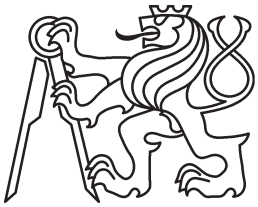
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
FAKULTA ELEKTROTECHNICKÁ
Katedra Teorie obvodů



Modelování neřečových událostí pro rozpoznávání řeči v reálných podmínkách

Dizertační práce

Josef Rajnoha

Praha, únor 2013

Studijní program: Elektrotechnika a informatika

Obor studia: Teoretická elektrotechnika

Školitel: Doc. Ing. Petr Pollák, CSc.

Abstrakt

Předložená disertační práce se zabývá tématem robustního rozpoznávání řeči v reálném prostředí, kdy je nutné brát v úvahu zvýšenou úroveň okolního rušení a přítomnost neřečových jevů v promluvě. Hlavním cílem práce je analýza existujících postupů zpracování řečového signálu s ohledem na dosažení robustní reprezentace řeči a následná optimalizace nastavení předzpracování řečového signálu. Analogicky byly použity a porovnány techniky vedoucí k potlačení vlivu rušení na úrovni modelování řeči.

Obsah práce lze tématicky rozdělit do základních bloků, které se věnují jednotlivým řešeným úlohám. První část práce zkoumá standardně používané parametrizace MFCC a PLP a z nich odvozené techniky pro získání robustní reprezentace řečového signálu. V rámci analýzy těchto technik bylo hledáno nejen optimální nastavení jednotlivých parametrů, ale zároveň byly jednotlivé bloky zpracování signálu podrobeny analýze z pohledu zvýšení robustnosti získané reprezentace signálu v kombinaci s metodami pro potlačování vlivu šumu – spektrální odečítání a detekce řečové aktivity. Srovnání těchto metod v různých šumových podmínkách ukázalo výhodné vlastnosti technik založených na LPC analýze při použití dodatečných algoritmů pro potlačování šumu. Navržený postup zpracování signálu dosáhl výsledků srovnatelných se standardy používanými pro potlačování šumu v řeči (ETSI standard).

Druhá část práce je věnována oblasti akustického modelování řečového signálu, především z pohledu přizpůsobení modelů řeči na šumové podmínky. Byly studovány podrobněji standardní adaptační techniky na bázi MLLR a proces modelování neřečových událostí. Pro analýzu přizpůsobení modelů řeči šumovým podmínkám byla navržena metoda, která zahrnuje optimalizovaný proces předzpracování signálu na úrovni parametrizace i modelování. Dosažená úspěšnost rozpoznávání přesáhla 98,5 % v podmínkách jedoucího automobilu (oproti původním méně než 90 %). Modelování neřečových událostí řečníka bylo realizováno v úloze rozpoznávání spontánních promluv, kde vedlo ke snížení chybovosti systému o cca 10 %.

Pro tyto účely vznikla v rámci práce databáze spojitých promluv spontánního charakteru.

Abstract

This thesis deals with the robust speech recognition in real environment, when it is necessary to take into account the presence of high level of ambient noise and non-speech events. The main objective of this work is to analyze existing procedures of speech signal processing with the purpose to achieve a robust representation of processed speech and optimized settings of speech signal parameterization. Also the techniques leading to interference suppression at speech signal modelling level were applied and compared.

The content of the work can be divided into basic blocks which are devoted to particular tasks solved. The first part examines the standardly used parameterizations MFCC and PLP and derived techniques for obtaining robust representation of speech signal. We were searching not only for the optimal setting of parameters, but individual signal processing blocks were analyzed from the viewpoint of increasing robust representation of the signal obtained. The techniques were studied in combination with methods for noise suppression, i.e. mainly Spectral subtraction and Voice activity detection. Advantageous properties of the techniques based on LPC analysis were observed when they are used together with additional techniques for noise suppression. Proposed approach reached results comparable to the standards used for noise suppression in speech (ETSI standard).

The second part is devoted to the modelling of acoustic speech signal, especially from the viewpoint of the adaptation of acoustic models for noisy-speech conditions. The standard adaptation techniques based on MLLR and the modelling of non-speech events were studied in more details. Proposed method involves optimized signal preprocessing at the level of the signal parameterization and modelling. Finally, recognition accuracy of more than 98.5 % was achieved in the case of driven car (compared to previous results below 90 %). Also the modelling of speaker non-speech events was implemented in large vocabulary continuous speech recognition task, which led to the reduction in error rate of about 10 %.

For this purpose, continuous database of spontaneous utterances character was createded within the work.

Poděkování

Tato dizertační práce vznikala pod vedením školitele Petra Polláka, kterému děkuji za pečlivou a trpělivou asistenci, cenné rady, profesionální a zároveň přátelský přístup při tvorbě práce, pokud možno vždy s červenou tužkou po ruce a poučným slovem na rtech.

Velký dík patří také všem kolegům na pracovišti, kteří podíleli na tvorbě a provozu zázemí, v němž byly realizovány početně náročné úlohy. Stejně tak děkuji doktorandům z katedry za zajímavé podněty v rámci diskusí, ať už na půdě akademické, nebo mimo ni a za jejich osobní příspěvek ke vzniku databáze spontánních promluv.

V neposlední řadě děkuji mé rodině a přátelům, kteří mi vytvářeli příjemné a klidné zázemí pro práci, za jejich duchovní podporu, pochopení mého časově náročného koníčka a ochotu být vždy nablízku s povzbuzením a úsměvem.

Obsah

1 Úvod **1**

2 Robustní rozpoznávače řeči - stav problematiky **3**
2.1 Princip rozpoznávání řeči na bázi HMM 3
2.1.1 Statistický model řeči 4
2.1.2 Předzpracování signálu – Extrakce příznaků 5
2.1.3 Akustické modelování 6
2.1.4 Výslovnostní slovník 9
2.1.5 Jazykové modelování 9
2.2 Robustnost rozpoznávače řeči 10
2.2.1 Robustnost při extrakci příznaků 11
2.2.2 Robustnost v akustickém modelování 15

3 Cíle práce **19**

4 Nastavení experimentů **23**
4.1 Použité nastavení rozpoznávače řeči 23
4.1.1 Nástroje pro parametrizaci signálu 24
4.1.2 Trénování akustických modelů 24
4.1.3 Rozpoznávací systémy 24
4.1.4 Míry pro hodnocení úspěšnosti rozpoznávače 27
4.2 Použité řečové databáze 27
4.2.1 Databáze SPEECON 28
4.2.2 Databáze CZKCC 31
4.2.3 SNR-set 32
4.2.4 AURORA 3 33
4.2.5 Databáze spontánních promluv 33

5 Databáze spontánních promluv **35**
5.1 Nahrávky přednášek 35
5.2 Nahrávací zařízení 35
5.3 Segmentace a anotace signálu 37
5.3.1 Segmentace dlouhých nahrávek 37

5.3.2 Ortografická transkripce . . . 37
5.3.3 Fonetická transkripce . . . 38
5.3.4 Anotace neřečových událostí . . . 38
5.4 Výsledný obsah databáze . . . 39
5.5 Srovnání kvality nahrávek . . . 40
5.5.1 Srozumitelnost řeči . . . 40
5.5.2 Výskyt neřečových událostí . . . 40

6 Příznaky pro rozpoznávání řeči 43
6.1 Standardní parametrizace . . . 43
6.2 Modifikované metody . . . 47
6.2.1 Popis modifikovaných metod . . . 47
6.2.2 Shrnutí vlastností parametrizačních technik . . . 49
6.3 Rozšířené spektrální odečítání . . . 50
6.4 Experimentální část . . . 51
6.4.1 Segmentace signálu pro krátkodobou analýzu . . . 51
6.4.2 Robustnost param. metod v nepřizpůsobených podmínkách . . . 52
6.4.3 Reálné prostředí - přizpůsobené podmínky . . . 54
6.4.4 Spektrální odečítání . . . 54
6.4.5 Spektrální odečítání – AURORA3 . . . 55

7 Detekce řečové aktivity pro účely rozpoznávání řeči 59
7.1 Kepstrální detektor řečové aktivity . . . 59
7.1.1 Kepstrální vzdálenost . . . 60
7.1.2 Prahování . . . 61
7.1.3 Vyhlazení výsledků detekce . . . 63
7.2 Možnosti nastavení VAD algoritmů . . . 64
7.3 Experimentální část . . . 64
7.3.1 Přesnost detekce řečové aktivity . . . 65
7.3.2 Selektivní trénování akustických modelů s VAD . . . 66
7.3.3 Spektrální odečítání s detekcí řeči . . . 70

8 Přizpůsobení akustických modelů na rušivé prostředí 77
8.1 Metody přizpůsobení modelů řeči na šumové prostředí . . . 77
8.1.1 Přetrénování na cílových podmínkách . . . 78
8.1.2 Trénování na obecných šumových podmínkách . . . 79
8.1.3 Adaptace modelů metodou MLLR . . . 79
8.2 Výchozí modely pro adaptaci . . . 80
8.2.1 Experimentální část . . . 81
8.2.2 Shrnutí . . . 83
8.3 Srovnání technik pro přizpůsobení modelů . . . 83
8.3.1 Experimentální část . . . 86
8.3.2 Shrnutí . . . 89

8.4 Adaptační schéma s MLLR 90
8.4.1 Experimentální část 91
8.4.2 Shrnutí 94

9 Modelování neřečových událostí **95**
9.1 Výskyt neřečových událostí v řečových korpusech 96
9.2 Klasifikace událostí v trénovací databázi 97
9.2.1 Inicializace modelů neřečových událostí 97
9.2.2 Rozšíření sady modelů 98
9.2.3 Trénování na celém řečovém materiálu 98
9.2.4 Využití spontánních promluv pro modelování neřeč. událostí . 100
9.3 Hodnocení kvality neřečových událostí v řečových korpusech 101
9.4 Odolnost vůči rušení v úloze LVCSR 102
9.4.1 Nastavení LVCSR systému 103
9.4.2 LVCSR s modely neřečových událostí 104
9.4.3 Experimentální část 104
9.4.4 Shrnutí 106

10 Závěr **107**

Vlastní publikace autora **111**

Literatura **113**

Příloha A **125**

Příloha B **127**

Příloha C **129**

Příloha D **131**

Seznam obrázků

2.1 Blokové schéma standardního rozpoznávače řeči založeného na HMM 4
2.2 Příklad 5-stavového HMM s dvěma neemitujícími stavy 7
2.3 Příklad jednoduché gramatiky rozpoznávače povelů 10
2.4 Blokové schéma potlačení šumu pomocí Wienerovy filtrace 13
2.5 Blokové schéma potlačení šumu spektrálním odečítáním 13
2.6 Blokové schéma potlačení šumu metodou ESS 14

5.1 Nahrávací schéma s bezdrátovou přenosovou sadou 36

6.1 Blokové schéma výpočtu standardních parametrizací MFCC a PLP . 44
6.2 Banky filtrů pro analýzu řeči - Barkova BF, Melovská BF 45
6.3 Spektrogram řečového signálu před a po filtraci Mel BF 45
6.4 Blokové schéma výpočtu standardních parametrizací RPLP 48
6.5 Blokové schéma výpočtu standardních parametrizací MFLP 48
6.6 Blokové schéma výpočtu standardních parametrizací BFCC 49
6.7 Blokové schéma výpočtu jednotlivých parametrizací 50
6.8 Chybovost *WER* pro rozdílná nastavení segmentace a délky banky filtrů 52

7.1 Blokové schéma použitého detektoru řeči 60
7.2 Průběh nastavení prahu pro různé typy prahování v rámci průchodu signálem 61
7.3 Blokové schéma pro algoritmus fixního prahování 62
7.4 Blokové schéma pro algoritmus adaptivního prahování 62
7.5 Blokové schéma pro algoritmus adaptivního prahování na bázi dynamiky 63
7.6 *WER* pro jednotlivé parametrizace MFCC a PLP a rozdílný algoritmus prahování za různých šumových podmínek 68
7.7 *WER* při různých volbách nastavení prahování 69

8.1 Výchozí řečové modely pro zvýšení efektivity přizpůsobení na šumové podmínky 80
8.2 Blokové schéma rozpoznávacího rámce pro analýzu robustnosti ASR systému v šumovém prostředí 81
8.3 Blokové schéma rozpoznávacího rámce pro analýzu robustnosti ASR systému v šumovém prostředí s přetrénováním/adaptací 84
8.4 *WER* pro rozdílné množství dat pro přizpůsobení modelů 85

8.5 Průměrné *WER* v jednotlivých fázích procesu potlačení vlivu rušení . 88
8.6 Blokové schéma vývoje akustických modelů 91
8.7 *WER* pro zpětnou adaptaci 93
8.8 *WER* pro inkrementální adaptaci 93

9.1 Neřečové události v použitých databázích pro různé šumové podmínky 96
9.2 Fáze trénování pro čtené řečové databáze 98
9.3 Rozdělení akustického skóre pro zarovnané neřečové události řečníka . 102
9.4 Zlepšení přesnosti rozpoznávače v závislosti na výskytu události `FIL` . 105

A.1 Odhad průměrného SNR v trénovací části fragmentů SPEECON . . . 126

Seznam tabulek

4.1 Základní popis akustických modelů 24
4.2 Základní popis akustických modelů 25
4.3 Nastavení rozpoznávače plynulých promluv 25
4.4 Rozložení výslovnostních variant ve statistickém jazykovém modelu pro rozpoznávač číslovek 26
4.5 Trénovací sady dat 30
4.6 Objem dat jednotlivých fragmentů databáze SPEECON 30
4.7 Průměr odhadnutého SNR ve fragmentech databáze SPEECON . . . 30
4.8 Objem dat v hodinách v jednotlivých fragmentech databáze CZKCC 32

5.1 Konvence pro anotaci typických jevů ve spontánní promluvě 38
5.2 Popis anotovaných neřečových událostí 39
5.3 Souhrn obsahu databáze z pohledu zastoupených slov 40
5.4 Rozdělení slov v jednotlivých databázích 41
5.5 Výskyt vyplněných pauz v použitých databázích 41

6.1 Popis přizpůsobení trénovacích podmínek 53
6.2 Výsledná *WER* při trénování na rozdílných fragmentech SPEECON a testování na databázi SNR-set 53
6.3 Chybovost *WER* na databázi SPEECON bez doplňujících algoritmů potlačení šumu 55
6.4 Chybovost *WER* na databázi SPEECON s použitím ESS 56
6.5 Chybovost *WER* na databázi AURORA s použitím ESS v porovnání se standardy 57

7.1 Průměrné *ERS* a *ERP* pro jednotlivé kontrolní sady 66
7.2 Chybovost *WER* na databázi SPEECON s použitím algoritmů ESS a VAD pro trénovací data, a pouze ESS pro testovací data 71
7.3 Chybovost *WER* na databázi SPEECON s použitím algoritmů ESS a VAD pro trénovací i testovací data 72
7.4 Chybovost *WER* na databázi SPEECON s použitím algoritmů ESS a VAD pro trénovací data, a pouze ESS pro testovací data 72
7.5 Chybovost *WER* na databázi AURORA3 s použitím algoritmů ESS a VAD (s informací o energii) 73

7.6 Chybovost *WER* na databázi AURORA3 s použitím algoritmů ESS a VAD (bez informace o energii) 74

8.1 Vliv ESS algoritmu na *WER* pro různá prostředí při trénování modelů na čistých šumových podmínkách (OFFICE) 82
8.2 Vliv ESS algoritmu na *WER* pro různá prostředí při trénování modelů na obecných šumových podmínkách (ALL) 83
8.3 Průměrné hodnoty *WER* za oba kanály pro jednotlivé trénovací sady 83
8.4 Průměrné množství řečového materiálu v adaptačních množinách přes všechny fragmenty databáze SPEECON 85
8.5 Srovnání *WER* pro čisté a obecné výchozí modely – bez přizpůsobení 86
8.6 Srovnání *WER* pro čisté a obecné výchozí modely – s přetrénováním . 87
8.7 Srovnání *WER* pro čisté a obecné výchozí modely – s adaptací 88
8.8 *WER* – MLLR pro různé parametrizace s ESS+VAD při trénování . . 89
8.9 Srovnání průměrných hodnot *WER* pro čisté a obecné výchozí modely v různých fázích přizpůsobení modelů 89
8.10 *WER* při blokové adaptaci 92

9.1 Počet neřečových událostí řečníka v použitých databázích 97
9.2 Chybovost na úrovni slov (*WER*) a inzercí, jejich relativní zlepšení v jednotlivých fázích trénování modelů neřečových událostí oproti výchozímu systému 99
9.3 Průměrná délka trvání neřečových událostí 99
9.4 Rozdíl v množství zaměněných a smazaných neřečových událostí pro dvě varianty iniciálních modelů události `BRE` 100
9.5 Chybovost na úrovni mazání a vkládání vyplněných pauz pro čtené a kombinované trénovací sady 100
9.6 Výskyt jiných neřečových událostí ve zkoumaných databázích před a po zarovnání 101
9.7 Nastavení rozpoznávače plynulých promluv 103
9.8 Pravděpodobnost výskytu n-gramů s neřečovou událostí `FIL` 104
9.9 Výsledky rozpoznávání 105

B.1 Srovnání průměrných hodnot *WER* pro čisté a obecné podmínky trénování výchozích modelů – po adaptaci modelů 128

C.1 Zastoupení promluv v databázi pro jednotlivé mluvčí 130

D.1 Úspěšnost rozpoznávání pro parametrizaci RPLP 133

Seznam použitých zkratek

ASR	Automatic Speech Recognition *Automatické rozpoznávání řeči*
BFCC	Bark Frequency Cepstral Coefficients *Bark-frekvenční kepstrální koeficienty*
DCT	Discrete Cosine Transform *Diskrétní kosinová transformace*
DFT	Discrete Fourier Transform *Diskrétní Fourierova transformace*
ESS	Extended Spectral Subtraction *Rozšířené spektrální odečítání*
HMM	Hidden Markov Model *Skrytý Markovovský model*
IDFT	Inverse Discrete Fourier Transform *Inverzní diskrétní Fourierova transformace*
IDCT	Inverse Discrete Cosine Transform *Inverzní diskrétní kosinová transformace*
LDA	Linear Discriminant Analysis *Lineární diskriminační analýza*
LVCSR	Large Vocabulary Continuous Speech Recognition *Rozpoznávání spojité řeči s velkým slovníkem*
MAP	Maximum A Posteriori *Maximální odhad a posteriorní pravděpodobnosti*
MFCC	Mel-Frequency Cepstral Coefficients *Mel-frekvenční kepstrální koeficienty*
MFLP	Mel-Frequency Linear Prediction coefficients *Mel-frekvenční koeficienty s lineární predikcí*
ML	Maximum Likelihood *Maximalizace věrohodnosti*
MLLR	Maximum Likelihood Linear Regression *Maximálně věrohodná lineární regrese*

MMI Maximum Mutual Information
Maximalizace vzájemné informace

MMSE Minimum Mean Square Error
Minimalizace kvadratické chyby

PLP Perceptual Linear Prediction
Percepční lineární predikce

RASTA RelAtive SpecTrAl analysis
Analýza relativních spekter - RASTA filtrace

RPLP Revised Perceptual Linear Prediction
Revidovaná percepční lineární predikce

SS Spectral Subtraction
Spektrální odečítání

TRAP Temporal Pattern
Časový vzor (časová trajektorie)

TTS Text To Speech
Konverze textu na řeč (syntéza řeči)

VAD Voice Activity Detection
Detekce řečové aktivity

WER Word Error Rate
Míra chybovosti na úrovni slov

Kapitola 1

Úvod

Řeč je základní formou mezilidské komunikace a představuje nejpřirozenější způsob předávání informací. Za účelem zvýšení přirozenosti a pohodlí při komunikaci člověka se strojem je proto věnována významná pozornost hlasovým technologiím. V těchto technologiích jsou využívány systémy pro automatické rozpoznávání řeči (ASR – Automatic Speech Recognition), které představují hlasový vstup pro stroj či obecný systém a systémy pro syntézu řeči (TTS – Text To Speech), kde je realizován hlasový výstup strojem. Tato práce je zaměřena na dílčí úlohy automatického rozpoznávání řeči s aplikacemi jako jsou přepis promluv do textové podoby (diktáty), titulkování videa s textem (např. filmy, zpravodajství), vyhledávání v záznamech (rozsáhlé databáze dokumentů, video/audio banky) či pro obecné zjednodušení práce s PC a mobilními přístroji (ovládání aplikací, diktování SMS, apod.).

S rozvojem těchto hlasových technologií a jejich zpřístupněním i v přístrojích užívaných v běžném životě přichází potřeba provozovat tyto systémy v reálných podmínkách. Dnes již běžně dostupné rozpoznávače řeči obvykle dosahují vysokých úspěšností rozpoznávání v tichém prostředí nebo v málo se měnících (stacionárních) podmínkách. Reálné prostředí ovšem stále přináší situace, které více či méně snižují přesnost rozpoznávání Jedná se především o vliv šumového pozadí a působení rušivých jevů. Rozpoznávání řečového signálu navíc znesnadňuje proměnlivost promluvy různých mluvčích (inter-speaker variability) i proměnlivost promluvy od jednoho mluvčího (intra-speaker variability). Významný rozdíl v podobě promluvy také existuje mezi řečí čtenou a spontánně pronesenou, především vlivem nespojitostí v promluvě a použité intonace.

Za tímto účelem se na různé úrovni zpracování a modelování řečového signálu používají algoritmy potlačující vliv informace, která je pro rozpoznávání řeči irelevantní a na tyto algoritmy se tato práce zaměřuje.

Standardním vstupem rozpoznávačů řeči je parametrizovaná forma signálu, tedy příznaky vhodné pro rozpoznávání řeči s možností potlačení nežádoucích vlivů. V důsledku okolního rušení a dalších rušivých faktorů ale často dochází k degradaci signálu, kterou již nelze tímto způsobem eliminovat. Pro zvýšení robustnosti rozpoznávače v reálných podmínkách je proto aplikována řada algoritmů pro potlačování šumu a

zvýrazňování řeči. Ve fázi trénování akustických modelů mohou být vedle řečových elementů modelovány šumy a neřečové události, které jsou v signálu přítomny a není je možné odstranit předchozím předzpracováním. Při trénování těchto modelů se používají rozsáhlé řečové databáze, které vystihují variabilitu cílových podmínek, a tak umožňují přizpůsobit modely podmínkám, v nichž je rozpoznávač provozován.

Ačkoliv se jedná o metody standardně využívané, ideální řešení pro obecné podmínky reálného prostředí je obtížné nalézt, ať už z důvodu vysoké variability hluku prostředí, nebo z důvodu různorodosti vlivů, které na řečový signál působí. Omezujícím prvkem může být také složitost celého výpočetního postupu, případně čas potřebný k provedení potřebných kroků algoritmu. Problematika robustního rozpoznávání řeči se proto optimalizuje pro konkrétní podmínky. Příkladem může být rozpoznávač předem definované sady povelů nebo rozpoznávač řeči provozovaný v automobilu případně rozpoznávač adaptovaný na konkrétního mluvčího, resp. skupinu mluvčích.

Tato práce si klade za cíl objektivní zhodnocení vlivu standardních a modifikovaných parametrizačních technik v kombinaci s vybranými metodami pro potlačování šumu na úspěšnost automatického rozpoznávání řeči. Kapitola 2 dává přehled o tématice statistického rozpoznávání řeči, základních metodách používaných ve stávajících systémech a především o postupech využívaných v rámci této práce. Navazující shrnutí cílů práce a stručný přehled členění je uveden v kapitole 3.

V kapitole 4 jsou popsány základní parametry použitého rozpoznávače řeči a metodiky srovnávání chybovosti systému pro jednotlivá nastavení rozpoznávacích úloh. Dále navazuje popis použitých databází, vybíraných s důrazem na zastoupení reálného prostředí. Tuto kapitolu doplňuje kapitola 5 přibližující postup tvorby databáze spontánních promluv, která vznikla především jako doplněk stávajících databází s ohledem na větší zastoupení neřečových událostí řečníka.

Následující kapitoly se zabývají metodami pro zvýšení robustnosti rozpoznávače řeči. Každá kapitola obsahuje rozbor dané metody a poté experimentální část týkající se zkoumané problematiky. Analýzu nastavení ASR systému z pohledu předzpracování řečového signálu uvádí kapitola 6, kde jsou porovnány standardní a modifikované parametrizační techniky a jejich možné doplnění algoritmy pro potlačení šumu. Následuje kapitola 7 pojednávající o využití detekce řečové aktivity v rámci zpracování řečového signálu pro potlačení vlivu šumového pozadí přítomného v pauzách řeči. Experimentální část uzavírá kapitola 8, která doplňuje robustní nastavení v předzpracování signálu algoritmy v oblasti modelování. Kapitola 9 uzavírá tématiku robustnosti z pohledu modelování neřečových událostí.

Kapitola 2

Robustní rozpoznávače řeči - stav problematiky

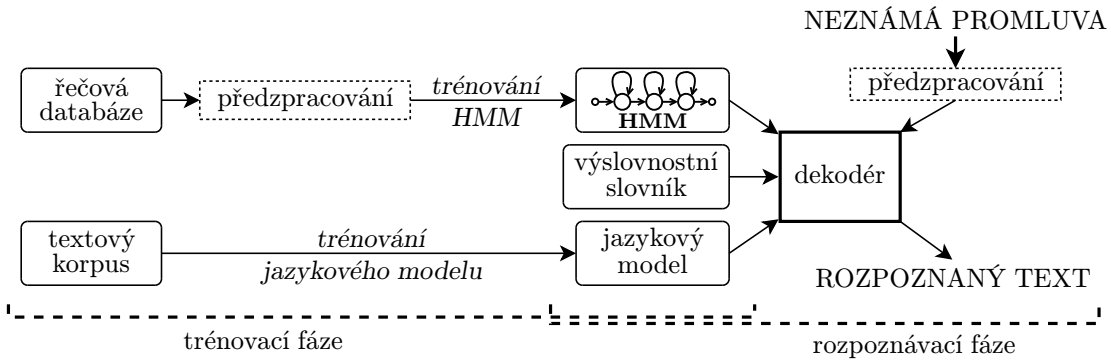
Teorie automatického rozpoznávání řeči zahrnuje problematiku z mnoha oblastí, od fyziologie hlasového traktu, např. [93], [35], přes analýzu akustických signálů [114] až po lingvistický rozbor jazyka ([14], [108]). Přestože existují publikace, které shrnují dílčí úlohy rozpoznávání řeči do jednoho celku ([88], [93]), jedná se i tak vždy jen o omezený výběr z rozsáhlé množiny aplikovaných procesů či používaných technik v rámci komplexní problematiky rozpoznávání řeči.

Úloha rozpoznávání řeči strojem sleduje rozvoj v oblasti výpočetních systémů a již s vývojem prvních digitálních počítačů v 50. letech se objevují také první systémy využívající dostupných a realizovatelných postupů rozpoznávací úlohy. Výzkum v oblasti rozpoznávání vzorů (pattern recognition) v 70. letech následně přispěl k uplatnění statistických přístupů a rozvoji rozpoznávání složitějších jazykových struktur. Podporovány stálým rozvojem výpočetních systémů, umožňují nynější ASR systémy realizovat potřebné operace ve velmi krátkých časech. Je tak možné výzkum orientovat nejen na základní úlohu rozpoznávání slov, ale také do oblasti rozpoznávání spojitých přirozených promluv (LVCSR – Large Vocabulary Continuous Speech Recognition). Mnoho takových systémů je již využíváno v praxi a to nejen pro angličtinu (např. [28]), ale i pro méně rozšířené jazyky, mezi něž lze řadit i češtinu ([1], [77], [76]).

Následující kapitola přináší přehled běžně využívaných principů rozpoznávání řeči s užším zaměřením na potřeby výzkumu a úloh řešených v rámci této práce.

2.1 Princip rozpoznávání řeči na bázi HMM

Dnešní rozpoznávače řeči jsou nejčastěji založeny na statistických metodách na bázi skrytých Markovovských modelů (HMM – Hidden Markov Model) [92]. Ty popisují signál z hlediska jeho statistických vlastností a umožňují tak modelovat i signály, jejichž podobu nelze vyjádřit deterministicky. Následující text přibližuje teorii rozpoznávání řeči na bázi HMM.



Obrázek 2.1: Blokové schéma standardního rozpoznávače řeči založeného na HMM

2.1.1 Statistický model řeči

Typickou strukturu rozpoznávače řeči založeného na statistickém modelování pomocí HMM lze znázornit blokovým schématem 2.1. Zde je vyznačena část akustického modelování, jejímž výstupem jsou modely akustické reprezentace promluvy. Část jazykového modelování pak popisuje skladbu jazyka na úrovni slov.

Základní princip rozpoznávání řeči je založen na hledání nejvěrohodnější posloupnosti slov $\tilde{\boldsymbol{W}} = w_1, w_2, ..., w_N$ ze všech možných posloupností $\mathcal{A}$ daných jazykovým modelem, které odpovídají naměřenému vektoru parametrů $\boldsymbol{O} = o_1, o_2, ..., o_T$ s použitím daných akustických modelů. Tedy

$$\tilde{\boldsymbol{W}} = \arg\max_{\boldsymbol{W} \in \mathcal{A}} \{P(\boldsymbol{W}|\boldsymbol{O})\}. \tag{2.1}$$

Pro určení pravděpodobnosti $P(\boldsymbol{W}|\boldsymbol{O})$ lze použít Bayesův vztah, který převádí tento problém na úlohu

$$\tilde{\boldsymbol{W}} = \arg\max_{\boldsymbol{W} \in \mathcal{A}} \left\{ \frac{P(\boldsymbol{O}|\boldsymbol{W})\, P(\boldsymbol{W})}{P(\boldsymbol{O})} \right\}. \tag{2.2}$$

Jelikož $P(\boldsymbol{O})$ je pro všechny kombinace $\boldsymbol{W} \in \mathcal{A}$ shodná, lze ve výsledku určit nejvěrohodnější kombinaci slov vztahem

$$\tilde{\boldsymbol{W}} = \arg\max_{\boldsymbol{W} \in \mathcal{A}} \{P(\boldsymbol{O}|\boldsymbol{W})\, P(\boldsymbol{W})\}. \tag{2.3}$$

První část ve vztahu (2.3), věrohodnost $P(\boldsymbol{O}|\boldsymbol{W})$, je dána modely základních akustických elementů řeči (*akustické modely*). Druhou část, pravděpodobnost $P(\boldsymbol{W})$ popisující pravděpodobnost výskytu dané kombinace slov $\boldsymbol{W}$, pak definuje *jazykový model*. Takto je popsán systém, který pro popis akustických vlastností řeči používá modely celých slov.

V případě, že jsou jako základní akustické elementy použity jednotky menší než slovo, např. slabiky či fonémy, je nutné doplnit systém také o informaci o posloupnosti těchto elementů v jednotlivých slovech. Tak je definována pravděpodobnost $P(\boldsymbol{S}|\boldsymbol{W})$

výskytu sekvence akustických elementů $\boldsymbol{S}$ v kombinaci slov $\boldsymbol{W}$. Tuto pravděpodobnost popisuje výslovnostní model, na schématu 2.1 reprezentovaný výslovnostním slovníkem. Celou rozpoznávací úlohu lze pak zapsat vztahem

$$\tilde{\boldsymbol{W}} = \arg\max_{\boldsymbol{W}\in\mathcal{A}} \{P(\boldsymbol{O}|\boldsymbol{S})\, P(\boldsymbol{S}|\boldsymbol{W})\, P(\boldsymbol{W})\}\,, \tag{2.4}$$

kde

- $P(\boldsymbol{O}|\boldsymbol{S})$ vyjadřuje věrohodnost generování vektoru $\boldsymbol{O}$ pro sekvenci akustických elementů $\boldsymbol{S}$,
- $P(\boldsymbol{S}|\boldsymbol{W})$ vyjadřuje pravděpodobnost sekvence akustických elementů $\boldsymbol{S}$, je-li dána posloupnost slov $\boldsymbol{W}$ pro daný výslovnostní model,
- $P(\boldsymbol{W})$ vyjadřuje pravděpodobnost posloupnosti slov $\boldsymbol{W}$ z pohledu daného jazykového modelu.

Zmíněné akustické a jazykové modely potřebné pro realizaci rozpoznávací úlohy se získávají v trénovací fázi přípravy rozpoznávače, jak je zachyceno na obr. 2.1.

Tyto tři výrazy vystihují základní funkční vstupy pro proces rozpoznávání řeči založený na modelování řečových jednotek kratších, než slovo. Vymezují proto také oblasti pro aplikaci algoritmů, které jsou používány pro zajištění robustnosti daného systému.

V následujícím textu jsou stručně popsány jednotlivé funkční oblasti procesu rozpoznávání řeči na bázi statistického modelu dle schématu 2.1.

2.1.2 Předzpracování signálu – Extrakce příznaků

Základní reprezentací řečového signálu v digitální podobě je vývoj akustického tlaku v časové oblasti snímaný mikrofonem do podoby elektrického signálu. Záznam v této podobě ale obsahuje velké množství redundantní informace, která není využita při procesu rozpoznávání [10].

Parametrizace řečového signálu popsaná níže převádí tento signál do podoby, která vystihuje charakteristiky významné pro rozpoznávání řeči a zároveň snižuje objem dat, který je předkládán klasifikačnímu systému. V této podobě (vektor $\boldsymbol{O}$ v předchozích vztazích) je řečový signál dále modelován akustickými modely. Navíc je průběh zpracování doplňován algoritmy, které mohou potlačovat rušení signálu.

Robustní předzpracování řečového signálu je tak významnou součástí ASR systémů pracujících v reálných podmínkách a v hlučném prostředí.

Standardně užívané reprezentace řečového signálu jsou založeny na znalosti produkce řeči hlasovým traktem, který se chová jako lineární systém buzený zdrojem měnícím se periodicky či náhodně v čase [95], a zpracování řeči lidským uchem, tedy spektrální analýze signálu ve vnitřním uchu [93]. Mezi nejčastěji používané parametrizace, které těchto vlastností využívají, patří mel-frekvenční kepstrální koeficienty (MFCC – Mel-Frequency Cepstral Coefficients) [16] a PLP koeficienty (PLP – Perceptual Linear Prediction) [40]. Tyto metody využívají podobných principů zpracování

signálu na bázi Fourierovy a kepstrální analýzy [94], a tak umožňují popsat řečový signál pomocí malého počtu parametrů. Díky nízké korelovanosti kepstra oproti spektru je tak lépe modelován vliv jednotlivých částí vokálního traktu případně i zdrojů šumu na výsledný řečový signál.

Rozsáhlou studii nastavení systémů založených na těchto technikách přináší např. [87]. Zde se autor vedle analýzy nastavení základních parametrů těchto metod věnuje také možnostem optimalizace objemu sady parametrů pomocí dekorelace a redukce dimenzí, viz také např. [38]. Tím může být dosaženo snížení počtu parametrů bez výrazného ovlivnění úspěšnosti rozpoznávače. Především ve výpočetně náročných LVCSR systémech tak vhodný přístup k redukci objemu dat může znamenat velmi výrazné snížení doby zpracování řečového signálu.

Vedle těchto metod byly vyvinuty další postupy, které pomáhají zvýšit robustnost ASR systému. Oproti výše uvedeným příznakům, které jsou založeny na krátkodobé spektrální analýze, využívají některé metody informace v delším časovém kontextu. Ukazuje se, že využití dlouhodobého vývoje signálu může přinášet informaci nejen o vlastním řečovém signálu, ale také umožňuje lépe pracovat s okolním ruchem, který signál doprovází [41]. Toho využívají například techniky TRAP [45] a TANDEM [43] [42]. I tyto metody alespoň částečně obsahují bloky zpracování používané u dříve uváděných standardních metod v kombinaci s využitím neuronových sítí.

V rámci výše zmíněných technik bylo publikováno mnoho dalších modifikací založených na algoritmech pro potlačování vlivu zkreslení signálu (PLP-RASTA [44]), doplnění či kombinace jednotlivých příznaků ([8]) nebo analýze fyziologie vnímání řeči lidským uchem [58].

Výsledky dosažené s těmito technikami ukazují, že optimalizací vybraných částí již standardizovaného procesu parametrizace lze dosáhnout zlepšení provozu systému ASR v daných podmínkách. Analýza a experimentální zhodnocení přínosu významných funkčních bloků metod MFCC a PLP je proto jednou z částí předkládané práce.

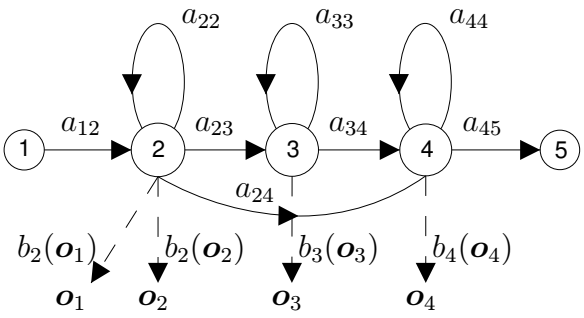
2.1.3 Akustické modelování

Akustické modelování umožňuje klasifikovat na bázi vybraných parametrů, který akustický element je předloženým signálem reprezentován. Při rozpoznávání řeči na bázi HMM je k tomu použito modelů, které reprezentují právě tyto jednotlivé akustické elementy, např. hlásky.

Obrázek 2.2 znázorňuje příklad 5-stavového Markovova modelu se dvěma neemitujícími stavy. Ten je charakterizován pravděpodobností přechodu mezi jednotlivými stavy a_{12}, a_{23} atd. a gaussovskou pravděpodobnostní funkcí $b_2(\boldsymbol{o}_t)$, $b_3(\boldsymbol{o}_t)$ a $b_4(\boldsymbol{o}_t)$ pro generování vektoru $\boldsymbol{o}_t$. Výstupní hustotní funkce $b_j(\boldsymbol{o}_t)$ je obecně charakterizována kombinací Gaussovských hustotních funkcí

$$\mathcal{N}(\boldsymbol{o};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{o}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{o}-\boldsymbol{\mu})} \tag{2.5}$$

kde n je dimenze příznakového vektoru $\boldsymbol{o}$, zatímco $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$ jsou parametry jedné komponenty modelu (střední hodnota a kovarianční matice).



Obrázek 2.2: Příklad 5-stavového skrytého Markovova modelu s dvěma neemitujícími stavy

Akustický model na bázi HMM popisuje pravděpodobnost $P(\boldsymbol{O}|M)$ s jakou je generován vektor parametrů $\boldsymbol{O}$ tvořený vektory $\boldsymbol{o}_t$ za předpokladu průchodu tímto modelem M přes jednotlivé stavy X, tedy

$$P(\boldsymbol{O}, X|M) = a_{12}b_2(\boldsymbol{o}_1)a_{22}b_2(\boldsymbol{o}_2)a_{23}b_3(\boldsymbol{o}_3)\ldots \tag{2.6}$$

V případě použití skrytých Markovových modelů jsou jednotlivé stavy průchodu neznámé, s pomocí rekurzivních postupů lze ale efektivně určit věrohodnost $P(\boldsymbol{O}|M)$ generování daného vektoru $\boldsymbol{O}$ modelem M. Pokud model M_i odpovídá slovu w_i resp. akustickému elementu s_i pak lze rovnice (2.3) resp. (2.4) řešit s předpokladem, že

$$P(\boldsymbol{O}|w_i) = P(\boldsymbol{O}|M_i) \tag{2.7}$$

resp.

$$P(\boldsymbol{O}|s_i) = P(\boldsymbol{O}|M_i) \tag{2.8}$$

Známe-li parametry modelu M, je možné řešit úlohu rozpoznávání řeči nalezením modelu M s maximální věrohodností pro dané pozorování $\boldsymbol{O}$. Získání parametrů modelu je řešeno procesem trénování.

Trénování akustických modelů

Parametry HMM se získávají procesem trénování, kdy se na základě trénovacích dat s pomocí Baum-Welchova algoritmu (více viz [123]) upřesňují hodnoty pravděpodobnosti přechodů mezi stavy a_{ij} a parametrů $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$ výstupní hustotní funkce b_i (viz vztah (2.5)).

Toto upřesnění může být realizováno na základě několika kritérií, z nichž nejčastějším kritériem je maximalizace věrohodnosti klasifikace řeči při použití správného modelu (ML – Maximum Likelihood). Vzhledem k omezením, která toto kritérium

předpokládá pro optimální odhad ([75]), existují i další kritéria, která zohledňují vzájemné rozdíly jednotlivých tříd akustických elementů a umožňují tak na rozdíl od ML zároveň snižovat chybnou klasifikaci, například kritérium minimalizace chybné klasifikace řeči přiřazením chybného modelu (MCE - Minimum Classification Error) [49] či maximalizace věrohodnosti klasifikace s ohledem na všechny možné výsledky klasifikace (MMI – Maximum Mutual Information) [6]. Na základě těchto kritérií jsou pak modely nejen trénovány, ale pomocí dalších algoritmů také adaptovány na konkrétního řečníka či prostředí.

Vlastní rozpoznávací proces je realizován Viterbiho algoritmem, který nalézá nejvěrohodnější cestu průchodu HMM sítí, více viz např. [123].

Subslovní akustické elementy pro modelování řeči

Je-li ASR úloha stanovena jako rozpoznávání omezeného počtu slov, která jsou neměnná, viz např. rozpoznávač číslovek, lze pomocí HMM modelovat celé slovo. Vektoru $\boldsymbol{O}$ pak odpovídá celé dané slovo, které je postupně generováno modelem. Trénovací databáze v takovém případě musí obsahovat dostatečné a reprezentativní zastoupení každého slova, aby byl model schopen generalizovat. Ve chvíli, kdy je potřeba rozšířit množinu rozpoznávaných slov nebo ji operativně přizpůsobovat konkrétnímu úkolu, je již nevhodné trénovat pro každé slovo vlastní model. Nejen že neúměrně narůstá počet modelů, rozsah a požadavky na trénovací databázi, navíc stále není možné rozpoznávat slova, pro která neexistuje model.

Z toho důvodu jsou pro složitější úlohy využívány subslovní akustické elementy, nejčastěji hlásky (fóny). Skládáním těchto elementů za sebou jsou pak tvořena potřebná slova. Pro češtinu používáme 44 základních hlásek [104]. Jejich podoba je ovšem velmi variabilní a závisí nejen na daném mluvčím, ale vlivem setrvačnosti mluvidel i na okolních hláskách. Tento jev, nazývaný koartikulace, je proto často vystihován modelováním monofónů s kontextovou závislostí, které již zahrnují informaci o sousedních fonémech, např. trifónů. Ukazuje se, že především ve složitějších úlohách rozpoznávání spojitých promluv je tato informace důležitá a napomáhá vyšší kvalitě rozpoznávání. Při některých jednodušších úlohách nebo v situaci, kdy není dostatek materiálu pro dostatečné natrénování extenzivní sady takových modelů, však může být žádoucí použití méně složitých modelů [124].

I v této práci jsou proto využívány varianty modelovaných elementů dle charakteru úlohy. V úlohách, které analyzují vliv algoritmu extrakce příznaků pro zpracování řečového signálu, je potřeba především vystihnout vliv zkoumaného faktoru, bez nutnosti zabývat se více vlastním nastavením složitějšího modelování. Proto je pro tyto úlohy často využíváno rozpoznávače řeči s malým slovníkem s využitím modelů monofónů. Až pro zhodnocení přínosu použitých metod na reálný běh rozpoznávače řeči je vhodné rozšiřovat také složitost vlastního rozpoznávače, např. použitím trifónů či zvýšením komplexnosti jazykového modelu.

2.1.4 Výslovnostní slovník

Je-li ASR systém založen na modelování akustických jednotek kratších než slovo, je potřeba definovat, jakým způsobem tyto jednotky tvoří výsledné slovo. Tuto informaci obsahuje výslovnostní slovník, který umožňuje vícenásobnou definici výslovnosti slova (např. pro rozlišení spisovné a hovorové podoby výslovnosti).

Výslovnostní slovník také definuje množinu všech slov, jež se mohou vyskytnout na výstupu rozpoznávače. Jedná-li se o rozpoznávač omezeného počtu slov, například povelů pro ovládání konkrétního zařízení, je tento slovník velmi malý a jeho rozsah připouští manuální zásahy pro zvýšení přesnosti modelování výslovnosti. S rostoucím objemem slovníku se ovšem možnosti ručního zásahu do definice výslovnosti snižují a pro rozpoznávače spojitých promluv je již nutná automatická tvorba slovníku na základě pevných pravidel. S tím je ovšem spojeno vysoké riziko špatné definice kvůli nepravidelnostem, jakou je například výslovnost cizích slov. Kvalita takového modelu pak klesá a s ní se snižuje i úspěšnost rozpoznávání.

2.1.5 Jazykové modelování

Je-li definována akustická podoba jednotlivých slov na bázi výslovnostního slovníku, lze jejich skládáním získat konečnou podobu informace v řečovém projevu. Skutečnost, jak jsou tato slova složena dohromady, je ovlivněna gramatikou daného jazyka, kterou vystihuje jazykový model.

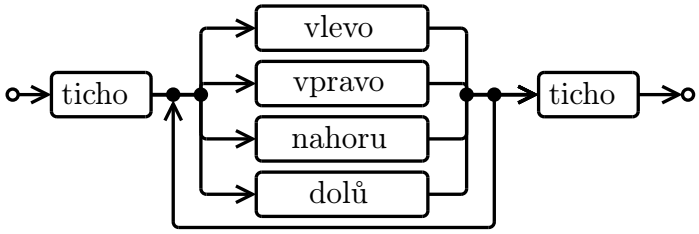
Jazykový model popisuje, jak pravděpodobná jsou jednotlivá spojení dvou či více slov definovaných ve výslovnostním slovníku a které kombinace slov jsou v dané gramatice přípustné. Reflektuje tak skutečnou gramatiku jazyka, ale umožňuje také přizpůsobit rozpoznávač jazykovým podmínkám, v nichž je ASR systém používán.

Gramatika

Tak jako výslovnostní slovník definuje povolenou posloupnost jednotlivých hlásek tvořících dané slovo, gramatika definuje povolený sled jednotlivých slov na výstupu rozpoznávače. V případě rozpoznávače řeči s malým slovníkem (rozpoznávač omezeného počtu slov, povelů) jsou pro tuto definici většinou použity gramatiky v podobě slovních smyček. Ty definují, která slova mohou po sobě následovat. Předpokladem pro takový model je, že pravděpodobnost výskytu všech slov smyčky je shodná.

Častým uplatněním definice jazykového modelu v podobě gramatiky je rozpoznávač omezené množiny povelů. Gramatika je v této úloze popsána smyčkou s jedním průchodem, která obsahuje všechny známé povely, z nichž jeden je vybrán jako nejvěrohodnější vzor pro rozpoznávané slovo.

Komplexnější podobu má gramatika pro rozpoznávač sekvence povelů nebo číslovek. Jedná se již o model, který simuluje běh řeči s malým slovníkem. Proto je takové nastavení často používáno jako první krok při analýze algoritmů pro rozpoznávání řeči. Většina experimentů této práce je z tohoto důvodu založena právě na úloze



Obrázek 2.3: Příklad jednoduché gramatiky rozpoznávače povelů

rozpoznávání sekvence číslovek. Tento přístup také koresponduje s nastavením referenčního rozpoznávače řeči pro projekt Aurora [83] definovaný jako standardizovaný rámec pro porovnání systémů, který je v práci také použit.

Statistický jazykový model

Pro úlohu rozpoznávání spojitých promluv s velkým slovníkem se již využívají statistické modely jazyka. Ty modelují pravděpodobnost výskytu konkrétní kombinace slov na základě jazykového modelu trénovaného na textech, které věrohodně vystihují charakter daného jazyka. Vzhledem ke složitosti obecného jazyka pak dochází k optimalizaci tohoto modelu, aby bylo jeho použití v rámci rozpoznávače možné z pohledu výpočetní náročnosti i úspěšnosti klasifikace. Významný vliv na složitost modelu má i charakter jazyka, jehož typickou vlastností může být ohebnost slov. Přístup k tvorbě jazykových modelů je proto velmi často vlastnímu jazyku přizpůsoben. To je případ i modelů používaných pro slovanské jazyky ([48], [103]).

Charakter spojitých promluv modelovaných statistickým jazykovým modelem se v mnohém odlišuje od promluv jednoduchých, jako je sekvence povelů. Proto lze očekávat, že i vliv jednotlivých algoritmů pro zvýšení robustnosti rozpoznávače řeči může vykazovat jisté odlišnosti pro různé typy promluv.

V této práci je pro analýzu vlastností algoritmů pro zpracování řeči použito nejprve základní gramatiky obsahující pouze číslovky. Následně jsou získané závěry ověřeny na obecných spojitých promluvách s použitím statistického jazykového modelu [86]. To vede ve výsledku k závěrům, které popisují možnost využití navrhovaných postupů v LVCSR systémech.

2.2 Robustnost rozpoznávače řeči

Proces parametrizace popsaný v sekci 2.1.2 nalézá příznaky v řečovém signálu, které jsou významné pro rozpoznávání řeči. Tak je možné eliminovat některé jevy, jejichž přítomnost v signále je náhodná nebo není spjatá s informací obsaženou v řeči, např. variabilita promluv jednotlivých mluvčích, intonace apod. V případě zašuměného signálu ale není toto předzpracování dostatečné, a proto existuje mnoho algoritmů, které potlačují vliv nežádoucího rušení na výsledek rozpoznávání [34].

Tato sekce popisuje vybrané oblasti, v rámci kterých je možné ASR systém rozšířit pro zvýšení účinnosti rozpoznávání v reálných šumových podmínkách a které jsou dále analyzovány v rámci této práce. V první části se jedná o algoritmy aplikované při předzpracování řečového signálu, druhá část pak popisuje možnosti zvýšení robustnosti systému na úrovni modelování řeči.

2.2.1 Robustnost při extrakci příznaků

Na rozpoznávaný signál může působit mnoho rušivých vlivů, které degradují kvalitu signálu a tím snižují schopnost systému správně klasifikovat obsaženou informaci. Jejich účinek lze pozorovat v několika místech cesty signálu od řečníka k rozpoznávači. Souhrnně lze při zahrnutí všech aditivních a konvolučních složek modelovat zašuměný signál jako

$$\hat{s}[n] = s[n] * h[n] + n[n], \tag{2.9}$$

kde $\hat{s}[n]$, $s[n]$, $h[n]$ a $n[n]$ reprezentují postupně výsledný zašuměný signál, původní nezašuměný signál, vliv konvoluční složky na signál a aditivní složku rušení.

Jsou-li akustické modely trénovány na obdobných podmínkách, v jakých je rozpoznávač provozován, a jsou-li tyto podmínky v čase neměnné (shodný mikrofon, shodné prostředí, apod.), je možné analyzovat signál jen s ohledem na aditivní složku rušení. Ve většině případů lze toto zjednodušení akceptovat, případně použít metody, které působí proti změnám signálu konvolučního charakteru. Ty se používají mimo jiné pro potlačení vlivu rozdílných charakteristik přenosové cesty (např. odlišný mikrofon) a patří sem například metoda kepstrálního odečítání (CMS - cepstral mean subtraction)[4]. Odlišný problém také tvoří úloha, kdy je ve snímané řeči ozvěna, jejíž vliv je snižován metodami pro potlačování echa, např. [9], případně je-li v dosahu řečníka jiný mluvčí. Poté je potřeba použít např. techniky pro separaci zdrojů [39].

S ohledem na rozsah jednotlivých oblastí potlačování rušení se zaměřuje tato práce jen na první typ - aditivní šum. Signál $s[n]$ ovlivněný pouze touto *aditivní* akustickou složkou lze zapsat jako

$$x[n] = s[n] + n[n]. \tag{2.10}$$

Ve frekvenční oblasti se pak tato rušivá složka projeví také jako aditivní, tedy

$$X(e^{j\Theta}) = S(e^{j\Theta}) + N(e^{j\Theta}), \tag{2.11}$$

kde $X(e^{j\Theta})$, $S(e^{j\Theta})$ a $N(e^{j\Theta})$ jsou krátkodobé odhady spektrálního obrazu signálů $x[n]$, $s[n]$ a $n[n]$.

Metody potlačující aditivní rušení jsou proto nejčastěji založeny na předpokladu, že se charakteristiky signálu a rušení významně liší. Na základě této odlišnosti může být jedna ze složek potlačena nebo zvýrazněna. Přesnost odhadu pak určuje, do jaké míry je tímto zásahem ovlivněna druhá složka. Odhad charakteristik jedné ze složek signálu je proto důležitou součástí algoritmů pro potlačování šumu. Nejčastějším přístupem bývá analýza segmentů řečových pauz získaných na základě detekce řečové

aktivity. Vedle této metody mohou být použity například sady předpokladů o šumovém prostředí, které popisují předpokládanou podobu charakteristik šumu, například pásmové omezení a tvar spektra šumu.

Nejjednodušším způsobem může být nalezení vhodné převodní funkce mezi podobou čisté a zašuměné řeči, tzv mapování parametrů [72, 71]. S její pomocí lze následně zašuměný signál transformovat na odpovídající čistý řečový signál a ten poté použít v rozpoznávači nezašuměných promluv. Nevýhodou této metody je ale potřeba čisté podoby řečového signálu a vysoká závislost na typu rušení. Naopak obdobný přístup na úrovni modelování řečového signálu je s výhodou využíván v metodách pro adaptaci modelů, které jsou analyzovány i v této práci, viz kap. 2.2.2.

V oblastech se známým prostředím, u kterého lze předem definovat jeho vlastnosti, nachází místo metoda filtrace nežádoucích složek signálu. S pomocí jednoduchého filtru lze docílit omezení těch složek signálu, v nichž je očekáván vyšší vliv rušení. Příkladem může být filtr pro potlačení rušení v automobilu, které se vyznačuje harmonickou strukturou s násobky na polovině základní frekvence [54]. Ve standardních parametrizačních metodách, jejichž analýza je součástí kapitoly 6 této práce, je také často využito preemfázového filtru. Ten naopak pomáhá zvýraznit složky řečového signálu na vyšších frekvencích a vyrovnat tak energetickou hladinu signálu v celé šířce zkoumaného spektra.

Vztah (2.11) ukazuje, že aditivní rušení lze ve frekvenční oblasti pozorovat také jako aditivní složku. To vede k metodám, které se snaží tuto složku odhadnout a na základě tohoto odhadu šumovou složku odstranit. Jednotlivé metody se proto liší především ve způsobu, jakým je odhad proveden.

Wienerova filtrace

Princip Wienerovy filtrace [60, 121] spočívá ve filtraci zašuměného signálu tak, aby došlo k odstranění šumové složky signálu na základě minimalizace střední kvadratické chyby, viz obr. 2.4. Pro přenosovou funkci Wienerova filtru platí

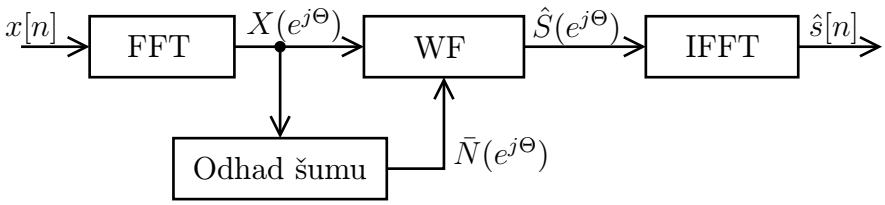
$$H(e^{j\Theta})^2 = \frac{|S(e^{j\Theta})|^2}{|S(e^{j\Theta})|^2 + |N(e^{j\Theta})|^2} \tag{2.12}$$

Za předpokladu, že šum a řeč jsou statisticky nezávislé stacionární procesy, Wienerův filtr poskytuje optimální odhad čistého signálu ze zašuměného.

Máme-li k dispozici odhad výkonového spektra šumového pozadí $\bar{N}(e^{j\Theta})$, lze původní nezašuměný signál rekonstruovat ze zašuměného signálu Wienerovou filtrací dle vztahu

$$H(e^{j\Theta})^2 = \frac{|S_n(e^{j\Theta})|^2 - |\bar{N}(e^{j\Theta})|^2}{|S_n(e^{j\Theta})|^2} \tag{2.13}$$

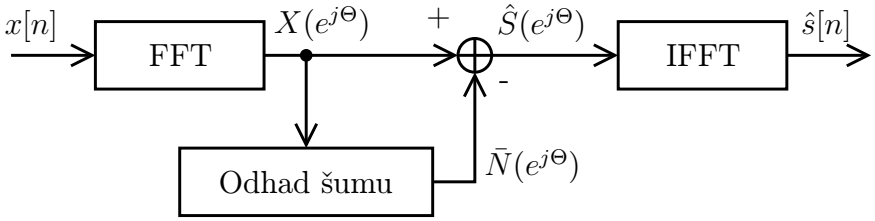
V adaptivní verzi této metody je pak možné sledovat i šumové charakteristiky v různých podmínkách a vlastnosti filtru přizpůsobovat změnám prostředí. I tato metoda má však v reálných podmínkách omezenou schopnost potlačit rušení přítomné v signálu, neboť reálné prostředí nesplňuje podmínky, za kterých algoritmus dostatečně přesně odhadne charakteristiky rušivého pozadí.



Obrázek 2.4: Blokové schéma potlačení šumu pomocí Wienerovy filtrace

Spektrální odečítání

Aditivní šumová složka signálu je ve spektrální oblasti charakterizována také aditivní složkou, viz vztah (2.11). Spektrální odečítání ([121], [50]) odhaduje čistý signál tak, že odečítá odhadnuté amplitudové spektrum šumu od zašuměného signálu (při zachování fázové složky). Základní blokové schéma metody je znázorněno na obr. 2.5



Obrázek 2.5: Blokové schéma potlačení šumu spektrálním odečítáním

Odhad šumové složky pro aplikaci spektrálníhho odečítání může být proveden například na základě analýzy pauzy v řeči za pomoci detektoru řečové aktivity (VAD). Možné selhání tohoto algoritmu vlivem nesprávné detekce v zarušených podmínkách vede ale k negativnímu ovlivnění řečového signálu a degradaci úspěšnosti rozpoznávače. Kvalitní VAD algoritmus navíc zvyšuje komplexnost systému, což může být překážkou k jeho použití.

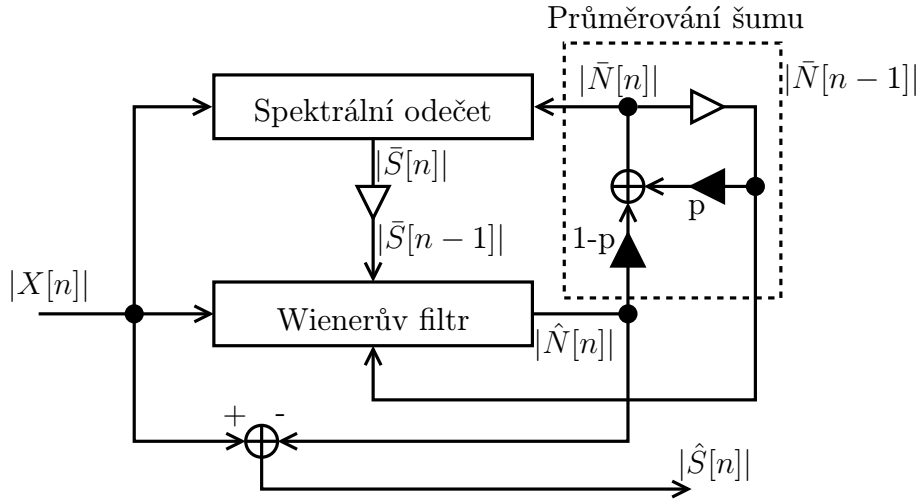
Jinou metodou pro odhad charakteristik šumové složky signálu je využití odhadu pomocí některých iterativních metod, např. minimální statistiky ([62], [63]). Výhodou této techniky je vedle její jednoduchosti také snadná implementovatelnost do procesu extrakce příznaků.

Rozšířené spektrální odečítání

Metoda ESS je kombinací Wienerovy filtrace a spektrálního odečítání, kdy je pro odhad šumové složky signálu využit adaptivní Wienerův filtr, jehož koeficienty jsou aktualizovány dle výstupního signálu. Tento odhad je následně ve použit v rámci spektrálního odečtu. Algoritmus proto nevyužívá VAD a tím eliminuje i nevýhody přístupů, jejichž úspěšnost závisí na přesnosti detekce řeči. Principiální blokové schéma algoritmu ESS znázorňuje obr. 2.6.

Technika ESS zahrnuje možnost nastavení parametrů, které upravují chování algoritmu – konstantu α pro kompenzaci vlivu odhadu charakteristik šumového pozadí na nastavení Wienerova filtru a konstantu p pro nastavení rychlosti reakce na změny v signálu. Tak je možné algoritmus přizpůsobit daným podmínkám.

Vzhledem k zaměření práce na rozpoznávání řeči v obecných reálných podmínkách je algoritmus ESS analyzován z pohledu jeho přínosu k úspěšnosti rozpoznávání řeči v těchto podmínkách.



Obrázek 2.6: Blokové schéma potlačení šumu metodou ESS

Jiné metody

Vedle výše uvedených metod existují i další algoritmy pro odhad charakteristik šumového pozadí, např. MMSE – Minimum Mean Square Error estimation [19], MLE – Maximum Likelihood Spectral Estimation [67], MPE – Maximum a Posteriori Spectral Estimation [122] či zvýrazňování řeči na bázi modelů [54]. Článek [8] ukazuje také využití informace z VAD jako doplněk existujícího vektoru parametrů, kde je v prostředí automobilu dosaženo dvouciferných zlepšení chyby rozpoznávače. Znalost vlastností rušivého prostředí lze výhodně využít různým postupem zpracování signálu v jednotlivých frekvenčních pásmech [20].

Spolu s přínosem ke snížení SNR v řečovém signálu ale mohou tyto metody, především ty jednodušší, ovlivnit signál vlastním šumem algoritmu. Tak vzniká například tzv. hudební šum [12]. Vyhodnocení vlivu těchto algoritmů je proto nutné provádět nejen na úrovni měření SNR. V této práci je přínos metody ESS hodnocen z pohledu přínosu metody k zlepšení klasifikace rozpoznávačem řeči.

Detekce řečové aktivity při odhadu charakteristik šumového pozadí řeči

Výše zmíněné algoritmy často v rámci zpracování signálu využívají detekce řečové aktivity k odhadu charakteristik šumového pozadí. Algoritmy pro detekci řeči proto často bývají významným prvkem ovlivňujícím úspěšnost celého systému v daných podmínkách.

Algoritmy detekce řeči jsou založené typicky na výkonové analýze signálu, spektrální či kepstrální analýze resp. koherenční analýze. Nejjednodušší formy detekce řeči zkoumají energii signálu nebo počet průchodů nulou [64], [52]. Jejich výhodou je velmi nízká výpočetní náročnost, naopak nevýhodou je vysoká chybovost v případě detekce řeči v šumovém prostředí. Spolehlivější algoritmy pro detekci jsou založeny na spektrálních (kepstrálních) vzdálenostech mezi řečovým signálem a pozadím řeči [37]. V případě zmíněných detektorů se obvykle zjišťuje míra odlišnosti daného bloku signálu od pozadí v dané oblasti (energie, spektrum, entropie spektra [119]). O vlastním výsledku detekce lze pak rozhodnout porovnáním této míry s prahovou hodnotou, kterou lze stanovit globálně jako fixní práh, či ji adaptivně obnovovat [111] podle aktuálních charakteristik pozadí řeči, případně lze využít více sofistikovaných rozhodovacích stromů [64]. Pro telekomunikační systémy se využívají algoritmy, které kombinují několik různých prvků pro zvýšení efektivity detekce [47], [21], to však v úloze rozpoznávání řeči není vhodné.

Pro prostředí s velmi vysokou hladinou rušení (např. jedoucí automobil) se dále používají vícekanálové metody [102]. S jejich nasazením ale současně vzrůstá výpočetní náročnost detekčních algoritmů a potřeba vícekanálového nahrávání zvyšuje také hardwarové požadavky systému.

Vedle výše zmíněných postupů mohou být využity i metody na bázi statistického zpracování signál, např. využití GMM modelů řeči a šumu [52], případně umělých neuronových sítí pro nelineární mapování mezi vektorem řečových příznaků a přítomností řeči [33].

Použití VAD ovlivňuje míra zarušení signálu, ale také požadavky na výpočetní náročnost. Jelikož tyto požadavky jdou většinou proti sobě, použitý VAD může významně ovlivnit úspěšnost celého procesu. Vzhledem k využití kepstrální reprezentace signálu pro účely rozpoznávání jsou v této práci jsou použity algoritmy detekce řeči založené na těchto charakteristikách signálu, které při zachování nízké náročnosti algoritmu dosahují úspěšnost detekce použitelné v úloze zpracování řeči.

2.2.2 Robustnost v akustickém modelování

Podobně jako volba vhodné parametrizace přispívá ke zvýšení robustnosti rozpoznávače především v podmínkách s rušivými vlivy, lze kvalitou akustických modelů významně ovlivnit schopnost systému správně rozpoznat zkreslenou příchozí promluvu.

Trénování/adaptace modelů

Metody zvyšování robustnosti ve fázi trénování akustických modelů či vlastní klasifikaci lze rozdělit do dvou tříd. Do první skupiny lze zahrnout metody, které využívají předchozí znalosti o podmínkách, ve kterých bude rozpoznávač provozován. V takovém případě lze akustické modely přizpůsobit těmto podmínkám předem v rámci procesu **trénování** a získat robustní modely použitelné pro standardní rozpoznávací proces. Druhou skupinu tvoří metody využívající zašuměná řečová data z reálného provozu rozpoznávače, na kterých jsou akustické modely dále **adaptovány**. Typicky se tyto metody využívají pro adaptaci systému nezávislého na mluvčím na konkrétního řečníka. Vhodnou volbou adaptační sady dat lze ale docílit také přizpůsobení modelů konkrétnímu prostředí.

- **Trénování**
  Mezi metody využívající proces trénování patří především:

  - (pře)trénování akustických modelů čisté řeči na řeči získané v koncových podmínkách,
  - (pře)trénování na signálech nahraných ve více typech rušivých podmínek [70], [61],
  - kompozice a dekompozice HMM [120],
  - paralelní kombinace akustických modelů [31].

  Ačkoliv ve výsledku přináší tyto metody často velmi efektivní řešení, možnost jejich nasazení je výrazně ovlivněna potřebou velkého množství trénovacího materiálu pořízeného v daném cílovém prostředí. Sběr takových dat je spojen s vysokými náklady a navíc výsledný materiál nezaručuje dostatečnou reprezentativitu. Například nahráváním řečových dat v automobilu lze získat poměrně rozsáhlý vzorek dat, ale ani ten nevystihne všechny aspekty reálného provozu automobilu. Nejedná se jen o zastoupení rušivého pozadí, ale také o potřebu vyváženosti databáze s ohledem na výskyt foneticky bohatého materiálu.

  Potřebného vyvážení databáze bez neúměrného zvýšení náročnosti sběru dat lze docílit využitím existujících foneticky bohatých databází čisté řeči a jejich kombinací se samostatně nahrávaným šumem. Zmíněný postup se využívá například při tvorbě rozpoznávače pro prostředí automobilu [11], [2]. I v tomto případě se ale často využívá alespoň z malé části dat reálně nahraných.

- **Adaptace**
  Mezi standardně používané adaptační techniky patří především metody založené na transformaci modelů na základě lineární regrese (Maximum-Likelihood Linear Regression – MLLR) [57] a metody na bázi MAP adaptace (Maximum A Posteriori adaptation) [32].

Zatímco metoda MAP je blízká metodě přetrénování modelů na cílových podmínkách a vyžaduje relativně objemný adaptační materiál, metoda MLLR umožňuje získat adaptované modely na malém počtu záznamů. Například autoři v [105] dosahují zlepšení chyby na úrovni slov o 9,2 % oproti výsledku s použitím trénování na rozličných podmínkách a zlepšení o 25 % oproti trénování na čistých podmínkách.

Modelování neřečových událostí

Významným a v dnešní době již standardně zohledňovaným aspektem rozpoznávání řeči jsou také neřečové události generované řečníkem (smích, kašel...), či specifické nestacionární či impulsivní jevy z okolí (bouchnutí dveří, vyzvánění telefonu).

Především při modelování rušení od řečníka lze využít faktu, že toto rušení přichází typicky v pauzách promluvy a lze k němu proto přistupovat jako k ostatním řečovým událostem (hláskám). Při trénování modelů segmentů řeči jsou v trénovacím materiálu zohledňovány i tyto události a je jim přiřazen speciální model. Z toho také plyne požadavek na spolehlivé označení těchto událostí v trénovací databázi.

Jelikož jazykový model v podobě gramatiky je velmi jednoduchý, jeho doplnění o tyto události v úloze rozpoznávání řeči s malým slovníkem je snazší [25]. V dnešní době jsou již řečové databáze o tyto informace doplněny, proto bývá v úloze ASR s jednoduchou gramatikou modelování neřečových událostí standardně zahrnuto.

Na druhou stranu při rozpoznávání spojitých promluv je nutné komplexnější řešení vzhledem ke složitosti statistického jazykového modelu ([17], [110], [115]). Úloha je pak často spojena s potřebou detekce těchto událostí [51], případně s modelováním dalších vlastností, které jsou se spojitou promluvou svázány [107].

Detekce řečové aktivity při modelování řeči - frame dropping

Detekce řečové aktivity byla zmíněna v sekci 2.2.1 jako metoda pro nalezení neřečových segmentů, na základě nichž jsou odhadovány charakteristiky šumového pozadí řeči. Další možné využití detektoru je v úloze odstranění neřečových segmentů signálu. Tak je možné v ideálním případě rozpoznávači předložit jen řečový signál, aby nedocházelo na chybně klasifikovaných úsecích v pauzách řeči k nesprávně vkládané informaci – tzv. vypouštění segmentů (frame dropping), viz např. [3]. Podobně může být této metody použito i pro odstranění jiných nežádoucích segmentů, které obsahují neřečovou informaci, jako je například vyplněná pauza [116].

I tak je ovšem v hlučných podmínkách výsledná detekce zatížena chybou, která může být při použití VAD pro odstraňování segmentů signálu kritická a negativně ovlivnit kvalitu systému, viz např. [68]. Proto je potřeba zajistit, aby byla tato chyba potlačena alespoň za cenu snížení citlivosti detektoru. V této práci je proto metoda vypouštění segmentů použita v souladu s tímto předpokladem tak, aby případný negativní vliv metody nepřispíval k navýšení chybovosti systému.

Kapitola 3

Cíle práce

Obecným cílem této práce je podrobná analýza vlastností algoritmů pro zvýšení robustnosti rozpoznávače řeči a návrh vhodných metod či optimalizovaného nastavení rozpoznávače pracujícího v reálných podmínkách. Následující sekce shrnuje hlavní body práce z pohledu významu pro robustní rozpoznávání řeči a přibližuje motivaci pro zkoumání dané oblasti. Poté následuje přehled konkrétních cílů práce.

Robustní předzpracování signálu přispívá k celkové robustnosti systému nejen nalezením příznaků vhodných pro správnou klasifikaci řečových elementů, ale také možností účinně potlačovat vliv nepříznivých podmínek, například rušení z okolí. Vlastnosti standardně využívaných metod MFCC a PLP jsou z mnoha pohledů důkladně zanalyzovány [87] a v rámci optimalizací jsou publikovány postupy, které vedou k dalšímu zvýšení robustnosti těchto parametrů. Mnohé z nich vedou na výraznější zásahy do vlastního procesu [58] a tím i na zvýšení výpočetní náročnosti algoritmu. V aplikacích, kde je kladen důraz na udržení nízkých nároků výpočtu, může být naopak vhodné využít stávajících funkčních bloků zpracování signálu a v rámci nich hledat možnosti optimalizace ([66], [46]).

Použití metod pro potlačení aditivního rušení může vést nejen ke zvýšení odolnosti rozpoznávače proti aditivnímu rušení, ale také k výraznému nárůstu složitosti celého systému. Proto je nalezení vhodných metod pro dané cílové podmínky důležitým bodem návrhu ASR systému. **Rozšířené spektrální odečítání** aplikuje zjednodušený proces spektrálního odečítání při zachování dobrých schopností potlačit rušení. Bez nutnosti použití VAD je tak možné získat parametry, které významně potlačují vliv aditivního šumu.

Při použití VAD jako zdroje informace pro techniku **vypouštění segmentů** signálu [3] lze vhodným nastavením parametrů detektoru výrazně snížit výskyt rušivých segmentů v signálu. Díky tomu je eliminován jejich vliv na chybnou klasifikaci těchto segmentů. Na druhou stranu je důležité, aby detekcí nebyla výrazně ovlivněna řečová část signálu.

Výběr vyvážené a přitom dostatečně obsáhlé sady trénovacích dat významně ovlivňuje kvalitu výsledných akustických modelů. I přesto však není možné trénováním postihnout rozmanitost reálných podmínek. Proto je důležitým bodem při tvorbě

rozpoznávače řeči možnost přizpůsobení již existujících modelů cílovým podmínkám. Transformace obecných akustických modelů na základě reprezentativních řečových vzorků z cílového prostředí je základní myšlenkou **adaptačních procesů**. Algoritmus MLLR [57] bývá využit jako první stupeň adaptace pro případ malého množství dat, resp. pro iniciální adaptaci před nasazením náročnějších technik. Přes tuto deklarovanou jednoduchost se jedná o metodu, která významným způsobem přináší možnost zahrnout do procesu modelování právě reálné podmínky provozu rozpoznávače. V kombinaci s efektivním předzpracováním je tak možné vytvořit systém odolný vůči rušivým podmínkám, které mají charakter převážně aditivní.

Modelování neřečových událostí je standardní součástí systémů pro rozpoznávání řeči. Při zahrnutí těchto událostí do rozpoznávacího procesu je proto nutné zohlednit také kvalitu trénování těchto modelů. To významně ovlivňuje nejen vlastní proces trénování, ale také nutnost získat kvalitní trénovací data, obsahující požadovanou informaci o neřečových událostech.

Realitou využití dnešních rozpoznávačů řeči jsou zejména aplikace rozpoznávání spojitých promluv. S tím jde ruku v ruce také potřeba využití **databází spontánních promluv**, které vystihují charakter souvislé promluvy lépe, než řeč čtená. Z pohledu vytváření kompletního trénovacího materiálu se jedná o náročnější proces, neboť na rozdíl od čtené řeči není předem znám obsah promluvy. Na druhou stranu je sběr těchto dat jednodušší díky možnosti využití volně dostupných zdrojů (rádio, TV, internet). Analýza procesů a podmínek, které je nutné při přípravě takové databáze zajistit, je proto důležitou znalostí pro správnou přípravu trénovacího materiálu.

Na základě výše uvedených shrnutí byly stanoveny základní cíle této práce.

- *Optimalizace nastavení parametrů standardní parametrizace MFCC a PLP s ohledem na robustnost vůči šumovému pozadí.*
  Nastavení základních parametrů předzpracování řečového signálu pomocí technik MFCC a PLP ovlivňuje schopnost rozpoznávače zvýraznit důležité příznaky pro klasifikaci řečových elementů, ale na druhou stranu také jiné elementy potlačit. Jedním ze základních cílů prezentované práce je proto analýza nastavení těchto parametrů a zhodnocení jejich vlivu na signál z reálného prostředí.
- *Analýza vlivu vybraných parametrizačních technik na robustnost rozpoznávače v reálném prostředí.*
  Navazujícím úkolem na předchozí bod je analýza vlivu modifikovaných parametrizačních technik a vybraných algoritmů pro potlačení šumu v signálu na robustnost rozpoznávače řeči. Jejich objektivní zhodnocení je provedeno na reálných nahrávkách řečového signálu. Modifikace jsou založeny na kombinaci standardních parametrizačních technik za účelem využití rozdílných přístupů ke zpracování signálu.
- *Analýza doplňkových algoritmů pro potlačování šumu v rámci předzpracování signálu z reálného prostředí.*
  Jako další cíl práce budou zkoumány možnosti rozšíření parametrizačních technik o metody pro potlačování šumu. Mezi zkoumané techniky patří algoritmus

rozšířeného spektrálního odečítání (ESS) a metoda vypouštění segmentů na bázi detekce řeči. Přínos jejich zahrnutí do procesu předzpracování řečového signálu bude analyzován především z pohledu eliminace chybovosti rozpoznávače v silně zašuměném prostředí automobilu a dále z pohledu omezení chybné klasifikace neřečových segmentů signálu.

- *Rozbor technik pro efektivní adaptaci modelů na šumové prostředí, návrh optimalizovaného procesu adaptace s využitím generických databází řeči.*
  Dalším cílem práce je rozbor základních parametrů techniky MLLR (dělení do regresních tříd, blokové vs adaptivní zpracování), analýza jejího přínosu v kombinaci s výše uvedenými schématy předzpracování signálu a následné možnosti využití metody především pro nahrávky pořízené v automobilu.
- *Zhodnocení možností modelování neřečových událostí s důrazem na neřečové události řečníka.*
  Vzhledem k četnému zastoupení neřečových událostí řečníka v promluvě, především ve spontánní řeči, ale také kvůli rozdílnému charakteru od šumu okolí, budou v této práci zanalyzovány dostupné databáze řeči s důrazem na výskyt těchto událostí. Tyto informace budou poté využity pro experimenty s modelováním neřečových událostí řečníka a analýze možností zvýšení efektivity trénování těchto modelů na dostupných databázích.

  Při rozpoznávání řeči na bázi statistického jazykového modelu vstupují do procesu rozpoznávání významné faktory – nastavení tohoto modelu a vlastnosti spojité promluvy. Vedle základních experimentů založených na rozpoznávání řeči s malým slovníkem bude proto zahrnuta také analýza přínosu modelování neřečových událostí v úloze rozpoznávání spojitých promluv.

  Dostupné řečové databáze čtených promluv obsahují značení neřečových událostí, které může být ovlivněno vysokou variabilitou subjektivního vnímání anotátorů i rozdílným nastavením podmínek nahrávání. Ukazuje se tak potřeba vzniku databáze spojitých promluv s vyváženým značením těchto událostí, která může být zdrojem nejen testovacích dat pro objektivní hodnocení metod pro potlačování vlivu neřečových událostí ale zároveň pro hodnocení úspěšnosti rozpoznávače spojitých promluv.

Kapitola 4

Nastavení experimentů

Experimenty provedené v této práci byly realizovány v několika konfiguracích, jejichž popis je uveden v první části následující kapitoly. Kapitola shrnuje také popis kritérií, kterými jsou popsány dosažené výsledky. Pro trénování akustických modelů jsou využívány rozsáhlé řečové databáze. Důraz při jejich tvorbě je kladen především na rovnoměrné zastoupení hlavních kategorií mluvčích, například podle pohlaví, věku či dialektu. V rámci přepisu obsahu promluvy již tyto databáze standardně obsahují také informaci o výskytu neřečových událostí (hluk pozadí, rušení od mluvčího). Taková databáze tak dává kompletní informaci pro tvorbu modelů řečových segmentů i pro testování systémů. Databáze použité v této práci přibližuje druhá část kapitoly. V rámci experimentů byly použity různé zdroje dat, především podle typu obsahu (čtené promluvy, spontánní řeč) a úrovně zašumění nahrávek (čisté nahrávky, obecné prostředí, automobil).

4.1 Použité nastavení rozpoznávače řeči

Použité řečové rozpoznávače byly realizovány pomocí sady nástrojů HTK Toolkit [123]. Výhodou použitého řešení je modulární struktura rozpoznávače, která umožňuje velmi flexibilně volit nastavení, nahrazení či vnoření jednotlivých bloků procesu rozpoznávání. To také umožnilo vznik dalších nástrojů, které doplňují systém HTK a pracují s odpovídajícím formátem dat.

Nevýhodou zvoleného obecného řešení může být především nižší míra optimalizace řešení na konkrétní úlohu, což se například projevuje delší dobou zpracování, především v komplexní úloze rozpoznávání řeči s velkým slovníkem.

V následujících kapitolách jsou popsány nejvýznamnější parametry rozpoznávacích systémů použitých v této práci. Detailní popis systému pro rozpoznávání řeči s malým slovníkem s využitím nástrojů HTK lze nalézt v [78], popisem vlastností rozpoznávače se statistickým modelem se zabývá [100].

4.1.1 Nástroje pro parametrizaci signálu

Ačkoliv sada nástrojů HTK obsahuje utilitu pro parametrizaci signálu HCopy, jedná se o nástroj, který poskytuje pouze základní volbu nastavení parametrizační techniky a algoritmů pro předzpracování signálu. V rámci práce byl proto použit parametrizační nástroj CtuCopy [15], který umožňuje detailnější možnosti nastavení parametrizačních technik a jejich doplnění o algoritmy potlačující šum. S jeho využitím byly pro reprezentaci řečového signálu generovány parametrizace blíže popsané v kapitole 6. Tyto parametrizace tvoří vektor 12 kepstrálních koeficientů a jeden koeficient logaritmu energie daného segmentu. Každý vektor je dále doplněn o dynamické a akcelerační koeficienty s konstantou M=2.

Pro srovnání dosažených výsledků rozpoznávání byl ve vybraných případech použit standard ETSI [22], který pro získání parametrizovaných dat využívá nástroj AdvFrontEnd.

4.1.2 Trénování akustických modelů

Akustické modely byly trénovány Baum-Welchovým algoritmem s využitím nástroje HERest. Níže uvedená tabulka shrnuje základní atributy akustických modelů. Tyto atributy se liší podle úlohy, pro kterou jsou modely použity.

Akustický model	levopravý HMM, tři emitující stavy, bez přeskoku stavů
Model krátké pauzy	jeden emitující stav svázaný s prostředním stavem modelu dlouhé pauzy (SIL)
Práh klestění (-t), inkrement a limit	250, 150, 1000

Tabulka 4.1: Základní popis akustických modelů

4.1.3 Rozpoznávací systémy

Pro zhodnocení vlivu navrhovaných postupů zpracování řečového signálu a rozpoznávání řeči byla realizována jak úloha rozpoznávání řeči s malým slovníkem, tak rozpoznávání spojitých promluv s velkým slovníkem. Jelikož se vlastnosti systémů pro realizaci těchto úloh liší, jsou v následující části popsána nastavení použitých rozpoznávačů.

Rozpoznávač na bázi gramatiky

Rozpoznávání řeči s využitím gramatiky je možné využít v případě nízkého počtu klasifikačních elementů (slov). V takovém případě lze považovat výskyt jednotlivých slov

v promluvě za stejně pravděpodobný. V této práci je pro posouzení vlivu navrhovaných postupů použit systém s gramatikou v podobě rozpoznávače sekvence číslovek. Tuto gramatiku představuje nekonečná smyčka obsahující jednotlivé číslovky oddělené modelem pauzy. Pro realizaci rozpoznávače Viterbiho algoritmem byl použit nástroj HVite s nastavením dle níže uvedených bodů v tab. 4.2.

Dekodér	HVite
Posílení vlivu vkládání slov (-p)	0,0
Posílení vlivu gramatiky (-s)	5,0
modelovaný element	monofón
počet modelů fonémů	43
počet směsí	32

Tabulka 4.2: Základní popis akustických modelů

Rozpoznávač se statistickým jazykovým modelem

Rozpoznávání spojitých promluv s velkým slovníkem již vyžaduje komplexnější přístup, např. z pohledu spojitosti řeči. V těchto úlohách jsou proto použity kontextově vázané fonémy – trifóny. Naopak vzhledem ke komplexnosti výpočetní úlohy bylo použito nižší množství směsí, než pro úlohu s gramatikou. Nastavení dekodéru shrnuje následující tabulka 4.3.

Dekodér	HDecode
modelovaný element	trifón
počet směsí	16
průměrný počet modelů po svázání	14232
prořezávání na úrovni modelů (-t)	200
prořezávání na úrovni slov (-v)	50
posílení vlivu gramatiky (-s)	10
posílení vlivu vkládání slov (-p)	-10

Tabulka 4.3: Nastavení rozpoznávače plynulých promluv

Statistický jazykový model byl použit ve dvou realizacích. Při adaptaci akustických modelů metodou MLLR (kap. 8) byl použit jazykový model obsahující pouze číslovky. V úloze rozpoznávání spontánních promluv s modelováním neřečových událostí (sekce 8.4) byl použit statistický model s velkým slovníkem.

Pro úlohu rozpoznávání číslovek s využitím statistického jazykového modelu byl použit jazykový model vytvořený na základě rozložení výslovnostních variant jednotlivých číslovek 0 – 9 v databázi SPEECON, viz tab. 4.4.

číslovka	výslovnostní varianta	počet výskytů
0	nula	1000
1	jedna	1000
	jeden	1
2	dva	509
	dvě	491
3	tři	1000
4	čtyři	310
	štyry	250
	štyři	220
	čtyry	210
5	pět	1000
6	šest	1000
7	sedm	410
	sedum	590
8	osm	380
	osum	620
9	devět	1000

Tabulka 4.4: Rozložení výslovnostních variant ve statistickém jazykovém modelu pro rozpoznávač číslovek

Pro účely rozpoznávání spojitých promluv byl vytvořen jazykový model [86] na bázi jazykového korpusu SYN2006PUB [126] ve variantách s různým počtem obsažených slov – 60k, 340k a různým stupněm složitosti – unigram, bigram, trigram.

Rozpoznávač pro databázi AURORA

V rámci této práce bylo ve většině experimentů využito výše uvedeného nastavení rozpoznávače. Výjimku tvoří experimenty s databází AURORA. Výše uvedené systémy jsou založeny na rozpoznávání subslovních akustických elementů, z nichž je možné modelovat libovolnou promluvu, naopak v úloze s databází AURORA je použito akustických modelů celých slov. Toto nastavení umožňuje trénovat modely na méně objemném řečovém materiálu, neboť není potřeba foneticky vyvážený materiál pro získání reprezentativního modelu každé hlásky. Trénování akustických modelů bylo provedeno nástrojem HERest se shodným nastavením, jako v případě rozpoznávače s gramatikou.

4.1.4 Míry pro hodnocení úspěšnosti rozpoznávače

Standardně používanou mírou pro hodnocení úspěšnosti rozpoznávače řeči je míra přesnosti rozpoznání jednotlivých slov (Accuracy)

$$ACC = \frac{N - I \ - D - S}{N} \cdot 100[\%] \tag{4.1}$$

kde N,I,D a S vyjadřují počet správně určených slov, počet chybně vložených, smazaných nebo zaměněných slov.

Doplněk k přesnosti pak vyjadřuje chybovost rozpoznávače na úrovni slov – Word Error Rate

$$WER = \frac{I \ + D + S}{N} \cdot 100[\%] \tag{4.2}$$

Vliv jednotlivých algoritmů na úspěšnost rozpoznávání je možné posuzovat porovnáním výše uvedených měr vůči základní (vztažné) hodnotě. Míra $ACCE$ (Accuracy Enhancement) určuje míru zlepšení přesnosti rozpoznávání oproti vztažné hodnotě $ACC_{baseline}$. Pro posouzení vlivu metody nezávisle na absolutních hodnotách lze použít i relativní míru $ACCE_r$.

$$ACCE = ACC_{new} - ACC_{baseline}[\%] \tag{4.3}$$

$$ACCE_r = \frac{ACC_{new} - ACC_{baseline}}{ACC_{baseline}} \cdot 100[\%] \tag{4.4}$$

Poměr

$$WERR = \frac{WER_{baseline} - \ WER_{new}}{WER_{baseline}} \cdot 100[\%] \tag{4.5}$$

kde $WER_{baseline}$ odpovídá chybovosti vztažného systému a WER_{new} značí chybovost nového systému, určuje míru zlepšení (snížení) chybovosti rozpoznávače.

4.2 Použité řečové databáze

Dostupnost databází pro účely trénování ASR modelů je významně ovlivněna mimo jiné i předpokládanými možnostmi jejich využití. Proto lze pro frekventovanější jazyky najít mnohem rozsáhlejší soubor řečových dat. Na druhou stranu jsou především v posledních letech zřejmé snahy rozšiřovat řečové databáze i pro jazyky méně zastoupené, např. češtinu. Zatímco první databáze řečových signálů obsahovaly čtené promluvy, s rozšiřujícím se využitím robustních rozpoznávačů spojitých promluv se také zvyšuje množství dostupných databází spontánních promluv, např. nahrávek televizních a rozhlasových zpráv, přednášek nebo rozhovorů. V současné době díky těmto snahám existuje mnoho zdrojů řečových a lingvistických dat, nejen pro účely

tvorby rozpoznávačů řeči, viz databáze poskytované skupinami ELRA [18] či LDC [56]).

Výhodou databází čtené řeči je možnost ovlivnit skladbu obsahu, což také umožní foneticky vyvážit obsah promluvy – to je důležité při tvorbě modelů jednotek řeči menších než slova (slabiky, fonémy, apod.). U čteného textu lze snáze přizpůsobit obsah promluv na specifické téma (např. číslovky, povely), čtená řeč bývá srozumitelnější a plynulejší a její nahrávání lze lépe koordinovat – pro mluvčího může být nahrávání při spontánní promluvě velmi stresující, což se odrazí i na výsledném projevu.

V rámci této práce byly použity dvě větší databáze čtených promluv nahrávané v různých prostředích, Czech Speecon Database [113] a CZKCC (databáze nahrávek v automobilu). Pro účely testování základních parametrů rozpoznávače číslovek pak byl soubor doplněn malou databází čtených číslovek (SNR set), doplněnou o aditivní složku hlukového pozadí nahrávaného v jedoucím automobilu. Pro srovnání s výsledky prezentovanými ve světě je pak využita databáze AURORA3, která obsahuje čtené číslovky pro několik evropských jazyků smíchané s různou úrovní šumu.

Databáze byly většinou rozděleny na několik částí s ohledem na použití (trénovací, testovací) a úroveň zašumění (čisté nahrávky, signály z konkrétního prostředí). Následující část popisuje jednotlivé databáze a použité dělení.

4.2.1 Databáze SPEECON

Databáze českých promluv SPEECON [113] obsahuje řečové nahrávky pořízené v různém prostředí – kancelář, automobil, uzavřené a otevřené veřejné prostory, domácí prostředí. Signál je snímán se vzorkovací frekvencí 16kHz, která je standardně využívána při realizaci rozpoznávačů řeči, s přesností 16 bitů na vzorek. Databáze obsahuje promluvy od 580 různých mluvčích dospělého věku, přepis obsahu jednotlivých nahrávek se zaznamenaným výskytem neřečových událostí (ruch okolí, neřečové události řečníka), dále informace o mluvčím a prostředí nahrávání.

Vzhledem k rozsahu databáze lze její obsah kategorizovat tak, aby co nejlépe popisoval informaci obsaženou v nahrávce a mohl tak být použit pro vyhodnocení výsledků rozpoznávání. Následující sekce popisuje použité dělení databáze na jednotlivé části, které mají společné vlastnosti.

Obsah promluv

Databáze obsahuje cca 300 promluv od každého mluvčího rozdělených do kategorií podle obsahu promluvy (jména, číslovky, úryvky z novin, apod.). Toto dělení bylo využito při výběru testovacích dat, např. izolované číslovky a sekvence číslovek pro případ rozpoznávače číslovek.

Nahrávací kanály

Promluvy byly nahrávány pomocí čtyř mikrofonů s různou snímací charakteristikou a různým umístěním od mluvčího. Signály z dvou mikrofonů umístěných v blízkosti mluvčího byly využívány v experimentech této práce.

V prvním případě se jedná o směrový mikrofon pro náhlavní soupravu, který snímá kvalitní řečový signál s minimálním zastoupením okolního rušení. Druhý mikrofon z hands-free sady zaznamenával signál s vyšší mírou rušení. Použití těchto kanálů simuluje podmínky použití standardního mikrofonu u PC – buď využití náhlavní sady nebo stolního mikrofonu. V následujícím textu budou tyto kanály značeny jako CS0 pro signál z náhlavní soupravy a CS1 pro signál z hands-free sady.

Čisté signály

Úvodní experimenty pro optimalizaci nastavení parametrizace signálu (sekce 6.4.1) byly provedeny na vybrané části databáze SPEECON, která obsahuje pouze položky s nízkým zastoupením rušení z okolí. Jedná se o promluvy z prostředí blízkého kancelářskému, kde je očekávána nízká úroveň šumu pozadí. Jednotlivé nahrávky byly navíc protříděny, aby výsledná sada dat neobsahovala položky nevhodné pro trénování rozpoznávače řeči, tedy např. přeřeknutí, hláskované položky, webové adresy, tedy položky s možným výskytem chyb v přepisu obsahu. Výsledný fragment, značený dále jako Office, obsahuje nahrávky od 190 mluvčích.

Doplněním fragmentu Office o data z prostředí domácích prostor (obývací pokoj) vznikl fragment Clean s 220 mluvčími. Z nich byly odstraněny položky s hudební reprodukcí v pozadí. Jedná se tak o část databáze, která obsahuje data se sníženou úrovní šumu.

Šumové pozadí

Signály z ostatních prostředí, které nebyly použity pro fragment Clean, tvoří množinu dat Noisy. I tato data byla protříděna, aby neobsahovala položky, které nejsou vhodné pro trénování rozpoznávače řeči. Zmíněné dva fragmenty tak tvoří množinu nahrávek All, která obsahuje všechny promluvy z databáze SPEECON vhodné pro použití v experimentech.

Databáze SPEECON obsahuje také informaci o úrovni okolního rušení v podobě odhadu SNR (Signal-to-Noise Ratio) pro jednotlivé nahrávky. Na základě této informace bylo provedeno dělení databáze na fragmenty HiSNR – signály s vysokým odstupem signálu a šumu, tedy nízkou úrovní zašumění a LoSNR – signály s nízkým odstupem šumu od signálu. Práh pro dělení do jednotlivých kategorií tvoří hodnota 20dB. Tyto fragmenty tak umožňují zkoumat vliv zašumění na výsledek rozpoznávání bez ohledu na typ rušení, pouze s přihlédnutím k míře zašumění.

Tabulka 4.5 popisuje dělení databáze SPEECON použité pro účely následujících experimentů a pro popis podmínek rozpoznávání. Obrázek A.1 v příloze pak ukazuje zastoupení různé úrovně zašumění nahrávek v jednotlivých fragmentech databáze.

Název	Popis prostředí
All	Všechna prostředí
Clean	Tiché prostředí
HiSNR	Prostředí s vyšší úrovní SNR

Název	Popis prostředí
Office	Prostředí kanceláře
Noisy	Hlučné prostředí
LoSNR	Prostředí s nižší úrovní SNR

Tabulka 4.5: Trénovací sady dat

Množiny dat pro trénování a testování

Výše uvedené dělení určuje fragmenty databáze, které mají podobné charakteristiky s ohledem na popis prostředí a vlastností zachyceného signálu. Pro testování rozpoznávače nezávislého na mluvčím pro dané podmínky bylo potřeba vyčlenit část dat, která bude použita pro testování kvality rozpoznávání. Mluvčí v testovací části se tak nesměl objevit v trénovacím fragmentu. Tabulka 4.6 zobrazuje objem dat, který byl pro jednotlivé fragmenty databáze SPEECON použit pro trénování a pro testování ASR systému.

Fragment	Trénovací část		Testovací část	
	počet mluvčích	délka [hod]	počet mluvčích	délka [hod]
All	531	141,7	59	0,63
Office	190	51,6	21	0,23
Clean	220	59,7	25	0,26
Noisy	273	71,7	30	0,32
HiSNR	495	103,3	55	0,59
LoSNR	417	36,3	46	0,50

Tabulka 4.6: Objem dat jednotlivých fragmentů databáze SPEECON

Tabulka 4.7 popisuje šumové podmínky v jednotlivých fragmentech databáze pomocí výčtu průměrné hodnoty odhadu SNR v daném fragmentu. Odhad SNR byl proveden pro všechny promluvy na základě nahrávek šumu z daného prostředí.

Kanál	SNR [dB]						
	All	Office	Car	Clean	Noisy	HiSNR	LoSNR
CS0	24,03	26,91	13,33	27,15	21,25	27,51	13,76
CS1	18,26	19,88	8,43	20,80	15,44	20,36	12,07

Tabulka 4.7: Průměr odhadnutého SNR ve fragmentech databáze SPEECON

4.2.2 Databáze CZKCC

Pro vybrané experimenty byla vedle databáze SPEECON použita také databáze řeči nahrávané v automobilu (CZKCC). Ta obsahuje české promluvy od 700 mluvčích zachycené dvěma mikrofony v automobilu za různých jízdních podmínek. Signály jsou v databázi uloženy se vzorkovací frekvencí 44,1 kHz, ale pro účely rozpoznávání byly podvzorkovány na kmitočet 16 kHz v souladu s parametry signálů v databázi SPEECON.

Obsah promluv

Databáze zahrnuje obdobné zastoupení obsahu promluv, jako databáze SPEECON. Tomu odpovídá i podobný průběh zpracování nahrávek - odstranění promluv s obsahem méně vhodným pro trénování rozpoznávače, apod.

Výrazným rozdílem je styl značení neřečových událostí, který je v porovnání s předchozí databází přesnější a klasifikuje tyto události do 10 základních skupin. Toto členění bylo využito pro přesnější popis neřečových událostí v kap. 9.

Nahrávací kanály

Při nahrávání promluv byly snímány vždy dva kanály s využitím tří různých mikrofonů s různou snímací charakteristikou. Pro účely této práce byly využity dva z těchto mikrofonů a vznikly tak dva fragmenty:

- *kanál Far-talk*
  Kanál snímaný vzdáleným AKG mikrofonem, který je vhodný pro použití právě v automobilu. Databáze obsahuje nahrávky od 700 mluvčích pořízené tímto mikrofonem.
- *kanál Close-talk*
  Kanál snímaný náhlavním mikrofonem Sennheiser, který je vhodný pro snímání kvalitního řečového záznamu. Databáze obsahuje 329 mluvčích zaznamenaných tímto mikrofonem[1].

Šumové pozadí

Každý mluvčí byl v průběhu nahrávání zaznamenán za tří různých jízdních podmínek. Nahrávky jsou proto v dalším textu klasifikovány dle úrovně zašumění signálu do tří skupin: Stojící vůz s vypnutým motorem (dále značeno jako OFF), stojící vůz s běžícím motorem (ON) a jedoucí vůz (DRV).

Tyto podmínky jsou v rámci popisu jízdních podmínek každé nahrávky přesněji popsány pomocí kritérií jako třída automobilu, typ silnice apod. To umožnilo blíže zkoumat například vliv jednotlivých podmínek na zastoupení neřečových událostí v promluvě, viz kap. 9.

[1] Zbývající část databáze byla nahrávána jiným typem vzdáleného mikrofonu (Peiker)

Množiny dat pro trénování, adaptaci a testování

Podobně jako v případě databáze SPEECON byly oba fragmenty rozděleny na více částí podle využití v procesu tvorby rozpoznávače. Jednotlivé množiny jsou disjunktní s ohledem na výskyt jednotlivých mluvčích a dělení také respektuje rovnoměrné zastoupení dalších kategorií, jako např. třída vozu, dialekt mluvčích apod. Výsledkem jsou pro každý kanál 3 množiny nahrávek:

- *train* – fragment pro přetrénování modelů (500/242 mluvčích v Far-talk resp. Close-talk kanálu)
- *adapt* – fragment pro adaptaci modelů (100/42 mluvčích v Far-talk resp. Close-talk kanálu)
- *test* – fragment pro testování systému (100/45 mluvčích v Far-talk resp. Close-talk kanálu)

Tabulka 4.8 ukazuje objem dat v jednotlivých fragmentech databáze rozdělený s ohledem na jízdní podmínky.

	Objem řečových dat [hod]		
	trénovací	adaptační	testovací
OFF	41,44	7,79	8,83
ON	32,6	6,05	7,37
DRV	148,55	29,5	30,99

Tabulka 4.8: Objem dat v hodinách v jednotlivých fragmentech databáze CZKCC

4.2.3 SNR-set

Pro rozšíření testovací databáze v případě rozpoznávače sekvence číslovek byla v práci využita i databáze promluv sesbíraných ve stojícím automobilu za podmínek blízkých podmínkám pro kanál CS0 databáze SPEECON. Tato sada obsahuje 121 promluv od 21 mluvčích a je použita především spolu se samostatnými nahrávkami šumu automobilu. Sada nahrávek obsahuje řečový materiál s celkovou délkou 1,15 hod.

Čisté signály byly smíchány s náhodně vybranými nahrávkami automobilového šumu snímaného za různých jízdních podmínek s cílem vytvořit sadu dat s požadovaným SNR, určeným na základě vztahu pro globální SNR

$$SNR = 10 \log_{10} \frac{P_s}{P_n} \; [dB] \tag{4.6}$$

Výsledná databáze je použita pro základní analýzu algoritmů pro potlačení aditivního rušení v řečovém signálu při rozpoznávání řeči.

4.2.4 AURORA 3

Výsledky získané na rozsáhlých řečových databázích sice dávají informaci o schopnosti rozpoznávače správně klasifikovat předloženou promluvu, ale i přes zevrubný popis podmínek databáze není možné jednoznačně porovnat na základě získaných výsledků dva rozpoznávače, pokud jsou testovány na různých databázích.

Z toho důvodu byla pro srovnání s jinými experimenty použita obecně využívaná databáze promluv AURORA3 [74]. Ta obsahuje nahrávky sekvence číslovek pro mluvčí v různých šumových podmínkách a s různými typy mikrofonů. V rámci databáze jsou pak data rozdělena do skupin podle míry odlišnosti trénovací a testovací části databáze.

Databáze je vytvořena pro čtyři jazyky - němčina, dánština, španělština a finština. Z těchto jazyků byly pro experimenty v této práci použity poslední tři jmenované jazyky. Pro každý jazyk je pak vytvořeno rozdělení na tři části:

- WM – well matched - nízká odlišnost nahrávacích systémů, tiché prostředí,
- MM – medium mismatch - středně ztížené podmínky rozpoznávání,
- HM – high mismatch - vysoká odlišnost prostředí v trénovací a testovací množině, hlučné prostředí.

Podrobnější popis je uveden v [74].

4.2.5 Databáze spontánních promluv

V předchozích sekcích byly popsány databáze, které z větší části obsahují čtené promluvy. Běžná řeč se ovšem v mnoha ohledech liší od čtené řeči a požadavek na spontánnost přináší problémy, které nemusely být v úloze rozpoznávání s malým slovníkem řešeny [109]. V první řadě se jedná o obsah řeči. Čtené databáze obsahují řeč foneticky vyváženou, s nízkým výskytem neplynulostí – předčasně ukončená věta, opakování, váhání, apod. Navíc je čtená promluva většinou monotónní, bez větších proměn zabarvení hlasu, a proto je i její variabilita nižší.

Naopak běžná řeč obsahuje mnoho rušivých prvků, s nimiž musí být při rozpoznávání počítáno, aby nezpůsobily chyby. Proto je pro tvorbu rozpoznávače plynulých promluv potřeba využít i dostupných databází spontánních promluv a tvorba takových databází je důležitou součástí mnoha projektů zaměřených na rozpoznávání obecných promluv ([118], [91]). Sběr spontánních dat je popsán v následující kapitole.

Kapitola 5

Databáze spontánních promluv

S rostoucím uplatněním rozpoznávačů řeči v praxi a poptávkou po systémech použitelných v obecném provozu je stále více zapotřebí využití zdrojů dat, které mnohem lépe vystihují styl běžné spojité řeči. V rámci této práce byla vytvořena databáze spojitých promluv, která doplňuje použitý řečový materiál. Tato kapitola popisuje aspekty sběru dat, proces kompletace databáze a základní analýzu obsahu promluv.

5.1 Nahrávky přednášek

V rámci pravidelných prezentací výsledků doktorského výzkumu na našem pracovišti byly nahrávány příspěvky prezentující různá témata v oblasti zpracování signálů (DSP - Digital Signal Processing). Ty byly doplněny o vybrané přednášky kurzů specializujících se na tématiku zpracování signálů.

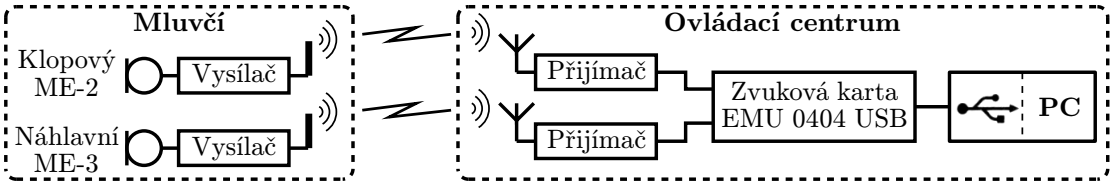
Každá přednáška obsahuje cca 20 – 30 minut promluv, které byly předem připraveny. Tím bylo zajištěno, že ačkoliv se jedná o spontánní přednes, promluvy byly plynulé a nahrávky tak obsahují poměrně kvalitní projev, který bylo možné snadno zpracovávat. Přednášky byly prezentovány jedním mluvčím, zřídka přerušeny dotazem z publika případně odpovědí na položený dotaz.

Rozpoznávače obecných promluv jsou často alespoň částečně přizpůsobovány tématu, neboť jazykový model pro obecnou češtinu nemůže dostatečně kvalitně modelovat rozmanitost jazyka. Z tohoto pohledu je pro testovací účely jednotné téma digitálního zpracování signálu nahrávaných přednášek výhodnou vlastností sbírané databáze.

5.2 Nahrávací zařízení

Pro nahrávání byl použit komerčně dostupný bezdrátový systém, který dovolil řečníkovi volný pohyb bez omezení, a to i pro případ vícekanálového snímání signálu. Systém byl navržen tak, aby poskytoval vysokou kvalitu signálu s možností přenositelnosti a připojení k libovolnému standardnímu PC pro ukládání snímaného záznamu.

Použití PC umožnilo průběžné sledování záznamu pro detekci případných poruch a snadné přizpůsobení základních nastavení systému pro získání potřebných vlastností signálu, např. dostatečné intenzity v případě slabého řečového signálu nebo naopak snížení intenzity při výskytu přebuzení.



Obrázek 5.1: Nahrávací schéma s bezdrátovou přenosovou sadou

Dvoukanálový přenos byl realizován pomocí dvou bezdrátových přenosových sad Sennheiser. Každá sada obsahovala vysílač, který byl připevněn např. na opasku, přijímač s LCD displejem pro sledování aktivity a mikrofonu. V prvním případě se jednalo o všesměrový klopový mikrofon, v druhém případě byl mikrofon se směrovou charakteristikou připojen na náhlavní sadě. Klopový mikrofon přichycený mluvčímu cca 10 cm pod bradou zachycoval vzhledem k jeho všesměrovosti vyšší úroveň rušení. Vzdálenost úst mluvčího od mikrofonu byla také proměnná (s pohybem hlavy mluvčího), proto byl signál velmi variabilní co do intenzity řečového signálu i šumu. Naopak signál z náhlavní sady poskytoval velmi kvalitní signál s velkou stabilitou intenzity a nízkou mírou zašumění.

Intenzitu přenášeného signálu bylo možné sledovat a pomocí ovládacích prvků zvukové karty nastavovat jeho úroveň. Intenzita signálu byla nastavena na ovládacích prvcích zvukové karty během prvních nahrávek tak, aby bylo dosaženo vyvážených vlastností signálu – minimalizace přebuzení a na druhou stranu dostatečně silný signál. Blokové schéma bezdrátového nahrávacího systému znázorňuje obr. 5.1.

Signál z nahrávacího systému byl ukládán v nejvyšší možné kvalitě – 48kHz, 16bitů na vzorek, aby byla dostupná vysoká kvalita signálu pro další využití. Tato frekvence je vhodná pro následné podvzorkování na standardně užívanou frekvenci 16kHz. Teprve pro účely použití v úloze rozpoznávání řeči byl signál následně podvzorkován pro kompatibilitu s ostatními použitými databázemi.

Nahrávání na PC bylo ovládáno volně dostupným nahrávacím programem *WaveLab Lite*. Ostatní kroky přípravy databáze, např. organizace databáze, segmentace dlouhých nahrávek či přepis obsahu nahrávek, byly provedeny později, nezávisle na nahrávání signálu.

V pozdější fázi projektu bylo pro nahrávání řeči navíc použito nahrávací zařízení (diktafon zn. Edirol) pro zachycení řečového signálu z větší vzdálenosti. Zařízení umožnilo dvoukanálové nahrávání z mikrofonů umístěných cca 6 cm od sebe a uložení signálů na paměťovou kartu ve WAV formátu ve kvalitě odpovídající výše popsanému nastavení (48kHz vzorkování, 16 bitů na vzorek). Takto zachycený signál obsahuje výrazně vyšší zastoupení okolního rušení a je vhodný pro simulaci náročných rozpoznávacích podmínek.

5.3 Segmentace a anotace signálu

Prezentace byly nahrávány v celku, bez předchozí segmentace. Tak vznikly signály o délce několika desítek minut, které bylo pro další použití potřeba rozdělit na kratší celky a anotovat. Tyto signály byly segmentovány a anotovány ručně, pro další případné rozšíření databáze jsou ale připraveny segmentační nástroje.

Pro segmentaci a ruční anotaci signálů byl použit volně dostupný nástroj pro segmentaci, popis a transkripci *Transcriber* (dodávaný LDC [7]). Ten poskytuje plnou podporu při tvorbě větších řečových databází. Další postup zpracování databáze je popsán v následujícím textu.

5.3.1 Segmentace dlouhých nahrávek

Každá promluva byla rozdělena na věty, jakožto vhodnou jednotku pro další zpracování a trénování ASR systémů. Ačkoliv byly přednášky předem připravené, nejednalo se o čtený text a obsah promluvy tak nebyl předem znám. Proto byly začátky a konce vět nalezeny manuálně na základě poslechu. Vzhledem k vlastnostem spontánních promluv [109], kterými jsou především výskyt opakování slov, nedokončení vět nebo naopak započetí věty několika způsoby, bylo určení hranice mezi větami často obtížné a vzniklé segmenty mohou být i relativně dlouhé, v řádu desítek vteřin.

Navíc docházelo na konci vět ke snížení hlasitosti projevu. V případech, kde to bylo možné, bylo proto potřeba zanechat před větou a za větou krátký úsek pauzy, aby byly obsaženy i tyto tiché úseky. Naproti tomu je v plynulé řeči častým jevem velmi rychlé navázání nové věty, kdy není mezi slovy obou vět žádná pauza. Správnému nalezení hranic mezi větami byla v databázi proto věnována vysoká pozornost.

Delší úseky pauzy bez řečové aktivity byly vyjmuty jako samostatné segmenty. Původní nahrávky v celku byly také ponechány v databázi pro další využití.

5.3.2 Ortografická transkripce

Obsah promluv byl přepsán v podobě ortografické anotace. Stejně jako v jiných databázových projektech ([91], [84]) byla i v této databázi použita standardizovaná pravidla. Celý proces transkripce byl rozdělen do několika kroků.

V prvním kroku byla zaznamenána pouze řečová informace a všechny ostatní jevy byly vynechány. Během přepisu se nezaznamenávala interpunkce, neboť tato informace není pro účely trénování AR systému zapotřebí. Doplnění interpunkčních znamének je další samostatnou a ne zcela jednoduchou úlohou, která v této fázi nebyla řešena. Pro přepis byla použita pouze malá písmena abecedy, což umožnilo zamezit případným nežádoucím rozdílům v situacích, kde toto rozlišení nebylo potřebné či jednoznačné.

Obsah promluv byl přepsán v přesné podobě, tedy s použitím např. hovorových variant slov či matematických výrazů, což jsou jevy, které technické promluvy často doprovází. Pro ostatní speciální jevy, které se v řeči vyskytují, jako např. hláskování,

přeřeknutí nebo cizí slova, byla použita konvence pro značení speciálních událostí pomocí XML tagu "Event" s atributy dle tab. 5.1. Tabulka zároveň ve sloupci 'Značka' uvádí značení, dle kterých je v tomto textu na událost odkazováno. Toto značení vychází z konvence použité v databázi SPEECON.

Jev	Značka	Přípustné hodnoty atributů		
		desc	type	extent
nesrozumitelné slovo	'**'	"xxx"	"pronounce"	"instantaneous"
cizí slovo	'∼'	"xxx"	"language"	"instantaneous"
přeřeknutí, část slova	'*'	značka '()' za slovem		
hláskované položky	'$'	značka '$' následovaná přesnou výslovností dané hlásky		

Tabulka 5.1: Konvence pro anotaci typických jevů ve spontánní promluvě

Malé výchylky od standardní výslovnosti nebyly zaznamenány, neboť mohou být snadno vystiženy variabilitou jednotlivých fonetických elementů. Vzhledem k tomu, že nebyl předem znám obsah promluv a v některých případech mohlo nesprávným porozuměním dojít k chybnému přepisu, byly promluvy kontrolovány více anotátory.

5.3.3 Fonetická transkripce

Fonetická transkripce není součástí databáze, ale pro další použití nahrávek byla automaticky vytvářena nástrojem *transc* [85] v další fázi anotací. Zmíněný nástroj generoval fonetický přepis na základě ortografické transkripce s ohledem na pravidla české výslovnosti. Nepravidelnosti ve výslovnosti byly z větší části pokryty slovníkem výjimek nebo přímou definicí výslovnosti v rámci ortografické transkripce.

5.3.4 Anotace neřečových událostí

Jak bylo naznačeno výše, spontánní promluvy se v mnoha ohledech liší od čtené řeči. Je to způsobeno tím, že mluvčí musí během řeči přemýšlet o následujících slovech. To má za následek snížení plynulosti řeči, projevující se především nárůstem výskytu pauz v řeči, prodlužováním hlásek a přítomností tzv. vyplněných pauz.

Tyto jevy byly v dalším anotačním kroku zaznamenávány spolu s ostatními neřečovými událostmi. Vzhledem k rozdílným vlastnostem přijímacích kanálů se výskyt těchto jevů může pro jednotlivé kanály mírně lišit.

Neřečové události, zaznamenané v přepisu podle tab. 5.2, mohou být rozděleny do několika základních kategorií:

- *Neřečové události řečníka* – Jelikož jsou tyto události generovány podobně jako řeč, většinou se objevují mezi dvěma slovy. Proto byly v anotaci zaznamenány v podobě klíčového slova vloženého do promluvy jako standardní slovo, např. '*slovo1* [*fil*] *slovo2*'. Typické neřečové události řečníka shrnuje tab. 5.2.

- *Události z šumového pozadí* – Ačkoliv byly nahrávky pořízeny v relativně tichém prostředí, mohou obsahovat rušení z okolí mluvčího, které bylo také nutné zaznamenat. Na rozdíl od neřečových událostí řečníka mohou tyto jevy ovlivňovat vlastní řeč. Proto byla pro jejich popis zavedena komplexnější metodika.

  Události vyskytující v pauzách mezi promluvami byly značeny podobně, jako neřečové události řečníka. Pokud byla ovlivněna promluva, bylo klíčovým slovem označeno počáteční a koncové slovo ovlivněného úseku promluvy, např. "*slovo1* [*int*−] *slovo2 slovo3* [−*int*] *slovo4*". Tento způsob zápisu byl použit i v obecném nastavení transkripčního programu *Transcriber*.

- Speciální případ představuje značka "[*sta*]" na začátku promluvy, která značí přítomnost stacionárního rušení v celém segmentu signálu.

- *Ostatní mluvčí* – Promluvy byly nahrávány v rámci prezentací, při nichž byla umožněna interakce s posluchači. Ačkoliv se jednalo o méně častý jev, mohl se vyskytnout případ, kdy byla zaznamenána promluva jiného mluvčího. V takovém případě je rozlišeno, zda se jedná o promluvu jiné osoby během pauzy hlavního mluvčího ([*other*]) nebo současnou promluvu více mluvčích, tzv. cocktail-party efekt ([*cockt*]).

Jev	Značka	Přípustné hodnoty atributů		
		desc	type	extent
cocktail party efekt	[*cockt*]	"cockt"	"noise"	"begin", "end"
nádech/výdech	[*dech*]	"r"	"noise"	"instantaneous"
vyplněná pauza	[*fil*]	"h"	"pronounce"	"instantaneous"
kašel/odkašlání	[*kasel*] / [*ehm*]	"k"	"noise"	"instantaneous"
mlasknutí	[*mlask*]	"m"	"noise"	"instantaneous"
smích	[*smich*]	"rire"	"noise"	"instantaneous"
stacionární šum	[*sta*]	"b"	"noise"	"begin", "end", "instantaneous"
nestacionární šum	[*int*]	"conv"	"noise"	"begin", "end", "instantaneous"

Tabulka 5.2: Popis anotovaných neřečových událostí

5.4 Výsledný obsah databáze

Objem řečových nahrávek tvoří 62 promluv o celkové délce necelých 22 hodin od 24 různých mluvčí – 23 mužů a jedné ženy. Nejdelší nahrávka obsahuje promluvu o délce necelých 40 minut. Tabulka C.1 v příloze ukazuje míru zastoupení jednotlivých mluvčích v databázi.

Celkový počet slov	63000
Počet rozdílných slov	7800
Počet hovorových/slangových výrazů	4,6 %
Oborově specializovaná slova	21 %

Tabulka 5.3: Souhrn obsahu databáze z pohledu zastoupených slov

5.5 Srovnání kvality nahrávek

Konečná podoba databáze, stanovená v průběhu vytváření prvních nahrávek, obsahuje nahrávky v celku, segmentované nahrávky, ortografickou transkripci obsahu promluv s vyznačením nepravidelností ve výslovnosti a označené neřečové události. V následujícím textu je databáze porovnána s dostupnými řečovými databázemi čtených promluv.

Databáze obsahuje 63000 slov, z toho 7800 různých slov. Spontánní charakter promluv je zřejmý ze zvýšeného výskytu hovorových a slangových výrazů v porovnání se čtenými databázemi – 4,6 % hovorových výrazů v DSP-setu oproti 0,08 % těchto výrazů v databázi Czech LC-Star2 [55]. Vzhledem k technickému zaměření přednášek obsahuje databáze cca 21 % specializovaných slov (méně používaní slova - porovnáváno k slovníku LC-Star2).

5.5.1 Srozumitelnost řeči

Jak již bylo zmíněno, spontánní charakter řeči má vliv na plynulost projevu a výskyt chyb různého charakteru, např. nesprávná výslovnost, nedokončené věty, zakoktání. Tabulka 5.4 ukazuje srovnání množství správných, špatně vyslovených a nesrozumitelných slov v prezentované databázi (CzLecDSP) oproti databázím čtené řeči (SPEECON, CZKCC).

Zatímco výskyt malých chyb ve výslovnosti je srovnatelný pro oba typy databází, výskyt slov s vyšší mírou nesrozumitelnosti je již pro spontánní promluvy vyšší. To je způsobeno především opakováním a opravami během promluvy, kdy mluvčí uprostřed slova přeruší větu, např. aby začal větu s malou úpravou od začátku. Na druhou stranu je navzdory těmto jevům výskyt nesrozumitelných slov v získaných signálech stále relativně malý a v tomto ohledu je databáze vhodným materiálem pro použití v oblasti trénování a testování rozpoznávačů řeči.

5.5.2 Výskyt neřečových událostí

Přítomnost neřečových událostí v trénovací databázi je významná pro tvorbu robustních ASR systémů. V následující části je porovnána databáze spontánních promluv s databázemi čtených promluv. Jak ukazuje např. [25], oba typy promluv mají specificky zastoupeny jednotlivé druhy neřečových událostí.

databáze	počet slov	malé chyby výslovnosti	nesrozumitelná/ nekompletní slova
SPEECON	561 716	1157 (0,21 %)	1768 (0,31 %)
CZKCC	1 067 412	1689 (0,16 %)	1902 (0,18 %)
CzLecDSP	63 000	85 (0,14 %)	445 (0,71 %)

Tabulka 5.4: Rozdělení slov v jednotlivých databázích

Malá část databáze SPEECON obsahuje nahrávky spontánních promluv. Tyto fragmenty byly proto pro účely srovnání odděleny.

Tabulka 5.5 ukazuje množství neřečových událostí označených v přepisu promluv jednotlivých databází (procenta jsou vztažena k počtu slov v daném fragmentu). Vzhledem ke kategorizaci neřečových událostí v databázi SPEECON pouze do dvou skupin – FIL (vyplněná pauza), SPK (ostatní neřeč. události řečníka), probíhá porovnání událostí právě na úrovni těchto dvou skupin, ačkoliv je u ostatních databází k dispozici přesnější značení.

Z porovnání výskytu vyplněných pauz v tabulce 5.5 je zřejmý vyšší výskyt tohoto jevu v promluvách spontánního charakteru, což souvisí s výše zmiňovaným jevem, kdy je promluva tvořena v průběhu řeči. Na druhou stranu je spontánní řeč více plynulá, s čímž souvisí i výskyt neřečových událostí typu mlasknutí, hlasitý nádech či odkašlání. Začátky vět při čtení jsou totiž často následovány právě přípravou na následující větu v podobě nadechnutí předcházeného otevřením úst, které může být doprovázeno zmíněným mlasknutím. To dokládá i tabulka 5.5 srovnávající ve sloupci "anotované značky" výskyt těchto neřečových událostí v jednotlivých fragmentech databáze.

Tabulka ukazuje také významný rozdíl ve výskytu neřečových událostí pro obě databáze čtené řeči. Vedle rozdílů v přístupu k vlastní transkripci těchto jevů totiž jejich výskyt ovlivňuje také kvalita mikrofonů, jejich pozice a nahrávací podmínky. Kvalitou anotace neřečových událostí se proto zabývá i část kapitoly 9.

databáze	počet slov	vyplněné pauzy	ostatní události
SPEECON čtený	146537	344 (0,23 %)	33125 (18,25 %)
SPEECON spont.	34954	1512 (4,33 %)	
CZKCC	244044	153 (0,06 %)	15728 (6,44 %)
CzLecDSP	54314	1449 (2,67 %)	203 (0,37 %)

Tabulka 5.5: Výskyt vyplněných pauz v použitých databázích

Kapitola 6

Příznaky pro rozpoznávání řeči

Standardní parametrizační techniky MFCC a PLP reprezentují řečový signál v kepstrální oblasti a jsou často používané pro rozpoznávání řeči. Optimalizací nastavení jednotlivých parametrizačních algoritmů pro získání nejvýhodnějších vlastností pro různé podmínky se zabývá mnoho prací, které ale také ukazují rozdílné vlastnosti těchto dvou metod v různých podmínkách. Nastavení těchto parametrů a případná modifikace parametrizace pro využití těchto vlastností v robustních systémech při šumových konkrétních podmínkách jsou důležitým krokem při tvorbě rozpoznávače řeči v reálném prostředí ([34], [13]).

Spolu s aplikací algoritmů pro potlačování šumu a zvýrazňování řeči je nalezení vhodné robustní parametrizace tématem následující kapitoly.

6.1 Standardní parametrizace

Parametrizační metody MFCC a PLP využívají principů tvorby hlasu řečovým traktem a vnímání řeči lidským uchem. Těchto znalostí využívá výpočetní postupu naznačený v blokovém schématu na obr. 6.1. Ten lze pro obě metody rozdělit do tří základních částí:

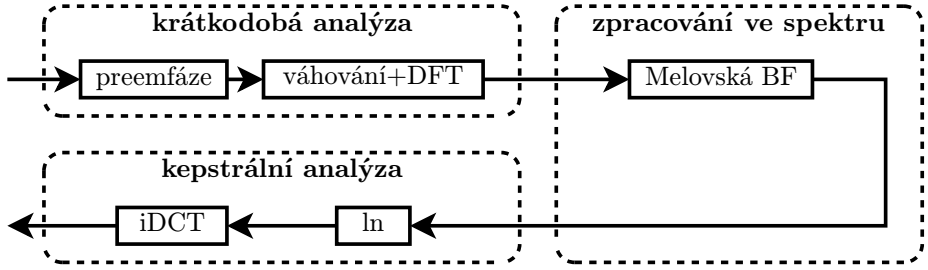
- krátkodobá Fourierova transformace signálu,
- nelineární zkreslení frekvenční osy,
- kepstrální analýza.

V rámci těchto částí jsou realizovány algoritmy, které upravují signál do podoby vhodné pro úlohu rozpoznávání řeči. Následující text přibližuje vlastnosti těchto procesů tak, aby mohly být následně využity pro hledání možných modifikací, které tyto vlastnosti využívají.

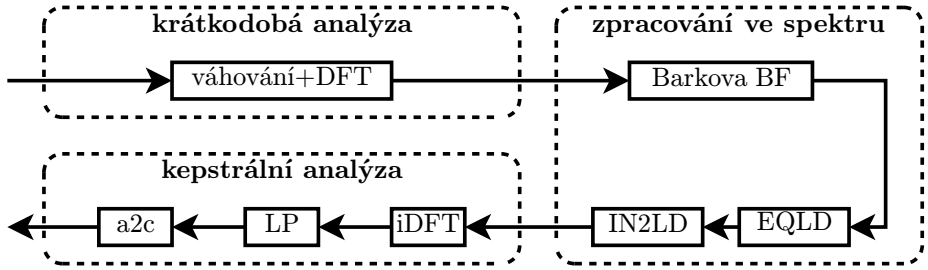
Krátkodobá Fourierova transformace signálu

Krátkodobá Fourierova transformace signálu poskytuje krátkodobý odhad spektrální podoby signálu. Použití krátkodobé Fourierovy transformace je založeno na před-

MFCC:



PLP:



Obrázek 6.1: Blokové schéma výpočtu standardních parametrizací MFCC a PLP

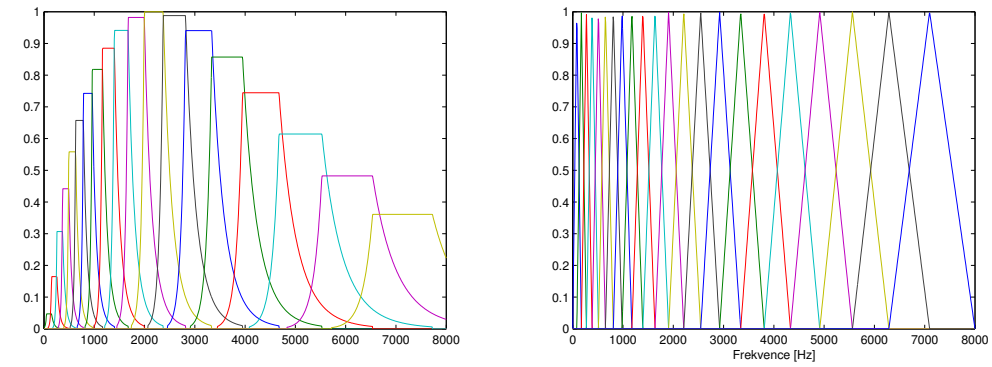
pokladu, že je řečový signál v analyzovaném úseku stacionární. Pro reálný signál není tento předpoklad splněn úplně a volbou délky zpracovávaného segmentu lze významně ovlivnit vlastnosti systému a jeho citlivost na charakter zpracovávaného signálu. V sekci 6.4.1 je proto analyzováno vhodné nastavení segmentace řečového signálu pro následné zpracování parametrizačními technikami.

Výpočetní postup je pro obě základní parametrizační techniky shodný a využívá nejčastěji algoritmů rychlé Fourierovy transformace (FFT - Fast Fourier Transform).

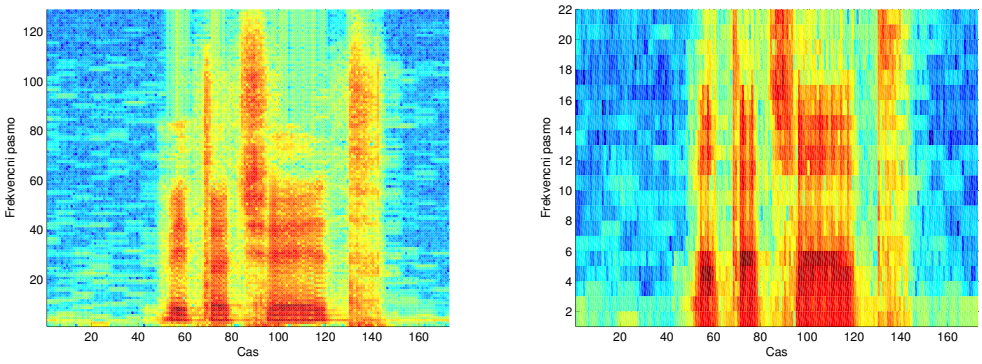
Nelineární zkreslení frekvenční osy

Je známo, že vnímání tónů lidským uchem není lineárně závislé na frekvenci poslouchaného tónu. S měnící se frekvencí se mění vnímání zvuku, mimo jiné například subjektivní výška tónu nebo schopnost rozlišovat blízké tóny. Tuto nelinearitu je možné zahrnout do procesu výpočtu parametrů signálu rozdělením signálu do jednotlivých pásem, která odpovídají tzv. kritickým pásmům. Ta jsou definována na základě experimentálních měření. V dalším zpracování signálu se poté pracuje pouze s energií v jednotlivých pásmech. Vzhledem k rozdílné šířce jednotlivých pásem je často tato energie normována k šířce pásma, aby nedocházelo k zvýšení vlivu širších pásem na vyšších frekvencích oproti pásmům užším.

Nelineární filtrace je v rámci standardních metod realizována pomocí tzv. Melovské (MFCC) resp. Barkovy (PLP) banky filtrů. Tyto banky se liší zejména ve tvaru použitých filtrů a v šířce jednotlivých pásem především na vyšších frekvencích (viz obr. 6.2). Výsledek zkreslení spektra ukazuje spektrogram na obr. 6.3.



Obrázek 6.2: Banky filtrů pro analýzu řeči - Barkova BF (vlevo), Melovská BF (vpravo)



Obrázek 6.3: Spektrogram řečového signálu před nelineární filtrací (vlevo) a po filtraci Melovskou BF s 22 pásmy (vpravo)

Vliv tvaru filtru na robustnost metody byl zkoumán např. v [90], který porovnává několik základních tvarů a mezi nimi také trojúhelníkový nebo tvar používaný v PLP zpracování. V práci [58] navíc analyzují autoři příspěvek filtru k robustnosti metody, pokud je filtr navržen s ohledem na charakteristiku vnímání zvuku lidským uchem. Výsledky těchto prací ale ukazují, že samotný tvar filtru nemá pro robustnost metody s ohledem na šumové prostředí velmi významný přínos.

Při výpočtu PLP parametrů je ve spektrální oblasti aplikován filtr pro snížení dynamiky řečového signálu a narovnání jeho typického průběhu – klesající výkon s rostoucí frekvencí a nelineární závislost vnímání hlasitosti. Filtr je založen na znalostech vnímání zvuku lidským uchem. Při výpočtu je použit tzv. equal-loudness filtr (EQLD), který koriguje výkonové charakteristiky křivkami stejné hlasitosti. Následně je aplikován intenzity-to-loudness filtr (IN2LD), který realizuje vztah mezi intenzitou zvuku a hlasitostí podle vztahu

$$\hat{P}(\Omega) = P(\Omega)^{0.3}. \tag{6.1}$$

Výsledným potlačením dynamiky je možné dosáhnout efektivnějšího popisu řečového signálu s omezeným počtem koeficientů.

Metoda MFCC používá pro úpravu dynamiky frekvenční reprezentace signálu tzv. pre-emfázovou filtraci. Tento filtr je aplikován v časové oblasti, většinou před krátkodobou transformací a je realizován zpravidla jednoduchou horní propustí 1. řádu.

Kepstrální analýza

Pro převod do kepstrální oblasti využívají zmíněné parametrizace rozdílný výpočetní postup, založený na základním vztahu pro výpočet výkonového kepstra signálu $x[n]$

$$c[n] = \mathcal{Z}^{-1}\{\ln|\mathcal{Z}\{x[n]\}|^2\}. \tag{6.2}$$

Melovské kepstrální koeficienty jsou určeny přímým výpočtem přes zpětnou Fourierovou transformaci logaritmu amplitudového spektra. Vzhledem k reálnému a symetrickému výkonovému spektru se pro výpočet používá jednodušší diskrétní kosinové transformace (DCT)

$$c[n] = \sum_{k=1}^{N} \ln E[k] \cos\left(\frac{\pi n}{N}(k - \frac{1}{2})\right), \quad \text{pro } n = 0, 1, ...M, \tag{6.3}$$

kde $E[k]$ je energie signálu v k-tém pásmu melovské banky filtrů, N je počet pásem banky filtrů a M je počet kepstrálních koeficientů, standardně volený na hodnotu 13.

Pro výpočet PLP kepstrálních koeficientů je použito rekurzivního výpočtu kepstra z autoregresních koeficientů. Pro tento výpočet je použit předem definovaný počet autoregresních koeficientů získaných Levinson-Durbinovým algoritmem na bázi LPC analýzy

$$c[n] = -a_n - \frac{1}{n}\sum_{k=1}^{n-1}(n-k)a_k c[n-k] \tag{6.4}$$

V případě čistého řečového signálu poskytují kepstrální koeficienty velmi kvalitní a výhodné vlastnosti pro účely rozpoznávání řeči. S rostoucím zašuměním signálu se ale kvalita generovaných příznaků snižuje. Vlivem rozdílných postupů výpočtu je ale také vliv šumu na jednotlivé metody různý [82]. Pro techniky založené na LPC analýze signálu, kde se pro odhad autokorelační funkce používá zašuměný signál, bývá proto vliv šumu výraznější. Naopak při použití těchto příznaků pro popis čistého signálu bývá dosahováno lepších výsledků.

Analýza vlastností standardních technik s ohledem na zvýšení robustnosti a možnost jejich doplnění o algoritmy pro potlačení vlivu rušení jsou tématem následujících sekcí. V oblasti porovnání metod MFCC a PLP byly na úrovni hledání optimálního nastavení parametrů výpočtu provedeny rozsáhlé studie, viz např. analýza optimálního počtu kepstrálních koeficientů a velikost banky filtrů v [89] a [87]. K základnímu

nastavení metod PLP a MFCC vzniklo mnoho variací, které mají pomoci popsat řečový signál robustněji a umožnit tak jejich využití i v podmínkách, kde standardní algoritmus poskytuje méně výhodné řešení. Tyto modifikace spočívají různě komplexních úpravách, od optimalizace tvaru a počtu pásem banky filtrů, viz např. [46], [58], až po parametrizaci na bázi RASTA-PLP [24] pro odstranění pomalu se měnících změn v signálu. Tato práce se proto zabývá dalšími možnostmi zvýšení robustnosti předzpracování signálu především na úrovni modifikace základních bloků výpočtu.

6.2 Modifikované metody

Podobnost výpočetních postupů standardních metod PLP a MFCC na jedné straně a rozdílné výsledky, které dosahují ASR systémy založené na těchto metodách na druhé straně jsou hlavními důvody, proč i přes mnoho publikovaných experimentů stále není zcela uzavřena otázka volby vhodné parametrizace pro dané podmínky.

Velmi obsáhlá experimentální práce [89] na téma volby optimálního nastavení parametrů těchto metod ukazuje, v jakých hodnotách je vhodné volit základní parametry - počet pásem banky filtrů a délku parametrizačního vektoru. Podobně lze metody porovnat z pohledu tvaru jednotlivých filtrů [90].

Tato nastavení pouze upravují již zvolené algoritmy a pomáhají vyladit systém pro jistý stupeň citlivosti. Na druhou stranu ponechává jednotlivým metodám jejich vlastnosti z pohledu robustnosti jednotlivých algoritmů zpracování. Cílem následující sekce je porovnání vlivu jednotlivých bloků zpracování s ohledem na jejich použití v standardním parametrizačním procesu.

6.2.1 Popis modifikovaných metod

Modifikace standardních metod záměnou hlavních funkčních bloků je využívána již v několika dříve navrhovaných úpravách základních parametrizačních technik ([66], [46], [125], [106]). V [46] autoři analyzují jednotlivé kroky výpočtu parametrizace a navrhují metodu RPLP, která kombinuje jednotlivé kroky výpočtu standardních parametrizací s ohledem na možný přínos k robustnosti metody zpracování signálu.

Podobně práce [125] a [106] kombinují Melovskou banku s PLP koeficienty a tuto modifikovanou metodu srovnávají za různých podmínek se standardními parametrizacemi a následně je i kombinují za účelem získání robustnějších systémů.

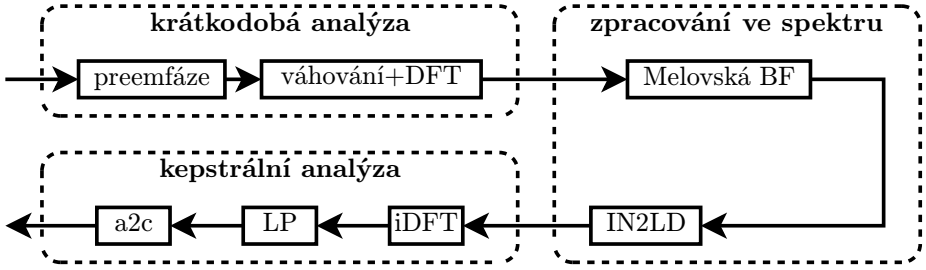
Výše zmíněné experimenty ukazují, že i přes obecně výhodné vlastnosti technik MFCC a PLP lze v rámci jejich výpočtu hledat optimalizace i méně komplexními přístupy. Proto byly navrženy metody, které kombinují jednotlivé bloky standardních metod tak, aby využily výhodných vlastností těchto metod.

Tato sekce se zabývá teoretickou analýzou a popisem tří modifikací s ohledem na jejich robustnost v úloze rozpoznávání řeči v šumovém prostředí. Experimentální ověření předpokladů při návrhu jednotlivých metod pak realizují experimenty v následujících sekcích a kapitolách.

RPLP - Revised PLP

První modifikace vychází z článku [46]. Tato metoda, jak je ukázáno na obr. 6.4, využívá pro spektrální zpracování signálu výpočetní postup parametrizace MFCC, pro přechod do kepstrální oblasti je ale použit výpočet přes LPC analýzu namísto DCT transformace. Důležitým bodem je ovšem zachování dvojitého potlačení dynamiky signálu, které umožňuje použití nižšího řádu pro následnou LPC analýzu. Zatímco první fáze potlačení je provedena pomocí preemfázového filtru, před LPC analýzou je aplikována IN2LD transformace zmíněná při popisu PLP parametrizace.

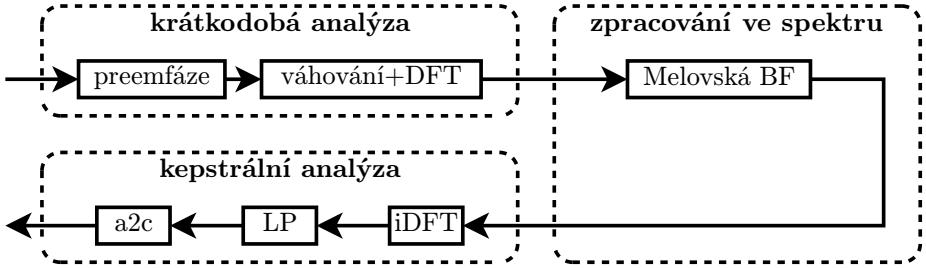
Při porovnání se standardními metodami MFCC a PLP dosahují autoři s použitím této úpravy na sadě dat z různých prostředí zlepšení *WER* z 20,5 % resp. 20,6 % na 19,9 %. Metoda RPLP tak realizuje modifikaci, která využívá předzpracování z metody MFCC pro potlačení citlivosti LPC analýzy.



Obrázek 6.4: Blokové schéma výpočtu standardních parametrizací RPLP

MFLP - Mel-Frequency Linear Prediction

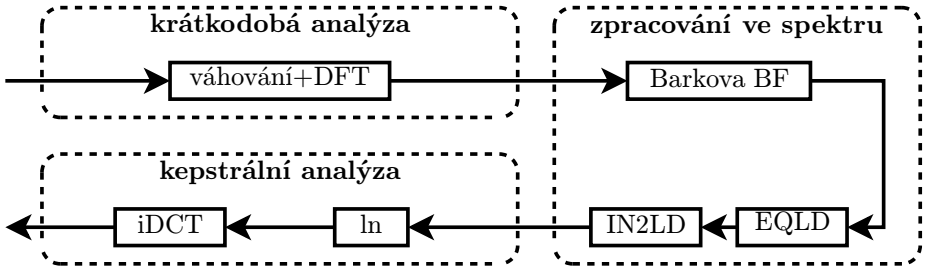
Jak název napovídá, také metoda MFLP na schématu 6.5 využívá předzpracování z parametrizace MFCC a pro převod do kepstrální oblasti je využita LPC analýza. Na rozdíl od metody RPLP ale nevyužívá bloku IN2LD. To znamená, že signál, který je zpracován LPC analýzou má zachovánu vyšší dynamiku, na druhou stranu tak mohou být zachyceny větší detaily signálu. Metoda tak odkrývá vliv dvojitého potlačení dynamiky signálu na LP zpracování při použití preemfázového filtru.



Obrázek 6.5: Blokové schéma výpočtu standardních parametrizací MFLP

BFCC - Bark Frequency Cepstral Coefficients

Poslední modifikovanou metodu BFCC (obr. 6.6) charakterizuje opačný postup výpočtu (s ohledem na výpočet standardních metod) oproti MFLP. Realizuje tak metodu, která zachovává postup inspirovaný vnímáním zvuku lidským uchem z metody PLP, výsledek spektrální analýzy je ale převeden do kepstrální oblasti DCT transformací. Metoda je inspirována snahou potlačit v parametrizačním procesu výše zmiňovanou citlivost LPC analýzy na šum v signálu při zachování zpracování signálu odvozeného ze znalosti lidského ucha.



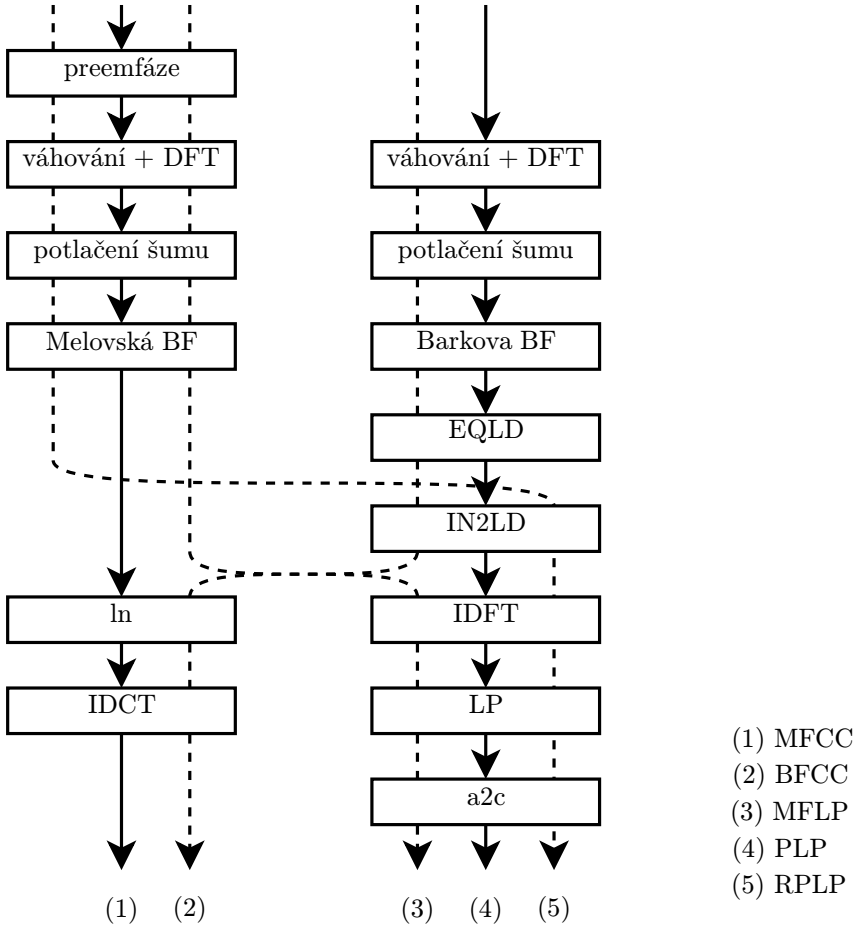
Obrázek 6.6: Blokové schéma výpočtu standardních parametrizací BFCC

6.2.2 Shrnutí vlastností parametrizačních technik

Ačkoliv existuje mnoho různých přístupů a možností modifikace standardně používaných parametrizačních metod, tato práce používá základní přístup záměny základních funkčních bloků. To umožňuje změnit vlastnosti metod a využívat výhodného chování jedné metody v rámci metody druhé bez navýšení výpočetní náročnosti procesu.

Navrhované modifikace vedou na 5 analyzovaných technik, které jsou dále porovnávány s ohledem na příspěvek k robustnosti ASR systému:

- PLP – standardní metoda založená na LPC analýze signálu s dobrými výsledky především v úloze rozpoznávání řeči v tichém prostředí,
- MFCC – velmi často používaná standardní metoda pro její dobré výsledky v obecném prostředí oproti PLP,
- RPLP – technika založená na LP zpracování signálu s prvky MFCC. Výsledky v [46] ukazují její vhodné vlastnosti ve zpracování signálu v obecných šumových podmínkách,
- MFLP – technika podobná RPLP ovšem zachovávající vyšší dynamiku řečového signálu pro zachování jeho kvality,
- BFCC – komplementární metoda k MFLP, která využívá předzpracování signálu odvozené od vnímání řeči lidským uchem a aplikující DCT namísto LPC analýzy pro přechod do kepstrální oblasti.



Obrázek 6.7: Blokové schéma výpočtu jednotlivých parametrizací

6.3 Rozšířené spektrální odečítání

V rámci výpočetního procesu standardní parametrizace je možné aplikovat další algoritmy, které přispívají k zvýšení robustnosti systému vůči šumu pozadí. V této práci je analyzován vliv algoritmu rozšířeného spektrálního odečítání (ESS) použitého v rámci jednotlivých parametrizačních technik.

Jak je uvedeno v sekci 2.2.1, lze algoritmus ESS nastavit pomocí parametrů α a p tak, aby dosahoval optimálních výsledků pro dané podmínky. V experimentech této práce jsou konstanty zvoleny na experimentální bázi na hodnotách $\alpha = 1$ a $p = 0,95$.

Stejně jako jiné metody, lze algoritmus spektrálního odečítání aplikovat v různých fázích zpracování spektrální podoby signálu. Jak ukazuje např. [79], při aplikaci algoritmu před bankou filtrů v rámci použité parametrizační techniky působí daná metoda na méně vyhlazený signál a může tak dosáhnou přesnějšího působení na šumovou složku signálu. Proto je i v experimentech této práce algoritmus použit před bankou filtrů, viz schéma 6.7, blok potlačení šumu.

6.4 Experimentální část

Následující sekce popisuje analýzu optimalizovaného nastavení základních vlastností parametrizačních technik. Pro porovnání byl použit fragment Clean a rozpoznávač číslovek. Výsledek této analýzy je poté využit pro obecné nastavení parametrů rozpoznávače v rámci této práce.

V další části je již analyzován rozpoznávač řeči se standardními a modifikovanými parametrizačními technikami, které jsou doplněny o metodu rozšířeného spektrálního odečítání. Pro objektivní srovnání vlivu jednotlivých nastavení předzpracování signálu jsou experimenty provedeny nejen na promluvách z databáze SPEECON, ale předzpracování je také použito pro standardizovanou úlohu na databázi Aurora. Metody jsou porovnávány na úloze rozpoznávání sekvence číslovek s využitím rozpoznávače fonémů. Výjimkou jsou experimenty na databázi AURORA3, kdy je použit rozpoznávač celých slov.

6.4.1 Segmentace signálu pro krátkodobou analýzu

Jak bylo zmíněno v předchozí části, vliv nastavení parametrizačních technik byl rozsáhle popsán v rámci práce [87]. Pro tyto rozsáhlé experimenty byla použita databáze telefonních promluv, které se v některých ohledech mohou lišit od řečových záznamů použitých v této práci. Následující analýza je proto zaměřena především na zkoumání dopadu nastavení těchto parametrů v podmínkách použitých v této práci, např. pro zvolené vzorkování 16 kHz a použité databáze dle kap. 4.

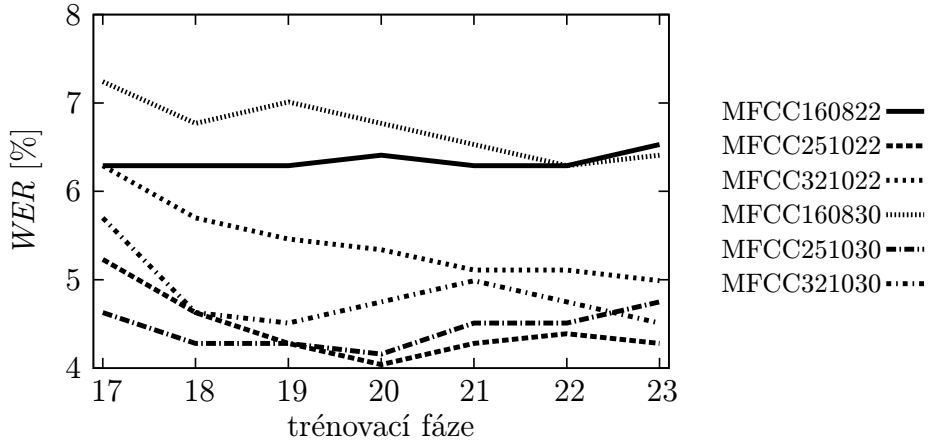
Analýza se zabývá nastavením základních parametrů s níže uvedenými volbami hodnot. Tyto volby jsou navrženy s ohledem na vlastnosti řečového signálu, především krátkodobou stacionaritu a rozložení informace v rámci frekvenčního pásma [93]:

- *Délka váhovacího okna:* 32 ms, 25 ms, 16 ms
- *Krok segmentace:* 8 ms, 10 ms, 16 ms
- *Počet pásem banky filtrů:*
  - 22 – standardizované nastavení pro vzorkovací frekvenci 8 kHz
  - 30 – nastavení, které zachovává rozložení 22 pásem v rozsahu 0 – 4 kHz při vzorkovací frekvenci 16 kHz

Nastavení každého experimentu je popsáno kódem mfccAABBCC, kde AA je dvoumístné vyjádření délky okna, BB je dvoumístné vyjádření kroku segmentace a CC značí nastavení počtu pásem banky filtrů. Pro modelování byla použita sada 44 3-stavových HMM modelů českých fonémů, model krátké a dlouhé pauzy. Modely jsou rozděleny na 32 mixtures.

Graf na obr. 6.8 ukazuje vývoj chybovosti rozpoznávače založeného na parametrizaci s různým nastavením v jednotlivých fázích cyklického přetrénování modelů ve fázi 17 až 23. Z tohoto vývoje lze vyčíst nejen výslednou úspěšnost rozpoznávače s použitím daných modelů, ale také trend, který trénování vykazuje. Tím lze do jisté míry odhadnout například přetrénovanost modelů.

Graf ukazuje, že parametrizace s oknem 16 ms a krokem 8 ms je v porovnání s ostatními volbami nejméně výhodná. Chybovost této segmentace výrazně ovlivňuje vyšší míra zastoupení chybně vložených slov, což mimo jiné souvisí také se závěry práce [96]. Pro tuto segmentaci není při zpracování tolik zastoupen vliv průměrování a dochází k vyšší citlivosti na rušení v signálu. Naopak použití parametrizace s 25 ms oknem a segmentací 10 ms vykazuje nejlepší výsledky rozpoznávání.



Obrázek 6.8: Chybovost *WER* pro rozdílná nastavení segmentace a délky banky filtrů

Vzhledem k dosaženým výsledkům je ve většině následujících experimentů dále používána následující volba:

- *Délka segmentačního okna:* 25 ms
- *Krok segmentace:* 10 ms
- *Počet pásem Melovské banky filtrů:* 22
- *Počet pásem Barkovy banky filtrů:* 19

6.4.2 Robustnost parametrizačních metod v nepřizpůsobených podmínkách

V prvním kroku byly standardní a modifikované metody porovnány v úloze rozpoznávání řeči v různě přizpůsobených šumových podmínkách. Cílem experimentu bylo získat základní přehled o citlivosti jednotlivých metod na rozdílné trénovací a testovací podmínky.

ASR systém byl trénován na trénovací části databáze Clean případně Noisy s použitím signálů z kanálu CS0 nebo CS1. Testování probíhalo na databázi SNR-set, tedy databázi signálů zachycujících sekvenci číslovek v tichém prostředí. Tím byly definovány čtyři základní případy: Zatímco použití trénovací sady Clean z kanálu CS0 simuluje shodné podmínky trénování a testování, trénování na sadě Noisy a použití

Šumové prostředí	Nahrávací kanál	Přizpůsobení	
		šumové podmínky	kanál
Clean	CS0	ano	ano
	CS1	ano	ne
Noisy	CS0	ne	ano
	CS1	ne	ne

Tabulka 6.1: Popis přizpůsobení trénovacích podmínek

signálů z kanálu CS1 simuluje rozdílné podmínky jak v šumovém pozadí, tak v použité přenosové cestě.

Tabulka 6.2 zobrazuje výsledky rozpoznávání pro jednotlivé trénovací podmínky. Z nich je patrný vysoký vliv nepřizpůsobeného nahrávacího kanálu, tedy nepřizpůsobení v konvoluční oblasti. Například pro parametrizaci MFCC a trénovací množinu Clean dosahuje chybovost hodnoty 4,55 % resp. 22,9 % pro náhlavní resp. pro hands-free mikrofon.

Naopak vzhledem k tomu, že v rámci databáze SPEECON nejsou mnohdy rozdíly v zašumění signálů v jednotlivých prostředích výrazné (viz analýza SNR na obr. A.1), jsou rozdíly pro rozdílné trénovací šumové podmínky menší, viz např. rozdíl hodnot 4,55 % a 6,06 % pro MFCC a kanál CS0.

		MFCC	PLP	MFLP	BFCC	RPLP
Clean	CS0	4,55	4,55	6,40	5.72	4,38
	CS1	22,90	22,56	30,81	31,99	21,38
Noisy	CS0	6,06	7,07	7,58	8,42	5,56
	CS1	26,94	31,14	34,34	42,59	20,71

Tabulka 6.2: Výsledná *WER* při trénování na rozdílných fragmentech SPEECON a testování na databázi SNR-set

Výsledek této analýzy lze shrnout v následujících bodech:

- Standardní parametrizační techniky MFCC a PLP dosahují srovnatelných výsledků spolu s technikou RPLP pro shodné i rozdílné podmínky trénování a testování.
- Metoda RPLP dosahuje mírně lepších výsledků, především ve srovnání s metodou PLP v nepřizpůsobených podmínkách.
- Metody MFLP (technika s vyšším stupněm dynamiky signálu pro LPC analýzu) i BFCC (DCT analýza signálu s perceptuálním předzpracováním) dosáhly oproti standardním technikám horších výsledků. Vliv zachování vyšší dynamiky signálu při LPC analýze a naopak potlačení dynamiky signálu při přímé kepstrální

analýze se tak ukazuje jako nevýhodný pro možné modifikace. Metoda MFLP ale v případě nižších rozdílů kanálů a vyššího zašumění signálu dosahuje hodnot podobných, jako metoda PLP.

- Na výsledcích rozpoznávání se ve velké míře projevilo především rozdíl přenosových kanálů trénovacích a testovacích signálů. Naopak přídavný šum ovlivnil výsledky rozpoznávání pouze v malé míře. Této vlastnosti je využito například při trénování rozpoznávače na obecných šumových podmínkách (8.1.2).

6.4.3 Reálné prostředí - přizpůsobené podmínky

Následující experiment porovnává navrhované metody s ohledem na robustnost již pro přizpůsobené podmínky, což je nejčastější způsob nastavení rozpoznávače. Systém byl trénován i testován na jednotlivých fragmentech databáze SPEECON s odpovídající trénovací a testovací sadou dat.

Výsledky jsou srovnávány pro jednotlivé fragmenty. Navíc je pro zhodnocení obecné citlivosti parametrizace na šumové podmínky uvedena průměrná chybovost pro fragmenty Office, Clean a Noisy. V experimentu na fragmentu All byly modely trénovány na všech dostupných podmínkách – to simuluje systém trénovaný na obecných podmínkách, viz sekce 2.2.2. Modely v tomto případě vystihují signál v jeho obecnější podobě a lze očekávat vyšší úspěšnost tohoto systému v zašuměných podmínkách (oproti hodnotě pro fragment Noisy), zatímco ve srovnání s hodnotou Clean bude chybovost vyšší.

Srovnání v tabulce 6.3 ukazuje mírně lepší výsledky pro PLP oproti MFCC v případě čistých signálů (tišší prostředí, směrový mikrofon), naopak ve více rušivých podmínkách je chybovost nižší pro techniku MFCC, pro nejvíce zašuměné prostředí (fragment Noisy, kanál CS1) téměř o 40 %.

V porovnání se standardními technikami přináší metoda RPLP další snížení chybovosti v zašuměném prostředí. Jak ukazují hodnoty v tab. 6.3, tato metoda dosahuje srovnatelného skóre s metodou MFCC. Výsledky pro metodu MFLP na zašuměných datech mírně překonávají PLP metodu a ukazují tak na možnou výhodnost zachování dynamiky signálu pro zašuměné signály.

Metoda BFCC opět ztrácí na standardní techniky a dosahuje výrazně nižší úspěšnosti rozpoznávání, především pro vysoce zašuměné podmínky hands-free mikrofonu a Noisy fragmentu databáze. Vliv potlačení dynamiky signálu tak přispívá především ke ztrátě robustnosti metody.

6.4.4 Spektrální odečítání

Pro snížení vlivu aditivního rušení z okolí na rozpoznávaný řečový signál byl použit algoritmus ESS popsaný v sekci 6.3 aplikovaný dle schématu 6.7. Vzhledem k reálnému charakteru prostředí a obecně nižší míře zašumění signálů lze očekávat, že vliv ESS v experimentálních výsledcích bude přinášet výraznější zlepšení v řádu 10 a více

náhlavní mikrofon (CS0)					
	MFCC	PLP	MFLP	BFCC	RPLP
Office	5,74	6,14	9,35	6,14	6,14
Clean	7,19	6,74	9,47	8,90	6,62
Noisy	13,98	18,97	16,47	19,95	11,84
Průměr	10,59	12,86	12,97	14,43	9,23
All	14,53	13,94	13,57	13,48	10,37

hands-free mikrofon (CS1)					
	MFCC	PLP	MFLP	BFCC	RPLP
Office	9,48	10,68	14,02	12,15	10,41
Clean	10,73	9,93	13,36	12,21	9,47
Noisy	15,14	24,67	20,04	31,43	16,74
Průměr	12,94	17,30	16,70	21,82	13,11
All	15,04	18,65	18,56	20,02	16,00

Tabulka 6.3: Chybovost *WER* na databázi SPEECON bez doplňujících algoritmů potlačení šumu

procent jen výjimečně. Analýza vlivu algoritmu ESS při použití těchto signálů na druhou stranu přináší také představu o vlivu této metody na čistý signál a jeho možnou degradaci. Výsledek rozpoznávání s použitím ESS ukazuje tabulka 6.4.

Při porovnání výsledků s předchozím experimentem je zřejmé, že algoritmus nepřináší pro čisté podmínky (především prostředí Office) zlepšení a naopak dochází ke snížení úspěšnosti rozpoznávání. To je způsobeno právě nižší mírou zašumění signálů, kdy algoritmus mnohem výrazněji ovlivňuje vlastní signál. Při zkoumání chování algoritmu v zašuměných podmínkách je již v mnoha případech význam potlačování šumu v signálu patrný. Například pro prostředí Noisy, kanál CS0 a parametrizaci PLP přispěl algoritmus ke snížení chybovosti z 18,97 % na 13,09 %, tedy o 31 %. Především pro metody, které dosahují vyšší úspěšnosti pro čisté signály (např. PLP), proto může být algoritmus vedoucí k potlačení zašumění signálu výhodným doplňkem.

6.4.5 Spektrální odečítání – AURORA3

Pro srovnání navrhovaných postupů byly provedeny experimenty na databázi AURORA3 pro 3 vybrané jazyky a pro různou míru shody trénovacích a testovacích podmínek, jak je popsáno v 4.2.4.

Chybovost rozpoznávače s použitím jednotlivých parametrizačních technik v tabulce 6.5 je porovnána k základní hodnotě (baseline), která je pro dané podmínky dosažena se standardním nastavením parametrizace dle [74]. Tomu odpovídají následující parametry:

náhlavní mikrofon (CS0)					
	MFCC	PLP	MFLP	BFCC	RPLP
Office	6,94	5,47	9,88	7,08	6,01
Clean	7,88	6,51	11,30	8,45	7,76
Noisy	11,40	13,09	17,36	16,30	11,67
Průměr	9,64	9,80	14,33	12,38	9,72
All	11,70	11,75	14,63	14,58	11,52

hands-free mikrofon (CS1)					
	MFCC	PLP	MFLP	BFCC	RPLP
Office	10,41	11,48	12,15	14,29	11,21
Clean	10,96	11,19	12,10	13,58	10,84
Noisy	15,85	24,13	16,30	34,55	12,91
Průměr	13,41	17,66	14,20	24,07	11,88
All	13,21	18,28	14,72	30,30	12,43

Tabulka 6.4: Chybovost *WER* na databázi SPEECON s použitím ESS

- *Parametrizační metoda:* MFCC
- *Koeficient preemfáze:* 0,97
- *Délka váhovacího okna:* 25 ms
- *Krok segmentace:* 10 ms
- *Počet pásem banky filtrů:* 23
- *Zlomová frekvence horní propusti:* 64Hz
- *Statické koeficienty:* 12 + 1 log. energie

Ve sloupci ETSI je navíc použit standard ETSI ES 202 050 [22]. Tato metoda používá dvoufázový algoritmus zvýrazňování řeči doplněný o detekci řečové aktivity. Jelikož je ve výše popsaných experimentech použit jen algoritmus ESS, je zřejmé, že rozpoznávač založený na uvedeném ETSI standardu vykazuje výrazně nižší chybovost. ETSI standard je zde proto zobrazen pouze pro srovnání původních výsledků s algoritmem s výrazně vyšší komplexností a srovnatelné výsledky jsou očekávány při použití algoritmů s VAD, které budou popsány později.

Použitá databáze AURORA obsahuje nahrávky z automobilu, kde se projevuje především přítomnost aditivního rušení, zatímco míra ostatních typů rušení (např. odrazy signálu) je nižší. Proto i příspěvek techniky ESS je zřejmý, převážně pro podmínky s nižší mírou rozdílnosti trénovacích a testovacích dat. V souladu s předchozí sekcí tak bylo dosaženo výrazných zlepšení chybovosti při použití technik na bázi LPC analýzy, až o 24 % pro PLP a Španělštinu, 48 % pro MFLP a Finštinu a 21 % pro RPLP a Dánštinu.

Španělština							
	baseline	MFCC	PLP	MFLP	BFCC	RPLP	ETSI
WM	13,15	15,18	10,04	12,14	10,74	10,72	6,58
MM	26,26	33,68	28,59	33,96	29,83	24,32	13,27
HM	57,77	56,78	60,30	56,24	62,65	56,63	15,79
Průměr	32,39	35,21	32,98	34,11	34,41	30,56	11,88

Finština							
	baseline	MFCC	PLP	MFLP	BFCC	RPLP	ETSI
WM	9,61	8,09	6,56	5,16	9,76	6,44	2,52
MM	27,63	30,16	60,19	22,02	33,31	46,31	12,72
HM	68,94	51,84	64,66	62,01	77,42	59,72	18,83
Průměr	35,39	30,03	43,80	29,73	40,16	37,49	11,36

Dánština							
	baseline	MFCC	PLP	MFLP	BFCC	RPLP	ETSI
WM	22,20	17,04	18,30	18,37	18,52	17,50	15,87
MM	53,60	55,93	49,72	48,87	51,55	50,00	38,98
HM	68,10	59,21	69,24	58,66	79,27	65,20	37,81
Průměr	47,97	44,06	45,75	41,97	49,78	44,23	30,89

Tabulka 6.5: Chybovost *WER* na databázi AURORA s použitím ESS v porovnání se standardy

Kapitola 7

Detekce řečové aktivity pro účely rozpoznávání řeči

Předchozí kapitola byla věnována technikám pro zvýšení robustnosti rozpoznávače řeči na úrovni parametrizace. Tato kapitola se zabývá analýzou možností přínosu detekce řečové aktivity v rámci předzpracování signálu v rozpoznávacích systémech.

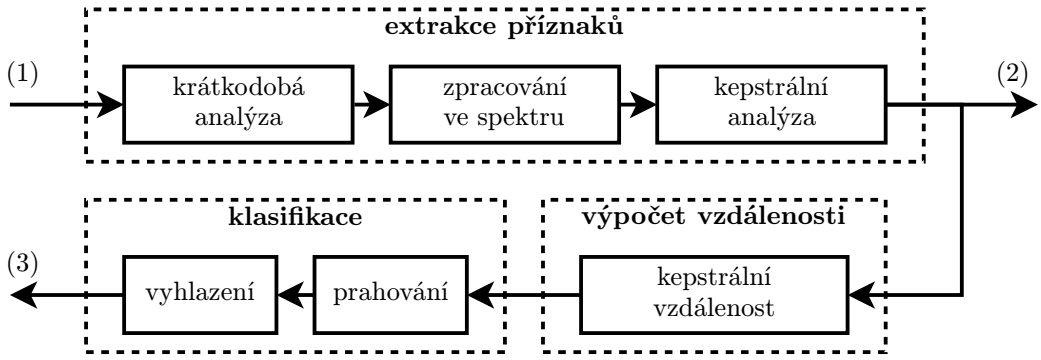
Detekce řeči je významnou součástí mnoha aplikací pro zpracování řeči. Nachází využití v systémech pro zvýrazňování řeči k aktualizaci parametrů modelu pozadí řeči, ve vokodéru pro přenos pouze řečového signálu a také v řečových rozpoznávačích pro detekci začátku a konce promluvy a pro odstranění neřečových částí signálu.

V rámci této kapitoly je popsán vliv použití VAD na bázi vypouštění segmentů. Tato metoda je založena na detekci neřečových segmentů v signálu a jejich odstranění ze signálu v dalších krocích rozpoznávacího procesu. Tím je možné potlačit chyby způsobené nesprávnou klasifikací segmentu jako řečový.

7.1 Kepstrální detektor řečové aktivity

Kepstrální detektor je postaven na struktuře standardního VAD systému na bázi kepstrální vzdálenosti aktuálního segmentu od charakteristik pozadí. Pro účely kepstrální analýzy řečového signálu se obvykle pracuje s reálným DFT kepstrem či LPC kepstrem. Při aplikaci v rámci rozpoznávání řeči se ukazuje jako výhodné použít parametrizace MFCC či PLP, neboť je-li detektor založený na těchto příznacích dále využíván v ASR, může se využít již existující reprezentace signálu použité pro rozpoznávání řeči.

Proces kepstrální detekce řečového signálu lze rozdělit do tří základních kroků podle obr. 7.1. V prvním kroku je nalezena kepstrální reprezentace signálu, dalším krokem je výpočet vzdálenosti pro popis odlišnosti daného segmentu signálu od odhadu charakteristik pozadí řeči, v posledním kroku je pak získaná vzdálenost porovnána s prahovou hodnotou pro konečné rozhodnutí o přítomnosti řeči. Tento výsledek může být mírně vyhlazen pro potlačení krátkých chybných rozhodnutí.



Obrázek 7.1: Blokové schéma použitého detektoru řeči

7.1.1 Kepstrální vzdálenost

V kepstrálních detektorech je obvykle využívána standardní kepstrální vzdálenost (CD) mezi aktuálním segmentem signálu a šumem pozadí. Potřebný odhad kepstra šumu pozadí je v tomto případě možno získat z průměrného kepstra signálu v pauzách řeči. Kepstrální vzdálenost pro i-tý segment lze vyjádřit jako

$$CD[i] = \sum_{k=1}^{p} (c_k[i] - \bar{c}_{o,k}[i])^2 \tag{7.1}$$

nebo v jednodušší podobě s nižší dynamikou jako

$$CD'[i] = \sum_{k=1}^{p} |c_k[i] - \bar{c}_{o,k}[i]| \tag{7.2}$$

pro p kepstrálních koeficientů. Detektory založené na této kepstrální vzdálenosti obvykle využívají výsledku detekce k získání průměrované kepstrální vzdálenosti $\bar{c}_{o,k}[i]$, což zavádí do algoritmu zpětnou vazbu. Tuto nevýhodu eliminuje použití kumulované kepstrální vzdálenosti (CDC) s využitím diferenční kepstrální analýzy.

Diferenční kepstrum je typicky aproximováno vztahem

$$\delta_k^{(M)}[i] = \left[\sum_{j=1}^{M} j(c_k[i+j] - c_k[i-j])\right] / \left[2\sum_{j=1}^{M} j^2\right] \tag{7.3}$$

kde M vyjadřuje řád odhadu diferenčního kepstra. Takto získaný odhad lze potom použít pro stanovení vzdálenosti pomocí kumulativních součtů (integrace) jednotlivých diferenčních kepstrálních koeficientů jako

$$CDC_k[i] = \sum_{j=0}^{i} \delta_k^{(M)}[j]. \tag{7.4}$$

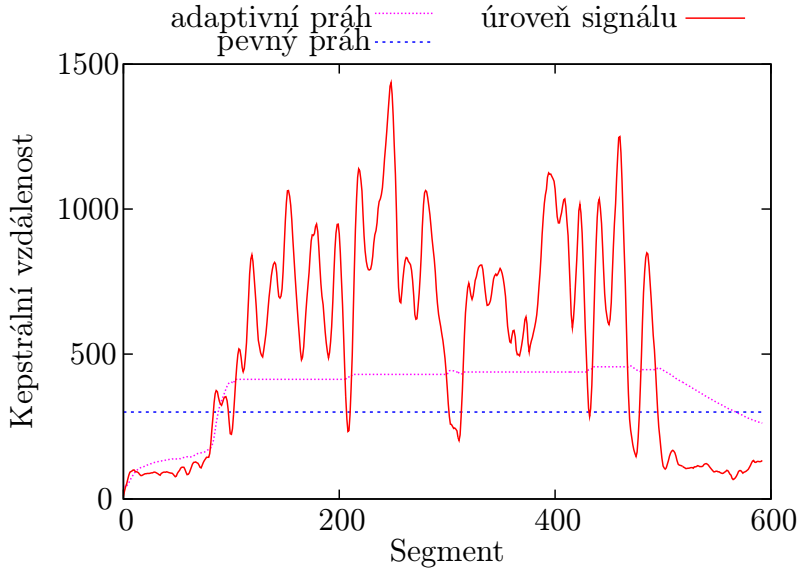
Vzhledem k nekorelovanosti jednotlivých kepstrálních koeficientů lze tyto koeficienty počítat jednotlivě a poté získat celkovou kepstrální vzdálenost jako součet jednotlivých vzdáleností

$$CD[i] = \sum_{k=1}^{p} |CDC_k[i]| = \sum_{k=1}^{p} \left| \sum_{j=0}^{i} \delta_k^{(M)}[j] \right|. \tag{7.5}$$

Takto v principu vyjádřená kepstrální vzdálenost pak odpovídá vztahu (7.2) s tím, že se počítá vzdálenost od prvního segmentu. Výše uvedená metoda pro určení kepstrální vzdálenosti přináší výhodu ve vyhlazení výsledků, které omezí vliv náhodných výchylek.

7.1.2 Prahování

Na základě získané kepstrální vzdálenosti lze již provést rozhodnutí o výsledku detekce. K tomu jsou typicky využívány dva níže uvedené základní přístupy k prahování.



Obrázek 7.2: Průběh nastavení prahu pro různé typy prahování v rámci průchodu signálem

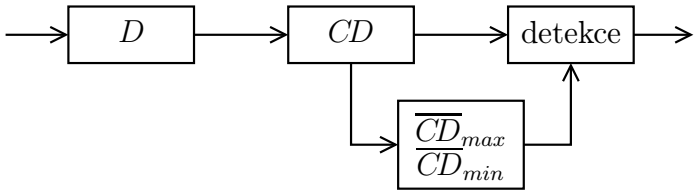
Dynamicky nastavený pevný práh (fixed)

Hodnota prahu THR_f

$$THR_f = \overline{CD}_{min} + \frac{p}{100}\left(\overline{CD}_{max} - \overline{CD}_{min}\right) \tag{7.6}$$

je určena na základě analýzy celé promluvy a toto počáteční nastavení prahu není v průběhu detekce řeči v dané promluvě modifikováno. Hodnota p je volena experimentálně na hodnotu typicky okolo 20 %.

Uvedené průměry $\overline{CD}_{min}$ a $\overline{CD}_{max}$ udávají průměr z 5 % nejnižších resp. nejvyšších hodnot získané kepstrální vzdálenosti CD. Toto průměrování omezuje vliv ojedinělých extrémních výchylek CD.



Obrázek 7.3: Blokové schéma pro algoritmus fixního prahování

Adaptivní práh řízený podle pozadí řeči (adapt)

Adaptivní práh THR_a

$$THR_a[i] = \bar{\mu}_{CD_n}[i] + 2\,\bar{\sigma}_{CD_n}[i] \tag{7.7}$$

je nastaven podle střední hodnoty kepstrální vzdálenosti v pauzách řeči CD_n zvýšené o dvojnásobek standardní odchylky. To umožňuje postihnout variabilitu charakteristik prostředí. K obnově hodnoty THR_a dochází pro každý vyhodnocený segment pauzy, případně na základě vyhodnocení několika po sobě jdoucích segmentů. Adaptivní nastavení prahu detekce již zavádí do algoritmu detekce zpětnou vazbu, neboť k obnově hodnot prahu dochází v pauzách řeči, vyhodnocených daným detektorem.

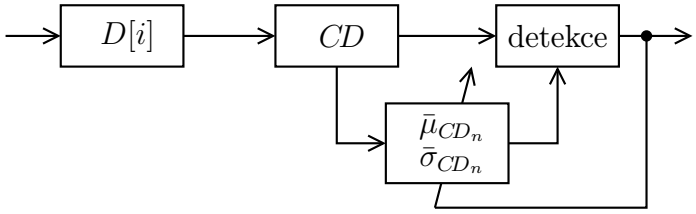
Hodnoty $\bar{\mu}_{CD_n}[i]$ a $\bar{\sigma}_{CD_n}[i]$ se je možné obnovovat také pomocí exponenciálního průměrování s relativně dlouhou časovou konstantou. To umožňuje zamezit vlivu náhodných výchylek v charakteristice pozadí řeči. Můžeme tedy psát

$$\bar{\mu}[i] = q\,\bar{\mu}[i-1] + (1-q)\,CD[i] \tag{7.8}$$

$$\bar{\sigma}^2[i] = q\,\bar{\mu}_2[i] - \bar{\mu}^2[i] \tag{7.9}$$

pro hodnotu q typicky 0,95 – 0,98.

Adaptivní algoritmus vyžaduje na začátku promluvy krátký úsek pauzy, ve kterém se inicializují hodnoty prahu. V této fázi jsou první segmenty signálu využity pro iniciální nastavení hodnot $\bar{\mu}_{CD_n}[i]$ a $\bar{\sigma}_{CD_n}[i]$.



Obrázek 7.4: Blokové schéma pro algoritmus adaptivního prahování

Adaptivní práh řízený dynamikou signálu

Volba adaptivního prahu s pevně nastavenou konstantou q pro obnovu hodnot je základním algoritmem, který je založen na analýze dynamiky signálu v pauzách řeči. Adaptivní práh řízený dynamikou signálu určuje práh pro klasifikaci signálu pro každý vzorek nikoliv na základě kepstrální vzdálenosti od odhadu charakteristik pozadí, ale na bázi zjištěné dynamiky signálu. Pro vlastní nastavení prahu pak platí

$$THR_{as}[i] = D_{min}[i] + \frac{p}{100}\,(D_{max}[i] - D_{min}[i]) \tag{7.10}$$

kde $\bar{D}_{min}[i]$ a $\bar{D}_{max}[i]$ jsou hodnoty hraniční hodnoty pro určení dynamického rozpětí signálu. Ty jsou průběžně upravovány dle vztahu

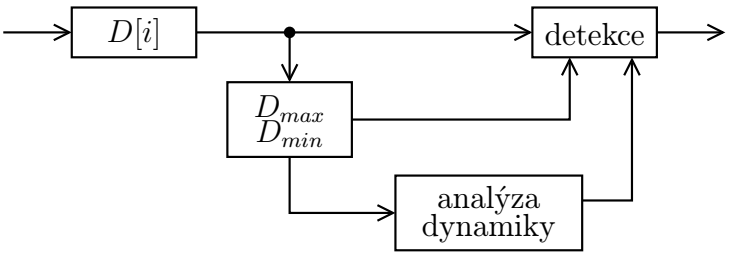
$$D_{min}[i] = \begin{cases} q_{min1}\, D_{min}[i-1] + (1 - q_{min1})\; D[i] & \text{pro } D[i] < D_{min}[i-1] \\ q_{min2}\, D_{min}[i-1] + (1 - q_{min2})\; D[i] & \text{v ostatních případech} \end{cases} \tag{7.11}$$

$$D_{max}[i] = \begin{cases} q_{max1}\, D_{max}[i-1] + (1 - q_{max1})\; D[i] & \text{pro } D[i] < D_{max}[i-1] \\ q_{max2}\, D_{max}[i-1] + (1 - q_{max2})\; D[i] & \text{v ostatních případech} \end{cases} \tag{7.12}$$

Hodnoty q_{min1}, q_{max1}, q_{min2} a q_{max2} vychází často z experimentálních měření a většinou se pohybují v rozmezí 0,95 – 0,98.

Zároveň je vhodné nastavit kritérium pro minimální dynamický rozsah signálu na základě odhadu dynamiky pozadí řeči, které zamezí výraznému poklesu hraničních hodnot v případě dlouhé pauzy.

Adaptivní práh na bázi dynamiky signálu eliminuje případnou chybu při odhadu charakteristik šumového pozadí a umožňuje aplikovat adaptivní algoritmus bez zpětné vazby, viz obr. 7.5.



Obrázek 7.5: Blokové schéma pro algoritmus adaptivního prahování na bázi dynamiky

7.1.3 Vyhlazení výsledků detekce

Výstup detektoru často obsahuje krátké úseky falešné detekce řeči nebo pauzy, které lze odstranit vyhlazením výsledků. To bývá prováděno typicky mediánovým filtrem.

V našem algoritmu je použit filtr řádu 3, přičemž je možné použít i řád 5. Vyšší řád již však může vést ke ztrátě informace, a proto není vhodný.

Je nutné dodat, že jisté míry vyhlazení je dosaženo již použitím vhodné parametrizace. Toho je dosaženo i pro parametrizace MFCC a PLP použitou nelineární filtrací. V případě PLP je navíc vyhlazení podpořeno provedenou LPC analýzou. K určitému vyhlazení dochází navíc také již při výpočtu diferenčních kepstrálních koeficientů. Výpočet kepstrální vzdálenosti založený na diferenčním kepstru ovlivňuje míru vyhlazení řádem M aproximace diferenčního kepstra $\delta_k^{(M)}$. Zvýšením řádu se docílí vyššího vyhlazení, které ale opět může způsobit ztrátu informace v podobě nedetekovaných krátkých segmentů.

Podobně může ovlivnit výsledné vlastnosti systému volba řádu mediánové filtrace výsledné detekce.

7.2 Možnosti nastavení VAD algoritmů

Jak je ukázáno v předchozím textu, celý detekční algoritmus má mnoho možných variant a nastavení, jimiž lze ovlivnit výsledné vlastnosti detektoru, které jsou ve velké míře určeny konečným využitím VAD. Při detekci pauz v řeči, které jsou využity pro následnou úpravu charakteristik šumového pozadí pro metody zvýrazňování řeči, je důležité korektní označení segmentů pauzy. Naopak pro použití v rozpoznávačích řeči, což je i případ využití VAD v této práci, je důležité, aby na základě chybné detekce nebyly odstraňovány řečové segmenty.

Pro různé aplikace detektoru je možné volit také algoritmus prahování. Pro online zpracování, typické při detekci začátku a konce promluvy pro ASR, je nutné použít adaptivní algoritmus. Zde není známa informace o celém signálu, a tedy jeho dynamice, podle níž se určuje hodnota prahu pro detekci. Tyto adaptivní algoritmy jsou však velmi citlivé na nastavení parametrů adaptace a pro obecné šumové podmínky není možné zajistit jejich optimální volbu. Je tedy důležité tyto parametry volit s ohledem na konkrétní aplikaci.

Na druhou stranu fixní prahování poskytuje výhodné vlastnosti pro off-line zpracování signálu, například při trénování rozpoznávače řeči nebo při blokovém zpracování řečových dat po detekci začátku a konce promluvy.

7.3 Experimentální část

Výše popsané algoritmy jsou v následující sekci analyzovány z pohledu kritérií, která umožňují popsat chování detektoru z pohledu detekce přítomnosti řeči v signálu. V další části je již detektor použit v procesu extrakce řečových příznaků pro úlohu rozpoznávání řeči. Důraz je kladen na nalezení vhodné volby detekčního algoritmu v závislosti na podmínkách rozpoznávání, ale analýza také zkoumá možnosti nasazení VAD s ohledem na trénovací proces.

Pro teoretické posouzení vlivu vlastností detektoru i pro následnou analýzu na základě výsledků rozpoznávání řeči byly použity MFCC a PLP příznaky s následujícím nastavením: délka okna 25 ms, krok segmentace 10 ms, 12 kepstrálních koeficientů + energie, dynamické a akcelerační koeficienty. Stejné kepstrální koeficienty jsou použity jak pro algoritmus detekce řečové aktivity, tak pro vlastní rozpoznávač řeči. Jedná se o rozpoznávač sekvence číslovek nezávislý na mluvčím založený na HMM modelech fonémů se standardní cyklickou gramatikou. Rozpoznávač byl testován v různých šumových podmínkách, aby byl lépe popsán vliv použitého VAD na robustnost systému.

Detekční algoritmus svou úspěšností při klasifikaci signálu na řečové a neřečové segmenty významně ovlivňuje podobu signálu, se kterou dále rozpoznávač řeči pracuje. Základním kritériem kvality takového algoritmu je proto chybovost této klasifikace. Použitý VAD byl proto analyzován s ohledem na tuto chybovost.

Pro testování přesnosti detekce řeči byly použity dvě rozdílné sady dat na bázi české databáze SPEECON:

Ref_manual - 150 vět od různých mluvčích v rozdílných šumových podmínkách s ručně označenými hranicemi hlásek

Ref_automat - cca 300 promluv obsahujících izolované číslovky s výraznější pauzou mezi slovy. Referenční detekce řečové aktivity je generována automatickým zarovnáním hlásek pomocí rozpoznávače řeči na bázi HMM.

Testy byly provedeny pro obě parametrizace MFCC i PLP s použitím fixního i adaptivního nastavení prahu detekce (fixed, adapt). Pro posouzení přesnosti detekce byla použita kritéria *ERS* – chyba detekce řeči a *ERP* – chyba detekce pauzy.

$$ERS = \frac{1}{L_S} \sum_{i=0}^{L-1} (\mathrm{vad}_{ref}[i] - \mathrm{vad}[i]) \mathrm{vad}_{ref}[i] \tag{7.13}$$

$$ERP = \frac{1}{L_P} \sum_{i=0}^{L-1} (\mathrm{vad}[i] - \mathrm{vad}_{ref}[i])(1 - \mathrm{vad}_{ref}[i]) \tag{7.14}$$

kde L je celkový počet segmentů, L_S a L_P počet segmentů řeči a pauzy.

7.3.1 Přesnost detekce řečové aktivity

Analýza přesnosti detekce umožňuje popsat chování detektoru s ohledem na možnou chybovost detekce přítomnosti řeči v signálu. Výsledkem této chybovosti může být následná nesprávná klasifikace řeči při použití detektoru v procesu rozpoznávání řeči.

Srovnání prezentovaných detektorů v tabulce 7.1 ukazuje lepší výsledky detekce řečových segmentů (*ERS*) než detekce pauzy (*ERP*) pro obě volby parametrizace i prahování. To je vlastnost, která je při použití detektoru v úloze rozpoznávání řeči upřednostňovaná, neboť následně dochází k nižšímu negativnímu ovlivnění řečových segmentů vlivem následných úprav signálu, např. metodou vypouštění segmentů.

Zároveň jsou hodnoty srovnány oproti standardu ITU-T G.729 [47], který je využíván v hlasovém kodeku. Výsledky ukazují vyšší úroveň chyby detekce řeči oproti standardu. Na druhou stranu je snížena chyba detekce pauz, což přispívá k omezení neřečových segmentů v signálu a tím i omezení možného vkládání slov v důsledku přítomnosti vyšší hladiny šumu v procesu rozpoznávání řeči.

Volba pevného prahu ale přináší výrazně nižší chybovost, především na úrovni detekce pauz. To je způsobeno rozdílnými vlastnostmi adaptivních algoritmů oproti pevnému prahu. Při volbě pevného prahu jsou předem známy vlastnosti celého signálu a práh je tak nastaven optimálně pro danou nahrávku. Naopak adaptivní algoritmus je nutné inicializovat a nastavit na vhodnou citlivost na změny signálu. To jsou parametry, které mohou při nevhodné volbě hodnot chybovost VAD výrazně zvýšit.

Ref_manual										
	G.729		MFCC-fixed		PLP-fixed		MFCC-adapt		PLP-adapt	
	ERS	ERP	ERS	ERP	ERS	ERP	ERS	ERP	ERS	ERP
mean	0,50	22,70	5,89	10,23	5,19	14,89	6,39	21,68	8,17	23,73
std	0,70	8,60	5,33	16,87	6,44	18,92	17,44	16,36	21,29	16,66

Ref_automat										
	G.729		MFCC-fixed		PLP-fixed		MFCC-adapt		PLP-adapt	
	ERS	ERP	ERS	ERP	ERS	ERP	ERS	ERP	ERS	ERP
mean	2,30	28,70	14,84	7,56	12,93	12,05	10,16	18,31	11,50	19,45
std	2,60	11,70	9,50	15,92	8,99	18,49	16,33	16,06	18,81	17,56

Tabulka 7.1: Průměrné *ERS* a *ERP* pro jednotlivé kontrolní sady

7.3.2 Selektivní trénování akustických modelů s VAD

Při rozpoznávání řeči se detekce řeči mimo jiné využívá pro úlohu vypouštění segmentů. Dochází tak k ovlivnění podoby výsledné parametrizace a případný vysoký podíl chybně odstraněných řečových segmentů, byť jen na okraji slova, může způsobit výrazný pokles schopnosti systému správně rozpoznat danou promluvu. Následující analýza ukazuje vliv použití VAD na úlohu rozpoznávání řeči a zkoumá možnosti nastavení algoritmu zpracování signálu pro snížení chybovosti systému.

Použití VAD bylo analyzováno pro následující nastavení:

- detekce aplikovaná pouze na testovací signály
- detekce aplikovaná na trénovací i testovací data
- detekce aplikovaná pouze na trénovací signály

Tato volba umožnila sledovat vliv použití VAD na kvalitu výsledných modelů, aniž by byla ovlivněna testovací množina, a zjistit tak přínos metody pro zvýšení kvality výsledných řečových modelů.

Analýza byla provedena na kombinaci fragmentů CLEAN a NOISY databáze SPEECON z kanálu CS0 (náhlavní mikrofon) a CS1 (hands-free mikrofon). Celkem tak byly vytvořeny 4 rozdílné sady trénovacích a odpovídajících testovacích dat.

Pro určení fixního prahu byla zvolena experimentálně určená hodnota p=20 % ze vztahu (7.6). Pro adaptivní algoritmus na bázi dynamiky signálu pak byly nastaveny parametry p=20 % - míra pro volbu prahu detekce , q_1=0,95 - rychlost adaptace horní meze dynamiky signálu, q_2=0,98 - rychlost adaptace dolní meze dynamiky signálu. Algoritmus využívá celý vektor příznaků odpovídající danému segmentu signálu. Při detekci je tak využita i informace o energii daného výseku signálu.

Výsledky experimentu jsou srovnány na úrovni *WER* a chybovost pro jednotlivá nastavení je porovnána k hodnotě dosažené bez použití detekce řečové aktivity.

VAD na testovacích promluvách

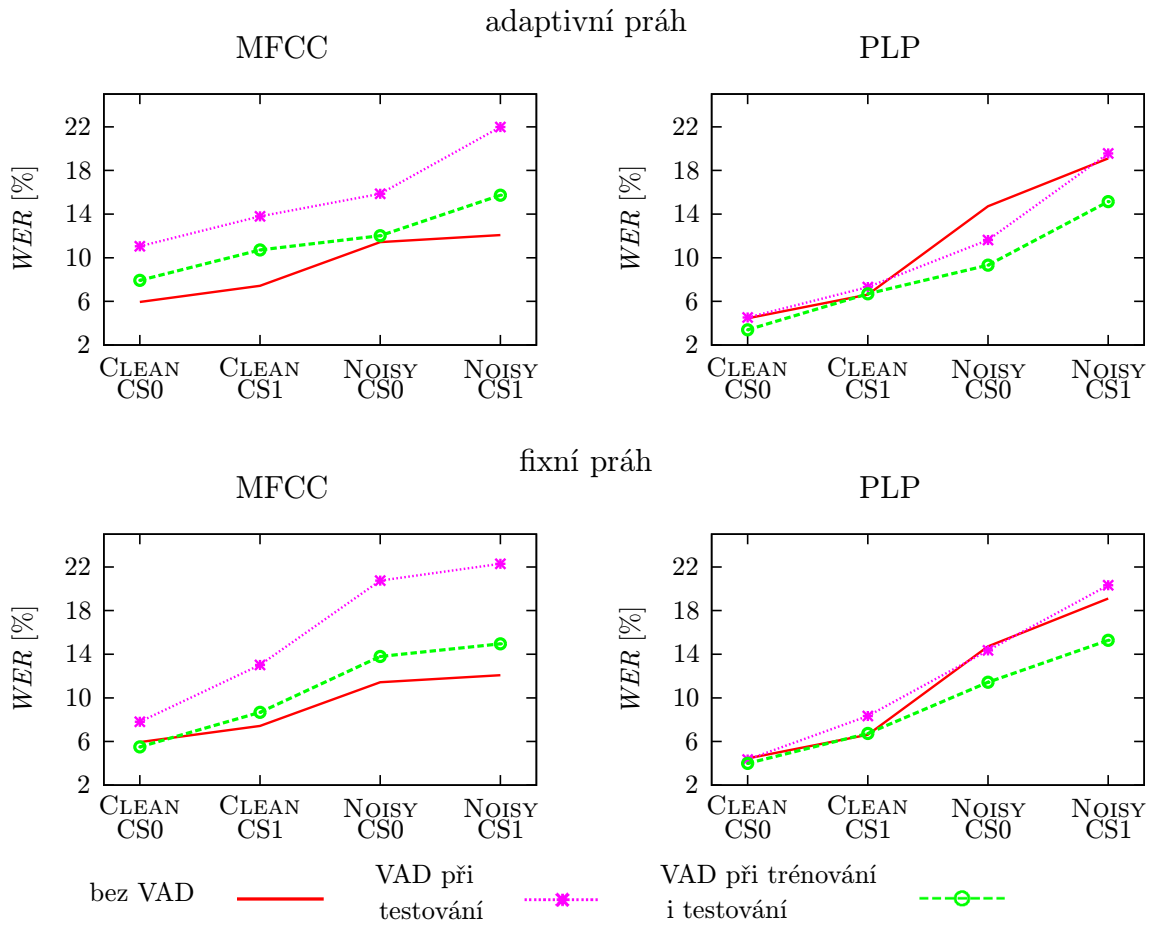
V prvním kroku analýzy přínosu VAD v procesu rozpoznávání řeči je VAD algoritmus aplikován pouze na data z testovací množiny. V případě ideální detekce by nemělo dojít ke snížení počtu správně detekovaných slov a naopak vlivem snížení počtu nesprávně vložených slov (inzercí) by výsledná chybovost měla klesnout. Grafy na obr. 7.6 zobrazují výsledky rozpoznávání pro obě parametrizace MFCC (vlevo) a PLP (vpravo) a pro jednotlivá nastavení volby prahu detekce - fixní (nahoře) nebo adaptivní (dole). Plnou čarou jsou zachyceny výsledky pro základní nastavení rozpoznávače bez použití VAD. Tečkovaná čára pak ukazuje výsledek pro množinu testovacích dat s VAD detekcí.

Na rozdíl od ideálního stavu došlo především pro parametrizaci MFCC k významnému zhoršení výsledků detekce. To souvisí s chybovostí detekčního algoritmu, který může způsobit odstranění i těch částí signálu, které jsou pro rozpoznávání řeči důležité. Nemusí se přitom jednat o odstranění hlasitých segmentů řeči. Jelikož jsou modely trénovány na celém signálu, který často obsahuje i tiché konce slov, a testovány na signálech, které mohou mít tyto části odstraněny, je pravděpodobně zdrojem chyby právě tato absence částí promluvy.

VAD při trénování i testování

V dalším kroku je proto použito VAD algoritmu na trénovacích datech a na této množině je systém přetrénován. Tím dojde k přizpůsobení modelů na takto upravenou parametrizaci.

Pro čistá řečová data dosahuje standardní systém chybovost cca 6 % resp. 4 % pro jednotlivé parametrizace MFCC a PLP. V případě velmi nepříznivých rušivých podmínek je to cca 12 resp. 19 %. Použitím detekce řečové aktivity se chybovost pro MFCC zvýšila, ale pro parametrizaci PLP došlo v případě velmi rušivých podmínek na snížení chybovosti k blízkosti hodnot pro MFCC. Algoritmus tak významně pomáhá potlačovat vliv rušení pro případ parametrizace PLP, která v experimentech vykazuje vyšší citlivost na šumové pozadí nahrávek.



Obrázek 7.6: *WER* pro jednotlivé parametrizace MFCC a PLP a rozdílný algoritmus prahování za různých šumových podmínek

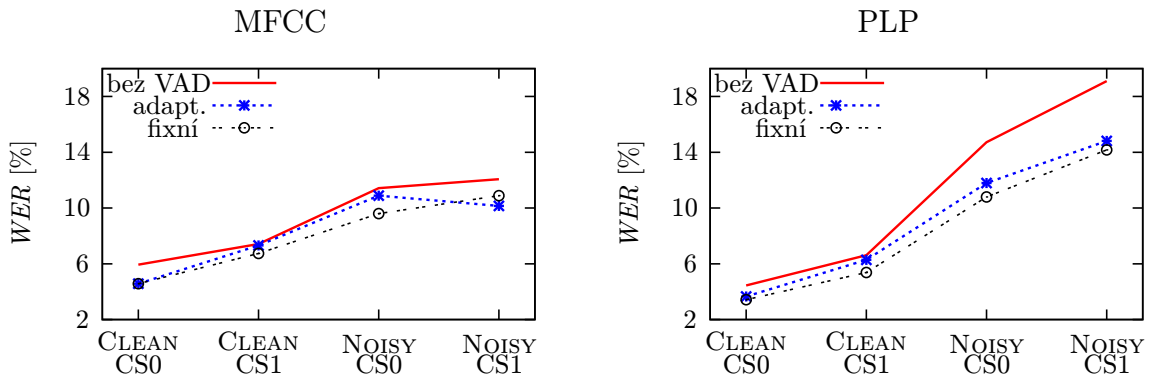
VAD - selektivní trénování

Vzhledem k vysoké chybovosti rozpoznávače způsobené chybou detekčního algoritmu, který může odstraňovat i tu část signálu, která nese řečovou informaci, ukazuje se jako vhodné použití modelů přetrénovaných na datech, na která byl aplikován algoritmus frame-dropping. Testovací data v tomto případě nebyla detektorem řeči zpracována. Tento postup umožňuje trénovat modely na datech, která obsahují mnohem méně neřečové informace a tím i případných artefaktů znehodnocujících trénovací data. Umožní to tak natrénovat modely přesněji.

Ačkoliv mohou být data deformována ztrátou některých řečových segmentů, algoritmus trénování v případě vyššího poškození umožní nahrávku přeskočit. Chybovost vlastního detekčního procesu tak není tolik významná, jako pro testovací data. Naopak chybovost detekce pauz způsobuje, že je i v trénovacích datech přítomna část signálu s neřečovým obsahem. Tyto segmenty tak mohou být využity pro trénování modelu pauzy.

Výsledky pro případ použití VAD jen u trénování ASR systému ukazuje obr. 7.7, opět pro obě parametrizace. Je zřejmé, že tento postup přináší ve všech případech snížení chyby rozpoznávání.

Rozdíl v úspěšnosti rozpoznávání pro jednotlivé případy nastavení volby prahování detektoru lze přisoudit především faktu, že pro některé nahrávky nemusí být volba adaptivního prahování dostatečně robustní, zvláště v případě krátkých nahrávek, kdy adaptační algoritmus nemá dostatečný počet vzorků pro správnou inicializaci koeficientů. Při nedostatku neřečových dat a nízké rychlosti adaptace prahu detekce pak není algoritmus schopen přizpůsobit adaptační práh daným podmínkám. Naopak při vyšší rychlosti adaptace může dojít k tomu, že algoritmus reaguje i na krátkodobé výchylky, které pak schopnost správné detekce řečové aktivity negativně ovlivní.



Obrázek 7.7: *WER* při různých volbách nastavení prahování

Shrnutí

Detekce řečové aktivity byla aplikována ve třech navazujících schématech. Nejdříve pouze na testovací úrovni, ve které přinesla zvýšenou chybovost především vlivem odstranění důležitých částí signálu. Poté byla detekce aplikována také na trénovací množině pro zvýšení kvality řečových modelů. To již přineslo pro případ parametrizace PLP, která je citlivější na šumové podmínky, výrazného zlepšení výsledků rozpoznávání. Při testování výsledných modelů na datech bez použití VAD pak dosáhl rozpoznávač nejlepších výsledků i pro čistší šumové podmínky. Hlavní závěry shrnují následující body:

- I přes vyšší hodnoty chybovosti VAD algoritmu může vhodné použití VAD vést na zvýšení robustnosti systému.
- Zvýšením kvality modelů při aplikaci detekce na trénovací data bylo dosaženo zlepšení výsledků rozpoznávače až o 20 %.
- Ke zvýšení robustnosti dochází i na datech z čistých reálných podmínek.
- Lze očekávat, že optimalizací nastavení algoritmu VAD by mohlo být dosaženo dalšího snížení chybovosti ASR systému.

7.3.3 Spektrální odečítání s detekcí řeči

V předchozí sekci byl navržen postup zvýšení robustnosti ASR systému použitím algoritmu vypouštění segmentů na základě VAD detekce. Algoritmus vychází z vypočtených kepstrálních charakteristik signálů a není proto výrazným zatížením procesu parametrizace signálu. Navíc, vzhledem k tomu, že je detekce aplikována pouze na trénovací množinu dat, lze provést výpočet off-line a při vlastním rozpoznávání již není detekce provedena.

Navrhované modelování řeči s využitím detektoru řeči je dalším krokem ke zvýšení robustnosti ASR systému z kapitoly 6. Použitá robustní parametrizace je doplněna o detekci řečové aktivity v trénovací fázi. Výsledkem je použití modelů, které jsou trénovány na datech s nižším zastoupením rušivých elementů v pauzách řeči.

Detektor řečové aktivity použitý v předchozích sekcích využívá informaci o energii signálu. Na bázi energie jsou postaveny základní algoritmy pro detekci řeči, které v tichých podmínkách mohou dosahovat dostatečně vysoké úspěšnosti. Na druhou stranu může v silně zarušených podmínkách docházet k tomu, že energie šumového signálu dosahuje úrovní energie řeči. A i když jsou charakteristiky šumu a řeči rozdílné, VAD založený na analýze energie v těchto případech zpravidla selhává. Pro experimenty jsou proto použity obě varianty zdrojových vektorů pro detekci řečové aktivity – s využitím koeficientu energie i bez jeho využití.

Analýza zahrnující VAD je opět provedena na jednotlivých podmnožinách databáze SPEECON a na databázi AURORA3 pro jednotlivé standardní i modifikované parametrizační techniky. V rámci nich je aplikován algoritmus spektrálního odečítání ESS. Mezi použité metody již nebyla zahrnuta metoda BFCC, která dosahovala výrazně vyšší chybovosti, než ostatní metody.

SPEECON – VAD s energií

Použití rozšířeného spektrálního odečítání a detekce řeči při trénování řečových modelů shrnuje tabulka 7.2. Oproti výsledkům bez použití metody ESS (tab. 6.3) příp. s využitím ESS (tab. 6.4) je zřejmý výrazný přínos navrhované metody. Vzhledem k výrazně vyšší chybovosti systému při použití VAD i pro testovací data není tato varianta uvedena.

Na čistých datech (Office, Clean) je dosaženo chybovosti i nižší než 2 %, v případě velmi zarušených podmínek je i pro kanál s vyšším zastoupením zaznamenaného rušení (Noisy) dosaženo chybovosti až 6,8 % pro parametrizaci RPLP. Použitím detekce řečové aktivity tak bylo dosaženo ve většině případů snížení chybovosti o více než 50 % oproti výchozímu nastavení bez ESS a VAD.

SPEECON – VAD bez energie

Na rozdíl od předchozích výsledků poskytuje detekční algoritmus bez využití informace o energii signálu srovnatelnou úspěšnost rozpoznávání i při použití VAD algo-

náhlavní mikrofon (CS0)				
	MFCC	PLP	MFLP	RPLP
Office	2,00	1,50	3,20	2,27
Clean	2,85	1,70	4,11	2,51
Noisy	8,46	8,55	8,99	7,30
Průměr	4,44	3,92	5,43	4,03
All	5,26	4,80	6,86	4,80

hands-free mikrofon (CS1)				
	MFCC	PLP	MFLP	RPLP
Office	4,94	4,40	6,41	4,54
Clean	5,37	5,30	7,42	5,30
Noisy	8,37	10,70	8,99	6,80
Průměr	6,22	6,80	7,61	5,55
All	6,54	7,91	8,55	6,10

Tabulka 7.2: Chybovost *WER* na databázi SPEECON s použitím algoritmů ESS a VAD pro trénovací data, a pouze ESS pro testovací data

ritmu na testovacích datech. Tabulka 7.3 shrnuje chybovost *WER* při použití VAD jak při trénování, tak při rozpoznávání.

Srovnáním s tabulkou 7.2 lze zjistit, že chybovost systému je v tomto případě vyšší. Jak bylo zmíněno dříve, v čistých podmínkách může algoritmus využívající informace o energii signálu využívat významné odlišnosti energie řeči a šumu a tím lze vysvětlit i lepší výsledky v předchozí sekci. Naopak v šumových podmínkách již může působit chybovost VAD algoritmu a tím poškozovat signál nebo zanechávat neřečové události v signálu pro rozpoznávání.

Tabulka 7.4 ukazuje tytéž výsledky pro situaci, kdy nebyl VAD použit pro rozpoznávání. V této situaci jsou výsledky rozpoznávání pro čisté signály stále méně přesné, než pro případ VAD s informací o energii, neboť zde také platí, že informace o energii může být významným přínosem pro detekci řeči v čistém prostředí. Pro šumové podmínky ale již dosahuje algoritmus v některých případech nižší chyby, než předchozí experimenty.

AURORA3 – VAD s energií

Pro srovnání byl test proveden také na standardizovaném rozpoznávači pro evaluace na databázi AURORA3, opět pro vybrané 3 jazyky. Stejně jako v předchozí části jsou srovnány oba algoritmy VAD s ohledem na využití informace o energii. Tabulka 7.5 zobrazuje výsledky pro navržený algoritmus využívající při detekci řeči informaci o energii.

náhlavní mikrofon (CS0)				
	MFCC	PLP	MFLP	RPLP
Office	3,74	4,14	4,81	3,87
Clean	4,22	4,79	5,59	4,00
Noisy	9,53	11,22	12,73	11,49
Průměr	6,88	8,01	9,16	7,7
All	7,27	7,27	9,83	7,54

hands-free mikrofon (CS1)				
	MFCC	PLP	MFLP	RPLP
Office	5,08	6,01	5,74	5,34
Clean	5,83	6,16	6,51	5,48
Noisy	10,77	12,06	10,95	10,33
Průměr	8,30	9,11	8,73	7,91
All	8,64	9,14	9,00	7,91

Tabulka 7.3: Chybovost *WER* na databázi SPEECON s použitím algoritmů ESS a VAD pro trénovací i testovací data

náhlavní mikrofon (CS0)				
	MFCC	PLP	MFLP	RPLP
Office	3,47	2,00	4,41	3,60
Clean	3,54	2,85	4,91	3,65
Noisy	8,64	10,15	10,77	8,73
Průměr	6,09	6,50	7,84	6,19
All	6,22	6,31	9,10	6,12

hands-free mikrofon (CS1)				
	MFCC	PLP	MFLP	RPLP
Office	4,81	5,07	6,54	5,07
Clean	4,57	4,79	6,96	5,37
Noisy	7,66	10,95	9,88	7,03
Průměr	6,12	7,87	8,42	6,20
All	6,99	7,72	8,18	6,44

Tabulka 7.4: Chybovost *WER* na databázi SPEECON s použitím algoritmů ESS a VAD pro trénovací data, a pouze ESS pro testovací data

Z hodnot je zřejmé, že v mnoha případech došlo ke zlepšení chybovosti oproti výchozím výsledkům z kap. 6. Na druhou stranu toto zlepšení není tak výrazné,

Španělština						
	baseline	MFCC	PLP	MFLP	RPLP	ETSI
WM	13,15	14,32	10,17	11,52	11,76	6,58
MM	26,26	26,11	21,15	23,90	20,07	13,27
HM	57,77	47,67	53,50	56,27	57,05	15,79
Průměr	32,39	29,37	28,27	30,56	29,63	11,88

Finština						
	baseline	MFCC	PLP	MFLP	RPLP	ETSI
WM	9,61	5,86	4,42	6,13	6,13	2,52
MM	27,63	27,56	44,80	44,25	22,30	12,72
HM	68,94	51,06	58,98	53,96	58,45	18,83
Průměr	35,39	28,16	36,07	34,78	28,96	11,36

Dánština						
	baseline	MFCC	PLP	MFLP	RPLP	ETSI
WM	22,20	15,67	15,72	17,73	15,50	15,87
MM	53,60	47,60	42,94	47,18	45,76	38,98
HM	68,10	61,36	66,93	68,69	67,36	37,81
Průměr	47,97	41,54	41,86	44,53	42,87	30,89

	baseline	MFCC	PLP	MFLP	RPLP	ETSI
Celkový průměr	38,58	33,02	35,4	36,62	33,82	18,04

Tabulka 7.5: Chybovost *WER* na databázi AURORA3 s použitím algoritmů ESS a VAD (s informací o energii)

jako v experimentu v předchozí sekci. Jedním z důvodů může být rozdílný přístup k modelování, kdy v úloze AURORA jsou použity modely celých slov, zatímco v experimentech předchozích jsou modelovány jednotlivé fonémy. Modely celých slov jsou méně ovlivněny změnami, které přináší použití VAD, především na okrajích slov. Vliv použití detekce v této úloze je proto nižší.

Srovnání se standardem ETSI ukazuje, že dvoustupňový algoritmus ETSI standardu výrazně lépe napomáhá systému ke snížení vlivu nepříznivých šumových podmínek.

AURORA3 - VAD bez energie

Opačná situace nastává při použití VAD algoritmu, který v rámci detekce nepoužívá informaci o energii. Tabulka 7.6 ukazuje, že ve většině případů dosahují konečné výsledky nižší chybovosti, než ETSI standard.

Zlepšení lze pozorovat především na WM případu, kdy např. pro Finštinu a PLP došlo ke zlepšení až o 54 %. Poslední řádek tabulky ukazuje celkový průměr pro všechny podmínky a všechny jazyky, kde je shrnutý přínos metody oproti výchozímu stavu. S výjimkou parametrizace PLP je průměrná chybovost nižší, než chybovost dosažená s parametrizací dle standardu ETSI.

Španělština						
	baseline	MFCC	PLP	MFLP	RPLP	ETSI
WM	13,15	4,77	5,10	6,76	4,45	6,58
MM	26,26	11,35	13,08	14,51	11,25	13,27
HM	57,77	18,34	17,49	20,83	18,52	15,79
Průměr	32,39	11,49	11,89	14,03	11,41	11,88

Finština						
	baseline	MFCC	PLP	MFLP	RPLP	ETSI
WM	9,61	4,60	3,42	3,46	3,51	2,52
MM	27,63	21,00	21,75	18,74	19,08	12,72
HM	68,94	23,82	31,55	18,27	29,82	18,83
Průměr	35,39	16,47	18,91	13,49	17,47	11,36

Dánština						
	baseline	MFCC	PLP	MFLP	RPLP	ETSI
WM	22,20	9,83	10,26	10,96	9,81	15,87
MM	53,60	29,59	30,82	30,40	28,84	38,98
HM	68,10	33,74	31,93	30,50	30,55	37,81
Průměr	47,97	24,39	24,34	23,95	23,07	30,89

	baseline	MFCC	PLP	MFLP	RPLP	ETSI
Celkový průměr	38,58	17,45	18,38	17,16	17,32	18,04

Tabulka 7.6: Chybovost *WER* na databázi AURORA3 s použitím algoritmů ESS a VAD (bez informace o energii)

Shrnutí

V úloze rozpoznávání řeči založené na modelech fonémů se ukazuje přístup s trénováním na datech s použitým VAD jako varianta přispívající významně k robustnosti systému. Především v případě použití rychlejších, ale méně přesných VAD algoritmů může být vzhledem k jejich možné vyšší chybovosti vhodné použít VAD pouze v trénovací fázi, kdy dochází k přesnějšímu natrénování modelů, zatímco vlastní rozpoznávaná řeč již není algoritmem deformován.

Pro úlohu založenou na modelech celých slov se pak vliv vlastního algoritmu pro detekci řečové aktivity projevuje velmi výrazně, což může být dáno vyšší variabilitou modelování slov z jednotlivých modelů fonémů.

Použití navrženého 1-krokového algoritmu s následným vyhlazením výsledků detekce překonalo i standardizovaný postup ETSI.

Zároveň využití detekce řeči nevyžadovalo speciální nároky na trénovací množinu dat s ohledem na zastoupení cílového prostředí. To na jednu stranu znamená snadnou použitelnost v procesu rozpoznávání řeči, na druhou stranu ale obecnost této metody z pohledu cílových šumových podmínek neumožňuje dosahovat vysoké efektivity. Okolní rušení v tomto případě sice není obsaženo v pauzách řeči, ale stále se vyskytuje v řečových segmentech. Další fází pro zvýšení robustnosti systému vůči šumovému prostředí může být přizpůsobení akustických modelů šumovým podmínkám popsané v následující kapitole.

Kapitola 8

Přizpůsobení akustických modelů na rušivé prostředí

Tato kapitola se zabývá možnostmi optimalizace podoby řečových modelů v konkrétním často hlučném prostředí. Při tom je použito trénovacích, resp. adaptačních dat, které charakterizují cílové podmínky, k přizpůsobení modelů řeči danému rušivému prostředí. Pro přizpůsobení modelů cílovým podmínkám jsou analyzovány dva základní přístupy. V prvním případě jsou modely čisté řeči přetrénovány na signálech z prostředí podobného cílovým podmínkám, v druhém případě jsou modely adaptovány na cílové podmínky metodou MLLR.

V rámci obou metod je zkoumána schopnost přizpůsobení akustických modelů na dané prostředí s ohledem na variabilitu šumového prostředí, množství adaptačního materiálu i s ohledem na volbu výchozích modelů. Jednotlivé přístupy jsou hodnoceny nejen s ohledem na jejich obecné vlastnosti při přizpůsobování modelů, ale je také posouzen vliv optimalizačních nastavení rozpoznávače řeči analyzovaných v předchozích kapitolách.

8.1 Metody přizpůsobení modelů řeči na šumové prostředí

Model čisté řeči popisuje charakteristiky řečového segmentu v čistých podmínkách bez rušivého pozadí. Metoda přizpůsobení modelů na šumové prostředí vychází z myšlenky, že lze analogicky nalézt takové parametry modelů, které popisují daný řečový element v daném šumovém prostředí. Máme-li k dispozici řečové nahrávky z cílového prostředí, můžeme je v procesu přizpůsobení použít, aby se podle nich modely čisté řeči transformovaly a vznikl tak nový model, vystihující zašuměnou řeč.

Výběr techniky pro přizpůsobení akustických modelů na šumové podmínky výrazně ovlivňují následující kritéria:

- *Dostupnost dat z cílového prostředí* – máme-li k dispozici velkou množinu dat z daného šumového prostředí (např. automobil), je možné použít metody, které

vyžadují více dat, ale mohou mnohem efektivněji modelovat vliv šumového prostředí na řečový signál. Naopak při nedostatku signálů je nutné použít metody, které generalizují a umožňují hledat obecnější formu přizpůsobení modelů.

- *Variabilita rušivého prostředí* – je-li prostředí málo proměnné, lze mnohem lépe postihnout jeho charakteristiky a jeho vliv na řečový signál je možné snáze zachytit v modelu a náročnost na použité signály není vysoká. Naopak velmi variabilní prostředí vyžaduje komplexnější řešení s dostatečně reprezentativní množinou dat pro přizpůsobení modelů a případnou schopností průběžné aktualizace modelů na dané podmínky.
- *Nezávislost na mluvčím* – podobně jako pro přizpůsobení modelů šumovému pozadí, lze při použití nahrávek pouze od jednoho mluvčího přizpůsobit systém také konkrétnímu mluvčímu. Aby k tomuto nedošlo v případě, že je potřeba zachovat nezávislost rozpoznávače na mluvčím, vyžaduje proces signály od dostatečně velkého množství mluvčích, což opět znamená vysoké nároky na použitá řečová data.

V této práci jsou analyzovány jednotlivé metody především z pohledu možného použití v obecném rušivém prostředí (databáze SPEECON), případně v prostředí automobilu (databáze CZKCC). Adaptovaný rozpoznávač je proto přizpůsoben danému typu prostředí (automobil, kancelář, apod.), ale nikoliv jeho konkrétní podobě (automobil dané značky, konkrétní místnost v konkrétní budově). Vzhledem k dostupnosti dat z různých typů prostředí je tedy možné akceptovat vyšší variabilitu takto definovaných rušivých podmínek. Zároveň je vzhledem k počtu mluvčích zastoupených v jednotlivých databázích splněn předpoklad nezávislosti systému na mluvčím.

8.1.1 Přetrénování na cílových podmínkách

V případě, že je cílové šumové prostředí známé a málo variabilní, lze použít nejjednodušší formu přizpůsobení modelů – standardní proces přetrénování modelů na datech nahraných v cílovém prostředí pomocí Baum-Welchova algoritmu.

Základem této metody je právě dostupnost poměrně rozsáhlé řečové databáze z různých prostředí. Tak jako modely řeči nezávislé na řečníkovi (SI – speaker independent) jsou trénovány na množině dat od velkého množství mluvčích, aby se dostatečně popsala variabilita řeči mezi jednotlivými řečníky, vyžaduje tato metoda velký objem dat, aby došlo k dostatečně robustnímu natrénování modelů.

Takto získané modely mají charakter přizpůsobených modelů pro dané podmínky a v případě dostatečně vysokého objemu adaptačních dat vedou na systém, který velmi kvalitně pracuje v konkrétních podmínkách. Na druhou stranu při jakékoliv změně těchto podmínek je potřeba zopakovat cyklus přetrénování na jiných datech. Obecnost šumových podmínek v signálech použitých pro přetrénování tedy určuje, do jaké míry je přípustná variabilita prostředí, ve kterém je rozpoznávač založený na takto přetrénovaných modelech řeči použit. Proto s vyšší variabilitou těchto podmínek metoda přechází do následující metody.

8.1.2 Trénování na obecných šumových podmínkách

Metoda přetrénování na obecném prostředí rozšiřuje předchozí metodu pro situaci, kdy není splněn předpoklad téměř neměnného cílového prostředí a zároveň je stále k dispozici dostatečný objem dat nahraných v daných podmínkách. Podobně jako předchozí metoda, i tento přístup využívá standardního trénovacího procesu, ale vzniklé modely charakterizují řečový signál ve více šumových podmínkách. Získané modely tak mohou vykazovat jistou nezávislost na prostředí, ve kterém je rozpoznávač provozován. I v tomto případě je důležitá dostupnost poměrně rozsáhlé řečové databáze z pohledu počtu mluvčích, zastoupeného prostředí a také fonetické bohatosti trénovacího materiálu. Proto je tato metoda používána především v případě, kdy je známo cílové prostředí a je k dispozici dostatek řečových dat.

Lze předpokládat, že takto získané modely nebudou mít vlastnosti nejkvalitnějších modelů čisté řeči a pro nezašuměnou promluvu může rozpoznávač založený na těchto modelech vykazovat nižší úspěšnost rozpoznávání, než s modely trénovanými na čisté řeči. V procesu přetrénování modelů na šumové podmínky ale mohou být prvním krokem k vytvoření modelů adaptovaných na konkrétní prostředí.

8.1.3 Adaptace modelů metodou MLLR

Proces přetrénování modelů je výrazně závislý na trénovací množině a v případě malého počtu dat se výrazně snižuje kvalita a variabilita modelů, která zaručuje nezávislost modelů na mluvčím a prostředí a schopnost popisovat řečový signál dostatečně robustně. Technika MLLR (Maximum-Likelihood Linear Regression) odhaduje lineární transformaci parametrů HMM modelů řečových i neřečových segmentů z adaptačních dat na základě kritéria maximalizace věrohodnosti modelů. Při tomto procesu je již pro malý objem dat nalezena transformace, která výrazným způsobem pomáhá adaptovat modely na cílové prostředí.

Výhodou algoritmu je možnost shlukování transformací pro jednotlivé modely. V nejjednodušší podobě tak existuje jedna globální transformace pro všechny modely, která tak vystihuje vliv šumového prostředí na řečový signál jednotně. Pro případ, kdy je k dispozici více dat je shlukování komplexnější a modely jsou děleny do regresních tříd na základě různých kritérií. Tím lze například pro případ rozdělení do dvou tříd vystihnout odlišný vliv šumu na charakteristiky signálu v pauze řeči od vlivu na charakteristiky řečového signálu.

Adaptace modelů na základě MLLR algoritmu tak tvoří variantu k metodě přetrénování modelů, kterou je možné použít pro mnohem menší objem adaptačních dat. V případě, že je k dispozici dostatečné množství materiálu pro adaptaci, dosahuje metoda MLLR srovnatelných výsledků, jako samotné přetrénování [57] při off-line zpracování, s klesajícím objemem dat je ale adaptační proces výrazně efektivnější.

Adaptace může probíhat dávkově, jsou-li k dispozici adaptační data před vlastním použitím rozpoznávače, nebo průběžně, pokud dochází k adaptaci na základě právě získaných dat.

8.2 Výchozí modely pro adaptaci

Důležitou součástí analýzy adaptačních technik je volba výchozích modelů, z nichž se vychází pro následnou adaptaci. Ta může výrazně ovlivnit rychlost adaptace na dané podmínky, stejně jako kvalitu výsledných modelů. Tyto výchozí modely mohou být jak modely čisté řeči, tak i modely řeči, která jistou míru zašumění obsahuje. Jak je popsáno v 8.1.2, jednou z metod pro zvýšení robustnosti modelů řeči pro aplikaci rozpoznávače řeči v rušivém prostředí je použití trénovacích dat nahraných v daném prostředí nebo v prostředích, která jsou cílovým podmínkám blízká ([70], [59]).

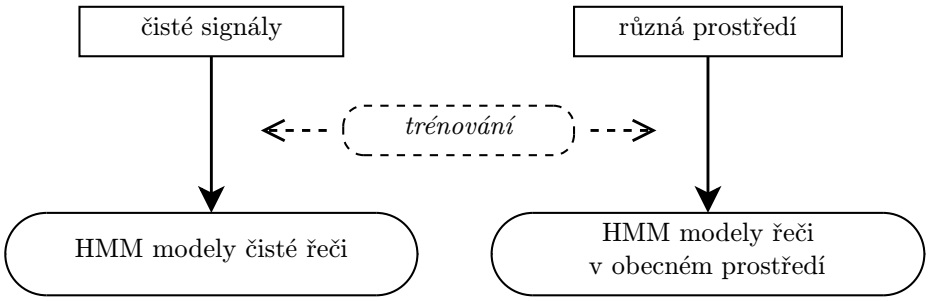
V [11] jsou například pro adaptaci modelů použita čistá řečová data smíchaná s nahraným šumem z automobilu. Výsledný signál tak neodpovídá zcela reálnému signálu, ale algoritmus umožňuje obejít potřebu získat velké množství řečových dat z daného prostředí a využít existujících databází čisté řeči. I přesto, že není použito reálných signálů, je ve výsledné práci s použitím adaptačních technik MLLR a MAP dosaženo snížení chybovosti systému z 14,38 % na 5,73 %. Podobně metodu mixování šumových a čistých řečových signálů využívá systém pro experimenty na úloze AURORA2 [69].

Jsou-li naopak k dispozici řečové signály z cílového prostředí, použití takových reálných dat umožňuje popsat vliv šumu na řečový signál nejen z pohledu aditivní složky, ale také konvoluční složku rušení (viz vztah (2.9)). Tak je dosaženo přesnějšího popisu dat a kvalitnějšího odhadu charakteristik zašuměného signálu.

Blokové schéma 8.1 zachycuje dvě základní sady modelů, které jsou v této sekci analyzovány s ohledem na jejich vhodnost pro použití jako výchozí modely pro následný proces přizpůsobení šumovým podmínkám. Jedná se o:

- modely čisté řeči,
- modely řeči v obecném šumovém prostředí.

První z přístupů umožňuje přetrénovat, případně adaptovat velmi kvalitní modely řeči a zajistit tak, že adaptační proces bude založen na schopnosti správně klasifikovat řečová data. Naopak druhý z přístupů využívá pro natrénování modelů již zašuměná data, která ale nemusí být nutně z cílového prostředí.



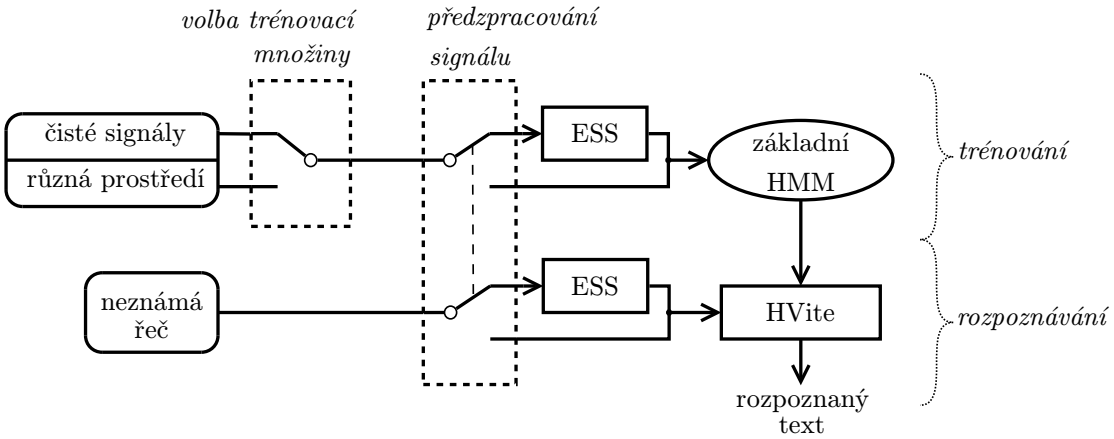
Obrázek 8.1: Výchozí řečové modely pro zvýšení efektivity přizpůsobení na šumové podmínky

Lze očekávat, že kvalita modelů trénovaných na datech ze zašuměného prostředí bude nižší, než při použití čistých trénovacích dat. To se může projevit především v případě rozpoznávání čistých promluv. Na druhou stranu mohou takto ovlivněné modely obsahovat jistou míru variability, která umožní mnohem snáze adaptovat modely na výsledné, často předem neznámé podmínky. Během adaptace tak může podpořit správnost klasifikace zašuměných adaptačních dat.

V rámci uvedené analýzy je také sledován vliv metody ESS na výslednou kvalitu modelů. Použitou metodou pro potlačení šumu lze v případě modelů trénovaných na obecných šumových podmínkách snížit vliv šumu na kvalitu modelu řeči ovlivněného přítomností šumu v trénovacích signálech.

Jak napovídají analýzy např. v [65], použití kombinace adaptivních algoritmů s metodami pro potlačování adaptivního rušení v signálu se ukazuje jako výhodná kombinace pro zvýšení robustnosti ASR systému v prostředích s vysokou úrovní rušení z okolí. Tyto techniky lze využít nejen pro blokové zpracování, ale také pro on-line adaptaci a jsou tak výhodným prvkem pro použití v prostředí, které má proměnné charakteristiky, případně pro prostředí s předem nespecifikovaným charakterem. Typickým příkladem je prostředí automobilu.

Schéma 8.2 zobrazuje celkový přehled výše popsaného trénovacího procesu s možným nastavením trénovacích dat a předzpracování signálu, jak je dále použito v experimentální části.



Obrázek 8.2: Blokové schéma rozpoznávacího rámce pro analýzu robustnosti ASR systému v šumovém prostředí

8.2.1 Experimentální část

Analýza proběhla na rozpoznávači sekvence číslovek nezávislém na mluvčím, který používá HMM modely fonémů a modely pro krátkou a dlouhou pauzu, dle popisu v sekci 4.1.3.

Pro trénování výchozích modelů byly použity fragmenty Office a All databáze SPEECON. V prvním případě se tak jedná o trénování na čistých signálech, druhý případ simuluje trénování na obecných podmínkách pro volbu výchozích modelů. V obou případech byl použit kanál CS0.

Testovací část fragmentů byla dále rozdělena. Vzhledem k úloze rozpoznávání číslovek byly nahrávky obsahující pouze číslovky použity jako testovací data, zbylá část sady byla použita pro účely přetrénování/adaptace v následující sekci. Z dostupných kanálů databáze SPEECON byly použity kanály CS0 a CS1.

Jsou-li řečové modely trénovány na čistých promluvách, přítomnost rušení v rozpoznávané řeči výrazně snižuje úspěšnost výsledného ASR systému. Jak ukazuje tabulka 8.1, použití algoritmu ESS může pomoci efekt rušivého pozadí významně potlačit. S použitím samotného ESS algoritmu tak bylo dosaženo až 26 % snížení chybovosti (Car; CS0). Na druhou stranu v případě, kdy je rušení již velmi výrazné (Car, Noisy; CS1), přínos metody je již nižší, pro tichá prostředí (Clean; CS0) dochází spíše k degradaci signálu a zvýšení chybovosti. Přesto v celkovém měřítku metoda přináší zvýšení robustnosti systému, především na kanále s více zarušenými signály.

Podobný přínos zaznamenává i postup, v němž je použito trénování na obecných podmínkách (tab. 8.2). Ačkoliv takto dochází k degradaci kvality modelů řeči a na čistých podmínkách je v tomto případě dosaženo nižších skóre, než v předchozím případě, průměrná chybovost je oproti použití čistých trénovacích dat výrazně nižší – 12,2 % oproti 8,6 %, a při použití ESS průměrná chybovost dále klesá, viz tab. 8.3.

Nejvýraznějšího zlepšení dosahuje prostředí automobilu, v němž díky výraznému aditivnímu charakteru rušení algoritmus ESS přináší zajímavá zlepšení, pro případ trénování na obecných podmínkách dokonce dosahuje výsledný systém nejlepších hodnot úspěšnosti.

náhlavní mikrofon (CS0)								
	All	Office	Car	Clean	Noisy	HiSNR	LoSNR	Průměr
bez ESS	8,32	3,47	7,03	4	13,09	7,03	11,88	7,83
ESS	8,46	4,17	5,2	4,45	11,31	6,83	11,76	7,45
Zlepšení [%]	-1,68	-20,17	26,03	-11,25	13,6	2,84	1,01	4,82

hands-free mikrofon (CS1)								
	All	Office	Car	Clean	Noisy	HiSNR	LoSNR	Průměr
bez ESS	14,72	8,14	35,17	11,07	18,79	10,2	17,65	16,53
ESS	13,53	8,01	29,97	9,82	16,56	9,41	16,31	14,80
Zlepšení [%]	8,08	1,6	14,79	11,29	11,87	7,75	7,59	10,48

Tabulka 8.1: Vliv ESS algoritmu na *WER* pro různá prostředí při trénování modelů na čistých šumových podmínkách (Office)

náhlavní mikrofon (CS0)								
	All	Office	Car	Clean	Noisy	HiSNR	LoSNR	Průměr
bez ESS	7,77	3,87	3,36	4,79	10,6	6,68	10,37	6,78
ESS	7,04	3,87	1,53	4,68	8,55	5,74	9,78	5,88
Zlepšení [%]	9,4	0	54,46	2,3	19,34	14,07	5,69	13,17

hands-free mikrofon (CS1)								
	All	Office	Car	Clean	Noisy	HiSNR	LoSNR	Průměr
bez ESS	10,65	9,61	8,26	11,19	10,95	10	12,81	10,5
ESS	10,65	7,21	8,87	9,7	10,77	9,55	12,58	9,9
Zlepšení [%]	0	24,97	-7,38	13,32	1,64	4,5	1,8	5,63

Tabulka 8.2: Vliv ESS algoritmu na *WER* pro různá prostředí při trénování modelů na obecných šumových podmínkách (All)

	Office	All
bez ESS	12,18	8,64
ESS	11,13	7,89
Zlepšení [%]	8,66	8,59

Tabulka 8.3: Průměrné hodnoty *WER* za oba kanály pro jednotlivé trénovací sady

8.2.2 Shrnutí

Myšlenka trénování modelů na datech z obecného šumového prostředí se ukazuje jako vhodný předstupeň pro snazší adaptaci systému na cílové podmínky. Ten umožňuje upravit parametry modelu o potřebnou variabilitu, díky níž probíhá proces adaptace snáze.

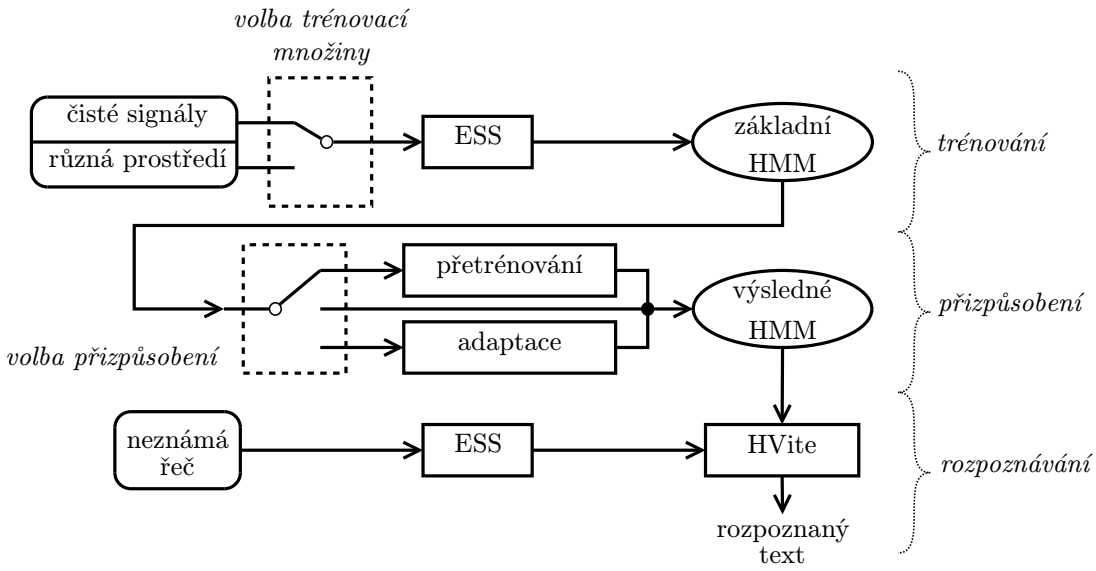
Trénování na obecných podmínkách přináší významné zlepšení výsledků rozpoznávače v reálných podmínkách i pro případ, kdy nejsou použity další metody pro potlačení vlivu rušení. V prezentovaných výsledcích bylo takto dosaženo o 22 % nižší chybovosti, než při použití modelů čisté řeči a algoritmu ESS.

Kombinace trénování na obecných podmínkách a ESS pak vedla na snížení chyby *WER* o 29 % oproti modelům čisté řeči a systému bez ESS. Pro data z automobilu bylo dosaženo zlepšení až o 70 %.

8.3 Srovnání technik pro přizpůsobení modelů

Následující analýza srovnává základní techniky pro přizpůsobení modelů řeči cílovým šumovým podmínkám. S ohledem na dosažený přínos metody ESS v předchozí sekci je v následující analýze vždy použita tato metoda při předzpracování signálu.

Schéma 8.3 zachycuje základní nastavení procesu trénování HMM modelů, do kterého je zahrnuto buď přetrénování nebo adaptace na bázi MLLR. Analýza byla opět provedena na několika fragmentech databáze SPEECON, aby bylo možné lépe postihnout vliv metody na různé úrovně zašumění signálu.



Obrázek 8.3: Blokové schéma rozpoznávacího rámce pro analýzu robustnosti ASR systému v šumovém prostředí s přetrénováním/adaptací

Pro účely přizpůsobení modelů byla z testovací části každého fragmentu vytvořena adaptační sada, viz sekce 8.2.1, která byla následně použita pro přetrénování výchozích modelů. Tím je dosaženo přizpůsobení modelů na podmínky obsažené v této sadě. Vzhledem k tomu, že se jedná o shodné podmínky, jako jsou obsaženy v testovacích sadách, dochází tak k výraznému přizpůsobení cílovým podmínkám. Pro přetrénování, které bylo provedeno vždy ve třech trénovacích cyklech, bylo použito vždy celé dostupné adaptační množiny, která čítá pro sadu Car cca 2400 promluv, pro sadu All téměř 11600 promluv.

Dílčí výsledky zahrnující adaptaci variancí nepřinesly významné zlepšení úspěšnosti rozpoznávání v úloze adaptace na šumové podmínky. Pro metodu MLLR byl proto zvolen algoritmus adaptace pouze středních hodnot parametrů, zatímco ostatní parametry zůstávají zachovány.

Omezená sada dat pro přizpůsobení modelů

Jak je zmíněno v sekci 8.1.3, algoritmus MLLR umožňuje použít pro adaptaci HMM modelů již relativně nízký objem adaptačních dat. Proto byly vytvořeny sady s 20, 50, 100, 200, 500 a 1000 promluvami pro každý fragment, které umožňují srovnat schopnost adaptace pro různé objemy dat. Tyto sady byly navrženy tak, aby obsa-

hovaly co nejvíce různých mluvčích a docházelo k minimálnímu ovlivnění nezávislosti výsledného rozpoznávače na mluvčím.

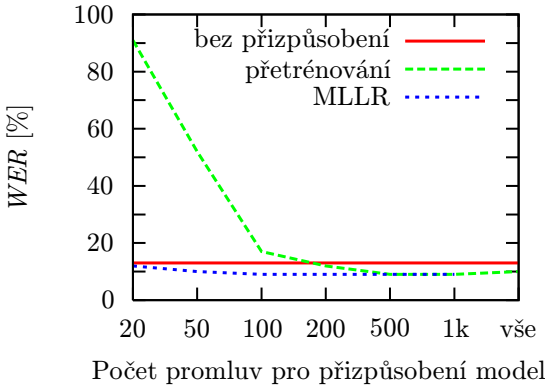
Ve výsledku obsahuje každá adaptační množina minimálně 18 mluvčích. S přihlédnutím např. k [11], kde je při použití 10–80 mluvčích dosaženo snížení chybovosti systému nezávislého na mluvčím, lze toto dělení považovat za dostatečné. Tabulka 8.4 ukazuje průměrnou délku adaptačních dat v jednotlivých fragmentech databáze.

počet promluv	20	50	100	200	500	1000	celý fragment
čas	65s	2,8min	5,7min	10,9min	26,9min	57,2min	7,8h

Tabulka 8.4: Průměrné množství řečového materiálu v adaptačních množinách přes všechny fragmenty databáze SPEECON

Graf na obr. 8.4 ukazuje vývoj chybovosti ASR systému s použitím různé velikosti adaptační sady dat pro přetrénování příp. MLLR adaptaci. Průběh vývoje chyby *WER* ukazuje, že použití malých adaptačních databází degraduje schopnost přizpůsobení modelů mechanizmem přetrénování, zatímco MLLR adaptace dosahuje lepších výsledků i pro malý počet adaptačních dat.

V další analýze bude pro srovnání obou technik použito maximálního počtu adaptačních dat, aby se ve výsledcích výrazně neprojevily vlastnosti algoritmu přetrénování na malém počtu nahrávek.



Obrázek 8.4: *WER* pro rozdílné množství dat pro přizpůsobení modelů

Klastrování pro MLLR

V rámci analýzy byl použit mechanizmus klastrování modelů do binárního regresního stromu [123]. Díky tomu je možné lépe vystihnout rozdílný vliv šumu na podobu výsledných modelů, neboť dojde k rozdělení skupiny modelů a jejich směsí na několik skupin, které sdílí transformační matici.

Vedle možnosti, kdy všechny modely sdílí právě jednu transformaci (globální transformace), bylo použito také nastavení 2, 4, 8, 16, 32 a 64 regresních tříd, do nichž byly modely a jejich směsi rozděleny dle akustické podobnosti analyzované na trénovací části databáze promluv.

8.3.1 Experimentální část

Výše popsané techniky přizpůsobení modelů šumovým podmínkám pomocí přetrénování a MLLR jsou v následující sekci použity na úloze rozpoznávání číslovek v různých typech prostředí a porovnány s ohledem na přínos využití dalších technik pro snížení vlivu rušivého prostředí. Není-li uvedeno jinak, je použita parametrizace MFCC.

Potlačení šumu metodou ESS

V tabulce 8.5 jsou shrnuty výsledky rozpoznávání číslovek pro jednotlivé fragmenty databáze SPEECON. V této fázi nebyly použity metody pro přizpůsobení modelů šumovým podmínkám, ale pouze potlačení šumu metodou ESS. Tabulka srovnává případ použití výchozích modelů (IM) trénovaných na čistých signálech (Čisté IM) nebo na obecných šumových podmínkách (Obecné IM).

Hodnoty odpovídají údajům v tabulkách 8.1 a 8.2 pro řádek s označením ESS. Je zde zřejmý přínos metody využívající trénování modelů na obecném šumovém prostředí, výrazný především v silně zarušených podmínkách. S využitím odšumování na bázi ESS bylo dosaženo pro fragment Car snížení chybovosti systému o více než 70 % pro oba pozorované nahrávací kanály. Možný pokles kvality řečových modelů diskutovaný výše se zde neprojevuje tak významně, jako přínos přítomnosti šumového pozadí ke schopnosti lépe klasifikovat řeč v předem nedefinovaných podmínkách. Ke zvýšení chybovosti oproti čistým výchozím modelům tak dochází pouze výjimečně v případě nízké úrovně zašumění testovacích promluv.

výchozí podmínky	náhlavní mikrofon (CS0)							
	All	Office	Car	Clean	Noisy	HiSNR	LoSNR	Průměr
Čisté	8,46	4,17	5,2	4,45	11,31	6,83	11,76	7,45
Obecné	7,04	3,87	1,53	4,68	8,55	5,74	9,78	5,88
Zlepšení [%]	16,78	7,19	70,58	-5,17	24,4	15,96	16,84	21,06

výchozí podmínky	hands-free mikrofon (CS1)							
	All	Office	Car	Clean	Noisy	HiSNR	LoSNR	Průměr
Čisté	13,53	8,01	29,97	9,82	16,56	9,41	16,31	14,8
Obecné	10,65	7,21	8,87	9,7	10,77	9,55	12,58	9,9
Zlepšení [%]	21,29	9,99	70,4	1,22	34,96	-1,49	22,87	33,09

Tabulka 8.5: Srovnání *WER* pro čisté a obecné výchozí modely – bez přizpůsobení

Vícenásobné přetrénování akustických modelů

Tabulka 8.6 shrnuje výsledky rozpoznávání číslovek po 3-násobném přetrénování akustických modelů na adaptačních datech. Pro inicializaci byly opět použity jak výchozí modely trénované na čisté řeči, tak výchozí modely trénované na obecných šumových podmínkách. Po přetrénování tak dochází k přizpůsobení modelů na konkrétní podmínky použité v adaptační sadě dat. Ve všech případech dochází ke zlepšení výsledků, celkově pak dosahuje pokles chybovosti 22,5 %.

Přínos použití obecných výchozích modelů lze sledovat na případu fragmentu ALLv kanálu CS0. Zatímco při použití čistých výchozích modelů a jejich přetrénování na obecných podmínkách vede na chybu 7,4 % (tab. 8.6), při použití obecných modelů je dosaženo nižší chyby 7,04 % i bez přetrénování (tab. 8.5). Pokud jsou pro přetrénování použity obecné výchozí modely, úspěšnost dále roste.

výchozí podmínky	náhlavní mikrofon (CS0)							
	ALL	OFFICE	CAR	CLEAN	NOISY	HISNR	LOSNR	Průměr
Čisté	7,4	3,74	3,36	5,02	8,9	6,24	11,24	6,56
Obecné	6,67	3,47	2,75	4,91	7,48	5,54	9,84	5,81
Zlepšení [%]	9,86	7,22	18,15	2,19	15,96	11,22	12,46	11,42

výchozí podmínky	hands-free mikrofon (CS1)							
	ALL	OFFICE	CAR	CLEAN	NOISY	HISNR	LOSNR	Průměr
Čisté	10,05	6,94	14,37	8,11	12,82	8,22	12,46	10,42
Obecné	8,32	6,01	5,2	7,42	7,57	7,92	9,03	7,35
Zlepšení [%]	17,21	13,4	63,81	8,51	40,95	3,65	27,53	29,46

Tabulka 8.6: Srovnání *WER* pro čisté a obecné výchozí modely – s přetrénováním

Adaptace modelů metodou MLLR

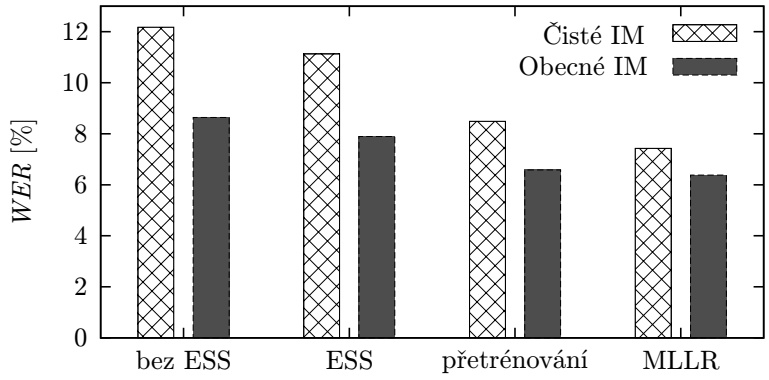
Výsledky pro přizpůsobení modelů řeči metodou MLLR shrnuje tab. 8.7. Pro možnost přehlednějšího porovnání nastavení počtu regresních tříd s předchozími výsledky je v tabulce použita vždy minimální hodnota z hodnot dosažených pro různé počty regresních tříd. Jednotlivé hodnoty *WER* jsou dále rozepsány v příloze v tabulce B.1.

Při srovnání s výsledky z tab. 8.6 je zřejmé, že pro kanál CS0 obsahující nižší míru zašumění jsou průměrné výsledky velmi podobné. Pro více rušivý kanál CS1 je již dosaženo znatelného zlepšení. Zároveň i pro případ adaptace se ukazuje výhodnost využití obecných výchozích modelů pro inicializaci adaptačního procesu. To ukazuje i shrnující graf na obrázku 8.5 znázorňující přínos použití jednotlivých analyzovaných přístupů z pohledu použití rozdílných výchozích modelů. Ve všech pozorovaných případech využití algoritmů pro zvýšení robustnosti bylo dosaženo lepších výsledků pro výchozí modely trénované na obecných šumových podmínkách.

výchozí podmínky	náhlavní mikrofon (CS0)							
	ALL	OFFICE	CAR	CLEAN	NOISY	HISNR	LOSNR	Průměr
Čisté	7,82	3,81	2,86	4,45	10,18	6,84	10,66	6,66
Obecné	7,18	3,54	2,4	4,53	8,46	5,99	9,53	5,95
Zlepšení [%]	8,29	7,05	16,11	-1,68	16,88	12,42	10,56	10,7

výchozí podmínky	hands-free mikrofon (CS1)							
	ALL	OFFICE	CAR	CLEAN	NOISY	HISNR	LOSNR	Průměr
Čisté	8,78	5,7	5,05	7,1	10,75	7,24	10,52	7,87
Obecné	7,44	5,36	3,01	6,68	7,7	7,2	8,83	6,6
Zlepšení [%]	15,21	5,88	40,37	5,89	28,31	0,58	16,03	16,14

Tabulka 8.7: Srovnání *WER* pro čisté a obecné výchozí modely – s adaptací



Obrázek 8.5: Průměrné *WER* v jednotlivých fázích procesu potlačení vlivu rušení

Srovnání parametrizačních technik v adaptačním schématu

Zatímco výše uvedená analýza byla provedena s použitím parametrizační techniky MFCC, tabulka 8.8 popisuje výsledky rozpoznávání číslovek na fragmentu ALL při použití různých parametrizačních technik. Pro toto srovnání bylo použito algoritmu ESS a zároveň byla při trénování modelů na signály aplikována detekce řečové aktivity.

Výsledná chybovost systému pro jednotlivá nastavení počtu regresních tříd v tabulce 8.8 ukazuje na významný přínos metody MLLR již při použití jedné regresní třídy. V porovnání jednotlivých parametrizačních technik pak dosahuje metoda RPLP nejvyššího skóre a v případě 32 regresních tříd minimální chybovosti systému ze všech srovnávaných nastavení.

trénovací fáze	počet tříd	MFCC	PLP	RPLP	MFLP
Monofóny	-	26,19	20,02	18,15	22,68
Trifóny	-	8,29	5,89	5,14	6,88
MLLR	1	5,9	3,98	4,16	5,24
MLLR	2	5,29	3,79	3,93	4,82
MLLR	4	5,05	3,84	3,27	4,3
MLLR	8	4,59	2,46	3,32	4,12
MLLR	16	4,4	3,41	2,95	4,12
MLLR	32	3,93	3,04	2,85	4,02

Tabulka 8.8: *WER* – MLLR pro různé parametrizace s ESS+VAD při trénování

8.3.2 Shrnutí

Přetrénování modelů představuje robustní off-line metodu pro zvýšení úspěšnosti rozpoznávače v šumovém prostředí pro případ dostatečného množství adaptačních dat. Největšího zlepšení dosáhl systém pro případ rozdílných nahrávacích kanálů trénovacích a testovacích dat. Při použití modelů z obecného prostředí bylo dosaženo dalších výrazných snížení chybovosti.

Výhody MLLR algoritmu oproti přetrénování spočívají především v možnosti aplikovat tento algoritmus s menší množinou dat nebo přímo v on-line zpracování signálu. Při použití shodné adaptační databáze tak systém dosahuje jen mírného zlepšení oproti systému s přetrénováním.

Tabulka 8.9 shrnuje dosaženou chybovost rozpoznávače v podobě průměrné hodnoty *WER* přes všechny použité fragmenty a nahrávací kanály. Po adaptaci akustických modelů metodou MLLR bylo s využitím výchozích modelů trénovaných na obecných šumových podmínkách dosaženo o 14 % nižší chybovosti v porovnání s použitím výchozích modelů z čistých šumových podmínek.

výchozí podmínky	bez přizpůsobení	přetrénování	MLLR
Čisté prostředí	11,13	8,49	7,27
Obecné prostředí	7,89	6,58	6,28
Zlepšení [%]	29,06	22,5	14,06

Tabulka 8.9: Srovnání průměrných hodnot *WER* pro čisté a obecné výchozí modely v různých fázích přizpůsobení modelů

8.4 Adaptační schéma s MLLR

Předchozí sekce 8.3 ukazuje přínos adaptace pro obecné podmínky. V následující analýze bude prezentováno schéma pro použití adaptačních technik v prostředí automobilu. Jedná se o velmi variabilní prostředí ovlivněné nejen jízdními podmínkami, ale i vlastním typem automobilu. Navržený systém používá extrakci příznaků na bázi standardních parametrizačních technik PLP a MFCC. S ohledem na dobré výsledky dosahované v prostředí automobilu je použit také algoritmus ESS.

Jak ukazuje i předchozí experiment, kombinace metod pro potlačování šumu s adaptivními algoritmy může být velmi výhodná v prostředí s rušivým pozadím. Adaptace tak může nejen zlepšit modely s ohledem na aditivní šum, ale také na zkreslení dané ESS.

Pro zachování vysoké kvality řečových modelů byly jako pro výchozí fázi experimentu použity výchozí modely natrénované na OFFICE části databáze SPEECON. Tyto modely je možné přetrénováním na části dat z databáze CZKCC přizpůsobit podmínkám prostředí stojícího automobilu.

Na rozdíl od předchozí analýzy jsou ale použity modely trifónů dle popisu v 4.1.3. Tím je dosaženo možnosti s výsledky dále navázat na rozpoznávání spojitých promluv. Následující rozpoznávání je proto provedeno s použitím dekodéru HDecode. Současně s použitím trifónů je použit unigramový statistický jazykový model, který obsahuje pouze číslovky 0–9, viz popis v sekci 4.1.3. Model je založen na informaci o výskytu číslovek v databázi SPEECON.

Výchozí modely tak byly získány 3-násobným přetrénováním výchozích modelů na fragmentu OFF trénovací části databáze CZKCC, viz obr. 8.6, fáze `I`, tedy signálů s čistým šumovým pozadím, ale s charakteristikami typickými pro automobilové prostředí. To zajistilo dostatečné přizpůsobení podmínkám v automobilu, zároveň ale také zachování nezávislosti řečových modelů na mluvčím.

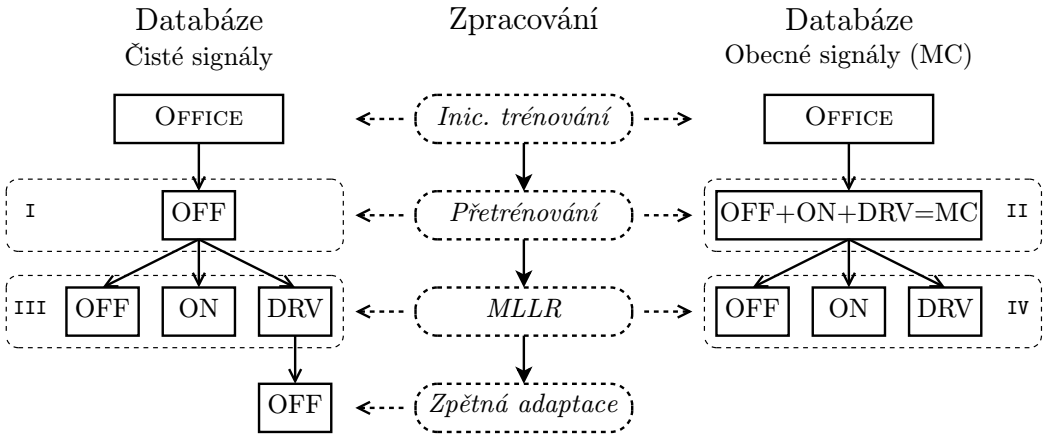
Vzhledem k výsledkům dosaženým pro obecné výchozí modely na obecném prostředí, bylo tohoto kroku použito i v následující analýze. Pro tento účel nebyly modely přizpůsobeny pouze na fragmentu OFF, ale pro přizpůsobení byly použity všechny tři fragmenty OFF, ON a DRV trénovací části databáze, viz obr. 8.6, fáze `II`.

Adaptace v dalším kroku byla provedena na adaptační části databáze CZKCC (fáze `III` a `IV`). Pro experiment byla použita celá adaptační část databáze, ale s ohledem na vlastnosti MLLR lze použít i menší část.

Důležitou vlastností adaptačního procesu, především v online módu, je vedle snížení vlivu šumových podmínek také schopnost dostatečně efektivně sledovat změny tohoto prostředí. Proto je bloková adaptace dále doplněna o procesy, které analyzují schopnost adaptace v daných podmínkách.

Zpětná adaptace

Proces zpětné adaptace zobrazený ve schématu 8.6 testuje úspěšnost rozpoznávače po adaptaci modelů přizpůsobených velmi zarušenému šumovému prostředí zpět na čisté



Obrázek 8.6: Blokové schéma vývoje akustických modelů

podmínky. Analýza tak ukazuje schopnost algoritmu přizpůsobit se výrazně rozdílným podmínkám a možnost použít tento postup pro on-line adaptaci.

Inkrementální adaptace

V rámci inkrementální adaptace je provedena adaptace modelů na základě výsledků z předchozího kroku rozpoznávání. Na rozdíl od blokové adaptace je tedy při inkrementální adaptaci použito výsledku předchozího rozpoznávání, nikoliv předem známých promluv a zároveň jsou k adaptaci použity pouze malé objemy dat. Vzhledem k tomu, že není implementován žádný algoritmus pro rozhodování o vhodnosti dané promluvy pro adaptaci, lze očekávat méně přesné výsledky.

Pro inkrementální adaptaci bylo použito stejné úlohy, ale dekodér HDecode neumožňuje inkrementální adaptaci. Proto byl použit dekodér HVite s cyklickou gramatikou. V tomto případě jsou rozpoznávány delší nahrávky, což umožňuje simulovat on-line průběh adaptace.

8.4.1 Experimentální část

Pro získání kvalitních výchozích modelů byla použita databáze SPEECON, fragment OFFICE. Takto získané modely nebyly použity pro rozpoznávání, neboť vlivem výrazně rozdílných podmínek nahrávání oproti testovacím datům se dá očekávat nízká úspěšnost výsledného systému.

Bloková adaptace

Výchozí modely byly přetrénovány na trénovací fragmentu OFF databáze CZKCC, aby došlo k přizpůsobení modelů na daný nahrávací kanál a prostředí automobilu (tab. 8.10, řádek I). Výsledný systém dosahuje až 1,3 % *WER* na OFF fragmentu testovací

části CZKCC s náhlavní soupravou, což je hodnota srovnatelná s hodnotami dosahovanými ve standardních systémech s nízkou úrovní šumu. Pro rozpoznávání v rušivých podmínkách (DRV) je ale pouze toto přetrénování nedostatečné a chybovost dosahuje hodnot 27,5 % a 12,6 % pro MLLR resp. PLP.

	engine OFF				engine ON				engine DRV			
	MFCC	PLP	MFCC +ESS	PLP +ESS	MFCC	PLP	MFCC +ESS	PLP +ESS	MFCC	PLP	MFCC +ESS	PLP +ESS
Far-talk												
I	2,56	2,3	2,14	2,24	2,28	2,15	2,68	1,9	16,9	17,08	14,58	15,39
II	2,69	2,14	2,24	2,59	2,9	3,24	1,71	2,34	12,46	8,91	9,96	7,73
III	3,17	4,03	2,3	2,34	1,75	1,56	2,34	2,06	7,36	6,25	6,48	6,82
IV	2,72	2,05	2,05	2,46	2,68	3,21	1,93	2,46	10,43	7,42	8,17	6,7
Close-talk												
I	2,09	1,3	1,39	0,96	1,43	2,04	0,95	1,09	27,57	12,6	11,65	4,91
II	1,57	1,3	1,22	0,96	1,15	3,67	1,02	0,68	4,75	2,8	5,62	2,45
III	2,09	1,57	1,39	0,87	1,29	1,97	0,95	0,88	2,77	6,22	2,14	2,26
IV	1,57	1,22	1,31	0,96	1,09	3,26	0,95	0,61	2,92	1,87	1,93	1,65

Tabulka 8.10: *WER* při blokové adaptaci

Dalšího zlepšení je dosaženo trénováním na obecných podmínkách z automobilu (fáze II), kdy je chybovost snížena na hodnotu 2,8 %. Tento přístup je ovšem závislý na použitých datech a vyžaduje velké objemy řečového materiálu.

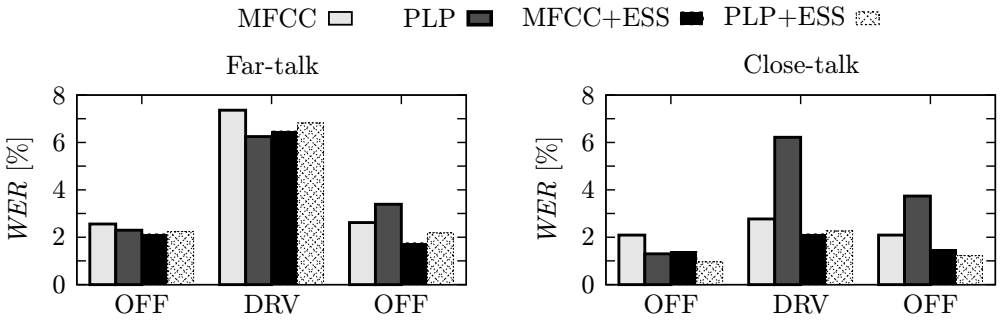
Na základě předchozích experimentů bylo pro prostředí automobilu použito také metody pro potlačení šumu ESS. Tento přístup vedl na výrazné snížení chybovosti rozpoznávání pro všechny analyzované podmínky, dokonce i v relativně tichých podmínkách stojícího automobilu (vlivem klimatizace, rušení z venkovního prostoru). Pro tiché podmínky je chybovost snížena na 0,96 %, při jízdě dosahuje *WER* až 2,45 %.

Po aplikaci blokové adaptace (fáze III a IV) vzrostla přesnost rozpoznávání i přes použití menších objemů adaptačních dat a výsledky překonávají hodnoty úspěšnosti z předchozích fází. Především pro jedoucí automobil a tedy nejvíce zarušené prostředí, je snížena chybovost až na hodnotu 1,65 % s náhlavní soupravou. Na druhou stranu již pro tiché prostředí (OFF) nedochází k dalšímu poklesu chybovosti.

Zpětná adaptace

Obr. 8.7 zobrazuje vývoj chybovosti systému při aplikaci zpětné adaptace. Zde se projevuje vliv vlastností modelů z fáze III, které jsou adaptovány na šumové podmínky jedoucího automobilu a jsou použity jako výchozí pro adaptaci na data z fragmentu OFF. Navzdory této skutečnosti dosahují výsledky před adaptací a po zpětné adaptaci podobných hodnot. V jediné případě při použití PLP parametrizace dosahuje systém po zpětné adaptaci vyšší chybovosti. S použitím ESS je vliv šumu na PLP snížen a výsledky jsou srovnatelné s MFCC. Tento jev lze přičítat vyšší citlivosti me-

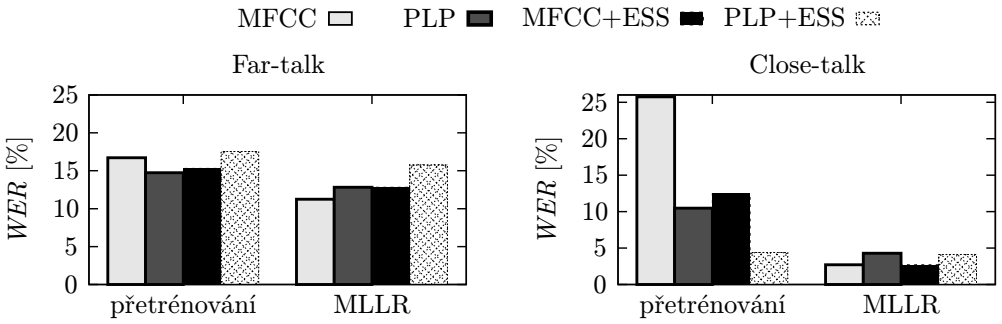
tody PLP na šum, která se projeví při adaptaci modelů přizpůsobených zašuměným podmínkám na podmínky nehlučného prostředí.



Obrázek 8.7: *WER* pro zpětnou adaptaci

Inkrementální adaptace

Chybovost dosažená při použití inkrementální adaptace dosahuje nižších hodnot, než při použití procesu přetrénování. To je patrné především pro náhlavní mikrofon, kde je adaptačním procesem bez dodatečných algoritmů pro rozhodování o úspěšnosti rozpoznávání dosaženo snížení průměrné chybovosti na hodnoty mezi 2,7 % a 4,3 %.



Obrázek 8.8: *WER* pro inkrementální adaptaci

Použití ESS algoritmu při inkrementální adaptaci v případě významně se měnících podmínek vedlo k mírně nižší úspěšnosti výsledného systému (obr. 8.8) oproti zpracování signálu bez ESS. To lze přičíst efektu maskování šumem [73], který se uplatňuje v případě metod nepoužívajících ESS. V tomto případě dochází ke snižování úrovně šumu a odkrývání artefaktů, které mohou mít degradující vliv na rozpoznávání.

Analýza průběžných výsledků systému s adaptací bez znalosti přepisu promluv ukázala PLP metodu jako více citlivou na šum, než systém založený na MFCC. Stejně tak průběh vývoje *WER* ukazuje metodu PLP jako citlivější na změnu mluvčího. Tyto vlastnosti korespondují s výsledky získanými v předchozích experimentech a ukázaly významný vliv odšumovacích technik na metody založené na LP analýze, díky kterým

je snížen vliv šumu na získané parametry signálu a tím i lepší schopnost klasifikace řeči.

8.4.2 Shrnutí

V rámci kapitoly byla analyzována úspěšnost rozpoznávání s použitím robustních parametrizací a MLLR adaptace modelů. Navrhovaná kombinace robustních metod přináší významné snížení chybovosti i ve velmi zarušených podmínkách reálného prostředí jedoucího automobilu. Závěry shrnují následující body:

- Použití algoritmu ESS v kombinaci se standardními metodami bylo zkoumáno v reálném prostředí automobilu. Algoritmus nevyužívá VAD, který může být významným zdrojem chyb, a dosahuje výrazných zlepšení oproti standardním metodám bez potlačení rušení a to až o 81 % pro případ parametrizace PLP.
- Adaptace na prostředí se ukazuje být vhodným řešením pro významné zlepšení úspěšnosti ASR systému i v případě jednodušší dělení transformačních funkcí na dvě skupiny - řeč, neřeč. Chybovost 2,14 % ve velmi hlučném prostředí jedoucího automobilu ukazuje na výhodu použití uvedené metody v těchto podmínkách. Při použití výchozích modelů trénovaných na obecných podmínkách bylo dosaženo hodnoty *WER* 1,65 %.
- Výsledky použití ESS a MLLR korespondují s výsledky z jiných analýz ([65]). Oproti uvedené analýze je v tomto případě výhodou použití algoritmu bez VAD.
- V souvislosti s on-line adaptací zpětná adaptace ukazuje, že přizpůsobení modelů na velmi odlišné podmínky je efektivní a schéma je vhodné pro použití pro průběžnou adaptaci.
- Inkrementální adaptace ukázala citlivost PLP na šum jako efekt ovlivňující průběh adaptace nezanedbatelnou měrou. Při použití ESS je možné ale tento efekt částečně potlačit.

Kapitola 9

Modelování neřečových událostí

Kapitoly 6 a 8 popisují možnosti zvyšování robustnosti ASR systémů vůči neřečovým událostem na úrovni potlačování šumu v řečovém signálu a nalezení vhodných příznaků pro popis řeči, případně vhodným modelováním hlučného pozadí na úrovni akustických modelů řeči. Tyto přístupy jsou efektivní především v případě vlivu stacionárních šumů s neměnnými nebo pomalu se měnícími vlastnostmi. Pokud se v promluvě vyskytne izolovaná neřečová událost nebo rychle se měnící rušivý signál na pozadí, schopnost rozpoznávače tyto jevy eliminovat je velmi nízká. V kapitole 7 je popsáno využití detekce řečové aktivity, která v kombinaci s algoritmy pro vyhlazování výsledků detekce umožňuje potlačit přítomnost izolovaných neřečových událostí. Přínos VAD lze sledovat především pro případ velmi krátkých nebo slabých signálů.

Odolnost systémů vůči rušivým neřečovým událostem je proto běžně zajišťována také metodami, které v rámci procesu dekódování využívají speciální modely pro dané typy událostí. Z tohoto pohledu lze metody rozdělit do dvou skupin.

V první skupině je použití modelů šumu a jejich kombinace s modely řeči tak, aby tvořily model zašuměné řeči. Tento přístup využívají metody tzv. paralelní kombinace modelů (např. [31], [53], [30]) a metody kompozice a dekompozice (např. [120], [29]). Tyto metody nachází uplatnění především v situaci, kdy jsou k dispozici dostatečné zdroje pro vytvoření modelu šumového prostředí. S ohledem na zaměření této práce na obecné šumové podmínky zde nejsou tyto metody dále analyzovány.

Druhou skupinu tvoří postupy, které využívají modely neřečových událostí řečníka. Tyto události se vyskytují mezi slovy promluvy a během klasifikace řeči mohou být nesprávně interpretovány jako hláska či skupina hlásek [108], a tak přispívat k chybnému výsledku celého rozpoznávání. Pokud jsou ale modely těchto událostí zahrnuty do procesu klasifikace, lze výskyt nesprávně rozpoznané události potlačit. Výsledky experimentů v [101] a [25] ukazují, že již jednoduché modely těchto událostí mohou významně napomoci ke snížení chybovosti systému.

Tato kapitola se zabývá analýzou vhodných podmínek a efektivního postupu trénování modelů neřečových událostí řečníka především v situaci, kdy není k dispozici dostatečné množství trénovacích dat. V další části kapitoly jsou analyzovány možnosti zahrnutí těchto modelů do procesu rozpoznávání řeči.

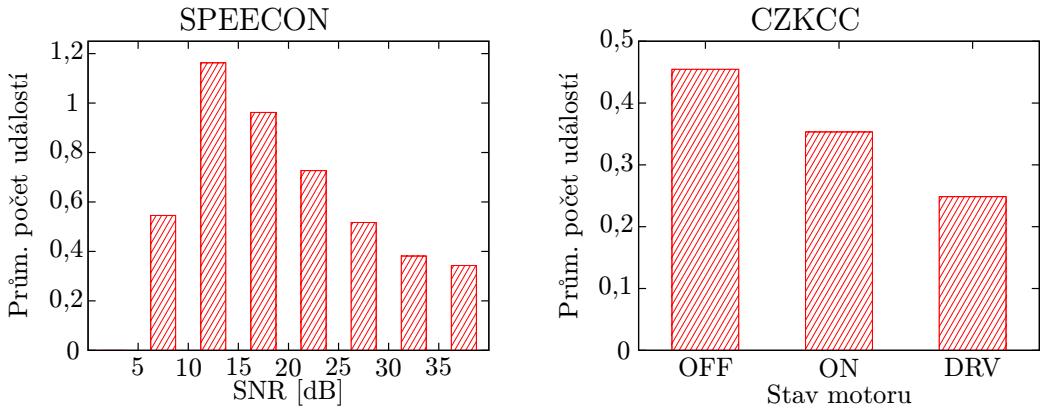
9.1 Výskyt neřečových událostí v řečových korpusech

Pro trénování modelů neřečových událostí řečníka je nutné mít k dispozici databázi promluv, v rámci jejichž přepisu je výskyt těchto neřečových událostí zaznamenán. To zvyšuje nároky na proces tvorby databáze a ne vždy jsou proto neřečové události anotovány, např. při automatickém přepisu obsahu promluv, případně nejsou zaznamenány v dostatečném detailu.

Stejně jako kvalitu modelů segmentů řeči tak i kvalitu modelů neřečových událostí významně ovlivňuje výběr trénovacích dat. Při výběru vhodné databáze je proto nutné brát v úvahu aspekty, které výskyt neřečových událostí ovlivňují. Těmi jsou především:

- povaha řečových dat – čtené promluvy / spojitá řeč,
- připravenost řečníka – předem připravený projev / spontánní promluvy,
- podmínky získávání dat – tiché prostředí / rušné okolí,
- požadovaný detail a schopnost rozlišit tento detail při pořizování přepisu.

Analýzou vlivů, které mohou ovlivnit výskyt neřečových událostí se detailněji zabývá např. článek [97]. Zde se ukazuje jako jeden z významných faktorů právě míra zarušení prostředí nahrávání řečových dat. Výsledkem je nárůst počtu neřečových událostí pro nahrávky s mírou SNR 10–25 dB, viz obr. 9.1. V tišším prostředí není mluvčí rozptylován okolním hlukem a má tendenci mluvit tišeji, klidněji, což se projevuje také na kvalitě projevu. Naopak v hlučném prostředí je velký počet těchto událostí maskován okolním rušením. To se odráží také v analýze počtu značek neřečových událostí v databázi promluv nahrávaných v automobilu, kde s vyšší mírou hlučnosti prostředí dochází k poklesu počtu zaznamenaných neřečových událostí.



Obrázek 9.1: Neřečové události v použitých databázích pro různé šumové podmínky

9.2 Klasifikace událostí v trénovací databázi

Pro účely této práce byly neřečové události rozděleny do tří základních skupin: vyplněná pauza (`FIL` – filled pause), hlasitý nádech/výdech (`BRE` – breath) a jiná událost řečníka (`SPK` – speaker non-speech event). Toto členění vychází z označení neřečových událostí v použité databázi SPEECON na skupinu `FIL` a `SPK` (viz sekce 5.5.2) a z požadavku na odlišení události `BRE` od ostatních událostí, které vykazují výrazně odlišný charakter (viz sekce 9.2.2).

Tabulka 9.1 shrnuje počet neřečových událostí v jednotlivých databázích. Je zde patrné, že databáze SPEECON obsahuje neřečových událostí přibližně dvojnásobné množství oproti databázi CZKCC. Proto může být vhodným zdrojem pro získání modelů neřečových událostí. Naopak databáze CZKCC obsahuje členění událostí na více typů, což může s ohledem na variabilitu těchto neřečových jevů být výhodou pro zvyšování počtu modelů neřečových událostí.

Databáze	Fragment	Počet promluv	Počet značek `SPK`	Počet značek `FIL`
SPEECON	Clean	63024	33138 (52,6 %)	1856 (2,94 %)
	All	180213	108474 (60,2 %)	7382 (4,10 %)
CZKCC	OFF	38391	11532 (30,0 %)	153 (3,99 %)
	celá DB	221318	46742 (21,1 %)	691 (3,12 %)

Tabulka 9.1: Počet neřečových událostí řečníka v použitých databázích (procenta jsou vztažena k počtu promluv v databázi)

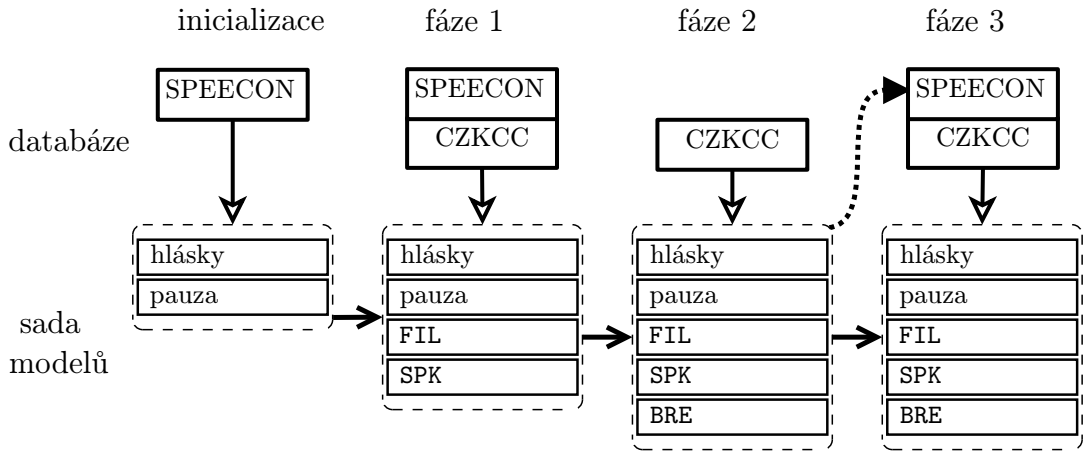
Následující sekce popisuje navržený postup využití databází s různou úrovní klasifikace pro získání požadované množiny modelů neřečových událostí v dostatečné kvalitě. Experimenty byly provedeny na nastavení rozpoznávače dle sekce 4.1.3 s použitím parametrizace MFCC.

Celý postup lze rozdělit na 3 fáze, které popisuje blokové schéma 9.2. Schéma zachycuje fragmenty databází použité v jednotlivých fázích a sadu modelů, která je v dané fázi trénována.

9.2.1 Inicializace modelů neřečových událostí

V první fázi jsou vytvořeny iniciální modely neřečových událostí řečníka – vyplněné pauzy (`FIL`) a jiné události (`SPK`) a přidány do sady HMM modelů fonémů pro rozpoznávač sekvence číslovek. Modely jsou získány jako kopie akusticky podobných fonémů "e" resp. "p". Výběr těchto tříd událostí je inspirován klasifikací událostí v použité trénovací databázi SPEECON.

Po přetrénování těchto modelů bylo dosaženo snížení chybovosti rozpoznávače číslovek z 4,99 % na 3,8 %, tedy o 20 % na databázi Clean.



Obrázek 9.2: Fáze trénování pro čtené řečové databáze

9.2.2 Rozšíření sady modelů

Vzhledem k vysoké variabilitě neřečových událostí ve skupině `SPK` byl vytvořen nový model, který charakterizuje události nádech/výdech (`BRE`). Tato událost je výrazně odlišná od ostatních událostí převážně plozivního charakteru. Pro získání iniciálních modelů je použita databáze CZKCC, která tyto události rozlišuje.

Pro inicializaci modelu události `BRE` bylo v souladu s experimenty v [25] použito hlásky "f". Podobně jako je v [108] analyzována délka neřečové události `FIL`, bylo následnou analýzou zjištěno, že inicializace tímto modelem nemusí být s ohledem na předpokládanou délku události vhodná. Při následném zarovnání databáze SPEECON bylo proto zkoumáno nastavení s více volbami výchozích modelů. Zatímco inicializací modelem hlásky "f" lze vystihnout akustickou podobu, výběr inicializace modelem dlouhé pausy byl motivován schopností modelovat události s delším průběhem. Vliv volby iniciálního modelu je blíže popsán v následující sekci.

9.2.3 Trénování na celém řečovém materiálu

Pro efektivní využití dostupného řečového materiálu je v dalším kroku nad promluvami databáze SPEECON provedeno zarovnání tak, aby s využitím modelů z druhé fáze byla původní událost `SPK` klasifikována jako plozivní (`SPK`) nebo frikativní (`BRE`). Takto upravený popis je dále použit k přetrénování celé sady modelů, čímž dochází ke zvýšení objemu trénovacího materiálu pro trénování modelů neřečových událostí.

Výsledkem celého procesu trénování je rozpoznávač číslovek, který oproti původnímu nastavení dosahuje v poslední fázi o 26,3 % nižší chybovosti na úrovni slov, viz tab. 9.2. Ačkoliv je výsledná chybovost mírně vyšší než v předchozí fázi, výhodou celého systému je možnost trénování na celé dostupné sadě dat, čímž je dosaženo snížení vlivu použité trénovací databáze na výsledné modely. Z pohledu chyby typu inzerce dosahuje rozpoznávač v poslední fázi nejnižší chybovosti, lze tedy předpokládat, že akustické modely vystihují dostatečně charakter modelovaných neřečových jevů.

	WER [%]		Inzerce [%]	
základní systém	4,99	–	0,95	–
fáze 1	3,44	(31,1 %)	0,22	(76,8 %)
fáze 2	3,56	(28,7 %)	0,22	(76,8 %)
fáze 3	3,68	(26,3 %)	0,11	(88,4 %)

Tabulka 9.2: Chybovost na úrovni slov (*WER*) a inzercí, jejich relativní zlepšení v jednotlivých fázích trénování modelů neřečových událostí oproti výchozímu systému

Inicializace modelu `BRE`

Pro porovnání obou přístupů zmíněných v sekci 9.2.2 bylo nad databázemi SPEECON a CZKCC provedeno zarovnání, které ohraničilo neřečové události. Tabulka 9.3 ukazuje průměrnou délku neřečové události po jednom a dvou cyklech přetrénování modelů neřečových událostí. Zatímco průměrná délka události `SPK` se po druhém přetrénování modelů změnila nanejvýš o 10 %, délka neřečové události `BRE` se především v případě inicializace modelem "f" mění výrazněji směrem k vyšším hodnotám. Lze tedy usuzovat, že volba modelu pauzy jako iniciálního modelu pro událost `BRE` může vést na rychlejší natrénování.

fáze, iniciální model	SPEECON		CZKCC	
	`SPK`	`BRE`	`SPK`	`BRE`
2, "f"	20,06	25,44	23,77	21,38
2, "pauza"	20,7	108,74	22,46	54,46
3, "f"	24,72	39,72	36,99	30,44
3, "pauza"	21,81	102,17	33,7	55,21

Tabulka 9.3: Průměrná délka trvání neřečových událostí

Podobně byl vliv iniciálního modelu porovnán s ohledem na stabilitu výsledku zarovnání. Situace, kdy se v různých fázích přetrénování výrazně změní i výsledek zarovnání, může značit nižší míru natrénovanosti modelu. V rámci analýzy byl porovnán výsledek zarovnání z druhé a třetí fáze. Pokud byla událost v druhé fázi označena jako `SPK` a ve třetí fázi jako `BRE` nebo naopak, byla situace označena jako Substituce. Pokud byla událost v právě jedné z fází smazána, je situace označena jako Výmaz.

Tabulka 9.4 ukazuje, že pro model iniciovaný modelem pauzy dochází k této záměně méně často. Zároveň je v rozpoznaných výsledcích mnohem méně událostí nerozpoznaných.

Závěrem lze tedy konstatovat, že pro inicializaci modelu události charakteru nádech/výdech může být výhodné použít model pauzy, který oproti modelu hlásky "f" vystihuje i charakteristickou délku události.

	pauza	hláska "f"
Substituce	4,44 %	10,44 %
Výmaz	3,86 %	4,14 %

Tabulka 9.4: Rozdíl v množství zaměněných a smazaných neřečových událostí pro dvě varianty iniciálních modelů události `BRE`

9.2.4 Využití spontánních promluv pro modelování neřečových událostí

V předchozích sekcích byly modely neřečových událostí trénovány na čtených promluvách. Jak ukazuje srovnání v tab. 5.5, výskyt události `FIL` je ve čtených promluvách výrazně nižší. Zároveň např. [118] ukazuje významnou roli této neřečové události v akustickém projevu řečníka. Schopnost tuto neřečovou událost rozpoznat a omezit tak chybovost rozpoznávače proto může být důležitou součástí robustního řešení rozpoznávače řeči. Přínos použití spontánních promluv tak může být právě ve zvýšení počtu těchto událostí v trénovacím materiálu. V této sekci je proto k množině trénovacích dat přidána část databáze spontánních promluv CzLecDSP. Tím je dosaženo zvýšení počtu událostí typu `FIL` v trénovacích datech.

Pro analýzu vlivu rozšíření trénovacího materiálu byla použita úloha rozpoznávání slov se slovníkem omezeným pouze na slova obsažená v promluvě a s možností zařazení události typu `FIL` mezi libovolná dvě slova promluvy. Tím bylo dosaženo možnosti hodnotit především schopnost rozpoznávače detekovat událost typu `FIL` v promluvě.

K trénování modelu neřečových událostí bylo použito nejprve pouze databáze čtených promluv Clean, poté byla trénovací množina doplněna o část databáze spontánních promluv. Pro testování je pak použita druhá část databáze spontánních promluv.

trénovací množina	vložené značky `FIL`		smazané značky `FIL`	
Clean	10196	(1222 %)	11	(1,32 %)
Clean+CzLecDSP	1927	(231 %)	249	(29,9 %)

Tabulka 9.5: Chybovost na úrovni mazání a vkládání vyplněných pauz pro čtené a kombinované trénovací sady (procenta jsou vztažena k množství anotovaných značek)

Tabulka 9.5 zobrazuje výsledek rozpoznání události `FIL` pro oba případy trénovacích množin. Ve výskytu chybně vložených značek (značka vložena do místa, kde v původní transkripci není) došlo k poklesu o 81 % a ačkoliv je výsledný počet neřečových událostí stále velký, pokles této chybovosti ukazuje na vyšší kvalitu modelů.

Počet smazaných neřečových událostí je sice zvýšen a nebylo rozpoznáno téměř 30 % původně označených událostí, ale z pohledu rozpoznávání se jedná o chybu, která může souviset s natrénováním modelu na více významné události, zatímco tyto nerozpoznané události jsou většinou modelovány obecným modelem ticha (viz předchozí sekce se zarovnáváním databáze).

Použití spontánních řečových dat tak významně pomáhá zvýšit kvalitu výsledných modelů pro použití v úloze rozpoznávání obecných promluv a bude nutnou součástí navazujících experimentů.

9.3 Hodnocení kvality neřečových událostí v řečových korpusech

Jak popisuje sekce 9.1, výskyt neřečových událostí řečníka v dostupných databázích se může vlivem mnoha různých činitelů výrazně odlišovat. Vedle vlivu charakteru promluv, který byl zmíněn již při popisu nově vytvářené databáze spontánních promluv v rámci této práce v sekci 5.5.2, se jedná především o vliv prostředí a anotátora. Při tvorbě databáze jsou prováděny kontroly, které omezují tento vliv, ale vzhledem k jejich náročnosti není takový postup efektivní. Výše vytvořené modely lze proto využít pro automatickou kontrolu databáze a ke sjednocení úrovně, pro kterou jsou neřečové události vyhodnoceny jako podstatné.

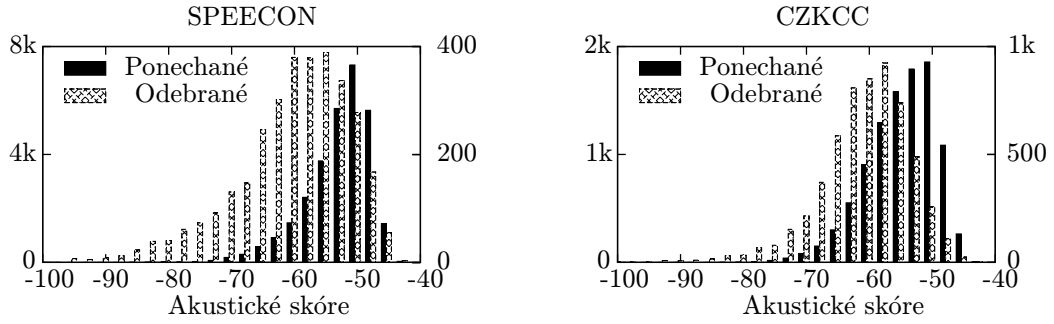
Proto byly všechny dostupné trénovací sady zarovnány standardním procesem zarovnání, který byl použit i v sekci 9.2.2. Vedle transkripce, která umožňuje předefinovat výslovnost jednotlivých slov a v konečném důsledku i zaměnit jedno slovo či neřečovou událost za jiné, případně událost úplně odstranit (náhradou za model ticha), poskytuje zarovnání také informaci o akustickém skóre. To udává statistickou věrohodnost, se kterou je událost modelována odpovídajícím modelem.

Tabulka 9.6 zobrazuje počet neřečových událostí před a po zarovnání. Procesem zarovnání tak bylo v každé množině odstraněno určité množství neřečových událostí. Grafy na obr. 9.3 zobrazují rozložení akustického skóre v jednotlivých databázích pro původní značky neřečových událostí řečníka, pro zarovnané značky a pro značky odstraněné na základě zarovnání. Z těchto histogramů lze odečíst, že odstranění se týká především značek, které získávají v rozpoznávací úloze nižší skóre a jsou tedy méně věrohodně modelovány. Tento rozdíl je zřejmý především na databázi CZKCC.

databáze	počet slov	ostatní neřečové události			
		anotované		zarovnané	
SPEECON	181517	33125	(18,25 %)	29934	(16,49 %)
CZKCC	244044	15728	(6,44 %)	9941	(4,07 %)
CzLecDSP	54314	203	(0,37 %)	134	(0,25 %)

Tabulka 9.6: Výskyt jiných neřečových událostí ve zkoumaných databázích před a po zarovnání

Vzhledem k malému rozsahu databáze spontánních promluv oproti databázím čtené řeči bylo možné provést poslechovou analýzu a vyhodnotit výsledek zarovnávání s vjemem lidským uchem. Tak lze zjistit, zda dochází k odstraňování opravdu jen méně významných událostí. Výsledek analýzy shrnují následující body:



Obrázek 9.3: Rozdělení akustického skóre pro zarovnané neřečové události řečníka

– *Odstranění nevýrazných nádechů a výdechů*: Mezi promluvami se objevují záznamy, v nichž je míra zastoupení události typu nádech/výdech výrazně častější, než u jiných nahrávek. V těchto případech je tendence při značení neřečových událostí mnohem častěji značit událost jako významnou, neboť s jejich vyšším výskytem roste i variabilita a je obtížné stanovit hranici pro posouzení, kdy už událost přestává být významnou.

  Z 33 odstraněných značek události nádech/výdech bylo 51,5 % od jednoho mluvčího s vysokým počtem těchto jevů v promluvě.

– *Ostatní události – nevýznamné nebo maskované*: Mezi další neřečové události patří velmi specifické neřečové události, které lze pozorovat pouze u jednoho mluvčího, kde působí jako významný prvek. Z pohledu obecného modelu jsou však méně významné.

  Z 16 odstraněných značek jiných neřečových událostí bylo 50 % podobného charakteru od jednoho mluvčího.

– *Chyba v zarovnání*: V některých případech nedošlo ke korektnímu zarovnání a výsledek zarovnání (hranice slov a událostí) neodpovídal reálnému výskytu těchto událostí v promluvě.

  Ze všech zarovnávaných promluv se ve dvou z nich objevovala nesprávné zarovnání, jehož vlivem byly odstraněny některé neřečové události z transkripce.

Vysoký podíl odstraněných neřečových událostí z databáze spontánních promluv může být navíc způsoben i rozdílnými trénovacími podmínkami pro modely neřečových událostí.

9.4 Odolnost vůči rušení v úloze LVCSR

Analýza v kapitole 6 a 7 ukazuje kladný vliv navrhovaných technik předzpracování řečového signálu na výsledky rozpoznávání řeči při použití malého slovníku. Převedení těchto algoritmů do procesu rozpoznávání spojitých promluv s velkým slovníkem

je dalším krokem k analýze možnosti využití těchto algoritmů pro rozpoznávače řeči v reálných podmínkách. Následující sekce se proto zabývá možností použití výše zkoumaných algoritmů v úloze LVCSR.

V rámci této části je realizováno rozpoznávání spojitých promluv s využitím výše zkoumaných parametrizačních technik a předzpracováním pomocí ESS a VAD. Následně je sada akustických modelů doplněna o model vyplněné pauzy `FIL` pro analýzu možnosti zvýšení odolnosti vůči neřečovým událostem řečníka.

9.4.1 Nastavení LVCSR systému

Pro realizaci rozpoznávače spojitých promluv s velkým slovníkem bylo použito nastavení v souladu s popisem v sekci 4.1.3. Jazykový model lze řadit do skupiny obecných jazykových modelů, nejedná se tedy o model přizpůsobený konkrétním podmínkám (např. tématu promluvy). Zároveň nebyla výrazně optimalizována sada modelů, se kterými LVCSR systém pracoval. Ve výsledku tak bylo při rozpoznávání spojitých promluv dosaženo průměrné hodnoty pro real-time faktor 12,62.

Parametry dekodéru HDecode byly zvoleny s ohledem na výsledky analýzy obecného nastavení rozpoznávače spojitých promluv v [100]. Pro testování úspěšnosti rozpoznávání byla použita databáze nahrávek CzLecDSP. Bližší údaje o nastavení systému jsou shrnuty v tabulce 9.7.

Akustické modely		
modelovaný element		trifón
počet směsí		16
počet modelů po svázání (z celkem 83250 kombinací)	MFCC	15555
	PLP	13642
	RPLP	13758
	MFLP	13976
trénovací databáze		All, kanál CS0
použité techniky potlačení šumu	trénování	ESS + VAD (vypouštění segmentů)
	rozpoznávání	ESS
Jazykový model		
stupeň modelu		2 – bigramový model
počet bigramů		340 000
Nastavení dekodéru HDecode		
prořezávání na úrovni modelů (-t)		200
prořezávání na úrovni slov (-v)		50
posílení vlivu gramatiky (-s)		10
posílení vlivu vkládání slov (-p)		-10

Tabulka 9.7: Nastavení rozpoznávače plynulých promluv

9.4.2 LVCSR s modely neřečových událostí

Zahrnutí modelů neřečových událostí řečníka do gramatiky rozpoznávače s malým slovníkem definované dle sekce 4.1.3 lze provést pouze rozšířením seznamu slov o řetězec vystihující tuto neřečovou událost. Výskyt takové neřečové události je následně při rozpoznávání považován za stejně pravděpodobný v rámci celé promluvy.

Pro rozpoznávač spojitých promluv využívající statistický jazykový model je situace složitější, neboť model vystihuje pravděpodobnost kombinace n-tic slov. Jedním z přístupů může být adaptace jazykového modelu podle databáze přepisů promluv, které obsahují označení těchto neřečových událostí. Nevýhodou tohoto přístupu je velká závislost úspěšnosti výsledného systému na těchto adaptačních datech.

Druhým přístupem, který je použit i v této práci, je zahrnutí neřečové události do statistického modelu tak, aby pravděpodobnost výskytu neřečové události v každém n-gramu modelu byla shodná (např. [80]). Takový model umožňuje zařadit neřečovou událost mezi libovolná slova a tak vystihuje charakter neřečové události `FIL`. Zatímco událost typu nádech se většinou vyskytuje na začátku promluvy, vyplněná pauza se objevuje v rámci promluvy téměř libovolně ([115]), většinou v okamžiku, kdy mluvčí "rozmýšlí" pokračování promluvy.

Pro zahrnutí události `FIL` do jazykového modelu byla proto zjištěna pravděpodobnost výskytu události v promluvě. Tato analýza proběhla nad promluvami databáze CzLecDSP. Výslednou analýzu výskytu značky `FIL` shrnuje tabulka 9.8.

typ n-gramu	označení n-gramu	logaritmus pravděpodobnosti
Unigram	`FIL`	-2,62
Bigram	`FIL`–`FIL`	-2,69
Bigram	`FIL` na začátku věty	-3,25
Bigram	`FIL` na konci věty	-1,53

Tabulka 9.8: Pravděpodobnost výskytu n-gramů s neřečovou událostí `FIL`

Bigramový jazykový model byl doplněn o model neřečové události `FIL`, jehož výskytu byla přisouzena pravděpodobnost na základě analýzy výskytu této události v databázi CzLecDSP. Akustický model byl získán trénováním modelu na databázi SPEECON. Pro zvýšení vlivu této neřečové události během procesu rozpoznávání byla hodnota pravděpodobnosti vynásobena koeficientem 0,3, čímž dojde ke zvýšení vlivu této události v jazykovém modelu.

9.4.3 Experimentální část

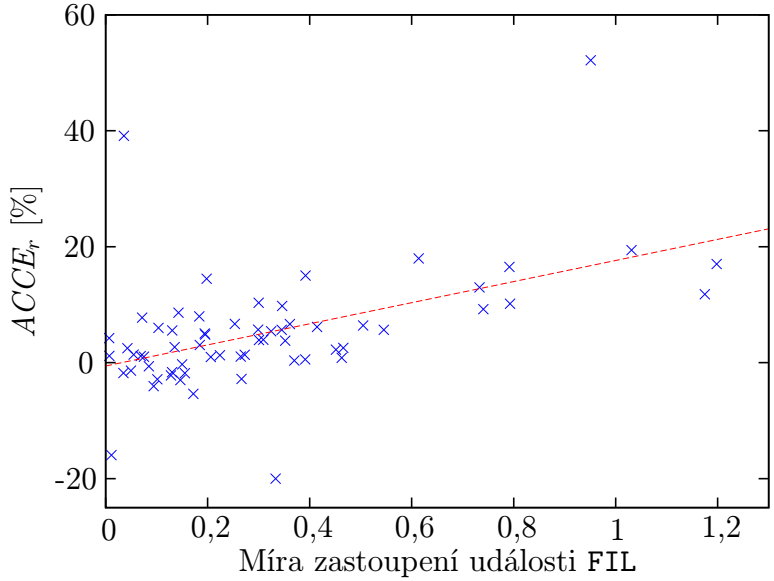
Výsledek rozpoznávání bez použití `FIL` v tabulce 9.9 opět ukazuje na výhodnost použití parametrizační techniky RPLP. Při použití modelu neřečové události `FIL` došlo ke zvýšení úspěšnosti rozpoznávání až na hodnotu 54,95 %.

	WER [%]			
	MFCC	PLP	MFLP	RPLP
bez modelu `FIL`	50,95	50,19	48,43	51,68
s modelem `FIL`	52,45	51,56	50,92	54,95

Tabulka 9.9: Výsledky rozpoznávání

Tabulka D.1 ukazuje výsledek rozpoznávání pro parametrizaci RPLP pro jednotlivé mluvčí. Je zde patrné, že pro vybrané mluvčí se úspěšnost pohybuje na úrovni přes 65 %. Jedná se především o příspěvky s obecnou tématikou. Naopak existují mluvčí, jejichž promluvy jsou rozpoznány s úspěšností pod 20 %. Při analýze jednotlivých promluv se ukázalo, že se jedná především o promluvy obsahující tématicky zaměřená slova, jako jsou přednášky na téma Zolotarevovy transformace (nahrávky 060 – 062) nebo Grangerovy kauzality (SPK4; nahrávky 028, 039, 052). Roli při rozpoznávání zároveň hraje i vlastní projev mluvčího. Proto jsou výsledky úspěšnosti rozpoznávání jednotlivých nahrávek stejného mluvčího většinou konzistentní.

Obecný trend příspěvku modelování neřečových událostí ve spontánní promluvě ukazuje graf 9.4. Zde je pro jednotlivé mluvčí vynesena hodnota zlepšení přesnosti rozpoznávání ($ACCE_r$) v závislosti na míře zastoupení značky `FIL`. Proložená přímka pak ukazuje vzestupný trend, který tato závislost vykazuje a potvrzuje tak výhodu použití modelů neřečových událostí v případě výskytu těchto událostí v promluvě.



Obrázek 9.4: Zlepšení přesnosti rozpoznávače v závislosti na výskytu události `FIL` v promluvě pro jednotlivé mluvčí

9.4.4 Shrnutí

Použití modelů neřečových událostí řečníka je již běžnou součástí dnešních rozpoznávačů řeči a v kombinaci s dalšími metodami potlačování neřečových jevů tak tvoří základní rámec pro robustní systémy. Analýza provedená v rámci této kapitoly ukazuje specifika trénování akustických modelů reprezentujících neřečové události řečníka a ukazuje využití některých závěrů v praktickém postupu nasazení těchto modelů při rozpoznávání promluv. Výsledkem je v některých případech zvýšení rozpoznávacího skóre v absolutní hodnotě o více než 10 %.

Kapitola 10

Závěr

Cílem této práce bylo prostudovat možnosti zvýšení robustnosti systému automatického rozpoznávání řeči v reálných podmínkách a navrhnout řešení využívající robustních metod pro zvýšení úspěšnosti rozpoznávání řeči v prostředí s přítomností šumu. Výzkum byl rozdělen do tématických částí, které na sebe logicky navazují a v rámci nichž jsou zkoumány algoritmy pro zvýšení odolnosti systému vůči rušivým jevům. Následující části shrnují nejvýznamnější dosažené závěry v jednotlivých oblastech.

Robustní parametrizační techniky

Důležitou oblastí pro tvorbu kvalitního rozpoznávače je volba parametrizační techniky, která umožní extrahovat z řečového signálu co nejvíce informace vhodné pro správnou klasifikaci. V této oblasti byly realizovány následující kroky.

- Prvním krokem při analýze parametrizačních technik bylo nalezení *optimalizovaného nastavení standardních metod* MFCC a PLP. Ukázalo se, že při volbě delších segmentačních oken jsou krátkodobá rušení např. v podobě impulsů potlačena, dochází ale současně ke snížení rozlišovací schopnosti i pro řečové prvky krátkodobého charakteru, např. plozivní hlásky. V souladu s nastavením většiny dnes používaných rozpoznávačů ukázaly i naše experimenty, že 25 ms je vhodná volba nastavení segmentace řečového signálu pro vyváženou citlivost systému.
- Standardní parametrizační techniky vykazují rozdílnou citlivost na přítomnost šumu v řečových datech vlivem rozdílného přístupu ke zpracování signálu v jednotlivých fázích výpočtu řečových příznaků. Byla provedena analýza *modifikovaných metod*, které kombinují jednotlivé funkční bloky standardních metod a umožňují studovat vliv a vlastnosti jednotlivých algoritmů použitých ve standardním zpracování.

  Především metoda RPLP se ukázala jako vhodná alternativa ke standardním postupům s využitím kombinace vlastností standardních technik. V čistých prostředích bylo dosahováno úspěšnosti srovnatelné s PLP metodou, v zašuměném prostředí byly výsledky srovnatelné s MFCC, zatímco PLP metoda dosaho-

vala výrazně vyšší chybovosti. V základní podobě tak na signálech z mikrofonu vykazujícího nižší úroveň zašumění dosáhly parametrizace MFCC, PLP a RPLP v průměru chybovosti 11,4 %, 11,8 % a 9,75 %. Další navrhované metody pomohly studovat význam potlačení dynamiky řečového signálu pro následné modelování. Toto potlačení se ukazuje jako důležité především v případě použití LPC analýzy, zatímco pro DFT analýzu může znamenat spíše ztrátu přesnosti. Ačkoliv parametrizační techniky založené na LPC analýze vykazují obecně vyšší citlivost na šum v signálu, při použití technik pro potlačování vlivu šumu je možné tuto citlivost významně potlačit. Výsledný rozpoznávač založený na těchto metodách může následně dosahovat úspěšnosti vyšší, než metody založené na přímé aplikaci DFT.

- Byl prokázán výrazný přínos *potlačení šumu technikou ESS*, s jejímž využitím bylo dosaženo zvýšení úspěšnosti rozpoznávání především v kombinaci s metodami založenými na LPC analýze v prostředí převážně aditivního šumu. Pro tyto podmínky se s metodou PLP snížila chybovost systému z 12,86 % na 9,8 %. Pro méně selektivní nahrávací kanál již nebylo výrazného zlepšení při použití samotného ESS dosaženo. Použití technik pro potlačení šumu je obecně limitováno předpoklady, které nebývají v reálném prostředí beze zbytku splněny. Pozorovatelné zlepšení bylo proto zaznamenáno především pro signály s nižší úrovní zašumění.

Adaptace modelů na rušivé prostředí

Techniky kompenzující vliv šumu na úrovni modelů navržené v této práci zahrnují několik algoritmů standardně používaných v jiných úlohách. Jejich vhodné použití při tvorbě akustických modelů vede k potlačení vlivu různých typů rušení.

- Navrhovaným postupem při aplikaci *detekce řečové aktivity* v rámci procesu rozpoznávání řeči je použití detekce pouze v průběhu trénování akustických modelů. Algoritmy VAD mohou pomoci odstranit neřečové části signálu a tím potlačit možnost jejich nesprávné klasifikace, na druhou stranu ve výrazně rušivých podmínkách může jejich špatná funkce přinést i chybné odstranění řečových segmentů. Použitím detekce pouze ve fázi trénování lze docílit zvýšení kvality získaných modelů řeči a zároveň neovlivnit kvalitu signálu při konečném použití rozpoznávače. V kombinaci s metodou ESS vedl tento postup na snížení chybovosti o více než 60 % v obecném prostředí pro metodu PLP. Na úloze AURORA3 bylo ve většině případů dosaženo úspěšnosti vyšší než pro 2-krokový algoritmus dle ETSI standardu.

- V rámci studia procesu přizpůsobení akustických modelů šumovým podmínkám byl analyzován vliv kvality výchozích modelů na proces přizpůsobení. Jsou-li akustické modely trénovány na signálech z konkrétního šumového prostředí, není jejich použití pro ostatní typy prostředí vhodné vzhledem k jejich přizpůsobení

na daný typ prostředí. Použití *obecných trénovacích podmínek* eliminuje tuto nevýhodu a přináší tak potřebnou variabilitu vůči změnám prostředí. Toho lze následně využít pro adaptaci modelů na dané šumové podmínky adaptačními technikami. Ačkoliv se trénováním na signálech z hlučného prostředí snižuje kvalita modelů, jejich použitím se podařilo snížit průměrnou chybovost systému pro obecné prostředí téměř o 29 %, zatímco jejich použití v případě velmi čistých nahrávek došlo ke zvýšení průměrné chybovosti pouze cca o 2 %. Modely natrénované na obecném prostředí mohou být s výhodou použity pro inicializaci v případě, že není známo cílové prostředí, případně pro urychlení procesu adaptace.

- Při použití *techniky MLLR* a iniciálních modelů trénovaných na čistých promluvách bylo i přes relativně jednoduché adaptační schéma dosaženo průměrné chybovosti 7,3 % oproti 8,5 % při přetrénování modelů standardním trénovacím procesem. Použití výchozích modelů trénovaných na podmínkách z obecného prostředí byla průměrná chybovost snížena na 6,38 %. V případě prostředí jedoucího automobilu s výrazně aditivním charakterem šumu bylo dosaženo chybovosti 1,65 %. Výsledné schéma je tak vhodným řešením pro využití při průběžném sledování a adaptaci na šumové prostředí, např. v automobilu. V další práci je možné se zaměřit na komplexnější adaptační techniky s přihlédnutím k charakteru adaptačních dat.

Řečová databáze spontánních promluv

V souladu s jinými zdroji se ukázalo, že čtený materiál, který je nejčastějším obsahem dostupných databází, obsahuje méně četné zastoupení neřečových událostí oproti promluvám spontánním. Proto byla v rámci práce vytvořena *databáze spojitých promluv spontánního charakteru*, která vystihuje rozdílné vlastnosti běžných promluv od čteného textu. Tato data jsou zdrojem důležitých informací a v kombinaci s rozsáhlými databázemi řeči jsou vhodným doplňkem při tvorbě rozpoznávače spojitých promluv.

V rámci tvorby databáze byly definovány a popsány některé základní aspekty zpracování spontánních promluv, které mohou být využity nejen při tvorbě obdobných zdrojů řečových dat, ale také v úloze rozpoznávání řeči. Z tohoto pohledu se jeví jako důležité především nespojitosti v řeči ovlivňující možnosti ruční a automatické segmentace nebo nepravidelnosti, které je nutné zaznamenat v přepisu obsahu nahrávek.

Modelování neřečových událostí řečníka

V souvislosti se snižováním citlivosti systému pro rozpoznávání řeči na rušení byl analyzován vliv neřečových událostí řečníka. Analýza použitých databází ukazuje, že řečové databáze čteného charakteru obsahují více událostí typu odkašlání, které souvisí často s uvolněním dýchacích cest před začátkem přerušované promluvy. Naopak vyplněná pauza provázející váhání a rozmýšlení dalšího pokračování promluvy je

méně častá. Důležitým prvkem je také vzdálenost mikrofonu od mluvčího, která ovlivňuje schopnost zachytit neřečové události slabšího charakteru, např. nádech mluvčího. První analýzy také ukázaly, že vysoká variabilita těchto jevů společně s jejich poměrně nízkým výskytem v trénovacím materiálu může být důvodem pro nedostatečně natrénovaný model. Na rozdíl od řečových elementů je ale velmi obtížné ovlivnit jejich přítomnost v trénovacím materiálu.

V rámci práce bylo navrženo schéma, které umožňuje výhodně využít databáze čtené řeči i s různým stupněm klasifikace značení neřečových událostí. Tím lze dosáhnout zvýšení počtu neřečových událostí v trénovacím materiálu bez nutnosti velmi složitých a nákladných manuálních zásahů do procesu trénování.

Použití neřečových událostí řečníka v procesu rozpoznávání řeči bylo analyzováno na úloze LVCSR. Již jednoduché zahrnutí modelu vyplněné pauzy do obecného statistického jazykového modelu vedlo ke zvýšení úspěšnosti rozpoznávání v absolutní hodnotě až o 3,27 %. Důležitým prvkem pro další zlepšení by byla analýza schopnosti jazykového modelu tyto události vložit do promluvy, aniž by tím docházelo ke ztrátě přesnosti rozpoznávání řeči.

Vlastní publikace autora

- Kočková-Amortová, L.; Pollák, P.; Rajnoha, J.; Ernestus, M. The Nijmegen Corpus of Casual Czech. (Opravená verze po 1. recenzi odeslaná do redakce) *Language Resources and Evaluation*, Springer Netherlands, 2013.

- Rajnoha, J.; Pollák, P. ASR systems in noisy environment: Analysis and solutions for increasing noise robustness. *Radioengineering 20*, 1 (2011), 74–84.

- Pollák, P.; Rajnoha, J. Multi-channel database of spontaneous Czech with synchronization of channels recorded by independent devices. In *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10)* Valletta, Malta, may 2010, N. Calzolari (Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds., European Language Resources Association (ELRA).

- Rajnoha, J.; Procházka, V.; Pollák, P. Tvorba rozpoznávače plynulých promluv v českém jazyce standardními nástroji HTK. *Akustické listy 16*, 1 (2010), 5–10.

- Rajnoha, J. Robustní rozpoznávání spojitých promluv kombinující metody potlačování šumu a průběžnou adaptaci akustických modelů na prostředí. In *Analýza a zpracování řečových a biologických signálů – sborník prací 2009* Praha: ČVUT v Praze, 2009.

- Rajnoha, J.; Pollák, P. Robust speech recognition in car environment combining noise reduction and acoustic model adaptation. In *Proceedings of the 19th Czech-German Workshop on Speech Processing* Prague: Academy of Sciences of the Czech Republic, Institute of Radioengineering and Electronics, 2009.

- Pollák, P.; Rajnoha, J. Long recording segmentation based on simple power voice activity detection with adaptive threshold and post-processing. In *SPECOM 2009 Proceedings* St. Petersburg: Institute for Informatics and Automation of RAS (SPIIRAS), 2009, pp. 55–60.

- Rajnoha, J. Multi-condition training for unknown environment adaptation in robust ASR under real conditions. In *POSTER 2009 [CD-ROM]* CTU in Prague, 2009, vol. 1, pp. 1–5.

- Rajnoha, J.; Pollák, P. Czech spontaneous speech collection and annotation: The database of technical lectures. In *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions*, A. Esposito and R. Vích, Eds., vol. 5641 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2009, pp. 377–385.

- Rajnoha, J. Multi-condition training for unknown environment adaptation in robust ASR under real conditions. *Acta Polytechnica 49*, 2-3/2009 (2009), 3–7.

- Rajnoha, J.; Pollák, P. Speaker non-speech event modelling in recognition of read and spontaneous speech. In *Proc. of Digital Technologies 2008* Zilina, SVK, 2008.

- Rajnoha, J. Sběr a anotace databáze spontánních promluv. In *Analýza a zpracování řečových a biologických signálů – sborník prací 2008* Praha: ČVUT v Praze, 2008.

- Rajnoha, J.; Pollák, P. Detektory řečové aktivity na bázi perceptivní kepstrální analýzy. In *Technical Computing Prague 2008* Prague, 2008, vol. 1, pp. 521–524.

- Rajnoha, J. Speaker non-speech event recognition with standard speech datasets. *Acta Polytechnica 47*, 4-5/2007 (February 2008), 107–111.

- Rajnoha, J. The analyses and improvements in non-speech event recognition task. In *POSTER 2007 [CD-ROM]* CTU in Prague, May 2007.

- Rajnoha, J.; Pollák, P. Modified feature extraction methods in robust speech recognition. In *Proceedings of 17th International Conference Radioelektronika 2007* Piscataway: Institute of Electrical and Electronic Engineers, 2007, vol. 1, pp. 521–524.

- Rajnoha, J.; Pollák, P. Voice activity detection in small vocabulary speech recognition. In *Proceedings of the 17th Czech-German Workshop on Speech Processing* Prague: Academy of Sciences of the Czech Republic, Institute of Photonics and Electronics AS ČR, 2007, pp. 43–48.

- Rajnoha, J.; Pollák, P. Modelling of speaker non-speech events in robust speech recognition. In *Proceedings of the 16th Czech-German Workshop on Speech Processing* Prague: Academy of Sciences of the Czech Republic, Institute of Photonics and Electronics AS ČR, 2006, pp. 149–155.

- Rajnoha, J. Spoken Czech digit recogniser at PC working in real environment. In *POSTER 2006 [CD-ROM]* CTU in Prague, March 2006.

Literatura

[1] Abad, A.; Meinedo, H.; Neto, J. Automatic classification and transcription of telephone speech in radio broadcast data. In *Proceedings of the 8th international conference on Computational Processing of the Portuguese Language (PROPOR '08)* Aveiro, Portugal, 2008, pp. 172–181.

[2] Abut, H.; Hansen, J. H. L.; Takeda, K. *DSP for In-Vehicle and Mobile Systems.* Springer Publishing, 2004.

[3] Andrassy, B.; Vlaj, D.; Beaugeant, C. Recognition performance of the Siemens front-end with and without frame dropping on the AURORA 2 database. In *Proc. EUROSPEECH* Scandinavia, 2001.

[4] Atal, B. S. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America 55*, 6 (1974), 1304–1312.

[5] Au Yeung, S.-K.; Siu, M.-H. Improved performance of AURORA 4 using HTK and unsupervised MLLR adaptation. In *Interspeech 2004 (ICSLP 2004)* October 2004.

[6] Bahl, L.; Brown, P.; de Souza, P.; Mercer, R. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.* 1986, vol. 11, pp. 49–52.

[7] Barras, C.; Geoffrois, E.; Wu, Z.; Liberman, M. Transcriber: A free tool for segmenting, labeling and transcribing speech. In *Proc. of the First international conference on language resources & evaluation (LREC)* Granada, Spain, 1998, pp. 1373–1376.

[8] Beaufays, F.; Boies, D.; Weintraub, M.; Zhu, Q. Using speech/non-speech detection to bias recognition search on noisy data. In *Proc. IEEE Conf. on Acoustic, Speech and Signal Processing (ICASSP'03)* 2003, vol. 1, pp. 424–427.

[9] Bees, D.; Blostein, M.; Kabal, P. Reverberant speech enhancement using cepstral processing. In *Proceedings of the Acoustics, Speech, and Signal Processing,*

(ICASSP-91) Washington, DC, USA, 1991, IEEE Computer Society, pp. 977–980.

[10] Benzeghiba, M.; Mori, R. D.; Deroo, O.; Dupont, S.; Erbes, T.; Jouvet, D.; Fissore, L.; Laface, P.; Mertins, A.; Ris, C.; Rose, R.; Tyagi, V.; Wellekens, C. Automatic speech recognition and speech variability: A review. *Speech Communication 49*, 10-11 (2007), 763–786.

[11] Bippus, R.; Fischer, A.; Stahl, V. Domain adaptation for robust automatic speech recognition in car environments. In *Proc. Eurospeech'99* Budapest, Hungary, 1999, pp. 1943–1946.

[12] Cappe, O. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Transactions on Speech and Audio Processing 2*, 2 (Apr. 1994), 345 – 349.

[13] Carlson, B. A.; Clements, M. A. Application of a weighted projection measure for robust hidden Markov model based speech recognition. In *ICASSP '91: Proceedings of the Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on* Washington, DC, USA, 1991, IEEE Computer Society, pp. 921–924.

[14] Clark, H. *Using Language*. Cambridge University Press, May 1996.

[15] CtuCopy [software package]. Ver. 3.0.11. `http://noel.feld.cvut.cz/speechlab`.

[16] Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on 28*, 4 (Aug 1980), 357–366.

[17] Duchateau, J.; Laureys, T.; Wambacq, P. Adding robustness to language models for spontaneous speech recognition. In *Proc. ISCA Workshop on Robustness Issues in Conversational Interaction* Norwich, UK, Aug 2004.

[18] ELRA – European Language Resources Association [online]. `http://www.elra.info/`.

[19] Ephraim, Y.; Malah, D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. In *IEEE Trans. Acoust., Speech, Signal Processing* April 1985, vol. ASSP-33, pp. 443–445.

[20] Ephraim, Y.; Trees, H. L. V. A signal subspace approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing 3*, 4 (July 1995), 251 – 266.

[21] ETSI. European Standard EN 300 965 - digital cellular telecommunications system (Phase 2+); Full rate speech; Voice Activity Detector (VAD) for full rate speech traffic channels, 2000.

[22] ETSI distributed speech recognition ES 202 050 standard [online]. `http://www.etsi.org/WebSite/Technologies/DistributedSpeechRecognition.aspx`.

[23] Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett. 27* (June 2006), 861–874.

[24] Fousek, P. *Extraction of Features for Automatic Recognition of Speech Based on Spectral Dynamics*. PhD thesis, CTU in Prague, 2007.

[25] Gajić, B.; Markhus, V.; Pettersen, S. G.; Johnsen, M. H. Automatic recognition of spontaneously dictated medical records for Norwegian. In *COST278 and ISCA Tutorial and Research Workshop - ROBUST2004* 2004.

[26] Gales, M. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language 12* (1998), 75–98.

[27] Gales, M. J. F. Maximum likelihood multiple subspace projections for hidden Markov models. *IEEE Transactions on Speech and Audio Processing 10*, 2 (Feb. 2002), 37 – 47.

[28] Gales, M. J. F.; et al. Progress in the CU-HTK broadcast news transcription system. *IEEE Transactions on Audio, Speech and Language Processing 14*, 5 (2006), 1513–1525.

[29] Gales, M. J. F.; Young, S. J. An improved approach to the hidden Markov model decomposition of speech and noise. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)* Los Alamitos, CA, USA, 1992, vol. 1, pp. 233–236.

[30] Gales, M. J. F.; Young, S. J. A fast and flexible implementation of parallel model combination. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP-95) 1* (1995), 133–136.

[31] Gales, M. J. F.; Young, S. J. Robust continuous speech recognition using parallel model combination. *IEEE Transactions on Speech and Audio Processing 4*, 5 (September 1996), 352–359.

[32] Gauvain, J.-L.; Lee, C.-H. Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing 2* (1994), 291–298.

[33] Gemello, R.; Mana, F.; De Mori, R. Non-linear estimation of voice activity to improve automatic recognition of noisy speech. In *Proc. of Interspeech 2005, 9-th European Conference on Speech Communication and Technology* Lisbon, Sep 2005.

[34] Gong, Y. Speech recognition in noisy environments: A survey. *Speech Communication 16*, 3 (1995), 261–291.

[35] Greenberg, S.; Popper, A. N.; Ainsworth, W. A.; Fay, R. R. *Speech Processing in the Auditory System.* SpringerVerlag, 2004.

[36] Grimm, M.; Kroschel, K., Eds. *Robust Speech Recognition and Understanding.* I-TECH Education and Publishing, Vienna, Austria, 2007.

[37] Haigh, J. A.; Mason, J. S. A voice activity detector based on cepstral analysis. In *Eurospeech'93 - Proceedings of the 3rd European Conference on Speech, Communication, and Technology* Berlin, Sept. 1993, pp. 1103–1106.

[38] Hain, T.; Woodland, P. C.; Evermann, G.; Gales, M. J. F.; Liu, X.; Moore, G. L.; Povey, D.; Wang, L. Automatic transcription of conversational telephone speech. *Speech and Audio Processing, IEEE Transactions on 13*, 6 (Nov. 2005), 1173–1185.

[39] Haykin, S.; Chen, Z. The cocktail party problem. *Neural Comput. 17*, 9 (2005), 1875–1902.

[40] Hermansky, H. Perceptual linear prediction (PLP) analysis of speech. *J. Acoust. Soc. Am. 87*, 04 (1990), 1738–1752.

[41] Hermansky, H. Should recognizers have ears? *Speech Commun. 25*, 1-3 (1998), 3–27.

[42] Hermansky, H. TRAP-TANDEM: Data-driven extraction of temporal features from speech. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, 2003. ASRU '03* Martigny, Switzerland, Nov.-3 Dec. 2003, pp. 255–260.

[43] Hermansky, H.; Ellis, D. P. W.; Sharma, S. Connectionist feature extraction for conventional HMM system. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP'00)* Istanbul, TR, 2000.

[44] Hermansky, H.; Morgan, N.; Bayya, A.; Kohn, P. Rasta-PLP speech analysis. ICSI Technical Report TR-91-069. Berkeley, California.

[45] Hermansky, H.; Sharma, S. Temporal patterns (TRAPs) in ASR of noisy speech. In *ICASSP '99: Proceedings of the Acoustics, Speech, and Signal Processing, 1999. on 1999 IEEE International Conference* Washington, DC, USA, 1999, IEEE Computer Society, pp. 289–292.

[46] Hönig, F.; Stemmer, G.; Hacker, C.; Brugnara, F. Revising perceptual linear prediction (PLP). In *Eurospeech 2005* 2005, pp. 2997–3000.

[47] International Telecommunication Union Recommendation G.729, annex b – A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70, 1996.

[48] Ircing, P. *Large vocabulary continuous speech recognition of highly inflectional language (Czech).* PhD thesis, University of West Bohemia, Faculty of Applied Sciences, Pilsen, Czech Republic, Plzeň, 2003.

[49] Juang, B. H.; Chou, W.; Lee, C. H. Minimum classification error rate methods for speech recognition. *IEEE Transactions on Speech and Audio Processing 5*, 3 (1997), 257–265.

[50] Kang, G.; Fransen, L. Quality improvement of LPC-processed noisy speech by using spectral subtraction. *Acoustics, Speech and Signal Processing, IEEE Transactions on 37*, 6 (Jun 1989), 939–942.

[51] Kennedy, L. S.; Ellis, D. P. W. Laughter detection in meetings. In *in Proc. NIST Meeting Recognition Workshop* 2004.

[52] Kida, Y.; Kawahara, T. Evaluation of voice activity detection by combining multiple features with weight adaptation. In *Proc. of Interspeech 2006, 9-th International conference on Spoken Language Processing* Pitssburgh, Sep 2006.

[53] Komori, Y.; Kosaka, T.; Yamamoto, H.; Yamada, M. Fast parallel model combination noise adaptation processing. In *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH'97)* September 1997, vol. 3, pp. 1523–1526.

[54] Krini, M.; Schmidt, G. Model-based speech enhancement. In *Speech and Audio Processing in Adverse Environments*, E. Hänsler and G. Schmidt, Eds., Signals and Communication Technology. Springer Berlin Heidelberg, 2008, pp. 89–134.

[55] LC-STAR II project site [online]. `http://www.lc-star.org/`.

[56] LDC – Linguistic Data Consortium [online]. `http://www.ldc.upenn.edu/`.

[57] Leggetter, C. J.; Woodland, P. C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language 9*, 2 (April 1995), 171–185.

[58] Li, Q.; Soong, F. K.; Siohan, O. An auditory system-based feature for robust speech recognition. In *Eurospeech 2001* 2001, vol. 1, pp. 619–622.

[59] Liao, Y. F.; Fang, H. H.; Hsu, C. H. Eigen-MLLR environment/speaker compensation for robust speech recognition. In *Proc. Interspeech'08* Brisbane, Australia, September 2008, pp. 1249–1252.

[60] Lim, J. S.; Oppenheim, A. V. Enhancement and bandwidth compression of noisy speech. In *Proc. IEEE* 1979, vol. 67, pp. 1586–1604.

[61] Lippmann, R.; Martin, E.; Paul, D. Multistyle training for robust isolated-word speech recognition. In *Proc. of IEEE ICASSP'87* 1987, pp. 709–712.

[62] Martin, R. Spectral subtraction based on minimum statistics. In *Proceedings of the European Signal Processing Conference (EUSIPCO'94)* Edinburgh, Scotland, 1994, pp. 1182–1185.

[63] Martin, R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. In *IEEE Trans. Speech Audio Process. 9 (5)* July 2001, pp. 504–512.

[64] Marzinzik, M.; Kollmeier, B. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Transactions on Speech and Audio Processing SAP-10*, 2 (FEB 2002), 109–118. ISSN 1063-6676.

[65] Matassoni, M.; Omologo, M.; Santarelli, A.; Svaizer, P. On the joint use of noise reduction and MLLR adaptation for in-car hands-free speech recognition. In *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on* 2002, vol. 1, pp. 289–292.

[66] Matsumoto, H.; Nakatoh, Y.; Furuhata, Y. An efficient Mel-LPC analysis method for speech recognition. In *Proc. ICSLP'98* 1998, pp. 1051–1054.

[67] McAulay, R.; Malpass, M. Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics, Speech and Signal Processing 28*, 2 (Apr. 1980), 137 – 145.

[68] Mihajlik, P.; Tobler, Z.; Tüske, Z.; Gordos, G. Evaluation and optimization of noise robust front-end technologies for the automatic recognition of Hungarian telephone speech. In *INTERSPEECH* 2005, ISCA, pp. 2677–2680.

[69] Ming, J.; Hou, B. Speech recognition in unknown noisy conditions. In Grimm and Kroschel [36], ch. 11, pp. 175–186.

[70] Ming, J.; Jancovic, P.; Hanna, P.; Stewart, D. Modelling the mixtures of known noise and unknown unexpected noise for robust speech recognition. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech'2001)* Aalborg, Denmark, September 2001, pp. 579–582.

[71] Mokbel, C.; Barbier, L.; Kerlou, Y.; Chollet, G. Word recognition in the car: Adapting recognizers to new environments. In *International Conference on Spoken Language Processing (ICSLP'92)* Banff, Alberta, Canada, October 1992, pp. 707–710.

[72] Mokbel, C.; Chollet, G. Word recognition in the car-speech enhancement/spectral transformations. In *ICASSP '91: Proceedings of the Acoustics, Speech, and Signal Processing* Washington, DC, USA, 1991, IEEE Computer Society, pp. 925–928.

[73] Morales, N.and Gu, L.; Gao, Y. Adding noise to improve noise robustness in speech recognition. In *Proc. Interspeech* Antwerp, Belgium, 2007, pp. 930–933.

[74] Moreno, A.; Lindberg, B.; Draxler, C.; Richard, G.; Choukri, K.; Euler, S.; Allen, J. Speechdatcar: A large speech database for automotive environments. In *Proceedings of the II LREC Conference* 2000.

[75] Nadas, A. A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood. *IEEE Transactions on Acoustics, Speech and Signal Processing 31*, 4 (Aug. 1983), 814 – 817.

[76] Nouza, J.; Silovsky, J.; Zdansky, J.; Cerva, P.; Kroul, M.; Chaloupka, J. Czech-to-Slovak adapted broadcast news transcription system. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association, (Interspeech'08)* Brisbane, Australia, September 2008, pp. 2683–2686.

[77] Nouza, J.; Zdansky, J.; David, P.; Cerva, P.; Kolorenc, J.; Nejedlova, D. Fully automated system for Czech spoken broadcast transcription with very large (300k+) lexicon. In *Proc. of International Conference on Speech and Language Processing (Interspeech 2005)* Lisboa, Portugal, September 2005.

[78] Novotný, J. *Robust speech recognition algorithms.* PhD thesis, CTU in Prague, 2004.

[79] Novotný, J.; Macháček, L. Noise reduction applied in real time speech recognition system. In *Proc. of Polish-Czech-Hungarian Workshop on Circuit Theory, Signal Processing, and Telecommunication Networks* Budapest, Hungary, Sept. 2001.

[80] Nunes, R. J. F.; Neves, L. M. L. Filled pause modeling. Tech. Rep. 35 / 2006 INESC-ID Lisboa, August 2006.

[81] Obuchi, Y.; Amano, A. Always listening to you: Creating exhaustive audio database in home environments. In *Proc. InterSpeech 2007* Antwerp, Belgium, August 2007, pp. 566–569.

[82] Openshaw, J.; Mason, J. On the limitations of cepstral features in noise. In *Proc. IEEE 1994 Conf. Acoust. Speech Signal Process.* 1994, vol. 2, pp. 49–52.

[83] Pearce, D.; Hirsch, H.-G. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *ISCA ITRW ASR2000* 2000, pp. 29–32.

[84] Pollák, P.; Černocký, J. Czech SPEECON adult database [on-line], November 2003. `http://www.speechdat.org/speecon`.

[85] Pollák, P.; Hanžl, V. Tool for Czech pronunciation generation combining fixed rules with pronunciation lexicon and lexicon management tool. In *Proc. of LREC'02, Third International Conference on Language Resources and Evaluation* Las Palmas, Spain, May 2002.

[86] Procházka, V.; Pollák, P.; Žďánský, J.; Nouza, J. Performance of Czech speech recognition with language models created from public resources. *Radioengineering 20*, 4 (Dec 2011).

[87] Psutka, J. *Techniky parametrizace, dekorelace a redukce dimenze příznaků v systémech rozpoznávání řeči.* PhD thesis, University of West Bohemia, Faculty of Applied Sciences, Pilsen, Czech Republic, Plzeň, 2007.

[88] Psutka, J.; Müller, L.; Matoušek, J.; Radová, V. *Mluvíme s počítačem česky.* Academia, Prague, 2006.

[89] Psutka, J.; Müller, L.; Psutka, J. V. Comparison of MFCC and PLP parametrizations in the speaker independent continuous speech recognition task. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech'01)* Aalborg, Denmark, September 2001, vol. 3, pp. 1813–1816.

[90] Psutka, J.; Müller, L.; Psutka, J. V. The influence of a filter shape in telephone-based recognition module using PLP parameterization. In *TSD '01: Proceedings of the 4th International Conference on Text, Speech and Dialogue* London, UK, 2001, Springer-Verlag, pp. 222–228.

[91] Psutka, J.; Radová, V.; Müller, L.; Matoušek, J.; Ircing, P.; Graff, D. Large broadcast news and read speech corpora of spoken Czech. In *Proc. Eurpospeech 2001* Ålborg, Denmark, 2001, pp. 2067–2070.

[92] Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE* 1989, pp. 257–286.

[93] Rabiner, L. R.; Juang, B.-H. *Fundamentals of Speech Recognition.* Prentice Hall, Inc., Englewood Cliffs, NJ, USA, 1993.

[94] Rabiner, L. R.; Schafer, R. W. Introduction to digital speech processing. *Found. Trends Signal Process. 1*, 1 (2007), 1–194.

[95] Rabiner, L. R.; Schaffer, R. W. *Digital Processing of Speech Signals.* Prentice Hall, Inc., Englewood Cliffs, NJ, USA, 1978.

[96] Rajnoha, J. Speech recognition in real environment on standard PC platform. Diplomová práce, CTU in Prague, 2006.

[97] Rajnoha, J. Speaker non-speech event recognition with standard speech datasets. *Acta Polytechnica 47*, 4-5/2007 (February 2008), 107–111.

[98] Rajnoha, J.; Pollák, P. Modelling of speaker non-speech events in robust speech recognition. In *Proceedings of the 16th Czech-German Workshop on Speech Processing* Prague: Academy of Sciences of the Czech Republic, Institute of Photonics and Electronics AS ČR, 2006, pp. 149–155.

[99] Rajnoha, J.; Pollák, P. Modified feature extraction methods in robust speech recognition. In *Proceedings of 17th International Conference Radioelektronika 2007* Piscataway: Institute of Electrical and Electronic Engineers, 2007, vol. 1, pp. 521–524.

[100] Rajnoha, J.; Procházka, V.; Pollák, P. Tvorba rozpoznávače plynulých promluv v českém jazyce standardními nástroji HTK. *Akustické listy 16*, 1 (2010), 5–10.

[101] Rodriguez, L. J.; Torres, I.; Varona, A. Evaluation of sublexical and lexical models of acoustic disfluencies for spontaneous speech recognition in Spanish. In *EUROSPEECH-2001* 2001, pp. 1665–1668.

[102] Rosca, J.; Balan, R.; Fan, N. P.; Beaugeant, C.; Gilg, V. Multichannel voice detection in adverse environments. In *Proc. of EUSIPCO 2002* Toulouse, France, Sep 2002.

[103] Rotovnik, T.; Maucec, M. S.; Kacic, Z. Large vocabulary continuous speech recognition of an inflected language using stems and endings. *Speech Communication 49*, 6 (2007), 437–452.

[104] SAMPA project webpage [on-line]. `http://www.phon.ucl.ac.uk/home/sampa/index.html`.

[105] Saon, G.; Huerta, H.; Jan, E. Robust digit recognition in noisy environments: the IBM AURORA 2 system. In *Proc. of Interspeech'01* 2001, pp. 629–632.

[106] Schluter, R.; Bezrukov, L.; Wagner, H.; Ney, H. Gammatone features and feature combination for large vocabulary speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007* Honolulu, HI, USA, Apr 2007, pp. 649–652.

[107] Schramm, H.; Aubert, X.; Bakker, B.; Meyer, C.; Ney, H. Modeling spontaneous speech variability in professional dictation. *Speech Communication 48*, 5 (2006), 493 – 515.

[108] Shriberg, E. Phonetic consequences of speech disfluency. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS-99)* San Francisco, USA, 1999, vol. 1, pp. 619–622.

[109] Shriberg, E. Spontaneous speech: How people really talk, and why engineers should care. In *Proc. Eurospeech 2005* Lisbon, Portugal, 2005, pp. 1781–1784.

[110] Siu, M.-H.; Ostendorf, M. Modeling disfluencies in conversational speech. In *Proceedings of Fourth International Conference on Spoken Language (ICSLP'96)* Philadelphia, PA, USA, October 1996, vol. 1, pp. 386–389.

[111] Sovka, P.; Pollák, P. The study of speech/pause detectors for speech enhancements methods. In *EUROSPEECH'95 - Proceedings of the 4th European Conference on Speech Communication and Technology* Madrid, Spain, September 1995, pp. 1575–1578.

[112] Sovka, P.; Pollák, P.; Kybic, J. Extended spectral subtraction. In *European Signal Processing Conference (EUSIPCO–96)* Trieste, Italy, September 1996.

[113] SPEECON project webpage [on-line]. `http://www.speechdat.org/speecon`.

[114] Stevens, K. N. *Acoustic Phonetics.* Current Studies in Linguistics 30. MIT Press, Cambridge, 1999.

[115] Stolcke, A.; Shriberg, E. Statistical language modeling for speech disfluencies. In *Proc. ICASSP '96* Atlanta, GA, 1996, pp. 405–408.

[116] Stouten, F.; Martens, J.-P. Benefits of disfluency detection in spontaneous speech recognition. In *Cost 278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction* Norwich, UK, 8 2004.

[117] Tanyer, S. G.; Özer, H. Voice activity detection in nonstationary noise. *IEEE Trans. on Speech and Audio Processing*, 4 (July 2000), 478–482.

[118] Trancoso, I.; Nunes, R.; Neves, L.; Viana, C.; Moniz, H.; Caseiro, D.; Mata, A., I. Recognition of classroom lectures in european Portuguese. In *Proc. Interspeech 2006* Pittsburgh, USA, 2006.

[119] Tüske, Z.; Mihajlik, P.; Tobler, Z.; Fegyó, T. Robust voice activity detection based on the entropy of noise-suppressed spectrum. In *Proc. of Interspeech 2005, 9-th European Conference on Speech Communication and Technology* Lisbon, Sep 2005.

[120] Varga, A.; Moore, R. Hidden Markov model decomposition of speech and noise. In *Proc. of IEEE Internat. Conf. on Acoustics, Speech and Signal Processing (ICASSP'90)* Albuquerque, Apr. 1990, pp. 845–848.

[121] Vaseghi, S. V. *Advanced Digital Signal Processing and Noise Reduction*, second ed. John Wiley & Sons, 2000.

[122] Wolfe, P. J.; Godsill, S. J. Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement. In *IEEE Workshop on Statistical Signal Processing* 2001, pp. 496–499.

[123] Young, S.; et al. *The HMM Toolkit (software and manual), ver. 3.2.1.* Cambridge University Engineering Department, 2002. `http://htk.eng.cam.ac.uk/`.

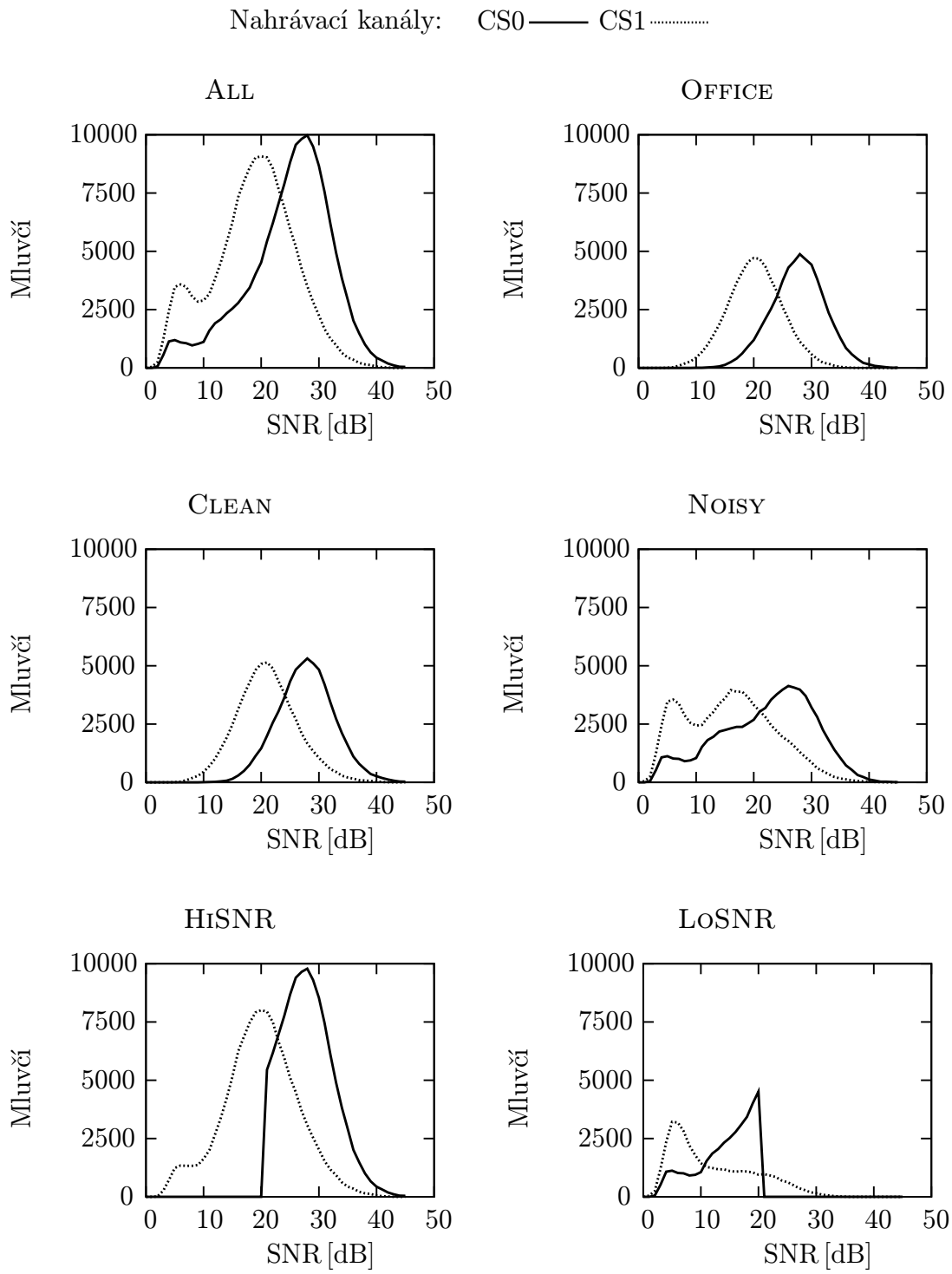
[124] Young, S. J.; Woodland, P. C.; Byrne, W. J. Spontaneous speech recognition for the credit card corpus using the HTK toolkit. *IEEE Transactions on Speech and Audio Processing 2*, 4 (October 1994), 615–621.

[125] Zolnay, A.; Schlüter, R.; Ney, H. Acoustic feature combination for robust speech recognition. In *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing* Philadelphia, PA, USA, March 2005, pp. 457–460.

[126] Ústav Českého národního korpusu – SYN2006PUB, 2006.

Příloha A

Rozložení záznamů dle SNR v databázi SPEECON



Obrázek A.1: Odhad průměrného SNR v trénovací části fragmentů databáze SPEECON (čárkované šrafování – náhlavní mikrofon, plný graf – hands-free mikrofon)

Příloha B

Výsledky rozpoznávání s adaptací pro různé počty regresních tříd

náhlavní mikrofon (CS0)									
výchozí modely	regr. třídy	All	Office	Car	Clean	Noisy	HiSNR	LoSNR	Průměr
Čisté	1	8,49	4,45	2,86	4,89	10,64	7,25	11,34	7,13
	2	8,09	4,39	2,96	4,59	10,52	7,1	11,23	6,98
	4	7,82	4,05	3,31	4,49	10,18	6,88	10,95	6,81
	8	7,92	4,03	3,52	4,55	10,42	6,93	10,79	6,88
	16	7,86	3,98	3,57	4,45	10,38	6,84	10,79	6,84
	32	7,88	3,81	3,77	4,49	10,31	6,89	10,72	6,84
	64	7,94	3,83	3,67	4,47	10,23	6,95	10,66	6,82
	min.	7,82	3,81	2,86	4,45	10,18	6,84	10,66	6,81
Obecné	1	7,47	3,69	2,4	4,66	8,9	6,05	9,84	6,15
	2	7,23	3,54	3,31	4,53	8,61	6,3	9,59	6,16
	4	7,21	3,58	3,31	4,64	8,76	6,11	9,6	6,17
	8	7,23	3,65	3,16	4,66	8,64	6,12	9,78	6,18
	16	7,22	3,71	3,01	4,63	8,49	6,1	9,62	6,11
	32	7,21	3,67	3,11	4,66	8,61	6	9,53	6,11
	64	7,18	3,63	3,21	4,66	8,46	5,99	9,6	6,1
	min.	7,18	3,54	2,4	4,53	8,46	5,99	9,53	6,1

hands-free mikrofon (CS1)									
výchozí modely	regr. třídy	All	Office	Car	Clean	Noisy	HiSNR	LoSNR	Průměr
Čisté	1	10,19	5,78	7,7	7,74	12,9	7,61	11,78	9,1
	2	8,78	6,07	5,35	7,19	10,75	7,4	10,52	8,01
	4	8,89	5,7	5,05	7,1	11,55	7,24	10,77	8,04
	8	8,91	5,79	5,25	7,23	11,7	7,34	11,07	8,18
	16	9,25	5,88	5,1	7,38	12,07	7,35	11,19	8,32
	32	9,38	5,81	5,4	7,19	12,22	7,33	11,55	8,41
	64	9,36	5,78	5,05	7,19	12,23	7,35	11,69	8,38
	min.	8,78	5,7	5,05	7,1	10,75	7,24	10,52	8,01
Obecné	1	8,3	5,43	3,93	7,57	8,93	7,9	9,63	7,39
	2	7,76	5,61	3,21	7,33	7,87	7,4	8,92	6,87
	4	7,53	5,36	3,21	6,76	7,88	7,2	8,87	6,69
	8	7,44	5,52	3,21	6,81	7,78	7,45	8,84	6,72
	16	7,48	5,56	3,16	6,95	7,7	7,47	8,91	6,75
	32	7,44	5,43	3,01	6,83	7,81	7,41	8,92	6,69
	64	7,46	5,43	3,11	6,68	7,72	7,39	8,83	6,66
	min.	7,44	5,36	3,01	6,68	7,7	7,2	8,83	6,66

Tabulka B.1: Srovnání průměrných hodnot *WER* pro čisté a obecné podmínky trénování výchozích modelů – po adaptaci modelů

Příloha C

Zastoupení promluv v databázi CzLecDSP

Mluvčí	Počet nahrávek	Celkový čas nahrávek	Tématický okruh	Označení session
SPK1	5	01:39:03	Zpracování řeči, neuronové sítě	005, 014, 023, 032, 046
SPK2	5	02:03:40	Zpracování řeči	004, 013, 027, 038, 050
SPK3	5	02:47:36	Zpracování řeči	003, 012, 022, 031, 049
SPK4	3	00:50:44	Zpracování EEG	028, 039, 052
SPK5	3	01:07:18	Zpracování EEG	002, 010, 042
SPK6	1	00:00:28	Zolotarevova transformace	062
SPK7	2	00:33:58	Zpracování řeči	006, 017
SPK8	4	00:52:46	Zpracování řeči	019, 026, 035, 047
SPK9	2	00:12:08	Zolotarevova transformace	060, 061
SPK10	1	00:15:31	Zpracování řeči	051
SPK11	3	01:14:16	Zpracování řeči, hardware	029, 040, 053
SPK12	1	00:10:06	Zpracování řeči	020
SPK13	2	00:22:13	Nanoskopie	041, 054
SPK14	1	00:13:31	Zpracování řeči	037
SPK15	1	00:20:24	Poruchy řeči	056
SPK16	1	00:23:08	Zpracování řeči	055
SPK17	1	00:14:32	Neuronové sítě	058
SPK18	3	01:27:32	Teorie digitálního zpracování řeči	000, 001, 009
SPK19	1	00:14:20	Biometrie	059
SPK20	4	01:22:26	Zpracování řeči	007, 016, 036, 048
SPK21	4	01:22:13	Zpracování EEG	018, 024, 033, 044
SPK22	4	01:48:05	Zpracování EEG	011, 021, 030, 043
SPK23	1	00:12:26	Poruchy řeči	057
SPK24	5	01:55:31	Zpracování řeči, neuronové sítě	008, 015, 025, 034, 045

Tabulka C.1: Zastoupení promluv v databázi pro jednotlivé mluvčí

Příloha D

Úspěšnost rozpoznávání s použitím modelů neřečových událostí

Mluvčí	ACC [%]		$ACCE$ [%]	$ACCE_r$ [%]	četnost `FIL` [%]
	bez modelu `FIL`	s modelem `FIL`			
000	60,96	64,03	3,07	5,04	0,194
001	60,45	65,66	5,21	8,62	0,142
002	40,92	43,65	2,73	6,67	0,361
003	56,86	57,44	0,58	1,02	0,074
004	64,6	68,29	3,69	5,71	0,298
005	47,05	44,52	-2,53	-5,38	0,171
006	40,21	40,64	0,43	1,07	0,264
007	54,21	62,36	8,15	15,03	0,391
008	46,52	45,13	-1,39	-2,99	0,146
009	57,7	60,91	3,21	5,56	0,13
010	40,23	41,76	1,53	3,8	0,351
011	57,22	66,67	9,45	16,52	0,791
012	62,78	65,44	2,66	4,24	0,007
013	62,63	65,67	3,04	4,85	0,195
014	48,07	49,53	1,46	3,04	0,184
015	46,8	46,67	-0,13	-0,28	0,149
016	59,78	63,36	3,58	5,99	0,103
017	40,36	40,91	0,55	1,36	0,272
018	40,73	42,94	2,21	5,43	0,324
019	68,3	67,18	-1,12	-1,64	0,13
020	63,53	62,4	-1,13	-1,78	0,034
021	55,13	65,84	10,71	19,43	1,031
022	63,28	68,2	4,92	7,77	0,071
023	50,15	51,43	1,28	2,55	0,466
024	52,13	55,09	2,96	5,68	0,345
025	52	49,89	-2,11	-4,06	0,094
026	65,78	68,38	2,6	3,95	0,309
027	62,76	66,78	4,02	6,41	0,504
028	18,22	18,43	0,21	1,15	0,068
029	62,45	73,67	11,22	17,97	0,613
030	61,24	68,46	7,22	11,79	1,174
031	60,87	61,71	0,84	1,38	0,055
032	50,68	50,97	0,29	0,57	0,391
033	50,44	51,08	0,64	1,27	0,224
034	56,42	55,4	-1,02	-1,81	0,155

pokračování na další straně ...

Mluvčí	ACC [%]		$ACCE$ [%]	$ACCE_r$ [%]	četnost `FIL` [%]
	bez modelu `FIL`	s modelem `FIL`			
035	63,78	64,31	0,53	0,83	0,462
036	61,02	67,33	6,31	10,34	0,3
037	42,02	49,17	7,15	17,02	1,198
038	61,04	67,24	6,2	10,16	0,793
039	16,76	23,32	6,56	39,14	0,035
040	64,35	68	3,65	5,67	0,545
041	57,47	55,83	-1,64	-2,85	0,101
042	40,13	41,7	1,57	3,91	0,3
043	56,16	63,45	7,29	12,98	0,733
044	50,31	48,9	-1,41	-2,8	0,266
045	45,43	45,15	-0,28	-0,62	0,084
046	24,33	24,42	0,09	0,37	0,37
047	61,92	63,31	1,39	2,24	0,451
048	60,16	64,96	4,8	7,98	0,183
049	61,59	63,13	1,54	2,5	0,042
050	62,33	68,43	6,1	9,79	0,346
051	43,79	43,19	-0,6	-1,37	0,049
052	20,52	17,25	-3,27	-15,94	0,011
053	63,38	67,3	3,92	6,18	0,414
054	59,07	63,02	3,95	6,69	0,253
055	49,49	50,82	1,33	2,69	0,135
056	53,36	52,18	-1,18	-2,21	0,127
057	64,22	64,97	0,75	1,17	0,007
058	43,2	47,19	3,99	9,24	0,74
059	29,24	44,49	15,25	52,15	0,951
060	23,61	27,03	3,42	14,49	0,198
061	26,45	26,71	0,26	0,98	0,206
062	27,78	22,22	-5,56	-20,01	0,333
Celkem	51,68	54,95	3,27	6,33	0,314

Tabulka D.1: Úspěšnost rozpoznávání pro parametrizaci RPLP