Decoding Melody: A Comparative Study of Whisper ASR and Human Accuracy in Transcribing Song Lyrics

Claudia Anna Narang-Keller
University of Zurich
claudiaanna.narang-keller@uzh.ch

Stylianos Psychias
University of Zurich
stylianos.psychias@uzh.ch

Abstract

This study examines the Whisper ASR system's performance as well as human transcription on song lyrics. The focus is on quantifying and understanding the discrepancies found. The analysis involves both quantitative metrics (i.e. WER) and qualitative evaluations. The results indicate that human transcriptions contain more errors than automatic transcriptions and that the WER varies substantially depending on the song and for manual transcription on the transcriber. This study highlights the potential and limitations of ASR in processing complex audio. Insights from this research can guide future enhancements in ASR technology.

1 Introduction

Automatic Speech Recognition (ASR) systems have transformed text transcription (Levis and Suvorov, 2012). They convert spoken language into written text efficiently. Yet, their effectiveness varies across different environments and speech types. This research focuses on automatic lyrics transcription (ALT) specifically the one of Whisper compared to human lyrics perception.

The Whisper ASR system, developed by OpenAI, shows promise in diverse audio environments. However, its effectiveness on song lyrics, with their distinct characteristics, remains less explored. This study aims to fill that gap. It compares the transcription accuracy of Whisper ASR to human transcription.

Understanding Whisper's performance can inform its training and optimization. Improved ASR systems can be incredibly beneficial. They aid in accessibility services and media production, enhancing content reach and inclusivity.

The methodology of our study involves quantitative and qualitative analyses. Word error rates (WER) form the quantitative basis. A qualitative review of mismatches offers deeper insights. This comprehensive approach sheds light on both the strengths and limitations of automatic speech recognition (ASR) technology.

2 Background

ASR systems have become ubiquitous in various applications. They transcribe speech into text efficiently. However, their performance in nonstandard settings, like song lyrics transcription, varies widely. This section explores the challenges and advancements in ALT.

Transcribing song lyrics presents unique challenges. Songs often mix complex vocals with varying background noises. The diversity in vocal styles and the presence of music complicate the recognition process. ASR systems must adapt to handle these complexities effectively. Additionally, human transcription of song lyrics can be particularly challenging due to the use of unfamiliar slang, the rapid delivery of text in genres like rap, and the extreme vocal pitches in opera-like settings.

Recent studies have highlighted these challenges. Kruspe (2024) notes the difficulty in phoneme recognition due to pitch variations and background noise in songs. These elements can drastically affect the ASR's accuracy. The necessity for robust ASR technologies in singing is stressed. This is particularly important for less popular or 'long tail' music genres where lyrics are not readily available.

The LyricWhiz (Zhuo et al., 2023) project illustrates an innovative approach, combining Whisper and ChatGPT to improve lyrics transcription. This system demonstrates significant advancements in handling multilingual lyrics transcription. It offers insights into integrating large language models with ASR. This integration leads to enhanced performance.

Deep learning has propelled significant progress in this field (Meseguer-Brocal et al., 2020). Neural networks have been particularly effective in improving phoneme recognition and overall transcription accuracy. Integrating a robust speech recognition model with a large language model shows promise in enhancing transcription accuracy. This improvement is evident across various languages.

Despite these advancements, many challenges remain. The variability in vocal execution and the presence of instrumental sounds continue to challenge the effectiveness of ASR in accurately transcribing song lyrics. Ongoing research (Molina Martinez et al., 2017) is crucial to overcome these barriers and enhance the capabilities of ASR systems in complex audio environments.

ASR applications in music go beyond transcription. They include real-time captioning for accessibility, music discovery, and assisting in music-related research. These applications not only enhance user interaction with music. But also, it promotes cultural exchange and improve accessibility for the hearing impaired.

In summary, ASR technology has made significant strides, particularly in the domain of music. However, the unique challenges of song lyrics transcription require continued research and development. This review underscores the importance of tailored approaches for ASR applications. It highlights the need to consider the specific challenges and requirements of each application.

3 Methods

The audio files, transcriptions, results in table and plot formats, and Python files used for preprocessing and analysis are stored in a private GitHub repository[1], accessible upon request.

3.1 ASR System Description

This study utilized the Whisper ASR system. Whisper was developed by OpenAI as a robust, multilingual ASR model (O'Sullivan et al., n.d.). It is designed to handle a wide range of audio inputs efficiently.

Whisper processes audio using advanced deep learning techniques. It supports multiple languages and adapts well to different accents and dialects. This makes it suitable for song lyrics, which often vary in language and style. The model operates on preprocessed audio data. It converts these into text by recognizing spoken words within the audio. For this project, the medium version of Whisper was used due to its balance between speed and accuracy.

3.2 Data Collection and Preprocessing

For the dataset, we utilized the Billboard Hot 100 Singles from the year 2020, as it provided a diverse range of genres and was readily accessible online[2]. These genres ranged from pop and rock to hip hop and R&B. Most of the songs are in English language.

For the first 30 songs, we gathered the complete lyrics by copying and pasting them from the internet. The song lyrics and information regarding genres were sourced from Musixmatch (mus). Using their API, we automatically retrieved the first 30% of each song's lyrics, as the API allows up to 2,000 free calls per day.

All songs underwent a standard preprocessing routine before transcription. Initially, each song file was in MP3 format. These were converted to WAV format for compatibility with the Whisper ASR system. The conversion process maintained a consistent bit rate and sampling rate across all files. This standardization was crucial for maintaining uniform audio quality. It ensured that any transcription errors could be attributed to the ASR system and not to variances in file quality. Audio files were also normalized to a standard volume level. This normalization helped minimize the impact of volume differences on the ASR's performance.

These preprocessing steps were automated using a script written in Python. The script utilized audio processing libraries such as `Pydub` (Robert, 2021) and `Librosa` (lib) for file conversion and normalization. This careful selection and preparation of the song dataset provided a robust foundation for testing and evaluating the Whisper ASR system.

3.3 Human Transcription

Several individuals proficient in English were assigned the task of listening to a selection of songs and transcribing the lyrics as heard. Three individuals completed the transcription on time; all of them are second language (L2) speakers with English certifications.

Participants were permitted to pause the songs to allow sufficient time for transcription but were restricted from replaying any portions of the songs to simulate real-life listening conditions more accurately and to make the transcriptions comparable among each other.

[1] https://github.com/cnaran/ASP_MusicLyrics

[2] https://archive.org/details/100Hits2021

3.4 Analytical Methods

The study measured transcription accuracy using the Word Error Rate (WER) (Ali and Renals, 2018). This metric is common in evaluating speech recognition systems.

WER assesses the percentage of errors at the word level. It counts substitutions, deletions, and insertions needed to correct the transcript relative to the length of the reference text.

$$WER = \frac{S + D + I}{N}$$

where:

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions, and
- N is the number of words in the reference.

The calculation of WER, as well as the identification of deletions, insertions, and substitutions, was automated using scripts. These scripts utilized the `jiwer` library, a tool popular for its effectiveness in calculating these metrics. This automation ensured accuracy and consistency in measurement across all transcriptions. As we were able to obtain 30% of the lyrics directly by calling the Musixmatch API, we compared the WER of the full lyrics to the WER obtained using 30% of the lyrics as the reference.

Mismatch analysis involved examining the transcripts to identify patterns in deletions, insertions, and substitutions for both the automatic and manual transcriptions.

Python's `pandas` library managed and analyzed the data. This library helped in organizing mismatched data and performing statistical computations. Visual representations of mismatches were created using `matplotlib`. These visualizations helped in quickly identifying trends and patterns in the data.

For detailed analysis, each mismatch was logged. This logging helped in understanding common error patterns. The scripts also tagged the type of error for each mismatch, facilitating further statistical analysis.

Visualizations were created using `matplotlib`. These graphs illustrated the distribution of WER across different songs and transcription methods.

4 Results

This study evaluated Whisper's performance in transcribing song lyrics, comparing it to human transcription accuracy by calculating the WER and conducting a mismatch analysis. In the following sections, we discuss the results in detail.

4.1 WER Analysis

The WER was substantially higher for human transcribers across all songs compared to Whisper. The average WER for Whisper transcriptions is 24%, whereas it is 53% for human transcriptions.

WER varied also significantly across different songs. Whisper achieved the lowest WER on the song "Falling" by Harry Styles (5%) and the highest WER on the song "Highest In The Room" by Travis Scott (54%). The highest WER for the manual transcription was also for the song by Travis Scott (66%) whereas the lowest WER was 30% for Shawn Mendes' "Señorita".

The transcription accuracy varied among human transcribers. They took approximately 5 minutes to transcribe each minute of the song. They noted that the task would have been easier if they could listen to the song multiple times or rewind a few seconds to re-listen to certain parts, especially during fast-paced rap sequences. Additionally, they found it challenging to listen to enough context to understand the lyrics while simultaneously remembering all the exact words.

WER calculated from Whisper transcriptions showed substantial differences when using the full lyrics as a reference compared to using only 30% of the lyrics as a reference.

Figure 1 shows a visual representation of WER for different transcription and reference lyrics (full vs. 30%).

4.2 Mismatch Analysis

A detailed analysis of discrepancies between ASR-generated and human transcripts versus the official lyrics revealed common transcription errors, indicating challenges in handling diverse accents, pronunciations, and musical backgrounds. For instance, the word "love" was frequently mistaken for "look", "heart" for "hard", "feel" for "fill", or "leave" for "live", highlighting typical phonetic confusions. Human transcription tended to produce more semantically coherent text, however.

Human transcribers particularly struggled with identifying words in sections where background

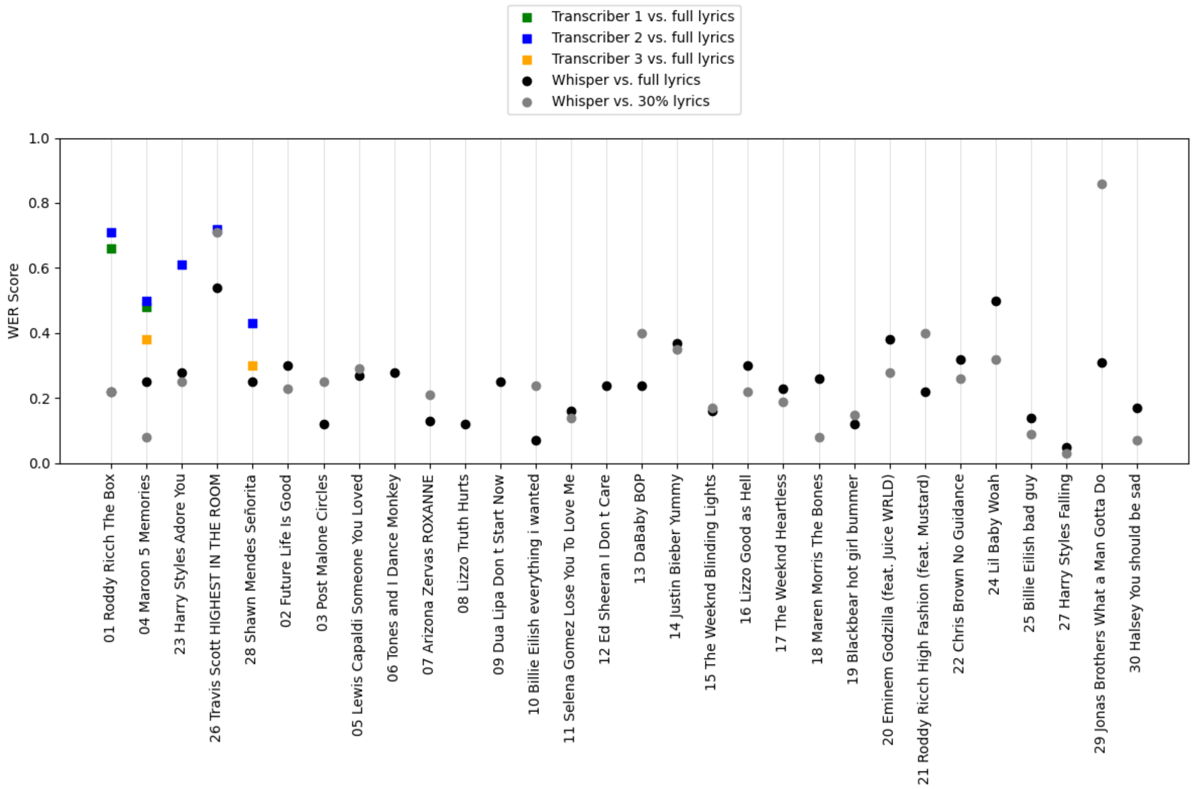

Figure 1: WER comparison for different transcripts and references

	Lyrics	Whisper (WER = 0.22)	Human transcriber 1 (WER = 0.66)
1		**E-boot, E-boot, E-boot, E-boot,**	
2		**E-boot, E-boot**	
3	Pullin' out the coupe at the lot	**Pulling** out the coupe at the lot,	**when i cook a delight**
4	Told 'em fuck 12, fuck SWAT	told em fuck 12, fuck Swat	told **them** fuck 12 fuck ~~**swat**~~
5	Bustin' all the bells out the box	Bustin' all the bells out the box,	**buzzing on** the bells **on** the box
6	I just hit a lick with the box	I just hit a lick with the box	i just **get** to **live** with the box
7	Had to put the stick in a box, mmh	Had to put the stick in **the** box, ~~**mmh**~~	**headin to stick** with a box ~~**mmh**~~
8	Pour up the whole damn seal,	pour up the whole damn seal	**put down** your whole damn seal
9	I'ma get lazy	I'ma get lazy,	ima get lazy
10	I got the mojo deals,	I got the mojo deals,	i got the mojo deals
11	we been trappin' like the '80s	we been trappin' like the 80s	~~**we been trappin'**~~ like **in** the 80s
12	She sucked a nigga soul,	She **such** a nigga, **so,**	she **got** a nigga song
13	gotta Cash App	**got the** cash **out**,	gotta cash app
14	Told 'em wipe a nigga nose,	told em wipe a nigga **No**,	**dont know white when nigga no**
15	say slatt, slatt	say **slat, slat**,	~~**say slatt, slatt**~~
16	I won't never sell my soul,	I won't **ever** sell my soul	i **will never** sell my soul
17	and I can back that	And I can back that,	and i can **take** that
18	And I really wanna know,	and I really wanna know	and i really wanna know
29	where you at, at?	where you at, **where**	**when** you **were where**
20	I was out back,	I was **at** back,	~~**I was out back,**~~
21	where the stash at?	where the stash at,	~~**where the**~~ **stackem, stackem**
22	Cruise the city in a bulletproof	cruisin' city in a bulletproof	cruise the city in your bulletproof
23	Cadillac (skrrt)	Cadillac,	cadillac
24	'Cause I know these niggas after	cause I know these niggas **out the**	cause ~~**I**~~ **you** ~~**these**~~ niggas **everywhere**
25	where the bag at (yeah)	**way** the bag at ~~**(yeah)**~~	**bad guys** ~~**at (yeah)**~~
26	Gotta move smarter,	Gotta move smarter,	gotta move smarter
27	gotta move harder	gotta move harder,	gotta move harder
28	Niggas try to get me for my water	**nigga** try to get me **fire** my water	**nigga tryin'** ~~**to**~~ **give** me ~~**for**~~ my water

Figure 2: Mismatch analysis for the song "The Box" from Roddy Ricch; insertions marked in blue, substitutions in violet and deletions in red

music was prominent, rap texts were fast-paced, or vocals heavily overlapped with instrumental sounds. This challenge was particularly evident in songs with complex musical arrangements or heavy electronic elements, where transcription errors frequently occurred due to compromised vocal clarity caused by effects or overlapping sounds.

Interjections such as "yeah," "hm," or "ay" were inconsistently transcribed compared to the reference, sometimes omitted or included erroneously when absent in the lyrics. Additionally, abbreviations were occasionally transcribed incorrectly, such as "them" as "em" or "you" as "ya".

The only song with code switching to Spanish was Shawn Mendes' Señorita which was handled fine by both the Whisper and by the transcriber.

Figure 2 illustrates a detailed mismatch analysis for the beginning of the song "The Box" by Roddy Ricch. This fast-paced track includes loud background music, extensive use of slang and profanity, and rapid pitch changes by the rapper. Whisper incorrectly transcribed the initial "I-U" background vocals as lyrics. Both Whisper and human transcriber 1 missed the interjections "mmh" and "yeah". The human transcriber noted difficulties in understanding and recalling the text passages for accurate transcription.

5 Discussion

This study explored the capabilities of the Whisper ASR system in transcribing song lyrics compared to human transcription.

5.1 Comparison to other ASR systems

According to Zhuo et al. (2023), the LyricWhiz ASR system achieved a WER of approximately 22% for the English songs tested when utilizing Whisper for ASR and ChatGPT for refining the lyrics. Without the assistance of ChatGPT and Whisper prompting, the overall WER increased to 33% for the same dataset. Our study, in comparison, achieved a WER of 24% for the selected songs. It's important to note that these numbers are not directly comparable as they were obtained from different datasets, but they provide insights into the approximate accuracy of the transcriptions.

5.2 Whisper ASR improvements

A common issue observed was Whisper's difficulty with phonetically similar words. For example, "love" was frequently misheard as "look," and "heart" as "hard". This pattern suggests that Whisper struggles with distinguishing between certain phonemes in the musical context. A potential solution could involve augmenting the training data with more musical samples containing their lyrics.

5.3 Manual transcriptions

Manual transcription proved to be cumbersome. The WER was notably higher compared to Whisper, indicating lower transcription accuracy that varied significantly among transcribers. ASR systems demonstrated better perfomance especially in transcribing fast-paced songs. Additionally, unfamiliarity with slang used in songs and differing cultural backgrounds from the artists made understanding lyrics more difficult for transcribers.

The transcription process guidelines influence WER; allowing transcribers to review passages would improve accuracy. ASR systems could streamline manual transcription by providing initial text, making the process more efficient for corrections. Furthermore, allow transcribers to transcribe songs in their native language and within their cultural context would improve their WER.

5.4 Improvements for the pipeline

To enhance the comprehensiveness of our dataset, it is recommended to improve the data collection pipeline. The analysis revealed that using only 30% of the readily available lyrics did not provide sufficient information, and utilizing the full lyrics as a reference is recommended. Specifically, integrating web scraping techniques to obtain complete song lyrics would streamline data analysis, though caution is necessary regarding copyright issues.

To obtain more meaningful results, interjections like "yeah," "oh," or "ay" could be excluded in a preprocessing step for both reference lyrics and transcriptions. This could be achieved by utilizing a predefined list of interjections to consistently filter them out. Furthermore, the impact of misspelled abbreviations, like "them" instead of "em," on the calculation of WER should be reduced, as these errors do not affect semantics.

Further, data analysis could focus on quantifying factors that influence recognition, such as the speed of spoken text, background noise levels, and the prevalence of slang words.

To differentiate between errors arising from music lyrics and those from unfamiliar vocabulary, lyrics should also be transcribed as normal speech.

This approach would provide insights into the specific challenges posed by lyrical content versus general language recognition.

Overall, more data need to be transcribed to gather more reliable insights.

6 Conclusion

This study evaluated the Whisper ASR system's and human transcribers ability to transcribe song lyrics. Results indicate that while Whisper can accurately transcribe certain songs, it struggles with others. The system often misinterpreted words with similar phonetic sounds. It also had difficulties with songs featuring complex musical backgrounds or fast-paced lyrics.

The transcription accuracy, i.e. the WER, was substantially higher for human transcribers compared to Whisper and varied across transcriber and songs.

The analysis showed that Whisper's performance varied significantly across different songs. This variation highlights the need for ASR systems to better adapt to the nuances of musical transcription.

The mismatch analysis provided valuable insights into common errors. These included frequent substitutions of phonetically similar words and challenges in noisy audio environments. These findings are crucial for future improvements in ASR technology.

In conclusion, the study confirms that while current ASR technology, like Whisper, has made significant strides, there remains room for improvement. Enhancing its ability to handle diverse and complex audio scenarios is essential.

Future research should focus on refining ASR models to be more sensitive to musical elements. Developing systems tailored to music transcription could bridge the current gaps in performance.

Overall, this research contributes to understanding ASR capabilities and limitations in a musical context. It lays the groundwork for further advancements in ALT technology.

References

librosa. https://librosa.org/doc/latest/index.html. Accessed on June 11, 2024.

Musixmatch. https://www.musixmatch.com/. Accessed on June 15, 2024.

A. Ali and S. Renals. 2018. Word error rate estimation for speech recognition: e-wer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

A. Kruspe. 2024. More than words: Advancements and challenges in speech recognition for singing. *arXiv preprint arXiv:2403.09298*.

J. Levis and R. Suvorov. 2012. Automatic speech recognition. *The Encyclopedia of Applied Linguistics*.

G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters. 2020. Creating dali, a large dataset of synchronized audio, lyrics, and notes. *Transactions of the International Society for Music Information Retrieval*.

E. Molina Martinez et al. 2017. Singing information processing: Techniques and applications.

J. O'Sullivan, G. Bogaarts, M. Kosek, R. Ullmann, P. Schoenenberger, C. Chatham, et al. n.d. Automatic speech recognition for asd using the open-source whisper model from openai. Accessed on June 11, 2024.

J. Robert. 2021. pydub. https://pypi.org/project/pydub/. Accessed on June 11, 2024.

L. Zhuo, R. Yuan, J. Pan, Y. Ma, Y. Li, G. Zhang, et al. 2023. Lyricwhiz: Robust multilingual zero-shot lyrics transcription by whispering to chatgpt. *arXiv preprint arXiv:2306.17103*.