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

ROBUSTNÍ DETEKCE ŘEČOVÉ AKTIVITY

ROBUST SPEECH ACTIVITY DETECTION

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. ANNA POPKOVÁ**
AUTHOR

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

BRNO 2019

Vysoké učení technické v Brně
Fakulta informačních technologií

Ústav počítačové grafiky a multimédií (UPGM) Akademický rok 2018/2019

Zadání diplomové práce

Studentka: **Popková Anna, Bc.**
Program: Informační technologie Obor: Počítačová grafika a multimédia
Název: **Robustní detekce řečové aktivity**
Robust Speech Activity Detection
Kategorie: Zpracování řeči a přirozeného jazyka

Zadání:

1. Prostudujte statistické techniky pro modelování řeči - zaměřte se převážně na Gaussovské modely a Neuronové sítě
2. Seznamte se s metody používanými pro detekci řečové aktivity
3. Seznamte se s daty, např. z veřejných evaluací řečových technologií.
4. Experimentujte s robustností systemu v šumu a multilingválním prostředí.
5. Navrhněte systém, který umí robustně detekovat i řeč na hudebním pozadí.
6. Vyhodnoťte na datech a analyzujte chování systému.

Literatura:

- NIST Open SAD Evaluation Plan - https://www.nist.gov/itl/iad/mig/nist-open-speech-activity-detection-evaluation
- M. Graciarena, L. Ferrer and V. Mitra, "The SRI System for the NIST OpenSAD 2015 Speech Activity Detection Evaluation," in *Proc. INTERSPEECH 2016*, pp. 3673-3677, September 2016. https://www.sri.com/work/publications/sri-system-nist-opensad-2015-speech-activity-detection-evaluation

Při obhajobě semestrální části projektu je požadováno:

- Body 1 až 3.

Podrobné závazné pokyny pro vypracování práce viz http://www.fit.vutbr.cz/info/szz/

Vedoucí práce: **Matějka Pavel, Ing., Ph.D.**
Vedoucí ústavu: Černocký Jan, doc. Dr. Ing.
Datum zadání: 1. listopadu 2018
Datum odevzdání: 22. května 2019
Datum schválení: 6. listopadu 2018

Abstrakt

Cílem této práce je navrhnout a vytvořit robustní detektor řečové aktivity, který je schopen detekovat řeč v různých jazycích, v prostředí se šumem a v prostředí s hudbou na pozadí. Tento problém jsem se rozhodla vyřešit použitím neuronové sítě jako klasifikačního modelu, který vstupním úsekům nahrávky přiřazuje jednu ze čtyř možných tříd — ticho, řeč, hudbu nebo hluk. Výsledný nástroj je schopný detekovat řeč v minimálně 12–ti jazycích. Řeč na hudebním pozadí až s 88 % úspěšností a výsledky úspěšnosti systému na zašuměných datech dosahují od 84 % (5 dB SNR) do 88 % (20 dB SNR). Tento nástroj je možné použít pro detekci řečové aktivity v různých výzkumných oblastech zpracování řeči. Hlavním jeho přínosem je eliminace hudby, která když odstraněna není, výrazně zvyšuje chybovost systémů na rozpoznávání mluvčího či řeči.

Abstract

The aim of this work is to design and create a robust speech activity detector that is able to detect speech in different languages, in a noise environment and with music on background. I decided to solve this problem by using a neural network as a classification model that assigns one of the four possible classes — silence, speech, music, or noise to the input of audio recording. The resulting tool is able to detect the speech in at least 12 languages. Speech with musical background up to 88 % accuracy and system success on noisy data reaches from 84 % (5 dB SNR) to 88 % (20 dB SNR). This tool can be used for speech activity detection in various research areas of speech processing. The main contribution is the elimination of music, which when not eliminated, significantly increases the error rate of systems for speaker identification or speech recognition.

Klíčová slova

Robustní detekce řečové aktivity, Hudba, Šum, Neuronová síť, SNR.

Keywords

Robust voice activity detection, Music, Noise, Neural Network, SNR.

Citace

POPKOVÁ, Anna. *Robustní detekce řečové aktivity.* Brno, 2019. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Pavel Matějka, Ph.D.

Robustní detekce řečové aktivity

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně pod vedením pana Ing. Pavla Matějky, Ph.D. Uvedla jsem všechny literární prameny a publikace, ze kterých jsem čerpala.

.....

Anna Popková

14. května 2019

Obsah

Kapitola 1

Úvod

Detekce řečové aktivity, *zkr. VAD z angl. Voice Activity Detection*, cílí na problém rozeznání řeči od okolního šumu či jiných neřečových projevů. Toto rozeznávání může probíhat za běhu — online — nebo až zpětně, kdy je k dispozici audio nahrávka — offline. Do okolního šumu zahrnujeme akustické prostředí, ve kterém se řečník vyskytuje. To jest jakýkoli hluk na pozadí (doprava, hudba, déšť, štěkot psů aj.), dále k okolnímu šumu přičítáme i vlastnosti média, přes které přenos řeči probíhá a ostatní zvukové projevy neřečového charakteru jako je smích, pláč, povzdechy, pískání atd.

Detekce řečové aktivity je základní operace, která nachází uplatnění v mnoha aplikacích v oblasti řečových technologií. Zpravidla totiž předchází všechny ostatní operace a je zasazena jako první blok v systémech pro kódování řeči, rozpoznávání řeči, rozpoznávání emocí, identifikace a verifikace mluvčího, identifikace jazyka a mnohá další. Z toho důvodu je od ní očekávána vysoká efektivita a robustnost. Dále existují i systémy, které poskytují detekci řečové aktivity samu o sobě, což slouží k profiltrování velkého množství dat a výsledkem jsou pouze úseky, ve kterých je řeč dominantní. Příkladem mohou být policejní odposlechy plné různých neřečových projevu, kde jejich manuální procházení by zabralo příliš času. Tento čas by díky použití robustního detektoru řeočvé aktivity mohl být ušetřen.

K volbě tohoto tématu mě přivedl zájem o možnosti využití znalostí o struktuře a vlastnostech řečového signálu v oblasti strojového učení. Jako první mě zaujala myšlenka rozeznávání řeči od hudby. Během studia tohoto tématu jsem však došla k názoru, že bude vhodnější pojmout nápad komplexněji, a tudíž navrhnout systém pro detekci řeči nejen na hudebním pozadí, ale i na obecnějším pozadí.

Mým cílem je navrhnout systém, který bude schopný robustně detekovat řeč v prostředí se šumem, v multilingválním prostředí nebo v prostředí s hudbou.

Text, který bude následovat, je členěn následujícím způsobem. První kapitola je věnována teoretickému úvodu do problému. Je v ní diskutováno, jak vzniká lidská řeč a jak je řeč možné modelovat počítačem. Dále je v ní shrnuta aktuální situace na výzkumném poli v oblasti detekce řečové aktivity. Jelikož je detekce řečové aktivity složena ze dvou hlavních částí — z extrakce příznaků a z rozpoznávacího modelu — je tato sekce rozdělena do dvou celků zmíněných významů.

Ve druhé kapitole jsou popsána data, která jsou pro vývoj detektoru řečové aktivity použita, a je předneseno jejich rozdělení do trénovací a testovací sady. Na její závěr jsou představeny evaluační metriky používané v této práci pro kvantitativní zhodnocení funkčnosti systému.

Další kapitola se věnuje návrhu a implementaci systému pro robustní detekci řečové aktivity. Je v ní přednesen podrobný plán práce včetně přípravy dat, výběru a konstrukce

klasifikátoru, návrhu experimentů a návrhu na finální zhodnocení. Druhou část tvoří implementační detaily, kde je zdůvodněna volba implementačního jazyka, jsou popsány použité knihovny a hlavně předneseny implementační detaily klasifikačního modelu — neuronové sítě.

Jádrem práce je kapitola následující poté, která je věnována vývoji úspěšnosti a robustnosti systému. Probíhá zde experimentování se systémem, analýza výsledků a následné ladění systému dle získaných závěrů.

V poslední kapitole jsou čtenáři představeny detaily finálního systému pro robustní detekci řečové aktivity. Celou práci uzavírá kvalitativní zhodnocení systému na poslechu nejlépe klasifikované a nejhůře klasifikované nahrávky.

Kapitola 2

Teoretický základ pro detekci řečové aktivity

Máme-li před sebou klasifikační problém, což rozhodně detekce řečové aktivity je, nabízí se řešení pomocí strojového učení. Ve strojovém učení existuje nepřeberné množství algoritmů představující různé matematické modely. Počínaje lineárním klasifikátorem, přes rozhodovací stromy až k neuronovým sítím. I když tyto modely sehrávají v řešení daného problému klíčovou roli, ve finálním systému zastávají pouze malou část. Jde sice o část neopomenutelnou, avšak je důležité si uvědomit, že velký vliv na úspěšnost konkrétního řešení mají data, kterými tyto modely plníme. Samotná data však nestačí, z dat je nutné dostat důležité informace — tento proces je nazýván extrakce příznaků. Po extrakci jsou s daty prováděny matematické operace. Většinou nechybí normalizace a také bývají využívány operace jako je Analýza hlavních komponent [27] nebo Lineárně diskriminativní analýza [32]. Cílem těchto operací je jak redukce dat, tak zvýšení jejich variabilnosti. Další operace nad daty jsou pak prováděny po výstupu z daného klasifikačního modelu.

Vše zmíněné však bude fungovat pouze tehdy, pokud jsme na začátku získali příznaky vhodné pro konkrétní problém. Snahou je vždy získat příznaky, které jsou diskriminativní pro zvolený klasifikační problém a zároveň v sobě neobsahují informace, které by klasifikaci mohly ztěžovat. Pro tento úkol je důležité dobře porozumět doméně, na které klasifikace probíhá. Tato kapitola se tedy bude zabývat teoretickým základem pro detekci řečové aktivity. Bude vysvětleno, jak řeč vzniká, jaké má vlastnosti a jak je tyto vlastnosti možné modelovat počítačem. Dále jak se liší od dalších akustických projevů jako je šum nebo hudba. Nakonec budou uvedeny modely, které se běžně používají při zpracování řeči.

2.1 Jak vzniká řeč

Tato kapitola čerpá z informací uvedených v [34]. Řeč začíná v plicích, z plic jde energie do hrtanu, kde dochází k její modulaci. Signál, který následně vznikne, je označován jako budící nebo excitační signál. Tento signál pak putuje do hlasivek a v závislosti na tom, zdali vyslovujeme znělé nebo neznělé tóny, hlasivky kmitají nebo jsou zavřené.

Když hlasivky kmitají, tvoří se periodický signál a vzniká tón řeči. Tento tón má každý člověk jiný, podle něj například poznáme, jestli nám po telefonu volá kamarádka nebo dědeček, protože jejich hlasy znějí různě. Různá znělost je způsobena různou frekvencí jejich základního tónu hlasu. U mužů se tato frekvence pohybuje kolem 100 Hz, u žen kolem 200 Hz a děti mívají tuto frekvencí nejvyšší — třeba až 350 Hz.

Naopak když jsou hlasivky otevřené, vzniká šum. Tento šum slyšíme, když vyslovujeme neznělé hlásky jako je například *s*, *š*, *r*, *ř*, atd.

Tento tón nebo šum vycházející z hlasivek však nestačí k tomu, abychom se dokázali dorozumět. K přeměně z těchto zvuků na vyslovitelné hlásky dochází v hlasovém neboli artikulačním traktu. Ten je tvořen hltanem, měkkým patrem, jazykem, ústní a nosní dutinou, zuby a rty. Všechny tyto části vědomě i nevědomě zapojujeme, když artikulujeme a vyslovujeme hlásky. V artikulačním traktu se během těchto pohybů tvoří rezonanční frekvence — formanty — ty formují spektrální obálku řeči. V závislosti na struktuře těchto formantů se dá rozpoznat, jakou hlásku, nebo přesněji foném, zrovna vyslovujeme. Což je důležité samozřejmě zejména u rozpoznávání řeči. Dá se ale tvrdit, že se tyto poznatky dají využít i při detekci hlasové aktivity.

Obě tyto části hlasového ústrojí se dají modelovat na počítači. Hlasivky generátorem impulsů nebo generátorem šumu a artikulační trakt lineárním filtrem. Výsledný signál v časové oblasti je konvolucí budícího signálu — jednotkových impulsů — a impulsní odezvy artikulačního traktu. Ve frekvenční oblasti se jedná o součin spektra jednotkového impulzu a spektra impulzní odezvy lineárního filtru. Výsledkem je spektrum řeči, a protože řeč je nestacionární, zajímá nás průběh tohoto spektra — nebo přesněji spektrální hustoty výkonu — v čase.

Pro další zkoumání charakteristik řeči je ale vhodné mít tyto dvě části spektra oddělené. K oddělení spektra buzení od spektra artikulačního traktu, které je pro nás zajímavější, nám slouží cepstrum. Cepstrum dokáže rozdělit operaci konvoluce pomocí nelineárních operací — konkrétně zpětnou Fourierovou transformací logaritmu spektrální hustoty výkonu. Výsledkem je součet cepstrálních koeficientů. První část cepstrálních koeficientů patří k buzení a ta zbylá část k filtru.

2.2 Techniky pro modelování řeči

V systémech pro zpracování řeči lze využít různé metody strojového učení. V první řadě je možné je rozdělit na metody s učitelem, *angl. supervised learning*, nebo bez učitele, *angl. unsupervised learning*. V této kapitole budou představeny ty nejčastěji používané a těmi jsou Gaussovské modely a Neuronové sítě. Obojí patří do skupiny metod učení s učitelem.

Gaussovské modely

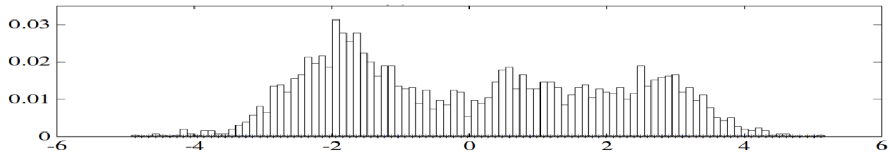
Gaussovské rozložení je nejběžnějším rozložením pravděpodobnosti a můžeme se s ním setkat při sledování různých jevů vyskytujících se v přírodě, mezi lidmi atd. Příkladem může být rozložení IQ, výšky, či váhy ve společnosti [5]. Gaussovský model je pak parametrický model, který toto rozložení modeluje, a který lze popsat pomocí dvou parametrů — střední hodnoty a rozptylu.

Parametry těchto modelů se odhadují z trénovacích dat patřící do stejné kategorie. V kontextu zpracování řeči se jedná o pravděpodobnostní rozložení dat patřící ke stejné klasifikační třídě. V případě detekce řečové aktivity si můžeme představit Gaussovské rozdělení s určitými parametry příznaků vygenerovaných z úseků nahrávky obsahující řeč a Gaussovské rozdělení s jinými parametry příznaků vyextrahovaných z dat, kde se řeč nevyskytuje nebo není dominantní.

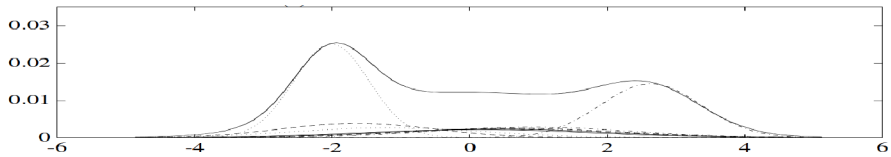
V realitě však nemáme pouze jeden příznak, ale máme jich více, tudíž potřebujeme vícerozměrné Gaussovské rozložení, kde se z jednorozměrných parametrů stanou také parametry vícerozměrné — ze střední hodnoty vektor středních hodnot a z rozptylu kovarianční

matice. Také většinou nestačí data modelovat pouze pomocí jednoho Gaussovského rozložení, ale používá se Směs Gaussovských rozložení, *zkr. GMM z angl. Gaussian Mixture Models* [26].

Směs Gaussovských rozložení představuje vážený součet několika Gaussovských rozložení. Používá se pro modelování pravděpodobnostního rozdělení příznaků, například těch vyextrahovaných ze spektra hlasového traktu. Celá směs Gaussovských rozdělení je reprezentována vektory středních hodnot, kovariančními maticemi a váhami pro jednotlivá Gaussovská rozložení. Směs Gaussovských rozložení umožňuje například modelování histogramu hodnoty konkrétního cepstrálního koeficientu extrahovaného z řeči člověka po dobu několika sekund. Příklad takového histogramu můžete vidět na obrázku 2.1 a k němu příslušnou Směs Gaussovských rozložení na obrázku 2.2. [26]



Obrázek 2.1: Histogram hodnot cepstrálního koeficientu extrahovaného z 25s řeči.



Obrázek 2.2: Směs Gaussovských rozložení modelující histogram hodnot příznaku zobrazeného na obrázku 2.1. Oba obrázky jsou převzaty z [26].

Na obrázku 2.2 můžeme vidět, že Směs Gaussovských rozložení umožňuje modelovat jakési skryté podtřídy. Těmi podtřídami se rozumí právě ona Gaussovská rozložení, jimiž je směs tvořena. Můžeme například mluvit o jedinečných konfiguracích hlasového traktu, ke kterým dochází při vyslovování samohlásek, nosovek nebo frikativ, kde každá z těchto skupin pak tvoří jedno nebo více Gaussovských rozložení v konkrétní směsi. Více informací lze dohledat v [26].

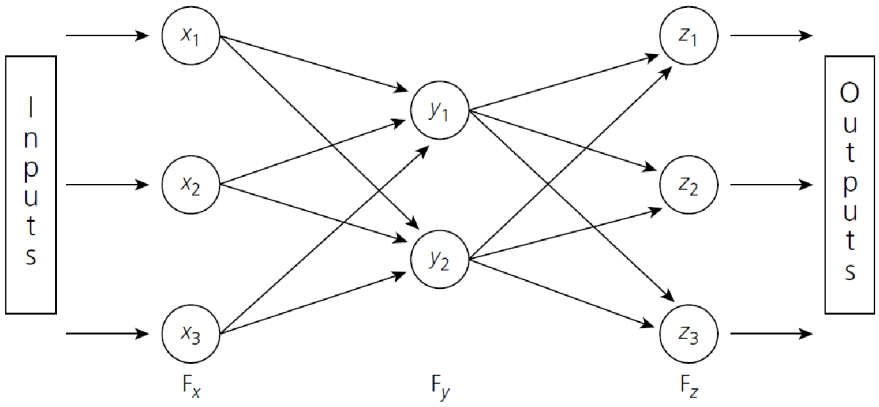
Parametry těchto modelů se odhadují pomocí speciálních algoritmů z trénovacích dat [26], ve výsledku je potom tolik směsí Gaussovských rozložení, kolik je v klasifikační úloze tříd. Klasifikace pak následně probíhá vyhodnocením, s jakou pravděpodobností ona vstupní data sedí pod onu konkrétní Směs Gaussovských rozložení.

Neuronové sítě

Dalším často používaným modelem při zpracovávání řeči jsou neuronové sítě. Na neuronovou síť je možno nahlížet jako na černou skříňku, *z angl. black box*, která přijímá vstupy a produkuje výstupy. Přesněji mapuje vstupní vektory na vektor výstupní. Jednou z možností pro využití takové sítě je právě klasifikace. Vstupní vektory v tom případě představují vstupní data — přesněji předpřipravené příznaky — a výstupní vektor udává příslušnost oněch dat ke konkrétní klasifikační třídě. [4]

Klasická neuronová síť se sestává z několika vrstev: Ze vstupní vrstvy, ze skrytých vrstev a z vrstvy výstupní. Každá z těchto vrstev je složena z několika zpracovávacích jednotek

— dále neuronů — a tyto neurony jsou v rámci jednotlivých vrstev propojeny váhovanými spoji. Počet těchto vrstev a neuronů v nich udává topologie sítě. Každý neuron zpracovává informace ze svých vstupů, provádí nad nimi předdefinovanou matematickou operaci a následně produkuje výstup pro neurony v další vrstvě, nebo část výstupu ze sítě celé [4]. Na obrázku 2.3 je možné vidět příklad neuronové sítě obsahující jednu skrytou vrstvu označenou jako *Fy*.



Obrázek 2.3: Možná topologie neuronové sítě se vstupní vrstvou (Fx), jednou skrytou vrstvou (Fy) a výstupní vrstvou(Fz), obrázek je převzat z [4].

2.3 Současný stav detekce řečové aktivity

Následující kapitola pojednává o tom, čeho bylo v poslední době dosaženo v oblasti detekce řečové aktivity. Nejprve jsou diskutovány různé příznakové sady, které byly pro tento účel testovány a dále pak celkové skladby systémů, které jsou pro detekci řečové aktivity používány. Pozornost bude také směřována na příznaky adresující problém robustnosti systému v rušivém prostředí a v prostředí s hudbou.

Rozbor příznaků používaných pro detekci řečové aktivity

Od příznaků čekáme, aby byly diskriminativní pro různé třídy — tzn. aby hodnoty příznaků pro řeč byli odlišné od hodnot příznaků extrahovaných z šumu na pozadí nebo čehokoli jiného, rušivého a přítomného v nahrávce.

Tyto příznaky se pak dají různě kombinovat — například lineárními kombinacemi, kde váhy jsou trénovány na základě minimální chyby. Nebo pokročilejšími technikami jako je Analýza hlavních komponent nebo Lineárně diskriminativní analýza. Všechny tyto metody mají za cíl snížení paměťové náročnosti a efektivnější využití potenciálu vybraných příznaků [3].

Nejprve lze uvažovat příznaky související s **energií zvuku**. Prvním identifikátorem toho, zdali je řeč přítomna, je hodnota krátkodobé energie signálu. Z této hodnoty zaručeně poznáme, jestli je v signálu ticho. Nepoznáme sice, jestli je v signálu přítomna řeč nebo hluk, protože jak jedno, tak druhé má vysokou energii, avšak tato hodnota nám může sloužit alespoň k vyfiltrování rámců, kde je naprosté ticho.

Příznaky odrážející harmonickou strukturu řeči

Z informací, které víme o tvoření řeči (kapitola 2.1), je možné tvrdit, že příznaky, které reprezentují harmonickou strukturu, jsou spolehlivým indikátorem přítomnosti řeči. Avšak nemůžeme čekat, že nám tyto harmonické příznaky sami o sobě objeví hlásky neznělé, které produkují místo harmonického tónu pouze šum. Naopak hudba, která je složena z velkého množství harmonických signálů — tónů — by mohla být těmito příznaky chybně označena jako řeč [9]. Tento problém by ale mohl být teoreticky řešitelný stanovením určitých prahů harmonicity, které by určily, zdali se ještě jedná o řeč, nebo už je přítomna hudba.

Pro určení znělosti řeči existuje množství příznaků, lišící se komplexitou. Takovým základním příznakem nízké úrovně komplexity je **počet průchodů nulou**, *zkr. ZCR z angl. Zero-crossing rate* [3]. Pro neznělou řeč je hodnota tohoto příznaku vyšší, než pro řeč znělou. Vysoká hodnota ZCR tedy může identifikovat neznělé rámce. Inverzí tohoto příznaku a jeho následnou kombinací s výše zmíněnou krátkodobou energií jsme schopni získat relativně dobré měřítko úrovně znělosti. Pro znělou řeč je hodnota tohoto kombinovaného příznaku vysoká, pro neznělou naopak nízká. Příznak, který dává podobné výsledky jako ZCR je **spektrální entropie signálu** [3].

Dále jsou sadou vhodných příznaků **hodnoty autokorelační funkce** [15]. Autokorelační funkce zachycuje harmonickou strukturu řečového signálu, ale také odráží vlastnosti artikulačního traktu. Pro periodické signály je hodnota autokorelační funkce maximální. Nicméně jsou situace, kde vysoké hodnoty tohoto příznaku způsobí šum, tudíž se na tento příznak nedá jednoznačně spoléhat.

Vhodný příznak tohoto charakteru je **cepstrální špička**, *angl. cepstral peak* [9]. Pro znělou řeč je detekce dobrá, a dokonce je z tohoto příznaku možné rozpoznat i řeč neznělou. Další příznak sloužící jako harmonický ukazatel je **polynomiální spektrální kontrast**, *angl. polynomial shape spectral contrast* [16]. Poměr délek samohlásek a souhlásek u řeči je obecně nižší než u zpěvu, což vede k nižší dlouhodobé periodicitě u řeči. Spektrální kontrast tak umožňuje přinejmenším do určité míry rozlišit zpěv od řeči.

Příznaky vhodné pro rušivé prostředí

Inspirací, jak přistupovat k detekci řečové aktivity v zašuměném prostředí je článek zabývající se odhalováním řeči ve filmech [16]. Filmy jsou obecně dobrým testovacím materiálem pro detekci řečové aktivity. Obsahují velké množství zvukových projevů, které klasifikujeme jako neřečové — střelba, šustění větru, déšť, doprava, často také hudba na pozadí nebo jako výplň mezi promluvami a další. Co se týče řeči, ta je ve filmech taky velice rozmanitá — různé emoční stavy mluvčích, od zloby přes smutek až například ke strachu a dále také řev, normální mluva a šeptání. To všechno jsou aspekty, které přímo ovlivňují korektnost systémů pro detekci řečové aktivity. [16]

Pro detekci řeči s nestacionárním šumem na pozadí navrhuje tento článek použití následujících setů příznaků — první, odrážející **vývoj řečového signálu v čase**. A druhý, již výše zmíněný polynomiální spektrální kontrast.

Jako další příznak pro rozlišení řeči na popředí od řeči nebo šumu na pozadí byl navrhnut příznak odrážející periodicitu řečového signálu filtrovaného metodou nulového kmitočtu, *angl. zero-frequency filtering*. Hodnota tohoto příznaku pro řeč na popředí je ve srovnání s řečí nebo šumem na pozadí vyšší. V tomto případě je ale důležité extrahovat příznak v závislosti na průměrné frekvenci základního tónu řeči. Více o tomto příznaku lze dohledat v [2].

Příznaky pro separaci hudby a řeči

Článek zabývající se rozlišováním mezi hudbou a mluveným slovem odrazuje od použití čistě hudebních příznaků. Tento argument má své důvody. Například skutečnost, že hudební signál není tak jednoduché generalizovat — existují různé druhy zvuků, melodie, žánry, ale také různé zdroje zvuku. Zatímco řeč pochází vždy z jednoho zdroje a tím je člověk. Obecně jsme schopní řečový signál uchopit lépe, protože bylo v této oblasti již provedeno mnoho výzkumů a jsme dobře obeznámeni s hlasovým traktem a celým řečovým ústrojím člověka. [14]

Příznaky odpovídající těmto návrhům prezentované v [14] jsou:

- **Spektrální těžiště** nebo centrum spektra, *angl. spectral centroid* [17]. Spektrální těžiště naznačuje, v jaké frekvenci se nachází největší část spektra a je definováno jako frekvenčně vážený součet energie spektra normalizovaný jejím neváhovaným součtem.
- **Spektrální roll-off koeficient** [17]. Spektrální roll-off koeficient je měřítkem šířky frekvenčního pásma analyzovaného rámce. Udává frekvenci, pod kterou je akumulováno určité procento daného signálu (obvykle 85-95 %).
- **Normalizovaná autokorelační funkce Zero-frequency filtrovaného řečového signálu** [2], protože tato autokorelační funkce umožňuje reprezentovat kvazi-periodickou povahu budícího signálu hlasivek, která je unikátní právě pro řeč.
- **Procento rámců s nízkou energií**.

Mimo výše zmíněných se často do systému pro detekci řečové aktivity zahrnují i následující příznaky: **Mel-frekvenční cepstrální koeficienty** [9], **spektrální tok**, *angl. Spectral Flux* [14] a také **sklon** (*angl. skewness*) a **kurtóza** (*angl. kurtosis*) **lineární predikce** [9].

Shrnutí klasifikačních modelů a metod pro detekci řečové aktivity

Typické klasifikační metody zahrnují jednodušší přístupy — například statistické modely s pravidly pro rozhodování o jakou třídu se jedná [31] — a také přístupy obsahující komplexnější modely jako jsou neuronové sítě [8], směsi Gaussovských rozložení [22, 6], Support Vector Machines [21, 13], skryté Markovovy modely [28] a další. Tyto klasifikační modely patří do skupiny metod učení s učitelem, což znamená, že prerekvizitou jsou data obsahující evaluace (informace o tom, do které klasifikační třídy patří). Se zvyšující se komplexitou modelu je potřeba i rozsáhlejší datová sada, a tudíž úkol hodnocení dat začíná být nelehký. Z toho důvodu se výzkumníci přiklánějí i k metodám učení bez učitele [11, 27], u kterých potřeba evaluací odpadá. Rozsáhlejší systémy jsou pak případně tvořeny kombinací dříve zmíněných.

Systém popsaný v [6] používá kombinaci metod učení bez učitele i s učitelem. Jedná se o systém robustní vůči šumu a skládající se ze dvou Směsí Gaussovských rozložení, *dále jako GMM*, reprezentující řečové a neřečové příznaky pro trénovací sadu. Jak bylo popsáno v kapitole 2.2, klasifikace pak určí hodnoty pravděpodobnosti pro GMM reprezentující řeč a GMM reprezentující rámce bez dominantní řeči. Následně se nad těmito pravděpodobnostními hodnotami provede Complete-linkage algoritmus, který zastoupí funkci tvrdého prahování, které je přítomno ve většině podobných systémů. [6]

Další případ, u kterého odpadá nutnost volby nebo nastavování prahů je popsán v [20]. Takový přístup je výhodný, protože hodnota těchto prahů je nestálá a odvislá od okolních vlivů.

Když klasifikujeme řečovou aktivitu, je výhodné mít k dispozici nějaký kontext. Jedním z hlavních limitů prvních detektorů řečové aktivity bylo, že algoritmy braly v potaz pouze aktuální rámec pro rozhodnutí, zdali se jedná o řeč nebo o šum. To je velice těžké, protože klasifikátor vidí pouze např. 20 ms řeči. Později však bylo zjištěno, že informace z minulosti může dost napomoci úspěšnosti systému [18]. Z počátku se tento problém řešil až po zpracování již klasifikovaných dat a to tak, že bylo provedeno vyhlazení výstupních dat, například pomocí mediánového filtru o určité velikosti. Tím bylo možné eliminovat chyby, které se vyskytovaly samostatně, ale bohužel tím mohlo docházet ke zvětšení chyb, které se vyskytly ve větším shluku.

Další možností, jak zahrnout kontextovou informaci je rozšíření vstupních dat tak, že jsou ke konkrétnímu rámci přidány i rámce z okolí, to však bohužel vede k velkému nárustu paměťové náročnosti. Právě tento problém se snaží adresovat v [12] za pomocí Rekurentních neuronových sítí, *dále jen RNN z angl. Recurrent Neural Network*. Hlavním rozdílem RNN oproti obyčejné neuronové síti je, že jednotlivé neurony mohou dostávat informace od ostatních uzlů nejen v aktuálním čase, ale i se zpožděním, tím je budována paměť systému a síť má díky tomu přístup k minulým hodnotám. RNN jsou příbuzné filtrům s nekonečnou impulsní odezvou (IIR filtry). [12]

Velmi komplexní model je použit v [8], kde pracují i s různými kanály pro přenos řeči. Tvoří jej 5 subsystémů, kde každý subsystém pracuje s fúzí tří typů příznaků: Akustické, hlasové a charakteristiky kanálu. Všechny subsystémy používají neuronové sítě se třemi skrytými vrstvami obsahující 500 neuronů. Používají zde taktiku pro spojení sousedních rámců pro vytvoření kontextu, jehož velikost je 31 rámců. Po výstupu z neuronové sítě používají průměrovací filtr o velikosti 41 rámců a k tomu navíc přidávají ke každému úseku klasifikovaném jako řeč ohraničení o velikosti 0.3 sekundy.

Jelikož jde o systém pro evaluaci, je přesně zadáno, jaká je cena chyby — větší váha je kladena na neopomenutí řeči. Pro určení prahů používají metodu nazvanou Test Adaptive Calibration [8], skládající se ze dvou kroků. Prvním krokem je nasazení výstupních hodnot z neuronové sítě na GMM, ve kterém všechny dílčí Gaussovské modely mají stejnou hodnotu rozptylu. A ve druhém kroku určí hodnotu prahu právě v závislosti na tom, aby bylo zamezeno opomenutí řeči. K tomu využívají techniku učení bez učitele, blíže popsanou v [8].

Kapitola 3

Výběr a příprava dat

Jak již bylo zmíněno v kapitole 2, pro klasifikační úkoly je volba vhodného datasetu zcela zásadní. Od dat bychom měli očekávat dostatečnou velikost, fonetickou vyváženost, přítomnost spontánních projevů a obecně dost variability, aby výsledný model byl schopný generalizace.

Po analýze různých data setů byla na základě předpokládané užitečnosti pro trénování robustního detektoru řečové aktivity vybrána následující kombinace databází.

1. Data set z rádiového vysílání Voice of America[1] (dále jen VOA) — obsahující zejména zpravodajství, hudbu a telefonní hovory v 16 kHz kvalitě. Mluveným jazykem je převážně angličtina. Celková velikost data setu činí 72 hodin, jednotlivé nahrávky trvají jednu hodinu.

2. A **Mu**sic, **S**peech, and **N**oise Corpus (dále jen MUSAN) [30] — obsahující nahrávky řeči, hudby a různých typů neřečových projevů — obecně označovaných jako hluk — v 16 kHz kvalitě. Hudební nahrávky obsahují různé hudební styly. Nahrávky obsahující řeč ji obsahují v 16–ti různých jazycích — angličtina, němčina, francouzština, italština, španělština, portugalština, ruština, čínština, japonština, polština, holandština, hebrejština, latina, maďarština, arabština a dánština. Celková velikost korpusu je 108.5 hodin a velikost jednotlivých nahrávek se pohybuje v rozmezí jedné sekundy až 17–ti minut.

3. GTZAN Music/speech collection (dále jen GTZAN) [19] — obsahující nahrávky řeči v různých jazycích (angličtina, němčina, čínština, řečtina, hindština, srbština) a prostředích a hudby různých hudebních stylů v 22.05 kHz kvalitě. Celková velikost této kolekce je 64 minut a každá nahrávka je dlouhá přesně 30 sekund.

U výše zmíněných data setů se liší způsob doručení evaluací (referenčních výstupů). Zatímco první databáze (VOA) je evaluována na přesnost stovek milisekund, zbývající korpusy obsahují evaluace na úrovni jednotlivých nahrávek. Celá nahrávka o velikosti 1 s – 17 min je označena jako řeč, hudba nebo hluk. Do přípravy dat bylo tedy zahrnuto také sjednocení těchto evaluací tak, aby pro každých 25 milisekund s 10–ti milisekundovým překryvem existovala referenční hodnota, zdali se jedná o ticho, řeč, hudbu či hluk. Evaluace pro VOA data set obsahovaly doplňující informace blíže popsané v následující kapitole 3.1.

[1] Voice of America News — https://www.voanews.com/

3.1 Analýza datasetu VOA

Tabulka 3.1: Detailní analýza zastoupení řeči v trénovacím data setu VOA. *Dva nebo více hlasů mluvících přes sebe.

základní dělení třídy řeč	délka celkem [sekundy]	další dělení	délka celkem [sekundy]	počet segmentů	průměrná délka segmentu [s]
ČISTÁ	43242.22	řeč	32342.73	1073	41.42
		crosstalk*	1930.72	148	13.05
		telefon	8968.78	196	45.76
S HUDBOU	13140.41	řeč s hudbou na pozadí	12707.11	917	67.93
		řeč s hudbou se zpěvem na pozadí	255.19	21	12.15
		crosstalk s hudbou na pozadí	84.97	9	9.44
		telefon s hudbou na pozadí	93.14	10	9.31
SE ŠUMEM	10346.13	řeč se šumem na pozadí	10139.53	445	91.81
		crosstalk se šumem na pozadí	111.63	14	7.97
		telefon se šumem na pozadí	94.97	4	23.74

Tabulka 3.2: Detailní analýza zastoupení řeči v testovacím data setu VOA.

základní dělení třídy řeč	délka celkem [sekundy]	další dělení	délka celkem [sekundy]	počet segmentů	průměrná délka segmentu [s]
ČISTÁ	49497.93	řeč	39073.00	1027	38.05
		crosstalk	2087.72	117	17.84
		telefon	8337.21	292	28.55
S HUDBOU	11213.47	řeč s hudbou na pozadí	10252.36	800	50.58
		řeč s hudbou se zpěvem na pozadí	467.72	24	19.49
		crosstalk s hudbou na pozadí	59.61	27	2.21
		telefon s hudbou na pozadí	433.78	12	36.15
SE ŠUMEM	5589.00	řeč se šumem na pozadí	5499.68	249	22.09
		crosstalk se šumem na pozadí	87.51	11	7.96
		telefon se šumem na pozadí	1.81	1	1.81

Anotace pro data set VOA byly vytvořeny pro výzkumné účely výzkumnou skupinou pro zpracování řeči Speech@FIT na Fakultě informačních technologií Vysokého učení technického v Brně. Evaluace tohoto data setu obsahuje celkem 22 různých tříd. Z toho jedna třída označující duplicitní data různých kategorií byla nepoužitelná a hned z počátku odstraněna. Zbývajících 21 tříd se dalo rozdělit do čtyř hlavních kategorií: Hudba, řeč, šum

a ticho. Do kategorie hudba byly zařazeny části označené jako čistá hudba, hudba s hlukem na pozadí, silná hudba se slabou řečí na pozadí, silná hudba se slabou řečí a slabým hlukem na pozadí, hudební klipy a hudba se zpěvem. Do kategorie hluk byla zařazena také třída nepotřebný zvuk (useless sound). Do kategorie ticho připadly pouze části označené jako ticho. A nakonec do kategorie řeč připadlo tříd nejvíce. Konkrétně se jednalo o dílčí třídy spadající pod čistou řeč, pod řeč s hudbou na pozadí a pro řeč se šumem na pozadí. Jednotlivé třídy spadající pod tuto kategorii jsou uvedeny v tabulkách 3.1 a 3.2, kde jsou také vypsaná konkrétní čísla udávající celkové množství konkrétní třídy v data setu, počet úseků s touto třídou a průměrnou délku takového úseku. Právě na úsecích různých typů řeči s hudbou na pozadí bude následně vyhodnocována robustnost detektoru řečové aktivity vůči hudebnímu pozadí.

Tabulka 3.3: Zastoupení jednotlivých tříd data setu VOA v trénovací a testovací sadě.

	TRÉNOVACÍ SADA			**TESTOVACÍ SADA**		
	celková délka [hh:mm:ss]	počet segmentů	průměrná délka segmentu	celková délka [hh:mm:ss]	počet segmentů	průměrná délka segmentu
HUDBA	13:23:19	1022	47 s	12:44:54	1003	46 s
ŘEČ	18:32:08	2837	24 s	18:25:00	2560	26 s
HLUK	0:03:44	55	4 s	0:01:49	53	2 s
TICHO	0:58:02	287	12 s	0:54:20	397	8 s
Celkem	**32:57:13**	**4201**	**87 s**	**32:06:03**	**4013**	**82 s**

3.2 Rozdělení dat na testovací a trénovací sady

K přípravě dat patří také jejich rozdělení na trénovací a testovací sadu. Data set VOA byl rozdělen v poměru 50 % na trénovací sadu a 50 % na testovací sadu. Za prvé se jednalo o rovnoměrné rozdělení mezi hlavní třídy — hudba, řeč, hluk a ticho, ale také o do jisté míry rovnoměrné rozdělení tříd dílčích. Zastoupení hlavních tříd pro trénovací a testovací sadu je možné vidět v tabulce 3.3. Detailnější rozdělení tříd pak v již zmíněných tabulkách 3.1 a 3.2.

V případě data setu MUSAN se jednalo o procentuální rozdělení 70 % na trénování a 30 % na testování. Detailní statistiky rozdělení je možné vidět v tabulce 3.4. Při tomto rozdělování byly brány v úvahu anotace, které jsou součástí data setu a obsahují informace o hudebním žánru a zdali nahrávka obsahuje vokály. Pro každou dílčí třídu (hudba, řeč, zpěv) je zajištěno, že jedna celá sada nahrávek stejného typu se nevyskytuje v trénovací sadě, tudíž pak výsledky na sadě testovací bude možné považovat za výsledky na neviděných datech. Pro třídu hudba se jedná o skupinu fma-western-art, což je západní hudba obsahující klasické styly, u třídy řeč jsou to nahrávky z portálu veřejných audioknih LibriVox[2]) a u třídy hluk se jedná o zvuky z portálu SoundBible[3].

Posledním data setem je GTZAN, který byl celý přiřazen do testovacích dat, tudíž v rámci testování se bude jednat o zcela neviděná data, na kterých bude možné ověřit robustnost a schopnost generalizace vyvinutého systému.

[2]LibriVox, free public domain audiobooks — https://librivox.org/
[3]SoundBible.com, Free Sound Clips — http://soundbible.com/

Tabulka 3.4: Zastoupení jednotlivých tříd data setu MUSAN v trénovací a testovací sadě.

	TRÉNOVACÍ SADA			TESTOVACÍ SADA		
V — s vokály, BV — bez vokálů M — moderní styl, K — klasický styl	celková délka [hh:mm:ss]	počet nahrá-vek	průměrná délka nahrávky	celková délka [hh:mm:ss]	počet nahrá-vek	průměrná délka nahrávky
HUDBA						
fma (M, V)	5:57:04	38	564	2:26:19	90	98
fma-western-art (K, V)	0:00:00	0	0	5:51:06	93	227
hd-classical (K, BV)	5:54:32	75	284	0:00:00	0	0
jamendo (M, V)	10:07:29	152	240	4:22:38	65	242
rfm (M, BV)	5:24:27	102	191	2:33:15	45	204
Celkem	**27:23:32**	**367**	**269**	**15:13:18**	**293**	**187**
ŘEČ						
librivox	0:00:00	0	0	20:22:32	173	424
os-gov	40:04:22	253	570	0:00:00	0	0
Celkem	**40:04:22**	**253**	**570**	**20:22:32**	**173**	**424**
HLUK						
free-sound	4:22:08	641	25	1:32:34	202	27
sound-bible	0:00:00	0	0	0:18:51	87	13
Celkem	**4:22:08**	**641**	**25**	**1:51:25**	**289**	**23**
CELKEM	**71:50:02**	**1261**	**179**	**37:27:15**	**755**	**173**

3.3 Data pro robustnost

Později byla původní testovací data zašuměna různými zdroji hluku pro zjištění, jak robustně systém pracuje. Jako zdroje hluku byly použity nahrávky HVAC z veřejné databáze zvuků Freesound[4], obsahující různé zvuky ventilátorů a obecně kancelářového a průmyslového hluku a také spojené stovky náhodných konverzací, které používá výzkumná skupina Speech@FIT jako data pro vývoj svých systémů. Tyto zdroje hluku byly rozděleny na část, kterou byla zašuměna data testovací a na část, kterou byla následně zašuměná také data trénovací za účelem zvětšení a obohacení trénovací sady a potenciální zvýšení robustnosti.

Pro zašumění byl použit nástroj FaNT (Filtering and Noise Adding Tool)[5] a kombinace dat a hluku byly provedeny v poměru signálu ke hluku, *angl. signal-to-noise ratio*, dále jen SNR, v hodnotách 20 dB, 15 dB, 10 dB a 5 dB.

3.4 Evaluační metriky

Aby bylo možné výsledky experimentů efektivně srovnávat, je třeba si určit jednotnou metriku pro určování úspěšnosti. Pro účely této práce jsou používány metriky dvě — úspěšnost a konfuzní matice.

Úspěšnost, *angl. accuracy*, je jednou z nejčastěji používaných evaluačních metrik jak pro binární klasifikační problémy, tak pro klasifikátory pracující s více třídami. Je velmi

[4]Freesound —https://freesound.org/

[5]Filtering and Noise Adding Tool — http://dnt.kr.hsnr.de/index964b.html

jednoduchá na výpočet a při rovnoměrném rozdělení tříd dává okamžitý obrázek toho, jak je daný systém přesný [10]. Je dána poměrem počtu úspěšně klasifikovaných rámců a celkového počtu rámců.

Pro detailnější představu o schopnosti systému je také často uvedena **konfuzní matice**, což je tabulka, ze které lze vyčíst, jak jsou třídy mezi sebou vzájemně zaměňovány. Díky tomu je možné konstatovat, která třída je více problematická a která je naopak rozpoznávána s největší úspěšností. Příklad takové konfuzní matice je tabulka 3.5. Na diagonále z levého horního rohu leží úspěšnosti jednotlivých tříd a mimo diagonálu jsou hodnoty chybně klasifikovaných tříd.

Tabulka 3.5: Příklad konfuzní matice.

		Predikce		
		Třída A	Třída B	...
Reference	Třída A	**Procentuální vyjádření množství rámců, které mají jako referenční třídu třídu A a tato třída jim byla klasifikátorem přiřazena**	Procentuální vyjádření množství rámců, které mají jako referenční třídu třídu A, ale místo ní jim klasifikátor přiřadil třídu B.	...
	Třída B	Procentuální vyjádření množství rámců, které mají jako referenční třídu třídu B, ale místo ní jim klasifikátor přiřadil třídu A.	**Procentuální vyjádření množství rámců, které mají jako referenční třídu třídu B a tato třída jim byla klasifikátorem přiřazena**	...
	...	...	...	...

Kapitola 4

Návrh a implementace robustního detektoru řečové aktivity

V následující kapitole bude přednesena analýza cílů a návrh na postup práce rozdělený do jednotlivých kroků. Dále se v této kapitole čtenář dozví o výběru implementačního jazyka a budou lehce popsány použité pomocné knihovny a nástroje. Dále budou popsány detaily některých zajímavých kroků při vývoji systému, a nakonec budou popsány implementační detaily neuronové sítě.

Hlavním cílem je vytvořit detektor řečové aktivity, který bude schopný rozpoznat řeč v různých jazycích, a to jak v zašuměném prostředí, tak v prostředí s hudbou na pozadí. Dalším cílem je rozpoznání více tříd, než je pouze řeč a vše ostatní. Zde navržený systém bude rozpoznávat třídy řeč, hudba, hluk a ticho. Dalším požadavkem je jednoduchost použití.

Takto vytvořený detektor bude možné využít pro rozmanité výzkumné účely v oblastech jako je rozpoznávání řeči, identifikace mluvčího, identifikace jazyka atd. Primárně výzkumnou skupinou Speech@FIT Vysokého učení technického v Brně místo dosud používaného fonémového rozpoznávače [29]. Tento fonémový rozpoznávač si totiž neumí poradit s hudbou, jak je následně i experimentálně prokázáno v kapitole 5.3. Vzhledem ke stoupající tendenci využívání audio nahrávek převzatých z kanálu youtube.com, je tento požadavek na schopnost nejen rozpoznat řeč na hudebním pozadí, ale také hudbu jako takovou, velmi aktuální. Nástroj pro detekci řečové aktivity bude možné využít také samostatně. Například při policejním vyšetřování je někdy nutné poslouchat dlouhé záznamy zvuku, kde řeč je přítomna pouze v některých úsecích. Použití detektoru řečové aktivity na tyto záznamy by značně ulehčilo práci a urychlilo proces vyšetřování.

V následujících sekcích bude podrobně popsán plán práce sestávající se z přípravy dat, výběru a konstrukce klasifikátoru, vývoje klasifikačního modelu, a nakonec zhodnocení použitelnosti systému.

4.1 Příprava dat

Prvním krokem je zpracování veškerých dat zahrnující jak data s nahrávkami, tak data s referenčními výstupy. Tuto práci lze dále dělit na:

1. **Rozdělení na trénovací a testovací sady**. Rozdělení musí být uděláno tak, aby vyhodnocení úspěšnosti na testovací sadě bylo vypovídající a bylo možné tvrdit, že systém je schopný generalizace. Toho lze docílit vybráním skupin nahrávek, které

jsou určitým způsobem jedinečné a odlišné od nahrávek, vyskytujících se v trénovací sadě. Například zvolením jednoho datového setu, který je celý zahrnut pouze do testovací sady. Nebo má–li data set svoje vnitřní rozdělení do skupin, umístit některé z těchto skupin pouze do testovací části. Detailní popis rozdělení splňující tento návrh je v kapitole 3.

2. **Sjednocení a úprava referenčních výstupů**. Vzhledem k tomu, že data sety vybrané pro vývoj zkoumaného systému mají rozdílně vytvořeny referenční výstupy, je třeba tyto referenční výstupy sjednotit. K data setu VOA jsou doručeny evaluace ve XML formátu s označením časového úseku v milisekundách a příslušné třídy. Naopak u data setů MUSAN a GTZAN jde pouze o evaluace na úrovni nahrávky — buď je celá nahrávka označena jako řeč, nebo jako hudba, případně jako hluk. Oba tyto případy je nutné sjednotit do formátu pro cílové rozpoznání. Tím je rozpoznání na úrovni rámců.

3. **Extrakce příznaků**. Aby bylo možné části audia klasifikovat, je nejprve nutné z něj vypočítat příznaky, na jejichž základě pak bude klasifikační model provádět rozhodnutí. Před samotnou extrakcí je vhodné se signálem udělat pár operací, jako je odstranění stejnosměrné složky a preemfáze. Dále se signál navzorkuje a vzorky se seskupí do překrývajících se rámců. Z rámců pak probíhá extrakce tak, že z každého rámce je vypočtena jedna hodnota příslušného příznaku. V aplikacích pracujících s řečovými signály se většinou používá délka rámce 25 ms s překryvem o délce 10 ms [34]. Tyto hodnoty budou použity také. Na základě informací získaných v literatuře jsem se rozhodla použít následující příznaky:

   - Absolutní hodnota amplitudy signálu — pro vyfiltrování úplného ticha.
   - Mel frekvenční cepstrální koeficienty, dále jen MFCC, 13 koeficientů.
   - Počet průchodů nulou z důvodů uvedených v 2.3 *(Zero–crossing rate)*.
   - Energie signálu.
   - Entropie energie.
   - Spektrální tok *(Spectral Flux)* a spektrální těžiště *(Spectral Centroid)*. Ráda bych navázala na [14], kde byly tyto příznaky testovány pouze na hudbě bez zpěvu, ale naznačují, že by mohly být užitečné i při aplikaci na hudbu se zpěvem.
   - Spektrální entropie *(Spectral Entropy)*.
   - Spektrální roll–off koeficient *(Spectral Rollof)*.
   - Spektrální rozpětí *(Spectral Spread)*.

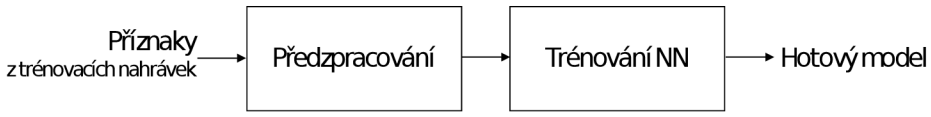
   Vzhledem k tomu, že vybraných příznaků není malé množství, pro začátek budou použity pouze MFCC příznaky spolu s absolutní hodnotou amplitudy signálu. Jelikož hlavním cílem je robustnost a komparativní analýza velkého množství příznaků uvedená v [9] uvádí pro tyto příznaky největší úspěšnost při detekci řečové aktivity v prostředí s různými úrovněmi šumu.

4.2 Výběr a konstrukce klasifikátoru

Dalším krokem je výběr klasifikačního modelu. Z nastudované literatury [22, 6, 8] se nabízejí dvě možnosti a těmi jsou Gaussovské modely nebo neuronové sítě. Dle mojí vlastní preference jsem se pro tento účel rozhodla použít neuronovou síť, která bude implementována

v jazyce Python ve verzi 2.7 [25] za použití knihovny TensorFlow ve stabilní verzi 1.13 [1]. TensorFlow je open source knihovna pro podporu implementace, trénování a vyhodnocování klasifikátorů pro strojové učení.

Aby bylo možné neuronovou síť použít pro vyhodnocení testovacích dat, je třeba ji nejprve natrénovat na trénovacích datech. V tomto okamžiku už budou připraveny příznaky pro trénovací i testovací sadu, tudíž následující postup se bude skládat z několika dílčích kroků: Nejprve dojde k předzpracování těchto příznaků — tím bude jejich normalizace a následně pospojování sousedních rámců pro získání kontextové informace. Takto předzpracovaná data již budou připravena na samotné trénování, které bude probíhat v cyklech — epochách — kdy během každé epochy projdou všechna trénovací data trénovaným modelem. Na základě jejich vyhodnocení budou upraveny parametry modelu tak, aby při příštím vyhodnocení byly výsledky blíže výsledkům očekávaným. Po několika epochách bude trénování ukončeno a výsledný model bude uložen a připraven na evaluaci dat testovacích. Schéma tohoto postupu je na obrázku 4.1. Implementačním detailům jednotlivých částí se budu věnovat v sekci 4.5.



Obrázek 4.1: Schéma návrhu trénování systému.

4.3 Vývoj klasifikačního modelu, experimenty

Trénování neuronové sítě má spoustu proměnných. Tyto proměnné je třeba identifikovat a experimentovat s jejich hodnotami, aby výsledný systém používal jejich optimální hodnoty a produkoval co nejpřesnější výsledky. Experimenty budou probíhat následujícím způsobem: Nejprve bude určená proměnná pro experimentování, následně bude provedeno trénování systému s různými hodnotami zvolené proměnné dle schématu 4.1. Výsledkem bude několik natrénovaných systémů. Systémy budou vyhodnoceny na testovacích sadách, jejich výsledky porovnány, a tím bude zjištěna optimální hodnota zkoumané proměnné. Jeden cyklus takového testování je vyobrazen na schématu 4.2.

Experimenty budou rozděleny do tří hlavních kategorií:

1. Experimenty s proměnnými pro předzpracování dat.
2. Experimenty s proměnnými, které vystupují během trénování klasifikačního modelu.
3. Experimenty s proměnnými upravující výstupy z klasifikačního modelu.

Obrázek 4.2: Schéma návrhu vyhodnocení testovacích dat.

Vývoj robustnosti

Na celkovou robustnost systému jsou kladeny tři hlavní požadavky:

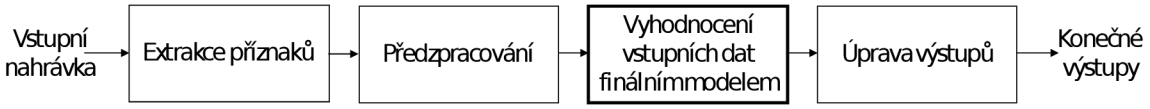
1. Detekce řeči v různých jazycích.
2. Detekce řeči v prostředí s různou úrovní šumu.
3. Detekce řeči na hudebním pozadí — zahrnující instrumentální i vokální hudbu.

Na základě výsledků získaných během experimentování bude zvolen model, na kterém budou vyhodnocena data obsahující (a) řeč v různých jazycích, (b) zašuměnou řeč, (c) řeč s hudbou na pozadí. Na základě dosažené úrovně úspěšnosti budou podniknuty kroky k v vylepšení těchto výsledků.

Vyhodnocení experimentů, konstrukce finálního systému

Nakonec budou vzaty v úvahu optimální hodnoty proměnných nalezené během experimentování a také kroky podniknuté během budování robustnosti a bude natrénován finální systém, u něhož bude kladen důraz na vyvarování se přetrénování. Přetrénování lze předcházet různými regularizačními metodami jako je například brzké zastavení, *angl. early stopping* [24], nebo vypouštěním neuronů ve skrytých vrstvách neuronové sítě, *angl. dropout* [33].

V momentu, kdy bude takový systém natrénován, bude zakomponován do programu, který bude jako vstup přijímat nahrávky v určitém formátu, následně z nahrávky extrahuje příznaky, které připraví na vstup do neuronové sítě. Neuronová síť provede evaluaci těchto dat, výstupy budou dodatečně upraveny a zapsány do výstupního souboru. Schéma dílčích kroků finálního systému je znázorněno na následujícím schématu 4.3.



Obrázek 4.3: Schéma návrhu finálního systému.

4.4 Zhodnocení použitelnosti systému

Posledním krokem bude zhodnocení použitelnosti finálního systému. Po celou dobu bude úspěšnost hodnocena porovnáváním referenčních výstupů s výstupy produkovanými systémem. Avšak pro tvrzení, že systém produkuje opravdu spolehlivé výsledky, je třeba provést ještě další pozorování, které proběhne manuálním vyhodnocením úspěšnosti vyprodukovaných výstupů. To znamená poslechem nahrávky a srovnáním výstupů s reálným obsahem nahrávky.

4.5 Implementační detaily

Jako implementační jazyk byl zvolen Python ve verzi 2.7 [25] a to kvůli dobré podpoře pro skriptování za účelem zpracovávání dat, popsané v kapitole 4.1. Pro jakékoli další manipulace s daty je Python taktéž vhodnou volbou. Většinově bude totiž pracováno s velkými maticemi čísel (příznaků). Pro maticové operace existuje v Pythonu velmi schopná knihovna Numpy [23].

Extrakce příznaků byla provedena pomocí nástroje pyAudioAnalysis [7]. Tento nástroj umožňuje extrahovat příznaky z navzorkovaného signálu a jelikož jde o open source projekt, lze si extrakci poupravit na míru. Před extrakcí je provedeno odstranění stejnosměrné složky signálu a preemfáze. Extrahované příznaky byly poté normalizovány na nulovou střední hodnotu a jednotkovou varianci, *angl. zero mean, unit variance*. Toho bylo dosaženo spočtením středních hodnot a směrodatných odchylek v rámci jednoho příznaku skrze jeden trénovací set a následným odečtením těchto hodnot v případě střední hodnoty a vydělením v případě směrodatné odchylky.

Vzhledem k tomu, že dat k trénování bylo poměrně dost na to, aby bylo možné je mít v jeden okamžik všechny načtené v paměti, bylo je potřeba rozdělit na několik trénovacích setů. Těchto setů bylo v první fázi vytvořeno 36, protože bylo nutné počítat s tím, že při spojování rámců k vytvoření kontextové informace dojde k nárustu dimenzionality dat. Trénování neuronové sítě pak probíhalo po těchto 36–ti částech, kdy jednotlivé sety byly načítány jeden po druhém. Jedna epocha pak znamenala projití všech těchto 36–ti data setů neuronovou sítí.

I když jednotlivé sety byly před prezentací neuronové síti promíchány, takto prováděné trénování nejprve dobré výsledky nepřinášelo. Dalo by se to přikládat faktu, že některé nahrávky byly dlouhé až 1 hodinu, a tím pádem bylo v jednom trénovacím setu mnoho dat podobného charakteru. To vedlo k nakalibrování neuronové sítě na ona data. Po načtení dalšího trénovacího setu zase pouze na data obsažená v setu dalším, a nakonec síť neměla schopnost generalizace. Z toho důvodu byla všechna data ještě před dalším zpracováním rozdělena do 10–ti sekundových úseků, které byly promíchány a následně opět spojeny do 36–ti trénovacích setů. Teď však již obsahující různorodá data. Tento krok přinesl výsledky, na kterých bylo možné systém dál budovat.

Implementační detaily neuronové sítě

Neuronová síť je implementována za pomoci knihovny TensorFlow. TensorFlow používá graf datového toku k reprezentaci operací jako jsou trénovací kroky, výpočty úspěšností a ztrát atd. Pro proměnné jsou vytvořena zástupná místa, *angl. placeholders*, která jsou plněna až později. Ke skutečnému vykonávání definovaných operací a naplnění zástupných míst pro proměnné nastává až po vytvoření sezení, *angl. session*. Program tedy neběží sekvenčně, ale nejprve je deklarativně vytvořen graf a jeho operace potom díky tomu mohou probíhat paralelně — je–li to možné. [1]

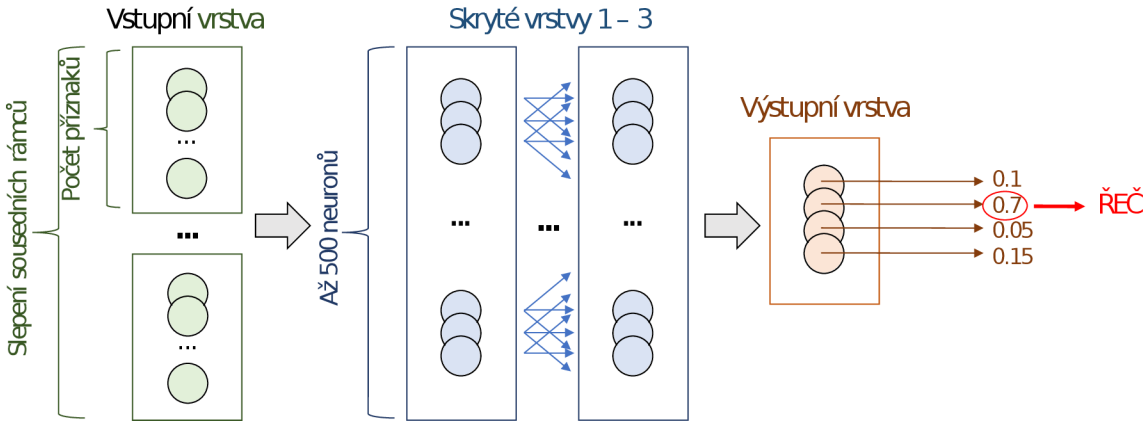
Vstupní vrstva neuronové sítě

Vstupem do neuronové sítě je vektor příznaků o různé velikosti. Tato velikost je závislá na počtu příznaků a na počtu a hustotě sjednocených sousedních rámců pro vytvoření kontextové informace. Dimenzionalita vstupních dat je dána následující formulí: $dimenzionalita = pocet\_priznaku * ((velikost\_kontextu)/s) * 2) + 1$, kde s je číslo udávající každý kolikátý rámec je zahrnut do vytvoření kontextové informace. Při přeskakování každého druhého rámce je $s = 2$.

Výstupní vrstva neuronové sítě

Neuronová síť obsahuje vstupní vrstvu o velikosti *dimenzionalita*, dále 1–3 plně propojených skrytých vrstev o velikostech v rozmezí $1/2 * dimenzionalita$ – 500 neuronů a nakonec výstupní vrstvou o velikosti počtu klasifikačních tříd. Výstupem neuronové sítě je v tomto

případě vektor o velikosti 4 — klasifikační třídy jsou řeč, hudba, hluk a ticho. Každý prvek tohoto vektoru představuje log–likelihood hodnotu pro jednu z klasifikačních tříd. Na tyto hodnoty je následně aplikována funkce softmax, která log–likelihood hodnoty převede na hodnoty pravděpodobnosti příslušnosti ke konkrétní třídě. Výsledná třída je získána aplikací funkce argmax na pravděpodobnostní hodnoty. Popsaná topologie sítě je znázorněna na schématu 4.4.



Obrázek 4.4: Schéma topologie neuronové sítě.

Trénování neuronové sítě

Neuronová síť je trénována po dávkách, *angl. batch*. Pro účely této práce je hodnota dávky 5 vstupních vzorků. Jeden trénovací krok je složen z představení těchto 5–ti vektorů neuronové síti a na základě porovnání referenčních výsledků s výstupy, které produkuje neuronová síť pro tyto vektory, je spočtena ztráta. Váhy neuronové sítě jsou poté upraveny tak, aby byla tato ztráta zmenšena.

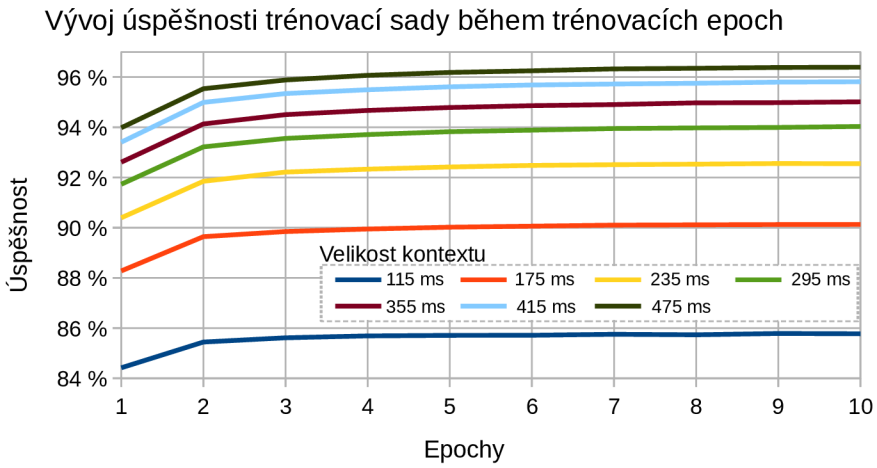
Jakmile neuronovou sítí projdou všechny dávky spadající do jednoho trénovacího setu, dojde k vyhodnocení úspěšnosti na tomto setu, set je uvolněn z paměti a je načten set další. Po zpracování všech setů je ukončena jedna epocha trénování a trénovaný model je uložen.

Kapitola 5

Experimenty se systémem pro detekci řečové aktivity

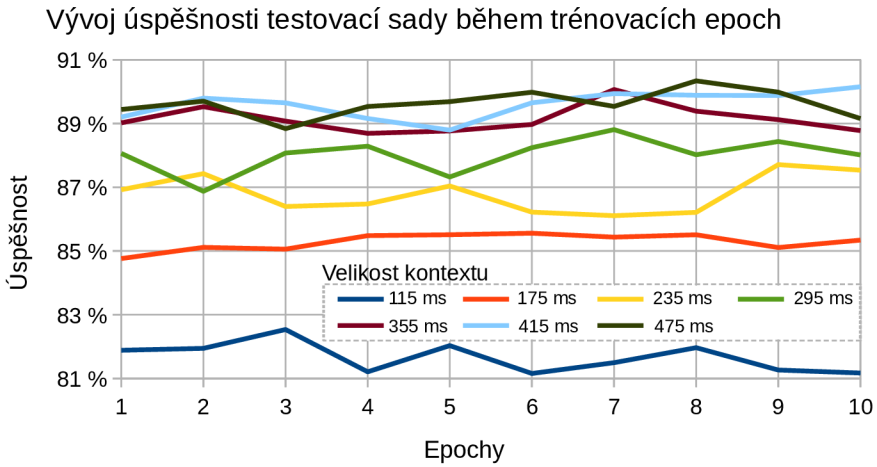
K vytvoření stabilního systému podávající solidní výsledky je třeba provést řadu experimentů, vyhodnotit jejich výsledky a na jejich základu systém formovat.

Vzhledem k tomu, že jako klasifikační model je použita neuronová síť, tak je pro ohraničení experimentů třeba vzít v úvahu, kolik je vhodné vykonat epoch pro dosažení srovnatelných výsledků. Tento počet byl pro většinu experimentů stanoven na 10. Po 10. epoše již k nijak zásadnímu zlepšení nedochází, proto je z časových důvodů trénovaní po 10. epoše vždy ukončeno. V některých experimentech, které byly časově více náročnější (sledování velikosti a počtu skrytých vrstev neuronové sítě) je tento počet epoch snížen na 5. V rámci jednoho typu experimentu je však počet epoch vždy jednotný, aby dosažené výsledky bylo možno porovnávat. Na obrázku 5.1 je například možné vidět vývoj úspěšnosti na trénovací sadě pro několik systémů (popsaných v kapitole 5.1) v průběhu zmíněných 10–ti epoch. Dále na obrázku 5.2 je možné vidět toto srovnání na sadě testovací.



Obrázek 5.1: Vývoj úspěšnosti systému během epoch — trénovací sada.

V následujících kapitolách je přednesen průběh a výsledky experimentů prováděných nad základním systémem pro detekci řečové aktivity, který jako klasifikační model používá neuronovou síť a byl naimplementován pro účely této práce. Nejprve bude diskutován časový kontext, dále postupy prováděné pro detekci ticha, následně zpracování výstupů z neuronové

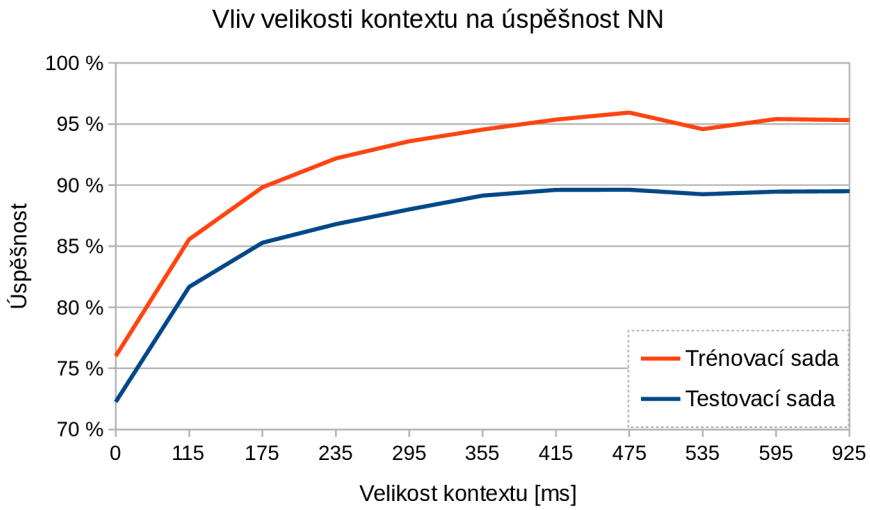


Obrázek 5.2: Vývoj úspěšnosti systému během epoch — testovací sada.

sítě a také experimenty s velikostí a počtem skrytých vrstev použité neuronové sítě. Nakonec bude konzultován vývoj robustnosti systému v multilingválním prostředí, v prostředí se šumem a v prostředí s hudebním pozadí.

5.1 Experimenty s časovým kontextem

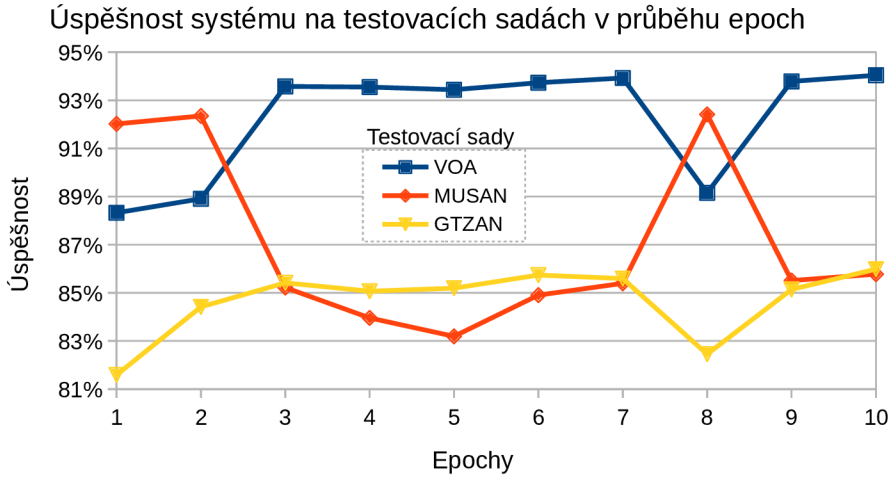
Klasická neuronová síť nemá žádný mechanismus pro uchování kontextové informace, proto před tím, než jsou data síti prezentována, je vhodné tuto kontextovou informaci vytvořit. Činí se tak spojováním sousedních rámců, čímž narůstá dimenzionalita vstupních dat. S tímto problémem lze však bojovat například přeskakováním některých rámců. Vzhledem k velikosti překryvu 10 ms z celkem 25 ms dlouhého rámce bylo nejčastěji zvoleno vypuštění každého druhého rámce. Bylo očekáváno, že na úspěšnost to mít moc velký vliv nebude, ale množství paměti, které tím bude ušetřené, bude veliké. Pro potvrzení této domněnky byl proveden experiment se systémy s velikostí kontextu 0.175 sekund (11 rámců) používající jak kontext úplný, tak kontext s vypuštěním každého druhého rámce. Systém používající všechny rámce byl pouze o 0.9 % úspěšnější.



Obrázek 5.3: Vliv velikosti kontextu na úspěšnost systému.

Cílem těchto experimentů je zjištění optimální velikosti časového kontextu s přihlédnutím k paměťové náročnosti.

Jak je možné vidět v grafu 5.3, optimální velikost kontextu pro systém je 0.415 sekund (27 rámců), tedy asi půl sekundy. Pokles u velikosti kontextu 0.535 sekund (35 rámců) lze přičítat tomu, že od této hodnoty byl spojován pouze každý třetí rámec a ne každý druhý jako tomu bylo do této hodnoty. Ke spojování pouze každého třetího rámce jsem musela přistoupit z důvodu paměťové náročnosti. Jak je pak ale možné vidět dále v grafu, toto zhoršení začíná byt kompenzováno se zvyšující se velikostí kontextu. Zaměříme-li se však na spodní křivku ukazující úspěšnost na testovacích datech, právě od velikosti kontextu kolem půl sekundy se úspěšnost systému již tak výrazně nezvětšuje, proto nemá smysl zvyšovat paměťovou náročnost nad tuto hodnotu.



Obrázek 5.4: Vývoj úspěšnosti během epoch.

Pro systém používající právě kontext o velikosti 27 rámců byla provedena následující hlubší analýza. V grafu 5.4 je možné vidět, jak se průběžně měnila úspěšnost na testovací sadě po jednotlivých trénovacích epochách. Z grafu je patrné, že různé datové sety mají různé nároky na parametry trénované sítě. Když vyroste úspěšnost některé z testovacích sad, úspěšnost těch dalších klesne a naopak.

V konfuzní matici 5.1 je možné vidět podrobnější analýzu, jak je systém schopný rozpoznávat jednotlivé třídy. Tabulka je rozdělena na tři části, kde každá z nich představuje úspěšnost na jiném testovacím data setu. Jak již bylo popsáno v sekci 3, ne u všech data setů jsou zastoupeny všechny klasifikační třídy (MUSAN, GTZAN). Zaměříme–li se blíže na klasifikační třídu pro ticho, vidíme, že systém ticho nerozpoznává u data setů MUSAN a GTZAN, což je dle referenčních výstupů správně, ale bohužel jej nerozpoznává ani u data setu VOA, kde by ale dle referenčních výstupů rozpoznáno být mělo. Tuto skutečnost lze přikládat stylu, kterými jsou referenční výstupy vytvořeny. U data setů MUSAN a GTZAN připadá jedna referenční třída celé nahrávce, zatímco u data setu VOA jsou referenční třídy přiřazovány datům na úseky desítek milisekund. Prvním pokusem pro řešení tohoto problému bylo zavedení tzv. preselektoru, který je popsán v následující kapitole.

Úspěšnosti rozpoznání ostatních tříd jsou uspokojivé, tudíž je tento systém dále braný jako referenční systém, se kterým jsou pak výsledky dalších vyvinutých systémů porovnávány.

Tabulka 5.1: Konfuzní matice pro systém používající kontext o velikosti 0.415 sekundy.

VOA

		Predikce			
		Ticho	Řeč	Hudba	Hluk
Reference	Ticho	**0.00 %**	94.85 %	4.65 %	0.51 %
	Řeč	0.00 %	**83.72 %**	15.60 %	0.68 %
	Hudba	0.00 %	3.25 %	**96.23 %**	0.52 %
	Hluk	0.00 %	42.70 %	52.16 %	**5.14 %**

MUSAN

		Predikce			
		Ticho	Řeč	Hudba	Hluk
Reference	Řeč	0.00 %	**94.74 %**	4.44 %	0.83 %
	Hudba	0.00 %	2.60 %	**96.05 %**	1.35 %
	Hluk	0.00 %	9.08 %	21.20 %	**69.72 %**

GTZAN

		Predikce			
		Ticho	Řeč	Hudba	Hluk
Reference	Řeč	0.00 %	**88.42 %**	9.76 %	1.81 %
	Hudba	0.00 %	9.26 %	**83.54 %**	7.20 %

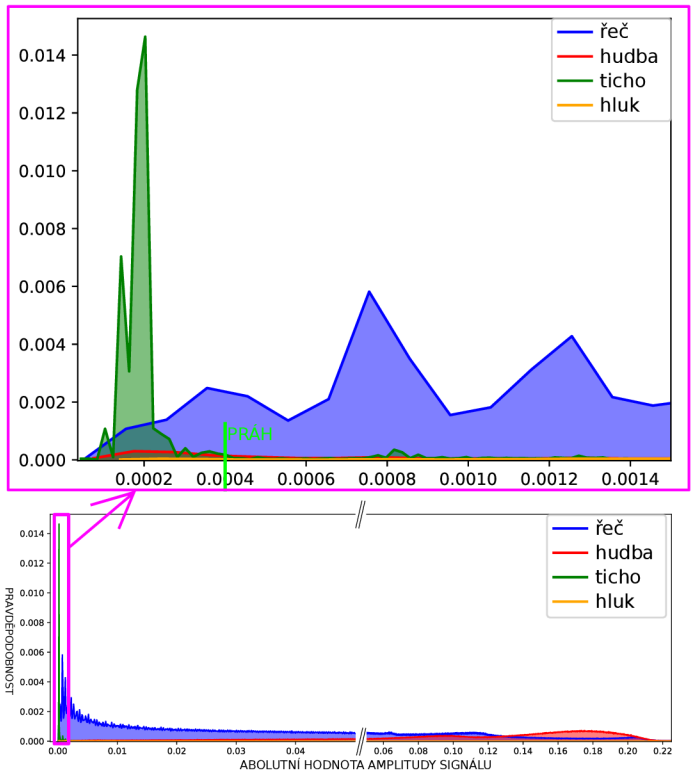
5.2 Preselektor

Cílem zavedení preselektoru je identifikace úplného ticha, tedy rámců, kde je amplituda nulová, nebo má velmi nízké hodnoty. Toho lze dosáhnout jednoduchou technikou prahování, kde hodnota prahu se určí podle histogramu trénovacích dat. V této kapitole bude popsáno, jaký byl postup k vytvoření takového preselektoru a jaké výsledky přineslo jeho začlenění do systému pro detekci řečové aktivity.

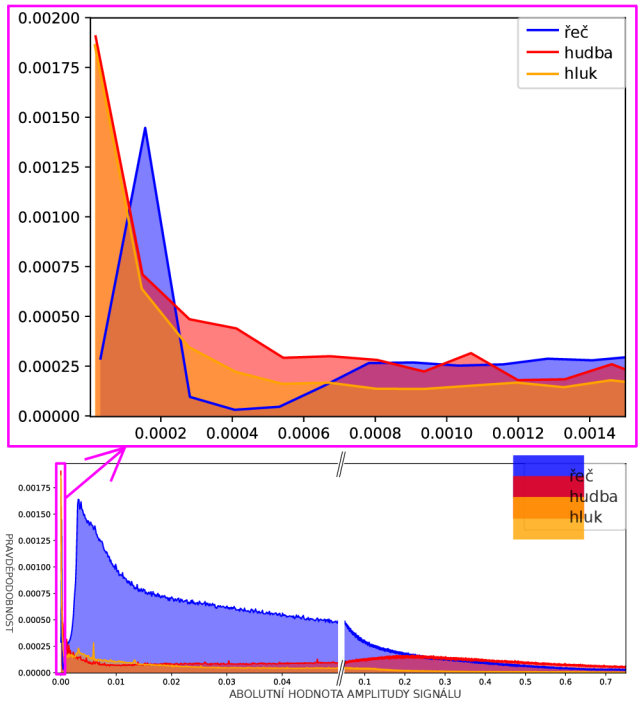
Pro odfiltrování ticha byl k základní sadě příznaků přidán příznak odrážející absolutní hodnotu amplitudy zpracovávaného signálu. Tento příznak byl získán následujícím způsobem: Vzorky zpracovávaného signálu byly sdruženy do rámců o velikosti 25 ms s překryvem o velikosti 10 ms. Následně byla z absolutních hodnot vzorků spadajících do jednoho rámce vybrána nejvyšší hodnota. Tato hodnota byla poté porovnána s prahem a na základě výsledku byla tomuto příznaku pro konkrétní rámec přiřazena jednička (pro hodnoty vyšší než hodnota prahu) nebo nula (pro hodnoty nižší než hodnota prahu).

Hodnota zmíněného prahu byla zvolena na základě histogramu jednotlivých tříd v trénovacích sadách data setů VOA a MUSAN. Tyto histogramy jsou k vidění na obrázcích 5.5 a 5.6. Histogram obou data setů je široký, pro účely nalezení vhodného prahu je však důležitá pouze malá část z počátku histogramu — tato část je zvětšena nad původní celý histogram a orámována růžovou barvou. U data setu VOA jsou rámce obsahující ticho velice pěkně nashromážděné pod hodnotou 0.0004 absolutní hodnoty amplitudy signálu. Množství řeči, které se nachází pod touto hodnotou společně s tichem, je zanedbatelné. Pro data set MUSAN je však situace jiná. Pod absolutní hodnotou amplitudy signálu 0.0004 je obsaženo nezanedbatelné množství jak řeči, tak hudby i hluku.

Výsledky, kterých systém dosáhl po aplikaci zmíněného prahu je možné vidět v konfuzní matici 5.2. Pro data set VOA došlo k očekávanému zlepšení rozpoznání ticha a to o 84 % absolutně. U data setu MUSAN ale bohužel došlo k poklesu úspěšnosti rozpoznání řeči o 4 % absolutně. Klasifikátor naopak rozpoznává ticho tam, kde dle referenčních výstupů ticho není. Je však velice pravděpodobné, že v data setech MUSAN i GTZAN se na nějaké



Obrázek 5.5: Histogram tříd data setu VOA.



Obrázek 5.6: Histogram tříd data setu MUSAN.

úseky milisekund ticho vyskytuje také. Vysvětlovalo by to i špičky na počátku histogramu pro data set MUSAN. Dá se tudíž předpokládat, že referenční výstupy nejsou zcela korektní a bylo by potřeba je opravit tak, aby i v data setech MUSAN a GTZAN bylo ticho korektně označeno.

Pro tento účel byl použit nástroj PhnRec [29] — Fonémový rozpoznávač založený na dlouhém kontextu, který je s úspěchem používán na detekci řeči ve výzkumné skupině Speech@FIT. Průběhu a výsledkům tohoto kroku je věnována následující kapitola.

Tabulka 5.2: Konfuzní matice pro systém používající preselektor.

VOA

		Predikce			
		Ticho	Řeč	Hudba	Hluk
Reference	Ticho	**84.38 %**	11.68 %	3.48 %	0.46 %
	Řeč	0.28 %	**83.46 %**	15.57 %	0.68 %
	Hudba	0.08 %	3.21 %	**96.19 %**	0.52 %
	Hluk	0.87 %	42.08 %	51.97 %	**5.08 %**

MUSAN

		Predikce			
		Ticho	Řeč	Hudba	Hluk
Reference	Řeč	4.01 %	**90.98 %**	4.25 %	0.75 %
	Hudba	0.67 %	2.59 %	**95.38 %**	1.35 %
	Hluk	2.20 %	8.71 %	20.70 %	**68.39 %**

GTZAN

		Predikce			
		Ticho	Řeč	Hudba	Hluk
Reference	Řeč	0.00 %	**88.42 %**	9.76 %	1.81 %
	Hudba	0.00 %	9.26 %	**83.54 %**	7.20 %

5.3 Vytvoření nových referenčních výstupů za pomoci nástroje PhnRec

Fonémový rozpoznávač PhnRec vyvinula výzkumná skupina Speech@FIT na Fakultě informačních technologií Vysokého učení technického v Brně. Tento nástroj slouží primárně pro výzkum, je jej možné použít na dílčí zpracování nahrávek a již byl úspěšně využit pro úkoly jako je detekce jazyka, indexace a hledání v nahrávkách a detekce klíčových slov. [29]

Výstupem z nástroje je soubor v HTK[1] formátu, který obsahuje jednotlivé časové intervaly a k nim příslušný popis, co se v daném intervalu vyskytuje. Popisem může být buď konkrétní foném (řeč), nebo označení pro pauzu (ticho), pro hluk v pozadí nebo neřečový projev mluvčího. Tento formát byl následně zpracován do stejného formátu jako jsou existující referenční výstupy. Tzn. označení ticha, řeči, hluku na pozadí nebo neřečového projevu mluvčího.

Ukázka takových výstupů pro konkrétní nahrávku obsahující hudbu z data setu MUSAN je v horní části obrázku 5.7. Z obrázku je patrné, že fonémový rozpoznávač si neumí poradit s hudbou, nemá pro ni žádnou konkrétní třídu, a tak je hudba rozpoznávána různě. V této konkrétní nahrávce jako řeč střídavě s neřečovým projevem mluvčího, u jiných nahrávek

[1] HT Speech Recognition Tool — http://htk.eng.cam.ac.uk/

se dokonce střídají třídy tři (řeč, hluk na pozadí a neřečové projevy). Srovnáme–li však graf v horní části obrázku s nahrávkou uprostřed obrázku vidíme, že ticho určuje fonémový rozpoznávač dobře — toto bylo ověřeno poslechnutím si většího počtu nahrávek.

Po uvážení všech těchto skutečností byly nové referenční výstupy vytvořeny následovně: Výstup z fonémového rozpoznávače převedený do formátu obsahující čtyři třídy (ticho, řeč, hluk pozadí a neřečový projev mluvčího) byl vyhlazený pomocí mediánového filtru o velikosti 775 ms — výsledek je znázorněn v horním grafu v obrázku 5.7. Pro další krok byly však použity pouze rámce obsahující ticho. Těmito rámci byly nahrazeny odpovídající rámce z původních referenčních výstupů. Ukázku je možné vidět ve spodním grafu obrázku 5.7. V grafu jsou zobrazeny i hodnoty preselektoru. Lze konstatovat, že použití fonémového rozpoznávače značně vylepšuje funkci preselektoru.

Od tohoto okamžiku byly výsledky získané nyní popsaným procesem brány jako nové referenční výstupy. Bylo tak učiněno pro všechny tři data sety — VOA, MUSAN a GTZAN.

Systém pracující s časovým kontextem o velikosti 415 ms z kapitoly 5.1, používající preselektor z kapitoly 5.2 byl natrénován na nové referenční výstupy a výsledky v podobě konfuzní matice je možné sledovat v tabulce 5.3. Ticho je rozpoznáváno u všech testovacích data setů s průměrnou 73 % úspěšností. Rozpoznání řeči u data setu MUSAN se vrátilo k původním 94 % — tak jako tomu bylo u prvního systému, který nepoužíval preselektor. Z toho se dá vyvodit, že nově vytvořené referenční výstupy odhalily ticho tam, kde jej předpovídal i preselektor. Úspěšnost rozpoznání řeči na data setu VOA klesla o 4 % absolutně, což je sice dost, ale vzhledem k tomu, že díky novým referenčním výstupům má systém lepší představu o tom, kde je ticho, jsou tyto ztráty tolerovatelné.

Tabulka 5.3: Konfuzní matice pro systém používající nové referenční výstupy.

VOA

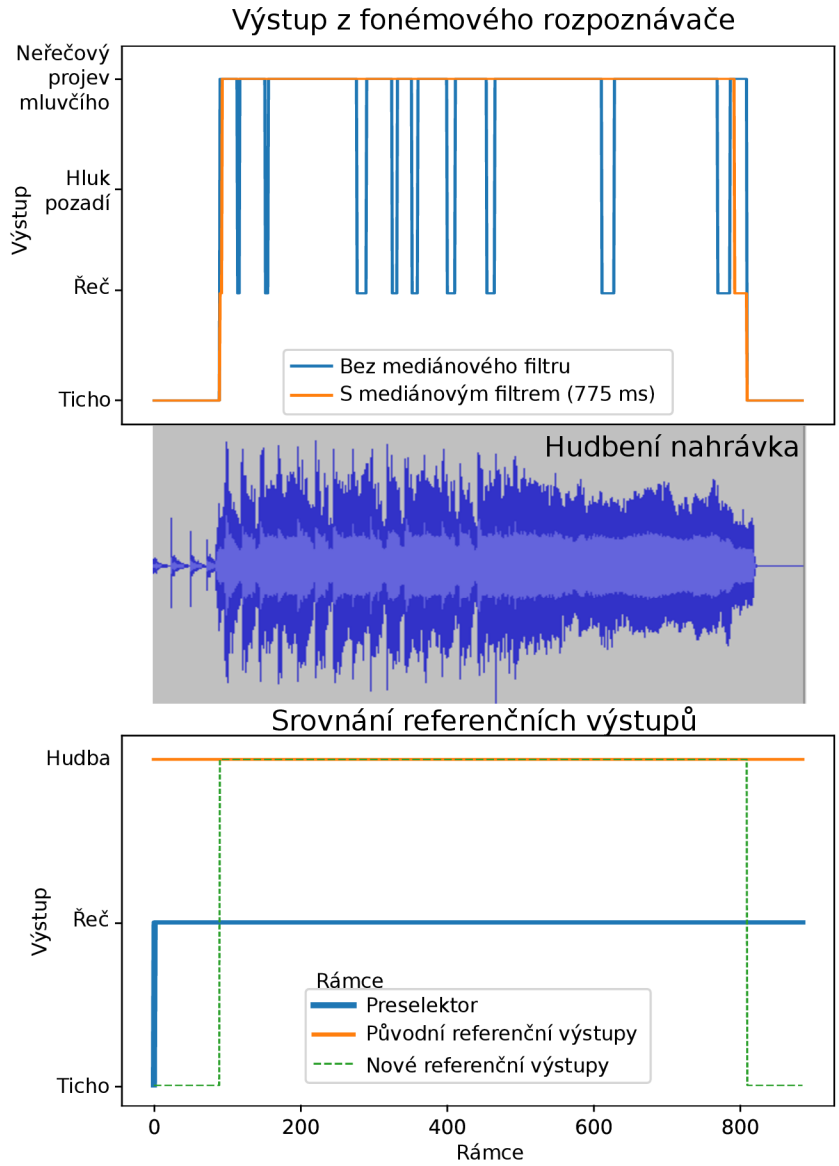
		Predikce			
		Ticho	Řeč	Hudba	Hluk
Reference	Ticho	**76.95 %**	17.71 %	5.21 %	0.13 %
	Řeč	9.52 %	**79.00 %**	10.65 %	0.83 %
	Hudba	1.45 %	4.14 %	**93.88 %**	0.53 %
	Hluk	8.45 %	32.50 %	53.23 %	**5.82 %**

MUSAN

		Predikce			
		Ticho	Řeč	Hudba	Hluk
Reference	Ticho	**75.20 %**	16.23 %	7.64 %	0.94 %
	Řeč	2.64 %	**94.03 %**	2.82 %	0.51 %
	Hudba	1.95 %	3.18 %	**93.11 %**	1.75 %
	Hluk	3.48 %	6.86 %	17.32 %	**72.34 %**

GTZAN

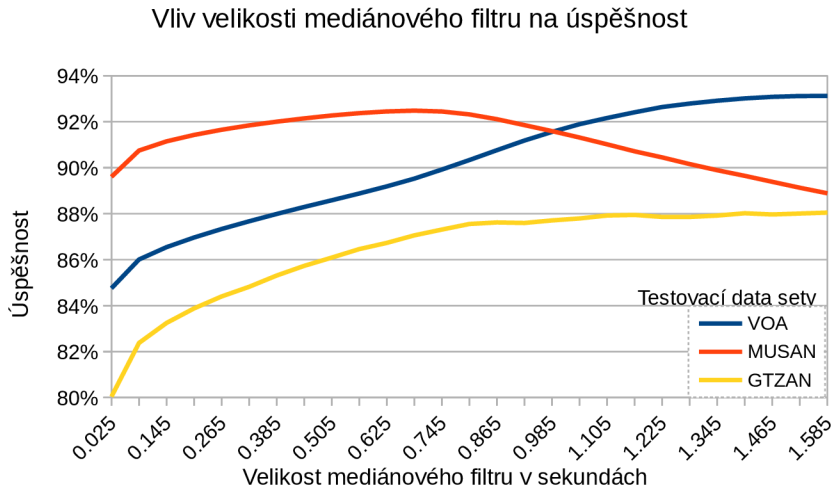
		Predikce			
		Ticho	Řeč	Hudba	Hluk
Reference	Ticho	**67.87 %**	20.19 %	11.44 %	0.49 %
	Řeč	3.80 %	**87.03 %**	6.79 %	2.38 %
	Hudba	6.59 %	9.78 %	**75.06 %**	8.57 %



Obrázek 5.7: Srovnání výstupů z fonémového rozpoznávače pro konkrétní nahrávku a její referenční výstupy.

5.4 Úprava výstupů z neuronové sítě

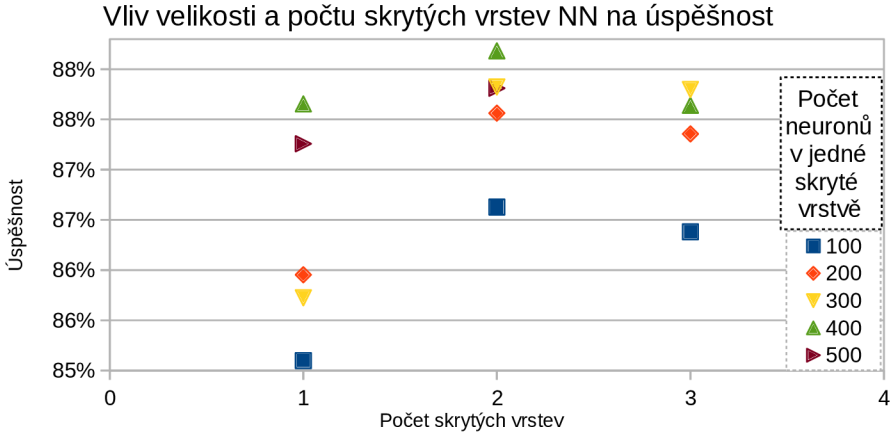
Nezpracovaný výstup z neuronové sítě byl dle očekávání chaotický. Je pravděpodobné, že třídy se nebudou měnit od rámce k rámci — tedy každých 25 milisekund. Z tohoto důvodu byl na výstupy aplikován mediánový filtr různých velikostí. Výsledky tohoto experimentu jsou vyneseny do grafu 5.8. Z grafu je patrné, že každému data setu vyhovuje jiná velikost mediánového filtru. Pro dosažení co nejlepší generalizace je potřeba vybrat kompromis. Tím kompromisem je hodnota 0.985 s (65 rámců). Celkovou úspěšnost systému na trénovací sadě to zvedne o 5 % absolutně.



Obrázek 5.8: Vliv velikosti mediánového filtru na úspěšnost systému.

5.5 Experimentování s velikostí a počtem skrytých vrstev neuronové sítě

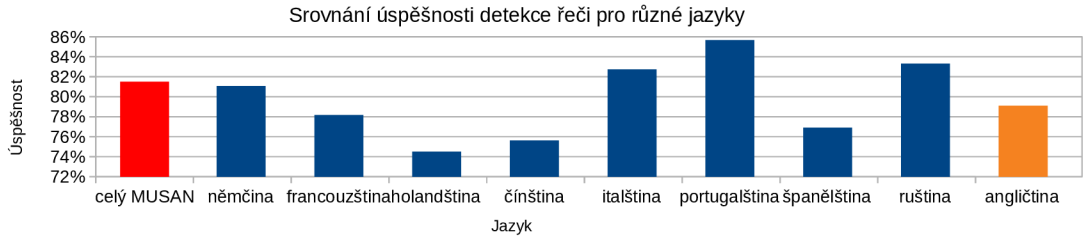
Dalším parametrem, který lze pozorovat je počet skrytých vrstev neuronové sítě a také jejich velikost. Několik výzkumníků se shoduje na tom, že jedna skrytá vrstva je dostačující pro jakékoli nelineární mapování [4]. Kolik však je vhodné, aby měla neuronů nebo zdali vyšší počet vrstev s méně neurony nebude pro daný úkol efektivnější, již jasné není. Proto byly provedeny experimenty, během nichž byla neuronová síť trénována s jednou, dvěma a třemi vrstvami o velikostech 100, 200, 300, 400 a 500 neuronů. Tyto experimenty byly časově poněkud náročnější, a proto byla každá z těchto sítí trénována pouze po dobu pěti epoch. Na obrázku 5.9 je možné sledovat, jaké nastavení sítě pasovalo nejlépe. Obecně vidíme, že nejlepších výsledků dosahují skryté vrstvy dvě. Konkrétně pak s 400 neurony v každé z nich. Rozdíl v úspěšnosti takto velké sítě oproti té nejmenší obsahující jednu vrstvu se 100 neurony jsou pouze 3 %. Tudíž pro praktickou implementaci by se dala použít i velice malá síť.



Obrázek 5.9: Vliv velikosti a počtu skrytých vrstev na úspěšnost neuronové sítě.

5.6 Analýza robustnosti detektoru řečové aktivity

Cílem této práce je vytvořit detektor řečové aktivity, který bude robustní jednak vůči šumu a hudbě na pozadí, a také bude schopný detekovat řeč v různých jazycích. Multilingvální robustnost byla otestována na data setu MUSAN, který celý obsahuje 16 různých jazyků včetně angličtiny. Otestovány byly pouze ty jazyky, pro které se v data setu vyskytuje alespoň 7 nahrávek. Aby byly výsledky vypovídající, je úspěšnost vyhodnocena vždy na právě sedmi nahrávkách patřící k danému jazyku. Na následujícím obrázku 5.10 je vidět srovnání němčiny, francouzštiny, holandštiny, čínštiny, italštiny, portugalštiny, španělštiny, ruštiny a také angličtiny, která je ve všech třech data setech obsažena nejvíce. Je vidět, že některé jazyky jsou problémovější, ale v průměru dosahují úspěšnosti velmi podobné angličtině, tudíž lze tvrdit, že systém detekuje řeč s multilingvální robustností a není přetrénovaný pouze na jazyk anglický. Na výsledky zde uvedené nebyl použit žádný mediánový filtr při výstupu z neuronové sítě, aby volba filtru neovlivnila možnost porovnání. Konečné výsledky na jednotlivých jazycích by tím pádem byly zhruba o 5 % vyšší.



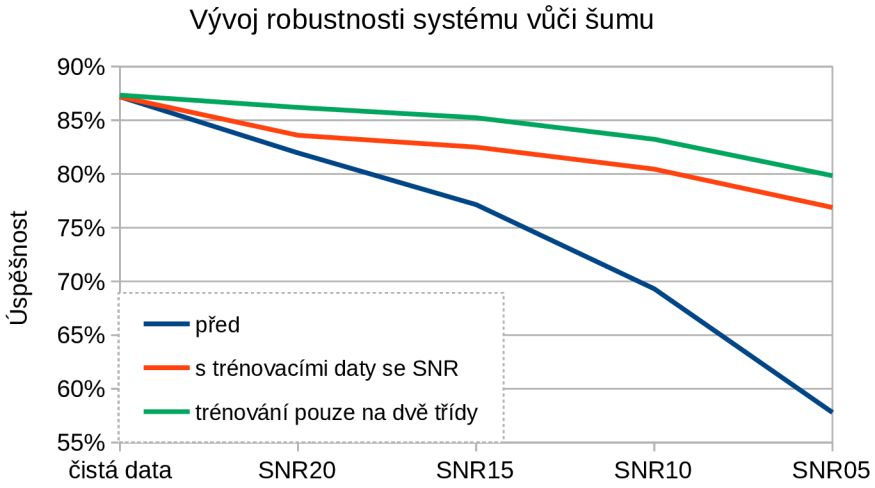
Obrázek 5.10: Úspěšnost detekce řeči pro různé jazyky.

Robustnost systému vůči šumu

Robustností vůči šumu se rozumí umět detekovat řeč, i když je na pozadí nějaký rušivý element. V těchto případech je nutné stanovit si, do jaké míry bude žádoucí, aby systém řeč ještě detekoval. Dle poslechu různých úrovní poměru signálu ke hluku, *z angl. signal-to-noise ratio*, dále jen SNR, jsem došla k závěru, že o řeči se šumem na pozadí lze mluvit do hodnoty 5 dB. Dále již je řeč velice těžko rozpoznatelná a v úsecích, kde je řeč tichá, už není slyšet vůbec.

Pro možnost sledování vývoje robustnosti byl základní systém používající kontext o velikosti 415 ms a nové referenční výstupy z kapitoly 5.3 vyhodnocen na testovacích datech, které byly zašuměny s hodnotami SNR 20 dB, 15 dB, 10 dB a 5 dB. Výsledek tohoto počinu je znázorněn modrou linkou v grafu 5.11 a není moc uspokojivý. Úspěšnost klesla z 87 % na pouhých 57 %. K vylepšení této situace bylo nutné podniknout další kroky. Nejprve byly ve výše zmíněných hodnotách SNR zašuměny i data trénovací a dále bylo potřeba nalézt vhodný poměr těchto zašuměných dat a dat čistých, nad kterými bude probíhat trénování, aby nedocházelo k úpadku úspěšnosti na datech čistých. Tento poměr byl ustálen na zachování původního množství čistých trénovacích dat a přidaní 1/4 náhodně vybraných trénovacích dat zašuměných v hodnotě SNR 20 dB, 1/4 náhodně vybraných trénovacích dat zašuměných v hodnotě SNR 15 dB a stejně tak zbylé 2/4 trénovacích dat pro hodnoty SNR 10 dB a 5 dB.

Srovnání je možné vidět v grafu 5.11. Původní systém (modrá linka), u kterého byl vidět pokles úspěšnosti se snižujícím se SNR až o 30 % absolutně a nový systém (červená



Obrázek 5.11: Vývoj robustnosti systému vůči šumu na pozadí.

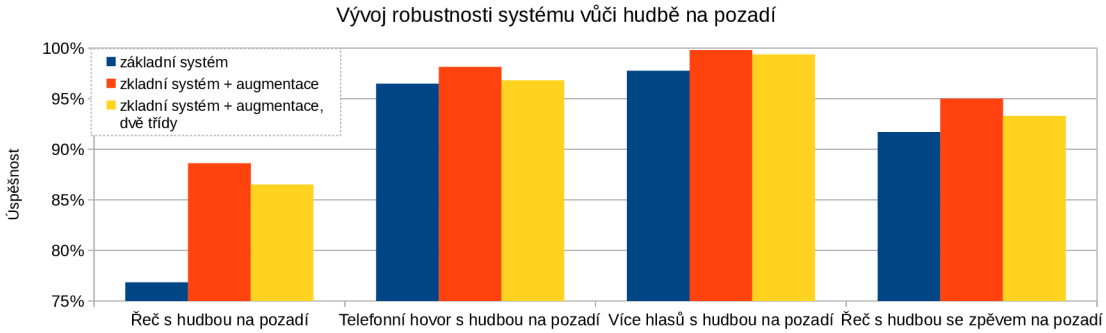
linka), který byl už vůči šumu značně odolnější. Pokles úspěšnosti od čisté řeči k řeči zašuměné s hodnotou SNR 5 dB je pouze o 10 % absolutně. Ještě o něco lepších výsledků bylo dosaženo při trénování systému pouze na dvě třídy — jednu třídu reprezentující řeč a druhou třídu reprezentující všechno ostatní. Tento vývoj je v grafu zaznačen zelenou linkou a jeho úspěšnost klesla pouze o 7 % absolutně. Pro možnost srovnání neovlivněné dalšími parametry nebyl na výstupy, které jsou zaznačeny v grafu, použit žádný mediánový filtr. Finální systém používající mediánový filtr proto dosahuje výsledků o něco vyšších.

Robustnost systému vůči hudbě na pozadí

Společně s robustností vůči šumu byla vyvíjena robustnost vůči hudbě na pozadí. Úspěšnost této robustnosti byla měřena na testovacích datech z data setu VOA, která obsahují různé formy řeči s různou formou hudby na pozadí. Kompletní popis těchto dat již byl představen v kapitole 3.1 v tabulce 3.2. Pro možnost porovnání byla tato data nejprve vyhodnocena systémem používající kontext o velikost 415 ms a nové referenční výstupy z kapitoly 5.3. Vzhledem k tomu, že v trénovacích datech už data obsahující řeč s hudbou na pozadí jsou, výsledky jsou uspokojivé a jsou v grafu 5.12 znázorněny modrými sloupci. Následně byla do trénovací sady společně se zašuměnými daty popsanými v předchozí kapitole 5.6, přidána uměle vytvořená data obsahující muziku na pozadí řeči — augmentace. Augmentace byla provedena přidáním dat obsahující hudbu z data setu MUSAN do dat obsahující řeč. Hudba byla přidána v náhodně zvolené hlasitosti tak, aby poměr signálu k hudbě byl v rozmezí 5–20 dB.

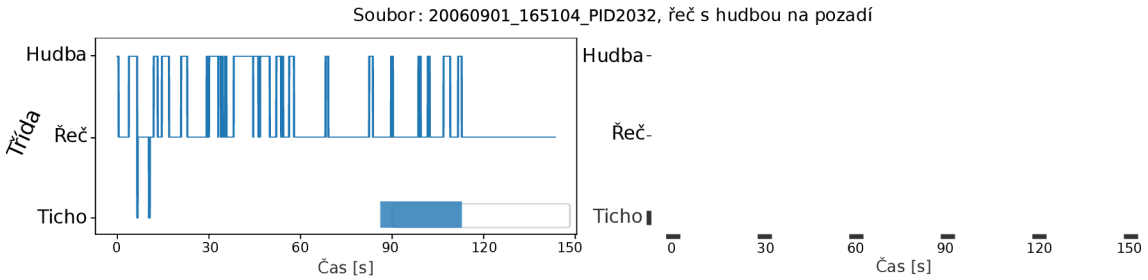
K jakým výsledkům vedl tento krok je možné sledovat v grafu 5.12. Zlepšení je patrné ve všech případech řeči s hudbou na pozadí. Pro nejpočetnější variantu (normální řeč s hudbou na pozadí — první skupinka sloupců) je zlepšení dokonce 12 % absolutně. Pro lepší představu je na obrázku 5.13 uvedena predikce systému pro jednu konkrétní nahrávku obsahující tuto normální řeč s hudbou na pozadí prvotním systémem (vlevo) a systémem trénovaným na větší robustnost (vpravo). Zlepšení je vidět například mezi 30. a 60. sekundou — predikce vpravo už neobsahuje delší úseky hudby tak jako je tomu v predikci vlevo.

V grafu 5.12 je možné ještě vidět další pokus o zlepšení zredukováním výstupních tříd pouze na dvě — na řeč a všechno ostatní. Úspěšnosti jsou znázorněny žlutými sloupci. Ro-



Obrázek 5.12: Vývoj robustnosti systému vůči hudbě na pozadí.

bustnosti detekce řeči na hudebním pozadí na rozdíl od robustnosti vůči šumu tato redukce klasifikačních tříd neprospěla. Oproti robustnímu systému používající třídy 4 úspěšnost klesla až o 2 % absolutně.



Obrázek 5.13: Predikce klasifikátoru pro rámce, kde je přítomna řeč s hudbou na pozadí. Před přidáním uměle vytvořených dat obsahující řeč s hudbou na pozadí (vlevo) a po přidání (vpravo).

Kapitola 6

Zhodnocení finálního systému pro robustní detekci řečové aktivity

V následující kapitole bude představen a zhodnocen finální systém pro robustní detekci řečové aktivity jak v prostředí se šumem, tak i v prostředí s hudbou. Finální systém čerpá z výsledků experimentů popsaných v kapitole 5 a umí detekovat nejen řeč, ale i hudbu, a to jak vokální, tak instrumentální. Dále bude detailně popsána úspěšnost finálního systému manuálním zhodnocením predikce pro dvě hodinové nahrávky z testovacího data setu VOA.

6.1 Popis finálního systému pro robustní detekci řečové aktivity

Finální systém pro detekci řečové aktivity byl sestaven dle poznatků získaných při experimentování, jehož průběh je popsán v kapitole 5. Používá celkem 14 příznaků. Z toho je 13 MFCC koeficientů a 1 příznak pro preselektor. Tato vstupní data jsou následně spojena do úseků o velikosti 0.415 sekund (27 rámců) pro vytvoření kontextové informace.

Klasifikaci provádí neuronová síť se dvěma skrytými vrstvami, kde každá z nich obsahuje 400 neuronů. Systém je natrénován na upravených původních referenčních výstupech, které jsou upraveny pomocí fonémového rozpoznávače PhnRec [29]. Ukončení trénování a vyvarování se přetrénování je zajištěno brzským zastavením, *angl. early stopping* [24], které používá 1/3 trénovacích dat pro validaci. Po ukončení každé epochy je tato validační sada vyhodnocena a dojde–li k poklesu úspěšnosti oproti epoše minulé, systém se vrátí k modelu, který byl natrénován v minulé epoše, a zkusí trénování znovu. Pokud je úspěšnost v další epoše vyšší, trénování pokračuje, v opačném případě dojde k zastavení a jako výsledný model je vybrán ten s poslední nejvyšší úspěšností na validační sadě. Data pro trénování obsahují 52 % originálních dat, 36 % dat zašuměných v hodnotách SNR 5, 10, 15 a 20 dB a 12 % dat, ke kterým byla uměle přidána hudba na pozadí v různé úrovni hlasitosti.

Po výstupu z neuronové sítě je na výsledek aplikován preselektor, který vezme poslední příznak a na základě jeho hodnoty přiřadí konkrétnímu rámci třídu pro ticho nebo ji nechá tak, jak byla klasifikována neuronovou sítí. Následně je na tyto výstupy aplikován mediánový filtr o velikosti 0.975 sekund (65 rámců).

6.2 Zhodnocení finálního systému pro robustní detekci řečové aktivity

Od kvantitativního měření úspěšnosti na celých sadách bude nyní pozornost přešunuta na měření kvalitativní. Pro vytvoření reálné představy, jak dobře detektor řečové aktivity funguje, byly vybrány dvě nahrávky z testovací sady data setu VOA. Data set VOA byl vybrán ze dvou důvodů. Prvním důvodem je forma dodání jeho původních anotací s přesností na milisekundy a druhým důvodem je obsah tohoto data setu. Data set VOA obsahuje záznamy rádiového vysílání, což je velice dobrý materiál jak pro trénování, tak pro testování detektorů řečových aktivit. Obsahuje v sobě totiž rozmanité množství zvukových projevů. Vyskytuje se v něm čistá řeč, rušivá řeč například z telefonního hovoru, řeč s hudbou na pozadí, samostatná hudba, či hudba se zpěvem. Rádiové vysílání je tak jedním z možných typů audia, pro které může být použití detektoru řečové aktivity přínosné.

Pro zhodnocení byly vybrány nahrávky s nejlepší a nejhorší úspěšností. Nejlepší nahrávka dosáhla úspěšnosti 97 % v čisté podobě. V podobě s největší úrovní zašumění (SNR 5 dB) byla její úspěšnost 86 %. Nejhorší nahrávka v čisté variantě dosáhla úspěšnosti 83 % a se zašuměním na hodnotu SNR 5 dB její úspěšnost klesla na 77 %. Nyní bude konzultováno, co stojí za těmito výsledky.

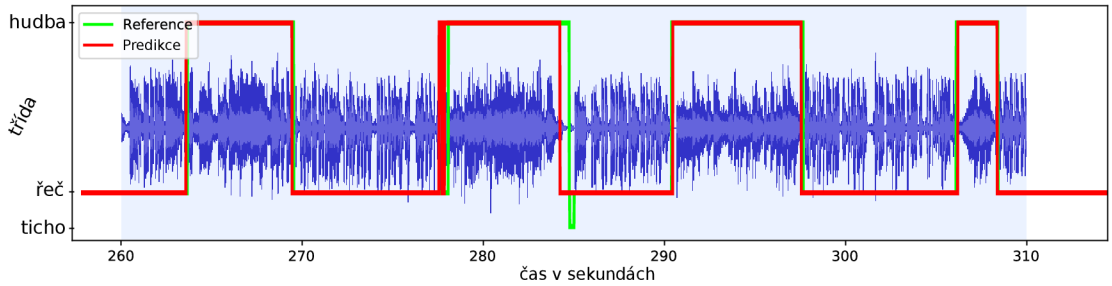
Zhodnocení nahrávky s nejvyšší úspěšností

Nahrávka dosahující nejlepší úspěšnosti obsahuje rádiové vysílání v arabštině a arabskou hudbu. Obsahuje 30 % řeči, 60 % hudby a zbylých 10 % je ticho a hluk. Mimo úseky ticha kratší než 5 sekund obsahuje tato nahrávka 15 přechodů mezi delšími úseky hudby a řeči. Část hudebního obsahu jsou na sebe navazující hudební klipy trvající v průměru 4 sekundy — tak jak to bývá v rádiovém vysílání u přehledu aktuálních hudebních hitů. Občas jsou tyto klipy proloženy krátkou promluvou moderátora, která trvá průměrně 5 sekund. Detektor tuto promluvu v čisté i zašuměné nahrávce identifikuje zcela přesně. Hudební klipy obsažené v této nahrávce obsahují velké množství zpěvu. U čisté nahrávky je tento zpěv rozpoznán zcela korektně jako hudba. U zašuměné nahrávky dochází v těchto okamžicích ke krátkým záměnám hudby za řeč, avšak ne na delší úsek než je půl sekundy. Čím melodičtější zpěv je a čím delší má tóny, tím lépe je rozpoznaný jako hudba i v zašuměné nahrávce.

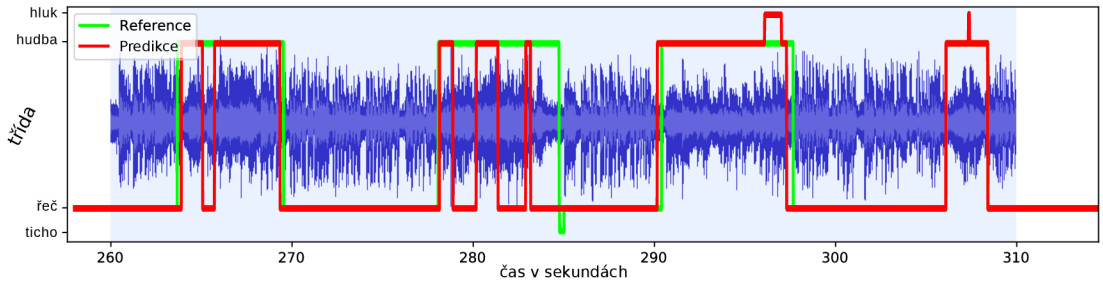
Dále nahrávka obsahuje krátké úseky obsahující na sebe navazující telefonní promluvy. Každý z těchto úseků má trvání kolem 5–ti sekund. Tyto úseky jsou opět rozpoznány korektně jak v čisté, tak v zašuměné nahrávce. Zdrojem chyb jsou však občasné přechody mezi těmito telefonními hovory obsahující slovo moderátora v délce 1–2 sekund. Detektor tento úsek rozpozná jako hudbu. Je však nutné podotknout, že toto slovo opravdu spíše zní jako nějaký hudební zvuk než jako řeč. Dalším zdrojem chyb jsou přechody mezi promluvami, kde referenční výstupy říkají, že by mělo být identifikováno ticho. Toto ticho je však velmi těžko pozorovatelné i pro člověka a trvá maximálně 1 sekundu. Úseky ticha, které trvají déle než 3 sekundy, pokud jde o nahrávku čistou, klasifikátor identifikuje. U zašuměné nahrávky se tato délka pro identifikaci ticha zvyšuje až na 5 sekund. Obecně, v zašuměné nahrávce jakýkoli úsek, který je delší než 5 sekund, klasifikátor identifikuje převážně správně. Delší úseky promluvy (30 sekund) jsou detekovány dokonce bezchybně jak v čisté, tak v zašuměné nahrávce.

Na následujících dvou obrázcích je možné vidět úseky, ve kterých se střídá řeč s hudbou, která v sobě obsahuje i zpěv. První obrázek 6.1 představuje rozpoznání na čisté nahrávce. Druhý obrázek 6.2 představuje ten stejný úsek, ale nahrávky zašuměné. U obou případů je

identifikace řeči téměř bezchybná. K chybám ovšem dochází u zašuměné nahrávky v místech, kde je hudba obsahující zpěv.



Obrázek 6.1: Ukázka rozpoznání řeči a hudby v čisté nahrávce.

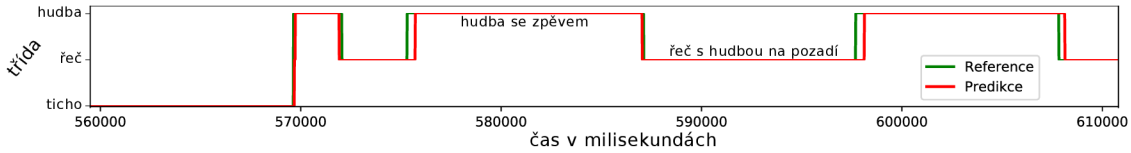


Obrázek 6.2: Ukázka rozpoznání řeči a hudby v zašuměné nahrávce.

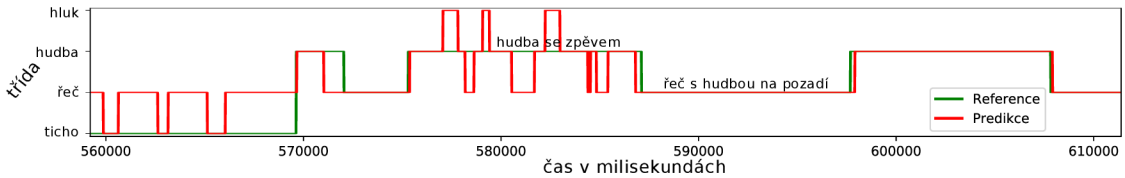
Zhodnocení nahrávky s nejnižší úspěšností

Nahrávka dosahující nejnižší úspěšnosti obsahuje rádiové vysílání v angličtině, zaměřené na hudební styl hip–hop. Nahrávka obsahuje 3 % ticha a hluku, 27 % řeči a 70 % hudby. Kromě krátkých úseků ticha (do 5–ti sekund) obsahuje tato nahrávka 16 přechodů mezi delšími úseky řeči a hudby. Jak již naznačovalo hodnocení nahrávky předchozí, čím nižší melodičnost a délka tónu zpěvu, tím hůře je pro klasifikátor rozpoznatelná jako hudba a klasifikátor si ji může občas plést s řečí. To je přesně případ hip–hopu, kde zpěv až tak moc melodický není. Někdy je dokonce kombinován s rapem, kde o tom, jestli se jedná o řeč nebo zpěv lze vést diskuze. Proto tuto nízkou úspěšnost přikládám tomu, že nahrávka obsahuje velké množství hudby hip–hopového stylu. I tak jsou ale tyto záměny za řeč tam, kde je hip–hopová hudba v čisté nahrávce, maximálně půl sekundy dlouhé. Majoritně je tato hudba rozpoznávána správně. U stejné nahrávky zašuměné na hodnotu SNR 5 dB jsou tyto chybně klasifikované úseky delší, třeba až 3 sekundy. Tomuto problému by šlo u nahrávek obsahujících větší množství hluku předejít delším mediánovým filtrem — pro finální systém je použit mediánový filtr o velikosti 1 sekundy.

V zašuměné nahrávce již samozřejmě není přítomno žádné skutečné ticho. Ticho a hluk splývá v jednu třídu a jelikož je v trénovacích datech větší množství referenčních výstupů označujících ticho než těch, které označují hluk, je hluku bez řeči či hudby často přiřazována kategorie ticha. To však nevidím jako moc velký problém, jelikož primárním cílem je identifikace řeči a na ni tato skutečnost negativní vliv nemá.



Obrázek 6.3: Ukázka rozpoznání různých tříd v čisté nahrávce



Obrázek 6.4: Ukázka rozpoznání různých tříd v zašuměné nahrávce.

Co se týče rozpoznávání řeči na hudebním pozadí, v obou poslouchaných nahrávkách jsou tyto úseky rozpoznávány s velkou přesností. Na obrázcích 6.3 a 6.4 je ještě dáno do kontrastu rozpoznání úseku, kde se střídají kategorie ticho, hudba, řeč, hudba se zpěvem a řeč s hudbou na pozadí. První obrázek je rozpoznání na čisté nahrávce a druhý je rozpoznání na tom samém úseku nahrávky, která je zašuměná na hodnotu SNR 5 dB. Je vidět, že s rozpoznáním řeči na hudebním pozadí výsledný systém problém nemá ani v jednom případě. Zatímco hudba se zpěvem a ticho je mnohem lépe klasifikováno v případě nahrávky čisté.

Celkové zhodnocení

Z analýzy provedené manuálním poslechem nahrávek s nejvyšší a nejnižší úspěšností je možné přednést několik závěrů o celkové funkčnosti detektoru řečové aktivity.

1. Detektor dokáže robustně detekovat řeč v různých jazycích.
2. Detektor je schopný detekce řeči na pozadí s hudbou.
3. Detektor identifikuje i řeč, která není v příliš dobré kvalitě — telefonní hovory a zašuměná řeč moderátorů rádia.
4. Detektor má problém s detekcí hudby, ve které je obsažený nemelodický zpěv (hip–hop, rap). Tato úspěšnost ještě více klesá s úrovní zašumění nahrávky.
5. Detektor v některých případech není schopen detekovat ticho nebo hluk, který je kratší než 3 sekundy. U nahrávek, které jsou hodně zašuměné je tato rozpoznatelnost o něco horší a systém je schopný detekovat úsek bez řeči či hudby, pokud je delší než 5 sekund.

Závěry ještě doplňují dvě tabulky obsahující konfuzní matice pro finální systém na různých datech. V první konfuzní matici 6.1 jsou výsledky systému na čistých datech. Pro třídu řeč a hudba dosahuje detektor vysoké úspěšnosti (až 99.92 % v případě testovacího data setu MUSAN). U tříd ticho a hluk je bohužel nutné konstatovat, že úspěšnost příliš vysoká není. Je však nutné podotknout, že některé úseky ticha i hluku, které se vyskytují

v referenčních výstupech jsou tak krátké, že jsou těžko detekovatelné i uchem člověka. U data setů MUSAN a GTZAN navíc původně ani ticho anotováno nebylo. Pro případy jako je zdlouhavé poslouchání nahrávek obsahujících dlouhé úseky ticha nebo hluku, ve kterých se čeká na promluvu, by však systém byl schopen detekovat řeč, která by se objevila, byť i na krátký moment. Na řádcích konfuzních matic, kde je referenčním výstupem řeč, je možné pozorovat, že spletení si řeči s jinou třídou je opravdu minimální — většinou do 1 %.

Tabulka 6.1: Konfuzní matice finálního systému na čistých datech různých testovacích data setů.

VOA

		Predikce				
		TICHO	ŘEČ	HUDBA	HLUK	celkem rámců
Reference	TICHO	**55.92 %**	39.66 %	4.42 %	0.01 %	571261
	ŘEČ	0.36 %	**97.66 %**	1.93 %	0.04 %	4122437
	HUDBA	0.15 %	2.53 %	**97.29 %**	0.02 %	3010104
	HLUK	0.43 %	54.06 %	44.01 %	**1.50 %**	20859

MUSAN

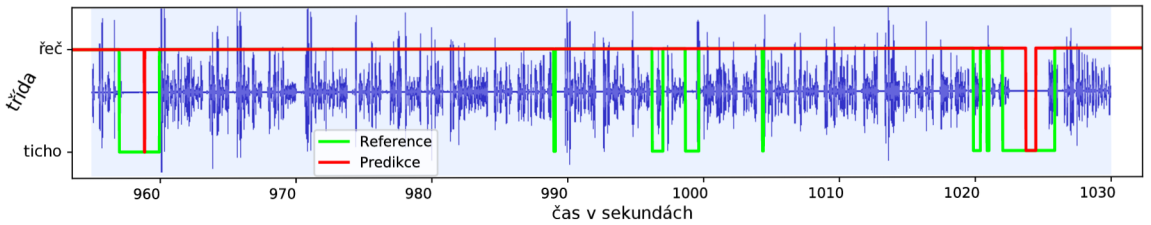
		Predikce				
		TICHO	ŘEČ	HUDBA	HLUK	celkem rámců
Reference	TICHO	**30.92 %**	58.98 %	10.01 %	0.09 %	1471888
	ŘEČ	0.04 %	**99.92 %**	0.03 %	0.00 %	3720507
	HUDBA	0.32 %	5.01 %	**94.62 %**	0.04 %	3376174
	HLUK	4.10 %	32.65 %	40.56 %	**22.68 %**	419808

GTZAN

		Predikce				
		TICHO	ŘEČ	HUDBA	HLUK	celkem rámců
Reference	TICHO	**12.73 %**	71.20 %	16.02 %	0.04 %	16407
	ŘEČ	0.12 %	**98.47 %**	1.37 %	0.04 %	116224
	HUDBA	0.66 %	11.78 %	**87.19 %**	0.36 %	122985

V další tabulce 6.2 jsou výsledky na datech s různými úrovněmi zašumění. Jakmile dojde k zašumění, tak přestává existovat ticho jako takové, ale třídy ticho a šum splynou v jednu. A vzhledem k tomu, že šumu je v trénovacích datech méně než ticha, detektor spíše určí, že se jedná o ticho než o šum. Opět je možné konstatovat, že detekce řeči má vysokou úspěšnost a záměna s jinou třídou je minimální. Většinou do 2 %. Jedinou výjimkou je zašumění s hodnotou 10 dB SNR, kde dochází ke snížení úspěšnosti identifikace řeči na 93 %, tudíž k rozpoznání řeči jako jiné třídy zde dochází v 7 %. Celková úspěšnost systému skrze všechny třídy se pak pohybuje od 84 % do 88 %.

Pro dokreslení výsledků o identifikaci ticha je přiložen obrázek 6.5, kde je vidět, jak systém ticho rozpoznává. Kamenem úrazu je v tomto případě mediánový filtr, který výsledky rozpoznání vyhlazuje. Jeho velikost by se případně dala nastavit individuálně pro potřeby konkrétního použití. Nicméně je vidět, že pokud je úsek delší než 3 sekundy, systém je schopný jej alespoň identifikovat. Obrázek 6.6 toto tvrzení potvrzuje, jelikož obsahuje velké množství úseků ticha, které jsou delší než 3 sekundy a systém je schopný toto ticho detekovat velice dobře.



Obrázek 6.5: Ukázka rozpoznávání úseků ticha v čisté nahrávce.

Obrázek 6.6: Ukázka rozpoznání různých tříd v čisté nahrávce obsahující úseky ticha.

Tabulka 6.2: Konfuzní matice finálního systému na zašuměných datech na testovacích data setech.

20 dB SNR		Predikce			
		TICHO	ŘEČ	HUDBA	HLUK
Reference	TICHO	**23.93 %**	69.86 %	5.92 %	0.30 %
	ŘEČ	0.54 %	**98.43 %**	0.99 %	0.04 %
	HUDBA	0.30 %	6.45 %	**93.20 %**	0.05 %
	HLUK	2.90 %	47.80 %	39.79 %	**12.51 %**

15 dB SNR		Predikce			
		TICHO	ŘEČ	HUDBA	HLUK
Reference	TICHO	**21.37 %**	72.20 %	6.13 %	0.30 %
	ŘEČ	0.54 %	**98.57 %**	0.85 %	0.04 %
	HUDBA	0.32 %	9.23 %	**90.38 %**	0.06 %
	HLUK	3.25 %	48.82 %	35.39 %	**12.54 %**

10 dB SNR		Predikce			
		TICHO	ŘEČ	HUDBA	HLUK
Reference	TICHO	**23.64 %**	61.09 %	12.02 %	3.25 %
	ŘEČ	2.57 %	**92.85 %**	3.50 %	1.09 %
	HUDBA	0.85 %	20.68 %	**77.44 %**	1.02 %
	HLUK	4.19 %	45.88 %	36.05 %	**13.88 %**

5 dB SNR		Predikce			
		TICHO	ŘEČ	HUDBA	HLUK
Reference	TICHO	**13.17 %**	78.03 %	7.68 %	1.12 %
	ŘEČ	0.49 %	**98.34 %**	1.05 %	0.12 %
	HUDBA	0.29 %	22.87 %	**76.28 %**	0.55 %
	HLUK	3.51 %	50.96 %	32.77 %	**12.77 %**

Dodatečné poznatky

Se systémem byly ještě provedeny následující operace, které buď nepřinesly očekávané výsledky, nebo je nelze zařadit do doposud prezentovaných kapitol. První operace se týká zvětšení příznakové sady. Jak již bylo zmíněno, počáteční systém pracoval pouze s MFCC příznaky, které byly později doplněny příznakem, odrážejícím přítomnost nebo nepřítomnost signálu na základě hlasitosti. V návrhu bylo příznaků ale zmíněno více, viz sekce 4.1. Tyto příznaky byly přidány do systému po ukončení experimentů pro zjištění, zdali toto obohacení zvýší úspěšnost systému. Po přidání dalších příznaků však již k žádnému zlepšení nedochází. Není tedy důvod, proč tyto příznaky přidávat a zvyšovat tak časovou náročnost evaluace. Dalším pokusem bylo vyhodnocení dat, které mají nižší vzorkovací frekvenci — konkrétně 8 kHz. Původní trénovací i testovací data totiž byly v kvalitě 16 kHz. Pokles úspěšnosti oproti vyšší vzorkovací frekvenci byl pouze o 0.1 %.

Kapitola 7

Závěr

Cílem této práce bylo navrhnout robustní detektor řečové aktivity, který bude umět detekovat řeč v různých jazycích, v prostředí se šumem a také na hudebním pozadí. Tento cíl byl naplněn a výsledkem je nástroj, který dosahuje úspěšnosti detekce řeči až 97 % na nahrávkách z rádiového vysílání, které představují právě onu přehlídku různých jazyků, nekvalitních úseků řeči v podobě telefonních vstupů, projevy moderátorů na hudebním pozadí a také hudbu jako takovou.

Cesta k vytvoření nástroje vedla přes studium modelů používaných pro detekci řečové aktivity, kde pozornost byla zaměřena především na Gaussovské modely a neuronové sítě. Jako model pro vývoj v rámci této práce byla zvolena neuronová síť. Dále byl proveden výběr a analýza dat, která byla pro vývoj detektoru řečové aktivity použita. Vybraná data splňují nároky na jazykovou pestrost, rozličnost hudebních stylů i rozmanitost prostředí, ze kterého pocházejí. Po vybudování základního systému následovala řada experimentů, která vedla k doladění jeho parametrů. Dále byly podniknuty kroky pro dosažení požadované robustnosti systému v šumu a na hudebním pozadí. Nakonec bylo provedeno důkladné zhodnocení úspěšnosti systému zahrnující jeho přednosti i nedostatky.

Systém prezentovaný v této práci je schopný detekovat řeč v minimálně 12–ti různých jazycích. Řeč na hudebním pozadí je schopen detekovat s 88 % úspěšností a co se týče úspěšnosti detekce řeči v zašuměné nahrávce, systém dosahuje úspěšnosti od 84 % do 88 % — v závislosti na velikosti hodnoty signálu k šumu, *angl. signal-to-noise ratio (SNR).*

Oproti prvnímu systému, který měl úspěšnost na zašuměných datech od 58 % do 82 %, systém po provedených experimentech a ladění, dosáhl zlepšení o 26 % absolutně při hodnotě 5 dB SNR a o 6 % absolutně při hodnotě 20 dB SNR. Zlepšení v rámci detekce řeči na hudebním pozadí bylo o 11 % absolutně — z původních 77 % na 88 %.

Představený systém provádí klasifikaci do čtyř tříd — ticho, řeč, hudba a hluk. Věřím, že kdyby třídy ticho a hluk byly spojeny do jedné a tato skutečnost byla vhodně reflektována v referenčních výstupech, systém by produkoval výsledky ještě lepší. Tímto směrem by tedy mohla pokračovat další dohledná práce. Co se týče práce do vzdálenější budoucnosti, určitě by stálo za pokus vytvořit fúzi tohoto systému a fonémového rozpoznávače Phnrec, který je v této práci použit na upřesnění původních referenčních výstupů. Od výsledné fúze bych si opět slibovala vylepšení úspěšnosti rozpoznání tříd pro ticho a hluk.

Literatura

[1] Abadi, M.; Agarwal, A.; Barham, P.; aj.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. [online], 2015, software dostupný na tensorflow.org.

[2] Deepak, K.; Sarma, B. D.; Prasanna, S. M.: Foreground speech segmentation using zero frequency filtered signal. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[3] Drugman, T.; Stylianou, Y.; Kida, Y.; aj.: Voice Activity Detection: Merging Source and Filter-based Information. *IEEE Signal Processing Letters*, ročník 23, 2016: s. 252–256.

[4] Eberhart, R. C.: *Computational Intelligence: Concepts to Implementations*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007, ISBN 1558607595, 9780080553832.

[5] Frank, S. A.: The Common Patterns of Nature. *Journal of evolutionary biology*, ročník 22, 08 2009: s. 1563–1585, doi:10.1111/j.1420-9101.2009.01775.x.

[6] Ghaemmaghami, H.; Dean, D.; Kalantari, S.; aj.: Complete-linkage clustering for voice activity detection in audio and visual speech. In *Interspeech*, Maritim International Congress Center, Dresden, Germany, Září 2015.

[7] Giannakopoulos, T.: pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. *PloS one*, ročník 10, č. 12, 2015.

[8] Graciarena, M.; Ferrer, L.; Mitra, V.: The SRI System for the NIST OpenSAD 2015 Speech Activity Detection Evaluation. In *Interspeech*, Září 2016, s. 3673–3677, doi:10.21437/Interspeech.2016-550.

[9] Graf, S.; Herbig, T.; Buck, M.; aj.: Features for voice activity detection: a comparative analysis. *EURASIP Journal on Advances in Signal Processing*, ročník 2015, č. 1, Říjen 2015: str. 91, ISSN 1687-6180, doi:10.1186/s13634-015-0277-z.

[10] Hossin, M.; Sulaiman, M. N.: A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, ročník 5, č. 2, Březen 2015, doi:10.5121/ijdkp.2015.5201.

[11] Huang, R.; Hansen, J. H. L.: Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora. *IEEE Transactions on Audio, Speech, and Language Processing*, ročník 14, č. 3, Květen 2006: s. 907–919, ISSN 1558-7916, doi:10.1109/TSA.2005.858057.

[12] Hughes, T.; Mierle, K.: Recurrent Neural Networks for Voice Activity Detection. In *ICASSP*, 2013, s. 7378–7382.

[13] Jo, Q. .; Chang, J. .; Shin, J. W.; aj.: Statistical model-based voice activity detection using support vector machine. *IET Signal Processing*, ročník 3, č. 3, Květen 2009: s. 205–210, ISSN 1751-9675, doi:10.1049/iet-spr.2008.0128.

[14] Khonglah, B. K.; Mahadeva Prasanna, S.: Speech / Music Classification Using Speech-specific Features. *Digit. Signal Process.*, ročník 48, č. C, Leden 2016: s. 71–83, ISSN 1051-2004, doi:10.1016/j.dsp.2015.09.005.

[15] Kristjansson, T.; Deligne, S.; Olsen, P.: Voicing features for robust speech detection. In *Interspeech'2005 - Eurospeech*, Lisbon, Portugal, 09 2005, s. 369–372.

[16] Lehner, B.; Widmer, G.; Sonnleitner, R.: Improving Voice Activity Detection in Movies. In *Interspeech*, Září 2015.

[17] Lerch, A.: *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. Wiley-IEEE Press, první vydání, 2012, ISBN 111826682X, 9781118266823.

[18] Ma, Y.; Nishihara, A.: Efficient Voice Activity Detection Algorithm Using Long-term Spectral Flatness Measure. *EURASIP J. Audio Speech Music Process.*, ročník 2013, č. 1, Prosinec 2013: s. 87:1–87:18, ISSN 1687-4714, doi:10.1186/1687-4722-2013-21.

[19] MARSYAS: GTZAN music/speech collection. [online], 2002, publikováno pod: CC0: Public Domain.
URL http://marsyas.info/downloads/datasets.html

[20] Meier, S.; Kellermann, W.: Artificial Neural Network-Based Feature Combination for Spatial Voice Activity Detection. In *Interspeech*, 2016.

[21] Mesgarani, N.; Slaney, M.; Shamma, S. A.: Discrimination of speech from nonspeech based on multiscale spectro-temporal Modulations. *IEEE Transactions on Audio, Speech, and Language Processing*, ročník 14, č. 3, Kveten 2006: s. 920–930, ISSN 1558-7916, doi:10.1109/TSA.2005.858055.

[22] Ng, T.; Zhang, B.; Nguyen, L.; aj.: Developing a speech activity detection system for the DARPA RATS program. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[23] Oliphant, T.: NumPy: A guide to NumPy. [online], 2006.
URL http://www.numpy.org/

[24] Prechelt, L.: *Early Stopping — But When?* Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, ISBN 978-3-642-35289-8, 53–67 s., doi:10.1007/978-3-642-35289-8_5.

[25] Python Software Foundation. Python Language Reference, version 2.7. [online].
URL https://www.python.org/

[26] Reynolds, D.: Gaussian mixture models. *Encyclopedia of biometrics*, 2015: s. 827–832.

[27] Sadjadi, S. O.; Hansen, J. H. L.: Unsupervised Speech Activity Detection Using Voicing Measures and Perceptual Spectral Flux. *IEEE Signal Processing Letters*, ročník 20, č. 3, Březen 2013: s. 197–200, ISSN 1070-9908, doi:10.1109/LSP.2013.2237903.

[28] Sarikaya, R.; Hansen, J. H.: Robust detection of speech activity in the presence of noise. In *Proc. ICSLP*, ročník 4, Citeseer, 1998, s. 1455–8.

[29] Schwarz, P.; Matějka, P.; Burget, L.; aj.: Phoneme recognizer based on long temporal context. [online].
URL https://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context

[30] Snyder, D.; Chen, G.; Povey, D.: MUSAN: A Music, Speech, and Noise Corpus. 2015, arXiv:1510.08484v1, 1510.08484.

[31] Sohn, J.; Kim, N. S.; Sung, W.: A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, ročník 6, č. 1, Leden 1999: s. 1–3, ISSN 1070-9908, doi:10.1109/97.736233.

[32] Soleimani, S. A.; Ahadi, S. M.: Voice Activity Detection based on Combination of Multiple Features using Linear/Kernel Discriminant Analyses. In *2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications*, Duben 2008, s. 1–5, doi:10.1109/ICTTA.2008.4530028.

[33] Srivastava, N.; Hinton, G.; Krizhevsky, A.; aj.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, ročník 15, 2014: s. 1929–1958.

[34] Černocký, H.: Zpracování řečových signálů — studijní opora. [online], 2016.
URL
https://www.fit.vutbr.cz/study/courses/ZRE/public/opora/zre_opora.pdf

Příloha A

Obsah CD

A.1 Soubory pro běh robustního detektoru řečové aktivity

1. **lists/** — Složka obsahující jeden soubor `list_example.list`, ve kterém je seznam nahrávek pro ukázkovou evaluaci.

2. **models/** — Složka obsahující data nutná pro načtení natrénované neuronové sítě.

3. **records/** — Složka se dvěma přichystanými nahrávkami, jedna čistá, druhá zašuměná v hodnotě 5 dB SNR.

4. **prepare_environment.sh** — Skript pro nachystání prostředí pro hladký běh detektoru řečové aktivity. Je doporučeno tento skript pustit nebo alespoň zkontrolovat, zda–li knihovny a nástroje, které skript stáhne a nainstaluje, již má uživatel nainstalované v příslušné verzi. Verze, které je nutné dodržet jsou ve skriptu explicitně zmíněné. U knihoven, kde verze zmíněna není, by na verzi nemělo záležet.

   Použití:

   ```
   ./prepare_environment.sh
   ```

5. **evaluate.sh** — Skript, který pouze sdružuje funkci skriptů `extract_features.sh` a `nn_evaluate.py`. Je připraven tak, aby extrahoval příznaky z nahrávek umístěných ve složce `records/` do složky `features/`, kterou vytvoří. Následně pustí evaluaci nad těmito příznaky a výsledky uloží do složky `outputs/`, která bude vytvořena také.

   Použití:

   ```
   ./evaluate.sh
   ```

6. **create_data_list.py** — Pomocný soubor pro tvorbu seznamu jmen souborů obsažené ve specifikované složce. Jako parametry bere tento skript jméno souboru pro seznam, cestu ke složce, kde jsou soubory, z nichž má být list vytvořen a posledním parametrem je velikost přípony souborů, která nemá být zahrnuta do jmenného seznamu včetně tečky. Pokud jde například o soubory ve formátu .raw, hodnota tohoto parametru bude 4.

Použití:

```
create_data_list.py <jmeno_listu> <jmeno_slozky> \
<delka_pripony_s_teckou>
```

7. **extract_features.py** — Skript, který přijme jako parametry seznam nahrávek, ze kterých mají být extrahovány příznaky, a cestu ke složce obsahující tyto nahrávky ve formátu .raw. Vytvoří složku `features/` a do ní extrahuje 13 MFCC příznaků a jeden příznak odrážející hodnotu preselektoru — pro odfiltrování ticha. Soubory s příznaky budou pojmenovány stejně jako zdrojové soubory s nahrávkami. Soubory s nahrávkami musí být ve formátu .raw, ve 8 kHz kvalitě, s 16–bit PCM.

   Použití:

   ```
   python extract_features.py <jmeno_listu_s_nahravkami> \
   <jmeno_slozky_s_nahravkami>
   ```

8. **nn_evaluate.py** — Skript, který přijme jako parametry seznam souborů obsahující příznaky a cestu ke složce, ve které jsou příznaky přítomny. Dle příznaků provede detekci řečové aktivity a výsledky uloží do složky `outputs/`, kterou před tím vytvoří. Výsledky jsou uloženy v binárním souboru jako Numpy pole[1] obsahující evaluace pro každý rámec o velikosti 25 ms s délkou překryvu 10 ms. Hodnoty evaluace jsou čísla od 0 do 3, kde 0 je pro ticho, 1 pro řeč, 2 pro hudbu a 3 pro hluk.

   Použití:

   ```
   python nn_evaluate.py <jmeno_listu_s_priznaky> \
   <jmeno_slozky_s_priznaky>
   ```

9. **utils.py** — Pomocný soubor obsahující pomocné funkce pro skripty `extract_features.py` a `nn_evaluate.py`.

10. **const.py** — Soubor obsahující konstanty pro jednotlivé klasifikační třídy.

A.2 Videa s ukázkou detekce řečové aktivity

Na přiloženém optickém médiu jsou k nalezení video nahrávky `ukazka_cista.mp4` a `ukazka_snr05.mp4`, které demonstrují funkci detektoru řečové aktivity na kousku čisté a zašuměné nahrávky.

[1] Numpy array — https://docs.scipy.org/doc/numpy/reference/generated/numpy.array.html