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 AUDIO HOVORU MEZI DVĚMA ÚČASTNÍKY

INTERVIEW ANALYSIS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **ALEXANDER POLOK**
AUTHOR

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

BRNO 2021

Vysoké učení technické v Brně
Fakulta informačních technologií

Ústav počítačové grafiky a multimédií (UPGM) Akademický rok 2020/2021

Zadání bakalářské práce

Student: **Polok Alexander**
Program: Informační technologie
Název: **Analýza audio hovoru mezi dvěma účastníky**
Analysis of Interview Audio
Kategorie: Zpracování řeči a přirozeného jazyka

Zadání:

1. Prostudujte statistické techniky pro modelování řeči.
2. Seznamte se s technikami pro detekci řečové aktivity
3. Navrhněte a implementujte alespoň 6 slabých klasifikátorů, které popisují hovor (např. základní tón řeči, energie, cross-talk, rychlost řeči, poměr řeči, reakční dobu, přepis řeči, sentiment, ...)
4. Vytvořte prezentační výstup (webová stránka, pdf, ...), který zobrazí statistiky z těchto klasifikátorů - například časový průběh, celkový průběh za celou nahrávku, histogram či vývoj v čase.
5. Vytvořte krátké video nebo plakát prezentující vaši práci.

Literatura:

- SIGMUND, M. *Zpracování řečových signálů, elektronická skripta.* FEKT: REL 07- 052. Brno: FEKT VUT, 2007.
- Online psychoterapie, hedepy.cz
- Dokumentace ke grantu TAČR DeePsy, deepsy.cz

Pro udělení zápočtu za první semestr je požadováno:

- Body 1 a 2, rozpracovaný bod 3.

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, doc. Dr. Ing.
Datum zadání: 1. listopadu 2020
Datum odevzdání: 12. května 2021
Datum schválení: 13. listopadu 2020

Abstrakt

Cílem této práce je analýza psychoterapeutických sezení. Z audionahrávek jsou extrahovány klasifikátory, které popisují proběhlou terapii. Ty jsou následně agregovány, porovnány s ostatními sezeními a graficky prezentovány v podobě zprávy shrnující daný rozhovor. Terapeutům je tímto způsobem k proběhlým sezením poskytnuta zpětná vazba, která může sloužit k profesnímu růstu a kvalitnější psychoterapii v budoucnu.

Abstract

The aim of this thesis is the analysis of psychotherapeutic sessions. Classifiers describing the therapy are extracted from the audio recordings. These are then aggregated, compared with other sessions, and graphically presented in a report summarizing the conversation. In this way, therapists are provided with feedback that can serve for professional growth and better psychotherapy in the future.

Klíčová slova

Analýza rozhovoru, Online psychoterapie, Klient a terapeut, Zpětná vazba, Detekce řečové aktivity, Diarizace, Zpracování řeči, Zpracování přirozeného jazyka, Paralingvistika

Keywords

Conversation analysis, Online psychotherapy, Client and therapist, Feedback, Speech activity detection, Diarization, Speech processing, Natural language processing, Paralinguistics

Citace

POLOK, Alexander. *Analýza audio hovoru mezi dvěma účastníky.* Brno, 2021. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Pavel Matějka, Ph.D.

Analýza audio hovoru mezi dvěma účastníky

Prohlášení

Prohlašuji, že jsem tuto bakalářskou 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
11. května 2021

Poděkování

Chtěl bych poděkovat vedoucímu práce, Ing. Pavlovi 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 Drahomíře Šloufové, Michalu Rozsívalovi a moji skvělé přítelkyni za korekturu gramatických chyb. Speciální poděkování patří Ladislavu Ondrisovi, s nimž jsem celou práci průběžně diskutoval a poskytl mi mnoho cenných rad a podnětů ke zlepšení.

Obsah

Kapitola 1

Úvod

Zpracování řeči a její analýza je v dnešní době celosvětově velmi rozvíjený a diskutovaný obor. Existuje nespočet odborných skupin věnujících se problematice identifikace řečníka, identifikaci jazyka, rozpoznávání řeči, detekci klíčových slov, rozpoznávání fonémů nebo zpracování přirozeného jazyka. Technologický rozvoj v posledních letech způsobil, že systémy věnující se problematice zpracování řeči dosahují velmi vysokých přesností a můžeme se s nimi setkat skoro na denní bázi, ať už třeba v podobě virtuálních asistentů Siri, Alexa nebo v rámci zlepšování kvality call center. Tato práce se zaměřuje na analýzu psychoterapeutického sezení. Psychoterapie je původem latinské slovo – skládající se z části psyché (výrazu pro lidskou duši či mysl) a therapeia (léčba). Jejím cílem je léčit mysl, obnovit její rovnováhu a vést ke zkvalitnění života klienta.

Pandemie virové choroby covid-19 změnila životy mnohých a zvýšila taktéž poptávku po psychoterapeutických sezeních, ať už z důvodu chybějící sociální interakce nebo zvýšeného stresu a nejistoty. Se zvyšující se poptávkou po psychoterapii a snahou o zamezení šíření choroby, což se projevilo v podobě mnoha vládních opatření, mezi něž patří především omezení kontaktu, se rozšířila alternativa známá jako online psychoterapie. Terapie pomocí video hovoru nebyla dosud v České republice příliš známá. Ve Spojených státech amerických, v severní Evropě nebo Velké Británii je však tato forma běžná a je často jedinou formou pomoci, kterou mohou využít invalidé, nemocní a lidé bydlící v odlehlých oblastech.

Forma video hovoru přináší možnost analyzovat psychoterapeutické sezení strojově a získat tímto způsobem nové informace o proběhlém sezení. Cílem není hodnotit úspěšnost terapie, ale nalézt klasifikátory a časové úseky, které dokážou terapeuta upozornit například na vyšší výskyt skákání do řeči nebo výskyt změny emočního stavu klienta.

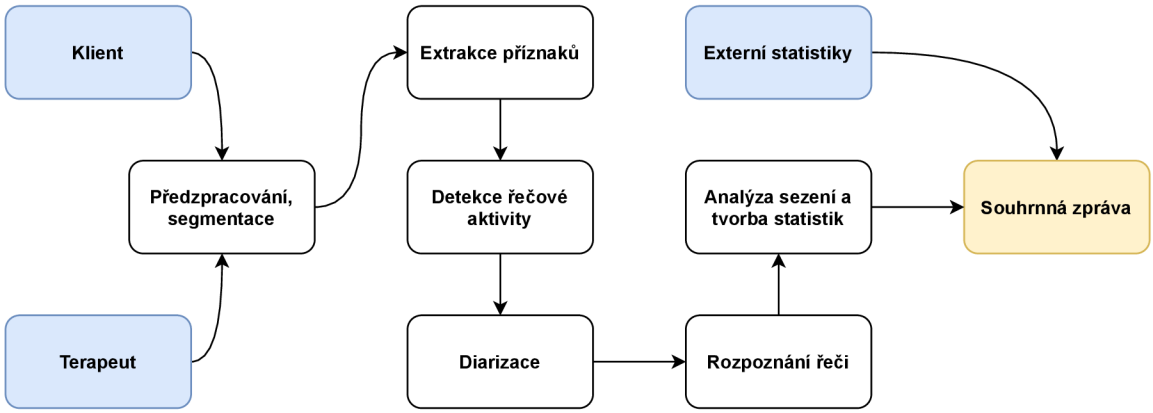
Kapitola 2 se zaměřuje na objasnění použitých technik. V rámci kapitoly 3 jsou představena data, pro jejichž zpracování byl systém navržen. Detaily návrhu tohoto systému jsou popsány v kapitole 4, v rámci této kapitoly jsou rovněž představeny klasifikátory obsažené ve výsledné souhrnné zprávě. Souhrnná zpráva, jež je výstupem této práce, může částečně nahrazovat drahou supervizi u mladých psychoterapeutů, kteří ukončili vzdělání a získávají praxi. Dokument přináší porovnání s ostatními terapeuty. Pro příslušné druhy terapie se však mohou metriky velmi lišit a interpretace hodnot už je na samotném terapeutovi, případně jeho zkušenějším kolegovi.

Kapitola 5 přibližuje implementační detaily navrženého systému. Obsahuje informace o použitých technologiích, algoritmech, knihovnách a externích nástrojích. V kapitole 6 jsou popsány experimenty, které vedly ke zlepšení přesnosti navrženého systému. V kapitole 7 jsou sumarizovány výsledky této práce a navržen další postup.

Kapitola 2

Analýza audio nahrávky a extrakce příznaků

V rámci této kapitoly je čtenář uveden do problematiky zpracování řečových signálů. Je mu postupně představen celý proces od zaznamenání vibrací vzduchu reprezentujících zvuk, až po metody použité pro rozpoznání řeči příslušných mluvčích. Obr. 2.1 demonstruje celý proces analýzy audio nahrávek klienta a terapeuta. Extrakcí statistik a tvorbou souhrnné zprávy se blíže zabývá kapitola 4.



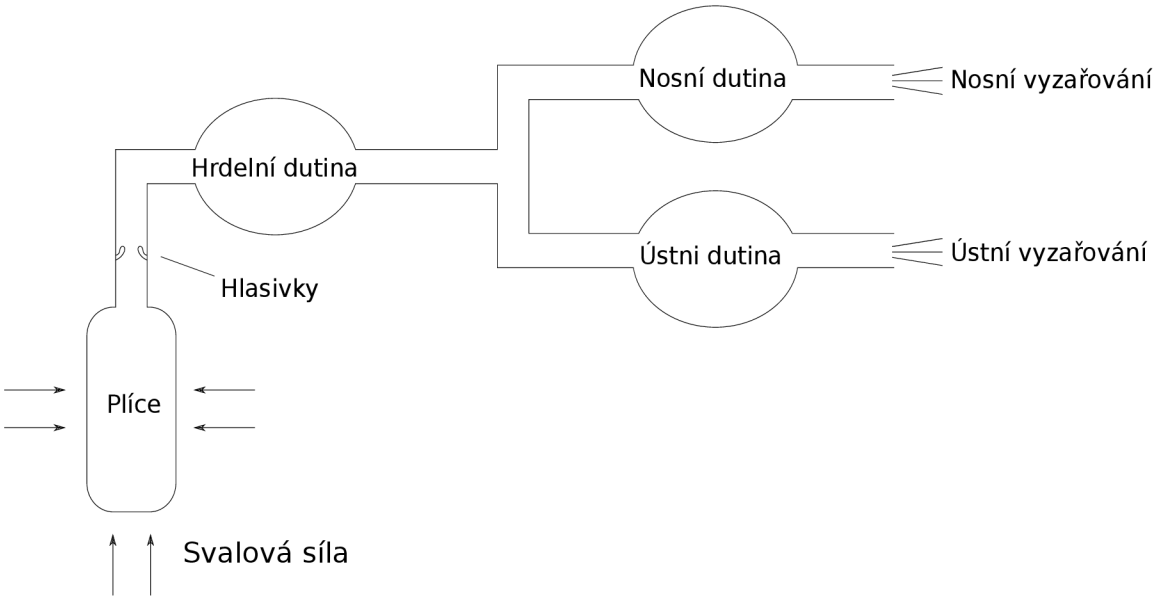
Obrázek 2.1: Proces zpracování vstupních nahrávek terapeuta a klienta. Prvním krokem analýzy je segmentace zvukové nahrávky. Ze segmentů jsou extrahovány příznaky, s jejichž pomocí dochází k určení řečově aktivních segmentů příslušných řečníků. Následně je vytvořen přepis z těchto řečově aktivní segmentů. Na závěr jsou vypočteny statistiky o sezení a je vytvořena souhrnná zpráva zahrnující porovnání s ostatními sezeními.

2.1 Řečový signál

Schopnost vyjadřovat se a rozumět mluvené řeči je dána pouze lidskému druhu jako doposud jedinému známému organismus žijícímu na Zemi. Občas sice říkáme, že mluví i papoušci, jedná se však jen o zvukové napodobení tvořené zcela jiným způsobem, než je artikulovaná řeč člověka. Zvuková řeč má významné postavení ve vývoji člověka jako druhu. Řeč je řízená mozkem a to zejména mozkovým centrem mluvení (řečového výkonu – centrum Wernickovo) a centrem slyšení řeči (dešifrování slyšeného signálu – centrum Brockovo). Dů-

sledkem poškození mozku mohou vzniknout poruchy, které řeč postihují. Mezi nejznámější patří afázie (neschopnost řeč tvořit nebo ji porozumět), koktavost, breptavost, řidčeji se může objevit oněmění, ev. mluvní negativismus [25].

Princip tvorby řeči, který je velmi podobný vytváření tónu u dechových nástrojů, je znázorněn na obr. 2.2. Samotná řeč vzniká proudem vzduchu, který je dodáván plícemi. Takto vzniklý proud prochází hlasivkovou štěrbinou, která je obklopena kmitajícími hlasivkami tvořícími budící signál. Tento signál však ještě nepovažujeme za řeč. Až artikulací, tedy příslušnou konfigurací řečových orgánů v dutině hrdelní, ústní a nosní, dochází ke vzniku zopakovatelných úseků řeči – hlásek [33].



Obrázek 2.2: Schematické znázornění hlasového ústroji.

Řeč je možné prezentovat jako soubor informací nebo jako fyzikální signál, který přenáší nejenom konkrétní zprávu, ale poskytuje rovněž informace o pohlaví, věku, zdravotním stavu, náladě, poruchách mluvy a identitě mluvčího. Z mluvy je možné taktéž odhalit prostředí a druh sdělení. Podle hlasu si dokážeme představit, jak daná osoba vypadá. Všechny tyto zmíněné charakteristiky se v posledních letech daří získávat i strojově. Je takto možné získat z audio nahrávky velmi širokou škálu informací o mluvčím.

Jelikož mluva je analogový signál daný spojitou funkcí spojitého času a aktuálně používané počítače jsou v absolutní většině digitální, je nutné pro další zpracování analogový signál zaznamenat a převést ho do nespojité posloupnosti digitálních (číselných) údajů. Tento proces začíná konverzí akustického tlaku na elektrický signál v zařízení známém jako mikrofon. Velmi slabý signál v jednotkách milivoltů je třeba zesílit a převést do číselné podoby. Za tímto účelem se využívá analogově digitální převodník na jehož výstupu je obdržena binární hodnota. Převod by nebyl možný bez určení vzorkovací frekvence[1] a rozlišení hodnot. Vzorkovací frekvence se nejčastěji pohybuje od 8 kHz u telefonních linek až po 44,1 kHz u kompaktních disků (CD). Při velmi nízkých vzorkovacích frekvencích může dojít k jevu zvanému aliasing (překrytí frekvenčních spekter vzorkovaného signálu). Aby k tomuto jevu nedošlo je nutné dodržet Shannonův teorém (vzorkovací frekvence musí být vyšší než dvojnásobek nejvyšší harmonické složky vzorkovaného signálu, aby došlo k přesné

[1] počet vzorků za jednotku času

rekonstrukci spojitého signálu). Rozlišení hodnot určuje přesnost, obvykle se používá 8 nebo 16 bitů, při příliš nízkém rozlišení digitalizovaného zvukového signálu je možné pozorovat slyšitelný kvantizační šum (ztráta informace, která vznikla vlivem zaokrouhlení okamžité hodnoty signálu do množiny hodnot rozlišení) [33].

2.2 Předzpracování

Lidská řeč je značně variabilní a je skoro nemožné dvakrát vyslovit jedno slovo totožně. Aby byla vyslovená slova naprosto stejná, bylo by nutné dodržet stejnou hlasitost, intonaci, výšku tónu, přízvuk a rychlost. To je z principu lidské řeči skoro nemožné. Při zaznamenávání řečové aktivity a převodu do digitální podoby do signálů můžeme zanést vlastnosti, které následnou analýzu značně ztíží. Cílem procesu předzpracování je potlačit rušivé prvky v podobě šumu okolí, neřečových událostí na straně řečníka či zkreslení vznikajících nekvalitním mikrofonem.

2.2.1 Preemfáze

Preemfáze je proces používaný v elektrotechnice pro zlepšení přenosových parametrů, přesněji dochází ke zdůrazňování amplitud spektrálních složek s jejich vzrůstající frekvencí. Podstatná část celkové energie řečového signálu (u některých mluvčích tvoří více než polovinu) leží v kmitočtovém pásmu pod hranicí 300 Hz a většina statisticky významných informací v pásmu nad 300 Hz. Cílem preemfáze je zvýraznit vyšší frekvence a vyrovnat energetické spektrum celého pásma. Provedením filtrace s horní propustí nad řečovým signálem docílíme:

- vyvážení frekvenčního spektra
- vyhneme se numerickým problémům během Fourierovy transformace
- možného vylepšení poměru signálu k šumu daného vztahem $\mathrm{SNR} = \frac{P_{signal}}{P_{noise}}$, kde P_{signal} je střední výkon signálu a P_{noise} střední výkon okolního šumu

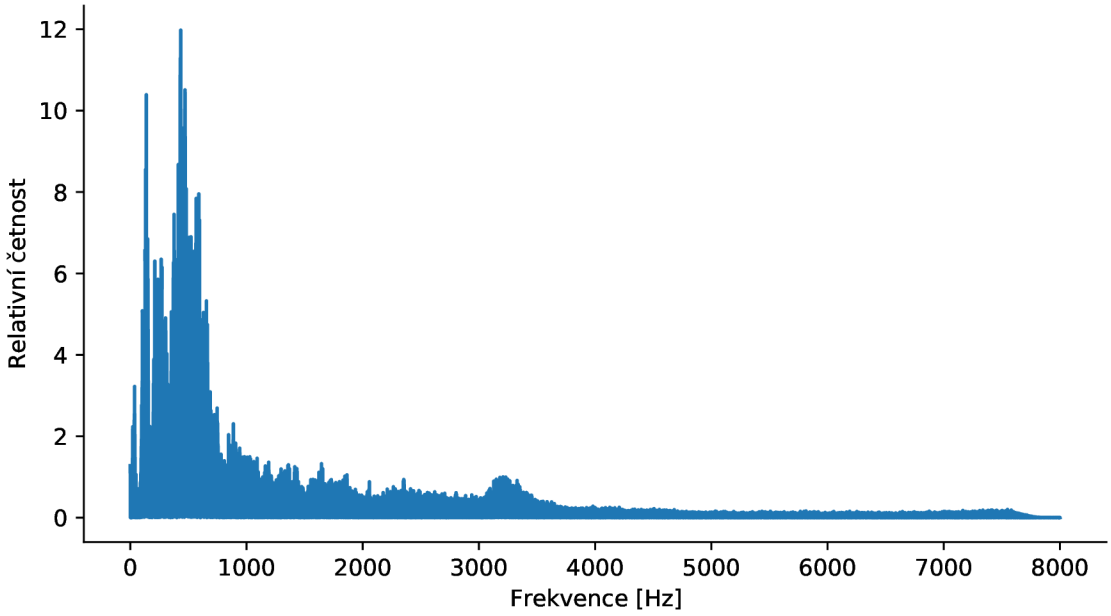
Preemfáze v časové doméně je daná vztahem

$$y_n = x_n - \alpha \cdot x_{n-1}, n \in \{1, \ldots, N\}, \tag{2.1}$$

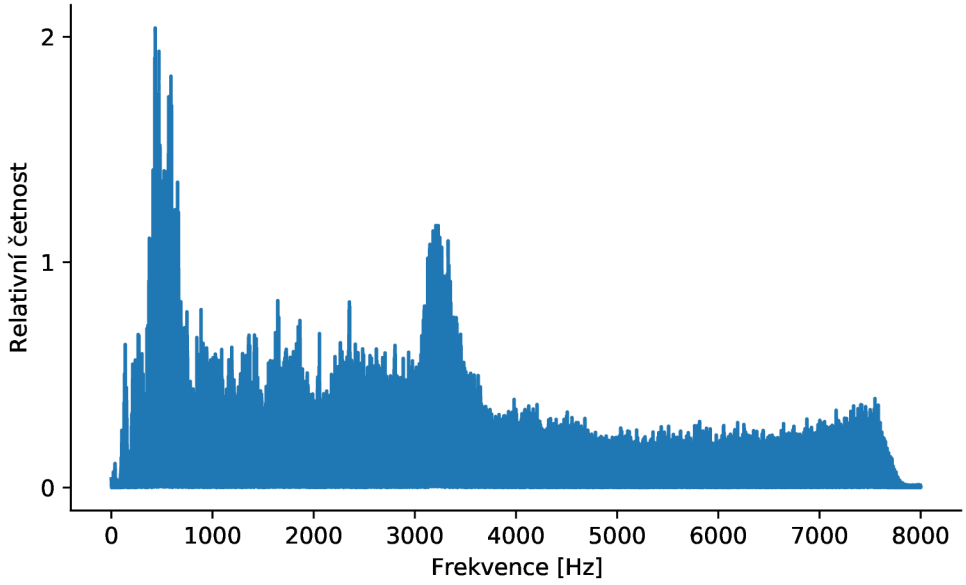
kde N je počet vzorků signálu a koeficient preemfáze α obvykle leží v intervalu $\alpha \in (0,9; 1,0)$. Nejčastěji používané hodnoty α bývají 0,95 nebo 0,97. Obr. 2.3 znázorňuje spektrum signálu a obr. 2.4 spektrum totožného signálu po aplikaci preemfáze, které je na první pohled vyváženější.

2.2.2 Segmentace

Řečový signál v průběhu času mění svůj charakter vzhledem ke své povaze. Přístup, v němž by extrakce příznaků audio nahrávky probíhala nad každým bitem signálů, je ale zcela nepředstavitelný. Řečový signál je však kvazistacionární, a tudíž je možné předpokládat, že N sousedních vzorků reprezentuje stejnou informaci. Je nutné nalézt takové N, které je dostatečně malé na to, abychom hledané vlastnosti signálu mohli bezchybně vyjádřit N vzorky a zároveň dostatečné velké, aby hledané vlastnosti nebyly ovlivněny lokálními změnami. Tyto protichůdné požadavky jsou vcelku splněny pro N vzorků, které reprezentují úsek o délce



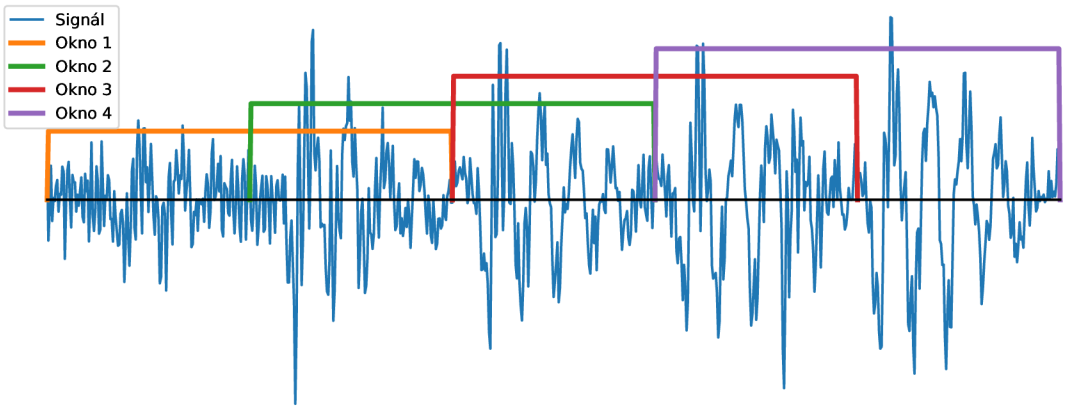
Obrázek 2.3: Spektrum signálu bez použití preemfáze.



Obrázek 2.4: Spektrum signálu po aplikaci preemfáze.

cca 10 až 25 ms. Zmíněné hodnoty odpovídají periodě změny hlasového ústrojí v lidském těle. Za účelem odstranění ostrých hran je vhodné sousední segmenty překrývat. Překrývání úseků je výpočetně náročnější, ale je dosaženo částečného vyhlazení časových průběhů získaných parametrů. V případě vyššího překrytí rámců je možné sledovat mezi sousedními rámci určitou závislost, která může způsobit selhání jistých klasifikačních technik před-

pokládajících nezávislost sousedních parametrů. Optimální hodnota překrytí segmentů se pohybuje okolo 50 % [33]. Princip segmentace je zobrazen na obr. 2.5.



Obrázek 2.5: Princip segmentace je znázorněný na signálu o délce 50 ms. Signál je rozdělen na 4 segmenty o velikosti 20 ms s překrytím 10 ms.

Segmentace je dána vztahem

$$\mathbf{y} = \mathbf{x} \times \mathbf{w}, \tag{2.2}$$

kde $\mathbf{x} \in \mathbb{R}^N$ je původní signál a $\mathbf{w} \in \mathbb{R}^N$ je vektor tzv. okénkové funkce, která je nulová mimo zvolený interval, symetrická kolem středu, maxima dosahuje většinou okolo středu intervalu a bývá klesající směrem od středu. V obou případech a v následujících vzorcích je N rovné počtu vzorků signálu. Pomocí segmentace je možné z nahrávky extrahovat příslušné vzorky a přidělit jim určitou váhu. Prakticky dochází nejdříve k izolaci příslušného segmentu a následně k násobení vzorků s příslušnými váhami. Existuje mnoho variant tzv. okénkových funkcí, patří mezi ně např. pravoúhlé okno, trojúhelníkové okno, kosinové okno, Gaussovo okno, Hannovo okno, Blackmanovo okno a Kaiserovo okno, Hammingovo nebo Parzenovo okno. Nejčastěji používanými typy oken při zpracování řeči jsou

- pravoúhlé okno

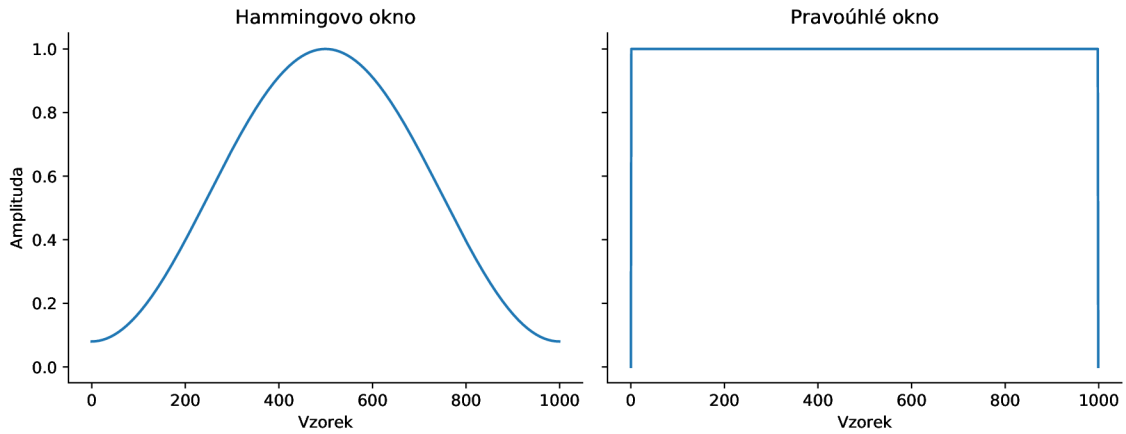
$$f_n = \begin{cases} 1 & n \in \{M+1, \ldots, M+K\} \\ 0 & \text{jinak} \end{cases}, n \in \{1, \ldots, N\} \tag{2.3}$$

- Hammingovo okno

$$f_n = \begin{cases} 0,54 - 0,46\cos(\frac{2\pi \cdot (n-M)}{K}) & n \in \{M+1, \ldots, M+K\} \\ 0 & \text{jinak} \end{cases}, n \in \{1, \ldots, N\}. \tag{2.4}$$

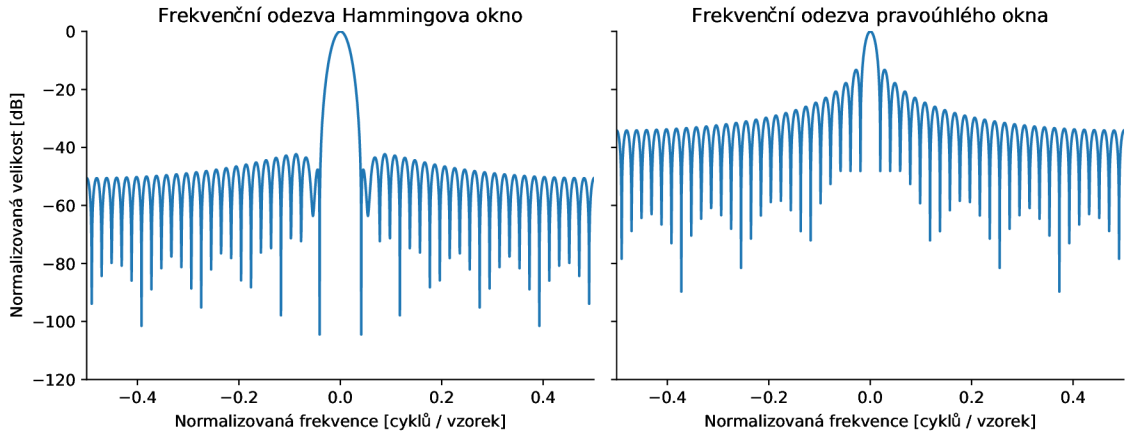
V obou případech M označuje první segment, ve kterém je okno nenulové a K délku okna.

Okna jsou zobrazena na obr. 2.6. V obou případech je N rovno délce okna. Přestože se pravoúhlé okno jeví jako velmi jednoduché a výpočetně zcela nenáročné, většinou je používáno okno Hammingovo, a to z důvodu vyšší stability výpočtů a nežádoucího rozmazání a rozptylu spektra, k němuž dochází při použití okna pravoúhlého, což je zapříčiněno frekvenční charakteristikou pravoúhlého okna. Spektrum obsahuje jeden hlavní lalok a velké



Obrázek 2.6: Tvar Hammingova a pravoúhlého okna o délkách 1000 vzorků.

množství laloků vedlejších. Konvolucí spektra okna se spektrem signálu se jedna frekvence vstupního signálu zpropaguje mezi sousední frekvence. Použitím okna Hammingova dojde k výraznějšímu „rozmáznutí“ spektra způsobeného dvojnásobnou šířkou hlavního laloku, které vede k nižší frekvenční selektivitě. Z praktických důvodů a kvůli vyšší stabilitě výpočtů je častěji používáno okno Hammingovo. Na obr. 2.7 jsou zobrazeny frekvenční odezvy pravoúhlého a Hammingova okna.



Obrázek 2.7: Frekvenční odezva pravoúhlého a Hammingova okna. Lze pozorovat zmiňovanou šířku hlavního laloku a rozdíl mezi energií hlavního a vedlejších laloků. Z výše uvedených důvodu je vhodnější použít okno Hammingovo.

Řečový signál je aplikováním W oken o délce K a následnou filtrací nulových hodnot rozdělen do W segmentů o konstantní délce segmentu K vzorků. Energetické spektrum segmentů je vyrovnanější a částečně je odstraněn šum. Takto získané segmenty můžeme dále analyzovat a aproximovat konstantními parametry.

2.3 Parametry řečového signálu

Jak již bylo zmíněno v podsekci 2.2.2, řeč je kvazistacionární a její segmenty můžeme považovat za stacionární, jelikož ke změnám dochází dostatečně pomalu. Tento předpoklad nám umožňuje segmenty reprezentovat konstantními parametry a není nutné provádět jejich výpočet nad každým bitem digitalizovaného signálu. Příslušné vzorky segmentů signálu obsahují vysoké množství informací. Některé z nich jsou pro účely této práce irelevantní. Cílem této sekce je představit nejdůležitější parametry aproximující úsek řeči pro potřeby psychoterapie a popsat způsob jejich extrakce.

2.3.1 Energie signálu

Střední krátkodobá energie signálu se v časové doméně definuje jako velikost skalárního součinu signálu se sebou samým. U diskrétního signálu $\mathbf{s}$ o délce segmentu N je dána vztahem

$$E = \sum_{n=1}^{N} |s_n|^2. \tag{2.5}$$

U řečových signálů neobsahujících šum a zvuky okolí může energie signálu sloužit k jednoduché detekci řečové aktivity. Energie signálu je u znělých hlásek podstatně vyšší než u neznělých. Podle tohoto kritéria lze fonémy rozdělit do pěti skupin: samohlásky, nosovky, znělé frikativy, neznělé frikativy a mezery v řeči [33].

Střední kvadratická energie

Dalším způsobem výpočtu energie signálu je jeho střední kvadratická energie (RMSE – root mean square energy). Vede k zvýraznění úseků, při nichž dochází k řečové aktivitě [19]. Je dána vztahem

$$RMSE = \sqrt{\frac{1}{N} \sum_{n=1}^{N} |s_n|^2}. \tag{2.6}$$

Obr. 2.8 znázorňuje rozdíl mezi energií a střední kvadratickou energií.

Normalizovaná energie

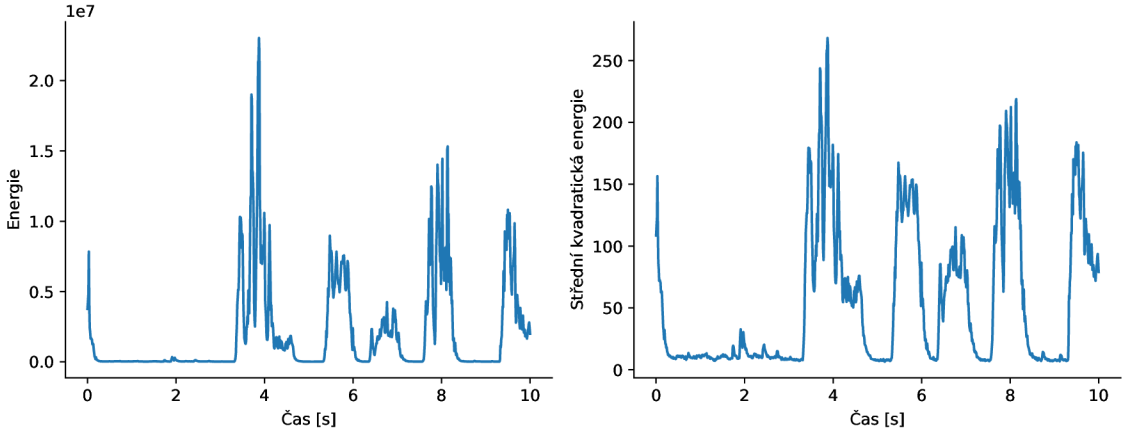
Abychom odstranili varianci mezi příslušnými nahrávkami, kanály a řečníky, je vhodné energii normalizovat do rozsahu $\langle -1, 1 \rangle$. Toho lze docílit následujícím vztahem

$$Enorm_n = \frac{E_n - \mu}{\lambda}, \tag{2.7}$$

kde $\mathbf{E}$ je vektor středních krátkodobých energií segmentů, μ jejich střední hodnota a λ jejich směrodatná odchylka.

2.3.2 Mel-frekvenční cepstrální koeficienty

Jak již bylo zmíněno v sekci 2.1, zvuk je buzení vznikající v hlasivkové štěrbině, které je následně filtrováno tvarem hlasového traktu jedince. Tento tvar určuje, jaký zvuk vychází. Pokud lze přesně určit tento tvar, pak je i možné poskytnout přesnou reprezentaci



Obrázek 2.8: Energie a střední kvadratická energie segmentů signálu. U střední kvadratické energie lze pozorovat vyšší energii řečově aktivních segmentů v poměru s tichem.

vytvářeného fonému. Cílem je oddělit buzení a získat tvar traktu. Ten se projevuje v krátkodobém výkonovém spektru. Úkolem Mel-frekvenčních cepstrálních koeficientů (MFCC – Mel-frequency cepstral coefficients) je co nejpřesněji reprezentovat tvar hlasového traktu řečníka. MFCC představili v 80. letech 20. století Steven B. Davis a Paul Mermelstein [4] a od té doby patří mezi nejpoužívanější příznaky využívané k rozpoznání řeči a řečníka.

Melova stupnice

MFCC díky nelineární transformaci do Melovy stupnice na rozdíl od DFT-cepstra aproximují lépe lidský sluch, který je výrazně citlivější na rozdíly v nižších frekvencích. Převodem do Melovy stupnice je pro lidské ucho rozdíl mezi 0 a 100 Mely stejný jako mezi 100 a 200 Mely, což neplatí pro frekvenci [13]. Převod z frekvence f do Melovy stupnice m je dán vztahem

$$M(f) = 1125 \ln(1 + \frac{f}{700}) \tag{2.8}$$

a podobně z Melovy stupnice na frekvenci

$$M^{-1}(m) = 700(e^{\frac{m}{1125}} - 1). \tag{2.9}$$

Melova banka filtrů

Výkonové spektrum signálu informuje o tom, jak silně je která frekvence v signálu zastoupená. Obsahuje však také mnoho informací, které jsou pro rozpoznání zbytečné. Lidské ucho téměř nedokáže rozlišit velmi blízké frekvence a zároveň, jak je popsáno výše, vnímá jinak rozdíl mezi frekvencemi určitého řádu. Vytvořením banky filtrů a jejím součinem s výkonovým spektrem jsou získány rovnoměrně rozložené koeficienty reprezentující zastoupení příslušných intervalů frekvencí v řeči [21].

Prvním krokem tvorby banky filtrů je určení spodní a horní hranice frekvencí, pro něž je banka tvořena. Tyto frekvence jsou převedeny do Melovy stupnice pomocí rovnice 2.8 a je nutné vytvořit N[2] rovnoměrně rozložených intervalů. Hranice intervalů jsou následně

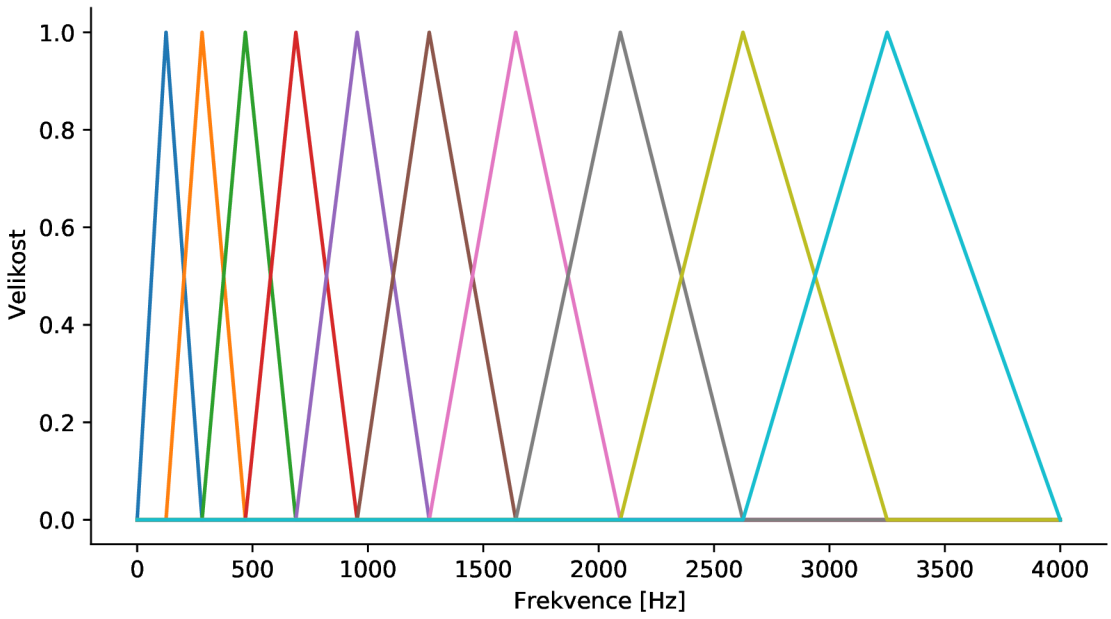
[2]počet MFCC koeficientů

převedeny zpátky do Hertzů podle vztahu 2.9. Tím získáváme hranice intervalů v Hertzích a zbývá pouze převést tato čísla do indexů vektoru výkonového spektra [21].

Nad získanými indexy vytvoříme N trojúhelníkových oken znázorněných na obr. 2.9 podle následujícího vztahu

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \frac{k-f[m-1]}{f[m]-f[m-1]} & f[m-1] <= k <= f[m] \\ \frac{f[m+1]-k}{f[m+1]-f[m]} & f[m] <= k <= f[m+1] \\ 0 & k > f[m+1] \end{cases}, \tag{2.10}$$

kde **f** je vektor indexů hranic bank, **m** příslušné okno a **k** index pole o velikosti rozlišení výkonového spektra.



Obrázek 2.9: Melova banka filtrů pro signál s vzorkovací frekvencí 8 kHz, s rozlišením výkonového spektra 512 vzorků pro 10 MFCC koeficientů.

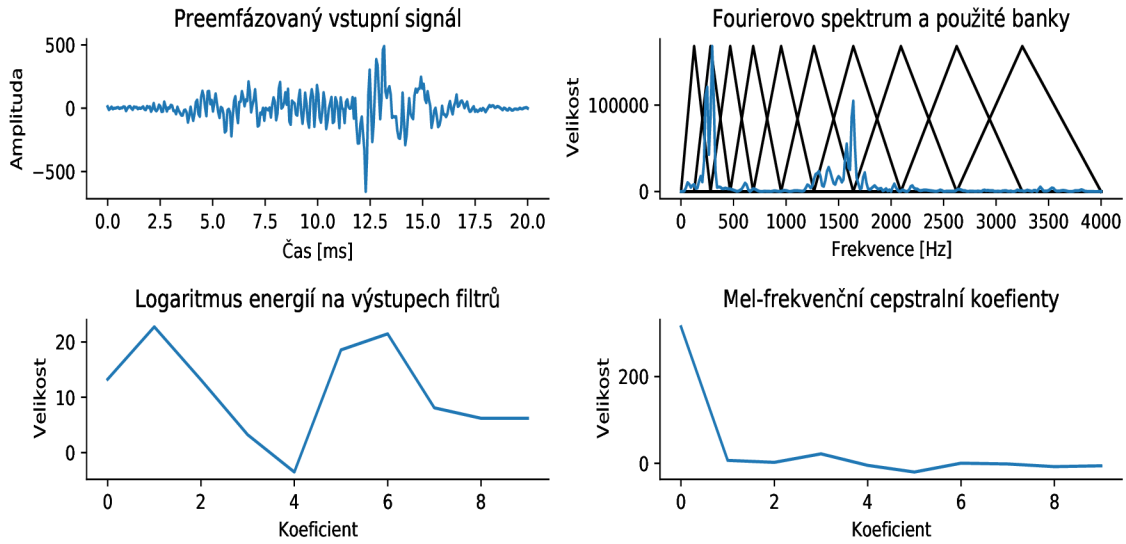
Získání MFCC

Mel-frekvenční cepstrální koeficienty jsou získány následujícím způsobem:

- signál je rozdělen do krátkých segmentů při použití okénkové funkce – nejčastěji je použito Hammingovo okno
- na segmenty je aplikována diskrétní Fourierova transformace
- pro každý segment je vypočteno jeho výkonové spektrum – normalizací mocniny absolutní hodnoty výstupu diskrétní Fourierové transformace
- je provedena nelineární transformace energií příslušných frekvencí do Melovy stupnice a je sečtena suma energií příslušných filtrů

- nad logaritmem energií banky filtrů je provedena diskrétní kosinová transformace

Obr. 2.10 zobrazuje proces extrakce MFCC příznaků jednoho segmentu o délce 20 ms signálu.



Obrázek 2.10: Extrakce 10 MFCC příznaků ze segmentu o délce 20 ms signálu vzorkovaného 8000 vzorků/s s rozlišením 512 vzorků frekvenčního spektra. Nad vstupním signálem je provedena preemfáze a následná segmentace s využitím Hammingova okna. Jeden z takto získaných segmentů je zobrazen nahoře vlevo. Nad signálem obsaženým v segmentu je provedena diskrétní Fourierova transformace a je vypočteno výkonové spektrum. Obrázek vpravo nahoře ukazuje výkonové spektrum a použitou banku filtrů. Následně na obrázku vlevo dole je zobrazen logaritmus sumy energií příslušných filtrů. Nad energií filtrů je provedena diskrétní kosinová transformace, která je zobrazena na obrázku vpravo dole.

Delta koeficienty

Získané MFCC koeficienty lze ještě rozšířit o delta, případně delta-delta MFCC, které obsahují informace o změnách příznaků mezi sousedními segmenty, definované jako

$$D[k][m] = \frac{\sum_{n=1}^{N} n(C[k+n][m] - C[k-n][m])}{2\sum_{n=1}^{N} n^2}, k \in \{1, \ldots, K\}, m \in \{1, \ldots, M\}, \quad (2.11)$$

kde K je počet segmentů signálu, M počet MFCC koeficientů, N počet sousedních segmentů, nad kterými má být výpočet dynamiky proveden a $\mathbf{C}$ je matice MFCC koeficientů o velikosti $K \times M$. Delta příznaky jsou podle vztahu 2.11 vypočítány nad MFCC příznaky, a obdobně delta-delta nad delta koeficienty [21].

2.4 Shlukovací algoritmy

Shlukovací algoritmy slouží k třídění dat do skupin (shluků) tak, aby si data náležející do stejné skupiny byla co nejvíce podobná a zároveň variance mezi třídami byla co nejvyšší.

Cílem shlukové analýzy je se co nejjednoznačněji rozhodnout, do které skupiny „dato“ náleží. V této práci jsou shlukovací algoritmy použity v procesu detekce řečové aktivity a diarizace.

2.4.1 Směs Gaussovských rozložení

Směs Gaussovských rozložení (GMM – Gaussian Mixture Model) je parametrická funkce hustoty pravděpodobnosti reprezentovaná jako vážený součet hustot Gaussovských komponent. Sekce vychází z následujícího článku [31].

GMM se běžně používají jako parametrický model distribuce pravděpodobnosti kontinuálních měření nebo v biometrických systémech. Parametry modelu (váhy, střední hodnoty a kovariační matice) jsou odhadnuty z trénovacích dat pomocí iteračního algoritmu Expectation-Maximization (EM), jelikož neexistuje žádné analytické řešení pro odhad těchto parametrů. V některých případech může být využit i Viterbiho algoritmus, který je velmi intuitivní, avšak kvůli vysoké pravděpodobnosti uvíznutí v lokálním minimu nemusí vždy vést k nalezení globálního optima.

Gaussovská směsice je reprezentována jejími středními hodnotami, kovariančními maticemi a váhami příslušných komponent. Tyto parametry jsou společně reprezentovány následující anotací

$$\mathbf{\Lambda} = \{\omega_i, \boldsymbol{\mu}_i, \mathbf{\Sigma_i}\} \quad \text{pro} \quad , i \in \{1, \ldots, M\}, \tag{2.12}$$

kde M je počet komponent Gaussovské směsice, $\boldsymbol{\mu}_i$ vektor středních hodnot a $\mathbf{\Sigma}_i$ kovarianční matice. Váhy komponent ω_i splňují následující rovnici

$$\sum_{i=1}^{M} \omega_i = 1. \tag{2.13}$$

Kovarianční matice $\mathbf{\Sigma}_i$ modelu $\mathbf{\Lambda}$ mohou být omezeny na diagonální matici, v takovém modelu předpokládáme, že vzorky modelované veličiny jsou na sobě statisticky nezávislé. Kovarianční matice mohou být taktéž sdílené mezi komponentami. Směs Gaussovských rozložení je vážená suma M komponent Gaussovských hustot daná vztahem

$$p(\mathbf{x}|\Lambda) = \sum_{i=1}^{M} \omega_i \cdot \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \mathbf{\Sigma}_i), \tag{2.14}$$

kde $\mathbf{x}$ je D-dimenzionální vektor (reprezentující objekt ze skupiny sledovaných dat), ω_i váha příslušné komponenty, a $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$ hustota Gaussovské komponenty. Příslušné hustoty komponent Gaussovské směsice odpovídají hustotám D-rozměrných normálních rozložení definovaných jako

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \mathbf{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_i|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{\Sigma}^{-1}(x-\boldsymbol{\mu}_i)}, \tag{2.15}$$

Jak již bylo zmíněno, směs Gaussovských rozložení je velmi často využívána v biometrických systémech, resp. systémech rozpoznání mluvčího, kde dokáže velmi dobře reprezentovat velkou třídu dat příslušné distribuce.

EM algoritmus

Již bylo vysvětleno, jaké parametry reprezentují GMM, ale otázkou zůstává, jak nalézt takové parametry, aby věrohodnost trénovacích dat byla jak nejvyšší. Existuje mnoho technik pro nalezení optimálních parametrů, nejznámější a nejvíce používanou je Expectation Maximization algoritmus. Expectation Maximization [6] je iterativní algoritmus pro trénování generativních modelů se skrytými proměnnými. Každá iterace tohoto algoritmu vede ke zvýšení věrohodnosti trénovacích dat. Algoritmus ovšem nezaručuje nalezení globálního optima.

Pro posloupnost T tréninkových dat $\mathbf{X} = \{\mathbf{x}_1, \cdot, \mathbf{x}_T\}$, věrohodnost GMM, za předpokladu nezávislosti mezi trénovacími daty, můžeme zapsat jako

$$p(\mathbf{X}|\mathbf{\Lambda}) = \prod_{t=1}^{T} p(\mathbf{x_t}|\mathbf{\Lambda}). \tag{2.16}$$

Tento výraz je bohužel nelineární funkcí parametrů $\mathbf{\Lambda}$ a přímá maximalizace není možná. Naštěstí je možné nalézt maximálně věrohodné (ML – Maximum Likelihood) parametry iterativně. Základní myšlenkou je, vycházeje z modelu reprezentovaným parametry $\mathbf{\Lambda}_{old}$, odhadnout nové parametry $\mathbf{\Lambda}_{new}$ tak, že $p(\mathbf{X}|\mathbf{\Lambda}_{new}) >= p(\mathbf{X}|\mathbf{\Lambda}_{old})$. Nový model bude dále inicializačním modelem dalšího kroku a iterační algoritmus pokračuje, dokud není splněna ukončující podmínka, kdy rozdíl věrohodnosti nového a starého modelu je menší než práh F.

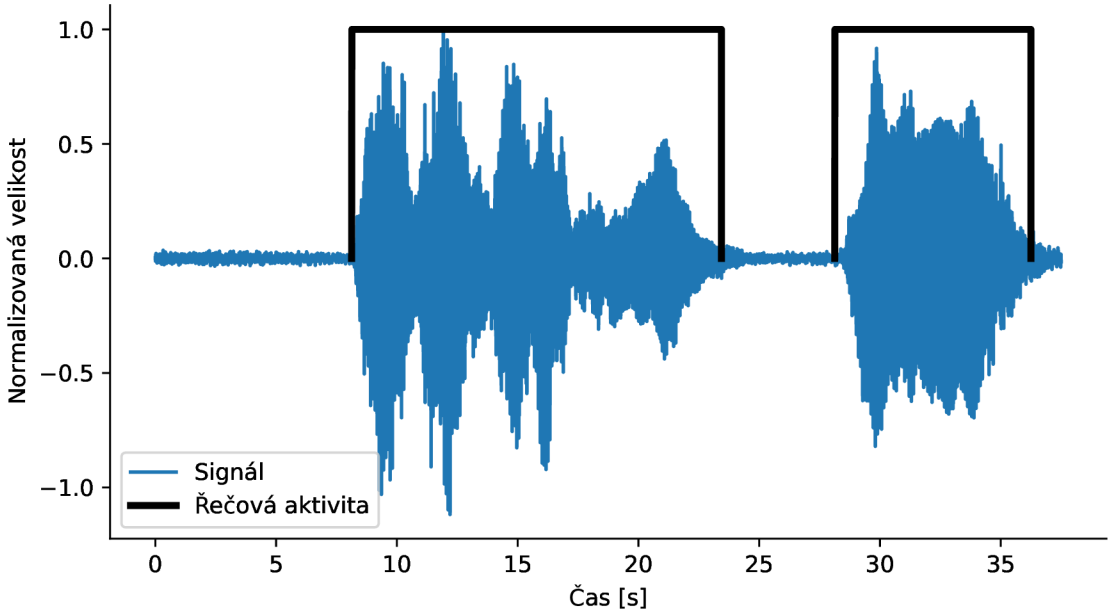
$$p(\mathbf{X}|\mathbf{\Lambda}_{new}) - p(\mathbf{X}|\mathbf{\Lambda}_{old}) < F \tag{2.17}$$

V každé iteraci dochází k výpočtu nových parametrů modelu. Nejdříve jsou v kroku E (Expectation step) vypočteny statistiky, konkrétně věrohodnosti příslušných komponent Gaussovské směsice vůči datům. V kroku M (Maximalization step) dochází k aktualizaci parametrů $\mathbf{\Lambda}$ podle následujících vztahů tak, aby věrohodnost modelu vůči trénovacím datům byla opět co nejvyšší.

1. Krok E:
   - Věrohodnost γ_{ct} Gaussovské komponenty c vůči vektoru dat $\mathbf{x}_t$ a parametrům $\mathbf{\Lambda}_{old}$
   $$P(c|\mathbf{x}_t, \mathbf{\Lambda}_{old}) = \frac{\omega_c^{old}\mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_c^{old}, \mathbf{\Sigma}_c^{old})}{\sum_{k=1}^{M}\omega_k^{old}\mathcal{N}(\mathbf{x_t}|\boldsymbol{\mu}_k^{old}, \mathbf{\Sigma}_k^{old})} = \gamma_{ct} \tag{2.18}$$
2. Krok M:
   - Nové váhy modelu
   $$\omega_c = \frac{1}{T}\sum_{t=1}^{T}\gamma_{ct}, c \in \{1, \ldots, M\} \tag{2.19}$$
   - Nové střední hodnoty
   $$\boldsymbol{\mu}_c = \frac{\sum_{t=1}^{T}\gamma_{ct}\mathbf{x}_t}{\sum_{t=1}^{T}\gamma_{ct}}, c \in \{1, \ldots, M\} \tag{2.20}$$
   - Nové rozptyly (kovarianční matice)
   $$\mathbf{\Sigma}_c^2 = \frac{\sum_{t=1}^{T}\gamma_{ct}\mathbf{x}_t^2}{\sum_{t=1}^{T}\gamma_{ct}} - \boldsymbol{\mu}_c^{-2}, c \in \{1, \ldots, M\} \tag{2.21}$$

2.5 Detekce řečové aktivity

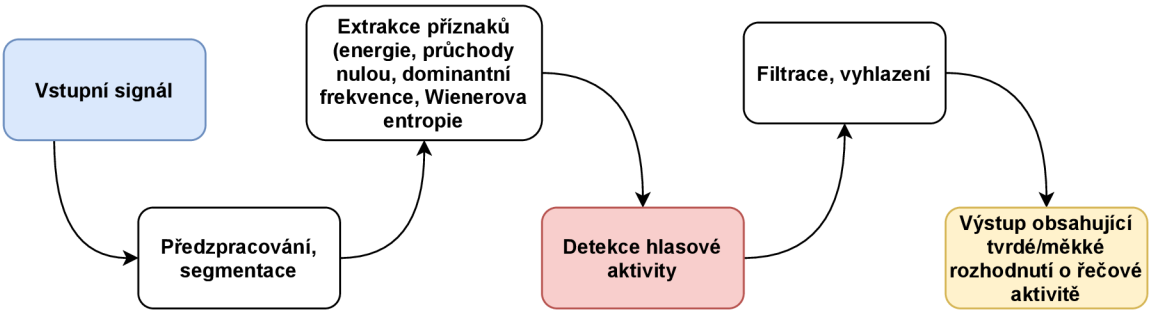
Segmentovaný signál nahrávky lidské řeči obsahuje vysoký podíl segmentů, kde se vyskytuje pouze šum okolí, případně nedochází k žádné řečové aktivitě mluvčího. Detekce řečové aktivity (VAD – Voice activity detecion) je úloha zabývající se určením, kdy opravdu dochází k mluvě. Tato úloha je základním kamenem všech systémů zabývajících se kódováním řeči, rozlišením řečníků v nahrávce, extrakci informací o mluvčích z nahrávky nebo systémů monitorujících zákaznická call centra. VAD je obvykle prvním stavebním kamenem těchto systémů. Cílem je co nejpřesněji a za co nejnižší výpočetní náklady rozlišit segmenty řečové aktivity od těch, které obsahují ticho, šum. Existuje mnoho řešení, začínajících od základních algoritmů, rozhodujících se pouze podle střední krátkodobé energie signálu, až po algoritmy používající spojení vícero komplexních modelů. Obr. 2.11 znázorňuje, jak by měl takový systém fungovat.



Obrázek 2.11: Detekce řečové aktivity na signálu o délce 8 sekund. Úseky, ve kterých je okno v hodnotě 1, jsou považovány za úseky řečové aktivity řečníka. Úseky okna v 0 reprezentují ticho.

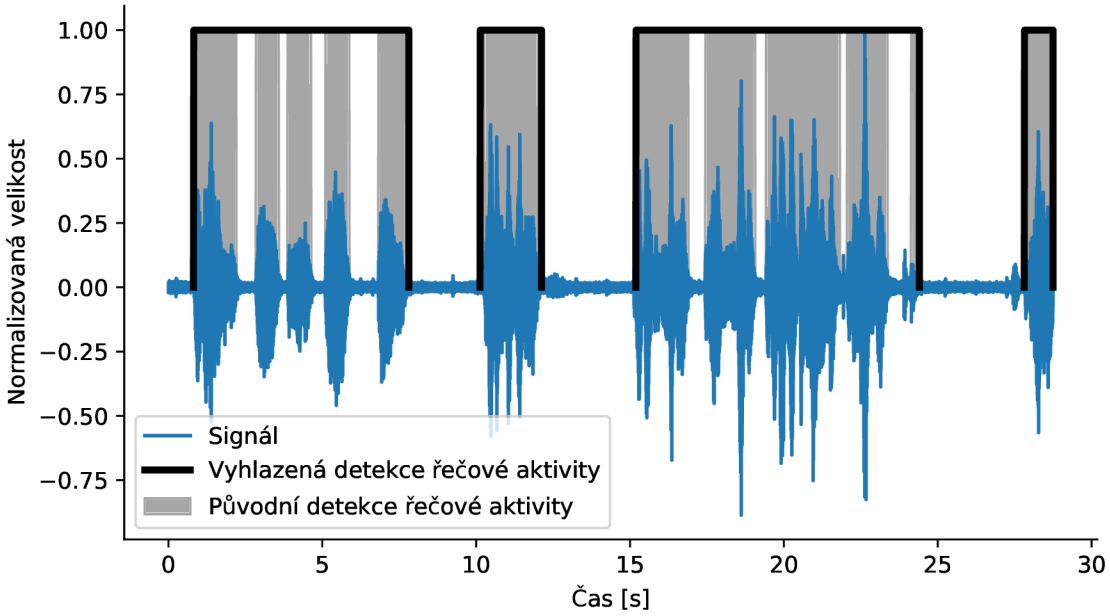
Obvyklé algoritmy detekce řečové aktivity nejdříve extrahují příznaky daného úseku nahrávky a ty se následně klasifikují pomocí diskriminačních modelů. Většina VAD systémů provádí tvrdé rozhodnutí na řeč a ticho. Obr. 2.12 znázorňuje koncepci základního detektoru řečové aktivity [2].

Některé systémy provádějí rozhodnutí nad rámci, které mohou reprezentovat úseky řeči o délce zhruba 10 ms, avšak délky fonému slovanských jazyků se pohybují někde mezi jednotkami až vyššími stovkami ms [16]. Z toho vyplývá, že statisticky v absolutní většině případů bude délka promluvy zahrnovat více než 1 segment. Taktéž nemusí být vždy úplně vhodné detekovat ticho o délce 10 ms v delším úseku plynulé řeči. Řešením je provést postprocessing a odstranit krátké úseky řeči/ticha. Existuje mnoho metod, kterými je možné vyhladit výstup systému VAD. Mezi běžně používané patří různé filtry jako třeba mediánový nebo průměrový filtr [11]. Lze taktéž využít předpokladu, že data pocházejí ze



Obrázek 2.12: Blokový diagram systému detekce řečové aktivity. Vstupní signál je předzpracován a segmentován do N úseků stejné délky. Ze segmentů jsou extrahovány příznaky, pomocí kterých je rozhodnuto, zda se jedná o řečově aktivní úsek. Výstup je následně vyhlazen a jsou odstraněny špičky.

skrytého Markovova modelu (HMM) [40] a použít forward-backward algoritmus s přechodovou maticí modelující pravděpodobnost přechodu ze stavu řeč/ticho a opačně [1]. Obr. 2.13 demonstruje jeden ze způsobů vyhlazení řečové aktivity.



Obrázek 2.13: Detekce řečové aktivity. Nad segmenty, které byly označeny za řečově aktivní, je provedeno filtrování. Hodnoty řečové aktivity v úsecích ticha, případně mluvy o délce menší než 500 milisekund byly negovány, čímž vznikly plynulé úseky mluvy.

2.5.1 Energetické VAD

Detekovat řečovou aktivitu lze velmi dobře podle energie signálu v případě, že signál není velmi výrazně „zašumělý“. Energeticky vyšší segmenty lze považovat za řeč, segmenty se střední energií jako šum a nízko-energetické segmenty jako ticho. V případě normalizované energie $E \in (-1,0; 1,0)$ lze předpokládat, že ticho má zápornou energii, šum se pohybuje

kolem 0 a řeč je reprezentována kladnou energií. Takto navržený systém však nedokáže rozpoznat energii mluvy a energii zvuku, který dokážou vyvolat jiné objekty kolem nás. Většina moderních mikrofonů však dokáže potlačit tyto okolní zvuky a lze předpokládat, že psychoterapeutická sezení probíhají v tichých, klidných prostorech.

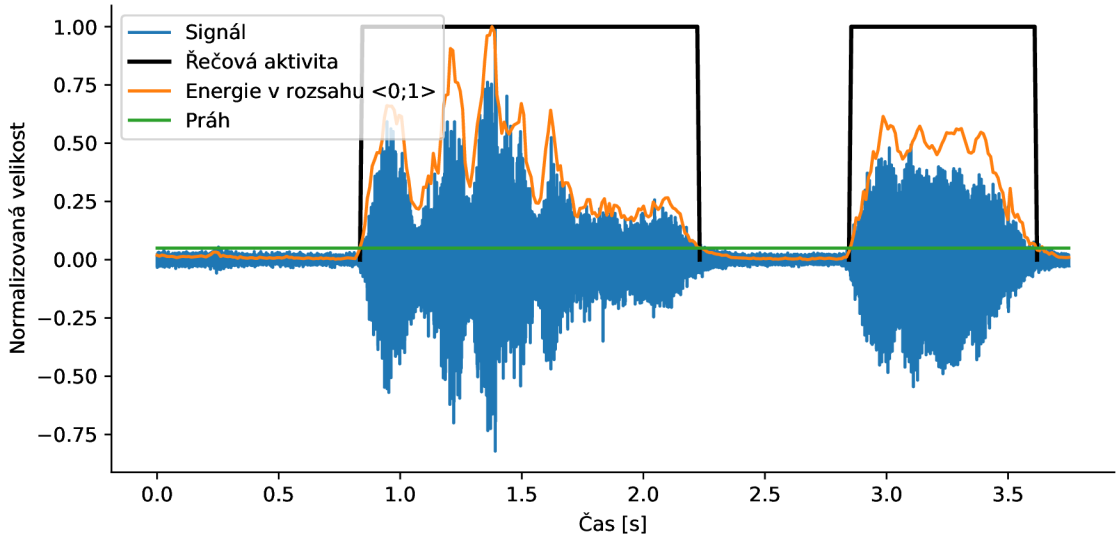
Prahování

Velmi jednoduchým, ale přesto celkem efektivním algoritmem je prahování podle energie segmentů. Prahovací funkce je definována jako

$$f_n = \begin{cases} 1 & E_n >= T \\ 0 & E_n < T \end{cases}, n \in \{1, \dots N\} \tag{2.22}$$

kde **E** je vektor energií segmentů signálu, T energetický práh a N počet segmentů signálu.

Tento algoritmus může být rozšířen o další příznaky, jako je počet průchodů nulou, dominantní frekvenci spektra nebo Wienerovu entropii [23]. Poté je nutné určit, zda je postačující, aby byla splněna jedna podmínka, nebo zda musí být splněny všechny, aby byl segment klasifikován jako řeč. Ačkoliv se jedná o vcelku jednoduchý algoritmus, je nutné nastavit všechny prahy manuálně, což v případě vysoké variance mezi nahrávkami může způsobovat špatné výsledky. Prahy lze dynamicky získat z prvních M segmentů (± 50 ms), u kterých lze předpokládat ticho na začátku nahrávky. Obr. 2.14 znázorňuje demonstrační příklad prahování signálu podle energie.



Obrázek 2.14: Detekce řečové aktivity pomocí prahování energie. Hodnota prahu je rovna $0,05$ v relativním měřítku, segmenty s relativní energií menší než práh jsou klasifikovány jako ticho a segmenty s vyšší energií jako řeč.

Adaptivní práh

Jak bylo zmíněno v předchozí podsekci 2.5.1 systémy založené na prahování nedokážou upravovat hodnotu prahu dynamicky a ne vždy je možné nastavit práh pro zcela odlišné

nahrávky. Hlavní myšlenkou tohoto algoritmu je adaptivní aktualizace prahu pomocí výpočtu minimální a maximální hodnoty energií doposud zpracovaných segmentů. Tímto odpadá nutnost inicializace prahu před spuštěním [19].

Algoritmus pracuje se střední kvadratickou energií popsanou v podsekci 2.3.1. Algoritmus postupně prochází segmenty a aktualizuje maximální energii **X** a minimální energii **M**. Všechny následující vzorce jsou definovány pro $n \in \{1, \ldots, N\}$, kde N je počet segmentů signálu.

$$Xn = max(E_n, X_{n-1}), \tag{2.23}$$

$$M_n = min(E_n, j_{n-1} \cdot M_{n-1}), \tag{2.24}$$

kde $max(a, b)$ je funkce vracející větší číslo z dvojice a, b a $min(a, b)$ funkce vracející menší číslo z dvojice a, b. Adaptivní koeficienty vektoru **j** jsou aktualizovány podle následujících pravidel.

$$j_n = \begin{cases} 1 & E_n < M_{n-1} \\ 1,0001 \cdot j_{n-1} & jinak \end{cases} \tag{2.25}$$

Vektor prahů **T** pro příslušné segmenty je definován následovně

$$\lambda_n = min(0,95; \frac{X_n - M_n}{X_n}), \tag{2.26}$$

$$T_n = X_n \cdot (1 - \lambda_n) + M_n \cdot \lambda_n. \tag{2.27}$$

Následné prahování již odpovídá algoritmu popsanému v předchozí podsekci 2.5.1. Obr. 2.15 znázorňuje hodnoty dynamického prahu příslušných segmentů signálu.

Gaussovská směsice tří komponent

Na rozdíl od algoritmů popsaných v předchozích sekcích je díky tomuto přístupu možné získat měkká rozhodnutí v podobě pravděpodobnosti, že se jedná o řeč, ticho nebo šum. Tyto pravděpodobnosti lze dále vyhladit použitím forward-backward algoritmu nebo jiné filtrační metody. Metoda je založena na směsi tří Gaussovských jednorozměrných rozložení reprezentujících řeč, šum a ticho. Model je trénován na energiích segmentů signálu **E** a jeho parametry jsou inicializovány následovně

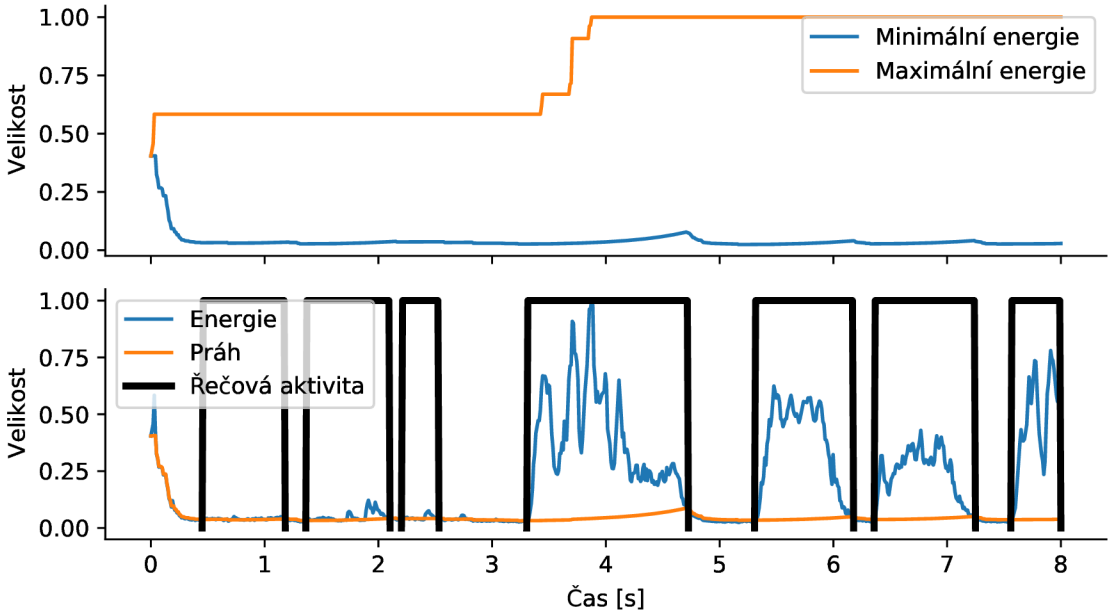
$$\omega_1 = \omega_2 = \omega_3 = \frac{1}{3} \tag{2.28}$$

$$\mu_1 = min(\mathbf{E}) \tag{2.29}$$

$$\mu_2 = mean(\mathbf{E}) \tag{2.30}$$

$$\mu_3 = max(\mathbf{E}) \tag{2.31}$$

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1 \tag{2.32}$$



Obrázek 2.15: Princip funkcionality adaptivního prahu. Na obrázku je zobrazena energie signálu, její minimální a maximální hodnota příslušných rámců, adaptivní práh a řečová aktivita vyhlazená mediánovým filtrem o velikosti 20 milisekund.

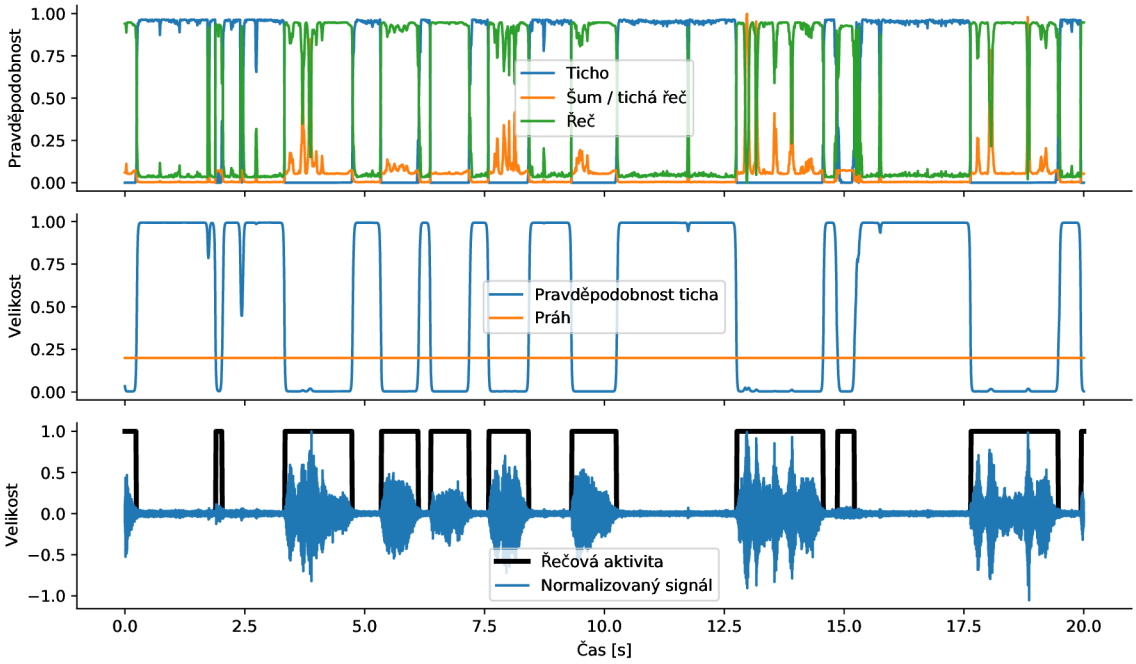
Parametry Gaussovské směsice dodržují značení definované v podsekci 2.4.1. Funkce označena jako $mean(\mathbf{X})$ vrací střední hodnotu položek vektoru $\mathbf{X}$, obdobně $min(\mathbf{X})$ nejmenší položku a $max(\mathbf{X})$ tu největší.

Model je trénován v několika iteracích dokud není dosažena konvergence. Následně je vypočtena věrohodnost každého segmentu vůči každé komponentě modelu. Takto je získán vektor o velikosti $3 \times N$, kde N je počet segmentů signálu. Vektor je vyhlazen forward-backward algoritmem s příslušnou přechodovou maticí.

Posledním krokem je určení řečově aktivních segmentů. Jako aktivní segmenty jsou zvoleny ty, kde je věrohodnost ticha nižší než práh. Aktivní segmenty jsou následně vyhlazeny podle potřeb následného zpracování. Obr. 2.16 znázorňuje funkcionalitu takového systému.

2.6 Diarizace

Diarizace je proces rozdělení audio nahrávky na homogenní úseky odpovídající příslušným mluvčím. Odpovídá na otázku „kdo kdy mluví“ v prostředí více mluvčích. Metoda je využívána v mnoha odvětvích zpracování a následné analýzy řeči. Přesnosti systémů se každým rokem zlepšují a přesnosti komerčně využívaných systémů se blíží ke 100 %. Správně označené úseky řeči mluvčích vedou k výraznému zlepšení systémů rozpoznání řeči (ASR – Automatic speech recognition), tedy systémů tvořících přepis (transkripci) řečových nahrávek. Typický diariya4n9 systém se skládá ze segmentace, extrakce příznaků, shlukování a případně resegmentace.



Obrázek 2.16: Detekce řečové aktivity pomocí směsice Gaussovských rozložení. Na horním obrázku jsou zobrazeny pravděpodobnosti příslušných komponent reprezentujících ticho, šum a řeč. Pravděpodobnosti jsou vyhlazeny forward-backward algoritmem s příslušnou přechodovou maticí. Tvrdé rozhodnutí o řečové aktivitě podle věrohodnosti komponenty reprezentující ticho demonstruje obrázek uprostřed. Spodní obrázek zobrazuje detekci řečové aktivity pro systémy velmi citlivé na aktivitu mluvčího.

Segmentace

Nahrávka se rozdělí do N segmentů o stejné délce M popsané v sekci 2.2.2. Ze segmentů jsou dále odfiltrovány ty, které neobsahují řečovou aktivitu podle sekce 2.5.

Extrakce příznaků

Z řečově aktivních segmentů jsou extrahovány příznaky. Nejčastěji se jedná o MFCC příznaky popsané v podsekci 2.3.2, faktory řečníka a jejich vlastní čísla [3], i-vektory [5], x-vektory [34] nebo případně d-vektory [35].

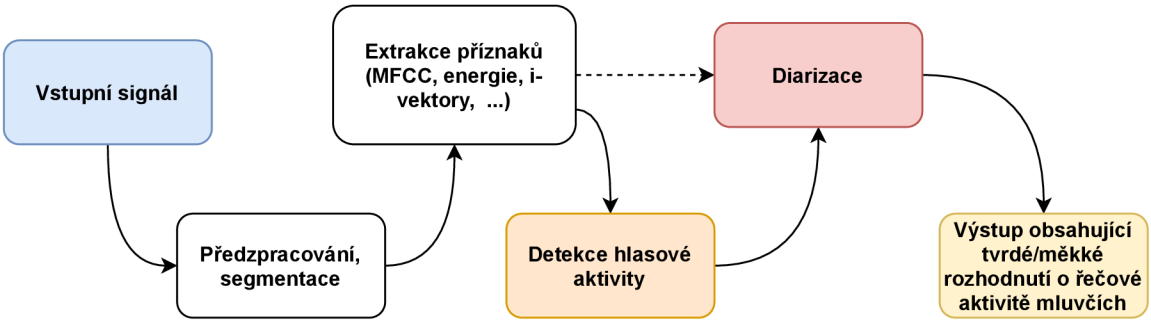
Shlukování

Příslušné segmenty signálu jsou přiřazeny řečníkům. Nejčastěji jsou za tímto účelem využity neuronové sítě [1] nebo směsi Gaussovských rozložení popsané v podsekci 2.4.1.

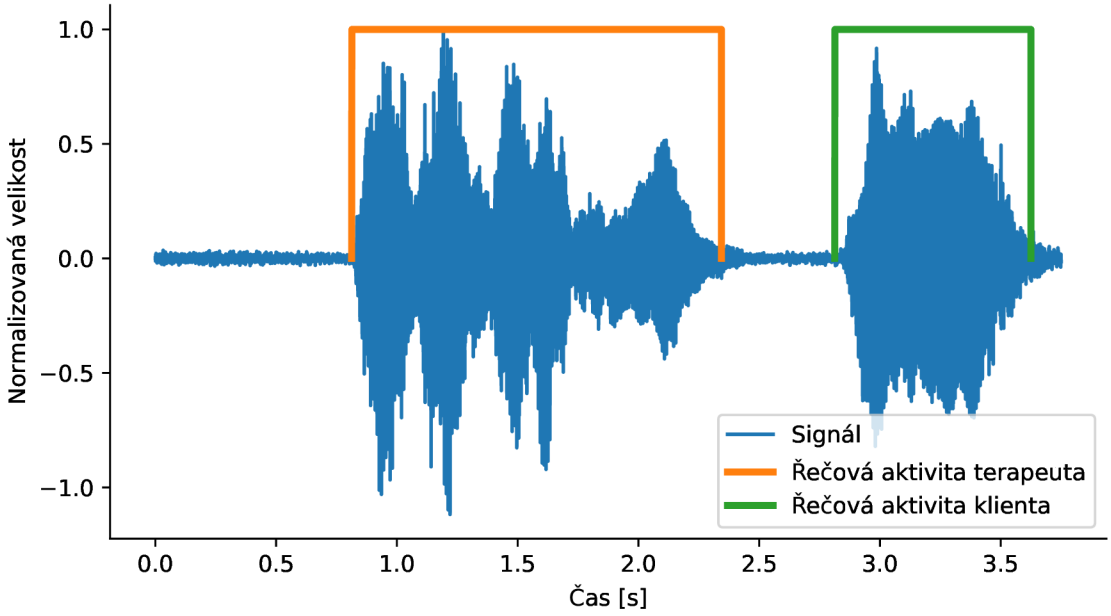
Resegmentace

Podobně jako u detekce řečové aktivity je vhodné provést vyhlazení výsledků pomocí filtrace nebo forward-backward algoritmu pomocí přechodové matice definující pravděpodobnosti změny řečníků.

Blokové schéma na obr. 2.17 znázorňuje základní architekturu diarizačního systému. Obr. 2.18 zobrazuje demonstrační výstup diarizačního systému.



Obrázek 2.17: Blokové schéma diarizace.



Obrázek 2.18: Detekce řečové aktivity příslušných řečníků vstupního signálu.

2.7 Rozpoznávání řeči

Následující sekce vychází z článku společnosti IBM[15], která je jedním z vedoucích vývojářů v oblasti ASR[3] (Automatické zpracování řeči). Automatické rozpoznávání řeči je další ze skupiny metod pro získání informací z vstupní nahrávky. Rozpoznáváním řeči se rozumí automatický převod mluvené řeči do textu. Získaný text je následně možné dále zpracovávat, analyzovat nebo vytvářet automatické odpovědi v textové, případně v hlasové podobě pomoci syntetizátoru řeči. Tímto způsobem dokáže v dnešní době komunikovat počítač s člověkem.

[3] Automatic speech recognition

Při návrhu systému ASR je nutno dbát na rozdíly v intonaci, výslovnosti, přízvuku, výšce či hlasitosti mluvčích. Některé metody mohou být velmi citlivé na hluk v pozadí, který může způsobovat špatné výsledky. Obecně rozlišujeme dva druhy ASR systémů:

- závislé na mluvčím (speaker dependent, SD),
- na mluvčím nezávislé (speaker independent, SI).

Systémy závislé na mluvčím dosahují lepších výsledků, ale jejich nevýhodou je, že mluvčí musí namluvit sadu nahrávek, aby bylo možné natrénovat příslušné modely. Existují však metody, které modely natrénované na velkém počtu mluvčích dokáží adaptovat na konkrétní osobu. Takové systémy můžeme pozorovat ve většině mobilních telefonů v podobě Asistenta Google[4] nebo systému Siri[5] firmy Apple. Značné využití systémů ASR můžeme sledovat i ve zdravotnictví nebo v automobilovém průmyslu.

Rozpoznání řeči je přímo závislé na diarizaci nahrávky. Pomocí skrytých Markových modelů [40], N-gramů [17] a neuronových sítí [1] jsou poté vytvořeny modely pro příslušný jazyk, odvětví.

V současné době existuje mnoho firem a univerzit vyvíjejících vlastní interní systémy. Některé z nich jsou poskytnuté v podobě online rozhraní. Mezi nejznámější patří:

- Google Cloud Speech API[6]
- Microsoft Bing Voice Recognition[7]
- IBM Speech to Text[8]
- Wit.ai[9]
- Houndify API[10]

Zároveň existuje mnoho volně dostupných systémů, které je možné natrénovat pro individuální účely. Mezi nejznámější patří DeepSpeech[11] společnosti Mozilla, sada nástrojů Kaldi[12] nebo starší systém CMUSphinx[13].

[4] https://assistant.google.com/
[5] https://www.apple.com/siri/
[6] https://cloud.google.com/speech-to-text
[7] https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/
[8] https://www.ibm.com/cloud/watson-speech-to-text
[9] https://wit.ai/
[10] https://www.houndify.com/
[11] https://github.com/mozilla/DeepSpeech
[12] https://github.com/kaldi-asr/kaldi
[13] https://cmusphinx.github.io/

Kapitola 3

Data

Automatické zpracování a následná analýza psychoterapeutických sezení vyžaduje databázi dat, vůči kterým je možné příslušná sezení porovnávat a strojově zpracovávat. Při zpracování nahrávek jsou využívány algoritmy strojového učení. Ty v datech hledají vzory, které dokáží formulovat předpovědi. S větším množstvím dat a více zkušenostmi jsou výsledky strojového učení přesnější – stejně jako se lidé zlepšují díky větší praxi. Psychoterapie je velmi citlivé odvětví, co se týče ochrany dat a tak neexistují žádné volně dostupné datasety. V rámci této práce jsou využita data projektu DeePsy[1], ta však obsahovala v době psaní této práce nedostatečný počet nahrávek, a byly proto doplněna o dataset CallHome[2]. Dataset CallHome byl vybrán z důvodu vysoké podobnosti struktury nahrávek k nahrávkám online psychoterapeutických sezení projektu DeePsy. V obou případech obsahují datasety nahrávky rozhovorů dvou osob, které jsou rozděleny do dvou kanálů bez přeslechů.

3.1 CallHome

Dataset CallHome American English Speech vytvořilo v roce 1997 mezinárodní konsorcium univerzit, knihoven, společností a vládních výzkumných laboratoří Linguistic Data Consortium[3]. Skládá se ze 120 telefonních nahrávek rodilých mluvčích v anglickém jazyce. Délka nahrávek se pohybuje okolo 30 minut a jedná se většinou o rozhovory mezi členy rodiny nebo rozhovory mezi blízkými přáteli. Vzorkovací frekvence nahrávek je 8 kHz. Dataset se skládá ze stereo nahrávek a jejich transkripcí. Transkripce však nepokrývají celou délku nahrávky a v této práci byly využity pouze anotované části datasetu, jedná se o 80 souborů o průměrné délce zhruba 10 minut. Příslušné kanály nahrávky od sebe oddělují mluvčí a neobsahují téměř žádný přeslech. V některých případech obsahují kanály navíc řečovou aktivitu dalších mluvčích. Většinou se však jedná o mluvu o délce jedné věty a pro účely této práce je skupina těchto mluvčích reprezentovaná jedním mluvčím.

3.2 DeePsy

DeePsy je projekt vedený týmem psychologů a psychoterapeutů z Masarykovy univerzity a odborníků na informační technologie z Vysokého učení technického v Brně. Zkoumají možnosti, jakými mohou informační technologie obohatit a zkvalitnit psychoterapeutickou

[1] https://www.deepsy.cz
[2] https://catalog.ldc.upenn.edu/LDC97S42
[3] https://www.ldc.upenn.edu

péči. Na projektu spolupracují Psychosomatická klinika, s.r.o., a Terapeutický přístav, z.ú., kterým účast umožňuje zlepšit kvalitu svých psychoterapeutických služeb. Všichni účastníci poskytují zpětnou vazbu o sezeních a souhlasí s jejich nahráváním. Cílem projektu je poskytnout terapeutům zpětnou vazbu a najít potenciálně problematické úseky k analýze a tím umožnit celkové zkvalitnění poskytované psychoterapeutické péče. Dataset obsahuje anonymizované nahrávky prezenčních a online sezení. Délka psychoterapeutických sezení většinou nepřekračuje 60 minut, obvykle se pohybuje okolo 50 minut.

3.2.1 Online sezení

Online sezení probíhají výhradně na platformě ZOOM[4]. Tato platforma umožňuje stáhnout video/audio záznam ihned po ukončení sezení. Při správném nastavení aplikace[5] jsou od sebe odděleny kanály mluvčích. Vzniká tak separátní nahrávka pro každého účastníka sezení, což vede k vyšší přesnosti diarizace blížící se k 100 %. Nejlepší výsledky jsou dosaženy na nahrávkách, ve kterých terapeut i klient používají náhlavní sady (sluchátka s mikrofonem). Takovéto nahrávky neobsahují skoro žádný okolní hluk a jsou ideální pro následné zpracování.

3.2.2 Prezenční sezení

Druhá skupina nahrávek je pořizována pomocí diktafonu ZOOM H2n. Diktafon je používán v režimu 4CH s maximální citlivostí. Toto nastavení vede k získání stereo nahrávky sezení, ve které je značně eliminován okolní hluk. Kanály však obsahují značný přeslech, který je demonstrován na obr. 3.1a, což vyžaduje náročnější zpracování pro určení, kdy kdo mluví. Pro optimální výsledky je vhodné diktafon umístit ve stejné vzdálenosti mezi terapeutem a klientem, jak znázorňuje obr. 3.1b. V opačném případě může být energie jednoho z řečníků značně vyšší.

3.3 Anotační soubory

Datová sada CallHome obsahuje kromě zvukových souborů taktéž anotační soubory s přepisem úseku nahrávky ve formátu `.txt`. Ty jsou použity pro validaci výsledků diarizace a automatického přepisu. Pro účely validace diarizace jsou převedeny do formátu Rich Transcription Time Marked (RTTM) [24], pomocí kterého je vyhodnocena přesnost systému. Pro nahrávky projektu DeePsy byly anotace vytvořeny ručně pomocí nástroje Transcriber[6] Ty jsou taktéž převedeny do formátu RTTM. Standardizovaný formát RTTM je znám především díky jeho využití v soutěžích rozpoznání mluvčích Národního institutu standardů a technologií (NIST)[7].

Formát RTTM je definován následujícími poli oddělenými mezerami. Každý řádek popisuje jeden úsek souboru:

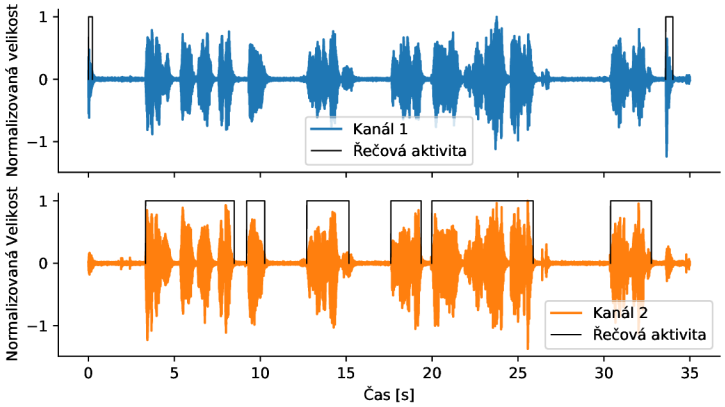
1	2	3	4	5	6	7	8	9	10
Type	File	Channel	Tstart	Tdur	Ortho	Stype	Name	Confidence	Slat

[4] https://zoom.us/

[5] https://support.zoom.us/hc/en-us/articles/201362473-Local-recording#h_96380e17-816d-4eda-a4b9-740d1498eac6

[6] http://trans.sourceforge.net/en/presentation.php

[7] https://www.nist.gov



(a) Přeslech mezi kanály nahrávky. Z obrázku je patrné, že řečová aktivita mluvčích je zaznamenána v obou kanálech zdrojové nahrávky. Je tedy nutné provést diarizaci nad vstupní nahrávkou.

(b) Optimální poloha diktafonu v průběhu sezení. Červená šipka označuje směr, ze kterého hledí terapeut na diktafon.

Obrázek 3.1: Nahrávky pořízené diktafonem ZOOM H2n.

Hodnoty polí *Ortho*, *Stype*, *Confidence* a *Slat* obsahují hodnotu `<NA>` a nejsou pro účely této práce využity. Hodnoty pole *Type* a *Stype* blíže definují původce mluvy nebo řečové aktivity a jsou zde použity konstantní hodnoty `SPEAKER` a `<NA>`. Hodnota *File* udává název zdrojového souboru, ke kterému se nahrávka váže (bez přípon a cesty). Pole *Channel* určuje kanál, ke kterému náleží příslušná anotace. *Tstart* určuje začátek řečové aktivity v sekundách a *Tdur* její délku. Hodnota *Name* obsahuje přidělený identifikátor řečníka.

Následující ukázka představuje dva segmenty řečové aktivity řečníků `B` a `A` zdrojového souboru **en_4077.wav**. V nahrávce se nejdříve vyskytuje řečová aktivita řečníka `B` o délce 1,55 s. Následuje 20 ms ticha a velmi krátká odpověď o délce 270 ms.

```
SPEAKER en_4077 2 0.000 1.550 <NA> <NA> B <NA> <NA>
SPEAKER en_4077 1 1.570 0.270 <NA> <NA> A~<NA> <NA>
```

Kapitola 4

Návrh systému

Analýza audio hovoru je velmi rozsáhlý pojem spojený s řadou technik pro dolování informací z nahrávky. V této kapitole je čtenáři postupně představen systém navržený pro analýzu psychoterapeutických sezení. V sekci 4.1 jsou sepsány všechny funkcionální požadavky, které by systém pro analýzu citlivých dat, jako jsou nahrávky psychoterapeutických sezení, měl splňovat. V následující sekci 4.2 jsou podrobně představeny funkcionální bloky navrženého systému. Systém by měl postupně zpracovávat vstupní nahrávku, extrahovat nezbytné informace a vytvářet souhrnný výstup. Ten je popsán v sekci 4.3. Funkcionalitu systému je potřeba validovat. Metriky k tomu určené jsou popsané v sekci 4.4.

4.1 Funkcionální požadavky

Systém pracuje s daty, která jsou velmi citlivá a jejich únik by mohl být pro uživatele velmi nepříjemný. Z tohoto důvodu musí být systém schopný provozu výhradně v **offline** režimu bez využití jakýchkoliv online nástrojů. Pro zajištění co největší bezpečnosti je taktéž nutné limitovat výstupy systému a to takovým způsobem, že může systém zanechávat soubory a datové výstupy pouze v místech k tomu explicitně určených. Pro vyšší **bezpečnost** je uživateli vřele doporučena anonymizace dat, kterou však systém explicitně nebude zajišťovat.

Jelikož dlouhá doba zpracování může být pro uživatele nepříjemná a poskytnutí zpětné vazby v době, kdy má terapeut sezení ještě čerstvě v paměti, může být značně přínosnější, dalším z požadavků je **rychlost** zpracování nahrávky. Z tohoto důvodu musí být **komplexita** systémů snížena na co nejmenší možnou míru a k implementaci musí být využity nástroje s co nejkratší dobu zpracování. Měl by být kladen důraz na **optimalizaci** příslušných (vysoce náročných) bloků systému.

Příslušné funkcionální bloky popsané v následující sekci 4.2 musí fungovat **odděleně** a musí být možná jejich snadná **nahraditelnost**. Bloky by měly být navzájem **nezávislé** a data načítat výhradně ze vstupních souborů a zapisovat do souborů výstupních.

Navržený systém musí extrahovat nejméně **6 slabých klasifikátorů**, které popisují uskutečněné sezení (např. základní tón řeči, energie, cross-talk, rychlost řeči, poměr řeči, reakční dobu, přepis řeči, sentiment, ...). Ne všechny získané klasifikátory musí být nutně zahrnuty do výstupní souhrnné zprávy.

Výstupní souhrnná zpráva dále popsaná v sekci 4.3 by měla být v co nejvyšší míře **přenositelná**. Z tohoto důvodu bude využit formát HTML [36], který může být v průběhu dalšího vývoje doplněn o elementy uživatelské interakce. Další možnou variantou je

PDF[1] soubor. Ten však nepodporuje elementy interakce v takové míře jak HTML dokument rozšířený o kaskádové styly a jazyk Javascript[10] a nebude proto využit.

4.2 Architektura systému

Jak bylo zmíněno v předchozí sekci, systém je rozdělen na jednotlivé bloky. Ty dovolují postupné zpracování nahrávky po částech. V rámci spuštění příslušných částí systému je možné zpracovávat více nahrávek najednou. Systém se sestává ze skupiny sedmi funkcionálních bloků, které jsou zodpovědné za:

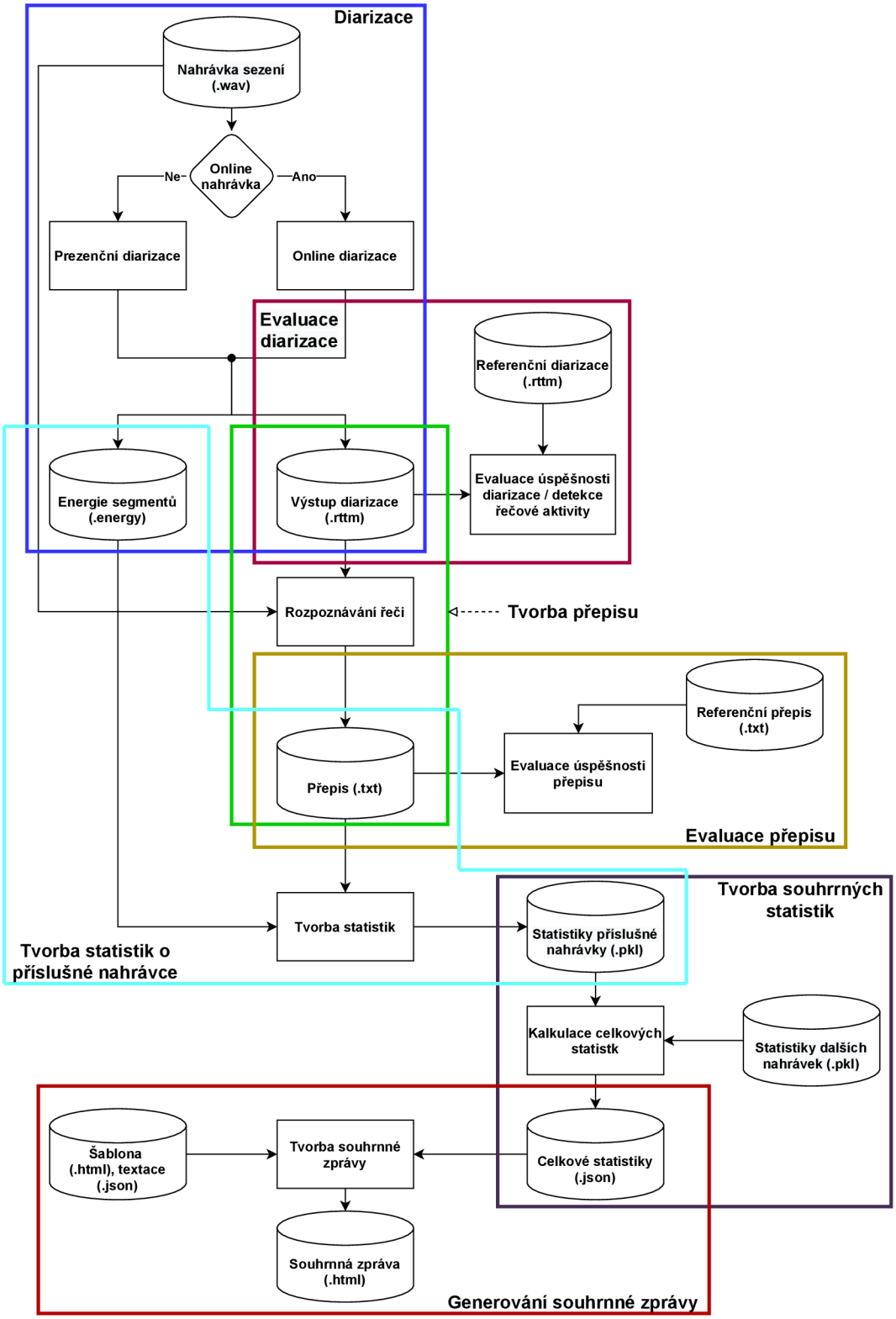
- diarizaci nahrávky online/prezenčního sezení,
- evaluaci diarizace (zahrnuje evaluaci detekce řečové aktivity),
- tvorbu přepisu z řečově aktivních úseků,
- evaluaci přepisu,
- kalkulaci statistik,
- kalkulaci celkových statistik všech dostupných nahrávek,
- tvorbu souhrnné zprávy příslušné nahrávky.

Výše zmíněné bloky je možné vzájemně řetězit. Aby nebyla nutná konverze při načítání a ukládání dat, příslušné bloky sdílejí formát vstupně-výstupních operací a formáty souborů. Obr. 4.1 zobrazuje architekturu navrženého systému v podobě diagramu závislostí funkčních bloků a vstupně-výstupních elementů navrženého systému.

Příslušné bloky jsou zároveň navrženy tak, aby v co nejvyšší míře sdílely společnou funkcionalitu, aby nebyla nutná duplicita kódu. Sdílená funkcionalita je implementována mimo příslušné bloky, aby nedošlo k neočekávaným změnám funkcionality bloku při modifikaci jiného bloku.

Systém je navržen pro analýzu dat popsaných v kapitole 3 ve formátu WAV (Waveform audio file format). Je schopný pracovat s libovolnou vzorkovací frekvencí, kterou dynamicky načte, a je možné ho použít pouze pro nahrávky obsahující dva kanály o celkové délce větší než 1 sekunda.

[1] Portable Document Format



Obrázek 4.1: Diagram architektury systému. Systém se sestává ze skupiny sedmi funkcionálních bloků, které jsou od sebe odděleny barevnými čárami. Navržený systém obsahuje 2 bloky, pomocí kterých je možné validovat výstupy.

4.2.1 Parametry

Zpracování audia a jeho analýza dovoluje použití vysokého počtu parametrů, které vedou k odlišnému chování systému. Parametry systému se nachází v jednom sdíleném souboru, a tak není nutné jejich složité dohledávání ve skupině mnoha zdrojových souborů. Návrh předpokládá spuštění celého bloku s neměnnou konfigurací. Změna parametrů po spuštění jednoho bloku může vést k zavádějícím výsledkům zpracování v dalším bloku. Z tohoto důvodu je nutná konfigurace před spuštěním celého systému nebo opětovné spuštění bloků s pozměněnými parametry. Nejzajímavější parametry systému a jejich výchozí hodnoty jsou uvedeny v tabulce 4.1.

Jméno	Hodnota	Popis
gmm_components	32	Počet komponent Gaussovské směsice pro rozpoznání řečníků
window_size	0,02	Velikost jednoho segmentu signálu [s]
window_overlap	0,01	Délka překrytí segmentů [s]
pre_emphasis_coefficient	0,97	Koeficient preemfáze
min_silence_likelihood	0,95	Minimální věrohodnost pro klasifikaci segmentu jako ticho

Tabulka 4.1: Parametry navrženého systému.

4.2.2 Detekce řečové aktivity a diarizace

Systém obsahuje dvě diarizační metody, které jsou optimalizovány pro použití pro příslušný druh psychoterapeutického sezení. Ze vstupní nahrávky s příponou `.wav` jsou extrahovány parametry použité pro detekci řečové aktivity a následnou diarizaci.

Obě metody jsou schopné pracovat se třemi detektory řečové aktivity (VAD) založené na energii signálu, jejichž funkcionalita je postupně popsána v sekci 2.5. Jako výchozí je použit systém směsice tří Gaussovských rozložení popsaný v podsekci 2.5.1, který byl v rámci testování (sekce 6.1) vyhodnocen jako nejúspěšnější. Tento systém používá pro rozhodnutí o řečové aktivitě směsici tří Gaussovských křivek modelujících ticho, šum a řeč. V následujících matematických výrazech považujme N za počet segmentů kanálu zdrojové nahrávky.

Online sezení

V případě diarizace online nahrávky (nahrávky neobsahující přeslech) je spuštěna detekce řečové aktivity příslušných kanálů nahrávky. Takto je získána matice $\mathbf{V} \in \{0,1\}^{N\times 2}$ booleovských hodnot aktivity segmentů příslušných kanálů. Získanou matici je možné vyhladit, případně odstranit krátkodobé změny řeč/ticho. Vyhlazenou matici $\mathbf{V_{smooth}} \in \{0,1\}^{N\times 2}$ je možné považovat za výstup systému. Na rozdíl od diarizace prezenčního sezení může takto získaný výstup obsahovat úseky, ve kterých je detekována souběžná mluva obou řečníků.

Prezenční sezení

Jak již bylo popsáno v podsekci 3.2.2 nahrávky prezenčních sezení obsahují značný přeslech mezi kanály. Z tohoto důvodu není možná jednoduchá diarizace pomocí výstupu řečové aktivity příslušných kanálů a po zvážení všech okolností vzniku nahrávky jsem navrhl čtyři

metody, pomocí kterých je možno rozhodnout, který z řečníků aktuálně mluví. Příslušné metody jsou vyhodnoceny v kapitole 6.

Prvním krokem použitých diarizačních metod je rozhodnutí o řečové aktivitě příslušného segmentu. Nad kanály nahrávky je spuštěna detekce řečové aktivity. Tím je získána matice $\mathbf{V} \in \{0,1\}^{N\times 2}$ booleovských hodnot aktivity segmentů příslušných kanálů. Následně je provedena logická disjunkce (or) segmentů mezi kanály. Takto je získán vektor $\mathbf{u} \in \{0,1\}^N$ booleovských hodnot označující přítomnost řečové aktivity alespoň jednoho z mluvčích daného segmentu.

$$u_n = V_n^1 \vee V_n^2, n \in \{1,\ldots,N\} \tag{4.1}$$

Zbývá rozhodnout, kdo je původcem mluvy. Příslušné metody se v tomto aspektu liší a jejich funkcionalita je popsána v následujících bodech. Všechny metody však na výstupu vracejí vektor $\mathbf{d} \in \{0,1,2\}^N$. Příslušné hodnoty znamenají:

- **0** – ticho,
- **1** – mluví řečník reprezentující kanál 1 (terapeut),
- **2** – mluví řečník reprezentující kanál 2 (klient).

Vektor **d** je po skončení diarizace možné vyhladit odstraněním krátkodobých změn řečníků. *Kanál 1* u všech typů diarizace (prezenční, online) odpovídá řečové nahrávce mluvy terapeuta. V případě, že jsou kanály prohozeny, výstup systému není možné považovat za validní.

1. **Energetická diarizace** – Vstupem je matice energií segmentů $\mathbf{E} \in \mathbb{R}^{N\times 2}$ a vektor řečově aktivních segmentů $\mathbf{u} \in \{0,1\}^N$.

   Výstupní vektor $\mathbf{d} \in \{0,1,2\}^N$ je vytvořen následujícím způsobem podle hodnoty segmentu v poli **u** a matici **E**.

$$d_n = \begin{cases} 0 & u_n = 0 \\ 1 & E_n^1 > E_n^2 \\ 2 & jinak \end{cases}, n \in \{1,\ldots,N\} \tag{4.2}$$

2. **Mel-frekvenční diarizace s využitím jednoho kanálu** – Obdobně jako u předchozí metody je vstupem matice energií segmentů $\mathbf{E} \in \mathbb{R}^{N\times 2}$, vektor řečově aktivních segmentů $\mathbf{u} \in \{0,1\}^N$ a navíc tenzor Mel-frekvenčních cepstrálních koeficientů $\mathbf{M} \in \mathbb{R}^{N\times K\times 2}$, kde K je počet koeficientů segmentu určený nastavitelným parametrem v konfiguračním souboru.

   Z vstupního tenzoru **M** jsou extrahovány segmenty, ve kterých se vyskytuje řečová aktivita jednoho z kanálu. Dimenzionalita nově vzniklého tenzoru **F** je rovna $A \times K$, kde $A = \sum_{n=1}^{N} u_n$.

   S takto získaným tenzorem **F** je natrénována směsice Gaussovských rozložení (GMM) blíže popsaná v sekci 2.4.1. Počet komponent C, konvergenční práh a maximální počet iterací je opět nastavitelný v konfiguračním souboru. Jelikož v této úloze nelze

předpokládat korelaci mezi Mel-frekvenčními cepstrálními koeficienty, je využita diagonální kovarianční matice pro optimalizační účely. Následně je proveden výpočet rozdílů energií segmentů příslušných kanálů, čímž vzniká vektor $\mathbf{e}_d$.

$$\mathbf{e}_d = \mathbf{E}^1 - \mathbf{E}^2 \tag{4.3}$$

Z natrénovaného univerzálního modelu $\mathtt{G}$ jsou vytvořeny dvě kopie – $\mathtt{G}^1$ reprezentující terapeuta a $\mathtt{G}^2$ reprezentující klienta. Z tenzoru $\mathbf{F}$ je extrahováno X % nejkladnějších segmentů vektoru $\mathbf{e}_d$, čímž vzniká tenzor $\mathbf{S} \in \mathbb{R}^{A/5 \times K \times 2}$ a X % nejzápornějších segmentů vektoru $\mathbf{e}_d$, které vedou k vzniku tenzoru $\mathbf{T} \in \mathbb{R}^{A/5 \times K \times 2}$. X označuje zvolený percentil, který by neměl překročit hodnotu 50 %, v této práci byly dosaženy nejlepší výsledky s $X \in \langle 0; 10 \rangle$. Maximální hodnota percentilu je stanovena na 10 %, neboť poměr řeči hovořících může být v daných sezeních značně nevyrovnaný a mohlo by tak docházet k velké chybovosti.

Nad nově vzniklým tenzorem $\mathbf{S}$ je proveden jeden krok EM algoritmu vycházející z modelu $\mathtt{G}$. Z nové matice věrohodnosti příslušných segmentů $\mathbf{\Gamma}$ vůči příslušným komponentám jsou vypočteny nové střední hodnoty $\boldsymbol{\mu}$. Ty jsou normalizovány sumou věrohodností příslušných komponent $\boldsymbol{\gamma'} \in \mathbb{R}^C$ modelu $\mathtt{G}$, čímž vzniká nový vektor $\boldsymbol{\mu'} \in \mathbb{R}^C$.

$$\gamma'_c = \sum_{a=1}^{A} \Gamma_{ac}, c \in \{1, \ldots, C\} \tag{4.4}$$

$$\boldsymbol{\mu'} = \boldsymbol{\mu} \times \boldsymbol{\gamma'}^{-1} \tag{4.5}$$

Nové váhy komponent $\mathbf{w} \in \mathbb{R}^C$ jsou vypočteny následovně:

$$\mathbf{w} = \boldsymbol{\gamma'} \times \frac{1}{A}. \tag{4.6}$$

Z nových vah $\mathbf{w}$ je vypočten vektor posunu $\mathbf{p}$ podle vztahu:

$$\mathbf{p} = P \times \mathbf{w}, \tag{4.7}$$

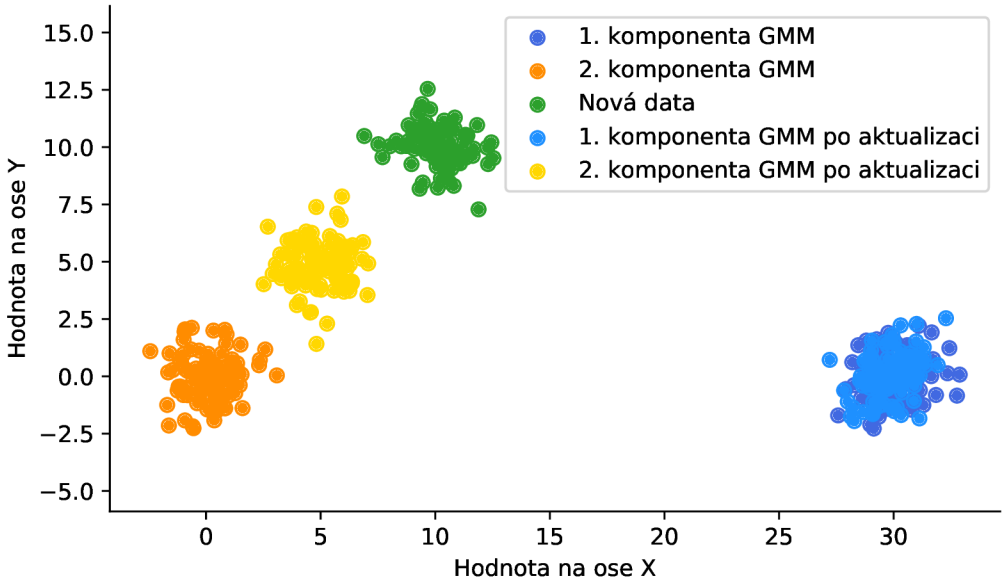
kde P je parametr posunu, který je vhodné nastavit v rozmezí 0-5 %.

Nové střední hodnoty $\boldsymbol{\mu}^1$ modelu $\mathtt{G}^1$ jsou vypočteny následovně:

$$\boldsymbol{\mu}^1 = (1 - \mathbf{p}) \times \boldsymbol{\mu} + \mathbf{p} \times \boldsymbol{\mu'}. \tag{4.8}$$

Nové střední hodnoty $\boldsymbol{\mu}^2$ modelu $\mathtt{G}^2$ jsou vypočteny totožným způsobem nad maticí $\mathbf{T}$. Obr. 4.2 demonstruje posun středních hodnot GMM modelující 2-dimenzionální data.

Následně je vypočtena věrohodnost MFCC koeficientů příslušného segmentu tenzoru $\mathbf{M}$ vůči modelu $\mathtt{G}^1$ a $\mathtt{G}^2$, čímž vznikají vektor $\mathbf{p}^1 \in \mathbb{R}^N$ a $\mathbf{p}^2 \in \mathbb{R}^N$. Jelikož je velmi nepravděpodobná změna mluvčího o délce několika milisekund uprostřed souvislé



Obrázek 4.2: Aktualizace středních hodnot 2-rozměrné směsice Gaussovských komponent s parametrem posunu $P = 0,5$.

několikasekundové promluvy druhého mluvčího, je vhodné nad těmito vektory provést forward-backward algoritmus, pomocí kterého jsou vyhlazeny krátkodobé změny mluvčích.

Následně je vypočten vektor rozdílu $\mathbf{r}$ věrohodností segmentů $\mathbf{p}^1$ a $\mathbf{p}^2$.

$$\mathbf{r} = \mathbf{p}^1 - \mathbf{p}^2 \tag{4.9}$$

Výsledný vektor $\mathbf{d}$ je vytvořen podle následujícího vztahu:

$$d_n = \begin{cases} 0 & u_n = 0 \\ 1 & r_n > 0 \\ 2 & jinak \end{cases}, n \in \{1, \ldots, N\} \tag{4.10}$$

3. **Mel-frekvenční diarizace s využitím obou kanálů** – Tato metoda funguje téměř identicky s tím rozdílem, že namísto vytvoření tenzoru $\mathbf{F}$ obsahující MFCC aktivních segmentů kanálu 1, je tenzor $\mathbf{F}$ vytvořen spojením řečově aktivních MFCC koeficientů kanálu 1 a 2. Dimenzionalita nově vzniklého tenzoru $\mathbf{F}$ je rovna $A \times 2K$. Zbytek metody je již totožný.

4. **Mel-frekvenční diarizace ve dvou iteracích s využitím obou kanálů** – Následující metoda se značně neliší oproti metodě popsané výše. K rozdílnému chování dochází po vytvoření vektorů $\mathbf{p}^1$ a $\mathbf{p}^2$. Pomocí těchto vektorů je vytvořen nový tenzor $\mathbf{S}'$ a $\mathbf{T}'$. Tenzor $\mathbf{S}'$ obsahuje 20 % nejvěrohodnějších segmentů vektoru $\mathbf{p}^1$ tenzoru $\mathbf{F}^1$ a obdobně tenzor $\mathbf{T}'$ obsahuje 20 % nejvěrohodnějších segmentů vektoru $\mathbf{p}^2$ tenzoru $\mathbf{F}^2$. Následně dochází k dalšímu posunutí středních hodnot směsic blíže ke středním

hodnotám MFCC příznaků extrahovaných tenzorů. Takto by bylo možné vykonat i vícero iterací, dosažené výsledky však neprokazovaly po více iteracích vyšší přesnost.

V rámci obou navržených podsystémů je po skončení diarizace energie signálu uložena do výstupního souboru s příponou `.energy` a výstup diarizace do souboru s příponou `.rttm`, jehož formát je detailněji popsán v předchozí kapitole v sekci 3.3.

4.2.3 Přepis

Automatický přepis nahrávek je možný až po provedení diarizace a je tak přímo závislý na předchozím bloku. Aby nebylo nutné znovu spouštět diarizaci, soubory s příponou `.rttm` jsou postupně načteny a následně je vytvořen přepis pro řečově aktivní úseky nahrávky. Ten je uložen do souboru s příponou `.txt`. Jelikož systém neobsahuje vlastní řešení rozpoznávání řeči a jsou využity externí řešení, byla tato část systému navržena za účelem co možná nejjednodušší záměny ASR modulu v budoucnosti. Pro nahrávky datasetu CallHome, ve kterých mluvčí hovoří anglickým jazykem, jsou k dispozici moduly CMUSphinx a Google Cloud Api. Nahrávky projektu Deepsy jsou velmi citlivé a jejich přepis je proveden interním ASR systémem pro český jazyk skupiny Speech@fit[2] [18].

Ze vstupní nahrávky jsou postupně extrahovány úseky odpovídající řečové aktivitě příslušného kanálu. Nad každým úsekem je spuštěno rozpoznání řeči a výsledný text je uložen do výstupního souboru.

4.2.4 Získání statistik

S dostupnými soubory `.rttm`, `.txt` a `.energy` je již možné provést kalkulaci statistik příslušné nahrávky. Sledované statistiky jsou detailně popsány v následující sekci 4.3. Tyto hodnoty jsou uloženy pomocí vysokoúrovňového binárního serializačního protokolu pickle [27] do souboru s příponou `.pkl`, aby je bylo možné načíst bez nutnosti manuální serializace.

Z existujících souborů obsahujících statistiky s příponou `.pkl` o daných sezeních je následně možné vytvořit pomocí dalšího funkcionálního bloku soubor obsahující průměrnou hodnotu sledovaných statistik a jejich varianci mezi nahrávkami. Aby bylo možné do souboru nahlédnout a tyto statistiky interpretovat, jsou hodnoty uloženy ve formátu JSON [9] s příponou `.json`.

4.2.5 Generování souhrnné zprávy

S využitím souboru obsahujícího textace s příponou `.json` a HTML šablonou s příponou `.html` (název souboru popisujícího kaskádové styly šablony by měl být pojmenován totožně jako HTML šablona a měl by být ukončen příponou `.css`) je možné ze získaných statistik vytvořit souhrnnou zprávu s příponou `.html` popsanou v následující sekci 4.3. Všechny výše zmíněné výstupní soubory jsou pojmenovávány podle identifikátoru vstupní nahrávky. Jelikož souhrnná zpráva obsahuje vysoký počet grafů, jsou vygenerované obrázky uloženy v externí složce vytvořené až za běhu, která je určená uživatelem.

[2] https://speech.fit.vutbr.cz/

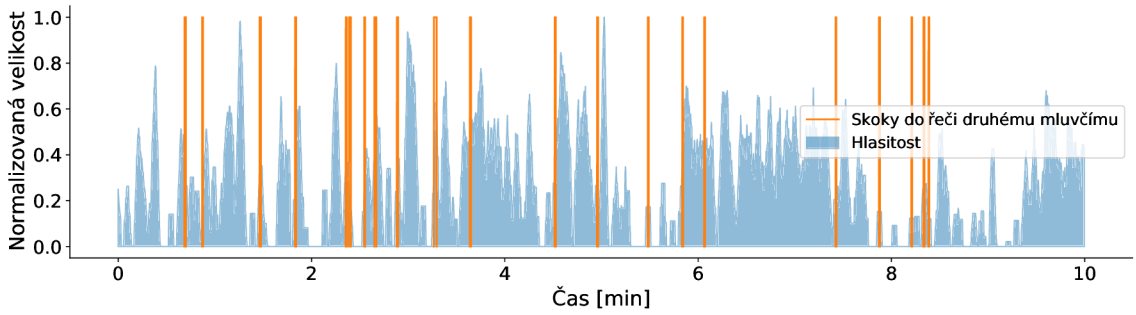
4.3 Dokument shrnující proběhlé sezení

Ač to není na první pohled patrné, audio nahrávky obsahují enormní množství informací. Cílem této sekce je vytvořit souhrnnou zprávu obsahující klasifikátory, které jsou pro účely psychoterapie nejzajímavější. V souhrnné zprávě jsou tyto klasifikátory představeny v podobě tabulek a grafů tak, aby bylo na první pohled patrné, co příslušné hodnoty představují, jak se sledovaná hodnota mění v průběhu sezení nebo jak se hodnota liší mezi sezeními. Systém všechny grafy automaticky generuje do předem určené složky podle daného identifikátoru nahrávky, texty obsahující další informace o sezení jsou taktéž automaticky generované za běhu. Aby nebylo nutné při každém spuštění pracně generovat HTML dokument a aby jeho editace byla co nejpříjemnější, je vytvořena šablona zprávy. Definuje vzhled a obsah souhrnné zprávy. V rámci tvorby výstupu pro příslušnou nahrávku jsou do kopie šablony dogenerovány nové grafy a doplněny hodnoty spolu s příslušnými texty. Předdefinovaná šablona taktéž dovoluje značně jednodušší úpravu pro konkretní účely. Její strukturu a vzhled je možné měnit bez zásahů do zdrojových kódů. Je však nutné dodržet pojmenování elementů ve zdrojových kódech i v šabloně, aby byla na správné místo dosazena správná hodnota. Tento dokument může terapeutovi pomoct odhalit souvislosti, které nemusí být na první pohled patrné. Náhled souhrnné zprávy se nachází v příloze B.

Zkoumáním zvukové stránky verbální komunikace se zabývá věda známá jako Paralingvistika [20]. Vlastnosti řeči mohou být v průběhu rozhovoru pozměňovány vědomě i nevědomě. V následujících podsekcích jsou představeny paralingvistické charakteristiky řeči.

4.3.1 Hlasitost

Hlas netlumočí jen obsah sdělení, nýbrž vypovídá i o psychickém rozpoložení hovořící postavy. Tento psychický stav se odráží především v tónu řeči a hlasitosti slovního projevu. U pacientů trpících bipolární poruchou [26] lze snadno rozlišit, v jakém aktuálním stavu se pacient nachází – smutek, skleslost během depresivní polohy či naopak živost a vzrušení při fázi manické. Obdobně lze v mírnějších podobách tyto aspekty zaznamenat i u osob bez vážnějších psychických potíží [20]. Terapeut se během terapie zaměřuje na mnoho věcí a změny hlasitosti mu mohou v průběhu sezení uniknout. Cílem je poskytnout terapeutovi tuto informaci strojově. Je k tomu využita energie signálu (podsekce 2.3.1), která je průměrována mezi N sousedními segmenty. Výstup, který je terapeutovi prezentován, je zobrazen na obr. 4.3.



Obrázek 4.3: Měnící se hlasitost v průběhu sezení. Hlasitost úseků, ve kterých se nenachází řečová aktivita je rovná nule. Nejvyšší hodnoty se nacházejí mezi 4–6 minutou sezení.

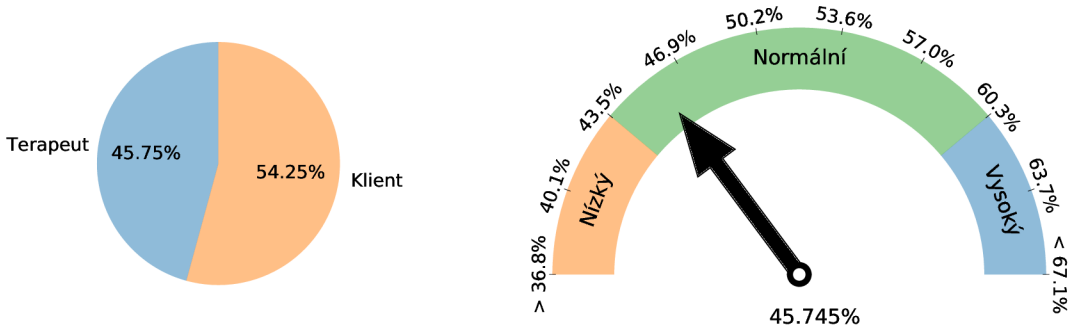
4.3.2 Poměr řeči

Mezi další paralingvistické charakteristiky řeči náleží poměr řeči. Prozrazuje nám, kolik toho daný mluvčí řekl v rámci rozhovoru. Poměr řeči můžeme klasifikovat do tří kategorií charakterizujících dané sezení.

- Terapeut vede sezení a klient stručně odpovídá.
- Poměr řeči mezi terapeutem a klientem je vyrovnaný.
- Klient vypráví o nastalé situaci a terapeut naslouchá a klade krátké otázky, které vedou klienta k rozvinutí dané myšlenky.

Terapeutovi je poskytnut poměr řeči mezi ním a klientem a poměr řeči a ticha obou účastníků sezení v podobě kruhových grafů [38]. Je mu taktéž poskytnuto porovnání s ostatními sezeními v podobě „rychloměru“ [14], aby byl na první pohled patrný vztah mezi sledovanými veličinami. Obr. 4.4 znázorňuje demonstrační výstup.

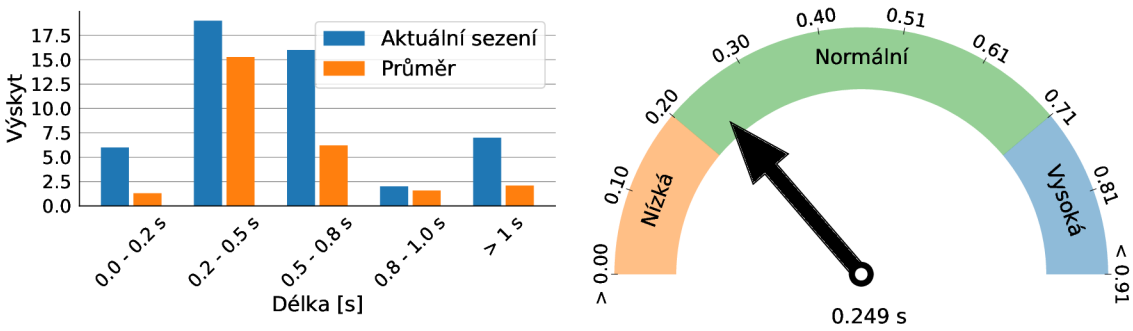
Dále můžeme zkoumat, jaký informační objem posluchač doopravdy získal z přednesu mluvčího. Tuto informaci lze orientačně získat pomocí odstranění spojek, předložek, částic a obdobných slov z přepisu nahrávky a vydělení délkou příslušných úseků mluvy. V průběhu psychoterapeutického sezení tyto hodnoty mohou značně kolísat a nepřinášejí terapeutovi až tak zásadní zjištění, a proto nejsou zahrnuty do souhrnného výstupu.



Obrázek 4.4: Poměr doby mluvy mezi terapeutem a klientem. Hodnoty se velmi neliší, oba řečníci mluvili zhruba stejně dlouho. Doba mluvy terapeuta odpovídá dlouhodobému průměru a je k němu přirovnána na obrázku vpravo.

4.3.3 Skákání do řeči

Mezi další charakteristiky rozhovoru patří skákání do řeči. U malých dětí je tento jev spojen s hyperaktivitou, ve skupině vícero mluvčích se jedná o aspekt pohrdání nebo rozčílení se z řečového projevu druhé osoby. Avšak v kontextu psychoterapeutických sezení tento pojem nabývá úplně jiných rozměrů. Psychoterapeuti dokážou skoky do řeči kontrolovat průběh terapie a vést ji správným směrem. Nadměrný výskyt skoků do řeči však může vést k nespokojenosti jednoho z mluvčích. V souhrnné zprávě je proto zahrnut histogram [22] skoků do řeči porovnávající konkrétní sezení se statistikami získanými mezi sezeními. Histogram atypického sezení je zobrazen na obr. 4.5.

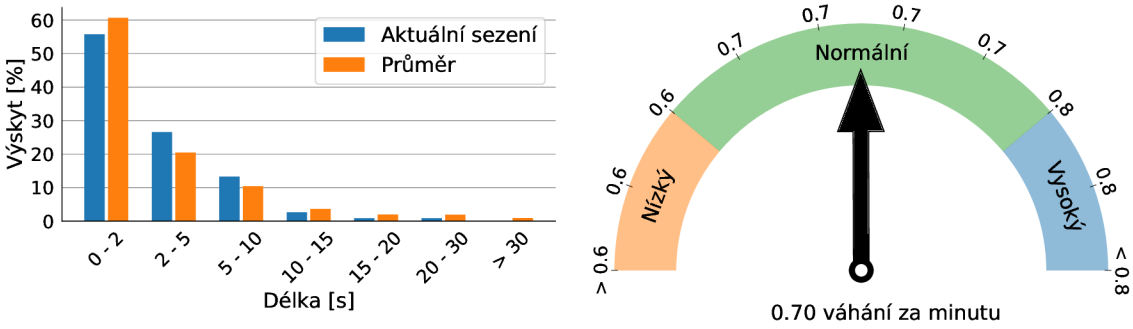


(a) (b)

Obrázek 4.5: **(a)** Počty skoků do řeči zobrazené v podobě histogramu. Hodnoty odpovídají dlouhodobému průměru. **(b)** Reakční doba je poměrně nízká, blíží se mimo hranici průměrných hodnot.

4.3.4 Váhání a délka souvislých úseků řeči

Pomlky, pauzy v řeči, zmlknutí a podobné projevy narušují plynulost řeči. Větné pomlky (mezi jednotlivými větami – trvají 0,5 až 1 sekundu) jsou potřebné, řeč se tak nestává lavinou či směsicí slov a je srozumitelná. Pomlky v jiných místech jsou tzv. „pomlky váhání“. Delší pomlky naznačují, že mluvčí více váhá a není si úplně jist svým sdělením. Mluvčí hovoří nejen tehdy, když mají jasno v tom, co chtějí sdělit, ale i tehdy, když plánují, o čem budou v další fázi hovořit. Běžný posluchač dokáže zaregistrovat náznak zaváhání již při délce pomlky 200 milisekund [20]. Větné pomlky (na rozdíl od pomlk váhání) souvisí i s pohyby účastníků rozhovoru (např. pokývnutí hlavou, přitakání) [8]. Fáze ticha mohou naznačovat nejistotu, rozpaky a váhání, naopak rychlý proud řeči zaujetí a citové vzrušení. Délky souvislých úseků mluvy mohou taktéž ledacos nastínit o průběhu sezení. Počet váhání a délka souvislých úseků je terapeutovi poskytnuta v podobě znázorněné na obr. 4.6.



(a) (b)

Obrázek 4.6: **(a)** Délka souvislých úseků řeči. V nahrávce se nejčastěji objevují krátké věty o délce 0–2 sekundy. Nebyla detekována žádná věta delší než 30 sekund. **(b)** Počet váhání za minutu v mluvě terapeuta. Hodnota 0,7 váhání za minutu přesně odpovídá měřenému průměru. Je však stále poměrně nízká v porovnání s průměrem váhání na straně klienta.

4.3.5 Reakční doba

Podobně jako úseky váhání můžeme taktéž klasifikovat podle jejich délky úseky čekání na reakci. Tyto hodnoty se mohou pohybovat od desítek milisekund až po několik sekund. Aby bylo možné tuto hodnotu co nejrychleji interpretovat, je terapeutovi poskytnut graf v podobě „rychloměru“ poskytující porovnání mezi průměrnou hodnotou reakční doby daného sezení a průměrnou hodnotou napříč všemi sezeními v databázi, případně pouze všemi sezeními daného terapeuta. Tento graf je znázorněný na obr. 4.5. V ideálním případě by se reakční doby měly pohybovat kolem dlouhodobého průměru, avšak určité druhy terapie, kdy terapeut pokládá složité otázky, vyžadují delší čas na konstrukci odpovědi a dochází tak k výraznému odklonění od dlouhodobého průměru.

4.3.6 Nejčastěji používané výrazy

Nejčastěji používané výrazy extrahované z přepisu nahrávky terapeutovi přinášejí určitý souhrn o daném sezení. Reprezentace těchto dat v podobě histogramu nebo v textové podobě může přinést potřebný náhled do probírané problematiky, avšak technika známá jako „Slovní mrak“ (Word Clouds) [37] povznáší tento náhled na ještě vyšší úroveň. Obr. 4.7 znázorňuje demonstrační mrak získaný z nahrávky sezení. Terapeut díky tomuto grafu taktéž dokáže zdokonalit svoje vyjadřovací schopnosti omezením nadužívání některých výrazů, které se mohou opakovat v mracích více sezení.

Obrázek 4.7: Slovní mrak vytvořený z nahrávky psychoterapeutického sezení. Je možné sledovat, že mluvčí velmi často používal výplňová slova jako např. „vlastně“, „prostě“ nebo „jako“.

4.3.7 Mimoslovní složky hlasového projevu

O rozpoložení hovořící osoby něco vypovídají i různé těžko popsatelné zvuky při slovním projevu. Mezi nejčastěji používané výplně mluvy patří zvuky jako „ehm ehm...“, „hmmm...“, „éééé...“. Tyto zvuky jsou ve většině případů využívány řečníky pro ujištění posluchačů, že promluva bude pokračovat a nemají ihned reagovat [20]. Tyto mimoslovní složky jsou v oficiálním přednesu považovány za chyby a mohou být vnímány negativně. Jsou nejčastěji vyvolány vnitřní tísní hovořícího nebo vztahem daným k osobě naslouchající (před někým se cítíme uvolněně, před jiným napjatě). Snižující se trend výskytu mimoslovních složek hlasového projevu klienta může reprezentovat zvyšující se důvěru mezi klientem a terapeutem, což může vést k lepším výsledkům terapie. Podobně jako reakční doba je tato hodnota zobrazena pomocí „rychloměru“ ukazujícího poměr mezi statistikou aktuálního a ostatních sezení.

4.3.8 Rychlost řeči

Dalším z charakteristických ukazatelů je rychlost řeči neboli tempo řeči. Vyjadřuje se nejčastěji počtem vyslovených slabik za určitou časovou jednotku. Pro češtinu se jako průměrné tempo udává hodnota 5 nebo 6 slabik za sekundu, rychlost řeči je však velmi proměnlivá mimo jiné v závislosti na konkrétním mluvčím, situaci, výstavbě výpovědi, stylu a funkci projevu. Průměrná hodnota charakteristická pro konkrétního člověka se nazývá osobní mluvní tempo.

Pokud mluvčí zvolí příliš vysoké tempo řeči (např. nad 10 slabik za sekundu), dochází obvykle k celkové deformaci projevu, protože artikulační orgány nezvládají tak rychle zaujímat patřičné pozice, chybí náležité frázování a odpovídající větná melodie. Posluchač vystavený takovéto překotné mluvě bude mít problémy s porozuměním. Avšak rovněž přehnaně pomalé tempo (pod dvě slabiky za sekundu) je pro vnímání obtížné. Důležité pro kultivovaný mluvený projev je také umět s rychlostí řeči funkčně pracovat, tj. nemluvit stále stejně rychle, ale přizpůsobovat tempo řeči jejímu obsahu [39].

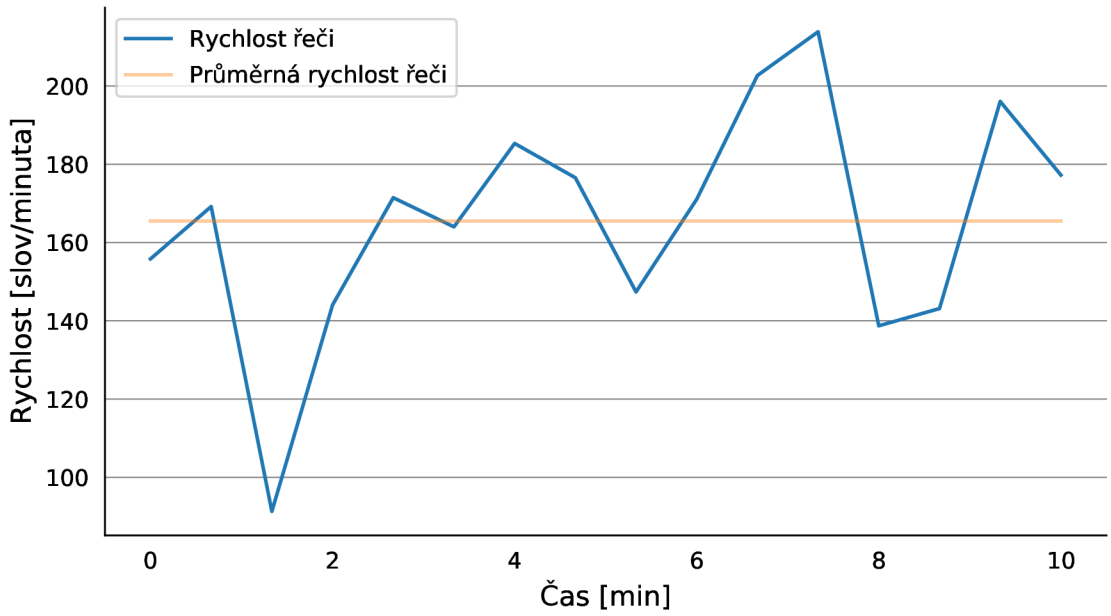
Terapeutovi je představena tato statistika v podobě liniového grafu obsahující rychlost N stejně dlouhých úseků řeči dané nahrávky. Vektor příslušných hodnot rychlostí úseků $\mathbf{s}$ je vypočten podle následujícího vztahu:

$$s_n = \frac{1}{N} \sum_{i=1}^{K} \frac{T_i^1}{T_i^2}, n \in \{1, \ldots, N\}, \tag{4.11}$$

kde $\mathbf{T} \in \mathbb{R}^{K \times 2}$ je vektor úseků obsahující pole segmentů řečové aktivity daného úseku n nahrávky obsahující hodnotu počtu slov T^1 a počtu minut daného segmentu T^2, K je variabilní velikost pole segmentů daného úseku. Průměrná hodnota rychlosti úseků by měla kolísat mezi 150–200 slovy za minutu [32]. Demonstrační ukázka je zobrazena na obr. 4.8.

4.3.9 Nálada v průběhu sezení

Mezi experimentální statistiky náleží nálada mluvčího v průběhu sezení. Ta je zkoumaná z přepisu terapie. Je dostupná pouze pro nahrávky v anglickém jazyce, jelikož se mi nepodařilo najít hotové řešení pro český jazyk a implementace takového systému by vydala na samostatnou bakalářskou nebo diplomovou práci. Odhad emocí funguje na základě detekce klíčových slov spojených z různým typem emocí v textovém přepisu. Tímto způsobem je získán poměr následujících emocí příslušných segmentů textového přepisu:



Obrázek 4.8: Měnící se rychlost řeči mluvy terapeuta v průběhu sezení. Průměrná hodnota leží v intervalu 150–200 slov za minutu. Vidíme však, že byla detekována výrazná odchylka mezi 1–2 minutou nahrávky. Jelikož se jedná o začátek sezení, je pravděpodobné, že klient mohl mít problém vysvětlit, co ho trápí a co by chtěl řešit.

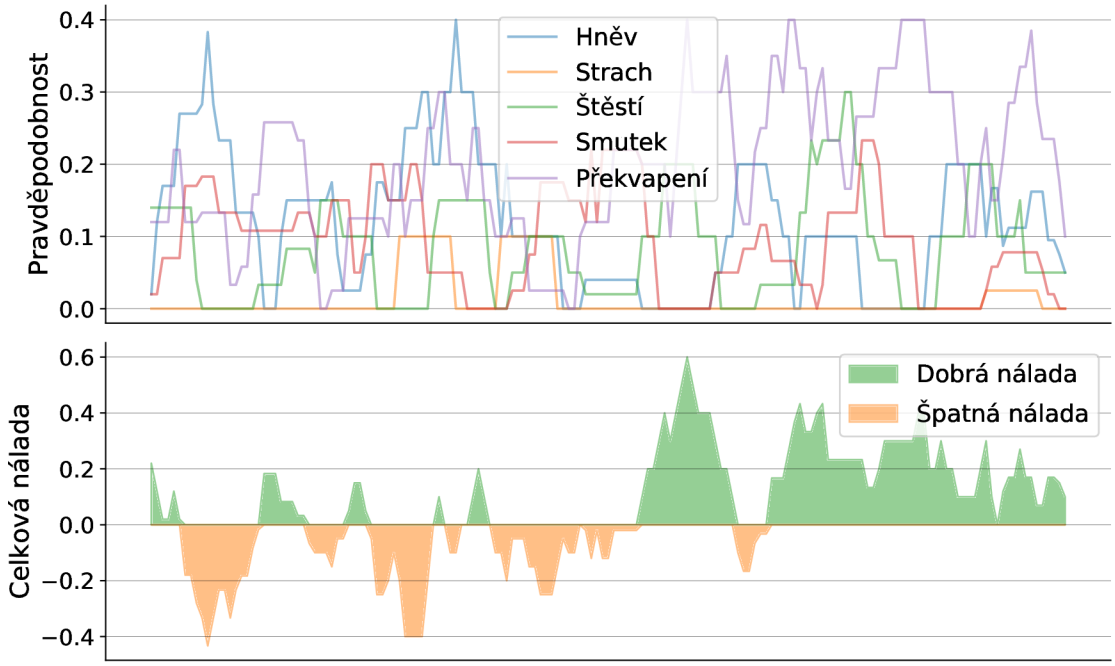
- štěstí,
- hněv,
- překvapení,
- smutek,
- strach.

Emoce je možné taktéž klasifikovat jako pozitivní a negativní, čímž je terapeutovi poskytnut přehled změn rozpoložení. Jelikož některé segmenty mají délku několika ms, je výstup systému detekce emocí vyhlazen průměrovým filtrem o velikosti 10 segmentů. Výstup systému pro náladu klienta je znázorněn na obr. 4.9.

4.3.10 Dlouhodobé statistiky

Aby bylo možné terapeutovi poskytnout informaci o tom, jak se příslušné sezení liší oproti jiným sezením (sezením terapeuta se stejným klientem, všem sezením daného terapeuta nebo všem dostupným sezením) jsou vypočteny průměrné hodnoty a jejich střední odchylky skupiny zvolených nahrávek. Postupně jsou vypočteny statistiky následujících veličin:

- poměr řeči terapeut:klient
- reakční doba
- počty skoků do řeči a jejich délka



Obrázek 4.9: Nálada klienta v průběhu sezení. Jsou zkoumány příslušné emoce a celková nálada. Lze pozorovat kladný trend. Klient z terapie pravděpodobně odchází s dobrou náladou.

- počet váhání za časovou jednotku
- procentuální výskyt úseků řeči o určité délce
- počet výplňových slov za minutu

Tyto statistiky je následně možné porovnat. Ku příkladu nepřiměřený počet skoků do řeči ze strany terapeuta může naznačovat, že sezení bylo vedeno až příliš konfrontačně.

4.4 Validace výsledků

Samotné získání statistik není příliš náchylné na chyby, ale předchází mu proces detekce řečové aktivity, diarizace a rozpoznání řeči, kde chybovost může být velmi vysoká. Z tohoto důvodu jsou výstupy částí systému validovány vůči referenčním anotacím, aby zvolené řešení bylo co nejúspěšnější. Dataset CallHome obsahuje anotační soubory, vůči kterým je systém možné rovnou validovat. Pro nahrávky projektu DeePsy byly vytvořeny anotace ručně, aby bylo možné ověřit správnou funkcionalitu systému na odlišném typu dat. Byl k tomu použit volně dostupný nástroj Transcriber[3].

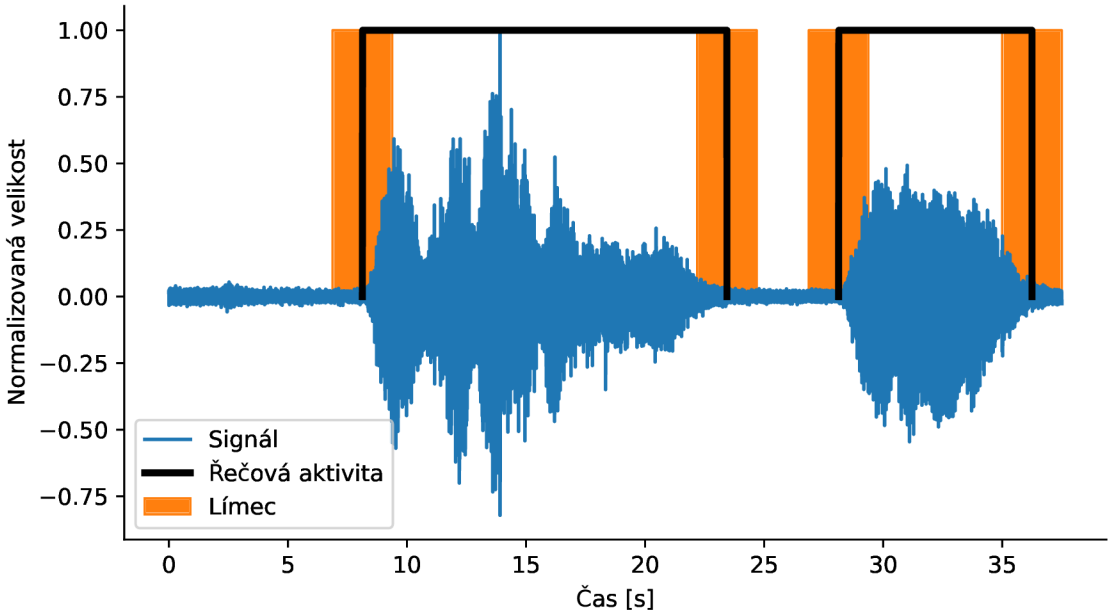
Nahrávky je možné rozdělit do dvou skupin – nahrávky s přeslechem (pořízené prezenčně) a nahrávky pořízené online. V obou případech jsou k dispozici skripty zjišťující

[3] http://trans.sourceforge.net/en/presentation.php

přesnost použitého systému. Chybovost systému diarizace DER (Diarization error rate) na nahrávkách pořízených prezenčně je definována následovně

$$\mathrm{DER} = \frac{M + F + C}{\mathrm{N}}, \tag{4.12}$$

kde N je počet segmentů příslušné nahrávky, M (Miss) označuje počet segmentů, ve kterých alespoň jeden z mluvčích mluvil, ale systém tento úsek klasifikoval jako neaktivní. F (False alarm) je počet segmentů, ve kterých byla detekována řečová aktivita, ale nikdo v nahrávce nemluvil, a konečně C (Confusion) je počet segmentů, ve kterých systém zvolil špatného řečníka. Situace, kdy jsou v referenční anotaci označeni oba mluvčí jako aktivní v daném úseku a systém detekuje pouze jednoho, není považována za chybu. Jelikož ruční anotace nejsou vždy zcela přesné, evaluační systém dovoluje použití „límce“ (collar). „Límec“ je úsek referenční anotace, kdy dochází ke změně hodnoty z ticha na mluvu a opačně. Segmenty nacházející se v prostoru „límce“ nejsou započítány do výsledné chyby diarizace. Standardní velikost „límce“ se pohybuje okolo 250 milisekund. Obr. 4.10 ukazuje jakým způsobem probíhá vyhodnocení DER na demonstrační nahrávce.



Obrázek 4.10: „Límec“ o velikosti 250 milisekund. Segmenty nacházející se v oblasti „límce“ jsou podbarveny a nejsou vyhodnocovány jako chybné při validaci diarizace.

Online nahrávky neobsahují žádný přeslech a mluvčí jsou od sebe odděleni v příslušných kanálech, což dovoluje definovat chybovost diarizace o něco striktněji. Chybu diarizace online nahrávky můžeme vypočítat jako průměr chyby detekce řečové aktivity příslušných kanálů nahrávky a definujeme ji jako

$$\mathrm{DER} = \frac{M^1 + F^1 + M^2 + F^2}{2 \cdot N}, \tag{4.13}$$

kde M^1, M^2, F^1 a F^2 označují chyby příslušných kanálů nahrávky. Diarizační chyby komerčně využívaných softwarů nepřekračují hranici 10 %.

Aby bylo možné porovnat přesnost rozpoznání řeči a tedy jeho strojového přepisu, je chyba takového systému WER (Word error rate) definována jako

$$\mathrm{WER} = \frac{S + I + D}{N}, \tag{4.14}$$

kde S (Substitutions) označuje počet záměn slov mezi referenční transkripci a výstupem systému, I (Insertions) počet slov, které jsou ve výstupu systému, ale nevyskytují se v referenčním výstupu, D (Deletions) počet slov, které jsou v referenční anotaci, ale nejsou zahrnuty v přepisu získaném systémem a N počet slov referenčního přepisu. Nevýhodou této metody validace je, že každé slovo má stejnou váhu, což je ale pro účely hrubého odhadu přesnosti systému v rámci této práce postačující.

Kapitola 5

Implementace

V této kapitole jsou představeny nástroje a knihovny použité při implementaci systému pro analýzu rozhovoru dvou osob. Sekce 5.1 zdůvodňuje použití právě jazyka Python[1] a jeho knihoven. V sekci 5.2 jsou popsány nejdůležitější části systému.

5.1 Použité nastroje

Pro implementaci systému popsaného v kapitole 4 jsem využil skriptovací jazyk Python ve verzi `3.8.5`. Jazyk Python se v posledních letech řadí mezi vůbec nejpoužívanější programovací jazyky. S tím je spojeno množství dostupných knihoven, návodů a dostupných řešení opakujících se problémů. Ačkoliv se jedná o skriptovací jazyk, Python plně podporuje objektově orientované a procedurální paradigma. Navíc podporuje některé aspekty funkcionálního a aspektově orientované programování. Výše zmíněné výhody a další[2] vedou k tomu, že je programování v Pythonu velmi komfortní a řešení problémů zabírá zlomek času, jenž by byl nutný k vyřešení stejného problému v jiném jazyce. Jistou alternativou pro účely tohoto systému může být implementace v jazyce Matlab, který je vhodný pro operace nad velkými tenzory. Osobně jsem se ale rozhodl tento jazyk nevyužít a dát přednost Pythonu a knihovnám, které tuto funkcionalitu přinášejí. Python je jazyk interpretovaný a výchozí interpret CPython nemusí být postačující u výpočetně náročných aplikací. Možnou alternativou je programovací jazyk a překladač Cython, který se snaží dosáhnout vyššího výkonu překladem do nativního kódu a obohacením jazyka o typový systém.

K implementaci problémů popsaných v kapitole 4 však využití jazyku Cython nebylo nutné, jelikož existuje knihovna NumPy[3], která podporuje více dimezionální pole – tenzory a operace nad nimi. Hodnoty v polích numpy obsahují datový typ a tím dochází k značné úspoře alokované paměti a vede to k vyšší rychlosti operací nad daty a je tak možné se vyvarovat velmi neefektivním smyčkám. Knihovna NumPy obsahuje nespočet matematických či statistických funkcí a operací nad tenzory [12]. Je základem mnoha dalších knihoven pro

[1] https://www.python.org/

[2] https://medium.com/@mindfiresolutions.usa/python-7-important-reasons-why-you-should-use-python-5801a98a0d0b

[3] https://numpy.org/

analýzu dat – Pandas[4], statistiku – SciPy[5], strojové učení – scikit-learn[6], nebo vizulaizaci – Matplotlib[7]. Je využita v co nejvyšší míře při operacích nad daty vstupní nahrávky.

Právě knihovna Matplotlib je použita pro vizualizaci všech grafů analýzy sezení. Je to poměrně nízkoúrovňová knihovna, která však s trochou praxe dovoluje vizualizovat velmi obsáhlé grafy. Opět existuje vysoký počet knihoven, které vycházejí z Matplotlibu. Za zmínku stojí vysoko-úrovňová vizualizační knihovna seaborn[8], která byla v rámci vývoje používána, ale v průběhu byla nahrazena právě knihovnou Matplotlib. Pro účely vizualizace nejpoužívanějších slov je navíc využita knihovna word_cloud[9].

Trénování směsic Gaussovských rozložení usnadňuje již zmíněnaá knihovna scikit-learn a práci s audio nahrávkami knihovna SciPy.

Při tvorbě výsledné zprávy je využita knihovna Beautiful Soup[10], která dovoluje rychlé a jednoduché parsování dokumentů rodiny XML.

Pro účely získání statistik je využita knihovna pro rozpoznání řeči SpeechRecognition[11] a knihovna pro detekci emocí text2emotion[12].

Postup při zpracování nahrávek je uživateli znázorněn pomocí knihovny progress[13].

Rovněž jsou využity standardní knihovny jazyka Python jako re, json, pickle, os nebo argparse. Hojně využívány jsou f-řetězce [29], které jsou dostupné v jazyce Python od verze 3.6 a generátorová notace polí [30], která dovoluje vytváření polí bez nutnosti pracného „appendování“.

5.2 Systém pro analýzu nahrávek

Navržený systém pro analýzu nahrávek sezení se skládá ze skupiny skriptů. V kořenové složce **src** se nacházejí spustitelné skripty odpovídající funkcionalitě bloků popsaných v kapitole 4. Bloky postupně načítají vstupní soubory, provádějí výpočty a ukládají výstupy, případně tisknou textové řetězce na standardní výstup. Následující pseudokód demonstruje standardní průchod vstupních dat jedním z bloků.

```
# Importování potřebných knihoven
import libs

# Načtení vstupních argumentů skriptu a jejich validace
args = load_arguments()
# Vytvoření výstupní stromové struktury
check_destination_directory_existence(args.dest)
# Načtení zdrojových souborů
source_files = load_files(args.src, args.extension)

# Postupné zpracování vstupních souborů
```

[4] https://pandas.pydata.org/
[5] https://www.scipy.org/
[6] https://scikit-learn.org/stable/
[7] https://matplotlib.org/
[8] https://seaborn.pydata.org/
[9] https://github.com/amueller/word_cloud
[10] https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[11] https://github.com/Uberi/speech_recognition
[12] https://github.com/aman2656/text2emotion-library
[13] https://github.com/verigak/progress/

```
for file in source_files:
    # Výpočet výstupních hodnot
    output = process_file(file)
    # Tisk dodatečných informací na standardní výstup
    print(output.text)
    # Uložení výstupních informací
    save_outputs(args.dest, file, output.data)

# Případný tisk stastik o~běhu skriptu a ukončení
print('Additional info')
```

Příslušné skripty čerpají z funkcí, které jsou dostupné v externích souborech. Ty jsou podle funkcionality rozděleny do příslušných složek. Následující blok demonstruje průběh diarizace nahrávky.

```
...
# Načtení vstupní nahrávky
wav_file, sampling_rate = read_wav_file(join(args.src, file))

# Preemfáze a segmentace pomocí Hammingova okna
signal = process_pre_emphasis(wav_file, params.pre_emphasis_coefficient)
segmented_tracks = process_hamming(signal, sampling_rate,
        params.window_size, params.window_overlap)

# Kalkulace energie segmentů
energy = calculate_energy(segmented_tracks)

# Extrakce Mel-frekvenčních koeficientů
_, mfcc, _ = calculate_mfcc(segmented_tracks, sampling_rate,
        params.cepstral_coef_count)

# Detekce řečové aktivity
vad = energy_gmm_based_vad_propagation(calculate_energy)

# Diarize pomocí směsice Gaussovských rozložení
diarization = gmm_mfcc_diarization(mfcc, vad, calculate_energy)
...
```

Systém obsahuje více funkcí, pomocí kterých je možné provést diarizaci. Ve většině případů funkce sdílejí vstupní a výstupní parametry, aby byla záměna co nejjednodušší.

Z řečově aktivních úseků je vytvořen přepis a poté jsou extrahovány všechny sledované statistiky blíže popsané v kapitole 2. Cílová tvorba souhrnné zprávy, ke které celý proces zpracování nahrávek spěje, probíhá následujícím způsobem.

```
...
for file in files:
    # Načtení přepisu a html šablony
    texts = load_texts(args.text)
    html = load_template(args.template)
```

```
    # Načtení statistik konkrétního a skupiny sezení
    stats = load_stats(file)
    stats_overall = load_stats_overall(args.stats)

    # Nahrazení textací a příloh cílového dokumentu
    add_texts(html, stats, stats_overall, texts, file)
    add_attachments(html, file)

    # Tvorba tabulek a grafů
    add_volume_table(html, stats)
    add_plots(html, stats, stats_overall, file)

    # Uložení editovaného html kódu do příslušného souboru
    with open(f'{join(args.path, file)}.html', 'w') as output:
        output.write(str(html))
...
```

Zajímavé úseky kódu

Jelikož zpracování poměrně dlouhých nahrávek je výpočetně náročné, byl pro účely testování rychlosti kódu implementován dekorátor [28] **@timeit**. Ten získává statistiky o volání příslušné funkce, což výrazně pomohlo při profilování kódu a vedlo k jeho následným optimalizacím.

```
import functools
import time

def timeit(func):
    """Dekorátor sloužící k~výpočtu doby běhu funkce"""

    # Tvorba nového dekorátu obalujícího funkci
    @functools.wraps(func)
    def new_func(*args, **kwargs):
        # Detekce času volání funkce
        start_time = time.time()
        # Provedení těla funkce
        ret_val = func(*args, **kwargs)
        # Délka běhu = konec -  start
        elapsed_time = time.time() - start_time
        # Tisk časových údajů na standardní výstup
        print(f'function [{func.__name__}] '
              f'finished in {int(elapsed_time * 1000)} ms')
        # Vrácení výsledků funkce
        return ret_val

return new_func

# Použití dekorátoru pro kalkulaci délky běhu funkce sum()
```

```
@timeit
def sum(a,b):
    return a + b
```

Většina kódu pracuje s daty reprezentovanými v podobě N-dimenzionálních polí (tenzorů) vytvořených v knihovně NumPy. Operace nad těmito poli a obecně vektorový přístup může být na první pohled nejasný a bylo nutné detailně prostudovat tuto problematiku. Knihovna NumPy má však skvělou dokumentaci a je tak možné ihned dohledat manuál příslušné funkce/operace. Následující kód demonstruje výpočet procentuálního výskytu úseků mluvy o příslušné délce pomocí knihovny NumPy.

```
def calculate_segments_len_distribution(bounds):
    """Výpočet histogramu délky úseků mluvy.
    Vstupem jsou hranice úseků v~podobě dvou dimenzionálního NumPy pole.
    bounds = [[0,10],[12,15]]
    označuje 2 úseky řeči mající délku 10 a 3 segmenty"""

    # Výpočet délky intervalů
    speech_bounds_len = (bounds[:, 1] - bounds[:, 0])

    # Inicializace intervalů distribuce
    bins = np.array([0, 2, 5, 10, 15, 20, 30, np.iinfo(np.int16).max])

    # Získání počtu úseků náležejících příslušným intervalům
    current_counts, _ = np.histogram(speech_bounds_len, bins=bins)

    # Normalizace hodnot do rozsahu 0 - 100 %
    current_counts = (current_counts / np.sum(current_counts)) * 100

    return current_counts
```

Další poměrně zajímavou pasáží je metoda **update_centers** třídy *MyGmm*, která dědí chování z třídy *GaussianMixture* knihovny scikit-learn. Tato metoda implementuje aktualizaci středních hodnot směsice Gaussovských rozložení o X % vzhledem k novým středům. Kód je podrobněji vysvětlen v části **Mel-frekvenční diarizace s využitím jednoho kanálu** sekce 4.2.2. Pro plné pochopení kódu je vhodné taktéž nahlédnout do sekce 2.4.1.

```
import numpy as np
from sklearn.mixture import GaussianMixture
from scipy.special import logsumexp


# Nová třída dědící z~třídy GaussianMixture
class MyGmm(GaussianMixture):

    def update_centers(self, data, shift):
        """Metoda pro aktualizaci středních hodnot GMM směrem
        k~novým středním hodnotám vypočtených ze vstupní proměnné data.
        Parametr shift určuje koeficient posunu k~novým středům."""
```

```
# Kalkulace věrohodnosti příslušných vzorků dat
weighted_log_gamma = self._estimate_weighted_log_prob(data)
log_evidence = logsumexp(weighted_log_gamma, axis=1)
gamma = np.exp(weighted_log_gamma - log_evidence[:, np.newaxis])
gamma_sum = gamma.sum(axis=0) + \
        10  * np.finfo(weighted_log_gamma.dtype).eps

# Výpočet nových středních hodnot a jejich vah
new_centers = np.dot(gamma.T, data)
new_centers_normalized = new_centers / gamma_sum[:, np.newaxis]
new_weights = gamma_sum / len(data)

# Aktualizace středních hodnot modelu
shift_center = shift * new_weights
shift_center = shift_center[:, np.newaxis]
self.means_ *= (1 - shift_center)
self.means_ += shift_center * new_centers_normalized
```

Kapitola 6

Experimenty

Detekce řečové aktivity, diarizace a automatické rozpoznání řeči jsou oblasti, ve kterých se mohou lišit i výstupy různých osob provádějící přepis nahrávek. Odhadnout, co přesně která osoba v daný moment říká, je velmi náročný úkol. Z tohoto důvodu je vcelku nemožné navrhnout systém, který by dosahoval přesnosti 100 %. V této kapitole jsou porovnány metody pro detekci řečové aktivity popsané v sekci 2.5, diarizační metody navržené v podsekci 4.2.2 a dostupné modely tvořící přepis představené v sekci 2.7.

Pro experimenty nad navrženými metodami byla dostupná data (dataset CallHome a data projektu DeePsy) rozdělena do tří skupin podle charakteru jejích původu. Metody byly postupně vyhodnoceny nad skupinou online, telefonních a prezenčních nahrávek pořízených diktafonem ZOOM H2n. Dataset Callhome obsahuje anotace, vůči kterým je možné validovat. Pro data projektu DeePsy bylo nutné vytvořit referenční anotace. Anotace online nahrávek byly vytvořené ASR systémem pro český jazyk skupiny Speech@fit, přesnost tohoto systému je velmi vysoká a tento výstup je možné považovat za referenční. Pro nahrávky pořízené diktafonem byly referenční anotace vytvořené manuálně pomocí nástroje Transcriber[1].

6.1 Detekce řečové aktivity

Detekce řečové aktivity je první a nejdůležitější částí navrženého systému. Provádět rozpoznávání řeči nad segmenty obsahujícími ticho je nežádoucí. Tento požadavek vede k nutnosti dosáhnout co nejvyšší přesnosti detekce. Implementované metody jsou v různých konfiguracích spuštěny nad dostupnými nahrávkami a je vyhodnocen počet falešných poplachů – „False alarm“ (segment obsahující ticho klasifikován jako řeč) F a segmentů obsahující řeč klasifikovaných jako ticho M – „Miss“.

Testování proběhlo s různými hodnotami parametrů a s použitím různých algoritmů pro vyhlazení výstupu. Následující tabulky ukazují hodnoty přesnosti příslušných běhů. Metody, jejichž výstup se může lišit podle pseudonáhodnosti inicializace počátečního stavu, jsou vyhodnoceny ve více iteracích. Přesnost systému U je vypočtena jako

$$U = 1 - \frac{F + M}{N}, \tag{6.1}$$

kde N je počet segmentů příslušné nahrávky. Pro nahrávky, ve kterých jsou mluvčí odděleni, je celková přesnost počítána jako průměr přesností kanálů. V případě nahrávek pořízených

[1] http://trans.sourceforge.net/en/presentation.php

diktafonem je přesnost počítána pro oba kanály zároveň tak, že dojde k logické disjunkci řečové aktivity příslušných kanálů.

V následujících tabulkách se objevují metody, které byly použity pro vyhlazení. Negací úseků se rozumí změna z řeči na ticho a opačně na úsecích o délce menší než X, které jsou zároveň obklopeny úseky odlišné hodnoty. p_{loop} označuje pravděpodobnost přetrvání v aktuálním stavu HMM při použití forward-backward algoritmu. Pro všechny níže použité metody je vypočtena střední kvadratická energie E, která je následně zobrazena do rozsahu $E \in \langle 0; 1 \rangle$.

Konstantní práh

Práh	**Vyhlazení**	**Miss [%]**	**False alarm [%]**	**Přesnost [%]**
CallHome				
Dynamicky vypočtený	-	23,973	2,639	73,388
0,05	-	27,746	0,226	72,028
0,005	-	7,768	3,960	88,272
0,004	-	6,766	4,856	88,378
0,004	Mediánový filtr (100 ms)	5,036	4,859	90,105
0,004	Mediánový filtr (300 ms)	3,881	4,601	91,518
0,004	Mediánový filtr (300 ms) Negace úseků o délce $L < 250$ ms	3,691	4,552	91,757
0,006	Mediánový filtr (300 ms) Negace úseků ticha o délce $L < 250$ ms Negace úseků řeči o délce $L < 100$ ms	4,628	3,043	92,329
Online nahrávky				
0,004	-	5,037	5,481	89,483
0,004	Mediánový filtr (300 ms)	3,127	5,272	91,601
0,006	Mediánový filtr (300 ms) Negace úseků ticha o délce $L < 250$ ms Negace úseků řeči o délce $L < 100$ ms	3,205	3,996	92,799
Nahrávky pořízené diktafonem				
0,004	-	3,099	7,104	89,797
0,004	Mediánový filtr (300 ms)	3,052	6,854	90,095

Adaptivní práh

Koeficient růstu minimální energie	λ	Vyhlazení	Miss [%]	False alarm [%]	Přesnost [%]
CallHome					
1,0001	0,95	-	22,815	0,755	76,430
1,0001	0,99	-	9,948	3,101	86,950
1,000001	0,99	-	9,326	3,546	87,128
1,000001	0,99	Mediánový filtr (100 ms)	7,284	3,476	89,240
1,000001	0,99	Mediánový filtr (300 ms)	5,589	3,290	91,120
1,000001	0,99	Mediánový filtr (300 ms) Negace úseků ticha o délce $L < 400$ ms Negace úseků řeči o délce $L < 100$ ms	5,680	2,566	91,754
Online nahrávky					
1,000001	0,99	-	7,766	3,123	89,111
1,000001	0,99	Mediánový filtr (300 ms)	5,169	2,918	91,914
1,000001	0,99	Mediánový filtr (300 ms) Negace úseků ticha o délce $L < 400$ ms Negace úseků řeči o délce $L < 100$ ms	4,309	2,975	92,716
Nahrávky pořízené diktafonem					
1,000001	0,99	-	10,062	3,428	86,510
1,000001	0,99	Mediánový filtr (300 ms)	9,549	3,507	86,944
1,000001	0,99	Mediánový filtr (300 ms) Negace úseků ticha o délce $L < 400$ ms Negace úseků řeči o délce $L < 100$ ms	8,917	3,527	87,555

Směsice tří Gaussovských rozložení

Prahována složka	Práh [%]	Vyhlazení	Miss [%]	False alarm [%]	Přesnost
CallHome					
Ticho	20	-	7,862	3,445	88,693
Ticho	90	-	6,282	4,444	89,275
Šum	5	-	13,716	5,724	80,560
Řeč	1,5	-	10,659	3,883	85,457
Řeč	90	-	30,740	0,330	68,930
Řeč	1	$p_{loop} = 0,9$	14,341	4,373	81,286
Ticho	95	$p_{loop} = 0,9$	3,299	4,989	91,712
Ticho	95	$p_{loop} = 0,95$	3,440	4,551	92,009
Ticho	95	$p_{loop} = 0,95$ Mediánový filtr (100 ms)	3,043	4,750	92,207
Ticho	95	$p_{loop} = 0,95$ Mediánový filtr (500 ms)	2,453	3,995	93,552
Ticho	95	$p_{loop} = 0,95$ Mediánový filtr (500 ms) Negace úseků ticha o délce $L < 400$ ms Negace úseků řeči o délce $L < 100$ ms	2,410	4,016	93,574
Online nahrávky					
Ticho	90	-	4,789	4,273	90,937
Řeč	1,5	-	8,401	7,809	83,790
Ticho	95	$p_{loop} = 0,95$	3,329	3,293	93,378
Ticho	95	$p_{loop} = 0,95$ Mediánový filtr (500 ms) Negace úseků ticha o délce $L < 400$ ms Negace úseků řeči o délce $L < 100$ ms	2,145	3,190	94,665
Nahrávky pořízené diktafonem					
Ticho	90	-	8,351	4,349	87,300
Ticho	99	$p_{loop} = 0,9$	4,555	5,900	89,546
Ticho	96	$p_{loop} = 0,8$ Negace úseků ticha o délce $L < 400$ ms Negace úseků řeči o délce $L < 100$ ms	2,427	7,875	89,698

Nejlepší varianty navržených systémů dosahují na datasetu CallHome průměrné přesnosti v rozmezí 90–95 %. Této přesnosti dosahují taktéž na nahrávkách online sezení, což odpovídá předpokladu, že přepis vytvořený ASR systémem pro český jazyk skupiny Speech@fit lze předpokládat za referenční. Horší přesnost je detekována na nahrávkách pořízených diktafonem. Systémy založené na energii nedokážou zcela rozlišit okolní šum od mluvy, a proto se na některých nahrávkách vyskytuje poměrně vysoký „False alarm“, který vede k nižší přesnosti celého systému. U nahrávek, které neobsahují značný šum okolí, se přesnost pohybuje v rozmezí 95–100 %.

Z použitých systémů nejlepší výsledky detekce řečové aktivity poskytuje systém směsice tří Gaussovských komponent s „prahováním“ ticha a následným vyhlazením s použitím forward-backward algoritmu, mediánového filtru a negování krátkých úseků řeči/ticha. Systém selhává na nahrávkách obsahujících vysoký výskyt okolních zvuků. Psychoterapeutická sezení však většinou probíhají v tichých prostorech.

6.2 Diarizace

Z řečově aktivních segmentů je dále možné určit, kdo je původcem mluvy. U nahrávek, kde jsou mluvčí odděleni kanály, lze za výstup rovnou považovat detekovanou řečovou aktivitu příslušných kanálů. V případě nahrávek pořízených diktafonem je nutné provést diarizaci jedním z algoritmů popsaných v sekci 2.6. Jelikož referenční anotace nemusí být vždy přesné, jsou systémy vyhodnoceny s „límcem“ o velikosti 250 ms a bez „límce“. Následující tabulky ukazují úspěšnosti nejlepších variant navržených systémů.

6.2.1 Telefonní a online nahrávky

Pro online nahrávky je evaluační metrika přísnější a není vyhodnocována diarizační chyba (DER) společně pro oba kanály, ale je vyhodnocen pro úspěšnost detekce řečové aktivity pro oba kanály a tato hodnota je průměrována, čímž vzniká o něco přísnější metrika.

Typ systému	**Miss [%]**	**False alarm [%]**	**Přesnost [%]**
`Konstantní práh`, práh = 0,006			
Mediánový filtr (300 ms)	4,628	3,043	92,329
Negace úseků ticha o délce $L < 250$ ms	3,070	2,177	94,752
Negace úseků řeči o délce $L < 100$ ms			
`Adaptivní práh`, růst = 1,000001, $\lambda = 0,99$			
Mediánový filtr (300 ms)	5,680	2,566	91,754
Negace úseků ticha o délce $L < 400$ ms	3,828	1,797	94,375
Negace úseků řeči o délce $L < 100$ ms			
`Směsice tří Gaussovských rozložení`			
složka – ticho, práh = 95%, $p_{loop} = 0,95$	2,453	3,995	93,552
Mediánový filtr (500 ms)	1,623	2,355	96,021
Negace úseků ticha o délce $L < 400$ ms			
Negace úseků řeči o délce $L < 100$ ms			

Tabulka 6.1: Úspěšnost diarizace závislá na použité metodě detekce řečové aktivity datasetu CallHome. Příslušné buňky metrik systému obsahují dvě hodnoty. Horní hodnota odpovídá evaluaci bez použití „límce“ a dolní evaluaci s použitím „límce“ o velikosti 250 ms.

Typ systému	Miss [%]	False alarm [%]	Přesnost [%]
`Konstantní práh`, práh = 0,006			
Mediánový filtr (300 ms)	3,205	3,996	92,799
Negace úseků ticha o délce $L < 250$ ms	1,961	3,459	94,579
Negace úseků řeči o délce $L < 100$ ms			
`Adaptivní práh`, růst = 1,000001, $\lambda = 0,99$			
Mediánový filtr (300 ms)	4,309	2,975	92,716
Negace úseků ticha o délce $L < 400$ ms	2,860	2,589	94,551
Negace úseků řeči o délce $L < 100$ ms			
`Směsice tří Gaussovských rozložení`			
složka – ticho, práh = 95%, $p_{loop} = 0,95$	2,145	3,190	94,665
Mediánový filtr (500 ms)	1,227	2,551	96,222
Negace úseků ticha o délce $L < 400$ ms			
Negace úseků řeči o délce $L < 100$ ms			

Tabulka 6.2: Úspěšnost diarizace závislá na použité metodě detekce řečové aktivity nahrávek online psychoterapeutických sezení projektu DeePsy. Obdobně jak v předchozí tabulce příslušné buňky metrik systému obsahují dvě hodnoty. Horní hodnota odpovídá evaluaci bez použití „límce“ a dolní evaluaci s použitím „límce“ o velikosti 250 ms.

6.2.2 Nahrávky pořízené diktafonem

S nahrávkami pořízenými diktafonem ZOOM H2n byly provedeny experimenty pomocí navržených metod. Nahrávky prezenčních psychoterapeutických sezení však byly do systému projektu DeePsy nahrány až v době odevzdání této práce, a tak nebylo zcela možné vytvořit ruční anotace většího množství těchto nahrávek. Optimální parametry navržených metod tak byly validovány pouze na části dostupných nahrávek, které byly do té doby anotovány. Metody jsou taktéž porovnány se značně robustnějším systémem VBHMM x-vectors Diarization (VBx) [7], který provádí diarizaci na nahrávce s jedním kanálem (pro tyto účely jsou kanály vstupních nahrávek spojeny do jednoho kanálu pomocí nástroje SoX[2]). Aby nebyla chyba diarizace DER (definována v sekci 4.4) zkreslená chybou způsobenou nepřesnou detekcí řečové aktivity, je načtená její referenční podoba ze souborů s příponou `.rttm` a nad ní je provedena diarizace. Tento způsob dovoluje pouze vyhodnocení chyby C (Confusion), kdy systém označí jako původce řeči terapeuta, avšak je jím klient a opačně. Hodnoty F (False alarm) a M (Miss) jsou rovné nule.

Navržené metody nejsou schopné detekovat více řečníků v jednom segmentu, jelikož provádějí diarizaci tvrdým rozhodnutím na základě pravděpodobností mluvy jednotlivých řečníků. Toto omezení znemožňuje detekci skoků do řeči, a z tohoto důvodu nejsou skoky do řeči v případě prezenčního sezení zahrnuty do souhrnné zprávy. Aby bylo možné tyto statistiky zahrnout, výstupem diarizace by musela být pravděpodobnost mluvy každého z řečníků a u segmentů, kde by byla pravděpodobnost podobně veliká, by bylo možné pomocí dalšího modelu detekovat pravděpodobnost, že mluví oba řečníci.

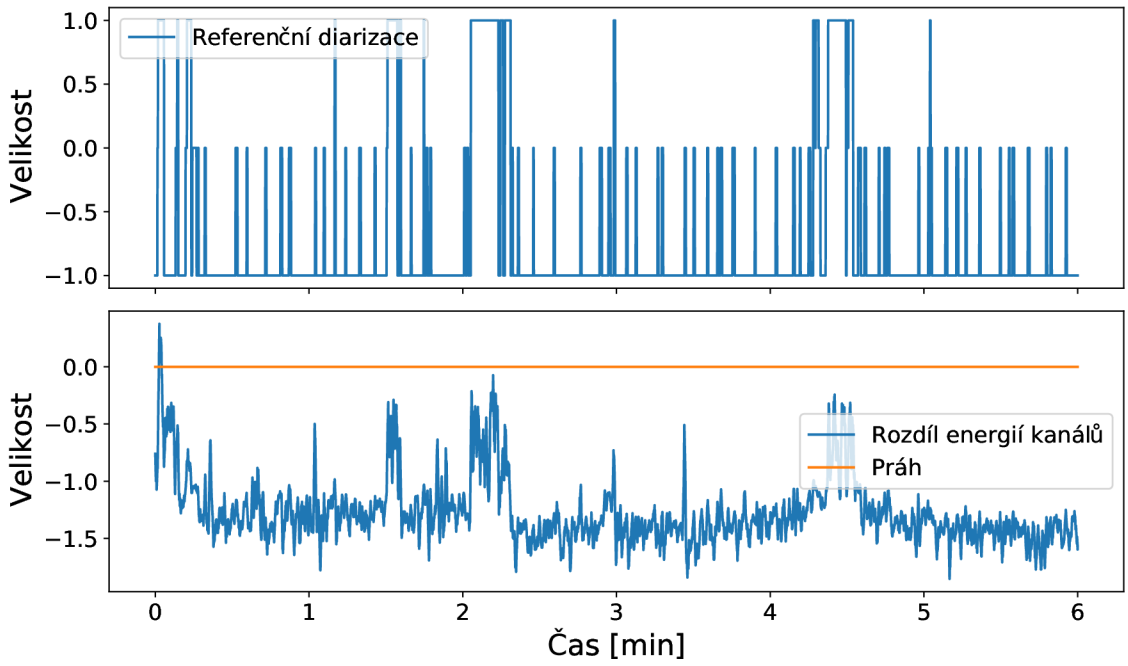
Následující tabulka obsahuje tuto chybu označenou v rovnici chybovosti diarizace 4.12 jako C (Confusion). Zároveň je vyhodnoceno P_1 jako procento počtu segmentů mluvy terapeuta, kdy byl terapeut klasifikován jako klient a P_2 jako procento počtu segmentů mluvy klienta, kdy byl klient klasifikován jako terapeut. Obdobně jako v předchozím případě jsou metody vyhodnoceny s „límcem“ o velikosti 250 ms a bez „límce“.

[2] http://sox.sourceforge.net/sox.html

Metoda	Vyhlazení	Límec	P_1 [%]	P_2 [%]	C [%]
Energetická diarizace – $E \in \langle 0; 1 \rangle$	-	0	28,953	18,435	21,997
Energetická diarizace	-	0	38,231	16,227	13,378
Energetická diarizace	Mediánový filtr (300 ms)	0	36,182	9,822	9,009
Energetická diarizace	Mediánový filtr (300 ms)	250	35,261	9,057	7,594
Energetická diarizace	Mediánový filtr (1 s)	0	36,548	9,139	8,604
Energetická diarizace	Mediánový filtr (1 s)	250	35,433	8,065	7,043

Tabulka 6.3: Úspěšnost diarizačního systému založeného na rozhodnutí podle vyšší z energií příslušných segmentů kanálů nahrávky.

Ačkoliv je chybovost systému energetické diarizace C poměrně nízká, systém naprosto selhává v případě, kdy energie jednoho z kanálu je vyšší v průběhu celého sezení, jak je zobrazeno na obr. 6.1, což odpovídá velmi vysoké hodnotě chyby P_1.



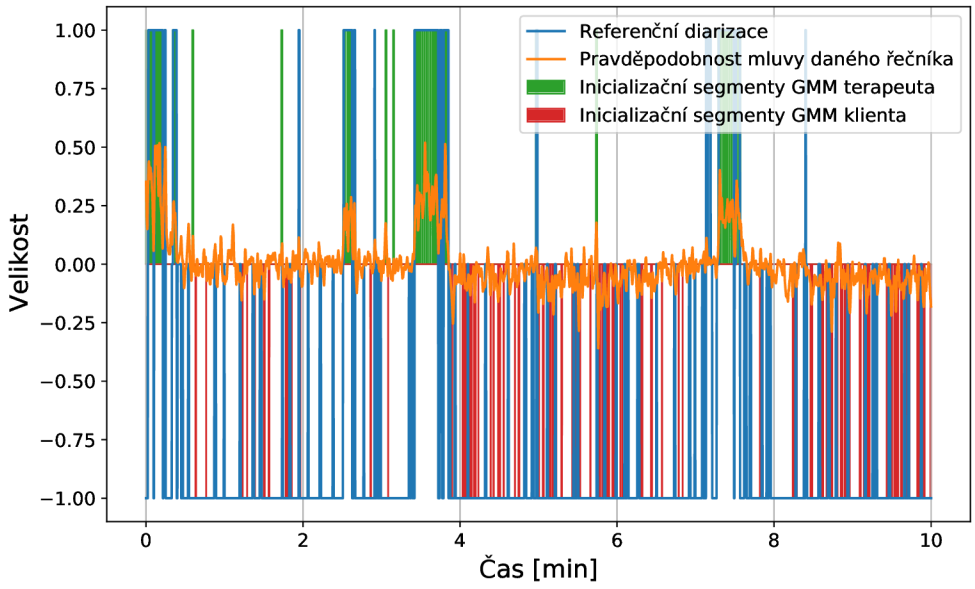
Obrázek 6.1: Na horním obrázku jsou zobrazeny hodnoty referenční diarizace příslušných segmentů. Na dolním obrázku je vidět rozdíl energie kanálu terapeuta oproti energii kanálu klienta, podle kterého následně dochází k určení řečníka. Hodnota 1 označuje segmenty, ve kterých hovořil terapeut, 0 ticho a −1 segmenty, ve kterých hovořil klient. Je pravděpodobné, že byla špatně zvolena pozice diktafonu, a tak je energie kanálu klienta po celou dobu vyšší, což vede ke špatné detekci mluvčího. Možným řešením může byt normalizace energie do stejného rozsahu nebo prahování pomocí průměrné hodnoty rozdílu energií, tato řešení však selhávají na ostatních nahrávkách, kde byla správně nastavena poloha diktafonu.

Tabulka 6.4 zobrazuje chybovosti metod založených na shlukování segmentů podle mel-frekvenčních koeficientů, které byly navrženy za účelem diarizace nahrávek prezenčních sezení.

MFCC koeficienty	Adaptace GMM podle	Vyhlazení	Límec	$\mathbf{P}_1$ [%]	$\mathbf{P}_2$ [%]	C [%]
10 MFCC jednoho kanálu	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 5 %	-	0	34,240	42,742	31,838
20 MFCC jednoho kanálu	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 5 %	-	0	33,026	42,082	31,463
20 MFCC jednoho kanálu	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 5 %	$mean(\mathbf{e}_d; 0,5\ s)$	0	30,212	40,889	30,438
20 MFCC jednoho kanálu	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 5 %	$mean(\mathbf{e}_d; 0,5\ s)$ $fb(\mathbf{p}^1\|\mathbf{p}^2; 0, 9)$	0	14,497	32,379	21,935
20 MFCC jednoho kanálu	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 5 %	$mean(\mathbf{e}_d; 0,5\ s)$ $fb(\mathbf{p}^1\|\mathbf{p}^2; 0, 9)$ $mean(\mathbf{r}; 0,5\ s)$	0	12,974	25,623	17,715

Tabulka 6.4: Úspěšnost diarizačního systému založeného na shlukování segmentů podle mel-frekvenčních koeficientů. Notace $\mathbf{p}^1|\mathbf{p}^2$ označuje konkatenaci dvou vektorů do matice. $fb(\mathbf{v}; x)$ označuje provedení forward-backward algoritmu nad maticí $\mathbf{v}$ s hodnotou $p_{loop} = x$. Obdobně $mean(\mathbf{v}; x)$ reprezentuje aplikaci průměrového filtru nad vektorem $\mathbf{v}$ o velikosti x sekund.

Je patrné, že takto navržený systém selhává. Důvodem je nejistota systému v některých segmentech způsobená nízkým rozdílem pravděpodobností vektoru $\mathbf{r}$. Obr. 6.2 zobrazuje princip fungování navrženého systému a ukazuje, kde systém selhává.



Obrázek 6.2: Princip fungování systému diarizace pomocí mel-frekvenčních koeficientů. Hodnota 1 odpovídá segmentům, ve kterých mluví terapeut a −1 segmentům, ve kterých mluví klient. 0 odpovídá tichu. Oranžová křivka zobrazuje pravděpodobnost mluvy příslušných osob. Je patrné, že systém dokáže velmi dobře detekovat segmenty, pomocí nichž byla provedena adaptace – tyto segmenty jsou podbarvené zeleně a červeně. V mnoha segmentech si však systém není jistý a hodnoty kolísají mezi kladnou a zápornou hodnotou na *ose y*, což způsobuje poměrně vysokou chybu diarizace.

Navrženou metodu a její deriváty, jejichž chybovost byla téměř identická, proto bylo nutné podrobit dalšímu zpracování. Je tedy použita metoda prahování příslušných složek. Dokud není překročen příslušný práh, zůstává hovořícím mluvčí předchozího segmentu. Tuto metodu je možné definovat následovně:

$$d_n = \begin{cases} 2 & r_n >= F_2 \\ 1 & r_n <= F_1 \\ 0 & u_n \\ d_{n-k} & jinak \end{cases}, n \in \{1, \ldots, N\}, \tag{6.2}$$

kde d_{n-k} je první aktivní segment, F_1 je práh vypočtený jako $\frac{\min(\mathbf{r})}{5}$ a F_2 jako $\frac{\max(\mathbf{r})}{5}$. Zbylé notace hodnot odpovídají těm, které byly zavedeny v sekci 4.2.2. Tabulka 6.5 ukazuje chybovost upravených metod systému.

MFCC koeficienty	**Adaptace GMM podle**	**Vyhlazení**	**Límec**	**P**$_1$ [%]	**P**$_2$ [%]	**C** [%]
20 MFCC jednoho kanálu	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 5 %	$mean(\mathbf{e}_d; 0,5\ s)$ $fb(\mathbf{p}^1 \vert \mathbf{p}^2; 0, 9)$ $mean(\mathbf{r}; 0,5\ s)$	0	8,753	14,873	10,262
20 MFCC jednoho kanálu + 20 delta koeficientů	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 5 %	$mean(\mathbf{e}_d; 0,5\ s)$ $fb(\mathbf{p}^1 \vert \mathbf{p}^2; 0, 9)$ $mean(\mathbf{r}; 0,5\ s)$	250	5,047	11,617	7,779
20 MFCC jednoho kanálu + 20 delta koeficientů	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 5 %	$mean(\mathbf{e}_d; 0,5\ s)$ $fb(\mathbf{p}^1 \vert \mathbf{p}^2; 0, 9)$ $mean(\mathbf{r}; 0,5\ s)$	250	3,434	10,071	6,043
(20 + 20) MFCC kanálů + (20 + 20) delta koeficientů	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 5 %	$mean(\mathbf{e}_d; 0,5\ s)$ $fb(\mathbf{p}^1 \vert \mathbf{p}^2; 0, 9)$ $mean(\mathbf{r}; 0,5\ s)$	0	5,144	10,933	6,629
(20 + 20) MFCC kanálů + (20 + 20) delta koeficientů	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 5 %	$mean(\mathbf{e}_d; 0,5\ s)$ $fb(\mathbf{p}^1 \vert \mathbf{p}^2; 0, 9)$ $mean(\mathbf{r}; 0,5\ s)$	250	1,936	9,855	5,519
(20 + 20) MFCC kanálů + (20 + 20) delta koeficientů	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 5 % 2 iterace adaptace	$mean(\mathbf{e}_d; 0,5\ s)$ $fb(\mathbf{p}^1 \vert \mathbf{p}^2; 0, 9)$ $mean(\mathbf{r}; 0,5\ s)$	0	5,990	14,652	9,420
(20 + 20) MFCC kanálů + (20 + 20) delta koeficientů	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 5 % 2 iterace adaptace	$mean(\mathbf{e}_d; 0,5\ s)$ $fb(\mathbf{p}^1 \vert \mathbf{p}^2; 0, 9)$ $mean(\mathbf{r}; 0,5\ s)$	250	4,398	13,413	7,611

Tabulka 6.5: Chybovost metod po aplikaci prahování. Nejlepší výsledky jsou dosaženy pomocí metody používající mel-frekvenční koeficienty obou kanálu bez následné adaptace v druhé iteraci.

Nejlépe fungující metoda byla dále podrobena sérii dalších pokusů, jejichž výsledky ilustruje tabulka 6.6.

Navržené metody dosahují nejlepší výsledky chybovosti C okolo 5 %, nebyly však testovány na dostačujícím počtu nahrávek tak, aby bylo možné objektivně posuzovat jejich úspěšnost. V případě pochybností o výsledcích diarizace je možné využít systém VBx, který je značně robustnější a otestován na velkém množství dat. Jeho chybovost na stejných nahrávkách zobrazuje tabulka 6.7.

MFCC koeficienty	Adaptace GMM podle	Vyhlazení	$\mathbf{P}_1$ [%]	$\mathbf{P}_2$ [%]	$\mathbf{C}$ [%]
(20 + 20) MFCC kanálů + (20 + 20) delta koeficientů	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 10 %	$mean(\mathbf{e}_d; 0,5\ s)$ $fb(\mathbf{p}^1\|\mathbf{p}^2; 0, 9)$ $mean(\mathbf{r}; 0,5\ s)$	5,490 4,069	8,361 6,806	5,945 4,397
(25 + 25) MFCC kanálů + (25 + 25) delta koeficientů	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 10 %	$mean(\mathbf{e}_d; 0,5\ s)$ $fb(\mathbf{p}^1\|\mathbf{p}^2; 0, 9)$ $mean(\mathbf{r}; 0,5\ s)$	4,475 3,194	16,009 14,619	9,497 7,662
(15 + 15) MFCC kanálů + (15 + 15) delta koeficientů	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 10 %	$mean(\mathbf{e}_d; 0,5\ s)$ $fb(\mathbf{p}^1\|\mathbf{p}^2; 0, 9)$ $mean(\mathbf{r}; 0,5\ s)$	4,260 3,240	9,491 7,826	6,282 4,723
(10 + 10) MFCC kanálů + (10 + 10) delta koeficientů	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 10 %	$mean(\mathbf{e}_d; 0,5\ s)$ $fb(\mathbf{p}^1\|\mathbf{p}^2; 0, 9)$ $mean(\mathbf{r}; 0,5\ s)$	3,180 2,197	9,606 7,938	6,055 4,538
(10 + 10) MFCC kanálů + (10 + 10) delta koeficientů	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 15 %	$mean(\mathbf{e}_d; 0,5\ s)$ $fb(\mathbf{p}^1\|\mathbf{p}^2; 0, 9)$ $mean(\mathbf{r}; 0,5\ s)$	3,447 2,548	8,489 6,725	5,464 3,947
(10 + 10) MFCC kanálů + (10 + 10) delta koeficientů	10 % segmentů $\mathbf{e}_d$ posun $\boldsymbol{\mu}$ o 20 %	$mean(\mathbf{e}_d; 0,5\ s)$ $fb(\mathbf{p}^1\|\mathbf{p}^2; 0, 9)$ $mean(\mathbf{r}; 0,5\ s)$	5,481 4,262	7,633 5,891	5,121 3,602

Tabulka 6.6: Chybovost nejlépe fungující metody diarizace prezenčních nahrávek s různou konfigurací. Chybovosti v příslušných řádcích obsahují postupně hodnoty chyb pro „límec“ rovný 0 ms a „límec“ o velikosti 250 ms.

	$\mathbf{P}_1$ [%]	$\mathbf{P}_2$ [%]	$\mathbf{C}$ [%]
Límec 0 ms	8,724	1,431	2,572
Límec 250 ms	6,638	0,712	1,449

Tabulka 6.7: Chybovost systému VBx. Lze prohlásit, že daný systém funguje velmi dobře a je možné ho použít pro následné získání statistik.

6.3 Validace statistik

Statistiky, které jsou získány z psychoterapeutických sezení a jejich odvozené průměrné hodnoty byly validovány poslechem audionahrávek. Statistiky založené na zpracování přirozeného jazyka byly validovány manuální analýzou přepisu a kontrolou přesnosti použitého rozpoznávače řeči. Dostupné online rozpoznávače řeči pro anglický jazyk na datasetu CallHome dosahovaly poměrně nízké úspěšnosti (70–90 %), avšak pro psychoterapeutická sezení, jejichž analýza je primární, byly přepisy vytvořeny ASR systémem pro český jazyk, jehož chybovost se pohybuje v řádech jednotek procent. Textové přepisy byly validovány implementovaným skriptem a subjektivním poslechem příslušných nahrávek. Neobsahovaly zásadní chyby, které by nějakým způsobem ovlivňovaly kontext sezení a z tohoto důvodu bylo možné je použít pro následnou extrakci slovních statistik.

Kapitola 7

Závěr

Cílem této bakalářské práce bylo vytvořit systém pro analýzu audio hovorů mezi dvěma účastníky. Tvorba tohoto systému vyžadovala studium a hlubší pochopení technik pro zpracování řečových signálů a paralingvistických charakteristik řeči.

Postupně byla čtenáři představena teorie nutná pro pochopení dané problematiky a data, se kterými systém pracuje. V dalších částech byl popsán navržený systém, způsob jeho implementace v jazyce Python a jeho úspěšnost. Vytvořený systém splňuje body dané zadáním a je přizpůsoben k použití pro analýzu psychoterapeutických sezení. Pro příslušné nahrávky je vygenerována zpráva v podobě HTML dokumentu obsahující vybrané charakteristiky, jejich změny v čase a porovnání s ostatními sezeními. Byly vygenerovány referenční statistiky pro psychoterapeutická sezení, vůči kterým je možné nové sezení porovnávat. V příloze se nachází plakát prezentující provedenou práci a náhled dokumentu obsahujícího souhrnnou zprávu proběhlého psychoterapeutického sezení.

Poskytnutý systém pro analýzu nahrávek psychoterapeutických sezení dosahuje přesnosti diarizace okolo 95 %. Pro tvorbu přepisu nahrávek byl využit externí rozpoznávač řeči. Analýza přepisu rozšířila souhrnnou zprávu o další metriky, což vedlo k ještě vyšší přidané hodnotě této zprávy pro terapeuty.

Souhrnnou zprávu by bylo možné rozšířit o další charakteristiky získané metodami zpracování přirozeného jazyka. Mezi tyto charakteristiky může patřit detekce klíčových slov, detekce probíraných témat nebo sumarizace témat proběhlé konverzace. Systém by bylo rovněž vhodné doplnit o detekci nálad pro český jazyk, která je aktuálně dostupná pouze pro angličtinu. Po grafické stránce by bylo možné implementovat další vylepšení, ať už v podobě přehrávání a vizualizace důležitých úseků nahrávky přímo v prohlížeči nebo výraznějšího použití kaskádových stylů pro lepší zážitek uživatele. Taktéž by mohla být velmi zajímavá implementace vlastního interaktivního prohlížeče audio nahrávky, ve kterém by byly zvýrazněny problematické části s možností jejich automatického přehrávání. Dalším vylepšením by mohlo být nasazení existujících skriptů na webový server, aby byl téměř kdokoliv schopen provádět analýzu hovoru.

Literatura

[1] Bishop, C. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. říjen 2007. ISBN 0387310738.

[2] Bäckström, T. *Voice activity detection (VAD)*. Srpen 2020 [cit. 2021-04-15]. Dostupné z: https://wiki.aalto.fi/pages/viewpage.action?pageId=151500905.

[3] Castaldo, F., Colibro, D., Dalmasso, E., Laface, P. a Vair, C. Stream-based speaker segmentation using speaker factors and eigenvoices. In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2008, s. 4133–4136. DOI: 10.1109/ICASSP.2008.4518564.

[4] Davis, S. a Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 1980, sv. 28, č. 4, s. 357–366. DOI: 10.1109/TASSP.1980.1163420.

[5] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P. a Ouellet, P. Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*. 2011, sv. 19, č. 4, s. 788–798. DOI: 10.1109/TASL.2010.2064307.

[6] Dempster, A. P., Laird, N. M. a Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*. 1977, sv. 39, č. 1, s. 1–38.

[7] Diez, M. S., Burget, L., Landini, N. F. a Černocký, J. Analysis of Speaker Diarization based on Bayesian HMM with Eigenvoice Priors. *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING*. 2020, sv. 28, č. 1, s. 355–368. DOI: 10.1109/TASLP.2019.2955293. ISSN 2329-9290. Dostupné z: https://www.fit.vut.cz/research/publication/12139.

[8] Dittmann, A. a Llewellyn, L. The phonemic clause as a unit of speech decoding. *Journal of personality and social psychology*. 1. vyd. Červenec 1967, sv. 6, č. 3, s. 341—349. DOI: 10.1037/h0024739. ISSN 0022-3514.

[9] Ecma International. *Standard ECMA-404*. Prosinec 2017 [cit. 2021-03-20]. Dostupné z: https://www.ecma-international.org/publications-and-standards/standards/ecma-404.

[10] Ecma International. *Standard ECMA-262*. Červen 2020 [cit. 2021-03-20]. Dostupné z: https://www.ecma-international.org/publications-and-standards/standards/ecma-262.

[11] Gupta, G. Algorithm for image processing using improved median filter and comparison of mean, median and improved median filter. *International Journal of Soft Computing and Engineering*. 2011, sv. 1, č. 5, s. 304–311.

[12] Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P. et al. Array programming with NumPy. *Nature*. Září 2020, sv. 585, č. 7825, s. 357–362. DOI: 10.1038/s41586-020-2649-2. Dostupné z: https://doi.org/10.1038/s41586-020-2649-2.

[13] Huang, X., Acero, A. a Hon, H.-W. *Spoken language processing: a guide to theory, algorithm, and system development*. Second. Prentice Education Taiwan, 2005. ISBN 0-13-022616-5.

[14] IBM. *Gauge Charts*. [cit. 2021-03-19]. Dostupné z: https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=types-gauge-charts.

[15] IBM Cloud Education. *What is Speech Recognition?* Září 2020 [cit. 2021-03-20]. Dostupné z: https://www.ibm.com/cloud/learn/speech-recognition.

[16] Igras, M., Ziółko, B. a Ziółko, M. Length of phonemes in a context of their positions in Polish sentences. In: *2013 International Conference on Signal Processing and Multimedia Applications (SIGMAP)*. 2013, s. 59–64.

[17] Jurafsky, D. a Martin, J. H. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. 3. vyd. Pearson, 2020. ISBN 9780131227989.

[18] Karafiát, M., Baskar, M. K., Szöke, I., Vydana, H. K., Veselý, K. et al. BUT Opensat 2019 Speech Recognition System. In:. 2020. ArXiv:2001.11360v1. Dostupné z: https://arxiv.org/abs/2001.11360.

[19] Kirill, S., Ekaterina, V. a Simak, B. Approach for Energy-Based Voice Detector with Adaptive Scaling Factor. *IAENG International Journal of Computer Science*. Listopad 2009, sv. 36.

[20] Křivohlavý, J. Jak si navzájem lépe porozumíme: kapitola Paralingvistika - svrchní tóny. In: *Jak si navzájem lépe porozumíme*. 1. vyd. Svoboda, 1988.

[21] Lyons, J. *Mel Frequency Cepstral Coefficient (MFCC) tutorial*. 2013 [cit. 2021-15-02]. Dostupné z: http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/#eqn1.

[22] McNamara, A. a Lunzer, A. *Exploring Histograms*. Září 2020 [cit. 2021-03-19]. Dostupné z: https://tinlizzie.org/histograms/.

[23] Moattar, M. H. a Homayounpour, M. M. A simple but efficient real-time Voice Activity Detection algorithm. In: *2009 17th European Signal Processing Conference*. 2009, s. 2549–2553.

[24] National Institute of Standards and Technology. *RT-06S Transcription Evaluation Plan*. 2006.

[25] Prof. PhDr. Marie Krčmová, C. *Fonetika* [online]. Filosofická fakulta MU Brno, 2007 [cit. 2021-15-02]. Dostupné z: http://is.muni.cz/elportal/estud/ff/js07/fonetika/materialy/ch05s01.html.

[26] Psychiatrická nemocnice Havlíčkův Brod. *Bipolární afektivní porucha.* [cit. 2021-03-10]. Dostupné z: http://www.plhb.cz/content/bipolarni-afektivni-porucha.

[27] Python Software Foundation. *Pickle - Python object serialization.* 2001-2021 [cit. 2021-03-20]. Dostupné z: https://docs.python.org/3/library/pickle.html.

[28] Real Python. *Primer on python decorators.* Duben 2021 [cit. 2021-04-15]. Dostupné z: https://realpython.com/primer-on-python-decorators/.

[29] Real Python. *Python 3's f-Strings: An Improved string formatting Syntax (Guide).* Březen 2021 [cit. 2021-04-15]. Dostupné z: https://realpython.com/python-f-strings/.

[30] Real Python. *When to Use a List Comprehension in Python.* Březen 2021 [cit. 2021-04-15]. Dostupné z: https://realpython.com/list-comprehension-python/.

[31] Reynolds, D. Gaussian Mixture Models. In: *Encyclopedia of Biometrics.* 2009.

[32] Rodero, E. A comparative analysis of speech rate and perception in radio bulletins. *Text & Talk.* 2012, sv. 32, č. 3, s. 391–411. DOI: doi:10.1515/text-2012-0019. Dostupné z: https://doi.org/10.1515/text-2012-0019.

[33] Sigmund, M. *Analýza řečových signálů.* 24-2000. MJ Servis: MJ Servis, červen 2001.

[34] 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.

[35] Wang, Q., Downey, C., Wan, L., Mansfield, P. A. a Moreno, I. L. Speaker Diarization with LSTM. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2018, s. 5239–5243. DOI: 10.1109/ICASSP.2018.8462628.

[36] Web Hypertext Application Technology Working Group. *HTML Living Standard.* 2021 [cit. 2021-15-02]. Dostupné z: https://html.spec.whatwg.org/.

[37] Wikipedia contributors. *Tag cloud — Wikipedia, The Free Encyclopedia.* 2021 [cit. 2021-03-19]. Dostupné z: https://en.wikipedia.org/w/index.php?title=Tag_cloud&oldid=1002276039.

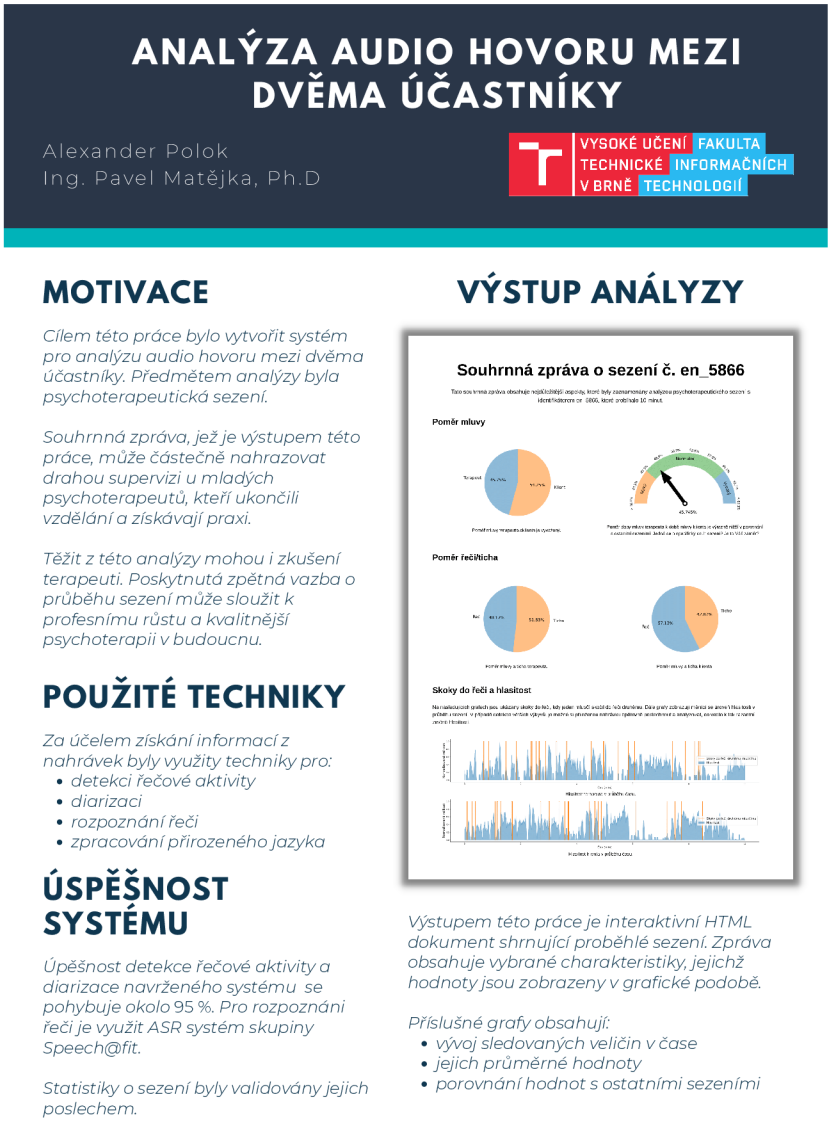
[38] Wilke, C. O. *Fundamentals of Data Visualization.* 2020 [cit. 2021-03-19]. Dostupné z: https://clauswilke.com/dataviz/visualizing-proportions.html#visualizing-proportions.

[39] Ústav pro jazyk český AV ČR, v. v. i.. *Internetová jazyková příručka.* 2008-2021 [cit. 2021-03-19]. Dostupné z: https://prirucka.ujc.cas.cz/.

[40] Černocký, J. *Zpracování řečových signálů – studijní opora.* 2016 [cit. 2021-03-10]. Dostupné z: https://www.fit.vutbr.cz/study/courses/ZRE/public/opora/zre_opora.pdf.

Příloha A

Plakát



Obrázek A.1: Plakát prezentující navržený systém a jeho výstupy.

Příloha B

Souhrnná zpráva

Obrázek B.1: Náhled části souhrnného dokumentu, který je automaticky vygenerován pro příslušná psychoterapeutická sezení.