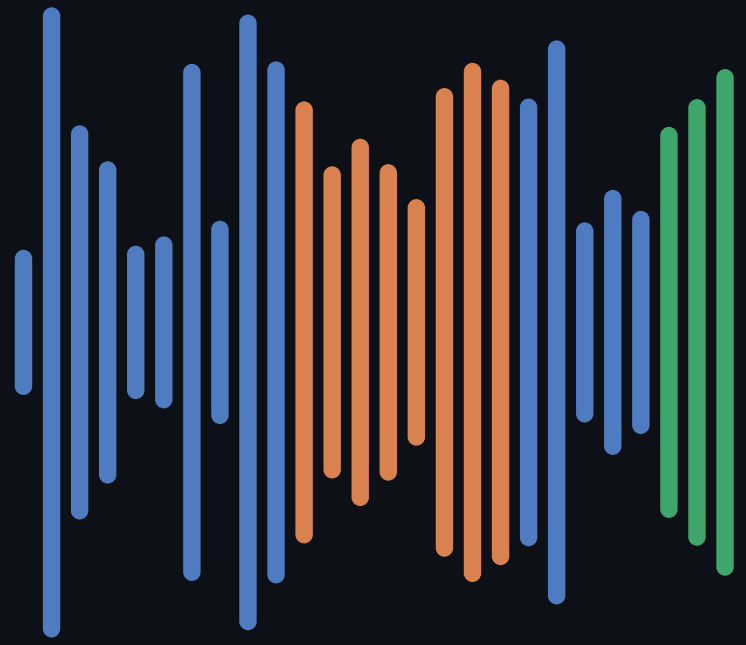
NVIDIA AI DAYS 2026 · 10. ROČNÍK

Jak se měnil vývoj AI aplikací během posledních let

Od HMM k vícejazyčným LLM modelům: 30 let rozpoznávání řeči a služba mCaptions



Ing. Josef Vopička, Ph.D.
spoluzakladatel, The MAMA AI

O mně

Téměř 30 let s rozpoznáváním řeči — od výzkumu po vlastní produkt

ČVUT FEL



Ph.D. — rozpoznávání řeči

- Fakulta elektrotechnická
- Rozpoznávání řeči (STT)
- Keyword spotting

1997



IBM Praha, R&D laboratoř

- Tým pod vedením Jana Šedivého
- Rozpoznávání řeči (STT)
- Od serverů po embedded

2021



The MAMA AI

- Spoluzakladatel
- Zpracování médií: zvuk, obraz, signál
- Produkt mCaptions

Evoluce technologie rozpoznávání řeči

Od integer aritmetiky na ARM po CUDA na NVIDIA GPU

1 GMM-HMM systémy

Směs Gaussových rozdělení + skrytý Markovův model; přepis telefonních hovorů

`C`

2 Embedded STT

Edge zařízení, gramatiky, ARM bez FPU → algoritmy v integer aritmetice; automotive

`C++`

3 Sjednocený produkt

Jeden HMM engine pro embedded i telefonní svět

`C++`

4 DNN-HMM hybrid

Neuronová síť (DNN) extrahuje příznaky, dekóduje HMM dekodér

`C++ → Python`

5 Plně neurální modely

RNN-Transducer s pomocnou vrstvou CTC (Connectionist Temporal Classification)

`Python + C++`

6 Vícejazyčné LLM modely

OpenAI Whisper, NVIDIA Parakeet a Nemotron (detail na dalším slidu)

`Python + CUDA`

Dnešní modely: Whisper, Parakeet, Nemotron

Tři přístupy k řeči v éře velkých modelů

Whisper

OpenAI

KATEGORIE TECHNOLOGIE

Samostatný model pro rozpoznávání řeči (STT / ASR) a překlad

ZÁKLADNÍ ARCHITEKTURA

Encoder-Decoder Transformer (Seq2Seq)

IDEÁLNÍ NASAZENÍ

Dávkový přepis podcastů, titulků, přednášek a překlady cizích jazyků

Parakeet

NVIDIA

KATEGORIE TECHNOLOGIE

Vysoce specializovaný, rychlý ASR model

ZÁKLADNÍ ARCHITEKTURA

FastConformer + CTC / TDT / RNN-T dekodéry

IDEÁLNÍ NASAZENÍ

Real-time přepis telefonních hovorů, call centra, rychlé diktování

Nemotron

NVIDIA

KATEGORIE TECHNOLOGIE

Rodina modelů: LLM, multimodální modely a řečové služby (NIM)

ZÁKLADNÍ ARCHITEKTURA

MoE / Dense Transformer + NIM Speech pipelines

IDEÁLNÍ NASAZENÍ

Komplexní autonomní hlasoví agenti, firemní boti a interaktivní asistenti

Co se změnilo ve vývoji

Dříve hodně specializovaných modelů na CPU, dnes jeden velký model na GPU

	Tehdy · GMM-HMM	Dnes · LLM-based STT
Jazyky	Model pro každý jazyk a doménu	Jeden vícejazyčný model
Slovník	Gramatiky, omezený slovník	Otevřený slovník, interpunkce
Hardware	CPU, ARM, integer aritmetika	NVIDIA GPU, CUDA
Vývoj	C / C++	Python + CUDA
Nasazení	Knihovna v zařízení nebo na serveru	Kontejnery, Kubernetes, SaaS i on-prem

Méně vlastních algoritmů, více práce s modely, daty a provozem

mCaptions

mcaptions.themama.ai

Příklad dnešní AI aplikace: přepis a shrnutí schůzek a hovorů



Nahrávka

Mobil na stole, web nebo nahraný soubor

Přepis

Vícejazyčný LLM STT model na NVIDIA GPU

Diarizace

Kdo kdy mluví

Mluvčí

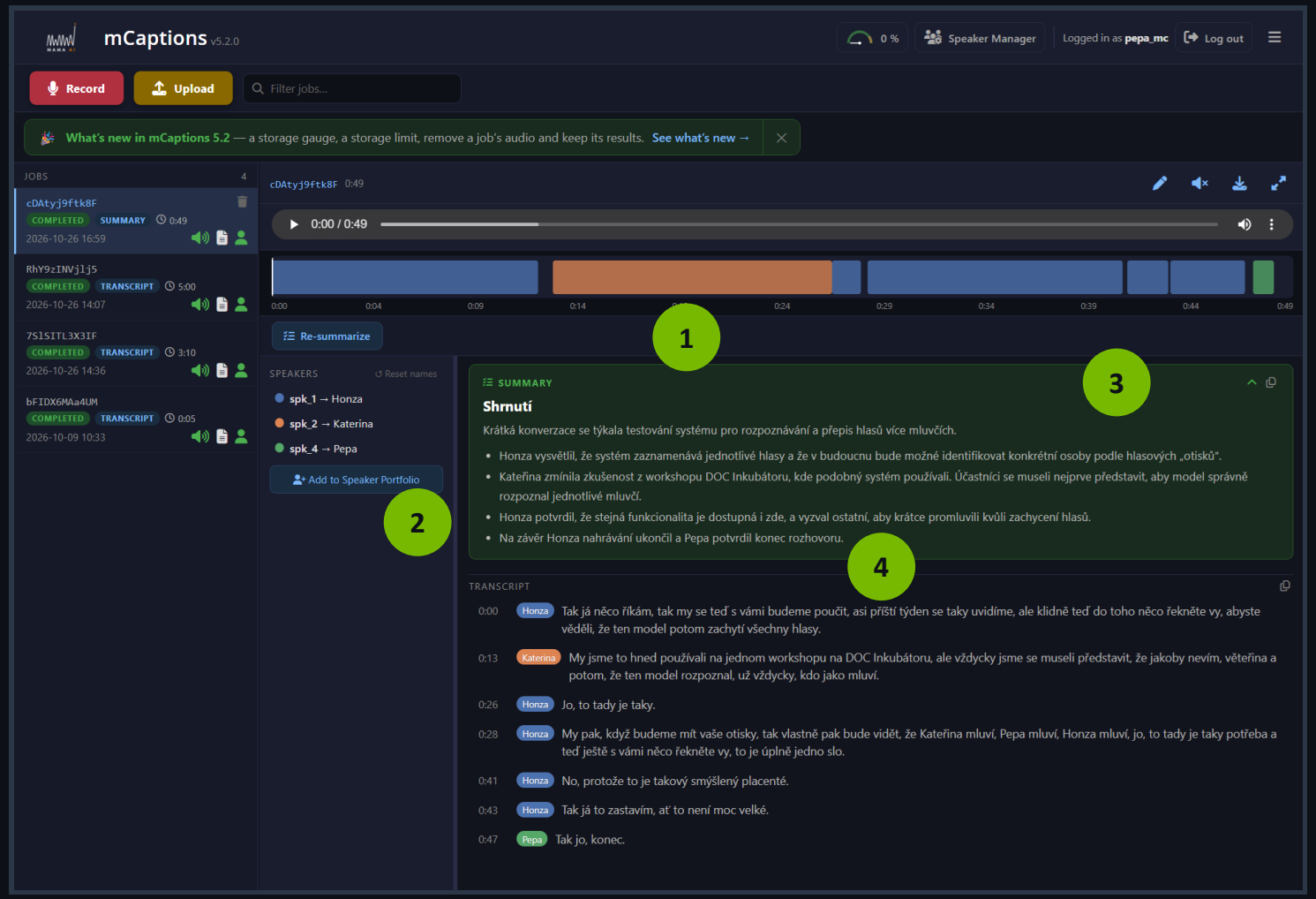
Každý uživatel si mluvčí pojmenuje sám

Shrnutí

LLM velkého poskytovatele nebo lokální

mCaptions v praxi

Krátká schůzka tří lidí: časová osa mluvčích, shrnutí a přepis



1. **Časová osa mluvčích**

   Diarizace barevně ukazuje, kdo kdy mluví

2. **Pojmenovaní mluvčí**

   spk_1 → Honza; profil mluvčího se uloží pro další nahrávky, vidí ho jen autor

3. **Shrnutí jazykovým modelem**

   Klíčové body schůzky během chvilky; uživatel může zadat vlastní prompt

4. **Přepis s časovými značkami**

   Každá promluva přiřazená mluvčímu

SaaS i on-prem: architektura a soukromí

Stejné kontejnery běží v našem cloudu i u zákazníka

SaaS · náš cloud, zákazník nemusí mít vlastní hardware

On-prem · plně autonomní, vlastní NVIDIA GPU

Kubernetes · náš cloud nebo infrastruktura zákazníka

Klienti

Web, mobil na stole

mCaptions

Web UI, API, správa uživatelů, metering

GPU workery

Scheduler: STT, diarizace a mluvčí na NVIDIA GPU

LLM

Lokální model, v SaaS i LLM velkého poskytovatele

Vše v kontejnerech
škálování přes Scheduler a Kubernetes

Privátní režim

Obsah se neukládá

V systému zůstává pouze záznam o délce zpracovaného audia

Soukromí mluvčí

Jména vidí jen uživatel, který je vytvořil

On-prem: ven nic

Audio, přepis ani shrnutí neopouští infrastrukturu zákazníka

Model je jen část aplikace

Co dnes kolem modelu vyvíjíme v mCaptions

Kontejnerizované řešení

Celý systém včetně modelů zabalený v kontejnerech

Škálování

Scheduler a Kubernetes přidávají GPU workery podle zátěže

Hierarchická správa uživatelů

Víceúrovňová struktura účtů pro celou organizaci

Metering

Měří se přepsané minuty, nikdy obsah přepisu

Dimenzování GPU

Paměť GPU a propustnost podle modelu, souběhu a objemu audia

Lokální LLM

Menší model na shrnutí sdílí GPU se STT

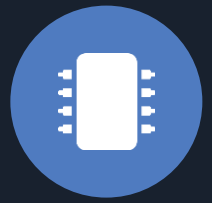
Co si odnést

Jak se za téměř 30 let změnil vývoj AI aplikací



Modely

Jeden vícejazyčný model nahradil desítky specializovaných modelů pro jazyky a domény



Hardware

Od ARM bez FPU a integer aritmetiky k NVIDIA GPU a CUDA



Aplikace

Hodnota je dnes v aplikaci kolem modelu: škálování, soukromí, SaaS i on-prem

OTÁZKY?

Děkuji

Ing. Josef Vopička, Ph.D.
josef.vopicka@themama.ai

