TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta mechatroniky a mezioborových inženýrských studií

ROZPOZNÁVÁNÍ AKUSTICKÉHO SIGNÁLU ŘEČI S PODPOROU VIZUÁLNÍ INFORMACE

DISERTAČNÍ PRÁCE

2005 **JOSEF CHALOUPKA**

Rozpoznávání akustického signálu řeči s podporou vizuální informace

Disertační práce

Disertant:	Ing. Josef Chaloupka
Studijní program:	2612V Elektrotechnika a informatika
Studijní obor:	2612V045 Technická kybernetika
Pracoviště:	Katedra elektrotechniky a zpracování signálů Fakulta mechatroniky a mezioborových inženýrských studií Technická Univerzita v Liberci
Školitel:	Prof. Ing. Jan Nouza, CSc.

Rozsah práce a příloh

Počet stran:	120
Počet obrázků:	66
Počet tabulek:	20
Počet vzorců:	101
Počet příloh:	7

Anotace

Rozpoznávání akustického signálu řeči s podporou vizuální informace

Disertační práce

Ing. Josef Chaloupka

Tato disertační práce pojednává o rozpoznávání akustického signálu řeči s podporou vizuální informace, neboli o audio-vizuálním zpracování a rozpoznávání řeči. Vzhledem ke stanovenému tématu se jedná o práci interdisciplinární, zasahující do více oborů.

V kapitole č.1 a č.2 je představena problematika audio-vizuálního zpracování a rozpoznávání řeči.

Kapitola č.3 pojednává o předzpracování, segmentaci, parametrizaci a rozpoznávání samostatného akustického signálu řeči metodou skrytých Markovových modelů. V této kapitole jsou uvedeny pouze vybrané části z oblasti zpracování a rozpoznávání akustického signálu řeči, které mají úzký vztah k následným testům pro audio-vizuální rozpoznávání izolovaných slov, popsaných v kapitole č. 8.

V kapitolách č.4 až č.6 jsou popsány metody a algoritmy pro vytvoření systému pro parametrizaci vizuálního signálu řeči. V kapitole č.4 jsou uvedeny metody a navržený algoritmus pro detekování lidského obličeje v obraze. Kapitola č.5 pojednává o metodách nalezení oblasti zájmu se rty v detekované oblasti obličeje a v kapitole č.6 jsou uvedeny vlastní metody pro vytvoření vizuálních příznaků z nalezené oblasti zájmu.

Kapitola č.7 popisuje vytvoření a zpracování audio-vizuální databáze, která je nezbytně nutná pro následující experimenty při audio-vizuálním rozpoznávání řeči.

V kapitole č.8 jsou popsány experimentální práce, týkající se audio-vizuálního rozpoznávání izolovaných slov, včetně srovnání se samostatným rozpoznáváním akustického a vizuálního signálu řeči. Také jsou zde uvedeny výsledky testů pro audio-vizuální rozpoznávání izolovaných slov v hlučných podmínkách.

Kapitola č.9 pojednává o dalších aplikačních možnostech využití audio-vizuálního zpracování a rozpoznávání řeči.

V závěrečné kapitole č.10 jsou popsány dosažené výsledky disertační práce a je naznačen směr, kam by se mohl další výzkum v oblasti audio-vizuálního zpracování a rozpoznávání řeči ubírat.

Annotation

Speech Recognition of the Acoustic Speech Signal Supported by Visual Information

Dissertation thesis

Ing. Josef Chaloupka

This dissertation thesis deals with the recognition of the acoustic speech signal supported by the visual information, or with the audio-visual speech processing and recognition. In regard of the given theme it is an interdisciplinary work touching more scientific domains.

The problems of audio-visual speech processing and recognition are introduced in Chapters 1 and 2.

Chapter 3 contains the details about processing, parametrization, and recognition of a single acoustic speech signal by the method of the hidden Markov models. In this chapter only chosen parts from the area of the acoustic speech processing and recognition are presented that have a close relationship to the resulting tests for audio-visual speech recognition of the isolated words. These tests are described in Chapter 8.

The methods and algorithms for the creation of the system for the visual speech signal parametrization are described in Chapters 4, 5, and 6. The designed algorithms for human face detection are stated in Chapter 4. Chapter 5 deals with the methods of finding the region of interest containing the lips, and the methods of the separation of the visual speech features from the region of interest are introduced in Chapter 6.

Chapter 7 describes the creation and processing of the audio-visual speech database that is necessary for the following experiments with the audio-visual speech recognition.

The experimental work is presented in Chapter 8. This experimental work concernes the audio-visual speech recognition of the isolated words including the comparison with the recognition of the sole acoustic and the sole visual speech signal. This chapter introduces also the results of the test of the audio-visual speech recognition in the noisy conditions.

Chapter 9 deals with the other possibilities of utilizing of the audio-visual speech processing and recognition.

The attained results of this dissertation thesis are described in the concluding Chapter 10. The trend of the further research in the area of the audio-visual processing and speech recognition is indicated here.

Obsah

Kapitola 1

Úvod

Audio-vizuální počítačové zpracování a rozpoznávání řeči patří v současné době k intenzivně se rozvíjející aplikační oblasti moderních hlasových technologií. Do nedávné doby stála tato oblast, vzhledem ke zpracování a rozpoznávání samotného akustického signálu řeči, v pozadí zájmu výzkumných týmů. V posledních deseti letech však začaly ve světě vznikat větší výzkumné týmy zabývající se primárně zpracováním a rozpoznáváním vizuální složky řeči. Tento trend byl mimo jiné způsoben nástupem spolehlivých, levných a dostatečně rychlých osobních počítačů. Zpracování a rozpoznávání vizuálního signálu řeči je totiž řádově několikanásobně časově náročnější, než je tomu u zpracování a rozpoznávání akustického signálu řeči. Pro příklad, kdyby akustický signál řeči byl digitalizován vzorkovací frekvencí 8 kHz a velikost jednoho vzorku by byla 16 bitů, tak by akustický signál o délce 1s měl 128 000 bitů, oproti tomu pro příslušný vizuální signál o snímkovací frekvenci 30 snímků za sekundu, kde jeden barevný video snímek by měl velikost 640 x 480 obrazových bodů a obrazovému bodu by příslušela RGB barevná hodnota (24 bitů), tak by jednosekundový vizuální signál řeči měl již 221 184 000 bitů, což je 1728 x více vzhledem k akustickému signálu řeči.

Oblast audio-vizuálního zpracování a rozpoznávání řeči lze rozdělit na dva směry výzkumu. V první podoblasti výzkumu jsou vytvářeny systémy audio-vizuální syntézy řeči. V těchto systémech je použit modul převodu psaného textu na akustický signál řeči TTS (Text To Speech), ke kterému je připojen modul, nejčastěji reprezentovaný 3D počítačovým modelem mluvící hlavy, u které je při promluvě animována tvář, tj. dochází u tohoto modelu k pohybu čelistí, jazyka, rtů, mimických svalů atd. Vlastní výzkum je v této oblasti zaměřen především na vytvoření kvalitního vizuálního 3D modelu, který by při promluvě co nejvíce odpovídal reálnému mluvčímu. V druhé podoblasti výzkumu jsou vytvářeny metody a algoritmy pro zpracování, parametrizaci a rozpoznávání vizuálního signálu řeči. Tato disertační práce je zaměřena především na tuto druhou podoblast audio-vizuálního zpracování a rozpoznávání řeči. Obě tyto podoblasti výzkumu mají celou řadu aplikačních možností, některé z nich jsou popsány v kapitole 9.

Prvním předpokladem pro vytvoření systému pro audio-vizuální rozpoznávání řeči je vytvoření audio-vizuální databáze videonahrávek promluv od různých mluvčích. Tato audio-vizuální databáze musí být navržena pro konkrétní národní jazyk, pro který je následně vytvořen systém automatického rozpoznávání řeči. Vytvořené metody a algoritmy pro předzpracování a parametrizaci vizuálního signálu řeči jsou však využitelné pro jakýkoliv jazyk, obdobně jako je tomu u zpracování a rozpoznávání samostatného akustického signálu řeči.

Pro záznam audio-vizuálního signálu řeči se v současné době (2005) nejčastěji používají digitální kamery zaznamenávající barevný obraz a příslušný akustický signál. Akustický signál řeči by bylo možné zaznamenávat i samostatně pomocí mikrofonu, poté by se však musela zajistit synchronizace mezi akustickým a vizuálním signálem, což není úplně triviální úloha, proto se spíše pro záznam akustického signálu využívá mikrofon integrovaný v kameře.

1.1 Současný stav výzkumu problematiky

Výzkum v oblasti audio-vizuálního rozpoznávání řeči je již teoreticky řešen více než 20 let, přesto teprve přibližně v posledních deseti letech vznikaly ve světě větší výzkumné týmy, které se touto oblastí zabývají, a to jak teoreticky, tak i prakticky. Zřejmě nejznámějším týmem je laboratoř IBM - Audio Visual Speech Technologies vedená Dr. Chalapathy Neti. Tato laboratoř se zabývá především audio-vizuálním zpracováním a rozpoznáváním řeči pro anglický jazyk, i když algoritmy navržené touto laboratoří pro parametrizaci a rozpoznávání audio-vizuálního signálu řeči jsou za určitých podmínek použitelné i pro jiné národní jazyky. V poslední době jsou v oblasti audio-vizuálního rozpoznávání řeči dělány pokusy zaměřené na off-line rozpoznávání slov z velkého slovníku [POT01c], rozpoznávání spojité řeči [POT04b] a on-line rozpoznávání slov z malého slovníku [CON03]. Tento výzkum si však zatím mohou dovolit pouze laboratoře, které mají rozsáhlejší a kvalitní audio-vizuální databázi promluv. Kromě anglického jazyka je vývoj a výzkum v oblasti audio-vizuálního rozpoznávání řešen i pro francouzštinu [OBR97], němčinu [KRO97], japonštinu [NAK00] a další jazyky technologicky vyspělých zemí. V České republice se ve větší míře kromě Laboratoře počítačového zpracování řeči na TUL zabývá problematikou audio-vizuálního zpracování a rozpoznáváním řeči pro český jazyk i tým oddělení umělé inteligence na Katedře kybernetiky Fakulty aplikovaných věd Západočeské univerzity v Plzni. Pro větší přehlednost této práce je další podrobný popis současného stavu v jednotlivých dílčích úlohách audio-vizuálního zpracování a rozpoznávání řeči uveden na začátku kapitol 2 – 7 a v kapitole 9.

1.2 Cíle disertační práce

Cíle této disertační práce následující:

- Navrhnout, vytvořit a anotovat dostatečně reprezentativní audio-vizuální databázi videonahrávek promluv pro český jazyk. V této databázi by se měly nacházet nahrávky slov i vět od různých mluvčích. Pro zpracování nahrávek z této databáze vytvořit vhodné nástroje (programy).
- Navrhnout a vytvořit systém pro parametrizaci vizuálního signálu řeči, který by byl složen z podsystémů pro detekci lidské tváře v obraze, pro nalezení rtů v detekované oblasti zájmu (tváře) a podsystému pro vlastní parametrizaci oblasti rtů. Podsystémy pro detekování obličeje, nalezení rtů a vizuální parametrizaci by měly být pokud možno výpočetně co nejrychlejší.
- Navrhnout vhodnou fúzi parametrizovaného akustického a vizuálního signálu řeči a provést experimentální otestování audio-vizuálního rozpoznávání izolovaných slov, při srovnání se samostatným rozpoznáváním akustického a vizuálního signálu řeči. Navrhnout, realizovat a experimentálně vyhodnotit test pro audio-vizuální rozpoznávání izolovaných slov v podmínkách proměnného hlučného pozadí.

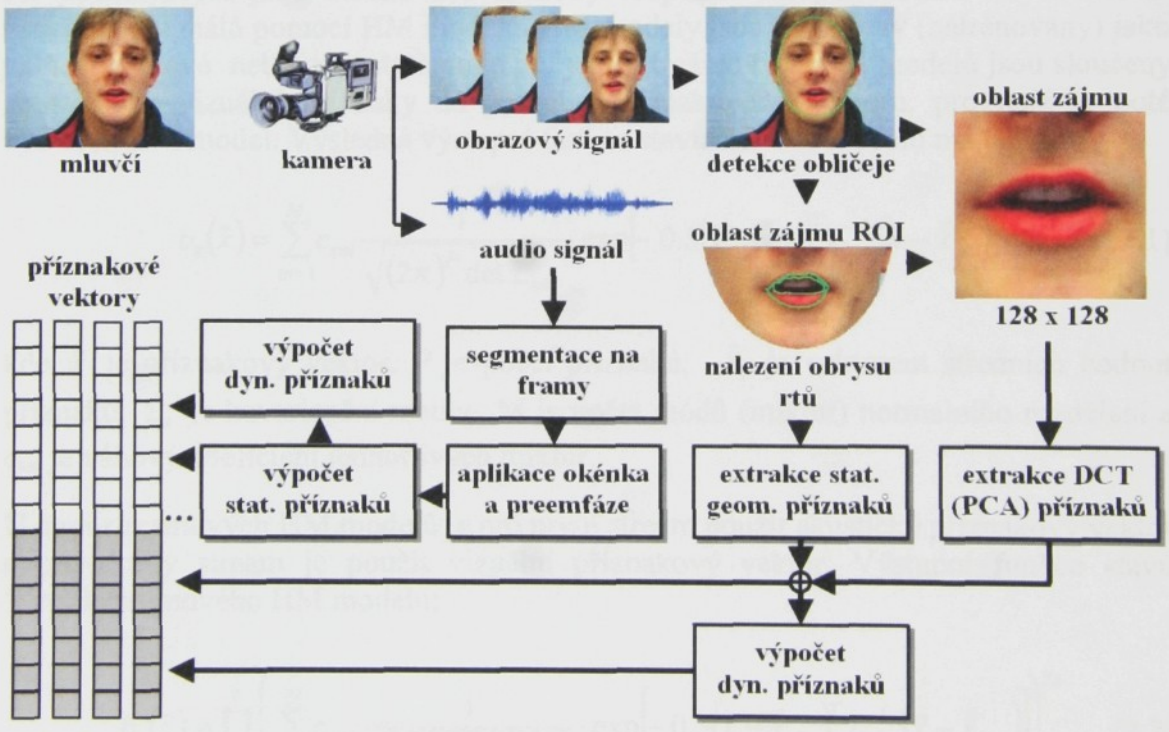
Kapitola 2

Audio-vizuální rozpoznávání řeči

Úloha rozpoznávání audio-vizuálního signálu řeči se skládá ze dvou částí. V prvním kroku je audio-vizuální signál předzpracován a parametrizován a v druhém kroku probíhá vlastní rozpoznávání.

2.1 Parametrizace audio-vizuálního signálu řeči

Pořízený audio-vizuální signál je rozdělen na akustický a vizuální signál a každý z těchto signálů je následně parametrizován. Parametrizace akustického signálu je již v dnešní době poměrně dobře vyřešena. Jako akustické příznaky se nejčastěji používají příznaky získané z kepstra akustického signálu (kapitola 3).



Obr. 2.1: Princip parametrizace audio-vizuálního signálu

Vizuální signál je složen z časového sledu obrazů (2D signálů), ve kterých jsou zaznamenání obličeje mluvčích. Pro parametrizaci vizuálního signálu se nejčastěji používají tvarové (geometrické) vizuální příznaky získané z nalezeného tvaru rtů nebo vizuální příznaky (DCT, PCA, ...) popisující informační obsah oblasti zájmu, ve které se

nacházejí rty a jejich bezprostřední okolí (kapitola 6). Pro vytvoření vizuálních příznaků je tak nutné nejdříve v obraze nalézt oblast zájmu se rty. Přímé nalezení rtů v obraze by bylo velmi složité, proto je v pořízeném obraze nejdříve detekován obličej mluvčího (kapitola 4) a z detekovaného obrazu obličeje je separována oblast zájmu se rty (kapitola 5), z níž jsou následně extrahovány vizuální příznaky. V některých dříve publikovaných pracích byla snímací kamera zaměřena přímo na oblast rtů mluvčího, nemusely se tak provádět operace pro detekování obličeje a nalezení obrazu oblasti zájmu se rty. Toto zjednodušení je však použitelné pouze v laboratorních podmínkách, kde se mluvčí příliš nepohybuje, a dnes se již příliš nepoužívá.

2.2 Rozpoznávání audio-vizuálního signálu řeči

Rozpoznávání parametrizovaného audio-vizuálního signálu řeči je dnes nejčastěji realizováno pomocí skrytých Markovových modelů (HMM – Hidden Markov Models) nebo pomocí umělých neuronových sítí (ANN – Artificial Neural Networks). Ve své práci jsem použil metodu HMM (kapitola 3) s využitím programu HTK [STE97]. Program HTK mimo jiné umožňuje vytvářet (natrénovat) HM modely a rozpoznávat parametrizované signály za využití těchto HM modelů.

2.2.1 Fúze a rozpoznávání audio-vizuálních signálů pomocí HTK

Při použití HTK jsou možné dva způsoby rozpoznávání parametrizovaných audio-vizuálních signálů pomocí HM modelů. HM modely jsou vytvořeny (natrénovány) jako jednostreamové nebo dvoustreamové[1]. U jednostreamových HM modelů jsou sloučeny akustické a vizuální příznaky do jednoho příznakového vektoru, pro nějž je poté vytvořen HM model. Výsledná výstupní funkce stavu S u HM modelu má podobu[2]:

$$b_s(\vec{x}) = \sum_{m=1}^{M} c_{sm} \frac{1}{\sqrt{(2\pi)^P \det \Sigma_{sm}}} \cdot \exp\left[-0.5\left(\vec{x} - \vec{\bar{x}}_{sm}\right)^T \Sigma_{sm}^{-1}\left(\vec{x} - \vec{\bar{x}}_{sm}\right)\right] \qquad (2.1)$$

kde $\vec{x}$ je příznakový vektor, P je počet příznaků, $\vec{\bar{x}}_s$ je vektorem středních hodnot příznaků, Σ_s je kovarianční matice, M je počet módů (mixtur) normálního rozdělení a c_{sm} je váhový koeficient jednotlivých mixtur.

U dvoustreamových HM modelů je pro první stream použit akustický příznakový vektor a pro druhý stream je použit vizuální příznakový vektor. Výstupní funkce stavu S dvoustreamového HM modelu:

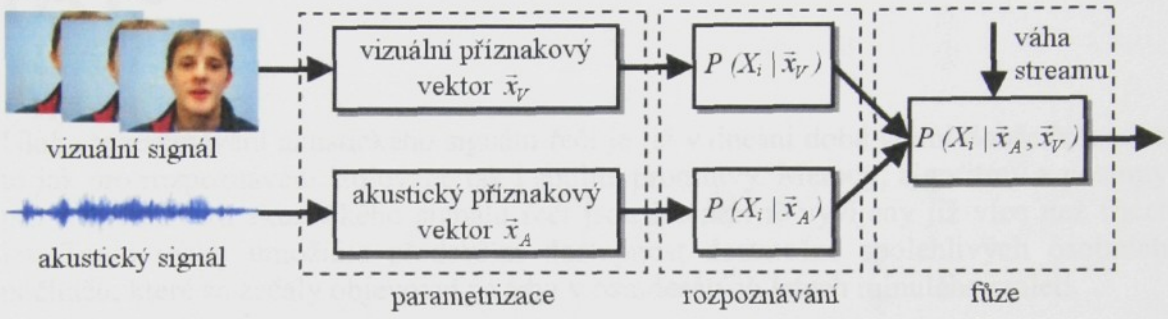
$$b_s(\vec{x}) = \prod_{t=1}^{T} \left(\sum_{m=1}^{M} c_{stm} \frac{1}{\sqrt{(2\pi)^P \det \Sigma_{stm}}} \cdot \exp\left[-0.5\left(\vec{x} - \vec{\bar{x}}_{stm}\right)^T \Sigma_{stm}^{-1}\left(\vec{x} - \vec{\bar{x}}_{stm}\right)\right] \right)^{\gamma_t} \qquad (2.2)$$

kde T je počet streamů ($T = 2$) a γ_t je váhový koeficient jednotlivých streamů, pro $T = 1$ a $\gamma_1 = 1$ (jednostreamový HM model) je vztah 2.2 identický se vztahem 2.1.

1) Tok akustických nebo vizuálních dat je dále označován převzatým anglickým slovem stream.
2) Struktura a definice HM modelů jsou popsány v kapitole 3.

2.2.1 Další způsoby fúze při rozpoznávání audio-vizuálních signálů

Do dnešní doby vniklo a bylo publikováno několik metod pro fúzi a rozpoznávání audio-vizuálního parametrizovaného signálu. Nejčastěji používanou metodou při využití HMM nebo ANN je metoda, kdy je akustický i vizuální signál rozpoznáván samostatně. Pro každý signál se poté počítá (odhaduje) pravděpodobnost $P(X_i \mid \vec{x}_A)$, $P(X_i \mid \vec{x}_V)$ určující, nakolik odpovídá dané slovo X_i ze slovníku akustickému $\vec{x}_A$ nebo vizuálním $\vec{x}_V$ příznakovému vektoru. Z těchto dvou pravděpodobností je na základě fúze určena výsledná pravděpodobnost $P(X_i \mid \vec{x}_A, \vec{x}_V)$.



Obr. 2.2: *Fúze a rozpoznávání akustického a vizuálního signálu*

Při fúzi se je možné jednotlivé pravděpodobnosti vynásobit váhovými koeficienty (2.5) nebo váhové koeficienty nepoužívat (2.3).

$$P(X_i \mid \vec{x}_A, \vec{x}_V) = \frac{P(X_i \mid \vec{x}_A) \cdot P(X_i \mid \vec{x}_V)}{P(X_i)} \cdot \eta \tag{2.3}$$

kde η je normalizační faktor počítaný z N slov z použitého slovníku pro rozpoznávání:

$$\eta = \frac{1}{\sum_{j=1}^{N} \frac{P(X_j \mid \vec{x}_A) \cdot P(X_j \mid \vec{x}_V)}{P(X_j)}} \tag{2.4}$$

$$P(X_i \mid \vec{x}_A, \vec{x}_V) = \frac{P^{\gamma}(X_i \mid \vec{x}_A) \cdot P^{1-\gamma}(X_i \mid \vec{x}_V)}{\sum_{j=1}^{N} P^{\gamma}(X_j \mid \vec{x}_A) \cdot P^{1-\gamma}(X_j \mid \vec{x}_V)} \tag{2.5}$$

kde γ je váhový faktor, který se obvykle volí 0 až 1. Někdy se používají i složitější fúzní vztahy, např. fúze s geometrickým vážením:

$$P(X_i \mid \vec{x}_A, \vec{x}_V) = \frac{P^{\alpha}(X_i \mid \vec{x}_A) \cdot P^{\beta}(X_i \mid \vec{x}_V)}{P^{\alpha+\beta-1}(X_i)} \cdot \varepsilon(\alpha, \beta) \tag{2.6}$$

kde α a β jsou váhové koeficienty a $\epsilon(\alpha, \beta)$ je normalizační faktor:

$$\varepsilon(\alpha, \beta) = \frac{1}{\sum_{j=1}^{N} \frac{P^{\alpha}(X_j \mid \vec{x}_A) \cdot P^{\beta}(X_j \mid \vec{x}_V)}{P^{\alpha+\beta-1}(X_j)}} \tag{2.7}$$

Kapitola 3

Rozpoznávání akustického signálu řeči

Úloha rozpoznávání akustického signálu řeči je již v dnešní době velmi dobře řešena, a to jak pro rozpoznávání izolované tak i spojité promluvy. Metody, algoritmy a postupy pro rozpoznávání akustického signálu řeči jsou intenzivně vyvíjeny již více než třicet let. Tento vývoj umožnila především dostupnost dostatečně spolehlivých osobních počítačů, které se začaly objevovat na trhu v osmdesátých letech minulého století.

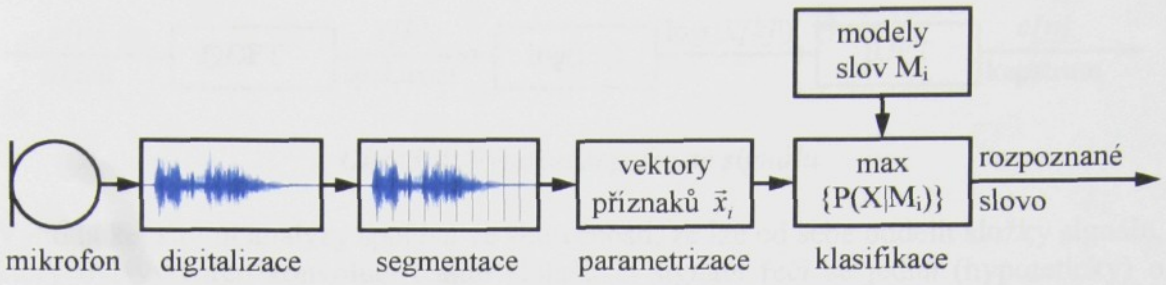
V úloze rozpoznávání izolovaných slov s velkým slovníkem (10 000 až 1 000 000 slov) se dnes již dosahuje rozpoznávacích skóre přes 90% a v úloze rozpoznávání spojité řeči je rozpoznávací skóre větší než 80%[1].

Ve své práci se chci zabývat především audio-vizuálním rozpoznáváním řeči pro český jazyk. Pro audio-vizuální rozpoznávání izolovaných slov a spojité promluvy založené na modelech menších stavebních jednotek řeči (fonémy, vizémy) je však potřeba vytvořit především vhodné modely českých vizémů. Nalezením vhodných českých vizémů se již zabývám a naše audio-vizuální databáze byla pořízená a zpracovávána s tímto záměrem (kapitola 7). Přesto největších pokroků jsem zatím dosáhl při audio-vizuálním rozpoznávání izolovaných slov založených na celoslovních modelech. Tato úloha vedla k nalezení a extrakci vizuálních příznaků a nalezení metodologie pro vlastní audio-vizuální rozpoznávání řeči.

V této kapitole je proto popsána strategie rozpoznávání řeči především pro akustické rozpoznávání izolovaných slov založená na celoslovních modelech. Zde popsaná digitalizace, segmentace a parametrizace akustického signálu řeči je však obdobná i pro rozpoznávání spojité řeči. O metodách pro rozpoznávání izolovaných slov a spojité řeči založené na modelech menších stavebních jednotek řeči se lze dočíst např. v [HUA01, PSU95, KOL01].

Na obrázku 3.1 je zobrazen princip rozpoznávání akustického signálu řeči. Nejprve je analogový signál z mikrofonu digitalizován, poté je digitalizovaný signál segmentován na menší segmenty (framy), každý fram je parametrizován a nakonec je provedena klasifikace, kde je sled parametrizovaných framů porovnáván s jednotlivými modely slov a na základě předem daného kritéria je rozhodnuto, který model slova patří ke vstupnímu signálu řeči.

1) Tyto hodnoty platí pro promluvu s dobrou výslovností v nehlučném prostředí.



Obr. 3.1: *Jednotlivé metody pro rozpoznávání akustického signálu řeči*

3.1 Digitalizace, segmentace a parametrizace signálu řeči

3.1.1 Digitalizace akustického signálu řeči

Vzhledem k technickému a vědeckému pokroku posledních desetiletí je digitalizace akustického signálu řeči (vzorkování a kvantování) v současné době již velmi dobře vyřešena a digitalizování signálu řeči dnes zajistí „každá" zvuková karta v PC. Z různých dřívějších pokusů se signály řeči bylo zjištěno, že k porozumění obsahu digitalizovaného signálu mluvené řeči postačuje frekvenční pásmo 0 až přibližně 3,4 kHz. Z Shannonova vzorkovacího teorému pak plyne, že pro digitalizaci signálu řeči by měla být vzorkovací frekvence Fs větší než 6,8 kHz. Pro účely rozpoznávání řeči se nejčastěji volí vzorkovací frekvence Fs ≥8 kHz a počet úrovní kvantování se obvykle volí 2^{16} [PSU95, HUA01].

3.1.2 Segmentace akustického signálu řeči

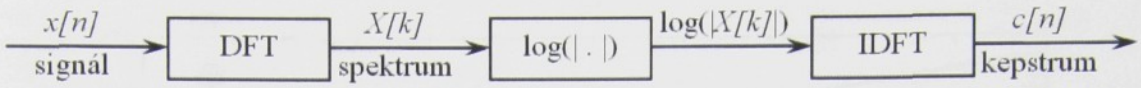
Před parametrizací je digitalizovaný signál řeči nejprve rozsegmentován na jednotlivé framy (segmenty). Využívá se toho, že se frekvenční parametry signálu řeči v průběhu několika málo ms (milisekund) nemění [PSU95, HUA01]. Pro zpracování a rozpoznávání řeči se nejčastěji volí délka jednoho framu 10 – 30 ms a obvykle se využívá i překrývání při posunu jednotlivých framů (kolem 10 ms). Parametrizace se poté provádí pro každý frame zvlášť.

3.1.3 Parametrizace akustického signálu řeči

Do dnešní doby (2005) bylo navrženo velké množství metod pro extrakci příznaků z akustického signálu řeči [HUA01]. Dnes se pro účely zpracování a rozpoznávání řeči nejčastěji používají příznaky získané z kepstra signálu, viz obr. 3.2 Kepstrální koeficienty *c[n]* signálu řeči se vypočtou:

$$c[n] = IDFT\left(\log\left(\left|DFT\{x[n]\}\right|\right)\right) \tag{3.1}$$

kde *x[n]* jsou vzorky digitalizovaného signálu řeči, *DFT* je diskrétní Fourierova transformace a *IDFT* je inverzní Fourierova transformace.



Obr. 3.2: *Výpočet kepstra ze signálu*

Výhoda kepstrální analýzy spočívá ve skutečnosti, že lze od sebe oddělit složky signálu, který byl vytvořen konvolucí těchto složek. U signálu řeči se jedná (hypoteticky) o konvoluci složky buzení hlasového traktu s impulsní odezvou hlasového ústrojí [PSU95, HUA01]. Před vlastním výpočtem kepstrálních koeficientů je vhodné upravit vzorky nacházející se v parametrizovaném framu. V prvním kroku je dobré zvýraznit hodnoty vyšších frekvencí v signálu (preemfáze). Preemfáze se realizuje pomocí číslicového filtru:

$$x'[n] = x[n] - a.x[n-1], 0 \le n < N \tag{3.2}$$

kde *x[n]* je původní hodnota vzorku, *x'[n]* je hodnota vzorku po preemfázi, *a* je konstanta, která se volí v rozsahu 0.95 – 0.98 a *N* je počet vzorků ve framu.

Po výpočtu preemfáze se na hodnoty vzorků framu aplikuje Hammingovo okénko. Tato operace potlačí hodnoty vzorků na okrajích framu, čímž se potlačí vliv náhlého uříznutí vzorků při segmentaci signálu. Nová hodnota vzorku z framu *x'[n]* se vypočte prostým vynásobením hodnoty vzorku *x[n]* (hodnota vzorku po preemfázi) s hodnotou váhové funkce okénka *w[n]*:

$$x'[n] = x[n].w[n], 0 \le n < N \tag{3.3}$$

kde hodnota váhové funkce okénka je vypočtena [STE97]:

$$w[n] = 0.54 + 0.46\cos\left[\frac{2\pi(n-1)}{N-1}\right] \tag{3.4}$$

kde *N* je délka Hammingova okénka = počet vzorků ve framu.

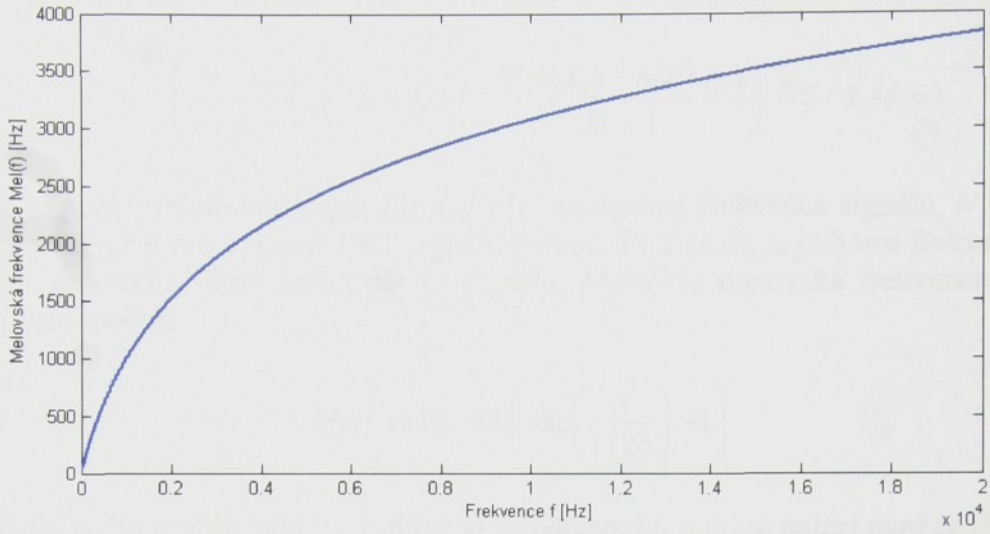
V naší Laboratoři počítačového zpracování řeči se prozatím nejlépe osvědčily MFCC kepstrální příznaky, které se s výhodou používají pro zpracování a rozpoznávání řeči i na jiných pracovištích u nás i ve světě. Tyto příznaky byly použity i v této práci.

3.1.3.1 MFCC kepstrální příznaky

MFCC (Mel-Frequency Cepstral Coefficients) příznaky jsou extrahovány z frekvenční oblasti akustického signálu. Pro vytvoření těchto parametrů se využívá skutečnosti, že člověk vnímá frekvence zvuku přibližně v logaritmické (melovské) stupnici, viz obr. 3.3. Převodní vztah mezi frekvencí a melovskou frekvencí je dán:

$$Mel(f) = 1125 \cdot \ln\left(1 + \frac{f}{700}\right) \tag{3.5}$$

kde *Mel(f)* je melovská frekvence, *f* je původní frekvence a ln je přirozený logaritmus.

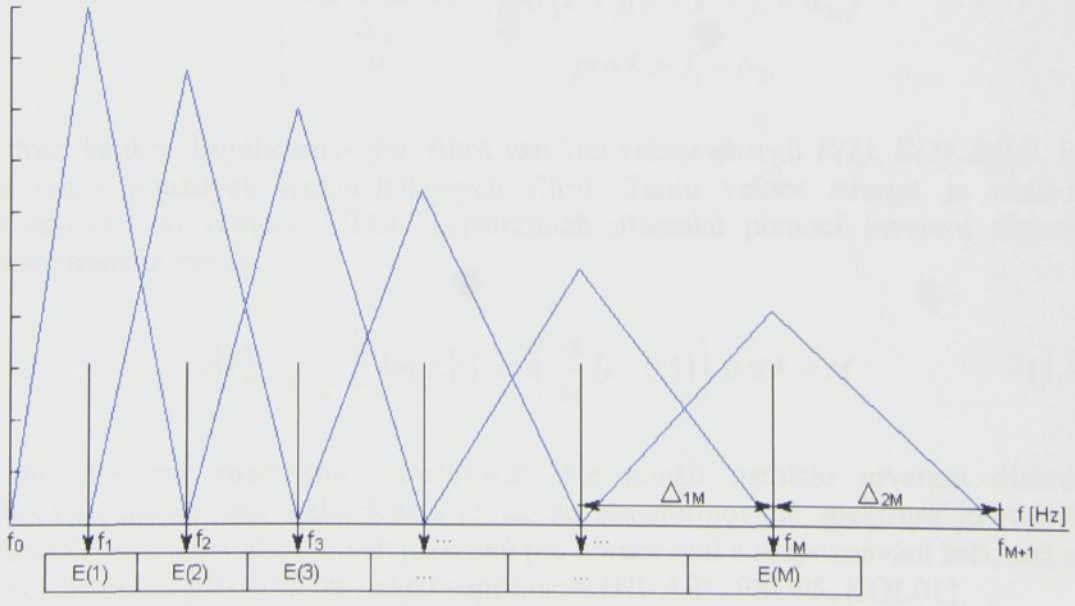


Obr. 3.3: *Převod frekvencí na melovské frekvence*

Pro stanovení MFCC příznaků se nejprve vypočte pro každý frame (segment akustického signálu) frekvenční spektrum pomocí diskrétní Fourierovy transformace DFT (3.6), kde DFT je obvykle počítána pomocí algoritmu rychlé Fourierovy transformace FFT (Fast Fourier Transform).

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-j2\pi nk/N}, 0 \le k < N \tag{3.6}$$

Z frekvenčních vzorků *X[k]* se spočítá amplitudové spektrum, které je poté filtrováno bankou M trojúhelníkových filtrů, viz obr. 3.4.



Obr. 3.4: *Pásma banky filtrů pro aplikaci melovské stupnice frekvencí*

Frekvence f_i pro návrh jednotlivých trojúhelníkových filtrů jsou počítány dle:

$$f_i = \left(\frac{N}{F_S}\right) Mel^{-1}\left(Mel(f_d) + i\frac{Mel(f_h) - Mel(f_d)}{M+1}\right), 0 \le i \le M+1 \qquad (3.7)$$

kde M je počet trojúhelníkových filtrů, F_S je vzorkovací frekvence signálu, N je počet vzorků použitých pro výpočet DFT signálu pomocí FFT (3.6), f_h je horní frekvence a f_d je dolní frekvence filtrovaného pásma signálu, *Mel(x)* je melovská frekvence (3.5) a $Mel^{-1}(x)$ se spočte:

$$Mel^{-1}(x) = 700\left(\exp\left(\frac{x}{1125}\right) - 1\right) \qquad (3.8)$$

Stanovení počtu trojúhelníkových filtrů M a frekvencí f_h a f_d lze nalézt např. v [HUA01, KOL01]. Po filtraci amplitudového spektra trojúhelníkovými filtry se z každého filtrovaného pásma amplitudového spektra spočítá energie:

$$E[i] = \sum_{k=f_i-\Delta_{1i}}^{f_i+\Delta_{2i}} |X[k]|^2 \cdot U_{\Delta_i}[k], 1 \le i \le M \qquad (3.9)$$

kde $\Delta_{1i} = f_i - f_{i-1}$ a $\Delta_{2i} = f_{i+1} - f_i$ a $U_{\Delta_i}[k]$ je váhová funkce příslušného trojúhelníkového filtru:

$$U_{\Delta_i}[k] = \begin{cases} 0 & pro\ k < f_i - \Delta_{1i} \\ \dfrac{k - (f_i - \Delta_{1i})}{\Delta_{1i}} & pro\ (k \ge f_i - \Delta_{1i}) \cap (k \le fi) \\ \dfrac{(f_i - \Delta_{2i}) - k}{\Delta_{2i}} & pro\ (k > fi) \cap (k > f_i - \Delta_{2i}) \\ 0 & pro\ k > f_i - \Delta_{2i} \end{cases} \qquad (3.10)$$

Po filtraci bankou trojúhelníkových filtrů vznikne vektor energií *E(1), E(2)..E(M)*, kde M je počet použitých trojúhelníkových filtrů. Tento vektor energií je následně transformován do vektoru MFCC kepstrálních příznaků pomocí inverzní diskrétní kosinové transformace:

$$c[k] = \sqrt{\frac{2}{M}} \sum_{i=1}^{M} \log E[i] \cdot \cos\left(\frac{\pi k}{M}(i - 0{,}5)\right), 0 \le k < M \qquad (3.11)$$

Inverzní diskrétní kosinovou transformaci lze použít namísto inverzní diskrétní Fourierovy transformace vzhledem k tomu, že transformované spektrum je reálné a symetrické. Popis extrakce jiných příznaků pro zpracování a rozpoznávání řeči, než zde popsaných MFCC příznaků lze nalézt například v [HUA01, PSU95, KOL01].

3.1.3.2 Liftrace kepstrálních příznaků

Vypočtené hodnoty kepstrálních příznaků nabývají se vzrůstajícím indexem stále nižších úrovní. K eliminaci tohoto jevu se používá liftrace kepstrálních příznaků (3.12), pomocí které se přibližně vyrovnají jednotlivé úrovně hodnot.

$$c'[k] = \left(1 + \frac{L}{2}\sin\left(\frac{\pi k}{L}\right)\right).c[k] \tag{3.12}$$

kde *c[k]* jsou původní hodnoty příznaků kepstra, *L* je délka liftračního okna a *c'[k]* jsou výsledné liftrované hodnoty kepstra.

3.1.3.3 Odečítání kepstrálního průměru

Metoda odečítání kepstrálního průměru CMS (3.13) (nebo také odečítání průměru příznaků – FMS – Feature Mean Subtraction) eliminuje různou střední hodnotu příznaků řečového signálu pocházejících z více různých zdrojů (různí mluvčí, mikrofony, hlasitost, ...).

$$V' = \left(\begin{bmatrix} v_{11} \\ v_{12} \\ \vdots \\ v_{1M} \end{bmatrix} - \bar{\bar{v}}, \begin{bmatrix} v_{21} \\ v_{22} \\ \vdots \\ v_{2M} \end{bmatrix} - \bar{\bar{v}}, \cdots, \begin{bmatrix} v_{N1} \\ v_{N2} \\ \vdots \\ v_{NM} \end{bmatrix} - \bar{\bar{v}} \right), \quad \bar{\bar{v}} = \frac{1}{N} \begin{bmatrix} \sum_{i=1}^{N} v_{i1} \\ \vdots \\ \sum_{i=1}^{N} v_{iM} \end{bmatrix} \tag{3.13}$$

kde *V'* je normalizovaná (pomocí FMS) posloupnost vektorů příznaků z jednotlivých parametrizovaných framů, *v* jsou jednotlivé příznaky, *N* je počet framů, *M* je počet příznaků v příznakovém vektoru framu a $\bar{\bar{v}}$ je vektor středních hodnot jednotlivých příznaků.

3.1.3.4 Dynamické příznaky

Dynamické příznaky vyjadřují změnu statických příznaků v čase a častokrát mají větší vypovídající schopnost než vlastní statické příznaky [NOU95]. V oblasti zpracování a rozpoznávání řeči se obvykle používají dynamické příznaky spočítané z první diference (delta příznaky) a dynamické příznaky z druhé diference (delta-delta příznaky, akcelerační příznaky). Pro vlastní výpočet dynamických příznaků se nejčastěji volí zpětná kauzální diference (3.14), nekauzální diference (3.15) nebo se používají složitější regresní aproximace (3.16).

$$x'[n] = x[n] - x[n-1] \tag{3.14}$$

$$x'[n] = \frac{x[n+1] - x[n-1]}{2} \tag{3.15}$$

kde *x[n]* jsou původní hodnoty vzorků, *x'[n]* je vlastní diference.

$$x'[n] = \frac{\sum_{i=1}^{N} i.(x[n+i] - x[n-i])}{2\sum_{i=1}^{N} i^2} \tag{3.16}$$

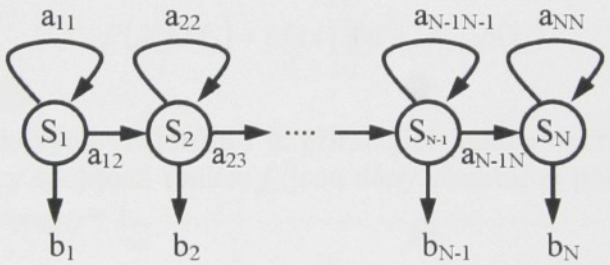
kde $x[n]$ jsou původní hodnoty vzorků, $x'[n]$ je diference a N je délka okna aproximace.

3.2 Rozpoznávání řeči metodou skrytých Markovových modelů

Rozpoznávání řeči metodou skrytých Markovových modelů HMM (Hidden Markov Model) je dnes nejčastěji používaná metoda pro rozpoznávání řeči. U této klasifikace se parametrizovaný akustický signál řeči porovnává s předem vytvořenými HM modely slov (hlásek...) a zjišťuje se, který model nejlépe charakterizuje parametrizovaný signál.

3.2.1 Celoslovní HM model

HM modely jsou tvořeny jednotlivými stavy uspořádanými do určité struktury. Pro rozpoznávání slov pomocí celoslovních modelů se obvykle používá lineární struktura stavů a přechod mezi stavy je možný pouze z leva do prava, viz obr. 3.5.



Obr. 3.5: *Levo-pravé lineární uspořádání stavů pro rozpoznávání slov založeného na celoslovních modelech*

Počet stavů HM modelu je obvykle menší než je počet framů ve slově, akustický signál řeči totiž sestává z různě dlouhých stacionárních úseků, tvořených několika sousedními framy, kde se jednotlivé příznaky příliš nemění. Pomocí statistických metod je poté pro každý stav hledáno statistické rozložení hodnot příznakových vektorů z „informačně podobných" framů. Nejčastěji používaným statistickým rozložením je normální rozložení. Při vlastním vytváření (trénování) HM modelu lze následně zjistit kolik framů obvykle reprezentuje konkrétní stav, z této informace je poté spočítána pravděpodobnost setrvání modelu ve stavu s, která je vyjádřená hodnotou a_{ss}. Pravděpodobnost a_{ss+1}, že model přejde do dalšího stavu se vypočte $a_{ss+1} = 1 - a_{ss}$. Vlastní stav modelu je charakterizován pravděpodobnostní výstupní funkcí $b_s(\vec{x})$ s vícerozměrným normálním rozdělením s vektorem středních hodnotou $\vec{\bar{x}}_s$ a kovarianční maticí Σ_s.

$$b_s(\vec{x}) = \frac{1}{\sqrt{(2\pi)^P \det \Sigma_s}} \cdot \exp\left[-0.5\left(\vec{x} - \vec{\bar{x}}_s\right)^T \Sigma_s^{-1}\left(\vec{x} - \vec{\bar{x}}_s\right)\right] \tag{3.17}$$

kde $\vec{x}$ je příznakový vektor a P je počet příznaků.

Pokud máme rozsáhlý soubor dat pro trénování jednotlivých HM modelů, tak lze výsledný HM model ještě více zkvalitnit například použitím vícemodálního normálního rozdělení. Při trénování se poté určí pro každý mód vektor středních hodnot, příslušná kovarianční matice a váhový koeficient *c*. Výsledná výstupní funkce ze vztahu 3.17 při využití vícemodálního normálního rozdělení má tvar:

$$b_s(\vec{x}) = \sum_{m=1}^{M} c_{sm} \frac{1}{\sqrt{(2\pi)^P \det \Sigma_{sm}}} \cdot \exp\left[-0.5(\vec{x} - \vec{\bar{x}}_{sm})^T \Sigma_{sm}^{-1} (\vec{x} - \vec{\bar{x}}_{sm})\right] \qquad (3.18)$$

kde M je počet módů normálního rozdělení.

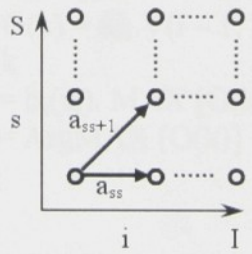
3.2.2 Klasifikace metodou skrytých Markovových modelů

Po vytvoření celoslovních HM modelů lze přistoupit k vlastní klasifikaci metodou skrytých Markovových modelů. U této klasifikace je nejprve počítána pravděpodobnost $P(X|M_i)$, nakolik parametrizovaný akustický signál (neznámé slovo X) odpovídá příslušnému modelu M_i, tj. je zjišťována maximální pravděpodobnost vypočítaná přes všechny přípustná přiřazení stavů a framů slova:

$$P(X \mid M_i) = \max_f \prod_{i=1}^{I} a_{f(i-1)f(i)} b(\vec{x}_i) \qquad (3.19)$$

kde I je počet příznakových vektorů a f je přiřazující funkce operující v rovině *is* (s – číslo stavu), podmínky spojitosti funkce f jsou dány strukturou použitého HM modelu, při $f(0)f(1)$ zavádíme $a_{f(0)f(1)} = 1$.

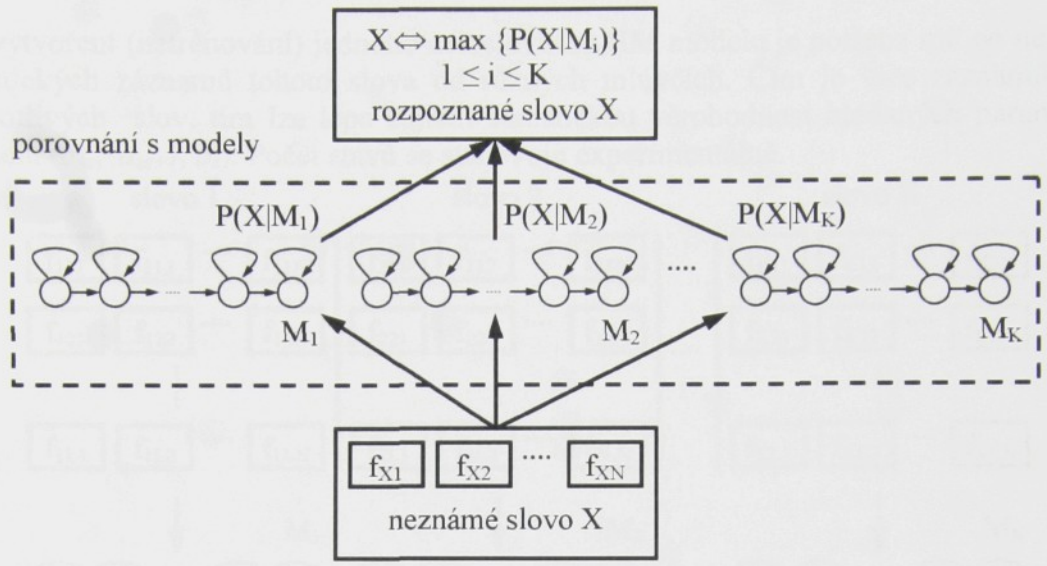
Pro levo-pravé lineární uspořádání stavů jsou možné pouze dva typy přechodů, setrvání v původním stavu nebo přechod do dalšího stavu.



Obr. 3.6: *Znázornění dvou možných přechodů v rovině is*

Ze všech vypočtených pravděpodobností $P(X|M_i)$ přes všechny modely M_i je vybrána maximální pravděpodobnost přiřazení framu neznámého slova stavům modelu. Neznámé slovo poté odpovídá modelu, pro který byla vypočtena tato maximální pravděpodobnost, viz obr. 3.7. Pro rychlý výpočet vztahu (3.19) se dnes obvykle používá Viterbiho algoritmus [HUA90, KOL01]. V tomto algoritmu je rekurzivně počítán kumulovaný součin $V(i,s)$.

$$V(i,s) = b_s(x_i).Max\left[a_{ss}V(i-1,s), a_{s-1s}V(i-1,s-1)\right] \qquad (3.20)$$



Obr. 3.7: *Princip klasifikace slov pomocí celoslovních skrytých Markovových modelů*

Struktura Viterbiho algoritmu je poté následující:

1) Inicializace

$V(1, 1) = b_1(x_1)$
$B(1, 1) = 1$ (pole zpětných ukazatelů)
$V(1, s) = -\infty$ pro s = 2, .. S, kde S je počet stavů HM modelu

2) Rekurze

```
FOR i = 2, .. I          kde I je počet framů
    FOR s = 1, .. S
        FOR k = s – 1, s
            O(k) = aks.V(i – 1, k)
        NEXT k
        V(i, s) = bs(xi). MAX [O(k)]      pro k = s – 1, s
        B(i, s) = ArgMAX [O(k)]           pro k = s – 1, s
    NEXT s
NEXT i
```

3) Ukončení

$P(X|M_i) = V(I, S)$

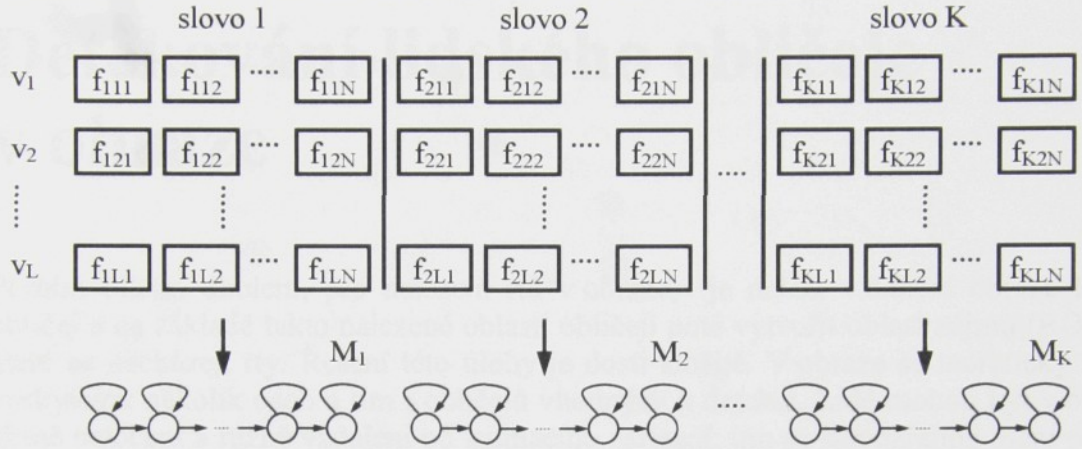
Pro účely stanovení a zobrazení Viterbiho cesty lze poté určit posloupnost stavů:

4) Určení posloupnosti stavů

```
f(I) = S
FOR i = I – 1, .. 1
    f (i) = B(i+1, f (i+1))
NEXT i
```

3.2.3 Trénování celoslovních HM model

Při vytvoření (natrénování) jednoho celoslovního HM modelu je potřeba mít co nejvíce akustických záznamů tohoto slova od různých mluvčích. Čím je více záznamů (v_i) jednotlivých slov, tím lze lépe zajistit statistickou věrohodnost hledaných parametrů modelu (a_{ss}, a_{ss+1}, b_s). Počet stavů se stanovuje experimentálně.



Obr. 3.8: *Trénování celoslovních HM modelů*

Na obr. 3.8 je zobrazen zjednodušený princip trénování jednotlivých celoslovních HM modelů, f jsou parametrizované framy, L je počet záznamů každého slova a N je počet jednotlivých framů záznamu slova v_i, přičemž N je obvykle pro každou variantu slova jiné a obdobně i L může být pro každé slovo jiné. Do dnešní doby vzniklo několik algoritmů pro natrénování HM modelů. Popis těchto algoritmů lze nalézt např. v [HUA90, PSU95, KOL01, HUA01].

Kapitola 4

Detekování lidského obličeje v obraze

Prvním dílčím úkolem, pro nalezení rtů v obraze, je nalézt v daném obraze lidský obličej a na základě takto nalezené oblasti obličeji poté vytvořit oblast zájmu (ROI), ve které se nacházejí rty. Řešení této úlohy je dosti složité. V obraze se teoreticky může vyskytovat několik osob a tím i obličejů vhodných k detekci. Lidé mohou být v obraze různě natočeni a různě vzdáleni od snímacího zařízení, tím se samozřejmě mění plocha zaznamenaného obličeje v obraze. Navíc se v průběhu snímání scény může měnit i osvětlení. Ve své práci se zabývám především rozpoznáváním obrazů, ve kterých se nachází pouze jeden mluvčí, je čelně natočen ke snímací kameře a osvětlení scény se příliš nemění. Pro tyto účely byla prostudována dostupná literatura a byly zváženy jednotlivé metody a algoritmy pro nalezení lidského obličeje.

4.1 Metody pro detekování lidského obličeje

Až dosud (r.2005) bylo navrženo a popsáno velké množství metod vhodných pro detekci obličeje v obraze. V následujících statích jsou uvedeny některé z těchto detekčních metod [YAN01]. Roztřídění některých metod není úplně jednoznačné, jelikož tyto metody by se daly zatřídit do více statí najednou.

4.1.1 Příznakově orientované metody

Následně popsané metody jsou zaměřeny na nalezení vhodných příznaků, které by vhodně charakterizovaly lidský obličej. Na základě těchto příznaků je poté provedena detekce lidského obličeje.

4.1.1.1 Obličejové příznaky

Tyto metody jsou založené na hledání celého obličeje nebo na hledání jeho částí – obočí, očí, nosu, rtů, vlasů. Jedna z těchto metod vychází z nalezení očí nebo očí a obočí [HAN98]. Pořízený obraz je segmentován pomocí vhodně zvolených prahů a v segmentovaném obrazu jsou poté nalezeny oči. Na základě vyhodnocení vzdálenosti a umístění očí v obraze je poté stanovena oblast, ve které se s největší pravděpodobností nalézá obličej. Tato metoda je však použitelná pouze při vhodně stanovených podmínkách pořízení obrazu.

Další používaná metoda pro detekci obličeje je založena na nalezení hran v obraze, např. [SIR83]. U této metody byl použit Cannyho hranový detektor [CAN83] pro nalezení hran. Na základě heuristické analýzy se poté jednotlivé hrany odstraní nebo sloučí tak, aby v hranovaném obraze zůstala pouze ta linie, která reprezentuje obrys obličeje. Tato hrana je poté nahrazena vhodnou elipsou, která odděluje nalezený obličej od pozadí obrazu.

Jiná metoda používá algoritmus kapek a linií [CHE93]. Pro tuto metodu byl vytvořen obličejový model, který se skládá ze dvou tmavých a tří světlých kapek, které reprezentují oči, lícní kosti a rty. Model dále obsahuje linie, reprezentující obrysy obličeje, obočí a rtů. Pro určení prostorového vztahu mezi kapkami jsou použity dvě trojúhelníkové geometrické konfigurace. Obraz je upraven pomocí Laplaceova gradientního operátoru s nízkým rozlišením. Tato operace usnadňuje nalezení jednotlivých kapek a linií. Na základě rozmístění jednotlivých kapek je obraz prohlédnut, zda se v předem odhadnutých místech nacházejí linie. Detekční algoritmus pak vyhodnocuje, na základě nalezení a rozmístění kapek a linií, zda se v pořízeném obraze nachází obličej.

V dalších metodách je využívána tvarová segmentace obrazu pomocí operací matematické morfologie [SON98] a to jak binární, tak šedotónové. Pomocí morfologických operací otevření a uzavření lze hledat celý obličej [ATH03] nebo jeho jednotlivé části [GRA96] – oči, obočí, nos, rty. Využívá se vlastnosti, že jednotlivé části obličeje mají určitý tvar (oči-kruh, obličej a rty-elipsa) a pokud je vhodně zvolena velikost a tvar strukturního elementu pro morfologické operace, tak lze nalézt celý obličej nebo jeho příznaky (oči, rty, nos, nosní dírky, obočí).

Při detekci obličeje se využívá i toho, že jednotlivé části obličeje (oči, rty, obočí, nosní dírky) jsou od sebe navzájem relativně stejně vzdálené [LEU95]. Pro tuto metodu je nejdříve vytvořena (ručně nebo poloautomaticky) databáze rozmístění a relativních vzdáleností částí obličeje. Z pořízeného obrazu se poté postupně vyseparují (pomocí postupné segmentace nebo různými obrazovými filtry) jednotlivé části obličeje a zjistí se jejich vzájemné relativní vzdálenosti, které se pak porovnají s hodnotami z databáze. Tato metoda je spolehlivější než metoda pouhého nalezení oblasti očí, jenž je popsána výše.

4.1.1.2 Obličejová textura

Lidský obličej je v pořízeném obraze reprezentován texturou, která se u lidí podobné barvy kůže příliš nemění. Této vlastnosti lze využít pro detekci lidského obličeje v obraze [AUG93]. Texturou je zde chápána pravidelná struktura obrazových bodů, kde každý obrazový bod má svou barevnou, popř. jasovou hodnotu. Velikost této textury (počet obrazových bodů) závisí na velikosti lidského obličeje v obraze a na celkovém rozlišením obrazu. Příznaky jsou zde chápány poté jednotlivé obrazové body nebo se často původní obraz převzorkuje na nižší rozlišení a tím se zmenší počet příznaků a celková výpočetní náročnost. Pořízený obraz se porovnává s obličejovými texturami z databáze, která se předem vytvoří z ručně nebo poloautomaticky zpracovaných snímků s lidským obličejem. Někdy se obraz neporovnává s celými texturami, ale postupně s výřezy textury.

4.1.1.3 Barva kůže

Nejčastěji používaná metoda pro detekci lidského obličeje je založena na zjištění barvy kůže. Na základě této informace se poté dá, s jistou pravděpodobností, nalézt lidský obličej v obraze. Ve světě existuje několik skupin lidí (etnik), kteří mají přibližně stejnou barvu kůže. Skupinou zde není chápána ani rasa či národnost. V současnosti se pro detekci lidského obličeje vymezují dvě skupiny lidí s různou barevností kůže - lidé se světlou barvou kůže (běloši, někteří asiaté, míšenci...) a lidé s tmavou barvou kůže (černoši, část asiatů, původní obyvatelé Austrálie, někteří z původních obyvatelé amerického kontinentu...). Při rozpoznávání se většinou dělají barevné modely pro každou z těchto skupin zvlášť nebo se vytváří vícemodální modely.

Vytváření barevných „kůžových" modelů závisí na použité snímací kameře. Dříve hojně rozšířené černo-bílé (šedotónové) kamery se dnes již příliš nepoužívají a je také více obtížnější detekovat obličej v šedotónovém obraze než v barevném. V následujících statích tak bude pojednáno pouze o barevných obrazech. Nejčastěji se barevný obraz reprezentuje v RGB barevném prostoru, kde R (red) je červená složka obrazu, G (green) je zelená složka obrazu a B (blue) je modrá složka obrazu. Složením těchto tří barevných složek vznikne barevný obraz.

Pro vytvoření barevného „kůžového" modelu se používá buď RGB barevný prostor, normalizovaný barevný prostor RGB nebo se RGB obraz převede do jiného barevného prostoru. Převedením RGB obrazu do jiného barevného prostoru se ve složkách (ve složce) může zvýraznit nebo potlačit určitá barva nebo její odstín, což může být výhodou při hledání určité skupiny barev. Nejvíce používané barevné transformace pro detekci obličeje z barvy kůže jsou převody z RGB barevného prostoru do barevného prostoru: HSV(I) (H – hue, S – saturation, V – value, I – intensity) [KJE96], YCrCb [ATH03], YIQ [DAI96], YES [SAB98], CIE XYZ [CHE95], CIE LUV [YAN98].

Existuje několik metod pro stanovení, zda obrazový bod v originálním nebo barevně transformovaném obraze má barvu kůže. Nejednoduší je segmentace obrazu prahováním, kdy všechny obrazové body, jejichž barevná hodnota je větší (nebo menší), než je hodnota prahu, budou mít novou hodnotu 1 (1 – obrazový bod má barevnou hodnotu jako barva kůže) nebo hodnotu 0 (0 – obrazový bod nemá barevnou hodnotu jako barva kůže). Velmi často se stanovují prahy dva a obrazový bod má barevnou hodnotu jako je barva kůže, pokud se jeho barevná hodnota BH nalézá v intervalu $BH_1 \leq BH \leq BH_2$, kde BH_1 a BH_2 jsou vhodně stanovené prahy. Hodnoty prahů se většinou stanovují experimentálně nebo na základě analýzy obrazového histogramu. Často používaná je i metoda, kdy se z obrazu vyseparují (ručně nebo poloautomaticky) pouze obrazové body, které příslušejí lidské kůži (v obraze). Všechny barevné hodnoty takto vybraných bodů se statisticky zpracují, tj. zjistí se střední hodnota, rozptyl atd. a tyto odhadnuté statistické parametry jsou použity pro rozpoznávání, s jakou pravděpodobností má neznámý obrazový bod barevnou hodnotu jako je barva kůže [YAN98]. Použít pouze odhad jedno-mixturového normálního rozdělení by bylo nevhodné, jelikož jak bylo popsáno na začátku existuji lidé se světlou a tmavou barvou kůže, proto se volí spíše více-mixturové normální rozdělení, nejčastěji však dvou-mixturová.

Barva kůže samo o sobě jako příznak pro detekci obličeje je v některých případech nedostatečná (nestejnoměrné a proměnlivé osvětlení, šum v obraze atd.), proto se tato metoda kombinuje s dalšími metodami pro detekci obličeje, jako je například tvarová segmentace nebo detekce pohybu ve video sekvenci.

4.1.1.4 Vícenásobné příznaky

Pro větší spolehlivost detekování lidského obličeje je vhodné použít více různých příznaků z různých metod, např. použít metodu pro barevnou segmentaci obrazu na základě vyhodnocení barvy kůže (kapitola 4.1.1.3.) a po této barevné segmentaci použít tvarovou segmentaci (kapitola 4.1.1.1.), která nám pomůže odstranit všechny obrazové body, které sice měly podobnou barvu jako je barva lidské kůže, ale tyto body patřili do okolí obrázku s lidským obličejem [ATH03]. Často se k těmto příznakům přidávají další, které jsou zaměřeny na zjištění velikosti, geometrie a pozice jednotlivých částí obličeje [YAN98]. Nebo jsou jako příznaky použity lidský obličej a vlasy a jejich vzájemné umístění v obraze. Obecně lze říct, že u této metody se využívají různé kombinace příznaků, které mají jakýkoliv vztah k obličeji a mohou tak napomoci jeho detekci. V některých experimentech se použily i příznaky získané po nalezení vousů a brýlí v obraze.

4.1.2 Porovnávání se vzory

Pro metodu porovnávání se vzory je potřeba předem vytvořit databázi vzorů, kde vzorem je chápána oblast, ve které jsou obrysy lidského obličeje a částí obličeje (oči, nos, rty). Nejčastěji se používají čelní pohledy na lidský obličej, ale je i možné vytvořit databázi vzorů nasnímaných při rozdílných natočeních lidské hlavy. Pro daný vstupní pořízený obraz s obličejem jsou nezávisle počítány korelační hodnoty pro obrysy obličeje, oči, nos a rty. Nevýhodou této metody je, že v porovnávací databázi musí být velké množství vzorů a ne vždy lze postihnout různé změny měřítka, natočení a tvaru, proto se vytvářejí databáze, kde jsou vzory a výřezy vzorů v různých měřítkách, pořízené při různých natočeních hlavy.

4.1.2.1 Předdefinované vzory

U této metody je vytvářena databáze vzorů a výřezů ze vzorů (podvzorů). Vzory z takto vytvořené databáze se poté porovnávají se vstupním obrazem, např. při využití korelace. Vzory se někdy upravují, aby se zvýraznila informace, kterou jsou tyto vzory charakteristické. Jednou z úprav je například předzpracovat vzory a vstupní obraz pomocí hranového obrazového detektoru [CRA87]. Aby se nemusela vytvářet příliš velká databáze vzorů, tak se jednotlivé vzory upravují před porovnáním se vstupním obrazem pomocí metod geometrických transformací [SON98], především s využitím transformací pro rotaci a změnu měřítka.

4.1.2.2 Deformovatelné vzory

Tyto metody se snaží najít univerzálnější vzory pro detekci obličeje [KWO94, LAM94]. Pomocí různých optimalizačních algoritmů se u této metody hledají univerzální (deformovatelné) vzory pro jednotlivé části obrazu. Poté se tyto univerzální vzory

porovnávají se vstupním obrazem (nebo jeho postupným výřezem) a hledá se optimální přiřazení mezi částí obrazu a vzorem, například pomocí minimální vzdálenosti. Při tomto porovnávání a přiřazení tak dochází k deformaci vzorů.

4.1.3 Učící se metody pro detekci obličeje

Učící se metody jsou založeny na podobném principu, jako metody porovnávání se vzory. Zde však nejsou jednotlivé vzory vytvořené experimentátorem, ale hledají se pomocí učících se algoritmů na základě předložené databáze obrázků, ve kterých jsou lidské obličeje.

4.1.3.1 Vlastní plochy (obličeje – Eigenfaces)

Pro tuto metodu se používá databáze normalizovaných obrazů, ve kterých je lidský obličej. Metoda vlastních ploch využívá pro rozpoznávání vlastní vektory, které byly získány z autokorelační matice obrazů z databáze. Tyto vlastní vektory se dnes označují jako vlastní plochy (eigenfaces) [YAM02]. Pro vypočítání koeficientů vlastních ploch se dnes používá Karhunen-Loèveova transformace [KAR46], pro kterou se spíše používá označení PCA (Principal Component Analysis - analýza nejdůležitějších součástí). Nejprve se vytvoří trénovací databáze obrázků s obličeji, které mají M x N obrazových bodů a z těchto obrázků jsou vytvořeny matice o rozměru M x N. Poté jsou vytvořeny základní vektory pocházející z optimálních podprostorů matic tak, že střední kvadratická chyba mezi projekcí trénovacích obrazů do těchto podprostorů a trénovacími obrazy je minimalizována. Množina optimálních základních vektorů se nazývá vlastní obrazy. Z takto vektorizovaných obrazů obličejů v trénovací databázi jsou počítány vlastní vektory. Pomocí této metody lze zakódovat větší množství obrazů s obličeji do menšího prostoru vlastních ploch, vytváří se tak modely, které poté slouží při rozpoznávání lidského obličeje. Pro detekci obličeje se ze vstupního obrazu postupně (změnou souřadnic x, y) vybírají menší předem zvolené oblasti a měří se vzdálenost mezi těmito oblastmi a předem natrénovanými vlastními plochami (modely). Z těchto vzdáleností se vytváří „obličejová mapa". Lidský obličej (obličeje) je poté dekován v místech lokálních minim obličejové mapy [TUR91]. Někdy je potřeba i zjistit, zda se v obraze lidský obličej vyskytuje nebo nevyskytuje. Pro tyto účely jsou poté vytvářeny neobličejové modely. Celkové vyhodnocení, zda se v obraze lidský obličej vyskytuje, je však více komplikované, jelikož je velmi obtížné nalézt vhodné reprezentativní obrázky, kde se lidský obličej nevyskytuje, teoreticky jich existuje nekonečné množství.

4.1.3.2 Statistické metody vytvoření (ne)obličejových modelů

Zde je využita matematická statistika a statistické metody pro vytvoření obličejových a neobličejových modelů. U jedné z charakteristických statistických metod pro detekci obličeje [SUN98] jsou nejprve obrázky s lidskými obličeji a bez nich normalizovány na určitou, předem danou velikost M x N pixelů. Z těchto matic se vytvoří vektory o velikosti k = (M x N) a z těchto vektorů se vytvoří určitý počet shluků (clusters) pro obličej a neobličej zvlášť. Shluky jsou representovány jako vícedimensionální normální rozdělení, charakterizované středními hodnotami a kovarinční maticí. Ze vstupního pořízeného obrazu se postupně (změnou souřadnic x, y) vybírají menší předem zvolené

oblasti. Tyto oblasti se normalizují a vektorizují a zjišťuje se při porovnání se shlukovými modely, zda oblast patří k obličejovému nebo k neobličejovému modelu. Opět je u této metody velmi obtížné vytvořit dostatečně reprezentativní neobličejové modely. Jedním ze zásadních úkolů u těchto metod je vytvoření normalizované matice, která by obsahovala menší počet dat, než matice obrazu s obličejem (pozadím). Nejjednodušší metoda je zmenšení původního obrazu pomocí obrazové geometrické transformace pro změnu měřítka (zmenšení obrazu) nebo lze využít různých ztrátových nebo bezztrátových obrazových transformací, např. výše popsanou transformaci PCA nebo lineární Fisherovu diskriminační analýzu (FLDA) [YAN00, CHE04] atd.

4.1.3.3 Neuronové sítě

Zhruba v polovině 80. let došlo k velkému rozvoji použití neuronových sítí ve výzkumu a aplikacích. Neuronové sítě se používají pro velké množství aplikací, kde chceme rozpoznávat různé vzory (signály, obrazy...), není proto divu, že se neuronové sítě uplatňují i při detekci obličeje [PRO92, JAI02]. Vytvoření neuronových sítí bylo inspirováno nervovou a řídicí soustavou vyšších organismů. Základní stavební jednotkou neuronových sítí je model neuronu. Neurony jsou vzájemně propojeny sítí, kterou se poté šíří jednotlivé signály, které jsou v neuronu akumulovány a pokud je překročen určitý práh, tak je z neuronu vyslán signál, který se sítí šíří dál [WID60]. V současné době existují různé architektury neuronových sítí, výběr jedné z nich pro různé aplikace pak spočívá v různých pravidlech, experimentování nebo v získaných zkušenostech experimentátora.

Po vytvoření neuronové sítě (softwarovém nebo hardwarovém) pro detekování obličeje je nutné nejdříve tuto neuronovou síť naučit, co má rozpoznávat. Pro tyto účely se používá databáze s lidskými obličeji, které jsou před vstupem do neuronové sítě různým způsobem modifikovány, např. jsou vstupní obrazy upraveny hranovým detektorem nebo jsou zmenšeny, komprimovány pomocí PCA, FLDA, DCT (diskrétní kosinová transformace – používaná například při ztrátové kompresi obrazu JPG) atd. Pokud je síť dostatečně natrénována, tak ji lze následně použít pro detekování obličeje.

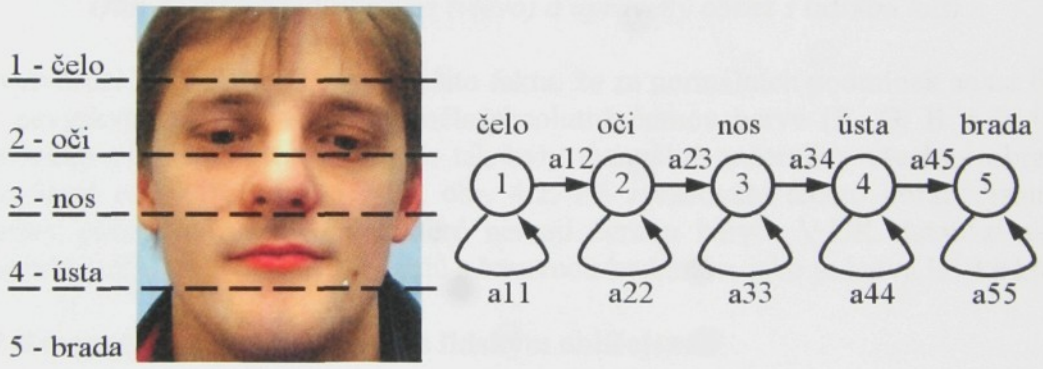
4.1.3.4 Prosívací řídké sítě

Prosívací řídké sítě (Sparse Network of Winnows - SNoW) využívají kombinaci lineárních funkcí a prosívacích pravidel [LIT88] a jsou především určeny pro doménové učení, kde v jednotlivých doménách je velké množství různých příznaků. SNoW se v současné době uplatňují i v oblasti detekování lidského obličeje [YAN01, YAN00b] snímaného v různých natočeních, v rozdílných světelných podmínkách a popsaných různými příznaky.

4.1.3.5 Skryté Markovovy modely

Použití skrytých Markovových modelů (HMM) pro detekci obličeje [MAR99] je stejné jako u rozpoznávání řeči pomocí HMM (Kapitola 3), namísto řečového jednorozměrného signálu je zde použit obraz (2D signál). Prvním úkolem je nalézt vhodné příznaky, které by dobře reprezentovaly lidský obličej a druhým úkolem je vhodné stanovení počtu stavů HM modelů(u). Pro natrénování HM modelů se používá,

tak jako u předchozích metod, databáze lidských obličejů a to jak originálních, tak upravených pomocí různých obrazových transformací, např. použitím hranového detektoru. Jedna z metod [SAM94] používá jeden model složený z pěti stavů, které jsou řazeny za sebou z leva do prava a které reprezentují oblasti s čelem, očima, nosem, ústy a bradou v obrazu s lidskou tváří, tato metoda dostatečně dobře reprezentuje myšlenku detekce lidského obličeje v obraze při využití skrytých Markovových modelů. Pro trénování je databáze obrazů s obličeji rozdělena na pět dílů, které budou poté reprezentovat oněch pět stavů. jednotlivé framy zde tvoří řádky nebo skupiny řádků v obraze (to je častější případ). Používá se i překryv jednotlivých framů obdobně, jako je tomu při rozpoznávání řečového signálu. Příznaky pak mohou být barevné hodnoty jednotlivých barevných bodů nebo to mohou být koeficienty získané z rozličných transformací (PCA, FLDA, DCT...). Pro trénování HM modelu se obraz skenuje odshora dolů, viz obr. 4.1. Při detekci obličeje se poté ze vstupního obrazu postupně (změnou souřadnic x, y) vybírají menší předem zvolené oblasti, které se porovnávají z modelem a zjišťuje se pravděpodobnost souhlasu výřezu s modelem. Podle hodnoty výsledné pravděpodobnosti pak můžeme stanovit, zda výřez je lidským obličejem, hodnota tohoto stanovení (prahu) se určí experimentálně. Můžeme samozřejmě vytvořit modelů více pro lidský obličej (pro různá etnika) a pro různá pozadí (neobličej), opět je zde ovšem problém s vytvořením dostatečného počtu reprezentativních modelů neobličeje.



Obr. 4.1: Použití skrytých Markovových modelů pro detekci obličeje

4.2 Detekování lidského obličeje

V předchozích statích byl uveden základní přehled dnes používaných metod pro detekování lidského obličeje v obraze. Pro účely audio – vizuálního rozpoznávání řeči bylo potřeba nalézt metodu pro detekování obličeje, která by byla jednoduchá (tj. výpočetně časově nenáročná) a zároveň dostatečně spolehlivá. V mnou pořízené audio – vizuální databázi AVDB2cz (kapitola 7) jsou videonahrávky, ve kterých je vždy pouze jedna osoba v obraze. Tím se problém nalezení lidského obličeje v obraze částečně zjednodušil. V databázi se však nachází 494 780 obrázků (dekódované nahrávky od 35 mluvčích) s lidským obličejem, proto byl velký důraz kladen na výpočetní nenáročnost použité metody. Po prostudování kladů a záporů jednotlivých metod pro detekci obličeje, byla nakonec vybrána kombinovaná metoda založená na barevné segmentaci obrazu (stať 4.1.1.3.) následovaná tvarovou segmentací (stať 4.1.1.1.).

4.2.1 Databáze obrazů s lidskou kůží

Před vlastní barevnou segmentací obrazu s obličejem bylo potřeba získat informaci o barvě lidské kůže. Pro tyto účely byla manuálně vytvořena databáze obrazů (DK), ve kterých se nacházela pouze lidská kůže. Tato DK databáze byla vytvořena z obrázků 52 osob, které se nacházely v mé původní audio – vizuální databázi AVDB1cz (kapitola 7). Tyto obrázky byly získány po dekódování videonahrávek, které byly pořízené v jiných světelných podmínkách a od jiných mluvčích než zpracovávané obrázky z databáze AVDB2cz, ale byly pořízeny stejnou kamerou a ukázalo se, že pro vytvoření barevných kůžových modelů jsou vyhovující.

Obr. 4.2: *Originální obraz (vlevo) a upravený obraz s lidskou kůží*

Pro vytvoření DK databáze bylo využito faktu, že za normálních podmínek se na lidské kůži nevyskytuje oblast, která by měla absolutně černou barvu (R, G, B = 0, 0, 0). V obrazech s obličeji z této databáze tak byly manuálně začerněny všechny obrazové body, které netvořily lidskou kůži, obr. 4.2. Při zpracování těchto obrazů jsou pak vybírány pouze obrazové body, které nemají černou barvu. V DK databázi se tak nacházelo 1 652 440 obrazových bodů s barevnou hodnotou, jako je barva lidské kůže.

4.2.2 Barevná segmentace obrazu s lidským obličejem

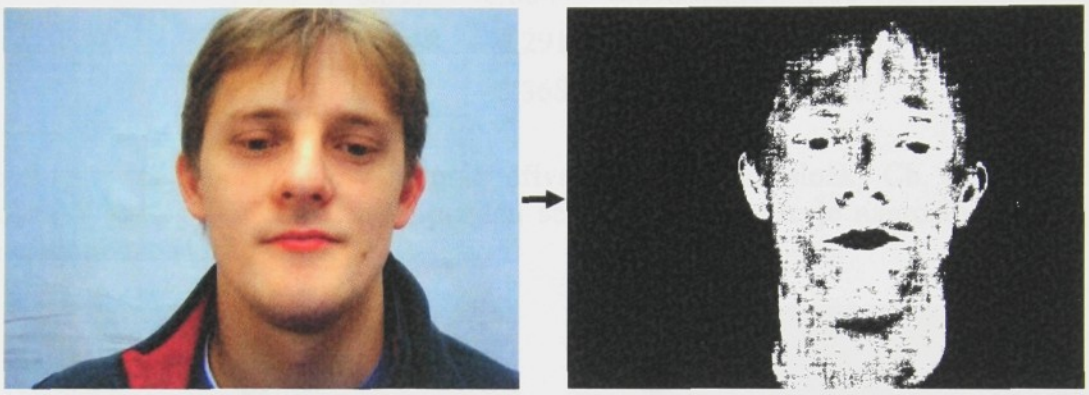
Následující popsané metody pro barevnou segmentaci obrazu s lidským obličejem slouží k oddělení objektu (obličeje) od okolí, při využití znalosti o barvě lidské kůže. Po této barevné segmentaci vznikne nový (binární) obraz, ve kterém obrazové body, které s určitou pravděpodobností patří k objektu, mají hodnotu 1. Body, které patří k okolí, mají hodnotu 0. Zpracovávané obrazy z databáze AVDB2cz byly uloženy jako bitmapy ve dvaceti-čtyř bitovém formátu. V tomto formátu je každému obrazovému bodu přiřazena barevná hodnota – tří bytový (24 bitů) vektor, ve kterém jsou uloženy barevné složky obrazu: R (Red – červená složka), G (Green – zelená složka) a B (Blue – modrá složka). Níže popsané metody využívají tyto RGB barevné složky pro segmentaci obrazu s lidským obličejem.

4.2.2.1 Segmentace obrazu pomocí převodní tabulky

První z metod pro barevnou segmentaci obrazu s obličejem byla založená na vytvoření převodní tabulky (Look-up table), která přímo převáděla barevné RGB hodnoty jednotlivých obrazových bodů na hodnoty 0 a 1.

Nejprve byla vytvořena tří-rozměrná tabulka o velikosti 256 (0..255) x 256 (0..255) x 256 (0..255) a všechny hodnoty v tabulce byly vynulovány. V této tabulce byly zaznamenávány četnosti barevných hodnot obrazových bodů z DK databáze. Barevná hodnota RGB obrazového bodu byla využita jako index v převodní tabulce. Pokud měl například obrazový bod hodnotu (R = 123, G = 14, B = 254), tak se hodnota v tabulce na pozici (123, 14, 254) zvýšila o 1. Poté byla všem hodnotám v tabulce, které byly větší než 1, přiřazena hodnota 1, ostatní hodnoty byly vynulovány.

Takto vytvořená tabulka pak byla použita pro převod barevného obrázku na binární, kdy každému obrazovému bodu z obrázku byla přiřazena hodnota 0 nebo 1, na základě barevné RGB hodnoty tohoto barevného bodu, která opět sloužila jako index tabulky.



Obr. 4.3: *Originální obraz (vlevo) a transformovaný binární obraz*

Tato metoda je při svém použití velmi rychlá, jelikož v jednom kroku dostaneme rovnou segmentovaný binární obraz. Na obrázku 4.3 je zobrazen výsledek po převodu barevného obrazu na binární (hodnoty 0 jsou zobrazené jako černé body a hodnoty 1 jsou bíle body). Z binárního obrazu je patrné, že ne všechny obrazové body, které tvoří lidskou kůži, mají hodnotu 1. To je nevýhodou použití této metody, protože pro dostatečnou přesnost potřebujeme velikou databázi s obrazy s lidskou kůží. Proto byla hledána jiná metoda pro barevnou segmentaci, která by používala více spojitější barevný model pro segmentaci. Přesto, pokud je vytvořena dostatečně velká DK databáze, tak je tato metoda pro barevnou segmentaci lidského obličeje jednou z nejefektivnějších.

4.2.2.2 Segmentace obrazu s využitím jednoduchého statistického modelu

Prvním úkolem u této metody bylo převést RGB barevný prostor do jiného (menšího) barevného prostoru. Pro tyto účely existuje celá řada barevných transformací, viz stať 4.1.1.3. Po počátečních pokusech jsem si vybral YCbCr barevnou transformaci, která byla použita např. v [ATH03] a v mnoha jiných odborných pracích zaměřených na detekování obličeje v obraze.

YCbCr barevná transformace byla původně určena pro barevné televizní vysílání, kde Y označuje jas a Cr, Cb jsou chromizační složky, které nesou informaci o barvě. Cr složka je reprezentována jako rozdíl mezi červenou složkou obrazu a referenční hodnotou a Cb je složka, která je reprezentována rozdílem mezi modrou složkou obrazu a referenční

hodnotou. Modulovaný televizní signál tak obsahoval všechny potřebné informace pro zobrazení v barevném i černobílém televizoru, jelikož ze složek Y, Cb a Cr lze zpětně vypočíst barevné hodnoty R, G a B.

Výhodou této transformace je, že je lineární a proto se snadno a rychle dají převést jenotlivé barevné složky R, G, B. Do dnešní doby vzniklo několik variant vztahů pro přepočet barevných hodnot do tohoto prostoru, liší se především hodnotami jednotlivých váhových koeficientů. Zde je uveden vztah, který byl odvozen z normovaného vztahu [KEI97] a který je použit i v programu MATLAB, ve funkci rgb2ycbcr:

$$\begin{bmatrix} Y \\ C_b \\ C_r \end{bmatrix} = \begin{bmatrix} 0.257 & 0.504 & 0.098 \\ -0.148 & -0.291 & 0.439 \\ 0.439 & -0.368 & -0.071 \end{bmatrix} \cdot \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \tag{4.1}$$

Pro detekování obličeje byly v mé práci dříve používány obě složky Cb, Cr [CHA03d] (složka Y není příliš použitelná), později se však ukázalo, že je naprosto postačující používat pouze složku Cr.

Obr. 4.4: *Transformace barevného obrazu do Cr složky*

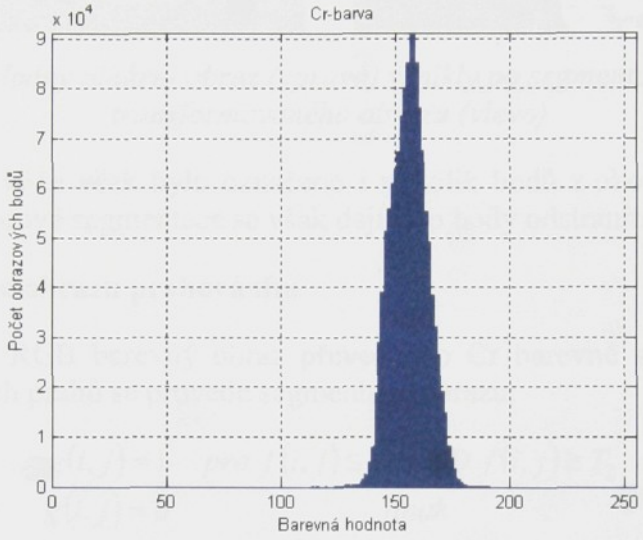
Na obrázku 4.4 je zobrazena Cr složka výsledného obrazu po YCbCr barevné transformaci, tmavá místa v obraze představují nízkou hodnotu složky Cr a světlá místa reprezentují vysokou hodnotu. Z obrázku je patrné, že čím je obrazový bod v originálním obraze více červený, tím je větší jeho hodnota v Cr složce.

Před vlastní barevnou segmentací obrazu byl vytvořen barevný pravděpodobnostní model. Barevné obrazy lidské kůže z DK databáze byly převedeny do Cr složky a ze všech Cr hodnot obrazových bodů lidské kůže byla vypočtená střední hodnota $\hat{\mu}_k$ (4.2) a rozptyl $\hat{\sigma}_k^2$ (4.3). Na obrázku 4.5 je zobrazen histogram, který byl vytvořen ze souboru Cr hodnot (1652440) z DK databáze.

$$\hat{\mu}_k = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{4.2}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{i=1}^{N}(x_i - \hat{\mu}_k)^2}{N-1} \tag{4.3}$$

kde x_i je Cr barevná hodnota obrazového bodu kůže z DK databáze a N je počet obrazových bodů z DK databáze. Výsledné hodnoty byly následující: střední hodnota $\hat{\mu}_k = 156.2$ a rozptyl $\hat{\sigma}_k^2 = 62.4$.



Obr. 4.5: *Histogram z Cr barevných hodnot obrazových bodů z DK databáze*

Pro segmentaci Cr-barevně transformovaného obrazu byla počítána pro každý obrazový bod *f* o souřadnicích (i, j) míra pravděpodobnosti p_k, která určuje, nakolik má příslušný obrazový bod *f(i, j)* barevnou hodnotu shodnou s barvou kůže *k*:

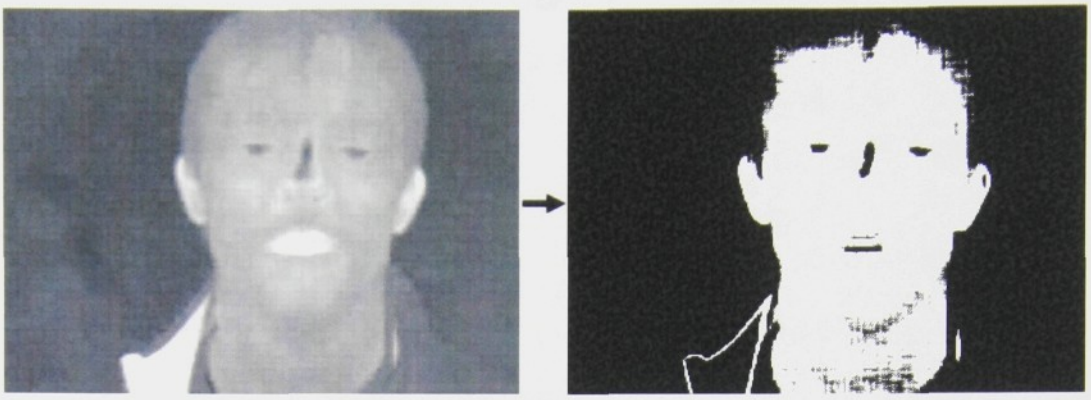
$$p_k(f(i,j)) = \frac{1}{\sqrt{2\pi}.\hat{\sigma}_k} \exp\left[-\frac{(f(i,j) - \hat{\mu}_k)^2}{2\hat{\sigma}_k^2} \right] \tag{4.4}$$

Vlastní segmentace obrazu pak byla následující:

$$\begin{aligned} g(i,j) &= 1 \quad pro\ p_k(f(i,j)) \geq P_k \\ g(i,j) &= 0 \quad pro\ p_k(f(i,j)) < P_k \end{aligned} \tag{4.5}$$

kde *f(i,j)* je obrazový bod v původním Cr barevném obraze, *g(i,j)* je obrazový bod ve výsledném binárním obraze po segmentaci a P_k je „pravděpodobnostní" práh, který byl stanoven experimentálně na hodnotu 0.008. Při takto zvoleném prahu by bylo zpětně označeno 95.44 % obrazových bodů z DK databáze hodnotou 1.

Binární obraz, který vznikl po segmentaci za využití statistického modelu barev lidské kůže, viz obr. 4.6. Oproti předchozí metodě s převodní tabulkou bylo nalezeno více obrazových bodů, které mají barvu kůže.



Obr. 4.6: *Výsledný binární obraz (vpravo) vzniklý po segmentaci Cr-barevně transformovaného obrazu (vlevo)*

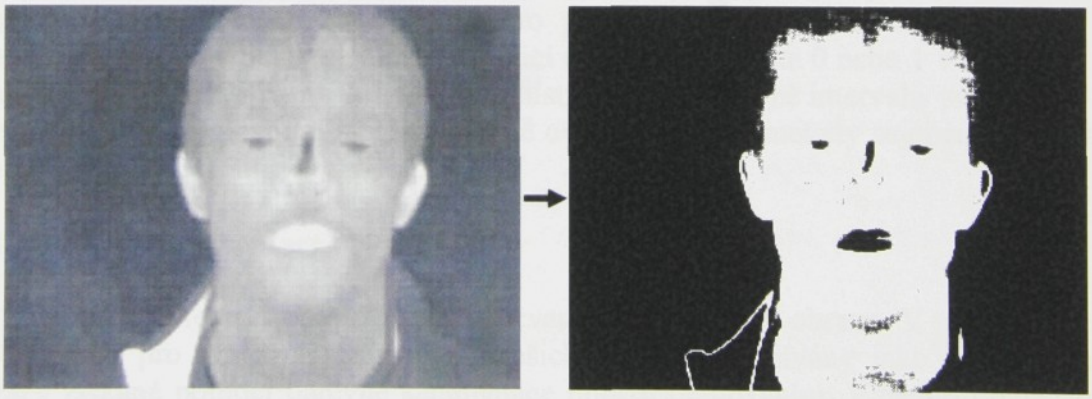
Za body s barvou kůže však bylo označeno i několik bodů z okolí lidského obličeje. Pomocí metody tvarové segmentace se však dají tyto body odstranit.

4.2.2.3 Segmentace obrazu prahováním

U této metody se RGB barevný obraz převede do Cr barevné složky a na základě vhodně stanovených prahů se provede segmentace obrazu:

$$\begin{aligned} g(i,j) &= 1 \quad pro\ f(i,j) \le T_1\ AND\ f(i,j) \ge T_2 \\ g(i,j) &= 0 \qquad\qquad jinak \end{aligned} \tag{4.6}$$

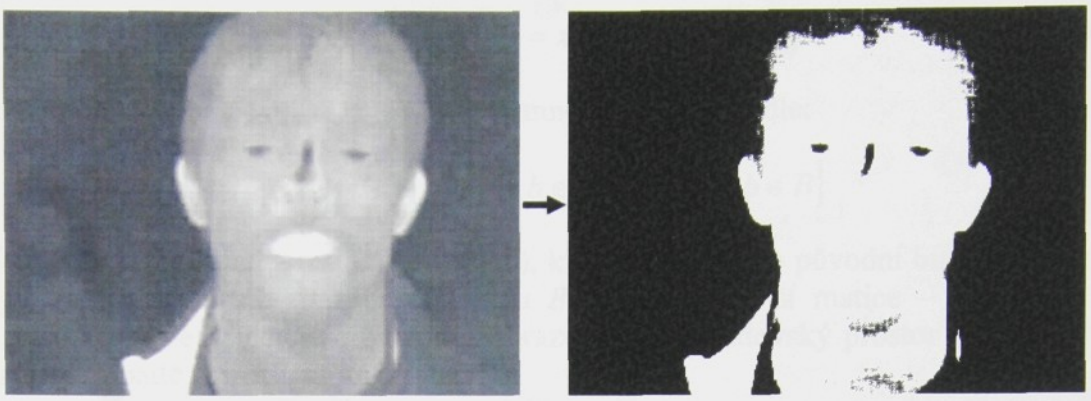
kde T_1 je dolní práh, T_2 je horní práh, $T_1 < T_2$, $f(i,j)$ je barevná hodnota obrazového bodu o souřadnicích i, j v původním Cr barevném obraze, $g(i,j)$ je nová binární hodnota obrazového bodu ve výsledném binárním obraze.



Obr. 4.7: *Binární obraz (vpravo) vytvořený segmentací prahováním z Cr-barevně transformovaného obrazu (vlevo)*

Hodnoty prahů byly stanovena (4.7) za využití odhadnutých parametrů střední hodnoty a rozptylu, viz předchozí stať. Pro toto stanovení je však použita směrodatná odchylka σ, která se vypočítá jako odmocnina z rozptylu.

$$\begin{aligned} T_1 &= \hat{\mu}_k - 2.\hat{\sigma}_k = 140.4 \\ T_2 &= \hat{\mu}_k + 2.\hat{\sigma}_k = 172 \end{aligned} \tag{4.7}$$



Obr. 4.8: *Binární obraz (vpravo) vytvořený segmentací prahováním, kdy byl použit pouze dolní práh T_1*

Na obrázku 4.8 je zobrazen výsledný segmentovaný obraz u kterého byl použit pouze dolní práh T_1. V tomto obraze se objevily i objekty (červená část bundy), které nemají přímý vztah k lidskému obličeji. Naopak při použití pouze horního prahu by se v obraze objevilo i pozadí, které by již nešlo odlišit od objektu (obličeje), to je důvod, proč jsou při segmentaci použity prahy dva.

Výsledky z metod segmentace obrazu s využitím jednoduchého statistického modelu a segmentace prahováním jsou velmi podobné, viz obr. 4.5 a 4.6. U metody segmentace s využitím jednoduchého statistického modelu se sice používá pouze jeden práh, ale zato se musí počítat míra pravděpodobnosti (4.4), takže tato metoda je výpočetně časově náročnější než metoda segmentace prahováním.

Algoritmus výpočtu segmentace obrazu prahováním se dá ještě podstatně zrychlit, při využití převodní tabulky o rozměrech 256 x 256 x 256, která je vytvořená tak, že se všechny hodnoty celého RGB barevného prostoru přepočítají do Cr složky a do výsledné převodní tabulky se uloží na pozici (R, G, B) hodnota 0 nebo 1 v závislosti na tom, jestli se tato Cr-barevná hodnota umístí uvnitř nebo vně intervalu prahů T_1 a T_2 (4.6). Použití této převodní tabulky je poté obdobné jako u metody segmentace obrazu pomocí převodní tabulky (stať 4.2.2.1.).

4.2.3 Tvarová segmentace obrazu

Po barevné segmentaci obrazu následuje tvarová segmentace obrazu. V současné době (2005) jsou pro tvarovou segmentaci našich obrázků používány jako vstup binární obrazy získané pomocí barevné segmentace obrazu prahováním. V těchto obrazech je větší množství detailů (obr. 4.6), které je potřeba odstranit, aby jsme nalezli oblast, ve které se nachází pouze lidský obličej. Pro tyto účely je použita tvarová segmentace obrazu založená na metodách matematické morfologie [SER82, SON98, HLA00]. Metody matematické morfologie jsou založeny na algebře nelineárních operací. Zde byla využita binární matematická morfologie a její dvě morfologické operace otevření a uzavření. Tyto dvě operace jsou založeny na elementárních morfologických operacích erozi a dilataci. Operace dilatace $\oplus$ skládá body dvou matic (množin) pomocí vektorového součtu (4.8).

$$X \oplus B = \left\{ p \in \varepsilon^2 : p = x + b, x \in X, b \in B \right\} \tag{4.8}$$

Morfologická operace eroze $\ominus$ skládá dvě množiny (matice) dle:

$$X \ominus B = \left\{ p \in \varepsilon^2 : p + b \in X \text{ pro každé } b \in B \right\} \tag{4.9}$$

kde x je obrazový bod z matice (množiny), která reprezentuje původní binární obraz X, b je bod z matice strukturního elementu B (obvykle menší matice – vzor) a p je obrazový bod ve výsledném binárním obrazu, ϵ^2 je euklidovský prostor – 2D matice, v našem případě reprezentující obraz

Morfologická operace otevření $\circ$ množiny X strukturním elementem B je definována jako morfologická operace eroze následovaná dilatací.

$$X \circ B = (X \ominus B) \oplus B \tag{4.10}$$

Morfologická operace uzavření $\bullet$ je definována jako morfologická operace dilatace následovaná erozí.

$$X \bullet B = (X \oplus B) \ominus B \tag{4.11}$$

Operace otevření a uzavření nám zjednoduší vstupní binární obraz, ve kterém poté bude menší množství detailů. Operace otevření oddělí objekty spojené úzkou šíjí a odstraní z obrazu všechny objekty, které jsou „menší" než strukturní element.Operace uzavření naopak spojí objekty, které jsou „blízko" u sebe, dále vyhladí obrys objektu a zaplní díry v objektu, které jsou „menší" než použitý strukturní element. Vlastnosti „menší" a „blízko" závisejí na velikosti a tvaru použitého strukturního elementu.

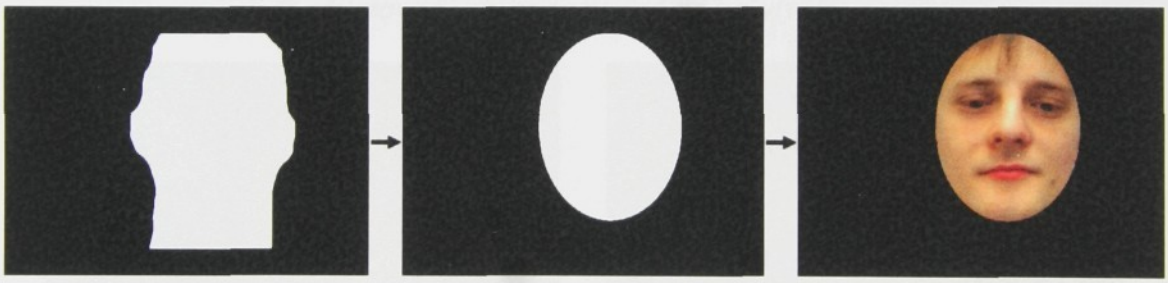
Pro vlastní použitou tvarovou segmentaci byla použita operace otevření následovaná operací uzavřením následovaná upravenou operací otevřením. Pro každou operaci byl použit jiný strukturní element.

Jako první z morfologických operací byla použita operace otevření s podlouhlou elipsou jako strukturním elementem. Pomocí této operace byla z původního binárního obrazu odstraněna většina detailů, které se nacházely v okolí (pozadí) lidského obličeje, viz obr. 4.9. Kdyby se namísto operace otevření použila operace uzavření, tak by tyto detaily byly ve výsledném obraze „zesíleny".

Obr. 4.9: *Vstupní binární obraz vytvořený barevnou segmentací (vlevo), binární obraz po prvním otevření (uprostřed) a následném uzavření (vpravo). Nad šipkami jsou zobrazené příslušné použité strukturní elementy.*

Po operaci otevření byla provedena operace uzavření. U této operace byl použit strukturní element ve tvaru větší podlouhlé elipsy. Tato operace odstranila z binárního obrazu díry způsobené stíny, rty, očima atd. a zjednodušila obrys objektu. Poslední použitá operace byla upravená operace otevření.

U předchozích dvou operací byl rozměr a tvar strukturního elementu zvolen experimentálně a pro všechny zpracovávané obrazy byl stejný. Vstupní strukturní element pro upravené otevření byla podlouhlá elipsa o rozměrech 362 (šířka) x 480 (výška) obrazových bodů, velikost zpracovávaných obrazů byla 640 x 480 obrazových bodů. Rozměry tohoto strukturního elementu byly stanoveny na základě analýzy rozměrů (šířky a výšky) lidského obličeje u několika desítek osob. Algoritmus pro operaci upravené otevření se skládal ze dvou částí. Za prvé byl binární obraz erodován strukturním elementem (elipsou), který byl postupně zmenšován. Pro každý krok eroze s měnícím se strukturním elementem byl vstupem binární obraz po uzavření. Cyklus se zastavil, když výsledkem eroze obrazu se zmenšovaným strukturním elementem byla hodnota jedna. Poté se provedla dilatace s tímto zmenšeným strukturním elementem. Vzhledem k tomu, že po (neukončené) erozi byl v binárním obraze pouze jeden obrazový bod s hodnotou jedna, tak i ve výsledném dilatovaném obraze byla pouze jedna automaticky vybraná elipsa. Tento výsledný obraz sloužil jako maska pro vybrání oblasti, ve které se nachází pouze lidský obličej, viz obr. 4.10. Z této oblasti byla poté vybrána oblast zájmu (ROI – region of interest), ve které se nacházely pouze rty.



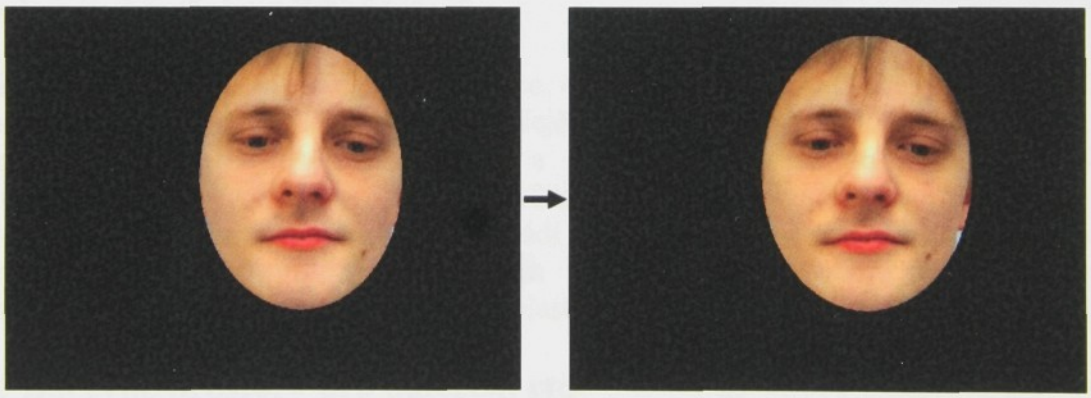
Obr. 4.10: *Binární obraz po uzavření (vlevo), po upravené operaci otevření (uprostřed) a výsledný obraz , ve kterém je již vybrána pouze oblast s obličejem.*

Pro účely přesnější představy o vybrané oblasti s obličejem byl vytvořen obrázek 4.11. V tomto barevném obraze je zobrazena hranice elipsy, která odděluje nalezený obličej od okolí. Obrys elipsy vznikl jako rozdíl výsledného binárního obrazu a binárního obrazu (výsledného), který byl erodován strukturním elementem 5 x 5. Obrys elipsy je pro vetší kontrast zobrazen zeleně.

Obr. 4.11: *Původní barevný obraz s vyznačenou detekovanou oblastí*

Použití algoritmů matematické morfologie pro detekování lidského obličeje je výpočetně poměrně náročné. Vzhledem k tomu, že se v našich nahrávkách nacházel pouze jeden mluvčí a lidský obličej tvořil v obraze nezanedbatelnou část, tak mohly být barevně segmentované obrazy a strukturní elementy před tvarovou segmentací 10 x zmenšeny. Vlastní výpočet morfologických operací pak byl několikanásobně rychlejší (2000 x na PC Barton 3000+, 1GB RAM). Posloupnost jednotlivých morfologických operací je stejná jako u nezmenšeného obrazu, viz obr. 4.12, pouze u poslední operace upraveného otevření se po erozi zjistí relativní poloha RP a relativní měřítko zmenšení RMZ výsledného zmenšeného strukturního elementu (elipsy). Dilatace se pak neprovádí, ale původní velký strukturní element (elipsa 362 x 480) se zmenší dle měřítka RMZ a umístí se ve vynulované matici (640 x 480) do polohy RP. Takto vytvořená maska pak opět slouží pro vybrání oblasti, kde se nachází pouze lidský obličej. Chyba, která vzniká zmenšením obrázků, je vzhledem k následné oblasti zájmu ROI a hledání rtů v této oblasti zanedbatelná, viz obr. 4.13.

Obr. 4.12: *a – zmenšený binární obraz po barevné segmentaci, b – obraz po otevření, c – obraz po uzavření, d – výsledný obraz po upraveném otevření (zde zobrazený dilatovaný obraz se normálně nevytváří)*



Obr. 4.13: *Srovnání výsledné detekce obličeje metodou barevné a tvarové segmentace původního obrazu (vpravo) a obrazu, který byl při tvarové segmentaci 10 x zmenšen*

V příloze č. 1 jsou zobrazeny výsledky detekce lidského obličeje pro různé mluvčí pomocí Cr-barevné segmentace prahováním a tvarové segmentace s využitím zmenšení binárního obrazu a strukturních elementů. U takto navržené metody trvá nalezení obličeje v jednom obraze přibližně 0.05 s na PC Barton 3000+, 1 GB RAM.

Kapitola 5

Nalezení oblasti rtů v detekované oblasti zájmu

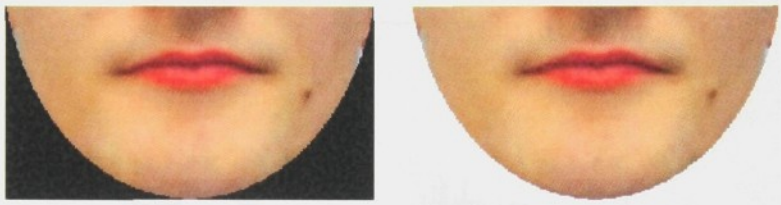
Nalézt rty v pořízeném obraze s mluvčím by byla velmi obtížná úloha. Pokud je totiž v obraze celá osoba popřípadě více osob, tak je velikost rtů oproti velikosti celého obrazu velmi malá a tím i hůře detekovatelná. Proto je v obraze nalezen nejdříve lidský obličej a z takto nalezené oblasti je vybrána určitá část – oblast zájmu (ROI – Region Of Interest), ve které se nacházejí rty. Pro audio-vizuální rozpoznávání řeči se pak používají vizuální příznaky, které byly získány z jednotlivých obrazových bodů z oblasti zájmu (kapitola 6) nebo geometrické příznaky, pro které je potřeba nalézt oblast rtů z oblasti zájmu. Existuje také několik metod, které využívají pro rozpoznávání kombinaci geometrických příznaků a příznaků získaných z obrazových bodů z ROI [HEN96]. Tato kapitola pojednává o metodách pro nalezení oblasti rtů v pořízené oblasti zájmu ROI.

Tak jako u detekování obličeje v obraze existuje v současné době (r.2005) velké množství metod pro nalezení oblasti rtů. Nejčastěji se používají metody pro segmentaci obrazu prahováním [GAO00, ZHA01] a segmentaci obrazu za pomocí různých barevných statistických modelů [DAU02]. Před vlastní segmentací obrazu se však velmi často vlastní barevný obraz (RGB) převádí do jiného barevného prostoru. Nalezení vhodného barevného prostoru je jedním z nejdůležitějších úkolů, pro následnou dostatečně spolehlivou segmentaci rtů z oblasti zájmu.

5.1 Barevné prostory pro nalezení rtů

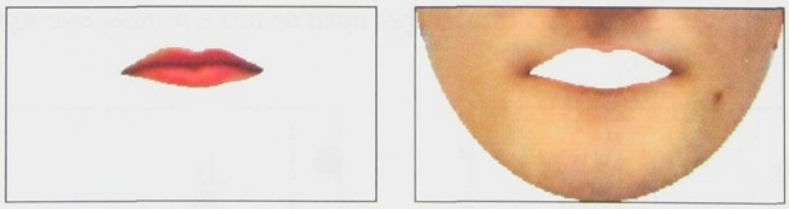
Po převodu RGB barevného obrazu do jiného barevného prostoru se využijí jenom některé složky z tohoto prostoru a to ty, ve kterých je velký kontrast barev mezi rty a okolím (kůží). U některých lidí je barevný kontrast mezi rty a kůží velice nízký a vlastní nalezení oblasti rtů je pak více obtížné, proto se hledají takové barevné transformace, které tento kontrast zesílí. Z pohledu nalezení rtů jsou pro rozpoznávání nejvhodnějšími objekty ženy, které si často barvu rtů zvýrazňují rtěnkou. V některých dřívějších ale i současných [DAU01] pracích se mluvčím záměrně zvýraznili rty kontrastní rtěnkou, např. modrou. U takto „upravených" mluvčích je pak velmi jednoduché provést segmentaci rtů. V této práci však byly zpracovávány nahrávky lidí, bez jakýchkoliv předběžných úprav. Pro účely nalezení oblasti rtů tak musela být stanovena vhodná a dostatečně spolehlivá barevná transformace. V následujících statích jsou uvedeny některé z barevných transformací používaných pro segmentaci rtů. Použití různých barevných transformací je však velmi závislé na pořízených video snímcích (pořizovací

technika, osvětlení scény atd.) a barevná transformace, která se ukázala jako vhodná v jedné práci nemusí být vhodná pro zpracování jiných video snímků. Při vytváření mého systému pro audio-vizuální rozpoznávání řeči bylo postupně použito několik vhodných barevných transformací, které jsou uvedeny v následujících statích. U každé barevné transformace je pro srovnání použit stejný snímek oblasti zájmu ROI, který byl vybrán z naší audio-vizuální řečové databáze (kapitola 7). Oblast zájmu pro nalezení oblasti rtů byla vytvořena ze spodní části automaticky nalezené elipsy (dolních 40 % z výšky elipsy), která reprezentuje nalezený detekovaný obličej v obraze (kapitola 4). Při předzpracování obrazu se ukázalo, že pro některé barevné transformace ROI je výhodnější nahradit černou barvu masky [R G B] = [0 0 0] barvou bílou [R G B] = [255 255 255], pro segmentaci takto upraveného obrazu ROI pak byl použit pouze jeden práh, viz obr. 5.1. Tato záměna je možná, jelikož při segmentaci oblasti zájmu jsou hledaným objektem rty a barva obrazových bodů, které tvoří obraz rtů není nikdy za „normálních podmínek" černá nebo bílá.



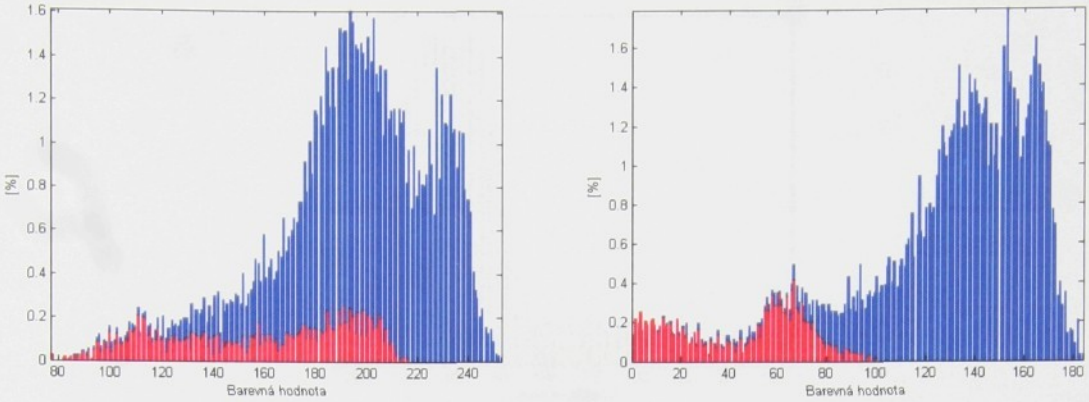
Obr. 5.1: *Oblasti zájmu ROI získané po detekci obličeje v obraze, s původní černou maskou (vlevo) a s bílou maskou (vpravo)*

Pro lepší představu o využitelnosti jednotlivých barevných transformací jsou dále místo barevně transformovaných obrazů oblasti zájmu uváděny obrazové histogramy. V těchto histogramech jsou navíc barevně odlišeny barevné body, které tvoří hledaný objekt – rty (červená barva) a pozadí – kůže obličeje (modrá barva). Pro takto vytvořené histogramy byly z ROI ručně odděleny obrazové body rtů a kůže, viz obr. 5.2. Vznikly tak dva obrazy, ve kterých se nachází pouze příslušný objekt (rty, kůže) a okolí je reprezentované body s bílou barvou, tyto bílé body se na vytvoření histogramu nepodílely. Výsledný histogram byl poté složen z histogramu obrazu rtů a obrazu kůže.

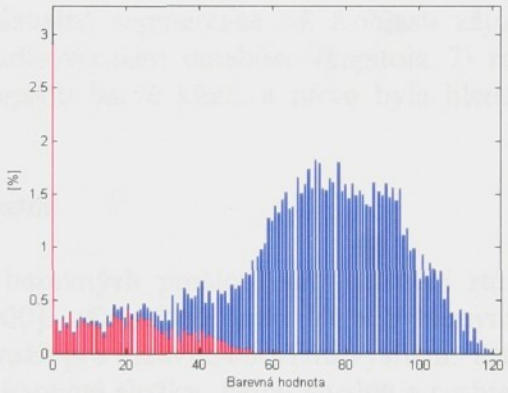


Obr. 5.2: *Rozdělení obrazu ROI na obraz se rty a s kůží pro účely vytvoření obrazového histogramu*

Na obrázků 5.3 jsou zobrazeny obrazové histogramy, které byly získány z jednotlivých barevných složek RGB barevného obrazu oblasti zájmu (obr. 5.1(2)). Z těchto histogramů je patrné, že ani u jedné z barevných složek by nebylo dostatečně možné oddělit obrazové body rtů od bodů kůže, aniž by bylo zároveň označeno větší množství obrazových bodů kůže jako obrazové body rtů a naopak.

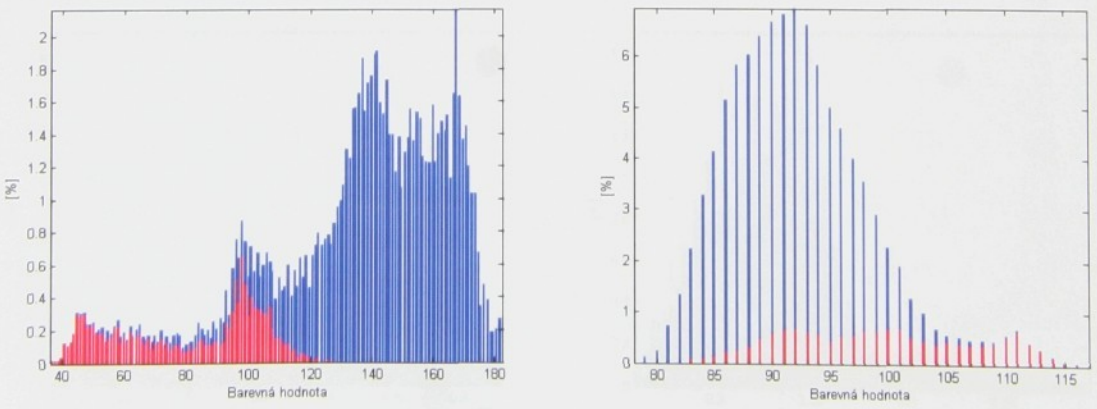


Obr. 5.3: *Obrazové histogramy barevných složek R (vlevo), G (vpravo) a B (dole) z barevného obrazu oblasti zájmu, na osa y jsou zobrazeny relativní četnosti*

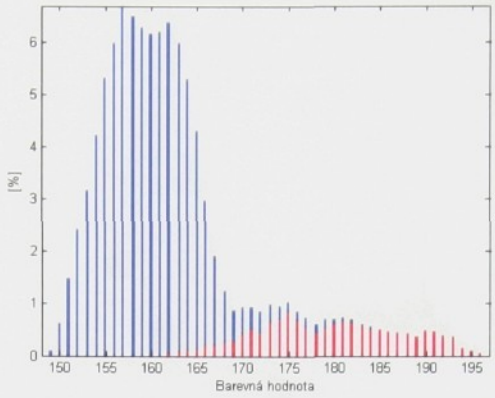


5.1.1 YCbCr barevný prostor

YCbCr barevný prostor (vztah 4.1) se spíše používá pro detekování lidského obličeje (kapitola 4), přesto ho lze za určitých podmínek využít i pro barevnou transformaci oblasti zájmu se rty [CHA03d, CHA04]. Výhoda použití stejné barevné transformace pro detekování obličeje a pro nalezení oblasti rtů plyne z toho, že je barevný obraz transformován pouze jednou a tím se ušetří výpočetní čas.



Obr. 5.4: *Obrazové histogramy barevných složek Y (vlevo) a Cb (vpravo)*



Obr. 5.5: *Obrazový histogram barevné složky Cr z YCbCr barevného prostoru*

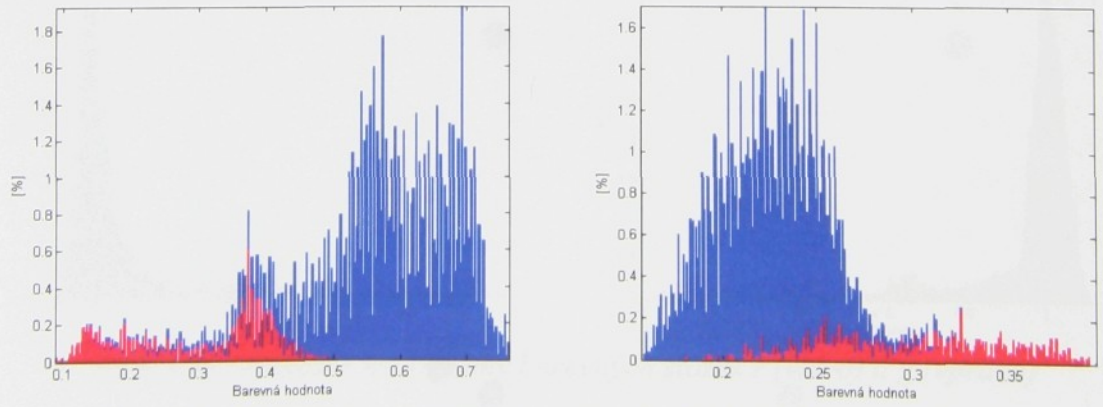
Obdobně jako u detekování obličeje je nejvhodnější použít Cr složku, viz obr. 5.5. Pokud ve videonahrávkách má mluvčí výrazně červené rty oproti barvě kůže, tak po použití Cr složky je následná segmentace rtů z oblasti zájmu dostatečně spolehlivá. V nahrávkách z naší audio-vizuální databáze (kapitola 7) měly však někteří mluvčí nevýraznou barvu rtů oproti barvě kůže, a proto byla hledána spolehlivější barevná transformace.

5.1.2 YIQ barevný prostor

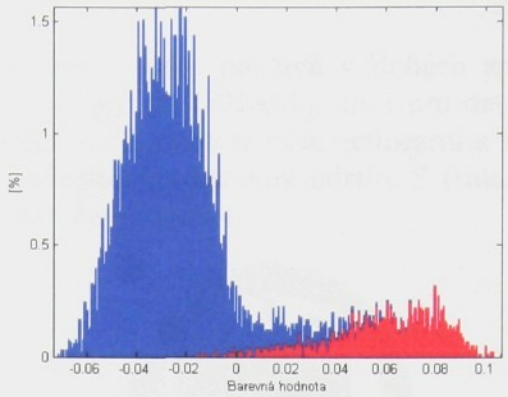
Dalším z používaných barevných prostorů pro nalezení rtů v oblasti zájmu je YIQ barevný prostor [GAO00]. Obdobně, jako YCbCr barevná prostor byl tento YIQ barevný prostor (5.1) využit pro barevné televizní vysílání. Tato transformace je lineární a tím se dají jednotlivé barevné složky velmi snadno a rychle transformovat. Složka Y představuje jas a I, Q jsou chrominanční složky.

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.275 & -0.321 \\ 0.212 & -0.523 & 0.311 \end{bmatrix} \cdot \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{5.1}$$

Pro nalezení rtů je nejvhodnější použít složku Q, viz obr 5.6. Barevný histogram ze složky jasu vypadá obdobně jako u YCbCr barevného prostoru a pro vlastní nalezení oblasti rtů není příliš vhodný.



Obr. 5.6: *Obrazové histogramy barevných složek Y (vlevo) a I (vpravo)*



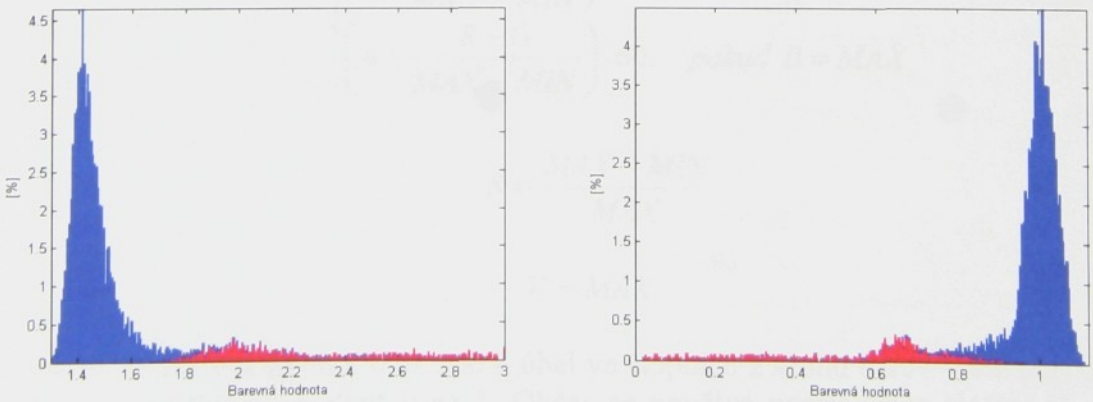
Obr. 5.7: Obrazový histogram barevné složky Q z YIQ barevného prostoru

5.1.3 rg barevný prostor

Transformace barevných složek pro tento barevný prostor je sice nelineární (5.2), ale vypočet této transformace je přesto dostatečně rychlý. Pro následnou barevnou segmentaci oblasti zájmu se rty se dají využít obě barevné složky r, g [SAN00] nebo lze použít pouze g složku [CHA04b].

$$[r\ g]=\left[\frac{3.R}{R+G+B}\ \frac{3.G}{R+G+B}\right] \tag{5.2}$$

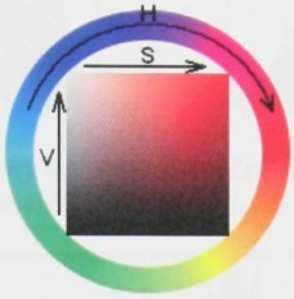
Před vlastní barevnou transformací je lepší převést černou masku v oblasti zájmu na bílou, viz obr. 5.1., tím se částečně vyhneme problému, kdy R = G = B = 0 a tím dojde ve vztahu 5.2 k dělení nulou, přesto pro převod RGB barevného obrazu do rg barevného prostoru se musí ošetřit případ, kdy R = G = B = 0, v tomto případě je výsledná barevná hodnota pro r barevnou složku nastavena na 0 a pro g na 3, čímž nemůže být následně označen absolutně černý bod (R = G = B = 0) jako bod tvořící obraz rtů. Pro rg barevnou transformaci je tak nejvhodnější použít převodní tabulky pro převod barev z RGB barevného prostoru do rg barevného prostoru, při využití převodní tabulky se nemusí převést černá maska na bílou, jelikož nám převodní tabulka přímo převede barvu [R G B] = [0 0 0] na [r g] = [0 3].



Obr. 5.8: Obrazové histogramy barevných složek r (vlevo) a g (vpravo)

5.1.4 HSV barevný prostor

HSV barevný prostor se velmi často používá v úlohách zpracování a rozpoznávání obrazu, slouží tak i pro nalezení rtů [ZHA01], ale i pro detekování obličeje v obraze [KJE96]. HSV transformace RGB barev je však nelineární a výpočetně více náročnější. Barevná složka H (hue) představuje barevný odstín, S (saturation) zobrazuje saturaci barvy a V (value) přibližně odpovídá jasu.



Obr. 5.9: *Schématické zobrazení barevného prostoru HSV*

Nejvhodnější pro nalezení rtů je složka H, složka V naopak příliš vhodná není, obdobně jako jas u YCbCr a YIQ barevného prostoru. Pro nalezení rtů se používá i HI [CIS03] (hue, intensity) a HSI [LIE98] (hue, saturation, intensity) barevný prostor. V těchto barevných prostorech se složka H počítá dle jiného vztahu než u HSV, ale výsledný transformovaný obraz H složky je po normalizaci přibližně stejný. Pro výpočet barevných složek HSV se nejdříve normalizuje RGB barevný prostor:

$$0 \le R,G,B \le 255 \Rightarrow [R\ G\ B]/255 \Rightarrow 0.0 \le R,G,B \le 1.0 \tag{5.3}$$

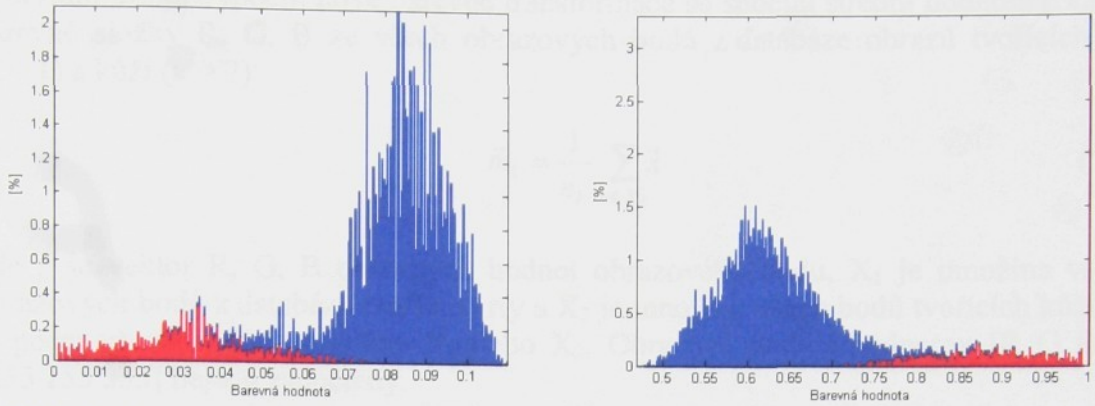
Poté se pro každý barevný obrazový bod zjistí z vektoru [R G B] minimální (MIN) a maximální hodnota (MAX). Nové HSV složky tohoto bodu se pak spočítají:

$$H = \begin{cases} \left(0 + \dfrac{G-B}{MAX-MIN}\right) \cdot 60, & pokud\ R = MAX \\ \left(2 + \dfrac{B-R}{MAX-MIN}\right) \cdot 60, & pokud\ G = MAX \\ \left(4 + \dfrac{R-G}{MAX-MIN}\right) \cdot 60, & pokud\ B = MAX \end{cases} \tag{5.4}$$

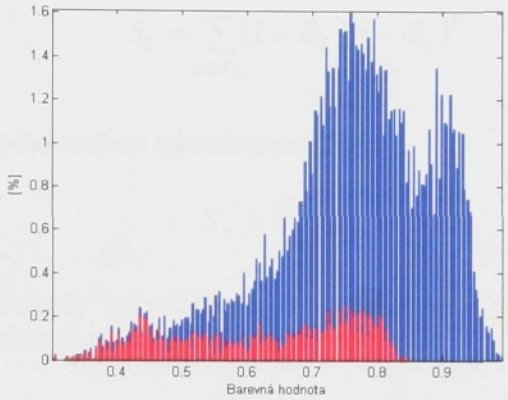
$$S = \frac{MAX - MIN}{MAX} \tag{5.5}$$

$$V = MAX \tag{5.6}$$

Složka H tak nabývá hodnot 0 až 360 – úhel ve stupních z kruhu barev – viz obr. 5.9 a složky S a V nabývají hodnot 0 až 1. Občas se používá normalizace složky H, jejíž hodnota je vydělena 360 a rozsah hodnot je poté 0 až 1. Takto normovaná složka H byla použita pro vytvoření barevného histogramu, viz obr. 5.10.



Obr. 5.10: *Obrazové histogramy barevných složek H (vlevo), S (vpravo) a V (dole) z barevného obrazu oblasti zájmu*



5.1.5 Barevný prostor získaný pomocí FLDA

Jak již bylo popsáno výše, tak použití některého ze známých barevných prostorů je dosti závislé na pořízených obrazech. Při snímání obrazu se velmi často mění osvětlení a kvalita pořízeného obrazu je také dosti závislá na snímací kameře. Transformace barev do jiného barevného prostoru, která se ukázala pro zpracování obrazů jako dobrá, nemusí již být použitelná u obrazů získaných z nahrávek, které byly pořízené v jiných světelných podmínkách nebo jinou snímací kamerou. Tento problém částečně řeší vytvoření vlastního barevného prostoru (vlastní transformace barev) pomocí Fisherovy lineární diskriminační analýzy FLDA [KAU98, CHA99], která se používá i pro nalezení vah jednotlivých příznaků pro rozpoznávání. Výhodou této metody je, že si jednotlivé koeficienty pro lineární transformaci RGB barev vypočítáváme předem a vlastní barevná transformace obrazu je tak velmi rychlá.

Prvním úkolem před vytvořením vlastní barevné transformace pomocí FLDA je zapotřebí vytvořit databázi (DAT_{rk}) obrazů oblastí zájmu se rty od různých mluvčích, pořízené pokud možno stejnou kamerou při různých světelných podmínkách. V těchto obrazech je potřeba co nejpřesněji nalézt rty a kůži a vzájemně je od sebe oddělit. V naší databázi byly vytvořeny z jednoho obrazu oblasti zájmu obrazy dva, v jednom se nacházely pouze rty (objekt) a ve druhém pouze kůže (okolí), ostatní body mají barevnou hodnotu [R G B] = [255 255 255] stejně jako na obr. 5.2. Pro tyto účely je vhodné mít v databázi pouze obrazy oblastí zájmu, ve kterých má mluvčí zavřené rty.

V prvním kroku výpočtu nové barevné transformace se spočítá střední hodnota z každé barevné složky R, G, B ze všech obrazových bodů z databáze obrazů tvořících rty (k = 1) a kůži (k = 2):

$$\vec{m}_k = \frac{1}{n_k} \sum_{x \in X_k} \vec{x} \tag{5.7}$$

kde x je vektor R, G, B barevných hodnot obrazového bodu, X_1 je množina všech obrazových bodů z databáze tvořících rty a X_2 je množina všech bodů tvořících kůži, n_k je počet všech bodů z množiny X_1 nebo X_2. Obrazové body s hodnotou [R G B] = [255 255 255] nejsou započteny.

Poté je spočítána matice rozptylů:

$$S_k = \sum_{x \in X_k} (\vec{x} - \vec{m}_k)(\vec{x} - \vec{m}_k)^T \tag{5.8}$$

Z matic rozptylů se vypočte matice významnosti:

$$S_w = S_1 + S_2 \tag{5.9}$$

Výsledné váhové koeficienty pro barevnou transformaci jsou spočítány z inverzní matice významnosti a vektorů středních hodnot:

$$(\alpha, \beta, \gamma) = S_w^{-1}(\vec{m}_1 - \vec{m}_2) \tag{5.10}$$

Nová barevná hodnota F se vypočte z původních R, G, B hodnot jako:

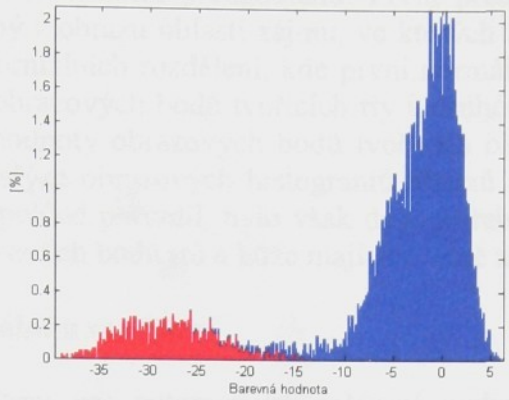
$$F = [\alpha \quad \beta \quad \gamma] \cdot \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{5.11}$$

Výsledná barevná transformace z obrázků pořízených PC web-kamerou LogiTech:

$$F = [-0.289 \quad 0.379 \quad 0.038] \cdot \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{5.12}$$

Po pořízení nových audio-vizuálních nahrávek se ukázalo, že barevná transformace (5.12) získaná pomocí Fisherovy lineární diskriminační analýzy je použitelná i pro nově pořízené obrazy oblastí zájmu s různými mluvčími, přesto při použití jiné kamery je vhodnější vypočíst novou transformaci z nově pořízených obrazů. Výhodou této transformace je, že je lineární a dá se velice snadno a rychle vypočítat. Ještě většího zrychlení pro transformaci RGB barevného obrazu do F složky se dá dosáhnout předem

vytvořenou převodní tabulkou, která je naprogramována na základě vypočtené převodní transformace a převádí tak přímo R, G, B barevné hodnoty do hodnoty F .



Obr. 5.11: *Obrazový histogram barevné složky F, která byla získána barevnou transformací získanou pomocí FLDA*

5.1.6 Vyhodnocení použitých barevných prostorů

V současné době existuje velké množství prací o nalezení a sledování rtů, ve kterých byly použity různé barevné prostory. V předchozích statích jsou popsány některé z nich, které se dnes používají a které jsem postupně použil ve své práci. Jako vhodné transformované složky pro zpracování našich videonahrávek byly: Cr složka z YCbCr, Q složka z YIQ, g složka z rg, H složka z HSV a vlastní F složka. Postupně byly hledány lepší barevné transformace a nakonec byla použita složka F z barevné transformace získané pomocí FLDA. Výhodou této transformace F bylo i to, že se v transformovaném obraze oblasti zájmu dal automaticky nalézt práh pro následnou segmentaci obrazu.

5.2 Automatické nalezení prahu pro segmentaci obrazu ROI

Po F-barevné transformaci obrazu oblasti zájmu je provedena segmentace tohoto obrazu pomocí prahování. Po této segmentaci vznikne binární obraz, ve kterém mají obrazové body hodnotu 0 nebo 1. Obrazové body patřící s jistou pravděpodobností objektu (rty) mají hodnotu 1 a obrazové body z pozadí (kůže, zuby...) mají hodnotu 0. U barevné segmentace obrazu při detekování obličeje byla hodnota prahu pro segmentaci předem experimentálně zjištěna a upravena tak, aby co nejvíce obrazových bodů, které měly barvu kůže, bylo označena hodnotou 1.

Pro segmentaci rtů z obrazu oblasti zájmu by pevné stanovení prahu nebylo příliš vhodné, jelikož v různých nahrávkách se velmi často mění osvětlení a kontrast mezi rty a kůží není ani po barevné transformaci příliš velký. Oproti lidskému obličeji je také plocha rtů v obraze menší a tím je i potřeba nalézt pokud možno co největší množství obrazových bodů, které tvoří rty. Proto je výhodnější a mnohdy i nezbytné stanovit automaticky hodnotu prahu pro každou videonahrávku zvlášť. Jednou z často používaných metod pro automatické nalezení prahu(ů) pro segmentaci obrazu je metoda založená na nalezení prahu(ů) v analyzovaném obrazovém histogramu. Obrazový

histogram je totiž velmi často jediná globální informace, kterou můžeme z pořízeného obrazu získat. Před vytvořením a navržením algoritmu pro automatického nalezení prahu muselo být splněno několik předpokladů. První předpoklad byl, že obrazový histogram (5.11) získaný z obrazu oblasti zájmu, ve kterých má mluvčí zavřené rty, je složen z dat ze dvou normálních rozdělení, kde první normální rozdělení je vytvořeno z F-barevných hodnot obrazových bodů tvořících rty a druhému normálnímu rozdělení odpovídají F-barevné hodnoty obrazových bodů tvořících okolí, kde okolí je tvořeno především kůží. Při analýze obrazových histogramů obrazů oblastí zájmu od různých mluvčích se tento předpoklad potvrdil, bylo však dále potřeba zjistit, zda rozdělení F-barevných hodnot obrazových bodů rtů a kůže mají skutečně normální rozdělení.

5.2.1 Hypotéza o normálním rozdělení

Před navržením algoritmu pro automatické nalezení prahu pro segmentaci obrazu oblasti zájmu bylo potřeba ověřit hypotézu, že se F-barevné hodnoty obrazových bodů tvořících rty nebo kůži řídí normálním rozdělení. Pro tyto účely byl použit parametrický χ^2 test dobré shody [ROG98].

Tento test byl proveden zvlášť pro F-barevné hodnoty obrazových bodů tvořících v obraze rty a pro F-barevné hodnoty obrazových bodů zobrazujících v obraze okolí (kůži). Vstupem pro tento test byly použity jako data F-barevné hodnoty obrazových bodů z obrazů oblastí zájmu z databáze DAT_{rk}, viz stať 5.1.5. Hodnoty testované hypotézy byly počítány pro 10 různých obrazů od 10 mluvčích.

V prvním kroku byly z dat spočítány odhady střední hodnoty $\hat{\mu}_k$ a rozptylu $\hat{\sigma}_k^2$:

$$\hat{\mu}_k = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5.13}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{i=1}^{N}(x_i - \hat{\mu}_k)^2}{N-1} \tag{5.14}$$

kde x_i jsou data – F-barevné hodnoty obrazových bodů tvořících v obraze rty (kůži), N je počet dat – počet obrazových bodů.

Poté bylo vytvořeno k intervalů (k bylo stanoveno na 10) a v těchto intervalech byly spočteny četnosti n_i:

$$n_1, n_2, \ldots, n_k; \quad \sum_{i=1}^{k} n_i = N \tag{5.15}$$

Dále byly vypočteny pravděpodobnosti p_i toho, že výsledek pokusu padne do jednotlivých intervalů $\langle t_{i-1}, t_i \rangle$, $f(x)$ je hustota pravděpodobnosti normálního rozdělení:

$$p_i = \int_{t_{i-1}}^{t_i} f(x)dx = F(t_i) - F(t_{i-1}) \tag{5.16}$$

Vztah (5.16) pro normované normální rozdělení $N(\mu,\sigma^2) = N(0,1)$:

$$p_i = F\left(\frac{t_i - \mu}{\sigma}\right) - F\left(\frac{t_{i-1} - \mu}{\sigma}\right) \tag{5.17}$$

e_i je poté počet výsledků, které teoreticky padnou do intervalu $\langle t_{i-1}, t_i \rangle$:

$$e_i = N.p_i \tag{5.18}$$

Z hodnot e_i a n_i je vypočtena testovaná charakteristika:

$$C = \sum_{i=1}^{k} \frac{(n_i - e_i)}{e_i} \tag{5.19}$$

Hypotézu o normálním rozdělení F-barevných hodnot obrazových bodů tvořících v obraze rty nebo kůži zamítneme na hladině významnosti α pokud je C > kvantil rozdělení $\chi^2_{(1-\alpha)}(sv)$, kde sv je počet stupňů volnosti: sv = k – 1 – q, q je počet odhadnutých parametrů a k je počet intervalů. V našem případě tedy sv = 10–1–2 = 7. Kvantil rozdělení pro 5% hladinu významnosti $\chi^2_{(0.95)}(7) = 14.0671$.

Ze souboru dat – F-barevných hodnot obrazových bodů zobrazujících v obraze oblasti zájmu (pro 10 oblastí zájmu) bylo vypočteno pro rty C_{rty} a pro okolí (kůži) $C_{okolí}$:

obr. č.	1	2	3	4	5	6	7	8	9	10
C_{rty}	10.3	13.4	12.1	11.5	12.9	13.9	13.7	12.3	10.5	12.8
$C_{okolí}$	7.8	11.9	7.5	7.7	8.5	12.3	8.7	7.9	7.5	10.7

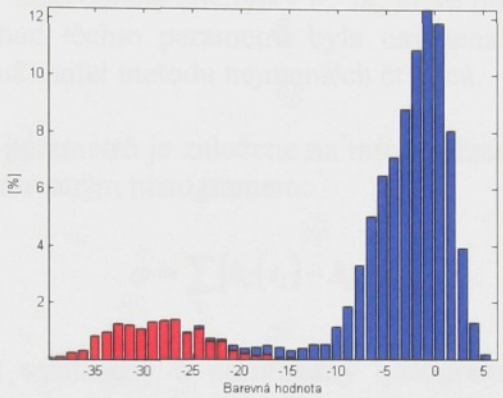
Tabulka 5.1: Vypočtené hodnoty testované charakteristiky z obrazů oblasti zájmu

Hodnoty vypočtené charakteristiky jsou menší než kvantil rozdělení $\chi^2_{(0.95)}(7)$, hypotézu tedy na 5% hladinu významnosti nezamítáme, jedná se tedy s určitou pravděpodobností o výběry z normálních rozdělení.

5.2.2 Algoritmus pro automatické nalezení prahu

Na základě předpokladu, že se obrazový histogram získaný z obrazu oblasti zájmu skládá ze dvou normálních rozdělení byl ve spolupráci s Doc. Ing. Vladimírem Kracíkem, CSc., z katedry aplikované matematiky (TUL) vytvořen algoritmus pro automatické nalezení prahu [CHA04c]. Podobný algoritmus pro nalezení prahu při segmentaci obrazů, jejichž obrazový histogram má tvar přibližně podobný dvoumodálnímu normálnímu rozdělení je popsán i v [SON98].

V prvém kroku byl tedy vypočten obrazový histogram z F barevných hodnot obrazových bodů obrazu oblasti zájmu. Pro výpočet tohoto histogramu se však musel předem stanovit počet intervalů I_p, ve kterých se poté počítaly jednotlivé četnosti. Pokud by měl obrazový histogram více intervalů, byla by výsledná hodnota nalezeného prahu více přesnější. V obraze oblasti zájmu se však nenalézá nekonečně mnoho obrazových bodů, a proto při větším množství intervalů je i výsledný histogram dosti členitý a tím i hůře použitelný, viz obr. 5.11, kde bylo použito 256 intervalů pro výpočet histogramů. Nakonec bylo experimentálně stanoven počet intervalů na 40. Pro tento počet intervalů je již výsledný histogram více „hladký" a nalezený práh z tohoto histogramu je dostatečně přesný pro následnou segmentaci obrazu, viz obr. 5.12.

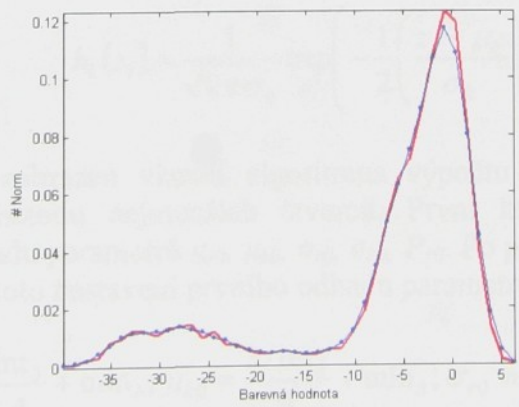


Obr. 5.12: *Obrazový histogram barevné složky F, kde bylo použito pro výpočet histogramu 40 intervalů*

Po výpočtu obrazového histogramu se ukázalo výhodným tento histogram před jeho dalším použitím vyhladit. Pro vyhlazení histogramu byla použita metoda filtrace klouzavým průměrem [HLA00]:

$$h_v(z_i) = \frac{1}{2K+1} \sum_{j=K}^{K} h(z_{i+j}) \tag{5.20}$$

kde $h(z_i)$ jsou původní hodnoty histogramu, $h_v(z_i)$ jsou nové hodnoty histogramu, z_i je příslušný interval a K je velikost filtrovaného okolí.



Obr. 5.13: *Obrazový histogram barevné složky F (červená křivka) a vyhlazený histogram (modrá křivka), na ose y je vynášena normovaná velikost četností histogramu*

Pozn: Obrazové histogramy budou dále pro lepší názornost zobrazovány namísto sloupcových grafů jako spojnicové grafy.

Pro vyhlazení histogramu byla zvolena velikost filtrovaného okolí K = 1 a histogram byl dále znormován tak, aby celková plocha histogramu byla 1 (5.21), viz obr. 5.13.

$$\sum_{i=0}^{Ip} h_v(z_i) = 1 \tag{5.21}$$

Jak již bylo uvedeno výše, výsledný histogram je složen ze dvou normálních rozdělení. Jedno je normální rozdělení $N(\mu_r, \sigma_r^2)$ popisující rozdělení F-barevných hodnot obrazových bodů zobrazujících rty v obraze oblasti zájmu a druhé je normální rozdělení $N(\mu_k, \sigma_k^2)$ F-barevných hodnot obrazových bodů zobrazujících okolí (kůži). Z analýzy vyhlazeného a normovaného histogramu tak chceme zjistit statistické parametry - střední hodnoty μ_r, μ_k a směrodatné odchylky σ_r, σ_k, které budou následně použity pro výpočet prahu. Pro odhad těchto parametrů byla navržena optimalizační gradientní matematická metoda využívající metodu nejmenších čtverců.

Tato metoda pro odhad parametrů je založena na minimalizaci kvadratické odchylky ϕ mezi skutečným a odhadovaným histogramem:

$$\varphi = \sum_i [h_v(z_i) - h_{vo}(z_i)]^2 \tag{5.22}$$

kde $h_v(z_i)$ je originální vyhlazený a normovaný histogram a $h_{vo}(z_i)$ je normovaný odhadovaný histogram.

Odhadovaný histogram je vypočten z parametrů μ_r, μ_k, σ_r, σ_k a P_r (5.23), kde P_r je relativní počet obrazových bodů zobrazujících rty. Relativní počet obrazových bodů zobrazujících okolí (kůži) $P_k = 1 - P_r$.

$$h_{vo}(z_i) = P_r \cdot h_r(z_i) + (1 - P_r) \cdot h_k(z_i) \tag{5.23}$$

$$h_r(z_i) = \frac{1}{\sqrt{2\pi}\sigma_r} \exp\left(-\frac{1}{2} \left(\frac{z_i - \mu_r}{\sigma_r} \right)^2 \right) \tag{5.24}$$

$$h_k(z_i) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2} \left(\frac{z_i - \mu_k}{\sigma_k} \right)^2 \right) \tag{5.25}$$

Na obrázku 5.14 je zobrazen vlastní algoritmus výpočtu optimalizační gradientní metody využívající metodu nejmenších čtverců. První krok tohoto algoritmu je nastavení prvního odhadu parametrů μ_{r0}, μ_{k0}, σ_{r0}, σ_{k0}, P_{r0}. Po provedení několika pokusů byly nakonec zvoleno toto nastavení prvního odhadu parametrů:

$$P_{r0} = 0.2;\ \mu_{r0} = \frac{\text{int}_h}{4} + \min_h;\ \mu_{k0} = \frac{3 \cdot \text{int}_h}{4} + \min_h;\ \sigma_{r0} = \frac{\text{int}_h}{9};\ \sigma_{k0} = \frac{\text{int}_h}{16} \tag{5.26}$$

kde $\text{int}_h = \max_h - \min_h$, $\max_h$ je maximální F-barevná hodnota z obrazu oblasti zájmu, ze které se počítá histogram a $\min_h$ je minimální F-barevná hodnota.

Poté je vypočten gradient:

$$\nabla\varphi = \left[\frac{\partial\varphi}{\partial P_r}, \frac{\partial\varphi}{\partial\mu_r}, \frac{\partial\varphi}{\partial\sigma_r}, \frac{\partial\varphi}{\partial\mu_k}, \frac{\partial\varphi}{\partial\sigma_k}\right] \tag{5.27}$$

Po dosazení rovnice 5.23 do vztahu 5.22 se jednotlivé složky gradientu spočítají jako:

$$\frac{\partial\varphi}{\partial P_r} = 2\sum_i \left(h_v(z_i) - \left(P_r h_r(z_i) + (1-P_r)h_k(z_i)\right)\right)\cdot\left(h_k(z_i) - h_r(z_i)\right) \tag{5.28}$$

$$\frac{\partial\varphi}{\partial\mu_r} = \frac{-2P_r}{\sigma_r^2}\sum_i \left(h_v(z_i) - \left(P_r h_r(z_i) + (1-P_r)h_k(z_i)\right)\right)\cdot(z_i-\mu_r)\cdot h_r(z_i) \tag{5.29}$$

$$\frac{\partial\varphi}{\partial\sigma_r} = \frac{2P_r}{\sigma_r}\sum_i \left(h_v(z_i) - \left(P_r h_r(z_i) + (1-P_r)h_k(z_i)\right)\right)\cdot\left(h_r(z_i) - \frac{(z_i-\mu_r)}{\sigma_r^2}h_r(z_i)\right) \tag{5.30}$$

$$\frac{\partial\varphi}{\partial\mu_k} = \frac{-2(1-P_r)}{\sigma_k^2}\sum_i \left(h_v(z_i) - \left(P_r h_r(z_i) + (1-P_r)h_k(z_i)\right)\right)\cdot(z_i-\mu_k)\cdot h_k(z_i) \tag{5.31}$$

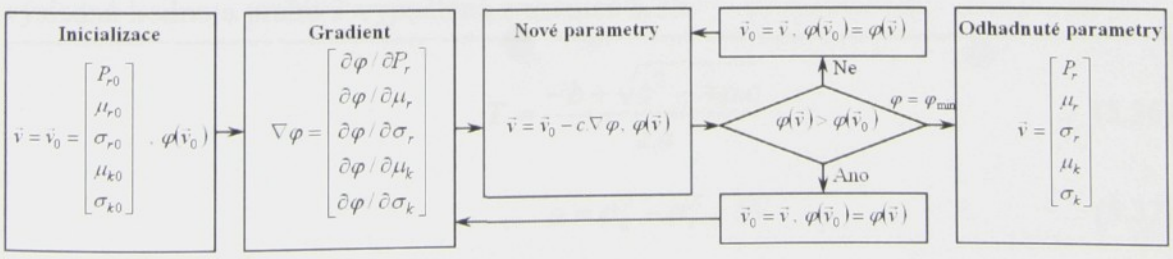
$$\frac{\partial\varphi}{\partial\sigma_k} = \frac{2(1-P_r)}{\sigma_k}\sum_i \left(h_v(z_i) - \left(P_r h_r(z_i) + (1-P_r)h_k(z_i)\right)\right)\cdot\left(h_k(z_i) - \frac{(z_i-\mu_k)}{\sigma_k^2}h_k(z_i)\right) \tag{5.32}$$

Na základě vypočteného gradientu jsou poté vypočteny nové hodnoty parametrů:

$$\vec{v} = \vec{v}_0 - c.\nabla\varphi \tag{5.33}$$

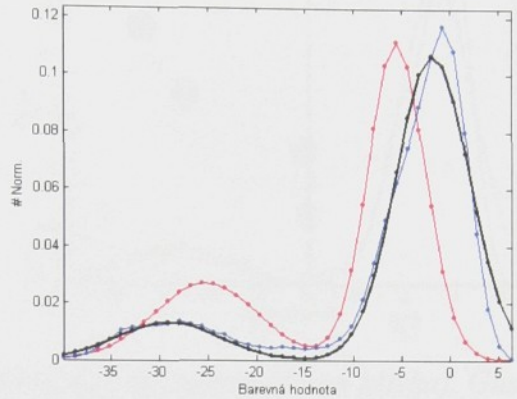
kde $\vec{v}_0 = [P_{r0}, \mu_{r0}, \sigma_{r0}, \mu_{k0}, \sigma_{k0}]$ a $\vec{v} = [P_r, \mu_r, \sigma_r, \mu_k, \sigma_k]$ jsou vektory hodnot parametrů a c je vhodně zvolená konstanta, např. c = 0.1.

Pokud nově vypočtená kvadratická odchylka $\varphi(\vec{v})$ (5.22) je menší než původní odchylka $\varphi(\vec{v}_0)$, tak se provede záměna hodnot $\vec{v}_0 = \vec{v}$, $\varphi(\vec{v}_0) = \varphi(\vec{v})$ a znovu se vypočtou nové hodnoty parametrů (5.33). V opačném případě se provede záměna hodnot vektorů parametrů a kvadratické odchylky a vypočte se nový gradient (5.27).



Obr. 5.14: *Algoritmus výpočtu parametrů pomocí optimalizační gradientní matematické metody využívající metodu nejmenších čtverců*

Vnitřní cyklus algoritmu se dá zastavit, pokud kvadratická odchylka, mezi skutečným a odhadovaným histogramem je minimální. Minimální odchylka by se dala určit předem, ale to by bylo nepraktické, jelikož se vstupní histogram pro každou z nahrávek odlišuje. Výhodnější je nechat proběhnou určitý počet cyklů a v každém cyklu si zapamatovat výsledný vektor parametrů. Z tohoto souboru vektorů parametrů je poté vybrán vektor, u kterého byla kvadratická odchylka nejmenší. Experimentálně bylo zjištěno, že pro dostatečně přesný odhad obrazového histogramu stačí přibližně 3000 cyklů.



Obr. 5.15: *Vyhlazený obrazový histogram (modrá křivka), histogram vypočtený z prvních nastavených parametrů (červená křivka) a výsledný histogram z odhadnutých parametrů (černá křivka)*

Z odhadnutých parametrů μ_r, μ_k, σ_r, σ_k, P_r je následně vypočten práh T v místě, kde se protíná Gaussova křivka vytvořená z odhadnutých parametrů μ_r, σ_r a P_r s Gaussovou křivkou vypočtenou z odhadnutých parametrů μk, σk a P_r, tj.:

$$\frac{P_r}{\sqrt{2\pi}\sigma_r}\exp\left(-\frac{1}{2}\left(\frac{T-\mu_r}{\sigma_r}\right)^2\right)=\frac{(1-P_r)}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{1}{2}\left(\frac{T-\mu_k}{\sigma_k}\right)^2\right) \tag{5.34}$$

po zlogaritmování vztahu 5.34:

$$\log P_r-\log\sigma_r-\frac{1}{2}\left(\frac{T-\mu_r}{\sigma_r}\right)^2=\log(1-P_r)-\log\sigma_k-\frac{1}{2}\left(\frac{T-\mu_k}{\sigma_k}\right)^2 \tag{5.35}$$

výsledná hodnota prahu T vypočtená z rovnice 5.35:

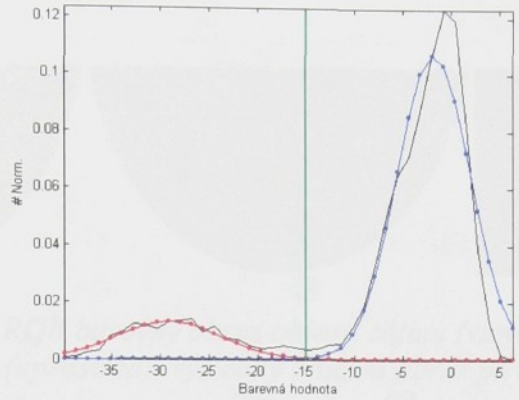
$$T=\frac{-b+\sqrt{b^2-4.a.c}}{2.a} \tag{5.36}$$

$$a=\sigma_k^2-\sigma_r^2 \tag{5.37}$$

$$b=2.\left(\sigma_r^2\mu_k-\sigma_k^2\mu_r\right) \tag{5.38}$$

$$c=\mu_r^2\sigma_k^2-\mu_k^2\sigma_r^2+2.\sigma_k^2\sigma_r^2.\log\left(\frac{\sigma_r(1-P_r)}{\sigma_k P_r}\right) \tag{5.39}$$

Po úpravě vznikne ze vztahu 5.35 kvadratická rovnice, která má dvě řešení. Správnou hodnotu prahu lze však získat pouze ze vztahu 5.36. Hodnota z druhého řešení kvadratické rovnice, kde ve vztahu 5.36 by před odmocninou bylo minus, není pro následnou segmentaci obrazu použitelná.



Obr. 5.16: *Původní obrazový histogram (černá křivka), Gussova křivka vypočtená z parametrů μ_r, σ_r a P_r (červená křivka) a Gussova křivka vypočtená z parametrů μk, σk a P_r (modrá křivka) a výsledný vypočtený práh (zobrazen zeleně)*

Původně byl tento algoritmus navržen pro obrazy oblastí zájmu, ve kterých má mluvčí zavřené rty [CHA04c]. Tento předpoklad byl u našich videonahrávek splněn, jelikož v nahrávkách měl mluvčí na začátku promluvy zavřené rty a práh se hledal pouze v prvním snímku nahrávky, pro ostatní snímky se poté použil tento práh. V jedné videonahrávce bylo pouze jedno slovo nebo jedna věta. Později se však ukázalo, že algoritmus najde docela spolehlivě práh i u snímků, kde má mluvčí otevřená ústa. Vnitřní oblast úst (zuby, jazyk..), která se v obraze objeví, je totiž vzhledem k celému zpracovávanému obrazu oblasti zájmu poměrně malá a tak nezanáší do algoritmu pro nalezení prahu příliš velkou chybu. Tato chyba by se dala ještě dále zmenšit tak, že by se v prvním snímku videonahrávky nalezl práh, provedla by se první segmentace rtů a ze všech segmentovaných snímků by se vybral takový, u kterého by šířka objektu (rtů) byla v horizontálním směru největší. U tohoto snímku se dá předpokládat, že má mluvčí s největší pravděpodobností zavřená ústa. Pro barevný originál tohoto snímku by se poté znovu nalezl práh a tímto prahem by se provedla znovu barevná segmentace. Z provedených pokusů však vyplynulo, že je hodnota prahu nalezená v prvním snímku videonahrávky naprosto dostatečná pro segmentaci dalších snímků z této nahrávky. Nalezení prahu pomocí tohoto algoritmu (při 3000 interakčních cyklech) trvá přibližně 0.07 s na PC Barton 3000+, 1 GB RAM.

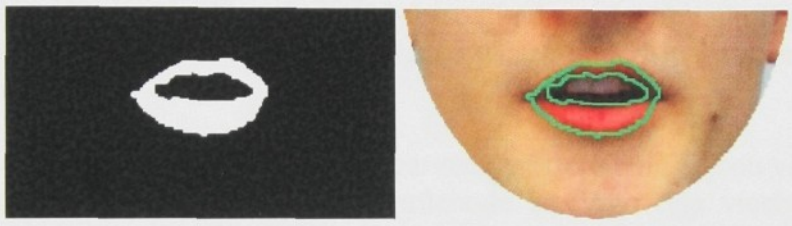
5.3 Segmentace obrazu ROI

Po nalezení vhodné hodnoty prahu T následuje segmentace obrazu prahováním (5.40) , viz obr. 5.17. Z výsledného binárního obrazu jsou poté odfiltrovány všechny shluky obrazových bodů, které se nacházejí v okolí rtů a které byly přesto označeny při prahování hodnotou 1. Pro tuto filtraci je použita morfologická operace otevření (4.10) se strukturním elementem o velikosti 3 x 3 obrazových bodů. Touto operací se také zjednoduší oblast objektu (rtů), viz obr. 5.18.

$$g(i,j)=1 \quad pro\ f(i,j)\le T$$
$$g(i,j)=0 \quad jinak \tag{5.40}$$

kde T je práh, $f(i,j)$ je barevná hodnota obrazového bodu o souřadnicích i, j v původním F barevném obraze, $g(i,j)$ je nová binární hodnota obrazového bodu ve výsledném binárním obraze.

Obr. 5.17: *Originální RGB barevný obraz oblasti zájmu (vlevo), obraz transformovaný do F-barevné složky (uprostřed), výsledný binární obraz po segmentaci prahováním (vpravo)*

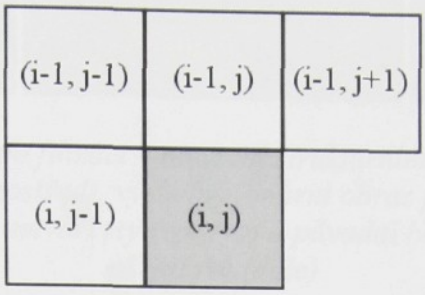


Obrázek 5.18: *Binární obraz po filtraci – otevření strukturním elementem (vlevo) a původní barevný obraz s nalezenými okraji rtů (vpravo)*

Na obrázku 5.17 je zobrazen výsledný binární obraz po segmentaci prahováním, u tohoto obrazu je patrné, že hodnotou 1 byly označeny i některé obrazové body, které netvoří rty. V následujícím kroku byly tyto body odstraněny pomocí filtrace založené na morfologické operaci otevření. Ve výsledném binárním obraze je tak pouze jeden objekt, který reprezentuje rty. Na lidském obličeji se však můžou v blízkosti rtů nacházet další souvislé oblasti obrazových bodů, které sice netvoří rty, ale přesto mají červený barevný odstín a tím se i objeví po segmentaci v binárním obraze. Tyto oblasti jsou většinou různé „defekty" kůže (jizvy, kožní vyrážky, akné...) a velmi často to jsou i oblasti obrazových bodů, které zobrazují nosní dírky. Použít vetší strukturní element pro odfiltrovaní těchto oblastí se ukázalo nepraktické a mnohdy i nemožné, jelikož by se tím zkreslila oblast nalezených rtů nebo by byla výrazně deformována. Proto pro odstranění těchto oblastí byla použitá metoda barvení oblastí [SON98].

Pomocí metody barvení oblastí se vytvoří víceúrovňový barevný obraz, kde obrazové body tvořící oblast mají stejnou číselnou hodnotu (barvu). Algoritmus této metody je následující: V prvním kroku je binární obraz procházen po řádcích od shora dolů (pokud jsou nulové souřadnice (0, 0) obrazu (matice) v levém horním rohu obrazu). Každému obrazovému bodu , který má nenulovou hodnotu f (i, j) > 0 je přiřazena nová hodnota v závislosti na hodnotách, které se nacházejí v okolí tohoto bodu. Okolí bodu (sousednost) je dáno maskou. Pro zpracování našich obrazů byla použita maska pro

8 - okolí bodu, viz obr. 5.19. Maska je zvolena tak, aby všechny obrazové body dané maskou byly již v z předchozích krocích obarveny.



Obr. 5.19: *Použitá maska pro barvení oblastí*

Pokud mají všechny obrazové body z okolí obrazového bodu nulovou hodnotu, je obrazovému bodu f (i, j) přiřazena nová číselná hodnota. Pokud má právě jeden bod z okolí nenulovou hodnotu je obrazovému bodu f (i, j) přidělena tato hodnota a pakliže se v okolí bodu f (i, j) nachází více bodů s nenulovou hodnotou, tak je přiřazena bodu f (i, j) kterákoliv z hodnot bodů z okolí. Když jsou hodnoty bodů v okolí různé (dochází ke kolizi barev), tak se navíc musí ekvivalence těchto hodnot zaznamenat do tabulky ekvivalence barev.

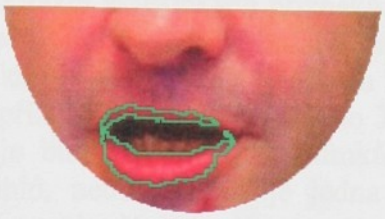
Po prvním průchodu jsou všechny oblasti obarvené, některé oblasti jsou však díky kolizi barev obarvené více barvami. V druhém průchodu se proto tyto vícebarevné oblasti obarví stejnou hodnotou na základě informace z tabulky ekvivalence barev.

V průběhu barvení oblastí je zároveň spočítán počet obrazových bodů náležející jednotlivým oblastem. Při zpracování obrazů ROI jsou oblasti, které mají menší počet obrazových bodů, než je předem stanovená hodnota, z binárního obrazu odstraněny. Tato metoda je limitována velikostí oblasti rtů, jelikož pokud jsou oblasti z okolí rtů větší než oblast rtů, tak nemohou být z binárního obrazu odstraněny. Tento případ je však dosti specifický a je dosti nepravděpodobné, že by se v obraze s „průměrnými" osobami objevily tak velké „patologické" oblasti.

Případ, kdy se v binárním obraze objevilo více oblastí je na obrázcích 5.20-21, kde byl použit obraz ROI od jiného mluvčího než pro obr. 5.17-18 a který byl pořízen v odlišných světelných podmínkách.

Obr. 5.20: *Originální RGB barevný obraz oblasti zájmu (vlevo), obraz transformovaný do F-barevné složky (uprostřed), výsledný binární obraz po segmentaci prahováním (vpravo)*

Obr. 5.21: *Binární obraz po filtraci – otevření strukturním elementem (vlevo), obraz s obarvenými oblastmi (uprostřed), výsledný binární obraz po odstranění oblastí, které s jistou pravděpodobností netvoří rty (vpravo) a původní barevný obraz s nalezenými okraji rtů (dole)*



Celkově byly po použití algoritmu barvení oblastí nalezeny čtyři oblasti označené čísly 1 až 4. Pro větší názornost byly tyto hodnoty obrazových bodů dodatečně nahrazeny RGB barvami 1 – modrá, 2 – zelená, 3 – červená, 4 – žlutá. Obrys rtů v původním barevném obraze vznikl stejně jako obrys detekovaného obličeje v obraze (kapitola 4), tj. jako rozdíl výsledného binárního obrazu a binárního obrazu (výsledného), který byl erodován strukturním elementem 5 x 5. Obrys je pro vetší kontrast zobrazen zeleně.

Kompletní navržená metoda pro nalezení rtů se skládá ze tří částí. Nejdříve je RGB barevný obraz nalezené oblasti zájmu (kapitola 4) převeden do F-barevné složky lineární transformací barev získanou z Fisherovy lineární diskriminační analýzy Poté je automaticky nalezen práh pro segmentaci obrazu a je provedena segmentace prahováním. Nakonec jsou z binárního obrazu odfiltrovány morfologickou operací otevření všechny drobné objekty a následně jsou odstraněny větší objekty z pozadí s využitím metody pro barvení objektů.

Takto navržená metoda pro nalezení oblasti rtů je dostatečně spolehlivá a zároveň i rychlá. Vytvoření výsledného binárního obrazu (pokud je vypočtena hodnota prahu pro segmentaci), ve kterém se nachází pouze rty trvá přibližně 0.01 s na PC Barton 3000+, 1 GB RAM. Automatické nalezení prahu pomocí výše popsané optimalizační gradientní metody využívající metodu nejmenších čtverců (při 3000 interakčních cyklech) pak trvá přibližně 0.06 s, práh se počítá pouze pro první snímek z nahrávky, pro segmentaci ostatních snímků z nahrávky se používá stejný práh.

V příloze č. 2 jsou zobrazeny zpracované obrazy oblasti zájmu s nalezenými oblastmi (obrysy) rtů pro různé mluvčí z naší audio-vizuální databáze.

Kapitola 6

Vizuální příznaky řeči

Obdobně, jako je tomu pro extrakci akustických příznaků, tak i pro extrakci vizuálních příznaků řeči existuje v dnešní době větší množství metod. Nejčastěji se vizuální příznaky extrahují z obrazů, kde je mluvčí čelně natočen ke snímací kameře, nověji se zkouší i extrakce vizuálních příznaků z aproximovaného 3D prostoru [GOE01, CIS04]. Aproximovaný 3D prostor je buď vytvořen ze snímků dvou kamer, které snímají mluvčího z dvou různých úhlů, nebo se použije jedna kamera a vhodně umístěné zrcadlo (popř. více zrcadel) a mluvčí je poté zaznamenán v jednom video snímku z dvou (i více) různých úhlů. Vlastní vizuální příznaky pro rozpoznávání řeči lze rozdělit do dvou kategorií:

- Tvarové vizuální příznaky
- Vizuální příznaky popisující informační obsah obrazu

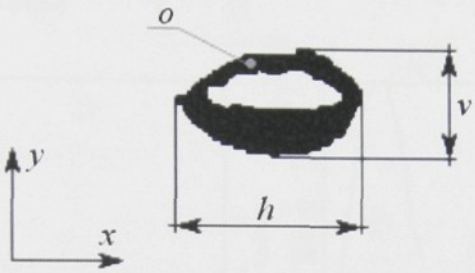
Tvarové příznaky, někdy též nazývané geometrické příznaky, jsou získány z nalezeného tvaru objektu rtů. Nejpoužívanější geometrické příznaky jsou: horizontální a vertikální rozšíření rtů, velikost oblasti rtů a zaokrouhlení rtů [PET84, GAO00 ,POT04]. Další tvarové příznaky lze vytvořit z analyzované hranice obrysu rtů [CHA99, KRO02], kde je buď sledován směr hrany obrysu rtů nebo je spodní a horní část tvaru rtů aproximována vhodnou křivkou, např. parabolou. Od použití příznaků z analyzované hranice rtů se však již v poslední době ustupuje, jelikož hlavním předpokladem pro jejich využití je použití kvalitních videonahrávek, kde obraz není příliš zatížen šumem a dalšími poruchami (hranice rtů musí být spolehlivě nalezená).

Dnes nejčastěji používanými vizuální příznaky jsou příznaky popisující informační obsah obrazu. Nejtriviálnější řešení by bylo použít jako příznaky přímo barevné hodnoty obrazových bodů z oblasti zájmu, to však není příliš praktické ani použitelné řešení, jelikož u oblasti zájmu o velikosti 128 x 128 bychom dostali pro jeden video snímek 16 384 příznaků. V takto rozsáhlého vektoru příznaků by bylo veliké množství různé informace o obraze a bylo by velice obtížné a časově náročné vytvořit spolehlivé modely, například metodou HMM. Proto se vlastní obraz oblasti zájmu transformuje některou vhodnou transformací a z transformovaného obrazu se vyberou pouze složky, které dobře reprezentují pořízený obraz. Dnes nejpoužívanějšími transformacemi pro nalezení vizuálních příznaků jsou: diskrétní kosinová transformace DCT (the Discrete Cosine Transform [DUC94, POT01, HEC02, SCA03, SCA04], PCA (the Principal Component Analysis) [BRE94,NET01, POT04]. Zřídka se používají i jiné obrazové transformace, například diskrétní vlnková transformace DWT (the Discrete Wavelet Transform) [MAT01] .

Existuje i několik prací, kde byly zkombinovány tvarové vizuální příznaky s vizuálními příznaky popisující informační obsah obrazu [CHN01, DUP00]. Ve své práci jsem použil geometrické a DCT vizuální příznaky, které jsou následně popsány.

6.1 Geometrické příznaky

Pro vytvoření vektoru geometrických příznaků byl použit binární obraz nalezené oblasti rtů (kapitola 5), viz obr. 6.1 (pro účely zobrazení byl obraz barevně invertován). Objektu rtů tak náleží obrazové body s hodnotou jedna (zde černá barva) a obrazové body z okolí mají hodnotu nula (bílá barva).



Obr. 6.1: *Binární obraz oblasti rtů s vyznačenými geometrickými příznaky*

Za geometrické příznaky byly vybrány: horizontální h (6.1) a vertikální v (6.2) rozšíření rtů, dále oblast rtů o (6.3) a zaokrouhlení rtů r (6.4), které nabývá největších hodnot při vyslovování fonémů u, o, ř.

$$h = \max_{y=0..N-1} \sum_{x=0}^{M-1} f(x,y) \tag{6.1}$$

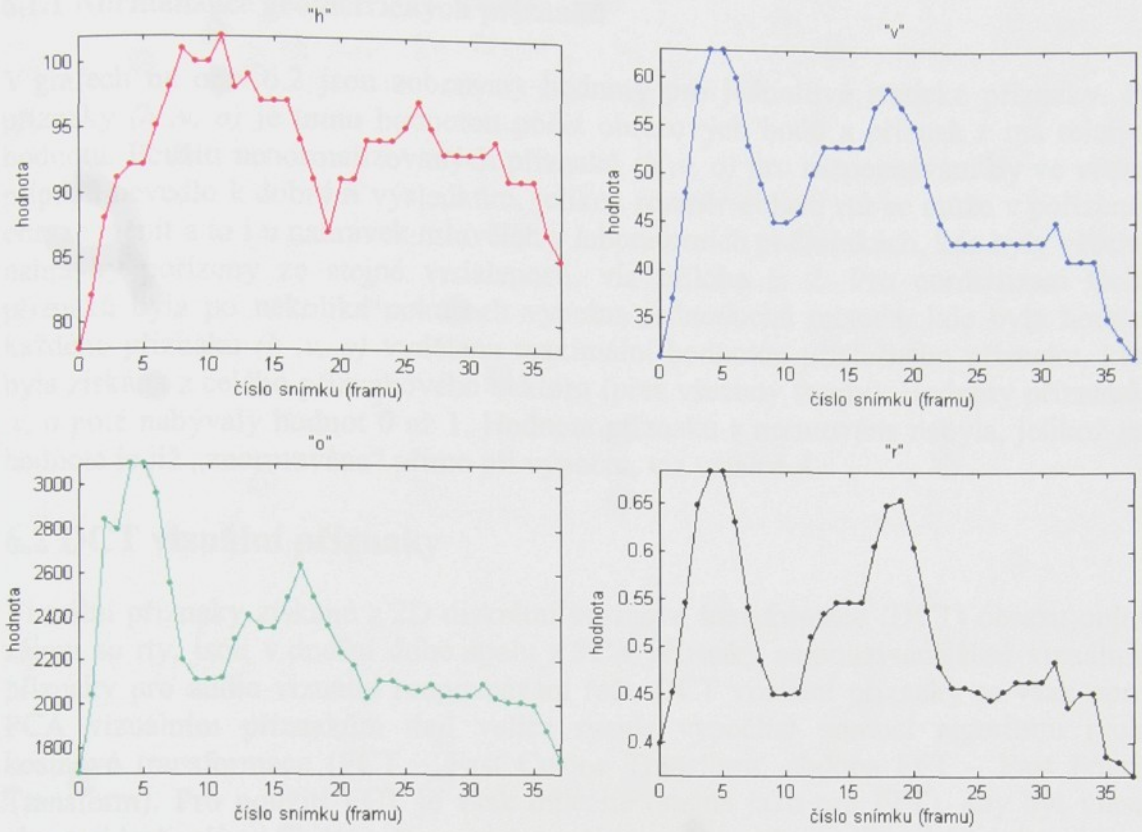
$$v = \max_{x=0..M-1} \sum_{y=0}^{N-1} f(x,y) \tag{6.2}$$

$$o = \sum_{y=0}^{N-1} \sum_{x=0}^{M-1} f(x,y) \tag{6.3}$$

$$r = v / h \tag{6.4}$$

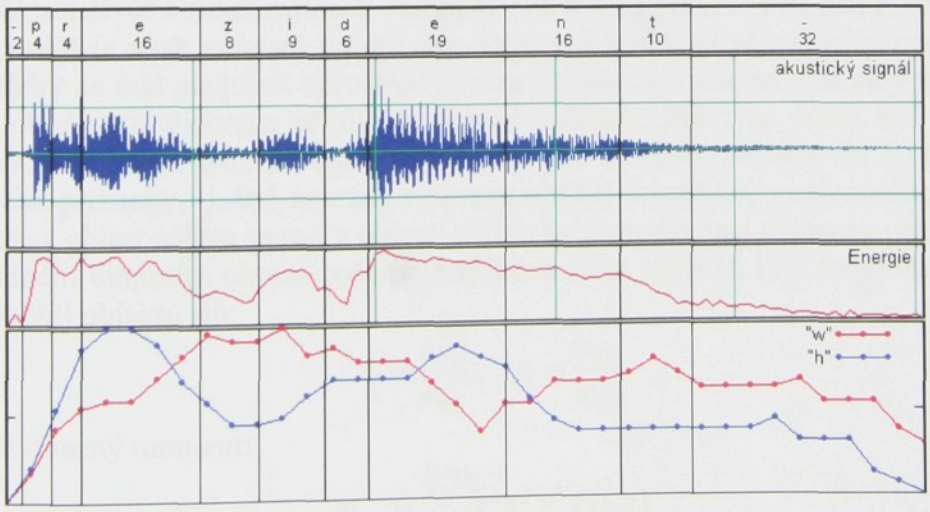
kde (x,y) jsou souřadnice obrazového bodu v binárním obraze nalezené oblasti rtů, $f(x,y)$ je obrazová funkce binárního obrazu nabývající hodnot 0, 1. M x N jsou rozměry binárního obrazu v obrazových bodech a max je funkce maxima.

Občas se geometrické příznaky získávají i z vnitřní oblasti úst, zde však musíme mít poměrně kvalitní nahrávky a daný mluvčí musí dostatečně artikulovat, navíc se ve vnitřní oblasti můžou (ale i nemusí) objevovat zuby a jazyk, který má leckdy podobnou barvu, jako je barva rtů, čímž jsou často při segmentaci obrazové body tvořící v obraze jazyk označeny hodnotou 1. Pro zjištění geometrických příznaků z vnitřní oblasti úst by se musel použít složitější (a tím i časově výpočetně náročnější) algoritmus pro segmentaci rtů a ani tento postup by nemusel vést ke spolehlivému nalezení příznaků.



Obr. 6.2: *Průběhy hodnot vizuálních příznaků (h ,v, o, r) pro slovo „prezident"*

Na obr. 6.2 jsou zobrazeny změny vypočtených hodnot příznaků (*h ,v, o, r*) v čase pro slovo „prezident". Slovo je tvořeno 38 video snímky (framy) při snímkovací frekvenci 30 snímků za sekundu, tj. délka jednoho framu je 33,33 ms, pro lepší představu jsou příznakové vektory (*h ,v*) zobrazeny společně s akustickým signálem u kterého byly nalezeny časové hranice jednotlivých fonémů, viz obr. 6.3.



Obr. 6.3: *Časové hranice jednotlivých fonému pro slovo „prezident", příslušný akustický signál, jeho energie a příslušné hodnoty vizuálních příznaků (h ,v). V horní části obrázku je fonetický přepis slova prezident, znak „-" označuje ticho, pod jednotlivými fonémy jsou zaznamenány časové délky fonémů v desítkách ms.*

6.1.1 Normalizace geometrických příznaků

V grafech na obr. 6.2 jsou zobrazeny hodnoty pro jednotlivé statické příznaky. Pro příznaky *(h ,v, o)* je touto hodnotou počet obrazových bodů a příznak *r* má relativní hodnotu. Použití nenormalizovaných příznaků *(h ,v, o)* pro rozpoznávání by ve většině případů nevedlo k dobrým výsledkům, jelikož rozměr oblasti rtů se může v pořízeném obraze měnit a to i u nahrávek mluvčích v laboratorních podmínkách, kde byly všechny nahrávky pořízeny ze stejné vzdálenosti, viz příloha č. 2. Pro normalizaci těchto příznaků byla po několika pokusech vybrána jednoduchá metoda, kde byla hodnota každého příznaku *(h ,v, o)* vydělena maximální hodnotou příslušného příznaku, která byla získána z celého příznakového vektoru (přes všechny framy). Hodnoty příznaků *h ,v, o* poté nabývaly hodnot 0 až 1. Hodnota příznaku *r* normována nebyla, jelikož tato hodnota je již „znormována" přímo při výpočtu, viz vztah 6.4.

6.2 DCT vizuální příznaky

Vizuální příznaky získané z 2D diskrétní kosinové transformace (DCT) obrazu oblasti zájmu se rty, jsou v dnešní době spolu z PCA příznaky nejpoužívanějšími vizuálními příznaky pro audio-vizuální rozpoznávání řeči. DCT vizuální příznaky se však oproti PCA vizuálním příznakům dají velice rychle vypočítat pomocí algoritmu rychlé kosinové transformace (FCT – Fast Cosine Transform, obdoba FFT – Fast Fourier Transform). Pro použití FCT je však důležité (stejně jako pro FFT), aby byl vlastní obraz oblasti zájmu nejlépe čtvercový a měl rozměr stran o velikosti 2^n, kde n je celé kladné číslo, pokud tato podmínka není splněna, tak se musí obraz na tuto velikost aproximovat. Proto je vhodné rovnou vytvářet obrazy oblasti zájmu, které by tuto podmínku splňovaly.

6.2.1 Vytvoření oblasti zájmu se rty pro FCT

Při analyzování obrazů mluvčích z videonahrávek databáze AVDB2cz (kapitola 7) jsem se rozhodl používat oblasti zájmu o velikosti 128 x 128 obrazových bodů. Pro využití DCT příznaků je však zároveň nutné, aby obraz rtů v oblasti zájmu byly normovány. Objekt rtů by se měl nacházet uprostřed obrazu oblasti zájmu a pro všechny mluvčí by měl mít objekt rtů v obraze přibližně stejnou velikost. Pro tyto účely byl vytvořen algoritmus, kde první krok probíhal stejně jako pro nalezení objektu rtů v obraze pro geometrické příznaky, tj. byl nalezen v obraze obličej mluvčího, z nalezeného obličeje byla vybrána oblast zájmu se rty, v oblasti zájmu byla segmentací nalezena oblast rtů.
Ve výsledném binárním obraze byla poté zjištěna šířka rtů h (6.1) a souřadnice těžiště x_t, y_t [SON98] objektu rtů:

$$x_t = \frac{m_{10}}{m_{00}}, \quad y_t = \frac{m_{01}}{m_{00}} \tag{6.5}$$

kde m_{pq} je obecný moment:

$$m_{pq} = \sum_{y=0}^{N-1} \sum_{x=0}^{M-1} x^p y^q f(i,j) \tag{6.6}$$

kde (x,y) jsou souřadnice obrazového bodu v binárním obraze nalezené oblasti rtů, $f(x,y)$ je obrazová funkce binárního obrazu nabývající hodnot 0, 1. M x N jsou rozměry binárního obrazu v obrazových bodech.

Pro zmenšení (zvětšení) obrazu s objektem rtů, bylo stanoveno, že výsledná šířka objektu rtů v obraze oblasti zájmu (128 x 128) je 100 obrazových bodů. Ze zjištěné šířky h objektu rtů byl poté určen koeficient zvětšení obrazu kz:

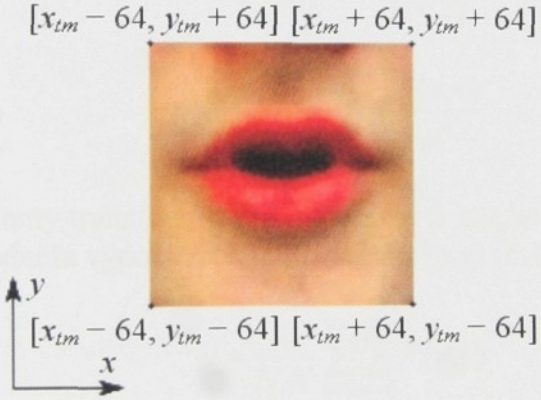
$$kz = h/100 \tag{6.7}$$

Dle koeficientu zvětšení (zmenšení) byl následně zvětšen (zmenšen) původní obraz. Pro zvětšení (zmenšení) obrazu byla použita geometrická transformace obrazu pro změnu měřítka [SON98]. Zároveň se změnou měřítka obrazu byly přepočítány souřadnice těžiště:

$$x_{tn} = x_t \cdot kz, \quad y_{tn} = y_t \cdot kz \tag{6.8}$$

kde x_{tn}, y_{tn} jsou nové souřadnice těžiště ve zvětšeném (zmenšeném) obraze a x_t, y_t jsou původní souřadnice těžiště.

Na základě nových souřadnic těžiště x_{tn}, y_{tn} byla vybrána ze zvětšeného (zmenšeného) obrazu oblast zájmu o velikosti 128 x 128 obrazových bodů, viz obr. 6.4.



Obr. 6.4: *Výsledný obraz oblasti zájmu (128 x 128) pro výpočet DCT viz příznaků*

6.2.2 2D Diskrétní kosinová transformace DCT

V současné době existuje několik různých definicí diskrétní kosinové transformace, ve své práci jsem využil algoritmus rychlé 2D kosinové transformace FCT, vycházející z definice DCT-II (6.9), která se pro zpracování obrazu používá nejčastěji [SON98].

$$F(u,v) = \frac{2c(u)c(v)}{N} \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} f(m,n) \cos\left(\frac{2m+1}{2N} u\pi\right) \cos\left(\frac{2n+1}{2N} v\pi\right) \tag{6.9}$$

kde $f(m,n)$ jsou hodnoty z původního obrazu o rozměrech N x N obrazových bodů, $F(u,v)$ jsou koeficienty transformovaného obrazu, $0 \leq u, v \leq N-1$ a c jsou koeficienty (6.10).

$$c(k) = \begin{cases} \frac{1}{\sqrt{2}} & pro\ k = 0 \frac{1}{\sqrt{2}} \\ 1 & pro\ k > 0\ 1 \end{cases} \tag{6.10}$$

6.2.3 Výběr a výpočet vizuálních příznaků z DCT

Použít celý transformovaný prostor DCT koeficientů jako vizuální příznaky by bylo značně problematické a časově výpočetně náročné, proto se v tomto prostoru hledá menší prostor příznaků, které poté slouží k vlastnímu rozpoznávání. Pro účely výběru těchto příznaků existuje dnes několik metod. Nejpoužívanější metody jsou založené na výpočtu energie *E* (6.11), rozptylu *R* (6.12) a normovaného rozptylu *NR* (6.13) z DCT koeficientů [KRO02]. Z těchto přepočtených koeficientů je poté vybíráno jako vizuální příznaky *P* koeficientů, které mají nejvyšší hodnotu. Novou metodou je pak metoda extrakce DCT vizuálních příznaků založená na vzájemné informaci [SCA04].

$$E(u,v) = F(u,v)^2 \tag{6.11}$$

$$R(u,v) = \frac{(F(u,v) - \mu)^2}{N^2 - 1} \tag{6.12}$$

$$NR(u,v) = \frac{R(u,v)}{\mu^2} \tag{6.13}$$

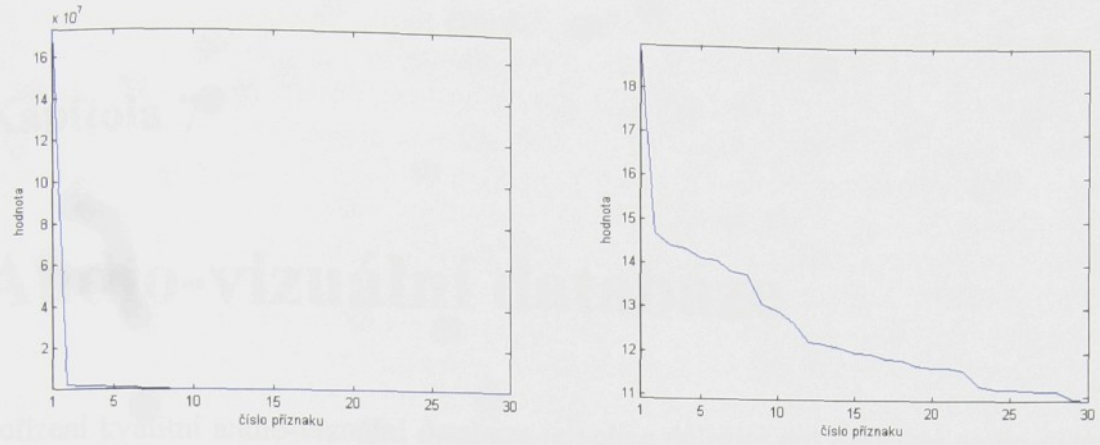
kde $F(u,v)$ jsou koeficienty transformovaného obrazu, $0 \leq u, v \leq N-1$, N x N je rozměr obrazu a μ je střední hodnota vypočtená z koeficientů $F(u,v)$ (6.14).

$$\mu = \frac{1}{N.N} \sum_{v=0}^{N-1} \sum_{u=0}^{N-1} F(u,v) \tag{6.14}$$

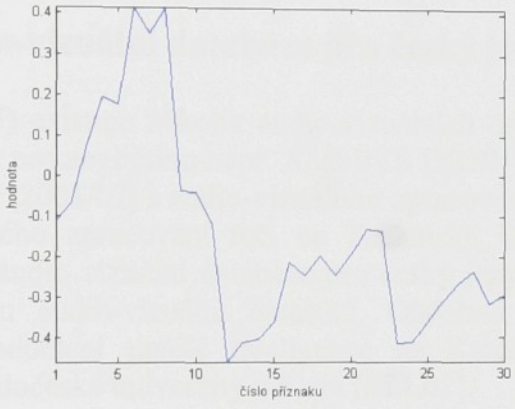
6.2.4 Normalizace DCT příznakového vektoru

Jak již bylo uvedeno v předchozí stati, tak z přepočtených koeficientů energie (rozptylu nebo normovaného rozptylu (6.11-13)) je vybráno za vizuální příznaky *P* koeficientů, které mají nejvyšší hodnotu. Úrovně vypočtených hodnot DCT vizuálních příznaků se stoupajícím indexem razantně klesají, což je nejpatrnější především u energie, viz obr. 6.5. Pro potlačení tohoto jevu se používají různé aproximační metody, které vyrovnávají úrovně hodnot jednotlivých příznaků.

V mé práci se mi nejvíce osvědčilo zlogaritmování (přirozeným logaritmem) všech hodnot z vizuálního příznakového vektoru. Další velmi vhodnou úpravou příznakového vektoru je odečtení střední hodnoty příznakového vektoru, obdobně jako ve vztahu (3.13), čímž se eliminuje různá střední hodnota vizuálního příznakového vektoru z jednotlivých videonahrávek.



Obr. 6.5: *Příznakový vektor tvořený třiceti nejvyššími hodnotami energie (vlevo) z DCT, zlogaritmovaný příznakový vektor (vpravo)*



Obr. 6.6: *Zlogaritmovaný příznakový vektor po odečtení střední hodnoty příznaků*

6.3 Dynamické vizuální příznaky

Samostatné statické geometrické a DCT vizuální příznaky, obdobně jako je tomu u akustických příznaků, nevedou při jejich použití pro rozpoznávání řeči k velkému rozpoznávacímu skóre. Proto se z geometrických a DCT vizuálních příznaků počítají dynamické příznaky. Pro výpočet dynamických příznaků se používají stejné vztahy jako pro výpočet dynamických příznaků u akustického signálu řeči (vztahy 3.14-16). Experimentálně bylo zjištěno, že nejlepší výsledky lze dosáhnout při výpočtu delta a akceleračních příznaků použitím jednoduché diference (3.14) a to jak pro geometrické, tak i DCT vizuální příznaky.

Kapitola 7

Audio-vizuální databáze

Pořízení kvalitní audio-vizuální databáze je velice důležité pro následné audio-vizuální zpracování a rozpoznávání řeči. Požadavky na vytvoření této databáze jsou obdobné, jako u vytvoření databáze pro akustické zpracování a rozpoznávání signálu řeči [POL02], tj. v databázi by se měly nacházet kvalitní nahrávky od velkého množství mluvčích a pokrývající co nejvíce jazyk, pro který je databáze vytvářena.

7.1 Vytvoření audio-vizuální databáze pro český jazyk

V současné době (2005) existuje několik audio-vizuálních databází pro různé národní jazyky, především však pro angličtinu např. XM2VTS [XDB] nebo AVICAR [LEE04]. V době prvních pokusů (2001/02) s audio-vizuálním zpracováním a rozpoznáváním řeči v Laboratoři počítačového zpracování řeči na Technické Univerzitě v Liberci však neexistovala dostupná audio-vizuální databáze pro český jazyk. Proto jsem se rozhodl vytvořit vlastní českou audio-vizuální databázi. Obdobná audio-vizuální databáze vznikla zároveň i v oddělení umělé inteligence na Katedře kybernetiky Fakulty aplikovaných věd Západočeské univerzity v Plzni [ZEL02].

7.1.1 Navržení slovníku a souboru vět

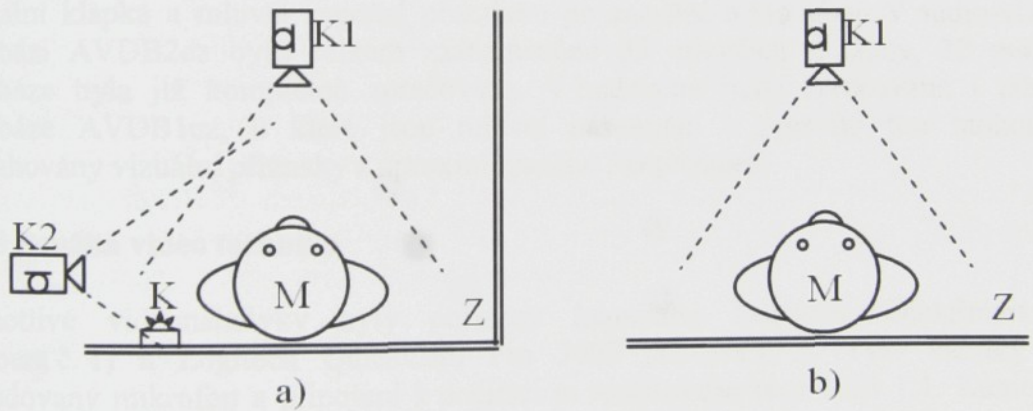
Před vlastním vytvořením videonahrávek musel být připraven slovník a soubor vět, které poté různí mluvčí namluvili. Nahrávky slov slouží pro natrénování a rozpoznávání izolovaných slov a nahrávky vět jsou použity pro rozpoznávání spojité řeči. Pro rozpoznávání izolovaných slov i pro rozpoznávání spojité řeči by bylo nejvhodnější, kdyby se v audio-vizuální databázi nacházelo pokud možno co nejvíce slov a foneticky bohatých vět. Tento požadavek však byl omezen skutečností, že jako mluvčí byli získáváni dobrovolníci, a proto nemohlo vlastní nahrávání trvat příliš dlouhou dobu. Zároveň delší nahrávání způsobuje únavu mluvčích a tím vzniká více chyb a také se postupně mění hlas osob. Jako optimum bylo nakonec vybráno padesát slov a padesát vět. Při takto zvoleném slovníku a souboru vět trvalo vytvoření nahrávek pro jednoho mluvčího přibližně 30 - 45 minut. Pokud se mluvčí přeřekl, tak byl požádán, aby slovo (větu) zopakoval znovu. Tím se zjednodušilo i následné zpracování nahrávek, jelikož chybovost byla omezena již na začátku.

Vlastní slovník padesáti slov byl sestaven ze slov, která byla náhodně vybrána z našeho textového korpusu. Jednalo se především o slova, která měla v tomto korpusu velkou četnost a která měla přibližně stejnou délku, viz příloha č. 3.

Pro vytvoření souboru vět bylo nejdříve shromážděno 100 vět různých českých přísloví. Přísloví byla vybrána, protože je většina lidí zná a dobře se vyslovují. Oproti prvotním pokusům, kde byly vybrány věty z novinových článků se u přísloví výrazně snížila chybovost způsobená různými přeřeknutími. Ze 100 vět bylo poté vybráno 37 vět, které byly foneticky bohaté. Pro tyto účely byl vytvořen program, který nejprve pro každou větu vytvořil její fonetickou transkripci [KOL01] a poté byly porovnány všechny kombinace těchto fonetických přepisů vět. Nakonec byla vybrána ta kombinace vět, v níž se nacházelo co největší množství výskytů jednotlivých českých fonémů. Některé speciální fonémy však v tomto souboru vět byly i tak nedostatečně zastoupeny. Konečný požadavek tedy byl, aby ve výsledném souboru 50 vět byl každý český foném zastoupen alespoň 6 x, proto bylo přidáno k 37 větám přísloví dalších 13 vět, ve kterých se tyto fonémy vyskytují. Ve vlastní vytvořené audio-vizuální databázi je poté jedno slovo nebo věta reprezentována jednou videonahrávkou. Tabulka použité fonetické abecedy s četnostmi jednotlivých fonémů ze souboru 50 vět, viz příloha č. 4.

7.1.2 Pořízení videonahrávek

Postupem času byly vytvořené dvě audio-vizuální databáze později označené AVDB1cz a AVDB2cz. Pro vytvoření audio-vizuální databáze AVDB1cz byly použity dvě kamery, kde jedna kamera snímala obličej mluvčího ze předu a druhá kamera z profilu, viz obr. 7.1. Rozlišení video snímků v této databázi je 240 (šířka) x 320 (výška) obrazových bodů.



Obr. 7.1: *Umístění kamer a nastavení scény: a) pro audio-vizuální databázi AVDB1cz, b) pro audio-vizuální databázi AVDB2cz, K1,2 – kamera č. 1,2, K – audio-vizuální klapka, M – mluvčí, Z – zástěna (pozadí)*

Pro synchronizaci nahrávek z těchto dvou kamer byla ve scéně s mluvčím umístěna audio-vizuální klapka – jednoduché elektronické zařízení složené z bzučáku a 6 mm čiré LED diody svítící zeleně. LED dioda byla elektronicky spojena s bzučákem a toto zařízení bylo připojeno k výstupu paralelnímu portu PC. Rozsvícení LED diody bylo zaznamenáváno v obraze obou použitých kamer. Na začátku nahrávky (před začátkem promluvy) byly do této audio-vizuální klapky poslány z PC tři pulsy dlouhé 0.5 s, pauza mezi pulsy byla 0.5 s, tím se 3 x ozval tón bzučáku a 3 x se rozsvítila LED dioda. Po zaznění těchto tří tónů začal mluvčí mluvit.

Vzhledem k tomu, že ve výsledné videonahrávce byl vizuální signál synchronizován s akustickým signálem a zároveň jsou ve videonahrávkách z obou kamer zaznamenávány v akustickém signálu na začátku stejné pulsy, tak lze poté na základě akustických signálů synchronizovat obě videonahrávky. Jelikož audio-vizuální klapka byla umístěna ve scéně na předem daném místě a LED dioda byla čirá (svítící zeleně), tak by se teoreticky dala synchronizace obou nahrávek provést i na základě vyhodnocení obrazových dat.

Použitá snímkovací frekvence obou kamer však byla 30 snímků za sekundu oproti 22050 vzorků za sekundu u akustického signálu, tím by se tedy mohla zavést do synchronizace obou nahrávek podstatně větší chyba, než při využití akustického signálu. LED dioda však byla u této klapky použita záměrně, aby se ověřil předpoklad, že je vizuální a akustická stopa skutečně ve výsledné videonahrávce synchronizována, tento předpoklad se později při analýze nahrávek potvrdil. Celkově bylo v databázi AVDB1cz zaznamenáno 52 mluvčích (5 žen, 47 mužů), uspořádání pracoviště pro vytvoření a-v nahrávek viz příloha č. 5.

Po několika pokusech při zpracování nahrávek z databáze AVDB1cz se ukázalo, že pro základní výzkum a navržení metody pro extrakci vizuálních příznaků z vizuální stopy videonahrávky by bylo výhodnější, kdyby obličej mluvčích zabíral v obraze větší plochu, proto byla vytvořená další audio-vizuální databáze AVDB2cz, ve které již rozlišení video snímků bylo 640 (šířka) x 480 (výška) obrazových bodů a mluvčí zde byli snímáni pouze z čelního pohledu, viz obr. 7.1. Ve scéně již nebyla umístěna audio-vizuální klapka a mluvčí započal promluvu po zaznění 0.5 s tónu. V audio-vizuální databázi AVDB2cz bylo celkem zaznamenáno 35 mluvčích (3 ženy, 32 mužů) a databáze byla již kompletně zpracována. V budoucnu bude zpracována i původní databáze AVDB1cz, u které jsou mluvčí nasnímáni i z profilu tím mohou být extrahovány vizuální příznaky z aproximovaného 3D prostoru.

7.1.3 Použitá video technika

Jednotlivé videonahrávky byly pořízeny kamerami Logitech ClickSmart 510 (kamera č. 1) a Logitech QuickCam Pro 3000 (kamera č. 2). Tyto kamery mají zabudovaný mikrofon a připojení k počítači je realizováno přes USB 1.1. Kamera č.1 zaznamenává obraz v rozlišení 640 x 480 (320 x 240) obrazových bodů při snímkovací frekvenci 30 snímků za sekundu a zvuk je zaznamenáván při vzorkovací frekvenci 11025 Hz (resp. 22050 Hz pro 320 x 240). Kamera č.2 zaznamenává obraz pouze do rozlišení 320 x 240 obrazových bodů při snímkovací frekvenci 30 snímků za sekundu a zvuk je zaznamenáván při vzorkovací frekvenci 22050 Hz.

Pro pořízení databáze AVDB1cz byly použity obě kamery s rozlišením 320 x 240 obrazových bodů, pro databázi AVDB2cz byla použita kamera č. 1 s rozlišením 640 x 480 obrazových bodů. Výsledné videonahrávky z obou kamer byly přenášeny pomocí rozhraní USB do PC a ukládány do jednotlivých souborů (*.avi), ve kterých byly obrazová data komprimovaná a zakódována kodekem Indeo® video 5.

7.2 Zpracování audio-vizuální databáze AVDB2cz

V této stati je popsáno zpracování nahrávek audio-vizuální databáze AVDB2cz, i když vlastní postup zpracování a vytvořené algoritmy jsou použitelné i pro zpracování audio-vizuální databáze AVDB1cz.

7.2.1 Dekódování nahrávek

V audio-vizuální databázi AVDB2cz jsou v současné době uloženy videonahrávky od 35 různých mluvčích, kde každý mluvčí namluvil 50 slov a 50 vět, viz stať 7.1.1. V každé videonahrávce je zaznamenáno pouze jedno slovo nebo věta. Původní záměr byl rozložit (dekódovat) jednotlivé nahrávky na zvukový soubor (*.wav) a video signál rozdělit na jednotlivé snímky v nekomprimovaném formátu (*.bmp), což by podstatně zjednodušilo další zpracování jednotlivých snímků, jelikož dekódování videonahrávek je velmi časově náročné. Dekódování celé databáze zabere přibližně 10 hodin. Po dekódování všech nahrávek však vznikne 494 780 obrazů, které při velikosti 640 x 480 obrazových bodů a 24-bitovém formátu bitmapy by zabraly na disku přibližně 500 GB(!), proto byly všechny dekódované snímky před uložením předzpracovány a to tak, že byl v dekódovaných snímcích nejprve nalezen obličej (kapitola 4) a na disk byla ukládána pouze oblast zájmu pro geometrické příznaky a oblast zájmu o velikosti 128 x 128 obrazových bodů pro DCT příznaky (kapitola 6) získané z detekované oblasti obličeje.

Pro každou nahrávku byl vytvořen samostatný adresář, ve kterém byly uloženy příslušné snímky oblastí zájmu, dále zde byl uložen zvukový soubor (*.wav), textový soubor (*.txt) a informační soubor (*.nfo). V textovém souboru bylo uloženo odpovídající slovo nebo věta a jejich fonetický přepis. V informačním souboru byly uloženy informace o dekódované nahrávce tj. vzorkovací (snímkovací) frekvence a počet vzorků (snímků) akustického a visuálního signálu, rozměry snímku v obrazových bodech a číslo prvního a posledního snímku z vizuálního signálu, tato informace byla později po automatickém upravení délky nahrávky změněna.

Akustický signál z originálního zvukového souboru, ve kterém byl zaznamenán signál při vzorkovací frekvenci 11025 Hz, byl převzorkován na 8000 Hz a uložen do nového zvukového souboru. Tato operace byla provedena, aby se následně daly, bez jakýchkoliv úprav, použít programy pro zpracování a rozpoznávání akustického signálu řeči, které byly vytvořeny v naší laboratoři a které byly primárně navrženy pro 8000 Hz signál.

7.2.2 Upravení délky videonahrávek

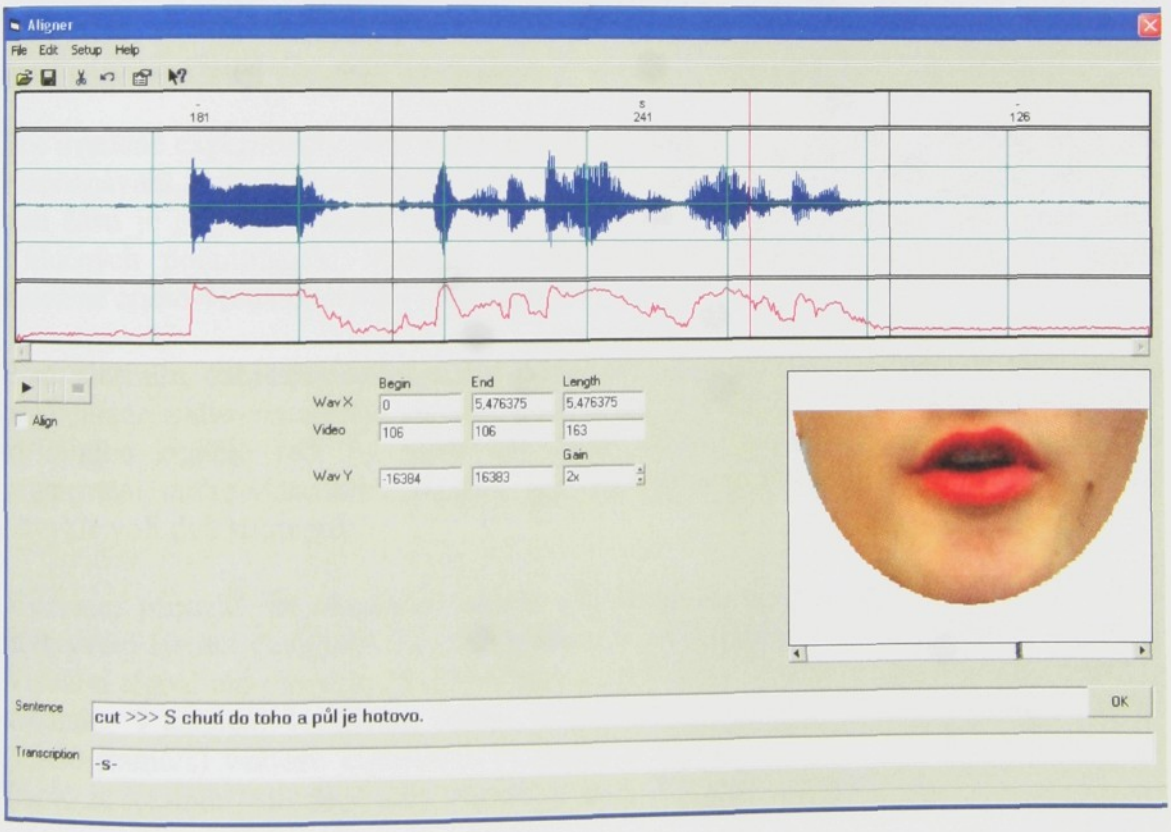
Z pořízených dekódovaných videonahrávek byl nejdříve odstraněn (odstřižen) počáteční 0.5 s tón, který upozorňoval mluvčí, že mohou začít promluvu. Pro nalezení tohoto tónu byla použita korelace záznamu tónu s postupným výřezem z akustického signálu, délka postupného výřezu (ve vzorcích) byla stejná jako délka zaznamenaného tónu. Po odstranění úvodního tónu byl použit jednoduchý stavový detektor, pomocí kterého byl nalezen začátek a konec promluvy. Původní akustický signál byl poté rozdělen do dvou zvukových souborů. V prvním byl vlastní záznam slova (věty) a ve druhém zvukovém souboru byl uložen signál od detekovaného konce promluvy, tj. signál, vněmž se

nacházel pouze šum. Tento signál pak později sloužil k odhadnutí SNR (odstup signálu od šumu). Podle upravené délky akustického signálu byla následně upravena délka vizuálního signálu. Z nalezeného začátku a konce promluvy v akustickém signálu bylo vypočteno příslušné číslo prvního a posledního snímku ve vizuálního signálu (7.1). Číslo prvního a posledního snímku bylo následně uloženo do příslušného informačního souboru.

$$vs_1 = round\left(\frac{as_1}{F_{sa}} \cdot F_{sv}\right), \quad vs_2 = round\left(\frac{as_2}{F_{sa}} \cdot F_{sv}\right) \tag{7.1}$$

kde as_1 (as_2) je číslo vzorku detekovaného začátku (konce) promluvy v akustickém signálu, Fsa je vzorkovací frekvence akustického signálu, Fsv je počet snímků za sekundu u vizuálního signálu, *round* označuje funkci zaokrouhlení na celočíselnou hodnotu a vs_1 (vs_2) reprezentuje výsledné číslo prvního (posledního) snímku příslušného upraveného vizuálního signálu.

Pro účely stříhání videonahrávek byl také vytvořen program Aligner. Tento program umožňuje automatické nebo poloautomatické stříhání audio-vizuálních nahrávek, viz obr. 7.2.



Obr. 7.2: *Program Aligner pro (polo)automatického stříhání audio-vizuálních nahrávek*

Kapitola 8

Experimentální práce – testy

V následujících statích jsou popsány testy a jejich výsledky při rozpoznávání izolovaných slov z audio-vizuální databáze AVDB2cz (kapitola 7). Pro rozpoznávání byl použit klasifikátor založený na technice skrytých Markovových modelů (kapitola 3), pro vlastní natrénování a rozpoznávání celoslovních modelů byl využit program HTK [STE97]. Celkově bylo v databázi AVDB2cz zaznamenáno 35 mluvčích, kde každý mluvčí namluvil 50 slov (příloha č.3). V trénovací databázi bylo 1500 slov od třiceti mluvčích a v testovací databázi se nacházelo 250 slov od zbylých pěti mluvčích.

I při takto relativně nízkém objemu slov bylo automatické zpracování a parametrizace vizuální části této databáze dosti časově náročné. Vlastní zpracovaná databáze AVDB2cz, ve které se po předzpracování nacházely již pouze oblasti zájmu vhodné pro parametrizaci vizuálního signálu zaujímala více než 70 GB diskového prostoru.

Zde uvedené experimenty jsou rozděleny do tří skupin. V první části jsou popsány testy rozpoznávání akustického signálu řeči, ve druhé jsou testy vizuálního signálu řeči a ve třetí části je popsáno vlastní rozpoznávání audio-vizuálního signálu řeči a jeho užití v hlučných podmínkách. Všechny experimentální testy jsou provedeny na stejné množině audio-vizuálních dat (1750 slov).

Před vlastním rozpoznáváním audio-vizuálního signálu řeči je nutné vyřešit otázku segmentace audio-vizuálního signálu, jelikož pro následnou fúzi a rozpoznávání audio-vizuálního signálu řeči by měla být délka framu (segmentu signálu) stejná. Při segmentaci audio-vizuálního signálu řeči na jednotlivé framy se v současné době obvykle volí dvě strategie.

V prvním případě je akustický signál segmentován po framech dlouhých 25 ms s překryvem 10 ms (kapitola 3)[1], tj. frekvence segmentace je 100 framů za sekundu. Vizuální signál má obvykle 25-50 snímků za sekundu, frekvence je tak 25-50 framů za sekundu. V každém vizuálním framu je provedena parametrizace a výsledný vektor (25-50 framů/s) vektoru vizuálních příznaků je aproximován na vektor, který má 100 framů za sekundu, obdobně jako segmentovaný akustický signál.

Ve druhém případě není provedena aproximace vizuálního příznakového vektoru, ale časová délka framu při segmentaci akustického signálu je zvolena stejně dlouhá, jako je délka vizuálního framu.

1) Délka framu a překryvu může být obecně i jiná, než zde uvedené hodnoty, přesto posun framu se obvykle volí 10 ms.

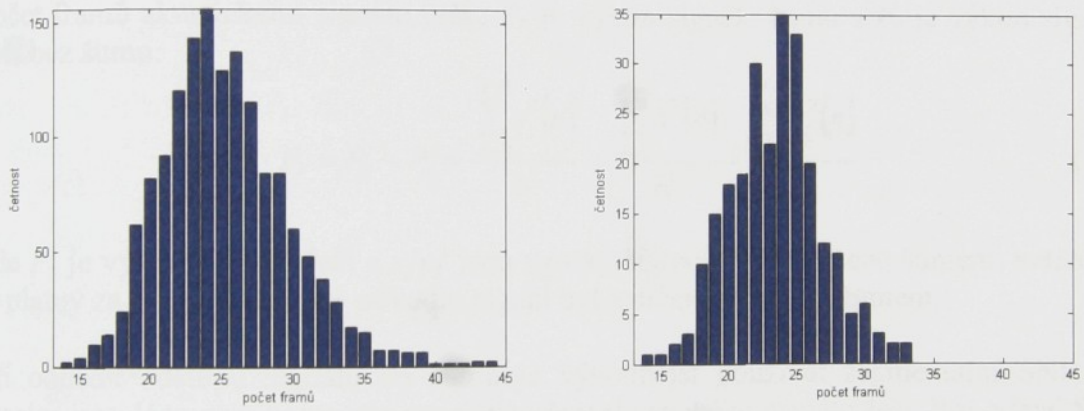
Ve své práci jsem zvolil druhý způsob, kdy snímkovací frekvence v našich audio-vizuálních nahrávkách byla 30 snímků za sekundu, délka framu tak byla stejná i pro segmentaci akustického signálu, tj. 33.3 ms bez překryvu.

8.1 Úloha rozpoznávání akustického signálu řeči

Pro rozpoznávání izolovaných slov byl použit klasifikátor založený na metodě skrytých Markovových modelů. Pro rozpoznávání slov byly předem vytvořeny (natrénovány) spojité celoslovní levo-pravé HM modely (kapitola 3). Před vlastním vytvořením těchto modelů bylo potřeba zjistit jaký je minimální počet framů (o délce 33.3 ms) tvořících jednotlivá slova z testovací databáze. Z tohoto minimálního počtu framů byl stanoven maximální počet stavů HM modelů.

8.1.1 Vyhodnocení počtu framů tvořících jednotlivá slova

Jak již bylo uvedeno v kapitole 7, tak akustické signály z adio-vizuálních nahrávek slov byly pomocí řečového detektoru rozděleny na akustický signál, ve kterém se nacházela pouze promluva (slovo) a na signál, který obsahoval šum. V každém signálu slova byl poté vypočten počet framů (o délce 33.3 ms), který tento signál tvoří:



Obr.8.1: Počty framů tvořících jednotlivá slova z trénovací části databáze 1500 slov (vlevo) a testovací části databáze 250 slov (vpravo)

Minimální počet framů tvořících jednotlivé akustické signály slov v testovací části databáze byl 14 framů (obdobně i v trénovací části databáze), viz obr. 8.1. Maximální počet stavů jednotlivých celoslovních HM modelů tak byl stanoven na 14 stavů.

8.1.2 Odhad SNR v akustickém signálu řeči

Před vlastním rozpoznávání slov z akustického signálu řeči bylo potřeba odhadnout odstup výkonu signálu od výkonu šumu SNR (Signal to Noise Ratio) v jednotlivých akustických nahrávkách slov. Pro odhad SNR se dnes obvykle používají vztahy počítající globální SNR (*GSNR* – 8.1), segmentální SNR (*SSNR* – 8.2) nebo segmentální aritmetické SNR (*SASNR* – 8.3) [POL02]. Zatímco globální SNR je počítáno z celého signálu, tak segmentální SNR je počítáno jako průměr z jednotlivých segmenttů (framů) signálu.

$$GSNR = 10.\log\frac{P_s}{P_n} = 10.\log\frac{\frac{1}{N}\sum_{n=0}^{N-1} s^2[n]}{\frac{1}{N}\sum_{n=0}^{N-1} n^2[n]} \tag{8.1}$$

$$SSNR = \frac{1}{F}\sum_{i=0}^{F-1}\left(10.\log\frac{\frac{1}{N}\sum_{n=0}^{N-1} s_i^2[n]}{\frac{1}{N}\sum_{n=0}^{N-1} n_i^2[n]}\right) \tag{8.2}$$

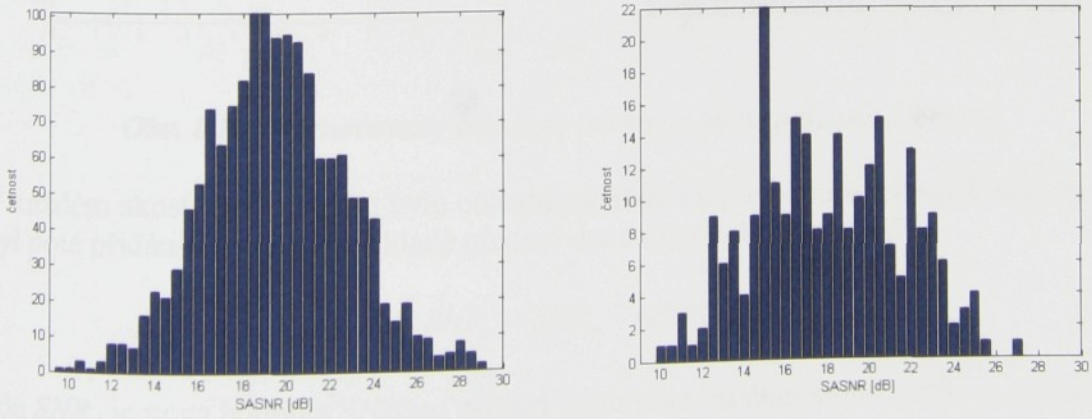
$$SASNR = 10.\log\left(\frac{1}{F}\sum_{i=0}^{F-1}\frac{\frac{1}{N}\sum_{n=0}^{N-1} s_i^2[n]}{\frac{1}{N}\sum_{n=0}^{N-1} n_i^2[n]}\right) \tag{8.3}$$

kde *s[n]* jsou vzorky signálu řeči (bez šumu), *n[n]* jsou vzorky příslušného šumu, N je počet všech vzorků signálu řeči nebo počet vzorků v jednom segmentu (framu), F je počet framů akustického signálu řeči., P_n je výkon signálu šumu a P_s je výkon signálu řeči bez šumu:

$$P_s = P_x - P_n = \frac{\sum_{n=0}^{N-1} s^2[n]}{N} = \frac{\sum_{n=0}^{N-1} x^2[n]}{N} - \frac{\sum_{n=0}^{N-1} n^2[n]}{N} \tag{8.4}$$

kde P_x je výkon signálu řeči a *x[n]* jsou vzorky signálu řeči zatížené šumem, vztah 8.4 je platný za předpokladu, že původní signál byl zatížen aditivním šumem.

Při odhadu odstupu signálu od šumu je výhodnější používat segmentální SNR, ve kterém jsou lépe postihnuty dynamické změny akustického signálu řeči. Pro odhad SNR v akustické části audio-vizuálních nahrávek jednotlivých slov z databáze AVDB2cz bylo použito aritmetické segmentální SNR, které poskytuje méně vychýlené hodnoty než *SSNR* [POL02].



Obr.8.2: *Odhad SASNR jednotlivých slov z trénovací části databáze 1500 slov (vlevo) a testovací části databáze 250 slov (vpravo)*

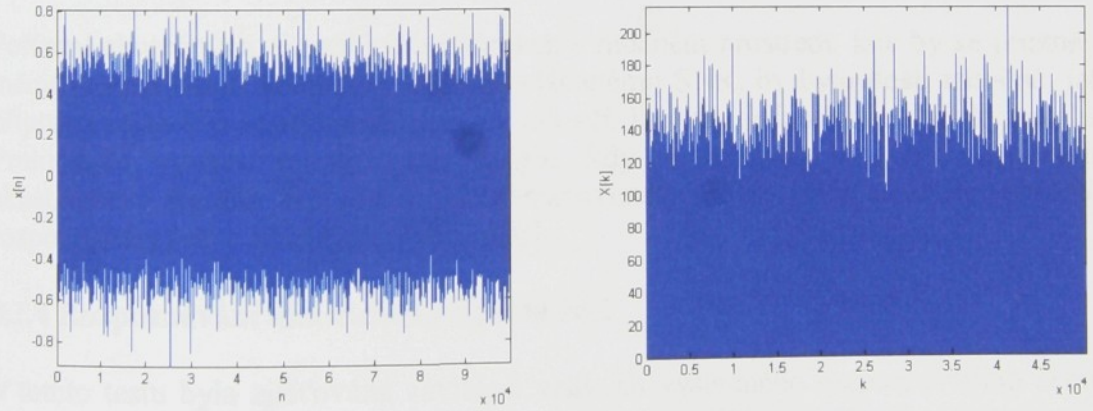
Část databáze	Trénovací											
Mluvčí č.	1	2	3	4	5	6	7	8	9	10	11	12
SASNR [dB]	26.3	17.1	16.9	20.3	17.3	20.9	20.5	22.6	18.2	22.4	20.5	18.4
Část databáze	Trénovací											
Mluvčí č.	13	14	15	16	17	18	19	20	21	22	23	24
SASNR [dB]	18.3	21.9	22.3	14.1	17.7	17.4	16.5	20.5	19.5	20.3	20.7	19.7
Část databáze	Trénovací						Testovací					X
Mluvčí č.	25	26	27	28	29	30	31	32	33	34	35	X
SASNR [dB]	19.2	20.7	21.0	18.5	17.1	18.6	20.3	16.0	16.9	14.7	22.1	X

Tab.8.1: *Průměrná hodnota odhadnutého SNR vypočtená z akustických nahrávek namluvených jednotlivými mluvčími*

Výsledná průměrná hodnota odhadnutého SASNR vypočtená z akustických nahrávek z trénovací části databáze byla 19.5 dB a průměrná hodnota SASNR z akustických nahrávek z testovací části databáze byla 18 dB.

8.1.3 Přidání aditivního šumu k akustickému signálu řeči

Pro účely následných testů rozpoznávání slov z akustického signálu byl vytvořen algoritmus, pomocí něhož bylo možné přidávat k akustickým nahrávkám aditivní šum. Tento algoritmus byl použit v testu, kde byla zkoumána závislost celkového rozpoznávacího skóre na šumu obsaženém v nahrávkách. Jako aditivní šum byl vybrán bílý šum uměle vygenerovaný pomocí generátoru náhodných hodnot, viz obr. 8.3.



Obr. 8.3: *Vygenerovaný bílý šum (vlevo) a jeho spektrum (vpravo)*

V každém akustickém signálu bylo odhadnuto SNR a k původnímu akustickému signálu byl poté přidán bílý šum na základě předem dané relativní změny SNR.

$$SNR_n = SNR_e - \Delta SNR \tag{8.5}$$

kde SNR_n je nová hodnota SNR po přidání aditivního (bílého) šumu, SNR_e je odhadnutá hodnota SNR (dle vztahu 8.3,4) v původním akustickém signálu a ΔSNR je předem stanovená relativní změna SNR.

Nová hodnota SNR_n je poté vypočtena:

$$SNR_n = 10.\log\frac{P_{Nx} - P_{Nn}}{P_{Nn} + c.P_{Nna}} \tag{8.6}$$

kde P_{Nx} je výkon původního akustického signálu řeči, zatíženého aditivním šumem o výkonu P_{Na}, P_{Nna} je výkon vygenerovaného bílého šumu, výkony jsou počítány ze signálů o délce N vzorků, kde N je délka původního akustického signálu řeči, ze signálů šumu (n a na) je pro výpočet příslušných výkonů bráno prvních N vzorků (příslušný aditivní šum n z akustických nahrávek byl delší než akustický signál řeči, obdobně i délka vygenerovaného šumu na byla stanovena větší než byla největší délka akustického signálu řeči z databáze), c je koeficient závislý na hodnotě SNR_n:

$$c(SNR_n) = \frac{P_{Nx} - P_{Nn}}{P_{Nna} \cdot 10^{\left(\frac{SNR_n}{10}\right)}} - \frac{P_{Nn}}{P_{Nna}} \tag{8.7}$$

Výsledný akustický signál řeči s přidaným bílým šumem dle předem odhadnutého SNR_n (8.5) je vytvořen dle:

$$x_n[n] = x[n] + na[n].\sqrt{c}, \quad 0 \le n \le N \tag{8.8}$$

kde $x_n[n]$ jsou nové hodnoty vzorků vypočtené z původních vzorků akustického signálu řeči $x[n]$ s přidaným aditivním bílým šumem na.

Pořízení skutečných akustických nahrávek v hlučném prostředí, kde by se pružně dala měnit úroveň šumu vzhledem k předpokládanému SNR, by bylo dosti náročné, jak na přípravu takového experimentu, tak na mluvčí, kteří by jednotlivé nahrávky namluvili. Proto bylo využito tohoto zjednodušení, kdy je přidáván bílý šum k původnímu akustickému signálu řeči dle předem stanoveného SNR, čímž je simulována úloha rozpoznávání řeči v hlučných podmínkách.

8.1.4 Rozpoznávání akustického signálu řeči

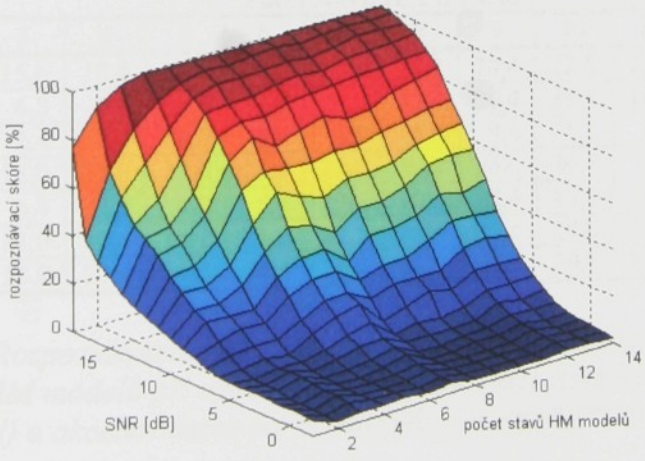
V tomto testu byla zjišťována závislost velikosti výsledného rozpoznávacího skóre na počtu stavů použitých (natrénovaných) celoslovních HM modelů při měnícím se SNR v akustickém signálu. Specifikace testu:

Klasifikátor založený na	celoslovní levo-pravé HM modely
Počet příznaků	39
Druh příznaků	13 x (MFCC + delta + akcelerační)
Počet stavů HM modelů	proměnný (1 – 14)
Délka framu	33.3 ms
Průměrná hodnota SNR [dB]	proměnná (18 – (- 2))
Slovník	50 slov
Trénovací databáze	30 mluvčích (1500 slov)
Testovací databáze	5 mluvčích (250 slov)

Výsledné rozpoznávací skóre [%] je vypočteno jako počet správně rozpoznaných slov ku počtu všech slov z testovací části databáze.

SNR [dB]	Počet stavů HM modelů													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
18	76,8	90,8	97,2	99,2	99,2	99,2	99,2	99,2	99,2	99,2	99,2	99,2	99,2	99,2
17	40,4	66,8	81,2	90,0	94,8	97,2	96,8	97,6	98,4	96,8	98,0	98,8	98,4	98,8
16	33,6	57,2	68,8	80,8	90,8	93,6	93,6	94,0	94,8	93,2	94,8	97,2	96,8	96,4
15	28,8	48,0	56,0	70,0	86,4	87,6	88,0	90,0	90,8	90,0	92,4	93,2	93,6	94,4
14	24,4	44,4	47,2	58,4	78,8	82,4	82,0	84,4	85,2	86,4	87,2	87,6	88,8	89,2
13	20,8	40,0	42,0	50,8	70,4	77,6	71,2	75,2	78,4	74,4	75,6	80,0	82,0	84,8
12	19,2	33,2	34,8	42,8	60,4	65,6	64,0	63,6	65,6	64,8	68,0	69,6	72,4	72,8
11	16,8	28,4	28,8	34,4	48,4	59,2	54,4	51,2	56,4	55,6	57,2	62,0	64,0	64,8
10	14,4	23,6	23,6	28,0	39,2	46,8	44,4	38,4	46,0	45,6	49,6	55,6	56,8	56,0
9	12,4	20,8	20,4	20,4	30,8	35,6	36,8	28,8	36,4	36,8	39,2	45,2	42,4	46,0
8	11,2	14,8	17,2	16,8	24,8	28,8	29,6	22,8	26,0	26,8	30,8	36,0	35,2	34,4
7	8,8	11,2	12,4	12,4	17,2	22,4	22,8	17,6	20,4	20,4	20,4	26,8	28,0	26,0
6	7,2	7,6	9,6	7,6	13,2	17,6	20,0	10,0	13,2	16,8	14,4	19,2	19,6	18,8
5	7,2	6,0	7,6	6,8	8,8	12,0	14,8	6,4	8,8	7,6	8,8	13,2	13,2	13,6
4	7,2	5,6	8,0	5,2	6,8	9,6	10,0	5,6	5,6	4,8	5,6	10,0	8,4	10,0
3	7,2	5,6	8,0	4,4	4,8	6,0	8,4	3,6	3,2	4,4	4,8	7,2	6,8	6,8
2	7,6	4,0	7,6	4,0	3,6	5,2	7,6	2,8	2,4	3,6	3,6	6,0	4,4	4,8
1	6,8	4,0	6,4	4,0	3,6	4,4	6,0	2,8	2,4	2,4	2,8	5,6	3,2	4,0
0	6,8	3,6	6,8	4,0	3,6	4,0	5,2	2,8	2,4	2,4	2,4	4,8	3,2	3,6
-1	6,4	3,2	5,6	4,0	4,0	3,2	4,4	2,8	2,4	2,4	2,4	4,0	3,2	3,2
-2	6,0	2,8	4,0	4,0	4,4	3,2	4,4	2,4	2,4	2,4	2,4	3,2	3,2	2,4

Tab. 8.2: *Rozpoznávací skóre [%] v závislosti na počtu stavů použitých (natrénovaných) HM modelů při měnícím se SNR v akustickém signálu*



Obr. 8.4: *Graf hodnot z tab. 8.2 – výsledné rozpoznávací skóre [%] v závislosti na počtu stavů HM modelů při měnícím se SNR v akustickém signálu*

Z tabulky 8.2 je patrné, že vyšších hodnot rozpoznávacího skóre v prostředí s vysokou hladinou šumu (přibližně do SNR 6dB) lze dosáhnout při použití vícestavových (10 a více) celoslovních HM modelů.

8.2 Úloha rozpoznávání vizuálního signálu řeči

V této úloze bylo testováno použití tvarových (geometrických) vizuálních příznaků a vizuální příznaky popisující informační obsah obrazu oblasti zájmu získaných pomocí DCT (kapitola 6) při automatickém rozpoznávání vizuálního signálu řeči.

8.2.1 Rozpoznávání vizuálního signálu s tvarovými příznaky

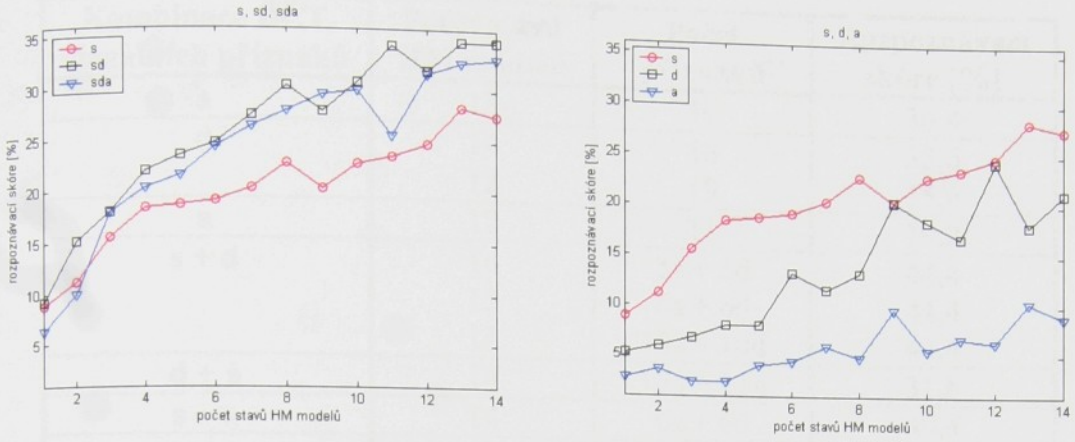
Jako tvarové příznaky byly v tomto testu požity vizuální příznaky horizontální (*h*) a vertikální (*v*) rozšíření rtů, dále oblast rtů (*o*) a zaokrouhlení rtů (*r*) (kapitola 6). Z těchto příznaků byly vypočteny dynamické a akcelerační příznaky, které jsou také při rozpoznávání akustického signálu řeči uplatněny. V tomto testu bylo zjišťováno rozpoznávací skóre na počtu stavů použitých celoslovních HM modelů, kde pro parametrizaci vizuálního signálu bylo použito sedm různých kombinací statických, dynamických a akceleračních příznaků (1. – pouze statické, 2. – pouze dynamické, 3. – pouze akcelerační, 4. – statické + dynamické, 5. – dynamické + akcelerační, 6. – statické + akcelerační, 7. – statické + akcelerační + dynamické). Specifikace testu:

Klasifikátor založený na	celoslovní levo-pravé HM modely
Počet příznaků	proměnný 1-12
Druh příznaků (kombinace)	*h*, *v*, *o*, *r*, delta, akcelerační
Počet stavů HM modelů	proměnný (1 – 14)
Délka framu	33.3 ms
Slovník	50 slov
Trénovací databáze	30 mluvčích (1500 slov)
Testovací databáze	5 mluvčích (250 slov)

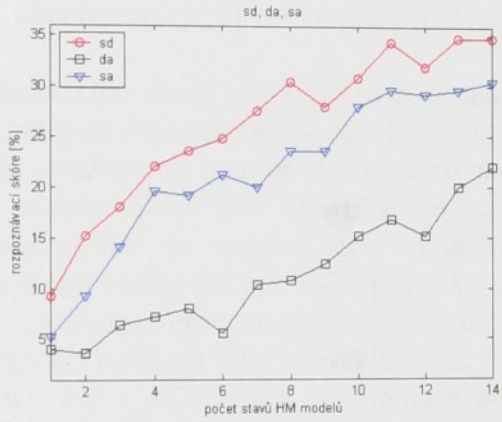
Viz. pří.	Počet stavů HM modelů													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
s	8,8	11,2	15,6	18,4	18,8	19,2	20,4	22,8	20,4	22,8	23,6	24,8	**28,4**	27,6
d	5,2	6,0	6,8	8,0	8,0	13,2	11,6	13,2	20,4	18,4	16,8	**24,4**	18,0	21,2
a	2,8	3,6	2,4	2,4	4,0	4,4	6,0	4,8	9,6	5,6	6,8	6,4	**10,4**	8,8
s+d	9,2	15,2	18,0	22,0	23,6	24,8	27,6	30,4	28,0	30,8	34,4	32,0	**34,8**	**34,8**
d+a	4,0	3,6	6,4	7,2	8,0	5,6	10,4	10,8	12,4	15,2	16,8	15,2	20,0	**22,0**
s+a	5,2	9,2	14,0	19,6	19,2	21,2	20,0	23,6	23,6	28,0	29,6	29,2	29,6	**30,4**
s+d+a	6,4	10,0	18,0	20,4	21,6	24,4	26,4	28,0	29,6	30,0	25,6	31,6	32,8	**33,2**

Tab. 8.3: *Rozpoznávací skóre [%] v závislosti na počtu stavů použitých (natrénovaných) HM modelů při využití různých kombinací statických (s), dynamických (d) a akceleračních (a) vizuálních tvarových příznaků*

Nejlepších výsledků rozpoznávacího skóre (34,8 %) bylo v tomto testu dosaženo při využití kombinace statických a dynamických vizuálních tvarových příznaků, které byly použity pro natrénování celoslovních HM modelů s třinácti (čtrnácti) stavy a následnému rozpoznávání slov klasifikátorem založeným na celoslovních HM modelech. Použití akceleračních příznaků v kombinaci se statickými a dynamickými příznaky nevedlo u našich nahrávek k lepšímu rozpoznávacímu skóre, viz tab. 8.3.



Obr. 8.5: *Grafy hodnot z tab. 8.3 – výsledné rozpoznávací skóre [%] v závislosti na počtu stavů HM modelů při využitá různých kombinací statických (s), dynamických (d) a akceleračních (a) vizuálních tvarových příznaků*



8.2.2 Rozpoznávání vizuálního signálu s DCT příznaky

Pro vizuální příznaky popisující informační obsah obrazu oblasti zájmu bylo použito prvních N příznaků s největší hodnotou energie (6.11) z obrazové matice oblasti zájmu transformované pomocí diskrétní kosinové transformace DCT (kapitola 6). V prvním testu byl zjišťován vliv počtu stavů celoslovních HM modelů a počtu N energetických příznaků na výsledné rozpoznávací skóre.

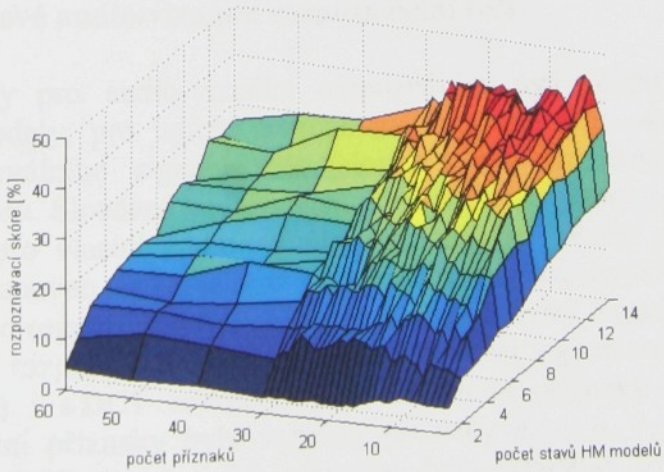
Počet vizuálních příznaků byl zvětšován o hodnotu 1 v rozsahu 1 až 30 příznaků a dále o hodnotu 10 v intervalu 30 až 60 příznaků. Výsledné tabulky z tohoto testu pro kombinace statických (s), dynamických (d) a akceleračních DCT-energetických vizuálních příznaků jsou dosti rozsáhlé, proto byly umístěny do přílohy č.6.

V následující tabulce 8.4 jsou uvedeny nejlepší výsledky rozpoznávacího skóre pro jednotlivé kombinace (s, d, a) energetických DCT vizuálních příznaků z tabulek 1 až 7 z přílohy č.6. U některých kombinací (pouze dynamické příznaky (d), statické a dynamické příznaky (sd)) vizuálních příznaků byla nejvyšší hodnota rozpoznávacího skóre dosažena pro různé kombinace počtu stavů HM modelu a počtu použitých vizuálních příznaků, proto jsou všechny tyto kombinace v tabulce uvedeny.

Kombinace DCT vizuálních příznaků	Počet stavů HM modelů	Počet příznaků	Rozpoznávací skóre [%]
s	13	10	36,4
d	14	16	35,6
	12	19	35,6
a	12	11	19,2
s + d	14	5s + 5d	44,4
	13	6s + 6d	44,4
	13	10s + 10d	44,4
d + a	14	16d + 16a	31,6
s + a	14	6s + 6d	40,4
s + d + a	14	5s + 5d +5a	**45,2**

Tab. 8.4: Vybrané nejvyšší hodnoty rozpoznávacího skóre v závislosti na počtu stavů použitých (natrénovaných) HM modelů a počtu použitých energetických DCT vizuálních příznaků při využitá různých kombinací statických (s), dynamických (d) a akceleračních (a) vizuálních příznaků

Nejvyšší hodnota rozpoznávacího skóre (45,2 %) byla dosažena při využití kombinace pěti statických a jím příslušných dynamických a akceleračních vizuálních energetických DCT příznaků, které byly použity pro natrénování čtrnácti stavových celoslovních HM modelů a k následnému rozpoznávání slov klasifikátorem založeným na celoslovních HM modelech. Na obrázku 8.6 je uveden graf průběhu rozpoznávacího skóre v závislosti na počtu použitých energetických DCT příznaků a počtu stavů HM modelů. Počet vybraných nejvyšších hodnot energetických příznaků byl zvětšován o hodnotu 1 v rozsahu 1 až 30 příznaků a o hodnotu 10 v intervalu 30 až 60 příznaků. Tento graf byl vytvořen z hodnot tabulky č.7 z přílohy č.6 (kombinace statických, dynamických a akceleračních příznaků). Průběh hodnot rozpoznávacího skóre pro jiné kombinace statických, dynamických a akceleračních vizuálních energetických DCT příznaků je přibližně stejný, jako je průběh v grafu na obr. 8.5, viz tabulky hodnot rozpoznávacího skóre v příloze č.6.



Obr. 8.6: *Graf hodnot výsledného rozpoznávací skóre (z tabulky č.7 v příloze č.6) v závislosti na počtu stavů HM modelů a počtu energetických DCT vizuálních příznaků při kombinaci statických (s), dynamických (d) a akceleračních (a) vizuálních příznaků*

Jak již bylo uvedeno výše, tak za vizuální příznaky bylo v tomto testu použito prvních N příznaků s největší hodnotou energie (6.11) z obrazové matice oblasti zájmu o velikosti 128 x 128 obrazových bodů, která byla transformována pomocí diskrétní kosinové transformace. V rámci dalších testů byl zkoumán vliv různě vypočtených vizuálních příznaků z obrazové matice transformované diskrétní kosinovou transformací, kde vizuální příznaky byly vybrány z nejvyšších hodnot energie E (6.11), rozptylu R (6.12) nebo normovaného rozptylu NR (6.13) z DCT koeficientů.

Pro každý z těchto zkoumaných vlivů byl vytvořen kompletní test rozpoznávání řeči v závislosti na počtu stavů použitých (natrénovaných) HM modelů a počtu použitých různě počítaných DCT vizuálních příznaků. Při porovnání těchto testů bylo dosaženo nejlepších výsledků pro energetické DCT vizuální příznaky. Pro názornost jsou v tab. 8.5 uvedeny hodnoty rozpoznávacího skóre pro různé varianty testů, kde byla použita kombinace pěti statických a jim přináležejících dynamických a akceleračních příznaků, tj. konfigurace u které bylo dosaženo nejlepšího rozpoznávacího skóre, viz tab. 8.4.

druh pří.	Počet stavů HM modelů													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
E	**5,2**	**11,2**	**13,2**	**19,2**	**22,8**	**26,4**	**30,8**	**32,4**	**32,0**	**34,8**	**36,8**	**38,8**	**37,6**	**45,2**
R	5,2	10,8	12,4	18,8	20,4	26,0	28,8	31,2	32,0	34,0	35,6	36,0	36,8	44,0
NR	4,8	8,0	12,0	12,8	16,8	21,2	24,4	21,6	22,4	24,0	25,2	27,2	28,8	31,6

Tab. 8.5: *výsledné rozpoznávací skóre [%] v závislosti na počtu stavů HM modelů pro vizuální příznaky vybrané z nejvyšších hodnot energie E, rozptylu R nebo normovaného rozptylu NR.*

8.3 Úloha rozpoznávání audio-vizuálního signálu řeči

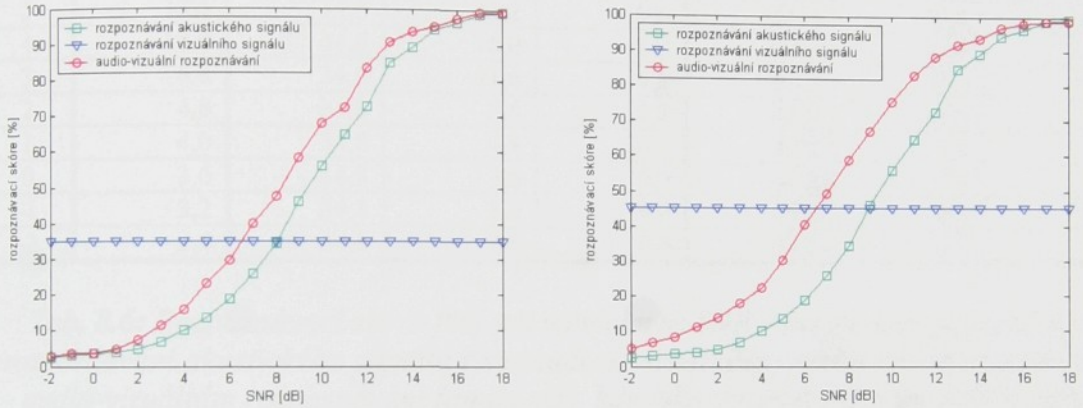
V této úloze byly provedeny testy pro vlastní audio-vizuální rozpoznávání řeči. Při vlastní fúzi akustických a vizuálních příznaků byly natrénovány čtrnáctistavové jednostreamové (dvoustreamové) celoslovní HM modely (kapitola 2), které poté sloužily k vlastnímu rozpoznávání pomocí klasifikátoru založeném na technice HMM.

8.3.1 Jednostreamové audio-vizuální rozpoznávání řeči

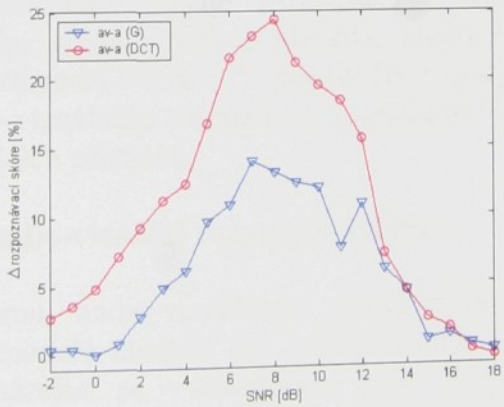
V tomto testu byly pro audio-vizuální rozpoznávání řeči použity jednostreamové celoslovní HM modely, pro jejichž natrénování byly sloučeny akustické a vizuální příznaky audio-vizuálního signálu řeči (slova) z trénovací databáze do jednoho příznakového vektoru. Sloučené audio-vizuální příznakové vektory z testovací databáze jsou pak použity pro vlastní audio-vizuální rozpoznávání. Počet stavů HM modelů v tomto testu byl stanoven na 14, jelikož při použití čtrnáctistavových HM modelů bylo dosaženo nejvyššího rozpoznávacího skóre jak u rozpoznávání akustického signálu řeči (tab. 8.2), tak pro rozpoznávání vizuálního signálu řeči s geometrickými vizuálními příznaky (tab. 8.3) i s DCT energetickými vizuálními příznaky (tab. 8.4,5). Jako geometrické vizuální příznaky byly vybrány příznaky h, v, o, r a příslušné delta příznaky a jako DCT energetické příznaky bylo vybráno 5 nejvyšších DCT energetických příznaků s příslušnými delta a akceleračními příznaky. Pro obě tyto kombinace vizuálních příznaků bylo dosaženo nejvyššího rozpoznávacího skóre, viz tab. 8.3-5.

Specifikace testu:

Klasifikátor založený na	jednostreamové celoslovní levo-pravé HM modely
Počet příznaků	47 aku. + geometrické (54 aku. + DCT)
Druh příznaků (kombinace aku. a viz)	39 aku. – 13 x (MFCC + delta + akcelerační) 8 viz – *h*, *v*, *o*, *r* + delta (15 viz – 5 x (ene. DCT + delta + akcelerační))
Počet stavů HM modelů	14
Délka framu	33.3 ms
Slovník	50 slov
Trénovací databáze	30 mluvčích (1500 slov)
Testovací databáze	5 mluvčích (250 slov)



Obr. 8.7: *Výsledné rozpoznávací skóre pro audio-vizuální rozpoznávání řeči s využitím geometrických vizuálních příznaků (vlevo) a DCT energetických vizuálních příznaků (vpravo) ve srovnání s rozpoznáváním samostatného akustického a vizuálního signálu řeči*



Obr. 8.8: *Rozdíl výsledného rozpoznávacího skóre mezi audio-vizuálním rozpoznáváním řeči (av) a rozpoznáváním řeči z akustického signálu (a), kde pro audio-vizuální rozpoznávání byly použity geometrické vizuální příznaky (G) nebo DCT energetické vizuální příznaky (DCT)*

SNR [dB]	Rozpoznávání						
	audio	viz (G)	audio-viz (G)	av-a (G)	viz (DCT)	audio-viz (DCT)	av-a (DCT)
18	99,2	34,8	99,2	0,0	45,2	98,8	-0,4
17	98,8	34,8	99,2	0,4	45,2	98,8	0,0
16	96,4	34,8	97,6	1,2	45,2	98,0	1,6
15	94,4	34,8	95,2	0,8	45,2	96,8	2,4
14	89,2	34,8	93,6	4,4	45,2	93,6	4,4
13	84,8	34,8	90,8	6,0	45,2	92,0	7,2
12	72,8	34,8	83,6	10,8	45,2	88,4	15,6
11	64,8	34,8	72,4	7,6	45,2	83,2	18,4
10	56,0	34,8	68,0	12,0	45,2	75,6	19,6
9	46,0	34,8	58,4	12,4	45,2	67,2	21,2
8	34,4	34,8	47,6	13,2	45,2	58,8	24,4
7	26,0	34,8	40,0	14,0	45,2	49,2	23,2
6	18,8	34,8	29,6	10,8	45,2	40,4	21,6
5	13,6	34,8	23,2	9,6	45,2	30,4	16,8
4	10,0	34,8	16,0	6,0	45,2	22,4	12,4
3	6,8	34,8	11,6	4,8	45,2	18,0	11,2
2	4,8	34,8	7,6	2,8	45,2	14,0	9,2
1	4,0	34,8	4,8	0,8	45,2	11,2	7,2
0	3,6	34,8	3,6	0,0	45,2	8,4	4,8
-1	3,2	34,8	3,6	0,4	45,2	6,8	3,6
-2	2,4	34,8	2,8	0,4	45,2	5,2	2,8

Tab. 8.6: *Rozpoznávací skóre [%] při měnícím se SNR v akustickém signálu, pro rozpoznávání akustického signálu řeči (audio), vizuálního signálu řeči (viz) nebo pro audio-vizuálním rozpoznávání (audio-viz), kde jako vizuální příznaky byly použity geometrické (G) nebo DCT energetické příznaky (DCT). Zároveň je zde uvedena hodnota rozdílu výsledného rozpoznávacího skóre mezi audio-vizuálním rozpoznáváním řeči a rozpoznáváním řeči z akustického signálu (av-a).*

Z grafů na obr. 8.7-8 a z příslušné tabulky tab. 8.6 je patrné, že vizuální složka řeči při audio-vizuálním rozpoznávání zlepšuje rozpoznávací skóre oproti samostatnému rozpoznávání akustického signálu řeči. Nejlepšího zlepšení bylo dosaženo u SNR v akustickém signálu v rozmezí 6-9 dB. Při použití DCT energetických příznaků bylo u jednostreamového audio-vizuálního rozpoznávání dosaženo v průměru lepších výsledků oproti použití geometrických příznaků.

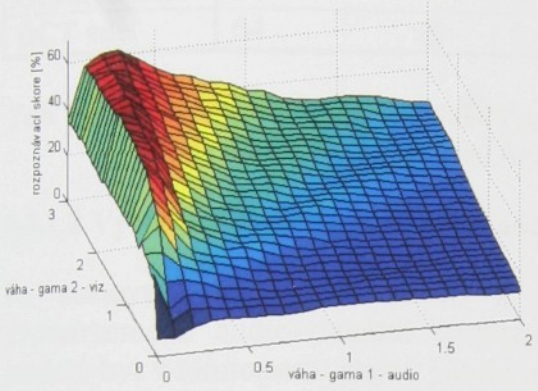
8.3.1 Dvoustreamové audio-vizuální rozpoznávání řeči

Oproti jednostreamovému audio-vizuálnímu rozpoznávání řeči lze při využití dvoustreamových HM modelů nastavit pro každý stream váhu a tím zvýraznit nebo potlačit informaci nacházející se v akustickém nebo vizuálním signálu řeči. Před vlastním dvoustreamovým a-v rozpoznáváním je potřeba stanovit hodnoty jednotlivých vah výstupní funkce HM modelů (2.2).

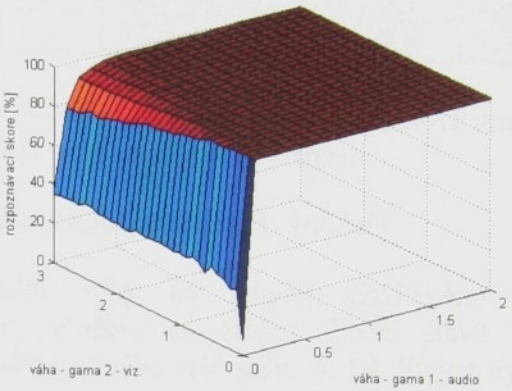
8.3.1.1 Stanovení vah pro dvoustreamové audio-vizuální rozpoznávání řeči

Stanovení vah se pro dvoustreamové audio-vizuální rozpoznávání řeči nejčastěji určuje experimentálně. Vzhledem ke stanovené úloze audio-vizuálního rozpoznávání řeči

v hlučných podmínkách jsem vytvořil experimentální test, u kterého bylo zjišťováno výsledné rozpoznávací skóre vzhledem k nastavení vah u akustického a vizuálního streamu, kde v akustické části jsou audionahrávky zatížené šumem o průměrném SNR 5 dB. Pro možnosti srovnání byl proveden stejný test i při použití originálních akustických nahrávek s průměrným SNR 18 dB. Tabulky hodnot rozpoznávacího skóre těchto dvou testů byly pro svou velkou rozsáhlost umístěny v příloze č.7, příslušné grafy viz obr. 8.9.

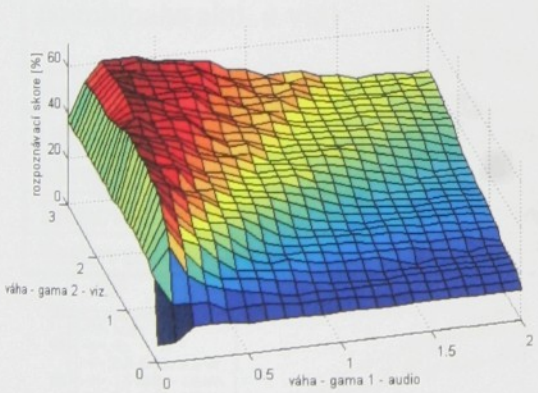


(g_1)

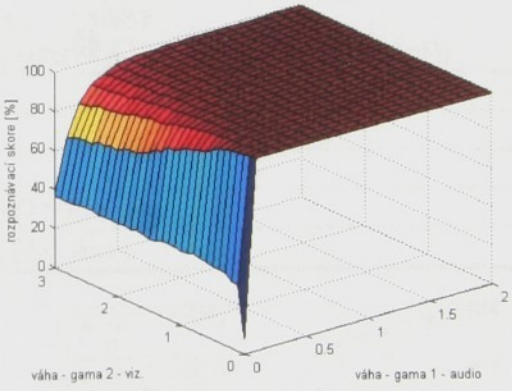


(g_2)

Obr. 8.9: *Grafy hodnot výsledného rozpoznávacího skóre v závislosti na použitých vahách akustického a vizuálního streamu při využití geometrických příznaků (g_1: 5dB SNR, g_2: 18dB SNR) nebo DCT energetických příznaků (g_3: 5dB SNR, g_4: 18dB SNR) pro audio-vizuální rozpoznávání.*



(g_3)



(g_4)

Výsledné hodnoty vah pro akustický a vizuální stream byly stanoveny na základě dvou kritérií, v prvním případě byly vybrány váhy na základě nejvyšší dosažené hodnoty rozpoznávacího skóre, kde byly akustické nahrávky s přidaným šumem o SNR 5dB (grafy g_1, g_3). V druhém případě (dle druhého kritéria) byly hodnoty vah vybrány na základě nejvyššího součtu hodnot rozpoznávacího skóre z testů při SNR 5dB a 18dB. Vybrané hodnoty vah jsou uvedeny v tab. 8.7, v tabulkách v příloze č.7 jsou tyto hodnoty silně zvýrazněny. Z tabulek v příloze č.7 je také patrné, že použitý HTK software zřejmě nemá naprogramovanou utilitu pro dvoustreamové rozpoznávání, přesně dle vztahu (2.2), který je uváděn v HTK manuálu [STE97], jelikož při nastavení vah 1.0 pro vizuální stream a 0.0 pro akustický stream by musely být hodnoty rozpoznávacího skóre při využití geometrických vizuálních příznaků 34,8% a při využití

DCT energetických vizuálních příznaků 45,2%, přesto hodnoty pro rozpoznávací skóre pro 5dB a 18dB u těchto vah (1.0 0.0) jsou stejné, tj akustický signál není včleněn při rozpoznávání (pro váhu akustického streamu 0.0). I přes tuto „drobnost" je HTK utilita plně funkční pro dvoustreamové rozpoznávání.

vizuální příznaky	SNR 5 dB			SNR 5dB vs. 18 dB		
	roz. skóre [%]	γ_1 (a)	γ_2 (v)	roz. skóre [%]	γ_1 (a)	γ_2 (v)
Geometrické	64	0,1	1,8	62,4 (5), 95,2 (18)	0,3	2,8
DCT ene.	64	0,1	1,3	57,6 (5), 97,6 (18)	0,6	2,3

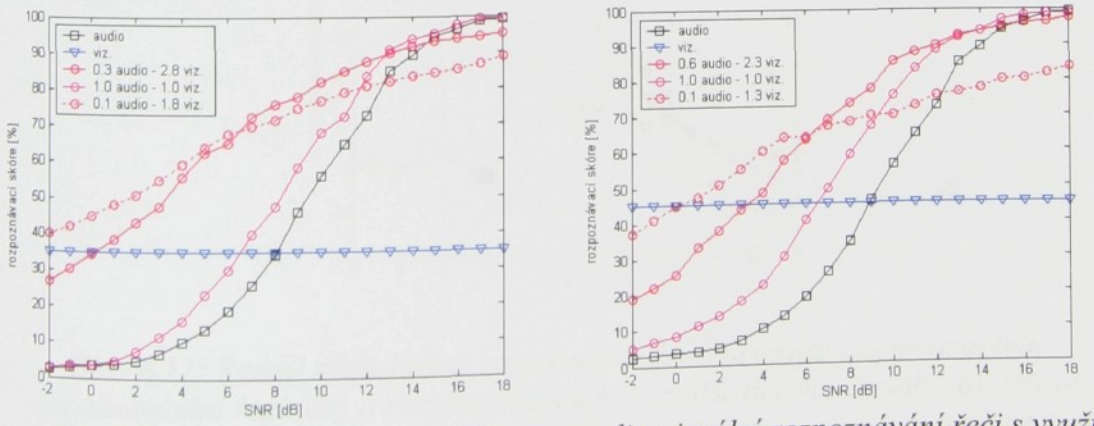
Tab. 8.7: Vybrané hodnoty akustické (γ_1) a vizuální (γ_2) váhy pro dvoustreamové audio-vizuální rozpoznávání řeči, dle prvního a druhého kritéria

8.3.1.2 Výsledky dvoustreamového audio-vizuálního rozpoznávání řeči

Po stanovení hodnot akustické a vizuální váhy lze použít klasifikátor pro dvoustreamové audio-vizuální rozpoznávání založený na HMM. Počet stavů HM modelů byl v tomto testu stejný jako u jednostreamového AV-ASR, tj. 14. Stejné byly i použité geometrické a DCT vizuální příznaky.

Specifikace testu:

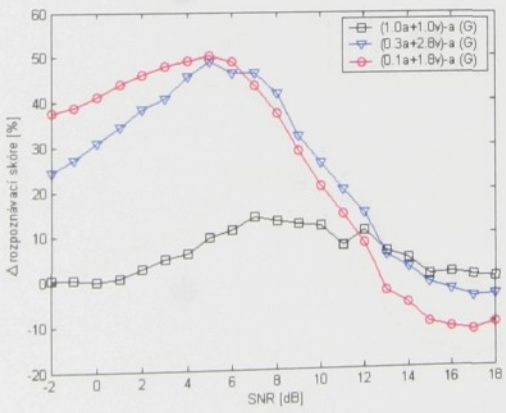
Klasifikátor založený na	dvoustreamové celoslovní levo-pravé HM modely
Počet příznaků	47 aku. + geometrické (54 aku. + DCT)
Druh příznaků (kombinace aku. a viz)	39 aku. – 13 x (MFCC + delta + akcelerační) 8 viz – *h*, *v*, *o*, *r* + delta (15 viz – 5 x (ene. DCT + delta + akcelerační))
Počet stavů HM modelů	14
Délka framu	33.3 ms
Slovník	50 slov
Trénovací databáze	30 mluvčích (1500 slov)
Testovací databáze	5 mluvčích (250 slov)



Obr. 8.10: Výsledné rozpoznávací skóre pro audio-vizuální rozpoznávání řeči s využitím geometrických příznaků (vlevo) a DCT energetických příznaků (vpravo) ve srovnání s rozpoznáváním samostatného akustického a vizuálního signálu řeči, kde pro dvoustreamové audio-vizuální rozpoznávání byly využity vybrané váhy z tab. 8.7 a pro porovnání byly navíc použity pro oba streamy stejné váhy 1.0, 1.0.

SNR [dB]	Rozpoznávání							
	audio	viz	1.0 audio 1.0 viz	av-a	0.1 audio 1.8 viz	av-a	0.3 audio 2.8 viz	av-a
18	99,2	34,8	99,2	0,0	88,8	-10,4	95,2	-4,0
17	98,8	34,8	99,2	0,4	86,8	-12,0	94,4	-4,4
16	96,4	34,8	97,6	1,2	85,2	-11,2	93,6	-2,8
15	94,4	34,8	95,2	0,8	84,4	-10,0	93,2	-1,2
14	89,2	34,8	93,6	4,4	83,6	-5,6	91,6	2,4
13	84,8	34,8	90,8	6,0	82,0	-2,8	90,0	5,2
12	72,8	34,8	83,6	10,8	80,8	8,0	87,6	14,8
11	64,8	34,8	72,4	7,6	79,2	14,4	84,8	20,0
10	56,0	34,8	68,0	12,0	76,8	20,8	82,0	26,0
9	46,0	34,8	58,4	12,4	74,8	28,8	78,0	32,0
8	34,4	34,8	47,6	13,2	71,6	37,2	76,0	41,6
7	26,0	34,8	40,0	14,0	69,6	43,6	72,4	46,4
6	18,8	34,8	29,6	10,8	67,6	48,8	65,2	46,4
5	13,6	34,8	23,2	9,6	64,0	50,4	62,4	48,8
4	10,0	34,8	16,0	6,0	59,2	49,2	55,6	45,6
3	6,8	34,8	11,6	4,8	54,8	48,0	47,6	40,8
2	4,8	34,8	7,6	2,8	50,8	46,0	43,2	38,4
1	4,0	34,8	4,8	0,8	48,0	44,0	38,4	34,4
0	3,6	34,8	3,6	0,0	44,8	41,2	34,4	30,8
-1	3,2	34,8	3,6	0,4	42,0	38,8	30,4	27,2
-2	2,4	34,8	2,8	0,4	40,0	37,6	26,8	24,4

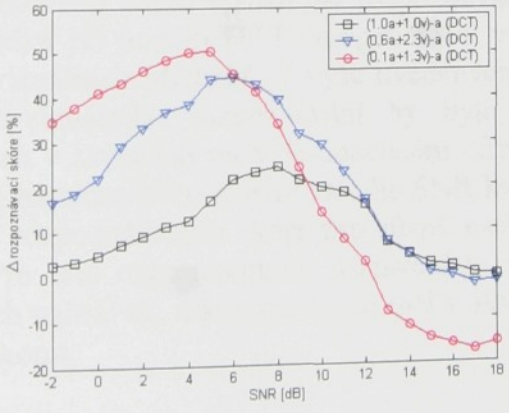
Tab. 8.8: *Rozpoznávací skóre [%] při měnícím se SNR v akustickém signálu, pro rozpoznávání akustického signálu řeči (audio), vizuálního signálu řeči (viz) nebo pro audio-vizuálním rozpoznávání (audio-viz), kde jako vizuální příznaky byly použity geometrické příznaky. Zároveň jsou zde uvedeny hodnoty rozdílu výsledného rozpoznávacího skóre mezi audio-vizuálním rozpoznáváním řeči a rozpoznáváním řeči z akustického signálu (av-a) pro různě zvolené váhy akustického a vizuálního streamu.*



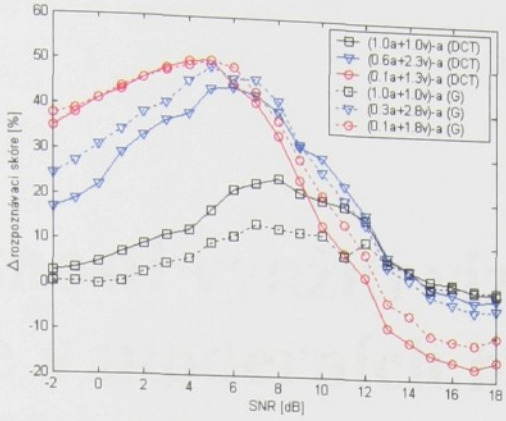
Obr. 8.11: *Rozdíl výsledného rozpoznávacího skóre mezi audio-vizuálním rozpoznáváním řeči (av) a rozpoznáváním řeči z akustického signálu (a), kde pro audio-vizuální rozpoznávání byly použity geometrické vizuální příznaky a různě zvolené váhy akustického a vizuálního streamu.*

SNR [dB]	Rozpoznávání							
	audio	viz	1.0 audio 1.0 viz	av-a	0.1 audio 1.8 viz	av-a	0.3 audio 2.8 viz	av-a
18	99,2	45,2	98,8	-0,4	83,6	-15,6	97,6	-1,6
17	98,8	45,2	98,8	0,0	81,6	-17,2	96,4	-2,4
16	96,4	45,2	98,0	1,6	80,4	-16,0	96,0	-0,4
15	94,4	45,2	96,8	2,4	80,0	-14,4	94,8	0,4
14	89,2	45,2	93,6	4,4	77,6	-11,6	93,6	4,4
13	84,8	45,2	92,0	7,2	76,4	-8,4	92,4	7,6
12	72,8	45,2	88,4	15,6	75,6	2,8	89,6	16,8
11	64,8	45,2	83,2	18,4	72,8	8,0	88,0	23,2
10	56,0	45,2	75,6	19,6	70,0	14,0	85,2	29,2
9	46,0	45,2	67,2	21,2	70,0	24,0	77,6	31,6
8	34,4	45,2	58,8	24,4	68,4	34,0	73,6	39,2
7	26,0	45,2	49,2	23,2	67,2	41,2	68,8	42,8
6	18,8	45,2	40,4	21,6	64,0	45,2	63,2	44,4
5	13,6	45,2	30,4	16,8	64,0	50,4	57,6	44,0
4	10,0	45,2	22,4	12,4	60,0	50,0	48,4	38,4
3	6,8	45,2	18,0	11,2	55,2	48,4	43,6	36,8
2	4,8	45,2	14,0	9,2	50,8	46,0	38,0	33,2
1	4,0	45,2	11,2	7,2	47,2	43,2	33,2	29,2
0	3,6	45,2	8,4	4,8	44,8	41,2	25,6	22,0
-1	3,2	45,2	6,8	3,6	41,2	38,0	22,0	18,8
-2	2,4	45,2	5,2	2,8	37,2	34,8	19,2	16,8

Tab. 8.9: *Rozpoznávací skóre [%] při měnícím se SNR v akustickém signálu, pro rozpoznávání akustického signálu řeči (audio), vizuálního signálu řeči (viz) nebo pro audio-vizuálním rozpoznávání (audio-viz), kde jako vizuální příznaky byly použity DCT energetické příznaky. Zároveň jsou zde uvedeny hodnoty rozdílu výsledného rozpoznávacího skóre mezi audio-vizuálním rozpoznáváním řeči a rozpoznáváním řeči z akustického signálu (av-a) pro různě zvolené váhy akustického a vizuálního streamu.*



Obr. 8.12: *Rozdíl výsledného rozpoznávacího skóre mezi audio-vizuálním rozpoznáváním řeči (av) a rozpoznáváním řeči z akustického signálu (a), kde pro audio-vizuální rozpoznávání byly použity DCT energetické vizuální příznaky a různě zvolené váhy akustického a vizuálního streamu.*



Obr. 8.13: *Porovnání rozdílu výsledného rozpoznávacího skóre mezi audio-vizuálním rozpoznáváním řeči (av) a rozpoznáváním řeči z akustického signálu (a), kde pro audio-vizuální rozpoznávání byly použity DCT energetické (DCT) nebo geometrické (G) vizuální příznaky a různě zvolené váhy akustického a vizuálního streamu*

8.3.1.3 Celkové zhodnocení experimentů pro audio-vizuální rozpoznávání řeči

Při dvoustreamovém audio-vizuálním rozpoznávání izolovaných slov (s hodnotami vah streamu dle tab. 8.7) bylo u našich nahrávek z AV databáze AVDB2cz (kapitola 7) v průměru dosaženo lepších výsledků zvýšení rozpoznávacího skóre než u jednostreamového audio-vizuálního rozpoznávání ve srovnání se samostatným rozpoznáváním akustického signálu, viz obr. 8.8 versus obr. 8.13. Přičemž hodnoty rozpoznávacího skóre u jednostreamového rozpoznávání jsou dle původního předpokladu shodné jako u dvoustreamového rozpoznávání se shodnými hodnotami vah 1.0 pro akustický i vizuální stream. U vah stanovených dle prvního a druhého kritéria (tab. 8.7) bylo v průměru dosaženo lepších výsledných hodnot rozpoznávacího skóre pro geometrické vizuální příznaky oproti DCT energetickým vizuálním příznakům, viz obr. 8.13. U jednostreamového audio-vizuálního rozpoznávání bylo naopak dosaženo v průměru lepších výsledků při použití DCT energetických vizuálních příznaků oproti použití geometrických vizuálních příznaků. Z výše uvedených testů dále vyplývá, že u dvoustreamového audio-vizuálního rozpoznávání by bylo výhodné pružně měnit hodnoty streamových vah v závislosti na vyhodnoceném odstupu výkonu signálu řeči od výkonu šumu SNR. Charakteristiku změny vah dle SNR lze vytvořit experimentálně dle maximální hodnoty rozpoznávacího skóre pro různě nastavené váhy streamů pro danou hodnotu SNR. Pro toto experimentální nastavení je však nutné vytvořit větší databázi audio-vizuálních nahrávek, u kterých by se měl v akustickém signálu nacházel reálný šum o různém výkonu.

Kapitola 9

Další možnosti využití vizuální informace v moderních hlasových technologiích

Současný výzkum a vývoj v oblasti zpracování a rozpoznávání řeči by se dal rozdělit přibližně do sedmi větších skupin:

- Audio-vizuální rozpoznávání řeči
- Audio-vizuální rozpoznávání řeči v hlučných podmínkách
- Audio-vizuální detekce mluvčího v prostoru
- Audio-vizuální identifikace mluvčích
- Audio-vizuální verifikace mluvčích
- Audio-vizuální segmentace řeči
- Audio-vizuální syntéza řeči

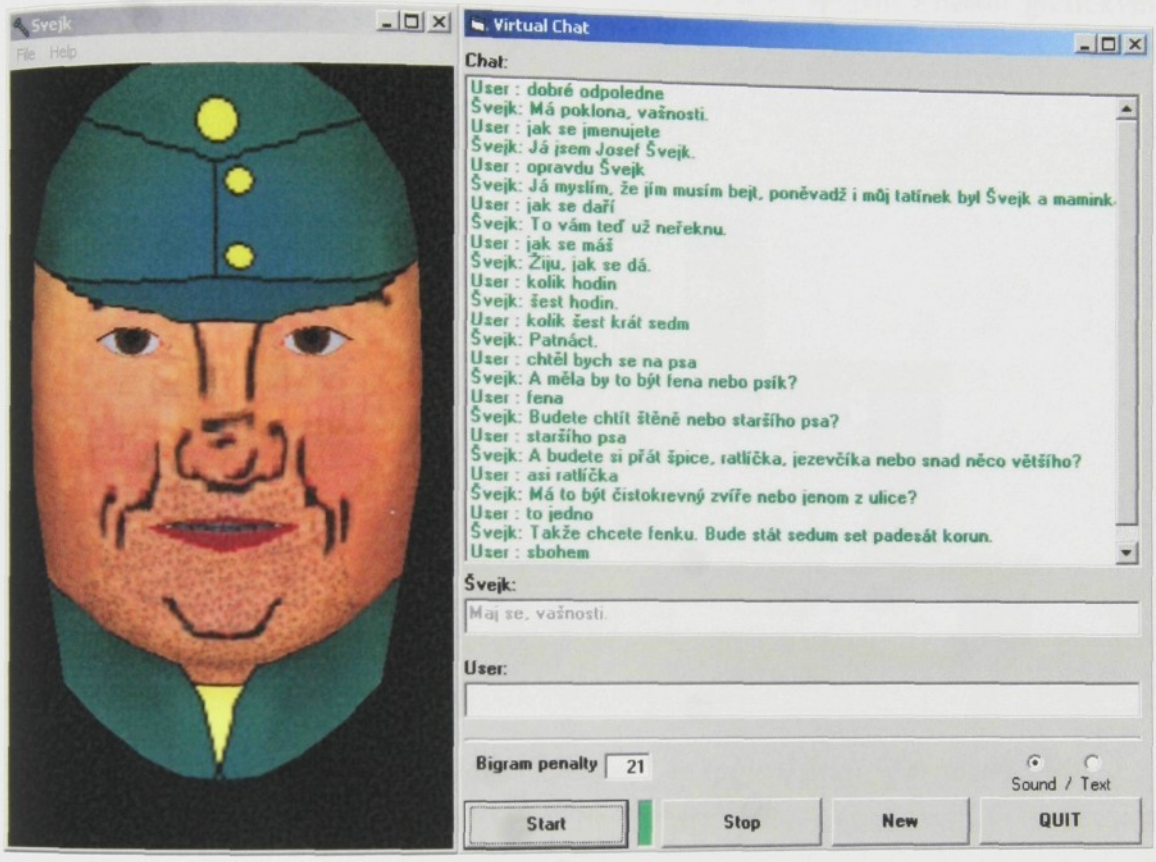
V oblasti audio-vizuálního rozpoznávání řeči v normálních podmínkách slouží obvykle vizuální složka řeči jako pomocná informace při rozpoznávání úseků řeči (např. fonémů), které je obtížné v akustickém signálu řeči rozpoznat a příslušná vizuální složka je schopné zlepšit výsledné rozpoznávací skóre u těchto úseků [POT01b].

Audio-vizuální rozpoznávání řeči v hlučných podmínkách je dnes zřejmě nejčastěji řešená úloha [HEC02, SCA03, POT03], kde se vizuální složka řeči při rozpoznávání s úspěchem používá. Testy s audio-vizuálním rozpoznáváním řeči v hlučných podmínkách tak byly provedeny i v této disertační práci.

U audio-vizuální detekce mluvčích v prostoru (ve scéně) je zjišťováno na základě vyhodnocení obrazového signálu, který mluvčí právě mluví [ZOT02]. Řešení této úlohy je dosti složité vzhledem k různým možnostem natočení mluvčích ke snímací kameře, dále může být ve snímané scéně větší množství mluvčích a někteří lidé, kteří nepromlouvají mohou intenzivněji dýchat ústy, což zhoršuje výsledné vyhodnocení.

Úloha audio-vizuální identifikace [FOX03] a verifikace [ZHA01] mluvčích bývá často zjednodušena, jelikož se ve snímaném obraze uvažuje pouze jedna osoba. Při identifikaci je poté osoba určena na základě porovnání s předem vytvořenými audio-vizuálními modely jednotlivých mluvčích. U verifikace je zjišťováno nakolik je audio-vizuální promluva podobná modelu osoby, za kterou se mluvčí vydává.

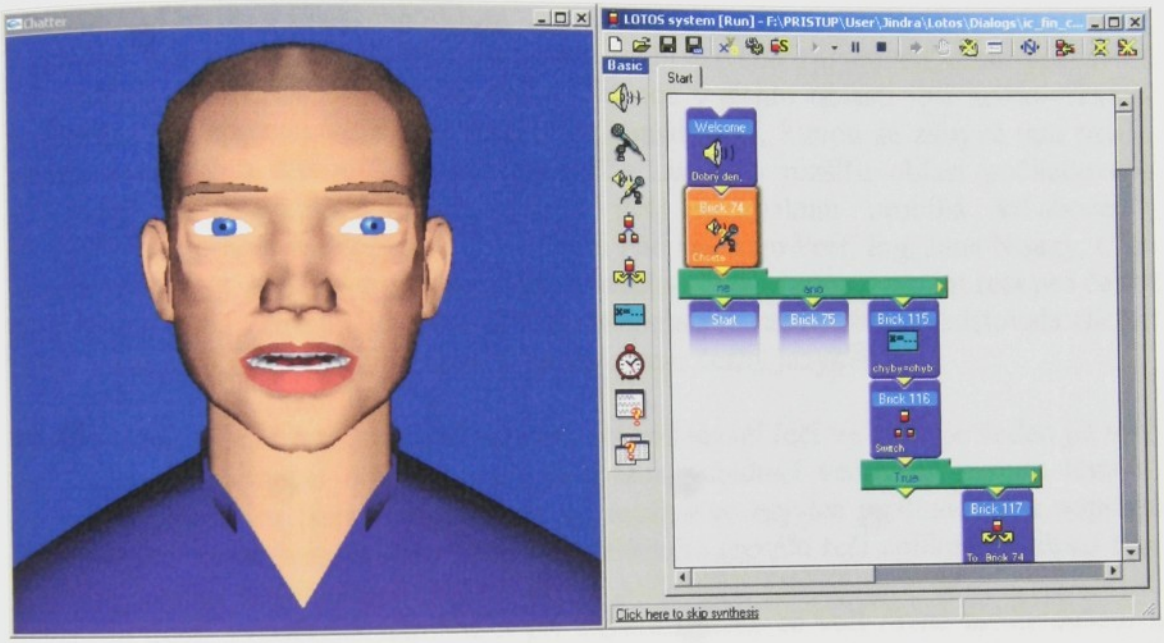
Audio-vizuální segmentace řečového signálu se dnes často používá pro segmentaci delších audio-vizuálních signálů na jednotlivé bloky, například segmentace pro přepis videonahrávek televizních zpráv [FRE04], kde vizuální složka většinou necharakterizuje přímo jednotlivé mluvčí, tj. odezírání ze rtů, ale spíše se ve vizuálním signálu sleduje změna scény. Pokud je menší množství mluvčích ve scéně (dva, tři), tak lze audio-vizuální signál segmentovat podle toho, kdo právě mluví, viz detekce mluvčích v prostoru.



Obr. 9.1: *Rozmluva s virtuální osobností - Švejkem*

Poslední větší oblastí je využití zpracování a rozpoznávání audio-vizuálního signálu řeči pro vytvoření audio-vizuálního TTS (Text To Speech) systému, kde je tento systém nejčastěji reprezentován 3D počítačovým modelem mluvící hlavy. Tento model se skládá z akustického systému TTS a 3D modelu hlavy, pro který je na základě zpracování a rozpoznávání videonahrávek mluvčích vytvářen algoritmus pro animaci modelu v závislosti na promluvě (pohyb rtů, čelistí, jazyka, mimických svalů). Ve světě dnes existují 3D modely mluvících hlav pro různé národnostní jazyky (angličtinu, španělštinu, francouzštinu, švédštinu...). Jedním z prvních pokusů o vytvoření systému audio-vizuální syntézy řeči bylo upravení anglického 3D modelu mluvící hlavy (Baldi) pro český jazyk [CHA01-02c]. Tento model vznikl na kalifornské univerzitě v Santa Cruz, ale na jeho vývoji se podílely i další laboratoře [COL99]. U tohoto anglického modelu byly přemapovány jednotlivé anglické vizémy na české, dle vzájemné podobnosti [CHA02c] a byl použit český akustický TTS systém, vytvořený v Ústavu radiotechniky a elektroniky Akademie věd ČR v Praze [PRI97]. Pro takto upravený model byl následně vytvořen test, u kterého bylo zjišťováno nakolik je tento model

srozumitelný pro lidi mluvící česky [CHA02]. Z tohoto testu vyplynulo, že je potřeba pro češtinu vytvořit vlastní model. Nový ryze český 3D model mluvící hlavy (Chatter) tak byl vytvořen v Laboratoři počítačového zpracování řeči na TU v Liberci [CHA02d-03c]. Obdobný český model vznikl i v Oddělení umělé inteligence na Katedře kybernetiky Fakulty aplikovaných věd Západočeské univerzity v Plzni [CIS02]. 3D model mluvící hlavy lze využít v různých komunikačně informačních systémech. V našich programech byl 3-D model mluvící hlavy použit v systému pro rozmluvu s virtuální osobností Švejkem [CHN02] (viz obr. 9.1) a ve spojení s naším grafickým návrhovým dialogovým systémem LOTOS [NOU01], viz obr 9.2.



Obr. 9.2: *3D česky mluvící hlava „Chatter" ve spojení grafickým návrhovým dialogovým systémem LOTOS*

Kapitola 10

Závěr

V průběhu posledních třiceti let se z oblasti počítačového zpracování a rozpoznávání řeči vydělila celá řada aplikačních oblastí, jednou z těchto oblastí je i audio-vizuální zpracování a rozpoznávání audio-vizuálního signálu řeči, kterou se zabývá tato práce. Jedním z hlavních důvodů vzniku této práce tak bylo rozšířit oblast počítačového zpracování a rozpoznávání řeči, jejíž vývoj a výzkum probíhá v Laboratoři počítačového zpracování řeči na TU v Liberci pod vedením Prof. Ing. Jana Nouzy, CSc. již více než 10 let. Dalším důvodem bylo, že audio-vizuální rozpoznávání řeči pro český jazyk bylo v době vzniku této práce (2001) v naprostých začátcích a neexistovala tak ani žádná dostupná kvalitní audio-vizuální databáze pro český jazyk.

Vývoj a výzkum v oblasti audio-vizuálního rozpoznávání řeči ve světě je veden již více než 20 let a do dnešní doby existuje celá řada publikací věnovaná tomuto tématu. Dostupnou literaturu jsem se snažil pokud možno co nejvíce prostudovat a popsané algoritmy pro zpracování a rozpoznávání vizuálního signálu řeči aplikovat, přesto tyto algoritmy nejsou stoprocentně přenositelné a použitelné, jelikož zpracování obrazového signálu je dosti výpočetně náročné a při experimentech se volí různá zjednodušení a dosti také záleží na použité audio-vizuální databázi. Proto většina zde popsaných algoritmů vznikla mou vlastní invencí nebo se jedná o modifikace algoritmů jiných autorů, kteří jsou citování v odkazech této práce. Celá problematika, především zpracování a rozpoznávání vizuálního signálu řeči, musela být v této práci komplexně řešena a zasahuje do celé řady oborů.

Všechny stanovené cíle této práce byly i přes velkou časovou náročnost postupně vyřešeny, tj. byly vytvořeny dvě audio-vizuální databáze promluv pro český jazyk (kapitola 7). Pro zpracování nahrávek z této databáze byl vytvořen komplexní program (příloha č.6). Dále byl vytvořen program pro parametrizaci vizuálního signálu řeči skládající se z detektoru lidské tváře v obraze (kapitola 4), systému pro nalezení rtů (kapitola 5) a vlastního systému pro parametrizaci vizuálního signálu, pomocí kterého jsou vytvářeny tvarové vizuální příznaky a vizuální DCT příznaky popisující informační obsah obrazu (kapitola 6). Dále byly řešeny otázky fúze akustických a vizuálních příznaků a byly vytvořeny experimentální testy pro audio-vizuálního rozpoznávání izolovaných slov, při srovnání se samostatným rozpoznáváním akustického a vizuálního signálu řeči. Závěrem byly navrženy testy pro audio-vizuální rozpoznávání izolovaných slov v hlučných podmínkách, při kterých se potvrdilo, že vizuální složka řeči může zlepšit rozpoznávací skóre v hlučných podmínkách (kapitola 8). V následující tabulce je uveden výběr výsledků z těchto testů na experimentální množině 1750 slov z audio-vizuální databáze AVDB2cz .

Odstup signálu od šumu SNR [dB]	Rozpoznávání akustického signálu řeči [%]	Rozpoznávání vizuálního signálu řeči [%]	Jednostreamové audio-vizuální rozpoznávání [%]	Dvoustreamové audio-vizuální rozpoznávání [%]
5	13.6	$^{DCT)}45.2(^{G)}34.8)$	$^{DCT)}30.4(^{G)}23.2)$	$^{DCT)}64.0(^{G)}64.0)$

Tab. 10.1: Vybrané hodnoty rozpoznávacího skóre [%] z experimentálních testů pro rozpoznávání akustického signálu (kapitola 8), vizuálního signálu řeči a pro jednostreamové a dvoustreamové audio-vizuální rozpoznávání řeči, kde jako vizuální příznaky byly použity geometrické (G) nebo DCT energetické vizuální příznaky.

Z výše uvedené tabulky je patrné, že vizuální složka řeči může výrazně zlepšit rozpoznávací skóre v hlučných podmínkách (zde např. pro 5 dB SNR).

10.1 Přínosy disertační práce

Nejdůležitější výsledky této disertační práce lze shrnout do následujících bodů:

- Vytvoření dvou audio-vizuálních databází AVDB1cz, AVDB2cz videonahrávek izolovaných slov i celých vět pro český jazyk (kapitola 7). Vytvoření programu pro zpracování nahrávek z této databáze (kapitola 7).

- Navržení a vytvoření systému pro detekování lidského obličeje, založeného na barevné a tvarové segmentaci obrazu (kapitola 4) a systému pro nalezení rtů v detekované oblasti zájmu (kapitola 5). Oba tyto systémy byly navrženy a naprogramovány tak, aby byly co nejspolehlivější a zároveň výpočetně časově rychlé. Jednotlivé části těchto systémů byly publikovány na mezinárodních konferencích, například na mezinárodní konferenci ICSLP 2004 pořádané v Jižní Koreji [CHA04c].

- Vytvoření komplexního systému pro parametrizaci vizuálního signálu, pomocí kterého jsou vytvářeny tvarové vizuální příznaky a vizuální DCT příznaky popisující informační obsah obrazu (kapitola 6).

- Návrh a experimentální ověření možností fúze a rozpoznávání audio-vizuálního signálu řeči v hlučných podmínkách (kapitola 8).

10.2 Aplikační oblasti

V dnešní době existuje celá řada aplikačních oblastí využití audio-vizuálního zpracování a rozpoznávání řeči, z nichž jmenujme například: vlastní audio-vizuální rozpoznávání řeči, audio-vizuální rozpoznávání řeči v hlučných podmínkách, audio-vizuální detekce mluvčího v prostoru, audio-vizuální identifikace mluvčích, audio-vizuální verifikace mluvčích, audio-vizuální segmentace řeči, audio-vizuální syntéza řeči. Tyto oblasti a jejich využití jsou podrobněji popsány v kapitole 9. Další z možných aplikačních oblastí při využití audio-vizuálního zpracování a rozpoznávání řeči je vytváření pomůcek pro výuku řeči nebo pomůcek pro sluchově nebo pohybově postižené lidi.

10.3 Náměty na další práci

Další výzkumná práce nyní směřuje k vývoji a vytvoření audio-vizuálního rozpoznávače řeči založeného na modelech menších stavebních jednotek řeči (fonémy a vizémy). V blízké budoucnosti bych také rád vylepšil svůj počítačový model česky mluvící hlavy, který vznikl v počátcích mé disertační práce (2002) a u kterého by bylo možné nově použít poznatky získané při tvorbě systému pro audio-vizuální rozpoznávání řeči.

Literatura

[ATH03] ATHOFF, F., McGLAUN, G., LANG, M., RIGOLL, G.: A real-time demonstrator for video-based recognition of dynamic head gestures, using discrete hidden Markov models. *In: Proc. of 7th World Multiconference on Systemics, Cybernetics and Informatics – SCI 2003*, July 2003, Orlando, USA, pp. 93-98, ISBN 980-6560-01-9

[AUG93] AUGUSTEIJN, M., F., SKUJCA, T., L.: Identification of human faces through texture-based feature recognition and neural network technology. *In Proc. of IEEE Conference on Neural Networks*, pp. 392-398., 1993

[BRE94] BREGLER, C., KONIG, Y.: Eigenlips for robust speech recognition. *In Proc. of International Conference Acoustic*, Speech and Signal Processing, Adelaide, Australia, pp.669-672, 1994

[CAN83] CANNY, J., F.: Finding edges and lines in images. *In Technical Report AI-TR-720*, MIT, Artificial Intelligence Laboratory, Cambridge, MA, 1983

[CAN86] CANNY, J., F.: A computational approach to edge detection. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 679-698, 1986

[CIS02] CÍSAŘ, P., KRŇOUL, Z., NOVÁK, J., ŽELEZNÝ, M.: Approach to an audio-visual speech synthesis using concatenation-based method. *In Proc. of 12th Czech-German Workshop „Speech Processing".* Prague, September 2002. pp. 64-65. ISBN 80-86269-09-4

[CIS03] CÍSAŘ, P., ŽELEZNÝ, M.: Feature Selection for the Czech Speaker Independent Automatic Lip-Reading. *In Proc. of 6th International Workshop on Electronics, Control, Measurment and Signals-ECMS 2003.* Liberec, June 2003. pp. 12-16. ISBN 80-7083-708-X

[CIS04] CÍSAŘ, P., ŽELEZNÝ, M., KRŇOUL, Z.: 3D Lip-tracking for Audio-Visual Speech Recognition in Real Applications. In Proc. of ICSLP 2004, October 2004, Jeju Island, Korea, ISSN 1225-441x

[COL99] COLE, R., MASSARO, D., W., DE VILLIERS, J., RUNDLE, B., SHOBAKI, K., WOUTERS, J., COHEN, M., BESKOW, J., STONE, P.,

CONNORS, P., TARACHOW, P., SOLCHER, D.: New tools for interactive speech and language training: Using animated conversational agents in the classrooms of profoundly deaf children. *In Proceedings of ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education*, London, UK, Apr 1999

[CON03] CONNELL, J., H., HAAS, N., MARCHERET, E., NETI, C., POTAMIANOS, G., VELIPASALAR, S.: A real-time prototype for small-vocabulary audio-visual ASR, *In Proc. Int. Conf. Multimedia Expo.*, vol. II, pp. 469-472, Baltimore, July 2003

[CRA87] CRAW, J., ELLIS, H., LISHMAN, J.: Automatic extraction of face features . *In Pattern Recognition Letters*, pp. 183-187, 1987

[DAI96] DAI, Y., NAKANO, Y.: Face-Texture model based on SGLD and its application in face detection in a color scene. *In Proc. of Pattern Recognition*, pp. 1007-1017, 1996

[DAU01] DAUBIAS, P., DELÉGLISE, P.: Evaluation of an Automatically Obtained Shape and Appearance Model For Automatic Audio Visual Speech Recognition. *In Proc. of Eurospeech 2001.* Aalborg, Sept. 2001, ISBN 87-90834-09-7. ISSN 1018-4074

[DAU02] DAUBIAS, P., DELÉGLISE, P.: LIP-READING BASED ON A FULLY AUTOMATIC STATISTICAL MODEL. *In Proc. of 6th Int. Conference on Spoken Language Processing,* Denver USA, September 2002, ISBN 1-876346-40-X

[DUC94] DUCHNOWSKI, P., MEIER, U., WAIBEL, A.: See me, hear me: Integration automatic speech recognition and lip-reading. In Proc. of ICSLP 94, Yokohoma, Japan, pp. 547-550, 1994

[DUP00] DUPONT, S., LUETTIN, J.: Audio-visual speech modeling for continuous speech recognition. *In IEEE Multimedia, Transactions on*, Sep 2000, pp. 141-151, Volume 2, Issue 3, ISSN: 1520-9210

[FOX03] FOX, N., REILLY, R., B.: Audio-Visual Speaker Identification Based on the Use of Dynamic Audio and Visual Features. *In Proc. 4th International Conference on Audio and Video Based Biometric Person Authentication*, June 2003

[FRE04] PEREZ-FREIRE, L., GARCIA-MATEO, C.: A MULTIMEDIA APPROACH FOR AUDIO SEGMENTATION IN TV BROADCAST NEWS, *In ICASSP 2004*, Montreal, Canada, 2004

[GAO00] GAO, W., MA, J., WANG, R., YAO, H.: Towards Robust Lipreading, *In International Conference on Spoken Language Processing*, pp. 15-19, Beijing, China, Oct, 2000, ISBN 7-80150-114-4

[GOE01] GOECKE, R., MILLAR, J., B., ZELINSKY, A., ROBERT-RIBES, J.: Stereo Vision Lip-Tracking for Audio-Video Speech Processing. *In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2001*, Salt Lake City, USA, 7-11 May 2001

[GRA96] GRAF, H., P., COSSATTO, E., GIBBON, D., KOCHEISEN, M., PETAJAN, E.: Multimodal system for locating heads and faces. *In Proc. of the Second International Conference on Automatic Face and Gestures Recognition*, pp. 88-93, 1996

[HAN98] HAN, C., C., LIAO, H., Y., M., YU, K.,C., CHEN, L., H.: Fast face detection via morphology-based pre-processing. *In Proceedings of the Ninth International Conference on Image Analysis and Processing*, pp. 469-476, 1998

[HEC02] HECKMANN, M., BERTHOMMIER, F., KROSCHEL, K.: Noise Adaptive Stream Weighting in Audio-Visual Speech Recognition, *In EURASIP Journal of Appl. Signal Processing*, vol. 2002, No. 11, Nov. 2002, p. 1260-1273

[HEN96] HENNECKE, M., E., STORK, D., G., PRASAD, K., V.: Visionary speech: Looking ahead to practical speechreading systems. *In book HENNECKE, M., E. and STORK, D., G., eds., Speechreading by Humans and Machines*, Springer, Berlin, pp. 331-349, 1996

[HEC02] HECKMAN, M., KROSCHEL, K., SAVARIAUX, C., BERTHOMMIER, F.: DCT-Based video features for audio-visual speech recognition. *In Proc. of the 7th ICSLP*, vol. 3, pp. 1925-1928. Denver, Colorado (USA), 2002

[HLA00] Hlaváč, V., Sedláček, M.: Zpracování signálu a obrazu. *Ve vydavatelství ČVUT*, Praha 2000, ISBN 80-01-02114-9

[HUA90] HUANG, X., D., ARIKI, Y., JACK, M., A.: Hidden Markov models for speech recognition. Edinburgh University Press, Edinburgh, United Kingdom, 1990

[HUA01] HUANG, X., ACERO, A., HON, H., W.: Spoken Language Processing. *In Prentice Hall PTR*, New jersey, United States of America, 2001, ISBN 0-13-022616-5

[CHA99] CHAN, M., T.: Automatic Lip Model Extraction for Constrained Contour-Based Tracking. *In Proc. of the IEEE International Conference on Image Processing*, Vol. 2 ,Kobe, Japan, October, 1999, pp. 848-851, ISBN 0-7803-5470-2.

[CHN01] CHAN, M., T.: HMM-BASED AUDIO-VISUAL SPEECH RECOGNITION INTEGRATING GEOMETRIC- AND APPEARANCE-BASED VISUAL FEATURES. In Proc. of IEEE Workshop on Multimedia Signal Processing, pp. 9-14, Cannes, France, Oct 3-5 2001

[CHE93] CHETVERIKOV, D., LERCH, A.: Multiresolution face detection. *In Theoretical Foundations of Computer Vision, volume 69 of Mathematical Research*, pp. 131-140., Akademie Verlg, 1993

[CHE95] CHEN, Q., WU, H., YACHIDA, M.: Face detection by fuzzy matching. *In Proc. of the Fifth IEEE International Conference on Computer Vision*, pp. 591-596, 1995

[CHE04] CHEN, S., LIU, J., ZHOU, Z.-H.: Making FLDA applicable to face recognition with one sample per person. *In Pattern Recognition*, pp. 1553-1555, 2004

[JAI02] JAIN, S., MOHSENI, M., RAJIV, J.: Automated Detection of Faces in an Image, *EE -368 Class Project*, http://ise.stanford.edu/2002projects/ee368/Project/reports/ee368group12.pdf, 2002

[KAR46] KARHUNEN, K.: Űber lineare methoden in der wahrscheinlichkeits-rechnung. *In Annales Academia Sciintiarum Fennicae, Series AI: Mathematica-Physica*, pp. 3-79, 1946

[KAU98] KAUCIC, R., BLAKE, A.: Accurate, Real-Time, Unadorned Lip Tracking. *In: Proc of the Sixth International Conference on Computer Vision*, Washington DC, USA, 1998, pp. 370-375, ISBN:81-7319-221-9

[KEI97] KEITH, J.: YCbCr to RGB Considerations. *In Application Note of Intersil Americas Inc.*, USA, 1997

[KJE96] KJELDSEN, R., KENDER, J.: Finding skin in color images. *In Proc. of the Second International Conference on Automatic Face and Gestures Recognition*, pp. 312-317, 1996

[KOL01] KOLEKTIV AUTORŮ, editor NOUZA, J.: Sborník článků – Počítačové zpracování řeči (cíle, problémy, metody a aplikace). *Ve vydavatelství Technické univerzity v Liberci*, Česká Republika, Liberec, 2001, ISBN 80-7083-551-6

[KRO97] KRONE, G., TALLE, B., WICHERT, A., PALM, G.: Neural architectures for sensorfusion in speech recognition. *In Proc. of Europ. Tut. Works: Audio-Visual Speech Processing*, Rhodes, Greece, pp. 57-60, 1997

[KRO02] KROSCHEL, K., HECKMANN, M.: Lip Parameter Extraction for Speechreading. *In Conf. El. Sprachsignalverarbeitung*, Dresden, pp.58-65, September 2002

[KWO94] KWON, Y., H., LOBO, V., N.: Face detection using templates. *In Proc. of IAPR International Conference on Pattern Recognition*, pp. 764-767, 1994

[LAM94] LAM, K., YAN, H.: Fast algorithm for locating head boundaries. *In Journal of Electronics boundaries,* pp. 351-359, 1994

[LEE04] LEE, B., HASEGAWA, M., B., GOUDESEUNE, C., KAMDAR, S., BORYS, S., LIU, M., HUANG, T.: AVICAR: Audio-Visual Speech Corpus in a Car Environment. *In INTERSPEECH 2004-ICSLP,* Jeju Island, Korea, October 2004, ISSN 1225-441x

[LEU95] LEUNG, T., K., BURL, M., C., PERONA, P.: Finding faces in cluttered scenes using random labeled graph matching. *In Proc. of the Fifth IEEE International Conference on Computer Vision*, pp. 637-644, 1995

[LIT88] LITTLESTONE, N.: learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *In Machine Learning*, 285-318, 1998

[LIE98] LIÉVIN, M., LUTHON, F.: Lip Features Automatic Extraction. *In Proc. of IEEE Conf. on Image Processing*, ICIP'98, Chicago, USA, vol. 3, pp. 168-172, oct. 1998.

[MAR99] MARTINEZ A., M.: Face Image Retrieval Using HMMs. *In Proc. of IEEE CVPR'99* (Workshop on Content-Based Access of Images and Video Libraries), , Ft. Collins, CO, USA. IEEE Computer Society 1999, ISBN 0-7695-0149-4, 23-25 June 1999

[MAT01] MATTHEWS, I., POTAMIANOS, G., NETI, C., LUETTIN, J.: A Comparison of Model and Transform-Based Visual Features for Audio-Visual LVCSR. *In Proc. of International Conference on Multimedia and Expo (ICME)*, Tokyo, 2001

[NAK00] NAKAMURA, S., ITO, H., SHIKANO, K.: Stream weight optimization of speech and lip image sequence for audio-visual speech recognition. In Proc. of Int. Conf. Spoken Language Processing, vol. III, Beijing, China, pp. 20-23, 2000

[NET01] NETI, CH., POTAMIANOS, G., LUETTIN, J., MATTHEWS, I., GLOTIN, H., VERGYRI, D.: LARGE-VOCABULARY AUDIO-VISUAL SPEECH RECOGNITION: A SUMMARY OF THE JOHNS HOPKINS SUMMER 2000 WORKSHOP, In Proc. of Workshop on Multimedia Signal Processing (Special Session on Joint Audio-Visual Processing), pp. 619-624, Cannes, 2001

[NOU95] NOUZA J: On the Speech Feature Selection Problem: Are Dynamic Features More Important Than the Static Ones? Proc. of EUROSPEECH'95 Conference, Madrid, Spain, Sept. 1995, pp. 919-923.

[NOU97] NOUZA, J., PSUTKA, J., UHLIŘ, J.: Phonetic Alphabet for Speech Recognition of Czech. *In Radioengineering*, vol.6, no.4, pp.16-20.,1997

[NOU00] NOUZA, J., MYSLIVEC, M.: Methods and Application of Phonetic Label Alignment in Speech Processing Tasks. *In Radioengineering*, vol.9, no.4, pp. 1-7 (ISSN 1210-2512)

[NOU01] NOUZA, T., NOUZA, J.: Graphic Platform for designing and developing practical voice interaction systems. *In Proc. of Eurospeech 2001*. Aalborg, Sept. 2001, pp.1287-1290. ISBN 87-90834-09-7. ISSN 1018-4074

[NOU04] NOUZA, J., NOUZA, T.: A Voice Dictation System for a Million-Word Czech Vocabulary. In: Proc. of ICCCT 2004, August 2004, Austin, USA, pp. 149-152, ISBN 980-6560-17-5

[OBR97] OBRECHT, R., A., JACOB, B., PARLANGEAU, N.: Audio-visual speech recognition and segmentation master slave HMM. *In Proc. of Europ. Tut. Works: Audio-Visual Speech Processing*, Rhodes, Greece, pp. 49-52, 1997

[OHM98] ÖHMAN, T.: An audio-visual speech database and automatic measurements of visual speech, *In Speech, Music and Hearing Quarterly Progress and Status Report.* Volume 34, KTH – Sweden, pp. 61-76, 1998

[PET84] PETAJAN, E., D.: Automatic lipreading to enhance speech recognition. In Proc. of Global Telecommunication Conference, Atlanta, USA, pp. 265-272, 1984

[POL02] POLLÁK, P.: Tvorba databází řečových signálů pro účely rozpoznávání a zvýrazňování. *Habilitační práce*, ČVUT, Praha, 2002

[POT01] POTAMIANOS, G., NETI, C.: Improved ROI and within frame discriminant features for lipreading, *In Proc. Int. Conf. Image Process.*, Thessaloniki, 2001

[POT01b] POTAMIANOS, G., NETI, C.: Automatic speechreading of impaired speech, *In Proc. Work. Audio-Visual Speech Process.*, Scheelsminde, 2001

[POT01c] POTAMIANOS, G., NETI, C., IYENGAR, G., HELMUTH, E.: Large-vocabulary audio-visual speech recognition by machines and humans, *In Proc. Eurospeech*, Aalborg, 2001.

[POT03] POTAMIANOS, G., NETI, C., GRAVIER, G., GARG, A.: Automatic Recognition of audio-visual speech: Recent progress and challenges. *In Proceedings of the IEEE*, vol. 91, no. 9, Sep. 2003

[POT04] POTAMIANOS, G.: Audio - Visual Automatic Speech Recognition: Visual Automatic Speech Recognition: Theory, Applications, and Challenges Theory, Applications, and Challenges. *In IBM T. J. Watson Research Center Yorktown Heights*, NY 10598, USA, 2004

[POT04b] POTAMIANOS, G., NETI, C., LUETTIN, J., MATTHEWS, I.: Audio-Visual Automatic Speech Recognition: An Overview. *In: Issues in Visual and Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier (Eds.), MIT Press (In Press), 2004

[PRI97] PŘIBIL, J.: Czech and Slovak TTS System Based on the Cepstral Speech Model. *In: Proc. of the 3th International Conference on Digital Signal Processing DSP '97*, Herl'any (Slovakia), September 3-4, 1997, pp. 23-26.

[PRO92] PROPP, M., SAMAL, A.: Artificial neural network architectures for human face detection. *In Intelligent Engineering Systems Through Artificial Neural Networks*, 1992

[PSU95] PSUTKA, J.: Komunikace s počítačem mluvenou řečí. *V nakladatelství Academia – Akademie věd České republiky*, Česká republika, Praha, 1995, ISBN 80-200-0203-0

[ROG98] ROGALEWICZ, V.: Pravděpodobnost a statistika pro inženýry. *Ve vydavatelství ČVUT*, ČR, Praha, 1998, ISBN 80-01-01740-0

[SAB98] SABER, E., TEKALP, A., M.: Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions. In Pattern Recognitions Letters, pp. 669-680, 1998

[SAM94] SAMARIA, F., YOUNG, S.: HMM based architecture for face identification. *In Image and Vision Computing*, pp. 537-583, 1994

[SAN00] SÁNCHEZ, M., U., R.: Aspects of facial biometrics or verification of personal identity. *PhD thesis*. Centre for Vision, Speech and Signal Processing, School of Electronic Engineering, Information technology and Mathematics, University of Surrey, Guildford, Surreychem, U.K., 2000

[SCA03] SCANLON, P., REILLY, R., B., DE CHAZAL, P.: Visual Feature Analysis for Automatic Speechreading. *In Audio Visual Speech Processing Conf.*, France, 2003

[SCA04] SCANLON, P., POTAMIANOS, G., LIBAL, V., CHU, S., M.: Mutual Information Based Visual Feature Selection for Lipreading. *In Proc. of ICSLP 2004*, October 2004, Jeju Island, Korea, ISSN 1225-441x

[SER82] SERRA, J. : Image Analysis and Mathematical Morphology. *In Academic Press*, GB, London, 1982

[SIR93] SIROHEY, S., A.: Human face segmentation and identification. *In Technical Report CS-TR-3176*, University of Meryland, 1993

[SON98] SONKA, M., HLAVÁČ, V., BOYLE, R.: Image Processing, Analysis, and Machina Vision, *In PWS Publishing*, 1998, ISBN 0-534-95393-X

[STE97] STEVE, Y., ODEL, J., OLLASON, D., VALTCHEV, V., WOODLAND, P.: The HTK Book, version 2.1. *In Cambridge University*, United Kingdom, 1997

[SUN98] SUNG, K., K., POGGIO, T.: Example-based learning for view-based human face detection. *In IEEE Transaction on Pattern Analysis and Machine Intelligence*, pp. 39-51, 1998

[TUR91] TURK, M., PENTLAND, A.: Eigenfaces for recognition. *In Journal of Cognitive Neuroscience*, pp. 71-86, 1991

[VER99] VERMA, A., FARUQUIE, T., NETI, C., BASU, S., SENIOR, A.: Late Integration in Continuous Audio-Visual Speech Recognition, *In ASRU*, Colorado, 1999

[WID60] WIDROW, B., HOFF JR., M., E.: Adaptive Switching Circuits. *In IRE Western Electric Show and Convention Record*, Part 4, pp. 96-104, 1960

[YAM02] YAMBOR, W., DRAPER, B., BEVERIDGE, R.: Analyzing PCA-based Face Recognition Algorithms: Eigenvector Selection and Distance Measures, *In Empirical Evaluation Methods in Computer Vision*, H. Christensen and J. Phillips (eds.), World Scientific Press, Singapore, 2002

[YAN98] YANG, M., H., AHUJA, N.: Detection human faces in color image. *In Proc. of IEEE International Conference on Image Processing*, volume 1, pp. 127-130,1998

[YAN00] YANG, M., H., AHUJA, N., KRIEGMAN, D.: Mixtures of linear subspaces for face detection. *In Proc. of the Fourth International Conference on Automatic Face and Gesture Recognition*, pp. 70- 76, 2000

[YAN00b] YANG, M., H., ROTH, D., AHUJA, N.: A SNoW-based face detector, *In Advanced in Neural Information Processing Systems 12*, pp. 855-861, MIT Press, 2000

[YAN01] YANG, M., H., AHUJA, N.: Face Detection and Gesture Recognition for Human-Computer Interaction. *In Kluwer Academic Publishers*, Boston, USA, 2001, ISBN 0-7923-7409-6

[ZEL02] Železný, M., Císař, P., Krňoul, Z., Novák, J.: Design of an Audio-Visual Speech Corpus for the Czech Audio-Visual Speech Synthesis. *In The 7th International Conference on Spoken Language Processing ICSLP2002*. Denver, U.S.A. 2002. pp. 1941-1944. (ISBN 1 876346 43 4), 2002

[ZHA01] ZHANG, X., BROUN, C., C.: Using lip features for multimodal speaker verification, *In A Speaker Odyssey - The Speaker Recognition Workshop*, Crete, Greece, pp.231-236., 2001

[ZOT02] ZOTKIN, D., N., DURAISWAMI, R., DAVIS, L., S. :Joint Audio-Visual Tracking Using Particle Filters, In EURASIP Journal on Applied Signal Processing (11), pp. 1154–1164, 2002

[XDB] The Extended M2VTS database, *In: http://xm2vtsdb.ee.surrey.ac.uk/*

Vlastní citované publikace:

[CHA01] CHALOUPKA, J., NOUZA, J.: Baldi (talking head) speaking Czech. *In Proc. of 11th Czech-German Workshop „Speech Processing".* Prague, September 2001. pp. 53-56. ISBN 80-86269-07-8

[CHA02] CHALOUPKA, J.: Talking Head: How Much Comprehensible Is It? *In Proc. of Radioelektronika 2002.* Bratislava, May 2002. pp. 202-205. ISBN 80-227-1700-2

[CHA02b] CHALOUPKA, J., NOUZA, J., PŘIBIL, J.: Czech-Speaking Artificial Face. *In Proc. of Biosignal 2002.* Brno, June 2002. pp. 403-405. ISBN 80-214-2120-7

[CHA02c] CHALOUPKA, J., NOUZA, J., DRÁBKOVÁ, J.: Developing an Artificial Talking Head for Czech Language. *In Proc. of SCI 2002.* Orlando USA, July 2002, Volume III. pp. 232-236. ISBN 980-07-8150-1

[CHA02d] CHALOUPKA, J.: Development of New Czech 3-D Talking Head. *In Proc. of 12th Czech-German Workshop „Speech Processing".* Prague, September 2002. pp. 54-58. ISBN 80-86269-09-4

[CHN02] NOUZA, J., KOLÁŘ, P., CHALOUPKA, J.: Voice Chat with a Virtual Character: The Good Soldier Švejk Case Project. In *Proc of TSD 2002.* Brno, September 2002. pp. 445-448. ISBN 0302-9743

[CHA03] CHALOUPKA, J.: Multimodal Signal Processing and Research. *In Proc. of Radioelektronika 2003.* Brno, May 2003. pp. 388-389. ISBN 80-214-2383-8

[CHA03b] CHALOUPKA, J.: The Czech Audio-Visual Speech Synthesizer System. *In Proc. of 6th International Workshop on Elektronics, Control, Measurment and Signals-ECMS 2003.* Liberec, June 2003. pp. 30-33. ISBN 80-7083-708-X

[CHA03c] CHALOUPKA, J.: The Czech Computerized Talking Head "Chatter". *In Proc. of 7th World Multiconference on Systemics, Cybernetics and Informatics-SCI 2003.* Orlando-USA, July 2003. Volume IV. pp. 320-323. ISBN 980-6560-01-9

[CHA03d] CHALOUPKA, J.: The Face Detection and Lips Tracking for Audio-Visual Speech Recognition. *In Proc. of 13th Czech-German Workshop „Speech Processing"*, September 2003, Prague, Czech Republic, pp. 141-145, ISBN 80-86269-10-8

[CHA04] CHALOUPKA, J.: Visual Signal Processing for Speech Recognition. *In Proc. of Radioelektronika 2004*, April 2004, Bratislava, Slovak Republic, pp. 406-409, ISBN 80-227-2017-8

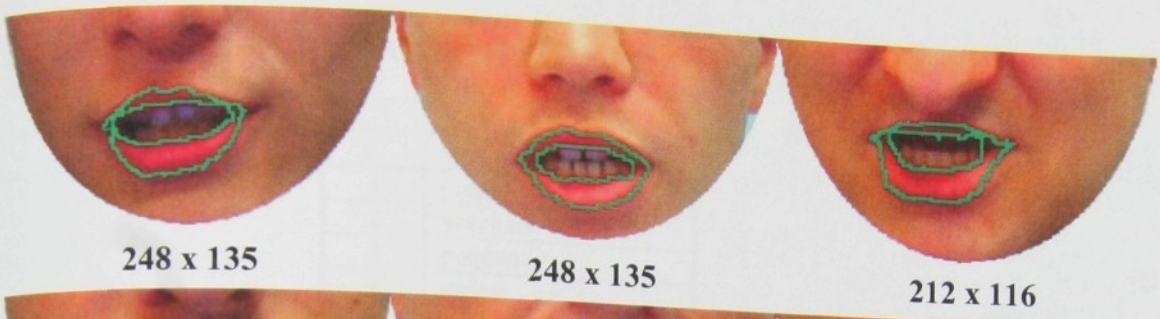
[CHA04b] CHALOUPKA, J., NOUZA, J.: Speech Recognition Supported by Camera Lips Reading. *In Proc. of ICCCT 2004*, August 2004, Austin, USA, pp. 116-119, ISBN 980-6560-17-5

[CHA04c] CHALOUPKA, J.: Automatic Lips Reading for Audio-Visual Speech Processing and Recognition. *In Proc. of ICSLP 2004*, October 2004, Jeju Island, Korea, pp. 2505-2508, ISSN 1225-441x

[CHA04d] CHALOUPKA, J.: Initial Experiments with Audio-Visual Isolated Words Recognition. *In Proc. of 14th Czech-German Workshop „Speech Processing"*, September 2004, Prague, Czech Republic, pp. 77-81, ISBN 80-86269-11-6

Příloha č. 1 – Detekování lidského obličeje

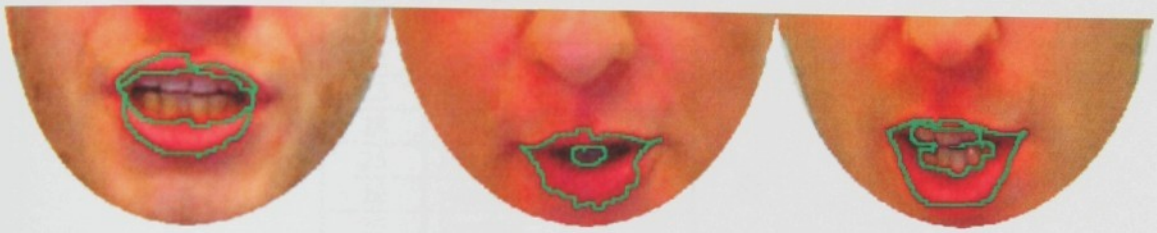
Příloha č. 2 – Nalezení hranic rtů



248 x 135 248 x 135 212 x 116

290 x 158 276 x 151 241 x 131

234 x 127 238 x 154 234 x 127



262 x 143 269 x 147 227 x 124

269 x 147 227 x 124 227 x 124

205 x 112 234 x 127

Příloha č. 3 – Slovník pro a-v nahrávky

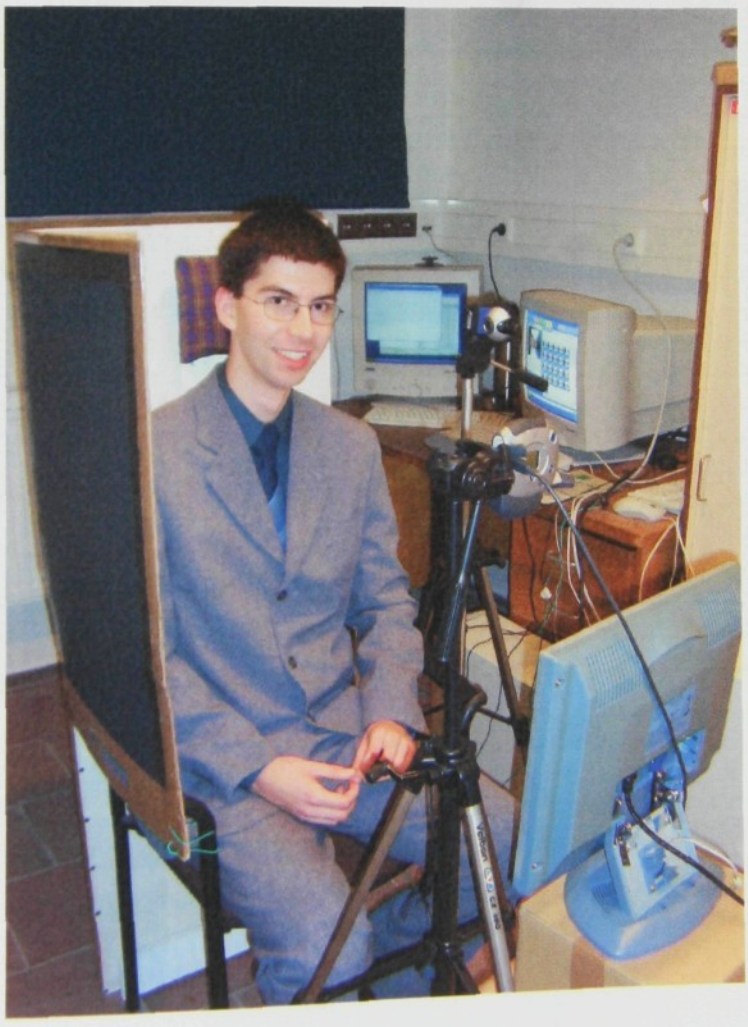
č.	slovo
1	protože
2	všechno
3	počítač
4	stejně
5	můžete
6	opravdu
7	skutečně
8	například
9	poslední
10	obvinění
11	peníze
12	dokonce
13	trochu
14	situace
15	problém
16	několik
17	vzpomínám
18	tisíce
19	taková
20	prostředí
21	prezident
22	hodiny
23	hlavní
24	doktor
25	řízení
26	rychle
27	kterého
28	abychom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

Příloha č. 4 – Tabulka četností fonémů z audio-vizuální databáze AVDB2cz

Č.	Foném[1]	PAC-CZ[2]	Příklad	Transkripce	Četnost	Relativní četnost [%]
1	„a“	a	táta	tát**a**	85	6,301
2	„á“	á	táta	t**á**ta	27	2,001
3	„b“	b	bába	**b**ába	23	1,705
4	„c“	c	ocel	o**c**el	24	1,779
5	„dz“	C	leckde	le**C**gde	6	0,445
6	„č“	č	čichá	**č**iXá	12	0,890
7	„dž“	Č	rádža	rá**Č**a	6	0,445
8	„d“	d	jeden	je**d**en	50	3,706
9	„ď“	ď	dělat	**ď**elat	12	0,890
10	„e“	e	lev	l**e**f	133	9,859
11	„é“	é	méně	m**é**ňe	12	0,890
12	„f“	f	fauna	**f**auna	12	0,890
13	„g“	g	guma	**g**uma	12	0,890
14	„h“	h	aha	a**h**a	25	1,853
15	„ch“	X	chudý	**X**udí	19	1,408
16	„i“ or „y“	i	bil, byl	b**i**l	87	6,449
17	„í“ or „ý“	í	vítr, lýko	v**í**tr, l**í**ko	57	4,225
18	„j“	j	dojat	do**j**at	36	2,669
19	„k“	k	kupec	**k**upec	44	3,262
20	„l“	l	dělá	ď**e**lá	57	4,225
21	„m“	m	máma	**m**á**m**a	48	3,558
22	„m“	M	tramvaj	tra**M**vaj	6	0,445
23	„n“	n	víno	ví**n**o	49	3,632
24	„n“	N	banka	ba**N**ka	6	0,445
25	„ň“	ň	koně	ko**ň**e	32	2,372
26	„o“	o	kolo	k**o**l**o**	87	6,449
27	„ó“	ó	óda	**ó**da	7	0,519
28	„p“	p	pupen	**p**u**p**en	37	2,743
29	„r“	r	bere	be**r**e	36	2,669
30	„ř“	ř	moře	mo**ř**e	11	0,815
31	„ř“	Ř	keř	ke**Ř**	12	0,890
32	„s“	s	sud	**s**ut	67	4,967
33	„š“	š	duše	du**š**e	15	1,112
34	„t“	t	dutý	du**t**í	43	3,188
35	„ť“	ť	kutil	ku**ť**il	23	1,705
36	„u“	u	duše	d**u**še	44	3,262
37	„ú“ or „ů“	ú	růže	r**ú**že	11	0,815
38	„v“	v	láva	lá**v**a	47	3,484
39	„z“	z	koza	ko**z**a	15	1,112
40	„ž“	ž	růže	rú**ž**e	14	1,038

1) Foném vyjádřený českými hláskami
2) Foném dle PAC [NOU97]

Příloha č. 5 – Pracoviště pro vytvoření a-v nahrávek



Příloha č. 6 – Tabulky

Tabulka č.1:

Rozpoznávací skóre [%] v závislosti na počtu stavů použitých (natrénovaných) HM modelů při měnícím se počtu **statických** energetických DCT vizuálních příznaků.

poč. pří.	Počet stavů HM modelů													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	8,8	8,4	8,8	8,4	10,8	11,6	11,2	14,8	13,6	14,0	18,0	21,2	19,6	20,8
2	5,6	12,4	16,4	17,6	17,2	18,4	20,0	22,0	24,4	22,0	24,0	27,2	27,6	26,8
3	6,4	14,8	14,4	19,2	22,0	19,6	23,6	24,0	23,6	27,6	27,6	28,4	31,2	33,6
4	6,8	11,6	14,4	18,0	18,8	23,2	22,4	25,6	23,6	26,8	32,0	32,4	29,6	34,0
5	6,0	12,4	13,2	18,4	22,4	22,0	20,8	26,0	27,6	27,6	30,8	31,6	30,8	34,0
6	6,8	12,8	16,4	19,6	20,8	20,8	26,0	26,0	28,8	31,6	29,6	34,0	34,0	33,6
7	7,2	13,6	15,2	18,0	18,8	23,6	28,0	27,2	28,4	30,4	29,2	34,0	33,6	33,6
8	6,4	12,4	16,0	18,4	22,8	24,4	25,2	27,2	29,6	29,2	31,2	35,2	33,6	32,0
9	6,4	12,8	18,0	20,4	24,4	26,4	28,8	29,6	29,2	30,4	33,6	34,8	32,0	30,8
10	5,6	12,0	14,8	18,8	20,8	26,4	23,6	28,4	28,0	30,8	31,6	34,0	**36,4**	34,0
11	5,6	13,2	14,0	18,0	20,4	22,8	26,0	29,2	29,2	28,8	29,6	30,8	32,0	34,8
12	6,0	12,0	14,0	14,0	22,0	24,4	22,4	26,8	31,6	29,2	28,0	30,0	30,8	30,4
13	5,6	11,2	11,2	13,2	20,4	25,6	23,6	27,2	24,0	29,2	28,4	30,0	30,8	30,4
14	5,2	10,8	12,8	16,0	22,4	24,0	26,8	26,8	24,8	29,6	29,6	31,2	31,6	32,4
15	4,4	11,6	12,0	15,2	19,6	22,4	23,2	26,8	27,2	31,2	29,6	31,2	32,4	31,2
16	4,8	11,6	11,6	14,0	21,2	20,4	26,0	28,4	29,6	28,0	27,6	31,2	33,2	30,8
17	4,8	10,4	12,0	16,4	18,8	20,8	22,8	26,4	28,4	28,0	26,0	30,4	32,4	30,8
18	4,8	13,2	12,8	16,0	21,2	22,8	22,0	27,2	26,0	25,6	25,2	31,2	30,0	30,4
19	4,4	11,6	10,0	14,4	20,8	21,2	23,6	22,4	28,4	24,0	26,4	27,6	29,2	27,6
20	4,0	12,4	11,6	14,4	17,2	20,4	24,4	22,8	25,2	24,4	26,4	29,6	29,6	28,4
21	4,0	10,4	11,6	13,2	16,4	18,0	23,6	22,0	25,2	28,4	24,8	29,6	30,4	31,2
22	3,6	10,0	13,2	15,6	17,6	16,8	21,6	25,2	25,6	27,6	25,2	28,0	31,2	31,6
23	3,2	10,0	14,0	16,0	16,4	20,0	22,8	24,8	26,4	24,8	27,2	31,6	30,0	31,2
24	3,6	8,8	13,6	14,8	18,4	18,4	24,4	26,0	26,0	28,4	27,2	32,4	28,0	31,2
25	4,0	8,8	13,2	15,6	16,0	20,8	21,2	24,4	25,6	27,6	27,6	30,4	28,8	27,2
26	3,6	9,6	14,4	14,0	17,2	19,6	22,0	24,8	26,8	27,2	26,4	30,0	31,2	27,2
27	3,6	11,2	14,0	13,2	16,0	18,4	21,2	22,4	23,6	24,4	25,2	29,2	31,2	30,0
28	3,2	10,8	12,8	13,2	16,8	19,2	22,4	23,2	23,6	22,8	24,8	27,6	28,0	27,6
29	3,2	11,2	14,4	14,8	16,4	18,8	23,6	22,4	23,2	20,8	24,8	26,8	29,2	27,2
30	3,6	10,4	14,8	14,0	15,2	18,8	21,6	21,6	24,0	20,8	25,6	24,8	26,8	28,0
40	4,0	10,8	11,6	11,2	12,4	16,0	14,8	21,2	19,2	19,2	24,4	21,2	22,8	22,8
50	3,2	8,8	8,8	11,2	10,4	16,0	12,8	16,4	17,2	19,2	19,2	22,4	22,8	22,8
60	2,0	6,4	7,6	12,0	12,8	13,2	12,8	14,0	12,8	18,0	19,2	20,4	22,0	21,6

Tabulka č.2:

Rozpoznávací skóre [%] v závislosti na počtu stavů použitých (natrénovaných) HM modelů při měnícím se počtu **dynamických** energetických DCT vizuálních příznaků.

poč.	Počet stavů HM modelů													
pří.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	4,8	4,8	7,2	4,8	6,8	6,4	8,0	6,8	8,8	12,4	10,4	11,6	15,2	13,6
2	5,2	8,4	10,4	16,0	14,0	19,2	21,2	21,2	24,4	24,0	24,4	25,6	25,2	24,8
3	4,8	6,8	11,6	15,2	17,2	20,0	22,0	24,0	26,4	28,8	27,6	27,6	29,2	31,6
4	6,0	7,2	10,0	15,2	17,6	18,8	17,6	25,2	26,4	29,6	27,2	29,6	30,4	32,4
5	6,0	6,8	8,8	14,8	16,4	24,0	19,6	26,8	32,0	30,8	30,8	30,4	32,0	33,2
6	5,6	5,2	9,2	16,0	18,8	23,2	27,2	27,2	28,0	30,0	30,0	28,4	28,0	30,0
7	6,0	6,8	7,2	15,6	20,4	21,6	25,6	31,2	28,4	28,8	30,4	33,6	31,6	32,4
8	6,4	7,6	9,6	13,6	18,8	20,0	22,4	28,0	28,4	28,0	30,4	30,4	32,8	31,2
9	6,4	9,2	9,6	13,2	17,2	20,0	19,2	27,2	25,2	29,6	34,0	30,0	32,8	33,6
10	6,0	7,2	9,2	12,4	16,8	19,2	23,6	26,0	26,4	25,6	30,4	30,8	33,2	35,2
11	5,6	6,8	7,6	10,8	17,2	22,0	23,2	25,2	29,6	30,4	29,2	30,8	33,6	33,2
12	6,0	6,8	8,4	12,0	16,8	19,2	23,6	20,4	30,4	28,0	30,4	30,4	34,4	31,2
13	6,4	7,2	10,0	12,0	14,8	22,8	24,4	23,2	25,6	26,8	26,8	30,0	32,8	35,2
14	6,0	5,2	7,2	11,6	14,8	22,0	22,8	24,4	28,0	24,4	26,0	32,0	34,8	31,6
15	6,4	6,8	9,2	12,0	16,0	23,2	22,8	26,4	25,6	28,4	32,0	34,4	33,2	33,6
16	6,4	8,0	8,4	12,0	19,2	20,0	22,8	22,4	27,2	27,6	29,2	33,6	32,4	**35,6**
17	7,2	6,4	10,8	10,8	17,2	19,2	20,0	26,8	26,0	28,0	31,6	33,2	31,6	30,0
18	7,6	8,0	8,8	9,6	18,8	19,2	22,4	23,6	24,8	28,4	27,6	34,8	27,2	34,0
19	6,0	7,6	9,2	12,0	17,2	19,2	18,0	22,0	27,2	27,6	30,8	**35,6**	31,6	32,4
20	4,4	8,8	9,6	13,2	19,6	20,0	21,2	20,4	24,4	27,2	31,2	30,0	32,0	30,4
21	4,0	7,6	10,0	13,6	18,8	19,2	19,6	19,2	21,2	26,4	32,4	26,0	33,2	30,4
22	3,6	7,2	8,0	13,2	15,6	18,0	21,6	20,0	23,6	22,4	27,2	28,8	31,2	29,6
23	4,0	6,0	8,0	12,8	15,6	17,6	18,0	16,8	25,2	27,2	29,2	27,6	30,4	29,6
24	4,0	7,2	9,6	14,4	14,8	19,2	18,4	22,0	24,4	25,2	30,0	24,8	30,4	30,4
25	4,0	7,6	9,6	14,0	15,2	16,4	14,0	19,2	23,6	22,8	26,4	26,0	30,8	32,4
26	4,4	8,4	8,4	13,6	15,6	16,4	16,4	17,2	21,6	24,4	26,0	26,8	28,4	26,8
27	4,0	8,4	8,0	12,8	14,0	16,0	14,8	15,6	24,0	21,2	22,4	25,6	30,0	26,4
28	4,0	8,0	8,4	12,4	14,8	16,4	15,6	18,4	23,2	22,4	22,8	24,4	27,6	26,0
29	3,6	7,6	10,0	12,4	11,6	15,2	14,0	18,8	22,4	20,0	24,0	24,0	26,4	27,2
30	3,6	6,8	9,6	12,4	12,8	14,4	14,8	16,8	22,8	20,0	22,0	24,0	26,8	26,0
40	5,2	8,8	8,4	11,2	10,8	14,8	11,6	15,6	18,0	18,4	18,0	21,6	24,4	22,4
50	4,8	7,6	5,6	10,0	11,2	12,0	11,6	14,8	14,4	15,6	16,4	18,8	22,8	16,8
60	4,8	6,4	7,6	9,2	8,8	10,0	11,6	13,2	14,4	13,6	16,0	14,8	17,6	19,6

Tabulka č.3:

Rozpoznávací skóre [%] v závislosti na počtu stavů použitých (natrénovaných) HM modelů při měnícím se počtu **akceleračních** energetických DCT vizuálních příznaků.

poč. pří.	Počet stavů HM modelů													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	3,6	3,2	3,2	2,0	3,2	4,0	4,4	3,2	3,6	4,0	4,0	5,2	4,0	2,8
2	4,4	4,4	6,0	7,2	4,0	6,4	6,8	8,0	8,8	10,8	12,0	13,2	11,2	13,6
3	2,0	3,6	4,4	5,2	6,0	9,2	8,0	9,2	7,6	10,8	8,8	11,6	12,0	15,2
4	2,0	4,0	3,2	6,4	9,2	7,2	7,6	8,8	8,4	8,8	8,8	10,4	12,0	11,6
5	2,4	2,8	4,0	6,8	8,8	8,8	7,6	8,8	9,2	10,4	10,8	11,2	10,4	10,0
6	2,8	3,2	4,8	6,0	5,6	6,8	8,4	10,8	10,4	9,6	9,6	11,6	12,8	14,0
7	3,2	3,6	2,8	4,8	6,4	6,4	8,8	9,2	10,4	8,4	11,2	12,8	11,2	12,8
8	3,2	4,0	5,6	5,2	6,0	4,0	8,4	7,6	8,8	10,0	10,0	12,4	14,8	17,6
9	4,0	4,4	4,4	6,0	7,2	7,6	6,0	9,6	8,8	10,8	10,4	15,6	13,6	16,0
10	3,6	4,4	6,0	8,8	5,6	7,6	8,0	8,4	9,6	11,6	15,6	16,0	14,8	18,8
11	3,6	4,4	6,0	8,8	7,2	8,4	8,0	10,4	10,4	12,0	12,8	**19,2**	15,2	14,8
12	3,2	4,0	6,0	8,0	7,6	8,4	7,6	10,0	11,2	12,0	12,4	14,0	17,6	16,0
13	3,2	4,4	5,6	9,6	6,4	8,8	8,4	8,8	7,2	10,0	10,8	16,8	17,2	14,8
14	3,2	4,4	6,4	8,4	7,2	10,4	9,2	9,2	8,8	8,8	14,8	14,4	15,2	18,0
15	2,8	5,6	4,4	7,2	6,0	10,0	6,4	8,0	7,6	8,4	10,8	16,0	15,6	16,8
16	4,0	4,8	5,6	8,8	8,0	7,2	8,0	8,4	9,2	9,2	10,0	13,2	16,0	14,0
17	3,6	4,4	4,4	8,4	5,6	8,0	7,6	6,8	10,4	8,8	9,2	11,6	14,4	14,8
18	3,6	3,6	4,4	11,2	8,0	7,6	6,8	8,4	7,6	8,4	10,0	12,4	14,8	14,8
19	4,0	3,2	4,0	9,2	6,8	7,6	9,6	8,4	6,8	6,8	11,2	14,0	15,2	13,2
20	3,6	4,0	5,2	4,8	5,6	6,0	7,2	8,0	8,8	7,6	9,6	10,8	13,6	13,6
21	3,6	3,2	3,2	6,4	5,6	6,0	8,0	8,4	8,0	5,6	10,8	11,2	15,6	14,8
22	3,6	4,0	3,2	7,2	6,4	5,6	7,6	8,8	6,8	8,8	9,6	10,4	14,0	14,8
23	4,0	4,0	4,0	8,4	6,4	7,2	6,0	7,6	7,6	13,2	9,6	9,2	14,0	14,8
24	3,6	3,6	2,8	8,0	6,4	8,4	6,4	10,4	8,4	13,2	10,0	12,0	16,4	15,6
25	3,6	4,0	4,0	6,8	6,8	7,2	6,4	11,2	8,8	12,8	9,6	12,0	15,6	14,8
26	3,6	4,0	3,6	8,4	4,8	8,0	8,0	9,6	7,6	10,8	9,6	12,8	12,0	15,2
27	3,6	3,6	4,0	7,6	4,4	8,0	6,8	10,4	8,0	10,0	9,6	11,2	12,4	14,4
28	3,2	5,6	5,6	5,6	7,2	6,4	6,4	10,4	6,8	10,8	6,8	11,2	12,0	12,0
29	3,2	4,4	6,4	6,4	5,6	6,0	5,6	10,0	8,8	9,2	9,6	12,4	14,4	12,4
30	3,2	5,6	5,2	7,2	8,0	4,4	6,4	10,8	8,4	10,4	8,8	10,8	12,4	14,0
40	4,0	2,4	4,8	5,2	5,6	5,2	5,2	8,0	7,6	8,0	6,8	5,2	12,4	10,0
50	3,6	3,6	4,8	4,8	6,4	4,8	7,2	7,2	7,6	6,8	8,8	8,0	10,8	9,2
60	3,2	4,0	4,8	4,8	4,4	5,2	7,6	6,0	7,6	6,4	6,4	8,0	7,6	8,4

Tabulka č.4:

Rozpoznávací skóre [%] v závislosti na počtu stavů použitých (natrénovaných) HM modelů při měnícím se počtu **statických a dynamických** energetických DCT vizuálních příznaků.

poč.	Počet stavů HM modelů													
pří.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	7,2	8,4	8,8	9,6	11,6	13,2	15,6	15,2	17,6	18,8	20,4	20,0	19,6	25,6
4	4,4	12,4	15,6	16,8	20,4	22,4	26,4	29,2	30,0	30,0	30,8	32,0	32,0	33,6
6	7,2	12,0	17,2	20,4	24,8	28,0	30,0	30,4	28,8	32,8	35,6	38,4	42,8	40,8
8	6,0	13,2	19,2	18,8	25,2	26,8	31,2	31,2	32,8	37,2	41,6	40,4	42,4	40,0
10	7,6	14,0	17,6	20,0	21,6	25,6	29,6	30,8	34,8	38,4	38,8	38,8	41,2	**44,4**
12	6,8	12,4	19,6	23,2	26,0	31,6	30,4	34,0	35,2	38,0	37,2	38,8	**44,4**	42,8
14	7,2	12,4	22,0	26,4	28,4	30,4	30,8	34,8	38,4	41,6	38,4	39,6	40,8	41,6
16	7,6	13,2	20,8	26,4	28,8	32,0	35,6	35,2	34,0	36,0	37,6	36,8	41,2	41,2
18	6,0	14,0	22,8	23,2	27,6	30,0	36,4	38,0	34,4	36,0	40,0	38,8	42,8	44,0
20	6,0	15,6	21,6	25,2	28,0	31,2	32,4	38,0	36,8	40,0	40,8	41,2	**44,4**	42,4
22	5,6	15,2	20,8	22,4	28,4	31,6	35,2	36,4	36,4	40,4	39,2	37,6	42,4	42,4
24	6,0	16,8	20,8	23,6	25,6	29,2	32,4	34,4	33,2	35,2	38,8	36,8	38,8	40,8
26	6,0	15,6	21,6	22,4	25,2	29,6	32,0	35,2	38,0	37,2	36,4	36,8	37,2	40,4
28	6,4	17,2	19,6	20,4	25,6	26,8	29,6	34,8	37,6	39,2	40,0	37,2	36,4	38,8
30	4,8	18,4	18,4	22,8	22,8	24,8	34,8	35,2	36,8	38,0	36,0	36,4	39,6	38,4
32	4,4	16,8	19,2	23,2	24,4	31,2	32,4	34,4	34,4	36,0	34,8	36,8	38,0	39,2
34	4,0	16,0	18,4	23,6	24,8	32,4	32,4	30,0	33,6	37,6	37,2	35,2	36,0	37,2
36	4,8	17,2	20,8	22,0	26,8	30,0	32,4	32,4	34,4	34,8	34,8	36,4	35,6	39,2
38	4,4	16,0	16,0	22,0	26,0	31,6	30,0	26,8	32,4	34,4	35,2	35,6	36,0	40,0
40	4,4	16,8	18,4	24,4	25,6	30,4	30,4	27,6	32,0	31,2	34,0	35,2	38,4	39,2
42	4,8	15,6	18,0	22,8	26,8	25,6	27,6	31,6	35,6	29,2	35,2	36,4	37,2	40,8
44	4,0	15,6	18,8	19,2	23,2	27,2	30,4	29,2	32,8	31,2	34,4	34,0	36,8	37,6
46	4,0	15,2	16,8	20,8	23,6	22,8	30,0	26,4	34,4	32,8	32,8	34,0	38,4	38,4
48	3,6	13,2	18,0	21,6	22,4	24,8	27,6	24,8	32,0	31,2	30,8	32,8	38,0	37,6
50	3,6	12,8	18,8	20,8	22,0	26,4	26,0	31,2	29,2	29,6	30,8	34,0	35,6	37,2
52	3,2	12,4	20,8	22,4	22,0	21,6	22,8	31,2	26,0	29,6	29,6	34,0	35,2	35,2
54	3,6	11,2	18,4	18,8	21,6	22,8	20,8	30,4	29,2	30,4	29,6	32,8	32,8	36,0
56	4,0	10,4	20,0	18,4	22,4	25,6	22,0	29,6	31,6	28,4	31,2	33,2	32,4	32,4
58	4,8	10,8	16,0	17,2	18,0	23,2	21,2	29,2	28,0	28,8	31,6	29,2	34,0	34,4
60	4,0	10,0	16,4	16,4	19,6	20,4	22,4	27,6	30,4	32,0	32,4	31,2	32,4	32,4
80	5,2	10,8	16,4	17,6	17,2	19,6	21,6	23,2	27,2	26,4	27,6	28,4	28,8	26,8
100	5,2	12,8	10,0	13,2	18,0	15,2	16,8	20,8	22,0	23,2	24,4	24,0	26,0	27,6
120	4,4	12,0	8,0	13,2	18,0	18,4	17,2	17,6	21,2	24,8	22,4	23,6	22,4	23,6

Tabulka č.5:

Rozpoznávací skóre [%] v závislosti na počtu stavů použitých (natrénovaných) HM modelů při měnícím se počtu **dynamických a akceleračních** energetických DCT vizuálních příznaků.

poč.	Počet stavů HM modelů													
pří.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	4,0	3,6	3,2	3,6	6,8	7,2	7,2	8,0	7,6	9,2	11,6	12,0	13,6	15,2
4	5,6	7,2	10,0	12,0	15,6	15,2	14,0	20,8	19,6	24,0	24,8	22,8	22,8	23,2
6	4,0	5,2	9,6	9,6	14,0	15,6	14,8	20,0	20,8	23,6	20,8	24,0	26,0	26,4
8	4,4	5,6	8,0	11,6	14,0	14,4	13,6	19,6	20,8	21,6	26,8	24,8	25,6	24,4
10	4,0	5,2	8,0	13,6	15,6	16,0	16,8	21,2	20,8	24,8	27,2	30,4	27,6	28,8
12	4,4	5,2	5,6	11,2	12,4	14,4	15,6	22,8	22,0	26,4	25,6	26,8	28,0	29,2
14	4,8	6,0	6,0	11,6	14,0	16,0	18,0	18,0	20,0	21,2	24,0	25,2	24,4	25,6
16	5,2	4,4	5,2	12,0	12,8	17,2	16,4	17,2	20,0	21,6	22,4	26,4	26,8	28,4
18	5,6	7,2	5,6	13,2	15,2	14,4	17,2	16,0	20,4	22,0	26,0	27,6	25,2	28,8
20	7,2	6,0	8,4	11,6	14,4	17,6	15,2	18,8	24,4	19,2	21,6	26,4	26,0	27,2
22	7,2	5,2	8,8	10,4	13,2	16,0	16,8	16,8	24,8	22,0	25,2	25,6	27,6	28,8
24	4,8	5,2	7,2	10,0	11,2	15,2	18,4	19,6	22,8	25,6	22,8	24,0	28,8	27,6
26	4,8	6,4	6,8	11,2	12,8	15,6	17,2	20,4	21,2	22,4	24,4	25,2	27,2	27,6
28	5,2	6,8	8,4	9,6	10,4	15,2	18,4	16,8	20,4	22,0	22,4	24,0	23,6	27,6
30	5,6	5,2	8,8	8,4	13,6	14,4	18,8	18,0	18,0	23,6	21,6	25,6	26,8	28,4
32	5,2	6,0	8,8	10,8	9,6	12,0	16,4	19,2	20,8	22,0	22,8	28,0	28,4	**31,6**
34	4,4	4,8	8,4	10,0	8,4	13,2	16,8	16,0	20,8	21,2	22,0	28,0	27,2	30,8
36	4,0	4,8	7,6	8,8	10,4	12,0	16,8	15,2	23,6	22,0	21,6	30,8	28,0	30,8
38	4,0	5,6	7,6	10,4	11,2	12,0	15,2	15,6	22,8	18,0	22,4	27,2	27,6	26,4
40	3,6	5,6	8,0	10,0	10,8	10,8	14,4	14,8	18,8	19,6	19,6	25,6	28,4	27,2
42	4,4	5,2	7,6	10,4	9,6	10,0	14,8	14,4	19,2	20,0	18,8	23,2	26,8	26,4
44	4,8	6,4	9,2	10,0	12,4	10,0	14,8	14,4	21,2	20,8	20,0	22,4	25,2	26,0
46	4,8	6,0	9,2	12,0	8,8	10,4	14,0	16,8	20,0	19,6	17,2	24,0	25,2	26,4
48	5,2	6,8	8,0	9,6	10,0	13,6	12,8	16,0	18,0	20,8	19,2	22,4	27,6	25,6
50	4,4	6,8	7,6	10,8	10,8	14,4	13,2	13,6	16,8	20,8	19,2	21,6	22,0	26,8
52	4,8	6,0	6,0	9,6	11,2	12,0	12,4	14,0	17,6	18,8	16,8	24,8	25,2	24,0
54	4,4	6,4	7,6	10,0	12,0	10,8	12,8	12,8	20,4	19,2	16,0	22,8	23,2	21,6
56	4,0	4,8	7,2	11,6	11,2	10,4	12,0	13,6	19,2	17,2	18,0	22,4	24,8	23,2
58	4,0	5,6	7,2	10,4	9,6	11,2	13,2	14,4	18,4	18,8	18,8	21,6	21,2	23,2
60	4,0	6,0	5,6	8,8	10,8	12,4	12,4	14,8	17,2	17,2	18,0	18,4	19,6	22,8
80	4,8	6,0	4,8	10,8	9,6	10,0	10,8	14,0	13,6	15,6	15,6	20,0	16,8	19,2
100	2,8	5,6	5,6	7,2	8,8	9,6	9,2	11,2	14,0	13,6	12,4	21,2	16,4	18,4
120	4,0	6,4	4,8	6,8	8,8	12,4	10,8	9,6	12,4	13,6	14,0	14,8	14,0	14,0

Tabulka č.6:

Rozpoznávací skóre [%] v závislosti na počtu stavů použitých (natrénovaných) HM modelů při měnícím se počtu **statických a akceleračních** energetických DCT vizuálních příznaků.

poč.	Počet stavů HM modelů													
pří.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	6,0	8,8	8,4	8,0	13,6	9,6	14,4	13,6	16,4	17,6	17,6	20,8	20,4	21,6
4	6,0	12,8	16,0	19,2	22,0	22,0	25,2	26,8	26,8	28,0	28,8	28,8	26,0	31,6
6	6,0	10,8	12,4	18,8	18,4	24,8	26,0	28,0	24,8	27,6	31,6	32,8	32,0	33,6
8	4,8	10,0	12,4	16,4	19,6	22,0	25,2	28,0	27,6	27,2	31,2	32,4	37,2	36,0
10	4,0	11,6	14,0	14,8	19,2	26,8	26,8	26,8	30,4	32,4	32,8	34,8	38,0	39,2
12	5,6	10,0	16,8	17,2	23,6	24,0	28,0	30,4	32,0	34,0	35,6	38,4	39,6	**40,4**
14	6,0	12,0	16,8	17,6	22,0	26,8	30,0	31,2	31,6	32,8	33,6	36,4	40,0	39,2
16	4,8	12,8	15,6	19,6	25,2	27,6	32,0	30,8	32,8	34,0	35,6	33,6	38,4	38,0
18	5,6	12,8	17,6	20,0	26,0	29,6	29,2	30,4	33,6	35,6	37,6	34,8	38,0	36,0
20	6,4	11,6	18,0	19,2	25,2	24,8	33,6	33,6	32,4	37,6	37,2	35,2	36,8	38,4
22	6,0	13,2	18,0	21,2	24,8	25,6	32,0	34,4	34,0	33,2	35,2	38,4	38,0	39,2
24	6,4	13,2	18,0	19,6	25,2	24,0	32,0	30,4	32,8	33,2	33,6	35,6	36,4	37,2
26	5,2	12,4	18,8	18,8	25,6	24,8	28,8	28,0	30,8	29,6	32,4	33,2	34,0	36,4
28	3,6	14,4	18,0	20,8	24,8	26,4	28,4	28,0	30,8	30,4	30,8	36,4	34,0	36,8
30	4,0	14,4	18,0	22,0	24,8	26,0	24,8	29,2	31,6	31,2	34,0	38,0	36,4	35,2
32	3,2	12,4	21,2	19,2	28,0	26,8	28,8	26,4	31,6	31,6	32,0	35,2	36,0	37,6
34	3,2	12,8	19,2	20,4	25,6	23,2	24,0	23,6	31,6	32,8	34,0	34,4	37,2	34,8
36	3,2	13,2	18,0	20,4	23,2	25,2	27,6	26,8	32,0	34,8	33,2	34,4	39,6	34,4
38	3,6	14,4	18,8	21,6	23,6	27,6	24,8	26,8	33,2	32,4	32,8	38,8	34,8	33,6
40	2,8	15,2	19,2	20,8	19,6	24,4	27,6	26,0	32,4	34,0	34,4	36,0	36,4	33,6
42	3,2	13,6	16,4	20,8	19,2	25,2	23,2	25,2	27,2	32,4	32,0	34,8	35,6	34,8
44	3,2	13,2	14,8	19,6	17,6	21,2	22,4	24,4	27,6	29,6	31,2	34,4	34,4	31,6
46	3,6	12,4	16,0	18,0	17,2	20,8	23,2	24,8	30,0	29,2	29,6	32,4	32,8	33,2
48	3,6	11,6	14,0	17,2	16,0	21,2	22,8	26,4	27,2	29,6	32,8	30,4	32,8	34,0
50	4,0	11,2	14,4	18,0	16,8	20,0	24,0	27,6	28,0	28,8	30,0	32,4	33,2	33,6
52	3,6	10,8	14,0	17,2	17,6	20,4	23,2	27,2	28,4	26,8	30,0	32,0	33,2	33,6
54	3,2	10,4	12,8	18,4	18,0	18,8	22,0	28,8	27,6	26,0	25,6	32,8	32,4	31,6
56	2,8	9,6	12,0	18,0	18,8	17,2	22,0	27,2	29,2	29,6	29,6	30,0	28,4	33,2
58	3,2	10,8	12,4	18,0	17,6	18,4	20,0	24,8	27,6	25,2	26,8	30,4	30,0	34,0
60	2,8	9,6	12,0	17,6	16,8	19,6	20,4	25,6	25,2	25,6	28,0	28,8	28,0	32,8
80	4,8	9,2	8,8	12,0	17,2	17,2	18,0	22,0	21,2	22,8	23,2	27,6	30,0	29,2
100	4,4	10,8	9,2	13,6	12,8	14,0	16,0	18,8	18,0	21,6	23,2	21,2	20,4	27,2
120	4,8	10,4	7,6	14,0	12,8	14,4	16,0	17,2	16,0	22,8	19,2	26,4	25,6	23,2

Tabulka č.7:

Rozpoznávací skóre [%] v závislosti na počtu stavů použitých (natrénovaných) HM modelů při měnícím se počtu **statických, dynamických a akceleračních** energetických DCT vizuálních příznaků.

poč. pří.	Počet stavů HM modelů													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
3	5,2	9,2	9,6	10,4	12,4	12,0	13,6	16,0	18,0	18,4	20,8	21,6	21,6	23,2
6	5,2	9,6	15,6	18,0	22,8	24,4	25,6	24,8	30,0	32,0	30,8	31,2	33,6	34,8
9	5,2	10,8	14,4	15,6	23,2	28,8	32,8	31,6	33,2	33,6	37,6	38,4	40,4	39,2
12	5,2	10,4	17,2	19,2	23,2	25,6	30,4	29,2	32,0	35,2	36,0	38,0	37,6	43,6
15	5,2	11,2	13,2	19,2	22,8	26,4	30,8	32,4	32,0	34,8	36,8	38,8	37,6	**45,2**
18	5,6	11,2	16,4	20,4	23,2	26,4	32,4	32,8	36,8	36,4	39,2	40,4	40,8	44,0
21	5,6	8,4	15,2	19,6	22,8	28,8	33,2	32,0	34,4	36,8	38,4	38,4	38,8	40,8
24	6,0	10,8	16,8	21,6	25,2	26,8	28,8	30,0	37,6	36,8	38,8	42,4	38,0	38,8
27	6,0	11,2	18,4	21,2	25,2	28,4	30,0	33,6	33,6	40,4	37,2	40,4	38,8	41,2
30	5,2	12,8	18,8	23,2	29,2	29,6	33,2	31,6	35,6	36,4	37,6	40,4	39,6	41,2
33	4,8	11,6	17,6	21,6	23,2	28,8	30,8	34,4	36,0	37,6	36,4	40,4	38,8	41,2
36	5,2	10,8	17,2	17,6	23,2	23,6	28,4	32,8	34,8	36,4	37,2	42,8	36,0	38,8
39	4,8	14,0	15,2	19,6	24,4	26,0	26,4	30,8	37,6	38,8	40,8	40,4	37,6	37,2
42	4,4	13,6	18,0	20,4	22,8	26,8	30,0	31,6	33,6	34,4	34,8	37,2	41,6	36,0
45	4,0	12,8	16,0	20,8	22,4	29,2	28,4	33,6	32,8	31,2	35,2	38,0	40,8	38,0
48	3,2	14,0	16,0	22,4	22,4	28,8	26,0	28,8	33,2	32,8	34,0	39,6	41,2	40,4
51	3,6	14,4	15,2	18,8	23,2	30,0	32,0	27,6	32,8	34,4	34,8	38,8	40,8	39,6
54	3,2	14,8	14,4	20,8	24,0	24,0	29,2	31,2	33,2	35,2	36,0	35,2	41,6	40,8
57	3,2	13,6	14,0	21,2	24,4	24,4	28,4	30,8	31,2	32,4	34,8	36,4	42,4	39,6
60	3,2	14,4	14,4	20,8	23,2	26,0	27,2	29,2	33,2	33,6	36,4	37,6	43,6	38,4
63	3,2	13,2	13,6	19,2	23,2	24,0	26,4	28,0	34,0	32,0	33,6	35,2	41,6	40,4
66	3,6	12,4	14,4	19,2	21,2	25,6	27,2	30,0	30,0	30,8	30,8	35,6	41,2	39,6
69	3,6	12,0	14,4	20,4	18,4	23,6	28,0	26,0	30,0	32,8	32,8	33,6	42,4	39,6
72	3,2	11,6	14,0	20,0	16,8	24,0	23,6	26,4	32,0	31,6	30,8	33,2	38,4	37,2
75	3,2	13,2	14,4	18,0	17,2	26,0	25,6	25,6	31,2	28,8	35,6	34,0	40,0	38,0
78	3,2	14,4	14,4	20,0	17,6	24,4	22,8	27,6	31,2	31,2	32,4	33,6	36,0	40,4
81	3,6	12,4	13,2	18,4	18,8	24,0	22,0	29,2	30,0	30,4	34,4	31,6	34,0	36,4
84	3,2	10,4	13,6	18,4	19,2	24,0	22,8	28,0	26,4	28,4	32,8	34,8	32,8	35,2
87	3,6	10,4	12,8	18,4	18,4	24,0	22,0	26,4	25,6	27,2	29,6	34,8	31,6	36,4
90	3,2	8,8	13,2	16,0	19,6	20,0	19,6	23,2	23,6	26,8	27,6	32,8	32,8	34,4
120	4,0	10,0	12,0	14,0	20,4	17,6	18,4	23,6	25,2	23,6	26,8	32,4	26,8	30,0
150	3,6	10,8	13,2	12,8	16,0	18,0	18,4	20,4	20,4	22,8	23,6	26,0	22,8	29,6
180	4,4	8,4	12,8	14,0	15,6	16,4	16,8	16,0	20,8	22,4	20,8	26,8	25,6	26,8

Příloha č.7: AUDIO-VIZUÁLNÍ ROZPOZNÁVÁNÍ ŘEČI – GEOMETRICKÉ VIZUÁLNÍ PŘÍZNAKY

γ_2 viz.	γ_1 audio																				
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0.0	7.6|7.6	7.2|99.2	10.8|99.2	12.8|99.2	13.2|99.2	12.4|99.2	12.4|99.2	12.0|99.2	12.4|99.2	13.6|99.2	13.6|99.2	13.6|99.2	13.6|99.2	13.2|99.2	13.2|99.2	12.8|99.2	12.8|99.2	12.8|99.2	12.0|99.2	12.0|99.2	12.0|99.2
0.1	32.8|32.8	12.0|99.2	16.4|99.2	16.0|99.2	14.8|99.2	14.8|99.2	14.4|99.2	13.6|99.2	14.0|99.2	14.4|99.2	14.8|99.2	14.8|99.2	14.8|99.2	14.8|99.2	14.4|99.2	14.0|99.2	14.0|99.2	14.0|99.2	12.8|99.2	12.8|99.2	12.8|99.2
0.2	30.8|30.8	14.8|98.8	20.4|99.2	18.8|99.2	16.8|99.2	15.2|99.2	15.2|99.2	14.8|99.2	14.8|99.2	14.8|99.2	14.8|99.2	15.2|99.2	15.2|99.2	14.8|99.2	14.8|99.2	14.4|99.2	14.4|99.2	14.4|99.2	14.0|99.2	14.0|99.2	14.0|99.2
0.3	34.4|34.4	24.4|98.8	24.4|99.2	22.4|99.2	18.8|99.2	18.0|99.2	16.8|99.2	15.2|99.2	15.2|99.2	15.2|99.2	15.6|99.2	15.2|99.2	16.0|99.2	15.6|99.2	15.6|99.2	14.4|99.2	14.4|99.2	14.4|99.2	14.0|99.2	14.0|99.2	14.0|99.2
0.4	35.2|35.2	34.8|98.4	28.0|98.8	25.2|99.2	22.8|99.2	20.4|99.2	18.0|99.2	17.6|99.2	17.2|99.2	16.0|99.2	16.0|99.2	15.6|99.2	15.6|99.2	15.2|99.2	15.6|99.2	14.8|99.2	14.8|99.2	14.8|99.2	14.8|99.2	14.4|99.2	14.4|99.2
0.5	38.0|38.0	44.4|97.6	31.2|98.4	27.2|99.2	25.2|99.2	23.2|99.2	20.4|99.2	19.2|99.2	17.6|99.2	17.6|99.2	16.8|99.2	16.0|99.2	15.6|99.2	15.2|99.2	14.8|99.2	14.8|99.2	14.8|99.2	15.2|99.2	14.8|99.2	14.8|99.2	14.8|99.2
0.6	32.8|32.8	48.0|97.6	38.0|98.4	28.8|98.4	26.4|99.2	24.4|99.2	23.6|99.2	20.0|99.2	19.2|99.2	17.6|99.2	18.4|99.2	17.6|99.2	16.4|99.2	15.6|99.2	15.6|99.2	14.8|99.2	14.4|99.2	15.2|99.2	15.2|99.2	14.8|99.2	14.8|99.2
0.7	36.8|36.8	51.2|98.0	42.0|98.4	33.6|98.4	29.2|99.2	25.6|99.2	26.0|99.2	23.2|99.2	20.4|99.2	19.2|99.2	19.2|99.2	18.4|99.2	18.0|99.2	17.2|99.2	16.0|99.2	15.2|99.2	15.2|99.2	15.2|99.2	15.2|99.2	15.2|99.2	14.8|99.2
0.8	38.0|38.0	54.0|96.8	44.0|98.4	35.6|98.4	31.2|98.8	27.6|99.2	25.2|99.2	25.2|99.2	23.2|99.2	20.4|99.2	19.2|99.2	19.6|99.2	18.4|99.2	18.0|99.2	17.6|99.2	16.8|99.2	16.0|99.2	16.0|99.2	15.6|99.2	15.6|99.2	15.6|99.2
0.9	36.0|36.0	58.4|95.2	47.2|98.4	36.0|98.4	31.6|98.4	28.8|99.2	26.0|99.2	26.0|99.2	24.8|99.2	23.2|99.2	22.0|99.2	19.2|99.2	19.6|99.2	18.4|99.2	18.4|99.2	17.6|99.2	17.2|99.2	17.2|99.2	16.0|99.2	16.0|99.2	15.6|99.2
1.0	36.8|36.8	58.8|94.4	51.6|98.0	38.8|98.4	34.0|98.4	31.2|98.8	28.4|99.2	25.6|99.2	25.6|99.2	24.4|99.2	23.2|99.2	22.0|99.2	21.2|99.2	19.6|99.2	19.2|99.2	18.8|99.2	17.6|99.2	17.6|99.2	17.6|99.2	16.8|99.2	16.4|99.2
1.1	36.0|36.0	59.6|94.0	52.0|98.0	41.2|98.4	36.0|98.4	30.8|98.4	28.8|99.2	27.2|99.2	25.6|99.2	25.6|99.2	24.4|99.2	23.2|99.2	22.0|99.2	21.2|99.2	19.2|99.2	19.6|99.2	18.0|99.2	17.6|99.2	17.6|99.2	17.6|99.2	17.6|99.2
1.2	36.8|36.8	59.6|93.2	54.8|98.0	42.0|98.0	37.2|98.4	32.4|98.4	30.8|99.2	28.8|99.2	26.0|99.2	25.2|99.2	26.0|99.2	24.4|99.2	23.2|99.2	21.6|99.2	20.8|99.2	19.6|99.2	18.8|99.2	18.0|99.2	17.6|99.2	17.6|99.2	17.6|99.2
1.3	36.8|36.8	61.2|91.6	57.2|97.2	44.8|98.0	40.0|98.4	34.4|98.4	30.8|98.8	29.2|99.2	27.6|99.2	25.6|99.2	26.0|99.2	26.0|99.2	24.0|99.2	22.4|99.2	21.2|99.2	20.8|99.2	19.6|99.2	18.8|99.2	18.8|99.2	18.0|99.2	17.6|99.2
1.4	36.8|36.8	61.2|91.2	58.4|97.2	47.6|98.0	40.8|98.4	36.8|98.4	32.4|98.4	30.8|99.2	29.6|99.2	26.8|99.2	25.6|99.2	26.0|99.2	25.6|99.2	23.6|99.2	22.8|99.2	21.2|99.2	20.4|99.2	20.0|99.2	18.8|99.2	19.2|99.2	18.0|99.2
1.5	36.4|36.4	62.0|90.4	58.0|96.4	47.6|98.0	42.4|98.4	37.6|98.4	33.6|98.4	30.8|98.8	30.4|99.2	29.2|99.2	26.4|99.2	25.6|99.2	26.0|99.2	25.2|99.2	23.6|99.2	22.4|99.2	20.8|99.2	20.4|99.2	20.0|99.2	18.8|99.2	18.8|99.2
1.6	36.4|36.4	61.2|89.6	58.4|96.4	50.4|97.6	44.4|98.0	39.6|98.4	35.2|98.4	31.6|98.8	30.8|99.2	30.4|99.2	28.0|99.2	26.0|99.2	25.6|99.2	25.6|99.2	24.8|99.2	22.8|99.2	22.4|99.2	20.8|99.2	20.4|99.2	20.0|99.2	19.2|99.2
1.7	37.2|37.2	62.4|90.0	59.6|95.2	50.8|97.6	45.6|98.0	40.8|98.4	37.2|98.4	32.8|98.4	30.8|98.8	31.2|99.2	29.6|99.2	26.4|99.2	25.6|99.2	25.6|99.2	25.2|99.2	24.8|99.2	22.8|99.2	22.4|99.2	20.8|99.2	20.4|99.2	20.4|99.2
1.8	34.8|34.8	**64.0**|**88.8**	60.0|95.2	52.0|97.6	46.8|98.0	41.6|98.4	37.6|98.4	33.6|98.4	31.6|98.8	31.6|99.2	30.4|99.2	28.4|99.2	26.8|99.2	25.2|99.2	25.6|99.2	25.2|99.2	24.8|99.2	22.8|99.2	22.4|99.2	20.8|99.2	20.4|99.2
1.9	35.2|35.2	60.8|87.2	60.8|94.8	53.6|97.6	46.8|98.0	42.8|98.0	38.8|98.4	35.6|98.4	33.2|98.4	31.2|98.8	31.2|99.2	29.6|99.2	26.8|99.2	25.6|99.2	25.2|99.2	25.6|99.2	25.2|99.2	24.4|99.2	22.8|99.2	22.4|99.2	21.2|99.2
2.0	36.0|36.0	60.4|85.6	62.0|93.6	55.2|97.6	47.6|98.0	45.2|98.4	41.2|98.4	37.2|98.4	33.2|98.4	31.2|98.8	31.2|99.2	30.8|99.2	28.4|99.2	26.0|99.2	25.2|99.2	24.8|99.2	24.8|99.2	24.8|99.2	24.0|99.2	22.8|99.2	22.4|99.2
2.1	35.6|35.6	60.0|84.4	62.8|93.2	55.6|96.8	49.2|97.6	45.2|98.0	42.0|98.0	38.0|98.4	35.2|98.4	32.0|98.4	31.2|98.8	31.2|99.2	29.6|99.2	27.6|99.2	26.0|99.2	25.2|99.2	25.2|99.2	24.8|99.2	24.8|99.2	23.6|99.2	22.8|99.2
2.2	36.4|36.4	60.0|84.4	61.2|93.2	55.6|96.8	51.6|97.6	46.8|98.0	42.0|98.0	39.2|98.4	36.4|98.4	32.8|98.4	31.2|98.8	31.2|99.2	30.4|99.2	28.0|99.2	26.4|99.2	25.2|99.2	24.8|99.2	24.8|99.2	24.4|99.2	24.8|99.2	23.6|99.2
2.3	36.0|36.0	58.8|82.8	62.4|92.8	57.2|96.8	51.2|97.6	46.8|98.0	43.6|98.0	40.0|98.4	37.6|98.4	33.6|98.4	31.6|98.8	31.2|99.2	31.2|99.2	29.6|99.2	27.2|99.2	26.0|99.2	25.2|99.2	24.8|99.2	24.8|99.2	24.4|99.2	24.8|99.2
2.4	37.6|37.6	60.0|81.6	62.4|92.4	57.6|96.4	52.4|97.6	46.8|98.0	45.2|98.0	42.0|98.0	38.0|98.4	35.2|98.4	32.8|98.4	31.2|98.8	31.2|99.2	29.6|99.2	28.4|99.2	27.2|99.2	26.0|99.2	24.8|99.2	24.8|99.2	24.4|99.2	24.4|99.2
2.5	38.0|38.0	58.4|80.4	62.8|92.0	57.2|96.0	54.0|97.6	48.0|98.0	45.6|98.0	42.8|98.0	40.0|98.4	36.8|98.4	33.2|98.4	31.6|98.8	31.2|99.2	30.8|99.2	29.6|99.2	28.0|99.2	26.4|99.2	24.8|99.2	24.8|99.2	24.8|99.2	24.4|99.2
2.6	35.6|35.6	60.0|81.6	62.4|92.0	59.2|95.2	54.4|97.6	49.2|97.6	46.4|98.0	42.8|98.0	40.8|98.4	37.6|98.4	35.2|98.4	32.8|98.4	31.2|98.8	31.2|99.2	30.4|99.2	29.2|99.2	26.0|99.2	25.6|99.2	24.8|99.2	24.8|99.2	24.8|99.2
2.7	36.4|36.4	60.0|80.0	61.2|92.0	60.0|95.2	54.8|97.6	50.0|97.6	47.2|98.0	43.2|98.0	42.0|98.4	38.4|98.4	36.0|98.4	33.2|98.4	31.2|98.8	30.8|99.2	31.2|99.2	30.0|99.2	28.0|99.2	26.0|99.2	25.6|99.2	24.8|99.2	24.8|99.2
2.8	36.0|36.0	58.4|78.0	61.2|91.6	**62.4**|**95.2**	55.2|97.2	50.4|97.6	47.6|98.0	45.2|98.0	42.8|98.0	39.6|98.4	36.8|98.4	34.0|98.4	32.4|98.4	30.8|98.8	31.2|99.2	30.8|99.2	29.6|99.2	26.8|99.2	26.0|99.2	24.8|99.2	24.8|99.2
2.9	35.6|35.6	58.8|78.0	61.6|91.2	62.8|94.8	55.6|96.8	52.0|97.6	47.6|98.0	45.2|98.0	42.8|98.0	41.2|98.4	37.6|98.4	35.2|98.4	33.2|98.4	31.2|98.8	30.8|99.2	31.2|99.2	30.0|99.2	28.8|99.2	26.8|99.2	25.6|99.2	24.8|99.2
3.0	34.8|34.8	57.2|77.2	61.6|91.2	62.0|94.4	57.6|96.8	52.4|97.6	48.4|98.0	46.8|98.0	43.2|98.0	42.0|98.4	38.0|98.4	36.8|98.4	32.8|98.4	31.6|98.8	30.8|98.8	31.2|99.2	30.8|99.2	30.0|99.2	28.0|99.2	26.0|99.2	25.6|99.2

Tabulka výsledných hodnot rozpoznávacího skóre pro audio-vizuální rozpoznávání řeči, kde jako vizuální příznaky byly použity geometrické (tvarové) příznaky. Rozpoznávací skóre bylo zjišťováno v závislosti na různých vahách pro akustickou část (γ_1) a vizuální část (γ_2). Výsledky jsou zapsány ve tvaru R1|R2 kde R1 je rozpoznávací skóre [%], kde v akustické části jsou audio nahrávky zatížené šumem o průměrném SNR 5 dB, R2 je rozpoznávací skóre [%] při použití originálních akustických nahrávek s průměrným SNR 18 dB.

Příloha č.7: AUDIO-VIZUÁLNÍ ROZPOZNÁVÁNÍ ŘEČI – DCT VIZUÁLNÍ PŘÍZNAKY

γ_2 viz.	γ_1 audio																						
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0.0	7.6|7.6	8.0|99.2	14.0|99.2	14.4|99.2	14.0|99.2	12.8|99.2	13.2|99.2	13.2|99.2	13.2|99.2	12.8|99.2	13.2|99.2	13.2|99.2	12.8|99.2	12.8|99.2	13.2|99.2	13.2|99.2	13.2|99.2	13.2|99.2	13.2|99.2	13.2|99.2	13.2|99.2	13.2|99.2
0.1	34.0|34.0	22.0|98.8	21.6|99.2	21.2|99.2	17.2|99.2	18.0|99.2	18.0|99.2	17.2|99.2	16.4|99.2	16.8|99.2	16.8|99.2	16.4|99.2	15.6|99.2	15.2|99.2	15.6|99.2	14.8|99.2	15.2|99.2	15.2|99.2	15.2|99.2	14.8|99.2	15.2|99.2
0.2	34.4|34.4	44.8|98.0	31.6|98.8	26.4|98.8	24.0|99.2	22.0|99.2	20.8|99.2	18.8|99.2	17.6|99.2	18.4|99.2	18.0|99.2	18.0|99.2	18.0|99.2	18.0|99.2	17.6|99.2	16.8|99.2	16.4|99.2	17.2|99.2	17.2|99.2	17.2|99.2	17.2|99.2
0.3	37.2|37.2	50.8|97.6	39.6|98.8	30.8|98.8	26.0|98.8	25.2|98.8	24.4|99.2	22.8|99.2	21.6|99.2	20.0|99.2	18.8|99.2	18.4|99.2	18.0|99.2	17.6|99.2	18.4|99.2	18.4|99.2	18.4|99.2	18.4|99.2	18.4|99.2	17.6|99.2	18.0|99.2
0.4	37.6|37.6	52.8|97.2	42.4|98.0	39.2|98.8	32.0|98.8	26.4|98.8	26.0|98.8	24.8|98.8	24.0|99.2	22.4|99.2	21.6|99.2	20.8|99.2	20.0|99.2	18.8|99.2	18.8|99.2	18.4|99.2	18.0|99.2	18.4|99.2	18.0|99.2	18.4|99.2	18.4|99.2
0.5	37.6|37.6	58.4|96.0	46.8|98.0	40.4|98.8	37.6|98.8	32.0|98.8	27.6|98.8	25.6|98.8	26.4|98.8	24.4|98.8	23.2|99.2	22.8|99.2	22.8|99.2	22.4|99.2	20.0|99.2	17.6|99.2	19.2|99.2	18.4|99.2	18.4|99.2	18.4|99.2	18.0|99.2
0.6	38.4|38.4	59.6|94.8	49.2|97.6	42.0|98.0	40.0|98.8	36.0|98.8	30.0|98.8	27.6|98.8	26.0|98.8	26.4|98.8	25.6|98.8	24.0|98.8	23.2|99.2	23.2|99.2	22.8|99.2	22.4|99.2	20.8|99.2	20.0|99.2	18.0|99.2	19.6|99.2	18.8|99.2
0.7	38.8|38.8	60.8|92.8	53.2|97.6	45.6|98.0	41.2|98.8	40.0|98.8	36.4|98.8	30.0|98.8	27.6|98.8	26.8|98.8	26.4|98.8	26.0|98.8	25.6|98.8	24.0|98.8	23.2|99.2	23.2|99.2	22.0|99.2	22.0|99.2	22.0|99.2	20.8|99.2	18.8|99.2
0.8	39.6|39.6	60.0|90.4	56.4|96.8	49.2|98.0	43.6|98.0	40.4|98.8	39.2|98.8	34.8|98.8	30.0|98.8	28.0|98.8	27.6|98.8	26.8|98.8	26.8|98.8	25.6|98.8	25.6|98.8	24.4|98.8	23.6|99.2	23.2|99.2	22.4|99.2	22.8|99.2	22.8|99.2
0.9	38.4|38.4	62.0|89.6	56.0|96.8	50.0|97.6	46.8|98.0	42.0|98.0	40.0|98.8	38.0|98.8	34.4|98.8	30.0|98.8	28.0|98.8	27.6|98.8	26.0|98.8	26.0|98.8	26.4|98.8	25.6|98.8	24.4|98.8	24.8|98.8	23.6|99.2	23.2|99.2	22.8|99.2
1.0	38.0|38.0	62.8|88.0	57.2|96.0	49.2|97.6	47.2|98.0	42.4|98.0	41.2|98.8	39.6|98.8	37.2|98.8	33.2|98.8	30.4|98.8	28.8|98.8	27.6|98.8	26.0|98.8	26.4|98.8	26.4|98.8	26.0|98.8	25.6|98.8	25.2|98.8	24.8|98.8	23.6|99.2
1.1	38.8|38.8	62.4|87.6	58.0|95.2	56.0|97.6	46.0|98.0	45.6|98.0	41.6|98.0	40.4|98.8	39.6|98.8	36.8|98.8	33.6|98.8	30.4|98.8	28.8|98.8	27.6|98.8	26.8|98.8	26.4|98.8	26.0|98.8	26.0|98.8	25.2|98.8	25.6|98.8	24.8|98.8
1.2	38.8|38.8	61.6|85.6	56.8|94.8	55.6|96.8	50.0|97.6	46.0|98.0	42.4|98.0	41.2|98.8	39.6|98.8	38.4|98.8	36.4|98.8	33.2|98.8	30.4|98.8	28.8|98.8	27.6|98.8	27.2|98.8	26.0|98.8	26.4|98.8	26.4|98.8	26.0|98.8	25.6|98.8
1.3	38.8|38.8	**64.0|83.6**	59.2|94.0	56.0|96.8	50.8|97.6	46.8|98.0	44.4|98.0	41.6|98.0	40.4|98.8	40.4|98.8	37.6|98.8	36.0|98.8	32.8|98.8	30.4|98.8	28.0|98.8	28.0|98.8	27.2|98.8	26.0|98.8	26.4|98.8	26.0|98.8	25.6|98.8
1.4	37.2|37.2	62.4|83.6	60.4|92.8	56.8|96.0	51.2|97.6	46.4|98.0	45.2|98.0	42.4|98.0	41.2|98.8	40.8|98.8	40.4|98.8	36.8|98.8	35.6|98.8	32.8|98.8	30.4|98.8	28.4|98.8	28.0|98.8	27.6|98.8	27.2|98.8	26.0|98.8	26.0|98.8
1.5	36.8|36.8	64.0|80.0	60.4|92.4	57.6|96.0	54.0|96.8	50.0|97.6	47.2|98.0	44.0|98.0	40.8|98.0	41.2|98.8	40.8|98.8	39.2|98.8	37.2|98.8	34.8|98.8	32.8|98.8	30.8|98.8	28.8|98.8	28.4|98.8	27.6|98.8	27.2|98.8	26.0|98.8
1.6	37.6|37.6	62.4|79.2	59.6|90.4	58.0|95.2	55.6|96.8	50.0|97.6	47.6|98.0	46.4|98.0	42.8|98.0	41.6|98.8	40.8|98.8	39.2|98.8	38.8|98.8	37.2|98.8	34.4|98.8	32.8|98.8	30.8|98.8	28.0|98.8	28.0|98.8	27.6|98.8	27.6|98.8
1.7	38.4|38.4	61.6|76.4	62.0|89.6	57.2|94.4	55.6|96.8	51.2|97.6	47.2|98.0	44.8|98.0	44.8|98.0	40.8|98.0	40.4|98.8	40.8|98.8	40.4|98.8	38.8|98.8	36.8|98.8	34.0|98.8	32.0|98.8	30.8|98.8	28.0|98.8	28.0|98.8	28.0|98.8
1.8	37.2|37.2	59.2|75.2	63.6|90.0	58.0|94.8	56.8|96.8	54.0|97.6	51.2|97.6	46.4|98.0	46.8|98.0	42.8|98.0	40.8|98.4	40.8|98.8	40.4|98.8	40.4|98.8	38.4|98.8	36.4|98.8	34.4|98.8	32.0|98.8	30.8|98.8	29.2|98.8	28.4|98.8
1.9	38.0|38.0	58.8|73.2	60.8|89.6	58.4|93.6	58.8|96.0	57.2|97.2	50.0|97.6	46.4|98.0	46.0|98.0	43.2|98.0	41.2|98.0	41.6|98.8	40.4|98.8	39.2|98.8	38.8|98.8	38.0|98.8	36.0|98.8	34.8|98.8	32.0|98.8	30.8|98.8	28.8|98.8
2.0	37.6|37.6	57.6|73.2	61.6|88.0	60.4|93.6	57.6|96.0	56.4|96.8	50.0|97.6	47.6|98.0	47.2|98.0	46.0|98.0	43.2|98.0	41.2|98.4	40.8|98.8	40.8|98.8	39.6|98.8	38.8|98.8	36.8|98.8	36.0|98.8	34.8|98.8	32.0|98.8	30.8|98.8
2.1	37.6|37.6	56.4|72.4	62.8|88.0	60.8|92.4	57.6|94.8	56.0|96.8	52.8|97.6	51.2|97.6	47.6|98.0	46.0|98.0	43.2|98.0	41.2|98.0	41.6|98.8	40.8|98.8	40.4|98.8	40.4|98.8	38.4|98.8	36.8|98.8	36.0|98.8	34.4|98.8	31.2|98.8
2.2	37.6|37.6	54.8|70.0	63.6|86.4	61.6|92.4	59.6|94.8	56.0|96.8	53.6|97.2	51.2|97.6	47.6|98.0	44.8|98.0	45.6|98.0	43.2|98.0	41.2|98.0	40.8|98.8	40.0|98.8	39.2|98.8	39.6|98.8	38.0|98.8	36.4|98.8	36.0|98.8	34.4|98.8
2.3	37.2|37.2	54.8|70.8	63.6|85.2	61.2|91.6	58.0|94.8	57.6|96.0	**57.6|97.6**	50.4|97.6	47.2|98.0	47.6|98.0	47.2|98.0	42.8|98.0	41.2|98.0	41.2|98.4	40.4|98.8	41.2|98.8	39.6|98.8	38.8|98.8	37.6|98.8	36.0|98.8	36.0|98.8
2.4	38.0|38.0	55.6|69.2	63.6|84.8	59.2|90.0	58.0|94.8	58.4|96.0	57.6|96.8	51.6|97.6	50.0|97.6	47.6|98.0	45.2|98.0	44.8|98.0	43.2|98.0	41.2|98.0	41.2|98.8	40.4|98.8	40.4|98.8	39.2|98.8	38.4|98.8	37.2|98.8	35.6|98.8
2.5	38.0|38.0	55.2|70.0	63.6|84.0	60.4|90.4	59.2|93.6	57.2|96.0	57.2|96.8	54.4|97.6	50.8|97.6	46.4|98.0	46.8|98.0	47.2|98.0	42.8|98.0	42.0|98.0	40.8|98.4	41.2|98.8	40.8|98.8	39.2|98.8	40.0|98.8	38.0|98.8	36.8|98.8
2.6	37.6|37.6	55.2|67.2	62.4|83.6	62.0|90.0	59.6|93.6	58.4|95.2	56.8|96.8	54.8|97.2	49.2|97.6	47.6|98.0	47.2|98.0	46.0|98.0	46.0|98.0	42.8|98.0	41.6|98.0	41.6|98.4	41.2|98.8	41.2|98.8	40.0|98.8	39.2|98.8	38.0|98.8
2.7	36.8|36.8	55.2|67.6	62.0|83.6	62.0|89.2	59.2|93.2	58.4|94.8	57.2|96.8	57.6|97.2	51.2|97.6	50.8|97.6	46.4|98.0	44.8|98.0	46.8|98.0	43.2|98.0	42.0|98.0	40.8|98.4	40.8|98.8	40.0|98.8	40.4|98.8	40.0|98.8	38.8|98.8
2.8	36.4|36.4	53.6|67.2	61.2|81.6	60.0|88.8	59.6|92.8	58.4|94.8	56.8|96.0	57.2|96.8	53.6|97.6	50.0|97.6	46.8|98.0	47.2|98.0	46.8|98.0	44.8|98.0	43.2|98.0	40.8|98.0	41.6|98.4	41.2|98.8	41.2|98.8	39.2|98.8	40.4|98.8
2.9	36.8|36.8	53.6|66.4	61.2|80.8	60.0|88.4	61.2|92.8	59.2|94.4	58.4|96.0	56.4|96.8	52.0|97.6	49.2|97.6	47.6|98.0	48.0|98.0	44.4|98.0	46.0|98.0	42.8|98.0	42.0|98.0	41.2|98.4	40.8|98.8	40.8|98.8	40.8|98.8	40.0|98.8
3.0	36.8|36.8	52.0|64.4	60.8|80.4	60.8|88.4	60.8|91.6	58.0|94.8	57.6|96.0	56.4|96.8	55.6|96.8	50.8|97.6	50.0|97.6	46.4|98.0	46.8|98.0	46.8|98.0	43.6|98.0	43.2|98.0	40.8|98.0	41.6|98.4	41.2|98.8	40.0|98.8	40.4|98.8

Tabulka výsledných hodnot rozpoznávacího skóre pro audio-vizuální rozpoznávání řeči, kde jako vizuální příznaky byly použity DCT příznaky. Rozpoznávací skóre bylo zjišťováno v závislosti na různých vahách pro akustickou část (γ_1) a vizuální část (γ_2). Výsledky jsou zapsány ve tvaru R1|R2 kde R1 je rozpoznávací skóre [%], kde v akustické části jsou audio nahrávky zatížené šumem o průměrném SNR 5 dB, R2 je rozpoznávací skóre [%] při použití originálních akustických nahrávek s průměrným SNR 18 dB.