audioshake

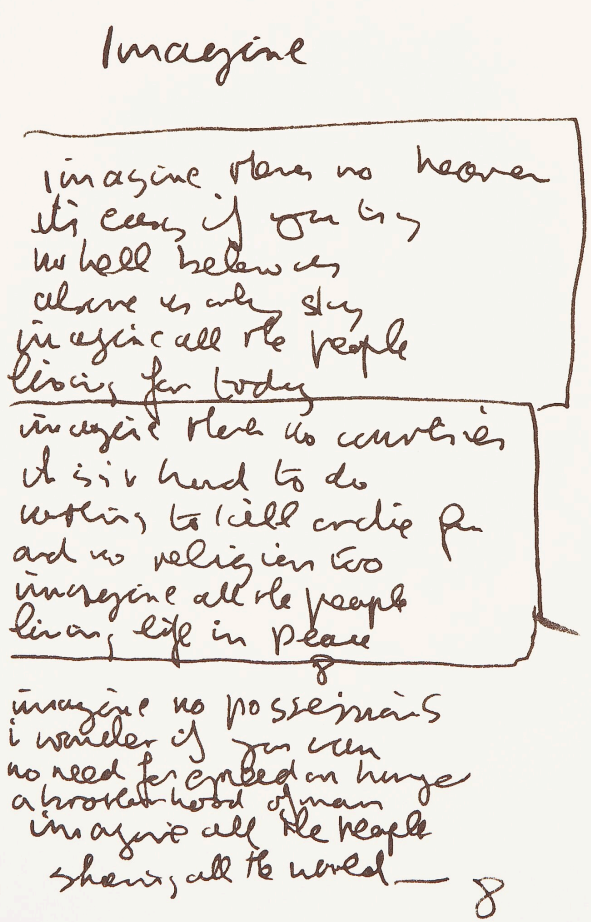
Formatting-Aware Automatic Lyrics Transcription: A Case Study on German Songs

Ondrej Cífka, Constantinos Dimitriou, Cheng-I Wang, Hendrik Schreiber, Luke Miner, Fabian Stöter

fabian@audioshake.ai 🏠

Automatic Lyric Transcription (ALT)

- task: recognize lyric text from singing
- vs ASR: singing is less intelligible than speech
- support multiple languages



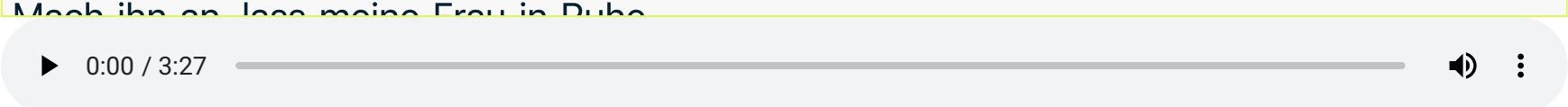
Application: Shorts

Application: Karaoke

Goal: lyrics should reflect both performance/recording and song-writing

Verdammt ich lieb Dich (Whisper)

Ich ziehe durch die Straßen bis nach Mitternacht
Ich hab das früher auch gern gemacht
Dich brauch ich dafür nicht
Ich sitz am Tresen, trinke noch ein Bier
Früher waren wir oft gemeinsam hier
Das macht mir, macht mir nichts
Gegenüber sitzt ein Typ wie ein Bär
Ich stell mir vor, wenn das dein neuer wär
Das joggt mich überhaupt nicht
Auf einmal packst mich, geh auf ihn zu
Mach ihn an, lass meine Frau in Ruhe



Problem: ALT benchmarks ignore letter case, punctuation and formatting (e.g. line break placement, parentheses around background vocals). This prevents the full evaluation of formatting-aware models.

Contributions

- `jam-alt` — A revised version of JamendoLyrics[1] MultiLang that follows industry standards for song lyrics transcription and formatting
- `alt-eval` — A set of automated evaluation metrics designed to capture and distinguish different types of errors
- `Benchmark` — The dataset and the metrics implementation are released online[2]

---

1. https://github.com/f90/jamendolyrics ↩
2. https://audioshake.github.io/jam-alt/ ↩

audioshake

jam-alt

jam-alt dataset

- we revised all lyrics from the JamendoLyrics dataset (20 songs for en, de, es, fr)
- applied a standardized set of rules
- matching recordings as closely as possible
- created a detailed annotation guide (released together with the dataset)
- each song was revised by a single annotator proficient in the language
- then reviewed by two other annotators

Rules

1. Only transcribe words and vocal sounds audible in the recording; exclude credits, section labels, style markings, non-vocal sounds etc.
2. Break lyrics up into lines and sections; separate sections by a single blank line.
3. Include each word, line and section as many times as heard. Do not use shorthands to denote repetitions.
4. Start each line with a capital letter; respect standard capitalization rules for each language.
5. Respect standard punctuation rules, but never end a line with a comma or a full stop.
6. Use standard spelling, including standardized spelling for slang/contractions where appropriate.
7. Transcribe background vocals and non-word vocal sounds if they contribute to the content of the song.
8. Place background vocals in parentheses.

German-Specific Rules

- We use Neue deutsche Rechtschreibung
- Apostrophes (elisions, contractions)
  - We follow the rules from Duden, but made a few additional rules to improve consistency:
  - Elision of `e` at the end of a word: we generally use apostrophes except in the case of imperatives. See rule D13/2.

    Example: `Komm her!` *but* `Ich komm' her`.
  - Elision of `e` in the middle of a word: we apply the rules in many cases except for when it is harder to read: See rule D13/1.

    Example: `seh'n` *but* `verstehn`.

Burnout Man - Abendblau

0:00 / 4:19

man o man Mannomann , am Ende des Tages ist immer noch Arbeit übrig

Wenn andere ihre Kinder betten
Brennt im Büro ein Licht
Man sieht dich dort die Firma retten
Die sonst zusammenbricht

naja Na ja , Kollege , einer muss sich ja drum darum kümmern

audioshake

alt-eval

Evaluation Metrics

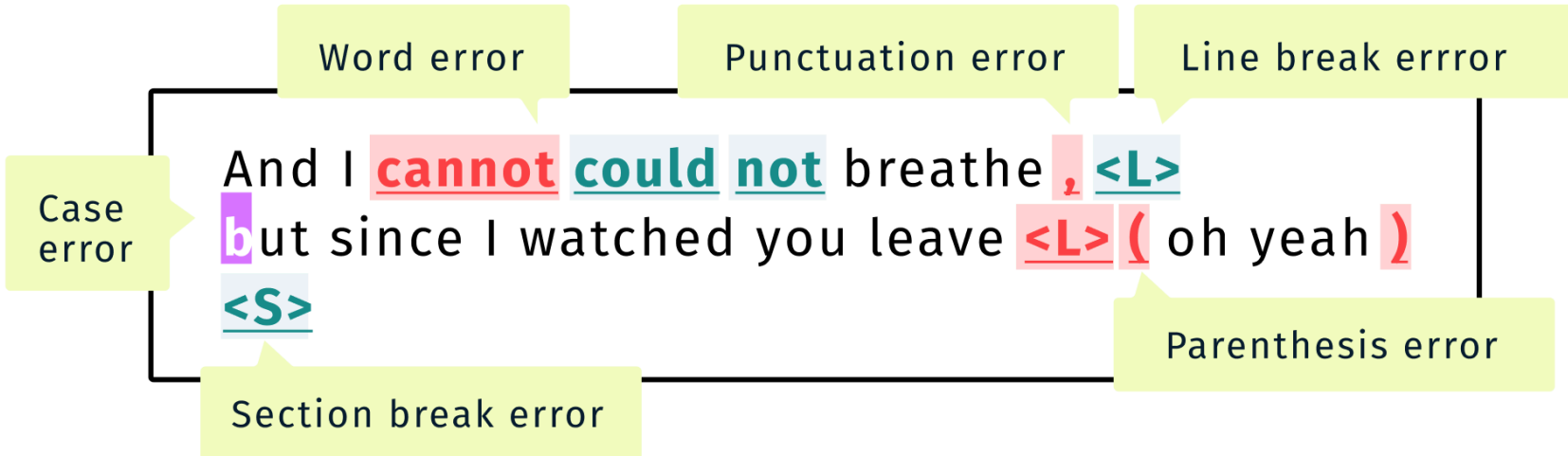
- Word error rate (WER) with lyrics-aware tokenization

```
"Sei's Melancholie" → ["sei", "'s", "melancholie"]
```

```
"Könnt' ich dir Schmerz erspar'n" → ["könnt'", "ich", "dir", "schmerz", "erspar'n"]
```

- Complementary to WER:
  - Case error rate to capture letter case errors (*melancholie* ≠ *Melancholie*)
  - Information retrieval metrics (precision, recall, F1) for tokens ignored by WER:
    - Punctuation
    - Parentheses (used to delimit background vocals)
    - Line breaks
    - Section breaks (i.e. double line breaks)

Tokenized metrics in detail



Legend

deletion insertion <L> – line break <S> – selection break

- TODO: german example

Usage

Setup

```
pip install datasets
pip install alt_eval
```

Loading `jam-alt`

```
from datasets import load_dataset
dataset = load_dataset("audioshake/jam-alt")["test"]
```

Your model

```
transcriptions = transcribe(dataset["audio"])
```

Compute metrics

```
compute_metrics(
  dataset["text"],
  transcriptions,
  languages=dataset["language"]
)
```

audioshake

Benchmark

Results

	All languages						English						Spanish		German		French	
	WER	E_{Aa}	F_{P}	F_{B}	F_{L}	F_{S}	WER	E_{Aa}	F_{P}	F_{B}	F_{L}	F_{S}	WER	E_{Aa}	WER	E_{Aa}	WER	E_{Aa}
Whisper v2	35.7	4.5	41.7	—	69.3	3.3	43.8	3.5	31.3	—	63.0	11.2	25.7	6.5	45.4	5.3	27.7	3.2
Whisper v2 +sep	44.0	5.3	28.0	—	61.2	—	32.3	5.3	39.2	—	53.8	—	38.8	7.1	65.2	5.9	43.3	3.2
Whisper v3	35.5	4.3	41.6	—	73.5	1.0	37.7	4.8	40.9	—	71.5	2.6	28.6	5.0	40.7	**4.0**	34.7	3.3
Whisper v3 +sep	47.9	3.8	29.0	—	65.7	—	43.0	4.1	23.3	—	66.8	—	61.5	**3.6**	43.5	4.4	44.9	3.2
LyricWhiz	—	—	—	—	—	—	24.6	3.5	34.0	—	74.0	1.4	—	—	—	—	—	—
AudioShake	**17.2**	**3.7**	**56.0**	**25.2**	**84.4**	**71.1**	**19.2**	**3.3**	**65.6**	**30.8**	**84.9**	**78.7**	**14.4**	5.0	**11.4**	4.4	**22.5**	**2.4**
JamendoLyrics	11.1	18.5	—	—	93.3	85.3	14.4	15.3	—	—	88.1	77.9	14.0	15.1	5.0	32.6	10.3	12.9

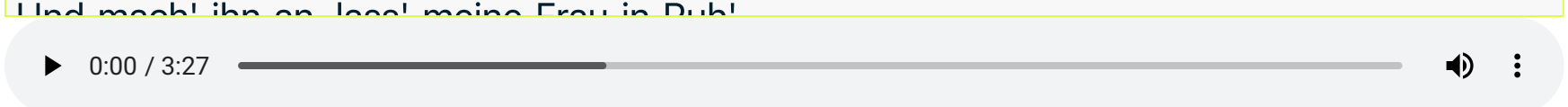
- `AudioShake` outperforms `Whisper` on all metrics

audioshake

Demo

Verdammt ich lieb Dich (AudioShake)

Ich ziehe durch die Straßen bis nach Mitternacht
Hab' das früher auch gern gemacht
Dich brauch' ich dafür nicht
Ich sitz' am Tresen, trinke noch 'n Bier
Früher waren wir oft gemeinsam hier
Das macht mir, macht mir nichts
Gegenüber sitzt 'n Typ wie 'n Bär
Ich stell' mir vor, wenn es dein neuer wär'
Das juckt mich überhaupt nicht
Auf einmal packt's mich, geh' auf ihn zu
Und mach' ihn an, lass' meine Frau in Ruh'



Ich hass dich (AudioShake)

Jede Nummer von Topmodeltypen in deinem Handy
Hast noch nie 'nen Cop bekommen, außer dem mit Präsenten
Würd' mich nicht mal wundern, wenn du plötzlich Präsident bist
Weil jeder trägt dich Huren so bedingungslos auf Händen
Und du weißt, wie das geht mit diesen Steuern und Finanzen
Hast von jedem gut laufenden Unternehmen Aktien
Kriegst alles geschenkt, kriegst alles hin
Bist Weihnachten nie zerrissen, weil deine Eltern noch zusammen sind
Striegelt und faltenlos, gewonnen der G-Lotterie
Und ich hab das falsche Los, sie immer nieten, C'est la vie
Sag mir, womit hast du das verdient?

0:00 / 3:13

Ich hass dich (Whisper)

Jede Nummer von Topmodel-Typen in deinem Handy
hast noch nie nen Cop bekommen, außer dem mit Präsenten
würd mich nicht mal wundern, wenn du plötzlich Präsident bist
weil jeder trägt dich hoch und so bedingungslos auf Händen
und du weißt, wie das geht mit diesen Steuern und Finanzen
hast von jedem gut laufenden Unternehmen Aktien
kriegst alles geschenkt, kriegst alles hin
bist weihnachten nie zerrissen, weil deine Eltern noch zusammen sind
du bist gestriegelt und feitenlos, gewonnen der G-Lotterie
und ich hab das falsche Los, sie im Manitens-Syllabie
sag mir, womit hast du das verdient?

0:00 / 3:13

Das ist die perfekte Welle (AudioShake)

Mit jeder Welle kam ein Traum

Träume gehen vorüber

Und dein Brett ist verstaubt

Deine Zweifel schäumen über

Hast ein Leben lang gewartet

Hast gehofft, dass es sie gibt

Hast den Glauben fast verloren

Hast dich nicht vom Fleck bewegt

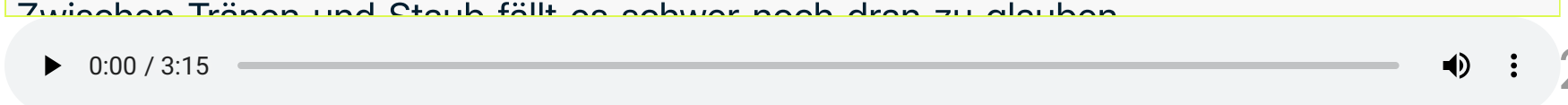
Jetzt kommt sie langsam auf dich zu

Das Wasser schlägt ihr ins Gesicht

0:00 / 3:15

Das ist die perfekte Welle (Whisper)

Mit jeder Welle kam ein Traum, Träume gehen vorüber und der Brett ist verstaubt.
Wenn die Zweifel schon über, ist ein Leben lang gewartet, ist gehofft, dass es sie gibt.
Hast den Glauben fast verloren, hast dich nicht vom Fleck bewegt.
Jetzt gucken sie langsam auf dich zu, das Wasser schlägt dir ins Gesicht.
Siehst dein Leben wie ein Film, du kannst nicht glauben, dass sie bricht.
Das ist die perfekte Welle, das ist der perfekte Tag.
Lass dich einfach von ihr tragen, denk am besten gar nicht nach.
Das ist die perfekte Welle, das ist der perfekte Tag.
Es gibt mehr als du weißt, es gibt mehr als du sagst.
Deine Hände sind schon taub, du hast Salz in deinen Augen.
Zwischen Tränen und Staub fällt es schwer noch dran zu glauben



Bonus: Unser Stammbaum (AudioShake)

Ich war 'ne stolze Römer
Komm' mit Cäsars Legion
Und ich bin 'ne Franzus
Komm' im Napoleon

Ich bin Buur, Schreiner, Fischer
Wetzlar und Edelmann
Sänger und Gaukler
Su fing alles an

Su es me alt, he hinjekumme

0:00 / 3:50

Bonus: Unser Stammbaum (Whisper)

Ich war ne stolze Römer, komm mit Cäsars Legion
Und ich bin ne Franzus, komm mit Napoleon
Schreiner, Fischer, Wetzlar und Edelmann, Sänger und Gaukler, so fing alles an
So sind wir alle hingekommen, wir sprechen Höck all dieselbe Sprache
Wir haben dadurch so viel gewonnen, wir sind wie wir sind
Wir Jäcke am Ring, das ist jetzt, wo wir stolz drauf sind
Ich bin aus Palermo, brat Spaghetti für euch mit
Und ich war ne Pimok, höck lach ich mit euch mit
Ich bin Grieche, Türke, Jude, Moslem und Buddhist
Wir all, wir sind nur Menschen, vom Herrn Gott sind wir nicht
So sind wir alle hingekommen, wir sprechen Höck all dieselbe Sprache

0:00 / 3:50

audioshake

Conclusion

- proposed `jam-alt`, a new benchmark for ALT
- results bring clarity into how existing systems differ in their performance on different aspects of the task
- best lyrics transcription model for German songs

Learn More

audioshake.github.io/jam-alt