VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

ANALÝZA ENTIT V PSYCHOTERAPEUTICKÝCH SEZENÍCH

HIGH LEVEL ANALYSIS OF THE PSYCHOTHERAPY SESSIONS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE **Bc. ALEXANDER POLOK**
AUTHOR

VEDOUCÍ PRÁCE **Ing. PAVEL MATĚJKA, Ph.D.**
SUPERVISOR

BRNO 2023

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Zadání diplomové práce

144979

Ústav:	Ústav počítačové grafiky a multimédií (UPGM)
Student:	**Polok Alexander, Bc.**
Program:	Informační technologie a umělá inteligence
Specializace:	Strojové učení
Název:	**Analýza entit v psychoterapeutických sezeních**
Kategorie:	Zpracování řeči a přirozeného jazyka
Akademický rok:	2022/23

Zadání:

Cílem práce je analýza

1. Prostudujte statistické techniky pro modelování řeči, přirozeného jazyka a techniky analýzy psychoterapeutického sezení mezi terapeutem a klientem.
2. Navrhněte a implementujte nejméně 3 příznaky modelující průběh sezení, které ho umožní terapeutovi analyzovat a vyvodit zpětnou vazbu (např. sentiment, jazyková podobnost klienta s terapeutem, detekce výplňových slov, sumarizace)
3. Nalezněte vhodnou grafickou prezentaci získaných příznaků - např. vývoj v čase, celkový počet za nahrávku, histogram atd. a zobrazte pro daný čas nebo jako průměr přes jedno či více sezení.
4. Integrujte alespoň 2 z těchto příznaků do produkční verze webu DeePsy.cz .

Literatura:

- HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, Abdelrahman Mohamed
- Czert -- Czech BERT-like Model for Language Representation Jakub Sido, Ondřej Pražák, Pavel Přibáň, Jan Pašek, Michal Seják, Miloslav Konopík

Při obhajobě semestrální části projektu je požadováno:
1. bod hotový a 2. rozpracovaný.

Podrobné závazné pokyny pro vypracování práce viz https://www.fit.vut.cz/study/theses/

Vedoucí práce:	**Matějka Pavel, Ing., Ph.D.**
Vedoucí ústavu:	Černocký Jan, prof. Dr. Ing.
Datum zadání:	1.11.2022
Termín pro odevzdání:	17.5.2023
Datum schválení:	31.10.2022

Abstrakt

Tato práce se zabývá analýzou psychoterapeutických sezení v rámci výzkumného projektu DeePsy. Jejím cílem je navrhnout a vytvořit sadu příznaků modelujících průběh sezení, jež mohou odhalit na první pohled nepatrné nuance. Zmíněné příznaky jsou automaticky extrahovány ze zdrojové nahrávky s využitím hlubokých neuronových sítí. Příznaky jsou zpracovány, porovnány napříč sezeními a graficky zobrazeny, čímž vzniká dokument plnící roli zpětné vazby o sezení pro terapeuta. Tato zpětná vazba může posloužit k profesnímu růstu a kvalitnější psychoterapii v budoucnu. Bylo dosaženo relativního zlepšení detekce řečové aktivity o 37,82 %. Byl zobecněn diarizační systém VBx ke konvergenci ke dvěma mluvčím s minimálním relativním zhoršením chybovosti o 0,66 %. Byl natrénován systém pro automatické rozpoznávání řeči, jehož chybovost je o 17,06 % relativně lepší než nejlepší dostupný hybridní model. Dále byly natrénovány systémy pro klasifikaci sentimentu, typu terapeutických intervencí a detekci překrývající se řeči.

Abstract

This work focuses on analyzing psychotherapy sessions within the DeePsy research project. This work aims to design and develop features that model the session dynamics, which can reveal seemingly subtle nuances. The mentioned features are automatically extracted from the source recording using neural networks. They are further processed, compared across sessions, and displayed graphically, creating a document that acts as a feedback document about the session for the therapist. Furthermore, this assistive tool can help therapists to professionally grow and to provide better psychotherapy in the future. A relative improvement in voice activity detection of 37.82% was achieved. The VBx diarization system was generalized to converge to two speakers with a minimum relative error rate degradation of 0.66%. An automatic speech recognition system has been trained with a 17.06% relative improvement over the best available hybrid model. Models for sentiment classification, type of therapeutic interventions, and overlapping speech detection were also trained.

Klíčová slova

strojové učení, analýza psychoterapeutických sezení, zpracování přirozeného jazyka, zpracování řeči, rozpoznávání řeči, detekce sentimentu, detekce klíčových slov, sumarizace, klasifikace terapeutických intervencí, překrývající se řeč, jazykové modely, transformery, neuronové sítě, wav2vec 2.0, whisper, hovorový jazyk, diarizace, učení s vlastním dozorem, kontrastivní učení

Keywords

machine learning, psychotherapy session analysis, natural language processing, speech processing, speech recognition, sentiment detection, keyword detection, summarization, therapeutic interventions classification, overlapping speech, language models, transformers, neural networks, wav2vec 2.0, whisper, colloquial language, diarization, self-supervised learning, contrastive learning

Citace

POLOK, Alexander. *Analýza entit v psychoterapeutických sezeních.* Brno, 2023. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Pavel Matějka, Ph.D.

Analýza entit v psychoterapeutických sezeních

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením pana Ing. Pavla Matějky, Ph.D. Uvedl jsem všechny literární prameny, publikace a další zdroje, ze kterých jsem čerpal.

.....
Alexander Polok
15. září 2023

Poděkování

Chtěl bych poděkovat Ing. Pavlu Matějkovi, Ph.D. za všechen čas, který byl ochoten věnovat konzultacím, za pomoc při řešení nejrůznějších problémů a za velmi přátelský a vstřícný přístup.

Zároveň bych chtěl poděkovat paní Mgr. Drahomíře Šloufové za korekturu gramatických a stylistických chyb. Poděkování také patří Tiboru Kubíkovi a Ladislavu Ondrisovi za cenné podněty a připomínky.

Tato práce proběhla se státní podporou Technologické agentury ČR v rámci Programu Éta (grant č. TL03000049). Výpočetní zdroje byly dodány v rámci projektu „e-Infrastruktura CZ“ (e-INFRA CZ LM2018140) podpořeného Ministerstvem školství, mládeže a tělovýchovy České republiky.

Obsah

1 Úvod **2**

2 Teorie **3**
2.1 Analýza psychoterapeutických sezení 3
2.2 Strojové učení 5
2.3 Zpracování řeči a přirozeného jazyka 6
2.4 Transformer 9
2.5 Jazykové modely 12
2.6 Řečové modely 15

3 Úlohy **22**
3.1 Detekce řečové aktivity 22
3.2 Diarizace 23
3.3 Detekce překrývající se řeči 24
3.4 Rozpoznávání řeči 25
3.5 Klasifikace sentimentu 28
3.6 Klasifikace typů terapeutických intervencí 28

4 Data **30**
4.1 Řečové korpusy 30
4.2 Textové korpusy 32

5 Experimenty **37**
5.1 Detekce řečové aktivity 37
5.2 Diarizace 38
5.3 Detekce překrývající se řeči 41
5.4 Rozpoznávání řeči 42
5.5 Klasifikace sentimentu 53
5.6 Klasifikace typů terapeutických intervencí 56
5.7 Souhrn 61

6 Extrahované příznaky **64**

7 Závěr **69**

Literatura **70**

A Typy terapeutických intervencí **83**

B Plakát prezentovaný v rámci konference Excel@FIT 2023 **86**

Kapitola 1

Úvod

V oblasti zpracování řeči a přirozeného jazyka v posledních letech došlo k enormnímu pokroku poháněného snahou o co největší otevřenost a možnost replikace zjištěných poznatků mezi vědci. Nejnovější systémy dosahují lidských přesností nebo je dokonce překonávají v nejrůznějších úlohách těchto odvětví. Díky zmíněné prosperitě dochází k aplikaci těchto systémů i v doposud nemyslitelných odvětvích. Jedním z nich je na příklad psychoterapie, původem latinské slovo – skládající se z části psyché (výrazu pro lidskou duši či mysl) a therapeia (léčba). Cílem tohoto oboru je léčit mysl, obnovit její rovnováhu a vést ke zkvalitnění života klienta.

V rámci mé diplomové práce jsou zkoumány možnosti aplikace metod strojového učení v oblasti psychoterapie za účelem získání příznaků modelujících průběh terapeutického sezení. Tato práce volně navazuje na mou bakalářskou práci „Analýza audio hovoru mezi dvěma účastníky“ obhájenou dne 16.6.2021 na FIT VUT v Brně. Má za cíl zaměřit se na entity vyššího řádu, jako nálada v rámci sezení, či měnící se distribuce témat.

V rámci kapitoly 2 je blíže představen kontext této práce, projekty zabývající se obdobnou tématikou a projekt DeePsy, v jehož rámci tato práce vznikla. Postupně je představen technický aparát využitý v nadcházejících kapitolách. Dále jsou v kapitole 3 exaktně definovány úlohy a z nich vycházející příznaky, které jsou v rámci této práce zkoumány. Na konci třetí kapitoly jsou diskutovány aktuální přístupy a vhodné metriky využité pro hodnocení kvality kandidátních řešení. V rámci kapitoly 4 jsou představena data, na nichž jsou příslušná kandidátní řešení trénována a následně ohodnocena. Je podrobně diskutován původ, příslušná specifika a možná předpojatost vycházející z těchto dat. Provedené experimenty jsou sepsány v kapitole 5. Je zde kladen důraz na interpretaci dosažených výsledků a jsou popsány postupy pro replikaci provedených experimentů. Získané příznaky a jejich integrace do systému DeePsy jsou popsány v rámci kapitoly 6. V kapitole 7 jsou sumarizovány výsledky této práce.

Kapitola 2

Teorie

V rámci této kapitoly je čtenáři nejdříve představen kontext práce a výzkumný projekt DeePsy, v jehož rámci tato práce mohla vzniknout. Dále je mu postupně představen teoretický aparát využitý v následujících kapitolách.

2.1 Analýza psychoterapeutických sezení

Masivní pokrok v oblasti strojového učení a neustále zpřesňování metod zpracování řeči a zpracování přirozeného jazyka, viz sekce 2.2, vede k začlenění těchto technik do širšího spektra oblastí. Psychoterapie není výjimkou. Nespočet studií se zabývá detekcí duševních poruch (např. bipolárních poruch nebo schizofrenie) s využitím zvukových záznamů, neurokognitivních dat či biomarkerů [140, 104]. Například autoři ve [46] se pokusili hledat přímou asociaci mezi hodnocením sezení a jeho jazykovým obsahem. Ačkoliv jsou tyto důkazní studie velmi povzbudivé a zajímavé, zatím se nepodařilo promítnout tato zlepšení do klinické praxe a slouží spíše jako pomocné nástroje pro psychoterapeuty [34]. Je nutno podotknout, že získávání a zpracování dat z odvětví psychoterapie je velmi problematické, sběr dat je časově náročný a především jakékoliv nakládání s daty v ČR je podmíněno Etickým kodexem ČAP[1] (České asociace pro psychoterapii). Samotná data jsou velmi citlivá, jelikož klienti velmi často rozebírají osobní prožitky a je nutno s takovými daty nakládat velmi obezřetně.

2.1.1 Komerční nástroje

V komerční sféře se již objevují první firmy poskytující automatické zpracování psychoterapeutických sezení. Za zmínění stojí především platforma Lyssn[2] nebo firma Upheal[3]. Obě řešení dovolují klinickým lékařům hodnotit a reflektovat svou praxi, přezkoumávat rozhovory, spolupracovat s kolegy, bezpečně shromažďovat data. Obsahují sadu nástrojů pro tvorbu automatických přepisů, souhrnů a rozpoznání entit vystupujících v rozhovorech v anglickém jazyce.

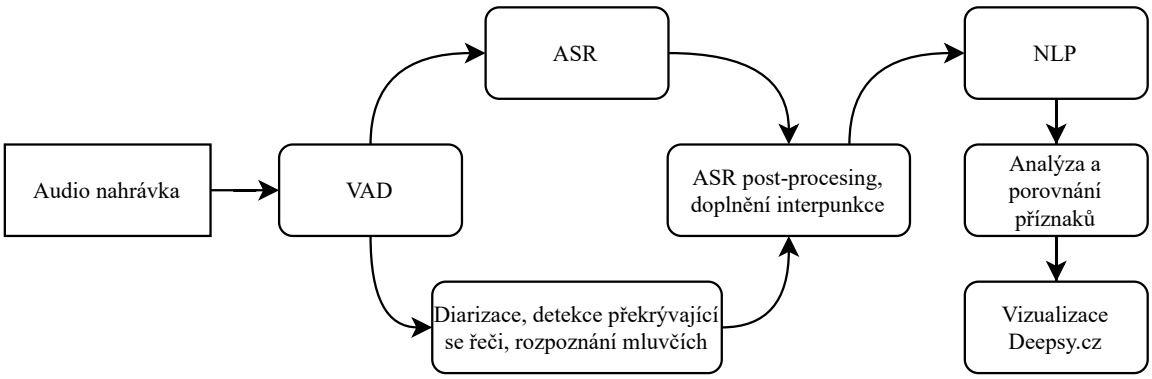
[1] https://czap.cz/Eticky-kodex
[2] https://www.lyssn.io/
[3] https://upheal.io/

2.1.2 DeePsy

Tato diplomová práce je součástí projektu DeePsy – *Deep Learning in Psychotherapy*. Název vznikl z pojmů „deep learning“ (hluboké učení) a psychotherapie. Jedná se o projekt se státní podporou Technologické agentury ČR v rámci Programu Éta[4]. Je veden týmem psychologů a psychoterapeutů z Masarykovy univerzity[5] a odborníků na informační technologie z Vysokého učení technického v Brně[6]. Je zaměřen na analýzu psychoterapeutických sezení v českém jazyce s využitím strojového učení. Na projektu spolupracují Psychosomatická klinika[7] a Terapeutický přístav[8], účast jim umožňuje zlepšit kvalitu svých psychoterapeutických služeb. Všichni účastníci projektu dobrovolně poskytují zpětnou vazbu o sezeních a souhlasí s jejich nahráváním.

Jedním z výstupů projektu je webová aplikace[9] usnadňující systematické získávání dat a zprostředkování zpětné vazby z proběhlých sezení v podobě dotazníků a automaticky vytvořených souhrnných zpráv. DeePsy je založeno na myšlence oboustranného propojování praxe a výzkumu v psychoterapii. V praktické rovině poskytuje psychoterapeutům nástroj pro systematické získávání zpětné vazby a zkvalitňování jejich práce. Tato data však zároveň mají významný potenciál pro výzkum, který umožní lépe porozumět zákonitostem psychoterapeutického procesu. Takto získané výzkumné poznatky spolu se zkušenostmi a podněty uživatelů pak opět slouží dalšímu vývoji a zlepšování aplikace DeePsy.

Schematický graf zpracování nahrávky v rámci systému je zobrazen na obr. 2.1. Nejdříve dochází k detekci řečové aktivity (*Voice Activity Detection* – VAD), nad aktivními úseky je spuštěno automatické rozpoznání řeči (*Automatic Speech Recognition* – ASR) a je provedena diarizace. Výstupy těchto systémů jsou zkombinovány a je doplněna interpunkce. Tato textová data jsou následně zpracována v sekci zpracování přirozeného jazyka (*Natural Language Processing* – NLP), je provedena analýza řečových a textových příznaků v rámci jednoho sezení, ale i mezi sezeními. Data jsou vizualizovaná na webu DeePsy.



Obrázek 2.1: Schematický graf zpracování nahrávky sezení v systému DeePsy.

[4] https://starfos.tacr.cz/en/project/TL03000049
[5] https://www.muni.cz/
[6] https://www.vut.cz/
[7] https://psychosomatika.cz/
[8] https://terapeutickypristav.cz/
[9] https://www.deepsy.cz/

2.2 Strojové učení

Strojové učení (*Machine Learning* – ML) je podobor umělé inteligence (*Artifical Inteligence* – AI) věnující se studiu, pochopení a tvorbě metod, které se dokáží učit. O počítačovém programu lze prohlásit, že *se učí* ze zkušeností E s ohledem na určitou třídu úloh T s mírou výkonnosti P, jestliže se jeho výkonnost v úlohách T, měřená pomocí P, zlepšuje ze zkušeností E [92].

Metody strojového učení jsou navrženy tak, aby byly schopné z množiny vzorových dat, dále trénovací množiny T_{train}, extrahovat informace, vztahy, které jim následně pomohou provádět předpovědi nebo rozhodnutí, aniž by k tomu byly explicitně naprogramovány.

Schopnost učit se je nejčastěji validována na testovací množině T_{test}. Systém je správně naučený tehdy, když se shodnou úspěšností vyhodnocuje trénovací i testovací množinu. O takovém systému můžeme prohlásit, že správně *generalizuje*. Pokud má vyhodnocení trénovací množiny výrazně vyšší úspěšnost, je systém *přeučený* (*overfitted*). Častým požadavkem je, aby množiny T_{train} a T_{test} byly disjunktní. Aby bylo možné ověřit průběh učení, je nejčastěji využita i třetí množina – validační množina T_{val}. Na této množině je v průběhu učení opětovně vyhodnocována výkonnost P. A v momentě, kdy tato výkonnost začne dlouhodobě klesat, je učení zastaveno, aby nedošlo k přeučení.

2.2.1 Základní klasifikace metod strojového učení

Algoritmy strojového učení lze rozdělit do následujících kategorií podle způsobu, jakým se učí. Některé publikace [13] uvažují jen 3 hlavní typy strojového učení, pro úplnost je zde však doplněna i čtvrtá kategorie.

Učení s učitelem – Supervised Learning

Množina trénovacích dat se skládá ze vstupů a odpovídajících výstupů. Příkladem učení s učitelem je *klasifikace* – přiřazení výstupní hodnoty (kategorie) z diskrétní množiny prvků. Dalším příkladem může být *regrese* – přiřažení výstupní hodnoty ze spojité veličiny [13].

Učení bez učitele – Unsupervised Learning

Množina trénovacích dat se skládá pouze ze vstupů bez odpovídajících výstupů. Cílem je odhalit ve vstupních datech skupiny vstupů, které mají obdobnou strukturu – vlastnosti a přiřadit jim identitu této skupiny. Takovýto přístup je označován jako *shlukování* (*clustering*). Dalším příkladem je odhadování distribuce (*density estimation*), ze které data mohou pocházet. Redukce dimensionality může být taktéž příkladem učení bez učitele [13].

Kombinace učení s učitelem a bez učitele – Semi-Supervised Learning

Je přístup kombinující malé množství označených (*labeled*) dat s velkým množstvím neoznačených dat. Tímto způsobem mohou neoznačená data výrazně zlepšit přesnost učení bez pracné anotace [149].

Posilované učení – Reinforcement Learning

Je technika zabývající se problémem nalezení vhodných akcí, které v dané situaci maximalizují odměnu v budoucnu. Algoritmus nemá k dispozici očekávané výstupy, místo toho je objevuje procesem pokusů a omylů, při němž obdrží odměnu nebo trest [13].

2.2.2 Další techniky strojového učení

V případě, že cílová doména úlohy neobsahuje dostatečný počet trénovacích vzorů, jsou často využívány následující techniky.

Přenesené učení – Transfer Learning

Je technika mající za cíl zlepšení výkonnosti cílového modelu v cílové doméně s využitím přenosu znalostí získaných jinými modely v jiných, ale souvisejících doménách. Tímto způsobem lze snížit závislost cílového modelu na velké trénovací sadě v cílové doméně [150].

Učení se pod vlastním dohledem – Self-Supervised Learning

Je paradigma, spadající do kategorie učení s učitelem. Liší se však v tom, že algoritmy nemají k dispozici žádné očekávané výstupy. Ty si však mohou automaticky vygenerovat ze vstupních dat. Samotná data tedy jistým způsobem poskytují supervizi. Algoritmy jsou nejčastěji v prvním kroce učeny k tomu, aby dokázaly odhalit strukturu a relace ve zdrojových datech. Cílovou úlohou může být předpovědět zakrytou část vstupu z jiných částí totožného vstupu. Získané vědomosti mohou být následně přeneseny a model může být dotrénován na jiné cílové úloze [6].

Kontrastivní učení pod vlastním dohledem – Contrastive Self-Supervised Learning (CSL)

V technice zvané kontrastivní učení pod vlastním dohledem je cílem tvorba interní reprezentace vstupu takovým způsobem, že podobné vstupy jsou v N-dimenzionálním prostoru blízko sebe a jsou zároveň daleko od odlišných vstupů. CSL vychází z předpokladu, že všechny instance v trénovací sadě $T_{train} = \{x_1, x_2, \ldots, x_N\}$ pocházejí z jiné třídy. Úlohou je prohlásit, zda dvojice příznaků (f_1, f_2) pochází ze stejné instance $x \in T_{train}$, či nikoliv [37].

Jedním ze způsobů učení je postupné přikládání modelu vstupní dvojice (f_1, f_2) spolu s automaticky vygenerovanou výstupní hodnotou y danou vztahem

$$y = \begin{cases} 1 & f_1 \in x \ \wedge f_2 \in x \\ -1 & f_1 \in x \ \wedge f_2 \in x'; x \neq x'. \end{cases} \tag{2.1}$$

2.3 Zpracování řeči a přirozeného jazyka

Cílem této práce je zpracování textu a řeči, což jsou sekvenční modality. Je vhodné je zpracovávat sekvenčním způsobem tak, aby nedošlo k ztrátám informací. V této sekci jsou představeny klíčové koncepty a architektury modelů použitých v kapitole 5.

2.3.1 Neuronové sítě pro zpracování sekvencí

V oblasti zpracování přirozeného jazyka či řeči je velmi častou úlohou zpracování sekvence elementů (tokenů, slov, zvukových jednotek), v níž existuje závislost mezi elementy. Tento fakt vede k tomu, že izolované zpracování není efektivní a dochází ke ztrátě globálních informací. Častým požadavkem je taktéž variabilní délka vstupu, což vede k tomu, že tradiční neuronové sítě (*Neural Networks* – NNs [13]) jsou pro tuto úlohu nevhodné.

Konvoluční sítě (C*onvolutional Neural Networks* – CNNs [39, 77, 70]) fungující na principu sdružení lokálních informací do vyšších logických úrovní fungují dobře pro klasifikaci,

kdy je sdružena celková globální informace, jež vede k výslednému rozhodnutí. CNNs však nejsou příliš vhodné pro ostatní typy úloh, jelikož postrádají schopnost uchování vztahů mezi vzdálenějšími elementy v sekvenci.

Oproti tomu rekurentní neuronové sítě (*Recurrent Neural Networks* – RNNs [118, 55]) zpracovávají vstup sekvenčně tak, že skrytý stav (*hidden state*) h_t v čase t je vyjádřen pomocí nelineární funkce f elementu vstupní sekvence x_t a předchozího skrytého stavu h_{t-1}

$$h_t = f(h_{t-1}, x_t). \tag{2.2}$$

Samotný princip fungování RNN je však pro zpracování dlouhých sekvencí nedostatečný, jelikož při zpracování delších sekvencí dochází k mizení gradientu (*vanishing gradient*). To vedlo k využití *Long Short-Term Memory* (LSTM) [54], či *Gated Recurrent Unit* (GRU) sítí [27]. Dalším krokem, který vedl ke zlepšení úspěšnosti, je zpracování vstupní sekvence z obou směrů, jelikož dopředný průchod předpokládá, že následující elementy závisí pouze na těch předchozích [124].

2.3.2 Z sekvence do sekvence – Sequence to Sequence (Seq2seq)

Speciálním případem zpracování sekvence je případ, kdy výstupem je taktéž sekvence. Příkladem mohou být např. strojový překlad, sumarizace textu nebo rozpoznání řeči. Za předpokladu, že vstupní sekvence $\mathbf{x} = (x_1, \ldots, x_T)$, jakož i výstupní sekvence $\mathbf{y} = (y_1, \ldots, y_{T'})$ je variabilní (odlišné) délky $T \neq T'$, je nemožné přímé mapování vstupu na výstup. Z tohoto důvodu byla navržena architektura enkodér-dekodér [132, 25]. Původní varianta obsahovala vícevrstvou RNN (LSTM) – enkodér, která mapuje vstupní sekvenci s využitím nelineární funkce q do vektoru fixní velikosti c podle vztahu

$$\mathbf{c} \ = q(\mathbf{h}_{1:T}) \tag{2.3}$$

a multivrstvé RNN (LSTM) pro dekódování podle následující faktorizace

$$p(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T} p(y_t|\mathbf{c}, \mathbf{y}_{1:t-1}). \tag{2.4}$$

Uvážíme-li jednoduchou RNN, podmíněnou pravděpodobnost $p(y_t|\mathbf{c}, \mathbf{y}_{1:t-1})$ lze vyjádřit jako

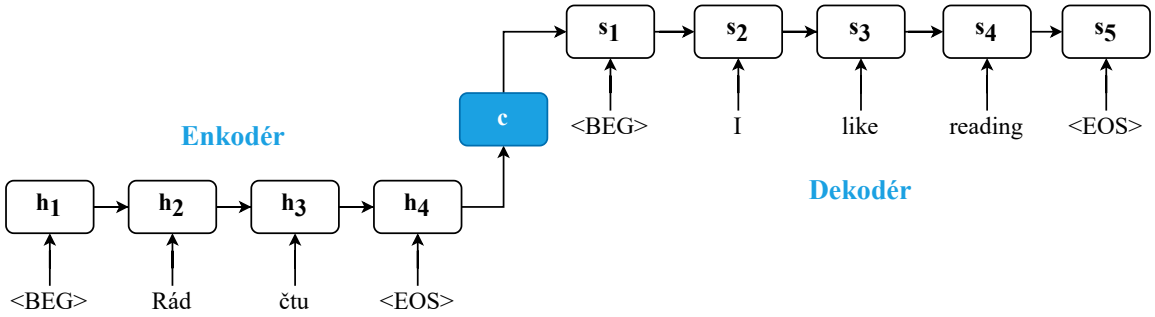
$$p(y_t|\mathbf{c}, \mathbf{y}_{1:t-1}) = g(y_{t-1}, s_t, \mathbf{c}), \tag{2.5}$$

kde g je nelineární (vícevrstvá) funkce a s_t je skrytý stav dekodéru v čase t.

Sekvence jsou nejčastěji zakončeny speciálním tokenem <EOS>, který taktéž indikuje konec generování. Vysokoúrovňový pohled na tuto architekturu je zobrazen na obr. 2.2.

2.3.3 Attention mechanizmus

Z předchozí sekce je patrné, že takto navržená architektura enkodér-dekodér obsahuje úzké hrdlo při abstrakci celé sekvence do vektoru $\mathbf{c}$. Jako řešení se jeví předat do dekodéru sekvenci všech skrytých stavů $\mathbf{h} = (h_1, \ldots, h_T)$ místo její zkomprimované podoby, což však vede k vysoké výpočetní náročnosti, v předchozích pracích bylo proto nejčastěji voleno $\mathbf{c} = h_T$. V roce 2014 Dzmitry Bahdanau navrhl elegantní způsob, jak využít celou sekvenci, výpočet částečně zjednodušit a zaměřit se na relevantní části vstupu – *attention* [7].



Obrázek 2.2: Architektura Seq2seq skládající se z enkóder a dekóder bloku použita pro překlad sekvence „Rád čtu“ z češtiny do angličtiny. Cílem enkodéru je postupně agregovat vstupní tokeny a vytvořit kontextový vektor **c**. Dekodér následně autoregresivně generuje tokeny, dokud nevygeneruje token reprezentující konec sekvence <EOS> nebo je generace zastavena.

S využitím bidireakcionální RNN (BiRNN) jako enkodéru podmíněnou pravděpodobnost $p(y_i|\mathbf{y}_{1:i-1}, \mathbf{x})$ vyjádřil jako

$$p(y_i|\mathbf{x}, \mathbf{y}_{1:i-1}) = g(y_{i-1}, \mathbf{s}_i, \mathbf{c}_i), \tag{2.6}$$

kde s_i je skrytý stav dekodéru v kroce i spočtený jako

$$\mathbf{s}_i = f(\mathbf{s}_{i-1}, y_{i-1}, \mathbf{c}_i). \tag{2.7}$$

Je zde viditelné, že na rozdíl od předchozího případu je skrytý stav $\mathbf{s}_i$ závislý na kontextuálním vektoru $\mathbf{c}_i$ pro příslušný krok i.

Kontextuální vektor $\mathbf{c}_i$ je definován jako vážená suma skrytých stavů enkodéru $\mathbf{h}$

$$\mathbf{c}_i = \sum_{j=1}^{T} \alpha_{ij}\mathbf{h}_j, \tag{2.8}$$

kde váha α_{ij}, definující významnost skrytého stavu enkodéru $\mathbf{h}_i$ vzhledem k předchozímu skrytému stavu dekodéru $\mathbf{s}_{i-1}$ při generování výstupu y_i a následujícího skrytého stavu $\mathbf{s}_i$, je vyjádřena jako

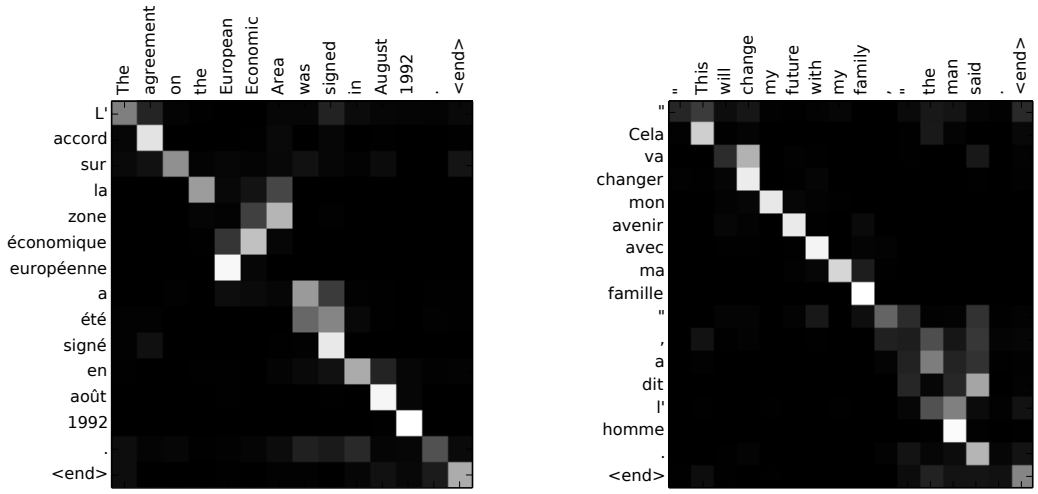
$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T} \exp(e_{ik})}, \tag{2.9}$$

kde $e_{ij} = a(\mathbf{s}_{i-1}, \mathbf{h}_j)$ je model zarovnání (*alignment model*) definující, jak je vstupní sekvence na pozici i zarovnána s výstupní sekvencí na pozici j. Vizuální reprezentace vah α_{ij} je zobrazena na obr. 2.3. V původním článku byl jako model zarovnání uvážen jednovrstvý perceptron [117, 88] tak, že

$$a(\mathbf{s}_{i-1}, \mathbf{h}_j) = \boldsymbol{\nu}_a^\intercal \tanh(\mathbf{W}_a\mathbf{s}_{i-1} + \mathbf{U}_a\mathbf{h}_j), \tag{2.10}$$

kde $\mathbf{W}_a \in \mathbb{R}^{n\times n}, \mathbf{U}_a \in \mathbb{R}^{n\times 2n}$ a $\boldsymbol{\nu}_a \in \mathbb{R}^n$ jsou matice váh.

Myšlenku *attention* v roce 2015 rozšířil Minh-Thang Luong a zavedl další možné varianty, jak definovat model zarovnání, jakož i pojem lokální *attention*, kdy je zarovnání



Obrázek 2.3: Vizuální reprezentace vah modelu zarovnání při úloze překladu sekvencí z francouzštiny do angličtiny, převzato z [7].

spočteno jen v rámci okna o fixní délce D [87]. Dalším rozdílem je, že Luong v rámci zarovnání uvažuje aktuální skrytý stav s_i a tedy definuje $e_{ij} = a(s_i, h_j)$.

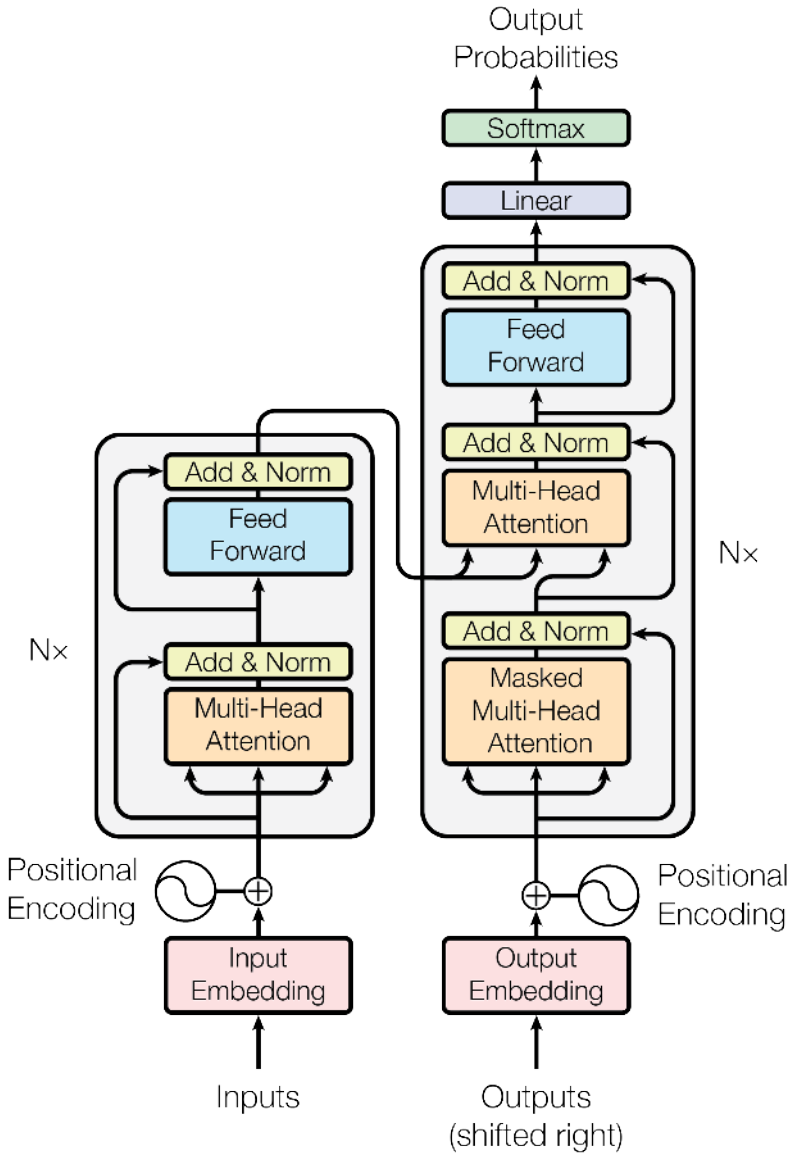
Attention mechanizmus v doposud představené podobě byl spolu s RNN využit v nespočtu prací. Za zmínění stojí například *Embeddings from Language Model* (ELMO) [103], které představilo nový způsob tvorby reprezentace slov, jejich *embeddingů*, uvažující jejich aktuální kontext na rozdíl od metod Word2Vec [90] a GloVe [102].

2.4 Transformer

Architektura Transformer [135] byla představena v roce 2017 jako alternativa k RNN, které do té doby patřily spolu s konvolučními neuronovými sítěmi mezi hlavní proudy výzkumu v oblasti zpracování přirozeného jazyka, řeči nebo obrazu. Jak již bylo zmíněno výše, zpracování sekvencí pomocí RNN je sekvenční úloha, která je zejména při trénování zpětným šířením chyby (backpropagation) na delších sekvencích paměťově velmi náročná – je nutno uchovat gradient přes celou sekvenci. Autoři článku „Attention Is All You Need“ [135] představili nový výpočetní *framework* dovolující masivní paralelizaci založený čistě na myšlence *attention* definované v sekci 2.3.3. Architektura tohoto *frameworku* je zobrazena na obr. 2.4.

Obdobně jak architektura Seq2seq se Transformer sestává ze dvou částí – enkodér a dekodér. Obě části obsahují vrstvu pro zakódování vstupu viz sekce 2.4.4 a N propojených bloků skládajících se z *self-attention*[10] a plně propojené neuronové sítě o 2 vsrtvách. Bloky dále obsahují reziduální spoje (*residual connections* [50]) a normalizaci po vrstvách (*layer normalization* [3]). Bloky dekodéru navíc obsahují další *attention* vrstvu, kde je přiveden výstup enkodéru. Na výstupu celého dekodéru se nachází plně propojena vrstva a softmax vrstva tak, aby výstupem byla distribuce mezi tokeny modelu.

[10] *Self-attention* je aplikace *attention* mechanizmu sekvence na sebe samu za účelem tvorby vlastní reprezentace [24].



Obrázek 2.4: Architektura modelu Transformer, převzato z [135].

2.4.1 Scaled Dot-Product Attention

Attention mechanizmus je v případě této architektury definován jako

$$\begin{aligned} \mathit{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) &= \mathrm{softmax}(\frac{\mathbf{Q}\mathbf{K}^\intercal}{\sqrt{d_k}})\mathbf{V}, \\ \mathbf{Q} &= \mathbf{W}_Q\mathbf{X}, \\ \mathbf{K} &= \mathbf{W}_K\mathbf{X}, \\ \mathbf{V} &= \mathbf{W}_V\mathbf{X}, \end{aligned} \tag{2.11}$$

kde $\mathbf{Q},\mathbf{K},\mathbf{V}$ jsou matice *Query*,*Key* a *Value*, $\mathbf{W}_Q \in \mathbb{R}^{d_{model}\times d_q}$, $\mathbf{W}_K \in \mathbb{R}^{d_{model}\times d_k}$ a $\mathbf{W}_V \in \mathbb{R}^{d_{model}\times d_v}$ jsou matice projekčních váh, $\mathbf{X} \in \mathbb{R}^{n\times d_{model}}$ je matice *embedingů*, n je délka vstupní sekvence, d_{model} je dimenzionalita *embedingů* a d_q, d_k, d_v jsou dimenzionality vektorů z matic *Query*,*Key* a *Value*. Použitím normalizační konstanty $d_k^{-\frac{1}{2}}$ autoři limitují efekt

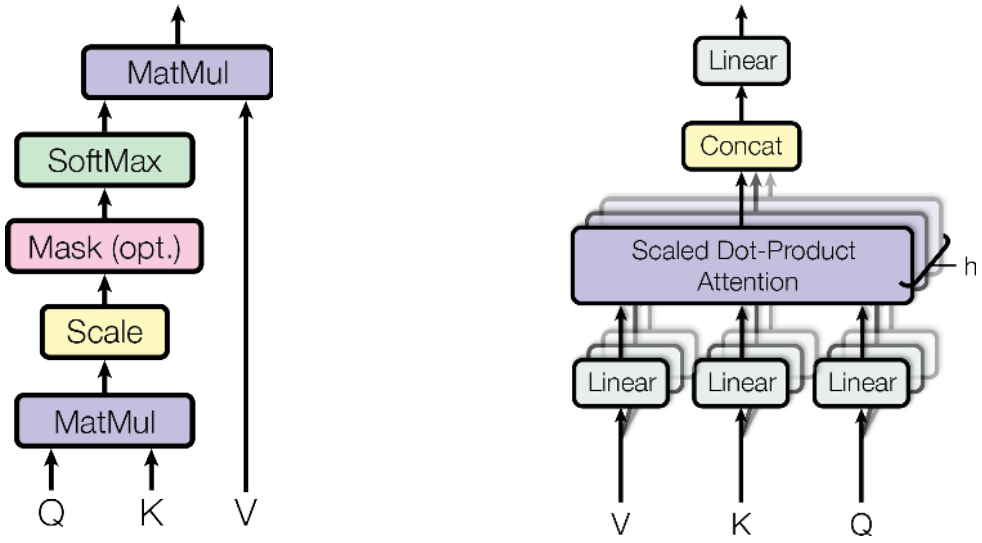
ztrácejícího se gradientu u funkce softmax pro dlouhé sekvence (*vanishing gradient* [100]). Výpočet je schematicky zobrazen na obr. 2.5a.

2.4.2 Multihead Attention

Autoři dále provedli analýzu a zjistili, že je výhodnější modelu povolit projekci do vícero prostorů $(\mathbf{Q}_1, \mathbf{K}_1, \mathbf{V}_1), \ldots, (\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h)$, a redefinovali použitý *attention* mechanizmus jako

$$MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [Attention(\mathbf{Q}_1, \mathbf{K}_1, \mathbf{V}_1), \ldots, Attention(\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h)]\mathbf{W}_O, \quad (2.12)$$

kde $\mathbf{W}_O \in \mathbb{R}^{hd_v \times d_{model}}$ je projekční matice do původního rozlišení a operátor $[\mathbf{A}, \mathbf{B}]$ označuje konkatenaci matic $\mathbf{A}$ a $\mathbf{B}$.



(a) *Scaled Dot-Product Attention* schéma, převzato z [135].

(b) *Multihead Attention schéma*, převzato z [135].

Obrázek 2.5: Schéma navržených variant *attention* mechanizmů v rámci práce „Attention Is All You Need“[135].

2.4.3 Poziční kódování – *positional encoding*

Jelikož takto navržený model neobsahuje žádný mechanizmus explicitně uvažující relativní či absolutní pořadí v sekvenci, autoři [135] přišli s myšlenkou dovolující využití této stěžejní informace, a tedy s myšlenkou pozičních embeddingů [42], které jsou přičteny k původním embeddingům

$$\mathbf{X} = \mathbf{X} + \mathbf{PE}, \quad (2.13)$$

kde $\mathbf{X} \in \mathbb{R}^{n \times d_{model}}$ je matice vstupní sekvence a $\mathbf{PE} \in \mathbb{R}^{n \times d_{model}}$ matice pozičních embeddingů, původně definovaná s využitím vlastností goniometrických funkcí sinus a cosinus jako

$$PE(i,j) = \begin{cases} sin(\frac{i}{10000^{2j/d_{model}}}) & j \mod 2 == 0 \\ cos(\frac{i}{10000^{2j/d_{model}}}) & jinak. \end{cases} \quad (2.14)$$

V rámci práce [135] se taktéž autoři pokoušeli model naučit tyto poziční embeddingy na přímo, avšak nevedlo to k žádnému významnému zlepšení[11].

2.4.4 Kódování vstupu

Nedílnou součástí moderního zpracování textového vstupu je jeho převedení do číselné podoby. Tato sekce vychází z [58]. Prvním krokem je rozdělení textu do sekvence tokenů (slov nebo jiných atomických prvků). Tento proces, nazývaný tokenizací, lze provádět na úrovni:

- celých slov – textový vstup může být rozdělen několika způsoby, např. podle mezer, bílých znaků nebo jakýchkoliv interpunkčních znamének, tento přístup však vede k obrovským slovníkům, byl využit v rámci již zmíněných algoritmů GloVe a Word2Vec [90, 102],
- znaků – slovník je mnohem menší a zároveň podíl neznámých slov na neviděných datech (*out-of-vocabulary*) je mnohem menší, nevýhodou je, že samotný znak nese mnohém méně informací než slovo a učení relací mezi vstupy a výstupy nemusí být efektivní, výhodné může být v jazycích jako čínština [101],
- nebo na úrovní podslov – tento přístup je založený na myšlence, že často používaná slova by neměla být rozdělena, naopak vzácná slova mohou být rozdělena – tedy například složenina „automobil“ může být rozdělena na tokeny „auto“ a „mobil“, na této myšlence jsou založeny algoritmy jako *Byte Pair Encoding* (BPE) [40], *WordPiece* [141] nebo *Unigram* [71].

Dalším krokem kódování je převedení tokenů do jejich číselné podoby. Lze rozlišit přístupy přímé (kódování 1 z n – *one hot encoding*), statické (vycházející z matice kookurence – *co-occurrence matrix*) a prediktivní – založené na neuronových sítích, trénovaných na (maskovaném) modelování jazyka.

2.5 Jazykové modely

Jazykový model je pravděpodobnostní distribuce nad sekvencí slov daného jazyka (nebo kombinace několika jazyků – vícejazyčný jazykový model – *multilingual language model*). Cílem jazykových modelů je co nejlépe aproximovat pravděpodobnost sekvence

$$P(\mathbf{x}_{1:N}) = \prod_{i=1}^{N} P(x_i|\mathbf{x}_{1:i-1}), \tag{2.15}$$

a tedy určit, jak pravděpodobné je, že zadaná sekvence pochází z daného jazyka [62]. Je patrné, že tedy jazykové modely mohou být taktéž užívany pro generování sekvencí z daného jazyka. V oblasti jazykového modelování lze rozlišit statistické jazykové modely (*Language Models* – LMs) a ty vycházející z neuronových sítí.

2.5.1 Statistické jazykové modely

Příklady statistických modelů jsou n-gramové LM [62]. Vycházejí z aproximace pravděpodobnosti sekvence $P(\mathbf{x}_{1:N})$ jako

[11] V úloze překladu z angličtiny do němčiny bylo dosaženo relativního zlepšení 0,39 % BLEU [97].

$$P(\mathbf{x}_{1:N}) \approx \prod_{i=1}^{N} P(x_i|\mathbf{x}_{i-n+1:i-1}), \tag{2.16}$$

kde n je faktor určující velikost uvažovaného n-gramu a $P(x_i|\mathbf{x}_{i-n+1:i-1})$ je definována jako

$$P(x_i|\mathbf{x}_{i-n+1:i-1}) = \frac{C(\mathbf{x}_{i-n+1:i})}{C(\mathbf{x}_{i-n+1:i-1})}, \tag{2.17}$$

kde $C(\mathbf{x}_{1:N})$ je počet výskytů příslušného n-gramu v trénovací množině [62]. Pro svoji jednoduchost jsou tyto modely nejčastěji využívany v aplikacích, kde je kritická doba odezvy systému.

Speciálním příkladem n-gramových LM je unigramový LM, předpokládající nezávislost slov v rámci sekvence a tedy modelující $P(\mathbf{x}_{1:N})$ jako

$$P(\mathbf{x}_{1:N}) = \prod_{i=1}^{N} P(x_i). \tag{2.18}$$

Mezi další statistické přístupy náleží například jazykové modely maximalizující entropii [12], či LM pracující v kontinuálním prostoru [125].

2.5.2 Jazykové modely vycházející z architektury Transformer

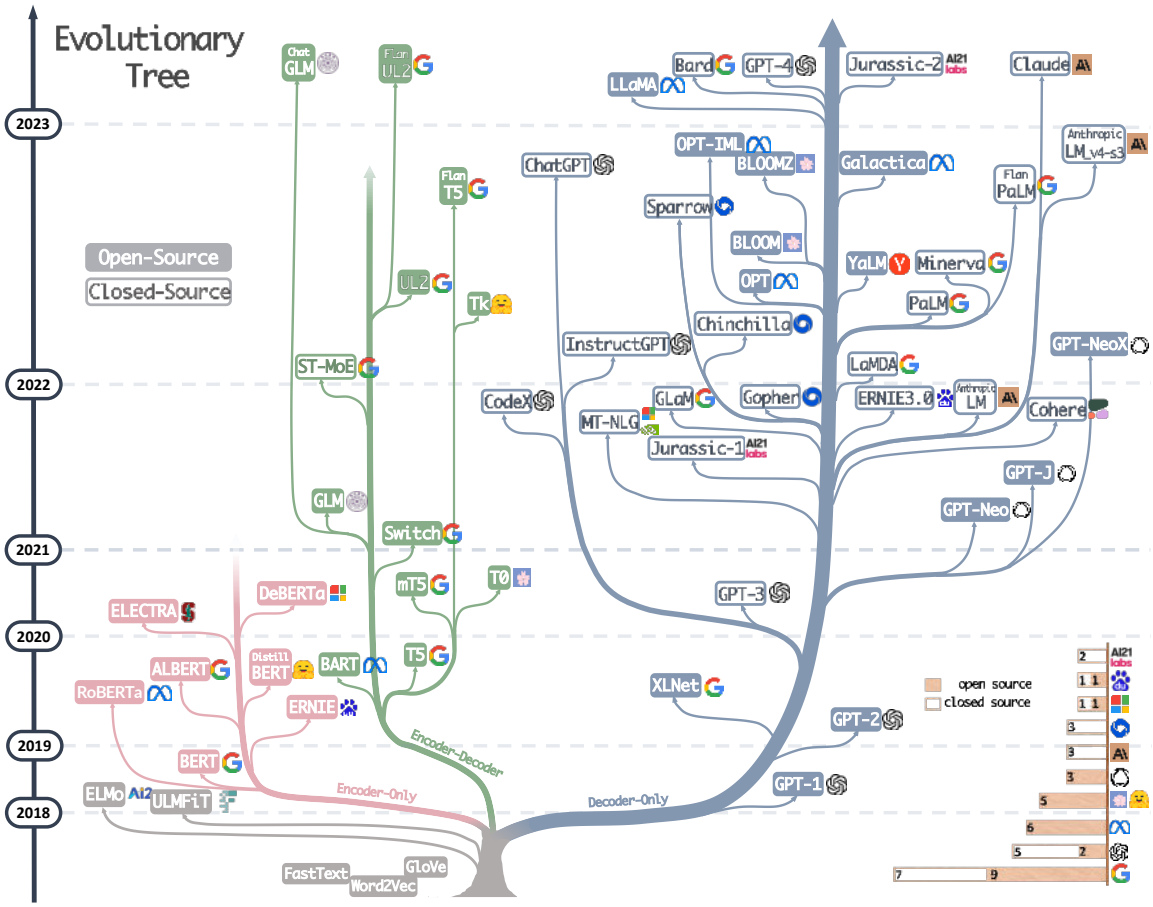
První úspěšné pokusy o jazykové modelování s využitím neuronových sítí byly provedeny již v roce 2003 [11]. V následujících letech byly úspěšnosti modelů postupně zlepšovány pomocí technik představených v sekcích 2.3.1 až 2.3.3. S využitím velkých předtrénovaných jazykových modelů však nastal v tomto odvětí velký rozmach. Tyto modely jsou nejčastěji předtrénovány učením se pod vlastním dohledem, viz sekce 2.2.2. V některých NLP úlohách došlo k přiblížení či dokonce překonání přesnosti lidských respondentů. Jazykové modely vycházející z architektury Transformer lze obecně rozdělit do tří kategorií viz obr. 2.6:

- obsahující pouze enkodér, často nazývané jako auto-kódující modely (*auto-encoding models*) – Tyto modely jsou nejčastěji předtrénovány tak, že se nějakým způsobem poškodí vstup (například se zamaskují náhodná slova, nebo poškodí návaznost vět) a úkolem modelu je danou sekvenci zrekonstruovat, či rozhodnout o návaznosti vět. Takto předtrénované modely jsou vhodné pro dotrénování na klasifikaci vět, extraktivní úlohy, rozpoznání entit, či sémantickou podobnost vět. Představiteli této kategorie jsou modely BERT [32], ALBERT [74], ELECTRA [28], DeBERTa [51], či RoBERTa [84].

- obsahující pouze dekodér, často nazývané jako auto-regresivní modely (*auto-regressive models*) – Tyto modely jsou nejčastěji učeny na predikování následujícího slova ve větě. Jsou používány k generování textu bez, či s využitím nějakého podnětu. Do této kategorie náleží modely jako GPT [111], GPT2 [112], GPT3 [18], GPT4 [95], LLaMa [134] CTRL [64], Transformer-XL [29], či XLNet [145].

- poslední kategorii tvoří modely obsahující jak enkodér, tak dekodér – Jedná se o modely jejichž vstupem je sekvence a výstupem také sekvence, odtud pochází také jejich název *Sequence-to-Sequence Models* (Seq2seq). Předtrénování těchto modelů lze provést pomocí postupů představených u kategorií výše, avšak obvykle jsou trénovány

složitěji. Například model T5 [113] se předtrénuje nahrazením náhodných úseků textu (které mohou obsahovat několik slov) speciálním *sentinel* tokenem a cílem je pak předpovědět text, který tato maskovaná slova nahrazuje. Jsou vhodné pro sumarizaci, překlad nebo generativní odpovědi na otázky [58]. Dalšími modely patřícími do této kategorie jsou například BART [80] nebo mBART [83].



Obrázek 2.6: Genealogický strom vývoje velkých jazykových modelů (*Large Language Models* – LLMs vertikální osa zobrazuje čas, šedá větev zobrazuje modely, jež nepracovaly s *attention* mechanizmem. Červená větev znázorňuje enkóder modely, zelená seq2seq varianty a modrá modely obsahující pouze dekodér, převzato z [144].

2.5.3 Maskované versus klasické jazykové modelování

Existují dva typy jazykového modelování, klasické (Causal Language Modeling – CLM) a maskované (Masked Language Modeling – MLM). Klasické jazykové modely se často používají pro generování textu. Klasické modelování jazyka předpovídá další token v posloupnosti. Model se může věnovat pouze předchozím tokenům. To znamená, že nevidí do budoucna. Příkladem klasického modelu je například GPT. Oproti tomu maskované modelování jazyka předpovídá maskovaný token v sekvenci a model může odvozovat kontext z minulých i budoucích tokenů. Maskované učení je skvělé pro úlohy, které vyžadují dobré

kontextové porozumění celé sekvenci jako například klasifikace sentimentu. Příkladem maskovaného jazykového modelu je BERT [58].

2.5.4 Perplexita – perplexity

Samotná pravděpodobnost sekvence $P(\mathbf{x}_{1:N})$ se může jevit jako dostatečná metrika hodnocení kvality modelu, avšak neuvažuje délku sekvence, a tedy není vhodná pro porovnání příslušnosti sekvencí různé délky. $P(\mathbf{x}_{1:N})$ je součinem podmíněných pravděpodobností příslušných slov, pro získání normalizované pravděpodobnosti P_{norm} je vhodné provést normalizaci délkou této sekvence N jako geometrický průměr příslušných faktorů

$$P_{norm}(\mathbf{x}_{1:N}) = (\prod_{i=1}^{N} P(x_i|\mathbf{x}_{1:i-1}))^{\frac{1}{N}}. \tag{2.19}$$

V praxi je však místo maximalizace normalizované pravděpodobnosti minimalizována perplexita PP, převracená hodnota P_{norm}.

$$PP(\mathbf{x}_{1:N}) = (\prod_{i=1}^{N} P(x_i|\mathbf{x}_{1:i-1}))^{-\frac{1}{N}} \tag{2.20}$$

V teorii informace je perplexita míra, jak dobře rozdělení pravděpodobnosti nebo pravděpodobnostní model předpovídá určitý vzorek. V kontextu NLP je jedním ze způsobů hodnocení kvality jazykového modelu. V případě generování textu zachycuje míru nejistoty modelu vůči své předpovědi.

2.6 Řečové modely

Obdobně jako je tomu v NLP, v oblasti zpracování řeči v posledních letech začínají taktéž dominovat předtrénované modely. Modely jsou nejdříve díky principu učení pod vlastním dohledem učeny k abstrahování relevantní informace z velkého kvanta dat – v tomto případě ze zvukových nahrávek[12]. Následně jsou tyto modely dotrénovány na cílovou úlohu jako například rozpoznání řeči, rozpoznání mluvčího či dokonce k detekci sentimentu. Tento proces je velmi výhodný zejména v situacích, kdy není k dispozici dostatek anotovaných dat pro příslušnou úlohu.

2.6.1 Wav2vec

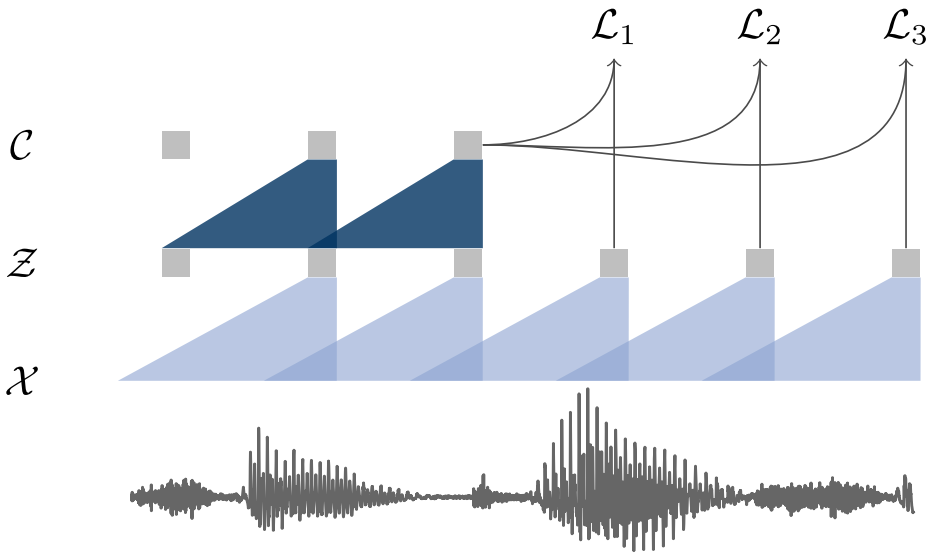
Wav2vec [123] je příkladem takového modelu. Jedná se o model představený v roce 2019 a jak úvod napovídá, klíčovou myšlenkou tohoto modelu je naučit se obecné reprezentace v prostředí, kde je k dispozici velké množství dat a tyto naučené reprezentace dále využít ke zlepšení výkonu v navazujících úlohách.

Architektura modelu

Architektura tohoto modelu je zobrazena na obr. 2.7. Skládá se ze dvou konvolučních neuronových sítí. První z nich, enkodér, promítá část vstupního audio signálu $\mathbf{x_i} \in \mathcal{X}$ do prostoru skrytých proměnných (*latent space*) $f : \mathcal{X} \mapsto \mathcal{Z}$. Tímto způsobem je snížena dimensionalita vstupních dat. Vzniklé skryté reprezentace $\mathbf{z}_{i-\nu}, \ldots, \mathbf{z}_i$ (dimensionalita těchto

[12] Někteří autoři tento proces přirovnávají k poslouchání nahrávky [57].

vektorů je v původním článku $d = 512$ a odpovídá 30 ms nahrávky) jsou dále pomocí druhé kontextuální sítě $g : \mathcal{Z} \mapsto \mathcal{C}$ zkombinovány do kontextualizovaného vektoru $\mathbf{c}_i = g(\mathbf{z}_{i-\nu:i})$ s dimenzionalitou 512, kde ν je parametr určující velikost receptivního pole sítě (*receptive field*). V provedených experimentech bylo uváženo $\nu = 7$, což zhruba odpovídá 210 ms.



Obrázek 2.7: Architektura modelu Wav2vec, vstupní signál $\mathcal{X}$ je postupně zpracován dvěma konvolučními neuronovými sítěmi. Kontextuální reprezentace $\mathcal{C}$ je spočtena s $\nu = \mathbf{2}$, $\mathcal{L}_{1:3}$ symbolizují hodnoty $k = 3$ částečných objektivních funkcí, převzato z [123].

Objektivní funkce

Model je učen kontrastivním způsobem, viz sekce 2.2.2, k tomu, aby dokázal rozlišit v kroku i nadcházející vzorky $\mathbf{z}_{i+k}$ od distraktorů $\tilde{\mathbf{z}}$, tedy vzorků pocházejících z pomocné uniformní distribuce $p_n(\mathbf{z}) = 1/T$, kde T je délka sekvence v prostoru $\mathcal{Z}$, $i \in \{1, \ldots, T\}$, $k \in \{1, \ldots, K\}$ a K je počet budoucích vzorků, jimž je predikována hodnota. Přesněji je model učen k minimalizaci objektivní funkce $\mathcal{L} = \sum_{k=1}^{K} \mathcal{L}_k$, kde $\mathcal{L}_k$ je definována jako

$$\mathcal{L}_k = -\sum_{i=1}^{T-k} (\log \sigma(\mathbf{z}_{i+k}^\intercal h_k(\mathbf{c}_i)) + \lambda \mathop{\mathbb{E}}_{\tilde{\mathbf{z}} \sim p_n} [\log \sigma(-\tilde{\mathbf{z}}_{i+k}^\intercal h_k(\mathbf{c}_i))]), \tag{2.21}$$

kde $\sigma(x) = 1/(1 + \exp(-x))$ je sigmodailní funkce, $\log \sigma(\mathbf{z}_{i+k}^\intercal h_k(\mathbf{c}_i))$ odpovídá pravděpodobnosti pozitivního vzorku, h_k je afinní transformace $h_k(\mathbf{c}_i) = \mathbf{W}_k \mathbf{c}_i + \mathbf{b}_k$ a λ je počet negativních vzorků – distraktorů.

Takto natrénované kontextuální reprezentace jsou následně předány do akustického modelu [108, 147], s jehož pomocí je dále možno provést automatické rozpoznání řeči (*Automatic Speech Recognition* – ASR). Je nutno podotknout, že takto navržený model uvažuje k predikci pouze předchozí vzorky $\mathbf{z}_{i-\nu}, \ldots, \mathbf{z}_i$.

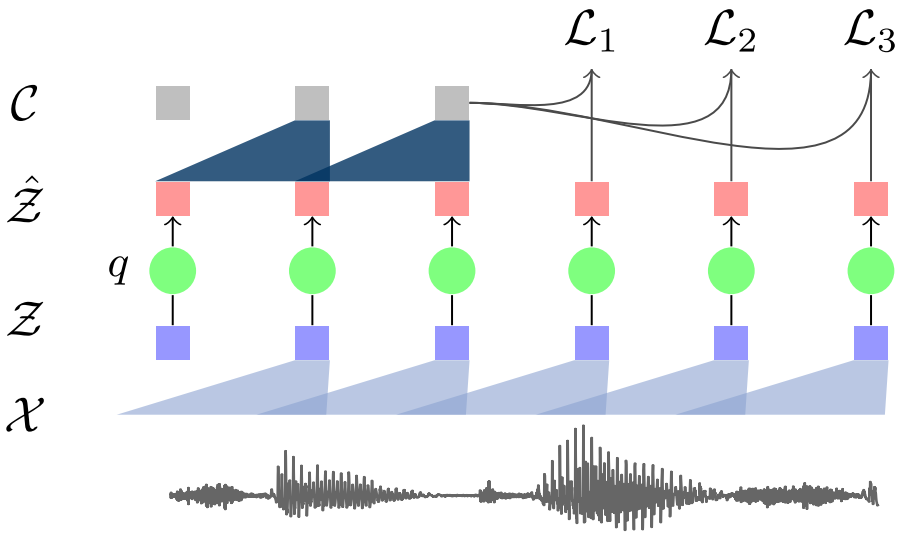
2.6.2 Vq-wav2vec

Na předchozí architekturu navazuje článek představující architekturu Vq-wav2vec [5] stejných autorů. Hlavní myšlenkou tohoto článku je převést vstupní signál do diskrétní podoby

tak, aby bylo možné efektivně využit Transformery pro tvorbu kontextuálních reprezentací. Autoři tohoto procesu diskretizace docílili tak, že představili nový kvantizační modul $q : \mathcal{Z} \mapsto \hat{\mathcal{Z}}$. Tento modul funguje na principu náhodně inicializované kódové knihy (*codebook*), jejíž prvky jsou následně učeny k tomu, aby co nejlépe reprezentovaly veličinu skrytých proměnných $\mathcal{Z}$.

Architektura modelu

Obdobně jako v předchozím případě je nejdřív vstupní audio zpracováno enkodérem $f : \mathcal{X} \mapsto \mathcal{Z}$. Následně jsou nahrazeny skryté reprezentace $\mathbf{z}$ odpovídajícími vektory $\hat{\mathbf{z}} = \mathbf{e}_i$ z kódové knihy $\hat{\mathbf{e}} \in \mathbb{R}^{V \times d}$, kde V je počet prvků kódové knihy (případně $\mathbf{e} \in \mathbb{R}^{V \times (d/G)}$, pokud je uváženo G kódových knih – $\hat{\mathbf{z}}$ je sestrojen konkatenací příslušných vektorů $\mathbf{e_1}, \ldots, \mathbf{e_G}$, což samozřejmě vede k více variantám, jak může být vektor $\hat{\mathbf{z}}$ sestrojen a zároveň zvyšuje využití příslušných kódových vektorů, což odstraňuje nutnost regularizace v objektivní funkci). Je nutno podotknout, že $\hat{\mathbf{z}}$ jsou stále vektory z hustého pole a pouze $\mathbf{i} \in [V]^G$ jsou diskrétní indexy. Dále je využita stejná kontextuální síť $g : \hat{\mathcal{Z}} \mapsto \mathcal{C}$ pro získaní kontextuálních vektorů. Architektura modelu je představena na obr. 2.8.



Obrázek 2.8: Architektura modelu Vq-wav2vec, vstupní signál $\mathcal{X}$ je postupně zpracován konvoluční sítí f, kvantizován s využitím kvantizačního modulu q a následně zpracován druhou konvoluční sítí g, čímž vzniká kontextuální reprezentace $\mathcal{C}$, převzato z [5].

Kvantizační modul

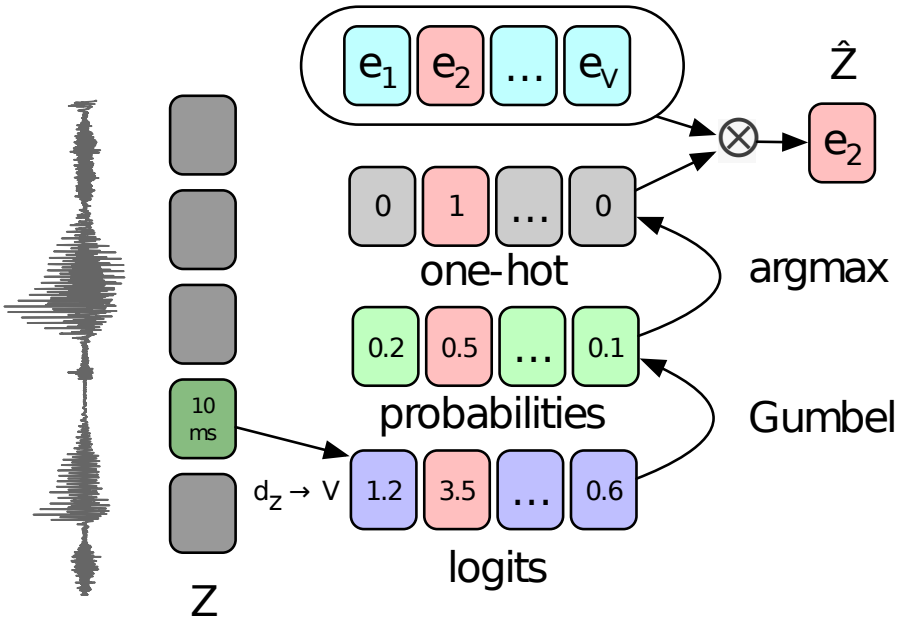
Doposud nebylo představeno, jak je samotná kvantizace provedena. Jedním ze způsobů se jeví využití kódového vektoru s nejmenší vzdáleností vůči vektoru $\mathbf{z}$, tedy hledání i splňujícího $i = \arg\min_j ||\mathbf{z} - \mathbf{e}_j||_2^2$. Příslušné úpravy objektivní funkce a další potřebné úpravy jsou detailněji představeny v původním článku v sekci 3.2 K-means.

Jako druhou variantu autoři uvážili aplikaci další lineární vrstvy na vektorech $\mathbf{z}$, aktivační vrstvy ReLU [39] a další lineární vrstvy s výstupem $\mathbf{l} \in \mathbb{R}^V$ (*logits*). Takto získané *logits* určují váhu, s jakou má být vybrán příslušný kódový vektor. Tyto váhy jsou následně

převedeny na pravděpodobnostní distribuci pomocí operace Gumbel-Softmax definované jako

$$p_j = \frac{\exp(l_j + \nu_j)/\tau}{\sum_{k=1}^{V} \exp(l_k + \nu_k)/\tau}, \tag{2.22}$$

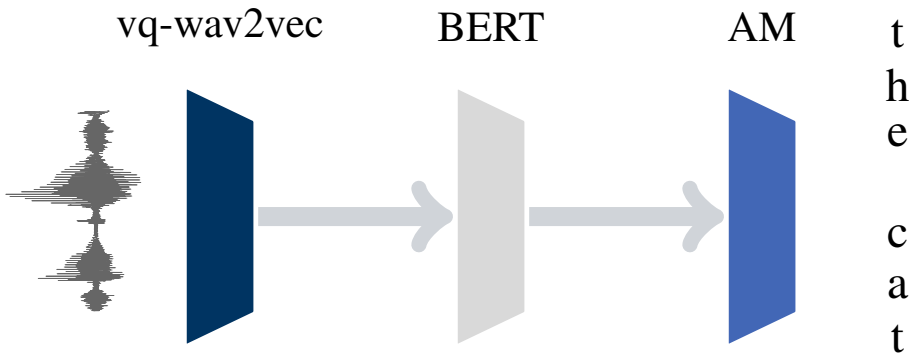
kde $\nu = -\log(-\log(u)), u \sim \mathcal{U}(0,1)$ reprezentuje Gumbelův šum (Gumbel noise) a τ je teplotní parametr, který určuje, jak moc se vzorky blíží kategorickému rozložení. Při nízké teplotě τ aproximuje výsledná distribuce operaci arg max. V rámci dopředného průchodu je $\hat{\mathbf{z}}$ sestrojen výběrem kódového vektoru $\mathbf{e}_i$ splňujícího $i = \arg\max_j p_j$. Při zpětném průchodu je předán gradient operace Gumbel-Softmax. Tímto přístupem je překonána nederivovatelnost operace arg max. Celý proces je znázorněn na obr. 2.9.



Obrázek 2.9: Znázornění funkce kvantizačního modulu $q : \mathcal{Z} \mapsto \hat{\mathcal{Z}}$ s využitím operace Gumbel-Softmax, převzato z [5].

Využití diskrétní reprezentace

Doposud byly indexy i uvažovány pouze pro kvantizaci vektorů $\mathbf{z} \mapsto \hat{\mathbf{z}}$, avšak tímto způsobem se model nepřímo naučil mapovat kusy audia na diskrétní indexy i. Ty byly dále využity pro natrénování jazykového modelu BERT [32], jehož maskované učení dovoluje vytvořit kontextuální reprezentace uvažující kontext z obou stran. Jelikož audio je kvazistacionární a každý z tokenů v této práci reprezentuje zhruba 10 ms, autoři byli nuceni maskovat 10 po sobě jdoucích tokenů, aby úloha nebyla příliš jednoduchá a model byl schopný se učit. Kontextuální reprezentace byly následně využity pro naučení akustického modelu obdobně jako v předchozím případě, viz obr. 2.10. Autoři se v rámci této práce rovněž pokusili natrénovat end-to-end model vynecháním akustického modelu, což vedlo k následující práci pojmenované wav2vec 2.0 [6].



Obrázek 2.10: Navržený princip trénování diskretizace audia za účelem rozpoznávání řeči. Vstupní nahrávka je diskretizována pomocí modulu Vq-wav2vec, na indexy je aplikován jazykový model BERT a výsledné reprezentace jsou dále předány do akustického modelu (AM), čímž vzniká výsledný přepis, převzato z [5].

2.6.3 Wav2vec 2.0

Další ze série článků stejných autorů představuje další architekturu postavenou na principu konstrastivního prediktivního učení. Na rozdíl od předchozí architektury autoři nahradili kontextuální konvoluční síť $g : \mathcal{Z} \mapsto \mathcal{C}$ Transformerem, což vedlo k dalšímu pokroku v modelování řeči [6].

Architektura

Architektura tohoto modelu je velice podobná předchozímu modelu. Blok $f : \mathcal{X} \mapsto \mathcal{Z}$ je funkčně ekvivalentní, je pouze nahrazena aktivační funkce *Rectified Linear Unit* (ReLU) za *Gaussian Error Linear Unit* (GELU) [53] a je přidána normalizace.

Kontextuální konvoluční síť $g : \mathcal{Z} \mapsto \mathcal{C}$ je nahrazena klasickým Transformerem, kde namísto fixního pozičního kódování byla využita konvoluční vrstva. Je nutno podotknout, že na rozdíl od architektury vq-wav2vec však vstup není kvantizován.

Kvantizační modul v tom případě plní odlišnou funkci. Je zde využitá kódová kniha o G skupinách viz sekce 2.6.1, příslušné kódové vektory $\mathbf{e}_1, \ldots, \mathbf{e}_G$ jsou konkatenovány a je aplikována lineární transformace $\mathbb{R}^d \mapsto \mathbb{R}^f$, čímž je získán kvantizovaný vektor $q \in \mathbb{R}^f$. V kvantizačním modulu je opět využit Gumbel-Softmax operátor s jedinou odlišností oproti předchozímu modelu. Model pracuje nad $\mathbf{l} \in \mathbb{R}^{G \times V}$. Na obr. 2.11 je zobrazena architektura tohoto modelu.

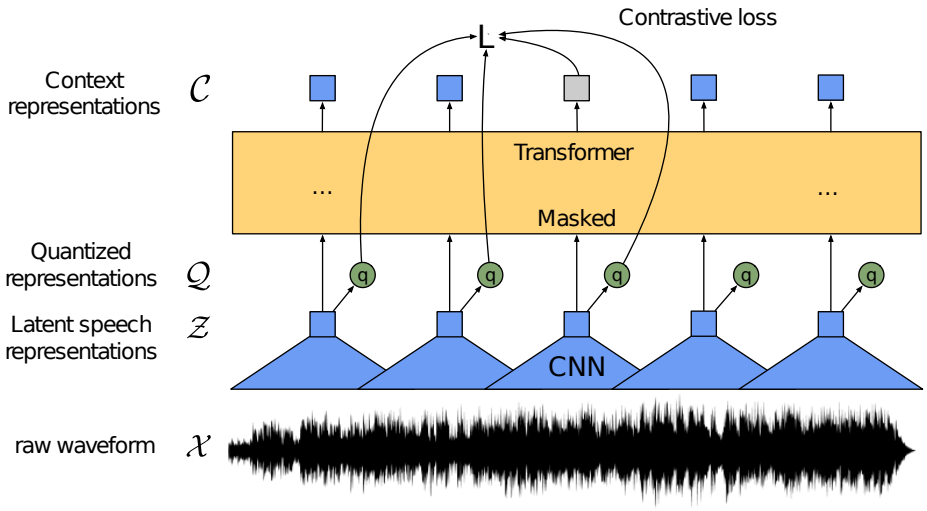
Objektivní funkce

Objektivní funkce $\mathcal{L}$ je v tomto případě velmi podobná a skládá se ze dvou částí

$$\mathcal{L} = \mathcal{L}_m + \alpha \mathcal{L}_d, \tag{2.23}$$

kde α je učitelný parametr.

$\mathcal{L}_m$ je kontrastivní objektivní funkce (*contrastive loss*). Má za cíl se znalostí kontextualizovaného vektoru $\mathbf{c}_t$ rozlišit správný kvantizovaný vektor $\mathbf{q}_t$ z množiny $K+1$ kvantizovaných kandidátních řešení $\tilde{\mathbf{q}} \in \mathbf{Q}_t$, kde K je počet distraktorů. Je definována jako



Obrázek 2.11: Architektura modelu wav2vec 2.0, převzato z [6].

$$\mathcal{L}_m = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa}, \tag{2.24}$$

kde $\kappa = 0,1$ je teplotní parametr a $sim(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\intercal \mathbf{b}/||\mathbf{a}||||\mathbf{b}||$ je kosinová podobnost vektorů $\mathbf{a}$ a $\mathbf{b}$.

$\mathcal{L}_d$ je objektivní funkce diverzity (*diversity loss*) navržená ke zvýšení využití všech položek kódové knihy. Jedná se o zprůměrovanou entropii distribuce $\bar{p}_g$ napříč kódovými položkami příslušné skupiny g v rámci daného batche se zanedbáním Gumbelova šumu ν a teploty τ daná vztahem

$$\mathcal{L}_d = \frac{1}{GV} \sum_{g=1}^{G} -H(\bar{p}_g) = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \bar{p}_{g,v} \log \bar{p}_{g,v}. \tag{2.25}$$

Takto definovaný model může být následně využit k přímému natrénovaní systému ASR a dosahuje s malým množstvím dat úctyhodné výsledky.

2.6.4 ClTRUS

Czech language TRransformer from Unlabeled Speech – ClTRUS [79] je monolinguální řečový transformer z rodiny wav2vec 2.0 předtrénovaný pro češtinu. Model je předtrénován na korpusu 80 tisíc hodin mluvené češtiny. Korpus se skládá z 22 tis. hodin nahrávek radiových přenosů, 18,7 tis. hodin nahrávek z datasetu VoxPopuli [136], 15 tis. hodin televizního vysílání, 12 tis. hodin obsahujících stínové řečníky, 5 tis. hodin sportovních přenosů, 2 tis. hodin telefonních dat, a menšího počtu dat z různých domén. Model byl trénován na nahrávkách nepřesahujících délku 30 sekund tak, aby bylo možné model trénovat s rozumnou velikostí batche. Autoři trénovali model s totožnými hyper-parametry, jak tomu bylo v původním článku představujícím rodinu modelů wav2vec 2.0 [6]. Model byl trénován na 400 tis. krocích s velikostí batche okolo 1,6 h, což odpovídalo zhruba 11 epochám. Předtrénovaný model bez tokenizeru a klasifikační vrstvy byl zveřejněn na portálu Hugging Face[13].

[13] https://huggingface.co/fav-kky/wav2vec2-base-cs-80k-ClTRUS

2.6.5 XLS-R

XLS-R [4] je rozsáhlý model pro učení mezijazykové reprezentace řeči založený na wav2vec 2.0. Autoři představili 3 varianty (300 mil., 1 mld. a 2 mld. parametrů[14]) modelu trénovaného na 436 tis. hodinách veřejně dostupných zvukových záznamů řeči ve 128 jazycích. Čeština je zastoupena 18,5 tis. hodinami. Autorům se podařilo zlepšit průměrnou úspěšnost překladu z 21 zdrojových jazyků do angličtiny o 7,4 BLEU [97]. V rozpoznání řeči dosáhli relativních zlepšení 14-34 % na datasetech BABEL [41], MLS [109], CommonVoice [2] a VoxPopuli [136]. Zlepšení bylo taktéž dosaženo v dalších odvětvích, jako třeba identifikace jazyka.

Aby autoři vyvážili zastoupení jazyků v procesu předtrénování, data v batchi jsou vzorkována z rozložení

$$p_l = (\frac{n_l}{N})^\alpha, \tag{2.26}$$

kde N je celkový počet vzorků, $l \in \{1, \ldots, L\}$, n_l je počet vzorků v daném jazyce a α je koeficient ovlivňující zastoupení jazyků obsahující vysoký versus nízký počet vzorků v procesu předtrénování. Trénovací vzorky byly oříznuty na maximálně 320K vzorků, což je v případě vzorkování na frekvenci 16 kHz rovno 20 sekundám. 300 mil. varianta byl trénován na 128 GPU s zhruba 4,3 h dat v jednom batchi. Větší varianty byly trénovány na 200 GPU s velikosti batche přibližně 2,8-3,6 hodiny.

2.6.6 Whisper

Whisper [110] je vícejazyčný multimodální model natrénovaný na 680 tis. hodinách supervizovaných dat. Dataset obsahuje 117 tis. hodin dat v jazycích jiných než angličtina a 125 tis. hodin dat pro překlad ze zdrojového jazyka do angličtiny. Jedná se o encoder-decoder Transformer, jehož váhy byly zveřejněny ve verzích počínaje 39 mil. po 1,55 mld. parametrů v anglické a vícejazyčné variantě[15]. Vstupní nahrávky jsou převzorkovány na frekvenci 16 kHz a je spočten 80-kanálový log Mel spektrogram s oknem o velikosti 25 ms s krokem 10 ms. Příznaky jsou dále normalizovány do rozsahu $\langle -1; 1 \rangle$ se střední hodnotou 0 napříč trénovacím datasetem. Data byla stažena z veřejně dostupných zdrojů a byla automaticky očištěna od strojově generovaných přepisů. Systém byl trénován na 30 s fragmentech nahrávek obsahujících i tiché segmenty, čímž se model naučil detekovat i řečovou aktivitu. Byl použit předtrénovaný tokenizer z modelu GPT2 [112], který byl pro vícejazyčné verze přetrénován, avšak bez změny velikosti. Model byl trénován s velikostí batche 256 segmentů přes 2^{20} kroků, což zhruba odpovídá 2-3 průchodům datasetem, čímž byla zajištěna dostatečná diverzita trénovacích dat, a tak i nepřímá regularizace, což vedlo k tomu, že nebyly využity žádné augmentace.

[14] https://huggingface.co/facebook/wav2vec2-xls-r-300m
[15] https://huggingface.co/openai/whisper-large-v2

Kapitola 3

Úlohy

V kapitole jsou představeny příslušné úkoly z oblasti zpracování řeči a přirozeného jazyka, jež byly v rámci této práce řešeny a zkoumány. Postupně je představena každá úloha a metriky použité pro hodnocení kvality kandidátních řešení.

3.1 Detekce řečové aktivity

Detekce řečové aktivity (*Voice Activity Detection* – VAD) je třída metod, jejichž cílem je určit, zda příslušný úsek řeči obsahuje řeč, či nikoliv. Ve většině případů jsou tyto metody využívané pro předzpracování vstupní nahrávky. V kódování řeči lze například použít VAD pro deaktivaci kodéru, a tedy pro snížení množství přenesených informací po kanálu. V rozpoznání řeči je VAD nejčastěji používáno k redukci výpočetní náročnosti vynecháním určitých úseků vstupní nahrávky. VAD může být také využito v procesu zlepšení kvality nahrávky, kdy může být z neřečových segmentů odhadnut okolní šum, který může být dále extrahován z nahrávky. Problém lze tedy formulovat tak, že hledáme y pro vstupní signál x takové, že

$$y = \begin{cases} 1 & \text{pokud je } x \text{ řeč} \\ 0 & \text{pokud } x \text{ není řeč.} \end{cases} \tag{3.1}$$

Se znalostí pravděpodobnosti řeči $p(y = 1)$ lze problém přeformulovat jako

$$y = \begin{cases} 1 & p(y = 1) \geq \theta \\ 0 & p(y = 1) < \theta, \end{cases} \tag{3.2}$$

kde θ je práh [19]. Ilustrační výstup VAD systému je zobrazen v rámci obr. 3.1.

Chybovost systému detekce řečové aktivity může být vyhodnocena několika způsoby. Nechť $\hat{\mathbf{y}} \in \{0, 1\}^N$ je vektor hypotéz a $\mathbf{y} \in \{0, 1\}^N$ vektor referenčních hodnot délky N. Potom počet falešných poplachů (*false alarm*) F a vynechaných (*miss detected*) M segmentů se rovnají

$$F = \sum_{i=1}^{N} \delta(\hat{y}_i - y_i), \tag{3.3}$$

$$M = \sum_{i=1}^{N} \delta(y_i - \hat{y}_i), \tag{3.4}$$

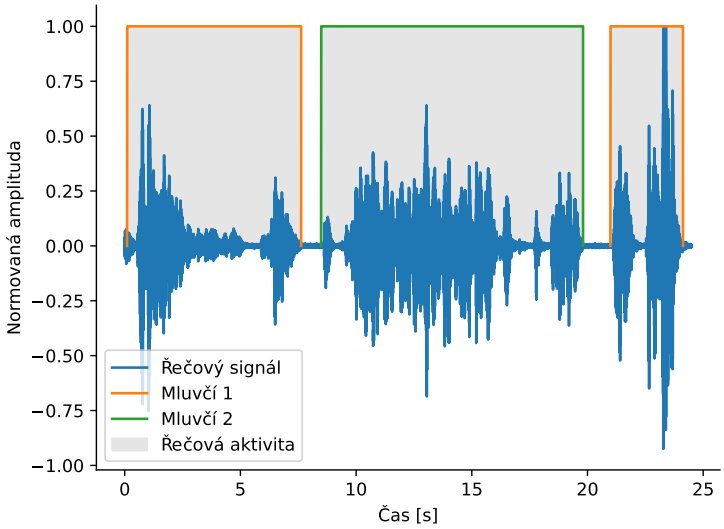
kde $\delta(x) = \begin{cases} 0 & x = -1 \\ x & jinak \end{cases}$. Celková chyba detekce řečové aktivity D – *detection error rate* je rovna

$$D = \frac{F + M}{N}. \tag{3.5}$$

3.2 Diarizace

Diarizace je jedním ze stavebních bloků pokročilých systémů zpracování řeči. Jedná se o proces detekce časových okamžiků změn mluvčích, segmentace nahrávky do úseků obsahujících příslušného řečníka a následné shlukování segmentů se stejnou identitou mluvčího.

Diarizace odpovídá na otázku „kdo kdy mluví“ v prostředí více mluvčích. Její výsledkem je taktéž detekce úseků, ve kterých nehovoří ani jeden z mluvčích. Tento proces nevyžaduje apriorní znalost identity nebo počtu řečníků, naproti tomu proces rozpoznání mluvčího odpovídá na otázku „kdo mluví“ se znalostí identity jedinců. Příklad výstupu diarizačního systému je zobrazen na obr. 3.1.



Obrázek 3.1: Příklad výstupu systému pro diarizaci. Podbarvena je část signálu označena systémem VAD jako řeč.

Tradičně se proces diarizace skládal z několika dílčích kroků, jako detekce řečové aktivity, segmentace, extrakce příznaků, shlukování a případného dalšího zpracování [99]. S nástupem hlubokých neuronových sítí se od tohoto přístupu přechází k *end-to-end* systémům [73].

Nechť $\mathbf{y} \in \mathbb{N}_0^N$ je vektor referenčních hodnot a $\hat{\mathbf{y}} \in \mathbb{N}_0^N$ vektor hypotéz. Chybovost systému diarizace je nejčastěji vyjadřována metrikou *Diarization Error Rate* (DER), která je definována jako

$$\mathrm{DER} = \frac{M + F \ + C}{\mathrm{N}}, \tag{3.6}$$

kde

$$F = \sum_{i=1}^{N} \delta(\text{sgn}(\hat{y}_i) - \text{sgn}(y_i)), \tag{3.7}$$

$$M = \sum_{i=1}^{N} \delta(\text{sgn}(y_i) - \text{sgn}(\hat{y}_i)), \tag{3.8}$$

$$C = \sum_{i=1}^{N} \text{sgn}(|y_i - \hat{y}_i|), \tag{3.9}$$

$$\text{sgn}(x) = \begin{cases} 0 & x = 0 \\ 1 & x > 0 \\ -1 & jinak \end{cases}, \tag{3.10}$$

kde N je počet rámců, M (*miss*) počet rámců, ve kterých alespoň jeden z mluvčích mluvil, ale systém tento úsek klasifikoval jako ticho. F (*false alarm*) je počet rámců, ve kterých byla detekována řečová aktivita, avšak nikdo nemluvil, a C (*confusion*) je počet rámců, ve kterých systém přiřadil špatného mluvčího.

Jelikož ruční anotace nejsou vždy zcela přesné, často je v rámci evaluace využit „límec“ (*collar*), což je úsek v referenční anotaci kolem časové události změny z ticha na mluvu a opačně. Segmenty nacházející se v prostoru „límce“ nejsou započítány do výsledné chyby diarizace.

3.3 Detekce překrývající se řeči

V prostředí psychoterapeutických sezení je velmi častým jevem skákaní do řeči, nebo-li překrývající se řeč. Jedná se o segmenty nahrávky, kdy hovoří vícero mluvčích najednou. Psychoterapeuti skoky do řeči často vyjadřují svoji přítomnost pro klienta. Nejčastěji se jedná o velmi krátké přitakávací úseky promluvy jako např. „hmmm“, „jo“.

Naivním řešením pro detekci překryvů se zdá porovnání vektorů hlasových otisků mluvčích s příznakovými vektory daných úseků a pokud u vícero mluvčích dochází k dostačující shodě (skalární součin, kosinová podobnost, *Probabilistic Linear Discriminant Analysis* (PLDA) skóre [60]), úsek je prohlášen za překrývající se řeč. Dalšími variantami je možnost použití neuronových sítí např. s koeficienty spektrální hustoty výkonu, nebo *Mel Frequency Cepstral Co-efficients* (MFCC) na vstupu [146, 72].

Pro vyhodnocení systémů pro detekci překrývající se řeči je nejčastěji využito preciznosti (*precision*) P a senzitivity (*recall*) R, případně jejich kombinace v podobě F1 skóre. Nechť $\hat{\mathbf{y}} \in \{0,1\}^N$ je vektor hypotéz a $\mathbf{y} \in \{0,1\}^N$ vektor referenčních hodnot délky N, potom

$$P = \frac{\sum_{i=1}^{N} \hat{y}_i}{|\hat{\mathbf{y}}|}, \tag{3.11}$$

$$R = \frac{\sum_{i=1}^{N} \hat{y}_i}{|\mathbf{y}|}, \tag{3.12}$$

$$\text{F1} = 2 \cdot \frac{P \cdot R}{P + R}. \tag{3.13}$$

3.4 Rozpoznávání řeči

Cílem systému pro automatické rozpoznávání řeči (*Automatic Speech Recognition* – ASR) je nalezení nejpravděpodobnější sekvence indexů slov $\hat{\mathbf{y}} \in \mathbb{N}^N$ ze slovníku V vzhledem k vstupnímu řečovému signálu $\mathbf{x} \in \mathbb{R}^M$. Problém rozpoznávání řeči je definován jako převod mluvených projevů na text. Hledané $\hat{\mathbf{y}}$ lze vyjádřit jako

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}). \tag{3.14}$$

Podle Bayesova pravidla lze posteriorní pravděpodobnost $P(\mathbf{y}|\mathbf{x})$ vyjádřit jako

$$P(\mathbf{y}|\mathbf{x}) = \frac{P(\mathbf{x}|\mathbf{y})P(\mathbf{y})}{P(\mathbf{x})}. \tag{3.15}$$

Jelikož $P(\mathbf{x})$ je konstanta vzhledem k hledanému $\hat{\mathbf{y}}$, lze výraz zjednodušit do následující podoby

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} P(\mathbf{x}|\mathbf{y})P(\mathbf{y}), \tag{3.16}$$

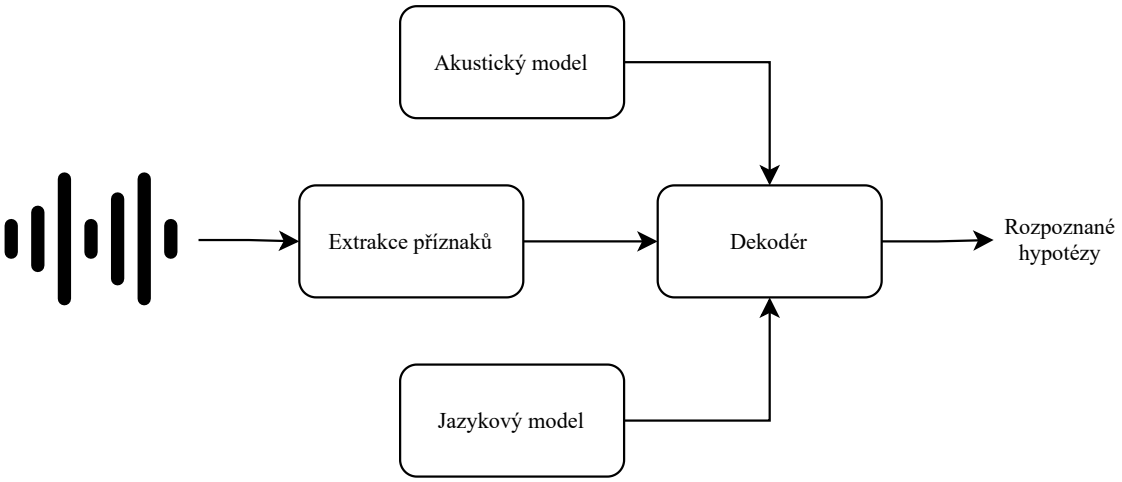
kde $P(\mathbf{y})$ lze modelovat jazykovým a $P(\mathbf{x}|\mathbf{y})$ akustickým modelem [19].

Architektura systému pro automatické rozpoznávání řeči se v průběhu let měnila. Jedna ze standardních architektur je zobrazena na obr. 3.2. Skládá se z následujících komponent:

- Extraktor příznaků: Převádí řečový signál na posloupnost vektorů akustických příznaků. Tyto vektory by měly být kompaktní a zároveň nést dostatek informací. Některé systémy pracují přímo s řečovými signály, které jsou pouze převzorkovány do vhodné vzorkovací frekvence, některé naopak používají velmi kompaktní podobu jako třeba MFCC.
- Akustický model: Statisticky modeluje pravděpodobnost fonémů, znaků podle jednotlivých vstupních zvuků, které tvoří slova v jazykovém modelu nebo gramatice.
- Jazykový model: Přiřazuje příslušným slovům, n-gramům, či sekvencím pravděpodobnost jejich výskytu v daném jazyku.
- Dekodér: Software, který přijímá akustické příznaky z extraktoru a postupně prohledává možné textové hypotézy. V případě prostředí s malým počtem výpočetních zdrojů může být využito chamtivé/lačné/hltavé prohledávání (*gready search*), nejčastěji je však nejlepší hypotéza vyhledávána pomocí paprskového algoritmu (*beam search*) [19].

Architektura modelů pro ASR se však výrazně změnila s nástupem hlubokých neuronových sítí. Původně hojně používané systémy využívající směsi Gaussovských rozložení (*Gaussian Mixture Model* – GMM) [13] a skrytých Markovových modelů (*Hidden Markov Model* – HMM [13]) byly postupně nahrazeny end-to-end (E2E) systémy. Hlavní výhodou tohoto přístupu je jediná objektivní funkce, která je konzistentní s požadavky na systém pro ASR. To dovoluje tuto síť optimalizovat napřímo na rozdíl od hybridních modelů, kde jsou příslušné komponenty optimalizované izolovaně. Návrh tradičních hybridních modelů je taktéž velmi komplikovaný a vyžaduje mnoho odborných znalostí s dlouholetými zkušenostmi s ASR [81].

Mezi nejpopulárnější techniky pro E2E ASR patří podle [81]:



Obrázek 3.2: Příklad standardního systému pro automatické rozpoznávání řeči.

1. **Connectionist Temporal Classification** (CTC) [48] – tato technika byla navržena s cílem, aby mapovala vstupní řečový signál $\mathbf{X} \in \mathbb{R}^{N \times D}$ na výstupní sekvenci $\hat{\mathbf{y}} \in \{0, \dots V\}^N$, kde D je dimensionalita řečového rámce, N délka sekvence a V je počet tokenů dané abecedy, a referenční výstupní sekvence $\mathbf{y} \in \{0, \dots, V\}^M$ tak, že $M \leq N$. Jelikož sousední rámce mohou obsahovat totožný znak, byl představen speciální *blank* token, jenž modeluje tiché segmenty, jakož i slouží k rozlišení opakujících se sousedních tokenů – například v případě slova *bezzubý* se systém musí naučit vložit *blank* token mezi sousední písmena z tak, aby bylo slovo rozpoznáno správně. Z předchozího předpokladu taktéž vychází, že referenční sekvence $\mathbf{y}$, může odpovídat několika hypotézám $\hat{\mathbf{y}}$. Pro definici objektivní funkce je nutné tuto n-tici hypotéz označit jako $B^{-1}(\mathbf{y})$, objektivní funkce je dále definována jako

$$\mathcal{L}_{CTC} = -\log P(\mathbf{y}|\mathbf{X}) \tag{3.17}$$

$$P(\mathbf{y}|\mathbf{X}) = \sum_{\mathbf{q} \in B^{-1}(\mathbf{y})} P(\mathbf{q}|\mathbf{X}) \tag{3.18}$$

a za předpokladu podmíněné nezávislosti mezi rámci

$$P(\mathbf{q}|\mathbf{X}) = \prod_{n=1}^{N} P(q_n|\mathbf{X}_n). \tag{3.19}$$

2. **Attention-based Encoder-Decoder** (AED) [25, 8] – je dalším typem end-to-end modelu, na rozdíl od předchozí varianty obsahuje navíc dekódovací síť a interní *Attention* vrstvu. Objektivní funkce je totožná, pravděpodobnost $P(\mathbf{y}|\mathbf{X})$ je již však modelována jako

$$P(\mathbf{y}|\mathbf{X}) = \prod_{m=1}^{M} P(y_m|\mathbf{y}_{1:m-1}, \mathbf{X}). \tag{3.20}$$

Enkodér zde pracuje stejným způsobem a je využit k extrakci vysokoúrovňových příznaků, jež jsou postupně převáděny pomocí *Attention* vrstvy na kontextové vektory $\mathbf{c}_m$, které jsou dále využity k autoregresivnímu generování výstupní sekvence.

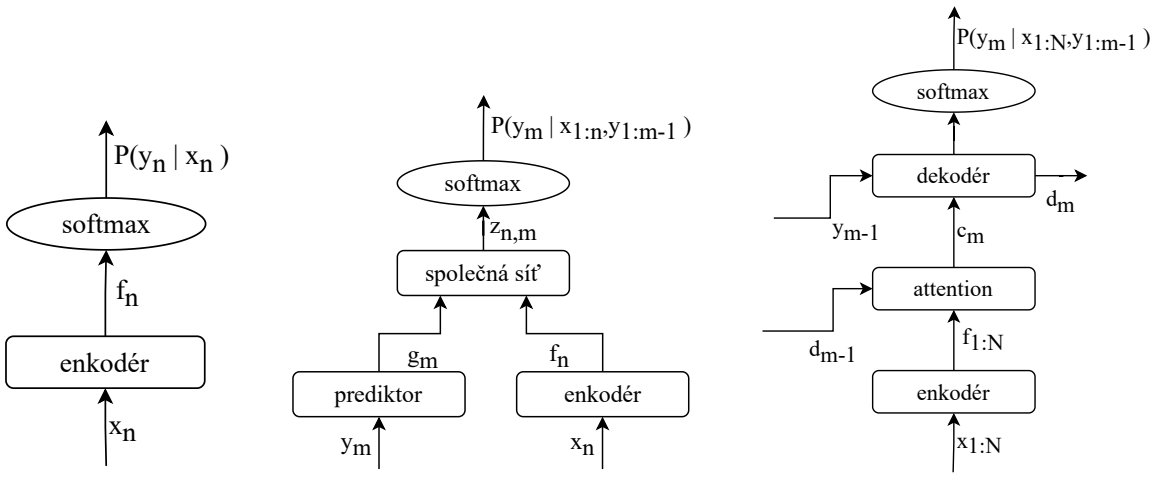
3. **Recurrent Neural Network Transducer** (RNN-T) [47] je technika modelování vhodná především pro streamovací ASR, jelikož při modelování pravděpodobnosti výstupní sekvence $\mathbf{y} \in \{0, \ldots, V\}^M$ lze uvažovat pouze předchozí výstupní tokeny a doposud viděné vstupní řečové rámce

$$P(\mathbf{y}|\mathbf{X}) = \prod_{m=1}^{M} P(y_m|\mathbf{y}_{1:m-1}, \mathbf{X}_{1:n}), \tag{3.21}$$

kde n je krok výpočtu. Z toho vychází, že v kroku n může být vygenerováno 0 až M tokenů. RNN-T obsahuje enkodér generující postupně vysokoúrovňové příznaky $\mathbf{f}_n$, predikční síť (*prediction network*) generující příznaky $\mathbf{g}_n$ se znalostí předchozích výstupních tokenů $(y_1, \ldots, y_{m-1})$ a společnou síť (*joint network*) kombinující tyto příznaky. Objektivní funkce je v tomto případě taktéž totožná. $P(\mathbf{y}|\mathbf{X})$ podobně jako v případě CTC uvažuje všechna zarovnání a je definována jako

$$P(\mathbf{y}|\mathbf{X}) = \sum_{\mathbf{a} \in A^{-1}(\mathbf{y})} P(\mathbf{a}|\mathbf{X}), \tag{3.22}$$

kde $A^{-1}(\mathbf{y})$ je n-tice možných zarovnání výstupní sekvence.



(a) CTC (b) RNN-T (c) AED

Obrázek 3.3: Příklady aktuálně nejčastěji používaných architektur pro rozpoznávání řeči.

Představené architektury jsou zobrazeny na obr. 3.3. Kvalita systému pro ASR je nejčastěji vyhodnocována pomocí metriky *Word Error Rate* – WER, případně *Character Error Rate* – CER. Tyto metriky jsou definovány jako

$$\text{WER} = \frac{S + D + I}{N}, \tag{3.23}$$

$$\text{CER} = \frac{S' + D' + I'}{N'}, \tag{3.24}$$

kde N je počet slov reference, S počet substitucí, D počet referenčních slov, jež se nenachází v hypotéze a I je počet slov, která se nenachází v referenci. Obdobně jsou definovány S', D', I', N', avšak tyto počty vycházejí z počtu znaků.

3.5 Klasifikace sentimentu

Jednou z disciplín zpracování přirozeného jazyka (*Natural Language Processing* – NLP) je detekce sentimentu. Spočívá v rozpoznávání kladného nebo záporného postoje vůči danému podnětu, který může být v některých případech nevědomý. Základním přístupem je klasifikace polarity daného textu na úrovni dokumentu, věty nebo tokenu do tří tříd – pozitivní, negativní, nebo neutrální. Pokročilá klasifikace se může zabývat emocionálními stavy, jako je radost, hněv, znechucení, smutek, strach a překvapení. Obecně lze rozlišit přístupy založené na strojovém učení a přístupy založené na lexikonech [126].

Nejlepší výsledky jsou aktuálně získávány s modely založenými na myšlence Attention mechanizmu a architektuře Transformer. Modely vycházející především z architektury a stylu trénování BERT dosahují velmi vysoké přesnosti klasifikace sentimentu recenzí v anglickém jazyce[1].

Systémy jsou nejčastěji vyhodnocovány přesností (*accuracy*), tedy podílem správně klasifikovaných prvků vůči celkovému počtu prvků. Velmi často jsou taktéž metody vyhodnocovány podle macro F1 skóre[2].

Problémem automatického vyhodnocení je však fakt, že v některých případech se ani lidští hodnotitelé nedokáží shodnout[3]. Systémy nemusí nutně dosahovat přesnost nad 95 %, aby je bylo možné považovat za dobré. Z tohoto důvodu je v některých případech uváděna taktéž Top-N *accuracy*.

3.6 Klasifikace typů terapeutických intervencí

Další z doménově specifických úloh je automatická detekce typů použitých terapeutických intervencí – souvislých, obsahově soudržných úseků řeči terapeuta, obvykle v délce jedné či několika vět, v lingvistice označovaných jako promluva. Klasifikace terapeutových či klientových promluv je oblastí, která díky možnostem strojového učení nově nabývá na významu [20, 36, 38]. Zatímco výše zmíněné typy analýz jsou poměrně nespecifické a lze je aplikovat na analýzu jakékoliv řeči, v rámci projektu DeePsy byl navržen oborově specifický klasifikačního protokol [86]. Pracuje s následujícími kategoriemi a podkategoriemi:

1. dotazování – divergentní, konvergentní
2. interpretace
3. reflexe – parafrázující reflexe, obohacující reflexe
4. potvrzování – procesové potvrzení, empatické ujištění, normalizace, empowerment
5. informování – edukace, perspektiva druhých
6. terapeutova direktivita – rada/úkol, procesové vedení
7. sebeodhalení
8. konfrontace

[1] https://paperswithcode.com/task/sentiment-analysis

[2] aritmetický průměr mezi F1 skóre příslušných tříd

[3] některé studie dokonce uvádí, že k tomu dochází až v 20 % případů – https://mashable.com/archive/sentiment-analysis/

9. ostatní – small talk, nezařaditelné, nesrozumitelné

Definice příslušných kategorií a podkategorii spolu s příklady se nachází v přiloze A. Jelikož se jedná o klasifikační úlohu do několika kategorií současně (*Multi Label Classification* – MLC [14]). Modely jsou nejčastěji evaluovány podle průměru (micro/macro) přesnosti (*accuracy*), preciznosti (*precision*), senzitivity (*recall*) či F1 skóre mezi kategoriemi.

Kapitola 4

Data

V rámci této kapitoly jsou čtenáři postupně představeny datové korpusy využité v rámci mé práce. Nejprve jsou popsány řečové korpusy využité k trénování systému pro řečové úlohy, dále jsou představeny textové korpusy určené pro trénování jazykových modelů. Na konci této kapitoly jsou popsána data využitá pro trénování a evaluaci systému pro cílové úlohy jako je klasifikace sentimentu či terapeutických intervencí.

4.1 Řečové korpusy

V této sekci jsou představeny řečové korpusy využité pro trénování systému pro VAD, diarizaci, detekci překrývající se řeči a ASR. Nejdříve je představen testovací korpus, následován postupným představením datasetů využitých pro učení pod vlastním dohledem a pro cílové dotrénování modelů.

4.1.1 DeePsyTest

Prezentované experimenty pro diarizaci, detekci skoků do řeči a automatického rozpoznání řeči popsané v následující kapitole byly evaluovány na testovací sadě DeeePsyTest, která byla částečně anotována v rámci této práce. Tato datová sada se skládá z 11 online, pěti sezení zaznamenaných na mobilní telefon a 32 psychoterapeutických sezení nahraných na diktafon ZOOM H2n[1]). Data byla anotována pomocí nástroje SpokenData[2]. Anotace obsahují textové přepisy vztahující se k příslušným časovým značkám a mluvčím. Taktéž jsou vyznačeny segmenty, kdy dochází k překrývající se řeči obou mluvčích. Tato data jsou exportována v podobě XML souboru.

Aby bylo možné zachytit co největší variabilitu mluvčích a prostředí, byly anotovány pouze části nahrávek o délce 5 minut. Celkově bylo extrahováno 48 úseků majících délku od 4 do 7 minut, což celkově činí 3,4 hodin mluvené řeči v rámci 4,1 hodin anotovaných nahrávek. Tyto segmenty byly zvoleny tak, aby pokrývaly jak začátek sezení, tak jeho průběh i konec. Zároveň byly účelně anotovány segmenty, ve kterých původní diarizační systém detekoval vysoký výskyt změn mluvčích a nebo byla dosažena nízká věrohodnost. Problematické úseky vhodné pro anotaci byly taktéž zvoleny odposlechem vybraných nahrávek.

V anotovaných nahrávkách je velmi patrný výskyt nespisovné češtiny, jakož i citelný vliv teritoriálního dialektu. V přepisech jsou velmi hojně zastoupena výplňová slova. Ně-

[1] https://zoomcorp.com/en/us/handheld-recorders/handheld-recorders/h2n-handy-recorder/
[2] https://www.spokendata.com/

které nahrávky obsahují vysoký podíl překrývající se řeči, jakož i poslechově velmi podobné jedince, a jsou vhodným zdrojem pro evaluaci výše zmíněných systémů.

4.1.2 DeePsyTrain

DeePsyTrain je datová sada 9 nahrávek o celkové délce 7,6 hodiny obsahující 5,5 hodiny řeči. Jedná se o celkově 5,2 tisíce segmentů. Zastoupení klienta a terapeuta je v poměru 2,6:1. Nahrávky byly taktéž nahrány na diktafon ZOOM H2n. Tato datová sada byla anotována stejným principem jako DeePsyTest. Navíc však obsahuje anotace typů terapeutických intervencí viz sekce 3.6.

4.1.3 DeePsyUnsupervised

Pro předtrénování či postupné předtrénování (*continued pretraining*) řečových modelů byla vytvořena interní trénovací sada DeePsyUnsupervised. Obsahuje 541 psychoterapeutických sezení mezi 37 mluvčími (10 terapeutů a 27 klientů). Celkově se jedná o 406 tisíc segmentů odpovídajících 696 hodinám řeči automaticky extrahovaných pomocí PyanNet[17] VAD systému natrénovaného v sekci 5.1.

4.1.4 BISON

Jedná se o databázi 22 hodin telefonních hovorů ze zákaznických call center vytvořených v rámci projektu BISON [151]. Jedná se o 18 tisíc segmentů v rozmezí 0,2 až 27,5 sekundy s průměrnou délkou 5s.

4.1.5 SpeeCon

SpeeCon je databáze řeči shromážděná v rámci projektu financovaného Evropskou komisí „Speech Driven Interfaces for Consumer Applications“. Databáze obsahuje 550 relací, které byly zaznamenány ve čtyřech různých prostředích: kancelář, zábava, veřejné místo, auto. Mluvčí byli vybráni s ohledem na dosažení co nejvyššího pokrytí z hlediska pohlaví, věku a dialektů mluvčích. Obsah korpusu je rozdělen do čtyř hlavních částí: volné spontánní položky, elicitované spontánní položky, čtené řeči a klíčových slov [45].

4.1.6 Temic

Temic je sbírka českých řečových dat zahrnující 710 mluvčích, která byla shromážděna pro účely TEMIC Speech Dialog Systems GmbH in Ulm na ČVUT v Praze ve spolupráci s Vysokým učením technickým v Brně a Západočeskou univerzitou v Plzni. Pokrytí mluvčích a obsah jsou velmi podobné k databázi SpeeCon. Nahrávky byly pořízeny v autě za různých podmínek a v různých situacích (např. zapnutý motor, vypnutý motor, zabouchnutí dveří, zapnuté stěrače atd.) [45].

4.1.7 ParCzech

ParCzech 3.0 je řečový korpus českých parlamentních projevů z Poslanecké sněmovny Parlamentu České republiky, které se konaly od 25. listopadu 2013 do 1. dubna 2021. V korpusu jsou zachovaná dostupná metadata jako identita mluvčího, pohlaví, webové odkazy doplněné o automatické morfologické a syntaktické anotace a rozpoznávání pojmenovaných entit

(*Named Entity Recognition* – NER). Dataset obsahuje celkově 154 tis. promluv 474 mluvčích o celkové délce 1332 hodin [15].

4.1.8 Common Voice

Mozzila Common Voice je multilinguální korpus přepsané řeči primárně vytvořený za účelem automatického rozpoznávání řeči. Může však být užitečný i v jiných oblastech, jako je např. identifikace jazyka. Korpus je založený na myšlence „crowdsourcingu“, kdy příslušní uživatelé nahrávky anotují a zároveň validují [2]. Dataset ve verzi 13.0 obsahuje 17,7 tis. validovaných hodin ve 108 jazycích. Česká část tohoto datasetu obsahuje 256 anotovaných hodin nahraných 876 mluvčími, ze kterých je 73 validovaných[3].

4.1.9 VoxPopuli

VoxPopuli je rozsáhlý vícejazyčný korpus obsahující 400 tisíc hodin neanotovaných řečových dat ve 23 jazycích. Dataset byl publikován jako nejrozsáhlejší dataset svého druhu. Dataset navíc obsahuje 1,8 tis. hodin přepsaných projevů ve zdrojovém jazyce s příslušnými překlady až do 15 dalších jazyků. Celkově tvoří 17,3 hodin data. Data jsou extrahována ze záznamů jednání na půdě Evropského parlamentu z let 2009-2020. Česká část obsahuje 18,7 tisíc vzorků o celkové délce 62 hodin [136].

4.1.10 ASRCorpora

Je pracovní jméno korpusu určeného pro trénování modelů pro rozpoznání řeči v češtině. Skládá se z dat pocházejících z výše zmíněných datových sad SpeeCon, Temic, BISON, ParCzech, Common Voice a DeePsyTrain doplněných o 143 tis. interních dat v poměru viz tabulka 4.1. Trénovací data pro rozpoznání řeči bylo nejdříve nutné pozbavit znaků, které se nevyskytují ve výstupní abecedě modelu. Z tohoto důvodů byly anotace nahrávek předzpracovány. Nejdříve byla normalizovaná diakritická znaménka jako `èäïłçüöńľ`, dále byla odstraněna interpunkční a jiná znaménka zapsaná dále ve formě regexu `[\,\?\.\!\-\;\:\"‘´¨\’]`. Další znaky jako `[§\%/+−]|c°|°c|<unk>|\d` byly nahrazeny tokenem `[UNK]`. Taktéž byly namapovány příslušné tokeny reprezentující váhání, případně hláskování na tokeny `[HES]` a `[SPELL]`. Nahrávky byly dále převzorkovány na frekvenci 16 kHz a z nahrávek byly vyextrahovány příslušné segmenty podle jejich časových značek. Aby bylo možné modely trénovat s dostatečnou velikostí batche, byly odstraněny segmenty delší než 20 s a kratší než 0,1 s. Tímto vznikla datová sada 754,4 tis. trénovacích vzorků, 80,2 tis. validačních, 5,2 tis. doménových vzorků určených pro trénování a 3,4 tis. testovacími vzorky.

4.2 Textové korpusy

Pro trénování jazykových modelů a předtrénování modelů pro navazující úlohy byla vytvořena obsáhlá textová datová sada kombinující data z několika zdrojů. Tabulka shrnuje použité datasety, počet jejich vzorků a zdrojovou doménu příslušných dat. Dataset je dále referován jako LMCorpora. Obsahuje celkově 77,39 milionů vzorků – souvětí. Níže jsou popsány datové sady, jež byly vytvořeny nebo alespoň modifikovány v rámci této práce. Některé z nich byly začleněny do LMCorpora, ostatní posloužily pro trénování modelů pro

[3] https://commonvoice.mozilla.org/en/datasets

Tabulka 4.1: Tabulka řečových počtu vzorků příslušných datasetů, jež byly zakomponovány do trénovacího datasetu AsrCorpora.

Zdroj	Počet vzorků [tis.]	Procentuální podíl [%]
SpeeCon	127	16,76
Temic	276	36,35
BISON	16	2,16
Common Voice	12	1,57
ParCzech	180	23,69
Interní data	143	18,78
DeePsyTrain	5	0,67

navazující úlohy. Dataset LMCorpora byl navíc doplněn o 83 tis. vzorků textových dat od partnerů projektu dále referovaných jako „Interní data“. Perplexita jazykových modelů byla evaluována na extrahovaných textových přepisech datasetu DeePsyTest. DeePsyASR obsahuje všechny dostupné textové přepisy terapeutických sezení (+10 sezení oproti DeePsyTrain) projektu DeePsy. Některá ze sezení nemohly být využita pro trénování řečových modelů, jelikož jejich časové značky byly posunuté a vyžadovaly další manuální opravy.

Tabulka 4.2: Tabulka textových datasetů, které byly využity pro trénování jazykových modelů.

ID	Zdroj	Počet souvětí	Doména	Přeloženo
1	The Prague Dependency Treebank 2.0 [10]	118 tis.	novinové články	
2	OpenSubtitles [82]	71,10 mil.	titulky	
3	Wikipedie [116]	2,940 mil.	odborné články	
4	AnnoMI	9,4 tis.	poradenské dialogy	✓
5	AlexanderStreet	54,9 tis.	poradenské dialogy	✓
6	Archív České televize[4]	2,19 mil.	titulky	
7	DeePsyASR	16 tis.	psychoterapeutická sezení	
8	TwitterEmotions	879 tis.	emočně orientované tweety	
9	Interní data	82,9 tis.	psych. sezení a rozhovory	
10	Diplomové práce[5]	6,1 tis.	rozhovory	

4.2.1 AnnoMI

Jedná se o 133 přepsaných motivačních rozhovorů (*Motivational Interviewing* –MI) rozdělených do dvou tříd podle kvality s kategoriemi inspirovanými kódováním MISC [122] a MITI [93]. Všechny nahrávky byly extrahovány z volně dostupných zdrojů a anotovány se souhlasem autorů. V době experimentů byla použita základní verze tohoto datasetu, jež obsahuje 9,7 tis. promluv anotovaných do 4 kategorií v případě, že se jedná o promluvu terapeuta – reflexe, otázka, vstup a jiné. V případě, že se jedná o promluvu klienta, jsou

[4] https://www.ceskatelevize.cz/

[5] Práce byly vyhledávány podle klíčových slov souvisejících s dialogy a rozhovory na portálu https://theses.cz/. Byly extrahovány pouze přepisy rozhovorů.

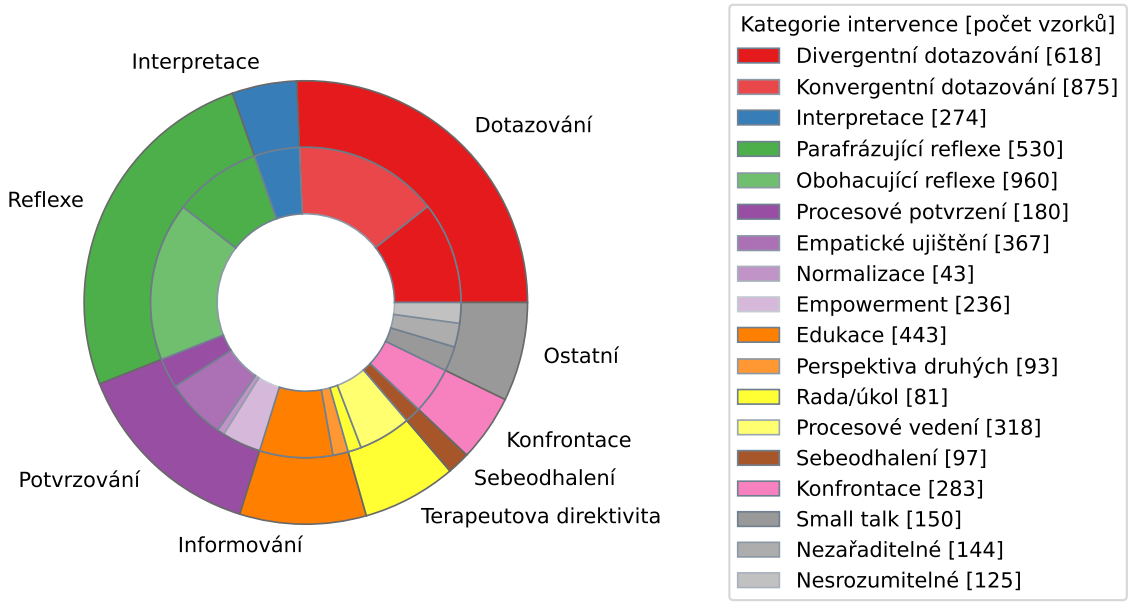
anotovány kategorie – změna, neutrální a beze-změny. Rozšířená verze datasetu obsahuje navíc podkategorie terapeutových promluv [143, 142]. Dataset byl automaticky přeložen pomocí online překladače[6] CUBBITT [107].

4.2.2 AlexanderStreet

Jedná se o databázi[7] 500 demonstračních videí terapie pokrývajících 245 různých témat. Sezení jsou vedena 146 terapeuty a je demonstrováno více než 100 různých terapeutických přístupů. Jelikož se jedná o databázi v anglickém jazyce, byly taktéž extrahována pouze textová data, která byla následně přeložena pomoci nástroje CUBBITT.

4.2.3 DeePsyInterventions

Je textová sada terapeutických intervencí rozdělených do kategorií dle klasifikačního protokolu navrženého v rámci projektu DeePsy [86]. V první fázi anotování sada obsahovala 7583 promluv, z čehož 2525 promluv patřilo terapeutovi a tedy bylo anotováno. Sada byla v další fázi doplněna o dalších 6821 promluv, ze kterých 1896 náleží terapeutovi. Příslušné promluvy mohou být přiřazeny do několika z 18 kategorií současně. Distribuce kategorií a podkategorií rozšířené sady je zobrazena na obrázku 4.1



Obrázek 4.1: Distribuce kategorií v rámci datasetu terapeutických intervencí vytvořeného v rámci projektu DeePsy. Vnější kruh symbolizuje obecné třídy jako dotazování, interpretace, načež vnitřní podtřídy jako divergentní, či konvergentní dotazování viz příloha A. V legendě jsou viditelné počty vzorků příslušných kategorií.

[6] https://lindat.mff.cuni.cz/services/translation/
[7] https://alexanderstreet.com/products/apa-psyctherapy%C2%AE

4.2.4 SentimentCorpus

V českém prostředí existují tři datové sady pro klasifikaci sentimentu [49]. Jedná se o datasety recenzí extrahovaných ze sociální sítě Facebook[8], Česko-Slovenské filmové databáze – ČSFD[9] a obchodního portálu MALL.cz[10]. Velikosti datasetů jsou detailněji popsány v tabulce 4.3. Data ve zmíněných datasetech jsou rozdělena do tří tříd – pozitivní, neutrální a negativní (případně bipolární).

Tabulka 4.3: Datové sady pro detekci sentimentu v českém jazyce, počet vzorků pro příslušné třídy a doména, ze které data pocházejí. Dataset z prostředí Facebooku byl anotován dvěma nezávislými anotátory, se shodou mezi anotátory $\kappa = 0,66$. Pro dataset ČSFD a MALL.CZ byly třídy automaticky odvozeny z uživatelských hodnocení [49].

Zdroj	Velikost	Doména
Facebook	2,6k pozitivní 5k neutrální 2k negativní	recenze značek, výrobků a služeb
ČSFD	31k pozitivní 31k neutrální 30k negativní	recenze filmů
MALL.CZ	103k pozitivní 32k neutrální 10k negativní	recenze produktů

4.2.5 TwitterEmotions

Problémem výše popsaných sad pro klasifikaci sentimentu je však to, že všechny zmíněné sady jsou založeny na recenzích. Ty však nepříliš odpovídají zkoumané doméně psychoterapeutických sezení a specifikům mluvené češtiny. Z tohoto důvodu byl v rámci této práce vytvořen dataset emočně orientovaných tweetů (krátkých příspěvků z sociální sítě Twitter) v českém jazyce.

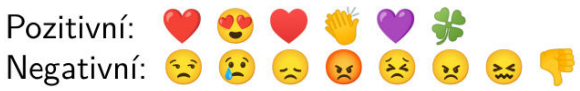
V prostředí sociální sítě Twitter lze rozlišit dva přístupy pro automatického přiřazení do tříd – podle použitých hashtagů (slovo nebo fráze začínající znakem „#“, nabývající formu klíčového slova) nebo podle použitých emotikonů [76]. Jelikož přiřazení hashtagů do příslušných tříd by vyžadovalo robustnější analýzu tweetu, byl dataset vytvořen s využitím emotikonů.

V rámci počáteční analýzy byla zkoumána množina nejpoužívanějších emotikonů na sociální síti Twitter. Pro tyto účely byla využita služba emojitracker[11], která sleduje výskyt emotikonů v tweetech v reálném čase. Z množiny těchto emotikonů bylo vybráno 31 emotikonů, které se na první pohled jevily jako patřící do kategorie pozitivní nebo negativní. Jelikož význam příslušných emotikonů je často velmi subjektivní byl taktéž kladen důraz na jejich textový popis na službě emojipedia[12].

[8] https://www.facebook.com/
[9] https://www.csfd.cz/
[10] https://www.mall.cz/
[11] https://emojitracker.com/
[12] https://emojipedia.org/

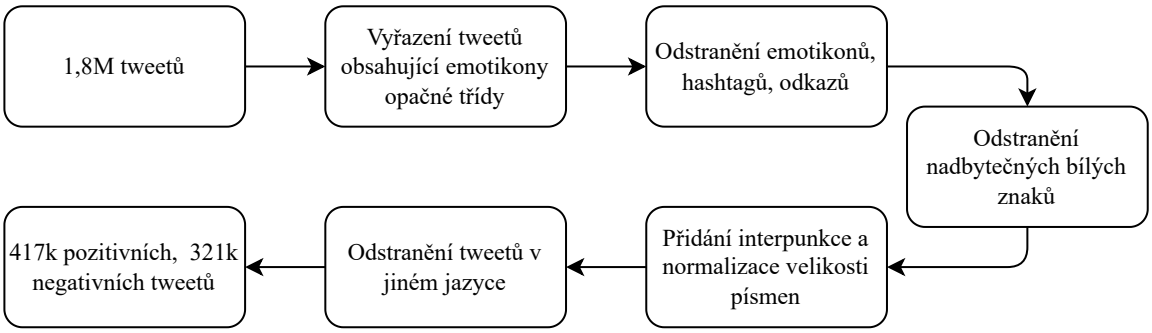
Pro každý emotikon ze zmíněné množiny 31 kandidátních emotikonů bylo staženo 200 náhodných tweetů prostřednictvím endpointu „Full-archive search“ služby Twitter API v2[13]. Tato data byla přiřazena dvěma nezávislými anotátory do zmíněných tříd a navíc také do třídy „nelze posoudit“. Emotikony, které obsahovaly více než 8% tweetů nepatřících do kategorie reprezentované daným emotikonem, byly vyřazeny. Takto vznikla podmnožina 14 emotikonů zobrazena na obr. 4.2.

Pro každý emotikon bylo staženo 200 tisíc tweetů v českém jazyce, duplicitní tweety byly odstraněny, retweety byly filtrovány, jakož i tweety obsahující přiložená média nebo odkazy. Byly zahrnuty i odpovědi na předchozí tweety, aby bylo co nejblíže simulováno prostředí konverzace.



Obrázek 4.2: Množina 14 emotikonů, jež byly využity pro automatickou tvorbu datasetu emočně orientovaných tweetů.

Tímto způsobem bylo získáno 1,8 milionu anonymizovaných tweetů. V surové podobě však nebyly přímo vhodné pro dotrénování jazykového modelu, a proto byla získaná datová sada podrobena předzpracování. Proces je zobrazen na obr. 4.3. Ze získané množiny byly odstraněny tweety, které obsahovaly emotikony náležící do opačné třídy sentimentu. Následně byly odstraněny samotné emotikony, hashtagy, odkazy a jiné nevhodné znaky. Tweety byly následně normalizovány do větné podoby, byla přidána interpunkce, byly odstraněny nadbytečné mezery a případně opraveno užití velkých písmen. Konečně byly odstraněny tweety v cizím jazyce s využitím jednoduchého modelu pro identifikaci jazyka (Language Identification – LID) Compact Language Detector v3[14]. Takto vznikla sada 417 tisíc pozitivních a 321
tisíc negativních tweetů, která byla pro účely experimentů doplněna o neutrální data pocházející z článků z české Wikipedie.



Obrázek 4.3: Schematický graf předzpracování a očištění datové sady stažené z Twitteru.

[13] https://developer.twitter.com/en/docs/twitter-api/
[14] https://github.com/google/cld3

Kapitola 5

Experimenty

V rámci této kapitoly jsou čtenáři představeny experimenty, které byly provedeny k natrénování modelů pro postupné zpracování nahrávky a získání příznaků popisujících psychoterapeutické sezení. Dílčí experimenty jsou rozděleny do sekcí odpovídajících příslušným úlohám z oblasti zpracování řeči a přirozeného jazyka. Sekce jsou uspořádány v chronologickém pořadí, v jakém je nahrávka zpracována v systému DeePsy.

5.1 Detekce řečové aktivity

Prvním, avšak velmi stěžejním krokem pro vybudování systému pro extrakci komplexních příznaků, je extrakce řeči z nahrávek. Je k tomu využit systém pro detekci řečové aktivity – VAD. Původně zaintegrovaný systém založený na dvouvrstvé NN natrénované na datech z projektu BABEL [105], dále referovaný jako *vad_baseline*, byl příliš agresivní a některé tiché segmenty v nahrávkách nebyly rozpoznány jako řeč. Doladění prahu nebylo dostačující, a tedy byly provedeny experimenty s cílem zlepšení úspěšnosti s různými architekturami pro detekci řečové aktivity. Tabulka 5.1 shrnuje učiněné experimenty, jež byly provedeny s límcem o velikosti 0 ms a 250 ms. Límec o velikosti 250 ms byl využit, jelikož i samotná reference není přesná a jedná se o standardní hodnotu při vyhodnocování diarizačních, či VAD systémů.

Byly analyzovány a dotrénovány různé architektury, počínajíc Směsicí Gaussovských rozložení trénované EM algoritmem [13] na energiích segmentu až po komplexnější architektury jako CRDNN [115] zahrnující konvoluční a LiGRU bloky [114]. Experimenty vyžadovaly integraci datové a evaluační sady do toolkitů NeMo[1], SpeechBrain[2] a pyannote[3] a tvorbu skriptů pro dotrénování příslušných modelů. Nejlepší výsledky byly získány adaptací PyanNet [17] na trénovací sadě DeePsyTrain. Původní model[4] byl dotrénován na datové sadě DeePsyTrain po 10 epoch s optimalizátorem Adam($\gamma = 0,001$, $\beta_1 = 0,9$, $\beta_2 = 0,999$) [65] a velikostí batche 128 vzorků. Došlo k relativnímu zlepšení *detection error rate* o 49% vůči *vad_baseline*. VAD je prvním systémem pro zpracování nahrávky v systému DeePsy viz obr. 2.1.

[1] https://github.com/NVIDIA/NeMo
[2] https://github.com/speechbrain/speechbrain
[3] https://github.com/pyannote/pyannote-audio
[4] https://huggingface.co/pyannote/voice-activity-detection

Tabulka 5.1: Chybovost systémů pro detekci řečové aktivity vyhodnocena na testovací sadě DeePsyTest s využitím metrik DER – Detection Error Rate, FA – False alarm a M – Miss rate.

Model	Trénovací data	Límec 0 ms			Límec 250 ms		
		DER [%] ↓	FA [%] ↓	M [%] ↓	DER [%] ↓	FA [%] ↓	M [%] ↓
vad_baseline [105]	BABEL [63]	10,63	9,35	1,29	5,87	4,67	1,20
Energetické GMM	-	20,05	1,34	18,71	16,84	0,81	16,03
GPVAD [33]	AudioSet [43]	8,92	**1,03**	7,90	6,05	**0,27**	5,78
CRDNN [115]	Libriparty [115]	14,59	13,68	0,90	10,58	9,78	0,79
PyanNet [17]	DIHARD3 [120, 119], VoxConverse [26] a AMI [21]	12,38	3,31	9,08	8,56	1,50	7,06
MarbleNet [61]	Google Speech Commands	27,00	15,58	11,42	24,00	11,92	12,08
multilingual MarbleNet [61]	Dataset V2[138]	14,46	10,91	3,56	10,38	7,22	3,16
PyanNet	DeePsyTrain	**6,61**	4,22	2,39	**2,99**	1,38	1,60
MarbleNet		12,74	12,42	**0,31**	8,54	8,25	**0,29**
CRDNN		9,50	7,42	2,07	5,58	4,00	1,58

5.2 Diarizace

V rámci psychoterapeutických seancí nejčastěji vystupují pouze dva mluvčí, klient a terapeut, což značně zjednodušuje problém diarizace, jelikož je předem známý počet mluvčích. Původně navržený systém [106] na bázi směsi Gaussovských komponent (*Gaussian Mixture Model* – GMM [13]) a Mel-frekvenčních cepstrálních koeficientů (*Mel-Frequency Cepstral Coeffiecients* – MFCCs [30]) však nedosahoval dostačující kvality a bylo nutné jej nahradit robustnějším systémem.

V rámci testovací sady DeePsyTest lze rozlišit dva scénáře sezení – prezenční a online. V případě online terapie se mezi kanály nenachází žádný přeslech a v navrženém systému je pouze spuštěna detekce řečové aktivity na příslušných kanálech, čímž je získána přesná informace o tom, kdo kdy mluví.

V případě prezenčních sezení nahrávaných na diktafón ZOOM H2n je audio nahráno na dva kanály. Ty však obsahují velmi vysoký přeslech, který se nepodařilo automaticky odstranit. V tomto případě tedy není možné provést diarizaci pouhým spuštěním VAD na daných kanálech. Obdobně nahrávky nahrané na mobilní telefon obsahují pouze jeden kanál. Pro experimentální účely byly tedy všechny nahrávky z testovací sady DeePsyTest smíchány do jednoho kanálu a převzorkovány na frekvenci 16 kHz.

Pro prvotní analýzu byl zvolen systém fungující na bázi Bayesovského skrytého Markovova modelu (*Bayesian Hidden Markov Model* – BHMM [44]) shlukujícího x-vektory [129] extrahované pomocí sítě ResNet101 [50]. Tento systém, pojmenovaný jako VBx [75], však předpokládá obecně neznámý počet mluvčích, a proto byl upraven tak, aby provedl diarizaci právě mezi dva mluvčí a následně jim přiřadil identitu.

Systém byl postupně evaluován ve třech prostředích – s límcem 0 ms, s límcem 0 ms včetně odstranění překrývající se řeči a s límcem 250 ms včetně odstranění překrývající se řeči. Chybovost systému je zobrazena v tabulce 5.2. Ve 47 ze 48 testovacích sezení systém s výchozími parametry detekoval dva mluvčí. Bohužel při inferenci na celých sezeních (délka $\approx 50\ min$) docházelo ke konvergenci k variabilnímu počtu mluvčích.

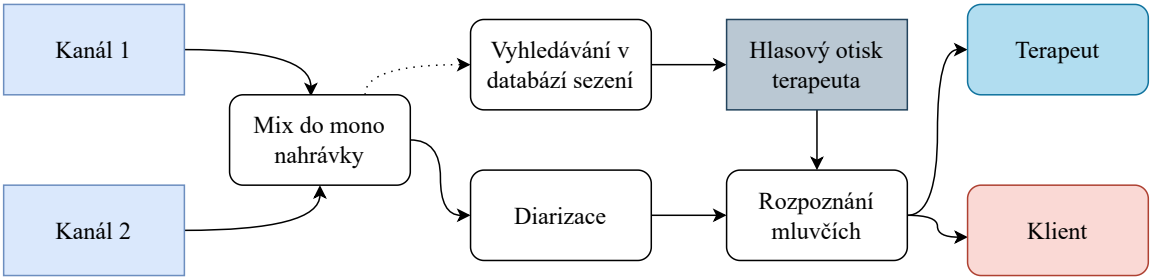
Pro účely konvergence ke dvěma mluvčím bylo dále experimentováno se systémem VBx a jeho parametry. Analyzované přístupy a jejich kvalita jsou zobrazeny v tabulce 5.2. Položka „VBx“ uvádí výsledky původního systému s výchozími parametry. Nejsou však žádným způsobem penalizovány výstupy, kdy je počet mluvčích různý od 2. „VBx regularizace počtu mluvčích“ je metoda, kdy je nastaven maximální počet mluvčích na 2 a systém má

nastavený regularizační koeficient Fa $= 0,001$, čímž je minimalizována pravděpodobnost spojení do jediného shluku. Bohužel takto navržena regularizace však drasticky zvýšila chybovost systému.

Bylo tedy dále uvažováno, jak využít znalosti o fixním počtu mluvčích. Byl implementován skript pro automatickou extrakci hlasových otisků terapeutů v podobě x-vektorů při registraci do systému. Tímto způsobem je předem známa identita terapeuta, která dovoluje při konvergenci do více než dvou shluků spočítat podobnost agregovaných vektorů reprezentujících shluky s řečovým otiskem terapeuta. Regularizační koeficienty byly nastaveny na hodnoty $F_a = 0,3, F_b = 2$. Pokud je splněna podmínka dostatečně vysokého skóre logaritmické věrohodnosti hypotézy, že příslušné vektory pocházejí od stejného mluvčího $\rho > 0$, dochází k empirickému rozpoznání shluků reprezentujících klienta a terapeuta – shluky jsou seřazeny podle nejvyššího skóre, ten s nejnižším je prohlášen za klienta. Následně je první shluk s pozitivním skóre prohlášen za terapeuta. Pokud je nalezen shluk s vyšším skóre mající alespoň 4krát (tento koeficient byl určen experimentálně) více segmentů, je uvážen jako nový shluk reprezentující terapeuta. Následně jsou spočteny podobnosti mezi ostatními shluky a shluky reprezentující terapeuta a klienta. Shluk mající nejvyšší podobnost s jedním z těchto shluků je zahrnut do nového společného shluku a celý proces pokračuje znova do doby, než vzniknou dva výsledné shluky. Celý přístup je naznačen na obr. 5.1 a v tabulce je označen identifikátorem „VBx + identita terapeuta“. „VBx + identita terapeuta + resegmentace“ uvažuje navíc resegmentaci po konvergenci do 2 shluků.

Dále byly provedeny experimenty s předtrénovaným *end-to-end* diarizačním systémem pyannote [16] postaveným na segmentaci PyanNet [17] a extrakci ECAPA vektorů [31]. Původní model natrénovaný na směsi dat z datasetů však nedosahoval lepší úspěšnosti než VBx. Systém byl tedy dotrénován na trénovací sadě DeePsyTrain s využitím optimalizátoru Adam($\gamma = 0,0001$, $\beta_1 = 0,9$, $\beta_2 = 0,999$) a velikostí batche 8 vzorků. Chybovost se podařilo snížit, avšak systém VBx byl ve všech evaluačních nastaveních lepší, a byl tedy zahrnut do výsledného systému.

Jak již bylo zmíněno, vstupní data byla nahrána na diktafón se dvěma kanály, ty však obsahují značný přeslech a jsou následně sloučeny do mono nahrávky. Tento předpoklad se jeví jako vhodný pro pokus o diarizaci s využitím separace mluvčích [94] do dvou kanálů a následné detekce řečové aktivity v příslušných kanálech. Může být předmětem budoucí analýzy.



Obrázek 5.1: Graf ilustrující zpracování stereo nahrávky prezenčního sezení v systému DePsy za účelem segmentace a rozpoznání mluvčích.

Tabulka 5.2: Chybovost diarizace závislá na použité metodě vyhodnocena na testovací sadě DeePsyTest s fixním modelem pro VAD PyanNet.

Metoda	Trénovací data	DER [%] ↓	False alarm [%] ↓	Miss rate [%] ↓	Confusion [%] ↓
Límec 0 ms, překrývající se řeč					
VBx [75]	-	16,57	3,73	9,84	3,00
pyannote [17]	DIHARD3 [120, 119], VoxConverse [26] a AMI [21]	17,95	3,51	11,22	3,21
pyannote	DeePsyTrain	18,23	3,30	10,83	4,10
VBx regularizace počtu mluvčích	-	19,65	3,67	9,94	6,04
VBx + identita terapeuta	-	16,97	3,67	9,94	3,56
VBx + identita terapeuta + resegmentace	-	16,68	3,67	9,94	3,08
Límec 0 ms, bez překrývající se řeči					
VBx	-	10,50	4,38	2,59	3,53
pyannote	DIHARD3, VoxConverse a AMI	14,17	4,14	6,25	3,78
pyannote	DeePsyTrain	12,31	3,88	3,59	4,83
VBx regularizace počtu mluvčích	-	14,05	4,32	2,61	7,11
VBx + identita terapeuta	-	10,89	4,32	2,61	3,96
VBx + identita terapeuta + resegmentace	-	10,56	4,32	2,61	3,62
Límec 250 ms, bez překrývající se řeči					
VBx	-	6,03	1,38	1,75	2,90
pyannote	DIHARD3, VoxConverse a AMI	9,71	1,94	4,26	3,52
pyannote	DeePsyTrain	7,65	1,15	2,24	4,26
VBx regularizace počtu mluvčích	-	9,77	1,35	1,74	6,68
VBx + identita terapeuta	-	6,44	1,35	1,74	3,35
VBx + identita terapeuta + resegmentace	-	6,10	1,35	1,74	3,01

5.3 Detekce překrývající se řeči

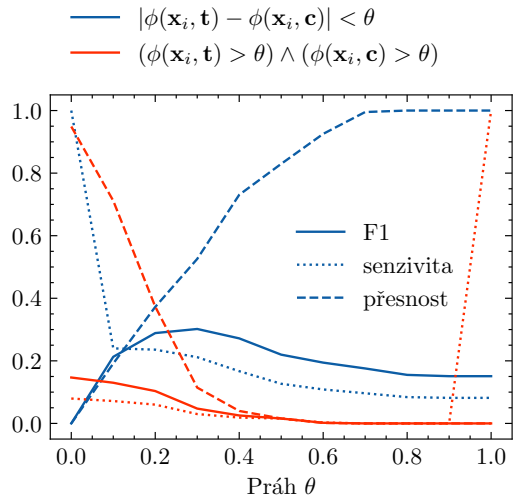
Jednou z navržených metrik popisujících psychoterapeutické sezení je překrývající se řeč. Prvotní sada experimentů, jejíž výsledky jsou zobrazeny v tabulce 5.3, vychází z předpokladu, že *embeddingy* (hlasové charakteristiky) extrahované ze segmentů nahrávky, kde hovoří oba mluvčí, by měly být podobné *embeddingům* reprezentujícím klienta a terapeuta. Naopak v případě segmentů, kde hovoří pouze jeden, by měla být podobnost vysoká pouze vzhledem k jednomu z mluvčích. Prvním pokusem o detekci překrývající se řeči bylo přímé využití referenční diarizace, tedy informace, kdy kdo mluví. Nejprve byly pro všechny nahrávky testovací sady DeePsyTest extrahovány ECAPA [31] *embeddingy*. Pro každý segment s_i testovací nahrávky byl extrahován *embedding* $\mathbf{x}_i \in \mathbb{R}^{192}$, kde $i \in \{0, \ldots, N\}$, kde N je počet segmentů dané nahrávky. Následně byl extrahován agregovaný *embedding* $\mathbf{c} \in \mathbb{R}^{192}$ reprezentující klienta a $\mathbf{t} \in \mathbb{R}^{192}$ reprezentující terapeuta ze segmentů přiřazených výlučně danému mluvčímu. Nechť $\phi(\mathbf{x}, \mathbf{y})$ je kosinová podobnost vektorů $\mathbf{x}$ a $\mathbf{y}$. Potom lze podobnost segmentu s_i vůči *embeddingu* terapeuta $\mathbf{t}$ vyjádřit jako $\phi(\mathbf{x}_i, \mathbf{t})$. S využitím tohoto vztahu byla navržena formulace detekce překrývající se řeči jako $(\phi(\mathbf{x}_i, \mathbf{t}) > \theta) \wedge (\phi(\mathbf{x}_i, \mathbf{c}) > \theta)$, kde θ je práh detekce.

Tabulka 5.3: Analýza úspěšnosti navržených metod pro detekci překrývající se řeči na testovací sadě DeePsyTest s využitím kosinové podobnosti vektorů charakterizujících daný segment nahrávky a vektorů příslušných mluvčích.

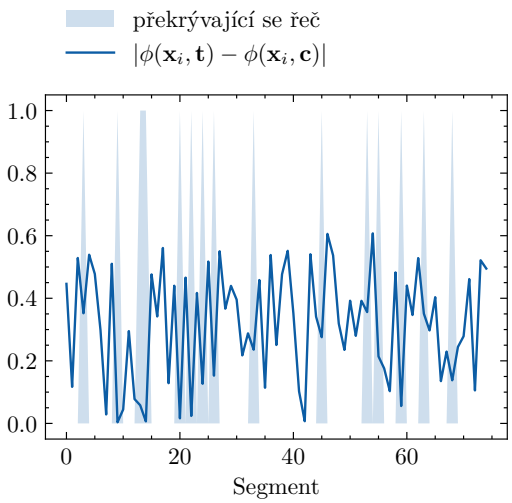
Metoda	θ	F1 ↑	Přesnost ↑	Senzitivita ↑
$(\phi(\mathbf{x}_i, \mathbf{t}) > \theta) \wedge (\phi(\mathbf{x}_i, \mathbf{c}) > \theta)$	$0,0$	$0,15$	$0,08$	$\mathbf{0,95}$
$\lvert\phi(\mathbf{x}_i, \mathbf{t}) - \phi(\mathbf{x}_i, \mathbf{c})\rvert < \theta$	$0,3$	$0,30$	$0,21$	$0,53$
$\psi(\mathbf{x}_{i-1}) > \frac{\psi(\mathbf{x}_i)}{\theta} < \psi(\mathbf{x}_{i+1})$	$1,3$	$\mathbf{0,40}$	$\mathbf{0,37}$	$0,43$

Postupně byly prohledány různé hodnoty prahu v rozmezí $\langle 0, 1\rangle$ s krokem $0,1$. Z tabulky je však patrné, že takto navržená metoda nefunguje. Jelikož nejlepší skóre bylo získáno označením téměř všech segmentů za překrývající se řeč. Po analýze podobností vektorů byla navržena nová metoda detekce jako $|\phi(\mathbf{x}_i, \mathbf{t}) - \phi(\mathbf{x}_i, \mathbf{c})| < \theta$. Tato metoda nutně nevyžaduje velkou podobnost obou vektorů, ale je zaměřena na rozdíl podobnosti. Je totiž pravděpodobné, že pokud segment obsahuje mluvu obou mluvčích, *embedding* může ležet v prostoru kolmo na oba vektory a podobnost tedy může být velmi nízká. Prohledáním různých prahů bylo dosaženo značného zlepšení F1 skóre na hodnotu $0,3$ viz obr. 5.2a. Získané skóre však bylo stále velmi nízké. Z tohoto důvodu byla provedena manuální analýza náhodné nahrávky z testovací sady viz obr. 5.2b. Po náhledu na obrázek byla navržena nová metoda $\psi(\mathbf{x}_{i-1}) > \frac{\psi(\mathbf{x}_i)}{\theta} < \psi(\mathbf{x}_{i+1})$, kde $\psi(\mathbf{x}_i) = |\phi(\mathbf{x}_i, \mathbf{t}) - \phi(\mathbf{x}_i, \mathbf{c})|$. Tato metoda předpokládá, že v případě překrývající se řeči je rozdíl podobnosti aktuálního vzorku mnohem nižší než rozdíl sousedních vzorků. V tomto případě θ plní funkci škálovacího koeficientu v rozsahu $\langle 1, 2\rangle$. Obdobně byly prohledány hodnoty koeficientu θ s krokem $0,1$ viz obr. 5.3. Z obrázku je viditelné, že úspěšnost detekce je sice stále nízká, avšak navržená metoda je nejlepší z navržených metod výše.

Z důvodu poměrně nízké úspěšnosti metod popsaných výše byly dále provedeny experimenty s předtrénovanou sítí PyanNet [17] pro detekci překrývající se řeči. Jedná se o totožnou síť z předešlé sekce 5.2. Jelikož je síť předtrénována na velkém objemu dat, byla nejprve provedena analýza různých prahovacích koeficientů sítě viz první část tabulky 5.4. Pro úplnost byly uváženy i optimální koeficienty nalazené přímo na testovací sadě. Nejlepší



(a) Analýza úspěšnosti detekce překrývající se řeči vzhledem k prahu θ a zvolené metodě detekce.



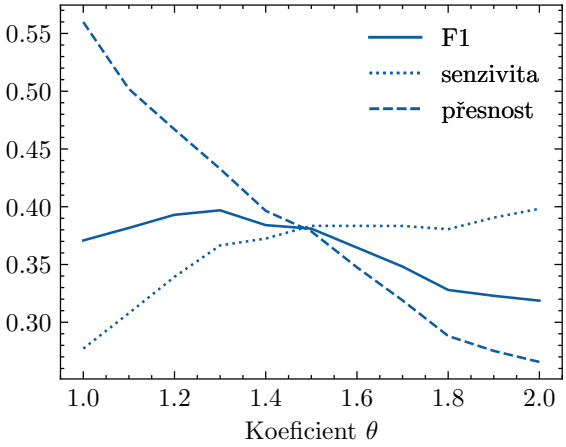
(b) Porovnání referenční anotace jedné z nahrávek testovací sady DeePsyTest vzhledem k absolutní hodnotě rozdílu podobností segmentu vůči mluvčím.

Obrázek 5.2: Experimenty provedené pro nalezení vhodného detektoru překrývající se řeči.

úspěšnost detekce překrývající se řeči byla dosažena s parametry nalazenými na datasetu DIHARD3 [120, 119]. Z tabulky je patrné, že vyhledání parametrů na trénovací sadě DeePsyTrain nevedlo k žádnému zlepšení. Důvodem je fakt, že v rámci anotace trénovací sady DeePsyTrain nebyl všemi anotátory dodržen anotační protokol a překrývající se řeč tak nebyla ve všech případech řádně označena. Což vyžaduje další kontrolu dat, která však doposud nebyla provedena. Zmíněný jev taktéž potvrzuje pokus o dotrénování sítě na těchto datech. Je viditelné markantní zhoršení. Z tohoto důvodu byl proveden následující experiment. Testovací sada DeePsyTest byla rozdělena na dvě poloviny. První polovina nahrávek posloužila k dotrénování sítě a druhá pro evaluaci. Uvedeným přístupem bylo získáno zlepšení úspěšnosti detekce na hodnotu F1 skóre 0,49. Tato hodnota však není zcela kompatibilní s experimenty uvažujícími pouhé doladění parametrů, jelikož testovací sada je odlišná. Ukazuje však na fakt, že s řádně anotovanými doménovými daty je možné docílit dalšího zlepšení detekčních vlastností. Model trénovaný na polovině testovacích dat byl zaintegrován do systému. K dalšímu zlepšení systému je nutné přeanotovat nahrávky z trénovací sady DeePsyTrain a natrénovat a vyhodnotit systém na větší sadě dat.

5.4 Rozpoznávání řeči

Klasifikace, či extrakce komplexnějších příznaků z dialogů vyžaduje kvalitní řečové a textové příznaky. Samotný automatický přepis řeči založený na hybridní architektuře CNN-TDNN-HMM [67] doplněné o n-gramový jazykový model však dosahoval poměrně vysoké chybovosti 28, 30% WER, z tohoto důvodu byly provedeny experimenty s modely založenými na architektuře Transformer, které jsou blíže popsány v sekci 2.6.



Obrázek 5.3: Analýza úspěšnosti detekce překrývající se řeči vzhledem ke koeficientu θ rovnice $\psi(\mathbf{x}_{i-1}) > \frac{\psi(\mathbf{x}_i)}{\theta} < \psi(\mathbf{x}_{i+1})$, kde $\psi(\mathbf{x}_i) = |\phi(\mathbf{x}_i, \mathbf{t}) - \phi(\mathbf{x}_i, \mathbf{c})|$.

Modely vycházející z architektury WAV2VEC2 byly doplněny o (klasifikační) lineární vrstvu viz obr. 5.4 velikosti 46 – počet znaků abecedy rozšířené o speciální tokeny `[UNK]`, `[HES]`, `[PAD]` a `[SPELL]`[5]. Dále byly analyzovány modely Whisper.

5.4.1 Analýza předtrénovaných modelů

Prvotní analýza vycházela z dotrénování veřejně dostupných modelů pro rozpoznávání řeči, či tvorby její reprezentace na datové sadě *ASRCorpora* bez doménové sady DeePsyTrain. Modely byly evaluovány na testovací sadě DeePsyTestV1. Pro prvotní analýzu byl zvolen multilinguální model XLS-R [4], seq2seq model Whisper [110] a model rodiny Wav2Vec2 ClTRUS [79] předtrénovaný na českých datech. Modely ClTRUS a XLS-R byly doplněny o klasifikační vrstvu viz výše. Architektura modelu Whisper byla ponechána beze změny. Aby byla prvotní analýza kompatibilní napříč modely, byly vybrány varianty „Whisper-small“ a „Whisper-base“ modelu Whisper. V případě XLS-R nejmenší varianta mající 300 milionů parametrů. Modely byly trénovány s optimalizátorem AdamW($\gamma = 0,00005$, $\beta_1 = 0,9$, $\beta_2 = 0,999$, $\lambda = 0,00$) [85] a velikostí batche 16 vzorků na grafikách NVIDIA A40[6]. V případě větších modelů byla využita akumulace gradientu.

V tabulce 5.5 jsou představeny názvy modelů s příslušným počtem parametrů a nejnižší dosaženou chybovostí v průběhu trénování. Z tabulky přímo vyplývá téměř dvojnásobná chybovost předtrénovaných modelů typu Transformer v porovnání s hybridním systémem. Dále je možné pozorovat fakt, že ačkoliv byl model ClTRUS trénován čistě na českých datech, viz sekce 2.6.4, jeho modelovací schopnosti nedosáhly úrovně multilinguálního modelu XLS-R-300m. Počet parametrů XLS-R-300m se zde jeví jako stěžejní. Průběh trénování ClTRUS a XLS-R-300m je zachycen na obr. 5.5a. Je nutno podotknout, že XLS-R-300m je nejmenší variantou modelu XLS-R, existují varianty s jednou a dvěma miliardami parametrů. Na rozdíl od ClTRUS a XLS-R modely Whisper-base a Whisper-small pochází z kategorie seq2seq. Modely tedy disponují explicitním jazykovým modelem, z tohoto dů-

[5]Token `[UNK]` v referenci označuje nerozpoznatelná slova, token `[HES]` je využit pro anotování neřečových příznaků jako povzdychnutí nebo zamlumlání, `[PAD]` je využit k zarovnání sekvencí do batche a tokenem `[SPELL]` jsou označeny segmenty, kdy mluvčí hláskuje.

[6]https://www.nvidia.com/en-us/data-center/a40/

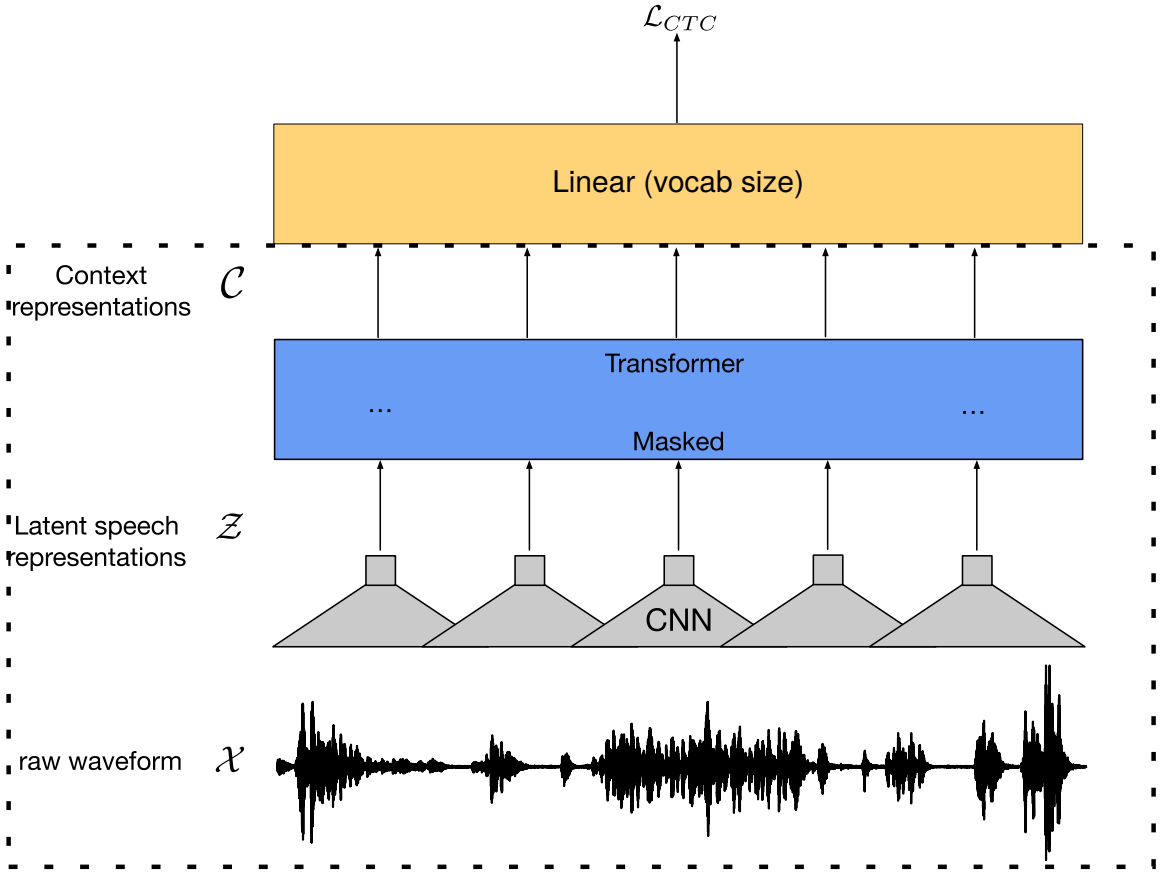
Tabulka 5.4: Úspěšnost detekce překrývající se řeči v závislosti na zvolených parametrech a natrénovaném modelu. Sloupec „onset“ odpovídá minimálnímu prahu pro začátek detekce překrývající se řeči. Jakmile věrohodnost segmentu nabude hodnoty nižší než „offset“, dochází k přepnutí detekce do režimu neaktivní. Detekované úseky jsou následně vyhlazeny pomocí minimální délky pro překrývající se řeč $\min_{on}$, respektive minimální mezery mezi překrývající se řečí $\min_{off}$. Všechny řádky, vyjma posledního, jsou evaluovány na testovací sadě DeePsyTest. Model natrénovaný na polovině sady DeePsyTest je vyhodnocen na druhé polovině této sady, jež nebyla zahrnuta mezi trénovací vzorky.

	Parametry				Metriky		
	On	Off	$\min_{on}$ [s]	$\min_{off}$ [s]	P ↑	S ↑	F1 ↑
Parametry doladěny na							
AMI Mix-Headset [21]	0,45	0,36	0,12	0,19	0,59	0,35	0,44
DIHARD3 [120, 119]	0,43	0,32	0,09	0,14	0,57	0,37	0,45
VoxConverse [26]	0,59	0,43	0,34	0,11	**0, 68**	0,28	0,40
DeePsyTrain	0,65	0,04	0	0	0,44	0,42	0,43
DeePsyTest	0,27	0,15	0	0	0,43	0,51	0,46
Dotrénováno na							
DeePsyTrain	0,01	0,21	0	0	0,13	0,41	0,19
DeePsyTest 1/2	0,14	0,13	0	0	0,44	**0, 56**	**0, 49**

vodu by jejich modelovací schopnosti měly být velmi dobré. Modely typu Whisper však selhávaly na tichých promluvách, kdy modely halucinovaly stejný token nebo predikovaly ticho, což vedlo k WER 52,4 %, resp. 56,86 %. Průběh WER na testovací sadě v procesu trénování modelů typu Whisper je zobrazený na obr. 5.5b. Z obrázku je patrné, že byl celý proces velmi nestabilní. Byla provedena analýza, jež potvrdila, že velikost WER na testovacím setu DeePsyTestV1 je způsobená zejména počtem vložených nebo odstraněných slov ve srovnání s náhodnou 5% částí z testovací sady datasetu Mozilla Common Voice viz tabulka 5.6.

Tabulka 5.5: Nejlepší dosažená úspěšnost předtrénovaných modelů ve výchozích konfiguracích na datové sadě DeePsyTestV1 po dotrénování na ASRCorpora. Výsledky byly získány pomocí *greedy search* dekódování.

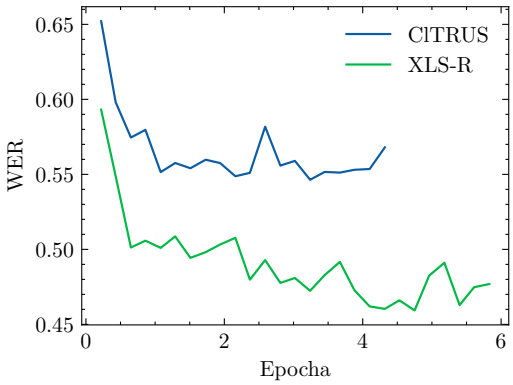
Model	Počet parametrů	WER [%] ↓	CER [%] ↓
CNN-TDNN-HMM	23 mil.	28.30	-
ClTRUS	95 mil.	54,64	33,72
XLS-R-300m	300 mil.	**45,93**	**28,64**
Whisper-base	74 mil	52,40	32,03
Whisper-small	244 mil	56,86	36,17



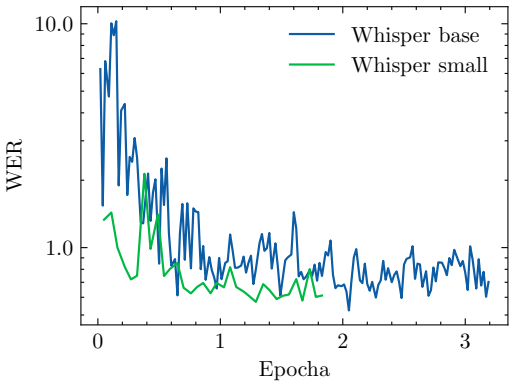
Obrázek 5.4: Graf ilustrující adaptaci architektury Wav2Vec2 pro úlohu rozpoznávání řeči. Vrstvy v čárkované části grafu zobrazují původní architekturu modelu [6], která je doplněna o klasifikační vrstvu. Model je trénován pomocí CTC objektivní funkce.

5.4.2 Přidání augmentací

Prvotní analýza měla za cíl exploraci dostupných *open-source* modelů. Získané výsledky však nedosahovaly požadované kvality. Bylo tedy dále zkoumáno, jak snížit chybovost systémů. Ačkoliv trénovací sada ASRCorpora obsahuje velký počet vzorků, prvotní experimenty indikovaly, že její doménový průnik z DeePsyTestV1 není dostačující. Z tohoto důvodu byla provedena analýza vlivů augmentací na schopnost rozpoznání řeči v psychoterapeutické doméně. Byly aplikovány standardní regularizační techniky, jako je frekvenční a časové maskování a časová deformace z knihovny SpecAugment [98]. Jelikož většina nahraných sezení obsahuje značný doslech, byla do signálu také dogenerována syntetická reverberace. Tato augmentace byla implementována s využitím knihovny pyroomaccustics [121]. Na začátku trénování je vytvořeno 500 náhodných impulzních odezev pokojů o velikosti $5 \times 2 \times 3$ metrů s náhodně rozmístěnými mikrofony a náhodným umístěním zdroje signálu. V rámci trénování sítě je pak s příslušnou pravděpodobností signál zkonvoluován s náhodnou impulzní odezvou s knihovny předgenerovaných odezev. Všechny augmentace byly implementovány jako online za běhu modelu. Celkově bylo provedeno 7 experimentů s dobou běhu pěti epoch. Pro účely rychlé explorace byl využit model ClTRUS. Experimenty jsou sumarizovány v tabulce 5.7. Nejlepší výsledky byly dosaženy v kombinaci všech



(a) Modely rodiny Wav2Vec2.



(b) Modely rodiny Whisper.

Obrázek 5.5: Experimenty provedené s předtrénovanými modely pro rozpoznávání řeči. Modely byly dotrénovány na datasetu ASRCorpora bez doménové sady DeePsyTrain a *checkpointy* byly přímo validovány na testovací sadě DeePsyTest. Trénování bylo předčasně ukončeno v době, kdy se WER na validační sadě přestala zlepšovat.

Tabulka 5.6: Analýza chybovosti modelu Whisper. Z tabulky je patrné, že nejvíce chyb je způsobeno vložením nových slov do hypotézy, což přímo odpovídá halucinaci stejného tokenu. Taktéž vysoký počet odstranění úměrný segmentům, kdy nebyla detekována řečová aktivita.

Trénovaný	Testovací sada	WER [%] ↓	CER [%] ↓	shod ↑	vložení ↓	Počet odstranění ↓	substitucí ↓
	DeePsyTestV1	106,68	71,34	6316	8455	4959	20711
✓	DeePsyTestV1	69,37	47,06	15821	6023	6492	9673
✓	Common Voice [2]	29,00	15,08	2630	72	162	811

augmentací s pravděpodobností $p = 0,2$. Celkový průběh chybovosti v rámci trénování je zobrazen na obr. 5.6a. S aplikací augmentací bylo dosaženo relativního zlepšení 11,25 %.

5.4.3 Vliv doménových dat

V době dokončení analýzy augmentací byla vytvořena doménová sada nahrávek sezení DeePsyTrain. Sada obsahuje celkově 9 nahrávek v celkové délce 7,6 hodin, viz sekce 4.1.2. Cílem této sady experimentů bylo odhalit vliv a nutnost další anotace dat. K následujícím experimentům byl využit nejlepší *checkpoint* získaný v předešlé sekci. Postupně bylo provedeno 5 experimentů s postupným přidáním 20 %, 40 %, 60 %, 80 % a 100 % dat ze sady DeePsyTrain. Model byl následně dotrénován po dobu pěti epoch. Graf na obr. 5.6b znázorňuje drastický posun s pouhým přidáním ≈ 1 h dat. Je však patrná stagnace v pozdější fázi tréningu v porovnání s experimentem uvažujícím celou trénovací sadu. Bylo zjištěno, že pouhá jedna hodina dat v doméně dokáže snížit relativně WER o 18,5 %. Kompletní doménová trénovací sada (7,6 hodiny) relativně snižuje WER o 26,00 %.

V rámci evaluace těchto experimentů byla odhalena chyba v testovací sadě. Dvě nahrávky obsahovaly posunuté kanály audio a při tokenizaci datasetu byla v kódu chyba, která neignorovala [PAD] token při dekódování. WER na nahrávkách s posunutými kanály dosahoval hodnoty $> 100\%$, proto pro další experimenty byla připravená opravená verze

Tabulka 5.7: Nejlepší dosažená úspěšnost modelu ClTRUS na datové sadě DeePsyTestV1 v závislosti na datových augmentacích v průběhu tří epoch. Sloupec pravděpodobnost určuje, s jakou pravděpodobností je provedena augmentace na daném vzorku. V případě kombinace augmentací je tato hodnota fixní pro dílčí augmentace.

Augmentace	Pravděpodobnost augmentace/augmentací	WER [%] ↓	CER [%] ↓
		53,09	32,24
Frekvenční a časové maskování	0,05	52,09	31,76
Reverberace	0,05	51,89	31,04
Časová deformace	0,05	52,51	31,85
Frekvenční a časové maskování, reverberace a časová deformace	0,05	51,29	30,95
	0,1	50,93	30,76
	0,2	**50,77**	**30,69**

testovací sady – DeePsyTestV2. Vliv změny testovací sady je upřesněn v rámci tabulky 5.8.

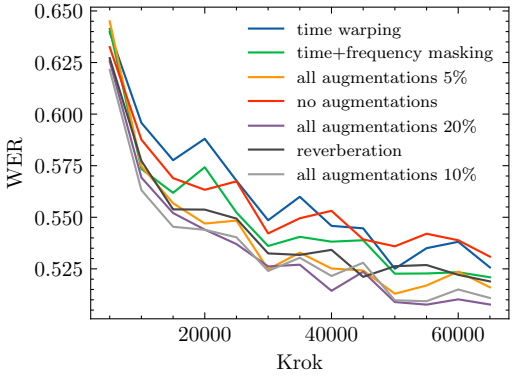
Tabulka 5.8: Vliv chybné nahrávky a chyby v tokenizaci na velikost WER a CER napříč testovacími sadami.

Model	Testovací dataset	WER [%] ↓	CER [%] ↓
ClTRUS	DeePsyTestV1	54,64	33,72
ClTRUS	DeePsyTestV2	49,40	23,96
XlS-R	DeePsyTestV1	45,93	28,64
XlS-R	DeePsyTestV2	40,76	19,51

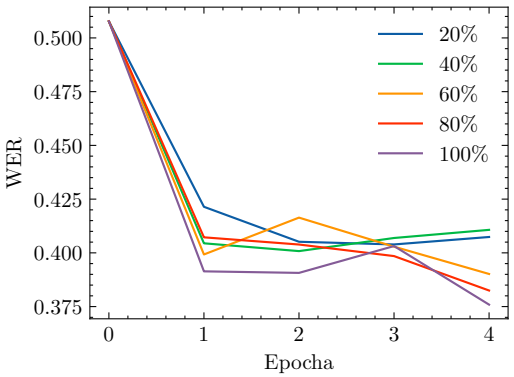
S takto upravenou testovací sadou byla provedena analýza, zda je vhodné model předtrénovávat na všech datech a zda náhodou není dostačující využít pouze 7 h doménových dat. Výsledky zobrazené v tabulce 5.9 ukazují na to, že optimální strategií je nejdříve model natrénovat na datové sadě ASRCorpora a poté pouze adaptovat na DeePsyTrain (řádky 3,6). Byly rovněž analyzovány varianty adaptace na příslušných sadách s modely ClTRUS a XLS-R-300m. Bylo zjištěno, že v případě modelu ClTRUS dotrénování na doménové sadě DeePsyTrain vede k lepším výsledkům než využití celé datové sady ASRCorpora (řádky 1,2). Tato vlastnost již však neplatí u modelu XLS-R-300m (řádky 4,5). Předtrénováním na datové sadě ASRCorpora a následným dotrénováním na DeePsyTrain bylo dosaženo relativního snížení WER o 22,06 % v případě modelu XLS-R-300m.

5.4.4 Jazykové modely využité u dekódování

Princip, jakým jsou předtrénovány modely rodiny Wav2Vec2, žádným způsobem nepředtrénovává jazykový model. Model má šanci naučit se distribuci příslušných sekvencí v rámci jazyka pouze v fázi dotrénování pro ASR. V rámci této fáze tréningu je však dostupná pouze malá množina textových dat, ze kterých model nemá šanci hlouběji pochopit vlastnosti jazyka, bylo taktéž využito dekódování po znacích, které úlohu modelování jazyka



(a) Vliv zvolené augmentace na WER v průběhů trénování.



(b) Vliv postupného přidání větší a větší části datasetu na dotrénování modelu pro příslušnou doménu.

Obrázek 5.6: Experimenty provedené pro analýzu vlivu přidání doménových dat, jakož i přidání datových augmentací.

Tabulka 5.9: Vliv doménových dat na úspěšnost rozpoznávání na testovací sadě DeePsyTestV2.

Model	Trénovací data	WER [%] ↓	CER [%] ↓
ClTRUS	ASRCorpora	49,40	23,96
	DeePsyTrain	45,24	20,58
	ASRCorpora → DeePsyTrain	38,69	17,96
XLS-R-300m	ASRCorpora	40,76	19,51
	DeePsyTrain	50,47	20,53
	ASRCorpora → DeePsyTrain	**31,77**	**14,39**

činí značně složitější v porovnání například s BPE tokenizací. Byly zkoumány možnosti, jak tedy přidat jazykovou informaci do procesu dekódování. Pro exploraci problému byly zvoleny n-gramové modely, které je možno rychle natrénovat, a dekódování s nimi je velmi rychlé. Byly postupně analyzovány modely s granularitou 2-4 viz první část tabulky 5.10. Analýza ukázala, že zvyšující granularita modelu vede ke snížení perplexity, avšak daný fakt nevede k významnému snížení WER. Dále byl zkoumán vliv váhy jazykového modelu ve fázi dekódování. Nejlepší výsledky byly dosaženy s váhou 0,4. Z tabulky je patrné, že příliš velká váha jazykového modelu vede k degradaci WER.

Dále bylo zkoumáno, jaký má vliv očištění dat a případné využití pouhé doménové části textového korpusu. Ukázalo se, že očištění datové sady vede k významnému poklesu perplexity, avšak příslušné změny mají minoritní vliv na výslednou chybu ASR, která je zde klíčová. Následně byla provedena analýza hypotetických možností systému. Ta ukázala, že se v rámci 100 nejpravděpodobnějších hypotéz modelu nachází promluvy s nižší chybovostí, a tedy v případě výběru hypotézy ze sady N-best listu podle orákula je možné dosáhnout významného snížení chybovosti systému. Dále byla provedena analýza s předtrénováním jazykového modelu na testovací datech. Díky tomu došlo k významnému snížení perplexity, avšak výsledný WER nebyl snížen. Nepodařilo se zjistit příčinu této inkonzistence.

Díky této sadě experimentů bylo dosaženo relativního snížení WER o 21,41 %. Jako nejlepší varianta se jeví předtrénování modelu na celé sadě bez očištění dat a přiřazení váhy 0,4. Pro následující experimenty byl zafixován 3-gramový jazykový model, jelikož se jeví jako nejlepší kompromis mezi velikostí a chybovostí modelu.

Tabulka 5.10: Analýza n-gramových LM použitých jako dekodér při automatickém rozpoznání řeči na testovací sadě DeePsyTestV2 s fixním akustickým modelem XLS-R-300m. LMCorpora označuje celou textovou sadu, LMCorpora* sadu očištěnou od vět obsahujících znaky, které se nenachází v abecedě modelu. LMCorpora $-\{4\}$ označuje datovou sadu LMCorpora, ze které byl extrahován dataset AnnoMI viz tabulka 4.2.

Řád LM	Perplexita ↓	WER [%] ↓	CER [%] ↓	Váha LM	Trénovací data
Vliv řádu n-gramového LM					
-	-	40,76	19,52	0,5	LMCorpora
2	302,958	32,48	18,35		
3	271,554	**32,32**	18,32		
4	**265,565**	32,34	**18,29**		
Vliv váhy LM					
3	271,554	32,19	**17,42**	0,3	LMCorpora
		32,05	17,60	0,35	
		32,03	17,73	0,4	
		32,32	18,32	0,5	
		32,83	18,86	0,6	
		35,64	20,81	0,8	
Vliv trénovacích dat					
3	271,554	**32,32**	**18,32**	0,5	LMCorpora
	282,409	32,36	18,29	0,5	LMCorpora $-\{1,3,4,5,6,8,10\}$
	306,56	32,58	18,36	0,5	LMCorpora $-\{7,9\}$
	262,944	32,35	18,36	0,5	LMCorpora*
Hypotetické možnosti systému					
4	256,683	32,06	17,80	0,4	LMCorpora*
4 orákulum	256,683	26,50	16,29	-	LMCorpora*
4	15,475	32,73	18,27	0,5	DeePsyTest

5.4.5 Reskórování hypotéz

Poznatky zjištěné z předchozích experimentů byly následně společně aplikovány a byl natrénován nový model XLS-R-300m. Přidáním augmentací do fáze prvotního trénování na sadě ASRCorpora bylo dosaženo zlepšení WER na 40,73 % viz tabulka 5.11. Další zlepšení na 32.03 % WER bylo dosaženo použitím 3-gramového jazykového modelu. Významného zlepšení bylo dosaženo pomocí finální adaptace systému pomocí 7 hodin doménových dat, čímž bylo dosaženo chybovosti WER 25,12 %, což je 45,31% relativní zlepšení oproti původnímu modelu XLS-R-300m a relativní zlepšení o 11,23 % vůči hybridnímu *baseline*.

Jelikož v rámci uvážení 100 nejpravděpodobnějších hypotéz systému a následného zvolení jako výstupu systému hypotézy s nejnižší chybou bylo dosaženo významného snížení chybovosti, byl dále proveden pokus o reskórování hypotéz externím jazykovým modelem. Pro tyto účely byla provedena analýza perplexity autoregresivních modelů natrénovaných

Tabulka 5.11: Postupné zlepšení dosažené ze zakomponováním příslušných technik.

System	WER [%] ↓
CNN-TDNN-HMM	28, 30
XLS-R-300m	45, 93
+ augmentace	40, 76
+ 3 gramový LM	32, 03
+ 7 hodin doménových dat	**25, 12**

na češtině viz tabulka 5.12. Po analýze dosažených výsledků a uvážení trénovacích dat se jevil model MU-NLPC/CzeGPT-2 jako nejvhodnější kandidát pro další experimenty. Model byl dotrénován pro účely projektu DeePsy na plné verzi datasetu LMCorpora a na čistě doménových datech. V případě doménové varianty bylo dosaženo snížení perplexity na 174,85 po 5 epochách. Tréning na všech datech nevedl k výraznému snížení perplexity a byl po 1,5 epoše zastaven.

Tabulka 5.12: Analýza předtrénovaných autoregresivních modelů pro klasické jazykové modelování v českém jazyce vyhodnocených na datové sadě DeePsyTest.

Model	Perplexita ↓
MU-NLPC/CzeGPT-2 [59]	533, 990
lchaloupsky/czech-gpt2-oscar [22]	2607, 64
spital/gpt2-small-czech-cs[7]	275, 67
fav-kky/gpt2-small-cs[8]	77090, 27
lchaloupsky/czech-gpt2-medical [22]	3089, 52

Následně byl tedy proveden pokus s reskórováním N-best hypotéz, kde $N = 250$. Všechny hypotézy byly postupně ohodnoceny akustickým modelemem XLS-R, 3-gramových LM, MU-NLPC/CzeGPT-2, MU-NLPC/CzeGPT-2 dotrénovaným na doménových datech, dále referovaným jako DeePsyGPT-2 a modelem dotrénovaným na LMCorpora LMGPT-2. Tímto byl získán dataset 3399 promluv, kde pro každou z promluv bylo vygenerováno 0–250 hypotéz, čímž vznikl dataset 174 tis vzorků obsahujících pět log-likelihoodů doplněných o referenci a odpovídající WER dané hypotézy. Dataset byl následně rozdělen na trénovací a testovací část. Prvním experimentem bylo zjištění, jak nízko je možné se dostat, pokud je využito Oracle reskórování – je vždy vybrána hypotéza s nejmenším WER. V rámci tohoto experimentu bylo s nově natrénovaným modelem XLS-R-300m dosaženo WER 17,95 %, což se jevilo jako pozitivní předpoklad k reskórování. Následně byl proveden experiment vybírající hypotézu vždy podle log-likelihoodu jednoho z modelů. Z tabulky 5.13 je patrné, že pouhé reskorování hypotéz jedním z dotrénovaných modelů nevede k žádnému zlepšení. Z experimentů bylo taktéž patrné, že obecné modely GPT2 dosahovaly horší úspěšnosti. Z tohoto důvodu byl pro další experimenty zvážen pouze model DeePsyGPT-2. Nejprve byl proveden pokus o predikci WER odhadem vah příslušných faktorů pomocí lineární regrese. Byly provedeny pokusy jak o lokální, tak globální normalizaci věrohodností a WER. Bylo

[7] https://huggingface.co/spital/gpt2-small-czech-cs
[8] https://huggingface.co/fav-kky/gpt2-small-cs

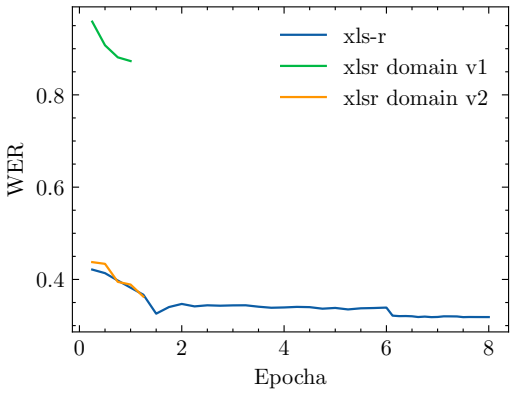
taktéž zkoumáno natrénování vah faktorů a ověření na stejné sadě, bohužel žádný z přístupů nevedl ke zlepšení WER a nepodařilo se odhalit, zda je v tomto přístupu logická chyba nebo jsou data zatížena přílišným šumem. Z tohoto důvodu bylo provedeno manuální prohledání vah modelů s krokem 0,1. Nejlepší WER bylo dosaženo s váhami 0,6 pro akustický, 0,4 pro 3-gramový LM a 0 pro DeePsyGPT-2. Smixováním log-likelihoodů se zmíněnými váhami bylo dosaženo WER 25,01 %. Dále byl proveden pokus o natrénovaní jednoduché 3 vrstvé neuronové sítě pro predikci WER, kde byly použity jako parametry délka sekvence a log-likelihoody XLS-R, 3-gramového LM a DeePsyGPT-2. Bohužel tento přístup taktéž nevedl k žádnému zlepšení.

Tabulka 5.13: Analýza reskórování 250-best hypotéz pro testovací sadu DeePsyTest a vliv na výslednou hodnotu WER. Příslušné řádky uvažují situaci, kdy pouze jeden z modelů určuje výstupní sekvenci podle jeho nejvyšší věrohodnosti.

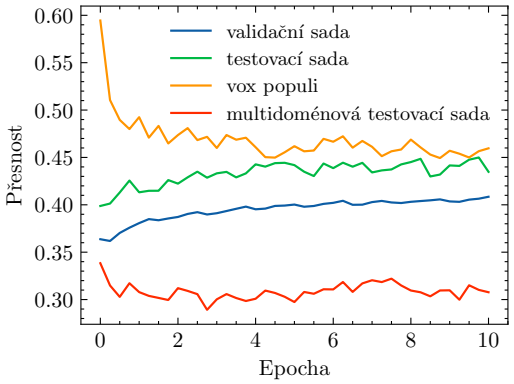
Reskorování podle	WER [%] ↓
XLS-R	27,35
3-gramový LM	27,69
MU-NLPC/CzeGPT-2	29,58
DeePsyGPT-2	28,75
LMGPT-2	28,97

5.4.6 Doménová adaptace na DeePsyUnsupervised

V době provedení tohoto experimentu bylo v rámci projektu DeePsy již nahráno celkově 696 hodin psychoterapeutických sezení, což je už dostatek pro doménovou adaptaci pomocí učení pod vlastním dohledem. Byly provedeny dva experimenty s následným dotrénováním na rozpoznávání řeči. Prvním pokusem bylo slepé dotrénování po dobu 20 tis. kroků s velikostí batche 2 a 8 kroky akumulace gradientu s optimalizátorem AdamW($\gamma = 0,001$, $\beta_1 = 0,9$, $\beta_2 = 0,999$, $\lambda = 0,01$). Bohužel učící konstanta byla velmi vysoká a vlastnosti modelu se přímo zhoršily na úloze automatického rozpoznávání řeči viz obr. 5.7a. Z tohoto důvodu byl trénovací skript přepsán a bylo přidáno monitorování napříč několika validačními sadami. Učící konstanta byla upravena na hodnotu $\gamma = 0,00001$. Na obr. 5.7b je vidět postupné zlepšování na validační a testovací sadě DeePsy a postupné zhoršování na sadě dat VoxPopuli, na které byl model původně trénován. Z grafu je patrné, že došlo k relativnímu zlepšení modelovacích schopností o 8,40 % relativně. Na úloze automatického rozpoznávání řeči nedošlo v první fázi tréningu, ve které bylo zlepšení nejvíce očekáváno, k žádnému zlepšení. Předtrénovaný model byl taktéž ověřen v seq2seq architektuře a dotrénován po 10 epoch. To však taktéž nevedlo k žádnému zlepšení výsledného systému viz obr. 5.8a, model je na tomto obrázku označen přerušovanou čárou. Jelikož předtrénování na doménových datech je stěžejní krok, který by celému rozpoznávači mohl výrazně pomoci, je nutné v budoucnu provést komplexnější analýzu a případně vyzkoušet jiné modely pro předtrénování řečových příznaků jako třeba HuBERT [56] či WavLM [23]. Věřím, že zde leží další skrytý potenciál a prostor pro zlepšení výsledného modelu.



(a) Pokusy o předtrénování XLS-R na doménových datech a následné trénování na rozpoznávání řeči. Z obrázků je patrné, že ani v jednom z případů nebylo dosaženo významného zlepšení v průběhu první fáze tréningu.



(b) Analýza průběhu přesnosti modelovacích schopností XLS-R v průběhu předtrénování na datasetu DeePsyUnsupervized. Graf ukazuje podíl správně rozpoznaných akustických jednotek mezi distraktory.

Obrázek 5.7: Experimenty provedené s předtrénováním modelu XLS-R na datové sadě DeePsyUnsupervized

.

5.4.7 Zakomponování jazykového modelu

Ačkoliv reskórování pomocí externího jazykového modelu architektury Transformer nevedlo k žádnému zlepšení, využití dekodéru typu Transformer je běžným způsobem, jak zlepšit systém pro automatické rozpoznávání řeči [6, 25]. V úlohách přímého překladu řeči je často využit předtrénovaný dekodér pro cílový jazyk a předtrénovaný enkodér pro zdrojový jazyk [137]. Modely jsou následně spojeny do jednotné architektury přidáním *cross-attention* vrstev a model je dotrénován pro cílovou úlohu. Architektura je naznačena na obr. 5.9. Po vzoru předchozího článku byly postupně provedeny experimenty uvedené v tabulce 5.14, zkoumající zakomponování výše představené modely XLS-R a GPT2 do společného modelu pro automatické rozpoznávání řeči. Byly postupně provedeny experimenty se zamrazením enkodéru a dekodéru, či ponecháním klasifikační CTC vrstvy a zakomponování společné objektivní funkce [139].

Prvotní experiment (řádek 4) se zamrazeným enkodérem běžel po dobu 5 epoch. Ostatní modely(řádky 5 a 7-9) byly trénovány 10 epoch. Evaluace na validační sadě běžela ve všech případech s využitím *greedy* dekódování. Průběh trénování je zobrazen na obr. 5.8a. Z porovnání řádků 4 a 7 je patrné, že trénování se zamrazeným enkodérem, který již byl natrénován pro rozpoznávání řeči, je výhodnější než trénování s odmrazeným enkodérem, což značně snižuje i dobu nutnou pro natrénování modelu. Řádky 5 a 8 znázorňují, že náhodná inicializace dekodéru se jeví jako lepší varianta oproti využití předtrénovaného dekodéru. Možným vysvětlením je, že enkodér a dekodér se na sebe navzájem pokouší adaptovat a tím jsou zároveň odstraněny předtrénované znalosti z obou modelů. Tento poznatek vedl k pokusu o postupné odmrazování parametrů (řádek 8) – na začátku trénování jsou zmrazeny vrstvy enkodéru kromě vrstvy poslední, na straně dekodéru jsou zamrazeny všechny vrstvy kromě klasifikační vrstvy. Zároveň jsou odmrazeny *cross-attention* vrstvy. Po 5 tisících trénovacích krocích jsou odmrazeny všechny vrstvy dekodéru a po 40 tis. krocích je odmrazena celá architektura. Tímto přístupem bohužel nebylo dosaženo zlepšení. Dále bylo provedeno přidání CTC vrstvy do enkodéru a zakomponování společné objektivní funkce

Tabulka 5.14: Analýza úspěšnosti natrénovaných modelů typu seq2seq na testovací sadě DeePsyTest.

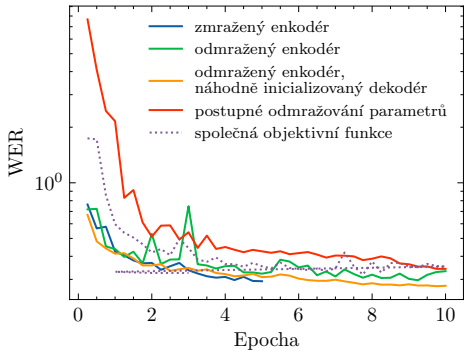
ID	Systém	WER [%] ↓
1	CNN-TDNN-HMM	28, 30
2	XLS-R-300m	31, 77
3	+ 3 gramový LM	25, 12
4	zamrazeny XLS-R-300m + GPT 2 dekodér	29, 31
5	XLS-R-300m + náhodně inicializovaný dekodér	27, 56
6	+ beam dekódování	**23, 47**
7	XLS-R-300m + GPT 2 dekóder	29, 67
8	+ postupné odmrazováni	34, 20
9	+ předtrénovaný XLS-R-300m + společná objektivní funkce	31, 82
10	Whisper-medium	24, 25

$\mathcal{L} = \lambda_{CTC}\mathcal{L}_{CTC} + (1 - \lambda_{CTC})\mathcal{L}_{ATT}$, kde λ_{CTC} (řádek 9). Do tohoto experimentu byl nedopatřením zahrnut předtrénovaný enkodér na DeePsyUnsupervised datech, což znemožňuje vyvození exaktních závěrů. Je nutno podotknout, že předchozí experimenty však nenaznačovaly žádné zlepšení předtrénováním modelů na DeePsyUnsupervised datech. Z tohoto předpokladu lze tedy částečně vyvodit, že začlenění společné objektivní funkce má pozitivní dopad na celkovou úspěšnost systému.

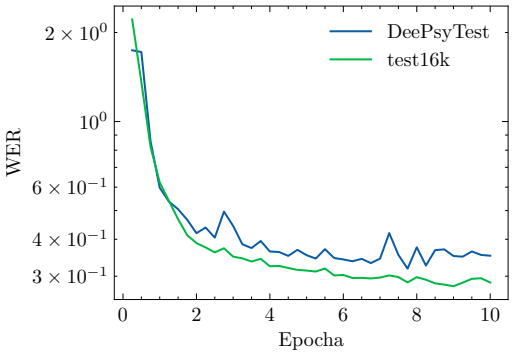
Je nutné podotknout, že výše popsané experimenty běžely na zmixované trénovací sadě obsahující doménová a obecná data. Nedošlo tedy k přímé adaptaci na doménu. S nejlepším checkpointem bylo dosaženo při *greedy* dekódování WER 27,56 %, což je horší než 3-gramový LM, avšak po začlenění *beam* dekódování s velikostí *beamu* 4 bylo dosaženo zlepšení na WER 23,47 % testovací sadě DeePsyTest, což je nejlepší dosažený výsledek v rámci této práce. Jedná se o relativní zlepšení 17,06 % vůči nejlepšímu dostupnému hybridnímu systému. Navzdory prvotním negativním výsledkům s dotrénováním architektury Whisper bylo po menším doladění parametrů s verzí Whisper-medium dosaženo 24,25 % WER. Tyto pozitivní výsledky motivují další analýzu modelů seq2seq a provedení detailnější analýzy. Je nutno podotknout, že model Wav2Vec2+GPT2 obsahuje o 300 mln. parametrů méně než model Whisper-medium a dosahuje lepších výsledků. Ačkoliv ≈ 25 % se jeví pořád jako vysoká nepřesnost, některé studie taktéž ukazují podobné výsledky na psychoterapeutických datech v angličtině [91].

5.5 Klasifikace sentimentu

Pro testování kvality klasifikace na reálných datech byla vytvořena malá testovací sada 400 anotovaných segmentů. Jako vstup posloužily automaticky vygenerované přepisy sezení nahrávek DeePsy, které byly anotovány do třech kategorií. Promluvy byly anotovány dvěma anotátory. Shoda mezi nimi byla rovna $\kappa = 0,89$ [89]. K evaluaci systému byly zvoleny promluvy, v nichž došlo ke shodě mezi anotátory. Příslušné kategorie v testovací sadě byly zastoupeny v poměru: neutrální 83,43 %, negativní 11,66 % a pozitivní 4,90 %. Je nutné poznamenat, že klasifikace probíhala na promluvách, které obsahují chybu zanesenou systémem ASR a samotná klasifikace promluv byla poměrně obtížná. Experimenty



(a) Analýza tréningu příslušných verzí navržené seq2seq architektury. Přerušovanou čárou je vyznačen model, jehož enkodér obsahuje předtrénovaný model XLS-R na doménových datech. U všech ostatní experimentů je využit enkodér XLS-R, který byl dotrénován pomocí CTC objektivní funkce.



(b) Analýza chybovosti WER napříč dvěma testovacími sadami. Z Obrázku je patrné, že trénovaný systém v tomto případě seq2seq se společnou objektivní funkci se postupně zlepšuje na obou doménách. Zlepšení na obecné testovací sadě je rychlejší než na doménové sadě DeePsyTest. Po odstranění přidaných tokenů označujících neřečové události a inkorporaci *beam* prohledávání bylo dosaženo WER 17,30 %.

Obrázek 5.8: Experimenty provedené se začleněním dekodéru do společné architektury seq2seq a analýza chybovosti ASR systému na DeePsyTest a obecné testovací sadě Testset16k.

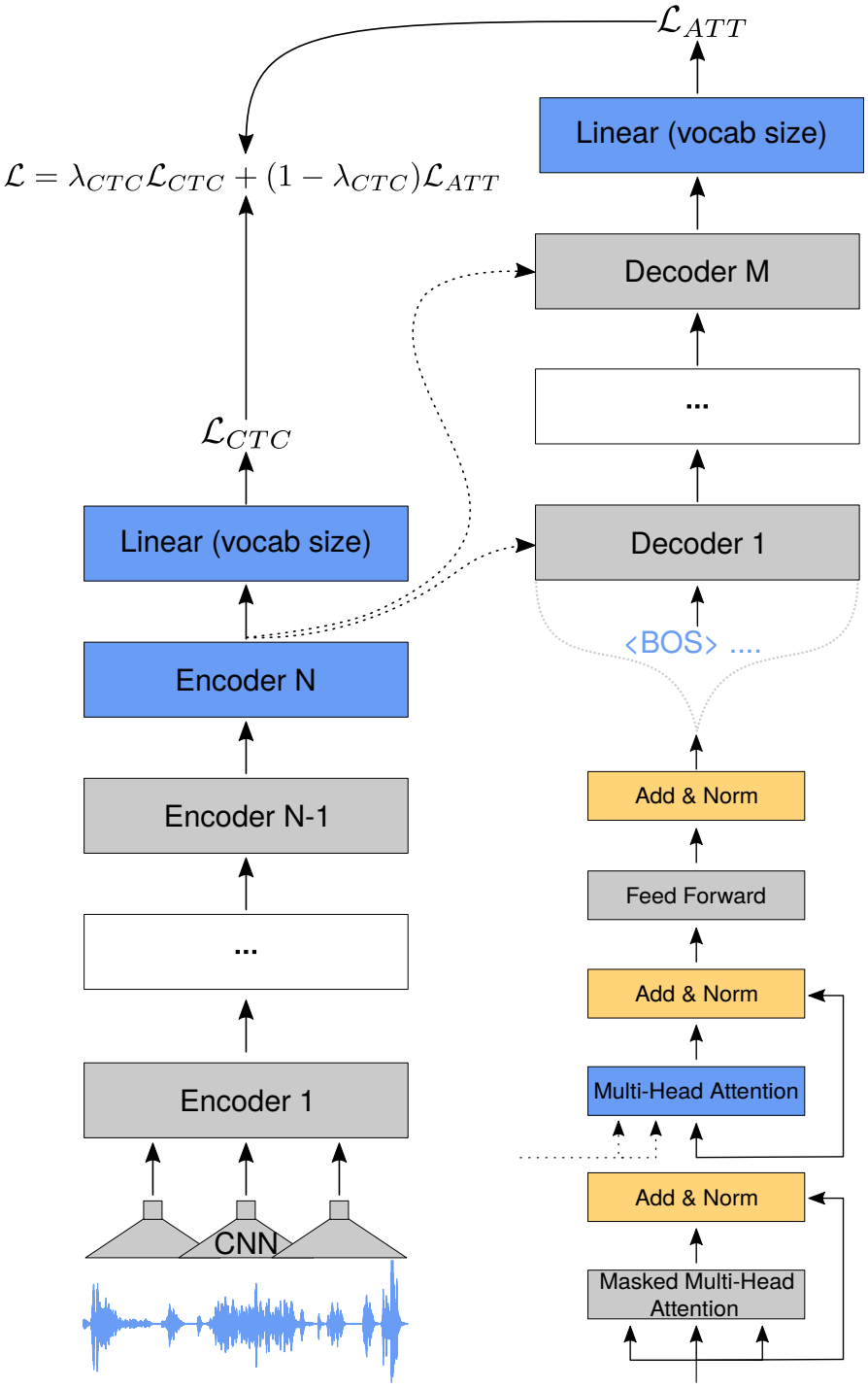
byly provedeny v době, kdy byla chybovost nejlepšího dostupného hybridního systému typu CNN-TDNN-HMM viz sekce 5.4 rovná 35,9 % WER.

5.5.1 Evaluace předtrénovaných modelů

Všechny níže popsané experimenty byly provedeny s předtrénovaným jazykovým modelem CZERT [127], který vychází z architektury BERT [32]. Model byl předtrénován na článcích z české Wikipedie, stažených zprávách v češtině a Českém národním korpusu [69]. Autoři tohoto modelu taktéž poskytli dvě dotrénované varianty pro klasifikaci sentimentu[9]. Ty byly natrénovány na datasetech recenzí z prostředí Facebooku[10] – CZERT-fb a hodnocení filmů z Česko-Slovenské filmové databáze[11] [49] – CZERT-csfd. Jelikož pro tuto úlohu nebyla anotována žádná doménová data, prvním z experimentů bylo pouhé otestování těchto *checkpointů* na výše popsané testovací sadě.

Tabulka 5.15 sumarizuje provedené experimenty, v sloupcích jsou uvedeny následující metriky: přesnost, top-2 přesnost a macro F1. Nejprve byla provedena evaluace předtrénovaných modelů. V případě CZERT-csfd bylo dosaženo přesnosti 0,55, což napovídá o doménové neshodě mezi trénovacími a testovacími daty. Příčinou může být taky fakt, že reference trénovací sady byly tvořeny automaticky. V případě CZERT-fb bylo dosaženo přesnosti 0,74, což je značně lepší výsledek, avšak makro průměr 0,46 indikuje fakt, že úspěšnost klasifikace napříč kategoriemi není vyvážená.

[9] https://huggingface.co/UWB-AIR/Czert-B-base-cased-long-zero-shot
[10] https://www.facebook.com
[11] https://www.csfd.cz



Obrázek 5.9: Navržený protokol trénování seq2seq modelů. V první fázi tréningu jsou zamrazeny všechny vrstvy enkodéru kromě vrstvy poslední. Na vrchol enkodéru je přidána CTC vrstva. Na straně dekodéru je taktéž celkově odmrazena pouze poslední vrstva, v příslušných vrstvách je odmrazena pouze *cross-attention* vrstva. Zamrazené parametry jsou označeny šedě, parametry, kde gradient proplouvá, modře. Po X krocích je následně odmrazený celý dekodér. Následně je po Y krocích odmrazena celá architektura. Tímto způsobem má model možnost postupné adaptace svých parametrů. Byla taktéž implementována společná objektivní funkce $\mathcal{L} = \lambda_{CTC}\mathcal{L}_{CTC} + (1 - \lambda_{CTC})\mathcal{L}_{ATT}$, kde λ_{CTC} je interpolační koeficient.

5.5.2 Kombinace dostupných dat a tvorba vlastní trénovací sady

V reakci na nepříznivé výstupy výše uvedené analýzy byl natrénován nový modelu na spojeném datasetu FB, ČSFD a Mall viz sekce 4.2.4. Bohužel kombinace dat z několika zdrojů vedla k razantnímu zhoršení přesnosti klasifikace na hodnotu 0,45. Z tohoto důvodu byla provedena analýza možnosti, jak získat další data bez nutnosti nákladné anotace dat. Byl tedy vytvořen dataset TwitterEmotions viz sekce 4.2.5. Na těchto datech byl natrénován nový model vycházející z předtrénovaného modelu CZERT. Model byl trénovaný na klasifikaci do tří tříd, kde neutrální třída byla reprezentována daty z Wikipedie (řádek 4). Bohužel s takto natrénovaným modelem se nepodařilo pokořit přesnost CZERT-fb. Z analýzy predikcí modelu bylo zjištěno, že predikce mající vysokou pravděpodobnost dané kategorie $\theta > 0,95$ ve většině případů odpovídají realitě a většina chyb byla učiněna v neutrální kategorii, kde byl model učen na datech z Wikipedie, která nemusí být nutně neutrální a mají zcela odlišnou strukturu než testovací data. Z tohoto důvodu byla neutrální data odstraněna a model byl učen k regresi do intervalu $\langle -1, 1\rangle$ s využitím pouze negativní a pozitivní třídy. Model, referovaný v tabulce jako *regrese Twitter*, byl evaluován s prahem pro detekce pozitivní, či negativní kategorie $\theta = \pm 0,95$, v případě, že skóre nepřesáhlo daný práh, byla kategorie prohlášena za neutrální. Takto byla získána přesnost 0,70, která však byla taktéž nižší než CZERT-fb.

Ve fúzi s modelem trénovaným čistě na datasetu FB a prahem nastaveným na hodnotu $\theta = 0,95$, model dosáhl nejlepší přesnosti, viz poslední řádek tabulky 5.15. Výsledkem této sady experimentů je fungující a nasazený systém pro klasifikaci sentimentu. Hodnoty jsou akumulovány v rozmezí 5 minutových segmentů. V rámci interní analýzy bylo empiricky zjištěno, že akumulované hodnoty odpovídají realitě a systém byla nasazen do provozu. Avšak pro natrénování systému, který by bylo možné určit pro věrohodnou klasifikaci příslušných promluv, je v budoucnu nutná anotace doménových trénovacích dat a tvorba obsáhlejší testovací sady.

Tabulka 5.15: Tabulka přesnosti klasifikace sentimentu pro model CZERT podle dat, na kterých byl model trénován.

Dataset/metoda	Přesnost ↑	Top-2 přesnost ↑	macro F1 ↑
ČSFD	0,55	0,79	0,36
FB	0,74	**0,97**	**0,46**
Mall + FB + ČSFD	0,45	0,91	0,31
Twitter (neutralní – data z Wikipedie)	0,63	0,95	0,33
regrese Twitter	0,70	-	0,34
Fúze (FB + regrese Twitter)	**0,79**	-	0,45

5.6 Klasifikace typů terapeutických intervencí

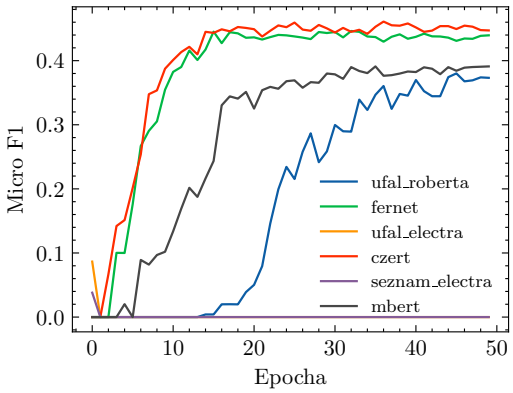
Pro účely natrénování a evaluace modelů pro klasifikaci typů terapeutických intervencí byl v rámci projektu DeePsy vytvořen dataset DeePsyInterventions, blíže popsán v rámci sekce 4.2.3. Dataset obsahuje promluvy klienta a terapeuta. Anotovány jsou pouze promluvy terapeutů. Ty byly pro prvotní experimenty z datasetu extrahovány a dále rozděleny v poměru 15:85 na validační *Val1* a trénovací část *Train1*. Některé kategorie však nebyly vůbec zastoupeny ve validační sadě, jelikož distribuce kategorií není rovnoměrná viz obr. 4.1.

Proto byly pomocí metody `IterativeStratification` z knihovny skmultilearn [133] vyextrahovány promluvy tak, aby byl zachován poměr kategorií a byla vytvořena trénovací sada *Train2* a validační sada *Val2*.

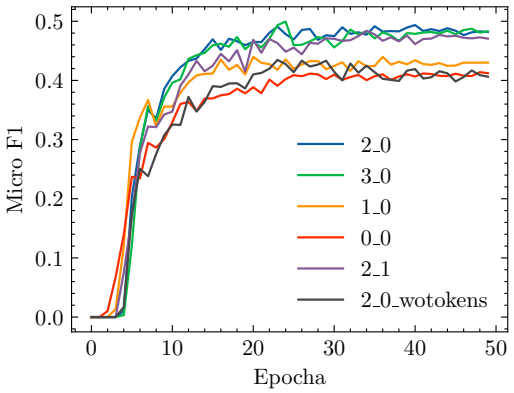
5.6.1 Prvotní experimenty

Pro prvotní experimenty byly analyzovány veřejně dostupné předtrénované jazykové modely typu enkodér trénované pomocí *Masked Language Modeling* – MLM viz sekce 2.5.3. Byly analyzovány následující modely předtrénované na českých datech: RobeCzech [130], FERNET-C5 [78], EleCzech-LC[12], CZERT [127], Small-E-Czech [66] a multilinguální BERT multilingual [32]. Modely byly doplněny o klasifikační vrstvu s výstupem do 18 kategorií doplněných o aktivační funkci logistické sigmoidy [35], jelikož se jedná o klasifikaci do více tříd současně (*Multi Label Classification* – MLC [14]). Všechny následující experimenty (pokud není řečeno jinak) byly trénovány s opitimalizátorem AdamW($\gamma = 0,00001$, $\beta_1 = 0,9$, $\beta_2 = 0,999$, $\lambda = 0,01$), lineární plánovačem (*linear scheduler*[13]) s velikostí batche 32 na 16 bitové přesnosti. Každý z experimentů běžel okolo jedné hodiny na NVIDIA RTX A5000[14].

Pro získání obecného přehledu o fungování modelů byla zvolena trénovací sada *Train1* a validační *Val1*. Modely byly evaluovány pomocí metriky micro F1. Porovnání modelů je zobrazeno na obr. 5.10a. Nejlepší výsledky byly dosaženy s modely CZERT a FERNET-C5. Článek [78], v němž byl FERNET-C5 představen, demonstruje, že model je lepší obecně v klasifikačních úlohách než CZERT. Model FERNET-C5 byl tedy zvolen jako fixní model pro další experimenty. Z grafu je patrné, že modely vycházející z kategorie Electra se nepodařilo dotrénovat na cílovou úlohu. Příčinu se nepodařilo odhalit. Je pravděpodobné, že chyba bude někde v tokenizaci, jelikož modely byly trénovány identickým zdrojovým kódem.



(a) Úspěšnost klasifikace porovnána mezi jazykovými modely předtrénovanými pro češtinu.



(b) Analýza vlivu kontextu na úspěšnost klasifikace mezi epochami.

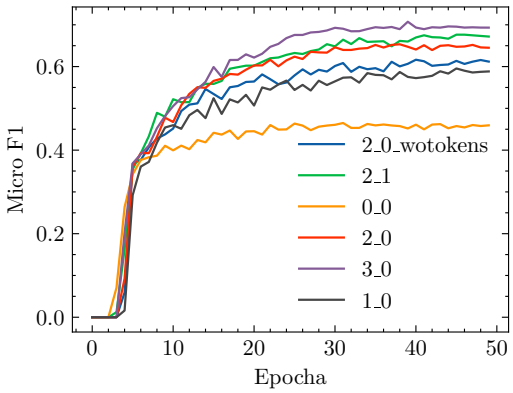
Obrázek 5.10: Prvotní analýza dostupných modelů a vhodného zahrnutí kontextu do trénování.

[12] https://huggingface.co/ufal/eleczech-lc-small
[13] https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.LinearLR.html
[14] https://www.nvidia.com/en-us/design-visualization/rtx-a5000/

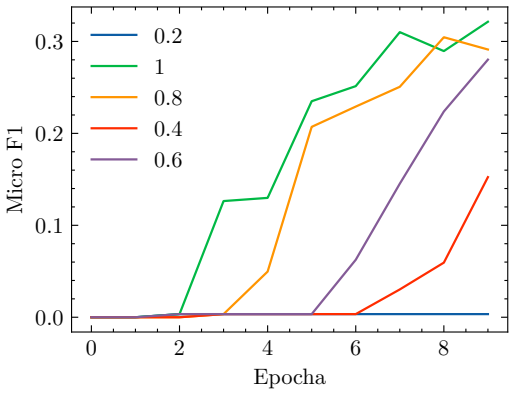
5.6.2 Přidání kontextu

Pro rozpoznání některých kategorií jako reflexe, dotazování nebo interpretace, je velmi stěžejní okolní kontext. Z tohoto důvodu byla provedena analýza velikosti kontextu na úspěšnost modelu viz obr. 5.10b. Z obrázku je patrné, že jako vhodný kompromis mezi délkou kontextu a úspěšností modelu se jeví varianta uvažující dvě předchozí věty. Je viditelné, že uvážení kontextu tří předchozích vět již nevede téměř k žádnému zlepšení. V provedených experimentech viditelných na obr. 5.10b byly automaticky přidány tokeny `[T]` a `[C]` reprezentující autora dané promluvy do vstupních sekvencí. Kontext a promluva byly rozděleny tokenem `[SEP]`, který je v modelech vycházejících z architektury BERT [32] používán k oddělení vět, u nichž je predikována sousednost. Aby byl ověřen přínos příslušného kódování, byl proveden pokus uvažující věty v původní podobě bez doplněných tokenů `[T]` a `[C]`. Z grafu je patrné, že tento přístup vedl k degradaci úspěšnosti klasifikace.

Pro získání lepšího porozumění dat byla provedena stejná analýza taktéž na trénovací sadě *Train2* a validační sadě *Val2*. Výsledky stejné sady experimentů jsou zobrazeny na obr. 5.11a Přístup uvažující stratifikovanou tvorbu validačního datasetu však není zcela spravedlivý, jelikož příslušné kategorie jsou často totožné mezi sousedními promluvami. Vzorky do validační sady byly vybrány náhodně, což modelu dovoluje v rámci trénování vidět kontext z validační sady a v rámci validace kontext z trénovací. Tento experiment však velmi dobře demonstruje schopnost pochopení kontextu. Bylo dosaženo mnohem vyšší úspěšnosti klasifikace intervencí. Dále byla provedena analýza vlivu velikosti datasetu při prvních 10 epochách tréningu. Tato sada experimentů je viditelná na obr. 5.11b Z grafu je patrné, že se sadou zmenšenou na 20 % původní velikosti model nedokáže vůbec klasifikovat vzorky z validační sady. Je poměrně zřetelné postupné zlepšování se zvětšující se trénovací sadou. Díky pozitivním výsledkům byla motivována další anotace dat, která vedla k vzniku druhé verze datasetu obsahující 40 nahrávek.



(a) Analýza vlivu kontextu na úspěšnost klasifikace mezi epochami na stratifikované verzi datasetu.



(b) Úspěšnost klasifikace se zvyšující se frakcí využitých dat pro trénování.

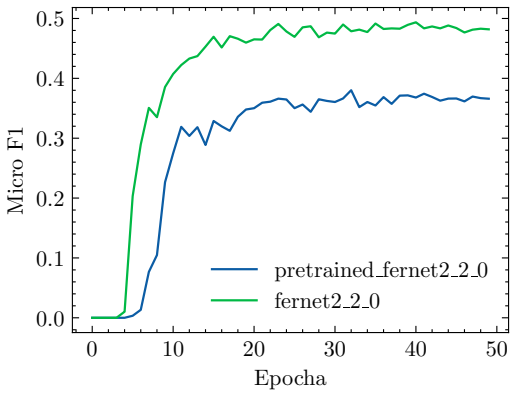
Obrázek 5.11: Porovnání hodnotu mikro průměru F1 mezi epochami při trénování klasifikace terapeutický intervencí.

5.6.3 Předtrénování modelu a regularizační techniky

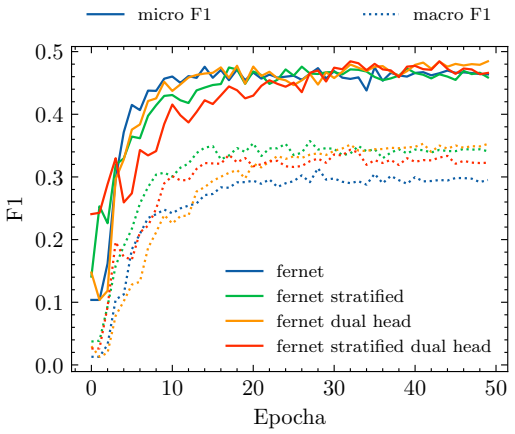
Jelikož v rámci projektu DeePsy existují neanotované promluvy, které jsou však významněji bližší anotovaným kategoriím než data, na kterých byly obecné modely pro češtinu, byl proveden pokus o předtrénování modelu na úloze maskovaného jazykového modelování. Byly k tomu použity datasety „Interní data“ a „DeePsyASR“ představené v rámci sekce 4.2. Model byl předtrénován z výchozích vah s optimalizátorem AdamW($\gamma = 0,001$, $\beta_1 = 0,9$, $\beta_2 = 0,999$, $\lambda = 0,00$) s velikostí batche 8, lineárním plánovačem (*linear scheduler*) a maximální délkou sekvence 512 tokenů po 10 epoch. Byl využit nejlepší *checkpoint* vyhodnocený na testovací sadě DeePsyTest. Bohužel po dotrénování pro klasifikaci nebylo dosaženo žádného zlepšení viz obr. 5.12a.

Dalším krokem byla implementace stratifikovaného vzorkování, jelikož zastoupení příslušných tříd v datasetu není vyvážené viz sekce 4.2.3. Toto rozšíření bylo implementováno pomocí třídy `WeightedRandomSampler`[15]. Dále byla analyzována regularizace přidáním druhé klasifikační vrstvy do 9 nadřazených kategorií a zakomponování do společné objektivní funkce v poměru $\mathcal{L} = \mathcal{L}_{sup} * \lambda_{sup} + \mathcal{L}_{sub} * (1 - \lambda_{sup})$, kde $\mathcal{L}_{sup}$ je *loss* rodičovské klasifikace a $\mathcal{L}_{sup}$ *loss* původní klasifikace do 18 tříd a λ_{sup} váhovací koeficient.

V následujících experimentech byl nastaven na hodnotu $\lambda_{sup} = 0,3$. Tyto regularizační experimenty jsou zobrazeny na obr. 5.12b. Byla taktéž zakomponována metrika makro F1 sloužící k monitorování průměru F1 mezi kategoriemi. Z experimentů vyplývá, že stratifikované vzorkování má pozitivní vliv na celkovou chybovost systému. Toto pozorování taktéž podporují grafy zobrazené na obr. 5.13, které rovněž přímo poukazují na špatnou schopnost klasifikace s malým počtem trénovacích vzorků v rámci kategorie. Experimenty taktéž poukázaly na nutnost vytvoření vyvážené validační sady, jelikož nezastoupení kategorie sebeodhalení ve validační sadě výrazně ovlivňuje výsledné metriky.



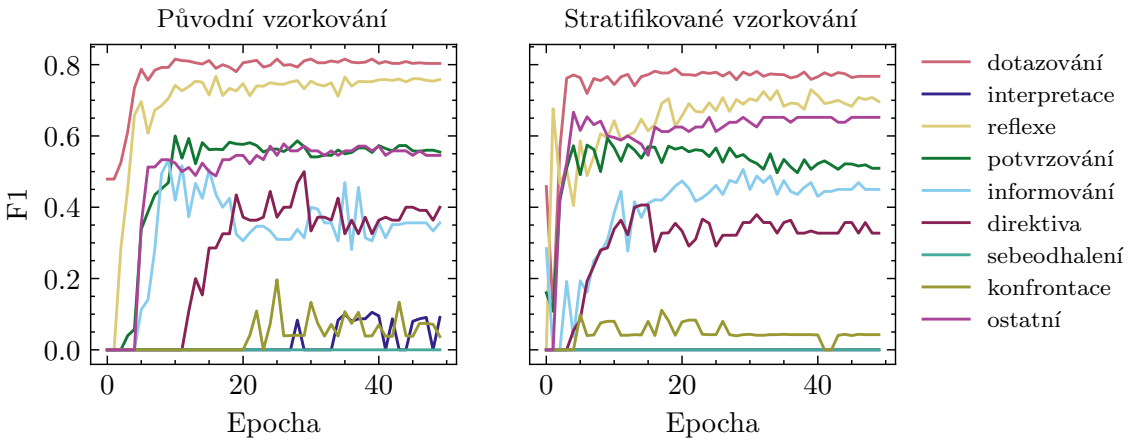
(a) Analýza vlivu předtrénování modelu pro maskované jazykové modelování na doménových datech na celkovou úspěšnost klasifikace



(b) Analýza vlivu implementování stratifikovaného vzorkování a přidání klasifikační vrstvy pro klasifikaci do rodičovských kategorií a zakomponování do společné objektivní funkce.

Obrázek 5.12: Experimenty provedené s předtrénovaním modelu Fernet na doménových textových datech a přidáním regularizačních technik na úspěšnost klasifikace.

[15] https://pytorch.org/docs/stable/_modules/torch/utils/data/sampler.html



Obrázek 5.13: Porovnání F1 rodičovských tříd se zakomponovaným stratifikovaným vzorkováním a bez něj.

5.6.4 Rozšíření datové sady

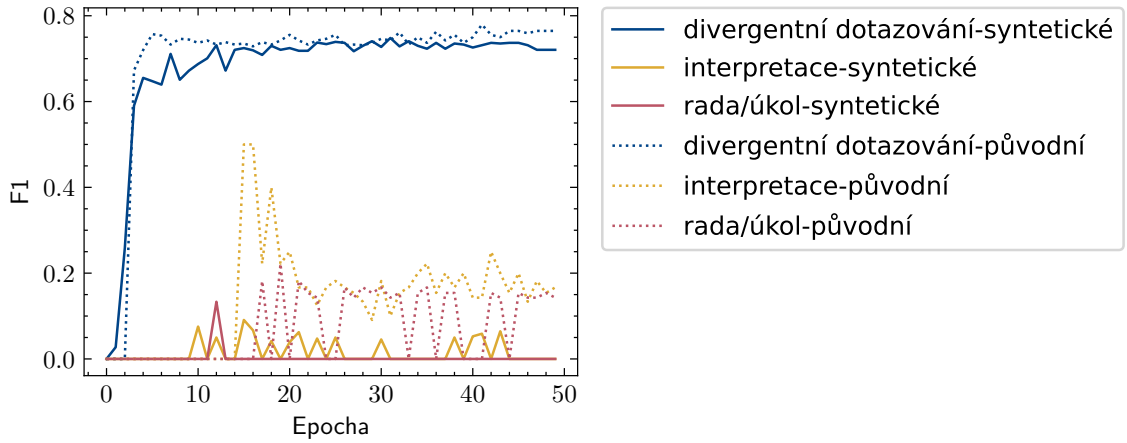
Jelikož prvotní experimenty poukázaly zlepšující se trend se zvyšujícím se počtem dat, bylo dále anotováno dalších 20 nahrávek sezení, které byly rovnou přidány do trénovací sady tak, aby byly experimenty kompatibilní a bylo provedeno porovnání. Přidání dat mělo samozřejmě pozitivní vliv a došlo ke zlepšení úspěšnosti klasifikace terapeutických intervencí. Graf na obr. 5.15a zobrazuje porovnání micro a macro F1 v průběhu trénování.

Jelikož anotace sezení je časově velice nákladná, byl proveden pokus o vygenerování syntetických dat pomocí nástroje ChatGPT[16]. Modelu byly přiloženy příkladové vzory a bylo dogenerováno dalších 400 příkladů celkově ze tří kategorií (divergentní dotazování [200], interpretace [100], rada/úkol[100]) tak, aby byly zastoupeny kategorie mající nízký počet vzorků a zároveň ty, na nichž již systém funguje dobře. Přidání těchto dat bohužel nevedlo ke zlepšení celkové úspěšnosti klasifikace viz obr. 5.15b. Obr. 5.14 ukazuje vývoj trénování na kategoriích, kde byla přidána syntetická data. Z analýzy je patrné, že přidání syntetických vzorků nemělo v žádné z kategorií pozitivní dopad na zlepšení úspěšnosti. Možnou příčinou je přílišná plynulost syntetických vět v porovnání s promluvami v rámci sezení.

5.6.5 Vyvážení datové sady

Jelikož v případě validační sady *Val1* nebyl zachován poměr kategorií vzhledem k celkové distribuci datasetu, byla vytvořena validační sada *Val2*. Ta však bohužel dovolovala modelu při evaluaci využívat kontextu, na kterém byl učen. Z tohoto důvodu byla vytvořena trénovací sada *Train3* a validační sada *Val3* v poměru 10:90 stejným způsobem jak *Train2*, resp. *Val2* s tím rozdílem, že z trénovací sady byly odstraněny všechny vzorky, u kterých kontext obsahoval nejméně jednu z promluv ve validační sadě. Tímto způsobem bylo zahozeno ≈ 300 vzorků. Avšak výsledná distribuce kategorií ve validační sadě odpovídá realitě a je zamezeno podvádění v rámci evaluace. Graf zobrazen na obr. 5.16 tedy nejlépe demonstruje schopnosti modelu klasifikovat dané promluvy do kategorií v reálném prostředí. Z grafu a pohledu na obr. 4.1 je taktéž patrná přímá korelace mezi počtem trénovacích vzorků a celkovou úspěšností klasifikace dané kategorie. Byly takto detekovány kategorie, které je dále třeba anotovat a natrénován systém, který demonstruje aktuální možnosti klasifikace

[16] https://chat.openai.com/

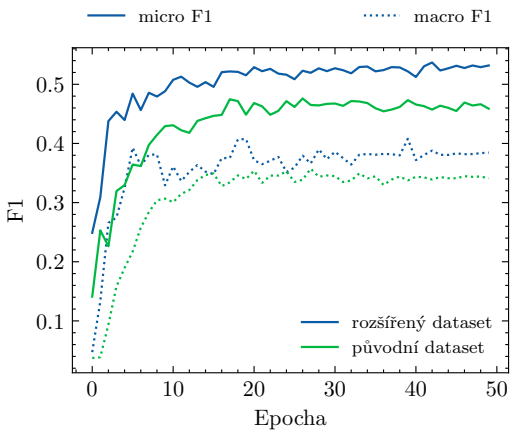


Obrázek 5.14: Porovnání F1 tříd, které byly rozšířeny o synteticky dogenerované vzorky.

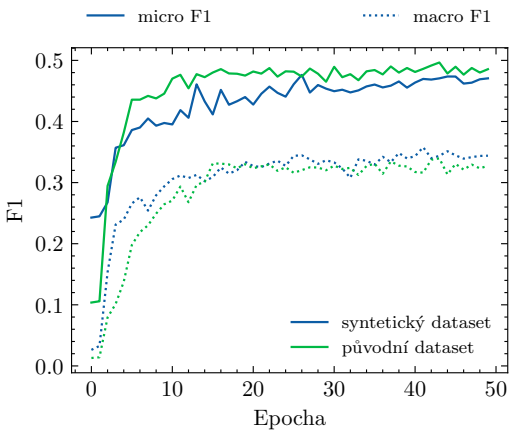
terapeutických intervencí. Pro nasazení systému je však nutné dále rozšířit trénovací sadu tak, aby výsledky byly dostatečně věrohodné. Je pravděpodobné, že s vylepšeným systémem ASR viz předchozí sekce 5.4, bude významně snížena doba nutná pro manuální opravu textového přepisu. Celkově tak bude anotace terapeutických intervencí mnohem rychlejší a bude tedy možné efektivněji získat data nutná pro dotrénování modelu pro věrohodnou klasifikaci terapeutických intervencí.

5.7 Souhrn

V této kapitole byly postupně představeny experimenty provedené s modely pro detekci řečové aktivity, diarizaci, automatické rozpoznávání řeči, detekci překrývající se řeči, klasifikaci sentimentu a klasifikaci typů terapeutických intervencí. Dosažené výsledky jsou shrnuty v tabulce 5.16. V rámci této kapitoly byla představena významná zlepšení v úlohách detekce řečové aktivity a rozpoznávání řeči a kroky, které k těmto zlepšením vedly. Dále byly dotrénovány *end-to-end* systémy pro diarizaci a detekci překrývající se řeči. Byla tak navržena a experimentálně ověřena adaptace systému VBx pro konvergenci ke dvěma mluvčím. Rovněž byly natrénovány modely pro klasifikaci sentimentu a terapeutických intervencí, u kterých byla diskutována aktuální omezení a byly popsány kroky, které je v budoucnu vhodné provést pro získání kvalitnějších modelů.

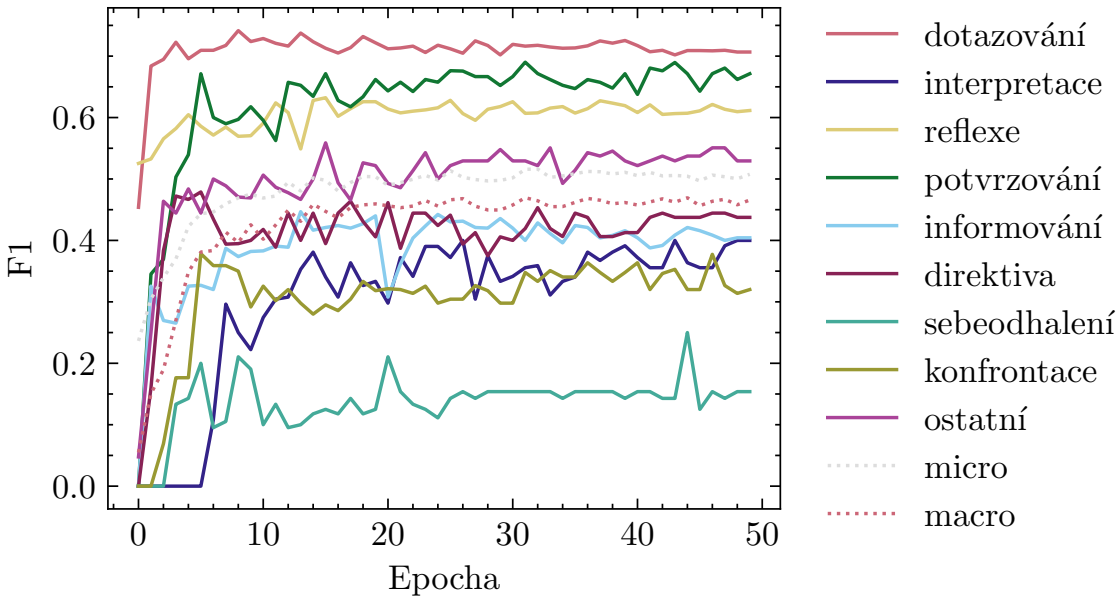


(a) Analýza vlivu přidání dalších 20 anotovaných nahrávek do trénovací sady.



(b) Analýza vlivu přidání synteticky dogenerovaných dat do trénovací sady.

Obrázek 5.15: Experimenty provedené s přidáním dat do původně navrženého trénovacího datasetu *Train1*. Validační dataset *Val1* byl zachován v původní formě tak, aby byly experimenty navzájem kompatibilní.



Obrázek 5.16: Porovnání F1 všech nadtříd a zároveň micro a macro průměru F1. Model byl trénován na trénovací sadě *Val3* a vyhodnocen na validační sadě *Val3*.

Tabulka 5.16: Souhrnná tabulka nejlepších dosažených výsledků v příslušných úlohách.

Úloha	Systém	Trénovací sada	Metrika	Hodnota
Detekce řečové aktivity	PyanNet	DeePsyTrain	Detection error rate [%] ↓	2,99
Diarizace	Adaptované VBx využívající identitu terapeuta	-	Diarization error rate [%] ↓	6,10
Detekce překrývající se řeči	PyanNet	DeePsyTest 1/2	F1 skóre ↑	0,49
Automatické rozpoznávání řeči	XLS-R-300m + GPT2	ASRCorpora + DeePsyTrain	Word error rate [%] ↓	23,47
Klasifikace sentimentu	CZERT	TwitterEmotions + Facebook	macro F1 ↑	0,45
Klasifikace terapeutických intervencí	FERNET	DeePsy-Interventions	macro F1 ↑	0,47

Kapitola 6

Extrahované příznaky

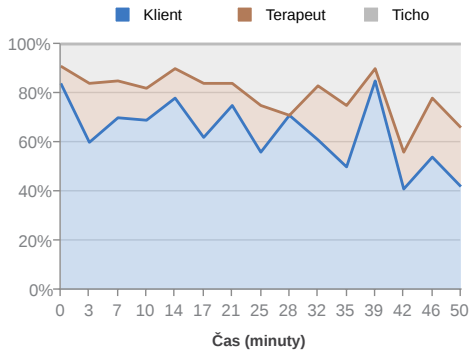
V rámci této kapitoly jsou čtenáři představeny extrahované příznaky z psychoterapuetických sezení, jež jsou jedním z výstupů této diplomové práce. Některé z příznaků byly již zakomponovány do systému DeePsy a příslušné ukázky pocházejí z automatické analýzy reálných sezení. U ostatních příznaků, které byly implementovány v rámci této práce, ale nebyly zatím nasazeny do systému, je představen návrh jejich grafické podoby. Grafické zobrazení získaných příznaků bylo iterativně diskutováno s psychoterapeuty a uživateli aplikace DeePsy. Příznaky jsou postupně extrahovány v pořadí odpovídajícím grafu implementovaného systému DeePsy viz obr. 2.1. V rámci této práce byl implementován back-end tohoto systému, který postupně zpracovává vstupní nahrávku a jehož výstupem je soubor v podobě XML dokumentu, který je uživateli k dispozici na webové stránce DeePsy.cz

Prvním z extrahovaných příznaků je **poměr řeči příslušných mluvčích**. Jedná se o příznak, který je získán po detekci řečové aktivity a následné diarizaci, kdy je známo kdo, kdy mluvil. Pro účely detekce řečové aktivity je využita dotrénovaná síť PyanNet viz sekce 5.1. Diarizace je provedena s využitím adaptovaného nástroje VBx viz sekce 5.2. Statistika poměru řeči byla nejdříve zobrazena v podobě koláčového grafu. Toto zobrazení však vedlo k příliš velké abstrakci údajů. Z tohoto důvodu bylo navrženo zobrazení v podobě spojnicového grafu viz obr. 6.1. Nahrávka je rozdělena do 15 segmentů a tato statistika je spočtena v rámci segmentů. Následně je vynesena do spojnicového grafu. Tento graf přináší terapeutovi první velmi obecnou charakteristiku proběhlého sezení.

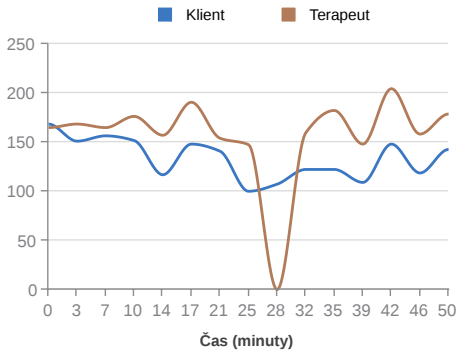
Dalšími možnými charakteristikami extrahovanými přímo ze signálu a časových značek rozlišujících příslušné mluvčí je **reakční doba** a **průměrná energie mluvčího**. Reakční doba byla vynesena do histogramu a byla analyzována doba s jakou příslušní mluvčí reagují na předchozí segmenty. Průměrná energie byla vynesena do spojnicového grafu porovnávajícího průměrnou energii terapeuta a klienta v průběhu sezení. Příznaky byly zaintegrovány do systému. Po důkladné analýze zpětné vazby terapeutů však bylo vyhodnoceno, že tyto charakteristiky pro terapeuty přinášejí až příliš obecnou informaci, kterou z jejich pohledu není možné přímo interpretovat a korelovat s průběhem daného sezení. Z tohoto důvodu jsou tyto příznaky aktuálně skryty v systému.

Mezi další charakteristiky rozhovoru, které je možné extrahovat přímo po diarizaci je **počet skoků do řeči** za jednotku času. V kontextu psychoterapeutických sezení je často skoky do řeči kontrolovat průběh terapie, čímž je sezení vedeno správným směrem. Jelikož dosažené výsledky detekce skoků do řeči v sekci 5.3 zatím nedosáhly potřebné věrohodnosti, tento příznak zatím nebyl zakomponován do systému.

Další z příznaků, které byly dále extrahovány, vyžadují již **textový přepis nahrávky**. Samotný přepis nahrávky je brán jako stěžejní prvek vybudované aplikace a lze ho pova-



Obrázek 6.1: Jak moc kdo mluví během sezení? Graf ukazuje v procentech, jaký prostor zabírala řeč klienta, terapeuta a nebo ticho.



Obrázek 6.2: Rychlost řeči. Graf ukazuje, jak rychle mluvili klient a terapeut. Rychlost řeči je vyjádřena jako průměrné množství slov za minutu.

žovat jako jeden z příznaků. Přepis terapeutické konverzace je velmi cenný a může sloužit nejen jako podklad pro pozdější reflexi průběhu sezení, ale také pro supervizní účely. Analýza komplexnějších jazykových příznaků je přímo závislá na extrakci automatického přepisu. Textový přepis je uživatelům prezentován v podobě bloků obsahujících časové značky, mluvčího a přepis daného segmentu. Je zde možnost přehrání záznamu dané nahrávky, které je synchronizováno s textovým přepisem viz obr. 6.3.

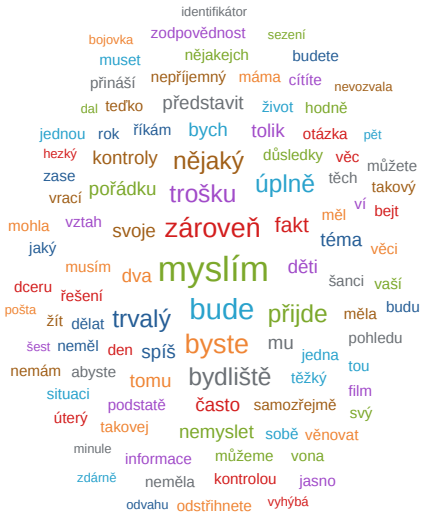
Po získání textového přepisu a znalosti výstupu diarizace dochází k přiřazení příslušných textových segmentů daným mluvčím. Následně dochází k postprocessingu výstupů ASR pomocí dvouvrstvé BiLSTM dotrénované v rámci tohoto projektu. Promluvy jsou doplněny o interpunkční znaménka a převedeny do větné podoby. Nad takto oddělenými promluvami příslušných řečníků nejdříve dochází k výpočtu **rychlosti řeči** jako průměrného počtu slov za minutu viz obr. 6.2.

Textové přepisy jsou dále lematizovány a je provedena morfologická analýza textu s pomocí nástroje MorphoDiTa [131]. Tímto způsobem je získán přehled o slovních druzích a slovesných časech příslušných vět. Souhrnou analýzou je taktéž získána informace o frekvenci příslušných slov v rámci sezení. **Klíčová slova**, tedy ta nejčetnější, jsou zobrazena ve formě „mračna slov“ [52]. Toto zobrazení viz obr. 6.4 lze chápat jako určitou formu shrnutí obsahu proběhlého sezení. „Mračno slov“ navíc poskytuje základní informaci o jazykových prostředcích použitých v rámci daného sezení. Terapeut může analyzovat, jaká výplňová slova používá přespříliš, případně si jejich zobrazování deaktivovat a zaměřit se pouze na slova s věcným významem.

Informace o slovesném čase příslušných slov, respektive vět, extrahována pomocí nástroje MorphoDiTa je taktéž zprůměrována napříč 15 segmenty sezení a zobrazena ve formě spojnicového grafu viz obr. 6.5. Samotná **informace o časovém horizontu** sezení neříká nic o tom, zda je kvalitativně jedno sezení lepší než druhé. Může být však dobrým vodítkem pro další rozvahu a plánování budoucích sezení. Slovesný čas je obzvláště důležité pro specifické terapeutické směry, které pracují s časovými perspektivami [152].

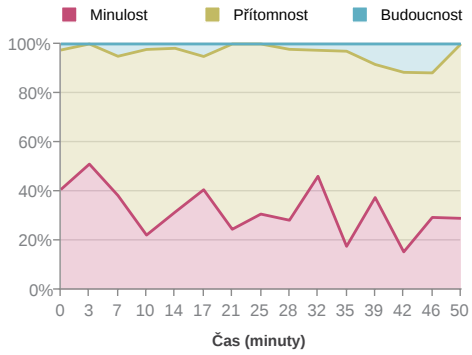
Další z příznaků, jež byly implementovány v rámci této práce a přímo zahrnuty do systému DeePsy, je detekce **základního emočního zabarvení** příslušných promluv [96]. K analýze segmentu dochází s pomocí dotrénované sítě CZERT viz sekce 5.5. Obdobně

Obrázek 6.3: Na obrázku je zobrazen přepis nahrávky terapeutického sezení v systému DeePsy. Přepis je synchronizovaný s přehrávanou nahrávkou, automaticky dochází ke zvýraznění věty, která je aktuálně přehrávána. Uživatel má k dispozici posuvník, kterým dokáže zakrýt slova, jež mají nízkou konfidenci.

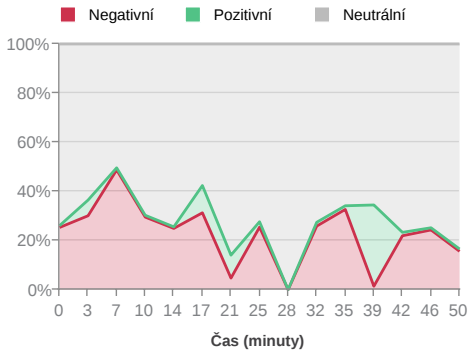


Obrázek 6.4: Nejčastěji používaná slova. Barva slova označuje slovní druh. Velikost písma označuje frekvenci použití daného slova v rámci psychoterapeutické seance.

jako slovesný čas je sentiment zobrazen agregovaně napříč segmenty v podobě spojnicového grafu viz obr. 6.6.



Obrázek 6.5: Časová perspektiva. Graf ukazuje v procentech, jaký podíl sloves byl formulován v minulém, přítomném nebo budoucím čase.

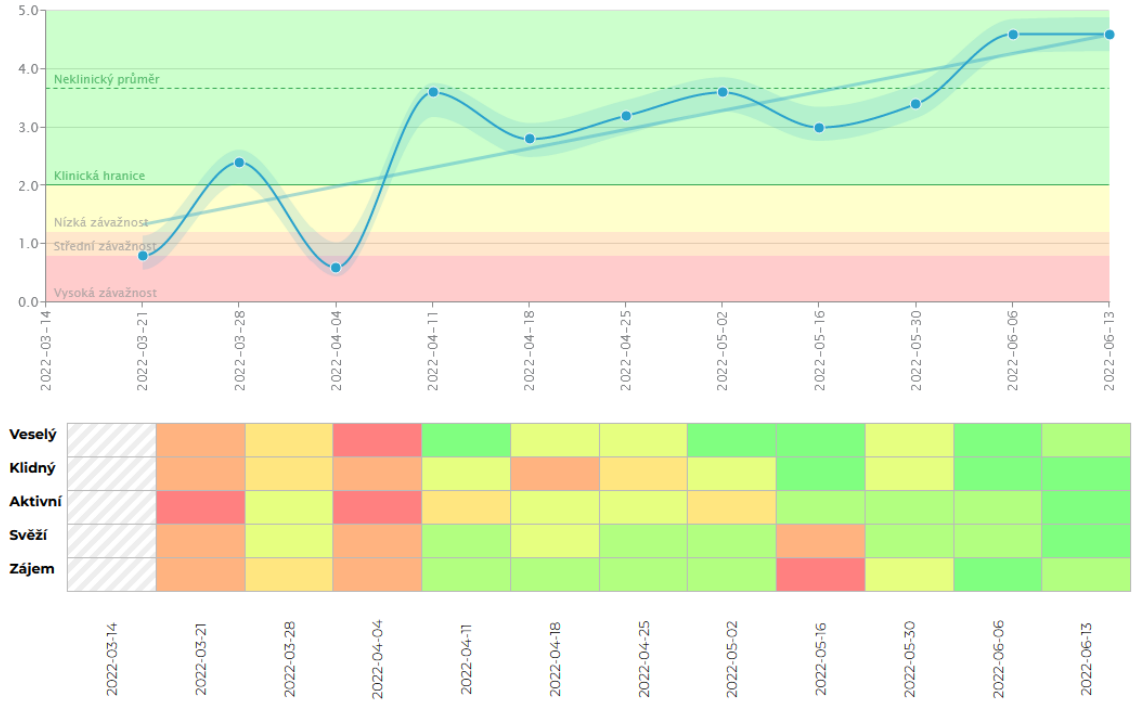


Obrázek 6.6: Emoce pojmenované v řeči. Graf ukazuje v procentech, jaký podíl řeči byl neutrálně, negativně, nebo pozitivně emočně zabarvený.

V rámci této práce byl taktéž navrhnut protokol **klasifikace terapeutických intervencí**. Model navržený v rámci sekce 5.6 zatím nebyl zakomponován do systému, jelikož dosažená úspěšnost modelu zatím není dostačující a je nutná anotace dalších dat. Obdobně jako v předchozích případech je navrženo zobrazení této metriky pomocí spojnicového grafu napříč kategoriemi a segmenty.

Dalším implementovaným příznakem je **jazyková variabilita** příslušných mluvčích, čili bohatost jazykových prostředků, které zejména terapeut během sezení využívá. Větší jazyková diverzita terapeuta může signalizovat přirozenější interakce a řešení specifických situací při terapii, což se následně promítá do efektivity terapeutické práce [148]. Příbuzným aspektem je také sledování jazykové koordinace mezi klientem a terapeutem, tedy míry používání podobných slov a vět, která může být chápána jako implicitní znak dobrého pracovního spojenectví [1, 68]. V rámci systému je spočtená **lexikální a sémantická podobnost** na úrovni slov a celých vět. Je k tomu využit model CZERT předtrénovaný na datasetu sémantických podobností novinových článků v českém jazyce [128]. Sémantická podobnost na úrovni slov je spočtena extrakcí skrytých stavů po N inferencích modelem CZERT a následným spočtením kosinové podobnosti $M \times M$, kde N je počet vět a M počet slov v rámci daného sezení. Pro získání sémantické podobnosti je provedeno N^2 inferencí modelem. Lexikální podobnost je taktéž spočtena na úrovní slov i vět, navíc je také spočtena na úrovní lemmatizovaných slov. Dále je spočtena agregována jazyková podobnost s kontextem dvou a pěti vět v minulosti. Je taktéž uvážena varianta s neomezeným kontextem do minulosti. Příznak je aktuálně navrženy pro integraci do systému a čeká na schválení ze strany terapeutů.

DeePsy navíc disponuje propracovaným **systémem dotazníků**, které pomáhají terapeutovi k základnímu skríninku nejběžnějších psychopatologií, průběžnému sledování míry klientových potíží, průběžnému sledování reakcí klienta na sezení a zjišťování klientových preferencí ve vztahu ke způsobu vedení terapie [152]. Jedním z dotazníku je WHO-5 [9], které měří duševní pohodu v rámci psychoterapeutické epizody viz obr. 6.7.



Obrázek 6.7: Duševní pohoda (WHO-5) sledována v rámci terapeutické epizody. Kromě samotné naměřené hodnoty znázorňuje také normy, které umožňují stav klienta orientačně porovnat s populací.

Kapitola 7

Závěr

Cílem této diplomové práce byla analýza entit v rámci psychoterapeutických sezení. Čtenář je nejprve seznámen s projektem DeePsy, v jehož rámci tato práce vznikla, následuje samotná motivace pro automatickou extrakci příznaků z psychoterapeutických sezení. Následně je uvedena potřebná teorie z oblasti zpracování řeči a přirozeného jazyka. Postupně jsou představeny řešené úlohy a data pro získání charakteristik terapeutických sezení.

Byla uskutečněna analýza volně dostupných modelů pro detekci řečové aktivity, jejich doménová adaptace a následné vyhodnocení na testovací sadě. Nejlepší natrénovaná varianta sítě PyanNet dosáhla chybovosti detekce řečové aktivity 2,99 %, což je relativní zlepšení o 37,82 % vůči *baseline* systému. Pro účely diarizace byla uskutečněna komparace standardního VBx a end2end systému pyannote. Zobecněním diarizačního systému VBx pomocí předem extrahovaných hlasových otisků terapeutů byla zajištěna konvergence systému ke dvěma mluvčím s minimálním relativním zhoršením chybovosti diarizace o 0,66 % vůči původní variantě. Dále proběhla studie předtrénovaných modelů typu Wav2Vec2 a Whisper pro automatické rozpoznávání řeči. Bylo natrénováno několik variant těchto systémů a byla provedena analýza vlivů faktorů jako augmentace, rozšíření trénovací sady o doménová data a inkorporace jazykových modelů. Byl realizován pokus o reskórování hypotéz ASR systému pomocí externího jazykového modelu GPT2. Tento experiment vedl k zakomponování tohoto modelu do vlastní seq2seq architektury. Bylo navrženo vlastní trénovací schéma pro sloučení předtrénovaných znalostí modelů XLS-R a GPT2. Nejlepší natrénovaný systém pro automatické rozpoznávání řeči dosáhl chybovosti WER 23,47 %, což je relativní zlepšení chybovosti o 17,06 % oproti nejlepšímu dostupnému hybridní modelu CNN-TDNN-HMM.

Mezi další experimenty provedené již nad textovými daty patří klasifikace sentimentu a terapeutických intervencí. Pro natrénování modelů byly vytvořeny nové trénovací sady TwitterEmotions a DeePsyInterventions. Dotrénovaný model CZERT dosáhl přesnosti klasifikace sentimentu v psychoterapeutických promluvách 79 %. Byly taktéž prozkoumány a dotrénovány modely pro maskované jazykové modelování a následnou klasifikaci terapeutických intervencí. Byl analyzován vliv kontextu, jakož i byla uskutečněna studie chybovosti modelů, která motivuje další sběr anotovaných dat. Byly provedeny pokusy o detekci překrývající se řeči, bylo dosaženo F1 skóre 0,49. Vyjma výstupů zmíněných systému byly taktéž extrahovány příznaky jako slovesný čas či jazyková podobnost řeči. Po četných konzultacích s terapeuty byla navržena sada příznaků, které byly zaintegrovány do systému DeePsy a jsou viditelné terapeutům po přihlášení do portálu. Provedené analýzy a experimenty ukazují významnost použití předtrénovaných modelů a zlepšení získána jejich doladěním na cílových datech. Celkově tato práce přispívá k vývoji spolehlivého asistenčního nástroje pro zvýšení kvality psychoterapeutických sezení a podporu profesního růstu terapeutů.

Literatura

[1] Aafjes Van Doorn, K., Porcerelli, J. a Müller Frommeyer, L. Language style matching in psychotherapy: An implicit aspect of alliance. Červenec 2020, sv. 67, s. 509–522. DOI: 10.1037/cou0000433.

[2] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M. et al. *Common Voice: A Massively-Multilingual Speech Corpus*. 2020.

[3] Ba, J. L., Kiros, J. R. a Hinton, G. E. *Layer Normalization*. arXiv, 2016. DOI: 10.48550/ARXIV.1607.06450. Dostupné z: https://arxiv.org/abs/1607.06450.

[4] Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q. et al. *XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale*. 2021.

[5] Baevski, A., Schneider, S. a Auli, M. Vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations. 2019, abs/1910.05453. Dostupné z: http://arxiv.org/abs/1910.05453.

[6] Baevski, A., Zhou, H., Mohamed, A. a Auli, M. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. 2020, abs/2006.11477. Dostupné z: https://arxiv.org/abs/2006.11477.

[7] Bahdanau, D., Cho, K. a Bengio, Y. *Neural Machine Translation by Jointly Learning to Align and Translate*. arXiv, 2014. DOI: 10.48550/ARXIV.1409.0473. Dostupné z: https://arxiv.org/abs/1409.0473.

[8] Bahdanau, D., Cho, K. a Bengio, Y. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016.

[9] Bech, P., Olsen, L. R., Kjoller, M. a Rasmussen, N. K. Measuring well-being rather than the absence of distress symptoms: a comparison of the SF-36 Mental Health subscale and the WHO-Five Well-Being Scale. United States: [b.n.]. 2003, sv. 12, č. 2, s. 85–91.

[10] Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V. et al. *Prague Dependency Treebank 3.0*. Prague, Czech republic: Univerzita Karlova v Praze, MFF, ÚFAL, 2013.

[11] Bengio, Y., Ducharme, R., Vincent, P. a Janvin, C. A Neural Probabilistic Language Model. JMLR.org. mar 2003, sv. 3, null, s. 1137–1155. ISSN 1532-4435.

[12] Berger, A. L., Pietra, V. J. D. a Pietra, S. A. D. A Maximum Entropy Approach to Natural Language Processing. Cambridge, MA, USA: MIT Press. mar 1996, sv. 22, č. 1, s. 39–71. ISSN 0891-2017.

[13] Bishop, C. M. a Nasrabadi, N. M. *Pattern recognition and machine learning*. Springer, 2006.

[14] Bogatinovski, J., Todorovski, L., Džeroski, S. a Kocev, D. Comprehensive comparative study of multi-label classification methods. 2022, sv. 203, s. 117215. DOI: https://doi.org/10.1016/j.eswa.2022.117215. ISSN 0957-4174. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0957417422005991.

[15] Bojar, O., Kratochvíl, J. a Polak, P. Large Corpus of Czech Parliament Plenary Hearings. In: *International Conference on Language Resources and Evaluation*. 2020.

[16] Bredin, H. *Pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe*. Oct 2022. Dostupné z: https://huggingface.co/pyannote/speaker-diarization/resolve/main/technical_report_2.1.pdf.

[17] Bredin, H. a Laurent, A. *End-to-end speaker segmentation for overlap-aware resegmentation*. 2021.

[18] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. et al. Language Models are Few-Shot Learners. 2020, abs/2005.14165. Dostupné z: https://arxiv.org/abs/2005.14165.

[19] Bäckström, T., Räsänen, O., Zewoudie, A., Zarazaga, P. P., Koivusalo, L. et al. *Introduction to Speech Processing*. 2. vyd. 2022. Dostupné z: https://speechprocessingbook.aalto.fi.

[20] Cao, J., Tanana, M., Imel, Z. E., Poitras, E., Atkins, D. C. et al. *Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes*. 2019.

[21] Carletta, J. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. May 2007, sv. 41, č. 2, s. 181–190. DOI: 10.1007/s10579-007-9040-x. ISSN 1572-8412. Dostupné z: https://doi.org/10.1007/s10579-007-9040-x.

[22] Chaloupský, L. *Automatic generation of medical reports from chest X-rays in Czech*. Praha, CZ, 2022. Diplomová práce. Univerzita Karlova, Matematicko-fyzikální fakulta, Ústav formální a aplikované lingvistiky. Dostupné z: http://hdl.handle.net/20.500.11956/176356.

[23] Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S. et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. Institute of Electrical and Electronics Engineers (IEEE). oct 2022, sv. 16, č. 6, s. 1505–1518. DOI: 10.1109/jstsp.2022.3188113. Dostupné z: https://doi.org/10.1109%2Fjstsp.2022.3188113.

[24] Cheng, J., Dong, L. a Lapata, M. *Long Short-Term Memory-Networks for Machine Reading*. arXiv, 2016. DOI: 10.48550/ARXIV.1601.06733. Dostupné z: https://arxiv.org/abs/1601.06733.

[25] Cho, K., Merrienboer, B. van, Gülçehre, Ç., Bougares, F., Schwenk, H. et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical

Machine Translation. 2014, abs/1406.1078. Dostupné z: http://arxiv.org/abs/1406.1078.

[26] Chung, J. S., Huh, J., Nagrani, A., Afouras, T. a Zisserman, A. Spot the Conversation: Speaker Diarisation in the Wild. In: *Interspeech 2020*. ISCA, Oct 2020. DOI: 10.21437/interspeech.2020-2337. Dostupné z: https://doi.org/10.21437%2Finterspeech.2020-2337.

[27] Chung, J., Gulcehre, C., Cho, K. a Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. Prosinec 2014.

[28] Clark, K., Luong, M., Le, Q. V. a Manning, C. D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. 2020, abs/2003.10555. Dostupné z: https://arxiv.org/abs/2003.10555.

[29] Dai, Z., Yang, Z., Yang, Y., Carbonell, J. G., Le, Q. V. et al. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. 2019, abs/1901.02860. Dostupné z: http://arxiv.org/abs/1901.02860.

[30] Davis, S. a Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. 1980, sv. 28, č. 4, s. 357–366. DOI: 10.1109/TASSP.1980.1163420.

[31] Desplanques, B., Thienpondt, J. a Demuynck, K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In: *Interspeech 2020*. ISCA, Oct 2020. DOI: 10.21437/interspeech.2020-2650. Dostupné z: https://doi.org/10.21437%2Finterspeech.2020-2650.

[32] Devlin, J., Chang, M., Lee, K. a Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018, abs/1810.04805. Dostupné z: http://arxiv.org/abs/1810.04805.

[33] Dinkel, H., Chen, Y., Wu, M. a Yu, K. *Voice activity detection in the wild via weakly supervised sound event detection*. 2020.

[34] Doorn, K. A. van, Kamsteeg, C., Bate, J. a Aafjes, M. A scoping review of machine learning in psychotherapy research. Routledge. 2021, sv. 31, č. 1, s. 92–116. DOI: 10.1080/10503307.2020.1808729. PMID: 32862761. Dostupné z: https://doi.org/10.1080/10503307.2020.1808729.

[35] Dubey, S. R., Singh, S. K. a Chaudhuri, B. B. *Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark*. 2022.

[36] Ewbank, M. P., Cummins, R., Tablan, V., Catarino, A., Buchholz, S. et al. Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. Routledge. 2021, sv. 31, č. 3, s. 300–312. DOI: 10.1080/10503307.2020.1788740. PMID: 32619163. Dostupné z: https://doi.org/10.1080/10503307.2020.1788740.

[37] Falcon, W. a Cho, K. A Framework For Contrastive Self-Supervised Learning

And Designing A New Approach. 2020, abs/2009.00104. Dostupné z: https://arxiv.org/abs/2009.00104.

[38] Flemotomos, N., Martinez, V., Chen, Z., Creed, T., Atkins, D. et al. Automated quality assessment of cognitive behavioral therapy sessions through highly contextualized language representations. Říjen 2021, sv. 16, s. e0258639. DOI: 10.1371/journal.pone.0258639.

[39] Fukushima, K. Neocognitron: A hierarchical neural network capable of visual pattern recognition. 1988, sv. 1, č. 2, s. 119–130. DOI: https://doi.org/10.1016/0893-6080(88)90014-7. ISSN 0893-6080. Dostupné z: https://www.sciencedirect.com/science/article/pii/0893608088900147.

[40] Gage, P. A new algorithm for data compression. 1994, sv. 12, s. 23–38.

[41] Gales, M., Knill, K., Ragni, A. a Rath, S. Speech recognition and keyword spotting for low-resource languages : Babel project research at CUED. In:. Květen 2014.

[42] Gehring, J., Auli, M., Grangier, D., Yarats, D. a Dauphin, Y. N. *Convolutional Sequence to Sequence Learning*. 2017.

[43] Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W. et al. Audio Set: An ontology and human-labeled dataset for audio events. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, s. 776–780. DOI: 10.1109/ICASSP.2017.7952261.

[44] Ghahramani, Z. An Introduction to Hidden Markov Models and Bayesian Networks. In: *Hidden Markov Models: Applications in Computer Vision*. USA: World Scientific Publishing Co., Inc., 2001, s. 9–42. ISBN 9810245645.

[45] Glembek, O., Karafiát, M., Burget, L. a Černocký, J. Czech Speech Recognizer for Multiple Environments. In: *Radioeletronika 2006*. 2006, s. 1–4. Dostupné z: https://www.fit.vut.cz/research/publication/8219.

[46] Goldberg, S. B., Flemotomos, N., Martinez, V. R., Tanana, M. J., Kuo, P. B. et al. Machine learning and natural language processing in psychotherapy research: Alliance as example use case. American Psychological Association. 2020, sv. 67, č. 4, s. 438.

[47] Graves, A. *Sequence Transduction with Recurrent Neural Networks*. 2012.

[48] Graves, A., Fernández, S., Gomez, F. a Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks. In:. Leden 2006, sv. 2006, s. 369–376. DOI: 10.1145/1143844.1143891.

[49] Habernal, I., Ptáček, T. a Steinberger, J. Sentiment Analysis in Czech Social Media Using Supervised Machine Learning. In: *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Atlanta, Georgia: Association for Computational Linguistics, červen 2013, s. 65–74. Dostupné z: https://aclanthology.org/W13-1609.

[50] He, K., Zhang, X., Ren, S. a Sun, J. *Deep Residual Learning for Image Recognition*. arXiv, 2015. DOI: 10.48550/ARXIV.1512.03385. Dostupné z: https://arxiv.org/abs/1512.03385.

[51] He, P., Liu, X., Gao, J. a Chen, W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. 2020, abs/2006.03654. Dostupné z: https://arxiv.org/abs/2006.03654.

[52] Heimerl, F., Lohmann, S., Lange, S. a Ertl, T. Word Cloud Explorer: Text Analytics Based on Word Clouds. In: *2014 47th Hawaii International Conference on System Sciences*. 2014, s. 1833–1842. DOI: 10.1109/HICSS.2014.231.

[53] Hendrycks, D. a Gimpel, K. Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units. 2016, abs/1606.08415. Dostupné z: http://arxiv.org/abs/1606.08415.

[54] Hochreiter, S. a Schmidhuber, J. Long Short-term Memory. Prosinec 1997, sv. 9, s. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

[55] Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. 1982, sv. 79, č. 8, s. 2554–2558. DOI: 10.1073/pnas.79.8.2554. Dostupné z: https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554.

[56] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R. et al. *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units*. 2021.

[57] Huang, P.-Y., Xu, H., Li, J., Baevski, A., Auli, M. et al. *Masked Autoencoders that Listen*. arXiv, 2022. DOI: 10.48550/ARXIV.2207.06405. Dostupné z: https://arxiv.org/abs/2207.06405.

[58] Hugging Face. *The Hugging Face Course, 2022*. 2022. Dostupné z: https://huggingface.co/course.

[59] Hájek, A. *Automatic text summarization [online]*. 2021 [cit. 2023-05-08]. SUPERVISOR : Aleš Horák. Dostupné z: https://is.muni.cz/th/jsw6t/.

[60] Ioffe, S. Probabilistic Linear Discriminant Analysis. In: Leonardis, A., Bischof, H. a Pinz, A., ed. *Computer Vision – ECCV 2006*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, s. 531–542. ISBN 978-3-540-33839-0.

[61] Jia, F., Majumdar, S. a Ginsburg, B. *MarbleNet: Deep 1D Time-Channel Separable Convolutional Neural Network for Voice Activity Detection*. 2021.

[62] Jurafsky, D. a Martin, J. H. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.: Pearson Prentice Hall, 2009. ISBN 9780131873216 0131873210. Dostupné z: http://www.amazon.com/Speech-Language-Processing-2nd-Edition/dp/0131873210/ref=pd_bxgy_b_img_y.

[63] Karafiát, M., Baskar, M. K., Matejka, P., Veselý, K., Grézl, F. et al. 2016

BUT Babel System: Multilingual BLSTM Acoustic Model with i-Vector Based Adaptation. In:. Srpen 2017, s. 719–723. DOI: 10.21437/Interspeech.2017-1775.

[64] Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C. a Socher, R. CTRL: A Conditional Transformer Language Model for Controllable Generation. 2019, abs/1909.05858. Dostupné z: http://arxiv.org/abs/1909.05858.

[65] Kingma, D. P. a Ba, J. *Adam: A Method for Stochastic Optimization*. 2017.

[66] Kocián, M., Náplava, J., Štancl, D. a Kadlec, V. *Siamese BERT-based Model for Web Search Relevance Ranking Evaluated on a New Czech Dataset*. 2021.

[67] Kocour, M., Umesh, J., Karafiat, M., Švec, J., López, F. et al. BCN2BRNO: ASR System Fusion for Albayzin 2022 Speech to Text Challenge. In: *Proc. IberSPEECH 2022*. 2022, s. 276–280. DOI: 10.21437/IberSPEECH.2022-56.

[68] Koole, S. L. a Tschacher, W. Synchrony in psychotherapy: A review and an integrative framework for the therapeutic alliance. Frontiers Media SA. červen 2016, sv. 7, s. 862.

[69] Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M. et al. *SYN v4: large corpus of written Czech*. 2016. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Dostupné z: http://hdl.handle.net/11234/1-1846.

[70] Krizhevsky, A., Sutskever, I. a Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira, F., Burges, C., Bottou, L. a Weinberger, K., ed. *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012, sv. 25. Dostupné z: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[71] Kudo, T. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. 2018, abs/1804.10959. Dostupné z: http://arxiv.org/abs/1804.10959.

[72] Kunešová, M., Hrúz, M., Zajíc, Z. a Radová, V. Detection of Overlapping Speech for the Purposes of Speaker Diarization. In: Salah, A. A., Karpov, A. a Potapova, R., ed. *Speech and Computer*. Cham: Springer International Publishing, 2019, s. 247–257. ISBN 978-3-030-26061-3.

[73] Lai, Y., Tang, X., Fu, Y. a Fang, R. End-to-end speaker diarization with transformer. 2021, abs/2112.07463. Dostupné z: https://arxiv.org/abs/2112.07463.

[74] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. 2019, abs/1909.11942. Dostupné z: http://arxiv.org/abs/1909.11942.

[75] Landini, F., Profant, J., Diez, M. a Burget, L. Bayesian HMM Clustering of X-Vector Sequences (VBx) in Speaker Diarization: Theory, Implementation and Analysis on Standard Tasks. GBR: Academic Press Ltd. jan 2022, sv. 71, C. DOI: 10.1016/j.csl.2021.101254. ISSN 0885-2308. Dostupné z: https://doi.org/10.1016/j.csl.2021.101254.

[76] Le, B. a Nguyen, H. Twitter Sentiment Analysis Using Machine Learning Techniques. In: Le Thi, H. A., Nguyen, N. T. a Do, T. V., ed. *Advanced Computational Methods for Knowledge Engineering*. Cham: Springer International Publishing, 2015, s. 279–289.

[77] Lecun, Y., Bottou, L., Bengio, Y. a Haffner, P. Gradient-based learning applied to document recognition. 1998, sv. 86, č. 11, s. 2278–2324. DOI: 10.1109/5.726791.

[78] Lehečka, J. a Švec, J. Comparison of Czech Transformers on Text Classification Tasks. In: Espinosa Anke, L., Martín Vide, C. a Spasić, I., ed. *Statistical Language and Speech Processing*. Cham: Springer International Publishing, 2021, s. 27–37. ISBN 978-3-030-89579-2.

[79] Lehečka, J., Švec, J., Prazak, A. a Psutka, J. Exploring Capabilities of Monolingual Audio Transformers using Large Datasets in Automatic Speech Recognition of Czech. In: *Interspeech 2022*. ISCA, Sep 2022. DOI: 10.21437/interspeech.2022-10439. Dostupné z: https://doi.org/10.21437%2Finterspeech.2022-10439.

[80] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A. et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. 2019, abs/1910.13461. Dostupné z: http://arxiv.org/abs/1910.13461.

[81] Li, J. *Recent Advances in End-to-End Automatic Speech Recognition*. 2022.

[82] Lison, P. a Tiedemann, J. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), Květen 2016, s. 923–929. Dostupné z: https://aclanthology.org/L16-1147.

[83] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S. et al. Multilingual Denoising Pre-training for Neural Machine Translation. 2020, abs/2001.08210. Dostupné z: https://arxiv.org/abs/2001.08210.

[84] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M. et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019, abs/1907.11692. Dostupné z: http://arxiv.org/abs/1907.11692.

[85] Loshchilov, I. a Hutter, F. *Decoupled Weight Decay Regularization*. 2019.

[86] Loulová Štěpánka. *Klasifikační systém pro počítačové zpracování terapeutovy řeči v rámci individuálních psychoterapeutických sezení*. 2022. Diplomová práce. Masarykova univerzita, Fakulta sociálních studií, Brno. Dostupné z: https://is.muni.cz/th/f0ux5/.

[87] Luong, M., Pham, H. a Manning, C. D. Effective Approaches to Attention-based Neural Machine Translation. 2015, abs/1508.04025. Dostupné z: http://arxiv.org/abs/1508.04025.

[88] Mcculloch, W. a Pitts, W. A Logical Calculus of Ideas Immanent in Nervous Activity. 1943, sv. 5, s. 127–147.

[89] McHugh, M. Interrater reliability: The kappa statistic. Říjen 2012, sv. 22, s. 276–82. DOI: 10.11613/BM.2012.031.

[90] Mikolov, T., Chen, K., Corrado, G. a Dean, J. *Efficient Estimation of Word Representations in Vector Space*. arXiv, 2013. DOI: 10.48550/ARXIV.1301.3781. Dostupné z: https://arxiv.org/abs/1301.3781.

[91] Miner, A. S., Haque, A., Fries, J. A., Fleming, S. L., Wilfley, D. E. et al. Assessing the accuracy of automatic speech recognition for psychotherapy. Jun 2020, sv. 3, č. 1, s. 82. DOI: 10.1038/s41746-020-0285-8. ISSN 2398-6352. Dostupné z: https://doi.org/10.1038/s41746-020-0285-8.

[92] Mitchell, T. *Machine learning*. McGraw-hill New York, 1997.

[93] Moyers, T. B., Rowell, L. N., Manuel, J. K., Ernst, D. a Houck, J. M. The Motivational Interviewing Treatment Integrity Code (MITI 4): Rationale, Preliminary Reliability and Validity. 2016, sv. 65, s. 36–42. DOI: https://doi.org/10.1016/j.jsat.2016.01.001. ISSN 0740-5472. Motivational Interviewing in Substance Use Treatment. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0740547216000143.

[94] Niu, S.-T., Du, J., Sun, L. a Lee, C.-H. *Separation Guided Speaker Diarization in Realistic Mismatched Conditions*. 2021.

[95] OpenAI. *GPT-4 Technical Report*. 2023.

[96] Pang, B. a Lee, L. Opinion Mining and Sentiment Analysis. Leden 2008, sv. 2, s. 1–135. DOI: 10.1561/1500000011.

[97] Papineni, K., Roukos, S., Ward, T. a Zhu, W.-J. Bleu: a Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, červenec 2002, s. 311–318. DOI: 10.3115/1073083.1073135. Dostupné z: https://aclanthology.org/P02-1040.

[98] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B. et al. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In: *Interspeech 2019*. ISCA, Sep 2019. DOI: 10.21437/interspeech.2019-2680. Dostupné z: https://doi.org/10.21437%2Finterspeech.2019-2680.

[99] Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S. et al. *A Review of Speaker Diarization: Recent Advances with Deep Learning*. arXiv, 2021. DOI: 10.48550/ARXIV.2101.09624. Dostupné z: https://arxiv.org/abs/2101.09624.

[100] Pascanu, R., Mikolov, T. a Bengio, Y. Understanding the exploding gradient problem. 2012, abs/1211.5063. Dostupné z: http://arxiv.org/abs/1211.5063.

[101] Peng, N. a Dredze, M. Named Entity Recognition for Chinese Social Media with Jointly Trained Embeddings. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for

Computational Linguistics, Září 2015, s. 548–554. DOI: 10.18653/v1/D15-1064. Dostupné z: https://aclanthology.org/D15-1064.

[102] Pennington, J., Socher, R. a Manning, C. GloVe: Global Vectors for Word Representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, říjen 2014, s. 1532–1543. DOI: 10.3115/v1/D14-1162. Dostupné z: https://aclanthology.org/D14-1162.

[103] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C. et al. Deep contextualized word representations. 2018, abs/1802.05365. Dostupné z: http://arxiv.org/abs/1802.05365.

[104] Pinto, J. V., Passos, I. C., Gomes, F., Reckziegel, R., Kapczinski, F. et al. Peripheral biomarker signatures of bipolar disorder and schizophrenia: A machine learning approach. 2017, sv. 188, s. 182–184. DOI: https://doi.org/10.1016/j.schres.2017.01.018. ISSN 0920-9964. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0920996417300233.

[105] Plchot, O., Matějka, P., Novotný, O., Cumani, S., Lozano, A. D. et al. Analysis of BUT-PT Submission for NIST LRE 2017. In: *Proceedings of Odyssey 2018 The Speaker and Language Recognition Workshop*. International Speech Communication Association, 2018, sv. 2018, č. 6, s. 47–53. DOI: 10.21437/Odyssey.2018-7. ISSN 2312-2846. Dostupné z: https://www.fit.vut.cz/research/publication/11762.

[106] Polok, A. *Analýza audio hovoru mezi dvěma účastníky*. Brno, CZ, 2021. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Dostupné z: https://www.fit.vut.cz/study/thesis/23343/.

[107] Popel, M., Tomkova, M., Tomek, J., Kaiser, Ł., Uszkoreit, J. et al. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. Sep 2020, sv. 11, č. 1, s. 4381. DOI: 10.1038/s41467-020-18073-9. ISSN 2041-1723. Dostupné z: https://doi.org/10.1038/s41467-020-18073-9.

[108] Pratap, V., Hannun, A. Y., Xu, Q., Cai, J., Kahn, J. et al. Wav2letter++: The Fastest Open-source Speech Recognition System. 2018, abs/1812.07625. Dostupné z: http://arxiv.org/abs/1812.07625.

[109] Pratap, V., Xu, Q., Sriram, A., Synnaeve, G. a Collobert, R. MLS: A Large-Scale Multilingual Dataset for Speech Research. In: *Interspeech 2020*. ISCA, Oct 2020. DOI: 10.21437/interspeech.2020-2826. Dostupné z: https://doi.org/10.21437%2Finterspeech.2020-2826.

[110] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022.

[111] Radford, A. a Narasimhan, K. Improving Language Understanding by Generative Pre-Training. In:. 2018.

[112] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. et al. Language Models

are Unsupervised Multitask Learners. In:. 2019.

[113] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S. et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. 2019, abs/1910.10683. Dostupné z: http://arxiv.org/abs/1910.10683.

[114] Ravanelli, M., Brakel, P., Omologo, M. a Bengio, Y. Light Gated Recurrent Units for Speech Recognition. Institute of Electrical and Electronics Engineers (IEEE). apr 2018, sv. 2, č. 2, s. 92–102. DOI: 10.1109/tetci.2017.2762739. Dostupné z: https://doi.org/10.1109%2Ftetci.2017.2762739.

[115] Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S. et al. *SpeechBrain: A General-Purpose Speech Toolkit*. 2021. ArXiv:2106.04624.

[116] Rosa, R. *Plaintext Wikipedia dump 2018*. 2018. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Dostupné z: http://hdl.handle.net/11234/1-2735.

[117] Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. 1958, sv. 65, č. 6, s. 386–408. DOI: 10.1037/h0042519. ISSN 0033-295X. Dostupné z: http://dx.doi.org/10.1037/h0042519.

[118] Rumelhart, D. E. a McClelland, J. L. Learning Internal Representations by Error Propagation. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. 1987, s. 318–362.

[119] Ryant, N., Church, K., Cieri, C., Du, J., Ganapathy, S. et al. *Third DIHARD Challenge Evaluation Plan*. 2020.

[120] Ryant, N., Singh, P., Krishnamohan, V., Varma, R., Church, K. et al. *The Third DIHARD Diarization Challenge*. 2021.

[121] Scheibler, R., Bezzam, E. a Dokmanic, I. Pyroomacoustics: A Python Package for Audio Room Simulation and Array Processing Algorithms. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr 2018. DOI: 10.1109/icassp.2018.8461310. Dostupné z: https://doi.org/10.1109%2Ficassp.2018.8461310.

[122] Schippers, G. a Schaap , C. The Motivational Interviewing Skill Code: Reliability and a Critical Appraisal. Červenec 2005, sv. 33, s. 285 – 298. DOI: 10.1017/S1352465804001948.

[123] Schneider, S., Baevski, A., Collobert, R. a Auli, M. Wav2vec: Unsupervised Pre-training for Speech Recognition. 2019, abs/1904.05862. Dostupné z: http://arxiv.org/abs/1904.05862.

[124] Schuster, M. a Paliwal, K. Bidirectional recurrent neural networks. 1997, sv. 45, č. 11, s. 2673–2681. DOI: 10.1109/78.650093.

[125] Schwenk, H. Continuous space language models. 2007, sv. 21, č. 3, s. 492–518. DOI: https://doi.org/10.1016/j.csl.2006.09.003. ISSN 0885-2308. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0885230806000325.

[126] Serrano Guerrero, J., Olivas, J., Romero, F. a Herrera Viedma, E. Sentiment analysis: A review and comparative analysis of web services. Srpen 2015, sv. 311, s. , 18–38. DOI: 10.1016/j.ins.2015.03.040.

[127] Sido, J., Pražák, O., Pribán, P., Pasek, J., Seják, M. et al. Czert - Czech BERT-like Model for Language Representation. 2021, abs/2103.13031. Dostupné z: https://arxiv.org/abs/2103.13031.

[128] Sido, J., Seják, M., Pražák, O., Konopík, M. a Moravec, V. *Czech News Dataset for Semantic Textual Similarity*. 2022.

[129] Snyder, D., Garcia Romero, D., Sell, G., Povey, D. a Khudanpur, S. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, s. 5329–5333. DOI: 10.1109/ICASSP.2018.8461375.

[130] Straka, M., Náplava, J., Straková, J. a Samuel, D. RobeCzech: Czech RoBERTa, a Monolingual Contextualized Language Representation Model. In: Ekštein, K., Pártl, F. a Konopík, M., ed. *Text, Speech, and Dialogue*. Cham: Springer International Publishing, 2021, s. 197–209.

[131] Straková, J., Straka, M. a Hajič, J. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland: Association for Computational Linguistics, červen 2014, s. 13–18. DOI: 10.3115/v1/P14-5003. Dostupné z: https://aclanthology.org/P14-5003.

[132] Sutskever, I., Vinyals, O. a Le, Q. V. Sequence to Sequence Learning with Neural Networks. 2014, abs/1409.3215. Dostupné z: http://arxiv.org/abs/1409.3215.

[133] Szymański, P. a Kajdanowicz, T. A scikit-based Python environment for performing multi-label classification. únor 2017.

[134] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A. et al. *LLaMA: Open and Efficient Foundation Language Models*. 2023.

[135] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L. et al. Attention Is All You Need. 2017, abs/1706.03762. Dostupné z: http://arxiv.org/abs/1706.03762.

[136] Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C. et al. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Srpen 2021, s. 993–1003. DOI: 10.18653/v1/2021.acl-long.80. Dostupné z: https://aclanthology.org/2021.acl-long.80.

[137] Wang, C., Wu, A., Pino, J., Baevski, A., Auli, M. et al. *Large-Scale Self- and*

Semi-Supervised Learning for Speech Translation. 2021.

[138] Warden, P. *Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition*. 2018.

[139] Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J. et al. *ESPnet: End-to-End Speech Processing Toolkit*. 2018.

[140] Wu, M.-J., Mwangi, B., Bauer, I. E., Passos, I. C., Sanches, M. et al. Identification and individualized prediction of clinical phenotypes in bipolar disorders using neurocognitive data, neuroimaging scans and machine learning. Elsevier BV. leden 2017, sv. 145, s. 254–264.

[141] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M. et al. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. arXiv, 2016. DOI: 10.48550/ARXIV.1609.08144. Dostupné z: https://arxiv.org/abs/1609.08144.

[142] Wu, Z., Balloccu, S., Kumar, V., Helaoui, R., Reforgiato Recupero, D. et al. Creation, Analysis and Evaluation of AnnoMI, a Dataset of Expert-Annotated Counselling Dialogues. 2023, sv. 15, č. 3. DOI: 10.3390/fi15030110. ISSN 1999-5903. Dostupné z: https://www.mdpi.com/1999-5903/15/3/110.

[143] Wu, Z., Balloccu, S., Kumar, V., Helaoui, R., Reiter, E. et al. Anno-MI: A Dataset of Expert-Annotated Counselling Dialogues. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, s. 6177–6181. DOI: 10.1109/ICASSP43922.2022.9746035.

[144] Yang, J., Jin, H., Tang, R., Han, X., Feng, Q. et al. *Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond*. 2023.

[145] Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R. et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding. 2019, abs/1906.08237. Dostupné z: http://arxiv.org/abs/1906.08237.

[146] Yousefi, M. a Hansen, J. H. L. Block-Based High Performance CNN Architectures for Frame-Level Overlapping Speech Detection. 2021, sv. 29, s. 28–40. DOI: 10.1109/TASLP.2020.3036237.

[147] Yu, D. a Deng, L. *Automatic speech recognition*. Springer, 2016.

[148] Zhang, J., Filbin, R., Morrison, C., Weiser, J. a Danescu Niculescu Mizil, C. Finding Your Voice: The Linguistic Development of Mental Health Counselors. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, červenec 2019, s. 936–947. DOI: 10.18653/v1/P19-1089. Dostupné z: https://aclanthology.org/P19-1089.

[149] Zhu, X. Semi-Supervised Learning Literature Survey. Červenec 2008, sv. 2.

[150] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y. et al. A Comprehensive Survey on Transfer Learning. 2019, abs/1911.02685. Dostupné z: http://arxiv.org/abs/1911.02685.

[151] Černocký, J. H., Luque, J., Segura, C., Karafiát, M., Szöke, I. et al. *Big speech data analytics for contact centers BISON*. Dostupné z: https://bison-project.eu/download/D4.3_Final_set_of_speech_technologies.pdf.

[152] Řiháček, T., Nehyba, J., Čevelíček, M., Polok, A., Matějka, P. et al. DeePsy: Představení online nástroje pro zpětnou vazbu v psychoterapii. v tisku.

Příloha A

Typy terapeutických intervencí

V rámci projektu DeePsy byla vytvořena klasifikační hierarchie typů terapeutických intervencí [86]. Definice spolu s příklady jsou doslovně vyňaty z manuálu pro anotování řeči terapeuta vytvořeného v rámci projektu DeePsy, jež vychází z výše citované diplomové práce.

1. Dotazování – Terapeut se dotazuje klienta s cílem získat od něj více informací nebo hlouběji prozkoumat jeho zkušenost.
    - Divergentní – Dotazování, které otevírá spektrum možností. Terapeut prozkoumává klientovu zkušenost. Má formu otevřené otázky nebo pobídky ke klientovu vyprávění. – „Jak se cítíte teď, když o tom mluvíte?“, „A když říkáte lidi, co přesně myslíte?“
    - Konvergentní – Terapeut se ujišťuje o tom, že správně porozuměl klientovu sdělení. Terapeut si ověřuje nějaké informace. Typicky má formu uzavřené otázky. – „Řekla jste jí to?“, „A vy takhle běžně ve vztazích se bojíte na něco zeptat a radši jste v nějaké nejistotě?“
2. Interpretace – Terapeut poskytuje výklad, vysvětlení chování a prožívání klienta. Terapeut předkládá klientovi nový pohled svými vlastními slovy, přináší nové porozumění. Účelem je umožnit nové porozumění a otevřít nové možnosti. – „A vlastně je to i vobraz toho, co se ňák děje teď ve vašem životě.“, „A ten strach vede k tomu, samozřejmě jste se asi nějak cítila v napětí a vede to k tomu, že jste raději do té práce nešla.“
3. Reflexe – Terapeut ve svém sdělení odráží nějaké aspekty klientových emocí, myšlenek nebo chování bez toho, aby rozebíral jejich důvody nebo souvislosti. Formuluje vlastními nebo klientovými slovy to, co klient sděluje, nebo co si myslí, že sděluje.
    - Parafrázující reflexe – Jedná se o prosté zopakování toho, co řekl klient. Není nutné, aby to byla poslední klientova slova v jeho sdělení, ale musí být obsažena v segmentu těsně předcházejícím tomuto terapeutovu sdělení. – Klient: „Ve čtvrtým měsíci.“, Terapeut: „Ve čtvrtým měsíci.“
      Klient: „. . . ona v tom vidí zase jako něco . . . “, Terapeut: „V tom hledá nějakou souvislost.“
    - Obohacující reflexe – Terapeut zkouší pojmenovat emoci, kterou klient může prožívat, dosud ji ale nepojmenoval. Terapeut obsáhleji shrnuje klientovo sdělení, kdy je obohacujícím prvkem to, že terapeut usiluje obsáhnout to podstatné

a použít výstižnější slova. – „Vám to taky bylo líto.“, „Jakože hodně hodně cest, když bych to zkusil jak kdyby říct za vás, hodně cest přede mnou, ale vlastně žádnou z nich úplně nevím, jestli ji chci jít, nebo kterou si vybrat.“

4. Potvrzování – Potvrzování shrnuje všechny terapeutovy věty, které mají za primární cíl klienta povzbudit.

   (a) Procesové potvrzení – Terapeut podporuje klienta v probíhajícím procesu (např. vyprávění). Patří sem pouze citoslovce a jednoduchá, obvykle jednoslovná vyjádření. – „Hm.“, „Uhm.“, „Ano.“

   (b) Empatické ujištění – Patří sem potvrzující promluvy s prvky vyjádření empatie a vztahových aspektů, kdy terapeut chce vyjádřit klientovi, že není v situaci sám. Vyjadřuje mu podporu, která je vztažená k emocím, které během sezení klient aktuálně prožívá. – „Všímám si toho.“, „Nemusíte říkat nic.“

   (c) Normalizace – Jedná se o ujištění klienta o tom, že to, co prožívá či co se mu děje, je v pořádku nebo obvyklé. – „Ehm, jo, jo, to tak ta hlava dělá.“, „Hmm, mnoho lidí to takhle dělá.“

   (d) Empowerment – Terapeut podporuje klientovy zdroje, které už má. Povzbuzuje ho a dodává mu naději tím, že upozorňuje na to, jaké dělá pokroky a co už dokázal. Podporuje jeho autonomii, osobní zodpovědnost a schopnost se rozhodovat. – „A to byly přitom docela odvážný otázky, prostě přijít za klukem a zeptat se ho na to (jestli se mnou teda chce chodit), to je docela kurážný.“, „V té době to je možná to nejlepší jako věc, kterou jste mohla pro sebe jako udělat.“

5. Informování – Tato kategorie shrnuje situace, kdy terapeut nabízí klientovi nové informace, například ohledně psychologických procesů na obecné rovině, různých principů fungování světa a toho, jak prakticky probíhá terapie.

   (a) Edukace – Terapeut klienta edukuje o obecných věcech a sděluje mu svůj názor, který by měl ovšem být podložen přinejmenším odbornými zkušenostmi. – „Ale třeba tak to ve vztahu často je, že opravdu jeden dělá to a ten druhý něco jinýho.“, „No tak já nevím, rodiče si nevybíráme. Si nevybereme.“

   (b) Perspektiva druhých – Psychoterapeut nabízí klientovi novou perspektivu na druhé lidi tím, že se pokouší vysvětlit, co by za jejich chováním mohlo být za důvody. – „Možná jí nic jinýho nezbývá.“, „Já myslím, že pro ni tohle to může bejt strašně důležitý.“

6. Terapeutova direktivita – Tato kategorie shrnuje věty, ve kterých je terapeut direktivnější a navrhuje klientovi, co by mohl nebo měl zkusit.

   (a) Rada/úkol – Terapeut dává klientovi pokyn či radu, která směřuje do života klienta mimo sezení. – „A napadá mě k tomu, že jedna ze strašně důležitejch věcí je, postarat se o to, abyste měla dost sil sama pro sebe.“, „Zkuste se zaměřit na to, jako, když mluvíte o tom zklidnění, co by mohlo pomoct jakoby i v tom.“

   (b) Procesové vedení – Terapeut využívá direktivnější intervence a techniky odpovídající nějakému psychoterapeutického směru. Terapeut výrazněji ovlivňuje proces terapie a vytváří tak strukturu sezení. – „Zůstaňme zpátky k tej vaší pubertální jako nějaké žačce, nebo ehmm svěřenkyni.“. „Kdybyste měla takovou

pomyslnou stupnici od 1 do 10, kdy 10 je maximum, mám tu schopnost vyvinutou na 100%, tak kde byste se teďkom nacházela v tý odolnosti vůči citovému nátlaku“

7. Sebeodhalení – Terapeut odhaluje vlastní prožívání nebo sděluje informace ze svého soukromí. Popisuje svoje pocity, názory, plány, přání a svoji osobní zkušenost. Může také sdělovat, jaké mu pomohly strategie ke zvládnutí podobné situace. – „Pro mě je to takový jako, jsem prostě rád, že, že jsem tohle to s vámi mohl sdílet. Nebo že jste se o to se mnou podělila.“, „Na mě vlezla nějaká rýma zase.“

8. Konfrontace – Terapeut poukazuje na obranné mechanismy a iracionální přesvědčení, které si klient nejspíše neuvědomuje. Všímá si rozporů mezi slovy a chováním, dvěma prohlášeními nebo mezi představami a realitou. –„Já bych si tímhle tím nebyl tak jistej.“, „A to jsou všechno vaše takový domněnky, který nějak nemůžete vědět.“

9. Ostatní – Každý segment potřebuje mít přiřazenou nějakou kategorii, a tak do této spadá všechno, co nejde zařadit do výše uvedených.

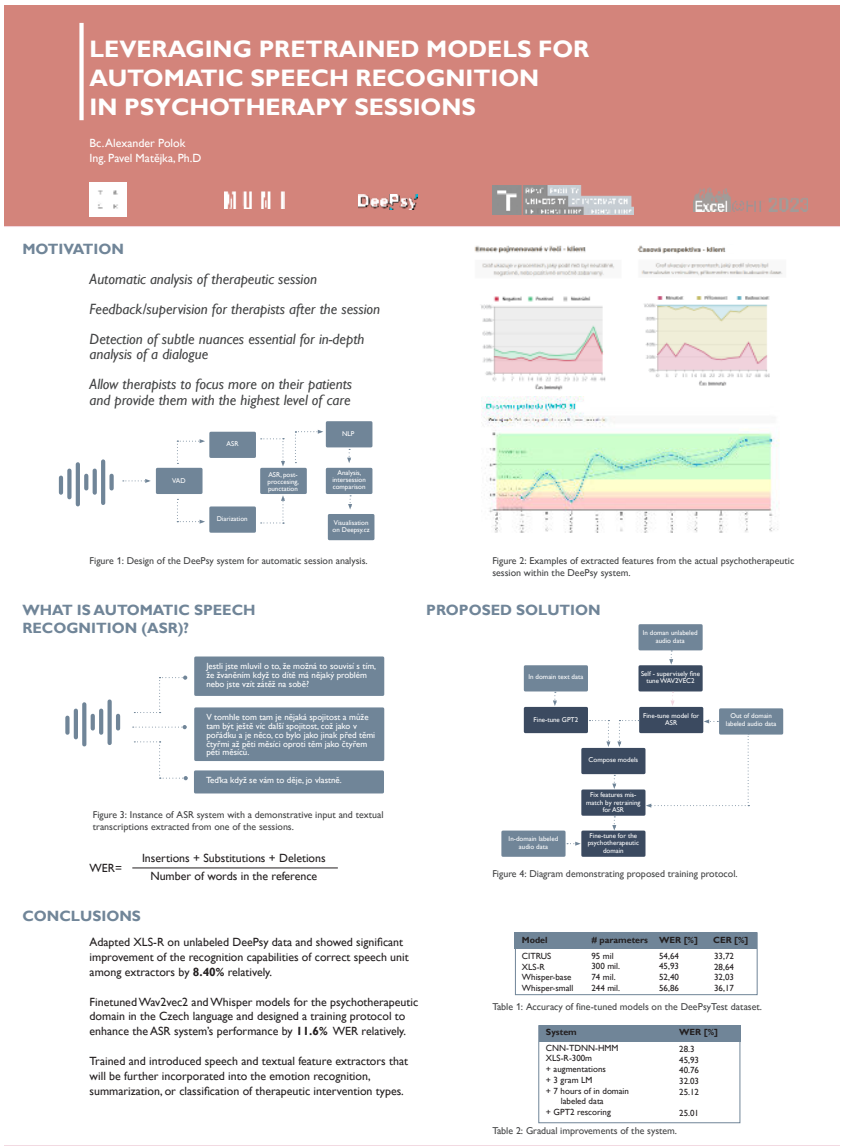
   (a) Small talk – Do této kategorie patří převážně věty na začátku a na konci sezení. Typicky sem ze začátku sezení zařazujeme pozdravení, přivítání klienta, poděkování, že přišel, nabídnutí vody/čaje a místa k sezení. Mohou sem také patřit zdvořilostní fráze jako je okomentování počasí nebo dopravní situace. Z průběhu sezení sem patří věty, které se nevztahují k tématům terapie a nepředpokládáme, že by měly význam pro terapeutický vztah. – „Tak vás vítám tady, znovu, skoro po měsíci.“, „Chcete dolít?“

   (b) Nezařaditelné – Zařazujeme sem věty, u kterých předpokládáme, že by mohly být terapeuticky důležité, ale nemáme pro ně kategorii. – „Teď jsem se v tom ztratil.“, „Jo, mně se to ňák pomotalo v té hlavě, že. Říkala jste, že byla, ale já jsem si to nějak spojil, že to bylo teď někdy nedávno.“

   (c) Nesrozumitelné – Do této kategorie patří to, co nelze zařadit výše, protože není terapeuticky důležité. Spadají sem nesrozumitelné věty a fragmenty vět přerušené klientem. – „Takže.“, „Nevím, jak jste se . . . “

Příloha B

Plakát prezentovaný v rámci konference Excel@FIT 2023



LEVERAGING PRETRAINED MODELS FOR AUTOMATIC SPEECH RECOGNITION IN PSYCHOTHERAPY SESSIONS

Bc. Alexander Polok
Ing. Pavel Matějka, Ph.D

MOTIVATION

Automatic analysis of therapeutic session

Feedback/supervision for therapists after the session

Detection of subtle nuances essential for in-depth analysis of a dialogue

Allow therapists to focus more on their patients and provide them with the highest level of care

Figure 1: Design of the DeePsy system for automatic session analysis.

Figure 2: Examples of extracted features from the actual psychotherapeutic session within the DeePsy system.

WHAT IS AUTOMATIC SPEECH RECOGNITION (ASR)?

Figure 3: Instance of ASR system with a demonstrative input and textual transcriptions extracted from one of the sessions.

$$\text{WER} = \frac{\text{Insertions} + \text{Substitutions} + \text{Deletions}}{\text{Number of words in the reference}}$$

PROPOSED SOLUTION

Figure 4: Diagram demonstrating proposed training protocol.

CONCLUSIONS

Adapted XLS-R on unlabeled DeePsy data and showed significant improvement of the recognition capabilities of correct speech unit among extractors by **8.40%** relatively.

Finetuned Wav2vec2 and Whisper models for the psychotherapeutic domain in the Czech language and designed a training protocol to enhance the ASR system's performance by **11.6%** WER relatively.

Trained and introduced speech and textual feature extractors that will be further incorporated into the emotion recognition, summarization, or classification of therapeutic intervention types.

Model	# parameters	WER [%]	CER [%]
CITRUS	95 mil	54,64	33,72
XLS-R	300 mil.	45,93	28,64
Whisper-base	74 mil.	52,40	32,03
Whisper-small	244 mil.	56,86	36,17

Table 1: Accuracy of fine-tuned models on the DeePsyTest dataset.

System	WER [%]
CNN-TDNN-HMM	28.3
XLS-R-300m	45,93
+ augmentations	40.76
+ 3 gram LM	32.03
+ 7 hours of in domain labeled data	25.12
+ GPT2 rescoring	25.01

Table 2: Gradual improvements of the system.

The development of the application is supported by the Technology Agency of the Czech Republic under the Eta Programme (grant no. TL03000049).
Computational resources were provided by the e-INFRA CZ project (ID:90140), supported by the Ministry of Education, Youth and Sports of the Czech Republic.