LyricWhiz: Robust Multilingual Zero-shot Lyrics Transcription by Whispering to ChatGPT

Le Zhuo, Ruibin Yuan, Jiahao Pan, Yizhi Li, Ge Zhang, Jiawen Huang, Si Liu, Roger Dannenberg, Jie Fu, Chenghua Lin, Emmanouil Benetos, Wenhu Chen, Wei Xue, and Yike Guo

1. Overview

- We propose LyricWhiz, the first automatic lyrics transcription system that can perform **zero-shot**, **multilingual**, **long-form** lyrics transcription.
- In LyricWhiz, Whisper functions as the "**ear**" 👂 by transcribing the audio; ChatGPT serves as the "**brain**" 🧠, acting as an annotator with a strong performance for contextualized output selection and correction (Fig. 1).
- We further use LyricWhiz to construct a large-scale **multilingual** lyric transcription dataset, MulJam.

Figure 1: Concept illustration of the working LyricWhiz.

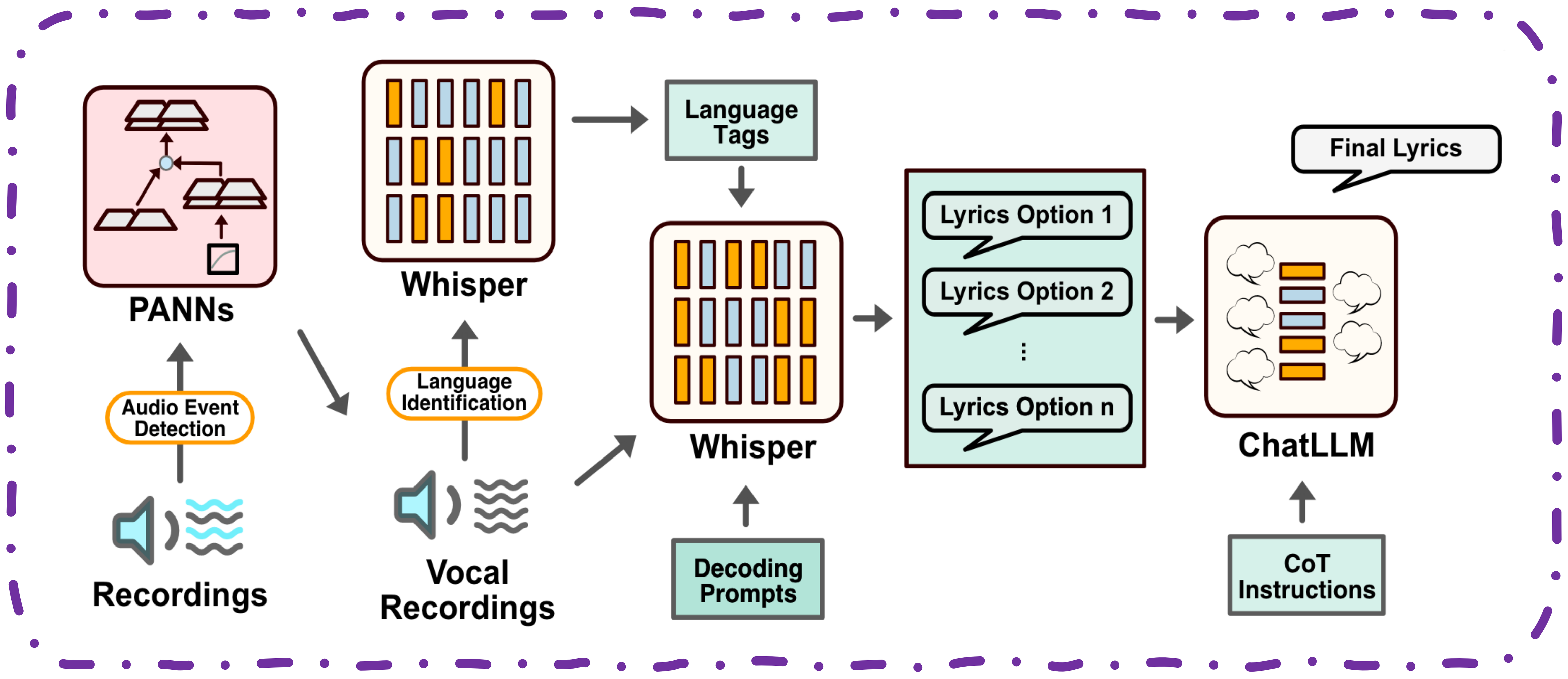

Figure 2: Framework of the proposed LyricWhiz.

2. Methodology

- LyricWhiz integrates two large-scale pre-train models from OpenAI -- Whisper and ChatGPT (Fig. 2).

Whisper - Zero-shot Lyrics Transcriptor

- Whisper, trained on speech data, excels in lyrics transcription within the music domain.
- We use the input prompt "**lyrics:**" as a prefix to guide it toward the ALT task.
- We leverage the **no speech probability** predicted by Whisper and drop predicted lines of lyrics with a no speech probability greater than 0.9.
- We generate **3 - 5 predictions** for each input music under identical settings.

ChatGPT - Effective Lyrics Post-processor

- We assign ChatGPT the role of a **lyrics transcription post-processor**.
- We stipulate that both input and output should be in **JSON format**.
- Inspired by **Chain-of-Thought** in LLMs, we decompose lyrics post-processing into three consecutive phases - analyze, make a choice, and output.

3. Dataset

- We further use LyricWhiz to construct the first **large-scale**, **weakly supervised**, and **copyright-free** multilingual lyric transcription dataset, MulJam.
- MulJam consists of 6,031 songs with 182,429 lines and a total duration of 381.9 hours (Tab. 1).

GPT-4 Instruction Prompt

Task: As a GPT-4 based lyrics transcription post-processor, your task is to analyze multiple ASR model-generated versions of a song's lyrics and determine the most accurate version closest to the true lyrics. Also filter out invalid lyrics when all predictions are nonsense.
Input: The input is in JSON format:
{"prediction_1": "line1;line2;...", ...}
Output: Your output must be strictly in readable JSON format without any extra text:
{
"reasons": "reason1;reason2;...",
"closest_prediction": <key_of_prediction>
"output": "line1;line2..."
}
Requirements: For the "reasons" field, you have to provide a reason for the choice of the "closest_prediction" field. For the "closest_prediction" field, choose the prediction key that is closest to the true lyrics. Only when all predictions greatly differ from each other or are completely nonsense or meaningless, which means that none of the predictions is valid, fill in "None" in this field. For the "output" field, you need to output the final lyrics of closest_prediction. If the "closest_prediction" field is "None", you should also output "None" in this field. The language of the input lyrics is English.

Figure 3: Instruction prompt for ChatGPT contextualized post-processing.

Dataset	Languages	Songs	Lines	Duraion
DSing [8]	1 (en)	4,324	81,092	149.1h
MUSDB18 [17]	1 (en)	82	2,289	4.6h
DALI-train [14]	1 (en)	3,913	180,034	208.6h
DALI-full [14]	30^*	$5,358^*$	-	-
MulJam (Ours)	6	6,031	182,429	381.9h

Table 1: Comparison between different lyrics transcription datasets.

Language	Songs_{train}	Songs_{test}	WER_{test}
English	3,791	20	21.86
French	1,030	7	26.64
Spanish	620	5	22.54
Italian	311	3	44.01
Russian	147	4	39.18
German	132	1	25.43
Overall	6,031	40	26.26

Table 3: The WERs (%) on our test set.

Method	Jamendo	Hansen	DSing
TDNN-F [8]	76.37	77.59	19.60
CTDNN-SA [45]	66.96	78.53	14.96
Genre-informed AM [12]	50.64	39.00	56.90
MSTRE-Net [13]	34.94	36.78	15.38
DE2-segmented [46]	44.52	49.92	-
W2V2-ALT [22]	33.13	18.71	**12.99**
LyricWhiz (Ours)	**24.25**	**7.85**	13.78
w/o ChatGPT Ens.	<u>28.18</u>	<u>8.07</u>	15.22
w/o Whis. Prompt	33.21	8.75	<u>13.40</u>

Method	a)	b)	c)
CTDNN-SA-mixture [17]	76.06	78.44	89.24
Ours-mixture	**50.90**	**47.04**	**50.70**
CTDNN-SA-vocals [17]	37.83	30.85	58.45
Ours-vocals	**26.29**	**25.27**	**33.30**

Table 2: The WERs (%) of various ALT systems, including ablation methods, on multiple datasets.

4. Results

- LyricWhiz **significantly reduces** Word Error Rate on various ALT benchmark datasets such as Jamendo and Hansen.
- Ablations indicate that both **Whisper prompt** and **ChatGPT ensemble** are essential for model performance.
- We manually create a multilingual test set of 40 songs for noise level estimation.
- Our model **achieves decent WER** without any post-processing tricks.

M-A-P multimodal art projection

Codes & Data