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

Dizertační práce

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

TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics
and Interdisciplinary Studies

The Institute of Information Technology and Electronics

Audiovisual Speech Recognition by Utilizing Methods for Automatic Lipreading

Dissertation

Study programme: P2612 Electrotechnics and informatics
Study branch: 2612V045 Technical cybernetics
Author: Ing. Karel Paleček
Supervisor: doc. Ing. Josef Chaloupka, Ph.D.

Prohlášení

Byl jsem seznámen s tím, že na mou dizertační práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména §60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci (TUL) nezasahuje do mých autorských práv užitím mé dizertační práce pro vnitřní potřebu TUL.

Užiji-li dizertační práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti TUL; v tomto případě má TUL právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Dizertační práci jsem vypracoval samostatně s použitím uvedené literatury a na základě konzultací s vedoucím dizertační práce a konzultantem.

Datum:

Podpis:

Poděkování

Děkuji svému školiteli doc. Ing. Josefovi Chaloupkovi, Ph.D. za užitečné konzultace, rady a připomínky k této práci a především za jeho trpělivost.

Děkuji také za podporu projektu SGS Technické Univerzity v Liberci s názvem „Aplikace zpracování číslicových signálů a návrh elektronických obvodů“ v letech 2013–2015.

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.

Text je rozdělen do 12 kapitol. Kapitoly 2–5 rozebírají současný stav problematiky rozdělením na několik dílčích podproblémů. V kapitole 2 je uveden přehled algoritmů pro zarovnání obličeje a detekce zájmové oblasti. Největší pozornost je věnována parametrizaci vizuálního signálu v kapitole 3. Následující kapitoly 4 a 5 popisují metody klasifikace a možnosti integrace vizuální informace do akustických řečových dekodérů. Přehled nejčastěji využívaných audiovizuálních databází je uveden v kapitole 6. Rešeržní část práce je uzavřena kapitolou 7, která porovnává nejlepší doposud dosažené výsledky v dostupné literatuře. Samostatně jsou posouzeny vizuální a audiovizuální systémy a navíc je problematika rozdělena dle typu rozpoznávaných promluv a závislosti na mluvčích. Zohledněn je rovněž vliv vizuálního předzpracování.

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 8, směřuje především k využití řečové dynamiky a 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 9. Databáze 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 10 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. Druhá část experimentů v kapitole 11 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.

The text of the dissertation consists of 12 chapters. Chapters 2–5 present the current state of the art and each focuses on one specific subproblem of visual and audiovisual speech recognition. Chapter 2 investigates methods for face alignment and detection of the region of interest. Commonly used features and algorithms of their extraction are examined in chapter 3, followed by an overview of classification methods in chapter 4, fusion of multiple sources of information in chapter 5, and existing audiovisual datasets in chapter 6. The first part of the thesis examining the state of the art is summarized in chapter 7, which compares currently the best results achieved on various commonly used datasets with respect to recognition grammar, vocabulary size, speaker dependency and visual preprocessing.

Three different robust visual parametrizations are proposed and explained in chapter 8: 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 9, 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 that minimizes the risk of overfitting when tuning the decoder.

Experiments in chapter 10 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 11.

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

Obsah

Seznam zkratek

AAM	Active appearance model (Aktivní vzhledový model)
ASM	Active shape model (Aktivní tvarový model)
ASR	Automatic speech recognition (Automatické rozpoznávání řeči)
AVSR	Audiovisual speech recognition (Audiovizuální rozpoznávání řeči)
CV	Cross validation (Křížová validace)
DAAM	Depth-extended Active Appearance Model (Hloubkový AAM)
DBN	Dynamic Bayesian net (Dynamická bayesovská síť)
DCT	Discrete cosine transform (Diskrétní kosinová transformace)
EI	Early integration (Brzká integrace)
EM	Expectation maximization
ESR	Explicit shape regression (Explicitní tvarová regrese)
GMM	Gaussian mixture model (Model gaussovské směsi)
HiLDA	Hierarchical linear discriminant analysis (Hierarchická LDA)
HMM	Hidden Markov model (Skrytý markovský model)
HOG	Histogram of oriented gradients (Histogram orientovaných gradientů)
HOGTOP	Histogram of oriented gradients from three orthogonal planes (Prostoročasový HOG)
KFCV	k-fold cross validation (k-násobná křížová validace)
LBP	Local binary pattern (Lokální binární vzor)
LBPTOP	Local binary pattern from three orthogonal planes (Prostoročasový LBP)
LDA	Linear discriminant analysis (Lineární diskriminační analýza)
LI	Late integration (Pozdní integrace)
LOOCV	Leave one out cross validation (Křížová validace vynech jeden)
LVCSR	Large vocabulary continuous speech recognition (Rozpoznávání spojité řeči s velkým slovníkem)
MF	Middle fusion (Střední fúze pomocí MSHMM)
MFCC	Mel frequency cepstral coefficients (Melovské kepstrální koeficienty)
MS	Multi-speaker (Víceuživatelský systém)
MSHMM	Multistream synchronous hidden Markov model (Vícekanálový synchronní HMM)
PCA	Principal component analysis (Analýza hlavních komponent)
ROI	Region of interest (Oblast zájmu)
SD	Speaker-dependent (Systém určený pro jednoho uživatele)
SI	Speaker-independent (Systém nezávislý na mluvčím)
SNR	Signal-to-noise ratio (Odstup signálu od šumu)
SVM	Support vector machine (Metoda podpůrných vektorů)
WAcc	Word accuracy (Slovní přesnost)
WCorr	Word corectness (Slovní korektnost)
WER	Word error rate (Slovní chybovost)

1. Úvod

Moorův zákon, tedy empirické pravidlo, dle něhož se počet tranzistorů umístitelných na integrovaný obvod při zachování stejné ceny přibližně zdvojnásobí každých 18 měsíců, stále od svého uvedení v roce 1965 do značné míry platí. Nárůst cenově dostupného výkonu především od uvedení osobních počítačů PC s procesory z rodiny x86 umožnil prudký rozvoj umělé inteligence, strojového učení, robotiky a automatizovaného zpracování velkých dat. Běžně dnes probíhá monitoring veřejného komunikačního prostoru, zejména internetu, s cílem vytěžit co nejvíce užitečných dat např. pro bezpečnostní či reklamní účely. Jazyková bariéra mezi lidmi z různých koutů světa částečně padá, jelikož i na běžném chytrém mobilním telefonu s přístupem k internetu lze zapnout překlad z mikrofonu v reálném čase. Roboti komunikují s člověkem a jsou schopni porozumět jeho povelům. V tzv. chytrých domácnostech mohou lidé hlasovými povely ovládat některé prvky, např. osvětlení či zábavní domácí centrum. Díky automatickému přepisu přednášek a výukových videí mohou studenti na školách a na internetu snáze vyhledávat informace. V neposlední řadě automatické titulkování internetových videí napomáhá neslyšícím ve vzdělávání a zábavě.

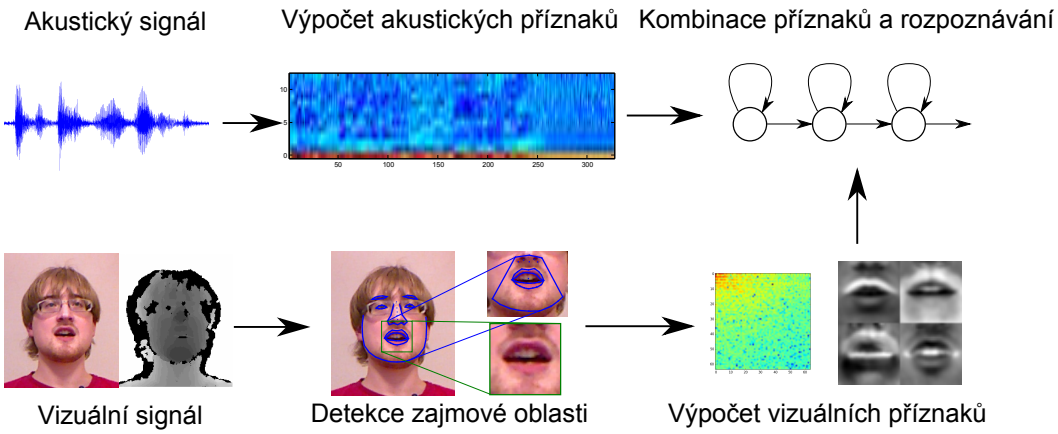
Jednu z hojně využívaných technologií, která všechno toto umožňuje, představuje automatické rozpoznávání řeči (Automatic Speech Recognition, ASR), tedy proces porozumění mluvené řeči umělou inteligencí. Jeho výstup je obvykle reprezentován jako textový přepis akustické nahrávky pokud možno ve srozumitelné a gramaticky bezchybné formě. Aby systém pro rozpoznávání mluvené řeči mohl dosahovat uspokojivé úspěšnosti, je pro jeho natrénování nezbytně nutné dostatečné množství dat, typicky stovky hodin nahrávek reprezentujících obě pohlaví, široké spektrum věkových skupin a za různé akustické podmínky. Jedním z pracovišť vyvíjejících komerčně úspěšný software NanoDictate pro automatické rozpoznávání češtiny a dalších slovanských jazyků je Ústav informačních technologií a elektroniky na Technické Univerzitě v Liberci. NanoDictate byl aplikován např. v projektu ministerstva kultury, jež si klade za cíl automatický přepis rádiového vysílání od 20. let minulého století až po současnost a momentálně se jeho úspěšnost pohybuje okolo 80 % v závislosti na stáří a kvalitě zvukové nahrávky [Nouza 2014]. Při přepisu běžných televizních a rádiových pořadů s nevýrazným hlukem na pozadí se úspěšnost pohybuje okolo 85 %. Jednu z nejjednodušších aplikací představuje diktát, kde za použití obecného akustického modelu a velkého slovníku se stovkami tisíc slov slovní přesnost dosahuje až 95 %. Naopak v obtížné úloze automatického přepisu přednášek, kde je akustický signál zarušen hluky na pozadí, max. úspěšnost dosahuje pouze 40–70% slovní přesnosti v závislosti na míře zarušení [Šeps 2014].

V oblasti ASR existuje řada podoborů, které se snaží využít specifických podmínek reálných aplikací za účelem snížení chybovosti. Pokud je např. vyvíjen diktovací systém pro lékaře, v každé jednotlivé ordinaci bude využíván pouze jedním uživatelem a zřejmě není tedy nutně zapotřebí co nejobecnější akustický model – výhodnější může být model přizpůsobit pro dané potřeby. Tzv. adaptaci akustického modelu lze samozřejmě využít i pro specifickou věkovou skupinu či

pohlaví, ne pouze pro jediného uživatele. Kromě akustického modelu lze adaptovat i model jazykový, tedy omezit množinu slov a větných konstrukcí, kterou je systém schopen rozeznat. Dle nároků na gramatickou korektnost lze také výstup dodatečně zpracovat a opravit případné chyby. Těmito a dalšími problémy se zabývá zpracování přirozeného jazyka (Natural Language Processing, NLP). Rozpoznávání robustní vůči nepříznivým akustickým podmínkám řeší oblast zvaná zvýrazňování řeči (angl. speech enhancement), jejíž cílem je očistit nahrávku od hluku na pozadí a ponechat pouze čistou řeč.

Podobný problém jako zvýrazňování řeči, avšak odlišným způsobem, pak řeší **audiovizuální rozpoznávání řeči** (Audio-Visual Speech Recognition, AVSR), které se místo cílené redukce nežádoucí informace v signálu naopak 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. Mezi výzkumníky pravděpodobně nejznámější demonstrací významu vizuální složky se od svého uvedení v roce 1976 stal tzv. McGurkův jev [McGurk 1976]. Experiment spočívá v informačním konfliktu mezi akustickou nahrávkou a videem řečníka. Posluchačům bylo přehráno video s řečníkem vyslovujícím VCV (Vowel-Consonant-Vowel, samohláska-souhláska-samohláska) sekvenci „aga“, avšak v doprovodné zvukové stopě znělo „aba“. Na takto pozměněné audiovizuální nahrávce pak většina posluchačů „slyší“ posloupnost „ada“, což jednoznačně dokazuje vliv vizuální složky na proces porozumění mluvené řeči u lidí. Další podobnou hláskovou konfigurací byla např. „aka“ (video), „apa“ (audio), resp. „ata“ (vjem posluchačů). Na práci navázal Summerfield [Summerfield 1987], který jev vysvětlil hypotézou VPAM (Visual: Place, Auditory: Manner). Podle ní jsou akustická a vizuální složka vzájemně komplementární. Zatímco vizuální složka dodává informaci o *místě* artikulace hlásky (place), tedy např. rty (bilabiála), zuby (dentála), či jazykem (alveolára), akustická složka informuje posluchače o *způsobu* artikulace (manner), tedy např. zněle, nezněle či nosově. Vizuální složku lze pro odezírání využít i samostatně a nezávisle na akustickém signálu, ovšem pouze s omezujícími podmínkami a malým slovníkem, např. pro jednoduché hlasové povely či v systémech pro vizuální verifikaci. Především vzhledem k variabilitě ve způsobu artikulace, kdy někteří lidé velmi zřetelně hýbou ústy, zatímco jiní spíše mumlají, však video nelze považovat za informačně plnohodnotnou alternativu k akustickému signálu. Avšak ani zřetelná artikulace by sama o sobě nedostačovala, jelikož mnoho informace podstatné pro porozumění řeči vzniká uvnitř artikulačních orgánů, lidskému oku či běžné kameře zakrytých.

S audiovizuálním rozpoznáváním řeči a odezíráním ze rtů jakožto hlavními tématy této dizertační práce úzce souvisí i několik dalších oborů. Mezi ně patří např. audiovizuální identifikace a verifikace, tedy oblast patřící pod biometrické ověřování identity např. pro bezpečnostní účely, nebo rozpoznávání pohlaví, věku, či emocí za účelem přirozenějších reakcí a citlivější komunikace stroje (např. robota) a člověka. Do jisté míry opačnou oblast výzkumu představuje audiovizuální syntéza řeči (Text-To-Speech Synthesis, TTS), jež oproti klasické akustické



Obrázek 1.1: Princip audiovizuálního rozpoznávání.

syntéze pracuje navíc grafickým modelem lidské tváře s cílem pomoci s porozuměním řeči především sluchově postiženým. Přestože zmíněným tématům se tato práce nevěnuje, mnoho dílčích problémů s problematikou audiovizuálního rozpoznávání řeči sdílejí a poznatky zde uvedené tak mají širší dosah.

1.1 Úloha audiovizuálního rozpoznávání řeči

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.

Vzhledem k frekvenční charakteristice lidského hlasu je obvykle akustický signál vzorkován s frekvencí 16 kHz a 16 bitů a dále segmentován na 10–25 ms dlouhé překrývající se stacionární úseky (framy), přičemž každý z těchto framů je parametrizován vektorem příznaků. V dnešní době se úlohou výběru optimální sady příznaků výzkum již příliš nezabývá. Ve velké většině systémů tvoří parametrizaci kepstrální příznaky Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Codes (LPC), či Perceptual Linear Prediction (PLP), jejichž přínos byl v průběhu let experimentálně ověřen. Obvykle se k těmto příznakům připojují jejich odvozeniny, tzv. delta a delta-delta (akcelerační) příznaky, které částečně zachycují řečovou dynamiku. Typický akustický příznakový vektor MFCC pak obsahuje 39 koeficientů, tj. 12-MFCC, 12-ΔMFCC, 12-$\Delta\Delta$MFCC a P, ΔP, $\Delta\Delta$P, kde P označuje energii vstupního signálu. Po výpočtu příznaků následuje odečítání kepstrálního průměru (Cepstral Mean Subtraction, CMS), které odstraňuje rozdíly ve střední hodnotě kepstra u signálů pocházejících z různých zdrojů. Proces výpočtu akustických

příznaků lze modifikovat např. adaptací na specifického řečníka či algoritmy zvýrazňování řeči pro potlačení šumu a ruchů na pozadí (např. slepá separace signálů, beamforming, ...). Přehled používaných příznaků a technik automatického rozpoznávání řeči nabízí např. [Huang 2001].

Zdrojem vizuálního signálu je videozáznam promluvy řečníka. Výzkum automatického odezírání ze rtů se soustředí nejčastěji na případ, kdy nahrávka zachycuje obličej řečníka z čelního pohledu [Heckmann 2001, Heckmann 2002b, Matthews 2002, Scanlon 2003, Saenko 2005, Lan 2009]. Tento zjednodušující předpoklad především významně zjednodušuje detekci zájmové oblasti v obraze a zároveň tak minimalizuje vliv vnějších zdrojů variability na celý proces rozpoznávání. Čelní pohled také zachycuje podstatnou část vizuální informace obsažené v promluvě. Množstvím informace zachycené bočním pohledem se zabývají např. práce [Lucey 2006a, Iwano 2007, Kumar 2007, Saitoh 2010], přičemž v článku [Kumar 2007] bylo automatickým rozpoznáváním z profilového pohledu dosaženo dokonce lepší úspěšnosti než u člověka. Závislost úspěšnosti automatického odezírání ze rtů na úhlu pohledu byla zkoumána v pracích [Lan 2012, Bowden 2013], přičemž v obou bylo dosaženo nejvyššího skóre pro 30° natočení kamery.

Jelikož v reálných aplikacích však může být podmínka fixního pohledu splněna jen obtížně, soustředí se výzkum i na automatické odezírání ze rtů nezávislé na natočení řečníkovy hlavy. Toho lze dosáhnout např. geometrickou transformací oblasti zájmu, která obraz narovná zpět do čelního pohledu [Lucey 2007, Lucey 2008, Lan 2012]. V článku [Pass 2010] byl naopak použitý přístup na úrovni příznaků, kdy byly vybírány koeficienty DCT tak, aby se minimalizoval jejich rozptyl v závislosti na úhlu pohledu.

Pro snížení závislosti příznaků na pozici a natočení řečníkovy hlavy lze rovněž využít více kamer a metod stereovidění [Loy 2000, Petr Císař 2004, Vorwerk 2010]. Kromě nezávislosti na relativní pozici kamery a hlavy je možné tímto způsobem rekonstruovat trojrozměrný povrch tváře a tím získat dodatečnou informaci nad rámec dvourozměrného obrazu. Problémem stereovidění je však značná citlivost na světelné podmínky a výpočetní náročnost, která pro běh v reálném čase vyžaduje implementaci na grafickém procesoru (GPU). Vizuální rozpoznávání řeči založené na využití více kamer tak zůstává spíše na okraji zájmu. Před několika lety se však na trh dostalo několik cenově dostupných zařízení, která implementují rekonstrukci disparitní/hloubkové mapy hardwarově s využitím infračerveného spektra, čímž do značné míry eliminují uvedené problémy. Příklady těchto zařízení jsou Microsoft Kinect, Asus Xtion či Creative Senz3D, všechna založená na referenčním návrhu PrimeSense[1]. Microsoft Kinect byl pro odezírání ze rtů úspěšně použitý v pracích [Galatas 2012, Yargic 2013]. K pořízení videozáznamu promluvy řečníka bylo v některých článcích využito také infračervené spektrum [Huang 2004].

Na rozdíl od akustického rozpoznávání řeči je při automatickém odezírání ze rtů vždy nezbytná fáze zpracování obrazu a detekce zájmové oblasti (Region of Interest, ROI). Úkolem je extrahovat tu část obrazu, kde se nachází většina informace

[1] `http://www.souvr.com/Soft/UploadSoft/201005/20100506172950 50.pdf`

spojené s řečníkovou promluvou. K tomu je nutné co možná nejpřesněji odhadnout především pozici a tvar úst. V AVSR literatuře jsou přitom nejrozšířenější tři základní způsoby: klasické metody založené na barevné segmentaci obrazu, posuvné okno využívající haarovské příznaky a Aktivní Vzhledový Model (Active Appearance Model, AAM). Metody založené na barevné segmentaci využívají barevné odlišnosti lidské pokožky a rtů oproti pozadí, přičemž nejčastěji zvýrazňují červenou složku některého z barevných prostorů (RGB, YCbYr, ...) [Lievin 1998]. Jejich zjevnou nevýhodou je závislost na barvě pozadí a pleti řečníka a také nasvětlení scény. Před více než deseti lety si ve velkou popularitu získaly metody založené na posuvném okénku s využitím haarovských příznaků. Nejznámějším příkladem je algoritmus Violy a Jonese [Viola 2001], který pracuje na principu vyčerpávajícího prohledávání obrázku a porovnání podobnosti každé podoblasti s naučeným vzorem, např. koutku úst. Pro detekci zájmové oblasti v úloze automatického odezírání ze rtů byla tato metoda použita např. v [Lucey 2007, Lucey 2008, Fu 2008, Zhao 2009, Zhou 2010]. Hodí se však spíše pouze na hrubý odhad pozice obličeje či některé z jeho částí. Pro zpřesnění odhadu pozice a tvaru úst jsou vhodnější metody pro zarovnání obličeje (angl. face alignment), které na obličeji detekují pozice tzv. klíčových bodů. Ty odpovídají např. rtům, nosu, očím apod. Nejznámějším zástupcem této kategorie metod je Aktivní vzhledový model. Vizuální příznaky extrahované na základě přesné pozice klíčových obecně dosahují lepších výsledků rozpoznávání [Matthews 2002, Lan 2009]. Kromě AAM však existuje celá řada mnohem efektivnějších metod, jež ovšem nebývají ve výzkumu automatického odezírání ze rtů příliš často využívány. Metody detekce a zarovnání obličeje jsou detailněji rozebrány v kapitole 2.

Zřejmě nejaktivnější oblastí výzkumu audiovizuálního rozpoznávání řeči je extrakce vizuálních příznaků. Hlavní motivace pro výpočet vizuálních příznaků spočívá v získání užitečné informace ze vstupního signálu. Z geometrického pohledu jde o transformaci vstupního vektoru hodnot do méně rozměrného prostoru, ve kterém se neprojevuje variabilita způsobená nežádoucími vlivy, tj. např. změnami v osvětlení či relativní pozicí kamery a obličeje. Příznaky v redukovaném prostoru by měly zachycovat pouze změny související s řečí a tedy co nejlépe odpovídat původní informaci vznikající při procesu tvorby vizuálního signálu před modifikací způsobenou externími vlivy. Příznaky lze hrubě kategorizovat do tří skupin:

1. příznaky extrahované z přibližně lokalizované oblasti zájmu,
2. příznaky odvozené z přesné pozice a tvaru obličejových částí,
3. příznaky využívající dynamiku řeči.

Toto dělení však samozřejmě není zcela jednoznačné a některé z algoritmů mohou spadat do více než jedné kategorie.

První kategorie metod nejčastěji využívá obdélníkovou oblast kolem úst, jenž je lokalizována pouze přibližně. Jelikož typickým rozměrem bývá 64×64 pixelů, tedy 4096 jasových hodnot v případě šedotónového obrázku, používá se

obvykle pro další zpracování signálu některá z metod redukce dimenze. Např. v práci [Bregler 1994] inspirované rozpoznáváním tváří byly takto extrahovány příznaky eigenlips, získané analýzou hlavních komponent. Velmi oblíbenou metodou používanou v experimentech jako baseline je též diskrétní kosinová transformace [Potamianos 2001b, Heckmann 2002b, Lan 2009]. Kromě uvedených metod byly pro redukci dimenze rozpoznávání použité také diskrétní Fourierova transformace [Duchnowski 1994], vlnkové transformace [Yu 1999, Puviarasan 2011], či lineární diskriminační analýza [Lan 2010].

Nevýhodou příznaků založených na hrubě lokalizované oblasti zájmu je nemožnost využití přesného tvaru úst. Tuto informaci lze využít pouze nepřímo skrze metody redukce rozměru dat. Příznaky odvozené z pozice klíčových bodů na tváři se snaží tento nedostatek odstranit a využít tvar úst přímo ve svém návrhu. Samozřejmě tím však vznikají vyšší nároky na přesnost a spolehlivost algoritmů pro detekci obličejových částí. Jednou z nejjednodušších metod je popis pohybu rtů pomocí jejich šířky, výšky a zaokrouhlení [Potamianos 1998b, Císař 2006]. Velmi oblíbenou metodou je použití parametrů AAM jako vizuální příznaky [Matthews 2002, Pitsikalis 2006, Lan 2009]. AAM příznaky využívají jak tvarovou, tak obrazovou informaci a dosahují vysoké úspěšnosti rozpoznávání.

Kromě statických příznaků se výzkum soustředí také na využití časových závislostí vizuálního řečového signálu. Příznaky tedy nemusí být nutně vypočteny pouze ze statického obrázku, je možné využít informaci o změně mezi jednotlivými snímky video signálu. Nejjednodušším způsobem je podobně jako v případě akustického rozpoznávání výpočet Δ a $\Delta\Delta$ koeficientů nad statickými příznaky, tj. rozdíl prvého či druhého řádu mezi sousledných snímky [Chaloupka 2008]. De facto standardem a zdaleka nejčastěji využívanou metodou dynamizace je pak redukce příznakových sekvencí pomocí lineární diskriminační analýzy [Matthews 2002, Lan 2010, Galatas 2012]. Dynamika však může být zohledněna přímo již v návrhu příznaků, ne pouze jako dodatečné zpracování statických příznaků. Např. v práci [Zhao 2009] autoři segmentovali nahrávky do překrývajících se úseků, přičemž na každém z nich vypočítali rozšířenou variantu lokálních binárních vzorů, která porovnává sousedící pixely i v časové ose. V [Ong 2011] byly pomocí boostingu extrahovány sekvence jednoduchých binárních příznaků. Vysoké úspěšnosti dosahují také v současné době populární metody tzv. manifold learningu, což je skupina algoritmů pro nelineární redukci rozměru dat. Tyto metody byly aplikované pro modelování časové závislosti sekvence vizuálních příznaků v článcích [Zhou 2010, Pei 2013, Zhou 2014]. Podrobně se parametrizací vizuálního signálu zabývá kapitola 3.

Posledními kroky automatického rozpoznávání řeči jsou fúze akustických a vizuálních příznaků a jejich klasifikace. Fúze může probíhat na dvou základních úrovních: příznaková a rozhodovací, někdy také označované jako brzká, resp. pozdní integrace. Fúze na příznakové úrovni probíhá před samotnou klasifikací. Nejjednodušší metodou kombinace obou kanálů je prosté vektorové spojení příznaků, čímž vznikne jediný hypervektor audiovizuálních koeficientů. V práci [Matthews 2002] byl tento vektor redukován lineární diskriminační analýzou. Složitější způsob

redukce na bázi hlubokých neuronových sítí byl aplikován v [Ngiam 2011]. Druhým způsobem integrace je oddělené zpracování i klasifikace obou kanálů, přičemž finální přiřazení sekvence příznakových vektorů k některé ze slovníkových položek dochází až na základě klasifikačního skóre obou kanálů. Podle toho, jakým způsobem je skóre vyjádřeno, se také volí pravidlo integrace. V případě, kdy je výstupem obou klasifikátorů pravděpodobnost, je nejčastějším integračním pravidlem jejich vážený součin. Lze ale také použít součtové pravidlo, které může za určitých okolností podávat lepší výsledky. Porovnáním různých integračních strategií se zabývá práce [Lucey 2005].

V současnosti je zdaleka nejpopulárnější metodou klasifikace skrytý Markovský model [Bregler 1994, Heckmann 2002b, Matthews 2002, Pitsikalis 2006, Lucey 2008, Lan 2009], který navíc umožňuje hybridní metodu fúze akustického a vizuálního kanálu. V některých pracích však byla pro rozpoznávání použita i metoda Support Vector Machines (SVM) [Zhao 2009, Ngiam 2011].

Především kvůli nárokům na velikost audiovizuální databáze a její manuální zpracování se literatura nejčastěji soustředí na rozpoznávání pouze izolovaných slov, např. jednoduchých hlasových povelů [Chaloupka 2008]. Částečně se směrem k rozpoznávání spojité řeči vydali autoři práce [Pachoud 2008], kde byly rozpoznávány číslovky v nahrávkách spontánních promluv. Rozpoznáváním delších úseků ve formě frází se zabývají práce [Zhao 2009, Ong 2011], i zde jsou však fráze nejmenší jednotkou rozpoznávání a nejedná se tak o plnohodnotnou spojitou řeč. Rozpoznáváním menších slovních jednotek (vizémů) se také zabývá např. práce [Zhou 2014], avšak bez ověření přínosu v systémech s větším slovníkem. V článku [Lan 2010] byly rozpoznávány věty na základě vizémových modelů s celkem 1000 slovy ve slovníku. Audiovizuální rozpoznávání spojité řeči za s využitím velkého slovníku a jazykových modelů stále není příliš rozšířené, výjimky představují např. práce [Potamianos 2003, Lucey 2008]. Spojitému audiovizuálnímu rozpoznávání češtiny se jako první ve své dizertační práci věnoval Petr Císař [Císař 2006], ovšem pouze s relativně malým slovníkem (344 slov).

1.2 Cíle dizertační práce

Cílem této 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 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. Detekce obličejových částí

Velmi důležitou [Potamianos 2004] součástí systému pro automatické odezírání ze rtů je zpracování obrazu a detektor zájmové oblasti (Region of Interest, ROI). Oproti fázi předzpracování akustického signálu představuje porozumění obrazu složitější úlohu z několika důvodů. Největší problém způsobuje variabilita vizuálního signálu, která závisí na mnoha faktorech jako jsou osvětlení, relativní pozice kamery a obličeje řečníka, stáří, pohlaví, barva pleti a rtů, vousy na tváři, kvalita snímacího zařízení, či artefakty způsobené kompresí videozáznamu. Úspěšnost rozpoznávání je pak kriticky závislá na návrhu algoritmu pro odstranění této variability z obrazového signálu tak, aby následně extrahované příznaky zachycovaly pouze změny týkající se řečové informace, nikoliv externích vlivů.

2.1 Barevná segmentace

Základní způsob detekce obličeje reprezentují metody založené na barevné odlišnosti lidské pokožky a zbytku obrazu. Segmentace je nejčastěji formulována jako klasifikace jednotlivých pixelů do dvou tříd: kůže versus pozadí. Pro snadnější rozlišení barev jsou pak obvykle využívány převody z RGB do různých barevných prostorů, např. HSV [Kjeldsen 1996], YCbCr [Althoff 2003], či CIE LUV [Yang 1998].

Obvykle probíhá klasifikace pixelů prahováním zpětné projekce histogramu. Např. v [Chaloupka 2005] je histogram vypočten na trénovací databázi obsahující manuálně segmentované oblasti kůže. Využita je k tomu červená Cr složka barevného prostoru YCbCr. Histogram hodnot v složky Cr je aproximován normálním rozdělením $p(v) = \mathcal{N}\left(v; \hat{\mu_k}, \sigma_k^2\right)$ se střední hodnotou $\hat{\mu_k}$ a rozptylem σ_k^2. Pixely $f(i,j)$ jsou pak označeny jako „pleťové“, pokud $p(f(i,j)) \geq 0.0456$, přičemž hranice prahu je stanovena experimentálně[1]. Pro odstranění děr a získání finální pozice obličeje je binární obraz filtrován morfologickými operacemi otevření a uzavření. Příklad prahování znázorňuje obr. 2.1.

Podobný postup barevné segmentace byl aplikován i např. v práci [Yang 1998], kde namísto YCbCr byl obraz převeden do prostoru CIE LUV. Využity byly přitom dvě složky U a V, tj. histogram a jeho aproximace normálním rozdělením

[1] Zvolená hodnota odpovídá prahování obrázku do intervalu $\hat{\mu_k} \pm 2\hat{\sigma_k}$.

Obrázek 2.1: Prahování Cr složky vstupního obrazu. Převzato z [Chaloupka 2005].

byly dvourozměrné. Algoritmus rovněž umožňoval detekci více obličejů, kdy každá celistvá oblast, jež obsahovala více než 70 % pleťových pixelů, byla označena jako nalezená tvář.

Pro zvýšení robustnosti vůči nepřesně anotovaným hranicím v trénovací databáze byl v článku [Kjeldsen 1996] namísto histogramu uveden tzv. barevný predikát. Jedná se o datovou strukturu podobnou histogramu, avšak inkrementováno je více buněk zároveň. Pro výpočet byly rovněž použity i negativní vzorky, tj. pixely náležící pozadí. Pro každý pleťový pixel byla odpovídající buňka barevného predikátu a její nejbližší okolí inkrementována o hodnotu pixelu váženou Gaussovým rozdělením se středem v této buňce. Pro pixely náležející pozadí byly buňky v barevném predikátu naopak dekrementovány, avšak s nižšími hodnotami a rozpětím gaussovských vah. Na rozdíl od předchozích prací byly také využity všechny složky barevného prostoru HSV, přičemž odstín H a sytost S byly kvantovány jemněji než jasová složka V. Tento postup minimalizoval počet chybných klasifikací a dosáhl lepších výsledků než pouhé hledání optimální hodnoty pro prahování zpětné projekce histogramu.

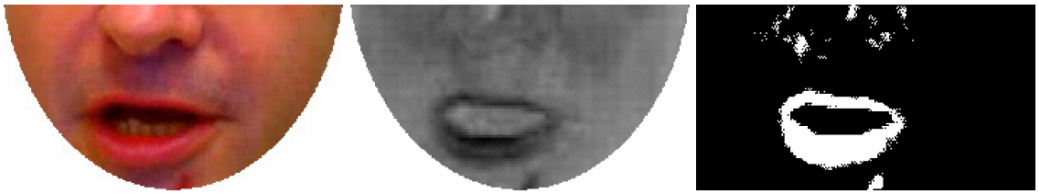
Pozice a tvar rtů lze určit podobným způsobem, tedy využitím barevných odlišností rtů a pokožky obličeje. Využívá se přitom především červená složka barevného spektra. Tu lze získat přímo z Cr složky YCbCr prostoru, či jako nelineární transformaci prostoru RGB. Např. v článku [Lievin 1998] byla použita transformace prostoru RGB na složky

$$H = 256 \times \frac{G}{R}, \quad I = \frac{R+G+B}{3}, \tag{2.1}$$

přičemž složka H byla následně prahována experimentálně stanovenými prahem. Rty byly detekovány také na základě jejich pohybu, a proto byly vypočteny rozdílový obrázek složek I dvou následujících snímků a odchylka složky H oproti prvnímu snímku sekvence. Na základě těchto příznaku byla každému pixelu přiřazena aposteriorní pravděpodobnost v rámci Markovského náhodného pole (Markov Random Field, MRF).

Transformaci do optimálního barevného prostoru lze také odhadnout ze statistických vlastností zkoumaných dat, tj. pixelů náležejících rtům a obličejové pokožce. Není tedy nutné používat některý z existujících barevných prostorů, jejichž podmnožina složek odpovídá svými charakteristikami barvě rtů spíše náhodou. Vhodnou metodou pro tento účel se jeví být **lineární diskriminační analýza** (Linear Discriminant Analysis, LDA). V případě hledání optimální barevné transformace tvoří datové vektory tři hodnoty složek R, G a B a pixely jsou označeny jako pleťové nebo náležející rtům. Výsledkem LDA transformace barevného prostoru RGB je jediná složka $F = (R, G, B) \cdot \boldsymbol{w}_0$, ve které se histogramy retních a pleťových pixelů maximálně odlišují.

V práci [Chaloupka 2005] byl vyhlazený histogram složky F aproximován váženým součtem dvou gaussovských funkcí. Odhad byl proveden metodou nejmenších čtverců, přičemž pro optimalizaci kvadratického kritéria byla zvolena metoda největšího spádu. Na základě aproximace byl zvolen optimální práh pro



Obrázek 2.2: Prahování F složky vstupního obrazu. Převzato z [Chaloupka 2005].

klasifikaci pixelů a výsledný obraz filtrován morfologickými operacemi. Oblast rtů pak byla detekována na základě celkového počtu do ní náležejících bodů. Příklad segmentace rtů uveden na obr. 2.2. V práci [Kaucic 1998] byla pro klasifikaci pixelů F složky použita Bayesova metoda maximální aposteriorní pravděpodobnosti.

Výhodou metod založených na barevné segmentaci obličeje a rtů je nezávislost na natočení řečníkovy hlavy. Algoritmy jsou navíc koncepčně jednoduché a z hlediska výpočetní složitosti nenáročné. Nevyužívají však příliš často apriorní informaci o relativních pozicích rtů a jiných obličejových částí a jsou tak velmi náchylné ke generování tzv. falešných alarmů, tedy oblastí nesprávně označených jako zájmové. Významnou nevýhodu představuje také přílišná závislost na světelných podmínkách. V neposlední řadě jsou často hodnoty prahů stanovovány ad hoc či pouhým subjektivním odhadem a není tedy zřejmé žádné objektivní kritérium, které by zvolené hodnoty optimalizovaly. Z těchto důvodů si postupem času oblibu získaly metody založené na strojovém učení a robustní klasifikaci.

2.2 Posuvné okno

V současnosti nejpopulárnějším způsobem automatické detekce lidské tváře v obraze jsou metody založené na posuvném okénku (angl. sliding window). Základní myšlenkou je vyčerpávající prohledávání všech obdélníkových podoblastí ve vstupním obrázku a posouzení, zda obsahují či neobsahují obličej. Zdaleka nejpoužívanějším a de facto standardem se pak od svého uvedení v roce 2001 stal kaskádní detektor Violy a Jonese (VJ) [Viola 2001].

Algoritmus VJ sestává ze tří klíčových součástí: efektivní výpočet příznaků pomocí součtového (integrálního) obrazu, trénování a kombinace slabých klasifikátorů pomocí boostingu a kaskádní klasifikace okének. Základní myšlenka detektoru spočívá v kombinaci velkého počtu tzv. slabých klasifikátorů pomocí adaptivního boostingu (AdaBoost) [Freund 1997]. Boosting představuje soubor hladových algoritmů, které postupně kombinují jednoduché klasifikátory tak, aby chybovost výsledné skupiny byla nižší než chybovost kteréhokoliv z jednotlivých klasifikátorů. V každé iteraci je vybrán právě jeden klasifikátor, který v kombinaci s předešlými vykazuje nejlepší rozpoznávací skóre (nejnižší chybovost). V případě Adaboostu má výsledná kombinace T slabých klasifikátorů $h_t : \mathbb{R}^d \rightarrow \{-1, +1\}$

podobu

$$C(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t h_t(\boldsymbol{x}), \tag{2.2}$$

kde $\boldsymbol{x} \in \mathbb{R}^d$. Koeficient α_t se odvíjí od chybovosti klasifikátoru h_t, která se počítá dle

$$\varepsilon_t = \sum_{n=1}^{N} d_n^{(t)} \mathbb{1}\left\{y_n \neq h_t(\boldsymbol{x}_n)\right\}, \tag{2.3}$$

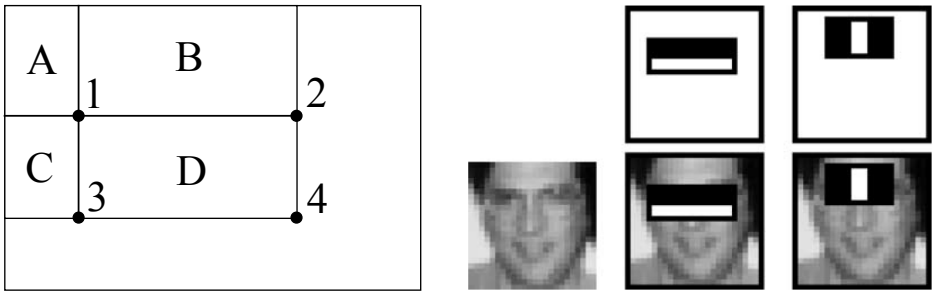
kde $d_n^{(t)}$ jsou váhy jednotlivých vzorků v iteraci t, $y_n \in \{-1, +1\}$ je třída vzorku $\boldsymbol{x}_n$ a $\mathbb{1}\{\cdot\}$ označuje funkci identity. Koeficient α_t se odvíjí od chybovosti ε_t tak, aby úspěšnější klasifikátory měly větší váhu ve výsledném kombinovaném klasifikátoru. Po každé iteraci jsou také upraveny váhy $d_n^{(t+1)}$ jednotlivých trénovacích vzorků $\boldsymbol{x}$ tak, aby se v dalších iteracích při výpočtu chybovosti proces více soustředil na nesprávně klasifikované vzorky.

V případě detekce lidské tváře algoritmem VJ jsou slabé klasifikátory tvořeny haarovskými příznaky f, jenž jsou znázorněny na obr. 2.3b. Pixely pod světlou oblastí se přičítají, pixely pod tmavou odečítají. Výsledný příznak je pak rozdílem těchto obdélníkových oblastí. Slabý klasifikátor má formu

$$h(x, f, p, \theta) = \begin{cases} +1 & pf(x) < p\theta \\ -1 & \text{jinak} \end{cases} \tag{2.4}$$

a porovná tedy hodnotu haarovského příznaku f s prahem θ. Polarita p určuje směr porovnání. Pro každý typ příznaku existuje v obrázku o rozměrech $N \times M$ celkem $N \cdot M \cdot (N-1) \cdot (M-1)/4$ dvojic levého horního a pravého dolního bodu, což např. pro obrázek 24×24 pixelů představuje cca 160000 dvojic. haarovské příznaky však lze vyhodnotit v konstantním čase pomocí tzv. součtového (integrálního) obrazu. Součtový obraz obsahuje v každém bodě (i, j) součet všech pixelů (i', j'), pro které $i' \leq i, j' \leq j$ a jde tedy o dvourozměrný kumulativní součet. Součet jasových hodnot v libovolně veliké obdélníkové oblasti lze vypočítat jako rozdíl 4 hodnot v součtovém obraze, jejich souřadnice odpovídají rohům oblasti. Viz obr. 2.3a, kde např. pozice 2 označuje součet pixelů obdélníků A a B. Součet pixelů obdélníka D lze spočítat jako $4 - 3 - 2 + 1$, kde čísla odpovídají pozicím na obrázku. Díky této vlastnosti je Adaboost schopný v trénovací fázi vyhodnotit velké množství slabých klasifikátorů h_t. Příznaky vybrané v počátečních iteracích trénování jsou znázorněny na obr. 2.3b. Např. první příznak odpovídá rozdílu jasových hodnot na tváři a kolem očí a zachycuje tak fakt, že oblast kolem očí je obvykle tmavší. Tímto způsobem je natrénováno několik tisíc klasifikátorů, dokud není dosaženo požadované max. chybovosti.

Počet obdélníkových okének je však příliš velký na to, aby se pokaždé vyhodnocovaly všechny příznaky. Pro běh v reálném čase tak detektor VJ využívá kaskádové klasifikace, kdy v každém kroku je vyhodnoceno pouze malé množství příznaků a v případě nízkého skóre se v klasifikaci dále nepokračuje. Takto je efektivně



(a) (b)

Obrázek 2.3: Součet oblasti pomocí integrálního obrazu (a) a nejlepší haarovské příznaky (b). Převzato z [Viola 2001].

odfiltrováno velké množství okének, jež neobsahují obličej a více času může být věnováno složitějším případům.

Na práci Violy a Jonese navázal např. Lienhart [Lienhart 2002], který rozšířil sadu haarovských příznaků o jejich varianty otočené o 45°. Další práce se věnovaly i jiným než čelním pohledům či použily jiné varianty boostingu [Li 2002]. Přehled současného stavu poznání v oblasti detekce obličeje popisuje článek [Zhang 2010].

Stejným způsobem jako obličej lze metodou posuvného okénka nalézt i jednotlivé obličejové části. Čím menší však daná oblast je, tím méně informace obsahuje a je tak i hůře rozlišitelná. Metody založené na posuvném okénku jsou tak vhodné spíše pro přibližný odhad pozice obličeje v obraze než pro přesnou lokalizaci jednotlivých obličejových částí.

2.3 Statistické modely vzhledu

V aplikacích, kde je nutná znalost pozice či tvaru např. očí či úst (např. monitorování bdělosti řidiče, odezírání ze rtů, ovládání PC pro postižené), metody založené na posuvném okénku nejsou dostatečně přesné. V takovýchto případech je pak nutné použít algoritmy pro tzv. zarovnání obličeje (angl. face alignment). Obvyklým způsobem formulace této úlohy je vyznačení předem daného fixního počtu klíčových bodů na obličeji a jejich následné hledání za nějakých omezujících podmínek. Tvar obličeje je zde vyjádřen jako vektor souřadnic

$$\boldsymbol{s} = (x_1, y_1, \ldots, x_v, y_v)^\top \tag{2.5}$$

kde každá z celkového počtu v dvojic (x_i, y_i) označuje pozici jednoho z předem definovaných klíčových bodů. Pokud jsou k dispozici obrázky s manuálně vyznačenými klíčovými body, je možné sestavit generativní statistický model [Cootes 2000], který umožňuje parametrický popis tvaru obličeje. Na základě tohoto modelu lze pak každému tvaru přiřadit určitou pravděpodobnost a tím významně omezit prostor, který je nutný při automatické lokalizaci klíčových bodů procházet. Zároveň takovýto model rovněž omezuje

Jedním z prvních statistických modelů vzhledu úspěšně aplikovaných pro odhad tvaru obličeje je **aktivní tvarový model** (Active Shape Model, ASM) [Cootes 1995]. Metoda ASM modeluje tvarovou variabilitu lidské tváře pomocí analýzy hlavních komponent (Principal Component Analysis, PCA) (viz sekci 3.1.2). Při trénovací fázi jsou nejprve všechny vektory $\boldsymbol{s}$ z databáze zarovnány Prokrustovou analýzou [Dryden 1998] do společného souřadného systému tak, aby výsledný model nezachycoval variabilitu způsobenou posunem, škálováním a rotací, nýbrž pouze tvarovými odlišnostmi v lidských tvářích. Trénování modelu spočívá ve výpočtu vlastního prostoru $\boldsymbol{S} = (\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$ množiny tvarových vektorů $\{\boldsymbol{h}_i\}_{i=0}^{N}$ metodou PCA a zachování pouze určitého procenta variability tvaru (typicky 95 %). Libovolný tvarový vektor $\boldsymbol{s}$ pak lze aproximovat modelem jako

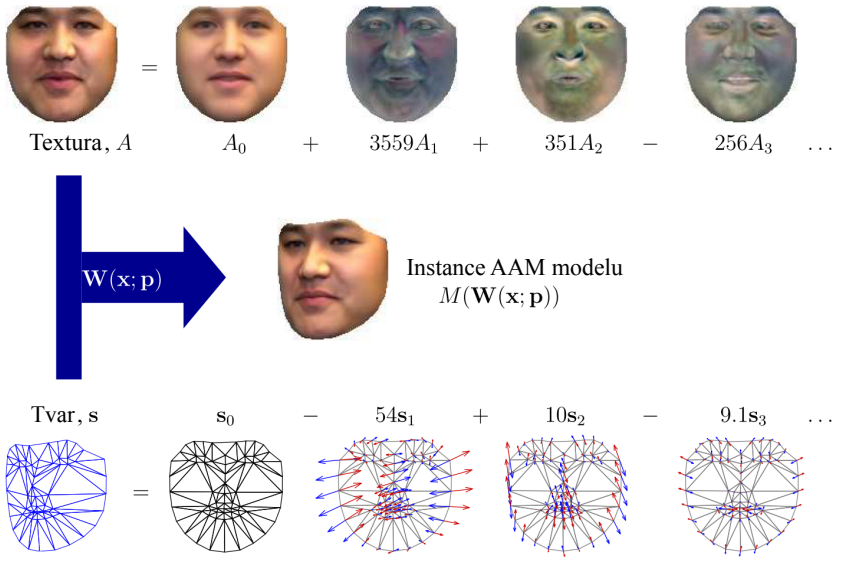
$$\boldsymbol{s} = \boldsymbol{s}_0 + \sum_{i=1}^{n} p_i \boldsymbol{s}_i, \tag{2.6}$$

kde $\boldsymbol{s}_0$ označuje aritmetický průměr všech tvarů v databázi a vektor $\boldsymbol{p} = (p_0, p_1, \ldots, p_n)$ tvoří parametrizaci tvaru. Jedná se tedy o lineární model variability, kde je každý vzorek $\boldsymbol{s}$ vyjádřen lineární kombinací bazických vektorů. Pro každý bod je navíc sestaven normalizovaný šedotónový profil podél úsečky kolmé vůči hraně modelu. Automatická lokalizace klíčových bodů (tj. zarovnání obličeje) pak probíhá iterativně střídáním dvou základních kroků. Nejprve se podél šedotónového profilu nalezne pro každý bod optimální pozice taková, která nejlépe odpovídá profilu zjištěném na trénovací sadě. Na základě nových pozic bodů jsou pak ze vztahu (2.6) v druhém kroku vypočteny nové parametry modelu $\boldsymbol{p}$ a omezeny na interval $|p_i| < 3\sigma_i$, kde σ_i je standardní odchylka asociovaná s i-tým módem variability $\boldsymbol{s}_i$ prostoru $\boldsymbol{S}$. Tento postup je opakován až do konvergence.

Problémem ASM je příliš jednoduchý model vzhledu okolí klíčových bodů, který není dostatečně reprezentativní a nezaručuje nalezení optimální pozice bodů. Nejznámějším a zdaleka nejpoužívanějším statistickým modelem vzhledu je **aktivní vzhledový model** (Active Appearance Model, AAM) [Cootes 1998, Matthews 2003], který tento problém řeší zahrnutím modelu variability textury lidské tváře, a tak zvyšuje diskriminační schopnost modelu. Trénování modelu AAM spočívá v trojnásobné aplikaci metody PCA, přičemž proces probíhá ve dvou hlavních fázích.

V první fázi trénování AAM jsou tvar a textura objektu modelovány odděleně. Tvar je modelován shodným způsobem jako u metody ASM. Textura objektu je reprezentována normalizovaným vektorem $A(\boldsymbol{x})$ pixelů $\boldsymbol{x}$, které se nacházejí uvnitř oblasti tvořené konvexním obalem klíčových bodů ve vektoru $\boldsymbol{s}_0$. Hodnoty pixelů jsou z této oblasti extrahovány geometrickou transformací, která je po částech afinní. Nad takto získanými množinami vektorů tvarů a textur je následně aplikována PCA. Tímto je umožněn lineární popis tvaru i textury každého objektu. Analogicky k (2.6) je libovolná textura $A(\boldsymbol{x})$ aproximována modelem jako

$$A(\boldsymbol{x}) = A_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i A_i(\boldsymbol{x}),\ \boldsymbol{x} \in \boldsymbol{s}_0 \tag{2.7}$$



Obrázek 2.4: Princip lineární variace tvaru a textury. Na obrázku je znázorněno, jak je z jednotlivých komponent vlastního prostoru sestaven výsledný vzhled objektu, v tomto případě lidské tváře. Převzato z [Matthews 2003].

kde $A_0(\boldsymbol{x})$ označuje průměrnou texturu, $A_i(\boldsymbol{x})$ značí vektory vlastního prostoru textur a vektor a $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \ldots, \lambda_m)$ tvoří parametrizaci textury $A(\boldsymbol{x})$. Jak je vzhled objektu reprezentován za předpokladu lineární variace tvaru a textury je znázorněno na obr. 2.4.

V druhé fázi trénování jsou tvar a textura kombinovány do jednoho statistického modelu. Toho je dosaženo třetí aplikací metody PCA nad vektory $\boldsymbol{p}$ a $\boldsymbol{\lambda}$. Před aplikací PCA je však nutné normalizovat energie obou parametrizací na stejnou hodnotu tak, aby nepřevážil vliv jedné z nich na úkor té druhé. Obvykle se aplikuje normalizační konstanta w_s spočítaná z poměru součtu vlastních čísel získaných z aplikace PCA v první fázi. Každý objekt je pak reprezentován lineárním modelem kombinovaného vzhledu

$$\boldsymbol{b} = \begin{bmatrix} w_s \boldsymbol{p} \\ \boldsymbol{\lambda} \end{bmatrix} = \boldsymbol{Q}\boldsymbol{c} \tag{2.8}$$

kde $\boldsymbol{Q}$ je vlastní prostor kombinace tvarových a texturových příznaků a vektor $\boldsymbol{c}$ parametrizuje kombinovaný vzhled. Výhodami kombinovaného vzhledu jsou dekorelace tvarových a texturových příznaků a při zachování pouze určitého procenta variability v datech (obvyklá hodnota je opět 95 % [Cootes 2000]) i další redukce celkového rozměru parametrizačního vektoru.

Automatickou lokalizaci bodů lze dosáhnout pomocí optimalizačních metod v rámci Lucas-Kanade frameworku pro zarovnávání obrázků. Minimalizováno je kvadratické kritérium popisující rozdíl mezi modelem generovanou texturou $A(\boldsymbol{x})$ (2.7) a texturou viděnou v obrázku $I\left(W\left(\boldsymbol{x}, \boldsymbol{p}'\right)\right)$ v závislosti na parametrech

$\boldsymbol{p}'$ a $\boldsymbol{\lambda}$, tj.

$$\boldsymbol{p}^*, \boldsymbol{\lambda}^* = \underset{\boldsymbol{p}', \boldsymbol{\lambda}}{\operatorname{argmin}} \left\| A_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i A_i(\boldsymbol{x}) - I\left(W\left(\boldsymbol{x}, \boldsymbol{p}'\right)\right) \right\|^2 \tag{2.9}$$

kde $\boldsymbol{p}' = (s, \theta, t_x, t_y, \boldsymbol{p})$ označuje spojení vektoru tvarových parametrů $\boldsymbol{p}$ a parametrů geometrické transformace podobnosti s, θ, t_x, t_y (posun, rotace, měřítko) a W je po částech afinní transformace, která slouží ke vzorkování pixelů v obrázku I. Finální pozice klíčových bodů v obrázku jsou rekonstruovány dosazením parametrů $\boldsymbol{p}^*$ do rovnice (2.6). V klíčové práci [Matthews 2003] je popsáno mnoho variant postupné optimalizace kritéria (2.9) pomocí Gauss-Newtonovy metody (a některých dalších), které se liší ve způsobu aktualizace parametrů $\Delta\boldsymbol{p}'$ potažmo linearizace (2.9) Taylorovým rozvojem. Jednu z variant představuje algoritmus Inverse Compositional (IC-AAM), který díky své efektivnosti umožňuje běh v superreálném čase (až stovky snímků za sekundu). Algoritmus IC-AAM, podobně jako některé další z rodiny Lucas-Kanade, však nelze použít pro optimalizaci parametrů AAM modelu s kombinovanými parametry $\boldsymbol{c}$. Jeho aplikace je možná, pouze pokud jsou tvarové a texturové parametry $\boldsymbol{p}$ a $\boldsymbol{\lambda}$ modelovány odděleně.

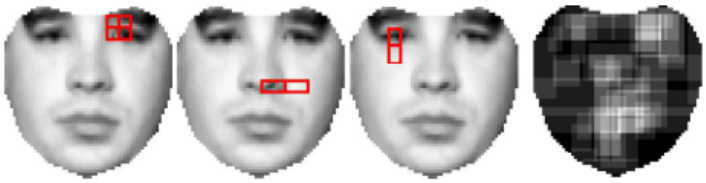
V práci [Liu 2007] byl pro lokalizaci klíčových bodů zvolen přístup inspirovaný detektorem Violy a Jonese. Posuzování podobnosti syntetizované a obrazové textury probíhalo pomocí kombinace slabých klasifikátorů haarovských příznaků s využitím součtových obrazů. Optimální tvarové parametry $\boldsymbol{p}^*$ byly odhadnuty maximalizací součtu odezev T slabých klasifikátorů, tj.

$$\boldsymbol{p}^* = \underset{\boldsymbol{p}}{\operatorname{argmax}} \sum_{t=1}^{T} \frac{2}{\pi} \operatorname{atan}\left\{ g_m \boldsymbol{A}_m^\top I\left(W(\boldsymbol{x}, \boldsymbol{p}') - t_m\right) \right\}, \tag{2.10}$$

kde $g_m = \pm 1$, t_m je prahová hodnota haarovského příznaku m. Součin $\boldsymbol{A}_m^\top I\left(W(\boldsymbol{x}, \boldsymbol{p}')\right)$ je jiná forma zápisu klasifikace textury $I\left(W(\boldsymbol{x}, \boldsymbol{p}')\right)$ vhodná pro výpočet gradientu (2.10). Kritérium bylo maximalizováno metodou největšího spádu (vzrůstu). Pro výběr optimálních příznaků autoři zvolili algoritmus GentleBoost, další z variant boostingu, přičemž negativní vzorky byly vygenerovány náhodným vychylováním jednotlivých tvarových parametrů $\boldsymbol{p}$ odpovídajících pozitivním vzorkům od své správné polohy. Na obrázku 2.5 jsou zobrazeny nejlepší tři příznaky vybrané algoritmem GentleBoost a nejlépe rozlišitelné oblasti na obličeji.

2.4 Lokální modely

Jednu z nevýhod Aktivního vzhledového modelu představuje závislost na osvětlení obličeje. Modelována je totiž textura celého povrchu obličeje, což při nedostatečném množství trénovacích dat způsobuje korelaci jinak nezávislých oblastí. Např. při osvětlení hlavy zprava bude levá strana obličeje tmavší. Pokud tento případ nenastal v trénovací databázi, stává se velmi obtížné nalézt optimální vzhledové parametry $\boldsymbol{\lambda}$. Pokud nastal pouze v několika málo případech, může shodou okolností korelovat



Obrázek 2.5: První tři příznaky vybrané GentleBoostem (vlevo) a histogram pozic nejlepších 50 příznaků (vpravo). Převzato z [Liu 2007].

s naprosto nesouvisejícími faktory, jako jsou např. úsměv či konkrétní obličej. Reprezentativnější trénovací databáze tento problém sice vyřeší, jenže zároveň zvětší složitost prostoru parametrů $\boldsymbol{\lambda}$. Tím vznikne více lokálních minim kritéria (2.9), což významně znesnadňuje jeho globální minimalizaci. Uvedené problémy se snaží odstranit lokální modely vzhledu.

V práci [Cristinacce 2006] byl představen tzv. **omezený lokální model** (Constrained Local Model, CLM), který na rozdíl od AAM modeluje variabilitu textury pouze v okolí klíčových bodů a tedy ne po celém povrchu tváře. Svojí koncepcí se tak nachází přibližně mezi aktivním tvarovým a aktivním vzhledovým modelem. Pro každý klíčový bod je vzhled modelován odděleně a nezávisle na ostatních. Během lokalizace je v okolí každého klíčového bodu vyhodnocena podobnost s naučeným vzorem a na základě této podobnosti pak určena pozice pro daný klíčový bod. Aby nedocházelo ke generování nesmyslných výsledných tvarů, sdružené pozice bodů jsou omezeny tvarovým modelem (2.6), podobně jako u ASM. Automatická lokalizace klíčových bodů je tedy formulována v prostoru tvarových parametrů jako

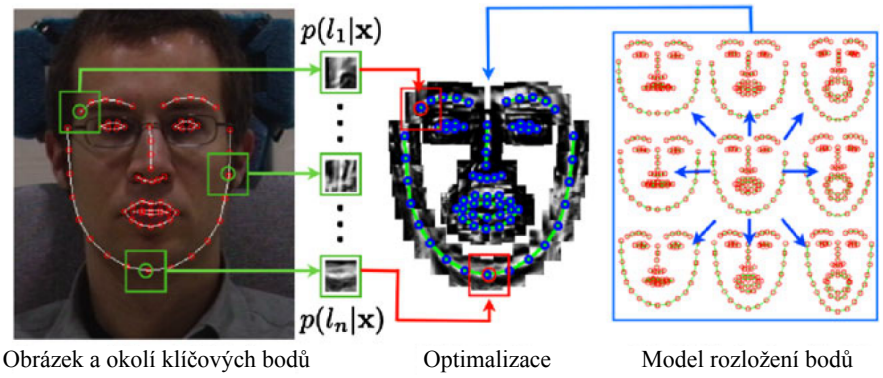
$$\boldsymbol{p}^* = \underset{\boldsymbol{p}}{\operatorname{argmin}}\, \mathcal{R}(\boldsymbol{p}) + \sum_{i=1}^{v} \mathcal{D}_i\left(\boldsymbol{x}_i\left(\boldsymbol{p}\right); \mathcal{I}\right), \tag{2.11}$$

kde $\mathcal{D}_i\left(\boldsymbol{x}_i\left(\boldsymbol{p}\right); \mathcal{I}\right)$ je rozdíl mezi okolím bodu $\boldsymbol{x}_i$ v obraze $\mathcal{I}$ a odpovídajícím naučeným vzorem a $\mathcal{R}(\boldsymbol{p})$ penalizuje odchylku od průměrného tvaru v modelu (2.6). Optimalizace probíhá střídáním dvou kroků: 1. nalezení optimální pozice pro každý klíčový bod a 2. optimalizace tvarových parametrů $\boldsymbol{p}$ na základě informace z předchozího kroku. Postup probíhá iterativně až do konvergence či dosažení maximálního počtu kroků. Proces znázorňuje obr.

Na práci [Cristinacce 2006] navázalo několik dalších, z nichž zřejmě nejvýraznější se stal článek [Saragih 2011]. Saragih zde formuloval pravděpodobnostní interpretaci úlohy (2.11) jako hledání parametrů maximalizací aposteriorní pravděpodobnosti (Maximum a posteriori, MAP):

$$\boldsymbol{p}_{\mathrm{MAP}} = \underset{\boldsymbol{p}}{\operatorname{argmax}}\, p\left(\boldsymbol{p}\right) \prod_{i=1}^{v} p\left(l_i = 1 \,|\, \boldsymbol{x}_i, \mathcal{I}\right), \tag{2.12}$$

kde $p\left(l_i = 1 \,|\, \boldsymbol{x}_i, \mathcal{I}\right)$ označuje pravděpodobnost, že i-tý klíčový bod je zarovnán na správné pozici a $p\left(\boldsymbol{p}\right)$ je apriorní pravděpodobnost tvarových parametrů $\boldsymbol{p}$. V



Obrázek 2.6: Ilustrace lokalizace bodů CLM. Převzato z [Saragih 2011].

případě uniformního rozdělení parametrů $\boldsymbol{p}$ se pak jedná o maximálně věrohodný odhad (Maximum likelihood, ML). PCA modelu tvarové variability (2.6) a penalizaci $\mathcal{R}(\boldsymbol{p}) = \|\boldsymbol{p}\|^2$ odpovídá vícerozměrné normální rozdělení $p(\boldsymbol{p}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ s diagonální kovarianční maticí $\boldsymbol{\Sigma}$.

Jedním z problémů maximalizace (2.12) je aproximace lokálních odezev klíčových bodů a jejich převod do spojité oblasti tak, aby optimalizace neprobíhala s celočíselným omezením souřadnic. Za předpokladu jednoho výrazného maxima lze odezvu i-tého bodu aproximovat např. dvourozměrným normálním rozdělením se střední hodnotou μ_i a kovariancí $\boldsymbol{\Sigma}_i$ [Cootes 1995] či paraboloidem [Wang 2008b], což při použití (2.6) a po zlogaritmování (2.12) vede na minimalizaci

$$\boldsymbol{p}^* = \underset{\boldsymbol{p}}{\operatorname{argmin}} \|\boldsymbol{p}\|^2 + \sum_{i=1}^{v} \|\boldsymbol{x}_i(\boldsymbol{p}) - \boldsymbol{\mu}_i\|^2_{\boldsymbol{\Sigma}_i^{-1}}, \tag{2.13}$$

Podobně jako v případě AAM, obvyklým způsobem řešení (2.13) je linearizace Taylorovým rozvojem a minimalizace Gauss-Newtonovou metodou. Lze však zvolit i složitější formu aproximace odezvy, např. v práci [Gu 2008] autoři použili gaussovskou směs. V článku [Saragih 2011] byla použitá neparametrická aproximace $p(l_i = 1 \,|\, \boldsymbol{x}_i, \mathcal{I})$ pomocí jádrového odhadu hustoty pravděpodobnosti (Kernel Density Estimation, KDE) s gaussovským jádrem. Kvalita jednotlivých kandidátů $\boldsymbol{y}_i$ na správnou pozici byla vyhodnocována pomocí Support Vector Machine (SVM) klasifikátoru natrénovaného pro každý bod zvlášť. Skutečné pozice klíčových bodů byly definovány jako skryté proměnné. Parametry byly získány maximalizací

$$\boldsymbol{p}_{\mathrm{MAP}} = \underset{\boldsymbol{p}}{\operatorname{argmax}}\, p(\boldsymbol{p}) \prod_{i=1}^{v} \sum_{\boldsymbol{y}_i \in \boldsymbol{\Psi}_i} p(l_i = 1 \,|\, \boldsymbol{y}_i, \mathcal{I})\, \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{y}_i, \rho \boldsymbol{I}), \tag{2.14}$$

kde $\boldsymbol{\Psi}_i$ je množina kandidátů na správnou pozici i-tého bodu a $\mathcal{N}(\boldsymbol{x}_i; \boldsymbol{y}_i, \rho \boldsymbol{I})$ označuje normální rozdělení se střední hodnotou $\boldsymbol{y}_i$ a uniformní diagonální kovariancí. Kritérium (2.14) bylo maximalizováno algoritmem Expectation Maximization (EM).

Belhumeur a kol. [Belhumeur 2011] definovali problém lokalizace bodů v rámci Bayesovského rozhodování bez parametrického popisu tvarové variability. Podobně jako v [Saragih 2011] byl pro ohodnocení kandidátů na správnou pozici klíčových bodů použitý SVM klasifikátor, zde nad příznaky Scale Invariant Feature Transform (SIFT) [Lowe 2004]. Lokalizace byla založena na předpokladu, že jakýkoliv tvar odpovídá některému vzorku $\boldsymbol{s}_{k,t} = (\boldsymbol{x}^1_{k,t}, \ldots, \boldsymbol{x}^v_{k,t})$ z trénovací databáze, kde t označuje eukleidovskou transformaci. Optimální konfigurace bodů $\boldsymbol{s}^*$ pak byla nalezena maximalizací aposteriorní pravděpodobnosti, která marginalizuje přes neznámý vzorek k a transformaci t, tj.

$$\boldsymbol{s}^* = a\boldsymbol{x}_1, \ldots, \boldsymbol{x}_v \sum_{k=1}^{N} \int_{t \in T} \prod_{i=1}^{v} p\left(\boldsymbol{x}_i - \boldsymbol{x}^i_{k,t}\right) p\left(\boldsymbol{x}_i \,|\, \mathcal{D}_i\left(\boldsymbol{x}_i; \mathcal{I}\right)\right) \mathrm{d}t, \tag{2.15}$$

kde $p\left(\boldsymbol{x}_i \,|\, \mathcal{D}_i\left(\boldsymbol{x}_i; \mathcal{I}\right)\right)$ je odezva detektoru v okolí bodu $\boldsymbol{x}_i$ a $p\left(\boldsymbol{x}_i - \boldsymbol{x}^i_{k,t}\right)$ penalizuje odchylku bodu $\boldsymbol{x}_i$ od modelového příkladu $\boldsymbol{x}^i_{k,t}$. Marginalizace v (2.15) je výpočetně velmi náročný problém, proto autoři použili metodu Random Sample Consensus (RANSAC) pro vygenerování množiny 100 nejlepších dvojic k a t, na jejichž základě byl řešen problém (2.15). Algoritmus dosahuje v současnosti jedny z nejlepších výsledků. I přes aproximaci integrálu Monte Carlo metodou však v praxi vyžaduje přibližně jednu sekundu na jeden klíčový bod, což znemožňuje jeho nasazení v systémech reálného času.

2.5 Diskriminační metody zarovnání obličeje

V současné době populární alternativou k lokalizaci klíčových bodů pomocí optimalizačních a generativních metod je diskriminační predikce tvaru. Algoritmy z této kategorie určují parametry modelu či přímo pozice klíčových bodů pouze na základě vyhodnocení nějakých příznaků a během testovací fáze tak nedochází k optimalizaci žádného kritéria.

Např. zarovnání aktivního vzhledového model prvně uvedeného v článku [Cootes 1998] původně nebylo formulováno v rámci Lucas-Kanade frameworku. Cootes a kol. předpokládali, že proces lokalizace probíhá pokaždé velmi podobně a tudíž existuje vztah mezi aktuální odchylkou modelu od obrazu $\Delta\boldsymbol{I}$ a optimální aktualizací kombinovaných parametrů $\Delta\boldsymbol{c}$. Tento vztah vyjádřili jako lineární závislost

$$\Delta\boldsymbol{c} = \boldsymbol{R}\Delta\boldsymbol{I} = \boldsymbol{R} \cdot \left[A(\boldsymbol{x}) - I\left(W\left(\boldsymbol{x}, \boldsymbol{p}'\right)\right)\right], \tag{2.16}$$

kde $A(\boldsymbol{x})$ je textura generovaná modelem (2.7) a $I\left(W\left(\boldsymbol{x}, \boldsymbol{p}'\right)\right)$ je textura viděná v obrázku. Koeficienty matice $\boldsymbol{R}$ lze získat v trénovací fázi ze známých dvojic odchylek $\Delta\boldsymbol{I}$ a optimálních aktualizací $\Delta\boldsymbol{c}$ pomocí metody nejmenších čtverců (vícerozměrné lineární regrese). Aktualizace parametrů $\boldsymbol{c}^{(t+1)} \leftarrow \boldsymbol{c}^{(t)} + \eta\Delta\boldsymbol{c}$ je prováděna iterativně, dokud není dosaženo konvergence. Výhodou tohoto postupu je optimalizace v prostoru kombinovaných parametrů, nikoliv oddělených jako v případě některých variant Lucas-Kanade registrace.

Na původní práci [Cootes 1998] navázal článek [Saragih 2007] zobecněním na nelineární vztah odchylky modelu od obrazu a optimální aktualizace parametrů. Inspirováni obličejovým detektorem Violy a Jonese (sekce 2.2) autoři sestavili výsledný silný regresor $R^k(\boldsymbol{f})$ pro k-tý parametr modelu (odhadovány byly pouze tvarové parametry $\boldsymbol{p}$) jako lineární kombinaci slabých regresních funkcí $g_t^k(\boldsymbol{f})$

$$R^k(\boldsymbol{f}) = \sum_{t=1}^{T} \alpha_t^k g_t^k(\boldsymbol{f}) \tag{2.17}$$

Příznakový vektor $\boldsymbol{f}$ je tvořený jednoduchými haarovskými příznaky akcelerovanými využitím součtového obrazu podobně jako v [Viola 2001]. Trénování probíhalo rovněž pomocí boostingu, ovšem s pozměněným kritériem pro výběr optimálního slabého klasifikátoru tak, aby nedocházelo k přetrénování.

V práci [Dantone 2012] autoři pro predikci optimální pozice klíčových bodů použili podmíněné regresní lesy (angl. conditional regression forest). Jako příznaky byly použity součty obdélníkových oblastí v obraze s využitím integrálních obrazů podobně jako v detektoru VJ (sekce 2.2). Lesy vyhodnocují pravděpodobnost, že obdélníková oblast odpovídá daným klíčovým bodům. Pro každý bod je odezva neparametricky aproximována KDE odhadem a následně je vypočítán posun její střední hodnoty. Tento postup je opakován až do dosažení předem definovaného počtu kroků. Pro zvýšení diskriminační schopnosti regresního lesu autoři trénovali několik různých lesů v závislosti na odhadnutém natočení obličeje. Výsledná pozice bodů je pak váženým průměrem odhadů jednotlivých lesů.

Xiong a De la Torre [Xiong 2013] navrhli diskriminační variantu obecné Newtonovy optimalizační metody a demonstrovali její využití pro lokalizaci klíčových bodů na obličeji. Na rozdíl od předchozích prací definovali regresi jako zobrazení z prostoru příznaků přímo do prostoru klíčových bodů, nikoliv parametrů. Úlohu formulovali jako hledání optimální aktualizace pozic klíčových bodů:

$$\Delta \boldsymbol{s}^* = \underset{\Delta \boldsymbol{s}}{\operatorname{argmin}} \left\| h\left(I(\boldsymbol{s} + \Delta \boldsymbol{s})\right) - h\left(I(\hat{\boldsymbol{s}})\right), \right\|_2^2 \tag{2.18}$$

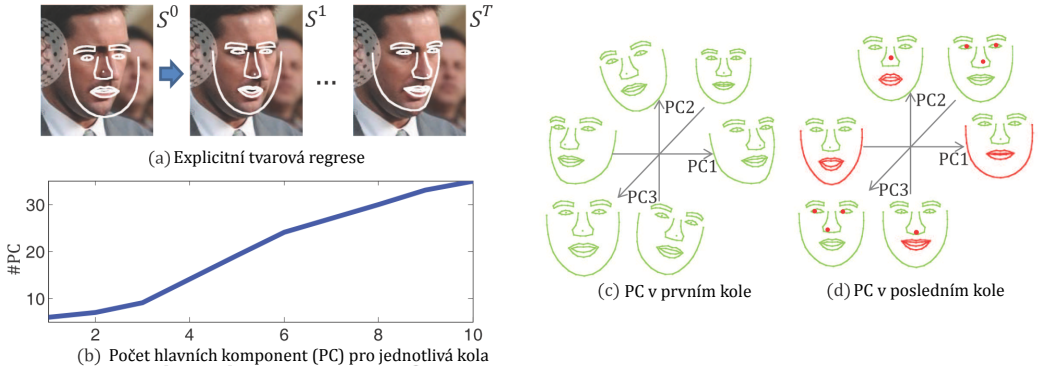
kde $h\left(I(\boldsymbol{s} + \Delta \boldsymbol{s})\right)$ označuje příznaky vypočtené z obrazu I na pozici $\boldsymbol{s} + \Delta \boldsymbol{s}$ a $\hat{\boldsymbol{s}}$ označuje pozice manuálně anotovaných bodů v trénovací databázi. Podobnost příznaků h tedy byla vyhodnocována pouze v okolí klíčových bodů podobně jako u lokálního omezeného modelu. Analogicky k [Cootes 1998] však optimalizace $\Delta \boldsymbol{s}$ probíhala diskriminačním způsobem, tedy natrénováním regresoru ve tvaru

$$\Delta \boldsymbol{s} = \boldsymbol{R} \cdot h\left(I(\boldsymbol{s})\right) + \boldsymbol{b} \tag{2.19}$$

Koeficienty matice $\boldsymbol{R}$ a translačního vektoru $\boldsymbol{b}$ jsou odhadnuty minimalizací

$$\boldsymbol{R}^*, \boldsymbol{b}^* = \underset{\boldsymbol{R},\boldsymbol{b}}{\operatorname{argmin}} \sum_{I^i} \int p\left(\boldsymbol{s}^i\right) \left\| \Delta \boldsymbol{s}^i - \boldsymbol{R} \cdot h(I(\boldsymbol{s})) - \boldsymbol{b} \right\|^2 \mathrm{d}\boldsymbol{s}^i, \tag{2.20}$$

kde i značí index obrázku v trénovací databázi a $p\left(\boldsymbol{s}^i\right)$ popisuje variabilitu obličejového detektoru. Pro urychlení integrace při výpočtu očekávané hodnoty



Obrázek 2.7: Princip odhadu pozice a tvaru obličeje pomocí metody ESR. Hlavní komponenty (PC) zachycující nejvýraznější módy variability jsou získány metodou PCA, viz sekci 3.1.2. Převzato z [Cao 2012].

inicializační odchylky 2.20 autoři použili Monte Carlo vzorkování. Na rozdíl od [Cootes 1998] bylo tímto způsobem natrénováno více vrstev regresorů pro postupné upřesňování odhadu.

Výpočetně nenáročný a efektivní algoritmus **explicitní tvarové regrese** (Explicit Shape Regression, ESR) byl navržen v práci [Cao 2012]. ESR je založená na dvouvrstvé kaskádní regresi s využitím velkého počtu primitivních regresorů. Tímto primitivním regresorem je tzv. fern[2] [Ozuysal 2010], což je kompozice F příznaků a prahů, které dělí prostor příznaků na 2^F přihrádek (angl. bin), přičemž každá přihrádka obsahuje výstupní hodnotu regresní veličiny. Jedná se tedy o zjednodušenou formu regresního stromu, kde všechny dělící nadplochy jsou vzájemně ortogonální a mají jeden společný bod. Na základě hodnot příznaků a jejich porovnání s prahy se při regresi vždy vybere jedna z přihrádek fernu a výstup nastaví na odpovídající hodnotu. V ESR je přitom jako příznak fernu použitý pouhý rozdíl dvou pixelů. Detekce klíčových bodů probíhá iterativně tak, že se v každé iteraci t z obrazu navzorkuje F pixelů, spočítají rozdíly jejich hodnot, které po porovnání s prahovými hodnotami indexují fern s optimální aktualizací parametrů $\Delta \boldsymbol{s}^{(t)}$. Nová pozice bodů v iteraci $t+1$ tedy bude $\boldsymbol{s}^{(t+1)} = \boldsymbol{s}^{(t)} + \Delta \boldsymbol{s}$. Opět se navzorkuje jiných F pixelů a použije se výstup z jiného fernu lépe uzpůsobeného přesnějšímu odhadu než v předchozí iteraci t. Opakováním těchto kroků se postupně aktuální odhad pozice klíčových bodů přibližuje jejich skutečné pozici.

Robustnost a přesnost ESR je však značně závislá na trénovací fázi, kdy dochází k výběru vzorkovacích pozic pixelů a výpočtu optimálních $\Delta \boldsymbol{s}^{(t)}$. K natrénování je nutná co největší databáze obličejů, kde má každý obrázek označeny pozice klíčových bodů. Každému obrázku je pak přidruženo několik startovacích pozic a nastává proces simulace detekce klíčových bodů. Trénovací fáze je rozdělena do K (např. $K = 10$) kol, přičemž v každém kole je vytvořeno N fernů (např. $N = 500$). V každém kole se nejprve náhodně vybere P pixelů (např. $P = 400$) a spočítají

[2] V době psaní textu nebyl známý český překlad tohoto termínu.

se jejich rozdíly. Z těchto $\binom{P}{2}$ rozdílů se pro každý fern vybere F (např. $F = 5$) takových, které v aktuálním kole nejlépe korelovaly s posunem ke správné pozici. Pro každou přihrádku ve fernu se výstupní hodnota nastaví na průměrný rozdíl mezi aktuální a správnou pozicí klíčových bodů vypočítaný ze všech trénovacích vzorků, jejichž rozdíly pixelů spadají do této přihrádky.

Příklad regrese je znázorněn na obrázku 2.7. Na obrázku 2.7a je zobrazen hrubý postup zpřesňování odhadu pozice a tvaru obličeje. Obrázky 2.7b, 2.7c a 2.7d pak znázorňují typickou podobu aktualizačních vektorů $\Delta \boldsymbol{s}^{(t)}$ v závislosti na aktuální fázi regrese. Zatímco na začátku regrese, kdy je odhad velmi nepřesný, se algoritmus snaží napravit chyby způsobené především posunem, rotací, či změnou měřítka, v závěrečných kolech se již upřesňují pouze detaily ve výrazu tváře. Vzhledem ke své rychlosti, robustnosti a přesnosti byl v této práci algoritmus ESR implementován a aplikován pro odezírání ze rtů a extrakci zájmové oblasti, viz sekci 9.3.

Na článek [Cao 2012] navázali ve své práci [Kazemi 2014] Kazemi a kol., kteří ferny vyměnili za plnohodnotné regresní stromy a navrhli odlišnou metodu před-výběru pixelových dvojic. Přestože jejich metoda neumožňuje korelační způsob výběru příznaků, dosáhli mírně lepších výsledků než [Cao 2012]. Richter a kol. [Richter 2014] pak rozšířili ESR o dodatečné příznaky a inspirováni článkem [Dantone 2012] i o odhad natočení hlavy.

3. Vizuální parametrizace

Klasifikovat vizuální signál promluvy přímo z jasových či RGB hodnot jednotlivých pixelů není praktické. Jednak v takovém případě mají data příliš vysokou dimenzi a jednak rovněž obsahují mnoho nadbytečné informace, jelikož kamera obvykle zachycuje kromě obličeje řečníka i nerelevantní pozadí. Pro rozpoznávání je tedy vhodné videa parametrizovat na co nejmenší množství příznaků, které pokud možno zachytí pouze užitečnou informaci a nic jiného. Následující text popisuje existující parametrizace a rozděluje je do několika kategorií podle typu informace, kterou se snaží z video signálu extrahovat.

3.1 Obrazové transformace

Jednu ze skupin představují příznaky extrahované z přibližně lokalizované oblasti zájmu (Region of Interest, ROI). Metody z této kategorie přistupují k jasovým hodnotám ROI jednotlivých snímků jako k vektorům (maticím) náhodných veličin, jejichž vysokou dimenzi se snaží redukovat bez využití jakýchkoliv dalších znalostí zkoumaných dat. Kvalita výsledné vizuální parametrizace tak závisí především na vhodné volbě metody redukce dimenze.

3.1.1 Integrální obrazové transformace

Klasickým způsobem redukce dimenze je analýza a následná filtrace frekvenčního spektra ROI. Základní, avšak v oblasti zpracování obrazu velice populární metodu představuje **diskrétní kosinová transformace** (Discrete Cosine Transform, DCT). Mezi její hlavní výhody patří rychlost výpočtu, který umožňuje použití obdobného algoritmu jako u rychlé Fourierovy transformace (Fast Fourier Transform, FFT) pro výpočet diskrétní Fourierovy transformace (Discrete Fourier Transform, DFT). Jako vizuální příznak má přitom dobrou rozlišovací schopnost [Obdržálek 2006]. Nevýhodou je její citlivost na změny v zarovnání oblasti, ze které se počítá.

Při použití 2D DCT jako vizuální je obvykle vybráno prvních několik koeficientů podle zig-zag řazení použitého např. ve ztrátovém obrazovém formátu JPEG. V oblasti audiovizuálního rozpoznávání řeči se osvědčil výběr koeficientů na základě jejich celkové energie nasčítané na trénovacích datech [Heckmann 2002b]. Kromě energie je však možné DCT koeficienty vybírat na základě jiných kriterií, např. vzdálenosti od nulté frekvence či vzájemné informace (mutual information) [Scanlon 2004].

DCT představuje speciální případ DFT, která je však definována v komplexním oboru hodnot a pro parametrizaci vizuálního signálu a rozpoznávání tak není příliš vhodná. V aplikacích komprese obrázků a videa se proto osvědčila spíše **diskrétní vlnková transformace** (Discrete Wavelet Transform, DWT). Na rozdíl od DFT, DWT lokalizuje signál nejen ve frekvenčním spektru, ale i v obrazové rovině (v čase pro jednorozměrný sekvenční signál). DWT rozkládá signál na množinu bazických

složek, které vznikají celočíselným posunem a změnou měřítka tzv. matečných funkcí. Matečné funkce jsou voleny tak, aby jejich odvozeniny byly vzájemně ortogonální. Výsledkem DWT transformace jsou tzv. aproximační a detailní koeficienty, které odpovídají dolní, resp. horní části spektra vstupního signálu. V případě dvourozměrné signálu (oblast ROI) vzniknou čtyři typy koeficientů (LL, LH, HL, HH), protože děleno je jak horizontální, tak vertikální spektrum obrázku.

Jednou z častých voleb matečné funkce jsou Haarovy vlnky, dalšími používanými matečnými funkcemi pak bývají např. funkce Mayerovy či Daubechiové. V úloze automatického odezírání ze rtů byla DWT s Haarovými vlnkami aplikována např. v článku [Puviarasan 2011], kde jako vizuální parametrizaci autoři zvolili aproximační koeficienty (LL). Pro redukci počtu koeficientů lze aplikovat podobné metody jako v případě DCT.

3.1.2 Analýza hlavních komponent

Velmi často používanou metodou pro redukci rozměru dat je analýza hlavních komponent (Principal Component Analysis, PCA). Kromě redukce dimenze se PCA používá také jako metoda dekorelace dat, která pomocí ortogonální lineární transformace převádí množinu (vektor) korelovaných veličin do prostoru, kde jsou tyto veličiny nekorelované. Jeden ze známých příkladů její úspěšné aplikace je také automatické rozpoznávání lidské tváře [Turk 1991]. Pro vizuální či audiovizuální rozpoznávání řeči byla aplikována např. v článcích [Bregler 1994, Lan 2009].

PCA může být definována[1] jako lineární ortogonální projekce, která převádí data do prostoru o nižším rozměru při maximálním zachování jejich rozptylu. Z geometrického pohledu se tedy podobně jako v případě lineárních integrálních transformací jedná o ortogonální projekci dat (rotaci) do jiného prostoru. PCA se však liší tím, že báze nového prostoru je závislá na rozložení a vzájemné kovarianci uvažovaných dat a není tedy konstantní. Pokud $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, $\boldsymbol{x}_i \in \mathcal{R}^d$ je $d \times N$ matice dat, jejichž kovarianci popisuje matice $\boldsymbol{C}$, pak lze ukázat [Bishop 2006], že sloupce hledané projekční matice $\boldsymbol{P} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n)$, $\boldsymbol{p}_i \in \mathcal{R}^d$ jsou vlastními vektory kovarianční matice $\boldsymbol{C}$:

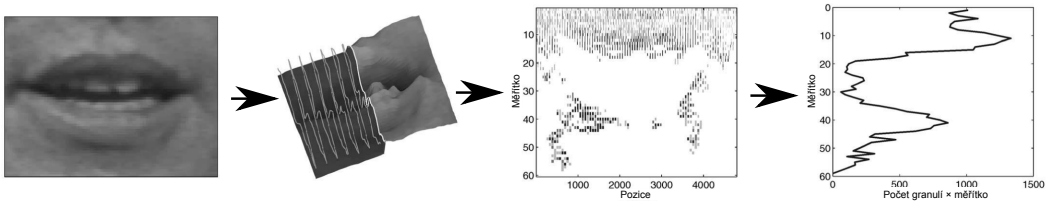
$$\boldsymbol{C}\boldsymbol{p}_i = \lambda_i \boldsymbol{p}_i, \; i = 1, \ldots, n \tag{3.1}$$

Počet n ortonormálních bazických vektorů $\boldsymbol{p}_i$ určuje výsledný rozměr dat po redukci. Obvyklým způsobem je volba n na základě poměru variability původních a transformovaných dat. Rozptyl dat po projekci na $\boldsymbol{p}_i$ určuje odpovídající vlastní číslo λ_i a tedy v případě, kdy je požadováno zachování alespoň 95 % původního rozptylu, je n zvoleno tak, aby

$$\frac{\sum_{i=1}^{n} \lambda_i}{\sum_{i=1}^{d} c_{ii}} \geq 0.95, \tag{3.2}$$

kde c_{ii} jsou prvky na diagonále $\boldsymbol{C}$. Počet bazických vektorů a tedy i počet výsledných koeficientů lze zvolit také na základě úspěšnosti výsledného klasifikátoru. Pro tento účel je vhodná technika křížové validace, aby nedocházelo k tzv. přeučení.

[1] PCA byla nezávisle na sobě objevena nejméně dvěma různými autory, přičemž každý ji definoval jinak.



Obrázek 3.1: Výpočet granulometrických příznaků. Převzato z [Matthews 2002].

Extrakce vizuálních řečových příznaků PCA probíhá obdobným způsobem jako u DCT. V první fázi jsou náhodně vybrány (dvourozměrné) ROI z promluv v trénovací množině, spojením řádků či sloupců převedeny na vektory a nad těmito daty jsou následně vypočteny projekční matice $\boldsymbol{P}$ a optimální rozměr n. V druhé fázi jsou řečové příznaky vypočteny jako

$$\boldsymbol{y} = \boldsymbol{P}^\top(\boldsymbol{x} - \bar{\boldsymbol{x}}) \tag{3.3}$$

kde $\boldsymbol{x}$ je vektor hodnot pixelů extrahovaných z ROI a $\bar{\boldsymbol{x}}$ průměrný vektor trénovací databáze $\boldsymbol{X}$.

3.1.3 Ostatní příznaky pro klasifikaci textur

Vizuální příznaky je možné z oblasti zájmu extrahovat také granulometrickými metodami. V článku [Matthews 2002] bylo dobré úspěšnosti použitím **síta** (angl. sieves) založeného na rekurzivním filtrování obrazu morfologickými operacemi či mediánovým filtrem. Pro tento účel je obrázek považovaný za graf, ve kterém pixely představují uzly a jejich sousednost reprezentují hrany. Rekurzivní aplikací filtrů jsou z obrázku postupně odstraněny lokální extrémy, přičemž v každé iteraci se sleduje jejich počet či součet hodnot. Vzhledem k předpokladu, že většina vizuální informace je zachycena svislým pohybem rtů, lze místo dvourozměrných variant filtrů vstupní obrázek vertikálním skenováním (tj. po sloupcích) nejprve převést na vektor a poté aplikovat standardní jednorozměrné operace. Filtry se aplikují rekurzivním způsobem, kdy jsou z obrázku postupně odstraňována lokální maxima vzrůstající velikosti a tím vzniká popis na úrovni prostoru měřítek. V každé iteraci, tj. pro každé měřítko r, se sleduje počet či součet amplitud prosetých extrémů (granularita) a na jeho základě se vypočte histogram měřítek. Tento histogram pak tvoří vizuální parametrizaci. Jeho dimenzi, která odpovídá vertikálnímu rozlišení obrázku, je možné redukovat metodou PCA. Postup pro ROI s rozměry 80×60 je ilustrován obrázkem 3.1. V článku [Matthews 2002] bylo dosaženo nejlepších výsledků s filtrem uzavření a sčítáním amplitud extrémů (tedy ne jejich pouhý počet). Např. v článku [Lan 2009] se však výsledky nepodařilo reprodukovat a klasické příznaky (DCT, AAM) dosahovaly vyšší úspěšnosti.

Jako vizuální parametrizaci lze v principu použít libovolný deskriptor vhodný pro klasifikaci textur. Např. v článku [Kricke 2008] autoři aplikovali pro tento účel **lokální binární vzory** (Local Binary Patterns, LBP). Operátor LBP porovnává intenzitu l_c každého pixelu v nějaké oblasti obrázku s P okolními pixely l_p

rovnoměrně rozprostřenými po kružnici s poloměrem R a středem v l_c. Řetězec výsledků porovnání $l_p > l_c$ tvoří P-bitové binární číslo

$$\mathrm{LBP}_{P,R} = \sum_{p=1}^{P} \mathbb{1}(l_p > l_c) \cdot 2^p. \tag{3.4}$$

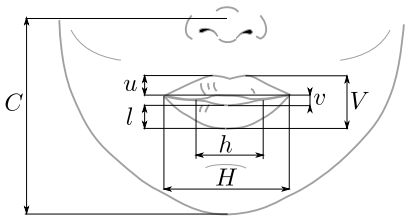
Histogram hodnot $\mathrm{LBP}_{P,R}$ v nějaké oblasti (např. ROI pro odezírání ze rtů) pak tvoří výsledný příznakový popis. Jelikož se některé bitové kombinace vyskytují častěji než jiné, používá se obvykle tzv. rovnoměrné LBP (angl. uniform LBP), které přiřazuje všem binárním číslům s více než dvěma přechody mezi nulou a jedničkou společnou hodnotu. Snížením počtu unikátních hodnot $\mathrm{LBP}_{P,R}$ je navíc redukován i výsledný histogram. Mezi další varianty pak patří rotačně invariantní LBP, u něhož jsou ekvivalentní všechna binární čísla, která lze nějakou bitovou rotací transformovat na stejnou hodnotu. V [Kricke 2008] bylo dosaženo nejlepších výsledků klasifikace jednotlivých vizémů pro rovnoměrné $\mathrm{LBP}_{8,3}$. LBP byly jako jeden z typů vizuální parametrizace použity v rovněž práci [Pei 2013].

Oblíbenou parametrizací v oblasti počítačového vidění je také **histogram orientovaných gradientů** (Histogram of Oriented Gradients, HOG), původně navržený pro detekci chodců. Myšlenka HOG vychází z deskriptoru SIFT (Scale Invariant Feature Trasnform), který byl vyvinut jako příznakový popis zájmových bodů pro jejich klasifikaci. Pří extrakci SIFT deskriptorů je okolí každého zájmového bodu rozděleno na několik podoblastí, přičemž na každé z nich je sestaven histogram orientací gradientů. Gradienty lze vypočítat např. Sobelovým filtrem. Výsledný deskriptor zájmového bodu je utvořen spojením vyhlazených histogramů z jednotlivých podoblastí jeho okolí. Deskriptor HOG pak oproti SIFT přidává několik kroků předzpracování obrazu a normalizace histogramů pro dosažení lepší robustnosti vůči změnám osvětlení. Oproti SIFT také pracuje na fixní mřížce vstupního obrazu, ne pouze v okolí zájmových bodů, a jedná se tedy o globální deskriptor textury. Jako jeden z typů vizuální řečové parametrizace pro rozpoznávání frází byl aplikován např. v práci [Pei 2013]. Pro rozpoznávání ruských slabik z video signálu aplikovali HOG autoři článku [Savchenko 2014].

S deskriptory HOG souvisí také model **bag of words** (BOW), který vychází z poznatků zpracování a indexování textů. Na nějaké trénovací množině je vektorovým kvantováním (např. k-means clustering) nejprve sestaven „slovník“ lokálních bodových deskriptorů. Každý obrázek, tj. např. ROI pro odezírání ze rtů, pak může být popsán vektorem četností těchto slov či jeho redukovanou variantou. Parametrizace BOW založená na deskriptorech HOG byla pro vizuální rozpoznávání řeči aplikována např. v článku [Ju 2013].

3.2 Tvarové a kombinované příznaky

Příznaky založené na přibližně lokalizované oblasti zájmu nevyužívají žádnou další znalost o zkoumaných datech a veškerou užitečnou informaci extrahují pouze nepřímo využitím metod redukce dimenze. Většina těchto metod však předpokládá



Obrázek 3.2: Geometrické vizuální příznaky.

zjednodušující vlastnosti, nejčastěji blízkost dat nějaké lineární nadploše jako např. PCA. Lineární model závislosti výsledné podoby ROI na původní informaci v signálu však obecně nemusí platit, což může až znemožnit automatickou extrakci informace z pozorovaných dat pouhou „slepou“ redukcí dimenze. Jednou z možností, jak tento nedostatek opravit, je využití více informace z fáze obrazového předzpracování a detekce obličejových částí. Pokud jsou např. známé pozice klíčových bodů na tváři odpovídající jednotlivým obličejovým částem, je možné tuto znalost využít pro výpočet příznaků a odstranění nadbytečné informace ze vstupního signálu.

Jedněmi z nejjednodušších příznaků popisujících tvarovou informaci jsou **vnitřní a vnější horizontální rozšíření rtů** h, resp. H, **vnitřní a vnější vertikální rozšíření rtů** v, resp. V, a **pozice brady** C definovaná např. vzdáleností od nosu. Geometrické vizuální příznaky ilustruje obrázek 3.2. Dalšími podobnými příznaky mohou být např. **výška horního a dolního rtu** u, resp. l, **zaokrouhlení rtů**, obvykle definováno jako poměr $R = V/H$, celková **plocha rtů**, či **viditelnost zubů**. Je zřejmé, že úspěšnost těchto jednoduchých heuristicky navržených příznaků závisí především na kvalitě lokalizace rtů popsané v kapitole 2. Nejčastěji využívané přitom bývají metody založené na barevné segmentaci [Potamianos 1998b, Chaloupka 2005, Císař 2006], které poměrně výrazně závisí na světelných podmínkách.

Saenko a kol. [Saenko 2005] navrhli použití tzv. **artikulačních příznaků** (Articulatory Features, AF), které klasifikují stav řečového traktu do několika kategorií. V případě odezírání ze rtů jde především o podobu rtů a viditelnost zubů. Autoři tak vybrali tři základní příznaky: horizontální rozšíření rtů H odstupňované do čtyř různých hodnot, zaokrouhlení rtů R kategorizované do dvou tříd „zaoblené“/„nezaoblené“ a labiodentální, zachycující dotyk spodních zubů a horního rtu. Pro každý příznak byl natrénován SVM, který klasifikoval ROI do jednoho z předurčených stavů a jehož skóre pak představovalo výslednou parametrizaci. Jelikož průběhy jednotlivých příznaků nemusí být dokonale synchronní, použili autoři pro klasifikaci řečových sekvencí vícevrstvou dynamickou bayesovskou síť, kde každý z uvedených příznaků reprezentoval jednu vrstvu.

Velmi rozšířený způsob využití tvarové informace představují **aktivní tvarový** či **aktivní vzhledový model**. Vizuální parametrizaci ASM představuje vektor tvarových parametrů $\boldsymbol{p}$ (2.6), tedy souřadnice tvarového vektoru po promítnutí do redukovaného vlastního prostoru. PCA modeluje tvar pro všechny klíčové

Obrázek 3.3: Různé konfigurace klíčových bodů AAM.

body sdruženě a obsáhne tak potenciálně více informace, než expertem stanovené příznaky. Tvarové koeficienty zároveň dekoreluje, takže jsou výsledné příznaky do určité míry nezávislé. AAM podobným způsobem navíc zahrnuje i texturu a oba zdroje informace kombinuje do jediného statistického modelu. Vizuální parametrizaci řeči pak může představovat spojení tvarových a texturových parametrů $\boldsymbol{b}$ před třetí aplikací PCA či kombinované parametry $\boldsymbol{c}$, viz (2.8).

U ASM/AAM se principiálně jedná o velmi podobný postup jako v případě PCA redukce ROI (3.3), avšak nad jinými daty. V případě použití výlučně texturových parametrů AAM se obě metody liší pouze ve způsobu extrakce ROI, kdy AAM na rozdíl od přibližně lokalizované ROI (obr. 9.4) extrahuje a normalizuje texturu pouze z konvexního obalu klíčových bodů. Tím je dosaženo odstranění nadbytečné informace, avšak pouze za cenu vyšší citlivosti na nepřesnosti lokalizace klíčových bodů. Zájmovou oblast AAM je možné zvolit libovolně, avšak vhodné je soustředit se na oblasti, kde se nachází nejvíce informace, tj. především kolem úst. Příliš malá ROI (např. pouze rty) však nemusí dostatečně diskriminovat mezi jednotlivými výrazy, a tak je vhodné nalézt kompromisní řešení, např. dolní část obličeje. Příklady konfigurace klíčových bodů, jež definují ROI pro výpočet AAM příznaků, ilustruje obr. 3.3.

Změnu tvaru ASM a AAM na rozdíl od geometrických příznaků popisují implicitně formou jednoduchého statistického modelu. Problém však představuje vizuální variabilita, kdy mohou změny v koeficientech PCA projekce zachycovat spíše rozdíly mezi řečníky než mezi jednotlivými hláskami. Jedním z tradičních způsobů nápravy je odečítání příznakového průměru (feature mean subtraction, FMS). Od výsledné parametrizace se odečte průměrný příznakový vektor odhadnutý na několika snímcích sekvence tak, aby byla zachycena pouze informace o změně vůči základnímu stavu. U AAM se však jako efektivnější jeví odečítání průměru na úrovni dat, ne příznaků, a to i při trénování modelu. Místo odečítání příznakového průměru lze závislost na řečníkovi eliminovat odečtením průměrného tvaru a textury a AAM příznaky vypočíst na takto upravených datech. Tento postup autoři článku [Papandreou 2009] nazvali jako vizemické AAM (angl. visemic AAM).

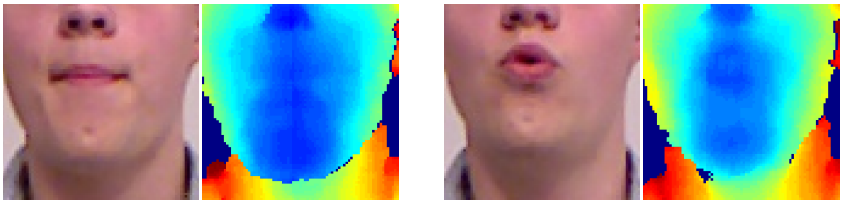
3.3 Využití prostorové informace

Kromě tvaru a vzhledu dolní části obličeje při čelním pohledu na mluvčího je vizuální informace obsažena i v hloubce, tedy vzdálenosti od pozorovatele. Při použití běžné kamery či webkamery však v obecném případě není možné tuto informaci z nasnímaných dat získat.

Jednou z cest, jak tuto překážku obejít, je použít některé omezující podmínky. Např. Blanz a kol. [Blanz 2003] na základě trojrozměrných skenů hlavy zhruba stovky dobrovolníků vytvořili deformovatelný trojrozměrný model (3D Morphable Model, 3DMM), který je postupným pyramidovým algoritmem s omezujícími podmínkami možné připodobnit obličeji na vstupním obrázku a tím kromě výrazu získat i odhad prostorového natočení. Algoritmus je však výpočetně velmi náročný a nehodí se tak pro zpracování v reálném čase. V [Xiao 2004] Xiao a kol. aplikací algoritmů pro odhad struktury z pohybu odhadli 3D souřadnice z 2D anotovaných obrázků obličejů a za pomocí omezujících podmínek podobně jako v článku Blanze a kol. detekovali pozici a natočení obličeje v obrázku. Žádný z těchto algoritmů však nebyl aplikován pro odezírání ze rtů.

V práci [Petr Císař 2004] autoři odvozovali trojrozměrnou podobu rtů pouze na základě dvourozměrného obrazu z videokamery o něco jednodušším způsobem. Pro tento účel byla vytvořena malá databáze foneticky vyvážených promluv pouze jednoho mluvčího. Tomu bylo na obličeji vyznačeno reflexivní barvou několik bodů podél vnější hranice rtů, což usnadnilo následné zpracování. Klíčové body na vnitřní hranici rtů byly anotovány manuálně až v nasnímaných obrazech. Trojrozměrný model rtů byl sestaven na základě více pohledů na řečníka, které byly získány systémem 4 zrcadel. V testovací fázi, kterou představovalo rozpoznávání řeči v automobilu, byl pomocí projektivní transformace dvourozměrnému tvaru v obrázku přiřazen nejpodobnější trojrozměrný z natrénované databáze. Sled těchto tvarů a z nich odvozených příznaků pak reprezentoval výslednou parametrizaci. Tímto postupem bylo dosaženo vyšší robustnosti vůči nepředpokládaným pohybům řečníka, přestože systém vyžadoval více pohledů pouze v trénovací fázi pro vytvoření trojrozměrného modelu.

Jiným způsobem odhadu pozice bodů v prostoru je plné využití více kamer a algoritmů **stereovidění**, pomocí kterých je následně možné hloubku rekonstruovat. Dvě kamery byly využity pro snímání řečníků při nahrávání australské audiovizuální databáze AVOZES [Goecke 2004], se kterou autoři následně pracovali v článku [Goecke 2005]. Celkem byly detekovány 4 body na řečníkových rtech (koutky úst a středy horního a dolního rtu), ze kterých byly odvozeny jednoduché geometrické příznaky, viz kapitolu 3.2. Klíčové body byly sledovány tříkrokovým algoritmem založeným na detektoru VJ (2.2). Výsledky však autoři porovnali s klasickým 2D rozpoznáváním z jediné kamery až v navazující práci [Goecke 2008], kde se přínos trojrozměrného sledování pohybu rtů prokázal jen ve specifických situacích, např. při pohybech hlavou. V běžných případech byla výsledná chybovost vyšší než při použití standardních dvourozměrných technik. Výsledky však byly ovlivněny volbou sledovacího algoritmu - pro dvourozměrné sledování autoři aplikovali několik variant



Obrázek 3.4: Ukázky interpolovaných hloubkových map při vyslovování souhlásky ‘m’ (vlevo) a samohlásky ‘u’ (vpravo) na databázi TULAVD.

robustnějšího AAM, pro trojrozměrné pouze uvedenou jednoduchou heuristickou metodu. Vzhledem k dosaženým výsledkům však další výzkum v oblasti odezírání pomocí stereovidění autoři přesto odložili.

Rozšíření automatického odezírání ze rtů s využitím stereovidění brání několik faktorů. Významnou překážku představuje náročnost na přípravu a konfiguraci více kamer. Algoritmy vyhledávání korespondencí mezi obrázky jsou navíc citlivé na světelné podmínky a tak potenciálně nespolehlivé. S rozvojem multimediálních a herních technologií v posledních dvou až třech letech se však situace změnila. Na trh se totiž dostala komerčně dostupná zařízení, která např. pomocí infračerveného spektra a techniky strukturovaného světla dokáží částečně zjednodušit náročnou úlohu stereovidění. Uživatel přímo ze snímače dostane pro každý pixel v obraze informaci o jeho vzdálenosti od senzoru. Podle typu informace o vzdálenosti tak vznikne buď hloubková, nebo tzv. disparitní mapa, která udává rozdíl v pozicích každého pixelu mezi dvěma náhledy na tutéž scénu. Kalibraci a složité a nespolehlivé vyhledávání řídkých korespondencí pak lze obejít, navíc s výhodou existence hloubkové informace pro každý bod v obraze, nejen pro několik klíčových bodů. Mezi zařízení schopná rekonstruovat tyto informace patří např. Microsoft Kinect, Asus Xtion či Creative Senz3D. V této práci je pro experimenty používáno zařízení Microsoft Kinect, popsané v sekci 9.1.1. Příklady bilineárně interpolovaných hloubkových map jsou znázorněny na obr. 3.4.

Hloubkovou či disparitní mapu lze využít jako dodatečný zdroj informace. První prací zabývající se přínosem informace rekonstruované z hloubkové mapy byla [Galatas 2012]. Autoři zde aplikovali **dvourozměrnou DCT** pro extrakci příznaků z hloubkové mapy, tedy shodný způsob jako u video snímků. Hloubkové příznaky byly využity v úloze rozpoznávání spojených číslovek. Dále byla hloubková data z Kinectu využita v práci [Pei 2013] jako jeden z mnoha zdrojů pro extrakci příznaků náhodnými lesy a v článku [Savchenko 2014] pro rozpoznávání ruských slabik.

3.4 Dynamické vizuální příznaky řeči

Velmi důležitou složku pro rozpoznávání řečového signálu představuje jeho dynamika, protože každý foném či vizém charakterizuje jiná souslednost pohybů vokálního traktu či rtů. Například někteří lidé přirozeně dýchají ústy, což znamená, že v klidové poloze mají rty mírně rozevřené. Bez zohlednění ostatních snímků

sekvence pak tento případ není možné rozlišit např. od hlásky 'e'. Tento problém lze zmírnit zahrnutím časové informace do příznakového popisu řeči, což umožní lépe rozlišit fonetické jednotky než při použití pouze statické parametrizace a díky sledování **informace o změně** zároveň sníží rozdíly mezi jednotlivými řečníky. V literatuře se přitom nejčastěji objevují dva základní způsoby využití řečové dynamiky:

1. lokální dynamizace statických příznaků,
2. prostoro-časová dynamická parametrizace.

V prvním případě jsou dynamické vlastnosti řeči odhadnuty na základě statického příznakového popisu nejbližšího okolí každého snímku a tato informace následně zahrnuta do výsledné parametrizace. Z hlediska využití se pak tyto příznaky nijak neodlišují od statických, pouze zachycují jiný typ informace.

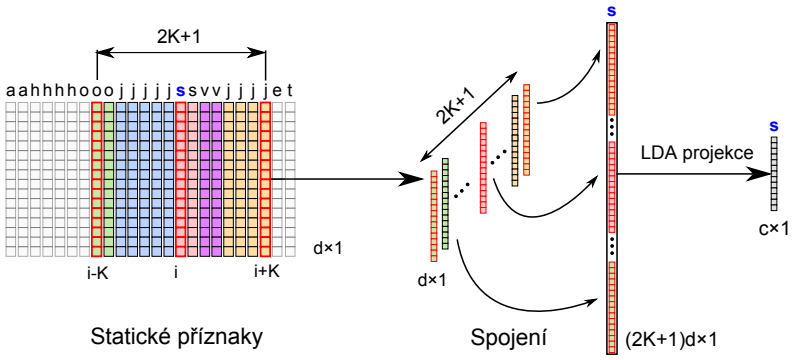
V druhém případě se dynamické vlastnosti promluvy zachycují již při návrhu příznaků. Na videosekvence je nahlíženo jako na trojrozměrná prostoročasová data a příznaky jsou extrahovány v závislosti na všech třech složkách. V některých případech může být návrh parametrizace uzpůsoben konkrétní úloze a časovým nárokům na extrakci, a tak se následné využití v klasifikátorech může lišit od statického popisu.

3.4.1 Lokální dynamizace statických příznaků

Klasickou metodou dynamizace statických příznaků využívanou především v oblasti automatického rozpoznávání řeči z akustického signálu je výpočet tzv. **delta** Δ a **akceleračních** $\Delta\Delta$ příznaků. Ze statické parametrizace $\boldsymbol{x}(n)$ se vypočte kauzální či nekauzální diference dvou snímků (Δ příznaky) a připojí ke statickému příznakovému vektoru, tedy např. $\boldsymbol{y}(n) = (\boldsymbol{x}(n), \boldsymbol{x}(n) - \boldsymbol{x}(n-1))^\top$. V případě $\Delta\Delta$ postupuje shodně, avšak navíc i s diferencí druhého řádu. Jelikož se jedná o numerickou aproximaci derivace příznakového popisu, lze použít i sofistikovanější způsoby výpočtu, např. proložením průběhu příznaků lokální regresní funkcí.

V oblasti automatického odezírání ze rtů se pro dynamizaci statického popisu osvědčila metoda spojení příznakových vektorů několika sousedních snímků a jejich následnou redukcí pomocí **lineární diskriminační analýzy** (Linear Discriminant Analysis, LDA). Tento postup byl aplikován v řadě prací, např. [Matthews 2002, Lucey 2007, Lucey 2006b, Lucey 2008, Lan 2010, Lan 2012, Estellers 2012b, Galatas 2012].

Lineární diskriminační analýza je soubor metod zabývajících se optimální projekcí dat vzhledem k jejich rozlišitelnosti v novém prostoru. Nejznámějším zástupcem je Fisherův lineární diskriminant (FLDA), který se od LDA jako skupiny algoritmů v literatuře obvykle nerozlišuje. FLDA se snaží najít takovou projekci dat, která maximalizuje poměr variability (rozptylu) mezi třídami a variability uvnitř tříd. To znamená, že shluky dat odpovídající jednotlivým třídám by měly být v nových souřadnicích co nejmenší a zároveň co možná nejdále od sebe. Pokud tedy $\boldsymbol{S}_W$ je součet kovariančních matic uvnitř jednotlivých tříd (vnitřní variabilita) a



Obrázek 3.5: Princip dynamických příznaků LDA.

$\boldsymbol{S}_B$ kovarianční matice průměrných vzorků jednotlivých tříd (vnější variabilita), pak FLDA hledá takovou projekční matici $\boldsymbol{W}$, která maximalizuje

$$\boldsymbol{W}_{\mathrm{LDA}} = \underset{\boldsymbol{W}'}{\operatorname{argmax}}\ Tr\left\{\left(\boldsymbol{W}'\boldsymbol{S}_W\boldsymbol{W}'^{\top}\right)^{-1}\left(\boldsymbol{W}'\boldsymbol{S}_B\boldsymbol{W}'^{\top}\right)\right\}, \tag{3.5}$$

kde $Tr\{\cdot\}$ označuje stopu matice. Řešením (3.5) jsou vlastní vektory matice $\boldsymbol{S}_W^{-1}\boldsymbol{S}_B$ [Bishop 2006]. Jelikož $\boldsymbol{S}_B$ má hodnost maximálně $J-1$, pouze $J-1$ vlastních vektorů bude přiřazeno k nenulovým vlastním číslům. Maximální výsledný rozměr redukovaných dat tak nikdy nemůže překročit počet tříd.

LDA je tedy metoda redukce dimenze s učitelem a vyžaduje, aby každý datový vektor byl přiřazen nějaké třídě. V případě řečové promluvy např. může být každý snímek anotován aktuálním vyslovovaným fonémem či vizémem. Při dynamizaci statických příznaků pomocí LDA se obvykle nekauzálním způsobem spojuje $2K+1$ sousledných příznakových vektorů, přičemž vzniklému hypervektoru se přiřadí foném či vizém prostředního snímku sekvence. Díky vlastnostem LDA pak bude mít po redukci výsledný příznakový vektor $\boldsymbol{y} = \boldsymbol{W}^{\top}(\boldsymbol{x} - \bar{\boldsymbol{x}})$ dimenzi vždy menší než počet fonetických jednotek. Postup dynamizace pomocí LDA znázorňuje obrázek 3.5.

3.4.2 Prostoro-časová dynamická parametrizace

Jedním z nejjednodušších způsobů, jak zahrnout dynamickou informaci procesu extrakce vizuálních příznaků, je výpočet **rozdílových obrazů** ze dvou či více sousedících snímků. Tento postup byl aplikován např. v [Gray 1996], kde autoři porovnávali několik variant příznaků extrahovaných z rozdílových obrazů. Nejlepších výsledků bylo dosaženo jednoduchým zmenšením obrázku filtrovaného dolní propustí a připojením shodným způsobem zpracovaného rozdílového obrazu do výsledné parametrizace. Sofistikovanější přístup k extrakci dynamiky ze dvou či více následujících obrázků představuje výpočet **optického toku** [Mase 1991, Gray 1996, Yoshinaga 2003, Tamura 2004, Shaikh 2010], jenž pro každý pixel udává směr a rychlost jeho pohybu vůči předešlému stavu. Např. v [Tamura 2004] byly z optického toku vypočteny horizontální a vertikální rozptyly a minima a maxima

integrálu optického toku. V [Shaikh 2010] pak byla vertikální složka optického toku rozdělena na 8 sloupců a na každém vypočítána průměrná hodnota. Vektor těchto hodnot představoval vizuální parametrizaci. Nevýhodu příznaků založených na optickém toku však představuje citlivost na šum a závislost na fázi předzpracování, tedy nízká robustnost.

Na videosekvence je také možné nahlížet jako na trojrozměrná data, kdy dvě obrazové souřadnice doplňuje navíc časová složka. Při extrakci příznaků je pak možné zohlednit lokální vlastnosti signálu i v třetím rozměru a tím postihnout informaci o dynamice. Tento postup byl aplikován např. v článku [Pachoud 2008], kde na video sekvence bylo nahlíženo jako na kvádry, jejichž hrany reprezentovaly obrazové souřadnice x, y a časová osa t. Tyto tzv. **makro-kvádry** (angl. macro-cuboid), jak je autoři nazvali, byly rozděleny na částečně se překrývající podbloky, z nichž byly následně extrahovány deskriptory SIFT (sekce 3.1.3), zobecněné pro trojrozměrný případ. Každé slovní jednotce (autoři rozpoznávali spojitě vyslovované číslovky 0–9) pak odpovídal nějaký makro-kvádr. Rozpoznávání vstupní sekvence spočívalo ve vyhledávání a zarovnávání makro-kvádrů na základě podobnosti a vzdálenosti jejich pod-bloků. Jelikož makro-kvádry byly extrahované vyčerpávajícím způsobem v různých pozicích a velikostech, tento postup zajistil nezávislost na měřítku (prostorovém i časovém). Navíc umožnil rozpoznávání libovolné sekvence fonémů či vizémů, ovšem pouze za předpokladu jejich časové nezávislosti a tedy bez možnosti sestavení slovních modelů.

Autoři Zhao a kol. [Zhao 2009] v rámci návrhu dynamických příznaků rozšířili **lokální binární vzory** (sekce 3.1.3) i do časové složky signálu. Podobně jako v [Pachoud 2008] na vstupní sekvence nahlíželi jako na kvádry s hranami odpovídajícími dvěma obrazovým a jedné časové složce. Vstupní sekvence byla opět rozdělena na překrývající se trojrozměrné segmenty a na každém z nich byly extrahovány prostorové LBP. Deskriptory byly vypočteny pouze ve třech rovinách: (x, y), (x, t) a (y, t) a nejednalo se tak o plně trojrozměrné LBP (zobecnění na elipsoid). Vektorově spojené histogramy LBP hodnot ze všech tří složek a všech segmentů pak představovaly výslednou parametrizaci vstupní sekvence, kterou autoři nazvali **LBPTOP** (Local Binary Patterns from Three Orthogonal Planes). Aby měl příznakový vektor vždy stejný rozměr, každá sekvence byla rozdělena na konstantní počet segmentů. Autoři aplikovali LBPTOP příznaky v úloze rozpoznávání anglických frází z databáze OuluVS a dosáhli poměrně dobrých výsledků, ovšem pouze pro rozpoznávání závislé na řečníkovi. Výsledky pro případ klasifikace nezávislé na mluvčím nebyly uvedeny.

Nevýhodami parametrizací [Pachoud 2008, Zhao 2009] jsou závislost na konkrétní úloze a částečná provázanost s návrhem klasifikátoru. Na rozdíl od Δ či LDA dynamizace jsou příznaky macro-cuboid a LBPTOP vypočteny pouze na řídké síti prostoročasu, a tak je vhodné použít specifický klasifikátor. Metody je nicméně možné modifikovat a parametrizaci extrahovat pro všechny snímky, tyto experimenty však nebyly ve zmíněných pracích provedeny.

3.5 Závislost na pozorovacím úhlu

Výzkum automatického odezírání ze rtů a audiovizuálního rozpoznávání se nejčastěji soustředí na případ, kdy kamera snímá řečníka z čelního pohledu (Frontal View, FV). Je tomu tak z několika důvodů. Tento předpoklad především významně zjednodušuje fázi detekce obličejových částí (kapitola 2), protože algoritmy nezatěžuje vnější variabilita způsobená perspektivní transformací. Eliminace nežádoucí variability je navíc vhodná i pro extrakci příznaků, které tak zachycují relevantnější informaci a umožňují lépe posoudit přínos navržené parametrizace. Obvykle se rovněž předpokládá, že právě z čelního pohledu je možné nejsnáze extrahovat stěžejní příznaky, tedy pohyb rtů, jazyka, či viditelnost zubů. Čelní pohled také představuje pro lidskou komunikaci nejpřirozenější způsob a tedy přímo odpovídá některým aplikacím, např. využití vizuální složky při automatickém přepisování televizních zpráv nebo při ovládání PC pomocí hlasových povelů pro tělesně postižené. V neposlední řadě pak neexistují dostatečně rozsáhlé volně dostupné audiovizuální databáze, na kterých by bylo možné ověřit úspěšnost příznaků v závislosti na zvoleném pohledu.

I přes zmíněné skutečnosti se však v literatuře objevují práce, které zkoumají množství informace zachycené v jiných než čelních pohledech. Ve zřejmě prvním článku zabývajícím se touto tématikou [Yoshinaga 2003] autoři extrahovali příznaky pomocí optického toku z **bočního pohledu** na mluvčího. Systém byl experimentálně ověřen v úloze rozpoznávání spojitých číslovek v Japonštině a ukázal přibližně 5–6 % přínos vizuální parametrizace extrahované z bočního pohledu oproti pouze akustickému rozpoznávání ve hlučném prostředí. Výsledky však nebyly porovnány s odezíráním z čelního pohledu. Lucey a Potamianos [Lucey 2006a] přímo porovnávali úspěšnost audiovizuálního rozpoznávání z čelního a profilového pohledu založeného na DCT příznacích dynamizovaných metodou LDA. V experimentech dosáhli pro profilové odezírání ze rtů o 60 % relativního zhoršení slovní chybovosti (Word Error Rate, WER) oproti čelnímu pohledu, avšak při kombinaci obou pohledů naopak 7 % zlepšení, což ukazuje na částečnou informační komplementaritu. Dokonce lepších výsledků oproti čelnímu pohledu dosáhli při profilovém odezírání Kumar a kol. [Kumar 2007]. Použili přitom jednoduché geometrické příznaky odvozené ze čtyř bodů na rtech. Příznaky byly založené na výšce, šířce a vyšpulení rtů, vše vypočteno relativně vůči referenčním bodům na nose a bradě, jenž byly detekovány pomocí barevného prahování profilových obrázků a následné detekce lokálních extrémů na kontuře obličeje. V úloze rozpoznávání izolovaných slov byly výsledky pro odezírání z profilu o 12–28 % relativní WER lepší než pří klasickém čelním pohledu. Výsledky ovšem byly ovlivněny volbou příznaků, kdy z obou pohledů nebyla extrahována shodná parametrizace. Z čelního pohledu lze navíc zachytit více informace než výlučně tvarovými příznaky, což nebylo v práci zohledněno. Na článek [Kumar 2007] navázali Saitoh a Konishi [Saitoh 2010], kteří vylepšili algoritmus extrakce kontury obličeje a rozšířili sadu geometrických příznaků. V úlohách rozpoznávání japonských samohlásek a izolovaných slov se svým algoritmem dosáhli dokonce lepší úspěšnosti než člověk.

V některých aplikacích předpoklad neměnné vzájemné pozice řečníkovy hlavy a kamery nerealistický. V takovém případě je pak nutné navrhnout **parametrizaci, která je vůči případným pohybům robustní**. Jednou z možností, jak toho dosáhnout, je **transformace obrazu či příznaků** z neznámého pohledu do unifikovaného prostoru společného pro všechny úhly náhledu. Tento postup zvolili Lucey a kol. [Lucey 2007], kteří definovali lineární závislost mezi příznaky extrahovanými z ánfasu a z profilu. Na základě trénovací parametrizace (DCT dynamizované metodou LDA) extrahované z čelních i profilových obrázků pak lze pomocí metody nejmenších čtverců odhadnout transformační matici

$$\boldsymbol{W} = \boldsymbol{T}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \lambda\boldsymbol{I})^{-1} \tag{3.6}$$

kde $\boldsymbol{X}, \boldsymbol{T}$ je původní, resp. cílová transformovaná podoba příznaků. Systém byl experimentálně ověřen v několika konfiguracích lišících se trénovacími daty: čelní pohled, profil, kombinovaný ánfas a profil, komb. ánfas a transformovaný profil a komb. profil a transformovaný ánfas. Nejlepších výsledků bylo dosaženo pro čtvrtý případ, přičemž relativní zlepšení WER oproti systému trénovaném pouze na ánfasu dosahovalo 19–50 %, v závislosti na tom, zda byla či nebyla testovací data transformována do čelního pohledu. Avšak v případě, že trénovací i testovací databáze obsahovaly pouze čelní pohled, model naučený na kombinovaných datech dosahoval o cca 7 % relativní WER horších výsledků než model natrénovaný pouze z ánfasu. V dalším článku [Lucey 2008] pak svůj systém Lucey a kol. rozšířili o fázi detekce natočení hlavy řečníka. Ná základě detekovaného natočení pak extrahovali příznaky do jediného společného skrytého markovského modelu a systém otestovali v úloze rozpoznávání spojitých číslovek na databázi CUAVE. I přes velmi malý slovník (11 položek, číslovka nula má v angličtině dvě výslovnosti „zero“ a „oh“) byla výsledná WER přes 60 %, což ukazuje na obtížnost při odstraňování nežádoucí variability ve vizuálních datech. Na práci [Lucey 2007] navázali také Estellers a Thiran [Estellers 2012b], kteří zpochybnili předpoklad lineární závislosti čelního a profilového pohledu, jelikož v globálním měřítku totiž může docházet např. zakrytí částí obličeje nosem. Lineární model platí pouze lokálně, a tak Estellers a Thiran rozdělili obrázky před extrakcí příznaků a transformací do unifikovaného na obdélníkové podoblasti a zpracovávali každou odděleně. Poněkud překvapivě se přínos lokalizované transformace projevil pouze při normalizaci zájmové oblasti, tj. na úrovni pixelů, avšak ne při transformaci na úrovni příznaků.

Závislostí úspěšnosti odezírání na relativním úhlu řečníka a kamery se zabývali Lan a kol. [Lan 2012]. Ke svým experimentům využili databázi LiLiR, ve které 38 mluvčích namluvilo 200 vět na několik kamer rovnoměrně rozmístěných v úhlech 0–90°. Jako parametrizaci přitom zvolili AAM příznaky dynamizované metodou LDA. V experimentech autoři došli k závěru, že optimálním úhlem snímání je cca 30°. V dalších experimentech zaměřených na návrh parametrizace robustní vůči natočení řečníka pak postupovali shodným způsobem jako předchozí práce [Lucey 2007, Estellers 2012b], tj. příznaky z jiných než požadovaného pohledu (30°) transformovali lineární projekcí $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ odhadnutou metodou nejmenších čtverců (3.6). Pro příznaky transformované z 60° úhlu na 30° bylo v nejlepším případě

dosaženo pouhého 3 % relativního zhoršení WER, tedy téměř identické úspěšnosti jako při ideálním záběru. Bohužel však výsledky vzhledem k různým databázím a aplikacím nejsou s předchozími pracemi přímo porovnatelné.

Pass a kol. [Pass 2010] zvolili jiný postup parametrizace robustní vůči pohybům řečníka. Z obrazů dvou různých záběrů vypočítali dvourozměrné DCT příznaky a z jejich rozdílu sestavili kovarianční matici. Jako optimální robustní příznaky pak vybrali takové, které se mezi oběma pohledy příliš nelišily, tj. **rozptyl jejich rozdílu je minimální**. Pro otestování algoritmu nahráli vlastní databázi QuLips dvou mluvčích, jejichž promluvy byly zachyceny dvěma kamerami s proměnnými vzájemnými pozicemi. Oba mluvčí namluvili 180 číslovek pro každou dvojici pohledů, dohromady 3600 číslovek. V experimentech autoři dosáhli až 50 % relativní zlepšení WER oproti případu, kdy je systém trénovaný pouze z jednoho pohledu. Je ovšem nutné dodat, že MCPV (Minimum Cross-Pose Variance) příznaky jsou vázané na konkretní dvojici pohledů a tedy nepředstavují parametrizaci robustní vůči libovolným změnám v natočení.

4. Metody klasifikace

Pokud je z obrazu extrahována užitečná informace v podobě sekvence příznakových vektorů $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$, lze úlohu rozpoznávání řeči z pravděpodobnostního hlediska formulovat jako

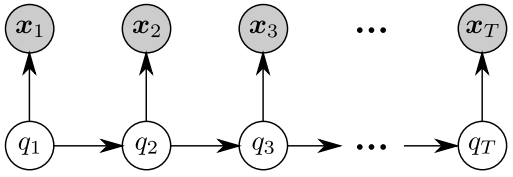
$$\boldsymbol{w}^* = \operatorname*{argmax}_{\boldsymbol{w}} p(\boldsymbol{w} \mid \boldsymbol{X}) = \operatorname*{argmax}_{\boldsymbol{w}} \frac{p(\boldsymbol{X} \mid \boldsymbol{w})\, p(\boldsymbol{w})}{p(\boldsymbol{X})}, \tag{4.1}$$

kde $\boldsymbol{w}^*$ je sekvence slov, jež nejlépe odpovídá pozorovaným datům. Pro jednoduchý model izolovaných slov s malým slovníkem lze $p(\boldsymbol{w} \mid \boldsymbol{X})$ vyhodnotit pro všechny možnosti a vybrat nejlepší. V principu pak lze použít libovolnou klasifikační metodu. U spojité řeči s neznámou délkou promluvy je však počet tříd (tj. různých sekvencí $\boldsymbol{w}$) příliš velký a tradiční, např. diskriminační, klasifikátory nejsou řešením.

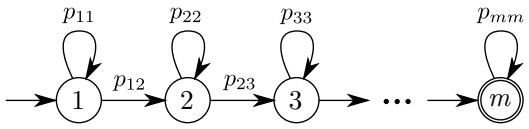
4.1 Skrytý markovský model

V oblasti automatického rozpoznávání řeči se zdaleka nejrozšířenějším způsobem klasifikace staly tzv. **skryté markovské modely** (Hidden Markov Model, HMM). Vycházejí z předpokladu, že lidskou řeč je možné rozdělit menší stacionární úseky, na nichž se statistické ani frekvenční vlastnosti signálu krátkodobě nemění. Tyto úseky lze považovat za stavy stochastického systému, jejichž správnou posloupností lze popsat či generovat libovolnou jazykovou jednotku. Úloha automatického rozpoznávání řeči pomocí HMM pak spočívá ve vyhodnocení pravděpodobnosti, s jakou mohl naučený model vygenerovat pozorovanou sekvenci řečových dat (např. příznaků), a jejím následném porovnání s modely ostatních jazykových jednotek.

HMM patří do skupiny dynamických generativních modelů, které popisují sekvenci pozorovaných dat na základě skrytých vnitřních stavů. Svojí strukturou odpovídá pravděpodobnostnímu grafickému modelu (Probabilistic Graphical Model, PGM) na obrázku 4.1. Šedě obarvená políčka značí viditelné (pozorované) proměnné, bílá pak skryté. V případě automatického odezírání ze rtů představuje pozorovanou posloupnost $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$, $\boldsymbol{x}_t \in \mathbb{R}^d$ vizuální parametrizace T snímků a $q_1, \ldots, q_T$, $q_t \in \{1, \ldots, m\}$ sekvenci vnitřních stavů, které příznakovou podobu generují. Na HMM je možné nahlížet jako na speciální případ diskrétní dynamické bayesovské sítě (Dynamic Bayesian Net, DBN), v níž každý stav q_t závisí právě na jednom předchozím stavu q_{t-1}. Tato závislost je vyjádřena jako pravděpodobnost $p_{ij} = p(q_t = j \mid q_{t-1} = i)$ přechodu ze stavu i do stavu j. Skryté markovské



Obrázek 4.1: HMM jako pravděpodobnostní grafický model.



Obrázek 4.2: Lineární HMM jako stochastický automat.

modely pro základní jazykové jednotky (fonémy, vizémy, celoslovní modely) mají p_{ij} nenulovou, pouze když $j \in \{i, i+1\}$, a označují se pak jako tzv. lineární levo-pravé HMM. Jejich forma odpovídá stochastickému konečnému automatu se stavovým diagramem 4.2. Jelikož HMM je generativní pravděpodobnostní model, každý stav má navíc asociovanou nějakou výstupní funkci popisující pravděpodobnostní hustotu dat. Obvyklou volbu v oblasti automatického rozpoznávání řeči představuje vícerozměrná gaussovská směs (Gaussian Mixture Model, GMM)

$$p\left(\boldsymbol{x}_t \,|\, q_t = i\right) = \sum_j \pi_{ij} \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}\right), \tag{4.2}$$

a výsledný model se pak běžně označuje jako HMM/GMM. Velmi populární pro modelování $p\left(\boldsymbol{x}_t \,|\, q_t\right)$ se pak v posledních několika letech staly hluboké neuronové sítě (Deep Neural Networks, DNN) [Bengio 2009], vedoucí na model HMM/DNN.

Pokud jsou přechodové pravděpodobnosti i parametry výstupních funkcí známé, optimální posloupnost vnitřních stavů nejlépe odpovídající sekvenci $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ se vzhledem k výše uvedeným vlastnostem grafického modelu 4.1 získá maximalizací

$$q_1^*, \ldots, q_T^* = \operatorname*{argmax}_{q_1,\ldots,q_T} \prod_{t=1}^{T} p\left(q_t \,|\, q_{t-1}\right) p\left(\boldsymbol{x}_t \,|\, q_t\right). \tag{4.3}$$

Úlohu (4.3) efektivním způsobem řeší Viterbiho algoritmus založený na dynamickém programování. Trénování HMM/GMM obvykle spočívá v maximálně věrohodném odhadu parametrů $\boldsymbol{\theta} = (p_{ij}, \pi_{ij}, \boldsymbol{\mu_{ij}}, \boldsymbol{\Sigma_{ij}})$ algoritmem Baum-Welch [Young 2006] na nějaké trénovací množině $\mathcal{X}_{\text{trn}}$:

$$\theta_{\text{ML}} = \operatorname*{argmax}_{\theta} p\left(\mathcal{X}_{\text{trn}} \,|\, \boldsymbol{\theta}\right) \tag{4.4}$$

Algoritmus je variantou EM a tedy kritérium (4.4) maximalizuje iterativně střídáním odhadového a maximalizačního kroku. Kromě maximálně věrohodného odhadu parametrů lze použít i diskriminační trénování HMM, např. optimalizaci vzájemné informace, minimalizaci klasifikačních chyb či maximalizaci odstupu [Heigold 2012].

4.2 Rozpoznávání izolovaných jednotek

Nejjednodušší úlohu automatického odezírání ze rtů představuje rozpoznávání izolovaných jednotek. Těmi mohou být např. jednoslovné hlasové povely, jednoduché fráze, ale i jednotlivé vizémy či fonémy.

V případě klasifikace pomocí **skrytých markovských modelů** se pak na základě příznakového popisu pro každou jednotku sestaví samostatný lineární HMM s m stavy a přechodovými pravděpodobnostmi p_{ji} dle obr. 4.2. Existují přitom dva základní způsoby tvorby lineárních HMM: **celoslovní** a **hláskové**. V prvním případě je HMM natrénován pro každou rozpoznávanou jednotku (např. slovo či fráze). V druhém případě se HMM trénují pro menší jazykové jednotky – **fonémy** pro akustické rozpoznávání řeči a **vizémy** pro automatické odezírání ze rtů – a výsledný model pro každou rozpoznávanou jednotku pak vznikne spojením těchto menších modelů za sebe. Jelikož se mohou vlastnosti shodných fonémů v různých slovech lišit, v oblasti automatického rozpoznávání řeči se častěji používají modely, jenž rozlišují fonémy podle jejich levém a pravém kontextu. Typicky se modelují např. všechny trojice, tedy tzv. trifóny. Zásadní nevýhodu ztěžující jejich aplikaci pro vizuální rozpoznávání však představuje nutnost velkého množství dat pro natrénování výsledného modelu. Při klasifikaci neznámé promluvy je pro každý naučený model odhadnuta optimální sekvence stavů (4.3) pomocí Viterbiho algoritmu a vybrán takový model, jehož pravděpodobnost je vzhledem k příznakové posloupnosti nejvyšší.

V úlohách rozpoznávání izolovaných jednotek je jednodušší použít celoslovní modely. Tento postup byl zvolen např. v pracích [Chaloupka 2005, Chaloupka 2011], kde autor řešil rozpoznávání 50 izolovaných českých slov pomocí 14-stavových celoslovních HMM. Pitsikalis a kol. [Pitsikalis 2006] použili osmistavové celoslovní HMM mimo jiné pro rozpoznávání izolovaných číslovek na databázi CUAVE. Stejnou úlohu řešili také Lucey a kol. v článku [Lucey 2008], pouze s jiným počtem stavů na model. Saitoh a kol. [Saitoh 2010] se zabývali rozpoznáváním japonských hlásek a slov z profilových videí, viz sekci 3.5, a pro rozpoznávání rovněž použili celoslovní HMM. Rozsáhlou prací porovnávající několik druhů příznaků je [Matthews 2002], kde autoři prováděli experimenty na databázi AVletters obsahující promluvy izolovaných anglických písmen.

Vzhledem k předpokladu, že rozpoznávané jednotky se ve videu vždy vyskytují a to pouze samostatně, lze v principu nahlížet na neznámou sekvenci holisticky a poté využít libovolnou metodu klasifikace. Způsob vypořádání se s časovým průběhem příznaků a měnící se řečovou dynamikou pak závisí zvoleném klasifikátoru.

Postup inspirovaný PCA projekcí ROI (sekce 3.1.2) byl aplikován v článku [Li 1995], kde autoři pomocí analýzy hlavních komponent redukovali celé sekvence obrázků. Pro zajištění stejné délky a normalizování řečové dynamiky byly sekvence časově zarovnány vůči nějaké referenci algoritmem DTW (Dynamic Time Warping). Analogicky ke statické variantě nazvali výsledné příznaky **eigensequences**. Tuto parametrizaci použili v úloze rozpoznávání izolovaných číslovek. Pro klasifikaci byla zvolena maximální projekční energie, kdy vstupní sekvence byla promítnuta na vlastní prostor každé číslovky. Pokud poměr L_2 norem vstupního a redukovaného vektoru byl blízko jedné, byla nalezena odpovídající číslovka.

Wang a kol. [Wang 2008a] modelovali časový průběh výšky a šířky rtů, ASM tvarových příznaků a viditelné plochy zubů pomocí **křivek b-spline** s přidaným gaussovským šumem. Křivky popisující stejnou jednotku (např. slovo) měly společné

základní parametry, jejichž hodnoty byly odhadnuty variantou algoritmu EM (Expectation Maximization). V případě pouze jednoho řečníka v systému byl pro každou třídu natrénován pouze jeden model, v opačném případě více modelů v závislosti na vizuální variabilitě. Rozpoznávání pak bylo založeno na proložení vstupní sekvence příznaků b-spline křivkou s parametry odhadnutými **metodou maximální věrohodnosti** pro každý naučený model a přiřazení toho nejpravděpodobnějšího.

V několika pracích se také objevil jednoduchý přístup spočívající v rozdělení nahrávek do konstantního počtu překrývajících se segmentů. Z každého úseku se vypočte reprezentativní příznakový vektor a následně pospojuje s parametrizací extrahovanou z ostatních segmentů, čímž vznikne jediný hypervektor popisující celou nahrávku. Jelikož kromě dělení nahrávky na úseky se nijak jinak s dynamikou a časovým průběhem nepracuje, lze zvolit libovolný klasifikátor. Např. v článku [Zhao 2009] byly na každém segmentu extrahovány příznaky LBPTOP a klasifikovány metodou SVM. Shodnou parametrizaci i klasifikátor použili i Zhou a kol. [Zhou 2011], avšak před dělením na segmenty zarovnávali nahrávky pomocí frekvenční interpolace. Ngiam a kol. [Ngiam 2011] extrahovali DCT příznaky z několika sousedících framů a následně je redukovali pomocí hlubokých neuronových sítí. Pro klasifikaci pak rovněž zvolili algoritmus SVM.

Ong a Bowden [Ong 2011] řešili úlohu rozpoznávání frází na databázi OuluVS pomocí tzv. **sekvenčních vzorů** (Sequential Pattern, SP). V článku byl navrhnut efektivní algoritmus jejich extrakce v rámci **boostingu** založeného na kombinaci velkého množství slabých klasifikátorů, viz sekci 2.2. Jako základní příznak autoři použili porovnání jasových hodnot několika párů pixelů, přičemž slabý klasifikátor zjišťoval, zda se nějaká konkrétní sekvence binárních porovnání vyskytuje ve vstupním videu. Přítomnost více takovýchto sekvencí pak tvořila základ pro výsledný silný klasifikátor. Jelikož kompletní prohledávání prostoru n-tic pixelových párů v každé iteraci boostingu není z důvodu časové náročnosti realizovatelné, navrhli autoři pro výběr optimálních dvojic algoritmus postupného prořezávání a stromového procházení v závislosti na horní a dolní mezi dosažitelné výsledné chybovosti.

Zhou a kol. [Zhou 2010] použili pro modelování časové závislosti **grafové vnořování** (angl. graph embedding). Na jednotlivé sekvence nahlíželi jako na grafy, ve kterých každý vrchol představuje jeden snímek. Hrany reprezentují sousednost a mají přiřazenou váhu na základě vzdálenosti snímků a jejich náležitosti ke konkrétní třídě. Grafové vnořování pak spočívá v projekci vrcholů, tedy např. příznakového popisu snímků, do nějakého prostoru, ve kterém jsou zachovány vzdálenosti vyjádřené vahami náležejících hran. Díky jednoduchosti navrženého grafu lze podobu báze hledaného prostoru odvodit analyticky. V navazujícím článku [Zhou 2011] autoři dokázali, že bazické vektory sestávají ze sinusových funkcí s různými frekvencemi. Systém natrénovaný pro rozpoznávání pouze od jednoho mluvčího (SD) klasifikoval neznámé promluvy na základě **korelace** promítnuté vstupní sekvence na prostor každé naučené fráze a výběr maxima. Nebyly přitom použity žádné příznaky, klasifikace probíhala přímo z jasových hodnot. V druhém

případě, tj. bez znalosti konečného uživatele systému (SI), autoři pro zvýšení robustnosti vůči vnější variabilitě použili příznaky LBPTOP a stejný postup klasifikace (SVM) jako v [Zhao 2009]. Před jejich extrakcí pouze normalizovali videa pomocí reprojekce z prostorů jednotlivých naučených frází.

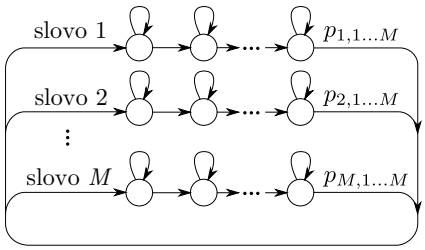
V zatím poslední práci [Zhou 2014] Zhou a kol. nahlíželi na proces generování příznakových či obrazových sekvencí jako na **lineární stochastický proces**. Identifikovali přitom dva hlavní zdroje variability: uživatelská a řečová. Uživatelská variabilita je způsobena rozdíly mezi jednotlivými řečníky, např. tvaru úst, ochlupení na tváři apod. Naopak řečová variabilita souvisí pouze s pohybem rtů a představuje tak zdroj užitečné informace. Pravděpodobnost pozorované sekvence $\boldsymbol{x}_t$ autoři definovali jako lineární model se skrytými proměnnými

$$p\left(\boldsymbol{x}_t \,|\, \boldsymbol{h}, \boldsymbol{z}_t\right) = \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu} + \boldsymbol{F}\boldsymbol{h} + \boldsymbol{G}\boldsymbol{z}_t, \boldsymbol{\Sigma}\right), \tag{4.5}$$

kde $\boldsymbol{F}$, $\boldsymbol{h}$ jsou prostor, resp. skryté parametry uživatelské variability a $\boldsymbol{G}$, $\boldsymbol{z}_t$ jsou prostor, resp. skryté parametry řečové variability. Časovou závislost skrytých proměnných $\boldsymbol{z}_t$ vyjádřili pomocí apriorní pravděpodobnosti $p\left(\boldsymbol{z}_t\right) = \mathcal{N}\left(\boldsymbol{z}_t; \boldsymbol{f}(t/T), \alpha\boldsymbol{I}\right)$, tj. sekvence skrytých proměnných by měla ležet blízko křivky $\boldsymbol{f}$ definované opět projekcí grafové reprezentace frází na sinusové složky. Pro natrénování modelu $\boldsymbol{\mu}, \boldsymbol{F}, \boldsymbol{h}, \boldsymbol{G}, \boldsymbol{z}_t$ autoři odvodili variantu algoritmu EM. Před klasifikací byly metodou maximální aposteriorní pravděpodobnosti (Maximum A Posteriori, MAP) odhadnuty skryté parametry $\boldsymbol{h}, \boldsymbol{z}_t$. Sekvence $\boldsymbol{z}_t$ pak byla klasifikována stejným způsobem jako v [Zhou 2011], tedy korelací se sinusoidami vypočtených z grafových reprezentací jednotlivých frází. I přes použití sofistikovanějšího modelu však Zhou a kol. dosáhli horších výsledků než v předchozích pracích.

Zatím nejlepších výsledků rozpoznávání frází a delších slovních jednotek dosáhli Pei a kol. [Pei 2013]. Příznaky extrahovali z lokálních oblastí kolem klíčových bodů v dolní části obličeje pomocí deskriptorů LBP a HOG (viz sekci 3.1.3) a ze změny jejich pozice. Na autorské databázi KinectVS, která to umožňovala, byly příznaky extrahované z obrazových i hloubkových dat získaných pomocí zařízení Kinect. Takto pro každý bod na obličeji vznikl příznakový popis, který byl redukován vícerozměrným škálováním (Multidimensional Scaling, MDS). MDS je nelineární metoda redukce dimenze, která promítá data do nového prostoru tak, aby párové vzdálenosti mezi jednotlivými body v původním prostoru byly maximálně zachovány. Pro výpočet efektivní párových podobností autoři použili **náhodné lesy** (angl. random forest). Jelikož každá promluva je pak reprezentována množinou vektorů odpovídajících redukovaným kombinovaným deskriptorům jednotlivých klíčových bodů, klasifikace probíhala na základě vyhledávání podobných skupin bodů v redukovaném prostoru, tzv. **zarovnáním nadploch** (angl. manifold alignment). Algoritmus byl otestován pro oba případy závislosti či nezávislosti na řečníkovi na několika databázích a dosahoval lepší úspěšnosti než metody Zhou a kol. [Zhou 2010, Zhou 2011, Zhou 2014] modelující časový průběh příznaků.

Uvedené metody vykazují podobné nevýhody jako parametrizace [Pachoud 2008, Zhao 2009] popsané v sekci 3.1.3, přičemž zde se projevuje především provázanost s konkrétní úlohou. Rozpoznávání celých slov či frází jako



Obrázek 4.3: Obecný stavový diagram HMM pro rozpoznávání spojité řeči.

nejmenších jednotek je totiž uzpůsoben i samotný návrh příznaků a klasifikátoru. Postup při aplikaci těchto sofistikovaných metod pro rozpoznávání založené na menších jednotkách, např. vizémech, nebo při kombinaci s akustickými příznaky, není zřejmý. V žádném článku nebyly uvedeny výsledky těchto klasifikátorů pro rozpoznávání spojité řeči, pouze v [Zhou 2014] autoři provedli experimenty s klasifikací jednotlivých vizémů.

4.3 Rozpoznávání spojité řeči

Složitější úlohu představuje rozpoznávání spojité řeči. V obecném případě se v neznámé sekvenci může vyskytovat libovolný počet slov a kromě samotné klasifikace je tedy nutné řešit zároveň i jejich lokalizaci. Pokud je struktura spojitých promluv nějak blíže specifikována, např. způsobem „figurka-na-pozice“ pro hru šachy, nutnost lokalizace sice částečně odpadá, i tak ale přetrvává problém kombinatorické exploze. Při rozpoznávání spojitých promluv jako celých frází by totiž i za použití malých slovníků vznikl obrovský počet slovních permutací a systém by pak potřeboval příliš velké množství trénovacích dat. Z těchto důvodů pro odezírání spojité řeči v literatuře jednoznačně převažují HMM, jelikož umožňují poměrně jednoduše rozšířit lineární levo-pravé modely pro případ spojité řeči bez nutnosti vytvářet holistickou reprezentaci každé fráze jako metody v sekci 4.2. Podobně jako lze z hláskových modelů sestavit modely pro celá slova či fráze jejich pospojováním za sebe, lze spojitou řeč modelovat jediným HMM vytvořeným pomocí zpětných vazeb mezi všemi dvojicemi naučených modelů, jak je schématicky znázorněno na obr. 4.3.

Protože samotný vizuální kanál neobsahuje dostatek informace pro spolehlivé odezírání ze rtů s libovolným slovníkem [Potamianos 2003], obvykle se v této úloze uvažuje pouze omezený slovník o několika málo položkách. Případně, jak bylo uvedeno výše, se předpokládá nějaká specifická struktura promluvy, např. „akce-předmět-pozice“, přičemž každá z částí promluv má definovaný vlastní slovník. Přechodové pravděpodobnosti ve stavovém diagramu 4.3 jsou v takovém případě nenulové pouze pro dvojice slov ze sousedních částí. Tento model byl zvolen např. při nahrávání audiovizuálních databází GRID a WAPUSK20.

Velmi často řešenou úlohou v oblasti spojitého odezírání ze rtů je **rozpoznávání sekvencí číslovek**. Např. v [Heckmann 2002b] autoři extrahovali z obrázků ROI

pravděpodobnostní parametrizaci založenou na klasifikaci DCT do vizémových tříd pomocí vícevrstvého perceptronu a výsledek klasifikovali **hláskovými HMM**. Experimenty však byly provedeny na neveřejné databázi s pouze jedním řečníkem. Rozsáhlejší přehledové experimenty pro rozpoznávání spojitých promluv číslovek provedli Potamianos a kol. [Potamianos 2003]. Pro rozpoznávání bylo na databázi, jež obsahovala přibližně 10 hodin spojitých promluv číslovek od 50 mluvčích ve studiovém prostředí, natrénováno 159 **kontextových** (až 11 fonémů) **hláskových HMM** pro celkem 22 fonémů. Svůj kaskádový algoritmus extrakce parametrizace založený na LDA dynamizaci DCT příznaků pak Potamianos a kol. otestovali kromě jiného na databázi spojitých číslovek namluvených v rušném prostředí 10 řečníky. Experimenty byly provedeny ve více-uživatelském schématu (MS). Rozpoznáváním spojených číslovek se zabývali také Fu a kol. [Fu 2008]. Pro extrakci příznaků navrhli vlastní algoritmus redukce dimenze založený na kombinaci grafového vnořování a Fisherova kritéria (3.5). Pro každou číslovku natrénovali 7stavový **celoslovní HMM**. Algoritmus testovali na databázi AVICAR, v níž jsou řečníci nasnímání čtyřmi různými kamerami, a dosáhli zde zatím nejlepších výsledků. Galatas a kol. [Galatas 2012] využili v úloze rozpoznávání spojených číslovek obrazová a hloubková data z Kinectu. Úlohu řešili na vlastní databázi BAVCD obsahující 6 nahrávek o 5 číslovkových sekvencích od 15 mluvčích. Klasifikace probíhala pomocí dvoukanálových **trifónových HMM** se třemi stavy.

Lan a kol. [Lan 2009] se zabývali rozpoznáváním **strukturované spojité řeči** s malým slovníkem na databázi GRID. Tato poměrně rozsáhlá databáze obsahuje cca 1000 nahrávek od každého z celkem 34 mluvčích, přičemž každá z vět je tvořena libovolnou kombinací slov ve formátu „příkaz-barva-předložka-písmeno-číslovka-příslovce“. Celková velikost slovníku je pouze 51 slov. V práci bylo porovnáno několik různých parametrizací (DCT, síta, AAM), přičemž pro rozpoznávání byly aplikovány **celoslovní HMM**. Z důvodu možných pauz ve větách byl natrénován navíc jeden model pro ticho. V rámci navazující práce [Lan 2010] pak autoři vytvořili vlastní databázi sestávající z nahrávek spojité řeči od 12 mluvčích. Slovník obsahoval přibližně 1000 položek. Pro rozpoznávání autoři opět použili HMM s modelem pro ticho navíc, tentokrát však na **vizémové** úrovni bez kontextu. Experimenty byly vyhodnoceny pouze z hlediska rozpoznávání jednotlivých vizémů, nikoliv slov, a to ve dvou konfiguracích MS a SI (závislost na řečníkovi). Nejvyšší dosažená úspěšnost byla cca 42 % pro SI a 50 % pro případ MS. Úspěšnost pro akustické MFCC příznaky v analogické úloze dosahovala přibližně 79 %.

Výjimku v rozpoznávání spojité řeči představuje článek [Pachoud 2008], kde místo skrytého markovského modelu autoři navrhli vlastní algoritmus. Jak bylo uvedeno v sekci 3.4.2, na video bylo nahlíženo jako na trojrozměrné kvádry s hranami odpovídajícími dvěma prostorovým a jedné časové složce. Nahrávky byly rozděleny na tzv. makro-kvádry, ze kterých byla extrahována vizuální parametrizace. S neznámou vstupní sekvencí se pracovalo obdobně. Porovnáním makro-kvádrů z naučeného modelu a ze vstupní sekvence pak vznikly histogramy nejpravděpodobnějších výskytů pro každé slovo ze slovníku. Algoritmus byl otestován na databázi CUAVE a úspěšnost pro jednotlivé číslovky se pohybovala v rozmezí 70–

100 % s výjimkou číslovky 8, u níž byla schopnost detekce a rozpoznání téměř nulová. Autoři to zdůvodnili nedostatkem pohybu rtů při vyslovování této číslovky v angličtině, díky čemuž navržená parametrizace nebyla schopná zachytit rozdíly oproti klidovému stavu. Postup umožňuje detekci libovolného počtu slov v sekvenci, avšak vzhledem k vyčerpávajícímu způsobu detekce není příliš vhodný pro práci s většími slovníky. Dalším problémem je nemožnost uvažovat různé přechodové pravděpodobnosti mezi jednotlivými slovníkovými položkami, neboli nějakou formu jazykového modelu. Nehodí se rovněž pro kombinaci více parametrizací, např. pro audiovizuální rozpoznávání. Algoritmus je tedy poměrně výrazně provázaný s cílovou aplikací.

5. Kombinace více zdrojů

Jak bylo uvedeno v předchozích kapitolách, vizuální signál sám o sobě neobsahuje dostatek informace pro spolehlivé odezírání ze rtů s větším slovníkem, a proto ho v této úloze může být vhodnější využít pouze jako doplňkový zdroj k akustickému rozpoznávání řeči, např. za účelem zvýšení robustnosti vůči hlukům na pozadí. Analogicky však lze zohlednit více zdrojů i v úloze čistě vizuálního odezírání ze rtů, kdy může být z video signálu extrahováno více druhů parametrizací zachycujících různé typy informace. V této kapitole jsou popsány metody integrace více informačních kanálů, jež se objevují v literatuře zabývající se problematikou audiovizuálního rozpoznávání řeči. Principiálně je lze rozdělit do tří hlavních skupin:

1. brzká (Early Integration, EI),
2. střední (Middle Fusion, MF),
3. pozdní (Late Integration, LI).

Metody **brzké integrace**, někdy označované také jako příznaková fúze (Feature Fusion, FF), kombinují informaci z více kanálů na úrovni parametrizace, tj. před samotnou klasifikací. Nejčastěji toho dosahují promítáním příznakových vektorů do společného prostoru a proces rozpoznávání dále neovlivňují. Významnou výhodu tohoto přístupu představuje možnost zkoumat korelace a jiné závislosti mezi všemi uvažovanými informačními kanály a tím potenciálně redukovat počet volných parametrů klasifikátoru. Jelikož však ke kombinaci dochází na parametrizační úrovni, je nutné, aby všechny kanály byly synchronní. Asynchronnost lze sice modelovat, ovšem pouze pokud je k dispozici dostatečné množství trénovacích dat. Pokud není trénovací sada dostatečně reprezentativní, anebo příznaky nejsou synchronní, vzájemné závislosti kanálů mohou být narušeny. Při příznakové fúzi je rovněž velmi obtížné modelovat spolehlivost, tedy například pokud jeden z kanálů chybí nebo je nějak zarušen.

Nevýhody brzké integrace do určité míry odstraňuje **pozdní integrace**, v některých pracích označované jako fúze rozhodnutí (Decision Fusion, DF). V tomto případě dochází ke kombinaci klasifikačního skóre z jednotlivých kanálů, které jsou zpracovány odděleně. Propojení je tedy prováděno vždy až na konci promluvy. Pokud některý z kanálů chybí, nebo je zarušen, může odpovídající klasifikátor promítnout tento fakt do svého výstupního skóre, např. v podobě konfidenčního intervalu či jiného typu nejistoty. Při fúzi pak lze dynamicky volit důležitost jednotlivých kanálů dle jejich spolehlivosti. Nevýhodu naopak představuje ztráta informace o vzájemných závislostech mezi informačními kanály a to jak statických, tak dynamických.

Kromě uvedených způsobů se v literatuře objevuje ještě hybridní metoda kombinace, tzv. **střední fúze**, která se snaží kombinovat vlastnosti obou předchozích metod. Toho je docíleno provázáním klasifikace a podmínky částečné synchronizace a z tohoto důvodu se objevuje prakticky výlučně v kombinaci

se skrytými markovskými modely či jejich zobecněním v podobě dynamických bayesovských sítí.

5.1 Brzká integrace

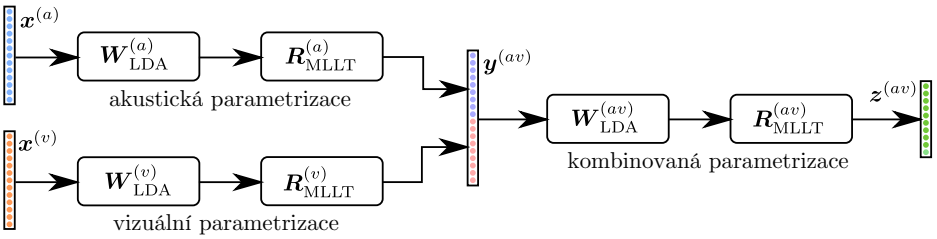
Nejjednodušší a zároveň nejčastější formou kombinace informace na příznakové úrovni je **vektorové spojení** $\boldsymbol{x} = [\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(S)}], \boldsymbol{x}^{(s)} \in \mathbb{R}^{d_s}$, viz např. [Adjoudani 1996, Lucey 2005, Galatas 2012]. Další možnosti zahrnují **vážení příznaků**: např. Heckmann a kol. [Heckmann 2001, Heckmann 2002b] extrahovali pravděpodobnostní parametrizaci pomocí jednoduché neuronové sítě naučenou klasifikovat příznaky akustického a vizuálního kanálu do fonetických tříd a kombinovat je jako geometrický průměr; Teissier a kol. [Teissier 1999] použili ke klasifikaci metodu maximální pravděpodobnosti za předpokladu gaussovského rozdělení a příznaky z obou kanálů vážili na úrovni kovariančních matic. V této práci zároveň shrnuli tři základní způsoby integrace na příznakové úrovni v závislosti na tom, zda se transformují vizuální příznaky do akustického prostoru, naopak, či do společného prostoru. Do kategorie příznakové fúze také spadají některé metody **zvýrazňování řeči** (angl. speech enhancement), které se snaží modifikovat akustickou parametrizaci na základě vizuálních či kombinovaných příznaků, viz např. [Deligne 2002, Goecke 2002].

Jelikož při nějaké formě vektorového spojování má výsledný vektor dimenzi rovnou součtu partikulárních rozměrů, mohou při nedostatku dat vznikat problémy s odhadem parametrů klasifikátoru v trénovací fázi a hrozí tedy riziko přeučení. V několika pracích [Neti 2000, Matthews 2001, Potamianos 2001b, Potamianos 2004] se proto autoři snažili riziko přeučení minimalizovat LDA projekcí příznakových vektorů. Redukcí dimenze se totiž obvykle sníží celkový počet volných parametrů klasifikátoru a úloha jejich odhadu (trénovaní modelu) se stane lépe podmíněnou. Jako třídy jednotlivých datových vzorků lze použít fonémy či vizémy. V [Matthews 2001] Matthews a kol. aplikovali LDA na spojené příznakové vektory ze sousedních snímků, viz sekci 3.4.1. Kromě LDA autoři zmíněných prací použili navíc maximálně věrohodnou lineární transformaci (Maximum Likelihood Linear Transformation, MLLT) dat, která maximalizuje věrohodnost dat v původním příznakovém prostoru za předpokladu diagonální kovariance v transformovaném prostoru. Ortogonální rotační matice $\boldsymbol{R}_{\mathrm{MLLT}}$ je získána maximalizací

$$\boldsymbol{R}_{\mathrm{MLLT}} = \underset{\boldsymbol{R}}{\operatorname{argmax}} \det(\boldsymbol{R})^{L} \prod_{c \in \mathcal{C}} \det\left(\operatorname{diag}\left(\boldsymbol{R}\boldsymbol{\Sigma}_c\boldsymbol{R}^{\top}\right)\right)^{-\frac{L^{(c)}}{2}}, \tag{5.1}$$

kde L je celkový počet vzorků v trénovací databázi, L_c počet vzorků náležející do třídy c a $\boldsymbol{\Sigma}_c$ kovarianční matice třídy c. Redukci LDA a rotaci MLLT autoři [Potamianos 2004] aplikovali na akustický i vizuální kanál a poté i na spojené příznaky, viz obr. 5.1, a výsledný algoritmus nazvali „**Hierarchická LDA**“ (HiLDA).

Několik autorů [Ngiam 2011, Huang 2013] se inspirovalo úspěchem hlubokých neuronových sítí (Deep Neural Network, DNN) v akustickém rozpoznávání řeči,



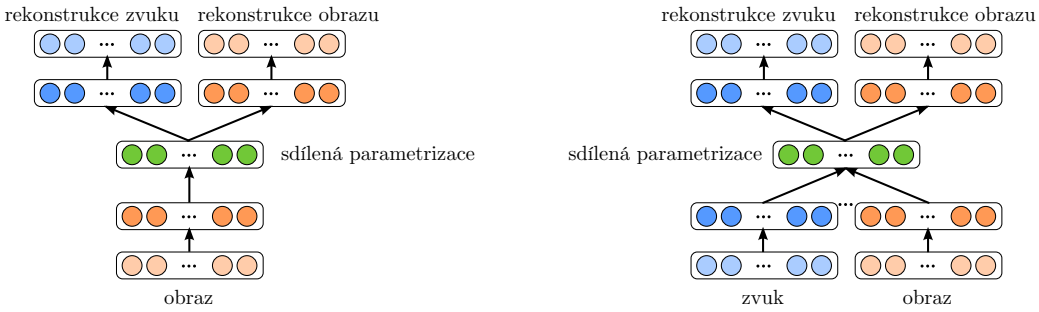
Obrázek 5.1: Hiearchická LDA (HiLDA) pro audiovizuální fúzi.

zpracování textu, klasifikaci neznámých objektů v obraze a dalších oblastech a navrhli pro integraci akustické a vizuální parametrizace **hluboké autoenkodéry** (Deep Autoencoder, DAE). Autoenkodér je hybridní způsob před-trénování neuronové sítě částečně bez učitele, u něhož nejnižší vrstva představuje samotná data a nejvyšší pak jejich rekonstrukci. Úkolem autoenkodéru je transformovat vstup do nějaké tzv. skryté reprezentace a následně ji rekonstruovat zpět do původního prostoru tak, aby bylo zachováno maximum informace dle nějakého kritéria. Skrytá vrstva sítě může obsahovat menší počet buněk než viditelné vrstvy a rekonstrukce je tedy v obecném případě ztrátová. Autoenkodér však nepředstavuje konkrétní metodu trénování, jedná se pouze o rámcovou formulaci učení neuronové sítě. Od PCA a příbuzných metod se tedy liší v několika bodech: volné kritérium, případná nelinearita projekce, možnost více vrstev a potenciálně vyšší rozměr transformovaných dat. Pouze za určitých podmínek tak lze na autoenkodér nahlížet jako na zobecnění PCA. Existují dva hlavní směry řešení této úlohy: pravděpodobnostní a optimalizační. První skupina nahlíží na neuronovou síť jako na generativní pravděpodobnostní model a během trénování se snaží maximalizovat věrohodnost dat vzhledem k vnitřním parametrům. Druhá skupina definuje nějakou explicitní rekonstrukční odchylku (kritérium), např. L_2, a parametry pak optimalizuje pomocí standardních numerických metod. Pro přehled viz [Bengio 2009].

Ngiam a kol. [Ngiam 2011] zvolili první způsob, kdy každou vrstvu sítě reprezentoval **omezený Boltzmannův stroj** (Restricted Boltzmann Machine, RBM). Výsledná mnohovrstvá síť pak patří do skupiny bayesovských a v angličtině se běžně označuje jako Deep Belief Net (DBN), bohužel se stejnou zkratkou jako odlišný pojem dynamické bayesovské sítě. RBM, přestože na něj lze nahlížet jako na generativní pravděpodobnostní model, má primárně definovanou funkci energie $E(\boldsymbol{v}, \boldsymbol{h})$, která exponenciálně souvisí se sdruženou hustotou pravděpodobností $p(\boldsymbol{v}, \boldsymbol{h})$ skrytých a viditelných proměnných $\boldsymbol{h}$, resp. $\boldsymbol{v}$:

$$-\log p(\boldsymbol{v}, \boldsymbol{h}) = E(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{\sigma^2}\left(\frac{1}{2}\boldsymbol{v}^\top\boldsymbol{v} - \boldsymbol{c}^\top\boldsymbol{v} + \boldsymbol{b}^\top\boldsymbol{h} + \boldsymbol{h}^\top\boldsymbol{W}\boldsymbol{v}\right), \tag{5.2}$$

kde $\boldsymbol{W}$ jsou váhy spojů mezi vrstvami. Autoři modifikovali RBM přidáním regularizujícího členu zajišťujícího řídkou aktivaci skryté reprezentace a parametry optima-



Obrázek 5.2: Příklady hlubokých autoenkodérů použitých v [Ngiam 2011].

lizovali dle

$$\boldsymbol{W}^*, \boldsymbol{b}^*, \boldsymbol{c}^* = \underset{\boldsymbol{W},\boldsymbol{b},\boldsymbol{c}}{\operatorname{argmin}} \; -\log p\left(\boldsymbol{v}, \boldsymbol{h}\right) + \lambda \sum_j \left\{ \rho - \frac{1}{m} \left(\sum_k E\left[h_j \,\middle|\, \boldsymbol{v}^{(k)} \right] \right) \right\}, \quad (5.3)$$

kde ρ je požadovaná průměrná aktivace skrytých buněk a $E[\cdot]$ značí očekávanou (střední) hodnotu. Pro optimalizaci (5.3) byl vrstvu po vrstvě aplikován upravený Hintonův Monte-Carlo algoritmus [Hinton 2006]. V článku bylo zváženo několik různých architektur hluboké sítě, z nichž dvě ilustruje obr. 5.2. DAE vlevo je trénován pro případ chybějících akustických dat, DAE vpravo pak pro standardní případ obou modalit. Vstupem (nejspodnější vrstva) byly v článku [Ngiam 2011] vektorově spojené parametrizace z několika sousledných snímků – vybělené spektrogramy pro audio a PCA koeficienty pro obraz. Huang a kol. [Huang 2013] zvolili MFCC, resp. DCT/LDA.

Algoritmus byl v [Ngiam 2011] otestován na dvou databázích AVletters a CUAVE v úloze rozpoznávání izolovaných anglických písmen a číslovek. V obou případech autoři dosáhli o 13–36 % relativní WER lepších výsledků než ostatní metody ověřené na stejných databázích. Bohužel však Ngiam a kol. místo tradičního lineárního HMM použili pro klasifikaci SVM, a přímé porovnání s HiLDA projekcí z jiných prací tak nemohlo být provedeno. Nejbližší srovnání, avšak v rámci dynamizace, nikoliv fúze, nabízí výsledky rozpoznávání pomocí vizuálního DAE (obr. 5.2 vlevo) a LDA dynamizovaných příznaků v článku [Lucey 2006b], kde však Lucey a kol. kromě HMM pro klasifikaci extrahovali i jiný typ vizuální parametrizace. Výsledky přesto naznačily, že tradiční metoda LDA dynamizace, nikoliv fúze, vykazuje lepší úspěšnost rozpoznávání (-36 % relativní WER). V práci [Huang 2013] nebyla úspěšnost DAE s ostatními metodami porovnána vůbec.

5.2 Pozdní integrace

Jak již bylo zmíněno, jedním z hlavním problémů metod brzké integrace je obtížné modelování spolehlivosti jednotlivých kanálů. Např. při zarušení akustického signálu šumem na pozadí je vhodné využít spíše vizuální informaci, naopak při selhání některého z algoritmů předzpracování obrazu spoléhat pouze na zvukovou složku.

Metody příznakové fúze přitom obvykle pracují s oběma kanály rovnocenným způsobem, a tak nereflektují případné změny v podmínkách.

Pokud jsou k dispozici příznakové sekvence $\boldsymbol{X}^{(s)} = \left(\boldsymbol{x}_1^{(s)}, \ldots, \boldsymbol{x}_T^{(s)}\right)$ extrahované z jednotlivých kanálů $s = 1, \ldots, S$ z celé neznámé promluvy $\boldsymbol{w}$, pak metody pozdní integrace provádějí klasifikaci na základě kombinace dílčích výstupních pravděpodobností odpovídajících klasifikátorů $p\left(\boldsymbol{w} \,\middle|\, \boldsymbol{X}^{(s)}\right)$. Odhad pravděpodobnosti nějaké sekvence slov $\boldsymbol{w}$ tedy je

$$\zeta\left(\boldsymbol{w} \,|\, \boldsymbol{X}\right) = F_{\text{comb}}\left\{p\left(\boldsymbol{w} \,\middle|\, \boldsymbol{X}^{(1)}\right), \ldots, p\left(\boldsymbol{w} \,\middle|\, \boldsymbol{X}^{(S)}\right)\right\}, \tag{5.4}$$

kde $F_{\text{comb}}\{\cdot\}$ představuje kombinační funkci. Odhad $\zeta\left(\boldsymbol{w} \,|\, \boldsymbol{X}\right)$ nereprezentuje v obecném případě skutečnou pravděpodobnost, ale pouze skóre, jenž je jí úměrné. Vytváří však stejné rozhodovací hranice pro výslednou klasifikaci, kterou lze na základě $\zeta\left(\boldsymbol{w} \,|\, \boldsymbol{X}\right)$ formulovat jako

$$\boldsymbol{w}^* = \operatorname*{argmax}_{\boldsymbol{w}} \zeta\left(\boldsymbol{w} \,|\, \boldsymbol{X}\right). \tag{5.5}$$

Úlohou je tedy nalézt takovou sekvenci slov $\boldsymbol{w}$, pro kterou je výsledné skóre $\zeta\left(\boldsymbol{w} \,|\, \boldsymbol{X}\right)$ kombinované přes všechny kanály maximální.

5.2.1 Metody kombinace skóre z více klasifikátorů

Za předpokladu statistické nezávislosti kanálů, potažmo skóre klasifikátorů, je optimální [Kittler 1998, Lucey 2005] kombinace dle pravidla součinu

$$\zeta_{\text{pr}}^*\left(\boldsymbol{w} \,|\, \boldsymbol{X}\right) = p\left(\boldsymbol{w}\right)^{-S+1} \prod_{s=1}^{S} p\left(\boldsymbol{w} \,\middle|\, \boldsymbol{X}^{(s)}\right), \tag{5.6}$$

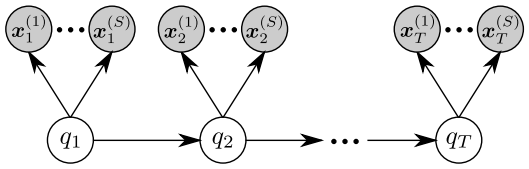
kde $p\left(\boldsymbol{w}\right)$ je apriorní pravděpodobnost sekvence $\boldsymbol{w}$. Tento způsob však nezohledňuje spolehlivost jednotlivých modalit a v praxi se tak častěji používá pravidlo **váženého součinu**

$$\zeta_{\text{pr}}\left(\boldsymbol{w} \,|\, \boldsymbol{X}\right) = p\left(\boldsymbol{w}\right)^{-S+1} \prod_{s=1}^{S} p\left(\boldsymbol{w} \,\middle|\, \boldsymbol{X}^{(s)}\right)^{\lambda^{(s)}}, \tag{5.7}$$

které každé modalitě přiřazuje váhu $\lambda^{(s)}$ reflektující její spolehlivost. Pravidlo (5.7) bylo pro audiovizuální kombinaci empiricky ověřeno v mnoha pracích a přestože se z pravděpodobnostního pohledu nejedná standardní operaci, podává lepší výsledky než za ideálních podmínek optimální součin (5.6).

Teoretický rámec vysvětlující rozdíly v úspěšnosti rozpoznávání v závislosti na zvoleném pravidlu navrhl Lucey ve svém článku [Lucey 2005]. Soustředil se přitom na neshodu mezi trénovací a testovací množinou, kdy věrohodnost $p\left(\boldsymbol{X} \,|\, \boldsymbol{w}\right)$ nutnou pro vyhodnocení aposteriorní pravděpodobnosti $p\left(\boldsymbol{w} \,|\, \boldsymbol{X}\right)$ dle Bayesova pravidla je možné rozdělit na dva členy

$$p\left(\boldsymbol{X} \,|\, \boldsymbol{w}\right) = p\left(\Omega\right) p\left(\boldsymbol{X} \,|\, \mathcal{X}_{\text{trn}}\right) + p\left(\bar{\Omega}\right) p\left(\boldsymbol{X} \,\middle|\, \bar{\Omega}\right), \tag{5.8}$$



Obrázek 5.3: Grafický model vícekanálového HMM pro kombinaci příznaků.

kde $p(\Omega)$ značí apriorní pravděpodobnost, že $\boldsymbol{X}$ pochází ze stejné distribuce jako trénovací množina $\mathcal{X}_{\text{trn}}$, značeno $\boldsymbol{X} \in \mathcal{X}_{\text{trn}}$. První člen tedy reprezentuje znalost klasifikátoru vzhledem k trénovacím datům, zatímco druhý člen představuje znalost statistických vlastností $\boldsymbol{X}$ pocházejících z jiného pravděpodobnostní distribuce. Z definice úlohy však není možné $p(\Omega)$ a $p\left(\boldsymbol{X} \,\middle|\, \bar{\Omega}\right)$ z dostupných dat odhadnout a druhý člen tak představuje nejistotu v rozhodování. Za ideálních podmínek je $p(\Omega)\, p(\boldsymbol{X} \,|\, \mathcal{X}_{\text{trn}}) \gg p\left(\bar{\Omega}\right) p\left(\boldsymbol{X} \,\middle|\, \bar{\Omega}\right)$ a Bayesovo pravidlo tak zůstává nezměněno. Avšak pokud je např. akustický signál při audiovizuálním rozpoznávání silně zarušen šumem, trénovací a testovací množiny si neodpovídají a nejistotu není možné zanedbat. V takovém případě se výsledná pravděpodobnost $p\left(\boldsymbol{w} \,\middle|\, \boldsymbol{X}^{(s)}\right)$ příliš neliší od apriorní pravděpodobnosti $p(\boldsymbol{w})$ a pravidlo součinu pak lze aproximovat **pravidlem součtu**

$$\zeta_{\text{sum}}(\boldsymbol{w} \,|\, \boldsymbol{X}) = (1-S)p(\boldsymbol{w}) + \sum_{s=1}^{S} \lambda^{(s)} p\left(\boldsymbol{w} \,\middle|\, \boldsymbol{X}^{(s)}\right), \tag{5.9}$$

kde váhy $\lambda^{(s)}$ podobně jako v případě váženého součinu mohou reflektovat nepoměr mezi oběma členy (5.8). Případně lze pro eliminaci nejistoty aplikovat výše uvedené pravidlo váženého součinu (5.7).

5.3 Střední integrace

Hybridní metodu kombinující vlastnosti příznakové i rozhodovací fúze představuje tzv. střední integrace. Vě většině případů je založená na rozšíření skrytých markovských modelů na více kanálů. V literatuře se přitom objevují dvě základní varianty: **synchronní** vícekanálový HMM a **asynchronní** vícekanálový HMM, jenž se odlišují podle toho, zda jsou jednotlivé modality (kanály v HMM) zpracovávány a trénovány synchronně či asynchronně.

5.3.1 Vícekanálové synchronní HMM

Jednodušší variantu s předpokladem vzájemné nezávislosti modálně specifických příznaků vzhledem k aktuálnímu stavu představují **synchronní vícekanálové HMM** (Multistream Synchronous HMM, MSHMM), znázorněné na obr. 5.3. MSHMM se díky své jednoduchosti a efektivitě staly velmi populárním způsobem integrace více typů příznaků. Výstupní pravděpodobnost každého stavu pro

příznakové kanály $\boldsymbol{x}_t^{(1)}, \ldots, \boldsymbol{x}_t^{(S)}$ je dle modelu na obr. 5.3 při $\lambda^{(s)} = 1$

$$p\left(\boldsymbol{x}_t^{(1)}, \ldots, \boldsymbol{x}_t^{(S)} \,\middle|\, q_t\right) = \prod_{s=1}^{S} p\left(\boldsymbol{x}_t^{(s)} \,\middle|\, q_t\right)^{\lambda^{(s)}}. \tag{5.10}$$

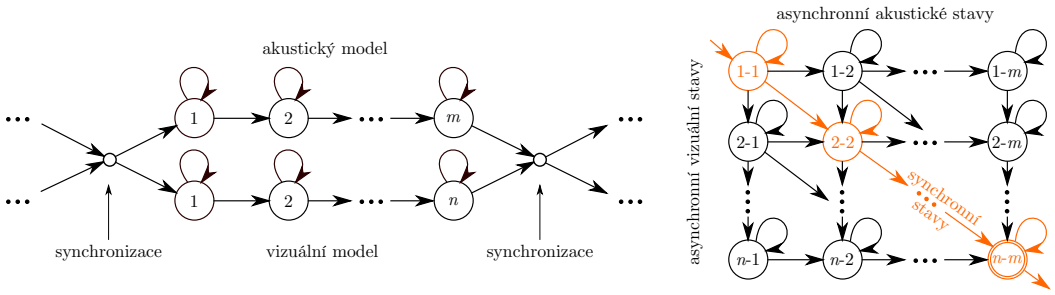
Jedná se tedy o shodný popis jako v případě obyčejného HMM natrénovaného nad vektorově spojenými parametrizacemi $\boldsymbol{x}_t = \left(\boldsymbol{x}_t^{(1)}, \ldots, \boldsymbol{x}_t^{(S)}\right)$, pouze s předpokladem vzájemné nezávislosti některých příznaků. Na MSHMM tak lze za určitých podmínek nahlížet i jako na specializaci klasického HMM.

V principu lze pro kombinaci pravděpodobností z jednotlivých kanálů použít libovolné z pravidel uvedených v sekci 5.2.1. Vztah (5.10) ovšem reprezentuje hustotu pravděpodobnosti pouze za použití pravidla součinu a při $\lambda^{(s)} = 1$ pro $s = 1, \ldots, S$. V rámci zohlednění spolehlivosti jednotlivých modalit se v literatuře častěji objevuje varianta s nejednotkovými vahami $\lambda^{(s)}$ a $p\left(\boldsymbol{x}_t^{(1)}, \ldots, \boldsymbol{x}_t^{(S)} \,\middle|\, q_t\right)$ pak představuje pouze skóre. Jelikož se MSHMM odlišuje od klasického jednokanálového HMM pouze předpokladem nezávislosti některých příznaků, lze pro trénování a rozpoznávání aplikovat téměř shodné postupy.

5.3.2 Asynchronní modely fúze

Při kombinaci různých typů parametrizace může docházet k situacím, kdy jednotlivé příznakové kanály **nejsou vzájemně synchronní**. Nejčastěji tento jev nastává při audiovizuální fúzi, tj. při kombinaci akustických a vizuálních příznaků, kdy obě modality nemusejí mít shodnou ani vzorkovací frekvenci. Zároveň se zde také projevuje variabilní prodleva mezi pohybem částí vokálního traktu, rtů a výsledným zvukem, která se v angličtině označuje jako „voice onset time“ (VOT) [Dupont 2000]. Je definována jako prodleva mezi zvukem z rázové a ze znělé části souhlásky, což přibližně odpovídá zpoždění či předstihu pohybu rtů vůči pohybu hlasivek. Různé práce se zabývaly časovými závislostmi akustického a vizuálního kanálu a jejich vlivem na srozumitelnost řeči. Z experimentů vyplynulo, že nejčastěji se prodleva pohybuje řádově do 100 ms [Grant 2001, Potamianos 2003, Grant 2004, Conrey 2006]. Při audiovizuálním rozpoznávání u člověka se zároveň ukazuje, že zpoždění akustického signálu oproti vizuálnímu v asymetrickém rozmezí přibližně -50 ms až +250 ms nemá na srozumitelnost řeči téměř žádný vliv. Např. dle [Conrey 2006] nejsou lidé schopni detekovat asynchronnost, pokud se vizuální signál oproti akustickému nezpožďuje o více než 300 ms nebo ho předbíhá max. o 500 ms. U lidského vnímání se tak ukazuje jistá asymetrie, kdy předbíhání vizuálního signálu oproti akustickému se projevuje v menší míře než opačný případ. Je zde také patrná poměrně výrazná tolerance vůči asynchronnostem, jež se pravděpodobně vyvinula v důsledku řečové variability mezi různými lidmi.

Synchronní vícekanálové skryté markovské modely lze zobecnit na asynchronní přidáním dodatečných stavů specifických pro jednotlivé modality. Tento způsob byl poprvé uveden v článku [Dupont 2000], kde Dupont a kol. kombinovali akustické a vizuální příznaky. V tomto případě na rozdíl od MSHMM model skutečně

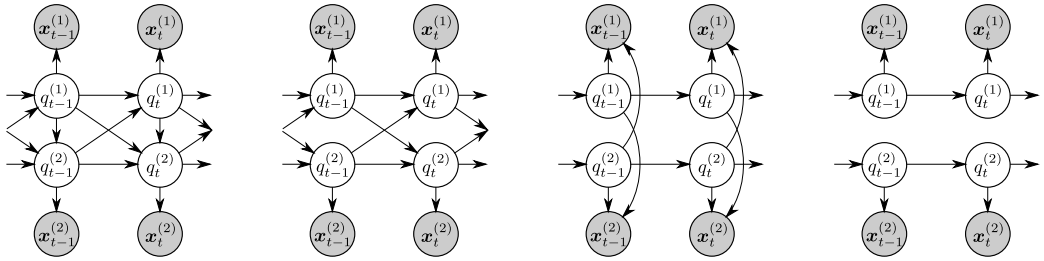


Obrázek 5.4: Dvoukanálový kompozitní HMM (vlevo) a jeho stavový diagram.

obsahuje samostatný kanál pro každou modalitu, čímž umožňuje asynchronní zpracování informace. V článku [Lucey 2005] byl tento model označen jako **vícekanálový asynchronní HMM** (Multistream Asychronous HMM, MAHMM) a ilustruje ho levá část obr. 5.4. Z pohledu grafických modelů se jedná o instanci tzv. **dynamické bayesovské sítě** (Dynamic Bayesian Net, DBN), jejíž počet markovských řetězců odpovídá počtu integrovaných modalit. Jednotlivé kanály modelu vlevo nemusejí mít nutně stejnou topologii a mohou tedy zahrnovat různý počet stavů. V předběžných experimentech však Dupont a kol. došli k závěru, že v případě neomezené asynchronnosti dochází především ve vizuálním kanálu k chybám zarovnání, a tak do modelu zahrnuli navíc synchronizační body. Ty nepředstavují plnohodnotný stav HMM, ale pouze omezení tabulky možných přechodů, a jejich úroveň (hláska, slovo, věta) nutné je definovat apriori.

Ekvivalentním způsobem popisu MAHMM je tzv. **vícerozměrný HMM** zobrazený na obr. 5.4 vpravo pro případ dvou modalit (zvuk, video). Tento model je založený na kompozitních stavech $\boldsymbol{q}_t = \left(q_t^{(1)}, \ldots, q_t^{(S)}\right)$ složených z jednotlivých modálně specifických stavů $q_t^{(s)}$ a přechodové pravděpodobnosti definuje mezi nimi. Při pohledu na HMM jako stavový automat vznikne standardním průnikem dílčích HMM a Potamianos a kol. ho tak označili jako tzv. **součinový HMM** (Product HMM, PHMM). Množinu kompozitních stavů tedy představují všechny možné uspořádané S-tice, kde S je počet kanálů. Např. stavový diagram synchronního HMM pak odpovídá diagonále dvourozměrného HMM, v tomto případě totiž oba kanály procházejí shodnou sekvencí. Stavy mimo diagonálu reprezentují asynchronnost, přičemž její míru lze kontrolovat maximální vzdáleností od diagonály.

V obecném případě nemá sdružená výstupní pravděpodobnost $p\left(\boldsymbol{x}_t^{(1)}, \ldots, \boldsymbol{x}_t^{(S)} \,\middle|\, \boldsymbol{q}_t\right)$ ani přechodová pravděpodobnost $p\left(\boldsymbol{q}_t \,\middle|\, \boldsymbol{q}_{t-1}\right)$ žádnou specifickou faktorizaci. Takový model se však stává ekvivalentní s brzkou integrací a z hlediska modelování asynchronnosti nezajímavý. Obvykle se tak termínem **součinový/kompozitní HMM** (PHMM) označuje nějaká jeho zjednodušená forma. Např. v kompozitním HMM uvedeném v [Potamianos 2003] jsou modální



součinový HMM sdružený HMM faktorový HMM nezávislý HMM

Obrázek 5.5: Varianty dekompozice vícerozměrného HMM pro $S = 2$.

příznaky $\boldsymbol{x}^{(s)}$ závislé pouze na stavu $q^{(s)}$ odpovídajícího markovského řetězce:

$$p\left(\boldsymbol{x}_t \,|\, \boldsymbol{q}_t\right) = \prod_{s=1}^{S} p\left(\boldsymbol{x}_t^{(s)} \,\middle|\, q_t^{(s)}\right)^{\lambda^{(s)}}, \tag{5.11}$$

kde $\boldsymbol{x}_t = \left(\boldsymbol{x}_t^{(1)}, \ldots, \boldsymbol{x}_t^{(S)}\right)$, opět s případným zahrnutím spolehlivosti kanálů v podobě vah $\lambda^{(s)}$. Kompozitní stavy $\boldsymbol{q}_t$ tak nejsou kompletně unikátní, parametry výstupních hustot pravděpodobnosti jsou sdílené po sloupcích i řádcích odpovídajících modalit. Model v této podobě je zachycen na obr. 5.5 vlevo.

Podle způsobů dekompozice přechodových a výstupních pravděpodobností lze z obecného součinového HMM odvodit další typy modelů, jejichž nejčastější varianty jsou rovněž znázorněny na obr. 5.5. V literatuře se poměrně často objevuje případ tzv. **sdruženého HMM** (Coupled HMM, CHMM), jenž předpokládá specifickou formu přechodových pravděpodobností

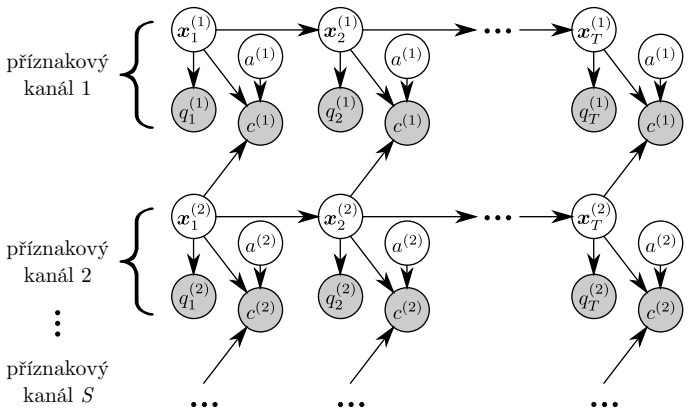
$$p\left(\boldsymbol{q}_t \,|\, \boldsymbol{q}_{t-1}\right) = \prod_{s=1}^{S} p\left(q_t^{(s)} \,\middle|\, \boldsymbol{q}_{t-1}\right). \tag{5.12}$$

Nejjednodušší formou MAHMM je **nezávislý kompozitní HMM** (Independent HMM, IHMM) zobrazený na obr. 5.5 vpravo. Faktorizace jeho přechodových pravděpodobností má tvar

$$p\left(\boldsymbol{q}_t \,|\, \boldsymbol{q}_{t-1}\right) = \prod_{s=1}^{S} p\left(q_t^{(s)} \,\middle|\, q_{t-1}^{(s)}\right) \tag{5.13}$$

a jednotlivé kanály jsou tedy nezávislé. Poměrně často využívanou formou DBN je tzv. **faktorový HMM** (Factorial HMM, FHMM), který používá stejný model přechodových pravděpodobností jako IHMM. Na rozdíl od něj však nezpracovává kanály nezávisle, ale synchronizuje je nepřímo skrze specifickou formu dekompozice výstupních pravděpodobností

$$p\left(\boldsymbol{x}_t \,|\, \boldsymbol{q}_t\right) = \prod_{s=1}^{S} \prod_{r=1}^{S} p\left(\boldsymbol{x}_t^{(s)} \,\middle|\, q_t^{(r)}\right)^{\lambda^{(s)}}. \tag{5.14}$$



Obrázek 5.6: DBN pro částečnou synchronizaci více příznaků [Saenko 2005].

V jistém smyslu se tedy jedná o opak synchronního vícekanálového HMM, kdy více stavů generuje jednu (či více) pozorovanou veličinu. Varianty vícekanálových asynchronních HMM včetně experimentů s audiovizuálním rozpoznáváním řeči shrnuje článek [Nefian 2002].

Saenko a kol. ve svých článcích [Saenko 2005, Saenko 2006] navrhli pravděpodobnostní model znázorněný na obr. 5.6. Model byl určen pro vyjádření časové závislosti několika artikulačních příznaků, viz sekci 3.2, a v dalším článku pak aplikován pro audiovizuální fúzi. Synchronizace je založená na pomocných proměnných a a c, přičemž a značí rozdíl mezi stavy přidružených kanálů a c tuto hodnotu kontroluje pomocí odpovídajícího pravděpodobnostního rozložení.

Nevýhodu při aplikací uvedených variant DBN/MAHMM představuje nutná úprava trénovacích a dekódovacích algoritmů. Stejně jako v případě lineárního HMM se přitom většinou jedná o nějakou formu metody EM pro trénování a Viterbiho algoritmu pro dekódování.

5.4 Modelování spolehlivosti kanálů

V předchozích sekcích byly uvedeny nejčastěji užívané metody integrace více příznakových kanálů do výsledného klasifikátoru. Mnoho z těchto metod je založeno na vážené kombinaci jednotlivých kanálů, např. pravidla pro pozdní integraci (5.7) a (5.9) či vícekanálový synchronní HMM (5.10). Je zřejmé, že klíčový vliv na výslednou úspěšnost rozpoznávání pak má volba integračních vah $\lambda^{(s)}$, jež by měly optimálně reflektovat spolehlivost odpovídajících modalit.

Integrační váhy mohou být nastaveny buď staticky, nebo se během promluvy dynamicky měnit v závislosti na aktuálním odhadu spolehlivosti kanálů. V případě **statických vah** se jejich optimální hodnoty obvykle získají učením na nějaké validační trénovací sadě. Vzhledem k formě výstupních pravděpodobností však není možné použít standardní techniky maximální věrohodnosti a je nutné přejít k diskriminačním metodám jejich odhadu [Potamianos 2003]. Ty zahrnují např. metodu pravděpodobnostního spádu (Generalized Probabilistic Descent,

GPD) [Potamianos 1998a, Nakamura 2000, Gravier 2002], maximální vzájemnou informaci (Maximum Mutual Information, MMI) [Jourlin 1997], či maximalizace entropie [Gravier 2002]. Při fúzi akustického a vizuálního kanálu pro audiovizuální rozpoznávání řeči v hlučném prostředí je velmi častým způsobem optimalizace vah přímá minimalizace výsledné chybovosti WER. Obvykle se přitom postupuje vyčerpávajícím způsobem, tedy ze všech možných kombinací vah se vybere ta s nejvyšší úspěšností rozpoznávání na nějakých validačních datech. Jelikož je požadován pouze argument maximální úspěšnosti, nikoliv její hodnota, pro zjednodušení problému lze uvažovat pouze takové kombinace, pro než $\sum_{s=1}^{S} \lambda^{(s)} = 1$. V případě audiovizuální rozpoznávání řeči se pak problém stává jednorozměrným, jelikož $\lambda^{(v)} = 1 - \lambda^{(a)}$. Toto zjednodušení navíc zachovává poměr logaritmů přechodových a výstupních pravděpodobností v HMM, jež společně představují jednotku míry v úloze Viterbiho dekódování (4.3). Za předpokladu hladkého průběhu WER v závislosti na $\lambda^{(s)}$ či v případě takového množství kanálů, které vyčerpávající způsob znemožňuje, lze rovněž použít numerické metody optimalizace.

Při audiovizuálním rozpoznávání řeči v hlučném prostředí může však dojít ke krátkodobým neočekávaným hlukům na pozadí, např. projíždějící automobil, výkřik, kýchnutí apod. Ke krátkodobému zašumění může dojít i v obrazovém signálu, např. změnou osvětlení scény, neočekávaným pohybem řečníka atd. Praktičtějším způsobem je v takových případech **dynamická regulace** kombinačních vah v závislosti na měnící se kvalitě jednotlivých kanálů. Na úrovni každého snímku lze vyhodnotit míru spolehlivosti některé či všech modalit a na jejím základě nastavit optimální integrační váhy. Vznikají zde však dva základní podproblémy: jakou zvolit míru spolehlivosti a jakým způsobem ji promítnout do nastavení optimálních vah.

5.4.1 Dynamický odhad spolehlivosti

Při audiovizuálním rozpoznávání řeči se jako míra spolehlivosti často využívá **poměr signálu a šumu** (Signal to Noise Ratio, SNR) [Adjoudani 1996, Teissier 1999, Shao 2008, Galatas 2012, Estellers 2012a], jenž je definovaný jako poměr výkonu zajímavého signálu vůči šumu. Pro dva stejně dlouhé úseky řeči $\boldsymbol{x}$ a šumu $\boldsymbol{\varepsilon}$ lze SNR v decibelech vypočítat jako poměr

$$\mathrm{SNR}_{\mathrm{dB}} = 10 \log \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{\varepsilon}\|}. \tag{5.15}$$

Obvykle se výpočet provádí na základě akustického signálu, kdy se na začátku promluvy předpokládá krátká chvíle ticha (tj. neobsahující řeč), na níž je možné vyhodnotit klíčové charakteristiky šumu. Problém tohoto postupu však představují některé typy hluků, např. lidská řeč na pozadí, jenž nejsou stacionární a v průběhu času mění. Lze ale použít i sofistikovanější metody výpočtu SNR s využitím detektorů hlasové aktivity (Voice Activity Detection, VAD). Takto postupovali např. Estellers a kol. v článku [Estellers 2012a], kteří při dynamické regulaci kombinačních vah zacházeli se SNR vypočítaným z řečových neřečových segmentů různým způsobem. Pro oba typy SNR natrénovali jinou mapovací funkci, jež

převáděla naměřené hodnoty na váhy pro akustický a obrazový signál. Shao a Baker [Shao 2008] natrénovali pro odhad SNR vícevrstvý perceptron. Jako příznaky použili výstupní pravděpodobnosti akustického a vizuálního HMM, jenž dekódovaly signál odděleně. Výstupní vrstva perceptronu tak pro regresi SNR využívala jak zvukovou, tak obrazovou složku. Kromě SNR lze srozumitelnost a kvalitu řeči v akustickém signálu odhadovat také pomocí tzv. hlasového indexu (Voicing Index, VI) [Glotin 2001, Heckmann 2002a], jenž měří poměr harmonických a neharmonických složek v signálu.

Dalšími poměrně často užívanými příznaky spolehlivosti [Potamianos 2003, Marcheret 2007, Estellers 2012a] jsou **entropie** aposteriorních pravděpodobností stavů

$$H_t^{(s)} = -\sum_{i=1}^{L} p\left(q_t^{(s)} = i \,\middle|\, \boldsymbol{x}_t^{(s)}\right) \log p\left(q_t^{(s)} = i \,\middle|\, \boldsymbol{x}_t^{(s)}\right), \tag{5.16}$$

logaritmické rozdíly N nejpravděpodobnějších stavů (angl. N-best log-likelihood difference)

$$L_t^{(s)} = \frac{1}{N-1} \sum_{n=2}^{N} \log \frac{p\left(\boldsymbol{x}_t^{(s)} \,\middle|\, q_{t,1}^{(s)}\right)}{p\left(\boldsymbol{x}_t^{(s)} \,\middle|\, q_{t,n}^{(s)}\right)} \tag{5.17}$$

a **disperze** N nejpravděpodobnějších stavů (angl. N-best log-likelihood dispersion)

$$D_t^{(s)} = \frac{2}{N\,(N-1)} \sum_{n=1}^{N} \sum_{m=n+1}^{N} \log \frac{p\left(\boldsymbol{x}_t^{(s)} \,\middle|\, q_{t,n}^{(s)}\right)}{p\left(\boldsymbol{x}_t^{(s)} \,\middle|\, q_{t,m}^{(s)}\right)}. \tag{5.18}$$

Na rozdíl od SNR jsou $H_t^{(s)}$, $L_t^{(s)}$ a $D_t^{(s)}$ modálně nezávislé a mohou být tedy vyhodnoceny na libovolném kanálu HMM. Stavové logaritmické rozdíly a disperze jsou založeny na úvaze, že u nezarušeného signálu bude výstupní pravděpodobnost nejpravděpodobnějšího stavu $q_{t,1}^{(s)}$ podstatně vyšší než výstupní pravděpodobnost n-tého nejpravděpodobnějšího stavu $q_{t,n}^{(s)}$, zatímco u signálů zatížených šumem budou hodnoty spíše uniformní. Disperze $D_t^{(s)}$ pak tuto úvahu rozšiřuje dále na rychlost klesání výstupní pravděpodobnosti v závislosti na pořadí stavu dle jeho pravděpodobnosti výskytu v čase t. Entropie pracuje s podobnou myšlenkou, avšak na úrovni aposteriorních stavových pravděpodobností.

Estellers a kol. [Estellers 2012a] si však všimli rozdílných výsledků $H_t^{(s)}$, $L_t^{(s)}$ a $D_t^{(s)}$ podle toho, jakým způsobem byly modelovány výstupní stavové pravděpodobnosti $p\left(\boldsymbol{x}_t^{(s)} \,\middle|\, q_{t,1}^{(s)}\right)$. Zatímco pro HMM využívající GMM model podávají lepší výsledky rozdíl a disperze, pro HMM založené na modelování výstupních pravděpodobností pomocí neuronových sítí vykazuje se jako vhodnější jeví entropie. Navrhli proto jinou metriku, jež sleduje a postupně akumuluje pouze **přechodové pravděpodobnosti** $p\left(q_t^{\mathrm{ML}} \,\middle|\, q_{t-1}^{\mathrm{ML}}\right)$ a není tak závislá na výstupním modelu:

$$C_t = C_{t-1} + p\left(q_t^{\mathrm{ML}} \,\middle|\, q_{t-1}^{\mathrm{ML}}\right) \tag{5.19}$$

Pro nespolehlivý signál se jako nejpravděpodobnější často vyhodnocují stavy, které neodpovídají lineárnímu HMM či omezení jazykovým modelem a v takových případech je tedy $p\left(q_t^{\mathrm{ML}} \,\middle|\, q_{t-1}^{\mathrm{ML}}\right) = 0$. Vysoké hodnoty C_t naopak značí čistý a nezarušený signál.

5.4.2 Nastavení vah na základě odhadu spolehlivosti

Pokud je k dispozici odhad spolehlivosti např. ve formě příznaků uvedených v předchozí sekci, je nezbytné jej převést na konkrétní hodnoty kombinačních vah. V zásadě je úkolem najít zobrazení $\lambda^{(s)} = f\left(\boldsymbol{y}\right)$, kde $\boldsymbol{y}^{(s)}$ je příznakový vektor a z důvodu přehlednosti je vynechán časový index t. Pro tento účel lze využít libovolnou funkci s vhodnými vlastnostmi. Např. v [Meier 1996, Glotin 2001] autoři použili po částech lineární funkci.

Poměrně častou volbou je sigmoid

$$\lambda^{(s)} = \frac{1}{1 + \exp\left(w_0 + w^\top \boldsymbol{y}^{(s)}\right)} \tag{5.20}$$

použitý např. v článcích [Potamianos 2003, Garg 2003, Marcheret 2007]. V čem se však jednotlivé práce liší, jsou metody odhadu parametrů funkce (5.20). Potamianos a kol. [Potamianos 2003] aplikovali a porovnali dva různé algoritmy: maximální podmíněná věrohodnost (Maximum Conditional Likelihood, MCL) a minimální klasifikační chyba (Minimum classification Error, MCE). Metoda MCL patří do skupiny pravděpodobnostních, kdy je maximalizována aposteriorní pravděpodobnost zvolených slovních jednotek (např. fonémů či vizémů) v závislosti na extrahované parametrizaci ze všech kanálů. MCE je naopak diskriminační způsob trénování, kdy je cílem minimalizovat výslednou klasifikační chybu slovních jednotek. Marcheret a kol. [Marcheret 2007] získali w_0, $\boldsymbol{w}$ standardní metodou nejmenších čtverců, kdy trénovací sadu tvořily vyčerpávajícím způsobem nalezené optimální hodnoty $\lambda^{(s)}$ pro různě zašuměná akustická data.

Autoři prací [Gurbuz 2002, Shao 2008] namísto analytické formy transformační funkce f zvolili vyhledávací tabulku (Lookup Table, LUT), jejíž hodnoty, podobně jako v práci [Marcheret 2007], byly nalezeny vyčerpávajícím způsobem.

Kromě sigmoidové funkce Marcheret a kol. [Marcheret 2007] navrhli také statistické modelování kombinačních vah v závislosti na příznacích spolehlivosti. Rozsah $0, \ldots, 1$ vah $\lambda^{(a)}$ pro akustický kanál nejprve kvantovali do několika úrovní a na každé z nich odhadli aposteriorní rozložení $p\left(\lambda_j^{(a)} \,\middle|\, \boldsymbol{y}^{(a)}\right)$ ve formě gaussovské směsi (GMM). Optimální $\lambda^{(a)}$ byla vybrána buď jako střední nebo maximální hodnota modelovaného rozložení.

V práci [Estellers 2012a] autoři navrhli transformační funkci ve tvaru $\lambda^{(s)} = A_1 e^{B_1 y^{(s)}} + A_2 e^{B_2 y^{(s)}}$, jejíž parametry A_1, A_2, B_1, B_2 byly odhadnuty metodou nejmenších čtverců. Trénovací data byla získána podobně jako v předchozích pracích vyčerpávajícím hledáním optimálních kombinačních vah pro jednotlivé SNR. Tvar funkce byl zvolen především pro spravedlivé porovnání různých příznaků

spolehlivosti tak, aby při odhadu parametrů A_1, A_2, B_1, B_2 vykazovaly podobnou průměrnou čtvercovou odchylku.

5.4.3 Ostatní metody zohlednění spolehlivosti

Kromě tradičního způsobu odhadu příznaků spolehlivosti a jejich transformace na kombinační váhy se v literatuře objevují i odlišné postupy. Např. Kolossa a kol. [Kolossa 2009] navrhli postup maskování modálních příznaků, kdy byly výstupní pravděpodobnosti sdruženého HMM/GMM $p(\boldsymbol{x}_t \,|\, q_t)$ (viz (4.2)) vypočteny pouze na základě podmnožiny původních vektorově spojených příznaků a jednalo se tedy o formu brzké integrace. Pro video byl natrénován jednoduchý klasifikátor, jenž určoval, zda je zájmová oblast extrahována správně či nikoliv. Bylo toho docíleno porovnáváním pravděpodobností dvou modelů $p\left(\boldsymbol{x}_t^{(v)} \,\middle|\, \mathcal{X}_{\text{trn}}^{(v)}\right)$ a $p\left(\boldsymbol{x}_t^{(v)} \,\middle|\, \bar{\mathcal{X}}_{\text{trn}}^{(v)}\right)$, kde $\mathcal{X}_{\text{trn}}^{(v)}$, $\bar{\mathcal{X}}_{\text{trn}}^{(v)}$ jsou množiny všech správně, resp. nesprávně zarovnaných trénovacích ROI. Akustické příznaky byly maskovány obdobným, byť jednodušším, způsobem, kdy hodnota každého koeficientu byla porovnána s odpovídající hodnotou extrahovanou z prvních neřečových 250 ms.

V práci [Stewart 2014] bylo porovnáno několik variant maxima aposteriorní stavové pravděpodobnosti pro kombinaci akustických a vizuálních příznaků. U těchto metod se ze všech možných $p(q_t \,|\, \boldsymbol{x}_t)$ vybírá ta, jejíž hodnota je maximální v závislosti na zvolené proměnné. Optimální $p(q_t \,|\, \boldsymbol{x}_t)$ pak aproximuje a nahrazuje emisní pravděpodobnost $p(\boldsymbol{x}_t \,|\, q_t)$. První porovnávanou variantou v článku je metoda MSP (Maximum Stream Posterior), která se soustředí na výběr nejméně zarušeného příznakového kanálu. Je zvolen takový kanál, jenž maximalizuje $p(q_t \,|\, \boldsymbol{x}_t)$, tj.

$$p(q_t \,|\, \boldsymbol{x}_t) = \max\left\{p\left(q_t \,\middle|\, \boldsymbol{x}_t^{(a)}\right), p\left(q_t \,\middle|\, \boldsymbol{x}_t^{(v)}\right), p\left(q_t \,\middle|\, \boldsymbol{x}_t^{(av)}\right)\right\}, \tag{5.21}$$

kde $\boldsymbol{x}^{(av)} = \left(\boldsymbol{x}_t^{(a)}, \boldsymbol{x}_t^{(v)}\right)$. Druhou porovnávanou variantou je MWSP (Maximum Weighted Stream Posterior), kdy je $p(q_t \,|\, \boldsymbol{x}_t)$ vyjádřena formou váženého součinu (5.7) a vybrána je taková váha λ, pro níž

$$p(q_t \,|\, \boldsymbol{x}_t) = \max_{\lambda^{(a)}} p\left(q_t \,\middle|\, \boldsymbol{x}_t; \lambda^{(a)}\right). \tag{5.22}$$

Optimální váha je vybrána pomocí vyčerpávajícího hledání z konečného počtu možností na intervalu $\langle 0, 1\rangle$. Třetí porovnávanou variantou je metoda PUM (Posterior Union Model), která vyjadřuje aposteriorní stavovou pravděpodobnost pomocí pravidel součtu (5.9) a součinu (5.6) a z obou vybírá maximum. Dle experimentů s audiovizuálním rozpoznáváním spojitých číslovek v hlučném prostředí na databázi XM2VTS se jako nejlepší jevila metoda MWSP.

Poměrně výjimečný počin v oblasti audiovizuální fúze pro robustní rozpoznávání řeči v hlučném prostředí představuje práce [Papandreou 2009], ve které Papandreou a kol. modelovali nejistotu přímo na příznakové úrovni. Na parametrizaci $\boldsymbol{x}_t$ nahlíželi jako na měření původních (skrytých) nezarušených hodnot $\boldsymbol{z}_t$ s přidaným

nezávislým aditivním šumem $\boldsymbol{e}_t$, tj. $\boldsymbol{x}_t = \boldsymbol{z}_t + \boldsymbol{e}_t$. Aposteriorní stavovou pravděpodobnost v MSHMM se součinovým pravidlem pak lze vyjádřit marginalizací skrytých proměnných

$$p\left(q_t \,|\, \boldsymbol{x}_t\right) \propto p\left(q_t\right) \prod_{s=1}^{S} \int_{-\infty}^{\infty} p\left(\boldsymbol{x}_t^{(s)} \,\middle|\, \boldsymbol{z}_t^{(s)}\right) p\left(\boldsymbol{z}_t^{(s)} \,\middle|\, q_t\right) \mathrm{d}x. \tag{5.23}$$

Pokud jsou $p\left(\boldsymbol{x}_t^{(s)} \,\middle|\, \boldsymbol{z}_t^{(s)}\right)$ i $p\left(\boldsymbol{z}_t^{(s)} \,\middle|\, q_t\right)$ gaussovská (GMM) rozložení, pak je $p\left(q_t \,|\, \boldsymbol{x}_t\right)$ rovněž gaussovské (GMM) a díky nezávislosti příznaků a šumu jeho střední hodnota i kovariance odpovídají součtu obou dílčích rozložení. Papandreou a kol. v práci ukázali, že za předpokladu, kdy kovariance šumu je násobkem kovariance příznaků a v aposteriorní GMM se jednotlivé gaussiány příliš nepřekrývají, lze z jejich modelu odvodit pravidlo váženého součinu (5.7) a tím získat jeho pravděpodobnostní interpretaci.

Klíčovým problémem modelu (5.23) je samotný odhad nejistoty, tedy střední hodnoty a kovariance šumu pro akustický a vizuální kanál. Pro detekci klíčových bodů na tváři i následnou parametrizaci autoři použili metodu AAM (viz sekce 2.3 a 3.2). Volba jim umožnila vyjádřit nejistotu pomocí hessiánu optimalizačního kritéria (2.9). V případě akustického kanálu použili autoři pro odhad nejistoty algoritmy zvýrazňování řeči. Svoji metodu Papandreou a kol. otestovali na databázi CUAVE v úloze audiovizuálního rozpoznávání izolovaných číslovek pro dva typy vícekanálových HMM: MSHMM a PHMM (viz sekce 5.3.1 a 5.3.2). I přes svojí sofistikovanost model podal lepší výsledky pouze pro poměrně vysoká SNR. Pro silně zarušený zvukový signál se jako vhodnější jevil standardní MSHMM s pravidlem váženého součinu. Metoda je rovněž svázaná s předzpracováním obrazu a následnou parametrizací – např. u diskriminačních metod detekce obličejových částí je modelování nejistoty obtížné.

6. Audiovizuální databáze

Většina algoritmů odezírání ze rtů a audiovizuálního rozpoznávání řeči je založena na strojovém učení a statistických metodách, které pro natrénování a správné fungování modelů vyžadují velké množství dat. Nároky na objem dat vznikají, rovněž pokud je nutné věrohodně porovnat efektivitu více různých přístupů, např. pro extrakci příznaků či metod audiovizuální fúze. V obou případech, pokud není k dispozici dostatek trénovacích, validačních a testovacích dat, hrozí riziko přeučení a dosažené výsledky tak nebudou reprodukovatelné v reálných aplikacích.

Zatímco v oblasti akustického rozpoznávání řeči byl tento problém již vyřešen, pro oblast audiovizuálního rozpoznávání stále neexistuje dostatek dostupných bimodálních dat, jež by bylo možné využít pro rozsáhlejší srovnávací experimenty. Mezi hlavní překážky patří především časová a finanční náročnost vytváření audiovizuálních dat. Jak bylo uvedeno k kapitolách 3, 4, 5, úspěšnost audiovizuálního rozpoznávání ovlivňuje celá řada různých faktorů: variabilita osvětlení, pohyb řečníka, úhel kamerového pohledu, zdrojová data, atd. Je velmi obtížné, aby jedna databáze pokrývala všechny tyto faktory. Pokud je např. cílem studovat závislost úspěšnosti rozpoznávání na světelných podmínkách, je nutné nahrát stejnou promluvu pro každé nasvícení zvlášť, čímž výrazně stoupá časová náročnost pořízení databáze. Zároveň také klesá ochota dobrovolníků takovou proceduru podstoupit. V oblasti klasického rozpoznávání řeči lze tento problém částečně obejít, protože různé hlukové podmínky je možné do jisté míry simulovat, např. v podobě aditivního šumu. Realistická simulace osvětlení či natočení kamery v obrazovém záznamu je však v porovnání mnohem obtížnější úloha.

Kromě úkolů spojených se zpracováním audia s sebou příprava audiovizuální databáze přináší i velký objem práce související s konfigurací systému a manuální anotací videa. Např. algoritmy využívající více kamer (stereovidění) jsou kriticky závislé na správné kalibraci a synchronizaci celého systému. Vizuální variabilita má rovněž významný vliv na lokalizaci zájmové oblasti a natrénování spolehlivého detektoru tak může vyžadovat anotaci dodatečných dat.

Většinu uvedených problémů lze vyřešit dostatkem financí, avšak v takových případech často vznikají licenční problémy. Rozsáhlejší databáze připravené s finanční podporou např. komerčních subjektů, zahrnující velké množství řečníků a spojitou řeč s velkým slovníkem, tak obvykle nebývají veřejně dostupné.

Existující audiovizuální databáze lze porovnávat na základě celé řady kritérií: typ obsažených promluv, velikost slovníku, počet mluvčích, charakter zdrojových dat, jazyk, či dostupnost. Tato kritéria jsou zohledněna v tabulce 6.1, jež předkládá přehled nejčastěji využívaných databází v novější literatuře.

Z hlediska promluv se nejčastěji objevují databáze izolovaných číslovek, v tab. 6.1 označené jako IČ. Velikost slovníku se pohybuje v rozmezí 10–13 podle toho, zda je zahrnutý spec. model pro ticho a jsou povoleny různé výslovnosti – např. v angličtině se nula může vyslovovat jako „zero“ nebo „oh“. Velmi podobnou úlohu představuje rozpoznávání izolovaných písmen (IP), kde je slovník o něco větší. Velmi populární především v oblasti odezírání ze rtů bez využití akustických dat je

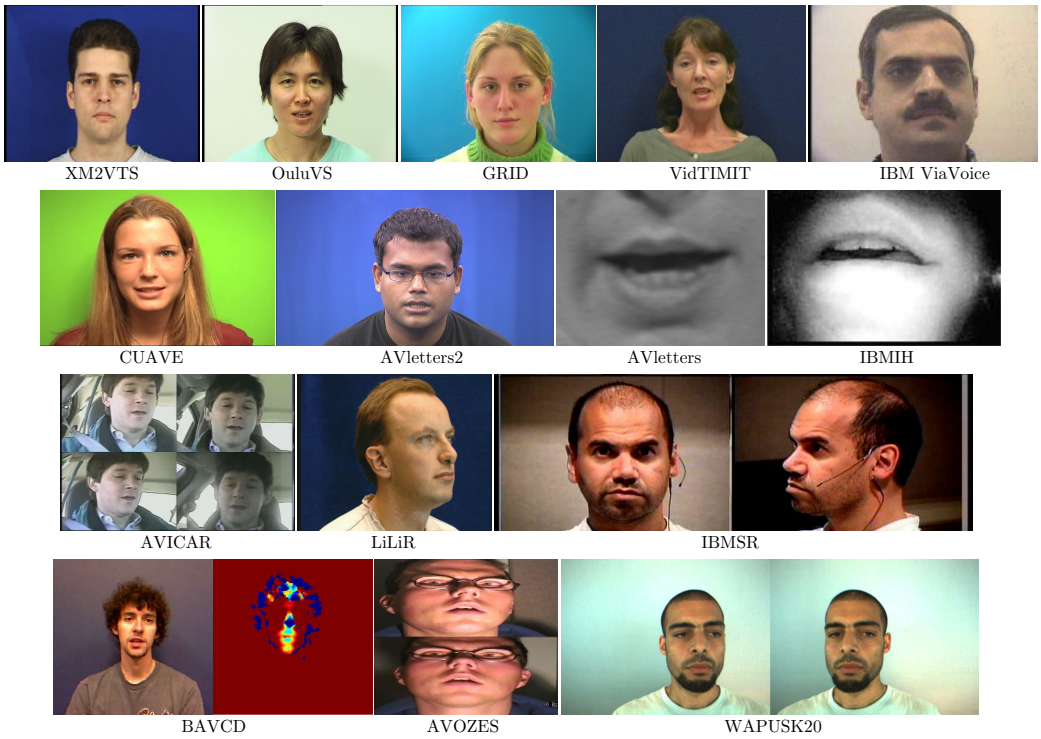
rozpoznávání izolovaných slov (IS) a kratších frází (R) jako celků. Např. databáze OuluVS [Zhao 2009] obsahuje 10 každodenních frází v anglickém jazyce, např. „How are you?“ či „Nice to meet you.“.

Kromě izolovaných jednotek výzkum klade velký důraz na rozpoznávání spojitých promluv. V oblasti pouze vizuálního odezírání ze rtů je častou úlohou rozpoznávání spojitých číslovek (SČ), např. telefonních čísel s pevnou délkou. Mezi typické a obvykle využívané zástupce patří databáze XM2VTS a CUAVE. První uvedená obsahuje nahrávky od rozsáhlého vzorku 295 mluvčích, bohužel je však dostupná pouze za úplatu (cca 1000£). O něco málo složitější úlohu pak představuje rozpoznávání strukturovaných vět (SV). Tento model byl uveden na databázi GRID [Cooke 2006], kde každá věta je tvořena sekvencí „příkaz(4)-barva(4)-předložka(4)-písmeno(25)-číslovka(10)-příslovce(4)“ (v závorce je uveden počet možností odpovídající kategorie) v angličtině s celkovým slovníkem o 51 slovech. Příklad takové věty je „Lay blue at C 7 please“.

Korpusy s malým slovníkem (písmena, číslovky, slova, fráze, ...) jsou populární z několika důvodů. Jednou z hlavních výhod je poměrně dobrá řešitelnost i za použití pouze vizuální parametrizace, tedy bez akustické informace, což umožňuje jednoduché a přímé porovnání. Pro rozpoznávání obvykle dostačují celoslovní HMM, a tak není nutná manuální anotace jednotlivých fonémů či vizémů pro natrénování hláskových modelů. Malé korpusy také nevyžadují užití jazykových modelů, jenž zanáší do výsledků další variabilitu. V neposlední řadě je manipulace s daty a průběh experimentů efektivní. Na druhou stranu je ovšem diskutabilní, do jaké míry jsou dosažené výsledky zobecnitelné na spontánní řeč s libovolně velkým slovníkem.

Databází, jež by obsahovaly spojitou řeč s neomezeným slovníkem od velkého množství řečníků, se za celou dobu výzkumu příliš mnoho neobjevilo. Zdaleka nejrozsáhlejším počinem se stala databáze IBM ViaVoice vytvořená v IBM Thomas J. Watson Research Center a následně představená na Johns Hopkins University summer 2000 workshopu. Databáze obsahuje přes 24000 nahrávek diktovaných promluv od 290 mluvčích s celkovým slovníkem 10403 slov. V IBM laboratořích pak vzniklo i několik dalších databází obsahující spojitou řeč, avšak žádná z nich není veřejně dostupná a to ani za poplatek. Ze zdarma dostupných databází je pravděpodobně nejrozsáhlejší AVICAR, obsahující necelých 7000 nahrávek od 86 mluvčích.

Z hlediska charakteru zdrojových dat je nejčastější případ jednoho čelního pohledu na řečníka pořízeného videokamerou se standardním rozlišením (např. PAL 720×576 px). Přestože příliš neodpovídá reálným podmínkám, přesto je tento způsob volen velmi často, zejména kvůli jednoduchosti fáze detekce ROI. Objevily se ale i databáze zaměřené na zkoumání závislosti rozpoznávání na kamerovém pohledu. Mezi ně patří např. IBM Smart Room databáze, v níž je každý dobrovolník nahraný dvěma různými kamerami z anfasu a z profilu. Databáze LiLiR (Language Independent Lip Reading) obsahuje nahrávky natočené dvěma HD a třemi SD kamerami zároveň rozmístěnými rovnoměrně od 0° do 90° kolem řečníka. Podobným způsobem byla pořízena i francouzská databáze LTS5, avšak s použitím pouze dvou



Obrázek 6.1: Ukázky vybraných audiovizuálních databází.

kamer a jejich postupným natáčením. Zřejmě největší databází zachycující více pohledů je AVICAR, nasnímaná v automobile s použitím 4 kamer a 7 mikrofonů. Částečně se problémem relativního pohybu řečníka a kamery zabývá i databáze CUAVE, kde část promluv je nahrána za pohybu hlavy a následně i z profilu. Opačným způsobem se vydali autoři databáze IBM Infra Headset, ve které měli mluvčí k dispozici headset s vestavěnou infračervenou kamerou vedle mikrofonu. Díky tomuto speciálnímu zařízení tak odpadla fáze detekce ROI a významně se snížila závislost na okolním osvětlení.

Některé databáze jsou navrženy pro využití informace z třetího rozměru, např. pomocí metod stereovidění. Příkladem je např. australská databáze AVOZES, jež obsahuje nahrávky fonémů (F), vizémů (V) a číslovek (Č), vždy zahrnutých v nějakých větách. Dalším příkladem je databáze WAPUSK20, jež modelem nahrávání kopíruje strukturované věty GRID. Několik autorů také vytvořilo databáze BAVCD a KinectVS, které již obsahují hloubková data vypočtená zařízením MS Kinect. Ukázkové snímky z vybraných databází jsou zachyceny na obr. 6.1.

Název	Typ promluv	|slovník|	# subj.	# promluv	Typ dat	Jazyk	Dostupnost
M2VTS [Pigeon 1997]	IČ	10	37	185	čelní	Franc.	Ne
AVletters [Matthews 2002]	IP	26	10	780	čelní ROI	Angl.	Zdarma
AVletters2 [Cox 2008]	IP	26	5	910	čelní HD	Angl.	Zdarma
OuluVS [Zhao 2009]	R	10	20	1000	čelní	Angl.	Zdarma
CUAVE [Patterson 2002]	IČ, SČ	10	36	~7000	čelní, profil, pohyb	Angl.	Zdarma
XM2VTS [Messer 1999]	SČ	10	295	7080	čelní	Angl.	Placená
IBMSR [Lucey 2006b]	SČ	11	38	1661	čelní, 2×profil, 2 mikrofony	Angl.	Ne
IBMIH [Huang 2004]	SČ, SŘ	11/?	79/113	4011/12186	čelní ROI	Angl.	Ne
GRID [Cooke 2006]	SV	51	34	34000	čelní	Angl.	Zdarma
VidTIMIT [Sanderson 2002]	SŘ	80	43	430	čelní	Angl.	Zdarma
AV-TIMIT [Hazen 2004]	SŘ	1973	223	4460	čelní	Angl.	Ne
IBM ViaVoice [Neti 2000]	SŘ	10403	290	24325	čelní	Angl.	Ne
UWB-05-HSCAVC [Císař 2006]	SŘ	?	100	2000	čelní	Češ.	Ne
AVDB2cz [Chaloupka 2005]	IS, SŘ	50/?	35	3500	čelní	Češ.	Ne
AVICAR [Lee 2004]	IP, IČ, SČ, SŘ	26/13/1317	86	~6800	4 čelní, 7 mikrofonů	Angl.	Zdarma
LiLiR [Lan 2010]	SŘ	~1000	20	4000	čelní, 30°, 45°, 60°, profil	Angl.	Ne
LTS5 [Estellers 2012b]	SČ	10	20	~200	čelní, 30°, 60°, profil	Franc.	Zdarma
AVOZES [Goecke 2004]	F/V, Č, SŘ	44/11/10/23	20	1160	stereo	Angl.	Placená
WAPUSK20 [Vorwerk 2010]	SV	52	20	2000	stereo, disparita	Angl.	Zdarma
BAVCD [Galatas 2011]	IČ, SČ	11	15	1200	čelní, hloubka	Angl., Řeč.	Ne
KinectVS [Pei 2013]	R	20	20	2400	čelní, hloubka	?	Ne
MIRACL-VC [Rekik 2014b]	IS, R	10/10	15	1500/1500	čelní, hloubka	?	Ne

Tabulka 6.1: Přehled vybraných audiovizuálních databází.

7. Shrnutí výsledků současného stavu poznání

Rozsáhlejší srovnávací práce, jaké jsou běžné v jiných oborech, se v literatuře audiovizuálního rozpoznávání řeči téměř nevyskytují. Výjimky reprezentují např. práce [Matthews 2002, Lan 2009] porovnávající různé druhy vizuálních příznaků. Především roztříštěnost a neexistence rozsáhlejších audiovizuálních databází způsobují, že je velmi obtížné objektivně kvantifikovat úspěšnost různých algoritmů současného stavu poznání. Volba testovací databáze bývá obvykle svévolná a spíše než pro objektivní porovnání se odvíjí od dostupnosti a množství práce, jež pro zpracování vyžaduje (např. detekce ROI).

I v případech, kdy autoři volí testovací databázi stejnou jako v souvisejících článcích od jiných výzkumníků, často použijí jiný protokol testování a přímé porovnání navržených algoritmů tak není možné. Typicky se liší např. velikost trénovacích a testovacích množin či závislost na mluvčích. Především pro rozpoznávání izolovaných jednotek neexistuje žádný standardní postup, a tak autoři často volí jednotlivé komponenty různě. Např. při uvedení nového způsobu integrace akustických a vizuálních příznaků Ngiam a kol. [Ngiam 2011] namísto HMM zvolili klasifikaci pomocí SVM a systém je tak možné porovnávat pouze jako celek, nikoliv přínos jeho jednotlivých částí. Na druhou stranu předpoklad vzájemné nezávislosti dílčích komponent nemusí být realistický, což může působit jako další faktor ve variabilitě dosažených výsledků. Výrazný vliv na výslednou úspěšnost má rovněž algoritmus detekce ROI. Např. v článku [Zhou 2011] použitím manuální anotace ROI klesla slovní chybovost (viz dále) o 76,5 %, což je mnohonásobně více než činily rozdíly mezi porovnávanými metodami. Tyto a další faktory ztěžují orientaci ve stavu poznání a je nutné je při hodnocení experimentů mít na paměti.

Obvyklým měřítkem úspěšnosti je **slovní přesnost** (angl. word accuracy, WAcc)

$$\mathrm{WAcc} = \frac{N - D - S - I}{N}, \tag{7.1}$$

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. Jelikož počet inzercí je teoreticky neomezený, může WAcc nabývat i záporných hodnot. V některých pracích bývá uvedena **slovní chybovost** (Word Error Rate, WER), pro níž platí $\mathrm{WER} = 1 - \mathrm{WAcc}$. Jako kritérium se poměrně často užívá relativní změna chybovosti

$$\delta_{\mathrm{WER}} = \frac{(\mathrm{WER2} - \mathrm{WER1})}{\mathrm{WER1}}, \tag{7.2}$$

např. při porovnávání navržené metody např. vůči stavu poznání nebo posuzování přínosu vizuální složky oproti akustické. Analogicky k WAcc jsou pak definovány úspěšnost, resp. chybovost jiných rozpoznávaných jednotek, např. fonémů či vizémů.

V literatuře se objevují tři základní způsoby testování navržených systémů v závislosti na rozdělení trénovací a testovací databáze vzhledem k mluvčím:

1. závislé na řečníkovi (Speaker Dependent, SD),

2. více-uživatelské (Multi-Speaker, MS),

3. nezávislé na řečníkovi (Speaker Independent, SI).

Systémy závislé na řečníkovi předpokládají pouze jednoho koncového uživatele. Tomuto předpokladu může být uzpůsoben i návrh příznaků, protože se zde nevyskytuje vizuální variabilita související se změnou řečníka. V cílové aplikaci tyto systémy před užitím vyžadují kalibrační fázi, kdy uživatel poskytne trénovací data např. v podobě několika nahrávek.

Více-uživatelské systémy pracují se stejnou množinou mluvčích v trénovací i testovací fázi. Na rozdíl od SD systémů se zde projevuje uživatelská vizuální variabilita, avšak pouze v omezené míře, jelikož navržená parametrizace nemusí zobecňovat na neznámé případy.

Nejsložitější případ představuje návrh **systémů nezávislých na řečníkovi**. Nemusí zde existovat žádný předpoklad ohledně množiny koncových uživatelů a uživatelská vizuální variabilita se tak zde projevuje v plné míře. Navržená parametrizace musí být dostatečně robustní a systém by měl mít schopnost zobecňovat na neznámé případy. Z aplikačního hlediska tyto systémy obvykle nevyžadují žádnou kooperaci koncového uživatele.

7.1 Vizuální rozpoznávání

Výzkum odezírání ze rtů, tedy pouze vizuálního rozpoznávání, se soustředí především na jednodušší úlohy s malým počtem slovníkových jednotek, typicky v rozmezí 10–50 položek. Nejzajímavější práce z posledních let a jejich dosažené výsledky shrnuje tabulka 7.1, kde **|S|** značí velikost slovníku **Z** závislost systému na mluvčích – závislý (SD), víceuživatelský (MS) a nezávislý (SI). Pro SD a MS je uveden počet mluvčích a pokud známo také poměr trénovacích a testovacích dat. Zkratka CV značí křížovou validaci (Cross Validation, CV). Pro SI systémy je uveden poměr trénovacích a testovacích mluvčích, přičemž u prací využívajících pro odladění parametrů validační sadu je uvedena její velikost jako prostřední hodnota. Ve sloupci **Pohled** značí č čelní pohled, p profil a h hloubkový obraz. Šipka označuje transformaci příznaků, např. z profilu do čelního pohledu: p →č. Výsledky nadepsané † byly dosaženy manuální anotací ROI, tedy bez použití automatických algoritmů.

V posledních několika letech se výzkum odezírání ze rtů zaměřil na rozpoznávání izolovaných jednotek – nejčastěji kratších frází – viz např. [Zhao 2009, Ong 2011, Zhou 2010, Zhou 2011, Pei 2013, Zhou 2014]. Objevují se především holistické metody narušující tradiční schéma oddělené extrakce příznaků a klasifikace a obě fáze tak do značné míry splývají. Populárními se staly především metody založené na grafovém vnořování a učení nadplochy. Nejlepších výsledků v současné době dosahuje algoritmus [Pei 2013] založený promítání promluv do nelineárního prostoru algoritmem MDS s párovými vzdálenostmi vypočtenými pomocí náhodných lesů. Značnou nevýhodou holistických metod je však provázanost s cílovou aplikací a způsob jejich aplikace pro rozpoznávání spojité řeči není zřejmý.

Úloha	Databáze	|S|	Z	Reference	# subj.	Pohled	Skóre [%]
IČ	CUAVE	10	SD	[Pei 2013]	2	č	100,00
				[Rekik 2014b]	36		90
			SI	[Papandreou 2009]	30:6 (6×CV)		83,00
				[Lucey 2006b]	28:8		77,08
				[Rekik 2014b]	36 (?:?)		70,10
				[Ngiam 2011]	18:18		68,70
				[Saenko 2006]	22:6:6		43,30
				[Lucey 2008]	25:8 (10×CV)	p→č	38,80
IP	AVletters	26	SD	[Pei 2013]	10	č	69,60
			MS	[Ngiam 2011]	10 (2:1)		64,40
				[Zhao 2009]			58,85
				[Matthews 2002]			44,60
			SI	[Zhao 2009]	9:1 CV		43,46
	AVletters2	26	SD	[Pei 2013]	5	č	91,80
				[Cox 2008]	5 (6:1 CV)		~87
			MS		5 (6:1 CV)		~85
			SI		4:1 CV		~8
R	OuluVS	10	SD	[Pei 2013]	20	č	97,30
				[Zhou 2011]	20 (4:1 CV)		96, 50†, 85,10
				[Rekik 2014b]	?		93,20
				[Zhou 2010]	20 (4:1 CV)		90, 60†
				[Ong 2011]			86,50
				[Zhao 2009]			70,20
			SI	[Pei 2013]	19:1 CV		89,70
				[Zhou 2014]			85, 60†, 76,60
				[Zhou 2011]			81,30
				[Rekik 2014b]			68,30
				[Ong 2011]			65,60
				[Zhao 2009]			62,40
	KinectVS	20	SD	[Pei 2013]	20	č, h	94,10
			SI		19:1 CV		87,70
SČ	LTS5	10	MS	[Estellers 2012b]	20 (2:1 CV)	č	71,10
						30°→č	67,70
						60°→č	64,30
						p→č	58,30
	IBMIH	11	SI	[Huang 2013]	70:37	č	64,8
	AVICAR	13	SI	[Fu 2008]	21:13	č	37,87
	BAVCD	11	SI	[Galatas 2012]	10:5 CV	č	43,60
						č, h	44,39
SV	GRID	51	SI	[Lan 2009]	14:1 CV	č	65,00

Tabulka 7.1: Srovnání posledních výsledků odezírání ze rtů.

Tradičním postupem pak je oddělená extrakce vizuálních příznaků a jejich následná klasifikace. Zde se výzkum zaměřil především na nalezení optimální vizuální parametrizace. Patrná je v současném výzkumu snaha využít dynamiku mluvené řeči, viz např. [Pachoud 2008, Zhao 2009, Ngiam 2011]. Největší problémem však stále zůstává závislost parametrizace na mluvčích [Lan 2009, Lan 2010]. Část výzkumu se rovněž zaměřila robustnost vůči pohybům a kamerovém pohledu na řečníka [Lucey 2006a, Lucey 2007, Lucey 2008, Estellers 2012b]. Spíše okrajově se v literatuře objevují práce využívající více pohledů či speciálních zařízení pro rekonstrukci třetího rozměru, viz např. [Galatas 2012, Rekik 2014b]. I přes uvedení nových typů vizuálních parametrizace se v článcích zabývajících se jinou dílčí problematikou rozpoznávání stále téměř vždy používají spíše tradiční příznaky. Jedním z důvodů je zřejmě absence srovnávacích experimentů, a tak se současný stav poznání rozvíjí jen pozvolna. Mezi klasifikátory jednoznačně převažuje lineární HMM, byť byly navrženy i jiné metody [Pachoud 2008, Zhao 2009, Ngiam 2011]. Jejich použití však bývá svázané se zvolenou parametrizací a tyto alternativní metody tak zůstávají spíše na okraji zájmu.

Výzkum se téměř nezabývá vlivem fáze detekce ROI. Přestože za posledních 10 let došlo k výraznému posunu v oblasti algoritmů pro zarovnání obličeje, viz např. [Saragih 2011, Cao 2012, Xiong 2013, Kazemi 2014], v literatuře odezírání ze rtů se tento vývoj prakticky neprojevil. Autoři volí buď tradiční metody jako AAM [Papandreou 2009, Pei 2013] a VJ [Lucey 2006b], nebo nějaký vlastní postup navržený ad hoc [Zhou 2011]. Jak již bylo řečeno v předchozí sekci, robustní detekce ROI má přitom na výsledek mnohdy daleko větší vliv než přínos navrhované metody, např. pro extrakci příznaků. Pro příklad viz rozdíly u [Zhou 2014] na databázi OuluVS. Vliv vizuálního předzpracování se projevuje i při porovnání mezi jednotlivými pracemi, kdy nejlepších výsledků je obvykle dosaženo v článcích, kde autoři využili sofistikovanější metodu zarovnání obličeje. Viz např. [Lan 2009, Papandreou 2009, Pei 2013], kde autoři aplikovali AAM, [Rekik 2014b], kde Rekik a kol. detekovali polohu a natočení obličeje v prostoru vlastní optimalizační metodou založenou na RGB-D datech [Rekik 2014a], či [Estellers 2012b], kde byly klíčové body vyznačeny přímo na obličejích řečníků reflexivní barvou. Z výsledků tedy není zcela zřejmé, do jaké míry výzkum srovnává metody parametrizace a klasifikace pro rozpoznávání řeči namísto zarovnání obličeje a detekce zájmové oblasti.

7.2 Audiovizuální rozpoznávání

Výsledky vybraných prací audiovizuálního rozpoznávání shrnuje tabulka 7.2. Sloupec **Ú** značí úlohu, **E** typ experimentů (závislost na mluvčích) a **P** pohled. V závorce za SNR je uveden typ hluku: bílý (angl. white, w) nebo babble. Výsledek nadepsaný $^{+}$ byl dosažen na řečově zvýrazněných nahrávkách. Horní index * značí použití optimálních vah, tedy se znalostí skutečného SNR či jiné metriky sloužící jako základ pro nastavování vah. V práci [Kratt 2004] byla pro vizuální kanál

použita menší trénovací databáze.

Oblast audiovizuálního rozpoznávání řeči je oproti vizuálnímu odezírání ze rtů méně aktivní. Výzkum se ubírá především směrem k dynamickému modelování spolehlivosti akustického a vizuálního kanálu, viz např. [Kolossa 2009, Estellers 2012a, Stewart 2014]. Tradičním způsobem je přitom odhad nějaké veličiny, např. SNR pro akustický kanál, či disperze pro modálně nezávislý odhad, a následná transformace na optimální váhy vícekanálového systému. Z teoretického pohledu je zajímavou prací článek [Papandreou 2009], kde je odhad nejistoty zanesen přímo do parametrizace a je s jeho pomocí možné odvodit pravidlo váženého součinu. Z praktického hlediska však nepřináší významný posun oproti tradičnímu MSHMM. Nejčastěji jsou akustický a vizuální kanál kombinovány pomocí hybridní fúze buď v HMM nebo DBN. Objevují se však i práce zabývající se brzkou integrací – dobrých výsledků dosáhli za použití hlubokých neuronových sítí autoři článků [Ngiam 2011, Huang 2013].

Přestože jedním z hlavních účelů vizuální informace je její doplňkové využití pro rozpoznávání spojité/spontánní řeči s velkým slovníkem, výzkum tohoto problému již dnes prakticky neprobíhá. Stav je zapříčiněn především nedostatkem rozsáhlých audiovizuálních databází, viz kapitolu 6. Nejvíce prací vzniklo v laboratořích IBM v letech 2000–2004 s využitím proprietární databáze ViaVoice [Glotin 2001, Potamianos 2001a, Potamianos 2003]. Významná část algoritmů, dodnes považovaných za stav poznání (např. HiLDA), přitom byla navržena na John Hopkins University summer workshopu 2000 [Neti 2000]. V českém prostředí se problematikou rozpoznávání spojité řeči zabýval Petr Císař ve své dizertační práci [Císař 2006].

Ú	Databáze	|S|	E	Reference	# subj.	P	SNR	Skóre [%]	
								A	AV
IČ	XM2VTS	10	SI	[Stewart 2014]	200:95	č	30dB	~98	~99
							0dB (w)	~44	~95
	CUAVE	10	SI	[Estellers 2012a]	30:3:3 (6×CV)		čistý	~99	~99
							0dB (b)	~67	88
				[Ngiam 2011]	18:18		čistý	95,8	94,4
							0dB (w)	75,8	82,2
				[Papandreou 2009]	30:6 (6×CV)		čistý	~95	~98
							5dB (b)	~58^{+}	~79, ~87*
				[Saenko 2006]	26:6:6		12dB (b)	92,3	97
							4dB (b)	67,7	80,7
SČ	LTS5	10	MS	[Estellers 2012b]	20 (2:1 CV)	č	7dB (b)	66,33	80,83
							0dB (b)	36,57	73,37
						30° →č	7dB(b)	64,2	77,0
							0dB(b)	36,5	68,3
	BAVCD	11	SI	[Galatas 2012]	14 (2:1 CV)	č, h	10dB (b)	~88	~89
							0dB (b)	~55	~48
	IBMIH	11	SI	[Huang 2013]	70:37	č	20dB	98,3	98,7
							7dB (b)	76,8	89,4
				[Marcheret 2007]	71:8		čistý	~99	~99
							5dB (b)	~26	~69
SV	GRID	51	SD	[Kolossa 2009]	2	č	čistý	~99	~99*
							10dB	~82	~89*
			SI	[Shao 2008]	34		čistý	~93	~91
							5dB (?)	~61	~74
SŘ	UWB-05-HSCAVC	344	MS	[Císař 2006]	100	č	čistý	81,47	84,86
							0dB (b)	60,79	74,31
	IBM ViaVoice	10403	SI	[Kratt 2004]	A 261:26 V 120:26	č	čistý	74,44	75,90
							zašuměný	46,06	52,63
				[Potamianos 2003]	239:25:26		čistý	~89	~89
							8.5dB (b)	~62	~79
				[Neti 2000], [Glotin 2001]			čistý	85,56	86,53
							8.5dB (b)	51,90	64,73

Tabulka 7.2: Srovnání vybraných výsledků audiovizuálního rozpoznávání.

8. Návrh vizuální parametrizace řeči

V práci je navrženo několik různých typů 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.

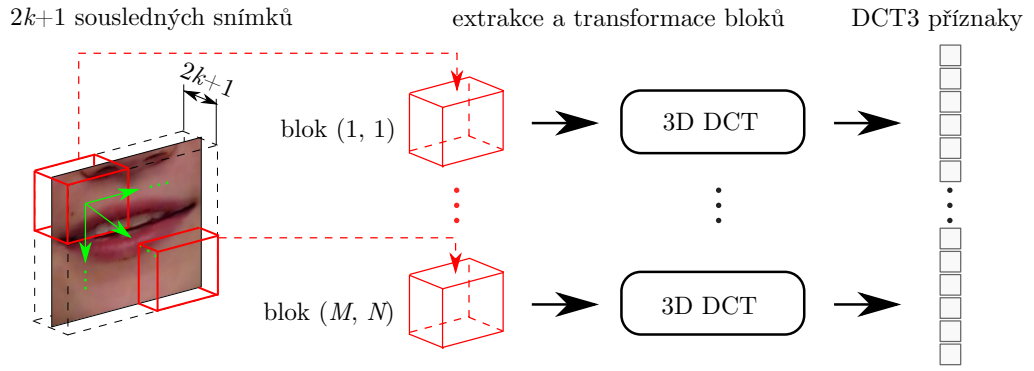
8.1 Trojrozměrná bloková DCT

Jednou z parametrizací navržených s cílem využít dynamiku řeči je trojrozměrná bloková diskrétní kosinová transformace. Extrakce příznaků probíhá následujícím způsobem. Na video je podobně jako např. v pracích [Pachoud 2008, Zhao 2009] nahlíženo jako na trojrozměrné útvary s dvěma obrazovými a jednou časovou osou. Každý snímek videa je zpracováván společně s k předchozími a k následujícími, tedy po sekvencích o délce $2k + 1$. Každý takto utvořený „kvádr“ je předmětem trojrozměrné redukce pomocí trojrozměrné DCT. Za účelem lepšího zachycení lokálních změn je podobně jako v práci [Lucey 2006b] obrazová rovina rozdělena na překrývající se bloky, které jsou parametrizovány odděleně. Výhodou toho přístupu je snížení závislosti příznaků na řečníkovi, protože jednotlivé podoblasti obsahují menší míru variability než kompletní ROI.

Celý postup je ilustrován na obr. 8.1. 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 10.

8.2 Histogram orientovaných gradientů s dynamikou

Deskriptory založené histogramech orientací gradientů představují v oblasti odezírání ze rtů poměrně oblíbenou a často užívanou parametrizaci, viz např. [Pachoud 2008, Pei 2013, Rekik 2014b, Savchenko 2014]. Jak bylo zmíněno v sekci 3.1.3, původně byla tato myšlenka aplikována pro rozpoznávání objektů na základě lokálních deskriptorů SIFT okolo zájmových bodů v průkopnické práci Davida Lowa [Lowe 2004] (rozšířená verze původního článku z roku 1999) a dále rozvedena článkem Dalala a Triggse [Dalal 2005] pro detekci lidské postavy pomocí histogramů orientovaných gradientů (Histogram of Oriented Gradients, HOG). Obraz je rozdělen na nepřekrývající se políčka o malém rozměru, typicky 8×8 pixelů. Pro

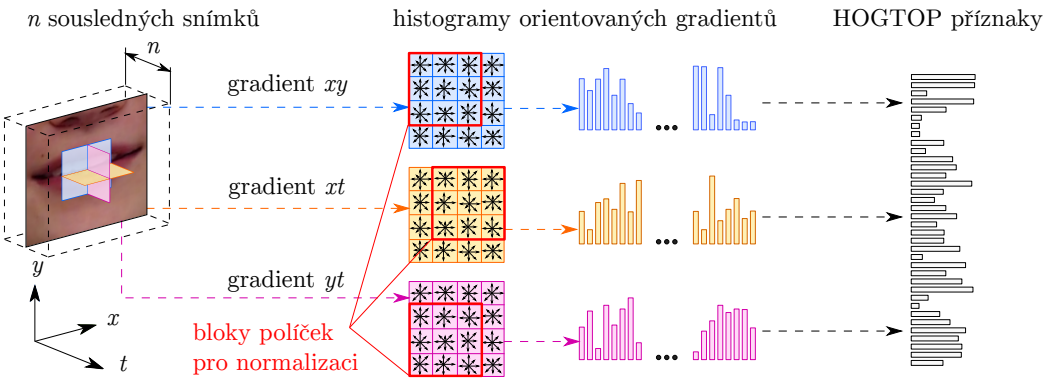


Obrázek 8.1: Princip extrakce DCT3 příznaků.

každý pixel každého políčka je vypočten lokální gradient (g_x, g_y), např. symetrickou diferencí či Sobelovým filtrem. Z orientací $\phi = \arctan(g_y, g_x)$ je sestaven histogram o b binech (v této práci $b = 8$), přičemž každý příspěvek je vážen magnitudou $A = \sqrt{g_x^2 + g_y^2}$. Sousední políčka jsou seskupeny do bloků např. po 3×3 a jejich histogramy vektorově spojeny za sebe. Výsledný vektor $\boldsymbol{v}$ je pro zvýšení robustnosti vůči změnám osvětlení normalizován multiplikativním faktorem $(\|\boldsymbol{v}\|_k^k + \varepsilon^k)^{-1/q}$, kde typicky $k = q = 2$ a $\varepsilon^k \ll \|\boldsymbol{v}\|_k^k$ zamezuje dělení nulou pro oblasti bez textury. Výsledný deskriptor celého obrazu je sestaven ze všech bloků vektorovým pospojováním. Všimněme si, že díky seskupování políček do bloků se jednotlivé dílčí histogramy ve výsledném popisu vyskytují několikrát (pro případ 3×3 až $9\times$, vždy s normalizací z jiného bloku), což má za následek vysokou dimenzi parametrizačního vektoru. Např. pro obrázek 64×64 pixelů, 8 binů v histogramu a 3×3 políček na blok činí rozměr vektoru 2592, bez seskupování do bloků pouze 512.

Standardní HOG jsou extrahovány pro každý snímek zvlášť a nijak tedy nezohledňují dynamiku řeči, která je přitom pro úspěšnost rozpoznávání kritická. Jak využít příznaky SIFT společně s řečovou dynamikou navrhli Pachoud a kol. v práci [Pachoud 2008]. Na vstupní videosekvenci nahlíželi jako kvádry a na různých podoblastech extrahovali trojrozměrné SIFT deskriptory, jež zobecňují standardní extrakci přidáním temporální složky gradientů. Jinou dynamickou modifikaci navrženou původními autory deskriptoru HOG Dalalem a Triggsem [Dalal 2006] využili pro automatické odezírání ze rtů Rekik a kol. [Rekik 2014b]. Jedná se o standardní deskriptor HOG, pouze vypočtený na x-ové a y-ové složce obrazu optického toku odděleně a poté spojen do jediného vektoru.

V této práci je navržena vlastní dynamizace deskriptoru HOG. Modifikace je inspirována příznaky LBPTOP [Zhao 2009], kdy jsou pro extrakci příznaků kromě obrazové roviny xy využity i roviny xt a yt. 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. Ty jsou dále zpracovány standardním způsobem, tedy sestavením histogramu orientací, normalizací přes bloky a následným pospojováním do parametrizačního vektoru. Pro zachycení



Obrázek 8.2: Princip extrakce HOGTOP příznaků.

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 tvořícího výsledný popis. Příznaky jsou označeny jako **HOGTOP** (Histogram of Oriented Gradients from Three Orthogonal Planes) a postup jejich extrakce ilustruje obr. 8.2. Jelikož deskriptory HOG mívají velmi vysokou dimenzi, která je v případě HOGTOP navíc trojnásobná, jsou příznaky podobně jako u DCT3 dále redukovány a zároveň dekorelovány metodou PCA na rozměr několika desítek koeficientů. Přesná hodnota je stanovena křížovou validací s cílem maximalizovat slovní přesnost, viz sekci 9.2.

8.3 Integrace hloubkových příznaků

Tato dizertační práce se rovněž zabývá využitím trojrozměrné informace pro odezírání ze rtů. Přibližně v roce 2010 se na trhu postupně začaly objevovat levná a dostupná zařízení schopná v reálném čase společně s RGB obrazem poskytovat i hloubkovou mapu a to se shodným či alespoň blízkým rozlišením a snímkovací frekvencí. V praxi to znamená, že pro každý RGB pixel každého snímku je dostupný ještě údaj o *vzdálenosti* od kamery a při znalosti kalibrační matice je tedy možné zpětně rekonstruovat 3D pozici každého viditelného bodu v prostoru. Nejznámější z těchto zařízení je bezesporu Microsoft Kinect, jehož technické specifikace a ukázky hloubkových map jsou uvedeny v kapitole 9 a který byl využitý i v této práci. Mezi další pak patří Asus Xtion, Creative Senz3D, SoftKinetic DS325, či PMD Camboard Pico.

V práci jsou prozkoumány tři základní způsoby integrace hloubkových dat. V principu je možné na hloubkovou mapu nahlížet stejným způsobem jako na RGB obraz a příznaky extrahovat analogicky, tedy např. PCA transformací ROI apod.

Takto extrahovanou parametrizaci je poté možné integrovat s video parametrizací jednoduchým vektorovým spojením, tedy pomocí tzv. **brzké integrace**, viz sekci 5.1. V případě skrytých markovských modelů je rovněž možné zohlednit relativní přínos jednotlivých kanálů díky **vícekanálovému synchronnímu HMM** 5.10, více v sekci 5.3.

Třetím 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. Je tedy sestaven **sdružený příznakový model** a teprve na něm je natrénován příslušný klasifikátor. Jednoduchou metodou z této skupiny je rozšířit 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

$$\boldsymbol{a} = \begin{bmatrix} w_s \boldsymbol{p} \\ \boldsymbol{\lambda} \\ w_d \boldsymbol{\gamma} \end{bmatrix} = \boldsymbol{\Phi} \boldsymbol{d}, \tag{8.1}$$

kde $\boldsymbol{\gamma}$ je PCA redukce „textury" extrahované z hloubkové mapy analogicky k (2.7) a multiplikativní konstanty w_s a w_d podobně jako u (2.8) 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. Jelikož Kinect poskytuje obrazová a hloubková data vzájemně zarovnané, samotné vyhledávání klíčových bodů může probíhat pouze ve video kanálu bez využití informace o hloubce.

9. Příprava dat a návrh testovacího protokolu

9.1 Audiovizuální databáze TULAVD

V rámci této práce jsem navrhl a vytvořil audiovizuální databázi TULAVD, která umožňuje testování navržených algoritmů a vizuálních řečových příznaků. Databáze obsahuje promluvy od celkem 54 mluvčích, z toho 23 žen a 31 mužů. Věk mluvčích se pohyboval v rozmezí 20–70 let s mediánem 25.5 roku.

9.1.1 Použitá zařízení

K nahrávání obrazové složky bylo použito více zdrojů, konkrétně 2 webkamery Logitech C920, 2 senzory Microsoft Kinect v první verzi a klopový mikrofon.

Microsoft Kinect je zařízení uvedené na trh koncem roku 2010 jako bezdotykové ovládání herní konzole Microsoft XBox. Je založeno na referenčním designu společnosti PrimeSense[1], viz obr. A.1, a je schopné snímat barevný obraz zároveň s hloubkou, obojí v rozlišení 640×480 pixelů při 30 snímcích za sekundu. Zároveň dokáže zaznamenávat zvuk 4 zabudovanými mikrofony v 16bitové kvalitě při vzorkovací frekvenci 16 kHz. Jelikož nejsou data při přenosu po sběrnici USB komprimována, vzniká v této konfiguraci velký datový tok. Obraz tedy není plně RGB, barvy jsou kódovány bayerovským uspořádáním RGGB, což znamená, že barevná složka má efektivně pouze poloviční rozlišení oproti jasové. Snímání hloubky je založeno na technice strukturovaného světla v infračervené oblasti. Kinect obsahuje emitor infračervených paprsků, které dopadají na scénu a vytvářejí nepravidelné obrazce, které jsou na povrchu objektů deformovány. Infračervená kamera tyto deformace zachytí a na základě pozice jednotlivých obrazů paprsků trianguluje pozici bodů v prostoru. Výsledné rozlišení je 11 bitů, tedy 2048 různých hodnot. Přesnost výpočtu je závislá na vzdálenosti objektu, přičemž minimum je podle specifikace 80 cm. Zařízení však dokáže snímat i nižší vzdálenosti až do 40

[1] `http://www.souvr.com/Soft/UploadSoft/201005/2010050617295050.pdf`

Obrázek 9.1: Ukázka obrazu a hloubkové mapy získané z Microsoft Kinect.

Obrázek 9.2: Microsoft Kinect a dvě kamery Logitech C920 použité k nahrávání.

cm, avšak s úměrně sníženou přesností či vyšším procentem nerozhodnutých bodů. K samotnému získání dat z Kinectu byla použita knihovna libfreenect[2] a operační systém Linux, protože ovladače dodávané firmou Microsoft pro první verzi snímače neumožňují dostatečnou konfiguraci a zároveň nedovolují snímat ze vzdálenosti nižší než 80 cm. Ukázka obrazu a příslušné hloubkové mapy z nahrávání databáze zařízením Kinect je k vidění na obr. 9.1. Ani přes použití techniky strukturovaného světla však není úloha výpočtu disparity vždy jednoznačně řešitelná. Senzor Kinect v takovémto případě vrací na odpovídající pozice v obraze nulovou hodnotu, což významným způsobem ovlivňuje výslednou parametrizaci. V této práci jsou proto chybějící hodnoty v hloubkové mapě ještě před výpočtem vizuálních příznaků odhadnuty podle nejbližšího okolí bilineární interpolací.

Kromě MS Kinect byly k nahrávání promluv použité dvě webkamery Logitech C920. C920 snímá plně RGB video v rozlišení FullHD, tj. 1920×1080 pixelů při 30 snímcích za sekundu. Obsahuje také dva mikrofony nahrávající audio s rozlišením 16 bitů a vzorkovací frekvencí 44 kHz. Kamery byly upevněny horizontálně vedle sebe ve vzdálenosti cca 12 cm tak, aby umožňovaly rekonstrukci hloubkové mapy pomocí stereovidění. Kalibrace stereo konfigurace byla provedena knihovnou OpenCV[3]. Nedostatkem webkamery C920 se ukázal být periodický hluk způsobený optikou snímače a jejím blízkým umístěním vedle jednoho z integrovaných mikrofonů. Zvukové nahrávky tak nejsou čisté, což komplikuje použití v experimentech audiovizuálního rozpoznávání se simulovaným hlučným prostředím. Obě kamery a Kinect tak, jak byly použité při nahrávání, jsou zachyceny na obr. 9.2.

9.1.2 Metodika nahrávání

Databáze byla nahrávána v běžných kancelářských prostorech během pracovní doby, místnost nebyla nijak odhlučněna. Na kvalitě výsledných nahrávek se však hluk na pozadí nijak zásadně neprojevil, mnohem silnější byl např. šum způsobený

[2] `http://openkinect.org`
[3] `http://opencv.org`

nekvalitní optikou na kameře C920, který byl zmíněn v kapitole 9.1.1. Světelné podmínky byly přibližně konstantní. Nahrávání probíhalo u obyčejného PC s instruktorem. Mluvčí byli usazeni ve vzdálenosti cca 80 cm od snímacích zařízení. Jak nahrávání probíhalo je zachyceno na obr. A.2 v příloze.

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 s využitím více kamer či hloubkové mapy ze senzoru Kinect.

Druhá část obsahuje od každého mluvčího 50 izolovaných slov. Těchto 50 slov 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 vět je rozděleno na dvě skupiny. Prvních 50 vět je společných pro všechny mluvčí, druhých 50 vět je pro každého mluvčího jedinečných. Výběr vět je popsán v sekci 9.1.3.

Pro snadnou spolupráci s knihovnami třetích stran jsou všechna videa uložena ve formátu Motion JPEG v rozlišení, v jakém byla získána ze zařízení. Nevýhodou tohoto algoritmu je nízký kompresní poměr, což v kombinaci s vysokým rozlišením znamená, že celková velikost databáze překračuje 1000 GB.

9.1.3 Textový korpus

Pro účely nahrávání bylo pro každého mluvčího sestaveno 50 vět, pro 54 mluvčích tedy celkem 2750 vět (50 bylo společných pro všechny). Každá skupina 50 vět byla vybírána z korpusu čítajícího přes 67000 vět a získaného z online zdrojů (např. zpravodajské servery). Výběr probíhal na základě relativní četnosti jednotlivých fonémů tak, aby odpovídaly rozložení v českém jazyce, přičemž fonetický přepis byl odhadnut automaticky pomocí softwaru NanoDictateT vyvíjeného na Ústavu Informačních Technologií a Informatiky na Technické Univerzitě v Liberci. K výběru vět byl použitý algoritmus [Radová 1999] navržený Radovou a Vopálkou na ZČU v Plzni.

Algoritmus je založen na iterativní proceduře, kdy postupně přidává věty do cílového výběru, dokud není dosaženo požadovaného počtu. Opakuje přitom dva základní kroky:

1. Pro každou větu je vypočteno skóre S, které vyjadřuje, jak vhodně ve smyslu cílového fonetického rozložení tato věta doplňuje dosavadní výběr.

2. Věta s nejvyšším skóre S je přidána do výběru.

Skóre S je definováno jako

$$S = \sum_{i=1}^{I} \left| \frac{m_i}{\sum_{j=1}^{I} m_i} - \frac{n_i + n'_i}{\sum_{j=1}^{I} (n_i + n'_i)} \right| \tag{9.1}$$

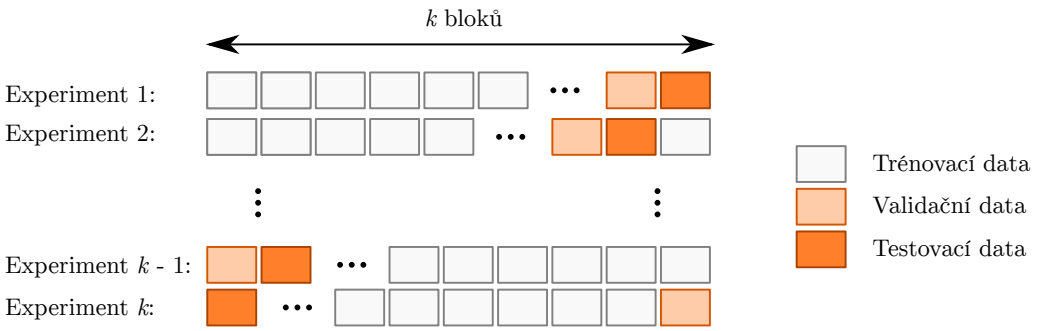
kde I je žádaný počet různých fonémů v cílovém výběru, m_i je četnost fonému i ve výchozím korpusu, n_i je četnost fonému i v dosavadním výběru a n'_i je četnost fonému i v aktuální větě.

9.2 Křížová validace

Úkolem strojového učení a rozpoznávání s učitelem je nalézt nějaký model, který dobře reprezentuje vztah mezi vstupní veličinou x, tedy např. parametrizací audiovizuální sekvence, a výstupní veličinou y, tedy např. textovým přepisem. Velmi jednoduše lze vztah zapsat jako $y = f(x) + \varepsilon$, kde ε reprezentuje náhodný (např. gaussovský) šum, jenž je nezávislý na datech a není možné ho ovlivnit. Učení modelu typicky spočívá ve stanovení parametrické formy $\hat{f}(x)$ a následném odhadu vnitřních (volných) parametrů, např. metodou maximální věrohodnosti. Vyhodnocení kvality modelu se provádí porovnáním skutečných y a predikovaných hodnot $\hat{y}$, nejčastěji čtvercovou odchylkou pro regresní problémy a počtem odlišností pro klasifikační problémy.

Nechť jsou kompletní dostupná data, jež jsou k dispozici pro trénování i testování navrženého modelu, označena jako $\mathcal{D}$. V nejjednodušším případě může být model natrénován i evaluován na této množině. To však s sebou především při malé velikosti $\mathcal{D}$, jež je pro výzkum v oblasti audiovizuálního rozpoznávání příznačná, přináší riziko tzv. **přeučení**, ke kterému dochází při příliš malém poměru omezujících podmínek a počtu volných parametrů modelu. Problém lze ilustrovat na příkladu polynomiální regrese, kde mezi dvěma veličinami x a y předpokládáme vztah $y = \varepsilon + \sum_{i=0}^{n} w_i x^k$. Pokud je závislost y na x ve skutečnosti lineární, tj. $n = 1$, dojde při nesprávné volbě hyperparametru $n \geq 2$ (a tedy příliš vysokému počtu volných parametrů w_i) k tomu, že model bude predikovat i nežádoucí šum ε. Takto lze na trénovací množině dosáhnout libovolně nízké regresní odchylky, ovšem pouze za cenu, že model přestane vystihovat skutečnou podstatu závislosti y na x. Model se v takovém případě příliš specializuje (přeučí) na konkrétní množinu $\mathcal{D}$ a přestane být relevantní pro jakýkoliv jiný vzorek $\mathcal{D}'$ ze stejné globální populace. Klasicky lze problém řešit rozdělením $\mathcal{D}$ na vzájemně disjunktní trénovací sadu $\mathcal{T}$ a validační sadu $\mathcal{V}$. Max. stupeň polynomiální regrese n pak lze zvolit tak, aby regresní odchylka byla minimální na validační sadě $\mathcal{V}$, přičemž samotný odhad parametrů však probíhá na trénovací sadě $\mathcal{T}$. Není tedy možné model neomezeně specializovat na $\mathcal{T}$, protože by tím přestal odpovídat $\mathcal{V}$. Pro omezení vlivu „šťastného“ nebo naopak „smolného“ rozdělení dat lze navíc aplikovat tuto metodu opakovaně s jiným rozdělením na $\mathcal{T}$ a $\mathcal{V}$. Optimální n je pak zvoleno tak, aby **průměrná regresní odchylka** pro všechny validační sady byla minimální.

Tzv. k-násobná křížová validace (K-Fold Cross Validation, KFCV) se od uvedeného postupu liší tím, že zaručuje, aby jednotlivé validační podmnožiny byly vzájemně disjunktní a tedy nezávislé. Datová množina $\mathcal{D}$ je rozdělena na k bloků, z nichž vždy právě jeden slouží jako validační sada a zbytek jako trénovací data. Každý experiment, tedy natrénování a následné otestování modelu, je opakováno



Obrázek 9.3: Princip vnější k-násobné křížové validace.

přesně k-krát, vždy s jiným validačním blokem. Jako optimální je pak vybrán takový model, jenž minimalizuje regresní či klasifikační chybu po zprůměrování přes všechny validační bloky.

Uvedená aplikace k-násobné křížové validace se však týká buď *volby modelu* nebo *odhadu jeho úspěšnosti* na neviděných datech, ne obojího zároveň. Jelikož křížovou validací se hyperparametr n sám stává předmětem optimalizace, neliší se v principu od vnitřních parametrů w_i. Na celý proces trénování a validace tak lze nahlížet jako na učení modelu s parametry $(\ldots, w_i, \ldots, n)$, které probíhá na jediné datové sadě $\mathcal{D} = \mathcal{T} \cup \mathcal{V}$ a tedy hrozí během něj stejné problémy jako v nejjednodušším případě s pevně zvoleným n a jedinou datovou množinou $\mathcal{D}$. Především při velkém počtu experimentů s různým nastavením hyperparametrů pak může být vysoká úspěšnost výsledkem pouhé náhody. Pro odhad stability a úspěšnosti celé trénovací procedury, tj. učení a ladění hyperparametrů pomocí KFCV, je tedy nutné aplikovat **metodu zádrže** (angl. hold-out), neboli z dostupných dat vyčlenit testovací množinu $\mathcal{S}$, která se předchozího nijak neúčastní. Předpokládané skóre modelu při aplikaci na neviděných datech pak lze na $\mathcal{S}$ odhadnout až po kompletním odladění na $\mathcal{T}$ a $\mathcal{V}$.

Jelikož u KFCV je nutné opakovat každý experiment (tedy trénování a testování modelu) k-krát, vyžaduje tento protokol především u složitějších modelů značnou výpočetní a pamětovou kapacitu. Pro odhad úspěšnosti navrženého modelu na neviděných datech využitím KFCV a metody zádrže tak lze s ohledem na tyto nároky postupovat více způsoby. Kompletní data $\mathcal{D}$ jsou přitom vždy rozdělena do k stejně velkých disjunktních bloků, které dle potřeby mohou zastávat jednu z následujících rolí: trénovací ($\mathcal{T}$), validační ($\mathcal{V}$) a testovací ($\mathcal{S}$).

1. „Vnitřní“ křížová validace:
   (a) Jeden z bloků vyčlenit jako zadržovanou (testovací) sadu $\mathcal{S}$.
   (b) Na zbylých datech aplikovat standardní KFCV (zde jako vstup $k-1$ bloků) pro výběr modelu, tedy $k-1$ experimentů.
   (c) Ohodnotit optimální model na zadržované (hold-out) sadě S.
2. „Vnější“ křížová validace:

(a) Natrénovat a vybrat optimální model pomocí $k-2$ bloků $\mathcal{T}$ a jednoho bloku $\mathcal{V}$ a vyhodnotit na bloku $\mathcal{S}$.

(b) Opakovat předchozí bod k-krát pro všechny ostatní $\mathcal{S}$.

(c) Úspěšnost vyhodnotit jako průměrné skóre přes všechny uvažované $\mathcal{S}$.

3. „Vnořená“ křížová validace:

(a) Aplikovat postup z bodu 1 k-krát pro všechny možné volby $\mathcal{S}$.

(b) Úspěšnost vyhodnotit jako průměrné skóre přes všechny uvažované $\mathcal{S}$.

Jelikož postup 1 příliš závisí na volbě hold-out množiny $\mathcal{S}$ a naopak metoda 3 je příliš výpočetně náročná, byl pro experimenty na databázi TULAVD jako kompromis zvolen protokol uvedený v bodě 2, tedy **vnější křížová validace**. Jedná se o zjednodušení plné vnořené křížové validace (bod 3), 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. Postup je ilustrován na obr. 9.3. Databáze TULAVD tedy byla rozdělena do šesti skupin po devíti mluvčích a velikosti trénovací, validační a testovací množiny tak činily 36, 9, resp. 9.

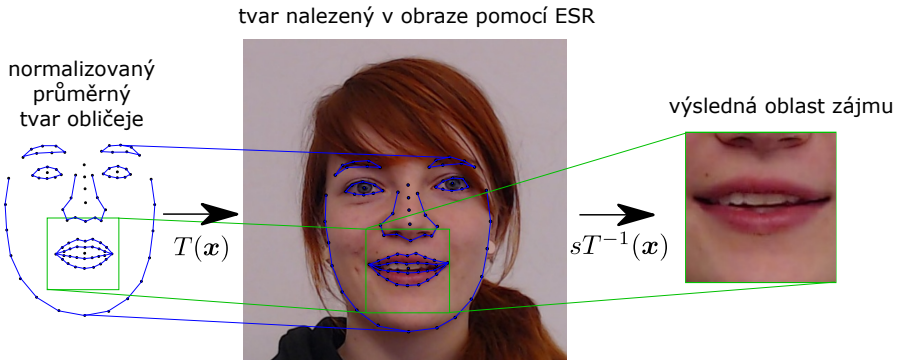
9.3 Extrakce zájmové oblasti

Oblast zájmu (Region of Interest, ROI) je obvykle definována jako obdélníková oblast pokrývající ústa a nejbližší okolí. Pokud není uvedeno jinak, v této práci je ROI definována jako čtvercová oblast o rozměrech 64×64 pixelů extrahovaná na základě lokalizace klíčových bodů na obličeji následujícím postupem, jenž je shodný pro všechny databáze uvedené v této kapitole.

Pro detekci klíčových bodů na obličeji je aplikován algoritmus **explicitní tvarové regrese** (Explicit Shape Regression, ESR) popsaný v sekci 2.5. Parametry tohoto diskriminačního modelu jsou odhadnuty na sadě manuálně anotovaných obrázků v trénovací fázi. Jelikož se práce věnuje především rozpoznávání nezávislém na řečníkovi, je ESR natrénován pouze na mluvčích nacházejících se v trénovací množině každého uvažovaného rozdělení KFCV, viz sekci 9.2. Množství trénovacích obrázků se pro jednotlivé testovací databáze liší a pohybuje se v rozmezí 300 (TULAVD) až 2000 (OuluVS, CUAVE). S cílem co možná nejlepší kompatibility s nejčastěji používanými modely v existujícím výzkumu byla zvolena konfigurace $v=93$ obličejových bodů, již znázorňuje obrázek 9.4. Každá konfigurace (tvar obličeje) je reprezentována vektorem zřetězených jednotlivých souřadnic dle modelu 2.5. Na trénovací je množině je vypočten průměrný tvar $\bar{\boldsymbol{s}}=(\bar{\boldsymbol{s}}_1,\ldots,\bar{\boldsymbol{s}}_v)^\top$, $\bar{\boldsymbol{s}}_i=(\bar{x}_i,\bar{y}_i)$ a normalizovaný tak, aby $\|\bar{\boldsymbol{s}}\|=1$. Velikost a pozice ROI jsou stanoveny relativně vůči normalizovanému průměrnému tvaru $\bar{\boldsymbol{s}}$ a tedy nezávisle na měřítku.

Pro každý snímek vstupní videosekvence je nejprve metodou **Violy a Jonese**[4] (viz sekci 2.2) nalezena přibližná pozice obličeje. Přesný tvar ve formě zřetězeného

[4]Byl aplikován model `haarcascade_frontalface_alt2` distribuovaný společně s knihovnou OpenCV.



Obrázek 9.4: Extrakce oblasti zájmu.

vektoru souřadnic $\boldsymbol{s} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_v)^\top$ klíčových bodů $\boldsymbol{s}_i = (x_i, y_i)$ je odhadnut pomocí natrénovaných ESR modelů. 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. Jelikož ESR při zarovnání nezohledňuje žádné objektivní kritérium a nesnaží se tedy nalézt minimum nějaké funkce, i při pouze málo odlišné inicializaci detektorem VJ se výsledek v každém dalším snímku vždy trochu liší od předchozího, čímž vzniká v klíčových bodech jistý šum. Za účelem minimalizace tohoto „chvění“ byla metoda ESR aplikována na každý snímek 10× a z výsledků spočítán medián. Díky efektivitě algoritmu bylo i přes opakování stále zachováno zpracování v reálném čase, obvykle detekce probíhala v 20–50 snímcích za sekundu.

Pro extrakci ROI je nutné znát geometrickou transformaci, jež popisuje pozici, velikost a natočení obličeje na snímku. Lze ji odhadnout na základě zarovnání průměrného $\bar{\boldsymbol{s}}$ a detekovaného tvaru $\boldsymbol{s}$, tj. minimalizací průměrné odchylky mezi odpovídajícími body obou vektorů souřadnic. Transformace $T(x)$ je tedy řešením

$$T = \underset{T'}{\operatorname{argmin}} \sum_{i=1}^{v} (\bar{\boldsymbol{s}}_i - T'(\boldsymbol{s}_i))^\top (\bar{\boldsymbol{s}}_i - T'(\boldsymbol{s}_i)) \tag{9.2}$$

V případě transformace podobnosti $T(\boldsymbol{x}), \boldsymbol{x} = (x, y)$ reprezentuje posun, otočení a změnu měřítka, tj.

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a & -b & t_x \\ b & a & t_y \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{9.3}$$

Minimalizace (9.2) a tedy nalezení parametrů a, b, t_x, t_y optimální transformace $T(x)$ pak odpovídá řešení soustavy

$$\sum_{i=1}^{v} \begin{bmatrix} \bar{x}_i & -\bar{y}_i & 1 & 0 \\ \bar{y}_i & \bar{x}_i & 0 & 1 \\ \bar{x}_i^2 + \bar{y}_i^2 & 0 & \bar{x}_i & \bar{y}_i \\ 0 & \bar{x}_i^2 + \bar{y}_i^2 & -\bar{y}_i & \bar{x}_i \end{bmatrix} \begin{bmatrix} a \\ b \\ t_x \\ t_y \end{bmatrix} = \sum_{i=1}^{v} \begin{bmatrix} x_i \\ y_i \\ x_i\bar{x}_i + y_i\bar{y}_i \\ y_i\bar{x}_i + x_i\bar{y}_i \end{bmatrix} \tag{9.4}$$

Pixely zájmové oblasti ROI jsou z obrázku I extrahovány dle $\mathrm{ROI}(\boldsymbol{x}) = I(sT^{-1}(\boldsymbol{x}))$ s bilineární interpolací jasových hodnot, přičemž hodnota s je zvolena tak, aby

Excuse me.	See you.
Goodbye.	I am sorry.
Hello.	Thank you.
How are you.	Have a good time.
Nice to meet you.	You are welcome.

Tabulka 9.1: Seznam frází zahrnutých v databázi OuluVS.

výsledná velikost činila požadovaných 64×64 pixelů. Pro zvýšení stability algoritmu jsou parametry transformace T průměrovány přes sousední snímky, tj. $T^{(t)} = (1 - \lambda)T^{(t)} + \lambda T^{(t-1)}$, přičemž faktor λ byl zvolen empiricky na hodnotu 0.7.

Oblast zájmu lze samozřejmě extrahovat i jednodušeji, např. na základě pozice koutků úst detekovaných algoritmem Violy a Jonese. Výše uvedený postup je však díky průměrování odchylky většího množství bodů robustnější vůči chybám detektoru a také zajišťuje neměnnou velikost ROI při různém otevření úst či pohybech řečníka.

9.4 Ostatní použité databáze

Kromě vlastní databáze TULAVD byly pro srovnávací experimenty využity dvě další volně dostupné databáze: OuluVS a CUAVE.

9.4.1 OuluVS

Audiovizuální databáze OuluVS[5], jež byla uvedena v článku [Zhao 2009] a pro výzkum dále použita v např. v pracích [Zhou 2010, Zhou 2011, Ong 2011, Pei 2013, Rekik 2014b, Zhou 2014]. 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). Jelikož pochází z univerzity ve finském Oulu, nejedná se o rodilé mluvčí a promluvy jsou obvykle zabarvené přízvukem. Seznam frází je uveden v tabulce 9.1.

Mluvčí jsou zachyceni z čelního pohledu v rozlišení 720 × 576 pixelů při 25 snímcích za sekundu v prokládaném režimu, viz obr. 6.1. V obrázcích se tedy poměrně často vyskytují artefakty související s nekvalitním zarovnáním řádků, viz obr. 9.5, které lze sice odstranit, avšak pouze za cenu snížení snímkové frekvence. Videa jsou uložena po jednotlivých snímcích jako sekvence bitmapových souborů. Kromě neupravených videonahrávek databáze rovněž obsahuje již předpřipravené zájmové oblasti o velikosti přibližně 120 × 70 pixelů ve čtyřech různých verzích v závislosti na algoritmu jejich extrakce. Nicméně mimo jiné i z důvodu velkého množství chybně zarovnaných ROI byl v experimentální části této práce použitý vlastní postup, popsaný v sekci 9.3. Zvuková stopa byla zachycena se vzorkovací frekvencí 48 kHz a rozlišením 16 bitů.

[5] `http://www.cse.oulu.fi/CMV/Downloads/OuluVS`

Obrázek 9.5: Artefakty způsobené prokládaným režimem.

Databáze OuluVS je určena především pro evaluaci algoritmů odezírání ze rtů bez využití akustické informace. Tomu je uzpůsobena velikost slovníku, jenž obsahuje pouhých 20 slov. Databáze je navíc distribuována bez anotace akustických dat (fonémové či alespoň slovní segmentace) a ve většině prací jsou tak uvedené fráze rozpoznávány vždy jako celek představující základní slovní jednotku. Nejvyšší úspěšnosti rozpoznávání nezávislého na řečníkovi dosáhli v této poměrně jednoduché úloze autoři Pei a kol. se skóre 89,7 %. Nejvýznamnějšími problémy při rozpoznávání se přitom jeví být závislost vizuální parametrizace na jednotlivých mluvčích a proměnná a často nedostatečná délka promluvy (např. pouze 8 snímků).

Hlavním důvodem pro využití OuluVS v této práci je porovnání navržených parametrizací a celkového postupu vizuálního rozpoznávání se stavem poznání. Jelikož ve všech článcích, kde se tato databáze vyskytla, autoři zvolili shodný postup evaluace dle protokolu Leave-one-out Cross Validation (LOOCV), byl tento postup aplikován i zde. Jedná se o limitní variantu opakované křížové validace, kde každý blok je tvořen právě jedním řečníkem. Každé rozdělení tedy obsahuje 19 trénovacích a 1 testovacího řečníka. Byl zvolen vlastní formát ROI o rozměrech 64×64 pixelů, shodný s databází TULAVD, a nebyly tedy využity předpřipravené soubory zmiňované výše. Extrakce ROI byla provedena dle postupu uvedeného v sekci 9.3, přičemž pro trénování ESR detektorů bylo využito něco málo přes 1763 manuálně anotovaných obrázků z databází TULAVD, CUAVE a GRID, a necelých 300 dalších od mluvčích přímo z OuluVS. Pro každé rozdělení LOOCV (celkem 20) byl přitom natrénován samostatný detektor tak, aby neobsahoval data od testovacího řečníka. Pro modely, jenž vyžadují anotaci jednotlivých snímků (např. dynamizace LDA, viz sekci 3.4.1), byly zvukové nahrávky nuceně zarovnány (forced alignment) s využitím volně dostupného akustického modelu [6] pro americkou angličtinu připraveného lingvistickou laboratoří na University of Pennsylvania.

9.4.2 CUAVE

Audiovizuální databáze CUAVE[7] byla vytvořena roku 2002 na Clemson University v Jižní Karolíně, USA, a za dobu své existence se díky své dostupnosti a designu stala poměrně populární ve výzkumu audiovizuálního rozpoznávání řeči, viz např. práce [Saenko 2006, Lucey 2006b, Lucey 2008, Papandreou 2009, Ngiam 2011, Estellers 2012a, Pei 2013]. Detailní popis návrhu a postupu při nahrávání je

[6] `http://www.ling.upenn.edu/phonetics/p2fa`
[7] `http://www.clemson.edu/ces/speech/cuave.htm`

k dispozici v článku [Patterson 2002]. Databáze obsahuje 4 verze promluv od dohromady 36 anglicky mluvících rodilých mluvčích (17 mužů a 19 žen). Ve všech verzích je slovník tvořen 10 číslovkami od nuly do devíti, liší se však způsob promluv. V první části každý mluvčí 5× opakuje číslovky vzestupně, v druhé 3× pozpátku a při pohybech hlavy, v třetí vzestupně jednou při pohledu zleva a jednou zprava, a konečně ve čtvrté v náhodném pořadí a s pohyby hlavy. V prvních třech částech jsou přitom číslovky vyslovovány odděleně, v poslední spojitě. Databáze CUAVE navíc obsahuje videosekvence, na kterých střídavě i současně mluví dva mluvčí.

Videa jsou uložena v kontejneru MPEG s rozlišením 720×480 pixelů při 29,97 snímcích za sekundu, podobně jako u databáze OuluVS v prokládaném režimu. Ukázkový snímek je znázorněn na obr. 6.1. Audio je součástí video souboru a nabízí standardní kvalitu 44,1 kHz a 16bitové rozlišení. Pro usnadnění práce s databází je přiloženo časové ohraničení jednotlivých slov v promluvách. Databáze rovněž obsahuje manuálně segmentovaný obličej a rty, ovšem u každého mluvčího pouze pro jeden snímek. Zájmová oblast byla tedy v této práci extrahována stejným způsobem jako u ostatních uvažovaných databází. Postup je uveden v sekci 9.3.

Ve výzkumu se zdaleka nejčastěji využívá první část databáze, tedy izolované číslovky. Nejlepších výsledků v současné době dosáhli Papandreou a kol. [Papandreou 2009] s úspěšností 83 % pro rozpoznávání bez využití akustické informace a Estellers a kol. [Estellers 2012a] pro audiovizuální rozpoznávání zašuměného signálu (0 dB) s celkovým skóre 88 %.

V této práci byla CUAVE zpracována obdobně jako ostatní databáze. Jelikož v dostupných publikacích neexistuje jednotný způsob rozdělení dat na trénovací a testovací část, bylo zvolen vlastní protokol dle popisu v sekci 9.2. Databáze tedy byla rozdělena do 6 skupin po 6 mluvčích, přičemž 4 skupiny sloužily jako trénovací množina, 1 jako validační a 1 jako testovací. Byl opět zvolen shodný formát ROI o rozměrech 64×64 pixelů a stejný postup její extrakce jako u databází TULAVD a OuluVS. Fonémová anotace jednotlivých snímků byla získána shodným postupem jako v případě OuluVS, tedy nuceným zarovnáním s americko-anglickým akustickým modelem[8].

[8] `http://www.ling.upenn.edu/phonetics/p2fa`

10. Rozpoznávání izolovaných slov a frází

Jako měřítko úspěšnosti byla pro všechny experimenty zvolena slovní přesnost [%] (7.1), jež v případě izolovaných slov vyjadřuje podíl správně rozpoznaných jednotek vůči celkovému počtu. Při porovnání dvou různých výsledků může být uvedena i relativní změna slovní chybovosti δ_{WER} (7.2). Ve všech experimentech kromě byly pro rozpoznávání aplikované celoslovní lineární levo-pravé skryté markovské modely. Počet stavů byl stanoven empiricky jako nejvyšší možný vzhledem k délce nejkratší promluvy v databázi. Pro TULAVD tedy bylo zvoleno 14 stavů, pro OuluVS a CUAVE pouze 8 a to i přesto, že OuluVS obsahuje oproti TULAVD celé fráze. Každý stav HMM je modelován gaussovskou směsí s jednou až dvěma komponentami a diagonální kovarianční maticí. Pokud není uvedeno jinak, konkrétní počet komponent je pro všechny stavy stejný a je jedním z optimalizovaných hyperparametrů v křížové validaci. Pro trénování modelů i klasifikaci byla využita volně dostupná knihovna HTK[1] [Young 2006] verze 3.4.1.

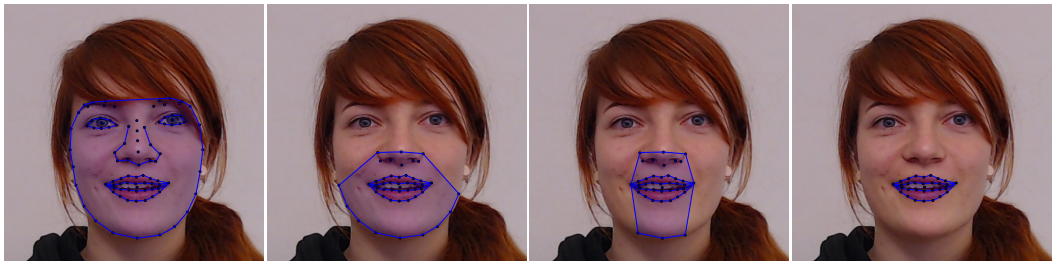
10.1 Vizuální rozpoznávání

Do experimentů byly zahrnuty tři databáze: vlastní TULAVD a volně dostupné OuluVS a CUAVE. Jejich detailní popis je uveden v kapitole 9. 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 9.2, přičemž počty mluvčích v trénovacích, validačních, resp. testovacích množinách jsou uvedeny v třetím sloupci tabulky 10.1. Ve všech experimentech na těchto dvou databázích jsou validační množiny využity pro odladění optimálních hodnot hyperparametrů, které jsou následně aplikovány při vyhodnocení na odpovídajících testovacích množinách. Pro každé rozdělení je tedy maximalizováno *validační* skóre, avšak do výsledků započteno odpovídající *testovací* skóre. Výslednou úspěšnost pak představuje průměrná hodnota slovní přesnosti dosažená pro jednotlivé testovací množiny. V případě OuluVS byla zvolena standardní křížová validace vynech jeden (LOOCV), kde hyperparametry jednotlivých modelů a algoritmů jsou vybrány tak, aby maximalizovaly průměrné skóre

[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 10.1: Rozdělení dat pro jednotlivé databáze.



(a) AAM-f (b) AAM-r (c) AAM-s (d) AAM-m

Obrázek 10.1: Konfigurace klíčových bodů AAM uvažované v experimentech.

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. Databáze navíc obsahuje pouze 20 řečníků a dělení na tři dostatečně velké skupiny by mohlo být problematické.

10.1.1 Srovnávací experimenty

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). DCT a PCA byly vybrány jako základní příznaky využívající pouze jasové hodnoty skrze lineární transformaci – DCT jako datově nezávislá projekce, PCA jako datově závislá projekce. Parametrizace AAM a její odvozeniny (tvarové, texturové, kombinované a vlastní hloubkově rozšířené parametry) byla zvolena jako zástupce tvarově orientovaných příznaků s přesnou lokalizací zájmové oblasti. LBPTOP patří mezi v současné době velmi populární a často užívané příznaky, ovšem pouze v kombinaci s jinými klasifikátory než klasickým HMM. Mimo porovnání s vlastní navrženou parametrizací bylo jedním z účelů v experimentech také zjistit, jak vhodné jsou pro extrakci na každém snímku, nikoliv pouze na několika málo shlucích sousledných snímků. DCT3 a HOGTOP jsou vlastní parametrizace navržené v této práci. 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.

Jak bylo zmíněno v sekci 3.2, u příznaků využívajících tvarový model (2.6) je možné zvolit libovolnou podmnožinu bodů tak, aby jejich konvexní obal zbytečně nezahrnoval oblasti, jež nenesou žádnou informaci. V praxi to znamená, že se obvykle uvažují pouze body na rtech, viz např. články články [Matthews 2002, Lan 2009, Shin 2011]. Lze ale nalézt i práce, kde je zájmová oblast volena jinak.

Typ	P	Video		Hloubka	
		L	C	L	C
AAM-f	48,4	48,6	54,9	42,6	54,4
AAM-r	**54,6**	**52,7**	**58,1**	**48,9**	**59,4**
AAM-s	50,1	49,9	56,4	27,4	46,0
AAM-m	47,1	51,5	54,2	28,8	42,3

Tabulka 10.2: Slovní přesnost [%] pro různé typy AAM v úloze rozpoznávání izolovaných slov na databázi TULAVD.

Např. v [Papandreou 2009, Chiţu 2012] byly zahrnuty i oblasti kolem úst. V práci [Matthews 2001] pak autoři využili dokonce celý obličej. Ve všech uvedených pracích však zájmové oblasti byly vybrány heuristicky a bez porovnání s ostatními modely. Část experimentů zde je proto věnována této problematice. Jsou uvažovány celkem čtyři různé kombinace klíčových bodů na obličeji, jejichž varianty jsou znázorněny na obr. 10.1.

Slovní přesnost dosaženou pro jednotlivé konfigurace na databázi TULAVD shrnuje tabulka 10.2. Experiment byl proveden samostatně pro tvarové (P), texturové (L) a kombinované (C) příznaky a opakován pro obrazová (video) a hloubková data. Výsledky na databázích OuluVS a CUAVE byly zaneseny do tabulky 10.3. Pro všechny typy příznaků byl postupem uvedeným na začátku této kapitoly optimalizován počet koeficientů a gaussovských komponent (1–2) v HMM tak, aby výsledné skóre bylo maximální. Ve většině experimentů nejvyššího skóre dosáhly AAM příznaky s regionem ‘r’, který zahrnuje dolní část obličeje. V několika případech bylo dosaženo vyššího skóre pro podoblasti ‘s’ (redukovaná varianta ‘r’) a ‘m’, avšak pouze pro tvarové příznaky. Zohlednění okolí úst tedy napomáhá rozpoznávání především díky textuře, zatímco tvarová informace se soustřeďuje zejména na rty. Jistou roli zde rovněž může hrát nejistota obličejového detektoru, kdy nekvalitní zarovnání úst způsobí nenávratnou ztrátu texturové informace, zatímco zahrnutím širšího okolí rty v konvexním obalu bodů zůstanou a budou pouze posunuté či zkreslené. Jak ovšem výsledky s užitím celého obličeje (AAM-f) ukazují, příliš velká oblast není pro rozpoznávání vhodná. Z experimentů je rovněž zřejmé, že se tvarové a texturové příznaky poměrně vhodně doplňují, kdy kombinované parametry dosahují vždy nejlepších výsledků.

Další experiment je opět zaměřen na **optimalizaci příznaků**, především z pohledu řečové dynamiky a postprocessingu. Je porovnána statická parametrizace, kdy jsou příznaky extrahovány pro každý snímek nezávisle, vůči Δ dynamizaci a LDA, viz sekci 3.4. U Δ dynamizace je vždy odečtena parametrizace předchozího snímku a připojena k aktuálnímu. U LDA je pospojováno $2K+1$ sousledných snímků a vzniklý hypervektor redukován metodou LDA (3.5). U některých příznaků je tato metoda ještě navíc následována Δ dynamizací. Optimální postprocessing byl pro každou parametrizaci stanoven zvlášť s cílem maximalizovat slovní přesnost

Typ	OuluVS			CUAVE		
	P	**L**	**C**	**P**	**L**	**C**
AAM-f	48,9	51,3	59,8	54,0	44,1	59,1
AAM-r	57,7	**65,5**	**72,8**	**61,4**	**55,3**	**63,8**
AAM-s	59,3	62,9	69,7	**61,4**	53,9	63,5
AAM-m	**61,1**	59,2	67,5	58,4	51,1	59,5

Tabulka 10.3: Slovní přesnost [%] pro různé typy AAM v úloze rozpoznávání izolovaných jednotek na databázích OuluVS a CUAVE.

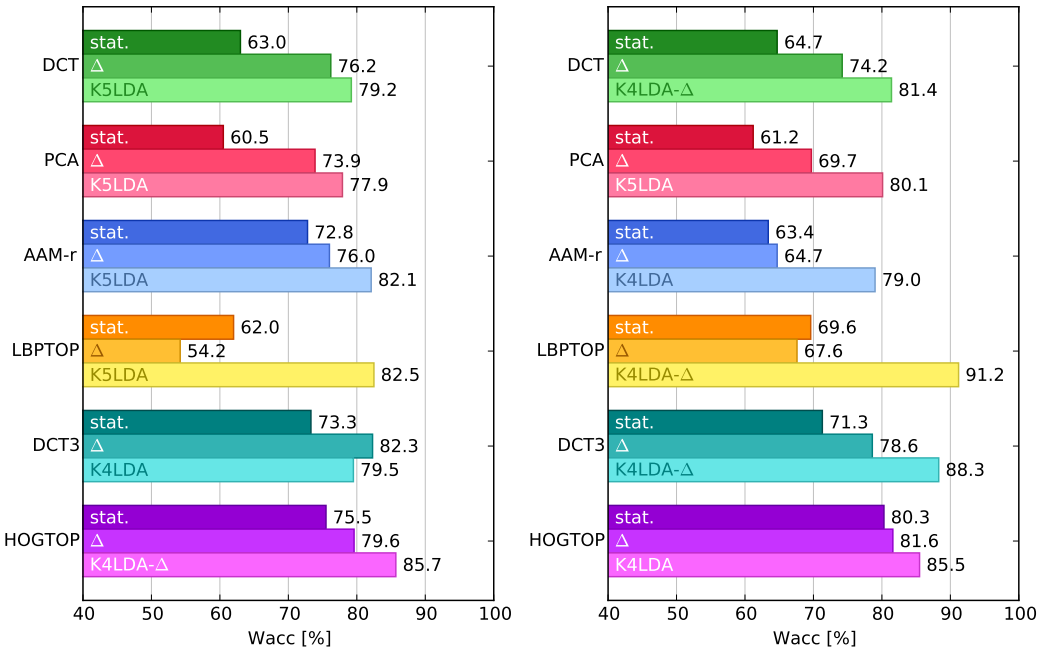
(nebyl považován za hyperparametr a tedy předmět křížové validace). Výsledky pro databázi TULAVD jsou uvedeny v tabulce 10.4 a pro OuluVS a CUAVE na grafech 10.2a a 10.2b, kde každá trojice sloupců odpovídá statickým, Δ, resp. LDA(+Δ) příznakům.

Dle očekávání dosahují dynamické varianty stejných příznaků vyšší úspěšnosti než statická parametrizace, přičemž ve většině případů vykazuje lepší výsledky dynamizace metodou LDA. Poměrně zajímavá je skutečnost, že LDA dynamizace maximální dosažená skóre sbližuje a její vhodnou aplikací se tak rozdíly mezi jednotlivými parametrizacemi snižují až mizí. Např. na databázi TULAVD je rozdíl mezi PCA a AAM-r ve statické variantě 7 % ($\delta_{\mathrm{WER}} = -14$ %), avšak po LDA dynamizaci pouze statisticky nevýznamných 0,2 % ($\delta_{\mathrm{WER}} = -0,4$ %). Přínos modifikace je navíc velmi výrazný, pro DCT, PCA, AAM či LBPTOP se zlepšení pohybuje v rozmezí 16–23 % (průměrná δ_{WER} cca -39 %), tedy mnohdy podstatně více než rozdíly mezi samotnými parametrizacemi. Vhodnější pro dynamizaci jsou samozřejmě příznaky, jenž při extrakci z aktuálního snímku nezohledňují okolí. S výjimkou LBPTOP na databázi OuluVS je zlepšení dynamicky navržených příznaků (LBPTOP, DCT3, HOGTOP) podstatně nižší, max. 10 % (δ_{WER} cca -43 %), jelikož po dynamizaci je využito již příliš velké okolí.

V experimentu se rovněž ukazuje přínos příznaků navržených v této práci. Oba navržené typy, DCT3 a HOGTOP, na vlastní databázi TULAVD dosahují ve statických variantách nejvyšší slovní přesnosti, 61,6 %, resp. 76,1 % pro video a 62,9 %, resp. 72,1 % pro hloubku. 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. Pouze v jednom případě na databázi CUAVE bylo vhodnou modifikací LBPTOP dosaženo lepšího rozpoznávacího skóre, 91,2 % oproti 85,5 %.

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



(a) (b)

Obrázek 10.2: Slovní přesnost [%] v závislosti na dynamizaci v úloze rozpoznávání izolovaných jednotek na databázích OuluVS (a) a CUAVE (b).

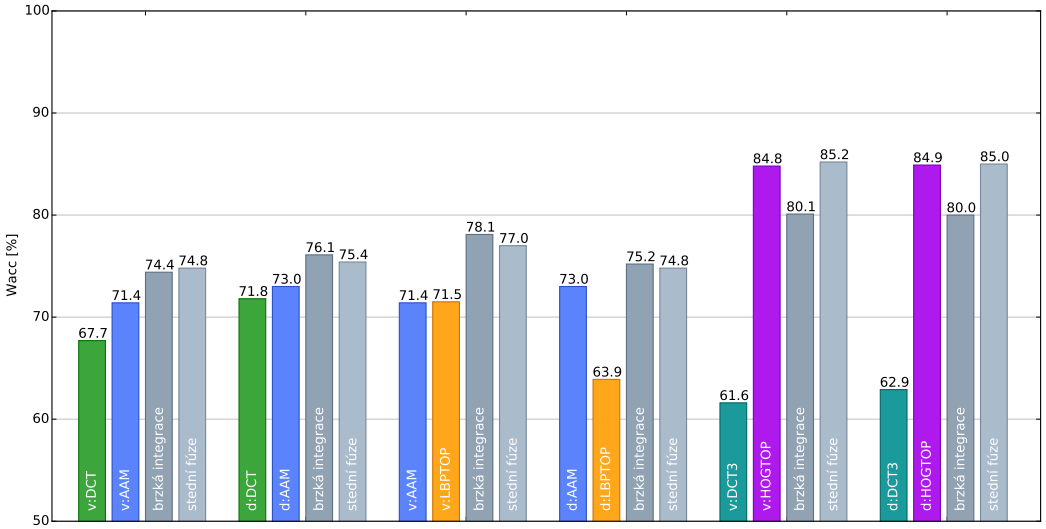
10.1.2 Kombinace příznaků

Různé parametrizace se zaměřují na odlišnou informaci obsaženou ve vstupním signálu. Např. obrazové transformace jako DCT a PCA lineárně kombinují jasové hodnoty zájmové oblasti za účelem snížení celkového rozměru. Příznaky AAM naproti tomu využívají kromě jasových hodnot i informaci o tvaru úst odhadnutého z pozic předem definovaných klíčových bodů. Parametrizace jako LBPTOP, DCT3, či HOGTOP se pak snaží využít především dynamiku řeči, tedy závislost vstupního signálu na čase. Lze tedy očekávat, že jednotlivé druhy příznaků by se mohly vzájemně doplňovat a jejich vhodnou kombinací by mohlo být možné navýšit úspěšnost rozpoznávání.

Kromě kombinace vizuálních parametrizací se tato práce zabývá i vlivem informace o hloubce (vzdálenosti jednotlivých pixelů od kamery). Neliší se tedy pouze jednotlivé parametrizace, ale i přímo zdroj. Informace o hloubce 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. Velký rozdíl v hloubkové informaci se rovněž projeví při otevřených ústech, tedy např. u samohlásek 'a' či 'e', ale i u navázaných souhlásek, např. při vyslovování slov „ale" nebo „protože".

Za účelem zjištění přínosu kombinace parametrizací a zdrojů byly příznaky kombinovány dvěma způsoby: brzkou a střední integrací, viz kapitolu 5. 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) (8.1), jež do modelu AAM přidává hloubkovou texturu. V případě střední integrace byl aplikován synchronní vícekanálový markovský model (MSHMM) dle (5.10), kde $\sum_s \lambda^{(s)} = 1$. Váhy byly považovány za hyperparametr modelu a tedy křížově validovány postupem uvedeným 10.1 (pro každé rozdělení odděleně). Všechny parametrizace byly optimalizovány dle výsledků z předchozích experimentů a kombinovány tak byly jejich modifikované verze již bez dalšího postprocessingu.

Výsledky fúze pro vybrané kombinace příznaků na databázi TULAVD znázorňuje sloupcový graf 10.3, kde každá čtveřice odpovídá jedné dvojici. Nejvyšší, přibližně 23% průměrné relativní redukce WER bylo dosaženo pro pár AAM a LBPTOP, jejichž samostatná skóre jsou 71,4 %, resp. 71,5 % a po vektorovém spojení (brzké integraci) až 78,1 %. Oba typy parametrizace totiž zachycují zcela jiný typ informace: AAM statickou, lokalizovanou a včetně zohlednění tvaru úst, zatímco LBPTOP dynamickou, včetně širšího okolí a pouze texturovou, a proto se vhodně doplňují. Naopak příznaky s podobnými charakteristikami dle očekávání příliš vhodné pro kombinaci nejsou. Např. pro dvojici DCT a PCA je relativní zlepšení WER pouze cca 7–8 % dle typu fúze. Navržená parametrizace HOGTOP propojením s jinými příznaky na databázi TULAVD dalšího zlepšení nedosahuje. S výjimkou statisticky nevýznamného zlepšení při střední fúzi s DCT3 ve všech kombinacích ve výsledku vykazuje nižší slovní přesnosti než samostatně. Je to dáno především velkým výkonnostním odstupem HOGTOP od ostatních parame-



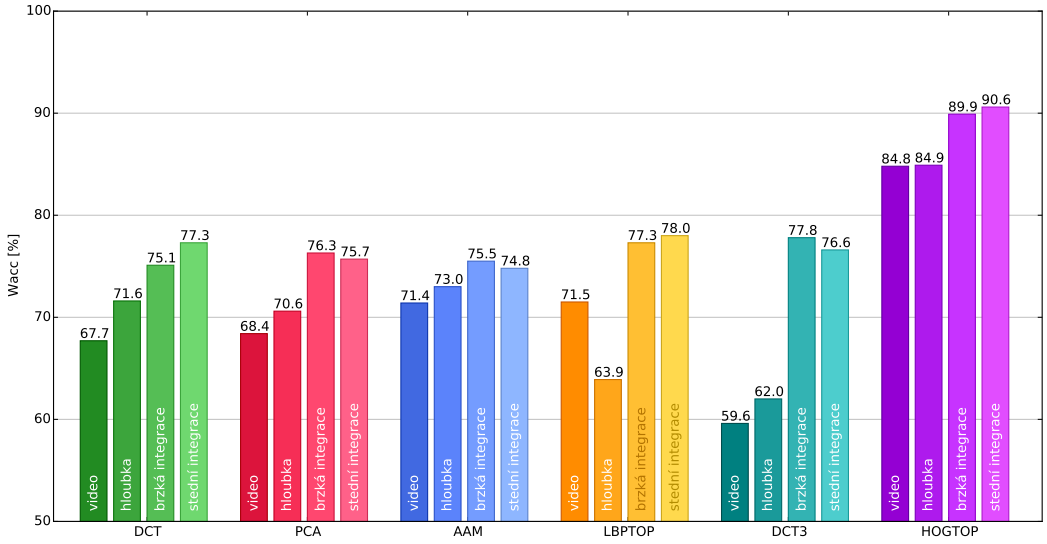
Obrázek 10.3: Dosažená úspěšnost v úloze rozpoznávání izolovaných slov na databázi TULAVD pro vybrané kombinace příznaků.

trizací, kdy ani jejich odlišný charakter nedokáže kompenzovat rozdíl. Podobný jev lze pozorovat u audio-vizuální kombinace, pokud jsou zvukové nahrávky čisté. V takových případech totiž rovněž dochází integrací vizuálních příznaků spíše k mírnému zhoršení úspěšnosti, více v sekci 10.2. Výsledky brzké a střední integrace pro všechny páry příznaků jsou uvedeny v příloze na obrázcích A.3, resp. A.4.

Sloupcový graf 10.4 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. Nejlepších výsledků bylo opět dosaženo aplikací příznaků HOGTOP, kde po integraci obou zdrojových modalit výsledná slovní přesnost překročila 90 %. Jak bylo zmíněno dříve, hloubková data extrahovaná pomocí zařízení Kinect s sebou přinášejí výhodu v necitlivosti na změnu světelných podmínek, což může i být jednou z příčin dobrých výsledků. Hlavní výhoda jejich aplikace však tkví v částečné komplementaritě vůči video signálu, jež je dána odlišnou podstatou zdrojových dat a především díky níž je po zkombinování kanálů dosaženo zlepšení. Nevýhodou tohoto postupu je nutnost extrahovat parametrizaci dvakrát, což může komplikovat zpracování v reálném čase.

10.1.3 Srovnání se stavem poznání

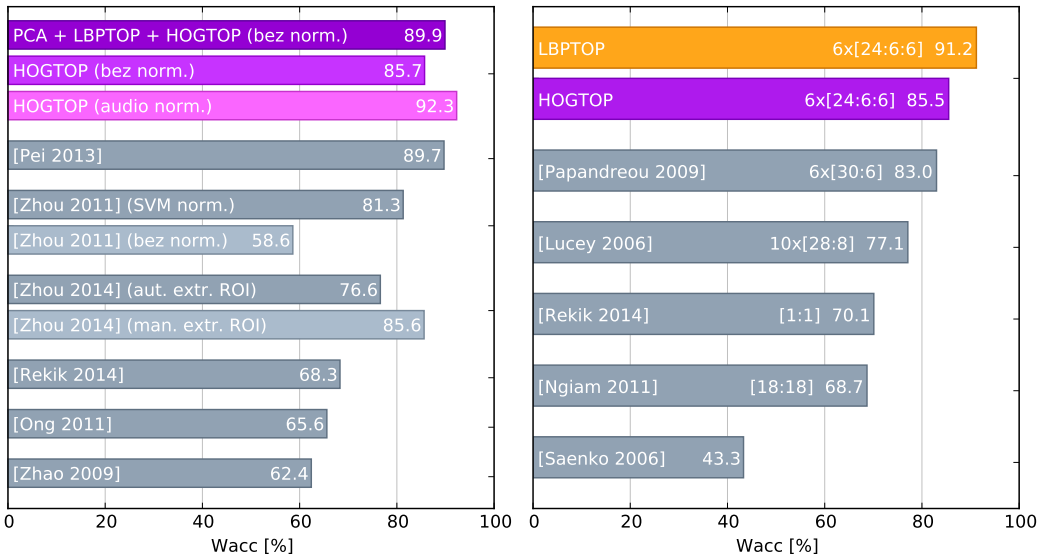
Výsledky v podobě slovní přesnosti **na databázi OuluVS** z této a jiných vybraných prací přehledně shrnuje graf 10.5a. Doposud nejlepších výsledků 89,7 % na databázi



Obrázek 10.4: Dosažená úspěšnost v úloze rozpoznávání izolovaných slov na databázi TULAVD pro kombinace obrazových a hloubkových příznaků.

OuluVS dosáhli Pei a kol. [Pei 2013], kteří při rozpoznávání využili několika různých typů příznaků (LBP, HOG, tvarové příznaky vycházející zarovnání obličeje pomocí AAM) a celé promluvy následně promítli do redukovaného prostoru pomocí vícerozměrného škálování (Multidimensional Scaling, MDS). Algoritmus je však uzpůsobený 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. V [Zhou 2011, Rekik 2014b, Zhao 2009] probíhá extrakce příznaků po blocích a každá promluva je reprezentována jako jediný spojený příznakový vektor klasifikovaný pomocí SVM. V [Ong 2011, Zhou 2014] autoři jiné algoritmy parametrizace a klasifikace, avšak i oni se zaměřili výhradně na rozpoznávání izolovaných jednotek. 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.

Z článků rovněž není zcela zřejmé, zda bylo z nahrávek odstraněno počáteční a koncové ticho. Pouze Zhou a kol. v článku [Zhou 2011] uvedli výsledky pro oba případy, ze kterých je zřejmá důležitost tohoto implementačního detailu. Při detekci řečových a neřečových segmentů pomocí SVM klasifikátoru (ovšem stále využívajícího pouze obrazová data) dosáhli podstatně vyšší slovní přesnosti (81,3 %) než bez této normalizace nahrávek (58,6 %). Jelikož ostatní autoři kromě Rekika a kol. [Rekik 2014b] se o ničem podobném vůbec nezmiňují, lze spíše předpokládat, že žádnou normalizaci neprovedli a výsledky tak odpovídají nahrávkám o celé délce. V této práci byly všechny experimenty provedeny pro oba případy, přičemž videa byla normalizována s využitím akustické informace (nucené zarovnání). Všechny doposud uvedené výsledky byly dosaženy bez jakékoliv normalizace. Po oříznutí



(a) (b)

Obrázek 10.5: Porovnání slovní přesnosti [%] dosažené v této práci (barevně zvýrazněné sloupce) s vybranými články od jiných autorů na databázích OuluVS (a) a CUAVE (b).

počátečního a koncového ticha z každé nahrávky max. slovní přesnost za použití příznaků HOGTOP vzrostla z 85,7 % na 92,3 %, tedy na nejvyšší hodnotu vůbec.

Jak bylo zmíněno v kapitole 7, přímé porovnání komplikují rozdíly v předzpracování vizuálních dat a extrakci zájmové oblasti. Ve většině prací byly aplikovány jednoduché ad hoc metody detekce na základě pozice očí apod., sofistikovanější algoritmy autoři využili pouze v [Pei 2013, Rekik 2014b]. Přímé porovnání provedli Zhou a kol. v [Zhou 2014], kde při manuální extrakci ROI vzrostla slovní přesnost z 76,6 % na 85,6 % ($\delta_{\mathrm{WER}} = -36$ %), tedy stejně dobrý výsledek jako v této práci při aplikaci příznaků HOGTOP. Zde se tedy ukazuje důležitost kvalitní detekce ROI, což bylo hlavním důvodem pro aplikaci explicitní tvarové regrese v této práci.

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 10.5b, 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. V základní nemodifikované verzi přitom LBPTOP dosahuje pouze 69,6 % (HOGTOP 80,3 %), Δ-dynamizací

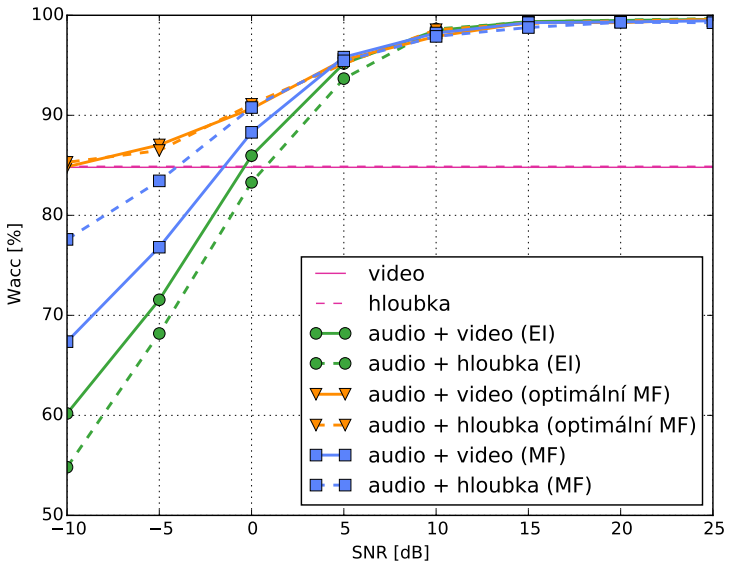
se pak skóre dokonce sníží na 67,6 %. Výrazný nárůst slovní přesnosti aplikací dynamické LDA je pro parametrizaci LBPTOP charakteristický. Proč k němu však v takovéto míře dochází, není zcela zřejmé.

Nejvíce se z hlediska extrakce příznaků, klasifikace a celkového testovacího protokolu této práci podobají články [Lucey 2006b, Papandreou 2009]. V [Lucey 2006b] autoři parametrizovali nahrávky pomocí blokové DCT, v [Papandreou 2009] pak Papandreou a kol. modifikovali AAM příznaky zohledněním nejistoty detektoru klíčových bodů. V obou pak byly pro rozpoznávání aplikovány celoslovní HMM. Rozdíly v celkové úspěšnosti oproti této práci tedy lze vysvětlit především kvalitou parametrizace, přesností a stabilitou vizuálního předzpracování a dalšími nespecifikovanými implementačními detaily. V ostatních článcích autoři zvolili např. jiný klasifikátor a přímé porovnání je tak složitější.

10.2 Audiovizuální rozpoznávání v hlučném prostředí

Vizuální příznaky samy o sobě neobsahují dostatek informace na spolehlivé odezírání mluvené řeči. Mohou ale sloužit jako podpora systémů pro rozpoznávání řeči v hlučném prostředí. V následujících experimentech byl tedy posuzován přínos obrazové informace v těchto 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, SNR) pohyboval v intervalu $[-10, 25]$ dB s krokem 5 dB. SNR byl vypočten jako poměr energií původního řečového signálu a aditivního šumu dle (5.15) z celé délky promluvy včetně krátkého ticha na začátku a v grafech je proto systematicky mírně podhodnocený. 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. Vektor akustických příznaků tedy měl celkový rozměr 39 koeficientů. Testování proběhlo podobně jako v předchozích případech využitím křížové validace. Pro kombinaci zvukových a vizuálních příznaků byly podobně jako v sekci 10.1.2 aplikovány vícekanálové HMM, přičemž váhy $\lambda^{(s)}$ jednotlivých kanálů s byly považovány za hyperparametry modelu.

Graf 10.6 porovnává tři varianty kombinace akustických a vizuálních příznaků HOGTOP: brzkou integraci (EI), střední fúzi (MF) s optimálními vahami pro každé SNR zvlášť (optimální MF) a střední fúzi s napevno nastavenými vahami (MF). V posledním zmiňovaném případě byly zvoleny váhy $\lambda^{(s)}$ tak, aby maximalizovaly skóre pro SNR = 5 dB. Jelikož byly váhy $\lambda^{(s)}$ považovány za hyperparametry modelu, bylo takto postupováno pro každé rozdělení křížové validace zvlášť a váhy maximalizovaly pouze validační skóre. V grafu je pak uvedeno skóre testovací. Zde

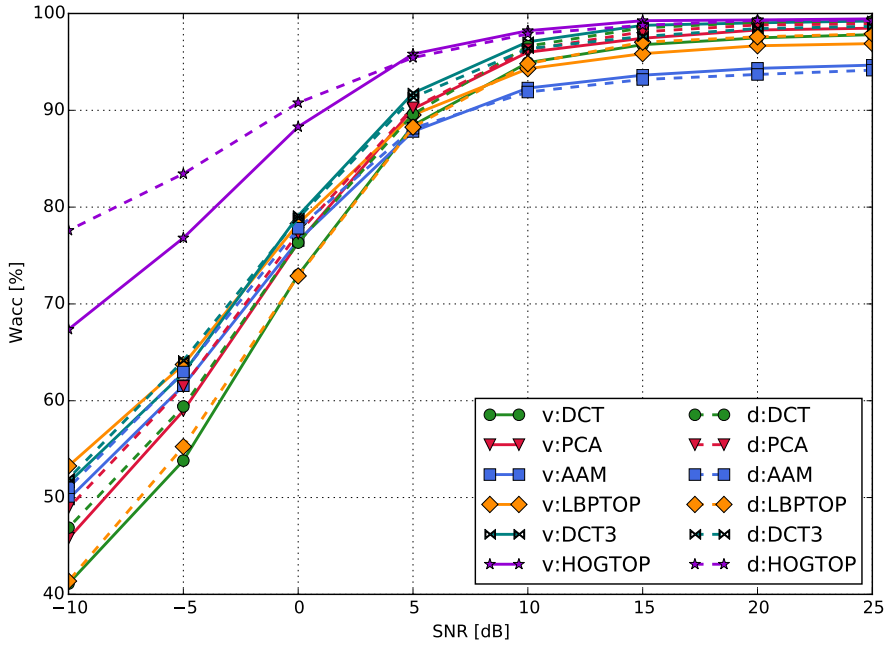


Obrázek 10.6: Audiovizuální rozpoznávání izolovaných slov na databázi TULAVD pomocí příznaků HOGTOP v prostředí s hlukem typu babble v závislosti na SNR a metodě integrace akustických a vizuálních příznaků.

se ukazuje výhoda vícekanálového modelu, který umožňuje skrze nastavení vah $\lambda^{(s)}$ optimalizovat využití jednotlivých kanálů. Váhy lze nastavit např. tak, aby byl maximalizován vážený průměr přes všechny SNR dle předpokládaného prostředí, ve kterém bude systém nasazen, či je měnit za běhu dle potřeby. V této práci nebyl implementován algoritmus pro odhad SNR z akustických dat, proto byly váhy nastaveny napevno. Horní mez slovní přesnosti, kterou by šlo teoreticky dosáhnout při optimálním odhadu SNR, zachycuje průběh optimální MF.

Na grafu 10.7 jsou prezentovány výsledky audiovizuálního rozpoznávání při zahlučení typu babble pro všechny porovnávané parametrizace. Váhy dvoukanálového HMM jsou opět nastaveny napevno tak, aby aby maximalizovaly úspěšnost při SNR = 5 dB. Ukazuje se, že většina příznaků dosahuje podobných výsledků, pouze s výjimkou HOGTOP, jejichž odstup při nejnižších SNR činí až 14–37 %. Stejné průběhy pro bílý šum a tovární hluk zobrazují grafy A.5, resp. A.6.

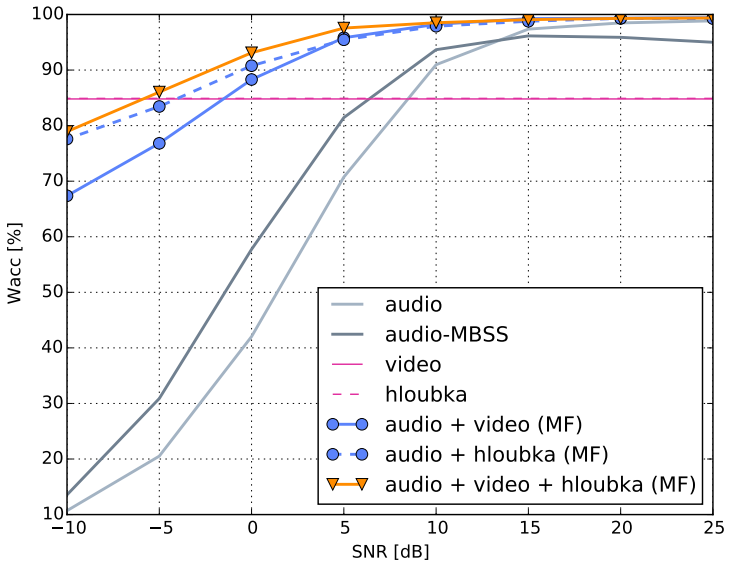
Grafy 10.8 a 10.9 porovnávají audiovizuální rozpoznávání s rozpoznáváním na zvukově opravených nahrávkách. V experimentech byly porovnávány samostatné audio příznaky aplikované na zahlučených nahrávkách, samostatné audio příznaky aplikované na nahrávkách se zvýrazněnou řečí, samostatné vizuální příznaky a kombinované audio-vizuální příznaky s pevnými vahami opět nastavenými tak, aby maximalizovaly úspěšnost při SNR = 5 dB, pro detaily viz výše. Pro zvýraznění řeči a odstranění šumu z akustických nahrávek byly zvoleny populární algoritmy vícepásmového spektrálního odečítání (Multiband Spectral Subtraction, MBSS) a logMMSE (log Minimum Mean Square Error) [Loizou 2013]. Metoda logMMSE byla aplikována na bílý šum, MBSS pak na hluk typu babble, protože v této konfiguraci



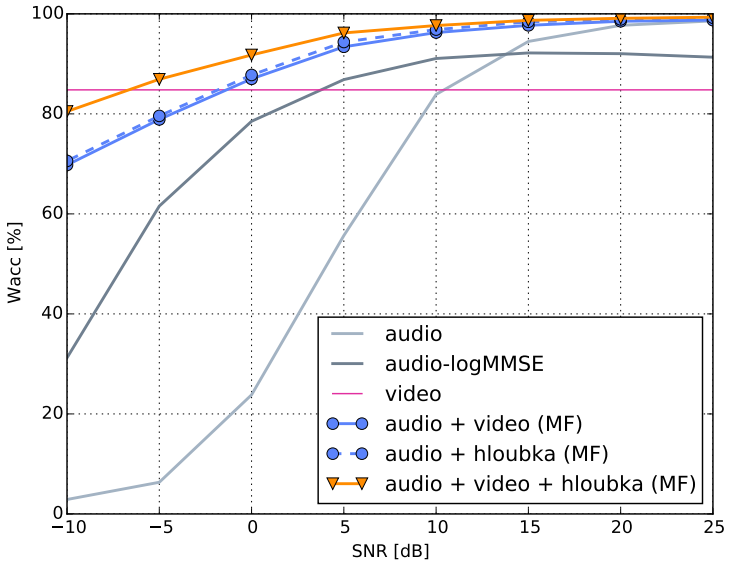
Obrázek 10.7: Audiovizuální rozpoznávání izolovaných slov na databázi TULAVD v prostředí s hlukem typu babble pro různé vizuální příznaky.

dosahovaly algoritmy pro zvýraznění řeči nejvyššího rozpoznávacího skóre.

Výsledky prokazují přínos vizuální informace v zahlučených prostředích. Ve všech experimentech bylo pomocí audiovizuálních příznaků dosaženo vyššího rozpoznávacího skóre než při použití samostatných audio příznaků a to pro zahlučené i opravené nahrávky. Přínos obrazové složky se dle očekávání projevuje především pro nižší hodnoty odstupu signálu od šumu. Nejmenšího rozdílu mezi samostatnými audio a audiovizuálními příznaky bylo dosaženo pro bílý šum, který je díky svým vlastnostem v porovnání s jinými typy hluků relativně snadno odstranitelný spektrálním odečítáním. Naopak poměrně nestacionární hluk typu babble je potlačitelný obtížněji, a v těchto prostředích je tak přínos vizuální informace výraznější. Speech enhancement algoritmy mají také nevýhodu v tom, že pro vysoké SNR rozpoznávací skóre spíše snižují, zatímco při užití vhodně zvolených vah k tomuto u audiovizuálního rozpoznávání nedochází v takové míře. Rovněž byl opět potvrzen přínos hloubkových příznaků, které při kombinaci s obrazovými daty zvýšily slovní přesnost pro nejnižší SNR až o 10 %.



Obrázek 10.8: Audiovizuální rozpoznávání s příznaky HOGTOP v prostředí s hlukem typu babble.



Obrázek 10.9: Audiovizuální rozpoznávání s příznaky HOGTOP v prostředí s hlukem typu bílý šum.

11. Audiovizuální rozpoznávání spojité řeči

V kapitole 10 byly pro klasifikaci promluv aplikovány celoslovní skryté markovské modely, kde pro každou výslovnost každého slova (či fráze v případě databáze OuluVS) ze slovníku byl natrénován speciální model. Tento postup je však aplikovatelný pouze u systémů s řádově desítkami, max. stovkami slovníkových položek, zatímco rozpoznávání běžného jazyka vyžaduje řádově min. desítky (angličtina) či stovky a víc (čeština) tisíc slov. Pro natrénování robustních modelů jsou však důležité množství a variabilita trénovacích dat, jenž při samostatném modelu pro každou možnou výslovnost každého slova lze v systémech s velkým slovníkem jen obtížně zajistit. Mnohem výhodnější je naučit samostatný HMM pro nějakou menší řečovou jednotku, kterou různá slova mohou sdílet, a tím zásadně snížit počet potřebných modelů potažmo nároky na trénovací množinu. V systémech pro rozpoznávání spojité řeči s velkým slovníkem (Large Vocabulary Continuous Speech Recognition, LVCSR) se tak obvykle používají hláskové modely, nejčastěji s monofóny či trifóny jako základními řečovými jednotkami. Každé slovo ve slovníku má stále svůj vlastní HMM, ten je však sestavený z menších hláskových modelů jejich zřetězením. Jednou z výhod tohoto postupu je např. možnost přidat do slovníku i slovo, které se v trénovací množině vůbec nevyskytuje. Pro detaily o této problematice viz např. [Huang 2001, Nouza 2009].

11.1 Hláskové modely

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. Data obsahují celkem 40 fonémů z české fonetické abecedy PAC-CZ [Nouza 1997] a 13 vizémů pro češtinu vybraných Císařem [Císař 2006]. Přiřazení fonémů a vizémů i s příklady slov popisuje tabulka A.2 v příloze. 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 9.1. 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.

U většiny parametrizací byl aplikován postprocessing, který se ukázal jako optimální v předešlých experimentech s rozpoznáváním izolovaných slov celoslovními HMM. Pouze v některých případech se dodatečné zpracování lišilo: u HOGTOP bylo zkráceno okolí vstupující do redukce LDA na 7 snímků ($K = 3$) a podobně se postupovalo i u hloubkového AAM (DAAM, sekce 8.3) snížením na 9 snímků ($K = 4$). Z experimentů byla vyřazena PCA, jelikož její výsledky i typy chyb velmi silně korelovala s DCT a pro experimenty se tak zdála být redundantní.

Pro natrénování hláskových HMM byla opět využita knihovna HTK[1]

[1] `http://htk.eng.cam.ac.uk/`

[Young 2006] verze 3.4.1. Každý foném či vizém modeluje třístavový lineární HMM s n-komponentovou gaussovskou směsí (GMM) pro každý stav. Tato konfigurace HMM vyžaduje data s frekvencí vyšší než 30Hz, jinak je minimální rozlišovací schopnost pouze cca 0,1 s na hlásku (0,033 s na stav), což modelu neumožňuje dostatečně přesné zarovnání stavů. Vizuální parametrizace získaná z videí o frekvenci 30 snímků za sekundu proto byla lineárně interpolována na frekvenci shodnou se zvukem, tedy 100 Hz. Druhý možný způsob, tedy extrakce příznaků z již interpolovaných zdrojových dat, s ohledem na výpočetní a paměťovou náročnost vyzkoušen nebyl. Trénování modelů proběhlo ve dvou fázích: pomocí Viterbiho algoritmu (`HInit`), které slouží jako přípravný krok, a následnou reestimací Baum-Welchovým algoritmem (`HRest`). `HInit` i `HRest` využívají informaci o časové pozici jednotlivých framů, která pochází z nuceného zarovnání akustických dat. Jelikož byla vizuální a akustická data pro spojitou řeč vzájemně desynchronizována, musela být časová anotace pro použití s vizuální parametrizací posunuta o empiricky zjištěnou konstantu (cca 0,23 s). Tento problém byl zohledněn i u vektorového spojování brzké a střední integrace, kde byly vizuální příznaky posunuty o stejnou konstantu a zároveň normalizovány na délku shodnou se zvukem. U rozpoznávání izolovaných slov celoslovními modely se přitom především díky nižší vzorkovací frekvenci (30 Hz vs 100 Hz) problém příliš neprojevoval. Program `HERest`, který se obvykle používá pro iterativní reestimaci akustických modelů a který již nevyžaduje informaci o časovém zarovnání, nebyl pro trénování modelů použitý. Jeho aplikací došlo pouze ke zhoršení výsledků, protože samotná vizuální parametrizace nenese dostatek informace pro nucené zarovnání dat. Proces reestimace pomocí `HERest`u tak postupně divergoval a s každou další iterací se slovní přesnost výsledného modelu zhoršila.

11.2 Rozpoznávání izolovaných slov

Jako předstupeň před dekódováním spojité řeči se první experiment zaměřuje na stejnou úlohu jako kapitola 10, tedy rozpoznávání izolovaných slov – zde však hláskovými, nikoliv celoslovními modely. Promluvy jsou před klasifikací opět ořízntuty tak, aby na začátku i konci řečové části bylo jen krátké ticho. Nicméně pro 100Hz parametrizaci i poměrně krátký segment může zkomplikovat správné zarovnání stavů jednotlivých hlásek, a do rozpoznávací gramatiky proto byla na začátek i konec přidána nepovinná ticha (model „hlásky“ reprezentující ticho).

Výsledky experimentu v podobě slovní přesnosti [%] prezentuje tabulka 11.1, kde pro celoslovní modely byl počet gaussovských komponent každého stavu HMM křížově validován ze dvou možností (1 nebo 2 komponenty na stav), viz kapitolu 10. 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 %). Nejméně přechodem na hláskové modely trpí příznaky založené na AAM: např. pro kombinovaný hloubkový AAM (DAAM) rozdíl činí cca 13 %. Zatímco pro celoslovní modely se úspěšnost AAM a

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

odvozených příznaků pohybuje pod skóre dynamicky navržených parametrizací, pro rozpoznávání založené na menších než slovních jednotkách se zdají být vhodnější. Důvody, proč tomu tak je, nejsou zcela zřejmé a budou předmětem dalšího výzkumu.

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

Experimenty byly provedeny ve čtyřech různých variantách lišících se základní fonetickou jednotkou a počtem komponent gaussovské směsi na stav. Dle očekávání vyšší slovní přesnosti dosahují modely založené na vizémech, které lépe odpovídají rozlišitelnosti jednotlivých hlásek v obrazových datech (samozřejmě platí opačně pro MFCC extrahované z audio zdroje). Nicméně podobně jako v případě počtu komponent gaussovské směsi na stav není přínos nikterak výrazný, max. 2,5 %. Jeden z problémů představuje kontextová závislost, kdy podoba vizémů se výrazně mění s okolními hláskami [Owens 1985]. Např. obě ‘c’ ve slově „Cecil“ nepochybně patří do vizémové skupiny ‘C’ (souhlásky ‘c’, ‘s’, ‘z’) s charakteristicky viditelnými

dotýkajícími se horními a dolními zuby, zatímco podobu ‘c’ ve slově „uculit“ či ‘z’ a ‘s’ ve slově „zůstat“ určuje spíše sousedící samohláska ‘u’. Problematiku velmi detailně rozebral ve své dizertační práci Ramage [Ramage 2013], který z uvedeného a několika dalších důvodů vizémy jako základní jednotku vizuální řeči zpochybnil. Dle jeho práce vizémy nesplňují tři základní požadavky, kterými jsou vysoký poměr variability vně (mezi) a uvnitř vizémových tříd, snadná (vizuální) zaměnitelnost fonémů uvnitř stejné vizémové třídy a nesměrovost fonémových substitucí uvnitř stejné vizémové třídy. Zdá se přitom, že všechny nedostatky uvedené v Ramagově práci způsobuje především předpoklad surjektivity zobrazení fonémů na vizémy, neboli mapování typu $M : 1$. Díky kontextové závislosti totiž může jeden foném mít více vizuálních podob a správnou vizémovou transkripcí trénovacích a testovacích dat by je tedy mohlo být možné minimalizovat. Vzhledem k časové náročnosti manuální tvorby výslovností jsem však tuto hypotézu neověřoval. Kromě fundamentálních problémů, které uvedl Ramage, existují i další praktické. Jedním z nich je i nekompatibilita s akustickými řečovými dekodéry, které samozřejmě využívají fonémové (typicky trifónové) hláskové modely. Aplikace vizémů pro jinou než pozdní integraci pak není zřejmá. S ohledem na množství a závažnost problémů spojených s jejich aplikací pro rozpoznávání spojité řeči tedy byly pro další experimenty i pro vizuální příznaky využité hláskové modely s monofóny a 8 komponentami na stav.

11.3 Rozpoznávání spojité řeči

Na rozdíl od izolovaných slov, kde má dekódovací gramatika jednoznačně danou jednoduchou formu (jedno slovo, příp. nepovinné ticho na okrajích nahrávky), v úloze rozpoznávání spojité řeči s velkým slovníkem (Large Vocabulary Continuous Speech Recognition, LVCSR) se na vstupu může vyskytovat libovolné množství slov v jakémkoliv pořadí. Je však ale zřejmé, že většina kombinací slov v daném jazyce je z gramatického hlediska nesmyslná a na výstupu dekodéru by se neměla objevovat. Při LVCSR se proto využívá jazykový model, který zajišťuje, aby rozpoznaná řeč co nejlépe odpovídala přirozenému jazyku a na výstupu dekodéru se tak neobjevovaly nesmyslné slovní sekvence. Nejrozšířenější formou je tzv. n-gramový jazykový model, kde pravděpodobnost výskytu každého slova w_i je určena na základě $n-1$ předchozích slov $w_{i-1}, \ldots, w_{i-n+1}$. Typicky se používá $n = 2$ (bigramy) či $n = 3$ (trigramy), v nejmodernějších systémech i vyšší. Pravděpodobnost $p\left(w_i \,|\, w_{i-1}, \ldots, w_{i-n+1}\right)$ se určuje na základě četnosti $c(w_i, w_{i-1}, \ldots, w_{i-n+1})$ výskytu n-gramů v nějakém trénovacím korpusu, který kvalitně reprezentuje rozpoznávaný jazyk. Platí tedy

$$p\left(w_i \,|\, w_{i-1}, \ldots, w_{i-n+1}\right) = \frac{c(w_i, w_{i-1}, \ldots, w_{i-n+1})}{c(w_{i-1}, \ldots, w_{i-n+1})}. \tag{11.1}$$

Aby nedocházelo k situacím, kdy nedostatečnou velikostí korpusu budou mít některé n-gramy nulovou pravděpodobnost, obvykle se jazykové modely vyhlazují. Mezi známé metody vyhlazování patří např. aditivní, Wittenovo-Bellovo, Goodovo-

Jazykový model	min	5k	50k	500k
Velikost slovníku	366	5 182	50 056	499 993
Počet bigramů	48 338	9 865 378	73 904 857	141 670 479

Tabulka 11.2: Velikosti slovníku pro uvažované jazykové modely

Turingovo, či Knesserovo-Nayovo. Detailně rozebírá problematiku zpracování přirozeného jazyka např. kniha [Jurafsky 2009].

V této práci byly sestaveny celkem 4 různé bigramové (tedy $n = 2$) jazykové modely lišící se velikostí základního slovníku od několika set do několika set tisíc slov. Přesná čísla udává tabulka 11.2. 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[2] [Stolcke 2002] ve verzi 1.7.1 se zapnutým Knesserovým-Nayovým vyhlazováním.

Jak bylo popsáno v sekci 9.1.3, prvních 50 vět bylo pro všechny mluvčí společných. Tyto v experimentech s rozpoznáváním spojité řeči sloužily jako testovací množina. Zbylých 50 vět od testovacích mluvčích nebylo pro stejné rozdělení křížové validace nijak využito. 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[3] [Lee 2009], který podporuje akustické (vizuální) modely natrénované v HTK. Nastavení parametrů dekodéru popisuje tabulka A.3 v příloze, přičemž jako důležité se ukázalo především zvýšení koeficientů penalizace vkládání slov (angl. insertion penalty).

Výsledky rozpoznávání v závislosti na použitém jazykovém modelu shrnuje tabulka 11.3. Pro každý experiment jsou uvedena dvě čísla: slovní přesnost WAcc (7.1) a v závorce slovní správnost (Word Correctness)

$$\mathrm{WCorr} = \frac{N - D - S}{N}, \tag{11.2}$$

která na rozdíl od WAcc nezapočítává chybně vložená slova (inzerce) a její hodnota tedy vždy leží v intervalu $\langle 0, 1 \rangle$. Sloupec **Z** označuje zdroj (a ... audio, v ... video, d ... hloubka). Vcelku dle očekávání platí, že se vzrůstající velikostí jazykového modelu klesá jak slovní přesnost, tak správnost, a to i pro akustické příznaky MFCC poměrně výrazně: ze 74 % při 366 slovech ve slovníku na 36,3 % s půlmilionovým slovníkem, což odpovídá přibližně 150 % nárůstu slovní chybovosti. Čistě vizuální rozpoznávání především s většími slovníky zcela selhává. Pouze paramet-

[2] `http://www.speech.sri.com/projects/srilm/`
[3] `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)
DCT	v	-17,3	(9,3)	-5,4	(2,4)	-1,7	(1,1)	-0,5	(0,7)
	d	-2,6	(6,3)	-0,3	(2,3)	0,1	(0,5)	0,1	(0,3)
AAM	v	**12,3**	(16,2)	2,8	(3,5)	1,3	(1,6)	0,8	(1,0)
	d	9,3	(12,4)	1,8	(2,3)	0,8	(1,0)	0,5	(0,5)
LBPTOP	v	6,31	(17,8)	1,8	(4,2)	1,3	(2,0)	0,8	(1,1)
	d	-0,1	(10,8)	-0,3	(2,0)	0,1	(0,8)	0,2	(0,4)
DCT3	v	-17,3	(10,0)	-6,2	(2,8)	-2,0	(1,2)	-0,6	(0,8)
	d	-8,3	(7,9)	-2,3	(1,8)	-0,6	(0,7)	-0,2	(0,4)
HOGTOP	v	3,8	(21,3)	0,7	(6,3)	1,0	(2,9)	1,0	(1,7)
	d	4,1	(15,8)	1,0	(3,9)	1,0	(1,9)	0,7	(1,1)
DAAM	(v, d)	12,2	**(24,5)**	**4,0**	**(7,5)**	**2,5**	**(3,7)**	**1,6**	**(2,1)**

Tabulka 11.3: Unimodální rozpoznávání spojité řeči s různými slovníky a jazykovými modely. Výsledky jsou uvedeny v [%] ve formátu „slovní přesnost (slovní správnost)“.

rizace založené na AAM dosahují alespoň pro nejmenší možný slovník statisticky zajímavých hodnot WAcc.

Z důvodu velkého počtu inzercí, které dostávají hodnoty WAcc až do záporných hodnot, je v tabulce 11.3 uvedena pro každý experiment v závorce i slovní přesnost. Na tu lze nahlížet jako na schopnost klasifikátoru nalézt slova ve vstupním signálu bez ohledu na výslednou chybovost. Jedním ze závažných problémů pro vizuální rozpoznávání je totiž nevhodnost příznaků pro rozlišení mezi řečovými a neřečovými úseky a dekodér má tudíž tendenci na výstup vkládat velké množství krátkých slov pro každý jemný pohyb úst. Především ve spojení s akustickými příznaky však může být zajímavá spíše schopnost objevit slovo, pokud se ve vstupním signálu vyskytuje, již lépe vystihuje správnost WCorr, než za každou cenu maximalizovat přesnost WAcc. Z pohledu slovní správnosti WCorr výsledky v podstatě kopírují předešlé experimenty s rozpoznáváním izolovaných slov a nejlepších výsledků tak dosahují příznaky HOGTOP a DAAM (jež ovšem využívají i hloubková data).

Tabulka 11.4 uvádí výsledky ve stejné formě jako tab. 11.3 pro brzkou integraci akustických a vizuálních parametrizací. Ve sloupci udávajícím zdrojová data operace (x, y) označuje vektorové zřetězení příznaků z modalit x a y (brzkou integraci). Zde je spíše než správnost zajímavější slovní přesnost WAcc, která je vcelku stabilně nejvyšší pro příznaky LBPTOP. Zajímavých hodnot dosahují ještě parametrizace AAM, DAAM a HOGTOP, naopak selhávají DCT a DCT3, jež skóre pro všechny jazykové modely skóre výrazně zhoršují. U LBPTOP jako u jediných došlo

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)
DCT	(a, v)	37,7	(64,6)	19,8	(40,6)	13,4	(27,1)	10,8	(19,9)
	(a, d)	50,7	(69,0)	31,8	(45,8)	23,4	(32,6)	18,8	(24,8)
AAM	(a, v)	71,2	(81,4)	57,7	(66,6)	46,8	(53,5)	39,6	(44,5)
	(a, d)	71,6	(81,9)	57,8	(66,2)	47,1	(53,2)	40,0	(44,5)
LBPTOP	(a, v)	**75,0**	**(84,6)**	**60,6**	**(68,7)**	**49,5**	**(55,2)**	**42,4**	**(46,3)**
	(a, d)	72,1	(82,7)	56,5	(65,2)	46,0	(52,1)	38,7	(42,9)
DCT3	(a, v)	42,4	(67,2)	24,7	(43,7)	17,6	(30,2)	14,5	(22,6)
	(a, d)	46,0	(65,8)	27,8	(42,8)	20,2	(29,4)	16,2	(22,0)
HOGTOP	(a, v)	71,4	(83,8)	56,8	(67,7)	46,9	(53,9)	38,9	(44,8)
	(a, d)	68,5	(81,3)	53,5	(64,4)	42,8	(50,7)	36,4	(42,1)
DAAM	(a, v, d)	65,9	(79,4)	51,9	(63,0)	41,9	(50,0)	35,3	(41,2)

Tabulka 11.4: Audiovizuální rozpoznávání spojité řeči s brzkou integrací akustických a vizuálních parametrizací pro různé slovníky a jazykové modely. Výsledky jsou uvedeny v [%] ve formátu „slovní přesnost (slovní správnost)“.

integrací vizuální informace ke zvýšení WAcc i pro nejmenší jazykový model, u ostatních (AAM, DAAM a HOGTOP) pouze u větších. Ve všech případech je však zlepšení nevýrazné v rozmezí 1–6 %, což odpovídá 4–10% relativnímu snížení WER. Výsledky pro audio-hloubkovou kombinaci jsou mírně horší, nicméně se však ukazuje, že hloubková data nesou užitečnou informaci i pro rozpoznávání s velkým slovníkem. Brzká integrace audia a videa i hloubky zároveň vedla k poměrně výraznému poklesu slovní přesnosti, např. přes 13 % v případě HOGTOP, ale až 33 % v případě DCT3. Nejpravděpodobnější příčinou je počet oproti audiu méně kvalitních obrazově založených příznaků, jenž tvoří až 80 % velikosti výsledného vektoru (např. 39 a:MFCC, 78 v:HOGTOP-LDA-Δ, 78 d:HOGTOP-LDA-Δ) a při výpočtu výstupní stavové pravděpodobnosti v HMM (4.2) tak výrazně převažují.

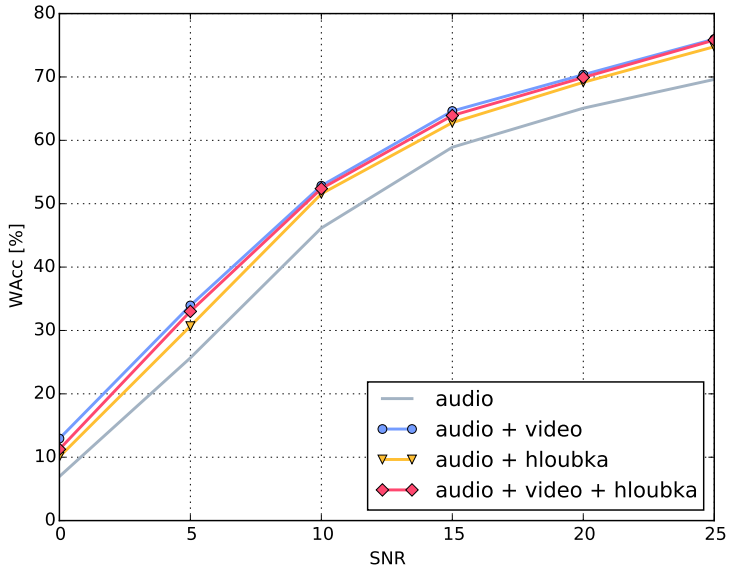
O něco lepší výsledky než brzká integrace podává střední fúze akustických a vizuálních příznaků, viz tabulku 11.5 pro vybrané typy parametrizací. Operace $[x, y]^{\lambda}$ zde značí střední fúzi příznakových vektorů z modalit x a y s vahami λ, jež byly křížově validovány tak, aby maximalizovaly slovní 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 (4.2) 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

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)

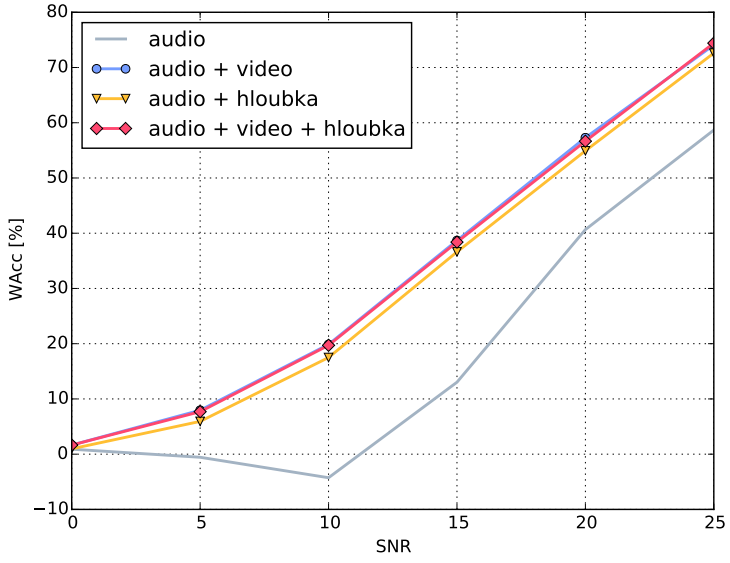
Tabulka 11.5: Audiovizuální rozpoznávání spojité řeči se střední fúzí akustických a vizuálních parametrizací pro různé slovníky a jazykové modely. Výsledky jsou uvedeny v [%] ve formátu „slovní přesnost (slovní správnost)“.

přitom lze interpretovat jako ukazatel významu a spolehlivosti jednotlivých kanálů. Pokles vizuálních vah $\lambda^{(v)} = 1 - \lambda^{(a)}$ 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 11.1 a 11.2 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 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.



Obrázek 11.1: Audiovizuální rozpoznávání spojité řeči se slovníkem o velikosti 366 slov za použití příznaků HOGTOP v prostředí s hlukem typu babble.



Obrázek 11.2: Audiovizuální rozpoznávání spojité řeči se slovníkem o velikosti 366 slov za použití příznaků HOGTOP v prostředí zarušeném bílým šumem.

12. 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 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 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 odladění 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. 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 %. Vybrané výsledky dosažené v této práci v úloze vizuálního rozpoznávání izolovaných jednotek porovnané se stavem poznání shrnuje tabulka 12.1.

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. 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

Par.	**TULAVD (IS)**	**OuluVS (F)**	**CUAVE (IČ)**
stav poznání	-	89,7	83,0
DCT	72,5	79,2	81,4
PCA	73,9	77,9	80,1
AAM	74,1	82,1	79,0
LBPTOP	74,2	82,5	**91,2**
DCT3	75,1	82,3	88,3
HOGTOP	**86,1**	85,7 (**89**, **9**†)	85,5

Tabulka 12.1: Slovní přesnost [%] rozpoznávání izolovaných slov (a frází) pomocí celoslovních modelů dosažená v této práci v porovnání se stavem poznání.

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.

12.1 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.

12.2 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 sub-slovní jednotku. Pro rozpoznávání by pak bylo možné využít vizémy, ovšem pouze za předpokladu správné vizémové transkripce, nikoliv jednosměrným přemapováním fonémů, viz sekci 11.1.

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

[Adjoudani 1996] A. Adjoudani a C. Benoît. *On the Integration of Auditory and Visual Parameters in an HMM-based ASR.* In DavidG. Stork a MarcusE. Hennecke, editeurs, Speechreading by Humans and Machines, volume 150 of *NATO ASI Series*, str. 461–471. Springer Berlin Heidelberg, 1996.

[Althoff 2003] Frank Althoff, Gregor Mcglaun, Manfred Lang, Gerhard Rigoll a Major Theme. *A real-time demonstrator for video-based recognition of dynamic head gestures, using discrete hidden Markov models*, 2003.

[Belhumeur 2011] Peter N. Belhumeur, David W. Jacobs, David J. Kriegman a Neeraj Kumar. *Localizing Parts of Faces Using a Consensus of Exemplars.* In The 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2011.

[Bengio 2009] Yoshua Bengio. *Learning Deep Architectures for AI.* Found. Trends Mach. Learn., vol. 2, no. 1, str. 1–127, January 2009.

[Bishop 2006] Christopher M. Bishop. Pattern recognition and machine learning (information science and statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[Blanz 2003] Volker Blanz a Thomas Vetter. *Face Recognition Based on Fitting a 3D Morphable Model.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 9, str. 1063–1074, 2003.

[Bowden 2013] Richard Bowden, Stephen Cox, Richard Harvey, Yuxuan Lan, Eng-Jon Ong, Gari Owen a Barry-John Theobald. *Recent developments in automated lip-reading*, 2013.

[Bregler 1994] Christoph Bregler a Yochai Konig. *"Eigenlips"for Robust Speech Recognition*, 1994.

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

[Chaloupka 2005] Josef Chaloupka. Rozpoznávání akustického signálu řeči s podporou vizuální informace. 2005.

[Chaloupka 2008] Josef Chaloupka, Jan Nouza a Jindrich Zdánský. *Audio-visual voice command recognition in noisy conditions.* In International Conference on Auditory-Visual Speech Processing 2008, Moreton Island, Queensland, Australia, September 26-29, 2008, str. 25–30, 2008.

[Chaloupka 2011] Josef Chaloupka. *Audio-Visual Isolated Words Recognition for Voice Dialogue System.* In Anna Esposito, Alessandro Vinciarelli, Klára Vicsi, Catherine Pelachaud a Anton Nijholt, editeurs, Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues, volume 6800 of *Lecture Notes in Computer Science*, str. 88–94. Springer Berlin Heidelberg, 2011.

[Chiţu 2012] A Chiţu a LJM Rothkrantz. *Automatic Visual Speech Recognition.* 2012.

[Conrey 2006] Brianna Conrey a David B Pisoni. *Auditory-visual speech perception and synchrony detection for speech and nonspeech signals*, 2006.

[Cooke 2006] Martin Cooke, Jon Barker, Stuart Cunningham a Xu Shao. *An audio-visual corpus for speech perception and automatic speech recognition.* The Journal of the Acoustical Society of America, vol. 120, no. 5, 2006.

[Cootes 1995] T. F. Cootes, C. J. Taylor, D. H. Cooper a J. Graham. *Active Shape Models&Mdash;Their Training and Application.* Comput. Vis. Image Underst., vol. 61, no. 1, str. 38–59, January 1995.

[Cootes 1998] Timothy F. Cootes, Gareth J. Edwards a Christopher J. Taylor. *Active Appearance Models.* In IEEE Transactions on Pattern Analysis and Machine Intelligence, str. 484–498. Springer, 1998.

[Cootes 2000] T.F. Cootes a C.J. Taylor. *Statistical Models of Appearance for Computer Vision*, 2000.

[Cox 2008] Stephen J. Cox, Richard Harvey, Yuxuan Lan, Jacob L. Newman a Barry-John Theobald. *The challenge of multispeaker lip-reading.* In AVSP, str. 179–184, 2008.

[Cristinacce 2006] David Cristinacce a Tim Cootes. *Feature detection and tracking with constrained local models.* str. 929–938, 2006.

[Císař 2006] Petr Císař. Využití metod odezírání ze rtů pro podporu rozpoznávání řeči. 2006.

[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.

[Dalal 2006] Navneet Dalal, Bill Triggs a Cordelia Schmid. *Human Detection Using Oriented Histograms of Flow and Appearance.* In Proceedings of the 9th European Conference on Computer Vision - Volume Part II, ECCV'06, str. 428–441, Berlin, Heidelberg, 2006. Springer-Verlag.

[Dantone 2012] M. Dantone, J. Gall, G. Fanelli a L. Van Gool. *Real-time Facial Feature Detection using Conditional Regression Forests.* In CVPR, 2012.

[Deligne 2002] S. Deligne, G. Potamianos a C. Neti. *Audio-visual speech enhancement with AVCDCN (audio-visual codebook dependent cepstral normalization).* In Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002, str. 68–71, Aug 2002.

[Dryden 1998] Ian L. Dryden a Kanti V. Mardia. Statistical shape analysis. Wiley series in probability and statistics. Wiley, Chichester [u.a.], 1998.

[Duchnowski 1994] P. Duchnowski a Uwe Meier and. *See Me, Hear Me: Integrating Automatic Speech Recognition and Lip-reading.* In Proceedings of the ICSLP 1994, 1994.

[Dupont 2000] S. Dupont a J. Luettin. *Audio-visual speech modeling for continuous speech recognition.* Multimedia, IEEE Transactions on, vol. 2, no. 3, str. 141–151, Sep 2000.

[Šeps 2014] Ladislav Šeps, Jiří Málek, Petr Červa a Jan Nouza. *Investigation of deep neural networks for robust recognition of nonlinearly distorted speech.* In INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014, str. 363–367, 2014.

[Estellers 2012a] V. Estellers, M. Gurban a J. Thiran. *On Dynamic Stream Weighting for Audio-Visual Speech Recognition.* Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 4, str. 1145–1157, May 2012.

[Estellers 2012b] Virginia Estellers a Jean-Philippe Thiran. *Multi-pose lipreading and audio-visual speech recognition.* EURASIP J. Adv. Sig. Proc., vol. 2012, str. 51, 2012.

[Freund 1997] Yoav Freund a Robert E. Schapire. *A Decision-theoretic Generalization of On-line Learning and an Application to Boosting.* J. Comput. Syst. Sci., vol. 55, no. 1, str. 119–139, August 1997.

[Fu 2008] Yun Fu, Shuicheng Yan a Thomas S. Huang. *Classification and Feature Extraction by Simplexization.* IEEE Transactions on Information Forensics and Security, vol. 3, no. 1, str. 91–100, 2008.

[Galatas 2011] Georgios Galatas, Gerasimos Potamianos, Dimitrios I. Kosmopoulos, Christopher McMurrough a Fillia Makedon. *Bilingual corpus for AVASR using multiple sensors and depth information.* In AVSP, str. 103–106, 2011.

[Galatas 2012] Georgios Galatas, Gerasimos Potamianos a Fillia Makedon. *Audio-visual speech recognition using depth information from the Kinect in noisy video conditions.* In Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments, PETRA '12, str. 2:1–2:4, New York, NY, USA, 2012. ACM.

[Garg 2003] A. Garg, G. Potamianos, C. Neti a T. S. Huang. *Frame-dependent Multi-stream Reliability Indicators for Audio-visual Speech Recognition.* In Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 3 (ICME '03) - Volume 03, ICME '03, str. 605–608, Washington, DC, USA, 2003. IEEE Computer Society.

[Glotin 2001] H. Glotin, D. Vergyr, C. Neti, G. Potamianos a J. Luettin. *Weighting schemes for audio-visual fusion in speech recognition.* In Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on, volume 1, str. 173–176 vol.1, 2001.

[Goecke 2002] Roland Goecke, Gerasimos Potamianos a Chalapathy Neti. *Noisy audio feature enhancement using audio-visual speech data.* In Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on, volume 2, str. II–2025–II–2028, May 2002.

[Goecke 2004] Roland Goecke a J Bruce Millar. *The Audio-Video Australian English Speech Data Corpus AVOZES.* In National University, str. 2525–2528, 2004.

[Goecke 2005] Roland Goecke. *3D Lip Tracking and Co-inertia Analysis for Improved Robustness of AudioVideo Automatic Speech Recognition.* In in Proceedings of the Auditory-Visual Speech Processing Workshop AVSP 2005, str. 109–114, 2005.

[Goecke 2008] Roland Goecke a Akshay Asthana. *A comparative study of 2d and 3d lip tracking methods for AV ASR.* In AVSP, str. 235–240, 2008.

[Grant 2001] Ken W. Grant a Steven Greenberg. *Speech intelligibility derived from asynchronous processing of auditory-visual information*, 2001.

[Grant 2004] Ken W. Grant, Virginie van Wassenhove a David Poeppel. *Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony.* Speech Communication, vol. 44, no. 1-4, str. 43–53, 2004.

[Gravier 2002] Guillaume Gravier, Scott Axelrod, Gerasimos Potamianos a Chalapathy Neti. *Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR.* In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2002, May 13-17 2002, Orlando, Florida, USA, str. 853–856, 2002.

[Gray 1996] Michael S. Gray, Javier R. Movellan a Terrence J. Sejnowski. *Dynamic Features for Visual Speechreading: A Systematic Comparison.* In Michael

Mozer, Michael I. Jordan a Thomas Petsche, editeurs, NIPS, str. 751–757. MIT Press, 1996.

[Gu 2008] Leon Gu a Takeo Kanade. *A Generative Shape Regularization Model for Robust Face Alignment.* In Proceedings of the 10th European Conference on Computer Vision: Part I, ECCV '08, str. 413–426, Berlin, Heidelberg, 2008. Springer-Verlag.

[Gurbuz 2002] Sabri Gurbuz, Z. Tufekci, E. Patterson a J.N. Gowdy. *Multi-stream product modal audio-visual integration strategy for robust adaptive speech recognition.* In Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on, volume 2, str. II–2021–II–2024, May 2002.

[Hazen 2004] Timothy J. Hazen, Kate Saenko, Chia-Hao La a James R. Glass. *A Segment-based Audio-visual Speech Recognizer: Data Collection, Development, and Initial Experiments.* In Proceedings of the 6th International Conference on Multimodal Interfaces, ICMI '04, str. 235–242, New York, NY, USA, 2004. ACM.

[Heckmann 2001] Martin Heckmann, Frédéric Berthommier a Kristian Kroschel. *A hybrid ANN/HMM audio-visual speech recognition system.* In Auditory-Visual Speech Processing, AVSP 2001, Aalborg, Denmark, September 7-9, 2001, str. 189–194, 2001.

[Heckmann 2002a] Martin Heckmann, Frédéric Berthommier a Kristian Kroschel. *Noise Adaptive Stream Weighting in Audio-Visual Speech Recognition.* EURASIP J. Adv. Sig. Proc., vol. 2002, no. 11, str. 1260–1273, 2002.

[Heckmann 2002b] Martin Heckmann, Kristian Kroschel, Christophe Savariaux, Fr d ric Berthommier a Universit t Karlsruhe. *DCT-based video features for audio-visual speech recognition.* In in 'International Conf. on Spoken Language Processing, 2002.

[Heigold 2012] G. Heigold, H. Ney, R. Schluter a S. Wiesler. *Discriminative Training for Automatic Speech Recognition: Modeling, Criteria, Optimization, Implementation, and Performance.* Signal Processing Magazine, IEEE, vol. 29, no. 6, str. 58–69, Nov 2012.

[Hinton 2006] G E Hinton a R R Salakhutdinov. *Reducing the dimensionality of data with neural networks.* Science, vol. 313, no. 5786, str. 504–507, July 2006.

[Huang 2001] Xuedong Huang, Alex Acero a Hsiao-Wuen Hon. Spoken language processing: A guide to theory, algorithm, and system development. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st édition, 2001.

[Huang 2004] Jing Huang, Gerasimos Potamianos, Jonathan Connell a Chalapathy Neti. *Audio-visual speech recognition using an infrared headset.* Speech Communication, vol. 44, no. 1-4, str. 83–96, 2004.

[Huang 2013] Jing Huang a B. Kingsbury. *Audio-visual deep learning for noise robust speech recognition.* In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, str. 7596–7599, May 2013.

[Iwano 2007] Koji Iwano, Tomoaki Yoshinaga, Satoshi Tamura a Sadaoki Furui. *Audio-Visual Speech Recognition Using Lip Information Extracted from Side-Face Images.* EURASIP J. Audio, Speech and Music Processing, vol. 2007, 2007.

[Jourlin 1997] Pierre Jourlin. *Word-dependent acoustic-labial weights in HMM-based speech recognition.* In ESCA Workshop on Audio-Visual Speech Processing, AVSP '97, Rhodes, Greece, September 26-27, 1997, str. 69–72, 1997.

[Ju 2013] Jeongwoo Ju, Heechul Jung a Junmo Kim. *Speaker Dependent Visual Speech Recognition by Symbol and Real Value Assignment.* In Jong-Hwan Kim, Eric T. Matson, Hyun Myung a Peter Xu, editeurs, Robot Intelligence Technology and Applications 2012, volume 208 of *Advances in Intelligent Systems and Computing*, str. 1015–1022. Springer Berlin Heidelberg, 2013.

[Jurafsky 2009] Daniel Jurafsky a James H. Martin. Speech and language processing (2nd edition). Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009.

[Kaucic 1998] R. Kaucic a A Blake. *Accurate, real-time, unadorned lip tracking.* In Computer Vision, 1998. Sixth International Conference on, str. 370–375, Jan 1998.

[Kazemi 2014] Vahid Kazemi a Josephine Sullivan. *One Millisecond Face Alignment with an Ensemble of Regression Trees.* In CVPR, 2014.

[Kittler 1998] J. Kittler, M. Hatef, R.P.W. Duin a J. Matas. *On combining classifiers.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 20, no. 3, str. 226–239, Mar 1998.

[Kjeldsen 1996] R. Kjeldsen a J. Kender. *Finding skin in color images.* In Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on, str. 312–317, Oct 1996.

[Kolossa 2009] Dorothea Kolossa, Steffen Zeiler, Alexander Vorwerk a Reinhold Orglmeister. *Audiovisual speech recognition with missing or unreliable data.* In Auditory-Visual Speech Processing, AVSP 2009, Norwich, UK, September 10-13, 2009, str. 117–122, 2009.

[Kratt 2004] Jan Kratt, Florian Metze, Rainer Stiefelhagen a Alex Waibel. *Large Vocabulary Audio-Visual Speech Recognition Using the Janus Speech Recognition Toolkit.* In CarlEdward Rasmussen, HeinrichH. Bülthoff, Bernhard Schölkopf a MartinA. Giese, editeurs, Pattern Recognition, volume 3175 of *Lecture Notes in Computer Science*, str. 488–495. Springer Berlin Heidelberg, 2004.

[Kricke 2008] Ralph Kricke, Thorsten Gernoth a Rolf-Rainer Grigat. *Local binary patterns for lip motion analysis.* In Proceedings of ICIP 2008 - 15th IEEE International Conference on Image Processing, str. 1472–1475. IEEE, 2008.

[Kumar 2007] Kshitiz Kumar, Tsuhan Chen a Richard M. Stern. *Profile View Lip Reading.* In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007, Honolulu, Hawaii, USA, April 15-20, 2007, str. 429–432, 2007.

[Lan 2009] Yuxuan Lan, Richard Harvey, Barry-John Theobald, Eng-Jon Ong a Richard Bowden. *Comparing Visual Features for Lipreading.* AVSP-2009, str. 102–106, 2009.

[Lan 2010] Y. Lan, B. Theobald, R. Harvey a R. Bowden. *Improving Visual Features for Lip-reading.* Proceedings of the International Conference on Auditory-Visual Speech Processing 2010, str. 142–147, 2010.

[Lan 2012] Yuxuan Lan, Barry-John Theobald a Richard Harvey. *View Independent Computer Lip-Reading.* In Proceedings of the 2012 IEEE International Conference on Multimedia and Expo, ICME 2012, Melbourne, Australia, July 9-13, 2012, str. 432–437, 2012.

[Lee 2004] Bowon Lee, Mark Hasegawa-Johnson, Camille Goudeseune, Suketu Kamdar, Sarah Borys, Ming Liu a Thomas S. Huang. *AVICAR: audio-visual speech corpus in a car environment.* In INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004, 2004.

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

[Li 1995] Nan Li, Shawn Dettmer a Mubarak Shah. *Lipreading Using Eigensequences.* In In Proc. of Workshop on Automatic Face and Gesture Recognition, str. 30–34, 1995.

[Li 2002] Stan Z. Li, Long Zhu, ZhenQiu Zhang, Andrew Blake, HongJiang Zhang a Harry Shum. *Statistical Learning of Multi-view Face Detection.* In Proceedings of the 7th European Conference on Computer Vision-Part IV, ECCV '02, str. 67–81, London, UK, UK, 2002. Springer-Verlag.

[Lienhart 2002] R. Lienhart a J. Maydt. *An extended set of Haar-like features for rapid object detection.* In Image Processing. 2002. Proceedings. 2002 International Conference on, volume 1, str. I–900–I–903 vol.1, 2002.

[Lievin 1998] M. Lievin a F. Luthon. *Lip features automatic extraction.* In Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on, str. 168–172 vol.3, Oct 1998.

[Liu 2007] Xiaoming Liu. *Generic Face Alignment using Boosted Appearance Model.* In 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA, 2007.

[Loizou 2013] Philipos C. Loizou. Speech enhancement: Theory and practice. CRC Press, Inc., Boca Raton, FL, USA, 2nd édition, 2013.

[Lowe 2004] David G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints.* Int. J. Comput. Vision, vol. 60, no. 2, str. 91–110, November 2004.

[Loy 2000] Gareth Loy, Roland Goecke, Sebastien Rougeaux, Alexander Zelinsky a Er Zelinsky. *Stereo 3D Lip Tracking.* In in '6th International Conference on Control, Automation, Robotics and Vision, 2000.

[Lucey 2005] Simon Lucey, Tsuhan Chen, Sridha Sridharan a Vinod Chandran. *Integration strategies for audio-visual speech processing: applied to text-dependent speaker recognition.* IEEE Transactions on Multimedia, vol. 7, no. 3, str. 495–506, 2005.

[Lucey 2006a] Patrick Lucey a Gerasimos Potamianos. *Lipreading Using Profile Versus Frontal Views.* In IEEE 8th Workshop on Multimedia Signal Processing, MMSP 2006, Victoria, BC, Canada, October 3-6, 2006, str. 24–28, 2006.

[Lucey 2006b] Patrick Lucey a Sridha Sridharan. *Patch-based Representation of Visual Speech.* In Proceedings of the HCSNet Workshop on Use of Vision in Human-computer Interaction - Volume 56, VisHCI '06, str. 79–85, Darlinghurst, Australia, Australia, 2006. Australian Computer Society, Inc.

[Lucey 2007] Patrick Lucey, Gerasimos Potamianos a Sridha Sridharan. *A unified approach to multi-pose audio-visual ASR.* In INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007, str. 650–653, 2007.

[Lucey 2008] Patrick Lucey, Sridha Sridharan a David Dean. *Continuous pose-invariant lipreading.* In INTERSPEECH 2008, 9th Annual Conference of

the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008, str. 2679–2682, 2008.

[Marcheret 2007] E. Marcheret, V. Libal a G. Potamianos. *Dynamic Stream Weight Modeling for Audio-Visual Speech Recognition*. In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, volume 4, str. IV–945–IV–948, April 2007.

[Mase 1991] Kenji Mase a Alex Pentland. *Automatic lipreading by optical-flow analysis*. Systems and Computers in Japan, vol. 22, no. 6, str. 67–76, 1991.

[Matthews 2001] I. Matthews, G. Potamianos, C. Neti a J. Luettin. *A comparison of model and transform-based visual features for audio-visual LVCSR*. In Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on, str. 825–828, Aug 2001.

[Matthews 2002] Iain Matthews, Tim Cootes, J. Andrew Bangham, Stephen Cox a Richard Harvey. *Extraction of Visual Features for Lipreading*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, str. 2002, 2002.

[Matthews 2003] Iain Matthews a Simon Baker. *Active Appearance Models Revisited*. International Journal of Computer Vision, vol. 60, str. 135–164, 2003.

[McGurk 1976] Harry McGurk a John MacDonald. *Hearing lips and seeing voices*. Nature, vol. 264, str. 746–748, 12 1976.

[Meier 1996] U. Meier, W. Hurst a P. Duchnowski. *Adaptive bimodal sensor fusion for automatic speechreading*. In Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, volume 2, str. 833–836 vol. 2, May 1996.

[Messer 1999] K. Messer, J. Matas, J. Kittler, J. Lüttin a G. Maitre. *XM2VTSDB: The Extended M2VTS Database*. In In Second International Conference on Audio and Video-based Biometric Person Authentication, str. 72–77, 1999.

[Nakamura 2000] Satoshi Nakamura, Hidetoshi Ito a Kiyohiro Shikano. *Stream weight optimization of speech and lip image sequence for audio-visual speech recognition*. In Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000, Beijing, China, October 16-20, 2000, str. 20–24, 2000.

[Nefian 2002] Ara V. Nefian, Luhong Liang, Xiaobo Pi, Xiaoxing Liu a Kevin Murphy. *Dynamic Bayesian Networks for Audio-visual Speech Recognition*. EURASIP J. Appl. Signal Process., vol. 2002, no. 1, str. 1274–1288, January 2002.

[Neti 2000] Chalapathy Neti, Gerasimos Potamianos, Juergen Luettin, Iain Matthews, Herve Glotin, Dimitra Vergyri, June Sison, Azad Mashari a Jie Zhou. *Audio-Visual Speech Recognition.* Rapport technique, The Johns Hopkins University, Baltimore, MD, Final Workshop 2000 Report, 2000.

[Ngiam 2011] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee a Andrew Y. Ng. *Multimodal Deep Learning.* In Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011, str. 689–696, 2011.

[Nouza 1997] J. Nouza, J. Psutka a J. Uhlíř. *Phonetic Alphabet for Speech Recognition of Czech*, 1997.

[Nouza 2009] J. Nouza, Z. Koldovský, R. Vích a Technická univerzita v Liberci. Ústav informačních technologií a elektroniky. Řeč a počítač: principy hlasové komunikace, úlohy, metody a aplikace : sborník článků. Technická univerzita v Liberci, 2009.

[Nouza 2014] Jan Nouza, Petr Červa, Jindřich Žďánský, Karel Blavka, Marek Boháč, Jan Silovský, Josef Chaloupka, Michaela Kuchařová, Ladislav Šeps, Jiří Málek a Michal Rott. *Speech-to-text technology to transcribe and disclose 100, 000+ hours of bilingual documents from historical Czech and Czechoslovak radio archive.* In INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014, str. 964–968, 2014.

[Obdržálek 2006] Štěpán Obdržálek a Jiří Matas. *Object Recognition Using Local Affine Frames on Maximally Stable Extremal Regions.* In Jean Ponce, Martial Hebert, Cordelia Schmid a Andrew Zisserman, editeurs, Toward Category-Level Object Recognition, volume 4170 of *Lecture Notes in Computer Science*, str. 83–104. Springer Berlin Heidelberg, 2006.

[Ong 2011] Eng-Jon Ong a Richard Bowden. *Learning Sequential Patterns for Lipreading.* In British Machine Vision Conference, BMVC 2011, Dundee, UK, August 29 - September 2, 2011. Proceedings, str. 1–10, 2011.

[Owens 1985] E. Owens a B. Blazek. *Visemes observed by hearing-impaired and normal-hearing adult viewers.* J Speech Hear Res, vol. 28, no. 3, str. 381–393, Sep 1985.

[Ozuysal 2010] M. Ozuysal, M. Calonder, V. Lepetit a P. Fua. *Fast Keypoint Recognition Using Random Ferns.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 32, no. 3, str. 448–461, 2010.

[Pachoud 2008] Samuel Pachoud, Shaogang Gong a Andrea Cavallaro. *Macro-cuboïd based probabilistic matching for lip-reading digits.* In 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA, 2008.

[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.

[Pass 2010] Adrian Pass, Jianguo Zhang a Darryl Stewart. *An investigation into features for multi-view lipreading.* In Proceedings of the International Conference on Image Processing, ICIP 2010, September 26-29, Hong Kong, China, str. 2417–2420, 2010.

[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.

[Petr Císař 2004] Miloš Železný Petr Císař a Zdeněk Krňoul. *3D lip-tracking for audio-visual speech recognition in real applications.* Journal of the Acoustical Society of Korea, str. 2521–2524, 2004.

[Pigeon 1997] Stéphane Pigeon a Luc Vandendorpe. *The M2VTS multimodal face database (Release 1.00).* In Josef Bigün, Gérard Chollet a Gunilla Borgefors, editeurs, Audio- and Video-based Biometric Person Authentication, volume 1206 of *Lecture Notes in Computer Science*, str. 403–409. Springer Berlin Heidelberg, 1997.

[Pitsikalis 2006] Vassilis Pitsikalis, Athanassios Katsamanis, George Papandreou a Petros Maragos. *Adaptive multimodal fusion by uncertainty compensation.* In INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006, 2006.

[Potamianos 1998a] G. Potamianos a H.P. Graf. *Discriminative training of HMM stream exponents for audio-visual speech recognition.* In Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, volume 6, str. 3733–3736 vol.6, May 1998.

[Potamianos 1998b] G. Potamianos, H.P. Graf a E. Cosatto. *An image transform approach for HMM based automatic lipreading.* In Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on, str. 173–177 vol.3, Oct 1998.

[Potamianos 2001a] Gerasimos Potamianos a Chalapathy Neti. *Automatic speechreading of impaired speech.* In Auditory-Visual Speech Processing, AVSP 2001, Aalborg, Denmark, September 7-9, 2001, str. 177–182, 2001.

[Potamianos 2001b] Gerasimos Potamianos, Chalapathy Neti, Giridharan Iyengar, Andrew W. Senior a Ashish Verma. *A Cascade Visual Front End for Speaker Independent Automatic Speechreading.* International Journal of Speech Technology, vol. 4, str. 2001, 2001.

[Potamianos 2003] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg a Andrew W. Senior. *Recent advances in the automatic recognition of audio-visual speech.* In PROC. IEEE, str. 1306–1326, 2003.

[Potamianos 2004] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin a Iain Matthews. *Audio-visual automatic speech recognition: An overview*, 2004.

[Puviarasan 2011] N. Puviarasan a S. Palanivel. *Lip Reading of Hearing Impaired Persons Using HMM.* Expert Syst. Appl., vol. 38, no. 4, str. 4477–4481, April 2011.

[Radová 1999] Vlasta Radová a Petr Vopálka. *Methods of Sentences Selection for Read-Speech Corpus Design.* In Václav Matousek, Pavel Mautner, Jana Ocelíková a Petr Sojka, editeurs, Text, Speech and Dialogue, volume 1692 of *Lecture Notes in Computer Science*, str. 165–170. Springer Berlin Heidelberg, 1999.

[Ramage 2013] Matthew David Ramage. Disproving visemes as the basic visual unit of speech. 2013.

[Rekik 2014a] Ahmed Rekik, Achraf Ben-Hamadou a Walid Mahdi. *Face pose tracking under arbitrary illumination changes.* In Computer Vision Theory and Applications (VISAPP), 2014 International Conference on, volume 3, str. 570–575, Jan 2014.

[Rekik 2014b] Ahmed Rekik, Achraf Ben-Hamadou a Walid Mahdi. *A New Visual Speech Recognition Approach for RGB-D Cameras.* In Aur lio Campilho a Mohamed Kamel, editeurs, Image Analysis and Recognition, Lecture Notes in Computer Science, str. 21–28. Springer International Publishing, 2014.

[Richter 2014] Matthias Richter, Hua Gao a Hazim Kemal Ekenel. *Extending explicit shape regression with mixed feature channels and pose priors.* In IEEE Winter Conference on Applications of Computer Vision, Steamboat Springs, CO, USA, March 24-26, 2014, str. 1013–1019, 2014.

[Saenko 2005] Kate Saenko, Karen Livescu, Michael Siracusa, Kevin Wilson, James Glass a Trevor Darrell. *Visual Speech Recognition with Loosely Synchronized Feature Streams.* In Proceedings of the Tenth IEEE International Conference on Computer Vision - Volume 2, ICCV '05, str. 1424–1431, Washington, DC, USA, 2005. IEEE Computer Society.

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

[Saitoh 2010] Takeshi Saitoh a Ryosuke Konishi. *Profile Lip Reading for Vowel and Word Recognition.* In 20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23-26 August 2010, str. 1356–1359, 2010.

[Sanderson 2002] Conrad Sanderson. *The VidTIMIT Database.* Idiap-Com Idiap-Com-06-2002, IDIAP, 0 2002.

[Saragih 2007] Jason Saragih a Roland Göcke. *A Nonlinear Discriminative Approach to AAM Fitting.* In IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007, str. 1–8, 2007.

[Saragih 2011] Jason M. Saragih, Simon Lucey a Jeffrey F. Cohn. *Deformable Model Fitting by Regularized Landmark Mean-Shift.* Int. J. Comput. Vision, vol. 91, no. 2, str. 200–215, January 2011.

[Savchenko 2014] A.V. Savchenko a Ya.I. Khokhlova. *About neural-network algorithms application in viseme classification problem with face video in audiovisual speech recognition systems.* Optical Memory and Neural Networks, vol. 23, no. 1, str. 34–42, 2014.

[Scanlon 2003] Patricia Scanlon, Richard B. Reilly a Philip de Chazal. *Visual feature analysis for automatic speechreading.* In AVSP 2003 - International Conference on Audio-Visual Speech Processing, St. Jorioz, France, September 4-7, 2003, str. 127–132, 2003.

[Scanlon 2004] Patricia Scanlon, Gerasimos Potamianos, Vit Libal a Stephen M. Chu. *Mutual Information Based Visual Feature Selection for Lipreading.* In in Proc. of ICSLP 2004, South Korea, str. 4–8, 2004.

[Shaikh 2010] A.A. Shaikh, D.K. Kumar, W.C. Yau, M.Z. Che Azemin a J. Gubbi. *Lip reading using optical flow and support vector machines.* In Image and Signal Processing (CISP), 2010 3rd International Congress on, volume 1, str. 327–330, Oct 2010.

[Shao 2008] Xu Shao a Jon Barker. *Stream weight estimation for multistream audio visual speech recognition in a multispeaker environment.* Speech Communication, vol. 50, no. 4, str. 337 – 353, 2008.

[Shin 2011] Jongju Shin, Jin Lee a Daijin Kim. *Real-time Lip Reading System for Isolated Korean Word Recognition.* Pattern Recogn., vol. 44, no. 3, str. 559–571, March 2011.

[Stewart 2014] D. Stewart, R. Seymour, A. Pass a Ji Ming. *Robust Audio-Visual Speech Recognition Under Noisy Audio-Video Conditions.* Cybernetics, IEEE Transactions on, vol. 44, no. 2, str. 175–184, Feb 2014.

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

[Summerfield 1987] Q. Summerfield. *Some preliminaries to a comprehensive account of audio-visual speech perception.* In Dodd, editeur, Hearing by eye: The psychology of lip-reading. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1987.

[Tamura 2004] Satoshi Tamura, Koji Iwano a Sadaoki Furui. *Multi-Modal Speech Recognition Using Optical-Flow Analysis for Lip Images.* J. VLSI Signal Process. Syst., vol. 36, no. 2/3, str. 117–124, February 2004.

[Teissier 1999] P. Teissier, J. Robert-Ribes, J.-L. Schwartz a A. Guerin-Dugue. *Comparing models for audiovisual fusion in a noisy-vowel recognition task.* Speech and Audio Processing, IEEE Transactions on, vol. 7, no. 6, str. 629–642, Nov 1999.

[Turk 1991] Matthew Turk a Alex Pentland. *Eigenfaces for recognition.* J. Cognitive Neuroscience, vol. 3, no. 1, str. 71–86, January 1991.

[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.

[Vorwerk 2010] Alexander Vorwerk, Xiaohui Wang, Dorothea Kolossa, Steffen Zeiler a Reinhold Orglmeister. *WAPUSK20 - A Database for Robust Audiovisual Speech Recognition.* In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta, 2010.

[Wang 2008a] S.L. Wang, A. Liew, W.H. Lau a S.H. Leung. *An Automatic Lipreading System for Spoken Digits With Limited Training Data.* Circuits and Systems for Video Technology, IEEE Transactions on, vol. 18, no. 12, str. 1760–1765, Dec 2008.

[Wang 2008b] Yang Wang, Simon Lucey a Jeffrey Cohn. *Enforcing Convexity for Improved Alignment with Constrained Local Models.* In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 2008.

[Xiao 2004] Jing Xiao, Simon Baker, Iain Matthews a Takeo Kanade. *Real-Time Combined 2D+3D Active Appearance Models.* In CVPR (2), str. 535–542, 2004.

[Xiong 2013] Xuehan Xiong a Fernando De la Torre. *Supervised Descent Method and Its Applications to Face Alignment.* In 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013, str. 532–539, 2013.

[Yang 1998] Ming-Hsuan Yang a N. Ahuja. *Detecting human faces in color images*. In Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on, volume 1, str. 127–130 vol.1, Oct 1998.

[Yargic 2013] A Yargic a M. Dogan. *A lip reading application on MS Kinect camera*. In Innovations in Intelligent Systems and Applications (INISTA), 2013 IEEE International Symposium on, str. 1–5, June 2013.

[Yoshinaga 2003] Tomoaki Yoshinaga, Satoshi Tamura, Koji Iwano a Sadaoki Furui. *Audio-Visual Speech Recognition Using Lip Movement Extracted from Side-Face Images*. In PROC. AVSP2003, ST JORIOZ, str. 117–120, 2003.

[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.

[Yu 1999] Keren Yu, Xiaoyi Jiang a Horst Bunke. *Lipreading Using Signal Analysis over Time*. Signal Process., vol. 77, no. 2, str. 195–208, September 1999.

[Zhang 2010] Cha Zhang a Zhengyou Zhang. *A Survey of Recent Advances in Face Detection*. Rapport technique MSR-TR-2010-66, June 2010.

[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.

[Zhou 2010] Ziheng Zhou, Guoying Zhao a M. Pietikainen. *Lipreading: A Graph Embedding Approach*. In Pattern Recognition (ICPR), 2010 20th International Conference on, str. 523–526, Aug 2010.

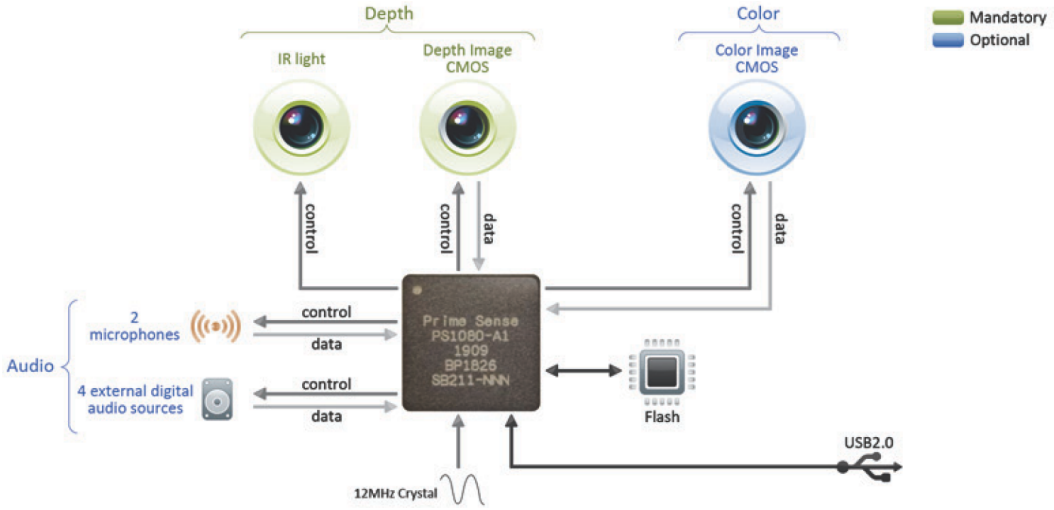
[Zhou 2011] Ziheng Zhou, Guoying Zhao a M. Pietikainen. *Towards a Practical Lipreading System*. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '11, str. 137–144, Washington, DC, USA, 2011. IEEE Computer Society.

[Zhou 2014] Ziheng Zhou, Xiaopeng Hong, Guoying Zhao a Matti Pietikäinen. *A Compact Representation of Visual Speech Data Using Latent Variables*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 1, str. 1, 2014.

A. Příloha

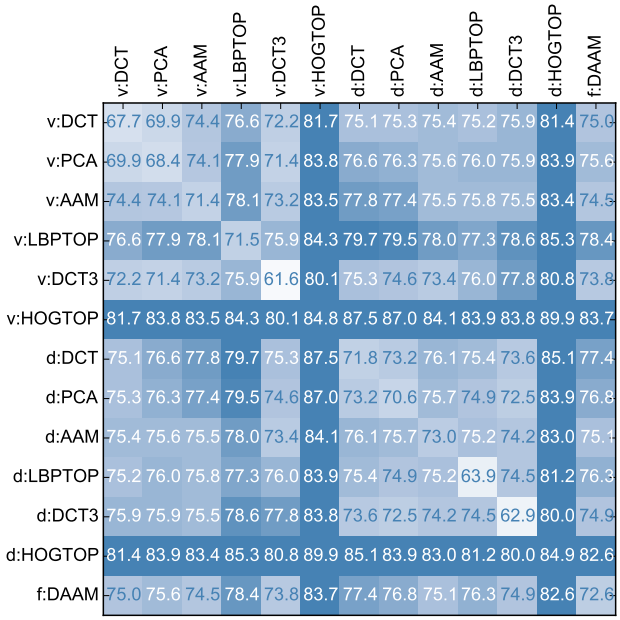
1	protože	26	rychle
2	všechno	27	kterého
3	počítač	28	abychom
4	stejně	29	zkoušky
5	můžete	30	získáte
6	opravdu	31	zejména
7	skutečně	32	naučil
8	například	33	našich
9	poslední	34	najednou
10	obvinění	35	českých
11	peníze	36	studenti
12	dokonce	37	ředitel
13	trochu	38	poprvé
14	situace	39	krátce
15	problém	40	kostel
16	několik	41	konečně
17	vzpomínám	42	děvčata
18	tisíce	43	žebřík
19	taková	44	zůstat
20	prostředí	45	zřejmě
21	prezident	46	znamená
22	hodiny	47	tradice
23	hlavní	48	státní
24	doktor	49	sportovní
25	řízení	50	skladatel

Tabulka A.1: Seznam izolovaných slov na databázi TULAVD.

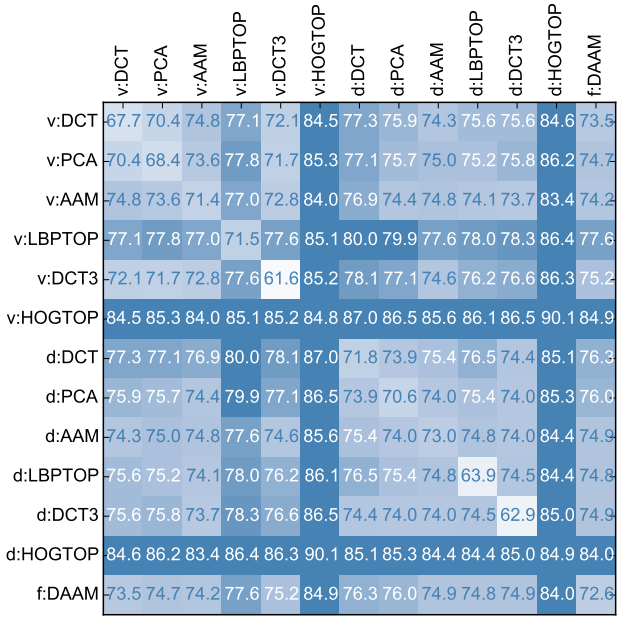


Obrázek A.1: PrimeSense specifikace, podle které je navržen Microsoft Kinect. Převzato z `http://www.souvr.com/Soft/UploadSoft/201005/2010050617295050.pdf`.

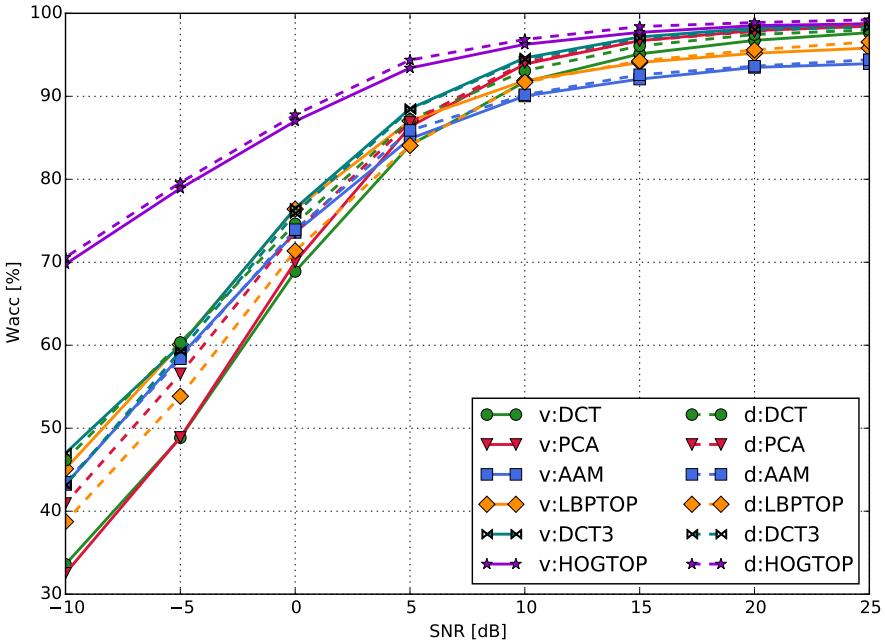
Obrázek A.2: Mluvčí při nahrávání databáze TULAVD.



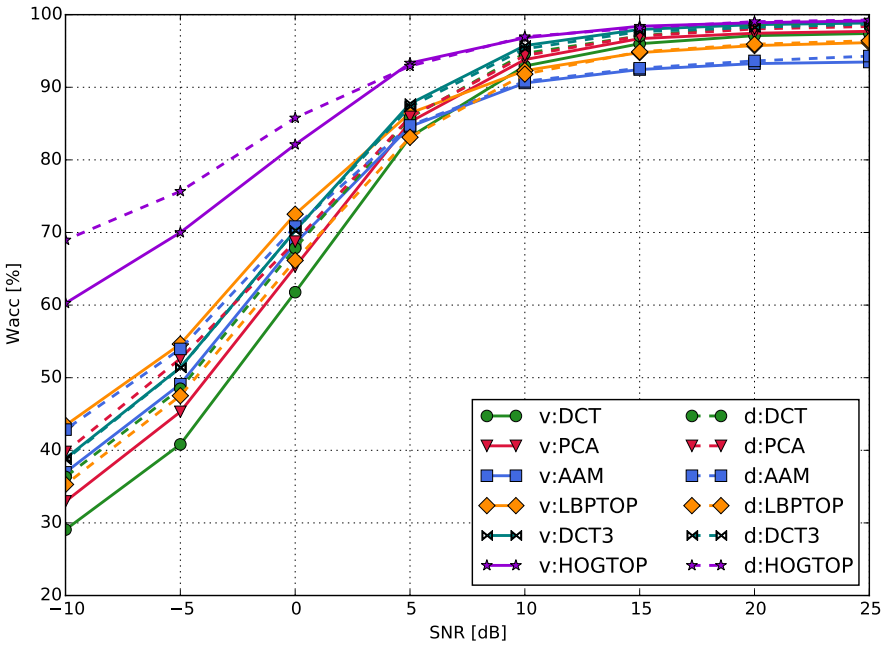
Obrázek A.3: Slovní přesnost [%] pro všechny kombinace příznaků na databázi TULAVD: brzká integrace.



Obrázek A.4: Slovní přesnost [%] pro všechny kombinace příznaků na databázi TULAVD: střední fúze.



Obrázek A.5: Audiovizuální rozpoznávání izolovaných slov na databázi TULAVD v prostředí zahlučeném bílým šumem pro různé vizuální příznaky.



Obrázek A.6: Audiovizuální rozpoznávání izolovaných slov na databázi TULAVD v prostředí s továrním hlukem pro různé vizuální příznaky.

No.	foném	PAC-CZ	příklad	transkr.	vizém
1	„a“	a	táta	tát**a**	A
2	„á“	á	táta	t**á**ta	A
3	„b“	b	bába	**b**ába	B
4	„c“	c	ocel	o**c**el	C
5	„dz“	C	leckde	le**C**gde	C
6	„č“	č	čichá	**č**iXá	Č
7	„dž“	Č	rádža	rá**Č**a	Č
8	„d“	d	jeden	je**d**en	D
9	„ď“	ď	dělat	**ď**elat	Ď
10	„e“	e	lev	l**e**f	E
11	„é“	é	méně	m**é**ňe	E
12	„f“	f	fauna	**f**auna	F
13	„g“	g	guma	**g**uma	G
14	„h“	h	aha	a**h**a	G
15	„ch“	X	chudý	**X**udí	G
16	„i“, „y“	i	bil, byl	b**i**l	I
17	„í“, „ý“	í	vítr, lýko	v**í**tr, l**í**ko	I
18	„j“	j	dojat	do**j**at	Ď
19	„k“	k	kupec	**k**upec	G
20	„l“	l	dělá	ď**e**lá	L
21	„m“	m	máma	**m**á**m**a	B
22	„m“	M	tramvaj	tra**M**vaj	B
23	„n“	n	víno	ví**n**o	D
24	„n“	N	banka	ba**N**ka	D
25	„ň“	ň	koně	ko**ň**e	Ď
26	„o“	o	kolo	k**o**l**o**	O
27	„ó“	ó	óda	**ó**da	O
28	„p“	p	pupen	**p**u**p**en	B
29	„r“	r	bere	be**r**e	L
30	„ř“	ř	moře	mo**ř**e	Č
31	„ř“	Ř	keř	ke**Ř**	Č
32	„s“	s	sud	**s**ut	C
33	„š“	š	duše	du**š**e	Č
34	„t“	t	dutý	du**t**í	D
35	„ť“	ť	kutil	ku**ť**il	Ď
36	„u“	u	duše	d**u**še	U
37	„ú“, „ů“	ú	růže	r**ú**že	U
38	„v“	v	láva	lá**v**a	F
39	„z“	z	koza	ko**z**a	C
40	„ž“	ž	růže	rú**ž**e	Č

Tabulka A.2: Tabulka fonémů z české fonetické abecedy PAC-CZ [Nouza 1997] a jim odpovídající vizémy dle [Císař 2006].

Název nastavení	**Parametr**	**Hodnota**
trellis beam width	`-b`	2000
score pruning thres	`-bs`	disabled
search candidate num	`-n`	100
search stack size	`-s`	500
search overflow	`-m`	2000
pass2 beam width	`-b2`	30
lookup range	`-lookuprange`	5
2nd scan beamthres	`-sb`	80.0
search till	`-n`	100
pass1 LM weight (insertion penalty)	`-lmp`	4.0 (-10.0)
pass2 LM weight (insertion penalty)	`-lmp2`	3.0 (-8.0)

Tabulka A.3: Nastavení dekodéru Julius pro rozpoznávání spojité řeči s velkým slovníkem. Oproti výchozím hodnotám bylo důležité především zvýšení koeficientů penalizace vkládání slov (poslední dva řádky).