Automatický detektor dialektu na základě audionahrávky

verze 1.0

Uživatelská příručka

a

Technická dokumentace

Jazyková paměť regionů České republiky.
Metody strojového učení pro uchování, dokumentaci a prezentaci nářečí českého jazyka.

Ministerstvo kultury ČR, NAKI III, DH23P03OVV010.
Program na podporu aplikovaného výzkumu v oblasti národní a kulturní identity 2023–2027.

Základní údaje

Název výsledku

Automatický detektor dialektu na základě audionahrávky

Verze

1.0

Druh výsledku

R – software

Lokalizace výsledku

https://jamap.cz/detektor

Lokalizace dokumentace

https://jamap.cz/detektor/dokumentace.pdf

Vlastník

Vysoké učení technické v Brně – Ústav pro jazyk český AV ČR, v. v. i.

Stručný popis produktu

Systém na základě vstupní zvukové nahrávky generuje automatické odhady konkrétního teritoriálního dialektu užívaného mluvčím (včetně vyčíslení pravděpodobnosti). Teritoriální dialekt je stanoven na úrovni nářečních podskupin; celkem se pracuje se 13 podskupinami.

Technické parametry produktu

Systém byl nejprve natrénován na velkém množství vícejazyčných audiálních dat, poté byl dotrénován na unikátních nářečních datech, čímž se zvýšila úspěšnost detektoru. Byl využit programovací jazyk Python (prostředí PyTorch) a toolkit Wespeaker (Wang 2023). Pro uživatelskou jednoduchost bylo pro tento systém vytvořeno webové rozhraní, které umožňuje otestovat detektor nářečí přes internetový prohlížeč jako „cloudovou službu“.

Ekonomické parametry výsledku

Software výrazným způsobem zefektivní zpracování audiálních dat v připravované *Databázi nářečních promluv pro odbornou veřejnost*. Kódy nářečních podskupin v ní byly dosud přiřazovány manuálně, nyní bude možné daný krok do velké míry zautomatizovat. Software bude rovněž uplatněn ke zpětné kontrole daných údajů u současných záznamů (k 5. 11. 2024 jde o 2 134 nahrávek), u nichž by jinak kontrola a manuální oprava byla z časových a personálních důvodů nemožná.

Druh možnosti využití výsledku jiným subjektem

A – k využití výsledku jiným subjektem je vždy nutné nabytí licence

Požadavek na licenční poplatek

N – poskytovatel licence na výsledek nepožaduje licenční poplatek

Licence

Software slouží k interní aplikaci pro potřeby projektu a Ministerstva kultury ČR. Po domluvě je možné využití i třetími stranami, a to při dodržení zásady uvedení zdroje a zachování licence. Celé znění licence:

Pro uvedení zdroje doporučujeme formulaci:
Automatický detektor dialektu na základě audionahrávky. Verze 1.0. Vysoké učení technické v Brně – Ústav pro jazyk český AV ČR, v. v. i., 2024.

Anglická verze:
Automatic Dialect Detector Based on Audio Recording. Version 1.0. Brno University of Technology – Czech Language Institute, Czech Academy of Sciences, 2024.

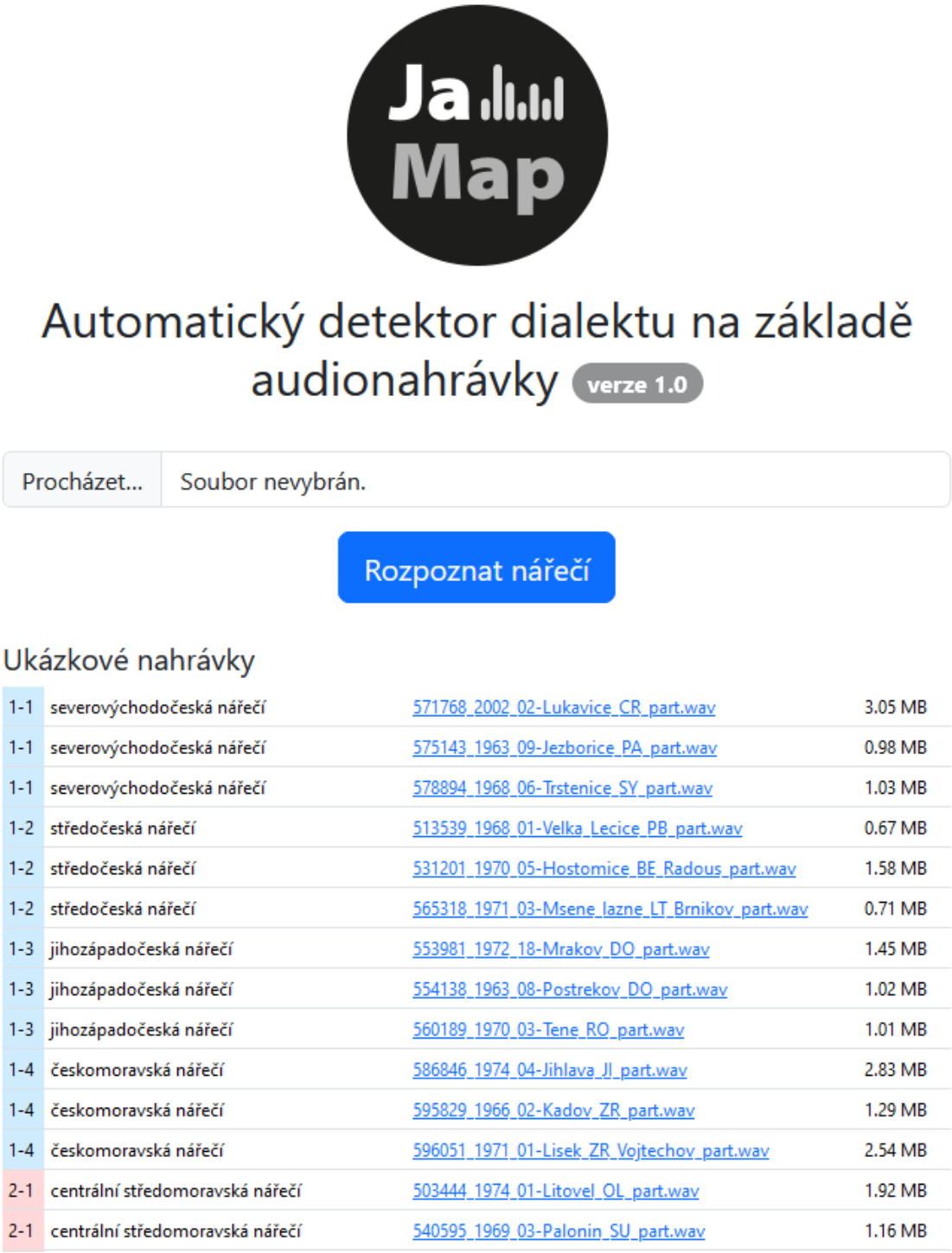
Zdrojový kód

https://jamap.cz/detektor/zdrojovy-kod.tar.gz

Návod k použití

Krok 1
Ovládání je velice intuitivní. Na webové stránce detektoru uživatel vybere soubor s nářeční nahrávkou ze svého počítače, a to pomocí tlačítka Procházet. Tlačítkem Rozpoznat nářečí pak spustí vlastní detektor. Pokud si chce uživatel software pouze otestovat, aplikace nabízí 39 ukázkových nahrávek k vyzkoušení.

Poznámka: Čísla v levém sloupci určují nářeční oblast, v níž byla nahrávka pořízena a kterou prezentuje jazykový kód mluvčího. Celkem jde o 13 nářečních skupin, konkrétně: 1-1 severovýchodočeská nářečí, 1-2 středočeská nářečí, 1-3 jihozápadočeská nářečí, 1-4 českomoravská nářečí, 2-1 centrální středomoravská nářečí, 2-2 jižní středomoravská nářečí, 2-3 západní středomoravský okrajový úsek, 2-4 východní středomoravský okrajový úsek, 3-1 jižní východomoravská nářečí, 3-2 severní východomoravská nářečí, 3-3 kopaničářská nářečí, 4-1 slezskomoravská nářečí, 4-2 slezskopolská nářečí. Týž číselný kód je využit i při vyhodnocení automatické detekce, viz krok 2.

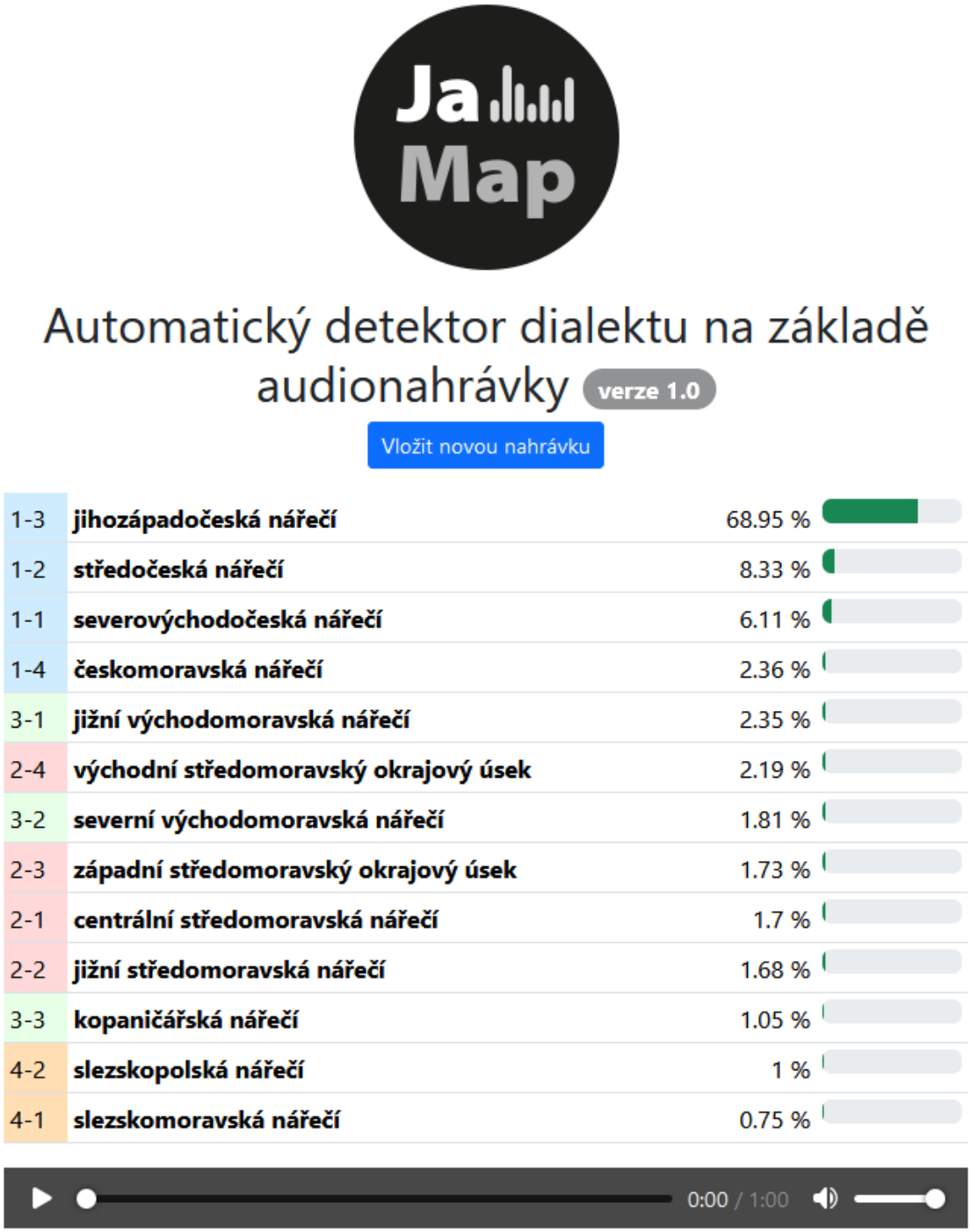


Obr. 1: Náhled rozhraní

Krok 2

Výstupem jsou automaticky generované pravděpodobnosti určující příslušnost nahrávky k určité nářeční podskupině (nebo podskupinám). Pravděpodobnost je vyjádřena procentuálně i graficky. Pro sluchovou kontrolu lze přehrát 60sekundový záznam, automaticky se načítající ze začátku nahraného audiosouboru.

Poznámka: Pro správné vyhodnocení je nezbytné pracovat se záznamy tradičních nářečních promluv (nikoliv smíšených nebo nenářečních projevů), a to jedné osoby (nikoliv více osob), která byla v době pořízení nahrávky zástupcem starší generace, přičemž splňovala i další požadavky kladené na ideálního nářečního mluvčího (k tomu Šimečková 2024). Rozpoznávač je v této verzi optimalizován na záznamy pořízené starší nahrávací technikou, zejména na audiální dokumenty z 60. a 70. let 20. století.



Obr. 2: Náhled výsledku detekce

Charakteristiky výsledku

Automatický detektor dialektu na základě audionahrávky je první elektronický výstup z řady softwarů vyvíjených v projektu Jazyková paměť regionů České republiky. Metody strojového učení pro uchování, dokumentaci a prezentaci nářečí českého jazyka. Ministerstvo kultury ČR, NAKI III, DH23P03OVV010. Software umožňuje u vstupní audionahrávky určit, ze které ze 13 předem definovaných nářečních oblastí (na úrovni nářečních podskupin) mluvčí pochází/pocházel, a to včetně míry pravděpodobnosti.

Jedinečnost a novost softwaru

Jde o první odborný detektor nářečí českého jazyka, a to jak po stránce architektury softwaru, tak trénovacími daty, na nichž byl nástroj vyvinut (viz popis technických parametrů níže).

Naplnění očekávaných dopadů programu NAKI III a využití softwaru

Software usnadňuje práci s nářečími, s jejich audiální podobou, umožňuje určovat teritoriální dialekt a nářeční příslušnost mluvčího, je využitelný např. při pořádání fonoték a databází mluveného slova, nahrávek terénních výzkumů apod., ale zapojit se dá prakticky kdekoli, kde se pracuje s neformálním mluveným projevem. Primárně je určen pro odbornou veřejnost zpracovávající audiální data (lingvisté, etnologové, orální historikové), jednoduché webové rozhraní a jeho snadné ovládání a srozumitelná prezentace výsledků však výrazně rozšiřuje paletu jeho možných využití (od pracovníků rozhlasu a televize přes studenty až po laické zájemce o jazyk). Značný potenciál má zvláště ve vzdělávání.
Těmito charakteristikami detektor vyhovuje důrazu, který je v programu NAKI III kladen na směřování výzkumu k výsledkům podporujícím rozvoj inteligentních specializací ve smyslu RIS3.
Odpovídá také následujícím bodům očekávaných dopadů programu NAKI III:
3) zpřístupnění kulturního dědictví široké obci zájemců a uživatelů,
4) posílení a integrace sociálně ekonomického uplatnění kulturního dědictví ve společnosti,
6) výzkum nástrojů a jejich ověření pro:

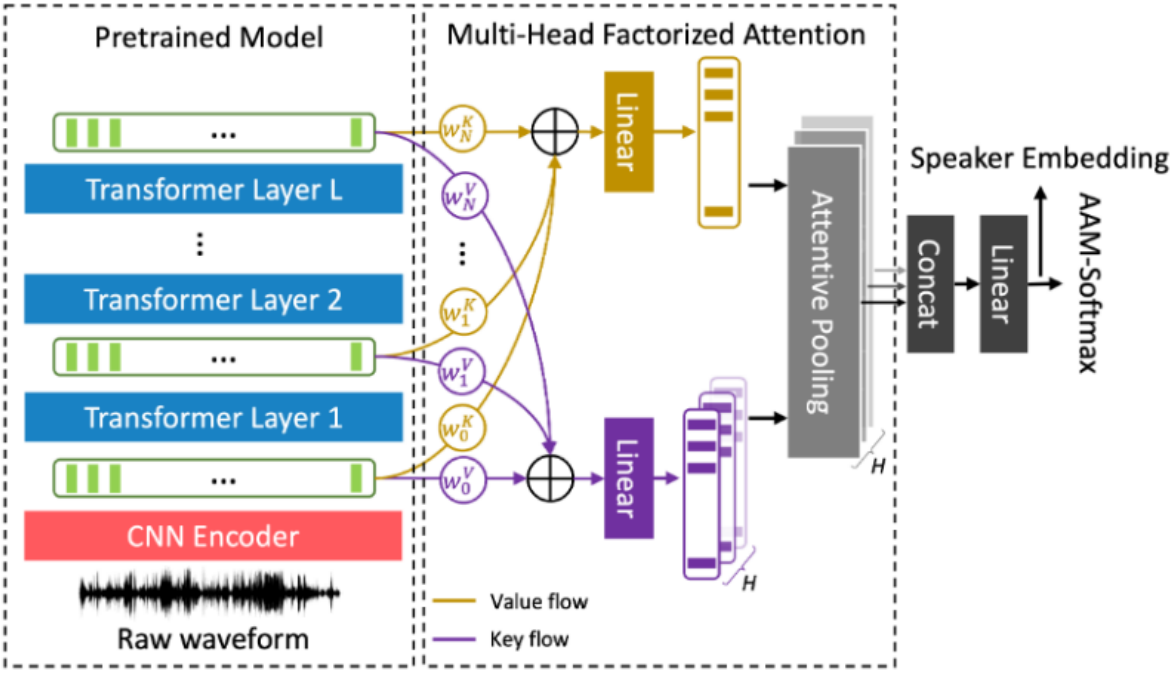
a) aktivní podíl na rozvíjení národní identity jako jednoho z elementů spoluvytváření evropské identity a kultury,
b) systematickou dokumentaci a prezentaci způsobu života a kultury menšin v minulosti i současnosti,
k) systematické a efektivní využívání prostředků technologií pro rozvoj národní kultury a kulturní identity obyvatelstva.

Technické parametry

Architektura

Architektura systému byla postavena na nejnovějších poznatcích v oblasti rozpoznávání mluvčích (Speaker Verification) a detekce jazyka (Language Identification). Ty využívají principu předtrénovaných modelů, které je možno trénovat na obrovském množství neanotované řeči a které se učí obecnou strukturu mluveného jazyka. Jednotlivé vrstvy tohoto modelu pak slouží jako vstupní parametry finálního detektoru, který je již trénován na koncových datech. Tento přístup je detailně popsán a analyzován pro oblast detekce mluvčích (Peng 2022). Zde byly analyzovány různé přístupy efektivní kombinace výstupů jednotlivých vrstev pro finální klasifikátor. Nejlepších výsledků bylo dosaženo pomocí techniky Multihead Factorized attention a WavLM base plus[1] modelu. Architektura systému je zobrazena na obr. 3.

[1] https://huggingface.co/microsoft/wavlm-base-plus



Obr. 3: Architektura systému pro verifikaci mluvčích, která byla použita pro detekci dialektu (převzato z Peng 2022)

Trénovací data

Systém byl nejprve natrénován na datech VoxLingua (Valk 2020), která obsahují 6 628 hodin audia ze 107 jazyků. Dále byl trénován na českých nářečních datech, poskytnutých dialektologickým oddělením ÚJČ AV ČR, v. v. i. Jedná se o unikátní data získávaná terénním výzkumem od 50. let 20. století do současnosti, která jsou uložena v Archivu zvukových záznamů nářečních promluv a nyní převáděna do budované Databáze nářečních promluv pro odbornou veřejnost (k těmto zdrojům viz Šimečková 2024). Pro trénování systému bylo použito 607 hodin nahrávek příslušejících do 13 nářečních podskupin. Tato data byla dále čištěna, přičemž pomocí techniky diarizace byli odstraněni vedlejší, tj. nenářeční mluvčí (zpravidla explorátoři vedoucí dialog s nářečními mluvčími). Tím byla získána koncová data o celkovém rozsahu cca 434 hodin.

Výstupní skóre detektoru

Detektor vrací skóre pro každou z 13 tříd (nářečních podskupin). Surové skóre vycházející z modelu ve formě věrohodnosti (likelihood) je transformováno do posteriorní pravděpodobnosti a vyjádřeno v procentech. Tato pravděpodobnost je následně ukázána uživatelům demonstrátoru.
Upozorňujeme, že konvertor na posteriorní pravděpodobnost nebyl kalibrován. Kalibrací se myslí naladění výstupních skóre tak, aby dávala smysl pro konkrétní nasazení (uživatelský přístup nebo cílová evaluační metrika). Důvodem je skutečnost, že detektor bude v budoucnu použit jako modul/komponenta celkového systému pro automatizované zpracování nahrávek. Pro tento účel bude později kalibrován včetně tvorby a aplikace kalibračních dat.

Ke skóre tedy doporučujeme přistupovat tak, že řadí jednotlivé kategorie. Uživatele by měla zajímat informace, která kategorie je vítězná a eventuálně která je vyhodnocena jako druhá. V žádném případě nelze například výsledek

interpretovat tak, že se v nahrávce vyskytuje projev obsahující 14 % slezskopolského nářečí. Daný výsledek lze interpretovat pouze tím způsobem, že automatický detektor nářečí s nadpoloviční pravděpodobností odhaduje, že se v nahrávce vyskytuje kopaničářské nářečí. Je však nezanedbatelná šance, že je na nahrávce zachyceno slezskopolské nářečí.

Pokud se stane, že výsledek detekce jsou 2 nebo 3 třídy s podobným skóre, napovídá to, že nářečí v nahrávce nelze jednoznačně určit.

Dále z principu fungování detektoru a trénovacích dat (historické záznamy) bude detektor náchylný k chybám na nových datech, která byla pořízena modernější, tudíž i kvalitnější technikou. Ke sběru takových dat dochází v současné době a budou přidána do trénování v pozdější fázi projektu. Tím dojde ke zvýšení robustnosti detektoru. Detektor se při tréninku učí hledat různé akustické jevy charakteristické pro danou třídu (nářečí). Pokud se například pouze pro určitou oblast používal specifický typ magnetofonu nebo mikrofonu, detektor se tuto informaci může naučit, a snižuje se tím jeho robustnost. Z toho důvodu je potřeba dbát na velkou variabilitu trénovacích dat, přičemž nová kvalitní data jsou velmi důležitá pro budoucí vývoj podobných systémů postavených na strojovém učení.

Modul detekce nářečí bude v této verzi sloužit zejména ke zpracování, sjednocení a roztřídění archivu historických nahrávek.

Popis demonstrátoru

Detektor nářečí je implementován jako nástroj v příkazové řádce v jazyce Python, daná forma je tudíž primárně vhodná pro použití softwaru v rámci vědeckovýzkumné činnosti, nikoliv pro testování méně technicky zdatnými uživateli. Z toho důvodu bylo implementováno i webové rozhraní a „cloudové“ zapouzdření detektoru. Zjednodušené schéma demonstrátoru zachycuje obrázek 4.

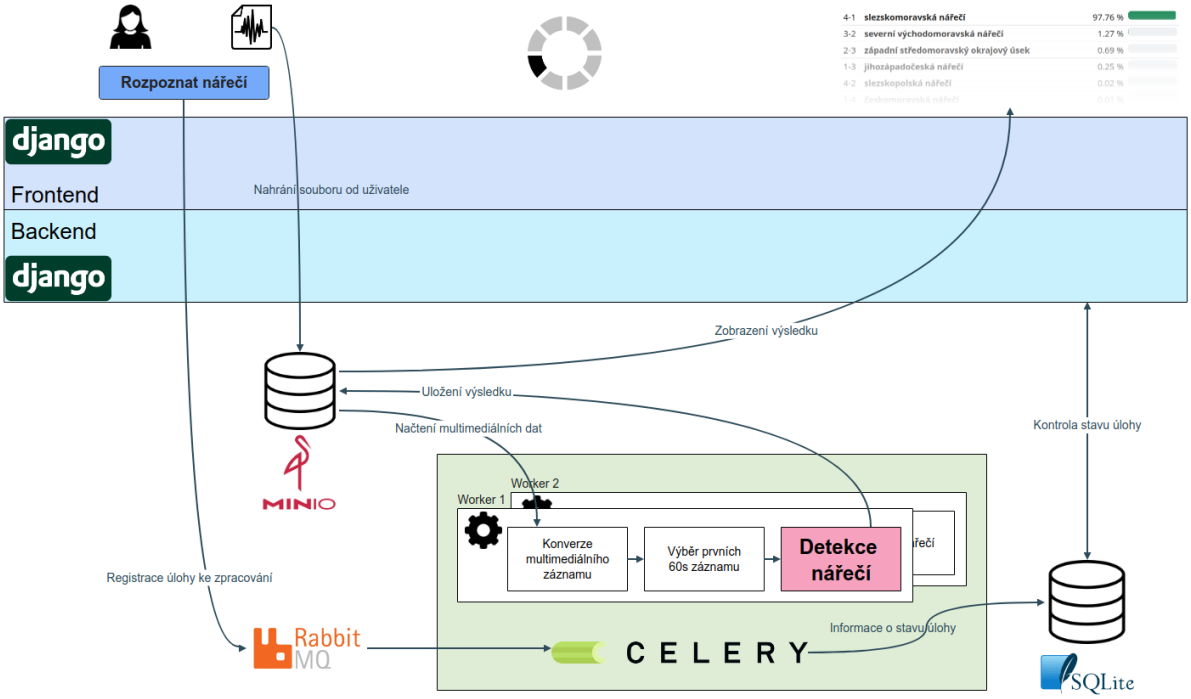
Demonstrátor se skládá z následujících komponent:

- ➢ Minio = datové úložiště
- ➢ RabbitMQ = obsluha fronty zpracovávaných úloh
- ➢ Celery = distribuovaný systém zpracování úloh
- ➢ Django = backend pro Celery; webový frontend

Výše uvedené komponenty běží jako služby na dedikovaném serveru na Fakultě informačních technologií Vysokého učení technického v Brně. Demonstrátor je přístupný z adresy https://jamap.cz/detektor. Po vložení nahrávky uživatelem a stisku tlačítka Rozpoznat nářečí (viz obr. 1) je tato nahrávka odeslána a uložena do Minio úložiště. Dále je webovým frontendem vytvořena úloha detekce nářečí z dané nahrávky. Uživateli se zobrazí rotující ikona, která znázorňuje, že se úloha zpracovává. Frontend se průběžně dotazuje backendu na výsledek zpracování.
Úloha je přes frontu RabbitMQ poslána ke zpracování. Pokud jsou k dispozici volné prostředky pro zpracování (worker), tak Celery přijme úlohu a je zahájeno zpracování.
V současné době je možné zpracovávat maximálně 2 úlohy současně. Pokud uživatelé vloží v krátkém sledu více úloh, musí počkat, než se dostanou na řadu.
Úloha je rozdělena do 3 po sobě jdoucích kroků:

- Konverze vstupních dat na požadovaný formát `PCM mono s16le 16kHz`. Vstupem tedy může být „libovolný“ multimediální soubor včetně videozáznamu.
- Extrakce prvních 60 sekund z nahrávky. K tomuto kroku se přistupuje z důvodu omezení paměťových nároků detektoru.
- Detekce nářečí z 60sekundové nahrávky a uložení výsledku do databáze backendu včetně indikace úspěšného dokončení úlohy.

V momentě, kdy frontend detekuje úspěšné dokončení úlohy, zobrazí výsledky uživateli (viz obr. 2).



Obr. 4: Zjednodušené schéma demonstrátoru detekce nářečí. Komponenta detektoru nářečí je zvýrazněna červeně

Literatura

Peng, Junyi et al. (2022). An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification. *arXiv*: 2210.01273. [online]. Dostupné z: https://arxiv.org/abs/2210.01273.

Šimečková, Marta (2024). *Archiv zvukových záznamů nářečních promluv*. Praha: Academia. Též online. Dostupné z: http://www.vedakolemnas.cz/sys/galerie-download/VKN-132_.pdf.

Valk, Jörgen a Alumäe, Tanel (2020). VoxLingua107: a Dataset for Spoken Language Recog-
nition. *arXiv*: 2011.12998 [eess.AS]. [online]. Dostupné z: https://arxiv.org/abs/2011.12998.

Wang, Hongji et al. (2023). Wespeaker: A research and production oriented speaker embedding learning toolkit. *arXiv*: 2210.17016. [online]. Dostupné z: https://arxiv.org/abs/2210.17016.