Figure 3: Average WER for time-stretched speech signals across various time stretch factors. The orange color denotes the unaltered speech.

The transcription performance decreases for very high or low time-stretch factors but remains fairly robust to small changes. This suggests that strong deviations from normal speech are problematic for Whisper, which seems reasonable given its speech-focused training. Quantifying phoneme duration deviations from normal speech depends on the music genre and language, but stretching factors of 3 are common for vowels (Duan et al., 2013; de Medeiros and Cabral, 2018). Our time-stretching experiment may explain why the SWD is challenging for Whisper. Further experiments, analyzing correlations between errors and stretched phonemes could help adapt ASR models to singing.

5 Conclusion and Future Work

Our study investigates Whisper's ALT performance on Western classical music using the SWD as a case study. We find a higher WER for the SWD compared to speech or other singing datasets, with significant fluctuations across songs and versions. Vocabulary has a minor impact, as spoken lyrics WER is comparable to other speech benchmarks. MSS-based vocal extraction has a negligible influence on the WER, indicating musical accompaniment is also not the primary issue. Preliminary experiments show that Whisper is robust against small speed variations but sensitive to larger variations in talking speed compared to normal speech.

We hope our study serves as a starting point for analyzing how characteristics of speech and singing influence ASR model performance. Our evaluation methodology, though applied to Whisper, is relevant beyond a single model. Applying this approach to other ASR models, such as Peng et al. (2024), could enhance understanding of their behavior. This perspective positions our work as a case study for evaluating large audio models and highlights the potential of the music domain for thorough analysis of pre-trained models.

[2] `https://github.com/meinardmueller/libtsm`

Acknowledgements: This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant No. 328416299 (DFG MU 2686/10-2) and Grant No. 401198673 (DFG MU 2686/13-2). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Ondřej Cífka, Constantinos Dimitriou, Cheng-i Wang, Hendrik Schreiber, Luke Miner, and Fabian-Robert Stöter. 2023. Jam-ALT: A Formatting-Aware Lyrics Transcription Benchmark. *arXiv*, abs/2311.13987.

Beatriz Raposo de Medeiros and Joao Paulo Cabral. 2018. Acoustic distinctions between speech and singing: Is singing acoustically more stable than speech? In *Proceedings of the International Conference on Speech Prosody*, pages 542–546, Poznań, Poland.

Alexandre Défossez. 2021. Hybrid spectrogram and waveform source separation. In *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, pages 1–13, Online.

Jonathan Driedger and Meinard Müller. 2014. TSM Toolbox: MATLAB implementations of time-scale modification algorithms. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 249–256, Erlangen, Germany.

Zhiyan Duan, Haotian Fang, Bo Li, Khe Chai Sim, and Ye Wang. 2013. The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2013, Kaohsiung, Taiwan, October 29 - November 1, 2013*, pages 1–9.

Xiaoxue Gao, Chitralekha Gupta, and Haizhou Li. 2023. Polyscriber: Integrated fine-tuning of extractor and lyrics transcriber for polyphonic music. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:1968–1981.

Xiaoxue Gao, Berrak Sisman, Rohan Kumar Das, and Karthika Vijayan. 2018. NUS-HLT Spoken Lyrics and Singing (SLS) Corpus. In *International Conference on Orange Technologies (ICOT)*, pages 1–6, Nusa Dua, Indonesia.

Chitralekha Gupta, Emre Yılmaz, and Haizhou Li. 2020. Automatic lyrics alignment and transcription in polyphonic music: Does background music help? In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 496–500, Barcelona, Spain.

Eric J. Humphrey, Sravana Reddy, Prem Seetharaman, Aparna Kumar, Rachel M. Bittner, Andrew Demetriou, Sankalp Gulati, Andreas Jansson, Tristan Jehan, Bernhard Lehner, Anna Krupse, and Luwei Yang. 2019. An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music. *IEEE Signal Processing Magazine*, 36(1):82–94.

Anna Kruspe. 2024. More than words: Advancements and challenges in speech recognition for singing. In *Conference on Electronic Speech Signal Processing (ESSV) Keynote*, Regensburg, Germany.

Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Heriberto Cuayáhuitl, and Björn W Schuller. 2023. Sparks of large audio models: A survey and outlook. *arXiv*, abs/2308.12792.

George List. 1963. The boundaries of speech and song. *Ethnomusicology*, 7(1):1–16.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80:9411–9457.

Aniruddh D Patel, John R Iversen, and Jason C Rosenberg. 2006. Comparing the rhythm and melody of speech and music: The case of british english and french. *The Journal of the Acoustical Society of America*, 119(5):3034–3047.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, et al. 2024. Owsm v3. 1: Better and faster open whisper-style speech models based on e-branchformer. *arXiv*, abs/2401.16658.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 28492–28518.

Che-Ping Tsai, Yi-Lin Tuan, and Lin-Shan Lee. 2018. Transcribing lyrics from commercial song audio: the first step towards singing content processing. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5749–5753.

Christina M Vanden Bosch der Nederlanden, Xin Qi, Sarah Sequeira, Prakhar Seth, Jessica A Grahn, Marc F Joanisse, and Erin E Hannon. 2023. Developmental changes in the categorization of speech and song. *Developmental Science*, 26(5):e13346.

Jun-You Wang, Chon-In Leong, Yu-Chen Lin, Li Su, and Jyh-Shing Roger Jang. 2023. Adapting pretrained speech model for mandarin lyrics transcription and alignment. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8, Taipei, Taiwan.

Jun-You Wang, Chung-Che Wang, Chon-In Leong, and Jyh-Shing Roger Jang. 2024. Mir-mlpop: A multilingual pop music dataset with time-aligned lyrics and audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1366–1370, Seoul, Korea, Republic of.

Christof Weiß, Frank Zalkow, Vlora Arifi-Müller, Meinard Müller, Hendrik Vincent Koops, Anja Volk, and Harald Grohganz. 2021. Schubert Winterreise dataset: A multimodal scenario for music analysis. *ACM Journal on Computing and Cultural Heritage (JOCCH)*, 14(2):25:1–18.

Le Zhuo, Ruibin Yuan, Jiahao Pan, Yinghao Ma, Yizhi Li, Ge Zhang, Si Liu, Roger B. Dannenberg, Jie Fu, Chenghua Lin, Emmanouil Benetos, Wenhu Chen, Wei Xue, and Yike Guo. 2023. LyricWhiz: Robust multilingual zero-shot lyrics transcription by Whispering to ChatGPT. In *Proceedings of the International Society for Music Information Retrieval Conference, ISMIR*, pages 343–351, Milano, Italy.