Potlačení halucinací ASR modelu Whisper-Large-V3

Maxim Plička, Michal Pyšík, Michal Bartošák

Brno University of Technology, Faculty of Information Technology

Božetěchova 1/2. 612 66 Brno - Královo Pole

May 7, 2024

Problematika halucinací

- **Definice:**

  Text generovaný modelem, který je sémanticky nesouvisející s referencí, ale přesto plynulý a smysluplný.

- Jeden z největších problémů dnešních ASR modelů, který přináší řadu rizik
- Příklad: *"everything is fine"* $\rightarrow$ *"everything is fine, please subscribe"*
- Nejnovější verze Whisperu (Large-V3) halucinuje více než předešlé verze

Vyvolání halucinací

- Experimentovali jsme s:
  - Špatnou kvalitou nahrávky a hlukem v pozadí
  - Vložením tichého úseku do různých míst nahrávek
  - Proložením celé nahrávky náhodným šumem či vysokofrekvenční sinusovkou
- Konzistentní výsledky přineslo vložení delšího tichého úseku na začátek i konec nahrávky
- Poznámka: model často halucinuje při přepisu nahrávek lidí trpící vadou řeči (afázie)

Metriky na detekci halucinací

1. Halucinace jsou často vloženy do jinak správného přepisu
   $\Rightarrow$ kontrola, že je přepis delší než reference (a WER $>$ 5 %)
   - **Potenciální halucinace**
   - Více false positives, ale zachytí hrubou většinu halucinací (good recall)

2. Určité halucinace jsou velice časté
   $\Rightarrow$ kontrola, že se jeden z běžných podřetězců nachází v přepisu, ale ne v referenci
   - **Běžná halucinace**
   - Vzácnější halucinace nedetekuje, ale méně false positives (good precision)

Metody potlačení halucinací

1. Menší model (Whisper-Tiny) halucinuje méně, navíc umí explicitně detekovat tiché úseky
   - Zarovnání výstupů obou modelů
   - Odstranění částí výstupu, které malý model označil za tiché
   - Jedná se tedy o postprocessing výstupu

2. Hlavním zdrojem halucinací jsou tiché úseky (či obecně neaktivita řečníka)
   - Vygenerování timestampů k dané nahrávkce spolehlivým VAD modelem (my použili Silero VAD)
   - Konkatenace jednotlivých úseků nahrávky s aktivitou řečníka
   - Jedná se tedy o preprocessing vstupu

Příprava experimentu

- Dataset: LibriSpeech ASR corpus (test - other)
  - Složitější věty v anglickém jazyce
  - 2939 nahrávek
  - Augmentace: vložení 20 sekund ticha před i po každou nahrávku
- Průběh experimentu
  - Původní Whisper-V3-Large, dále obě implementované metody
  - Celkový počet halucinací dle obou našich metrik
  - Celkový Word Error Rate (kontrola že se samotná kvalita přepisu nezhoršila)

Výsledky experimentu - halucinace

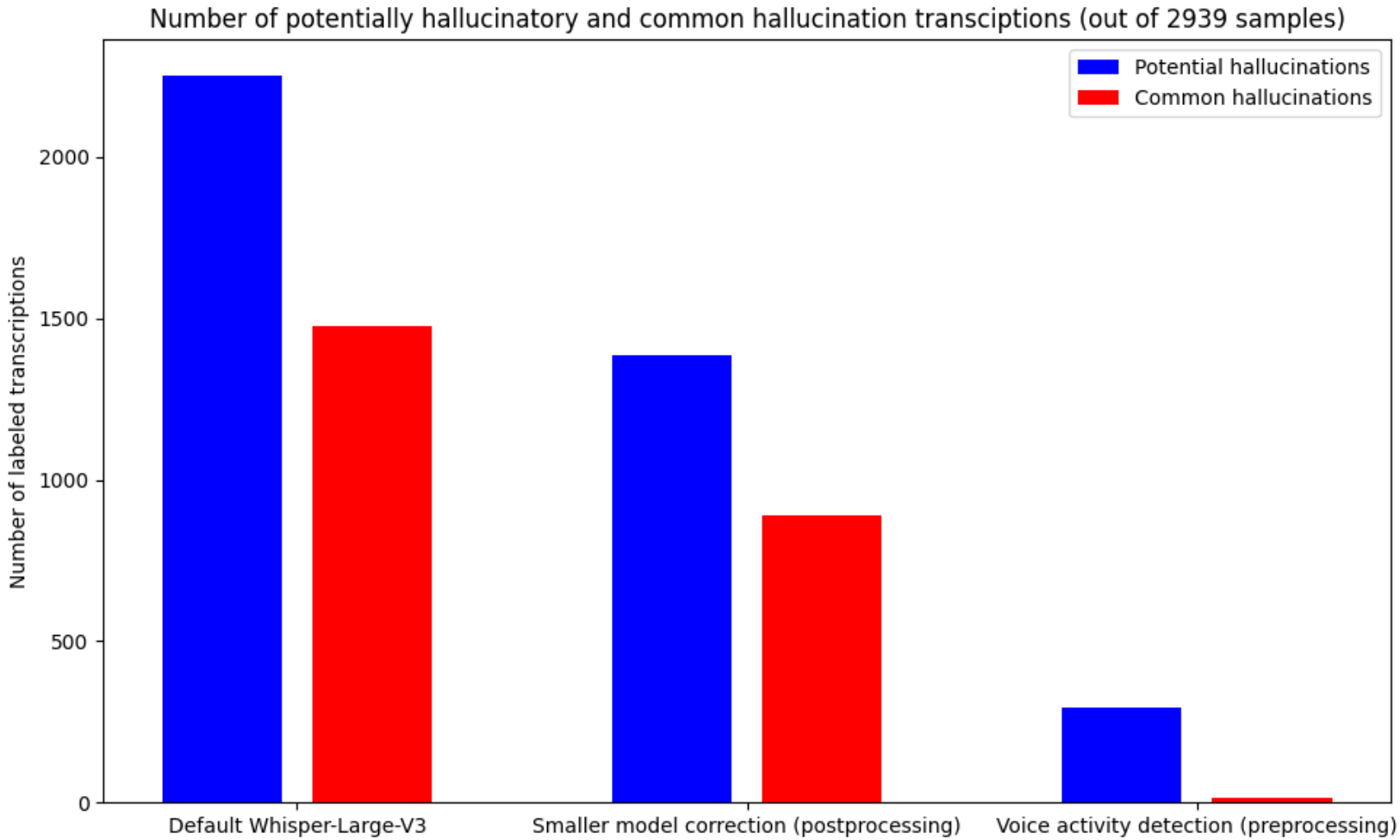

Figure: Počet detekovaných halucinací při použití jednotlivých metod.

Výsledky experimentu - WER

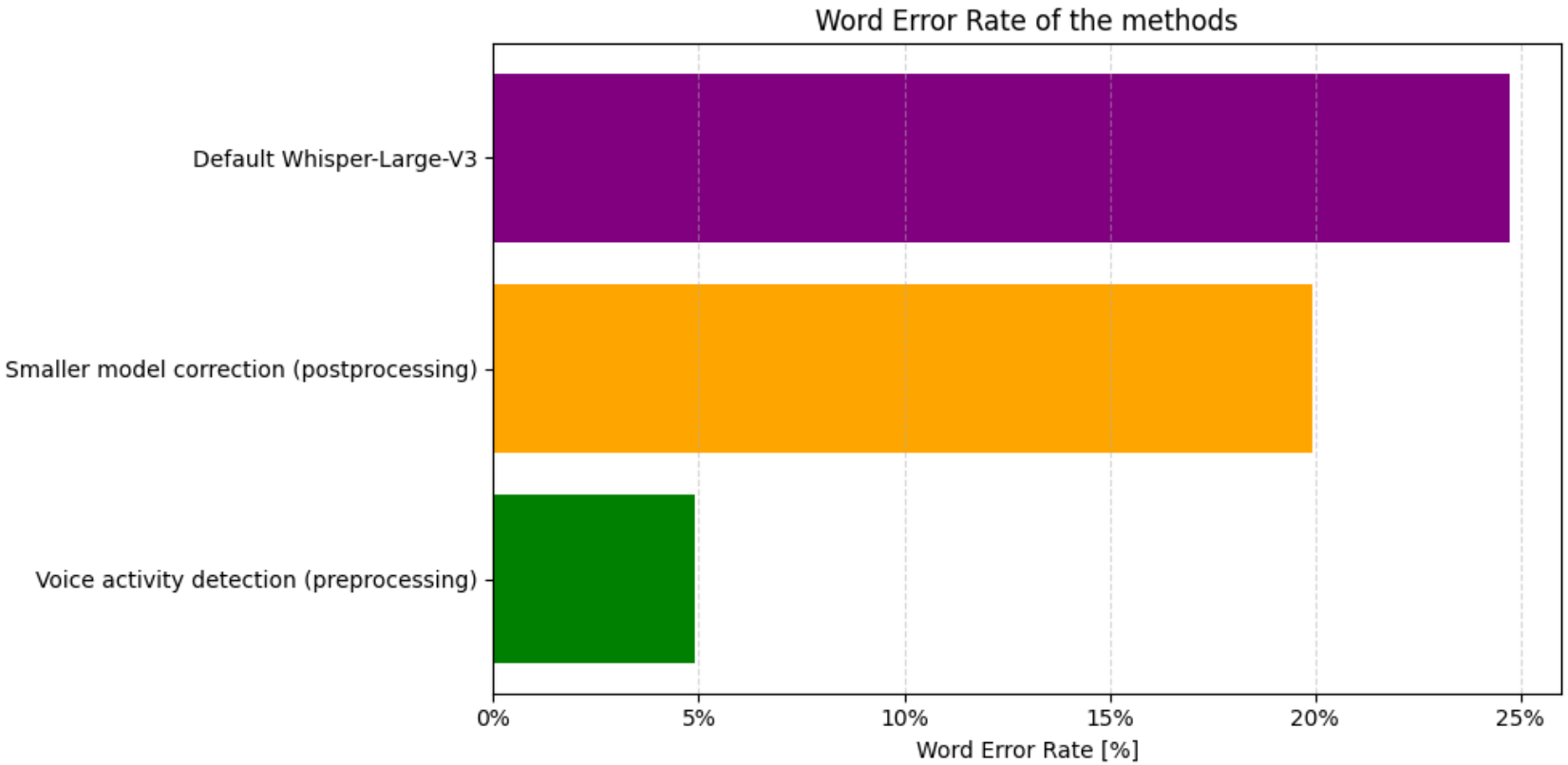

Figure: Celkový Word Error Rate při použití jednotlivých metod.

Interpretace výsledků

- Zdrojem halucinací tohoto modelu je zpravidla snaha o přepis částí vstupu bez aktivity řečníka
- Plně efektivním řešením je odstranění takových úseků vstupu—všechny "halucinace" při použití tohoto přístupu byly false positiva
- Tento proces lze automatizovat použitím spolehlivého Voice Activity Detection modelu
- Poznámka: tento přístup se osvědčil i na přepis řeči lidí trpící afázií, s čímž má Whisper normálně problém