FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

KATEDRA INFORMATIKY
A VÝPOČETNÍ TECHNIKY

Bakalářská práce

# Integrovaný systém pro automatizaci záznamu, úprav a publikaci přednášek

Martin Reich

PLZEŇ

2024

FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

KATEDRA INFORMATIKY
A VÝPOČETNÍ TECHNIKY

## Bakalářská práce

# Integrovaný systém pro automatizaci záznamu, úprav a publikaci přednášek

Martin Reich

**Vedoucí práce**
Doc. Ing. Roman Mouček, Ph.D.

PLZEŇ 2024

**Citace v seznamu literatury:**
REICH, Martin. *Integrovaný systém pro automatizaci záznamu, úprav a publikaci přednášek*. Plzeň, 2024. Bakalářská práce. Západočeská univerzita v Plzni, Fakulta aplikovaných věd, Katedra informatiky a výpočetní techniky. Vedoucí práce Doc. Ing. Roman Mouček, Ph.D.

ZÁPADOČESKÁ UNIVERZITA V PLZNI
Fakulta aplikovaných věd
Akademický rok: 2024/2025

Studijní program: Informatika a výpočetní technika
Forma studia: Prezenční
Specializace/kombinace: Informatika (INF18bp)

# Podklad pro zadání BAKALÁŘSKÉ práce studenta

| | |
|---|---|
| Jméno a příjmení: | **Martin REICH** |
| Osobní číslo: | **A22B0123P** |
| Adresa: | Plovární 1458/21, Plzeň – Jižní Předměstí, 30100 Plzeň 1, Česká republika |
| Téma práce: | Integrovaný systém pro automatizaci záznamu, úprav a publikaci přednášek |
| Téma práce anglicky: | An integrated system for automating the recording, editing and publication of lectures |
| Jazyk práce: | Čeština |
| Související osoby: | Doc. Ing. Roman Mouček, Ph.D. (Vedoucí)<br>Katedra informatiky a výpočetní techniky |

Zásady pro vypracování:

1. Seznamte se současnými metodami a možnostmi převodu řeči do textu a zpět a metodami parafrázování
2. Seznamte se současnými systémy, nástroji a knihovnami pro transkripci řeči na text, převod textu na řeč a parafrázování
3. Navrhněte integrovaný systém, který dokáže automatizovat převod zvukového a obrazového záznamu přednášky v češtině či angličtině do textu, tento následně parafrázovat a zpětně převést do zvukové a obrazové podoby přednášky
4. Navržený systém dle bodu 3 implementujte.
5. Ověřte výsledné řešení na dostatečném počtu přednášek v českém i anglickém jazyce.

Seznam doporučené literatury:

Dodá vedoucí práce

Podpis studenta: Datum:

Podpis vedoucího práce: Datum:

# Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného akademického titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Západočeská univerzita v Plzni má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V Plzni dne 31. prosince 2024

. . . . . .

Martin Reich

# Abstrakt

TODO

# Abstract

TODO in english

## Klíčová slova

Převod řeči na text • AI • Převod textu na řeč • Parafrázování • Automatizace

# Poděkování

Rád bych tímto poděkoval vedoucímu bakalářské práce Doc. Ing. Romanu Moučkovi, Ph.D. za pomoc a odborné vedení při vypracování této práce.

# Obsah

# Úvod 1

# Analýza problému 2

**Na schůzce probrat, jestli je to potřeba nebo to předělat(většina věcí je zmíněna v ostatních kapitolách)**

## 2.1 Studium současných metod převodu řeči do textu a zpět, včetně parafrázování

V první fázi je nutné porozumět současným technologiím pro:

- Převod řeči na text: Mezi nejznámější metody patří systémy založené na hlubokých neuronových sítích, konkrétně na architekturách jako jsou konvoluční neuronové sítě (CNN) a rekurentní neuronové sítě (RNN), nebo i pokročilejší transformery. Pro hodnocení efektivity a vhodnosti těchto metod se používají různé metriky. K těm nejčastěji využívaným patří standardní metriky hlubokého učení, jako je přesnost (accuracy), F1-skóre, recall (senzitivita nebo míra pravdivých pozitiv, TPR), precision (pozitivní prediktivní hodnota) a specificity (míra pravdivých negativ, TNR).
- Převod textu na řeč: Nejnovější metody TTS využívají neuronové sítě, jako třeba TTS model od OpenAI, ElevenLabs modely, Google TTS model atd.. Tyto modely zajišťují přirozenější syntézu řeči a mohou pracovat s intonací a rytmem, což je klíčové pro parafrázované texty. K hodnocení výkonnosti těchto modelů se využívají objektivní a subjektivní metriky jako např.: Mel cepstral distortion (MCD), měřící podobnost syntetizovaného a přirozeného zvuku. Root mean square error (RMSE) pro logaritmus základní frekvence (f0). Voiced/unvoiced (V/UV) error rate pro analýzu správnosti rozhodnutí o znělosti. Gross pitch error (GPE) pro hodnocení odchylek výšky hlasu. Mean opinion score (MOS), kde posluchači hodnotí kvalitu syntézy na stupnici od 0 do 5. Více zde [**KS2023**].

- Parafrázování: Parafrázování textu je možné provádět pomocí modelů NLP (natural language processing, zpracování přirozeného jazyka), které dokážou přepsat text jinými slovy, aniž by změnily význam. Zde se využívají transformer modely, například T5 nebo GPT-4. Při parafrázování je výzvou zachovat srozumitelnost a přirozenost textu, zejména pokud se bude převádět zpět na řeč.

## 2.2 Průzkum dostupných modelů, nástrojů a knihoven a jejich vyzkoušení

Druhá fáze zahrnuje detailní průzkum stávajících nástrojů, modelů a knihoven, které podporují každý krok procesu a jejich vyzkoušení. Pro STT:

- Modely a nástroje:
  - OpenAI Whisper: Open-source model s vysokou přesností a podporou více jazyků.
  - Google Speech-to-Text API: Komerční služba s možností přizpůsobení pro konkrétní domény.
  - DeepSpeech: Open-source model od Mozilly vhodný pro vlastní nasazení.
  - Microsoft Azure Speech-to-Text: Cloudová služba s podporou integrace do enterprise aplikací.
- Knihovny:
  - SpeechRecognition: Python knihovna pro jednoduchou práci s STT.
  - PyTorch a TensorFlow: Frameworky využívané pro implementaci a trénink vlastních modelů STT.

Pro TTS:

- Modely a nástroje:
  - ElevenLabs Speech Synthesis: Pokročilý model s podporou práce s emocemi a intonací.
  - Google TTS API: Komerční služba s realistickou syntézou řeči a možnostmi přizpůsobení.
  - Microsoft Azure Text-to-Speech: Služba s vysokou kvalitou syntézy a jazykovou podporou.

- Knihovny:
  - PyTorch a TensorFlow: Frameworky využívané pro implementaci a trénink vlastních modelů TTS.

Pro parafrázování:

- Modely a nástroje:
  - OpenAI GPT-4: Výkonný generativní model schopný sofistikovaného parafrázování.
  - BART (Facebook): Transformer zaměřený na generování textu a parafrázování.
  - T5 (Text-to-Text Transfer Transformer): Univerzální model od Googlu pro různé NLP úkoly, včetně parafrázování.
- Knihovny:
  - Hugging Face Transformers: Knihovna podporující širokou škálu NLP modelů, včetně BART, T5 a GPT.
  - PyTorch a TensorFlow: Frameworky využívané pro implementaci a trénink vlastních modelů.

**TODO Doptat se jak udělat úspěšnost modelů**

## 2.3 Návrh integrovaného systému a jeho implementace

Dále je potřeba si rozmyslet a navrhnout systém, který:

- Automatizuje převod mluveného slova na text: Tento krok zahrnuje přepis audia nebo video záznamů přednášky na text.
- Parafrázuje přepsaný text: Po přepisu je text předán k parafrázování, aby výsledný text byl přeformulovaný.
- Převádí parafrázovaný text zpět na řeč: Po parafrázování je třeba převést text na řeč, která by měla být pokud možno přirozená a příjemná k poslechu.
- Synchronizuje obraz a zvuk: Nakonec je potřeba synchronizovat nový zvuk s obrazem, který může být originální nebo nový ve formě animace či avatara.

Po navrhnutí takovéhoto systému je potřeba ho implementovat(viz. implementace**Dodat odkaz po přidání implementace**)

## 2.4 Testování systému

Na závěr je potřeba zvolit vhodnou metodu pro testování navrženého a implementované systému. Tento krok také zahrnuje testování podle kritérií na reálných přednáškách v českém a anglickém jazyce. Mezi tyto kritéria může patřit:

- Kvalita: Testování přesnosti rozpoznávání řeči, přirozenosti parafrázování a srozumitelnosti výstupu.
- Uživatelské hodnocení: Testování s cílovými uživateli, kteří posoudí užitečnost a přirozenost celého procesu.
- Výkon: Hodnocení rychlosti a efektivity systému, aby byla zajištěna schopnost zpracovat delší přednášky(maximálně 215 minut) bez výrazného zpoždění.

# Převod řeči do textu 3

**OTÁZKA: mám tam dopnit i ostatní metody(statistické a hybridní) a ne jen neuronky?** Převedení řeči do textu (Speech-to-Text, STT) je oblast, která se zabývá automatickým rozpoznáváním mluveného slova a jeho převodem na psaný text. Je to multidisciplinární obor zahrnující akustiku, jazykovědu, statistiku a strojové učení. Současné technologie pro převod řeči na text jsou poháněny pokročilými modely strojového učení. Tyto modely využívají akustické modely (AM) a jazykové modely (LM), které spolupracují na přesné interpretaci mluvené řeči. Hlavními pilíři moderních přístupů jsou hluboké neuronové sítě, transformery, a techniky přenosového učení (DTL), které umožňují přizpůsobení modelů specifickým podmínkám.

## 3.1 Metody

Současné metody převodu řeči na text lze rozdělit do tří hlavních kategorií:

1. Statistické metody - Skryté Markovovy modely (HMM) a Gaussian Mixture Models (GMM), které jsou tradičními přístupy k ASR.

2. Neuronové sítě - Pokročilé modely jako RNN, CNN a transformery, které nabízejí větší flexibilitu a přesnost.

3. Hybridní metody - Kombinují prvky tradičních modelů a hlubokého učení pro lepší výkonnost.

## 3.2 Umělé neuronové sítě

Umělé neuronové sítě hrají klíčovou roli v systémech pro převod řeči na text. Tyto sítě jsou inspirovány strukturou a funkcí lidského mozku a napodobují procesy rozhodování prostřednictvím propojení umělých neuronů. Každá neuronová síť se skládá z vrstev uzlů zahrnujících vstupní, skryté a výstupní vrstvy. V STT neuronové sítě dokáží analyzovat složité akustické vzory, které vznikají z mluvené řeči, a převádět je do strukturovaného textu. Tyto modely se učí z velkých množství

tréninkových dat, což jim umožňuje zlepšovat přesnost a přizpůsobit se různým jazykovým kontextům a akustickým prostředím. Optimalizované neuronové sítě mohou provádět úlohy, jako je rozpoznávání řeči, rychle a efektivně, což umožňuje nasazení v aplikacích, jako jsou virtuální asistenti nebo titulkovací systémy.
Výhody:

- Přesnost: Díky schopnosti identifikovat složité akustické a jazykové vzory jsou neuronové sítě schopny dosahovat vysoké přesnosti rozpoznávání.
- Adaptabilita: Sítě lze trénovat pro různé jazyky, akcenty a aplikace, což je činí univerzálními.
- Všestrannost: Neuronové sítě podporují různé přístupy, jako je klasická segmentace na fonémy i moderní end-to-end modely.

Nevýhody:

- Požadavky na tréninková data: Neuronové sítě potřebují velké množství kvalitních tréninkových dat, aby byly schopny efektivně se učit a generalizovat na nová data.
- Trénink neuronových sítí může být časově náročný a vyžaduje značné výpočetní zdroje, zejména při práci s velkými datovými sadami nebo složitými architekturami.

## 3.3 Hluboké učení

Hluboké učení je podmnožina strojového učení, která využívá více vrstvé neuronové sítě, nazývané hluboké neuronové sítě, k simulaci složitého rozhodovacího procesu lidského mozku. Hlavní rozdíl mezi hlubokým učením a strojovým učením spočívá ve struktuře architektury základní neuronové sítě. Tradiční modely strojového učení, které nejsou hluboké, používají jednoduché neuronové sítě s jednou nebo dvěma výpočetními vrstvami. Naopak modely hlubokého učení mají tři a více vrstev, často stovky nebo dokonce tisíce vrstev.

### 3.3.1 Rekurentní neuronové sítě

Rekurentní neuronové sítě (RNN) jsou modely hlubokého učení a jsou schopné zpracovávat sekvence dat tím, že mají vnitřní paměť, která uchovává informace o předchozích vstupech. RNN jsou široce využívané v ASR systémech, protože umožňují modelování časových závislostí ve zvukových datech. Díky své vnitřní paměti uchovávají informace o předchozích vstupech, což je klíčové pro rozpoznávání souvislé řeči. Výhody:

- Zpracování sekvenčních dat s ohledem na kontext.
- Flexibilita při práci s různě dlouhými vstupy a výstupy.
- Efektivní pro modelování přirozeného toku řeči.

Nevýhody:

- Problém mizícího gradientu při trénování dlouhých sekvencí.
- Vysoké výpočetní nároky při zpracování dlouhých sekvencí.
- Obtížná paralelizace oproti jiným modelům.

### 3.3.2 Transformer

Transformery jsou modely hlubokého učení a používají mechanismus vlastní pozornosti, který umožňuje modelu vážit různá slova v sekvenci na základě jejich relevance. To znamená, že model může posoudit, která slova mají vliv na ostatní slova, což je klíčové pro porozumění kontextu. Transformery se staly klíčovou technologií v ASR díky schopnosti efektivně modelovat dlouhodobé závislosti v řečových datech. Používají mechanismus pozornosti (self-attention), který umožňuje modelu zohlednit celou větu najednou, což zlepšuje přesnost transkripce. Výhody:

- Paralelizace výpočtů urychluje trénink.
- Lepší zachycení kontextu než tradiční RNN modely.
- Schopnost efektivně pracovat s dlouhými sekvencemi řeči.

Nevýhody:

- Vysoké nároky na paměť a výpočetní výkon.
- Vyžadují velké množství tréninkových dat.
- Složitější interpretovatelnost výstupů oproti tradičním modelům.

### 3.3.3 Konvoluční neuronové sítě

Konvoluční neuronové sítě (CNN) jsou specifickým typem neuronových sítí, které se ukázaly jako velmi efektivní při analýze vizuálních dat, jako jsou obrázky, videa a audia. CNN se ukázaly jako užitečné v ASR zejména při analýze akustických signálů a spektrogramů. CNN modely mohou efektivně extrahovat relevantní rysy ze zvukových dat a pomáhají redukovat šum, což je zásadní pro přesné rozpoznávání řeči. CNN má tři hlavní typy vrstev.

- Konvoluční vrstva: Tato vrstva aplikuje konvoluční operaci na vstupní data pomocí malých filtrů(filtr pokrývá jen část vstupních dat). Hodnoty(váhy) ve filtru se vynásobí s hodnotami na vstupních datech, preš který je filtr "umístěn". Filtr se poté přesune, aby vytvořil reprezentaci nové části vstupních dat, a proces se opakuje, dokud nejsou pokrytá celá vstupní data.
- Pooling vrstva: Podobně jako u konvoluční vsrtvy, tato vrstva aplikuje filtry na vstupní data. Jediný rozdíl je, že filtr nemá žádné váhy a ze vstupních dat se vezme hodnotu se největší nebo průměrnou hodnotou.
- Plně propojená vrstva: Na konci CNN se obvykle nachází jedna nebo více plně propojených vrstev, které kombinují výstupy z předchozích vrstev a předpovídají finální výstup.

Výhody:

- Efektivní extrakce relevantních rysů z akustických dat.
- Redukce výpočetní složitosti oproti jiným modelům.
- Robustnost vůči šumu a variabilitě vstupních řečových dat.

Nevýhody:

- Omezená schopnost zachytit dlouhodobé závislosti v datech.
- Vyžaduje kvalitní vstupní spektrogramy pro maximální výkon.
- Méně efektivní pro modelování složitějších jazykových struktur oproti RNN a transformerům.

## 3.4 Evaluační metriky

Pro hodnocení účinnosti a vhodnosti technik pro převod řeči na text se používají různé metody. Mezi běžně používané metriky patří:

- Přesnost (Accuracy): Měří celkový podíl správně rozpoznaných znaků nebo slov na celkovém počtu slov.
- F1-score: Kombinace přesnosti a citlivosti, která poskytuje rovnováhu mezi těmito dvěma metrikami.
- Recall (senzitivita nebo True Positive Rate - TPR): Měří schopnost systému správně detekovat skutečné pozitivní případy.

- Precision (pozitivní predikční hodnota): Udává, jak přesně model identifikuje pozitivní případy ve své predikci.
- Specificity (True Negative Rate - TNR): Měří schopnost systému správně detekovat negativní případy.

Kromě těchto běžně používaných metrik pro strojové učení, existují také specifické metriky pro ASR, které hodnotí různé aspekty rozpoznávání řeči, jako například:

- Word Error Rate (WER): Jedna z nejběžněji používaných metrik pro hodnocení přesnosti ASR, která měří počet chyb (výměna, přeskočení a vložení) na počet slov v referenčním textu.
- Character Error Rate (CER): Podobná WER, ale počítá chyby na úrovni znaků, což je užitečné pro jazyky, kde není vždy snadné definovat "slovo".
- Sentence Error Rate (SER): Měří, kolik celých vět bylo rozpoznáno chybně.

Výše uvedené metriky pomáhají vyhodnocovat nejen základní přesnost systému, ale i jeho schopnost adaptace na různé akustické a jazykové podmínky, což je klíčové pro efektivní nasazení ASR ve skutečných aplikacích. **Poznámka:** Hodnocení kvality ASR se může výrazně lišit v závislosti na typu úkolu (např. rozpoznávání jednotlivých slov versus rozpoznávání dlouhých konverzací) a použitém datasetu.

## 3.5 Vybrané modely

Bylo vybíráno mezi 3 modely pro převod řeči na text: Nova-2 od Deepgramu, Whisper-1 od OpenAI a Google TTS.

### 3.5.1 Nova-2 (Deepgram)

Nova-2 je model pro převod řeči na text vyvinutý společností Deepgram, který nabízí následující vlastnosti:
Výhody:

- Vysoká přesnost: Nova-2 dosahuje nižší míry chybovosti (Word Error Rate - WER) o 30 ve srovnání s konkurencí, což naznačuje vysokou přesnost transkripce.
- Rychlost: Model poskytuje rychlou transkripci s nízkou latencí, což je vhodné pro aplikace vyžadující reálný přepis řeči.
- Nákladová efektivita: Deepgram nabízí konkurenceschopné ceny, přičemž Nova-2 je cenově dostupný model pro převod řeči na text.

Nevýhody:

- Omezená podpora jazyků: I když Nova-2 podporuje několik jazyků, jeho pokrytí nemusí být tak široké jako u některých konkurentů.

### 3.5.2 Whisper-1 (OpenAI)

Whisper je model pro převod řeči na text vyvinutý společností OpenAI. I když byl navržen pro širokou škálu jazyků, existují určité obavy týkající se jeho přesnosti:
Výhody:

- Vícejazyčná podpora: Whisper podporuje širokou škálu jazyků, což umožňuje jeho použití v různých jazykových prostředích.
- Flexibilita: Model je navržen tak, aby byl použitelný v různých aplikacích a scénářích.

Nevýhody:

- Neúplná přesnost: Whisper může generovat nepřesné nebo smyšlené texty, zejména v tichých pasážích nahrávek.
- Bezpečnostní obavy: V citlivých prostředích, jako jsou nemocnice, mohou nepřesnosti vést k závažným důsledkům, jako jsou nesprávné diagnózy.

### 3.5.3 Google STT (Google Cloud Speech-to-Text)

Google Cloud Speech-to-Text je služba pro převod řeči na text vyvinutá společností Google. Nabízí širokou škálu funkcí a vlastností, které mohou být užitečné v různých aplikacích a scénářích:
Výhody:

- Široká podpora jazyků: Google Cloud Speech-to-Text podporuje více než 125 jazyků a dialektů, což umožňuje využití v různých jazykových prostředích.
- Pokročilé funkce přepisu: Služba nabízí identifikaci mluvčích, časové kódy, titulky a vlastní slovník, což umožňuje přesnější a přizpůsobený přepis.
- Možnost provozu na zařízení: Google Cloud Speech-to-Text umožňuje provozovat rozpoznávání řeči přímo na zařízeních bez nutnosti připojení k síti, což zajišťuje rychlost a ochranu soukromí.

Nevýhody:

- Omezení v některých jazycích: Kvalita přepisu může být nižší u méně běžných jazyků nebo dialektů.

- Omezené formáty: Google STT nepodporuje třeba video formáty, takže uživatel nejdřív musí převést audio např. z mp4 formátu do mp3 nebo podobných formátů.
- Velikostní omezení: Google STT podporuje v jednom požadavku maximálně audio velké 10MB.

### 3.5.4 Testování a výsledky modelů

**TODO dodat graf/y s výsledky testování modelů podle rychlosti a přesnosti a napsat který model bude použit k řešení probmlému.**

# Parafrázování 4

Parafrázování je proces přeformulování textu nebo myšlenek, při kterém zachováme původní význam, ale změníme slovní formulaci a strukturu. Cílem parafrázování je převést informace jinými slovy, často pro lepší pochopení, přizpůsobení určitému publiku nebo aby se text stal originálnějším a nevznikl tak problém s plagiátorstvím.

## 4.1 Tradiční metody parafrázování

Tradiční přístupy k parafrázování jsou metody, které nevyužívají neuronové sítě, a spoléhají na pravidla, tezaury[1] nebo statistické modely. Tyto přístupy byly v počátcích výzkumu parafrázování široce využívány a poskytly základ pro pozdější vývoj pokročilejších technik.

### 4.1.1 Pravidlové přístupy

Tyto metody parafrázování využívají pravidla, která mohou být buď ručně vytvořena, nebo automaticky generována. V počátečních výzkumech byla pravidla většinou sestavována manuálně, což bylo časově náročné. Následně byly navrženy postupy umožňující automatizované získávání těchto pravidel.

### 4.1.2 Přístupy založené na tezaurech

Tento přístup generuje parafráze nahrazením některých slov ve zdrojových větách jejich synonymy získanými z tezauru. Proces zahrnuje nejprve extrakci všech synonym pro nahrazovaná slova a následný výběr optimálního kandidáta podle kontextu zdrojové věty.

[1] Řízený slovník termínů (klíčových slov), které se používají pro vyjádření obsahu dokumentu.

### 4.1.3 Přístupy založené na statistickém strojovém překladu (SMT)

Tento přístup vychází z myšlenky, že parafrázování lze považovat za speciální případ strojového překladu, konkrétně za monolingvální překlad. Model statistického překladu hledá nejlepší překlad zdrojového textu na základě pravděpodobnostního a jazykového modelu. Analogicky při parafrázování hledá model nejlepší parafrázi zdrojového textu za pomocí pravděpodobnostního a jazykového modelu.

## 4.2 Neuronové přístupy

S rozvojem neuronových sítí, zejména frameworku sekvence na sekvenci (Seq2Seq), se začaly využívat neuronové modely pro generování parafrází. První použití těchto modelů pro parafrázování bylo popsáno Prakashem et al. (2016), což inspirovalo široké využití neuronových modelů v této oblasti. Níže jsou popsány hlavní přístupy využívající neuronové modely.

### 4.2.1 Architektura kódovač-dekódovač

Většina současných modelů pro generování parafrází je založena na modelech sekvence na sekvenci sestávajících z kódovače a dekodéru. Kódovač převádí vstupní text na vektorovou reprezentaci zachycující jeho význam, zatímco dekodér na základě této reprezentace generuje parafráze.

#### 4.2.1.1 Kódovací část

Cílem kódovače je extrakce sémantické informace potřebné pro generování parafrází. Mezi nejpoužívanější přístupy patří:

- Rekurentní neuronové sítě (RNN) – LSTM sítě byly prvními využitými modely pro zpracování dlouhých sekvencí.
- Konvoluční neuronové sítě (CNN) – Díky menšímu počtu parametrů jsou rychlejší na trénování.
- Transformery – Díky schopnosti zachytit dlouhodobé závislosti dosahují špičkových výsledků v generování textu.
- Předtrénované jazykové modely – Modely jako GPT-2 a BART jsou nyní využívány jako kódovač-dekódovač rámce.

#### 4.2.1.2 Dekódovací část

Dekodér využívá kontextovou reprezentaci při každém kroku dekódování a generuje výstupní text. Mezi nejčastější metody patří:

- Greedy decoding – Výběr slova s nejvyšší pravděpodobností.
- Beam search – Identifikuje k nejlepší cesty během dekódování.
- Blokovací mechanismy – Zabraňují generování stejných slov jako ve vstupním textu, čímž zvyšují rozmanitost parafrází.

### 4.2.2 Vylepšení architektury kódovač-dekódovač

Vylepšení těchto modelů lze rozdělit do dvou kategorií:

- Zlepšení modelu – Mechanismy jako Attention, Copy a variational autoencoder (VAE) umožňují lepší generování parafrází.
- Zaměření na atributy – Metody zaměřené na rozmanitost generovaných parafrází, syntaktickou kontrolu a granularitu.

## 4.3 Pokročilé metody parafrazování

Kromě standardního frameworku kódovač-dekódovač existují různé pokročilé metody pro generování parafrází:

- Variational Autoencoder (VAE) – zachycuje bohaté reprezentace v latentním prostoru a umožňuje větší rozmanitost v generovaných parafrázích.
- Generativní adversariální sítě (GAN) – diskriminátor a generátor se učí v adversariálním procesu, což může zlepšit kvalitu generovaného textu.
- Posilované učení (RL) – pomáhá řešit problém exposure bias a lépe odpovídá na metriky hodnocení kvality.

## 4.4 Evaluační metriky

Pro hodnocení generování parafrází se běžně používají dva základní typy evaluačních metrik: automatické hodnocení a hodnocení lidmi.
Automatické hodnocení – Mezi nejběžněji používané metriky pro automatické hodnocení generování parafrází patří:

- BLEU (Papineni et al., 2002), původně vyvinutý pro hodnocení strojového překladu.

- METEOR (Denkowski a Lavie, 2014), který řeší slabiny BLEU v měření sémantických ekvivalentů a lépe koreluje s hodnocením lidí na úrovni vět a segmentů.
- ROUGE (Lin, 2004), metrika založená na recallu, původně vytvořená pro hodnocení sumarizace textu.
- TER (Snover), která měří počet úprav potřebných k tomu, aby lidský překladatel změnil překlad tak, aby přesně odpovídal referenčnímu překladu. Výsledky TER jsou v rozsahu 0-1 a obvykle se prezentují jako procenta, přičemž nižší hodnoty jsou lepší.

Hodnocení lidmi – Automatické metriky se zaměřují hlavně na překryvy n-gramů, což nezohledňuje význam. Proto je pro přesnější a kvalitativní hodnocení generovaného výstupu využíváno hodnocení lidmi. Lidé hodnotí parafráze podle různých kvalitativních dimenzí, jako je podobnost, srozumitelnost a plynulost. I když je hodnocení lidmi časově náročné a nákladné, poskytuje lepší a reprezentativnější obraz o kvalitě generovaného výstupu.

## 4.5 Vybrané modely pro řešení daného problému

**TODO dopsat vybrané modely jejich výhody a nevýhody**

### 4.5.1 Testování a výsledky modelů

**TODO dodat graf/y s výsledky testování modelů podle rychlosti a napsat který model bude použit a proč k řešení probmlému.**

# Převod textu na řeč — 5

Převod textu na řeč (Text-to-Speech, TTS) je technologie, která umožňuje automatické generování syntetického hlasu z psaného textu. Cílem moderních TTS systémů je dosáhnout přirozené intonace, správné výslovnosti a srozumitelnosti na úrovni lidské řeči. Současné přístupy k TTS zahrnují tradiční metody, jako je artikulační syntéza, formantová syntéza a konkatenativní syntéza, stejně jako moderní metody založené na hlubokém učení. Nejnovější pokroky zahrnují autoregresivní modely, jako je WaveNet a Tacotron, a neautoregresivní modely, jako je ParaNet a FastSpeech.

## 5.1 Metody syntézy řeči

**OTÁZKA: stačí takto lehce, nebo to víc rozepsat?** Metody převodu textu na řeč lze rozdělit do několika hlavních kategorií:

- Artikulační syntéza – simuluje pohyby řečového traktu a proudění vzduchu, což umožňuje věrnou syntézu zvuku, ale je výpočetně náročná.
- Formantová syntéza – generuje řeč na základě matematických modelů vokálních traktů a parametrů, jako jsou formantové frekvence.
- Konkatenativní syntéza – využívá databázi nahraných řečových jednotek, které se spojují dohromady k vytvoření plynulé řeči.
- Statistická parametrická syntéza – zahrnuje modely jako HMM (Hidden Markov Model), které umožňují flexibilnější manipulaci s hlasem.
- Neuronové sítě – moderní přístupy využívající hluboké učení k přímé syntéze řeči ze vstupního textu.

## 5.2 Hluboké učení v TTS

V posledním desetiletí se v TTS výrazně prosadily metody hlubokého učení. Nejvýznamnějšími architekturami jsou:

- Rekurentní neuronové sítě (RNN) – umožňují modelování časových závislostí v řeči, ale trpí problémem mizících gradientů.
- Konvoluční neuronové sítě (CNN) – používají se v moderních přístupech, jako je Tacotron 2, ke zpracování mel-spektrogramů.
- Transformery – modely jako FastSpeech využívají pozornostní mechanismy a paralelní zpracování pro efektivní syntézu řeči.

## 5.3 Evaluační metriky

Pro hodnocení kvality syntetické řeči se používají objektivní i subjektivní metriky:

### 5.3.1 Objektivní metriky

- Mel cepstral distortion (MCD) – měří rozdíl mezi syntetizovaným a referenčním zvukem na základě mel-frekvenčního cepstra.
- Root Mean Square Error (RMSE) log f0 – hodnotí rozdíl v intonaci mezi syntetickou a referenční řečí.
- Voiced/Unvoiced (VUV) error rate – hodnotí správnost identifikace znělých a neznělých částí řeči.

### 5.3.2 Subjektivní metriky

- Mean Opinion Score (MOS) – škála 1–5, kde 5 znamená přirozenou lidskou řeč.
- Preference test – uživatelé porovnávají syntetizovanou řeč s referenční řečí a hodnotí preference.

## 5.4 Vybrané modely

**OTÁZKA: Je tato kapitola provedena lépe tady v TTS nebo je líp provedena v STT?**

### 5.4.1 Modely OpenAI

OpenAI představila své modely TTS, které umožňují generovat vysoce kvalitní mluvený zvuk z textu. Nabízí dvě varianty:

- tts-1 – optimalizovaný pro aplikace v reálném čase.
- tts-1-hd – optimalizovaný pro vyšší kvalitu zvuku.

Tyto modely jsou dostupné přes API a podporují streamování audia. Cena začíná na $0.015 za 1 000 vstupních znaků. Maximální velikost vstupu na požadavek je 4 096 znaků, což odpovídá přibližně 5 minutám audia při výchozí rychlosti.

### 5.4.2 Modely ElevenLabs

ElevenLabs nabízí pokročilé modely TTS s důrazem na přirozenost a emocionální výraz. Mezi hlavní modely patří:

- Eleven Multilingual v2 – poskytuje nejpřirozenější zvuk s bohatým emocionálním projevem a podporuje 29 jazyků. Omezení je 10 000 znaků na požadavek.
- Eleven Flash v2.5 – optimalizovaný pro ultra nízkou latenci ( 75 ms) a podporuje 32 jazyků. Omezení je 40 000 znaků na požadavek.

**TODO dodat informace o modelu z kybernetiky**

### 5.4.3 Testování a výsledky modelů

**TODO dodat graf/y s výsledky testování modelů podle rychlosti a přesnosti a napsat který model bude použit k řešení probmlému.**

# Zkratky 6

| | |
|---|---|
| TTS | text-to-speech |
| RNN | rekuretní neuronová síť |
| CNN | konvoluční neuronová sít |
| STT | speech-to-text |
| TPR | true positive rate |
| TNR | true negative rate |
| NLP | natural language processing |

Tabulka 6.1: Tabulka zkratek

# Bibliografie


[Fox23] FOX, Josh. Introducing Nova-2: The Fastest, Most Accurate Speech-to-Text API. 2023. Dostupné také z: https://deepgram.com/learn/nova-2-speech-to-text-api?utm_source=chatgpt.com.

[KS23] KAUR, Navdeep; SINGH, Parminder. Conventional and contemporary approaches used in text to speech synthesis: a review. *Artificial Intelligence Review*. 2023, roč. 56. Dostupné také z: https://link.springer.com/article/10.1007/s10462-022-10315-0.

[KHH24] KHEDDAR, Hamza; HEMIS, Mustapha; HIMEUR, Yassine. Automatic speech recognition using advanced deep learning approaches: A survey. *Information Fusion*. 2024, roč. 109. Dostupné také z: https://www.sciencedirect.com/science/article/pii/S1566253524002008.

[ZB21] ZHOU, Jianing; BHAT, Suma. Paraphrase Generation: A Survey of the State of the Art. In: 2021. Dostupné také z: https://aclanthology.org/2021.emnlp-main.414/.