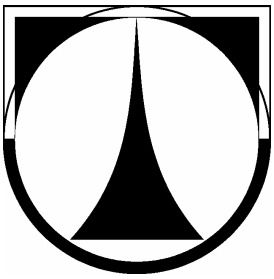
TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta mechatroniky a mezioborových inženýrských studií



ROZPOZNÁVÁNÍ AKUSTICKÉHO SIGNÁLU ŘEČI S PODPOROU VIZUÁLNÍ INFORMACE

AUTOREFERÁT DISERTAČNÍ PRÁCE

2005 **JOSEF CHALOUPKA**

Rozpoznávání akustického signálu řeči s podporou vizuální informace

Autoreferát disertační práce

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 Hálkova 6, 46 117, Liberec
Školitel:	Prof. Ing. Jan Nouza, CSc.
Oponenti:	Prof. Ing. Josef Psutka, CSc. (FAV ZČU Plzeň) Doc. Ing. Petr Pollák, CSc. (FEL ČVUT Praha) Doc. Ing. Jozef Juhár, PhD. (FEI TU Košice)

S disertační prací je možné se seznámit na děkanátu Fakulty mechatroniky a mezioborových inženýrských studií Technické univerzity v Liberci, Čížkova ulice č. 3, budova B, tel.: 485 535 3110.

Rozsah vlastní disertační práce a příloh:

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

ROZPOZNÁVÁNÍ AKUSTICKÉHO SIGNÁLU ŘEČI
S PODPOROU VIZUÁLNÍ INFORMACE

Obsah

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 disertační práci 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 [1], rozpoznávání spojité řeči [2] a on-line rozpoznávání slov z malého slovníku [3]. 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 [4], němčinu [5], japonštinu [6] 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 jednotlivých kapitol.

1.2 Cíle disertační práce

Cíle této disertační práce byly 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.

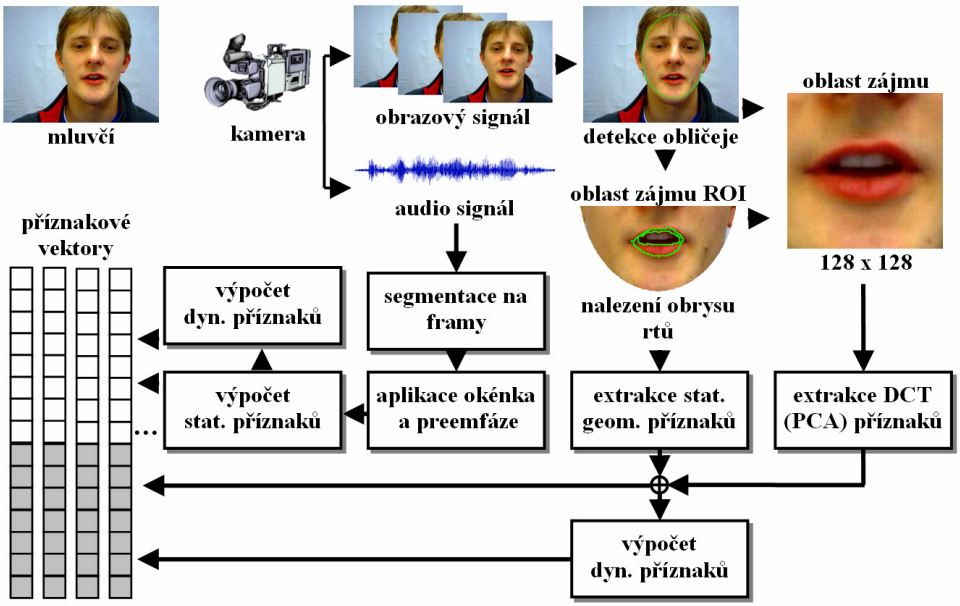
- Navrhnout vhodnou fúzi parametrizovaného akustického a vizuálního signálu řeči a provést experimentální otestování audio-vizuálního rozpoznávání izolovaných slov, při srovnání se samostatným rozpoznáváním akustického a vizuálního signálu řeči. Navrhnout, realizovat a experimentálně vyhodnotit test pro audio-vizuální rozpoznávání izolovaných slov v podmínkách proměnného hlučného pozadí.

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 signálu 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.



Obr. 2.1: *Princip parametrizace audio-vizuálního signálu*

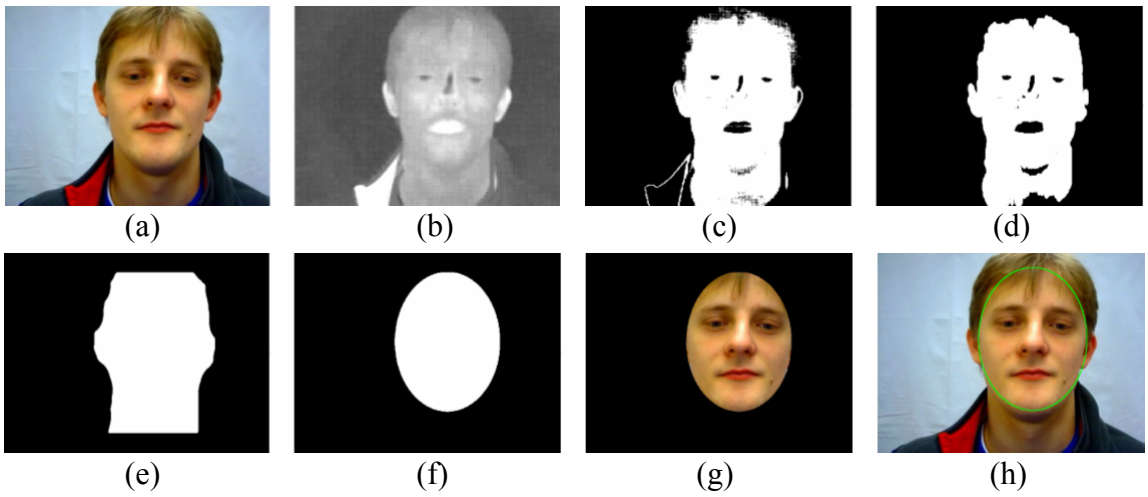
Vizuální signál je složen z časového sledu obrazů (2D signálů), ve kterých jsou zaznamenáni obličeje mluvčích. Pro parametrizaci vizuálního signálu se nejčastěji používají tvarové (geometrické) vizuální příznaky získané z nalezeného tvaru rtů nebo vizuální příznaky (DCT, PCA, ...) popisující informační obsah oblasti zájmu, ve které se nacházejí rty a jejich bezprostřední okolí. 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 a z detekovaného obrazu obličeje je separována oblast zájmu se rty, 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 s využitím programu HTK [7]. Program HTK mimo jiné umožňuje vytvářet (natrénovat) HM modely a rozpoznávat parametrizované signály za využití těchto HM modelů. Při použití HTK jsou možné dva způsoby rozpoznávání parametrizovaných audio-vizuálních signálů pomocí HM modelů. HM modely jsou vytvořeny (natrénovány) jako jednostreamové nebo dvoustreamové. U jednostreamových HM modelů jsou sloučeny akustické a vizuální příznaky do jednoho příznakového vektoru, pro nějž je poté vytvořen HM model.

3 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 obličeje poté vytvořit oblast zájmu (ROI-Region Of Interest), 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. Nakonec byl navržen vlastní algoritmus založený na principu barevné a tvarové segmentace obrazu [8]. Pro tvarovou segmentaci byly použity metody matematické morfologie, viz obr. 3.1.



Obr. 3.1: *Originální RGB barevný obraz (a), obraz transformovaný do Cr-barevné složky (b), výsledný binární obraz po segmentaci prahováním (c), binární obraz po použití morfologické operace otevření (d), po uzavření (e), po upravené operaci otevření (f), výsledný obraz , ve kterém je již vybrána pouze oblast s obličejem (g) a původní barevný obraz s vyznačenou detekovanou oblastí (h). U každé morfologické operace byl použit jiný strukturní element, pro (f) byl tento element vybrán automaticky.*

4 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 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 [9]. Tato kapitola pojednává o metodách pro nalezení oblasti rtů v pořízené oblasti zájmu ROI.

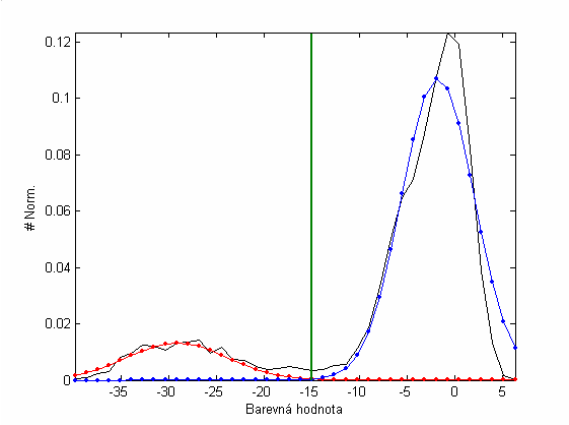
Tak jako u detekování obličeje v obraze existuje v současné době (r. 2005) velké množství metod pro nalezení oblasti rtů. Nejčastěji se používají metody pro segmentaci obrazu prahováním [10, 11] a segmentaci obrazu za pomocí různých barevných statistických modelů [12]. Před vlastní segmentací obrazu 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. Byly zkoumány výhody a nevýhody jednotlivých barevných prostorů a nakonec byl vybrán F - barevný prostor, který byl vytvořen pomocí Fisherovy lineární diskriminační analýzy [13]. Pro nalezení objektu rtů při segmentaci obrazu oblasti zájmu bylo potřeba přesněji nalézt práh. Na základě ověřeného 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 [14]. V tomto algoritmu byla použita gradientní metoda využívající metodu nejmenších čtverců, pomocí které byly odhadnuty parametry μ_r, μ_k, σ_r, σ_k a P_r, kde μ_r je střední hodnota a σ_r je směrodatná odchylka z normálního rozdělení $N(\mu_r, \sigma_r^2)$ popisujícího rozdělení barevných hodnot obrazových bodů zobrazujících rty v obraze oblasti zájmu a μ_k je střední hodnota a σ_k je směrodatná odchylka z normálního rozdělení $N(\mu_k, \sigma_k^2)$ popisujícího rozdělení barevných hodnot obrazových bodů zobrazujících okolí (kůži) v obraze oblasti zájmu. P_r je relativní počet obrazových bodů zobrazujících rty. Relativní počet obrazových bodů zobrazujících okolí (kůži) $P_k = 1 - P_r$. Z těchto parametrů byl poté počítán práh T pro segmentaci dle následujících vztahů:

$$T = \frac{-b + \sqrt{b^2 - 4.a.c}}{2.a} \tag{4.1}$$

$$a = \sigma_k^2 - \sigma_r^2 \tag{4.2}$$

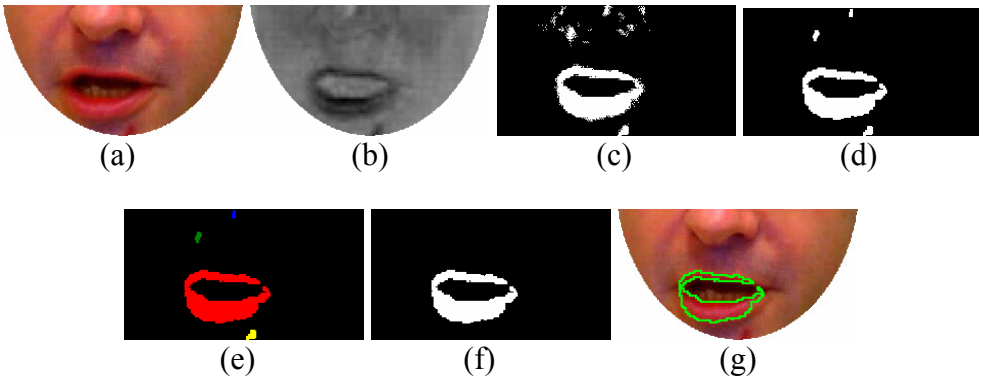
$$b = 2.\left(\sigma_r^2 \mu_k - \sigma_k^2 \mu_r\right) \tag{4.3}$$

$$c = \mu_r^2 \sigma_k^2 - \mu_k^2 \sigma_r^2 + 2.\sigma_k^2 \sigma_r^2 .\log\left(\frac{\sigma_r (1 - P_r)}{\sigma_k P_r}\right) \tag{4.4}$$



Obr. 4.1: *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 μ_k, σ_k a P_r (modrá křivka) a výsledný vypočtený práh (zobrazen zeleně)*

Po segmentaci obrazu oblasti zájmu je výsledný binární obraz filtrován pomocí morfologické operace otevření a následně byla použita filtrace drobných objektů (jizvy, akné, vyrážka ...) pomocí metody barvení objektů, viz obr. 4.2.



Obr. 4.2: *Originální RGB barevný obraz oblasti zájmu (a), obraz transformovaný do F-barevné složky (b), výsledný binární obraz po segmentaci prahováním (c), binární obraz po filtraci – otevření strukturním elementem (d), obraz s obarvenými oblastmi (e), výsledný binární obraz po odstranění oblastí, které s jistou pravděpodobností netvoří rty (f) a původní barevný obraz s nalezenými okraji rtů (g)*

5. 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 [15, 16]. 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 a vizuální příznaky popisující informační obsah obrazu.

Jako tvarové příznaky byly v této práci vybrány horizontální *h* (5.1) a vertikální *v* (5.2) rozšíření rtů, dále oblast rtů *o* (5.3) a zaokrouhlení rtů *r* (5.4), které nabývá největších hodnot při vyslovování fonémů u, o, ř.

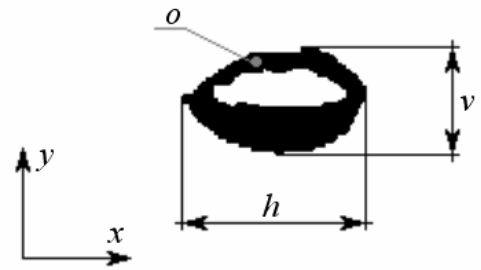
$$h = \max_{y=0..N-1} \sum_{x=0}^{M-1} f(x,y) \tag{5.1}$$

$$v = \max_{x=0..M-1} \sum_{y=0}^{N-1} f(x,y) \tag{5.2}$$

$$o = \sum_{y=0}^{N-1} \sum_{x=0}^{M-1} f(x,y) \tag{5.3}$$

$$r = v / h \tag{5.4}$$

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.



Obr. 5.1: *Binární obraz oblasti rtů s vyznačenými geometrickými příznaky*

V této práci byly použity i vizuální příznaky popisující informační obsah obrazu, jednalo se o DCT energetické příznaky [17]. Ke všem statickým vizuálním příznakům byly počítány příslušné dynamické příznaky.

6 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 [18], tj. v databázi by se měly nacházet kvalitní nahrávky od velkého množství mluvčích a pokrývající co nejvíce jazyk, pro který je databáze vytvářena.

V současné době (2005) existuje několik audio-vizuálních databází pro různé národní jazyky, především však pro angličtinu např. XM2VTS [19] nebo AVICAR [20]. 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 [21]. Postupem času byly vytvořené dvě audio-vizuální databáze později označené AVDB1cz a AVDB2cz. V první videodatabázi AVDB1cz bylo zaznamenáno 52 mluvčích dvěma kamerami (snímání ze předu a z profilu) při rozlišení 320 x 240 obrazových bodů a při snímkovací frekvenci 30 snímků za sekundu. 35 mluvčích bylo zaznamenáno v databázi AVDB2cz, kde byla použita pouze jedna kamera (čelní snímání) při rozlišení 640 x 480 obrazových bodů a při snímkovací frekvenci 30 snímků za sekundu. Každý mluvčí namluvil 50 slov a 50 foneticky bohatých vět. Každý český foném se v souboru vět nacházel alespoň 6x.

7 Experimentální práce – testy

V následujících statích jsou popsány testy a jejich výsledky při rozpoznávání izolovaných slov z audio-vizuální databáze AVDB2cz. Pro rozpoznávání byl použit klasifikátor založený na technice skrytých Markovových modelů, pro vlastní natrénování a rozpoznávání celoslovních modelů byl využit program HTK [7]. 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).

7.1 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. Před vlastním vytvořením těchto modelů bylo potřeba zjistit jaký je minimální počet framů (o délce 33.3 ms) tvořících jednotlivá slova z testovací databáze. Minimální počet framů byl v našich nahrávkách 14. Z tohoto minimálního počtu framů byl stanoven maximální počet stavů HM modelů. V originálních videonahrávkách (v akustickém streamu) z testovací části databáze byla průměrná hodnota SNR 18 dB.

Pro účely následných testů rozpoznávání slov z akustického signálu byl vytvořen algoritmus, pomocí něhož bylo možné přidávat k akustickým nahrávkám aditivní bílý

šum. Tento algoritmus byl použit v testu, kde byla zkoumána závislost celkového rozpoznávacího skóre na šumu obsaženém v nahrávkách.

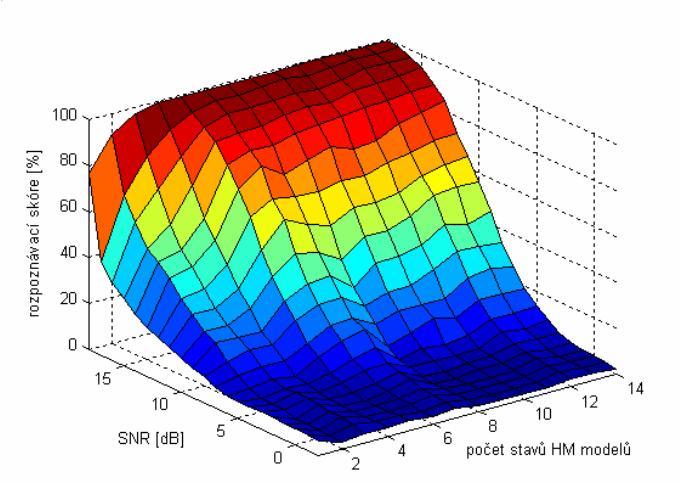
V tomto testu byla zjišťována závislost velikosti výsledného rozpoznávacího skóre na počtu stavů použitých (natrénovaných) celoslovních HM modelů při měnícím se SNR v akustickém signálu. Specifikace testu:

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



Obr. 7.1: *Graf hodnot z tab. 7.1 – výsledné rozpoznávací skóre [%] v závislosti na počtu stavů HM modelů při měnícím se SNR v akustickém signálu*

Z tabulky 7.1 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.

7.2 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 při automatickém rozpoznávání vizuálního signálu řeči.

7.2.1 Rozpoznávání vizuálního signálu s tvarovými příznaky

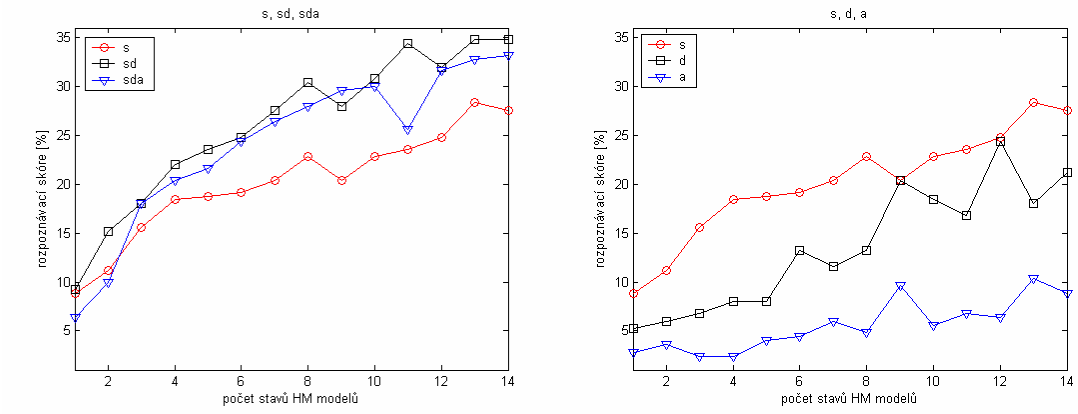
Jako tvarové příznaky byly v tomto testu použity vizuální příznaky horizontální (*h*) a vertikální (*v*) rozšíření rtů, dále oblast rtů (*o*) a zaokrouhlení rtů (*r*). Z těchto příznaků byly vypočteny dynamické a akcelerační příznaky, které jsou také při rozpoznávání akustického signálu řeči uplatněny. 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)

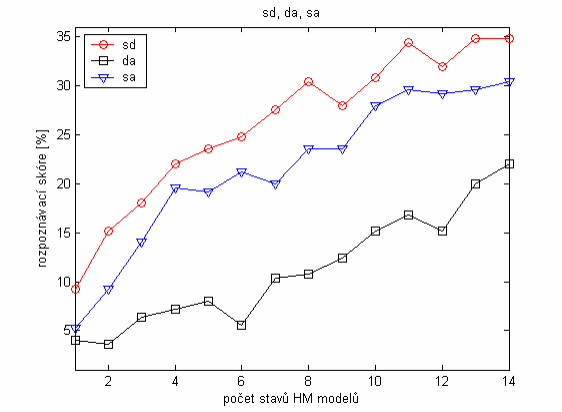
Viz. pří.	Počet stavů HM modelů													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
s	8,8	11,2	15,6	18,4	18,8	19,2	20,4	22,8	20,4	22,8	23,6	24,8	**28,4**	27,6
d	5,2	6,0	6,8	8,0	8,0	13,2	11,6	13,2	20,4	18,4	16,8	**24,4**	18,0	21,2
a	2,8	3,6	2,4	2,4	4,0	4,4	6,0	4,8	9,6	5,6	6,8	6,4	**10,4**	8,8
s+d	9,2	15,2	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. 7.2: *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. 7.2.



Obr. 7.2: *Grafy hodnot z tab. 7.2 – výsledné rozpoznávací skóre [%] v závislosti na počtu stavů HM modelů při využití různých kombinací statických (s), dynamických (d) a akceleračních (a) vizuálních tvarových příznaků*



7.2.2 Rozpoznávání vizuálního signálu s DCT příznaky

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

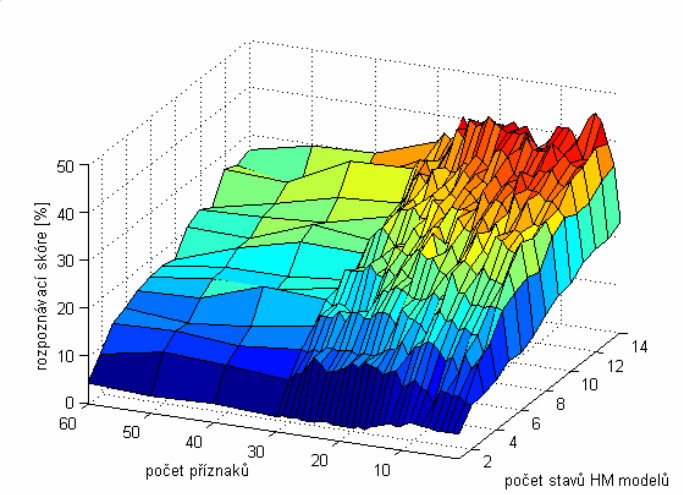
Počet vizuálních příznaků byl zvětšován o hodnotu 1 v rozsahu 1 až 30 příznaků a dále o hodnotu 10 v intervalu 30 až 60 příznaků. Výsledné tabulky z tohoto testu pro kombinace statických (s), dynamických (d) a akceleračních DCT-energetických vizuálních příznaků jsou dosti rozsáhlé, proto byly umístěny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

V následující tabulce 7.3 jsou uvedeny nejlepší výsledky rozpoznávacího skóre pro jednotlivé kombinace (s, d, a) energetických DCT vizuálních příznaků. 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. 7.3: 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 7.3 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 rozpoznávacího skóre při kombinaci statických, dynamických a akceleračních příznaků.



Obr. 7.3: *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 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*, rozptylu *R* nebo normovaného rozptylu *NR* 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. 7.4 uvedeny hodnoty rozpoznávacího skóre pro různé varianty testů, kde byla použita kombinace pěti statických a příslušných 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. 7.4: *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.*

7.3 Úloha rozpoznávání audio-vizuálního signálu řeči

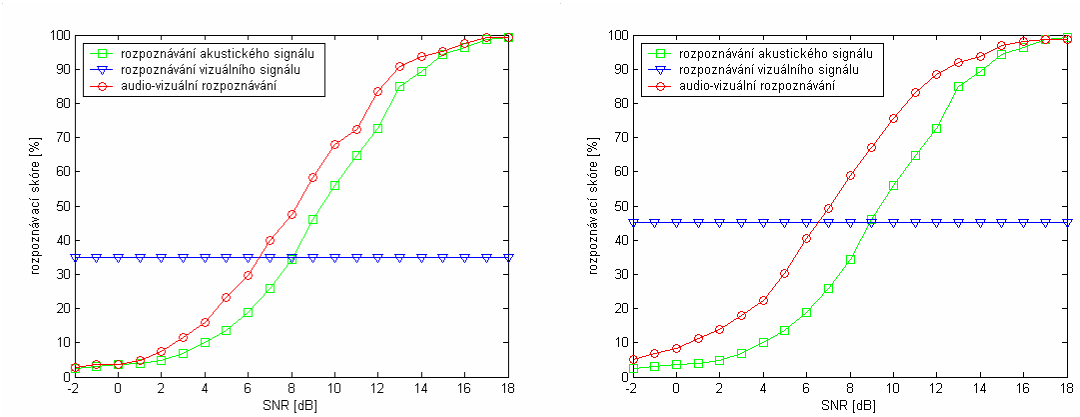
V této úloze byly provedeny testy pro vlastní audio-vizuální rozpoznávání řeči. Při vlastní fúzi akustických a vizuálních příznaků byly natrénovány čtrnáctistavové jednostreamové (dvoustreamové) celoslovní HM modely, které poté sloužily k vlastnímu rozpoznávání pomocí klasifikátoru založeném na technice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

7.3.1 Jednostreamové audio-vizuální rozpoznávání řeči

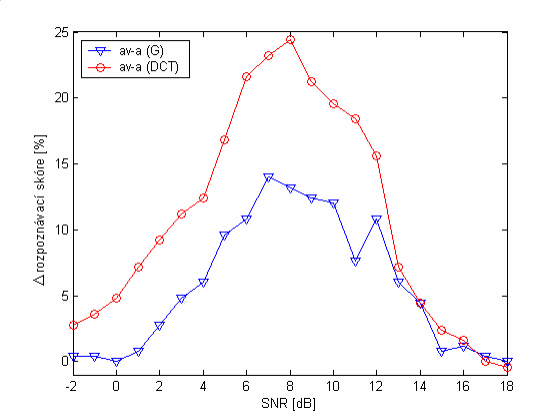
V tomto testu byly pro audio-vizuální rozpoznávání řeči použity jednostreamové celoslovní HM modely, pro jejichž natrénování byly sloučeny akustické a vizuální příznaky audio-vizuálního signálu řeči (slova) z trénovací databáze do jednoho příznakového vektoru. Sloučené audio-vizuální příznakové vektory 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, tak pro rozpoznávání vizuálního signálu řeči s geometrickými vizuálními příznaky i s DCT energetickými vizuálními příznaky. 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.

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. 7.4: *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. 7.5: *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. 7.5: *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).*

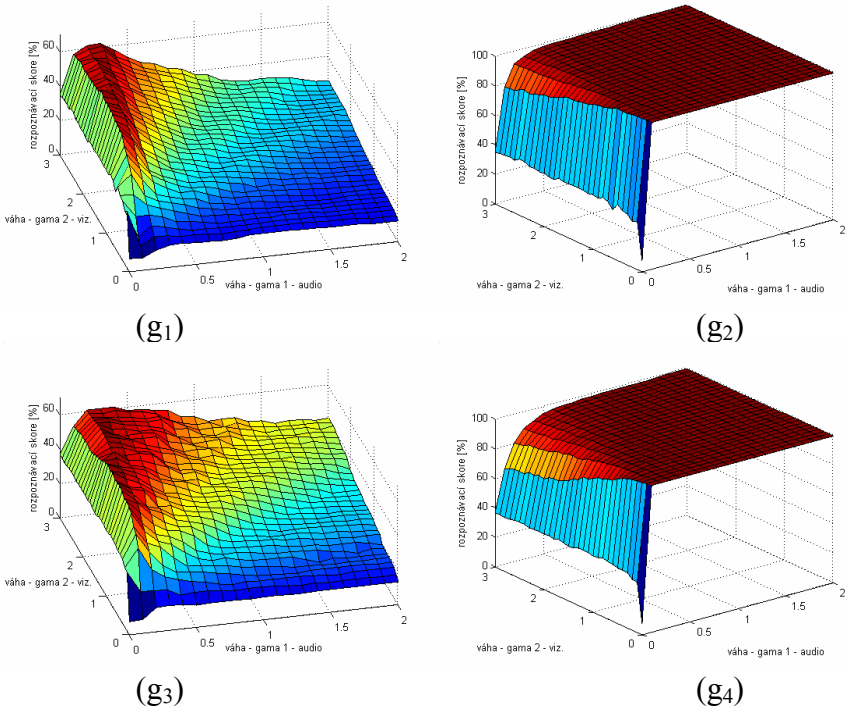
Z grafů na obr. 7.4-5 a z příslušné tabulky tab. 7.5 je patrné, že vizuální složka řeči při audio-vizuálním rozpoznávání zlepšuje rozpoznávací skóre oproti samostatnému rozpoznávání akustického signálu řeči. Nejvyššího zlepšení bylo dosaženo u SNR v akustickém signálu v rozmezí 6-9 dB. Při použití DCT energetických příznaků bylo u jednostreamového audio-vizuálního rozpoznávání dosaženo v průměru lepších výsledků oproti použití geometrických příznaků.

7.3.2 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.

7.3.2.1 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.



(g_1) (g_2)

(g_3) (g_4)

Obr. 7.6: *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.*

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.

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

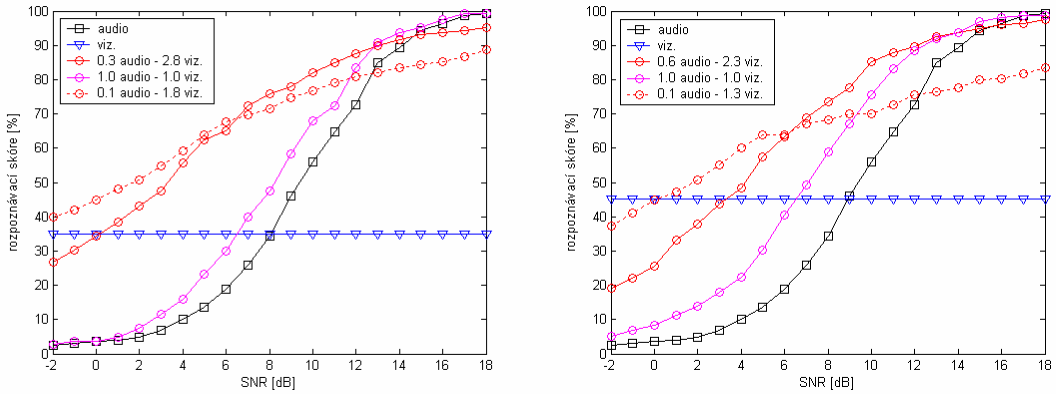
Tab. 7.6: *Vybrané hodnoty akustické (γ_1) a vizuální (γ_2) váhy pro dvoustreamové audio-vizuální rozpoznávání řeči, dle prvního a druhého kritéria*

7.3.2.2 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:

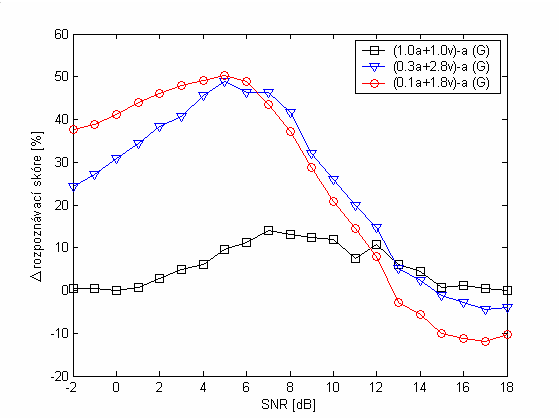
Klasifikátor založený na	dvoustreamové celoslovní levo-pravé HM modely
Počet příznaků	47 aku. + geometrické (54 aku. + DCT)
Druh příznaků (kombinace aku. a viz)	39 aku. – 13 x (MFCC + delta + akcelerační) 8 viz – *h*, *v*, *o*, *r* + delta (15 viz – 5 x (ene. DCT + delta + akcelerační))
Počet stavů HM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



Obr. 7.7: *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

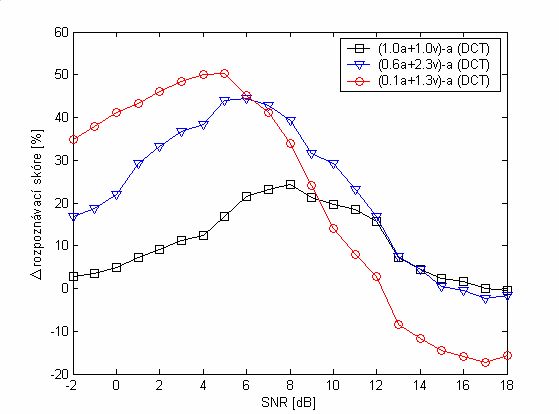
Tab. 7.7: *Rozpoznávací skóre [%] při měnícím se SNR v akustickém signálu, pro rozpoznávání akustického signálu řeči (audio), vizuálního signálu řeči (viz) nebo pro audio-vizuálním rozpoznávání (audio-viz), kde jako vizuální příznaky byly použity geometrické příznaky. Zároveň jsou zde uvedeny hodnoty 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.*



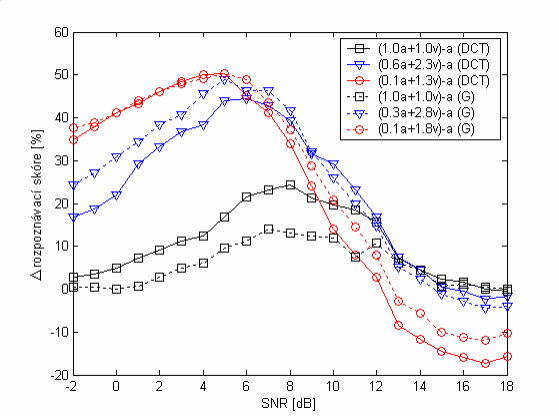
Obr. 7.8: *Rozdíl výsledného rozpoznávacího skóre mezi audio-vizuálním rozpoznáváním řeči (av) a rozpoznáváním řeči z akustického signálu (a), kde pro audio-vizuální rozpoznávání byly použity geometrické vizuální příznaky a různě zvolené váhy akustického a vizuálního streamu.*

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. 7.8: *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 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. 7.9: *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. 7.10: *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*

7.3.2.3 Celkové zhodnocení experimentů pro audio-vizuální rozpoznávání řeči

Při dvoustreamovém audio-vizuálním rozpoznávání izolovaných slov (s hodnotami vah streamu dle tab. 7.6) bylo u našich nahrávek z AV databáze AVDB2cz v průměru dosaženo lepších výsledků zvýšení rozpoznávacího skóre než u jednostreamového audio-vizuálního rozpoznávání ve srovnání se samostatným rozpoznáváním akustického signálu. 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. 7.6) bylo v průměru dosaženo lepších výsledných hodnot rozpoznávacího skóre pro geometrické vizuální příznaky oproti DCT energetickým vizuálním příznakům. U jednostreamového audio-vizuálního rozpoznávání bylo naopak dosaženo v průměru lepších výsledků při použití DCT energetických vizuálních příznaků oproti použití geometrických vizuálních příznaků.

8 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. Pro zpracování nahrávek z této databáze byl vytvořen komplexní program. Dále byl vytvořen program pro parametrizaci vizuálního signálu řeči skládající se z detektoru lidské tváře v obraze, systému pro nalezení rtů a vlastního systému pro parametrizaci vizuálního signálu, pomocí kterého jsou vytvářeny tvarové vizuální příznaky a vizuální DCT příznaky popisující informační obsah obrazu. 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. 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 vízuálního 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. 8.1: *Vybrané hodnoty rozpoznávacího skóre [%] z experimentálních testů pro rozpoznávání akustického signálu, 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).

8.1 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. Vytvoření programu pro zpracování nahrávek z této databáze.

- Navržení a vytvoření systému pro detekování lidského obličeje, založeného na barevné a tvarové segmentaci obrazu a systému pro nalezení rtů v detekované oblasti zájmu. 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 [14].

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

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

8.2 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. 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.

8.3 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.

9 Literatura

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

[2] 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

[3] 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

[4] 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

[5] 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

[6] 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

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

[8] 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, CZ, pp. 141-145, ISBN 80-86269-10-8

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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

[17] CHALOUPKA, J.:Extraction of the Visual Features by DCT for AVSR. *In Proc. of Radioelektronika 2005.* Brno, May 2005. pp. 467-470. ISBN 80-214-2904-6

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

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

[20] 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

[21] 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

Vlastní publikované práce:

- 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
- 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
- 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
- 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
- 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
- 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
- CHALOUPKA, J.: Multimodal Signal Processing and Research. *In Proc. of Radioelektronika 2003.* Brno, May 2003. pp. 388-389. ISBN 80-214-2383-8
- 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
- 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
- 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
- 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
- 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
- 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
- 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
- CHALOUPKA, J.: Extraction of the Visual Features by DCT for AVSR. *In Proc. of Radioelektronika 2005.* Brno, May 2005. pp. 467-470. ISBN 80-214-2904-6

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.

Anglická anotace z disertační práce. V tomto autoreferátu byly vybrány nejdůležitější části z kapitol 4 – 8, + úvod (kapitola 1) a závěr (kapitola 10).

Ing. Josef Chaloupka

ROZPOZNÁVÁNÍ AKUSTICKÉHO SIGNÁLU ŘEČI
S PODPOROU VIZUÁLNÍ INFORMACE
Autoreferát disertační práce

Technická univerzita v Liberci
Fakulta mechatroniky a mezioborových inženýrských studií

28 stran
Náklad 20 výtisků

2005