VĚDECKÉ SPISY VYSOKÉHO UČENÍ TECHNICKÉHO V BRNĚ

Edice PhD Thesis, sv. 650

ISSN 1213-4198

Ing. Petr Zelinka

Zvyšování účinnosti strojového rozpoznávání řeči

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV RADIOELEKTRONIKY

Ing. Petr Zelinka

ZVYŠOVÁNÍ ÚČINNOSTI STROJOVÉHO ROZPOZNÁVÁNÍ ŘEČI

ENHANCING THE EFFECTIVENESS OF AUTOMATIC SPEECH RECOGNITION

ZKRÁCENÁ VERZE PH.D. THESIS

Studijní obor:	Elektronika a sdělovací technika
Školitel:	prof. Ing. Milan Sigmund, CSc.
Oponenti:	prof. Ing. Jana Tučková, CSc. prof. Ing. Jan Nouza, CSc.
Datum obhajoby:	29. března 2012

KLÍČOVÁ SLOVA

Strojové rozpoznávání řeči, skryté Markovovy modely, nestacionární šum, hlasové úsilí, variabilita řeči.

KEYWORDS

Automatic speech recognition, hidden Markov models, nonstationary noise, vocal effort, speech variability.

DISERTAČNÍ PRÁCE JE ULOŽENA:

Ústav radioelektroniky
Fakulta elektrotechniky a komunikačních technologií
Vysoké učení technické v Brně
Purkyňova 118
61200 Brno

ISBN 978-80-214-4471-3
ISSN 1213-4198

OBSAH

1 ÚVOD 5
2 VLIVY SNIŽUJÍCÍ ÚSPĚŠNOST ROZPOZNÁVÁNÍ ŘEČI 6
2.1 Akustický šum 7
2.2 Změny v řečovém projevu mluvčího 8
3 NAVRŽENÉ METODY ZVYŠOVÁNÍ ŠUMOVÉ ODOLNOSTI ROZPOZNÁVAČE 10
3.1 Kubická interpolace GMM 10
3.2 PMC s vícestavovým ergodickým HMM šumu 12
3.3 Použití hierarchie HMM pro modelování šumu 15
4 NAVRŽENÉ METODY ZVYŠOVÁNÍ ODOLNOSTI ROZPOZNÁVAČE VŮČI ZMĚNÁM V PROMLUVĚ MLUVČÍHO 18
4.1 Databáze hlasového úsilí BUT-VE1 18
4.2 Klasifikátor hlasového úsilí nezávislý na slovníku 21
4.3 Zvyšování robustnosti rozpoznávače vůči změnám v hlasovém úsilí mluvčího 23
5 ZÁVĚR 25
VYBRANÉ VLASTNÍ PUBLIKACE 27
VYBRANÁ LITERATURA 28
ŽIVOTOPIS 29
ABSTRAKT 30
ABSTRACT 30

1 ÚVOD

Vývoj metod pro automatické rozpoznávání řeči probíhá již přes půl století, ovšem ani přes enormní úsilí vědeckých pracovišť po celém světě ještě stále nebylo dosaženo úspěšnosti rozpoznávání srovnatelné se schopnostmi lidí. Dnešní komerčně dostupné systémy jsou vždy omezeny předpokládaným rozsahem použití. Na jedné straně stojí rozpoznávače plynule diktované řeči s velkým slovníkem, jejichž nízká chybovost je podmíněna nízkou úrovní okolního hluku, použitím náhlavního mikrofonu a neutrální promluvou mluvčího. Diktovací systémy se též opírají o statistický jazykový model, který ale nelze uplatnit v případě rozpoznávání izolovaných slovních příkazů. Druhou skupinu pak tvoří systémy zaměřené na rozpoznávání omezeného slovníku v prostředí s nezanedbatelnou úrovní okolních ruchů. Nejčastěji se jedná o systémy určené k použití v automobilech, vojenských letadlech apod. Také zde je častou podmínkou pro dosažení vysoké úspěšnosti neutrální promluva mluvčího a nezřídka speciální hardware (mikrofonní pole). Navíc jsou tyto systémy vždy přijatelně použitelné jen v daném konkrétním prostředí.

Při mém setkání s Prof. Dr. med. Christopherem Nimsky a jeho lékařským týmem na neurochirurgickém pracovišti Uniklinikum Marburg v Německu jsem byl překvapen jejich konstatováním, že nevěří v možnost použití hlasem ovládaných přístrojů na operačním sále. Tato skepse vychází ze zkušenosti s komerčními systémy, které jim byly v minulosti nabízeny k otestování a které se ukázaly být z hlediska úspěšnosti rozpoznávání řeči nedostatečně spolehlivými v akustických podmínkách jejich pracoviště.

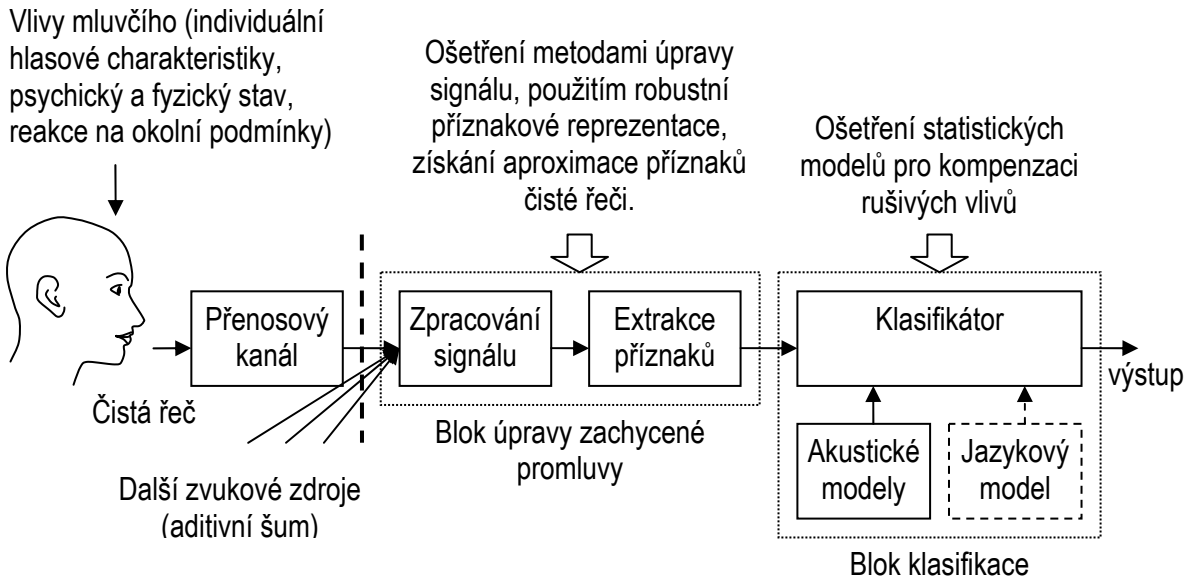
Cílem mé práce je analyzovat zdroje chybovosti u současných systémů pro automatické rozpoznávání řeči a formulovat vylepšení a nové přístupy vedoucí ke zvýšení robustnosti těchto systémů. Výzkum byl specificky zaměřen na akustické prostředí operačního sálu. V rámci řešení této problematiky byly vytčeny následující cíle:

- Prvním cílem bylo určit strukturu rozpoznávače, která nabídne nejvyšší potenciál možného zvyšování úspěšnosti rozpoznávání s uvážením malého slovníku izolovaně vyslovených slov od předem známé skupiny mluvčích.
- Za předpokladu použití optimální výchozí struktury rozpoznávače bylo dále prioritou formulovat postupy pro zvýšení šumové robustnosti rozpoznávače v reálném akustickém prostředí operačního sálu. Při zkoumání vhodných metod bylo potřeba zaměřit pozornost na nejpodstatnější část rozpoznávače, kterou jsou akustické statistické modely. Cílem bylo zlepšit schopnost těchto modelů adekvátně postihnout nestacionární šum typický pro dané pracovní prostředí. Předpokladem pro úspěšné řešení bylo získání reprezentativního záznamu zvuků na operačním sále při probíhajícím zákroku.
- Spolehlivá funkce rozpoznávače za všech podmínek vyžaduje robustnost vůči změnám ve způsobu promluvy mluvčího. Nejvýraznějšími změnami, které lze

předpokládat u uživatelů rozpoznávače v prostředí operačního sálu, jsou rozdílné úrovně hlasového úsilí. V dosavadní literatuře nebyl dopad změn v celém rozsahu hlasového úsilí (od šepotu až po křik) na úspěšnost rozpoznávání kvantifikován, bylo tedy nutné tento dopad experimentálně ověřit. Dále pak bylo cílem navrhnout vhodný přístup pro zvýšení robustnosti systému vůči těmto změnám a zajistit tak konzistentní spolehlivost rozpoznávání.

2 VLIVY SNIŽUJÍCÍ ÚSPĚŠNOST ROZPOZNÁVÁNÍ ŘEČI

Hlavním zaměřením disertace bylo hledat způsoby zvyšování úspěšnosti automatického rozpoznávání řeči. Pro dosažení tohoto cíle bylo nutné nejprve identifikovat příčiny selhávání automatických rozpoznávačů a zjistit míru dopadu příslušných vlivů na úspěšnost rozpoznávání. V nezkráceném textu disertace je uveden ucelený přehled nejvýznamnějších rušivých vlivů včetně známých metod pro jejich detekci analýzou řečového signálu a způsobů kompenzace. Základní představu o rušivých vlivech v souvislosti s konstrukcí rozpoznávače řeči si lze učinit z obr. 2.1.



Obr. 2.1 Obecný model rozpoznávače řeči, zdroje rušivých vlivů a oblasti jejich ošetření

Jádrem každého rozpoznávače řeči je sada akustických modelů řečových jednotek (slov, fonémů, trifonů apod.) vytvořená při konstrukci systému. U současných rozpoznávačů mají tyto modely nejčastěji strukturu skrytých Markovových modelů (hidden Markov models, HMM) se spojitými výstupními rozloženími tvořenými směsí gaussovských funkcí (Gaussian mixture model, GMM). V rámci hledání nejvhodnější konstrukce rozpoznávače pro možnost dalšího zvyšování účinnosti byla testována i alternativní struktura využívající přímé porovnávání se slovními vzory na základě borcení časové osy (dynamic time warping, DTW). HMM se však ukázaly být po všech stránkách výhodnější (úspěšnost, rychlost rozpoznávání,

potenciál pro další vylepšování). Výsledky rozpoznávání jsou primárně vždy závislé na míře shody mezi rozpoznávanou promluvou a použitými modely. Negativní dopad rušivých vlivů spočívá ve zvýšení rozdílnosti příslušných modelů oproti řeči a následně snížení schopnosti rozpoznávače oddělit jednotlivé rozpoznávané třídy. Záměnou modelů pak dochází ke zvýšení chybovosti a selhávání systému. Rušivé vlivy lze rozdělit do několika kategorií:

- Aditivní šum – k řečovému signálu se z okolí přidávají další rušivé signály.
- Konvoluční zkreslení – dáno frekvenční charakteristikou přenosového kanálu (akustika místnosti, poloha mikrofonu, elektrická přenosová cesta, aj.).
- Vliv mluvčího – specifika mluvčího, stres, fyzická a psychická únava, onemocnění, reakce na okolní hluk (Lombard efekt), úroveň hlasového úsilí, rychlost a pečlivost výslovnosti atd. – každý z těchto vlivů hraje roli při vytváření řeči a mění tak parametry řečového signálu.

V následujícím textu je uveden výtah z přehledu existujících metod pro ošetření rušivých vlivů.

2.1 AKUSTICKÝ ŠUM

Při rozpoznávání řeči se vstupní signál po provedení prvotního předzpracování vždy převádí na sled příznakových vektorů získaných metodami krátkodobé analýzy signálu. Nejčastěji používaným způsobem parametrického popisu řečového signálu jsou koeficienty **MFCC** (**mel-scale frequency cepstral coefficients**). Při nižších poměrech SNR (signal-to-noise ratio) však rozpoznávače založené na těchto koeficientech rychle ztrácí spolehlivost. V literatuře je popsáno velké množství alternativních příznakových reprezentací, přičemž MFCC jsou obvykle chápány jako výchozí varianta, s níž je srovnávána šumová odolnost dané alternativy. Typickým představitelem jsou koeficienty **PLP** (**perceptive linear predictive**) [8] založené na známých charakteristikách lidského sluchu (spektrální rozlišení kritických pásem, křivka stejné hlasitosti, vztah mezi hlasitostí a výkonem). Koeficienty **RASTA-PLP** (**relative spectra**) [9] využívají oproti PLP navíc filtrace modulačního spektra pro omezení aditivního i konvolučního šumu. Použitou filtrací se v každé spektrální složce eliminují časové změny vybočující mimo typický rámec projevů lidského hlasového traku. Tím lze dosáhnout až řádového snížení chybovosti při výrazném konvolučním zkreslení a taktéž významného omezení vlivu aditivního šumu.

Jako tradiční přístup pro potlačení aditivního šumu lze označit metody **spektrálního odčítání** v rozličných variantách [10][11]. Jsou založeny na odhadu průměrného šumového spektra a jeho následném odečtení od spektra zašuměné řeči. Tyto jednoduché metody jsou použitelné pouze pro stacionární aditivní šum nekorelovaný s řečovým signálem. Standardním postupem pro vykompenzování lineárního frekvenčního zkreslení přenosového kanálu jsou algoritmy normalizace kepstra [10] – **CMN** (**cepstral mean normalization**) a **CVN** (**cepstrum variance normalization**). Využívá se skutečnosti, že (v čase konstantní) charakteristika přenosového kanálu se v kepstru projeví přičtením konstantního vektoru ke kepstru

řeči. Mezi hlavní omezení patří požadavek na stacionaritu kanálu a minimální aditivní šum. Další skupinou přístupů pro získání čisté řeči ze zašuměné je použití adaptivních filtrů.

Nejvyšší účinnosti ve zvyšování šumové robustnosti rozpoznávačů dosahují metody zahrnující informaci o šumu do akustických modelů. Nejjednodušším přístupem k akustickému modelování zašuměné řeči je zahrnutí čisté řeči i řeči zarušené všemi předpokládanými typy šumu v předpokládaných poměrech SNR do trénovací množiny HMM rozpoznávače (**multi-condition training** [10]). Kvůli málo diskriminativním modelům je však smysluplné použití limitováno na malý rozsah variability šumu [12]. Pro zvýšení diskriminačních schopností modelů zašuměné řeči je možné natrénovat oddělené HMM pro různé typy a úrovně šumu, takže rozptyly uvnitř tříd se sníží (**multiple-model framework, MMF** [12]). Správnost výsledné funkce je silně závislá na pomocném šumovém klasifikátoru [12]. Jednou z nejpoužívanějších metod pro on-line adaptaci akustických modelů řeči na aktuální šumové podmínky je **MLLR** (**maximum likelihood linear regression**) [13]. Využívá trénovací nahrávku, pomocí které jsou upraveny vektory středních hodnot ve všech stavech akustického modelu.

Nejvyspělejší metody pro ošetření vlivu šumu na úrovni akustických modelů používají oddělené modely čisté řeči a šumu, ze kterých se při rozpoznávání vytváří modely zašuměné řeči přizpůsobené aktuálním podmínkám. Metoda **PMC** (**parallel model combination**) [14] transformuje parametry HMM čisté řeči a šumu z kepstrální do spektrální domény, kde se provádí jejich kombinace některou z existujících aproximací (log-normal, log-max, data-driven, atd.). Parametry kombinovaného modelu se pak zpětnou transformací vrací do kepstrální domény, ve které probíhá vlastní rozpoznávání. Používané aproximace zanedbávají korelaci řeči a šumu (tj. konvoluční šum). Na principu aproximace nelineárního vztahu řeči a šumu v kepstrální doméně pomocí Taylorova rozvoje je postavena metoda **VTS** (**vector Taylor series**) [10]. VTS oproti PMC umožňuje snížit nároky na výpočetní výkon spojené s modelováním zašuměné řeči, přičemž umožňuje zahrnout i lineární zkreslení přenosového kanálu.

2.2 ZMĚNY V ŘEČOVÉM PROJEVU MLUVČÍHO

Rozličné vnější a vnitřní vlivy mohou výrazně ovlivnit řečový projev mluvčího [15] (stres, únava, onemocnění, vliv léků, intoxikace atd.). Jakákoli změna v řečovém signálu, na kterou nebyl rozpoznávač řeči natrénován, snižuje úspěšnost rozpoznávání (neshoda akustických modelů). Současné komerční rozpoznávače s těmito jevy nepočítají, což snižuje jejich použitelnost v nestandardních podmínkách [10]. Mezi nejčastěji uváděné parametry řeči měnící se v důsledku změn stavu mluvčího patří [15][16]:

- **Intenzita hlasu, hlasové úsilí** – řečník zvyšuje intenzitu hlasu při zvýšené úrovni okolního hluku (Lombard efekt), při rozčilení a vysoké fyzické zátěži [15]. Naopak tišší řeč je typickým důsledkem stavu skleslosti a deprese [16].

- **Výška hlasu** – charakterizována frekvencí základního tónu f_0. Změny f_0 jsou velmi různorodé a závislé nejen na konkrétním vlivu, ale i na individuálním mluvčím [17]. Kromě fyziologických jevů se uplatňují i vědomé kompenzační mechanismy, které zvyšují nepředvídatelnost změn f_0.
- **Časové relace promluvy** – doba trvání slov, změny trvání jednotlivých fonémů, mezer mezi slovy a větami, průměrná rychlost řeči aj. Pomalá řeč je nejčastěji spojována s únavou a depresivními stavy [15][18]. Změny rozptylu trvání všech fonetických jednotek (fonémy, difony, slabiky, slova, fráze) jsou indikátorem pro stresové stavy [19]. Výrazné zvýšení rozptylu fonetických jednotek je typické pro stav rozhněvání [19]. Zvláštním časovým parametrem je doba reakce mluvčího na dotaz v rámci dialogu [18], která se mění např. v důsledku spánkové deprivace.
- **Spektrum hlasového traktu** – v důsledku psychického stavu mluvčího dochází ke změnám spektrální polohy a šířky formantů. Mění se též strmost poklesu energie směrem k vyšším frekvencím [15][19].
- **Průběh hlasivkového budícího signálu** – spolu s formanty určuje hlasivkový signál tvar spektra hlasového traktu.

Je tedy zřejmé, že vliv stresu a dalších faktorů ovlivňující řeč mluvčího je nutné vzít v úvahu, pokud má být zajištěna spolehlivá funkce rozpoznávače řeči za všech situací. V [15] jsou uvedeny výsledky experimentů s běžnými typy rozpoznávačů s využitím databáze SUSAS (Speech under simulated and actual stress) [19]. Testovaný rozpoznávač závislý na mluvčím v případě neutrální řeči fungoval s 88,3% úspěšností, průměrná úspěšnost pro všechny simulované změny hlasu byla 58%, nejhorší úspěšnost 20% vyšla pro řeč rozzlobeného mluvčího. V případě rozpoznávače nezávislého na mluvčím byl zaznamenán pokles úspěšnosti z 96% na 87% s nejhorším výsledkem 73%. Dopad změn hlasového úsilí mluvčích na úspěšnost systémů rozpoznávání řeči prozatím nebyl v celém rozsahu experimentálně zkoumán. Rozpoznávač nezávislý na mluvčím natrénovaný na normální řeči, který v případě testu na stejném typu řeči dosáhl úspěšnosti 80%, rozpoznával šepot s úspěšností pouze 40% [20]. Použitím MLLR (maximum likelihood linear regression) adaptace došlo k mírnému zlepšení (o 17% absolutně). U dalších typů změn řeči jako např. únavy či alkoholní intoxikace doposud nebyl dopad na úspěšnost automatického rozpoznávání řeči kvantifikován.

Konvenčním přístupem pro snížení vlivu změn v promluvě mluvčího na úspěšnost rozpoznávání je **multi-style training** [21]. Jedná se o metodu použitelnou pro rozpoznávač závislý na mluvčím, kdy mluvčí v tréninkové fázi simuluje větší množství emocí a akustický model jeho promluvy tak zahrnuje širší rozpětí možných realizací rozpoznávaných promluv. Nižší diskriminativnost vzniklých akustických modelů však omezuje dosažitelnou úspěšnost rozpoznávání. Aplikace tohoto přístupu na rozpoznávač nezávislý na mluvčím nebyla příliš úspěšná [22]. V [23] je popsán postup pro vygenerování modelů pozměněné řeči pomocí speciálního HMM-generátoru ze standardní neutrální řečové databáze, čímž odpadá

náročnost získávání dat při multi-style trénování, dosažitelné výsledky však nenabízejí výraznější zlepšení.

3 NAVRŽENÉ METODY ZVYŠOVÁNÍ ŠUMOVÉ ODOLNOSTI ROZPOZNÁVAČE

3.1 KUBICKÁ INTERPOLACE GMM

Jedním z možných postupů pro zvýšení šumové robustnosti rozpoznávače je použití sady akustických modelů natrénovaných na zašuměné řeči zvlášť pro několik odstupňovaných poměrů SNR (multiple-model framework, MMF [12]). Předpokladem úspěšné implementace je zachování charakteru šumu uvažovaného při tréninku i v testovacích podmínkách (tj. srovnatelný tvar spektrální hustoty výkonu, odpovídající statistické momenty 2. řádu). Při rozpoznávání se ovšem vyskytují i případy SNR v "mezerách" mezi natrénovanými hodnotami. Za tímto účelem Xu et al. [12] implementují lineární interpolaci hustoty pravděpodobnosti dvojice nejbližších modelů (uzlové body), čímž se získá aproximovaný model pro dané aktuální SNR. Lineární interpolace však nebere v úvahu globální trend daný ostatními uzlovými body a přesnost interpolace rapidně klesá se zvyšujícím se rozestupem uzlových bodů.

Pro odstranění těchto nedostatků jsem navrhl přístup využívající plynulého prokládání hustoty pravděpodobnosti ρ akustických modelů typu GMM pomocí Hermitova kubického splajnu [5]. Interpolované hodnoty hustoty pravděpodobnosti $\hat{\rho}(SNR)$ lze z uzlových bodů ρ na intervalu $\langle SNR_k, SNR_{k+1} \rangle$ získat [24]

$$\begin{aligned} \hat{\rho}(SNR) &= h_{00}(\tau)\rho_k + h_{10}(\tau)(SNR_{k+1} - SNR_k)d_k + \\ &+ h_{01}(\tau)\rho_{k+1} + h_{11}(\tau)(SNR_{k+1} - SNR_k)d_{k+1} \end{aligned}, \tag{3.1}$$

kde $\tau = (SNR - SNR_k) / (SNR_{k+1} - SNR_k)$ je normalizovaná nezávisle proměnná s ohledem na rozestup prokládaných uzlových bodů, d jsou tangenty v uzlových bodech a h jsou bázové funkce

$$\begin{aligned} h_{00}(\tau) &= 2\tau^3 - 3\tau^2 + 1, \\ h_{10}(\tau) &= \tau^3 - 2\tau^2 + \tau, \\ h_{01}(\tau) &= -2\tau^3 + 3\tau^2, \\ h_{11}(\tau) &= \tau^3 - \tau^2. \end{aligned} \tag{3.2}$$

Vhodným způsobem určování tangent lze zajistit, že výsledná interpolace neosciluje, zachovává monotónnost při monotónních SNR_k, kopíruje lokální maxima a minima prokládané funkce a $\hat{\rho}$ na žádném z úseků nevybočí z intervalu $\langle \rho_k, \rho_{k+1} \rangle$. Tyto vlastnosti jsou důležité, aby hustota prokládaných GMM plynule přecházela mezi jednotlivými uzlovými body zachovávaje globální trend bez vzniku abnormálních výkyvů. Výpočet tangent se řídí pravidlem [24]

$$d_k = \begin{cases} 0 & \text{pro } \operatorname{sign}\delta_k \neq \operatorname{sign}\delta_{k-1} \\ 0 & \text{pro } \delta_k = 0 \text{ nebo } \delta_{k-1} = 0 \\ \dfrac{1}{d_k} = \dfrac{1}{2}\left(\dfrac{1}{\delta_{k-1}} + \dfrac{1}{\delta_k}\right) & \text{pro } \forall \text{ ostatní } \{\delta_k, \delta_{k-1}\} \end{cases}, \tag{3.3}$$

kde

$$\delta_k = \frac{\rho_{k+1} - \rho_k}{SNR_{k+1} - SNR_k}. \tag{3.4}$$

Pro praktické ověření uvedené metody jsem provedl experimenty s detektorem hranic slov využívajícím dva bayesovské klasifikátory, každý se dvěma GMM o 30 gaussovkách natrénovaný na 32 ms segmentech s 16 ms překryvem:

1) klasifikátor rozlišující třídy „řeč“ a „ticho“ s příznakovým vektorem tvořeným 19 MFCC koeficienty a 38 dynamickými koeficienty (19 delta, 19 akceleračních) získaných derivováním interpolačního polynomu 3. řádu, který je určen pro každou dimenzi MFCC z 5 sousedních segmentů,
2) klasifikátor rozlišující třídy „začátek slova“ a „konec slova“ s příznakovým vektorem tvořeným koeficienty interpolačního polynomu 3. řádu určeného z 19 MFCC koeficientů v 5 sousedních segmentech a střední hodnotou rozdílu hustoty pravděpodobnosti GMM „řeč“ a GMM „ticho“ v 5 sousedních segmentech.

Druhý z uvedených klasifikátorů byl trénován specificky jen na segmentech signálu odpovídajících označeným okamžikům začátku a konce slov. Pokusy byly realizovány s řečovou databází německých slov (viz kap. 4.1.1 v disertaci), jako aditivní šum byl použit AWGN (additive white Gaussian noise), všechna slova byla výkonově normalizována. Bylo natrénováno celkem 6 skupin modelů odpovídajících $SNR \in \{0,3,6,9,12,15\}$. Hustoty pravděpodobnosti každého ze 4 natrénovaných GMM pro dané SNR tvořily uzlové body kubické interpolace. Interpolací byly získány akustické modely v podstatně jemnějším odstupňování SNR.

Výsledky testovaného klasifikátoru byly hodnoceny z pohledu schopnosti přesně určit pozici začátků a konců slov. Bylo ověřeno, že interpolované GMM konzistentně kopírují trend daný uzlovými GMM. Výhodou uvedené metody je, že stačí uchovávat mnohonásobně menší množství parametrů klasifikátoru při výsledné přesnosti modelů srovnatelné s mnohem hustší škálou GMM. Uvedený princip je možné rozšířit i na komplexnější akustické modely typu HMM, kde by se interpolace aplikovala zvlášť na GMM v rámci každého ze stavů HMM. Pro praktické použití by bylo nutné navíc použít detektor aktuální hodnoty SNR, který by na základě pozorovaného vstupního signálu vybíral požadovaný typ akustických modelů. Nevýhodou této metody (a celého MMF) je nutnost zaručit shodný

charakter aditivního šumu v trénovacích a testovacích podmínkách, tento šum musí mít navíc stacionární charakter.

3.2 PMC S VÍCESTAVOVÝM ERGODICKÝM HMM ŠUMU

Spolehlivá funkce rozpoznávače řeči v prostředí operačního sálu vyžaduje, aby byl systém pro toto specifické prostředí co nejlépe adaptován. Za tímto účelem byl pořízen kompletní zvukový záznam neurochirurgické operace na operačním sále Uniklinikum Marburg v Německu. Z analýzy získaného záznamu bylo patrné, že zvukové pozadí v tomto prostředí je značně různorodé a pro úspěšné rozpoznávání nestačí jen přičíst náhodně vybraný úsek šumu ke vzorkům řeči a takto natrénovat řečové modely. V další práci byl tedy pro účely modelování zašuměné řeči použit přístup založený na oddělených modelech čisté řeči a šumu využívající paralelní kombinace modelů (PMC). Díky dostupnému záznamu bylo možné natrénovat podrobný akustický model šumu.

Oproti řečovým HMM není při konstrukci HMM určeného pro modelování šumu možno vycházet z předem známých pravidelností ve struktuře signálu (fonémy, difony apod.). Inicializace HMM tedy musí být řízená trénovacími daty a má velký význam pro kvalitu výsledných modelů. Obvyklým přístupem pro konstrukci šumových HMM je daty řízené shlukování všech vzorků parametrizovaného signálu[1] a následná inicializace parametrů HMM na základě maximum likelihood (ML) estimace vycházející z dat přiřazených jednotlivým shlukům. Počet shluků se volí shodný se zamýšleným počtem stavů šumového HMM. Po inicializaci následuje reestimace parametrů HMM Baum-Welchovým (BW) algoritmem [25]. BW ovšem dokáže najít pouze lokální optimum; je tedy nutné zajistit, aby inicializované parametry byly dostatečně blízko optimálním. Při obvyklém inicializačním postupu však nejsou zohledněny lokální časové souvislosti v signálu, neboť při shlukování se vychází pouze ze vzájemné podobnosti vzorků signálu a nebere se v úvahu jejich pozice v trénovací nahrávce. Veškerou informaci o časových souvislostech pak musí podchytit matice pravděpodobností přechodů, která však vzhledem k omezenému počtu stavů šumového HMM poskytuje jen velmi hrubý popis. V rámci řešení tohoto problému jsem zkoumal možnosti segmentace šumové nahrávky na základě lokálně stacionárních úseků [4], takže jednotlivé stavy HMM jsou trénovány na množině úseků signálu lépe splňujících teoretický požadavek na stacionaritu signálu v rámci daného stavu HMM. Výstupní rozložení pravděpodobnosti jednotlivých stavů HMM následně věrněji reprezentují skutečné rozptyly v rámci úseků signálu připadajících na dané stavy HMM.

Pro účely identifikace statisticky konzistentních úseků jsem využil bayesovského informačního kritéria (Bayes information criterion, BIC [26]). BIC představuje nástroj pro optimální výběr statistických modelů na základě pozorovaných dat. Při aplikaci tohoto přístupu na určování hranic segmentů signálu se provádí statistický test hypotézy, zda je daný úsek signálu lépe charakterizován jedním normálním

[1] Vzorkem parametrizovaného signálu je jeden příznakový vektor určený metodami krátkodobé analýzy z nejkratšího uvažovaného úseku signálu o obvyklé délce 10-30 ms

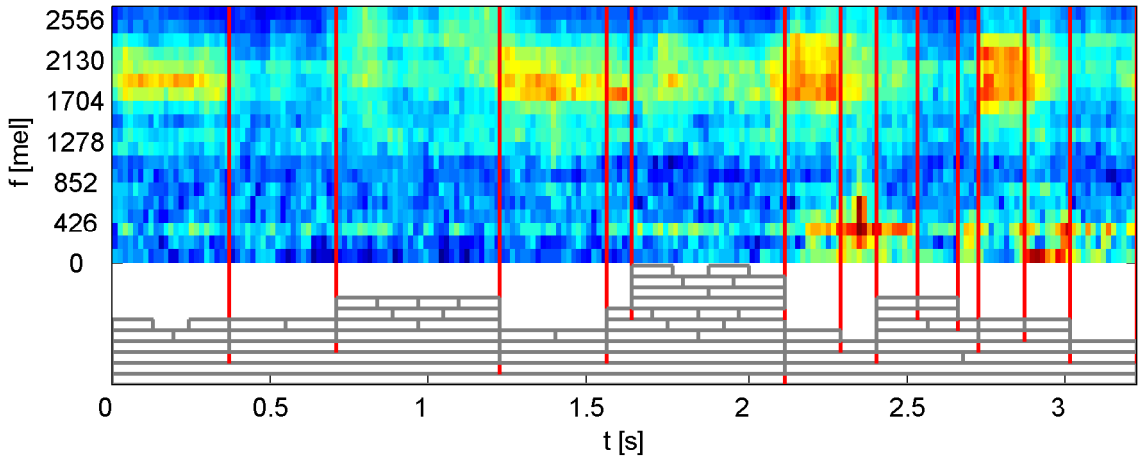
rozložením nebo zda je vhodnější jej rozdělit na dvojici navazujících podúseků, každý popsaný vlastním normálním rozložením. Při popisu signálu jsem vycházel z parametrizované verze definované sekvencí vektorů (každý o 19 MFCC koeficientech). Příslušné normální rozložení je tedy určeno 19-prvkovým vektorem středních hodnot a kovarianční maticí o rozměru 19×19. Máme-li úsek parametrizovaného signálu začínajícího v čase a, končícího v c a majícího možnou hranici na pozici b, je nutné pro tento úsek spočítat ΔBIC skóre [26]

$$\Delta\,\mathrm{BIC}(a,b,c) = \frac{1}{2}(c-a)\log\left|\boldsymbol{\Sigma}^{ac}\right| -$$

$$(b-a)\log\left|\boldsymbol{\Sigma}^{ab}\right| - (c-b)\log\left|\boldsymbol{\Sigma}^{bc}\right| - \frac{1}{2}\eta\left(N + \frac{1}{2}N(N-1)\right)\log(c-a), \qquad (3.5)$$

kde $\boldsymbol{\Sigma}^{\mathbf{ac}}$ je kovarianční matice normálního rozložení přiřazeného na základě ML odhadu intervalu (a,c), $\boldsymbol{\Sigma}^{\mathbf{ab}}$ je analogicky přiřazena intervalu (a,b) a $\boldsymbol{\Sigma}^{\mathbf{bc}}$ odpovídá intervalu (b,c). N je dimenze použitého příznakového vektoru. Koeficient η určuje míru detailů výsledné segmentace. Pokud je pro některé b skóre ΔBIC(a,b,c) větší než nula, nachází se na pozici b hledaná hranice v rámci intervalu (a,c).

Přestože BIC podává optimální rozhodnutí v daném intervalu (a,c), je potřebná vhodná strategie volby těchto intervalů, aby výsledná segmentace byla i z vyšší perspektivy smysluplná. Shaobing a Gopalakrishnan [26] zvolili sekvenční úpravy hranic intervalů, kde a začíná v nulovém čase a c se postupně zvětšuje; při nalezení hranice na pozici b se přiřadí $b \rightarrow a$ a od tohoto bodu se začíná se zvětšováním intervalu (a,c) opět od minima. Praktické experimenty s touto strategií však odhalily, že nalezené hranice často označovaly nepodstatné lokální detaily a mnohé významnější předěly byly ignorovány. V [4] jsem proto navrhl robustnější přístup založený na globální segmentaci celé nahrávky stromovým způsobem s větvením určeným pozicemi dosavadně nalezených hranic. Tento algoritmus je vhodný pro off-line segmentaci; z důvodu kauzality jej nelze aplikovat pro on-line určování hranic z postupně získávaných dat. Na obr. 3.1 je znázorněn výsledek běhu algoritmu na úseku šumového signálu z akustického pozadí operačního sálu. Spektrogram signálu ukazuje vývoj spektrálních vlastností šumu přes několik sekund, frekvenční osa je uvedena v logaritmické melové škále. Šedé obdélníky znázorňují jednotlivé testované intervaly (a,c) při běhu algoritmu. Svislé červené čáry pak označují nalezené hranice v daných úsecích. Není-li v daném úseku hranice nalezena, úsek se v následných iteracích dělí na menší části, v rámci nichž se opakovaně provádí BIC test až po minimální délku úseku. Minimální délka testovaného úseku je omezena konstantou 2ε, která zajišťuje potřebný objem vzorků pro smysluplný výpočet lokální statistiky úseku signálu.



Obr. 3.1 Spektrogram zvuků na operačním sále s vizualizací segmentačního algoritmu

Segmenty označené BIC segmentací jsou v dalším zpracování brány jako nedělitelné jednotky. Jejich shlukování do stavů výsledného HMM je analogické obvyklému postupu aplikovanému na vzorky parametrizovaného signálu. Oproti případu bez segmentace je však nutno počítat s menším absolutním počtem přechodů mezi stavy v rámci trénovacích dat, pro odhad matice přechodů a apriorních pravděpodobností HMM byl proto místo ML zvolen vhodnější odhad expected likelihood estimation (ELE) [27]. ELE narozdíl od ML předpokládá konečný počet pozorování při odhadu pravděpodobností, takže i řídce pozorovaným jevům je přiřazena nenulová pravděpodobnost.

Pro nalezení optimálního způsobu inicializace šumového HMM jsem provedl sadu srovnávacích experimentů zahrnujících shlukovací metody *k*-means a hierarchical agglomerative clustering (HAC) v několika modifikacích [3]. Bylo porovnáno jak použití BIC segmentace, tak přímé shlukování všech vzorků šumu. Navržený postup využívající BIC, ELE a *k*-means ve spojení se split-merge EM (expectation maximization) [28] estimací parametrů GMM v jednotlivých stavech (bez Baum-Welch reestimace) se ukázal být z testovaných přístupů nejlepší, tj. výsledný šumový HMM vykazoval maximální aposteriorní pravděpodobnost pro zvukovou nahrávku operačního sálu.

Reálnou použitelnost šumového modelu určeného BIC/ELE/*k*-means postupem jsem ověřil testem rozpoznávání izolovaných slov na pozadí hluku operačního sálu [4]. Byla použita databáze 8 mluvčích, jejichž hlasové vzorky pro 12 německých číslovek jsem shromáždil při mé stáži na Hochschule RheinMain ve Wiesbadenu v Německu. Celkový počet slov ve všech variantách byl 960; polovina tohoto množství byla použita pro trénování celoslovních levo-pravých HMM, druhá polovina slov byla následně zarušena aditivním přidáním nahrávky akustického pozadí operačního sálu a použita pro testování. Zvuková data byla reprezentována 19 MFCC koeficienty získávanými z 32 ms rámců s 10 ms posuvem; data byla vzorkována frekvencí 16 kHz. V testovací nahrávce byla jednotlivá slova oddělena

přibližně sekundovými intervaly, v celé nahrávce byl přítomen hluk operačního sálu s poměrem SNR odstupňovaným v rozsahu 3 až 18 dB.

Počet stavů slovních HMM byl určen na základě průměrné délky daného slova v rámci všech trénovacích variant a pohyboval se v rozsahu 7 až 9. Výstupní rozložení ve všech stavech bylo tvořeno GMM se 7 komponenty. Model šumu byl tvořen desetistavovým ergodickým HMM s počtem gaussovek v GMM jednotlivých stavů určených split-merge EM algoritmem v rozsahu 4 až 12. Při rozpoznávání byl v každém časovém kroku metodou PMC vytvářen kombinovaný model zašuměné řeči, který zahrnoval všechny slovní HMM a taktéž všechny stavy šumového modelu. Pravděpodobnosti přechodů mezi stavy modelu zašuměné řeči byly spočítány z příslušných pravděpodobností modelů čisté řeči a šumu a předem stanovené apriorní pravděpodobnosti odmlk mezi slovy. Při výpočtu se předpokládalo, že šum a řeč jsou statisticky nezávislé procesy.

Dosažené chybovosti WER jsou uvedeny v tab. 3.1. Rozpoznávač byl trénován na malé skupině mluvčích, jejichž řeč byla použita i při testování. Jedná se tedy principiálně o rozpoznávač na pomezí mezi systémem závislým a nezávislým na mluvčím. Pokud by bylo provedeno přizpůsobení na jediného mluvčího, lze očekávat další zvýšení úspěšnosti. Pro porovnání byla stejná testovací nahrávka též rozpoznávána pomocí rozpoznávače využívajícího pouze HMM čisté řeči (tj. bez přizpůsobení šumu). V tomto případě musel být pro dosažení obdobné chybovosti SNR o cca 30 dB vyšší, což dokumentuje vysokou robustnost navrženého PMC rozpoznávače.

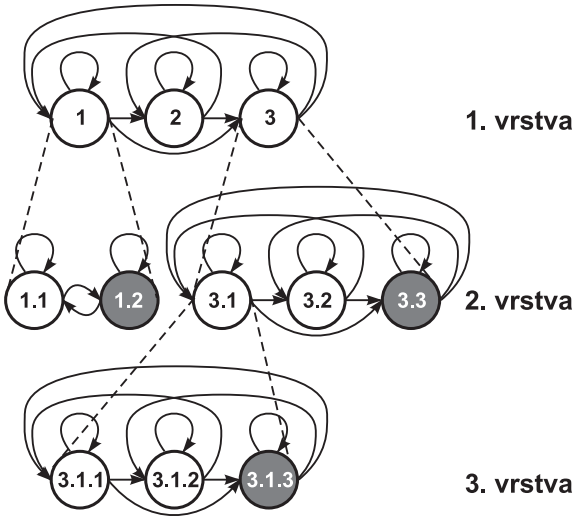
Tab. 3.1 WER [%] PMC rozpoznávače izolovaných slov s rušivým akustickým pozadím operačního sálu

SNR [dB]	**WER [%]**	
	Trénovací množina	Testovací množina
3	8,1	12,9
6	5,3	10,2
12	1,9	7,1
18	1,4	4,2

3.3 POUŽITÍ HIERARCHIE HMM PRO MODELOVÁNÍ ŠUMU

Zvyšování šumové robustnosti rozpoznávače je možné dosáhnout podrobnějším akustickým modelem šumu. Při použití metod paralelní kombinace modelů však s narůstajícím počtem stavů šumového HMM prudce rostou výpočetní nároky, neboť se násobně zvětšuje počet stavů výsledného HMM zašuměné řeči. V článku [1] jsem popsal inovativní strukturu šumového modelu, která využívá struktury klasifikačního stromu pro urychlení výpočtu modelu zašuměné řeči.

Klasifikační stromy jsou známy svou schopností poskytnout rychlá rozhodnutí o komplexních strukturách, čehož je dosaženo díky hierarchickému uspořádání rozhodovacího stromu. S každým rozhodnutím se tak zužuje množství zbývajících tříd, mezi kterými je potřeba rozhodnout. West a Cox [29] využili hierarchické struktury GMM modelů pro klasifikaci hudebních signálů. Výsledný model poskytl lepší výsledky než obvyklá plochá struktura porovnávající naráz všechny třídy. V [1] jsem popsal rozšíření tohoto přístupu na vícestavové HMM (místo jednoduchých GMM), kde všechny stavy každého HMM slouží jako větvící body. Výsledná struktura hierarchicky svázané sady HMM připomíná standardní hierarchický skrytý Markovův model (hierarchical hidden Markov model, HHMM [30]), ovšem zavedení principu klasifikačního stromu výrazně omezuje počet stavů, který je při rozpoznávání nutno vyhodnotit v každém časovém kroku. Zvýšení rychlosti zpracování za cenu jednoduššího vyhodnocení pochopitelně zavádí určitý kompromis v přesnosti akustického modelu. Obr. 3.2 ilustruje příklad navrženého stromu HMM.



Obr. 3.2 Struktura hierarchického modelu šumu

Všechny uzly v celém navrženém stromu jsou emitující (oproti tomu u HHMM bývá obvykle emitující jen nejnižší vrstva). Kořen stromu (1. vrstva) na obr. 3.2 je tvořen jediným třístavovým HMM. Každý ze stavů modelu v 1. vrstvě dále expanduje do samostatného HMM ve 2. vrstvě. Bíle označené stavy z 2. vrstvy mají přiřazeny další samostatné HMM ve 3. vrstvě. Tmavě označené uzly jsou speciální neexpandující stavy a je jim přiřazeno výstupní rozložení totožné s uzlem v nadřazené vrstvě. Přechod do tohoto stavu při rozpoznávání tak omezí hloubku zanoření a fakticky určuje, jak obecný model (ze všech dostupných v dané linii stromu) nejlépe vystihuje aktuálně rozpoznávaná data. Pokud tedy rozpoznávaná data budou značně odlišná od trénovací nahrávky, šumový model na základě maxima aposteriorní pravděpodobnosti automaticky zvolí obecnější HMM. Pokud naopak bude daný šumový model v dobré shodě s rozpoznávaným signálem, budou

použity nejpodrobnější HMM v nejnižších vrstvách. Tímto je zajištěna škálovatelnost modelu zaručující na základě Bayesova pravidla optimální volbu modelů z dostupné sady.

V hierarchické struktuře klasifikačního stromu jsou informace o modelovaném akustickém šumu podchyceny s rozdílným stupněm obecnosti v závislosti na jednotlivých vrstvách. HMM v první vrstvě zachycuje nejobecnější charakter signálu a přechody mezi stavy v této vrstvě jsou málo časté. S každou další vrstvou se obsažené HMM stávají více specializované na lokální prvky v šumovém signálu. Při sestavování klasifikačního stromu musí být tento princip brán v úvahu a správné přiřazení trénovacích dat jednotlivým stavům HMM je tudíž zásadní pro dobrou funkci rozpoznávače. Při formulaci algoritmu pro určování struktury stromu na základě dat jsem vycházel z BIC segmentace. Hierarchická struktura navrženého segmentačního algoritmu umožňuje snadné přiřazení dat jednotlivým stavům HMM se zvolenou průměrnou dobou trvání na základě horizontálního řezu v potřebné hloubce segmentační hierarchie. Při přiřazování dat stavům v 2. a dalších vrstvách je vždy nutné vycházet pouze z úseků signálu přiřazených nadřazenému stavu ve vyšší vrstvě. Tímto je zaručeno, že každá nižší část klasifikačního stromu je zobecněna ve vyšších částech. Díky tomu má při rozpoznávání smysl postupovat od shora dolů a prohledávat jen ty HMM, které vycházejí z již určených stavů ve vyšších vrstvách.

Při rozpoznávání řeči s využitím uvedeného hierarchického modelu šumu se vychází z paralelní kombinace modelů čisté řeči a šumu. Rozpoznávač generuje složený HMM zašuměné řeči spojením všech slovních HMM a určeného HMM šumu v aktuální pozici klasifikačního stromu. Tento složený model obsahuje všechny kombinace řečových a šumových stavů včetně samotných stavů šumu modelující pomlky mezi slovy. Oproti systému s jediným šumovým HMM jsou však v tomto uspořádání dynamicky vyměňovány šumové části kombinovaného modelu na základě aktuální pozice v hierarchii klasifikačního stromu. Finálním výsledkem rozpoznávání zašuměné řeči je sekvence stavů řečových HMM (a tomu odpovídající rozpoznaná slova). Řečová složka se určuje z Viterbiho dekódování zašuměné řeči spojené s poslední vrstvou v hierarchii šumového modelu.

V tab. 3.2 jsou uvedeny hodnoty negativního log-likelihoodu po Viterbiho dekódování při aplikaci samotného šumového modelu (tj. bez kombinace s řečovými HMM) na rozpoznávání nahrávky samotného šumu. Test byl proveden pro rozsah počtu gaussovek $1-40$ v GMM výstupních rozložení HMM. Je patrné, že likelihood třívrstvého modelu je podstatně lepší než v případě standardního HMM v ploché struktuře s ekvivalentním počtem stavů. Výsledky testu rozpoznávání zašuměné řeči pro rozsah SNR 3 až 18 dB jsou pak uvedeny v

tab. 3.3. Tabulka uvádí tři sady experimentů, kdy byl postupně zvyšován počet vrstev šumového modelu. Hierarchický třívrstvý model šumu použitý při experimentech měl nastaveno ve všech vrstvách maximum 5 stavů, skutečný počet v jednotlivých vrstvách byl však v průběhu konstrukce modelu automaticky zmenšován s ohledem na objem dostupných trénovacích dat. Je patrné, že komplexnější model umožňuje nižší chybovost, dosažitelné přírůstky

rozpoznávací schopnosti zároveň však s narůstajícím počtem vrstev klesají. Při konstrukci rozpoznávače tedy musí být zvolen vhodný počet vrstev šumového modelu s ohledem na dostupná trénovací data a tolerovatelnou výpočetní zátěž při rozpoznávání. Se zvyšujícím se počtem vrstev též narůstají požadavky na přesnost použité aproximace PMC (při experimentech byla použita pouze aproximace log-normal).

Tab. 3.2 Shoda šumového modelu s testovací částí šumové nahrávky

Max. počet gaussovek v GMM		1	5	10	15	20	25	30	35	40
Negativní log-likelihood (nižší je lepší)	vrstva 1	11,82	10,95	10,64	10,53	10,45	10,41	10,38	10,39	10,37
	vrstva 1+2	10,84	9,98	9,83	9,78	9,78	9,77	9,78	9,76	9,77
	vrstva 1+2+3	9,89	9,36	9,31	9,28	9,26	9,25	9,26	9,27	9,26
Plochý HMM s 15 stavy		10,92	10,07	9,93	9,81	9,90	9,86	9,85	9,85	9,84

Tab. 3.3 Výsledky rozpoznávání zašuměných slov s hierarchickým modelem šumu

SNR (dB)		3	6	9	12	18
WER (%)	vrstva 1	16,8	12,9	11,4	8,8	6,0
	vrstva 1+2	14,4	11,0	9,7	8,2	4,8
	vrstva 1+2+3	14,4	10,0	8,4	7,8	4,7

4 NAVRŽENÉ METODY ZVYŠOVÁNÍ ODOLNOSTI ROZPOZNÁVAČE VŮČI ZMĚNÁM V PROMLUVĚ MLUVČÍHO

Při snaze o zvýšení odolnosti rozpoznávače vůči změnám v hlase mluvčích jsem se zaměřil na úroveň hlasového úsilí, které bylo v dosavadní literatuře věnováno minimum pozornosti a přitom jde o jeden z nejčastějších typů modifikace hlasu. Zavedení odolnosti rozpoznávačů vůči celé škále změn v hlasovém úsilí má velký praktický význam a může významně přispět ke zvýšení použitelnosti technologie rozpoznávání řeči v oblastech, kde to dodnes nebylo možné.

4.1 DATABÁZE HLASOVÉHO ÚSILÍ BUT-VE1

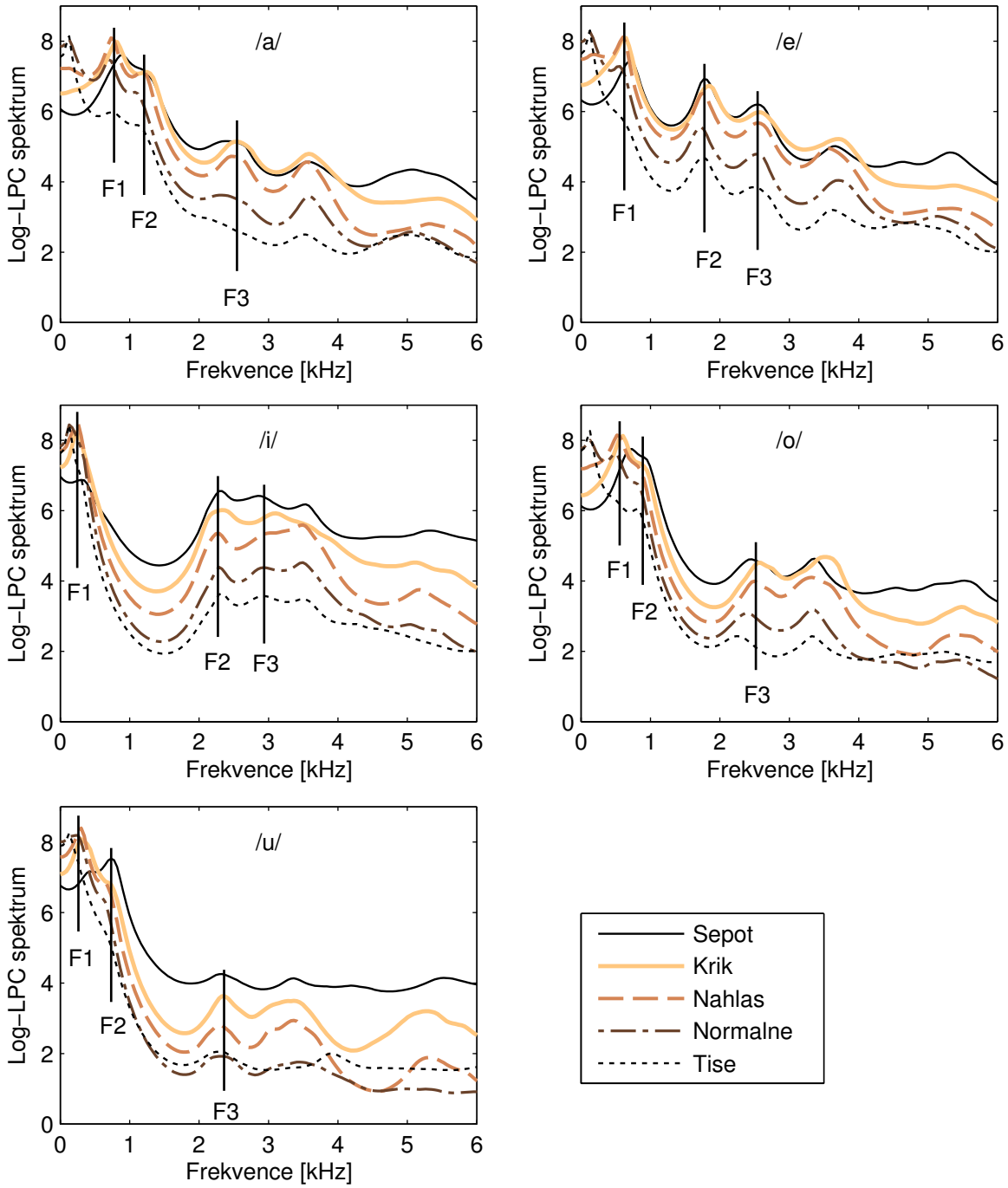
V současnosti neexistuje žádná komerčně dostupná řečová databáze pokrývající celou škálu hlasového úsilí, na které by bylo možno provést rozsáhlejší experimenty s rozpoznáváním řeči. Přistoupil jsem tedy ke shromáždění statisticky významného množství řečového materiálu od více mluvčích v kompletním rozsahu hlasového úsilí s využitím špičkové aparatury a v kontrolovaných podmínkách. Výsledkem je řečová databáze izolovaných slov **BUT-VE1**. Sestavování databáze probíhalo ve více etapách v době od září do prosince 2010 s následným zpracováním (extrakce,

fonetická segmentace, kompletace) až do března 2011. Všechny promluvy byly zaznamenány 2 mikrofony. Mikrofon Brüel & Kjaer 4189-B-001 byl použit pro záznam ve vzdálenosti 1 m od úst mluvčího v přímém směru, mikrofon Shure PG30 byl použit pro záznam ze vzdálenosti 5 cm od koutku úst (náhlavní mikrofon). Veškeré záznamy jsou kalibrovány pro možnost přesně určit hladinu akustického tlaku (sound pressure level) L_p v libovolné části signálu. Hodnoty L_p jsou vztaženy k pozici 1 m od úst mluvčího v přímém směru. Zvuková data jsou uložena ve formátu stereo PCM (pulse-code modulation) se vzorkovací frekvencí 44,1 kHz a 16 bitovým kvantováním. Levý kanál obsahuje záznam z náhlavního mikrofonu, pravý kanál ze vzdáleného mikrofonu. Oba kanály jsou časově synchronizovány, ovšem v pravém kanále je signál o 2 ms zpožděn, což je přirozený důsledek doby šíření akustické vlny na vzdálenost 1 m ke vzdálenému mikrofonu. Nahrávání proběhlo v izolované bezdozvukové komoře na Ústavu telekomunikací VUT v Brně.

Databáze umožňuje trénink a testování rozpoznávačů závislých i nezávislých na mluvčím pro úrovně hlasu od šepotu přes tichý hlas, normální promluvu, zvýšený hlas až po křik. Celkem bylo zaznamenáno přes 20 tisíc slov od 13 mluvčích, každý mluvčí namluvil 3 bloky slov s 10 opakováními, dále sadu typických příkazů a úsek plynulého textu. Všechny tyto promluvy byly zopakovány pro každý z pěti módů hlasového úsilí. Pro každého mluvčího byl též nahrán úsek neartikulovaného hlasového signálu (hláska /á/) s odstupňovanou úrovní měřeného L_p počínaje 50 dB po 5 dB až do 85 dB. Celkem je od každého mluvčího v každém hlasovém módu k dispozici cca 19 minut záznamu, celkový rozsah databáze přesahuje 4 hodiny. Pro každé zaznamenané slovo byla provedena fonetická segmentace, fonetické hranice jsou stanoveny s časovým rozlišením 12 ms.

Celý obsah databáze je k dispozici ve dvou variantách – originální a s redukovaným šumem. Pro redukci šumu byla použita metoda odčítání spektra vycházející ze vzorku samotného šumu bez obsahu řeči. Redukce šumu byla provedena nezávisle pro levý a pravý kanál. Použitý algoritmus nezpůsobuje degradaci řeči a nevnáší slyšitelný muzikální šum. V případě pravého kanálu (vzdálený mikrofon) je tak dosaženo snížení úrovně šumu výsledného signálu z ekvivalentní hodnoty L_p 44,5 dB na 10,6 dB, ve které je zahrnut i kvantizační šum daný 16-bitovým rozlišením. Díky tomu je i pro šepot dosaženo odstupu průměrného výkonu signálu od výkonu šumu lepšího než 20 dB. Levý kanál díky menší vzdálenosti mikrofonu od úst mluvčího dosahuje ještě podstatně vyššího SNR.

S využitím BUT-VE1 byly zkoumány akustické a fonetické změny v hlasovém signálu v závislosti na úrovni hlasového úsilí. Pro účely zvyšování úspěšnosti rozpoznávání řeči byly zejména zajímavé statistické parametry řeči společné pro všechny mluvčí. Průměrná log-LPC (linear predictive coding) spektra samohlásek na obr. 4.1 dokumentují výrazné odchylky tvaru spekter mezi jednotlivými hlasovými módy.



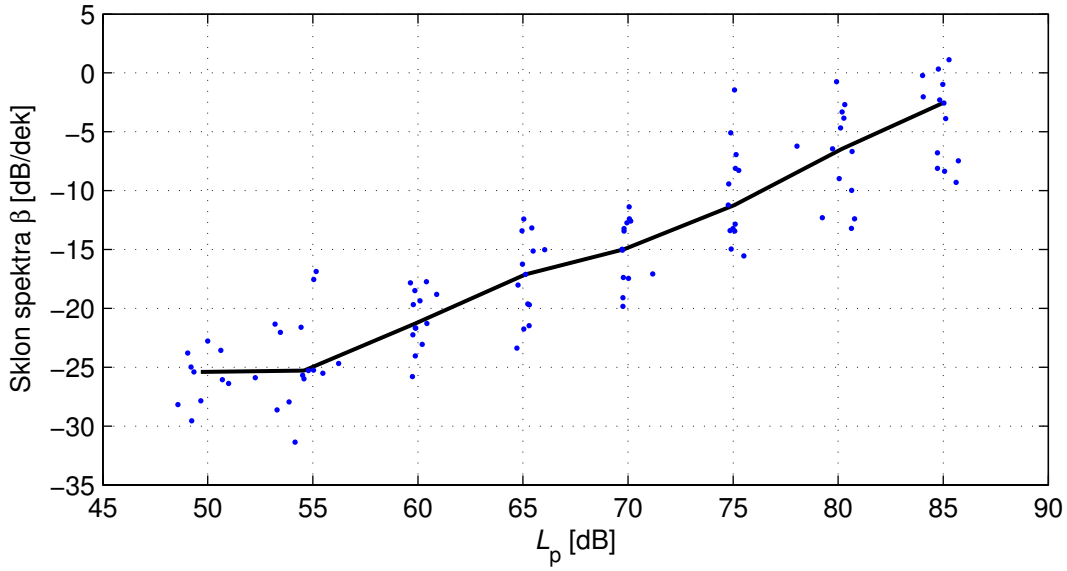
Obr. 4.1 Log-LPC spektra samohlásek s normalizovaným výkonem v pěti hlasových módech zprůměrovaná pro všechny mluvčí s vyznačenými prvními třemi formanty

Všechna uvedená spektra odpovídají signálu se stejnou energií (tj. po normalizaci). V některých samohláskách lze pozorovat menší změny v pozicích formantů, daleko výraznější je ale velký rozdíl úrovně u vyšších formantů. Důvodem je nárůst obsahu harmonických složek v budícím hlasivkovém signálu směrem k vyšším hlasovým intenzitám [31]. Tyto změny lze charakterizovat koeficientem sklonu spektra β určeném jako směrnici lineární regrese grafu

výkonového spektra v pásmu $0-4000$ Hz s oběma osami v logaritmické škále. Rovnice regresní přímky má tvar

$$10\log_{10}\left|S_f\right|^2 = \alpha + \beta \log_{10} f \ , \tag{4.1}$$

kde S_f je vzorek komplexního spektra řeči odpovídající frekvenci f. Výpočet spektra využívá FFT o délce 1024 vzorků, tj. při vzorkovací frekvenci 44,1 kHz se použije prvních 93 spektrálních složek. Pro určení β je použita metoda estimace nejmenších čtverců. Jednotkou β je dB/dekádu. Horní frekvenční limit byl zvolen 4 kHz, aby byl zahrnut průměrný spektrální rozsah hlasivkového excitačního signálu. V BUT-VE1 je k dispozici pro všechny mluvčí sada úseků neartikulované řeči (hláska /á/) s odstupňovanou hladinou akustického tlaku od 50 po 5 do 85 dB. Mediánovou filtrací β přes všechny mluvčí byla získána křivka závislosti β na L_p viz obr. 4.2. Jednotlivé body v grafu představují vzorky β pro jednoho mluvčího určené průměrem z cca 1 s záznamu analyzovaného v blocích po 24 ms s překryvem 12 ms. Graf má podobný průběh jako u příbuzného koeficientu „spectral balance“ uváděného v [31]; při dosahovaných úrovních hlasového úsilí však nebyl pozorován efekt saturace udávaný v [31] pro L_p nad 97 dB v 30 cm vzdálenosti (průměrná hodnota pro mužské hlasy). Tomu by při vzdálenosti 1 m odpovídala přibližná hodnota L_p 87 dB, což přesahuje rozsah použitý při vytváření databáze BUT-VE1.



Obr. 4.2 Závislost sklonu spektra na hladině akustického tlaku pro neartikulovaný hlas

4.2 KLASIFIKÁTOR HLASOVÉHO ÚSILÍ NEZÁVISLÝ NA SLOVNÍKU

Prvním krokem k umožnění spolehlivého rozpoznávání řeči ve všech hlasových módech je zavedení schopnosti rozpoznávače určit, jakým hlasem mluvčí momentálně mluví. Pro tento účel jsem vytvořil klasifikátor hlasového módu

pracující nezávisle na slovníku, který je možné předřadit před rozpoznávač řeči a umožnit tak rozpoznávači správně reagovat na měnící se hlas mluvčího.

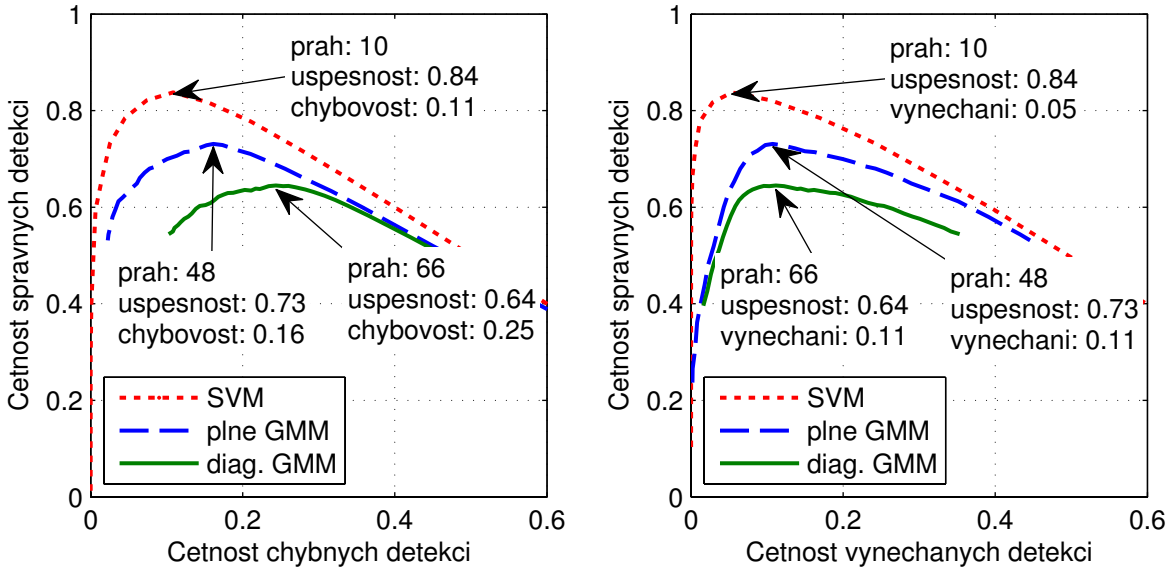
Z analýzy fonetických vlastností hlasu v různých módech vyplývá, že spektrum samohlásek lze použít jako spolehlivý zdroj informace pro určení úrovně hlasového úsilí. Je tedy nutné v rozpoznávané promluvě identifikovat samohlásky a následně pomocí jejich spektra rozhodnout o aktuálním hlasovém úsilí mluvčího. Jako příznaková reprezentace pro klasifikátor byly použity MFCC koeficienty. Při funkci klasifikátoru jsou tak periodicky analyzovány 24 ms segmenty příchozího signálu s 12 ms překryvem. Pokud klasifikátor vyhodnotí daný segment jako jednu ze samohlásek v jednom z 5 hlasových módů, je indikována příslušná detekce. Pro ostatní segmenty klasifikátor neposkytuje žádný výstup.

Byly testovány tři varianty klasifikátorů hlasového módu:

1) Bayesovský klasifikátor vycházející z GMM modelů jednotlivých samohlásek zvlášť pro každý hlasový mód s diagonálními kovariančními maticemi,
2) varianta bayesovského klasifikátoru s plnými kovariančními maticemi,
3) klasifikátor typu multi-class SVM (support vector machine) s oddělenými třídami pro jednotlivé hlasové módy zahrnující všechny samohlásky se zvlášť definovanou obecnou třídou pro celý zbylý příznakový prostor (zahrnuje ticho a ostatní fonémy).

Při tréninku klasifikátorů se vycházelo z úseků signálu v databázi BUT-VE1 obsahujících příslušné samohlásky. Počet gaussovek rozložení GMM u bayesovských klasifikátorů byl stanoven automaticky při estimaci parametrů algoritmem split-merge EM a pohyboval se v rozsahu 2 až 10 pro diagonální kovariance a 2 až 8 v případě plných kovariančních matic. Každé samohlásce v každém hlasovém módu je přiřazen vlastní GMM.

Při klasifikaci řečových módů je důležitá nejen správná identifikace úseků signálu obsahujících hledané samohlásky, ale i korektní vyhodnocení všech ostatních segmentů jako neurčených. Za tímto účelem jsou všechny testované typy klasifikátorů vybaveny explicitním určováním míry věrohodnosti, na základě které se rozhoduje, zda bude indikována detekce či nikoli. V případě GMM klasifikátorů je jako míra věrohodnosti použit negativní log-likelihood (aposteriorní pravděpodobnost), u SVM se uvažují odhady pravděpodobnosti příslušnosti vzorku k dané třídě (metoda one-against-one) převedené do negativní logaritmické míry. Je stanoven práh věrohodnosti, který musí být překročen, aby byla indikována detekce. Pro nalezení optimálního prahu u každého klasifikátoru byly zjišťovány relativní četnosti správných detekcí, chybných detekcí a vynechaných detekcí při zpracování trénovací množiny slov, viz obr. 4.3. Hodnota prahu při maximální četnosti správných detekcí byla vybrána pro použití při klasifikaci.



Obr. 4.3 Relativní četnosti detekcí klasifikátoru hlasových módů

Nejlepších výsledků dosahuje SVM klasifikátor módů s celkovou úspěšností 84% při 11% chybně rozpoznaných segmentů a 5% vynechaných detekcí. Srovnatelný klasifikátor módů řeči uvedený v [32] dosahoval průměrné úspěšnosti pouze 69,3% (s nejhorším výsledkem 44,5% pro neutrální řeč).

4.3 ZVYŠOVÁNÍ ROBUSTNOSTI ROZPOZNÁVAČE VŮČI ZMĚNÁM V HLASOVÉM ÚSILÍ MLUVČÍHO

S využitím databáze BUT-VE1 bylo možné prověřit účinnost přístupů zlepšujících funkci rozpoznávače jak pro případ systému závislého na mluvčím, tak pro rozpoznávač na mluvčím nezávislý. Dostupný rozsah variant slov byl rozdělen na poloviny; první část sloužila k tréninku rozpoznávačů, na druhé části byly prováděny testy. Slova byla před použitím výkonově normalizována, čímž se vyloučil triviální vliv úrovně signálu na úspěšnost rozpoznávání a šlo tedy čistě o fonetické změny ve struktuře slov. Pro dosažení realistických podmínek byl ke všem slovům přidán AWGN šum s průměrným odstupem SNR = 20 dB. Provedené experimenty lze rozdělit do následujících kategorií:

A) Rozpoznávač nezávislý na mluvčím:

1) Test multi-style HMM nezávislých na hlasovém módu,
2) HMM závislé na hlasovém módu + GMM klasifikátor hlasového módu s diagonálními kovariančními maticemi,
3) HMM závislé na hlasovém módu + GMM klasifikátor hlasového módu s plnými kovariančními maticemi,
4) HMM závislé na hlasovém módu + SVM klasifikátor hlasového módu

5) Test možnosti adaptace HMM trénovaných jen na normální řeči pro ostatní módy pomocí MLLR.

B) Rozpoznávač závislý na mluvčím:

1) Test možnosti přizpůsobení HMM nezávislých na mluvčím pro konkrétního mluvčího pomocí MLLR adaptace,
2) Použití HMM trénovaných přímo na konkrétním mluvčím + SVM klasifikátor módů.

Výsledky nejpodstatnějších experimentů jsou shrnuty v tab. 4.1. Všechny experimenty byly prováděny odděleně pro každého mluvčího, výsledky pak byly zprůměrovány. První experiment s multi-style HMM ukázal, že pouhé rozšíření trénovací množiny daného HMM zahrnutím všech hlasových módů nezajistí spolehlivé rozpoznávání ve všech úrovních hlasu nezávisle na mluvčím. Navrhl jsem využití přístupu multiple-model framework (MMF) [12] pro dosažení spolehlivé funkce rozpoznávače v celém spektru hlasového úsilí mluvčích. Pro každý z hlasových módů je zvlášť natrénována sada slovních HMM a výše popsaný klasifikátor módů v průběhu rozpoznávání vybírá nejvhodnější sadu s ohledem na aktuálně zachycenou řeč. Výsledky v tab. 4.1 shrnují dosažené chybovosti rozpoznávačů nezávislých na mluvčím (tj. testovaný mluvčí byl vždy vyloučen z trénovací množiny). Výchozí systém trénovaný obvyklým způsobem, tj. pouze na neutrální řeči, dosáhl průměrné chybovosti 38,1%, přičemž v šepotu a křiku byl prakticky nepoužitelný. Oproti tomu multi-style rozpoznávač (klasický způsob ošetření variability mluvčích) fungoval s chybovostí 33,2, což dává 4,9% absolutní pokles chybovosti (12,9% relativně). Navržený MMF přístup s SVM klasifikátorem módů dosáhl 19,3% WER, což dává 18,8% absolutní pokles (49,3% relativně) oproti výchozímu systému. Je tedy zřejmé, že navržený přístup dává největší potenciál zvyšování úspěšnosti rozpoznávání. Dalšího poklesu chybovosti by bylo možné dosáhnout rozšiřováním trénovací množiny, popř. přizpůsobením rozpoznávače konkrétnímu mluvčímu.

Tab. 4.1 Chybovosti WER rozpoznávačů řeči nezávislých na mluvčím

	Průměr WER [%] pro 13 mluvčích					
Testovaný mód řeči	Šepot	Tiše	Normálně	Nahlas	Křik	Průměr
Rozpoznávač trénovaný jen na neutrální řeči (výchozí systém)	69,9	37,0	11,5	23,7	48,4	**38,1**
Multi-style HMM rozpoznávač	53,2	35,8	24,4	26,4	26,1	**33,2**
MMF rozpoznávač s SVM klasifikátorem módů	28,6	33,0	13,1	10,3	11,4	**19,3**

V dalších experimentech byly zkoumány možnosti modifikace existujícího výchozího rozpoznávače nezávislého na mluvčím trénovaném jen na neutrální řeči tak, aby jej bylo možno použít v širším rozsahu hlasového úsilí. Použitím MLLR adaptace s využitím malého množství trénovacích vzorů došlo k úspěšnému rozšíření použitelnosti rozpoznávače, ovšem dosažené výsledky nedosahují kvalit MMF systému.

Ověřil jsem též principiální použitelnost navrženého MMF přístupu pro rozpoznávač závislý na mluvčím. Byl zaznamenán očekávaný pokles chybovosti oproti systému nezávislém na mluvčím, pro praktické využití by ovšem bylo nutné shromáždit větší objem trénovacích dat.

5 ZÁVĚR

Disertační práce podává ucelený přehled známých zdrojů chybovosti systémů pro automatické rozpoznávání řeči projevujících se při nasazení těchto systémů v náročných podmínkách. První část práce je věnována popisu způsobů detekce jednotlivých rušivých vlivů a metod pro minimalizaci jejich negativního dopadu na úspěšnost rozpoznávání. V první řadě jde o akustický šum; pro jeho ošetření existuje řada typů robustní příznakové reprezentace, které vychází z typických charakteristik řeči ve snaze potlačit projevy všech ostatních zvukových zdrojů. Dalším přístupem je úprava signálu a příznakových vektorů s cílem odfiltrovat neřečové složky. Poslední třídu metod tvoří algoritmy ošetřující vliv šumu na úrovni akustických modelů. Jako nejperspektivnější z nich se jeví použití speciálních modelů akustického šumu, které lze při rozpoznávání využít pro modelování zašuměné řeči v aktuálních podmínkách.

Kromě vnějších zvukových zdrojů má na úspěšnost rozpoznávání nezanedbatelný dopad i řečový projev samotného mluvčího. Vlivy jako stres, únava, popř. reakce na okolní podmínky (Lombard efekt, úroveň hlasového úsilí) vedou k podstatným změnám v řeči mluvčího. Podaný přehled konkrétních změn v hlase v souvislosti s jednotlivými vlivy včetně metod jejich detekce shrnuje současný stav vývoje v této oblasti. U vlivů, kde je to známo, je uveden i kvantifikovaný dopad na úspěšnost rozpoznávání.

Vlastní výzkum byl zaměřen na možnosti zvyšování účinnosti automatického rozpoznávání izolovaných slovních příkazů se zaměřením na specifické prostředí operačního sálu. V prvním kroku šlo o výběr nejvhodnější struktury rozpoznávače nabízející nejvyšší potenciál dalšího vylepšování v souvislosti s šumem i změnami v hlase mluvčích. Při základním testu rozpoznávání 23 slov vyslovovaných 6 mluvčími dosáhl DTW rozpoznávač chybovosti WER = 15,2%. Oproti tomu HMM rozpoznávač s celoslovními modely na stejné úloze vykazoval chybovost pouze 5,5%, přičemž čas potřebný pro rozpoznání vyšel 12× kratší. Vzhledem k jednoznačně lepším výsledkům i dalším výhodným vlastnostem HMM přístupu (schopnost využít informační obsah rozsáhlé databáze, možnost adaptace na šum a další vlivy) byl další výzkum orientován na tento typ statistického modelování.

Při výzkumu možností zvyšování šumové odolnosti rozpoznávače v prostředí operačního sálu jsem se zaměřil na konstrukci robustních akustických modelů rušivých zvuků. Pro tento účel jsem při mé stáži v Německu pořídil zvukový záznam několikahodinové neurochirurgické operace na Uniklinikum Marburg. S využitím této nahrávky jsem vyvinul nový přístup pro automatickou daty řízenou segmentaci signálu na lokálně stacionární úseky. Segmentovaný signál pak posloužil pro konstrukci robustního vícestavového markovského modelu zvuků na operačním sále. Spojením tohoto modelu a sady HMM čisté řeči metodou paralelní kombinace modelů vznikl rozpoznávač slovních příkazů dosahující při použití malou skupinou mluvčích chybovosti pod 10% při SNR nad 6 dB v prostředí nestacionárního šumu operačního sálu. Pro dosažení srovnatelné úspěšnosti systému používajícího pouze modely čisté řeči bez uvážení šumu by musel být poměr SNR cca o 30 dB vyšší.

V rámci dalšího vývoje šumově robustního rozpoznávače jsem formuloval novou strukturu akustického modelu rušivých zvuků využívající hierarchicky uspořádanou sadu skrytých markovských modelů ve struktuře klasifikačního stromu. Tento model umožňuje podrobnější popis šumového signálu, než by bylo dosažitelné s klasickým plochým HMM, nedochází však k enormnímu nárůstu výpočetních nároků.

Kromě akustického šumu má na spolehlivost rozpoznávače vliv i jeho schopnost přizpůsobit se změnám v hlase mluvčího. V dosavadní literatuře nebyla kvantifikována souvislost změn v celém rozsahu hlasového úsilí mluvčích (od šepotu až po křik) s úspěšností automatického rozpoznávání řeči. Provedl jsem tedy experimentální prověření spolehlivosti funkce rozpoznávače trénovaného jen na neutrální řeči při testu na celém rozsahu hlasového úsilí. Zjištěné zásadní selhání systému bylo motivací pro další výzkum v této oblasti.

V současnosti neexistuje komerčně dostupná databáze obsahující vzorky řeči v celém rozsahu hlasového úsilí, bylo tedy nutné přikročit k vytvoření nové databáze BUT-VE1. Jedná se o první databázi svého druhu obsahující statisticky významné množství řečového materiálu od více mluvčích zaznamenaného špičkovou aparaturou za přísně kontrolovaných podmínek. Databáze poskytuje fonetickou segmentaci všech slov a možnost přesně určit kalibrovanou hodnotu hladiny akustického tlaku v libovolném úseku řečového signálu. S použitím databáze BUT-VE1 jsem provedl reprezentativní sadu experimentů dokumentujících dopad změn v hlasovém úsilí mluvčích na úspěšnost rozpoznávání. Navrhl jsem koncept spojení klasifikátoru hlasových módů nezávislý na slovníku a multiple-model framework HMM, čímž vznikl první automatický rozpoznávač řeči konzistentně fungující v celém rozsahu hlasového úsilí mluvčích. Oproti systému trénovaném jen na neutrální řeči dosahuje 49,3% relativního poklesu WER (18,8% absolutně).

Shrnutí vlastního přínosu k rozvoji vědního oboru:

- Podání uceleného přehledu vlivů majících za následek pokles úspěšnosti automatických rozpoznávačů řeči včetně výčtu metod umožňujících jejich ošetření.

- Navržení nového přístupu pro automatickou segmentaci šumového signálu na lokálně stacionární úseky s využitím bayesovského informačního kritéria.
- Experimentální prozkoumání možností inicializace a tréninku vícestavového ergodického Markovova modelu nestacionárního šumu pro použití v rozpoznávači řeči kombinujícím modely čisté řeči a šumu (PMC).
- Navržení nové struktury statistického akustického modelu šumu využívajícího hierarchii HMM uspořádanou ve struktuře klasifikačního stromu. Model umožňuje podrobný popis šumového signálu s minimálními výpočetními nároky při dekódování a lze jej použít v PMC.
- Vytvoření unikátní řečové databáze BUT-VE1 zahrnující 4 hodiny řečového materiálu od 13 mluvčích ve všech hlasových módech (šepot, tichá řeč, normální řeč, hlasitá řeč, křik) s precizní fonetickou segmentací a kalibrací hladiny akustického tlaku.
- Stanovení vlivu změn v hlasovém úsilí mluvčích na úspěšnost systému pro automatické rozpoznávání řeči.
- Navržení konceptu rozpoznávače řeči vycházejícího z multiple-model framework a klasifikátoru řečových módů pro dosažení vysoké spolehlivosti při rozpoznávání řeči bez ohledu na momentální úroveň hlasového úsilí mluvčího.

Další vývoj zvyšování úspěšnosti automatických rozpoznávačů řeči by mohl být zaměřen na ošetření dalších vlivů mluvčího, které doposud nebyly zohledňovány – únava, zdravotní stav, laxní výslovnost apod. I v těchto případech by pravděpodobně bylo možno aplikovat přístup založený na oddělené klasifikaci typu promluvy a rozpoznávači slov využívajícím přizpůsobené markovské modely.

VYBRANÉ VLASTNÍ PUBLIKACE

[1] ZELINKA, P.; SIGMUND, M. Hierarchical classification tree modeling of nonstationary noise for robust speech recognition. *Information Technology and Control,* 2010, vol. 39, no. 3, p. 202 – 210. ISSN: 1392-124X.

[2] ZELINKA, P., SIGMUND, M. Automatic Vocal Effort Detection for Reliable Speech Recognition. In *Proceedings of the 2010 IEEE International Workshop on Machine Learning for Signal Processing*. Kittilä (Finland): Aalto University, 2010, p. 349 – 354. ISBN 978-1-4244-7876-7.

[3] ZELINKA, P.; SIGMUND, M. Building a Multi-State Noise HMM: Comparison of Approaches. In *Proceedings of the 33rd International Conference on Telecommunications and Signal Processing - TSP 2010.* Baden (Austria): 2010, p. 69 – 73. ISBN: 978-963-88981-0-4.

[4] ZELINKA, P., SIGMUND, M. Towards Reliable Speech Recognition in Operating Room Noise Environment. In *Proceedings of the 20th International Conference Radioelektronika 2010*. Brno: Dept. of Radioelectronics, BUT, 2010, p. 31 – 34. ISBN 978-1-4244-6319-0.

[5] ZELINKA, P. Smooth Interpolation of Gaussian Mixture Models. In *Proceedings of the 19th International Conference Radioelektronika 2009*. Bratislava (Slovak Republic): Dept. of Radioelectronics, BUT, 2009, p. 321 – 323. ISBN 978-1-4244-3536-4.

[6] ZELINKA, P.; SIGMUND, M., SCHIMMEL, J. Impact of vocal effort variability on automatic speech recognition. *Speech Communication,* 2012, vol. 54, no. 6, p. 732 – 742. ISSN: 0167-6393.

VYBRANÁ LITERATURA

[7] REYNOLDS, D. A. Experimental Evaluation of Features for Robust Speaker Identification. *IEEE Transactions on Speech and Audio Processing*. 1994, vol. 2, no. 4, p. 639 – 643. ISSN 1063-6676.

[8] HERMANSKY, H. Perceptual Linear Prediction (PLP) Analysis of Speech. *Journal of the Acoustic Society of America.* 1990, vol. 87, no. 4, p. 1738 – 1752. ISSN: 0001-4966.

[9] HERMANSKY, H., MORGAN, N. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*. 1994, vol. 2, no. 4, p. 578 – 589. ISSN 1063-6676.

[10] ROSE, R. Environmental Robustness in Automatic Speech Recognition. In *ITRW 2004*. Norwich (UK): ISCA, 2004, paper KRR.

[11] YUO, K., WANG, H. Robust Features for Noisy Speech Recognition Based on Temporal Trajectory Filtering of Short-Time Autocorrelation Sequences. *Speech Communication*. 1999, vol. 28, no. 1, p. 13 – 24. ISSN 0167-6393.

[12] XU, H., TAN, Z., DALSGAARD, P., LINDBERG, B. Robust Speech Recognition Based on Noise and SNR Classification – A Multiple-Model Framework. In *INTERSPEECH-2005*. Lisbon (Portugal): ISCA, 2005, p. 977 – 980. ISSN 1018-4074.

[13] LEGGETTER, C. J., WOODLAND, P. C. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech and Language*. 1995, vol. 9, no. 2, p. 171-185. ISSN 0885-2308.

[14] GALES, M. J. F., YOUNG, S. J. Robust continuous speech recognition using parallel model combination. In *IEEE Transactions on Speech and Audio Processing*, 1996, Vol. 4, No. 5, p. 352 – 359. ISSN: 1063-6676.

[15] STEENEKEN, H. J. M., HANSEN, J. H. L. Speech Under Stress Conditions: Overviews of the Effect on Speech Production and on System Performance. In *ICASSP-99.* Phoenix (USA): IEEE, 2002, p. 2079 – 2082. ISBN 0-7803-5041-3.

[16] BARRETT, J., PAUS, T. Affect-induced Changes in Speech Production. *Experimental Brain Research.* 2002, vol. 146, no. 4, p. 531 – 537. ISSN 0014-4819.

[17] DE BRUIJN, C., WHITESIDE, S. Use of Speech Recognition and Voice Fatigue: Measures of F0 and Spectral Slope. In *ICPhS XVI*. Saarbrücken (Germany): Universität des Saarlandes, 2007, p. 2061 – 2064.

[18] NWE, T. L., LI, H., DONG, M. Analysis and Detection of Speech Under Sleep Deprivation. In *INTERSPEECH 2006*. Pittsburgh (USA): ISCA, 2006, paper 1934-Wed2BuP.15.

[19] STEENEKEN, H. et al. The Impact of Speech Under 'Stress' on Military Speech Technology. NATO Project 4 Report AC/232/IST/TG-01. 2000. ISBN 92-837-1027-4.

[20] ITOH, T., TAKEDA, K., ITAKURA, F. Acoustic analysis and recognition of whispered speech. In *ASRU '01*. Madonna di Campiglio (Italy): IEEE, 2001, p. 429 – 432. ISBN 0-7803-7343-X.

[21] LIPPMANN, R., MARTIN, E. A., PAUL, D. B. Multi-style Training for Robust Isolated-word Speech Recognition. In *ICASSP 87*. IEEE, 1987, p. 705–708.

[22] WOMACK, B. D., HANSEN, J. H. L. Classification of Speech Under Stress Using Target Driven Features. Speech Communication. *Speech Communication.* 1996, vol. 20, no. 1-2, p. 131–150. ISSN 0167-6393.

[23] BOU-GHAZALE, S. E., HANSEN, J. H. L. HMM-Based Stressed Speech Modeling with Application to Improved Synthesis and Recognition of Isolated Speech Under Stress. In *IEEE Transactions on Speech and Audio Processing*. 1998, vol. 6, no. 3, p. 201 – 216. ISSN 1063-6676.

[24] MOLER, C.B. Numerical computing with Matlab. Siam, 2004. ISBN: 0-89871-560-1.

[25] PSUTKA, J., MÜLLER, L., MATOUŠEK, J., RADOVÁ, V. *Mluvíme s počítačem česky*. Praha: Academia, 2006. 752 pages. ISBN 80-200-1309-1.

[26] SHAOBING, S., GOPALAKRISHNAN, P. S. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop.* Virginia, 1998, p. 127 – 132.

[27] MANNING, C. D., SCHÜTZE, H. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999. 620 s. ISBN: 978-0262133609.

[28] BLEAKAS, K., LAGARIS, I. E. Split-merge incremental learning (SMILE) of mixture models. In *ICANN 2007*. Porto, 2007, p. 291 – 300. ISBN: 978-3-540-74689-8.

[29] WEST, K., COX, S. Features and classifiers for the automatic classification of musical audio signals. In *ISMIR '04*. Barcelona: Universitat Pompeu Fabra, 2004. ISBN: 84-88042-44-2.

[30] FINE, S., SINGER, Y., TISHBY, N. The hierarchical hidden Markov model: analysis and applications. *Machine Learning*, 1998, no. 32, p. 41 – 62. ISSN: 0885-6125.

[31] TERNSTRÖM, S., BOHMAN, M., SÖDERSTEN, M. Loud speech over noise: Some spectral attributes with gender differences. *Journal of the Acoustical Society of America*. 2006, vol. 3, no. 119, p. 1648 – 1665. ISSN 0001-4966.

[32] ZHANG, C., HANSEN, J. H. L. Analysis and Classification of Speech Mode: Whispered Through Shouted. In *INTERSPEECH-2007*. Antwerp (Belgium): ISCA, 2007, p. 2289–2292.

ŽIVOTOPIS

Osobní údaje:

Jméno:	Petr Zelinka
Datum narození:	24.12.1982
Adresa:	Zahraničního odboje 919, 67401 Třebíč
E-mail:	xzelin06@stud.feec.vutbr.cz

Vzdělání:

2008 – 2012 Doktorský studijní program Elektronika a sdělovací technika, Fakulta elektrotechniky a komunikačních technologií, Vysoké učení technické v Brně, dizertační práce – Zvyšování účinnosti strojového rozpoznávání řeči.

2006 – 2008 Magisterský studijní program Elektronika a sdělovací technika, Fakulta elektrotechniky a komunikačních technologií, Vysoké učení technické v Brně, diplomová práce – Realizace OFDM kodéru pro účely DVB-T.

2003 – 2006 Bakalářský studijní program Elektronika a sdělovací technika, Fakulta elektrotechniky a komunikačních technologií, Vysoké učení technické v Brně, bakalářská práce – Zařízení pro elektronickou volbu otázek u zkoušky.

Odborné stáže:

III – VII/2009 Stáž na Hochschule RheinMain ve Wiesbadenu v Německu. Spolupráce s lékařským pracovištěm Uniklinikum Marburg.

Pedagogická praxe:

2008 – 2010 Výuka laboratorních cvičení předmětu „Základy televizní techniky“ na Ústavu radioelektroniky.

Vedení bakalářských prací na Ústavu radioelektroniky VUT v Brně.

ABSTRAKT

V práci jsou identifikovány příčiny nedostatečné spolehlivosti současných systémů pro automatické rozpoznávání řeči při jejich nasazení v náročných podmínkách. U jednotlivých rušivých vlivů je popsán jejich dopad na úspěšnost rozpoznávání a je podán výčet známých postupů pro identifikaci těchto vlivů analýzou rozpoznávaného signálu. Je též uveden přehled obvyklých metod používaných k omezení dopadu rušivých vlivů na funkci rozpoznávače řeči. Vlastní přínos tkví v navržení nových postupů pro vytváření akustických modelů zašuměné řeči a modelů nestacionárního šumu, díky kterým je možné dosáhnout vysoké úspěšnosti rozpoznávání v náročných akustických podmínkách. Účinnost navržených opatření byla otestována na rozpoznávači izolovaných slov s využitím nahrávky reálného akustického pozadí operačního sálu pořízené na Uniklinikum Marburg v Německu při několikahodinové neurochirurgické operaci. Tato práce jako první přináší popis dopadu změn v hlasovém úsilí mluvčích na spolehlivost rozpoznávání řeči v celém rozsahu, tj. od šepotu až po křik. Je navržena koncepce rozpoznávače řeči, který je imunní vůči změnám v hlasovém úsilí mluvčích. Pro účely zkoumání změn v hlasovém úsilí byla v rámci řešení práce sestavena nová řečová databáze BUT-VE1.

ABSTRACT

This work identifies the causes for unsatisfactory reliability of contemporary systems for automatic speech recognition when deployed in demanding conditions. The impact of the individual sources of performance degradation is documented and a list of known methods for their identification from the recognized signal is given. An overview of the usual methods to suppress the impact of the disruptive influences on the performance of speech recognition is provided. The essential contribution of the work is the formulation of new approaches to constructing acoustical models of noisy speech and nonstationary noise allowing high recognition performance in challenging conditions. The viability of the proposed methods is verified on an isolated-word speech recognizer utilizing several-hour-long recording of the real operating room background acoustical noise recorded at the Uniklinikum Marburg in Germany. This work is the first to identify the impact of changes in speaker's vocal effort on the reliability of automatic speech recognition in the full vocal effort range (i.e. whispering through shouting). A new concept of a speech recognizer immune to the changes in vocal effort is proposed. For the purposes of research on changes in vocal effort, a new speech database, BUT-VE1, was created.