TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

Ústav Informačních technologií a elektroniky

Audiovizuální rozpoznávání řeči s využitím metod pro automatické odezírání ze rtů

Audiovisual Speech Recognition by Utilizing Methods for Automatic Lipreading

Autoreferát dizertační práce

Studijní program: P2612 Elektrotechnika a informatika
Studijní obor: 2612V045 Technická kybernetika
Autor: Ing. Karel Paleček
Školitel: doc. Ing. Josef Chaloupka, Ph.D.

Abstrakt

Automatické odezírání ze rtů je oborem vyvíjejícím se na pomezí automatického rozpoznávání řeči, strojového učení a počítačového vidění již více než 20 let. Ani přes významné pokroky od doby svého uvedení se však audiovizuální systémy rozpoznávání řeči v praxi výrazně neprosadily a to z několika důvodů. Jeden z klíčových předpokladů, návrh robustní parametrizace, zde navíc s využitím informace o trojrozměrné podobě povrchu úst, je předmětem této dizertační práce.

V práci jsou navrženy tři nové vizuální parametrizace řeči: trojrozměrná bloková diskrétní kosinová transformace (DCT3), prostoro-časově modifikovaný histogram orientovaných gradientů (HOGTOP) a rozšířený aktivní vzhledový model (DAAM). Jejich návrh, popsaný v kapitole 2, směřuje především k využití řečové dynamiky. Cílem příznaků DAAM je zrobustnění klasického AAM integrací hloubkových dat jakožto zjednodušené formy informace o trojrozměrné podobě rtů.

Za účelem vyhodnocení navržených i v současné době existujících parametrizací je vytvořena audiovizuální databáze TULAVD obsahující 54 mluvčích, viz kapitolu 3. Zatímco většina dostupných databází zahrnuje pouze promluvy s omezeným slovníkem a striktní gramatickou strukturou, TULAVD je navržena i s ohledem na automatické rozpoznávání spojité řeči s velkým slovníkem (LVCSR). Samostatná sekce je věnována návrhu testovacího protokolu, který zamezuje optimalizaci modelů na testovaná data a výsledky v experimentální části tak nejsou zatíženy pozitivní zaujatostí.

Experimentální část v kapitole 4 se věnuje především evaluaci navržených parametrizací a srovnání existujících na úloze rozpoznávání izolovaných slov. Kromě TULAVD je úspěšnost vlastní parametrizace demonstrována na dalších dvou známých databázích pro možnost přímého srovnání se stavem poznání. Rovněž je samostatně demonstrován pozitivní přínos hloubkových dat rekonstruovaných pomocí MS Kinect, jak v rámci brzké integrace a DAAM, tak při integraci v synchronních vícekanálových HMM. Druhá část experimentů v kapitole 5 je pak zaměřena vyhodnocení vlivu vizuální informace v úloze LVCSR s různě velkými slovníky od několika stovek do pěti set tisíc slov.

Klíčová slova: audiovizuální rozpoznávání řeči, odezírání ze rtů, rozpoznávání spojité řeči s velkým slovníkem, hloubková mapa, Kinect, skrytý markovský model

Abstract

Automatic lip reading is a research field closely related to automatic speech recognition, machine learning and computer vision. Despite being developed for more than two decades, systems for audiovisual speech recognition are still not widely used in practice due to several reasons. One critical component, namely the design of a robust and discriminative visual parametrization, here also with utilization of information about depth, is the main topic of this dissertation thesis.

Three different robust visual parametrizations are proposed and explained in chapter 2: block-based three-dimensional discrete cosine transform (DCT3), spatiotemporal histogram of oriented gradients (HOGTOP), and depth-extended active appearance model (DAAM). While the former two are ROI-based source-agnostic parametrizations designed mainly to exploit the speech dynamics, DAAM directly integrates depth data obtained via Kinect in order to achieve greater robustness against lightning variations and better phone discrimination.

In order to evaluate the existing and proposed features on both video and depth data, new database called TULAVD has been recorded. As described in chapter 3, each of the 54 speakers uttered 50 isolated words and 100 gramatically unrestricted sentences in Czech language. Special section is devoted to the design of the evaluation protocol in order to minimize the risk of overfitting and introduction of an optimistic bias when tuning hyperparameters of the parametrizations and the model.

Experiments in chapter 4 evaluate selected popular and proposed features in the task of isolated unit recognition. In order to compare the achieved results to the state of the art, two other commonly used datasets besides TULAVD are included: OuluVS and CUAVE. Experiments on multiple modality fusion show the benefit of adding the Kinect depth data into the recognition process for both feature fusion and integration via multistream hidden Markov model. As opposed to the vast majority of recent work on lipreading, the above mentioned evaluation is also performed in the task of large vocabulary continuous speech recognition with gradually increasing vocabulary size from several hundreds to half a million, see chapter 5.

Keywords: audiovisual speech recognition, lipreading, large vocabulary continous speech recognition, depth map, Kinect, hidden Markov model

Obsah

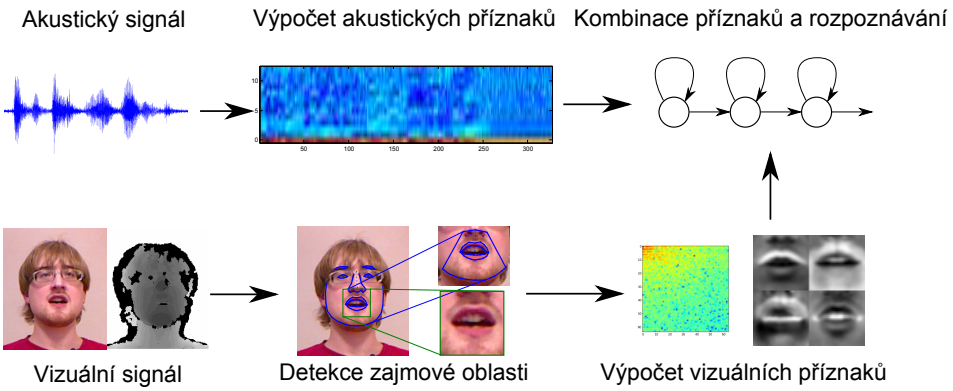
1. Úvod

Automatického rozpoznávání řeči (Automatic Speech Recognition, ASR) je obor, ve kterém aktivní výzkum probíhá již od 60. let minulého století. V dnešní době nachází široké uplatnění, např. v oblastech bezpečnosti, vzdělávání, interakce člověka s počítačem, ale i v chytrých domácnostech a zábavě. I přes nesporný pokrok však nelze považovat problém za vyřešený. Především v prostředích s hlukem na pozadí algoritmy často nedosahují uspokojivé přesnosti. ASR proto zahrnuje řadu pod-oblastí, které řeší některé dílčí problémy, případně se snaží využít specifických podmínek reálných aplikací za účelem snížení chybovosti. Jednu z těchto oblastí představuje audiovizuální rozpoznávání řeči (Audio-Visual Speech Recognition, AVSR), které se snaží využít dodatečná obrazová data, akustickým hlukem na pozadí nezatížená. Idea je přitom inspirována způsobem, jakým se s podobnými podmínkami běžně vypořádávají zdraví, ale i sluchově postižení lidé, tj. odezíráním pohybu rtů.

Proces audiovizuálního rozpoznávání řeči lze rozdělit do několika základních bloků, které jsou schematicky znázorněny na obrázku 1.1. Vstupem systému je řečníkova promluva v podobě akustického a vizuálního signálu, výstup pak představuje sekvence rozpoznaných slov. Zpracování akustického a obrazového kanálu probíhá do značné míry nezávisle a k fúzi informace dochází až ve fázi samotného rozpoznávání. Toto uspořádání zajišťuje modularitu automatického rozpoznávání řeči tak, aby bylo možné při absenci jednoho z kanálů zachovat funkci celého systému. Podrobně jednotlivé komponenty rozebírají kapitoly 2–5 hlavního textu předložené práce. Kvantitativní srovnání současného stavu poznání v závislosti na typu úlohy a dalších faktorech poskytuje kapitola 7.

1.1 Cíle dizertační práce

Cílem předložené dizertační práce je především návrh robustní a dostatečně diskriminační vizuální parametrizace vhodné pro rozpoznávání nezávislém na řečníkovi. Pro extrakci by se kromě klasické RGB textury a tvarových příznaků jako vhodná mohla ukázat i informace o trojrozměrné podobě úst, např. změny ve vyšpulení a zatažení. Pro extrakci takových příznaků lze rekonstruovat povrch oblasti zájmu z více pohledů, nebo využít některé z dostupných zařízení, jež úlohu řeší interně a problémy s případnou citlivostí



Obrázek 1.1: Princip audiovizuálního rozpoznávání.

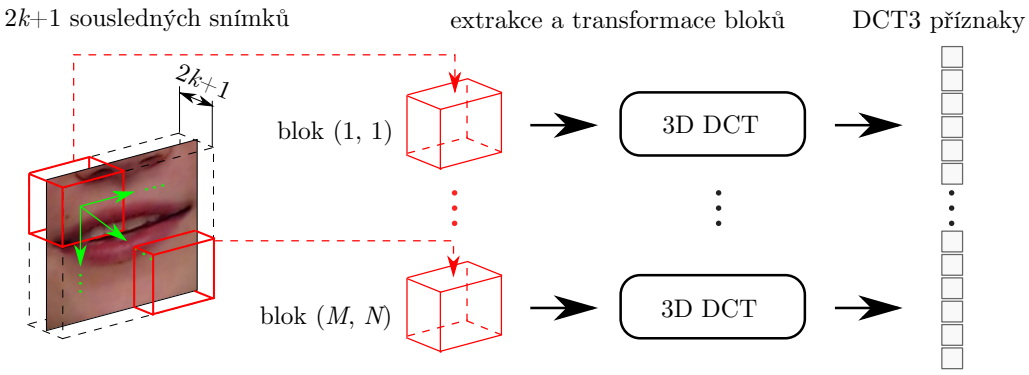
na změny osvětlení řeší přechodem do infračervené oblasti. Přínos navržené parametrizace by měl být ověřen nejen na zjednodušených specializovaných úlohách jako je např. rozpoznávání izolovaných slov a frází, ale i v reálnějších podmínkách se spontánní řečí a větším slovníkem. Samostatně by měl být vyhodnocen přínos trojrozměrné informace oproti jednoduššímu případu standardní RGB kamery. Protokol evaluace musí být navržen tak, aby nedocházelo k optimalizaci parametrů na testovací data a výsledky tak byly přímo porovnatelné a vypovídající. Jelikož žádná z dostupných audiovizuálních databází uvedené nároky nesplňuje, jedním z prvních úkolů musí být vytvoření vlastní. Přehledně hlavní cíle této práce shrnuje následující výčet.

- Vytvoření uceleného přehledu stavu poznání v problematice AVSR a úzce souvisejících oblastech.
- Návrh kvalitní vizuální parametrizace s využitím rekonstrukce trojrozměrné informace v podobě hloubkových map.
- Sestavení dostatečně rozsáhlé audiovizuální databáze pro otestování existujících a navržených metod.
- Srovnání nejrozšířenějších parametrizací na více audiovizuálních databázích v úloze rozpoznávání izolovaných jednotek.
- Systematické vyhodnocení přínosu integrace hloubkových dat.
- Srovnání parametrizací a posouzení přínosu vizuální složky v úloze rozpoznávání spojité řeči s velkým slovníkem.

2. Návrh vizuální parametrizace řeči

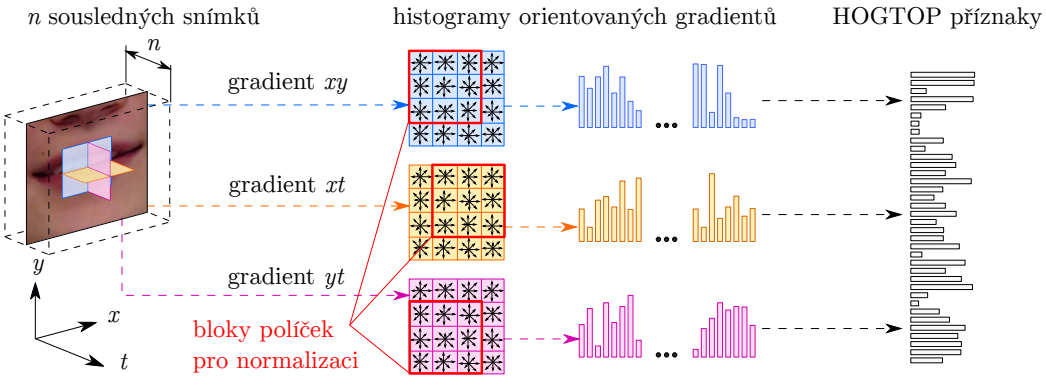
V kapitole 8 hlavního textu práce jsou navrženy tři typy vizuální parametrizace. Jedním z cílů bylo využít co nejlépe dynamiku řeči přímo příznakovém popisu a nespoléhat tak v tomto ohledu pouze na klasifikátor. Dalším záměrem pak bylo vyhodnotit přínos hloubkové informace, kterou nabízí stále více zařízení levně dostupných na trhu. Čím více požadované informace je ze signálu vytěženo během fáze parametrizace, tím více se zjednodušuje návrh klasifikačního modelu, protože v trénovací fázi totiž není nutné hledat složité nelineární korelace a časové závislosti. Jednodušší model znamená méně volných parametrů, nižší datovou náročnost, vyšší efektivitu a především nižší riziko přeučení.

Jednou z parametrizací navržených s cílem využít dynamiku řeči je **trojrozměrná bloková diskrétní kosinová transformace** (DCT3), jejíž extrakce z videa ilustruje obrázek 2.1. Na video je podobně nahlíženo jako na trojrozměrné útvary s dvěma obrazovými a jednou časovou osou. Sekvence $2k + 1$ sousledných snímků se středem v aktuálním snímku je rozdělena na $N \times M \times 1$ bloků v ose x, y, resp. t. Z každého bloku je trojrozměrnou DCT extrahováno d koeficientů a výsledek pospojován do jediného vektoru. Výsledná parametrizace je vzhledem k potenciálně vysoké dimenzi redukována a zároveň dekorelována metodou PCA. Optimální velikost bloku, překryv a počet DCT koeficientů jsou zjištěny pomocí křížové validace, viz kapitolu 4.



Obrázek 2.1: Princip extrakce DCT3 příznaků.

Další navrženou parametrizací je **prostoročasový histogram orientovaných gradientů** (Spatio-temporal Histogram of Oriented Gradients, HOGTOP). Zkratka HOGTOP odvozena od způsobu extrakce ze tří kolmých rovin xy, xt a yt (t je časová osa) analogicky k známé modifikaci LBPTOP, tedy z anglického spojení Histogram of Oriented Gradients from Three Orthogonal Planes. Postup je extrakce ilustruje obr. 2.2. Pro každý pixel vstupního snímku jsou vypočteny tři gradienty g_{xy}, g_{xt} a g_{yt} pro odpovídající roviny, čímž vzniknou tři samostatné gradientní obrazy. Pro každou složku zvlášť je sestaven histogram gradientových směrů, normalizován přes bloky a následně pospojován do parametrizačního vektoru. Pro zachycení dynamiky delší než jen mezi dvěma následujícími snímky je derivace aproximována konvolucí s derivovaným gaussovským jádrem o délce $2k + 1$ koeficientů a jedná se tak o nekauzální filtr. Hodnota k byla stanovena empiricky na $k = 3$. Na rozdíl od původní práce [Dalal 2005], kde autoři dosáhli nejlepších výsledků aplikací obyčejné diference, je zde stejný filtr použitý i pro obrazovou rovinu. Takto vytvořené příznaky odpovídající jednotlivým rovinám jsou nakonec spojeny do jediného vektoru a redukovány a dekorelovány metodou PCA na rozměr několika desítek koeficientů. Přesná hodnota je stanovena křížovou validací s cílem maximalizovat slovní přesnost, viz sekci 3.2.



Obrázek 2.2: Princip extrakce HOGTOP příznaků.

Práce se rovněž zabývá využitím hloubkové mapy pro odezírání ze rtů. Možným způsobem je varianta přímé integrace, kdy je ovšem zohledněna vzájemná korelace příznaků pomocí metod strojového učení a výsledná parametrizace tedy není pouhým spojením dílčích vektorů. Modifikován byl proto aktivní vzhledový model o texturu

extrahovanou z hloubkové mapy. Podobně jako jsou kombinovány tvar a textura trojnásobnou aplikací PCA, je možné do modelu přidat hloubkovou „texturu“ a dále pracovat se vzhledovým modelem jako v klasickém případě. Takto rozšířené AAM kombinuje parametry jako

$$\boldsymbol{a} = \begin{bmatrix} w_s \boldsymbol{p} \\ \boldsymbol{\lambda} \\ w_d \boldsymbol{\gamma} \end{bmatrix} = \boldsymbol{\Phi} \boldsymbol{d}, \tag{2.1}$$

kde $\boldsymbol{\gamma}$ je PCA redukce „textury“ extrahované z hloubkové mapy a multiplikativní konstanty w_s a w_d mají za úkol normalizovat všechny příznaky na stejný rozptyl pro následnou analýzu hlavních komponent. Výhodou tohoto postupu je, že výsledná parametrizace zachycuje většinu uvažovaných zdrojů informace: tvar, texturu i hloubku a příznaky ze všech tří modalit vzájemně dekoreluje pomocí metody PCA.

3. Příprava dat a návrh testovacího protokolu

3.1 Audiovizuální databáze TULAVD

V rámci práce byla vytvořena audiovizuální databáze TULAVD, která umožňuje testování navržených vizuálních řečových příznaků a přínosu hloubkových dat pro odezírání ze rtů. Databáze obsahuje promluvy od celkem 54 mluvčích, z toho 23 žen a 31 mužů. K nahrávání bylo použito více zdrojů, konkrétně 2 webkamery Logitech C920, 2 senzory Microsoft Kinect v první verzi a klopový mikrofon. Ukázka obrazu a příslušné hloubkové mapy z nahrávání databáze zařízením Kinect je k vidění na obr. 3.1. Databáze byla nahrávána v běžných kancelářských prostorech během pracovní doby, místnost nebyla nijak odhlučněna. Světelné podmínky byly přibližně konstantní. Nahrávání probíhalo u obyčejného PC s instruktorem. Mluvčí byli usazeni ve vzdálenosti cca 80 cm od snímacích zařízení.

Databáze obsahuje tři hlavní části. První část je pouze obrazová a tvoří ji databáze obličejů. Obličej každého mluvčího je zachycen v několika polohách, s různým nasvícením, nezakrytý či s částečným zakrytím a s různými výrazy jako např. úsměv, údiv či znechucení. Účelem je vytvoření databáze obličejů, pomocí které lze natrénovat robustní detektory. Druhá část obsahuje od každého mluvčího 50 izolovaných slov. Těchto 50 slov

Obrázek 3.1: Ukázka obrazu a hloubkové mapy získané z Microsoft Kinect.

je pro každého mluvčího stejných a tvoří základ pro experimenty s různými vizuálními řečovými příznaky. V třetí části bylo každým mluvčím namluveno 100 vět či souvětí o délce 5-20 slov. Těchto 100 foneticky vyrovnaných vět je rozděleno na dvě skupiny. Prvních 50 vět je společných pro všechny mluvčí, zbylé jsou pro každého jedinečné.

3.2 Křížová validace

Častým způsobem vyhodnocení úspěšnosti navrženého modelu je tzv. k-násobná křížová validace (K-Fold Cross Validation, KFCV), která kompletní dostupná data dělí např. po mluvčích do k stejně velkých disjunktních bloků, jež dle potřeby mohou sloužit jako trénovací ($\mathcal{T}$) či validační ($\mathcal{V}$) data. Tradičně se KFCV používá pro výběr modelu (tj. odladění všech parametrů) současně s porovnáním úspěšnosti vůči jiným algoritmům. Jelikož je však úspěšnost vyhodnocována na validačních datech, která zároveň slouží pro odlaďování, může docházet k přeučení a v důsledku k optimistické zaujatosti. Správnějším postupem je proto vnořená (dvojúrovňová) křížová validace, která zavádí navíc testovací bloky $\mathcal{S}$, jež se nijak procesu trénování neúčastní. Jejím problémem jsou však značné nároky na výpočetní a paměťovou kapacitu. V práci byl proto jako kompromis navržen protokol sestávající z následujících kroků:

1. natrénovat a vybrat optimální model pomocí $k-2$ bloků $\mathcal{T}$ a jednoho bloku $\mathcal{V}$ a vyhodnotit na bloku $\mathcal{S}$;

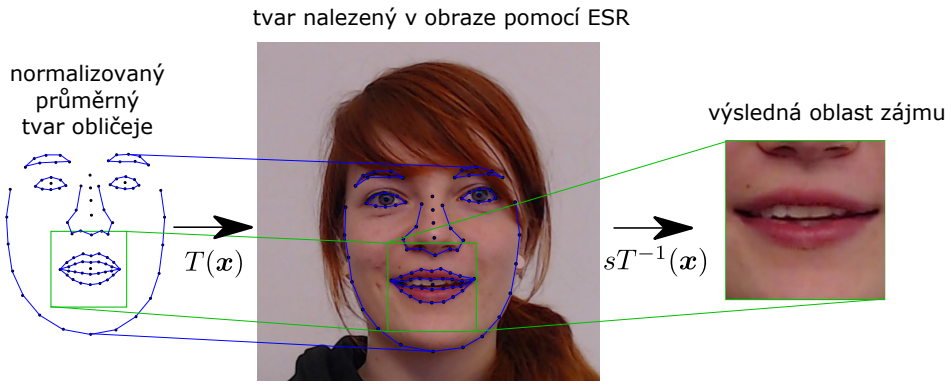
2. opakovat předchozí bod k-krát pro všechny ostatní $\mathcal{S}$;

3. úspěšnost vyhodnotit jako průměrné skóre přes všechny uvažované $\mathcal{S}$.

Jedná se o zjednodušení plné vnořené křížové validace, kdy vnitřní křížová validace zahrnuje vždy pouze jedno rozdělení na $(k-2)\times\mathcal{T}$ bloků a jeden $\mathcal{V}$ blok ze všech možných $k-1$ způsobů. Databáze TULAVD tedy byla rozdělena do šesti skupin po devíti mluvčích a velikosti trénovací, validační a testovací množiny tak činily 36, 9, resp. 9.

3.3 Extrakce zájmové oblasti

V předložené práci je oblast zájmu (Region of Interest, ROI) definována jako čtvercová oblast o rozměrech 64×64 pixelů extrahovaná na základě lokalizace klíčových bodů na obličeji algoritmem **explicitní tvarové regrese** (Explicit Shape Regression, ESR) [Cao 2012]. Jelikož se práce věnuje především rozpoznávání nezávislém na řečníkovi, je ESR natrénován pouze na mluvčích nacházejících se v trénovací množině každého uvažovaného rozdělení křížové validace, viz sekci 3.2. S cílem co možná nejlepší kompatibility s nejčastěji používanými modely v existujícím výzkumu byla zvolena konfigurace $v = 93$ obličejových bodů, již znázorňuje obrázek 3.2. Velikost a pozice ROI jsou stanoveny relativně vůči normalizovanému průměrnému tvaru obličeje a tedy nezávisle na měřítku. Jako inicializace procesu zarovnání obličeje slouží metoda Violy a Jonese [Viola 2001]. Algoritmus ESR byl implementován v jazyce C++ a na procesoru i7 2600k @ 3.4 GHz se 16 GB RAM trvá jedno zarovnání cca 2–5 ms.



Obrázek 3.2: Extrakce oblasti zájmu.

3.4 Ostatní použité databáze

Kromě vlastní databáze TULAVD byly pro srovnávací experimenty využity dvě další volně dostupné databáze: **OuluVS** [Zhao 2009] a **CUAVE** [Patterson 2002]. Databáze OuluVS obsahuje celkem 20 řečníků (17 mužů a 3 ženy), z nichž každý 5x opakuje 10 krátkých každodenních frází v angličtině, dohromady tedy přibližně 1000 promluv (výjimečně jsou promluvy opakovány 4x či 6x). Pro modely, jenž vyžadují anotaci jednotlivých snímků (např. dynamizace LDA), byly zvukové nahrávky OuluVS a CUAVE nuceně zarovnány (forced alignment) s využitím volně dostupného akustického modelu[1] pro americkou angličtinu připraveného lingvistickou laboratoří na University of Pennsylvania.

4. Rozpoznávání izolovaných slov a frází

Experimentální část dizertační práce je rozdělena do dvou kapitol, z nichž první (10) prezentuje výsledky dosažené v úloze rozpoznávání izolovaných slov a frází pomocí celoslovních modelů. Následující kapitola 11 hlavního textu se pak zabývá modely hláskovými a jejich aplikací v úloze audiovizuálního rozpoznávání spojité řeči s velkým slovníkem. V obou úlohách byla za kritérium úspěšnosti byla zvolena **slovní přesnost** (angl. word accuracy, WAcc)

$$\mathrm{WAcc} = \frac{N - D - S - I}{N}, \tag{4.1}$$

kde N je celkový počet slov v referenčním přepisu, D je počet vynechaných slov, tzv. delecí, S počet chybně rozpoznaných slov, tzv. substitucí, a I počet chybně vložených slov, tzv. inzercí. V případě izolovaných slov vyjadřuje WAcc podíl správně rozpoznaných jednotek vůči celkovému počtu. Při posuzování míry zlepšení jednoho modelu vůči jinému je rovněž uváděna relativní změna **slovní chybovosti** (Word Error Rate, WER)

$$\delta_{\mathrm{WER}} = \frac{(\mathrm{WER2} - \mathrm{WER1})}{\mathrm{WER1}} = \frac{\mathrm{WAcc1} - \mathrm{WAcc2}}{1 - \mathrm{WAcc1}}, \tag{4.2}$$

kde WER = 1 − WAcc. Pro trénování modelů i klasifikaci byla využita volně dostupná knihovna HTK[1] [Young 2006] verze 3.4.1.

[1] `http://www.ling.upenn.edu/phonetics/p2fa`
[1] `http://htk.eng.cam.ac.uk/`

Databáze	slovník	# mluvčích	Rozdělení	Opakování
TULAVD	50 slov	54	36:9:9	6×CV
OuluVS	10 frází	20	19:1	20×LOOCV
CUAVE	10 číslovek	36	24:6:6	6×CV

Tabulka 4.1: Rozdělení dat jednotlivých databází pro rozpoznávání izolovaných jednotek.

4.1 Vizuální rozpoznávání

Do experimentů byly zahrnuty tři databáze: vlastní TULAVD a volně dostupné OuluVS a CUAVE. Z databáze CUAVE byla využita pouze část s izolovanými číslovkami, přičemž hranice jednotlivých promluv byly určeny na základě nuceného zarovnání zvukové stopy s 0.5 s navíc na začátku i konci (typická délka jedné číslovky je něco přes 1 s). Při evaluaci algoritmů na databázích TULAVD a CUAVE byl aplikován postup vnější křížové validace ze sekce 3.2, přičemž počty mluvčích v trénovacích, validačních, resp. testovacích množinách jsou uvedeny v třetím sloupci tabulky 4.1. V případě OuluVS byla kvůli kompatibilitě se stavem poznání zvolena standardní křížová validace vynech jeden (LOOCV), kde hyperparametry jednotlivých modelů a algoritmů jsou vybrány tak, aby maximalizovaly průměrné skóre přes všechny testovací množiny. Toto skóre je pak zároveň uvedeno ve výsledcích jako dosažená slovní přesnost. Jelikož zde testovací množiny poskytují zpětnou vazbu pro proces ladění hyperparametrů, stávají se součástí trénování a hrozí tak riziko optimistické zaujatosti. Důvodem aplikace LOOCV však bylo dosažení co nejširší kompatibility se stavem poznání, jelikož byl tento postup zvolen ve všech známých pracích využívajících OuluVS.

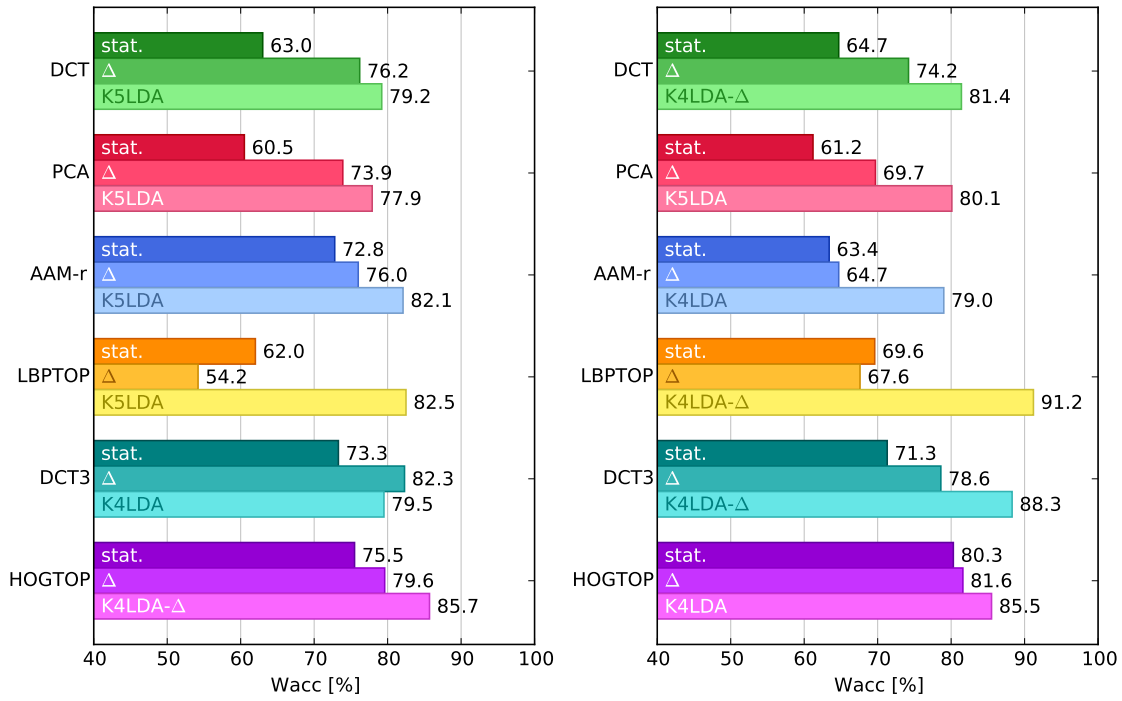
V experimentech byly zohledněny následující parametrizace: diskrétní kosinová transformace (DCT), analýza hlavních komponent (PCA), aktivní vzhledový model (AAM), lokální binární vzory s využitím časové složky (LBPTOP), trojrozměrná bloková DCT (DCT3) a histogram orientovaných gradientů s využitím časové složky (HOGTOP). Všechny parametrizace s výjimkou AAM jsou extrahovány ze šedotónových obrázků zájmové oblasti v případě videa a bilineárně interpolovaných dat v případě hloubkové mapy (týká se pouze databáze TULAVD). AAM textura je extrahována ze všech tří složek RGB a vektorově spojena za sebe. Pro extrakci AAM příznaků byla využita dolní část obličeje a tedy pouze podmnožina všech obličejových bodů. Experimentální porovnání úspěšnosti v závislosti na zvolené konfiguraci klíčových bodů je k dispozici v hlavním textu práce.

4.1.1 Srovnávací experimenty

Parametrizace jsou porovnávány ve třech variantách: statické, dynamické Δ a dynamické LDA (příp. $+\Delta$). U poslední varianty je pospojováno $2K+1$ sousledných snímků a vzniklý hypervektor redukován metodou LDA. Optimální postprocessing byl pro každou parametrizaci stanoven zvlášť s cílem maximalizovat slovní přesnost (nebyl považován za hyperparametr a tedy předmět křížové validace). Výsledky pro databázi TULAVD jsou uvedeny v tabulce 4.2 a pro OuluVS a CUAVE na grafech 4.1a a 4.1b. V experimentu se ukazuje přínos příznaků navržených v této práci. Oba navržené typy, DCT3 a HOGTOP, na vlastní databázi TULAVD dosahují ve statických variantách nejvyšší slovní přesnosti.

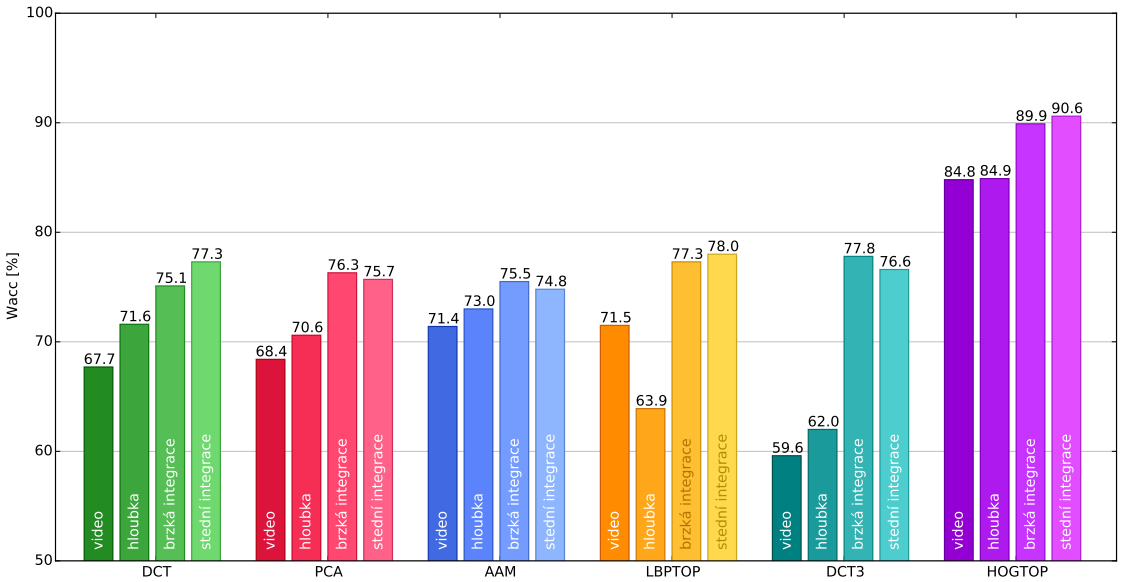
Param.	Video		Hloubka		Mod.
	Stat.	Dyn.	Stat.	Dyn.	
DCT	54,0	68,9	55,9	66,0	Δ
		72,5		74,4	K5LDA
PCA	51,4	64,4	55,7	65,3	Δ
		73,9		72,4	K5LDA
AAM-r	58,1	61,8	59,7	63,0	Δ
		74,1		75,2	K5LDA
LBPTOP	67,4	69,7	40,9	43,7	Δ
		74,2		64,3	K3LDA
DCT3	61,6	70,8	62,9	73,0	Δ
		75,1		70,3	K4LDA-Δ
HOGTOP	76,1	80,4	72,1	75,0	Δ
		86,4		84,4	K4LDA-Δ

Tabulka 4.2: Slovní přesnost [%] v závislosti na dynamizaci v úloze rozpoznávání izolovaných slov na databázi TULAVD.



(a) (b)

Obrázek 4.1: Slovní přesnost [%] v závislosti na dynamizaci v úloze rozpoznávání izolovaných jednotek na databázích OuluVS (a) a CUAVE (b).



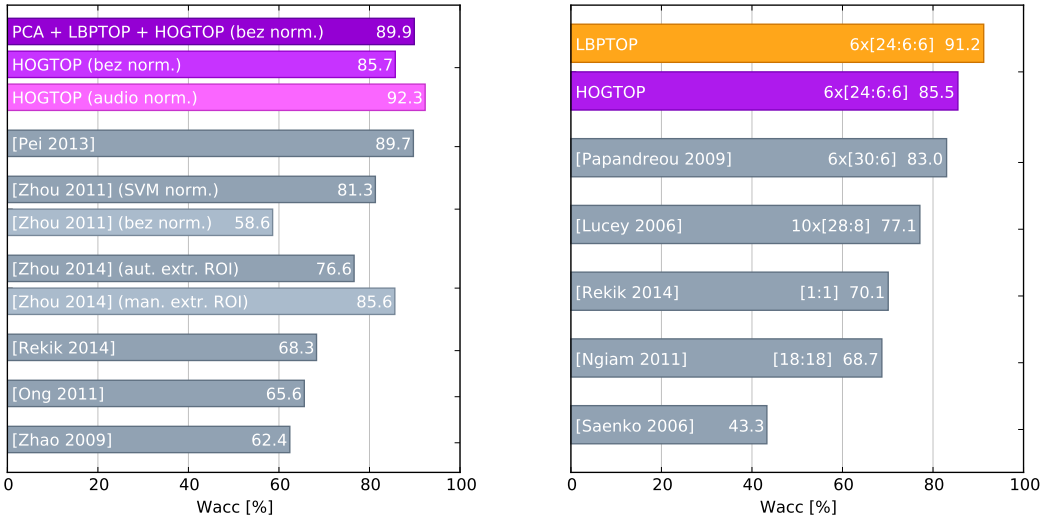
Obrázek 4.2: Dosažená úspěšnost v úloze rozpoznávání izolovaných slov na databázi TULAVD pro kombinace obrazových a hloubkových příznaků.

Parametrizace HOGTOP převyšuje ostatní typy i v základní statické variantě, přičemž však stále dokáže benefitovat z dodatečné LDA dynamizace až na 86,1 % pro video a 84,4 % pro hloubku, což je např. oproti v současné době velmi populární parametrizaci LBPTOP o 15 % ($\delta_{\mathrm{WER}} = -47$ %), resp. 20 % ($\delta_{\mathrm{WER}} = -56$ %) lepší výsledek. Na databázích OuluVS a CUAVE nejsou rozdíly tak výrazné, přesto však příznaky HOGTOP vykazují nejlepší výsledky.

4.1.2 Integrace hloubkových příznaků

Předložená práce se zabývá i vlivem informace o hloubce (vzdálenosti jednotlivých pixelů od kamery), která může být přínosná především pro zvýraznění rozdílů mezi hláskami jako ‘m’ či ‘b’, jež charakterizují zatažená ústa, a např. ‘u’ či ‘č’, pro něž jsou typické vyšpulené rty. Za účelem zjištění přínosu kombinace parametrizací a byly zohledněny dva typy integrace: brzká a střední. Základní typ brzké integrace spočívá v jednoduchém spojení dvou či více příznakových vektorů, přičemž další zpracování se nijak neliší od standardního postupu. Druhým zástupcem brzké integrace je hloubkový AAM (DAAM) (2.1) popsaný v sekci 8.3 hlavního textu práce. V případě střední integrace byl aplikován synchronní vícekanálový markovský model (MSHMM) se součtem vah rovným jedné. Váhy byly považovány za hyperparametr modelu a tedy křížově validovány postupem uvedeným 4.1 (pro každé rozdělení odděleně).

Sloupcový graf 4.2 zobrazuje výsledky po kombinaci stejných typů parametrizace z různých zdrojů, tj. videa a hloubky. Brzká integrace v podobě hloubkového AAM (DAAM) v experimentu dosáhla úspěšnosti 74,9 %. Ve všech případech se potvrzuje přínos integrace hloubkových dat, přičemž zlepšení se pohybuje v rozmezí 4–18 % (δ_{WER} −14 až −45 %). Ve většině experimentů dokonce příznaky extrahované z hloubkové mapy dosahují mírně lepších výsledků než tradiční obrazové. Výjimku představuje pouze LBPTOP, kde je maximum pro hloubková data o 8 % nižší než pro video.



(a) (b)

Obrázek 4.3: Porovnání slovní přesnosti [%] dosažené v této práci (barevně zvýrazněné sloupce) s vybranými články od jiných autorů na databázích OuluVS (a) a CUAVE (b).

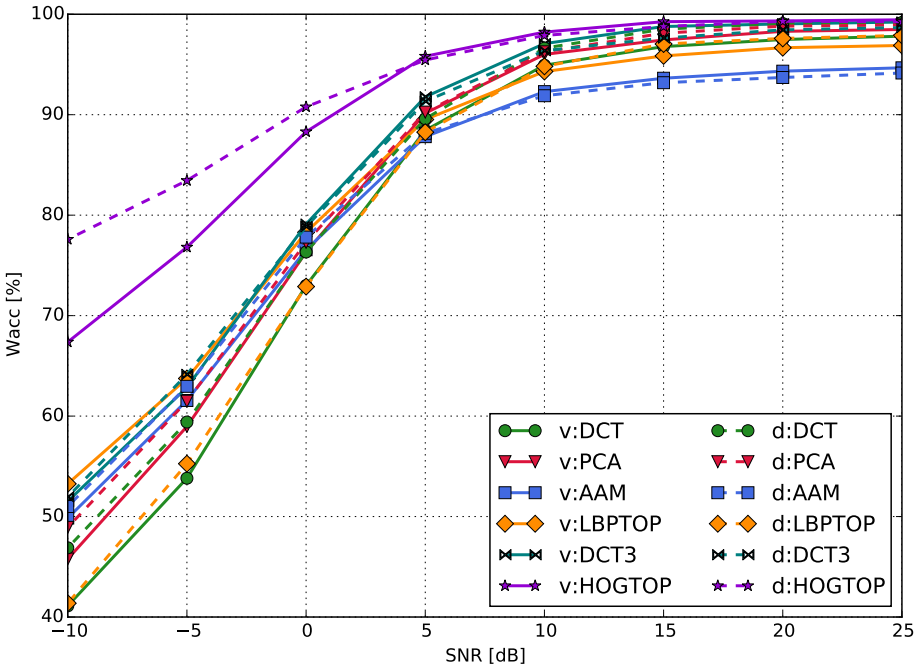
4.1.3 Srovnání se stavem poznání

Výsledky v podobě slovní přesnosti **na databázi OuluVS** z této a jiných vybraných prací přehledně shrnuje graf 4.3a. Doposud nejlepších výsledků 89,7 % na databázi OuluVS dosáhli Pei a kol. [Pei 2013], jejichž metoda je však uzpůsobená pouze na rozpoznávání izolovaných jednotek a způsob využití v reálném systému společně s akustickými příznaky není zcela zřejmý. S podobným problémem se potýkají i zbylé práce. Zde bylo dosaženo příznaky HOGTOP úspěšnosti pouze 85,7 %, tedy o 4 % méně ($\delta_{\mathrm{WER}} = +35$ %). Ovšem střední fúzí PCA, LBPTOP a HOGTOP úspěšnost vzrostla až na **89,9 %**, navíc za použití HMM a tedy s výhodou jednoduché aplikovatelnosti pro rozpoznávání spojité řeči.

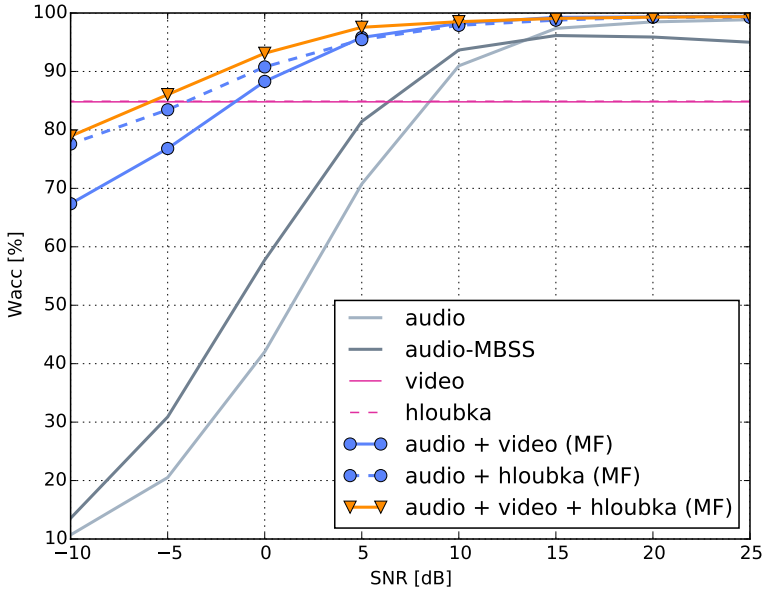
Srovnání výsledků **na databázi CUAVE** představuje poněkud obtížnější úkol. Na rozdíl od OuluVS se autoři neshodují na způsobu rozdělení dat a protokolu testování a kromě všech uvedených nesrovnalostí z předchozí databáze tak do výsledků vstupuje další zdroj variability v podobě poměru trénovací a testovací (příp. validační) množiny. Výsledky jsou opět shrnuty v grafu 4.3b, přičemž rozdělení dat dle mluvčích je uvedeno před skóre v hranatých závorkách. V [Saenko 2006] byli mluvčí rozdělení do trénovací, validační a testovací množiny v poměru 22:6:6. Nejlepšího výsledku 91,2 % bylo dosaženo v této práci, ovšem užitím LDA-dynamizovaných příznaků LBPTOP navržených Zhao a kol. [Zhao 2009]. Aplikací vlastní HOGTOP bylo dosaženo pouze 85,6 % slovní přesnosti, stále je to však v porovnání se stavem poznání nejlepší výsledek.

4.2 Audiovizuální rozpoznávání v hlučném prostředí

Provedeny byly i experimenty posuzující přínos obrazové informace v hlučných prostředích. Využita byla pouze vlastní databáze TULAVD, jejíž slovník obsahuje 50 položek oproti pouhým 10 v databázích OuluVS a CUAVE. Jelikož byla databáze TULAVD nahrána v relativně tichém prostředí, byl hluk k původnímu signálu přičítán uměle a s různou intenzitou tak, aby se odstup signálu od šumu (Signal to Noise Ratio,



Obrázek 4.4: Audiovizuální rozpoznávání izolovaných slov na databázi TULAVD v prostředí s hlukem typu babble pro různé vizuální příznaky.



Obrázek 4.5: Audiovizuální rozpoznávání s příznaky HOGTOP v prostředí s hlukem typu babble.

SNR) pohyboval v intervalu $[-10, 25]$ dB s krokem 5 dB. Jako zdroj hluků posloužila databáze NOISEX [Varga 1992], která obsahuje různé typy hluků. Pro demonstraci přínosu vizuálních příznaků v hlučném prostředí byly využity dva z nich: bílý šum (šum s plochým spektrem) a hluk typu babble, který simuluje prostředí s hlasy na pozadí. Ve všech experimentech byly zvukové nahrávky parametrizovány 13 kepstrálními koeficienty MFCC (včetně nultého koeficientu) a jejich delta a akceleračními odvozeninami. Výsledky pro hluk typu babble prezentuje graf 4.4, pro bílý šum a tovární hluk jsou výsledky k dispozici v hlavním textu práce. Váhy dvoukanálového HMM byly nastaveny napevno tak, aby aby maximalizovaly úspěšnost při SNR = 5 dB. Graf 4.5 pak porovnává audiovizuální rozpoznávání s rozpoznáváním na zvukově opravených nahrávkách pomocí metody vícepásmového spektrálního odečítání (Multi-Band Spectral Subtraction, MBSS). Grafu opět potvrzují přínos hloubkových příznaků, které při kombinaci s obrazovými daty zvyšují slovní přesnost pro nejnižší SNR až o 10 %.

5. Audiovizuální rozpoznávání spojité řeči

Druhá část experimentů je zaměřena na (audio-)vizuální rozpoznávání řeči na základě hláskových modelů. Vzhledem k množství trénovacích audiovizuálních dat jsou v této práci pro rozpoznávání spojité řeči jako základní řečová jednotka použity pouze bezkontextové modely, tedy samotné fonémy (monofóny) a vizémy. Zároveň jsou ze stejných důvodů data rozdělena pouze na trénovací a testovací a z protokolu tak odpadá validace na samostatné množině. V experimentech byla využita pouze vlastní databáze TULAVD obsahující 100 vět v běžné češtině od každého z 54 mluvčích, viz sekci 3.1. V každém ze šesti rozdělení křížové validace tedy trénovací množina sestává ze všech spojitých promluv od 45 mluvčích s celkovým průměrným časem 4h:52min včetně dat pro modelování ticha a neřečových hluků. K natrénování hláskových HMM byla opět využita knihovna HTK, přičemž každý foném či vizém modeluje třístavový lineární HMM s n-komponentovou gaussovskou směsí (GMM) na každý stav. Parametrizace byly dynamizovány (Δ, LDA) dle experimentů z předchozí kapitoly a lineárně interpolovány na 100 Hz.

5.1 Rozpoznávání izolovaných slov

Jako předstupeň před dekódováním spojité řeči se první experiment zaměřuje na stejnou úlohu jako kapitola 4, tedy rozpoznávání izolovaných slov – zde však hláskovými, nikoliv celoslovními modely. Výsledky experimentu v podobě slovní přesnosti [%] prezentuje tabulka 5.1, kde pro celoslovní modely byl počet gaussovských komponent každého stavu HMM křížově validován ze dvou možností (1 nebo 2 komponenty na stav), viz kapitolu 4. Dle očekávání pro všechny parametrizace s výjimkou akustických MFCC došlo k výraznému zhoršení slovní přesnosti oproti celoslovním modelům, nejčastěji mezi 20–30 % (δ_{WER} +50 až +200 %). Zhoršení slovní přesnosti hláskových modelů není překvapivé, protože modelování i klasifikace krátkých hlásek představuje kvůli mnohem slabší rozlišitelnosti oproti celým slovům podstatně složitější úlohu. Zde se tak ukazuje jeden z klíčových nedostatků současného stavu poznání, kde se výzkumníci téměř výhradně soustředí na klasifikaci dlouhých, obvykle dobře charakterizovatelných a navíc izolovaných jednotek, čemuž uzpůsobují i návrh parametrizace a klasifikace. Modelovány jsou obvykle

Par.	Z	Celoslovní	Fonémové		Vizémové	
počet komp. GMM:		1/2	8	16	8	16
MFCC	a	99,8	99,5	99,8	97,4	98,0
DCT	v	72,5	42,6	42,8	42,4	43,9
	d	74,4	39,3	42,5	38,6	43,1
AAM	v	74,1	57,5	58,5	59,0	59,3
	d	75,2	54,1	55,0	55,3	56,6
LBPTOP	v	74,2	54,6	56,4	54,6	56,3
	d	64,3	48,7	47,4	45,3	48,2
DCT3	v	75,1	42,6	43,1	43,4	45,6
	d	70,3	45,1	47,0	45,4	47,6
HOGTOP	v	**86,4**	59,5	61,0	59,8	60,1
	d	84,4	56,6	58,3	56,6	57,7
DAAM	(v, d)	74,9	**62,0**	**64,6**	**63,0**	**64,7**

Tabulka 5.1: Výsledky (slovní přesnost v [%]) rozpoznávání izolovaných slov hláskovými modely.

celé promluvy a není tak zřejmé, jaké úspěšnosti by algoritmy dosahovaly pro menší, hůře diskriminovatelné jednotky.

5.2 Rozpoznávání spojité řeči

V této práci byly sestaveny celkem 4 různé bigramové (tedy $n = 2$) jazykové modely pro LVCSR lišící se velikostí základního slovníku od několika set do několika set tisíc slov. Přesná čísla udává tabulka 11.2 v hlavním textu práce. Slova byla vybrána dle jejich četnosti v trénovacím korpusu s tím, že všech 366 slov vyskytujících se testovacích datech bylo do slovníku povinně přidáno. Jako trénovací data posloužily textové korpusy nasbírané na Ústavu Informačních technologií a elektroniky (ITE) na Technické univerzitě v Liberci a určené pro rozpoznávání mluvené češtiny programem NanoDictate. Přibližně 60 GB textů bylo získáno z internetových vydání známých českých periodik a částečně také manuálním přepisem televizních a rádiových zpravodajství. Pro tvorbu modelů byla využita knihovna SRILM[1] [Stolcke 2002] ve verzi 1.7.1 se zapnutým Knesserovým-Nayovým vyhlazováním. Jelikož rozpoznávání spojité řeči v HTK pomocí programu `HVite` je příliš pomalé a `HDecode` pracuje výhradně s trifónovými modely, zvolil jsem pro další experimenty dekodér Julius[2] [Lee 2009], který podporuje akustické (vizuální) modely natrénované v HTK. Pro každý experiment jsou uvedena dvě čísla: slovní přesnost WAcc (4.1) a v závorce slovní správnost (Word Correctness), která na rozdíl od WAcc nezapočítává chybně vložená slova (inzerce) a její hodnota tedy vždy leží v intervalu $\langle 0, 1 \rangle$.

Tabulka 5.2 shrnuje výsledky audiovizuálního rozpoznávání střední fúzí pro vybrané

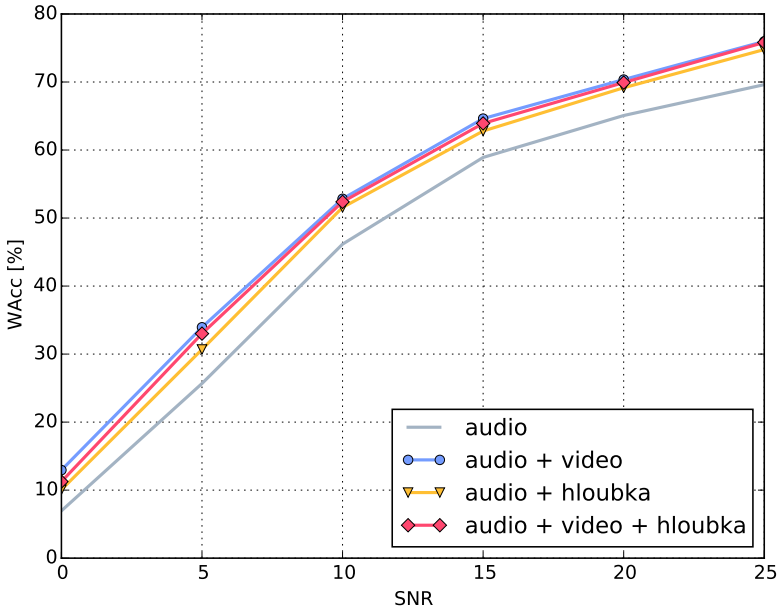
[1] `http://www.speech.sri.com/projects/srilm/`
[2] `http://julius.osdn.jp/en_index.php`

Par.	Z	Slovník							
		min		5k		50k		500k	
MFCC	a	74,0	(81,8)	55,9	(62,1)	43,9	(48,0)	36,3	(39,0)
AAM	$[a, v]^{\lambda}$	76,7	(82,2)	60,5	(64,2)	48,7	(50,5)	40,2	(41,8)
	$[a, d]^{\lambda}$	76,8	(82,3)	60,0	(63,8)	48,0	(50,2)	39,5	(41,0)
	$[a, v, d]^{\lambda}$	76,9	(82,2)	60,2	(64,0)	48,3	(50,6)	39,9	(41,4)
LBPTOP	$[a, v]^{\lambda}$	79,2	(84,1)	62,7	(66,3)	**50,1**	(52,3)	**41,7**	**(43,1)**
	$[a, d]^{\lambda}$	77,8	(82,3)	60,8	(64,3)	48,5	(50,6)	39,8	(41,1)
	$[a, v, d]^{\lambda}$	79,3	(83,6)	62,6	(66,0)	50,0	(52,2)	41,4	(42,8)
HOGTOP	$[a, v]^{\lambda}$	78,1	(83,2)	60,2	(63,7)	47,8	(50,5)	42,0	(43,9)
	$[a, d]^{\lambda}$	77,2	(82,2)	58,3	(62,6)	46,2	(48,8)	40,7	(42,6)
	$[a, v, d]^{\lambda}$	**79,4**	**(84,6)**	**62,9**	**(66,7)**	**50,1**	**(52,7)**	41,6	**(43,1)**
DAAM	$[a, (v, d)]^{\lambda}$	75,2	(81,4)	58,6	(62,9)	48,0	(50,2)	40,7	(42,7)

Tabulka 5.2: Audiovizuální rozpoznávání spojité řeči se střední fúzí akustických a vizuálních parametrizací pro různé slovníky a jazykové modely. Výsledky jsou uvedeny v [%] ve formátu „slovní přesnost (slovní správnost)“.

typy parametrizací v závislosti na slovníku. Výsledky vizuálního rozpoznávání a brzké audiovizuální integrace poskytují tabulky 11.3, resp. 11.4. Operace $[x, y]^{\lambda}$ zde značí střední fúzi příznakových vektorů z modalit x a y s vahami λ, jež byly křížově validovány tak, aby maximalizovaly slovní přesnost. Vyzkoušeny byly kombinace s $\sum \lambda^{(s)} = 1$ přes všechny kanály s (dvojice audio, video/hloubka) s krokem 0,1 a min. vahou pro audio $\lambda^{(\mathrm{a})} = 0,4$. Průměrná optimální váha $\lambda^{(\mathrm{a})}$ přes všechny parametrizace činila pro nejmenší slovník min $\lambda^{(\mathrm{a})} = 0,69$, pro půlmilionový 500k pak cca $\lambda^{(\mathrm{a})} = 0,74$. Díky optimálnímu nastavení vah při výpočtu výstupní stavové pravděpodobnosti nedochází současnou inkorporací obrazových i hloubkových dat k poklesu slovní přesnosti. Pro příznaky HOGTOP naopak úspěšnost roste, byť max. v řádu jednotek procent. Optimální poměr vah přitom lze interpretovat jako ukazatel významu a spolehlivosti jednotlivých kanálů. Pokles vizuálních vah $\lambda^{(\mathrm{v})} = 1 - \lambda^{(\mathrm{a})}$ pro větší slovník tak názorně demonstruje nižší míru užitečné informace obsažené ve vizuální složce, jež však ani pro velké slovníky a jazykové modely neztrácí svůj význam a relativně snižuje slovní chybovost až o 9 %. Výsledky jsou však přesto ovlivněny poměrně malým vzorkem testovacích dat a není tak zřejmé, nakolik by obrazová složka mohla být přínosná pro robustní akustické modely trénované na stovkách hodin dat.

Podobně jako v případě izolovaných slov byl proveden experiment s rozpoznáváním v hlučném prostředí i pro spojitou řeč. Do grafů 5.1 a 11.2 (v hlavním textu) je zanesena slovní přesnost dosažená při rozpoznávání spojité řeči v prostředí s hluky typu babble a bílý šum v rozmezí 0–20 dB pomocí navržených příznaků HOGTOP. Váha akustického kanálu byla empiricky nastavena na hodnotu 0,7, přičemž při kombinaci všech tří modalit (audio, video, hloubka) poměr činil 0,7 : 0,2 : 0,1. Rozpoznávání téměř čistého signálu (SNR odstup 25 dB) dosáhlo 70% slovní přesnosti na audio datech a 76 % na audiovizuálních datech a tedy integrace vizuální složky relativně snížila chybovost o 20 %. Pro signál zašuměný hlukem babble o relativní energii -10 dB oproti čistému klesla WAcc na 46 %, resp. 53 % a relativní zlepšení WER integrací vizuální složky tedy činilo cca



Obrázek 5.1: Audiovizuální rozpoznávání spojité řeči se slovníkem o velikosti 366 slov za použití příznaků HOGTOP v prostředí s hlukem typu babble.

13 %. Dodatečná integrace hloubkových příznaků již další zlepšení nepřinesla, naopak došlo spíše k mírnému zhoršení v řádu desetin až jednotek procent v závislosti na SNR.

6. Závěr

Předložená dizertační práce popisuje současný stav poznání v oblasti automatického audiovizuálního rozpoznávání řeči a odezírání ze rtů. Hlavní pozornost byla věnována parametrizaci vizuálního řečového signálu jakožto jedné z klíčových komponent problematiky. Spíše než z pohledu algoritmického byly metody rozděleny do skupin dle cílové aplikace a informace, které se snaží z obrazového signálu vytěžit. Pro případ odezírání z čelního pohledu se v současné době jeví jako nejnadějnější příznaky využívající dynamiku řeči, nejčastěji založené na promítání delších sekvencí do lineárních prostorů s lepší diskriminací pomocí metod grafového vnořování a strojového učení obecně. Naopak od klasických tvarově orientovaných či expertem stanovených parametrizací se spíše ustupuje. V tomto duchu se také vyvíjí problematika klasifikace, kdy především pro čistě vizuální odezírání ze rtů se obvykle využívají algoritmy specializované na cílovou klasifikaci a jí podřízenou automatickou extrakcí užitečné informace. Do značné míry tak již neplatí tradiční jasně oddělitelné schéma parametrizace a klasifikace, nýbrž se rozdíly mezi oběma fázemi zastírají. Výzkum ovšem v současnosti příliš neřeší, jak nové a sofistikované metody zacílené na vizuální rozpoznávání izolovaných jednotek zobecnit na spojitou řeč s velkým slovníkem a v kombinaci s akustickými příznaky.

Kromě obvykle zdůrazňované parametrizace obrazového signálu je v teoretické části v kapitole 2 hlavního textu rozebrána i problematika vizuálního předzpracování, především tzv. zarovnání obličeje a detekce zájmové oblasti, která má na výslednou úspěšnost rozpoznávání zásadní vliv. Výzkum v detekci obličejových částí se za posledních 10–15 let výrazně posunul vpřed, přičemž v současné době se největší pozornost soustředí na

diskriminační algoritmy lokalizace zájmových bodů na obličeji. Na rozdíl od tradičních optimalizačních metod se složitějším statistickým generativním modelem vzhledu jsou ty diskriminační obvykle méně časově náročné a díky metodám strojového učení dosahují vyšší přesnosti a spolehlivosti. Jeden z populárních algoritmů, explicitní tvarová regrese, byl pro detekci klíčových bodů na obličeji implementován i v této práci. Literatura AVSR se však pokrokům v oblasti zarovnání obličeje příliš nevěnuje. V kapitole 7 hlavního textu přitom bylo ukázáno, že v případě vizuálního odezírání dosáhly ve všech srovnávaných úlohách nejlepších výsledků systémy se sofistikovanou detekcí zájmové oblasti, přestože primárním cílem bylo vyhodnocení přínosu vizuální parametrizace či klasifikace. Naopak systémy s vizuálním předzpracováním navrženým ad hoc dosahovaly až o desítky procent horší slovní přesnosti.

V současnosti existuje poměrně velké množství volně dostupných audiovizuálních databází. Většina však obsahuje spíše malé množství řečníků (do 20) nebo je omezena typem promluv. Jedním z cílů práce přitom bylo navrhnout parametrizaci vhodnou i pro rozpoznávání spojité řeči s velkým slovníkem, ne pouze pro jednoduché systémy s několika málo izolovanými slovy (např. číslovky či jednoduché fráze). V rámci této práce jsem proto vytvořil vlastní audiovizuální databázi TULAVD s celkem 54 mluvčími, kteří namluvili přibližně 5,5 hodiny dat v podobě izolovaných slov a spojité řeči s neomezeným slovníkem. Databáze byla navržena s ohledem na využití hloubkových dat pro automatické odezírání ze rtů. K nahrávání proto byly využity dvě kamery a Microsoft Kinect, jenž problémy spojené s rekonstrukcí disparitní mapy řeší interně a jeho využití je tak mnohem snazší než implementace a odladňování metod stereovidění. Kromě audiovizuálních dat databáze obsahuje i 583 trojic obrázků (levá webkamera, RGB video, hloubková mapa) různých obličejových výrazů s manuálně vyznačenými 93 klíčovými body na tváři, které slouží pro trénování modelů vizuálního předzpracování.

Před provedením experimentů byl navržen testovací protokol tak, aby srovnání různých druhů parametrizace bylo vypovídající. Hlavním cílem bylo zamezit přeučení, tedy optimalizaci volných parametrů na testovací data, a z něj vycházející optimistickou zaujatost, viz sekci 9.2 v hlavním textu. Jako kompromis mezi statistickou relevancí a výpočetní náročnosti byla zvolena zjednodušená varianta vnořené křížové validace, která proces učení modelu, ladění parametrů a testování opakuje pro několik možných rozdělení dat. Protokol však bohužel nemohl být dodržen ve všech experimentech. Při experimentální evaluaci na databázi OuluVS byl zvolen postup s ohledem na kompatibilitu se stavem poznání, aby výsledky bylo možné přímo porovnat. V experimentech s rozpoznáváním spojité řeči pak přistoupeno ke standardní k-blokové křížové validaci s ohledem na množství trénovacích dat.

Jedním z hlavních přínosů práce je návrh vlastní parametrizace. V práci byly představeny tři nové parametrizace, trojrozměrná bloková DCT (DCT3), prostoročasový histogram orientovaných gradientů (HOGTOP) a kombinovaný hloubkový aktivní vzhledový model (DAAM). Zatímco hlavním cílem DCT3 a HOGTOP je využití řečové dynamiky jakožto důležité diskriminační informace, DAAM je navržen s cílem zahrnout do extrakce hloubková data. Všechny tři parametrizace dosáhly v experimentech s rozpoznáváním izolovaných slov dobrých výsledků, avšak jako jednoznačně nejkvalitnější se ukázala HOGTOP. Nejlepšího výsledku bylo s touto parametrizací dosaženo na databázích TULAVD a OuluVS, na CUAVE vyššího skóre dosáhly LBPTOP a DCT3, avšak ve všech případech byla slovní přesnost stále nad aktuálně nejlepším výsledkem

83 % WAcc v práci [Papandreou 2009]. Na databázi OuluVS byl stav poznání překonán pouze o 0,2 %. V článku [Pei 2013] autoři uvedli 89,7 % WAcc pro rozpoznávání založené kombinaci několika typů parametrizací, zde bylo střední fúzí příznaků PCA, LBPTOP a HOGTOP dosaženo 89,9 %, při použití pouze HOGTOP 85,5 %.

Téměř přehlížená se v literatuře zdá být problematika audiovizuálního LVCSR. Jak bylo uvedeno výše, obvykle se pracuje s celými promluvami jako nejmenšími řečovými jednotkami, díky čemuž si udržují dostatek diskriminační informace a hledání optimální příznakové projekce je tak algoritmicky zajímavější. Bohužel se však navržené metody stávají obtížně využitelné v systémech s bohatším slovníkem a v kombinaci s akustickou parametrizací. V této práci byl systém navržen s ohledem na využití i v LVCSR, jemuž se věnuje kapitola 11 hlavního textu. Testované parametrizace byly vyhodnoceny pro 4 různé slovníky s velikostí od 366 do 500 000 slov. Stanovením vhodných vah MSHMM bylo integrací vizuálních dat dosaženo pro parametrizace AAM, LBPTOP, HOGTOP a DAAM zlepšení o 1 až 6 % aboslutně a to jak pro RGB video, tak pro hloubková zdrojová data. Zlepšení se přitom projevilo pro všechny slovníky a jim odpovídající jazykové modely, z čehož lze dovodit přínos vizuální složky pro rozpoznávání běžného jazyka s neomezeným slovníkem. Nejlepších výsledků ve většině případů dosáhla parametrizace LBPTOP, pouze při střední fúzi audia, videa a hloubky byla slovní přesnost vyšší pro HOGTOP. Jako nevhodné se pro LVCSR ukázaly parametrizace založené na DCT (včetně DCT3) a PCA, jejichž aplikací došlo ke zhoršení výsledků v porovnání s MFCC. Obecně se však rozdíly mezi jednotlivými parametrizacemi oproti rozpoznávání izolovaných slov řádově snížily. Stěžejní roli nejen z hlediska vah jednotlivých kanálů MSHMM totiž v LVCSR hrají akustické příznaky společně s jazykovým modelem a variabilita ve vizuální parametrizaci se proto neprojeví v takové míře. Např. mezi AAM a LBPTOP činil rozdíl ve slovní přesnosti pro největší slovník pouze cca 2 %.

Část experimentů se věnovala vyhodnocení přínosu hloubkových dat pro vizuální a audiovizuální rozpoznávání řeči. Experimentálně bylo ukázáno, že hloubková data nesou podobné množství informace jako RGB video. Rozpoznávání izolovaných slov na základě příznaků extrahovaných z hloubkové mapy dosahovalo relativně vůči RGB ekvivalentu o 10 % horší až 2 % lepší slovní přesnosti, což odpovídá -7% až +38% relativní změně WER. Modality však lze kombinovat skrze MSHMM, čímž se výsledné skóre zlepší o 1–5 % absolutně, resp. 3–30 % relativně z hlediska chybovosti WER. Přínos hloubkové mapy lze tedy spatřit především v její částečné komplementaritě vůči obrazovým datům. Obdobně se hloubkově orientované parametrizace chovaly i v úloze LVCSR, kde však rozdíly potažmo přínos byly méně výrazné z důvodů uvedených v předchozím odstavci.

6.1 Souhrn hlavních přínosů práce

Pro lepší přehlednost je uveden následující seznam, který shrnuje nejdůležitější přínosy této dizertační práce. V práci byly

- navrženy tři typy vizuální parametrizace vhodné pro rozpoznávání izolovaných slov i spojité řeči: trojrozměrná bloková DCT, prostoročasový histogram orientovaných gradientů a hloubkově rozšířený aktivní vzhledový model,
- demonstrován přínos integrace hloubkových dat pro vizuální i audiovizuální rozpoznávání řeči,

- jednotnou a vůči přeučení robustní metodikou srovnány nerozšířenější typy parametrizací na více audiovizuálních databázích v úloze rozpoznávání izolovaných jednotek,
- vyhodnocen přínos vizuální složky i v obvykle přehlíženém audiovizuálním rozpoznávání spojité řeči s velkým slovníkem,
- sestavena středně rozsáhlá audiovizuální databáze TULAVD s 54 mluvčími obsahující RGB video a hloubková data.

6.2 Budoucí práce

Z krátkodobého pohledu mezi potenciální směry dalšího výzkumu patří např. automatická extrakce příznaků pomocí hlubokých neuronových sítí, jež je v současnosti populární především v počítačovém vidění. Hluboké neuronové sítě se nabízí pro využití také při integraci akustických a vizuálních příznaků či jako alternativa ke gaussovské směsi ve skrytých markovských modelech. Zřejmě největší překážku k jejich plnému využití představuje nutnost rozsáhlé (audio–)vizuální databáze, jejíž tvorba je časově velmi náročný úkol. Pro co možná největší vypovídající hodnotu by databáze ideálně měla obsahovat data z různých zdrojů a řečníky v různé relativní pozici vůči kameře. Aby se mohly audiovizuální systémy skutečně prosadit v praxi jako rozšíření stávajících akustických dekodérů, musí umožňovat modulární způsob integrace. Z dlouhodobého hlediska by se proto měl výzkum soustředit především na pozdní integraci, jež umožní trénovat vizuální modely nezávisle na akustických, nevyžadujíc stejná data a subslovní jednotku. Pro rozpoznávání by pak bylo možné využít vizémy, ovšem pouze za předpokladu správné vizémové transkripce, nikoliv jednosměrným přemapováním fonémů, viz sekci 11.1 v hlavním textu.

Seznam publikovaných prací

[Paleček 2014a] Karel Paleček. *Comparison of Depth-based Features for Lipreading.* In Proceedings of Telecommunications and Signal Processing (TSP) conference, Berlin, Germany, str. 648–651, 2014 (IEEE Xplore).

[Paleček 2014b] Karel Paleček. *Extraction of Features for Lip-reading Using Autoencoders.* In Proceedings of the International Conference on Speech and Computer (SPECOM), Novi Sad, Serbia, 2014 (WoS, SCOPUS).

[Paleček 2013] Karel Paleček a Josef Chaloupka. *Audio-visual speech recognition in noisy audio environments.* In Telecommunications and Signal Processing (TSP), 36th International Conference on, str. 484–487, 2013 (SCOPUS, IEEE Xplore).

[Červa 2012] Petr Červa, Jan Silovský, Jindřich Žďánský, Ondřej Smola, Karel Blavka, Karel Paleček a Jan Nouza. *Browsing, Indexing and Automatic Transcription of Lectures for Distance Learning.* In Proceedings of IEEE conf. on Multimedia Signal Processing (MMSP), Banff, Canada, str. 198–202, 2012 (WoS, IEEE Xplore).

[Paleček 2012a] Karel Paleček. *Detection of Similar Advertisements in Media Databases.* In Lecture Notes in Computer Science, Springer-Verlag Berlin, vol. 6800, str. 178–184, 2012 (WoS, SCOPUS).

[Paleček 2012b] Karel Paleček, David Gerónimo a Frédéric Lerasle. *Pre-attention cues for person detection.* In Proceedings of the 2011 international conference on Cognitive Behavioural Systems (COST'11), Springer-Verlag, Berlin, Heidelberg, str. 225–235, 2012 (SCOPUS).

[Červa 2011a] Petr Červa, Karel Paleček, Jan Silovský a Jan Nouza. *An Investigation into VTLN for Improved Transcription of Czech Broadcast Programs.* In Proceedings of 53rd International IEEE Symposium ELMAR-2011, Zadar, Croatia, str. 201–204, 2011 (SCOPUS, IEEE Xplore).

[Červa 2011b] Petr Červa, Karel Paleček, Jan Silovský a Jan Nouza. *Using Unsupervised Feature-Based Speaker Adaptation for Improved Transcription of Spoken Archives.* In Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011), Florence, Italy, str. 2565 – 2568, 2011 (WoS, SCOPUS).

[Hnilička 2010] Ondřej Hnilička, Jiří Málek, Karel Paleček a Zbyněk Koldovský. *A Fast C++ Implementation of Time-domain Blind Speech Separation Algorithm.* In Proceedings 20th Czech-German Workshop on Speech Processing, Prague, 2010.

Literatura

[Cao 2012] Xudong Cao, Yichen Wei, Fang Wen a Jian Sun. *Face alignment by explicit shape regression.* In in CVPR, 2012.

[Dalal 2005] N. Dalal a B. Triggs. *Histograms of oriented gradients for human detection.* In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, str. 886–893 vol. 1, June 2005.

[Lee 2009] Akinobu Lee a Tatsuya Kawahara. *Recent Development of Open-Source Speech Recognition Engine Julius.* 2009.

[Papandreou 2009] G. Papandreou, A. Katsamanis, V. Pitsikalis a P. Maragos. *Adaptive Multimodal Fusion by Uncertainty Compensation with Application to Audiovisual Speech Recognition.* IEEE Trans. on Audio, Speech and Language Process., vol. 17, no. 3, str. 423–435, March 2009.

[Patterson 2002] E.K. Patterson, S. Gurbuz, Z. Tufekci a J. Gowdy. *CUAVE: A new audio-visual database for multimodal human-computer interface research.* In Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on, volume 2, str. II–2017–II–2020, May 2002.

[Pei 2013] Yuru Pei, Tae-Kyun Kim a Hongbin Zha. *Unsupervised Random Forest Manifold Alignment for Lipreading.* In IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013, str. 129–136, 2013.

[Saenko 2006] Kate Saenko a Karen Livescu. *An Asynchronous DBN for Audio-Visual speech Recognition.* In SLT, str. 154–157, 2006.

[Stolcke 2002] Andreas Stolcke. *SRILM – an extensible language modeling toolkit.* In Proceedings of ICSLP, volume 2, str. 901–904, Denver, USA, 2002.

[Varga 1992] A. Varga, H. J. M. Steeneken, M. Tomlinson a D. Jones. *The NOISEX–92 Study on the Effect of Additive Noise on Automatic Speech Recognition.* Technical Report, DRA Speech Research Unit, 1992.

[Viola 2001] Paul Viola a Michael Jones. *Robust Real-time Object Detection.* In International Journal of Computer Vision, 2001.

[Young 2006] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev a P. C. Woodland. The HTK book, version 3.4. Cambridge University Engineering Department, Cambridge, UK, 2006.

[Zhao 2009] Guoying Zhao, Mark Barnard a Matti Pietikäinen. *Lipreading With Local Spatiotemporal Descriptors.* IEEE Transactions on Multimedia, vol. 11, no. 7, str. 1254–1265, 2009.