TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky a mezioborových studií

Studijní program: N2612 – Elektrotechnika a informatika
Studijní obor: 1802T007 – Informační technologie

Vliv použitých vizuálních příznaků řeči v úloze audio-vizuálního rozpoznávání řečového signálu

Influence of the visual feautures of speech in the audio-visual speech recognition

Diplomová práce

Autor: **Bc. Jan Matějka**
Vedoucí práce: Ing. Josef Chaloupka, Ph.D.
Konzultant: Ing. Jindřich Žďánský, Ph.D.

V Liberci 20. 5. 2010

TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky a mezioborových studií
Ústav informačních technologií a elektroniky
Akademický rok: **2009/2010**

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení: **Bc. Jan MATĚJKA**

Studijní program: **N2612 Elektrotechnika a informatika**

Studijní obor: **Informační technologie**

Název tématu: **Vliv použitých vizuálních příznaků řeči v úloze audio-vizuálního rozpoznávání řečového signálu**

Zásady pro vypracování:

1. Seznamte se s problematikou zpracování a rozpoznávání signálu řeči.
2. Seznamte se s vývojovým a programátorským prostředím Microsoft Visual C++ a programem HTK Toolkit pro trénování a rozpoznávání pomocí skrytých markovských modelů HMM..
3. Vytvořte systém pro audio-vizuální rozpoznávání řeči pomocí skrytých markovských modelů jednotlivých fonémů a vizémů.
4. Pro natrénování modelů vizémů použijte různé vizuální příznaky řeči. Vyhodnoťte vliv použitých vizuálních příznaků na výsledné rozpoznávací skóre.
5. Otestujte tento systém také na úloze audio-vizuálního rozpoznávání řeči v hlučných podmínkách.

Rozsah grafických prací:	**Dle potřeby dokumentace**
Rozsah pracovní zprávy:	**cca 40 - 50 stran**
Forma zpracování diplomové práce:	**tištěná/elektronická**

Seznam odborné literatury:

[1] 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

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

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

[4] HLAVÁČ V., SEDLÁČEK M.: Zpracování signálu a obrazu, Skripta FEL ČVUT, Praha 2000, ISBN 80-01-02114-9

[5] DAVIES, E., R.: Machine Vision - Theory, Algorithms, Practicalities. Morgan Kaufmann Press. UK, 2005, ISBN 0-12-206093-8

Vedoucí diplomové práce:	**Ing. Josef Chaloupka, Ph.D.** Ústav informačních technologií a elektroniky
Konzultant diplomové práce:	**Ing. Jindřich Žďánský, Ph.D.** Ústav informačních technologií a elektroniky
Datum zadání diplomové práce:	**16. října 2009**
Termín odevzdání diplomové práce:	**21. května 2010**

V. Kopecký

prof. Ing. Václav Kopecký, CSc.
děkan

L.S.

Novák

prof. Ing. Ondřej Novák, CSc.
vedoucí ústavu

V Liberci dne 16. října 2009

Prohlášení

Byl(a) jsem seznámen(a) s tím, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 o právu autorském, zejména § 60 (školní dílo).

Beru na vědomí, že TUL má právo na uzavření licenční smlouvy o užití mé diplomové práce a prohlašuji, že **s o u h l a s í m** s případným užitím mé diplomové práce (prodej, zapůjčení apod.).

Jsem si vědom(a) toho, že užít své diplomové práce či poskytnout licenci k jejímu využití mohu jen se souhlasem TUL, která má právo ode mne požadovat přiměřený příspěvek na úhradu nákladů, vynaložených univerzitou na vytvoření díla (až do jejich skutečné výše).

Diplomovou práci jsem vypracoval(a) samostatně s použitím uvedené literatury a na základě konzultací s vedoucím diplomové práce a konzultantem.

Datum 20. 5. 2010

Podpis

Děkuji vedoucímu mé diplomové práce Ing. Josefu Chaloupkovi, Ph.D za užitečné rady a čas, který mi věnoval při vzniku této práce. Zároveň děkuji mé rodině a přítelkyni za podporu a porozumění v době studia.

Abstrakt

Tato diplomová práce se zabývá rozpoznáváním signálu řeči na základě jeho akustické a vizuální složky, tedy audiovizuálním rozpoznáváním.

V prvních dvou kapitolách je popsáno zpracování akustické a vizuální složky řečového signálu a jeho parametrizace. Dále jsou podrobněji popsány nejčastěji využívané příznaky pro rozpoznávání řeči.

Kapitola č.3 popisuje klasifikaci pomocí metody Skrytých Markovských modelů a vysvětluje rozdíl mezi rozpoznáváním izolovaných slov a fonémově orientovaným rozpoznáváním. V kapitole č.4 jsou objasněny principy audiovizuálního rozpoznávání řeči, především potom fůze akustických a vizuálních příznaků.

Kapitola č.5 popisuje používanou audiovizuální databázi a ukazuje úpravy, které byly provedeny pro natrénování kvalitních modelů. V šesté kapitole jsou popsány experimentální testy prováděné na databázi. Jedná se rozpoznávání akustického, vizuálního signálu a audiovizuální rozpoznávání v hlučných podmínkách.

Abstract

This thesis deals with the recognition of speech signal based on the acoustic and visual components, it means audio-visual recognition.

In the first two chapters is described the processing of acoustic and visual components of speech signal and its parameterization. Further in details are described features most frequently used for speech recognition.

Chapter No. 3 describes the classification according to the Hidden Markov models method and explains the difference between isolated words recognition and phoneme oriented recognition. Chapter No. 4 illustrates the principles of audiovisual speech recognition, especially fusion of acoustic and visual features.

No.5 chapter describes used audio-visual database and shows the modifications that were made for training good models. The sixth chapter describes the experimental tests performed on the database. It is a recognition of acoustic, visual signal and audio-visual recognition in the noisy conditions.

Obsah

Úvod

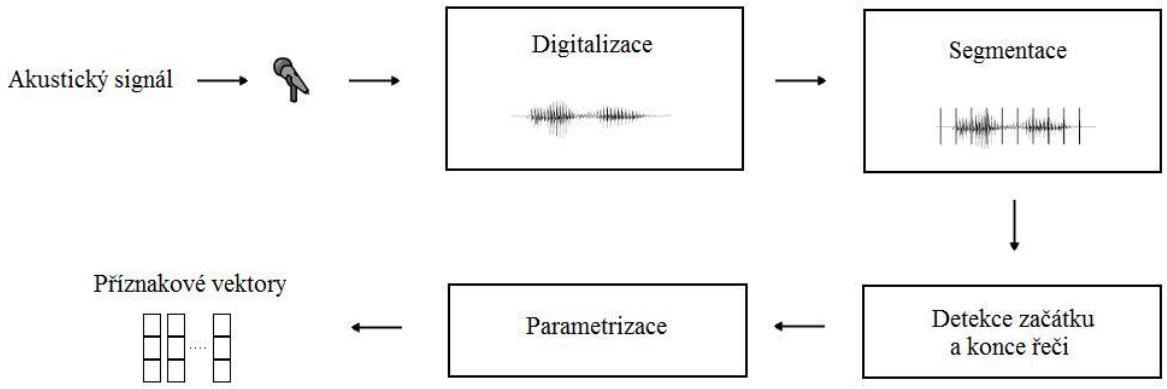
Oblast počítačového zpracování a rozpoznávání řeči přináší už řadu let možnost tvorby aplikací, kde může člověk komunikovat s počítačem přirozenou řečí. Tyto aplikace značně zpříjemňují a zrychlují komunikaci s přístrojem a umožňují pracovat na počítači i tělesně postiženým lidem. V současné době díky technickému pokroku nemusí tyto aplikace sloužit jen pro osobní počítače, ale vznikají například i aplikace pro diktování spojité řeči do takzvaných smart telefonů.

Rozpoznávání akustického signálu řeči je již v současnosti na velice dobré úrovni, je používáno například pro automatický přepis mluveného projevu do textové podoby, pro hlasové ovládání počítačových aplikací a robotů, nebo pro identifikaci mluvčího. Velikým problémem u tohoto rozpoznávání je ovšem okolní ruch. Možná i proto se v posledních letech začalo k rozpoznávání využívat i vizuální složky řečového signálu, na kterou okolní ruch žádný vliv nemá. Vznikla tak oblast audiovizuálního počítačového zpracování a rozpoznávání řeči. Rychlejšímu vývoji této oblasti zcela jistě napomohl i fakt, že dnešní osobní počítače disponují lepším výkonem a hlavně velkým paměťovým prostorem. Vizuální složka řeči totiž oproti akustické zabírá mnohokrát více místa v paměti počítače a její zpracování je náročnější. V dnešní době jsou již dostupnější i kvalitní digitální kamery.

Úkolem diplomové práce je seznámit se více s problematikou audiovizuálního zpracování a rozpoznávání řečového signálu. Upravit nahrávky v databázi mluvčích pro účely rozpoznávání. Navrhnout a vytvořit systém pro audiovizuální rozpoznávání řeči pomocí skrytých Markovských modelů za použití programového balíku HTK. Rozpoznávat řečový signál pomocí akustické, vizuální a obou složek dohromady. Dalším úkolem je provést testování rozpoznávače v hlučných podmínkách a zhodnotit a porovnat dosažené výsledky.

1 Akustická složka řečového signálu

Tato kapitola se zabývá akustickou složkou řečového signálu. Konkrétně potom zpracováním tohoto signálu pro potřeby rozpoznávání. Celý proces se dá rozdělit do několika kroků popsaných níže. Jedná se o digitalizaci signálu, jeho segmentaci, nalezení začátku a konce promluvy a samotnou parametrizaci (viz. Obr. 1).



Obr. 1 –Schéma zpracování akustického signálu řeči

1.1 Digitalizace

Na začátku se musí zpracovat analogový signál z mikrofonu. V dnešní době již tuto úlohu zvládne běžná zvuková karta, která je součástí každého počítače. Pro porozumění obsahu řeči stačí frekvenční pásmo přibližně od 0,3 do 3,4 kHz, proto podle Nyquistova Shannonova vzorkovacího teorému by měla být vzorkovací frekvence vyšší než 6,8 kHz. V databázi, kterou jsem měl k dispozici, měly audio nahrávky vzorkovací frekvenci 8 kHz s šestnáctibitovým rozlišením. Například u záznamu hudby na CD se používá vzorkovací frekvence 44 kHz (průměrné zdravé lidské ucho slyší frekvence maximálně do 20 – 22 kHz). Pro rozpoznávání se nejčastěji volí 8, nebo 16 kHz. Důležité je používat pro trénování i provozování klasifikátoru stejnou vzorkovací frekvenci.

1.2 Segmentace

Charakter akustického signálu se v čase mění, proto je nutné ho pro další práci rozdělit na krátké časové segmenty (tzv. framy). Segmentace musí být pravidelná a délka segmentu se volí tak, aby nebyla delší než trvání nejkratších hlásek (fonémů). Výhodou je, že v takto malých časových úsecích (několik ms) se příliš nemění frekvenční vlastnosti signálu. Framy se volí tak, aby se částečně překrývaly. Tím se dosáhne hladký průběh parametrů počítaných v jednotlivých framech. Pro vzorkovací frekvenci 8 kHz je typická délka framu 20 ms s překryvem 10 ms [2].

1.3 Detekce začátku a konce řeči

Pro nalezení promluvy v řečovém signálu se využívá energie signálu. Předpokládá se, že energie signálu pozadí (ticha) je menší než energie framů promluvy. Toto ovšem nemusí platit při velkém šumu a okolním hluku. Spočítá se energie ve všech framech nahrávky a určí se vhodný práh. Začátek a konec promluvy se potom najde například tak, že se testuje několik prvních (respektive posledních) po sobě jdoucích překročení tohoto prahu. Energii signálu počítáme obvykle v logaritmickém měřítku, s tím že L je počet vzorků v jednom framu:

$$E = \log\left(\sum_{n=1}^{L-1} x^2(n)\right) \tag{1,1}$$

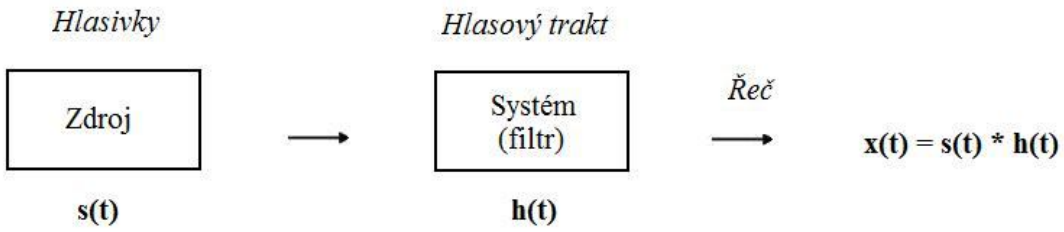
1.4 Parametrizace

Akustický signál v sobě nese hodně redundantních informací, cílem parametrizace je redukovat tuto redundanci a reprezentovat signál menším počtem dat vhodných pro rozpoznávání [1]. Tato data (parametry signálu) nazýváme příznaky. Na konci parametrizace bychom měli získat soubor příznaků pro každý frame signálu, neboli příznakové vektory.

Historicky první příznaky používané pro rozpoznávání byly energie signálu a počet průchodů nulou. Tyto dva parametry mají minimální výpočetní náročnost. Později se díky rychlé Fourierově transformaci (FFT) začalo používat spektrum signálu a spektrální příznaky. Jako další se objevily Lineárně prediktivní koeficienty (LPC). Dnes jsou nejvíce používány statické a dynamické kepstrální příznaky.

1.4.1 Kepstrální příznaky

Pokud se podíváme na řeč z pohledu modelu, vidíme, že hlasivky přestavují zdroj signálu a hlasový trakt jakýsi systém – filtr (viz obr. 2). Řeč je potom v krátkém úseku hypoteticky konvolucí zdrojového signálu a impulsní odezvy systému. Máme tedy dvě složky. První z nich, zdroj, jehož charakter závisí na výšce hlasu, intonaci apod. A druhou, impulsní odezvu systému, tu důležitější pro rozpoznávání řeči. Nastavení systému (filtru) se totiž mění v závislosti na hláskách. Je tedy potřeba oddělit informaci o zdroji a informaci o systému. K tomu se využije tzv. kepstra.



Obr. 2 –Systémový model řeči

Kepstrum c[n] posloupnosti vzorků digitalizovaného řečového signálu x[n] je definováno jako inverzní Fourierova transformace logaritmu absolutní hodnoty spektra signálu:

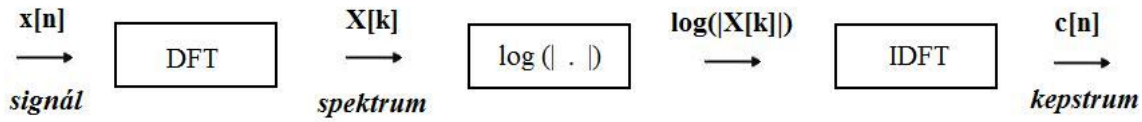
$$c[n] = IDFT\{\log(|DFT\{x[n]\}|)\} \tag{1,2}$$

Kepstrální analýzu lze použít k oddělení složek signálu, který vznikl konvolucí několika složek [1]. Oddělování je založeno na vlastnostech DFT a na pravidle o tom, že logaritmus součinu je roven součtu logaritmů.

DFT převádí konvoluci v časové oblasti na prostý součin v oblasti frekvenční a logaritmus tohoto součinu spektra buzení (zdroj) se spektrem impulsní odezvy systému lze převést na součet logaritmů transformací obou složek [1]. Po aplikaci IDFT je potom kepstrum takovéhoto signálu součtem kepstra buzení a kepstra impulsní odezvy hlasového traktu.

Pomocí tzv. liftrace tohoto kepstra se z něho vyberou pouze jeho nejnižší složky, s periodou podstatně menší než je perioda základního tónu buzení. Právě tyto kepstrální

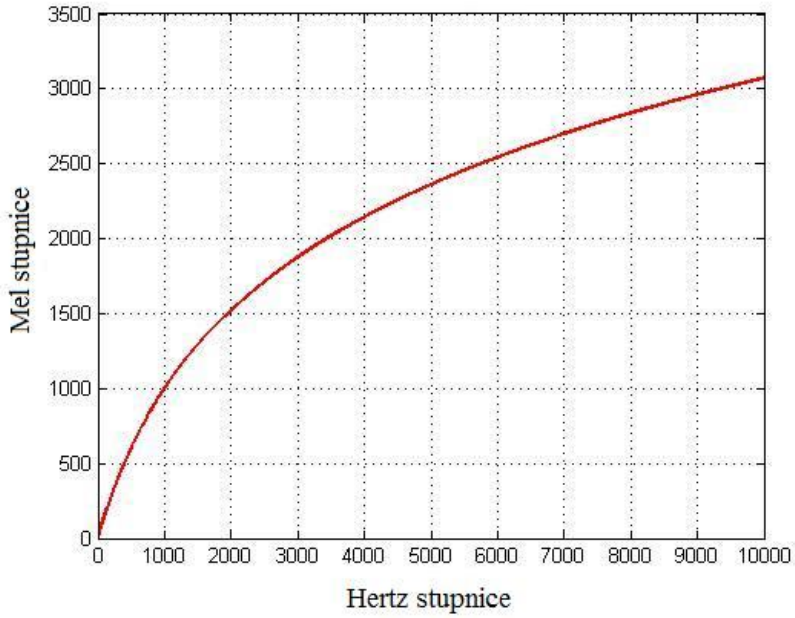
koeficienty, které určují nastavení hlasového traktu, se používají pro popis řečového signálu. Výrazy kepstrum a liftrace vznikly přesmyčkami ze slov spektrum a filtrace.



Obr. 3 – Výpočet kepstra

V této práci jsou využity MFCC (Mel-Frequency Cepstral Coefficients) kepstrální příznaky. Tyto příznaky jsou dnes nejčastěji používány pro zpracování a rozpoznávání řeči. Metoda výpočtu MFCC příznaků vychází z frekvenční oblasti a využívá faktu, že lidské ucho vnímá zvukové frekvence ne v lineární, ale zhruba v logaritmické stupnici. Ve vyšších frekvencích už není schopno tolik rozlišovat rozdíl. Tato stupnice se nazývá melovská (viz obr. 4). Převod mezi frekvencí v Hz a melovskou frekvencí je dán vztahem:

$$Mel(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \qquad (1,3)$$



Obr. 4 –Převodní křivka mezi frekvencí a Mel – frekvencí

Výpočet MFCC koeficientů lze potom rozdělit do několika kroků:

1) Vyříznutí jednoho framu signálu
2) Aplikace preemfázového filtru

   Signál projde číslicovým filtrem definovaným jako:

$$y[n] = x[n] - O{,}97 \cdot x[n-1] \tag{1,4}$$

   Tím jsou posíleny vyšší frekvence, které jsou zeslabeny cestou k mikrofonu a dynamicky potlačena stejnosměrná složka vznikající na zvukových kartách.

3) Aplikace Hammingova okna
4) Výpočet FFT
5) Výpočet spektrálního výkonu
6) Rozdělení spektrálního výkonu do pásem

   Na melovskou stupnici se pomocí trojúhelníkových oken definují částečně se překrývající pásma. Výkony jednotlivých složek FFT se vždy vynásobí příslušným koeficientem okna a uvnitř okna se sečtou. Tak se dostanou výkony v jednotlivých pásmech.

7) Logaritmus výkonu v každém pásmu
8) Výpočet IFFT

   V praxi se zpětná Fourierova transformace provede pomocí Diskrétní kosinové transformace (DCT). Jejím výsledkem jsou už kepstrální koeficienty, nejčastěji se využívá prvních třináct koeficientů.

9) Liftrace

Výsledné koeficienty se vynásobí okénkovou funkcí, danou vztahem:

$$c_n' = \left(1 + \frac{L}{2}\sin\left(\frac{\pi \cdot n}{L}\right)\right) \cdot c_n \qquad (1,5)$$

kde L je délka liftračního okna. Tím se dosáhne vyrovnání rozdílů v rozptylech koeficientů.

10) Výpočet Delta a Delta – Delta koeficientů

Ke statickým MFCC koeficientům se vypočtou ještě dynamické koeficienty (pomocí první a druhé derivace). K výpočtu dynamických koeficientů se používá okolí dvou framů na obě strany.

11) Normalizace MFCC

Tento krok je nutný, pokud jsou nahrávky získány z různých zdrojů a různých prostředí. Použije se operace zvaná CMS nebo CMN (Cepstral Mean Substraction/Normalization), která spočívá ve výpočtu středních hodnot všech koeficientů přes celou nahrávku a odečtení této hodnoty od koeficientů ve všech framech.

2 Vizuální složka řečového signálu

Tato kapitola popisuje vizuální složku řečového signálu. Zabývá se zpracováním vizuálního signálu, nalezením oblasti zájmu, extrakcí vizuálních příznaků ze signálu a jejich popisem.

2.1 Zpracování obrazu

V rozpoznávání řeči lze využít i vizuální složku zaznamenanou na kameru. Výslednou podobu promluvy ovlivňují pohyby jazyka, hlasivek a dalších ústrojí hlasového traktu. Zaznamenat tyto prvky hlasového ústrojí je ovšem velmi obtížné a rozpoznávat řeč pomocí jich je nereálné. Vizuální informace se nejlépe získá z obličeje mluvčího, konkrétně potom z aktuálního nastavení jeho úst. Proto je důležité na pořízených snímcích nejprve detekovat obličej mluvčího a následně i oblast zájmu (ROI – region of interest), v našem případě ústa. V diplomové práci je využívána audiovizuální databáze, kde byla mluvčím snímána hlava, řešila se tedy jen úloha nalezení oblasti zájmu ROI (viz. kapitola 5). Před výpočtem vizuálních příznaků se ještě obraz převádí z barevného prostoru RGB na šedotónový obraz.

2.2 Vizuální příznaky řeči

První co by přirozeně každého napadlo, by bylo použít jako vizuální příznaky řeči parametry mluvčího úst. Takovéto příznaky se nazývají tvarové a patří mezi ně například horizontální a vertikální rozšíření rtů, velikost a zaokrouhlení rtů, nebo geometrické příznaky získané z analýzy hranice obrysu rtů. 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 nalezena) [3].

V dnešní době nejpoužívanějšími vizuálními příznaky jsou příznaky popisující informační obsah obrazu. Získání příznaků není tak výpočetně náročné, protože zde odpadá předzpracování oblasti zájmu ROI. Použít jako příznakový vektor celý obrazový prostor by bylo nepraktické řešení, například pro oblast zájmu 128x128 obrazových bodů bychom dostali pro jeden snímek 16384 příznaků. Takovýto příznakový vektor by obsahoval velké množství zbytečných dat a natrénované HMM modely by nebyly příliš

spolehlivé pro rozpoznávání řeči. Snímek oblasti zájmu ROI se proto vhodně transformuje a z transformovaného obrazu se vyberou pouze složky, které dobře reprezentují pořízený obraz. Nejpoužívanějšími transformacemi jsou analýza hlavních komponent (PCA – the Principal component analysis), lineární diskriminační analýza (LDA – the Linear discriminant analysis) a diskrétní kosinová transformace (DCT – the Discrete cosine transform), využívaná i v diplomové práci. Zřídka se využívají i jiné obrazové transformace, například diskrétní vlnková transformace (DWT – the Discrete Wavelet trasnform) [3].

2.2.1 DCT vizuální příznaky

Příznaky získané 2D diskrétní kosinovou transformací (DCT) z oblasti zájmu ROI mají oproti PCA příznakům výhodu v možnosti rychlejšího výpočtu pomocí algoritmu rychlé kosinové transformace (FCT – Fast cosine transorm, obdoba FFT – Fast Fourier transform). Pro použití transformace 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 není tato podmínka splněna, musí se obraz na tuto velikost aproximovat [3]. Je proto vhodné volit velikost obrazů oblasti zájmu tak, aby rovnou tuto podmínku splňovaly.

Velikost oblasti zájmu ROI se nejčastěji volí 64x64 nebo 128x128 obrazových bodů. Zároveň je nutné zajistit normalizaci ROI, rty mluvčího by se měly nacházet přibližně uprostřed obrazu a pro všechny mluvčí by měl mít objekt rtů přibližně stejnou velikost. Detailněji popisuje nalezení oblasti ROI kapitola číslo 5.

2.2.1.1 Diskrétní kosinová transformace

Existuje několik definicí diskrétní kosinové transformace, v diplomové práci je využit algoritmus rychlé 2D kosinové transformace FCT, vycházející ze vztahu pro DCT-II:

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

kde *f(m,n)* jsou hodnoty z původního obrazu o rozměrech *NxN*, *F(u,v)* jsou koeficienty transformovaného obrazu, *0 <= u,v <= N-1* a *c* jsou koeficienty z (2,2).

$$c(k) = \frac{1}{\sqrt{2}} \quad pro \quad k = 0 \; \frac{1}{\sqrt{2}}$$
$$c(k) = 1 \quad pro \quad k = 0 \; 1 \tag{2,2}$$

2.2.1.2 Výpočet vizuálních DCT příznaků

Jak už bylo uvedeno výše, DCT transformace slouží k redukci obrazových dat na informaci dobře popisující vizuální složku řeči v obraze. Nepoužije se tedy celý transformovaný prostor DCT koeficientů, ale hledá se menší prostor příznaků, které poté slouží k vlastnímu rozpoznávání. Pro výběr příznaků existuje několik metod, používají se například metody založené na výpočtu rozptylu a normovaného rozptylu. V diplomové práci je použita metoda výpočtu energie E:

$$E(u,v) = F(u,v)^2 \tag{2,3}$$

Kde F(u,v) jsou koeficienty transformovaného obrazu, *0 <= u,v <= N-1* a *NxN* je rozměr obrazu.

Z vypočtených koeficientů je poté vybíráno jako vizuální příznaky *P* koeficientů, které mají nejvyšší hodnotu [3].

2.2.1.3 Normalizace příznakového vektoru

Nyní tedy máme matici příznakových vektorů o rozměru *PxN* , kde *P* je počet vybraných maximálních hodnot DCT koeficientů a *N* počet snímků v jedné nahrávce. Pro příklad: z videonahrávky o délce 1 sekundy při snímkovací frekvenci 30*Hz* vznikne 30 snímků, vybíráme 5 maximálních DCT koeficientů, matice má tedy rozměr 5x30.

Úrovně vypočtených hodnot DCT vizuálních příznaků se stoupajícím indexem razantně klesají. Pro potlačení tohoto jevu se používají různé aproximační metody, které vyrovnávají úrovně hodnot jednotlivých příznaků [3]. Používá se například zlogaritmování všech hodnot z vizuálního příznakového vektoru nebo odečtení střední hodnoty příznakového vektoru. Tím se eliminuje různá střední hodnota z jednotlivých videonahrávek, u akustického signálu řeči se tento krok normalizace provádí kvůli zbavení se závislosti na mluvčím.

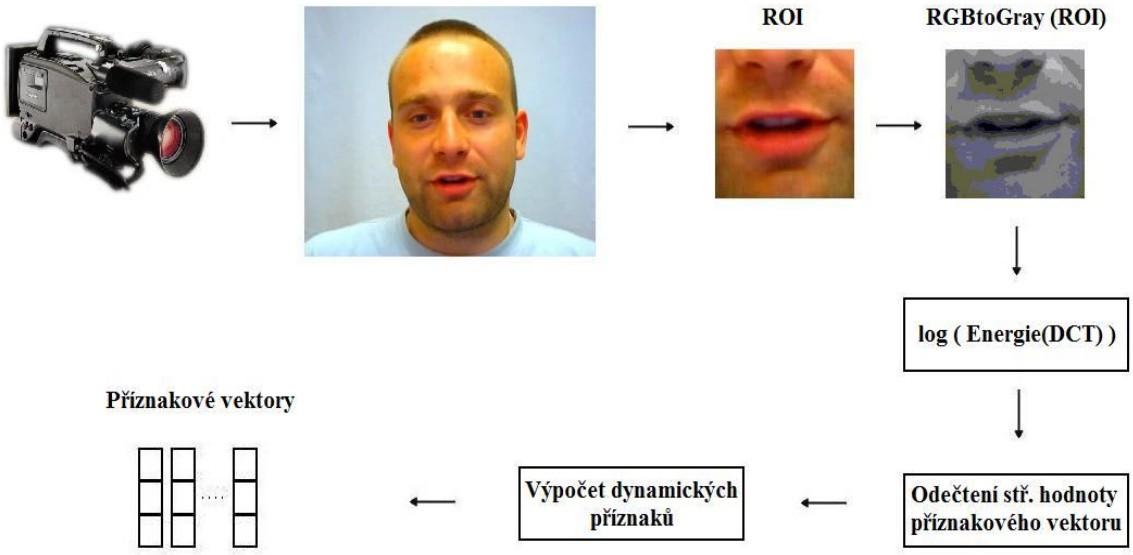
2.2.1.4 Dynamické a akcelerační DCT příznaky

Kvůli lepšímu popisu dynamického chování řečového signálu se obdobně jako u akustického signálu řeči a kepstrálních příznaků počítají dynamické DCT příznaky. Rozpoznávání řeči jen pomocí statických DCT příznaků nevede k velkému rozpoznávacímu skóre. Proto se k nim přidávají ještě dynamické (2,4), akcelerační (2,5) a někdy i tzv. triple-delta příznaky. Ty se vypočítají ze statických DCT příznaků jednoduchou diferencí:

$$x'[n] = x[n] - x[n-1] \tag{2,4}$$

$$x''[n] = (x[n] - x[n-1]) - (x[n-1] - x[n-2]) \tag{2,5}$$

kde *x[n]* jsou jednotlivé příznaky v čase *n*.



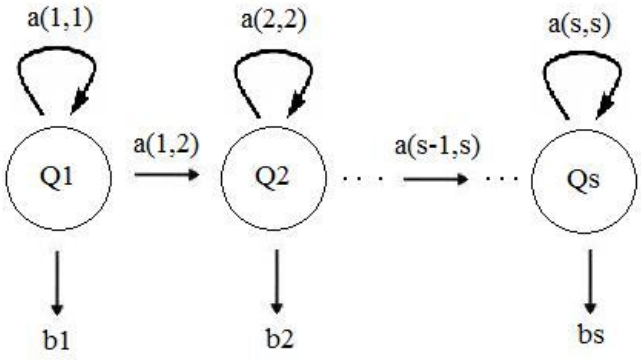
Obr. 5 – Schéma parametrizace vizuální složky řečového signálu

3 Rozpoznávání řečového signálu

Po zpracování řečového signálu, parametrizaci a extrakci příznaků následuje v rozpoznávání řeči důležitý krok klasifikace. Jako klasifikátor se dříve používala metoda DTW (Dynamic Time Warping), česky dynamické borcení času. Její nevýhodou je ovšem závislost referencí na mluvčím a tudíž nutnost velkého množství referencí od různých osob k zbavení se této závislosti. Není zde ani možnost přidávat jednoduše nová slova do slovníku. Každé slovo musí mít svojí referenční nahrávku. Proto se v devadesátých letech dvacátého století přešlo k parametrický modelům. Dnes nejvyužívanějším klasifikátorem je metoda skrytých Markovských modelů (Hidden Markov model – dále jen HMM). Využívají se i tzv. Neuronové sítě.

3.1 Metoda skrytých Markovských modelů (HMM)

Při snaze vytvořit univerzálnější referenční vzory se ukázalo, že není nutné, aby bylo slovo reprezentováno konkrétní a úplnou posloupností framových vektorů. Řeč se totiž skládá z kratších či delších stacionárních úseků, tvořených většinou několika sousedními framy, v nichž se parametry mění jen málo [2]. Přešlo se tedy k abstraktním modelům, kde jsou framy nahrazeny menším počtem stavů. Každý stav potom nese informaci o vlastnostech framů, které reprezentuje. Tato informace se získá použitím statistických metod, které každému stavu určí statistické rozložení hodnot příznakových vektorů. Obvykle je to normální Gaussovo rozložení.



Obr. 6 –HMM – struktura modelu

Typický model slova má tzv. levo-pravou strukturu (viz obr. 6). Přechod je možný pouze do sousedního stavu a to zleva doprava. Pravděpodobnost, že model setrvá v aktuálním stavu s je vyjádřena hodnotou ass. Pravděpodobnost přechodu do následujícího stavu hodnotou ass+1. Platí, že ass + ass+1 = 1. Každý stav je popsán pravděpodobnostní výstupní funkcí s normálním rozložením:

$$b_s(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma_s} \exp\left[-\frac{(x - \mu_s)^2}{2\sigma_s^2} \right] \tag{3,1}$$

kde μ je střední hodnota a σ je rozptyl. Vztah určuje míru pravděpodobnosti, že frame popsaný příznakem s hodnotou x patří ke stavu s.

Střední hodnota a rozptyl jsou dány vztahy:

$$\mu_s = \frac{1}{N_s} \sum_{i=1}^{N_s} x_i \tag{3,2}$$

$$\sigma_s^2 = \frac{1}{N_s} \sum_{i=1}^{N_s} (x_i - \mu_s)^2 \tag{3,3}$$

kde Ns je počet framů přiřazených stavu s.

Je-li k dispozici větší množství dat pro trénování, lze model ještě zkvalitnit tzv. vícemodálním normálním rozložením. Data, z kterých se určují parametry normálního rozložení, jsou často uspořádána ve více shlucích. Proto se přesnější popis rozložení dat dostane, má-li každý tento shluk přiřazenu vlastní Gaussovu funkci. Výsledné rozložení je potom váženou směsí jednotlivých Gaussových funkcí. Takováto „směs“ bývá označována jako Gaussian Mixture Model (GMM) a jednotlivé složky potom hovorově „mixtury“. GMM je dán vztahem:

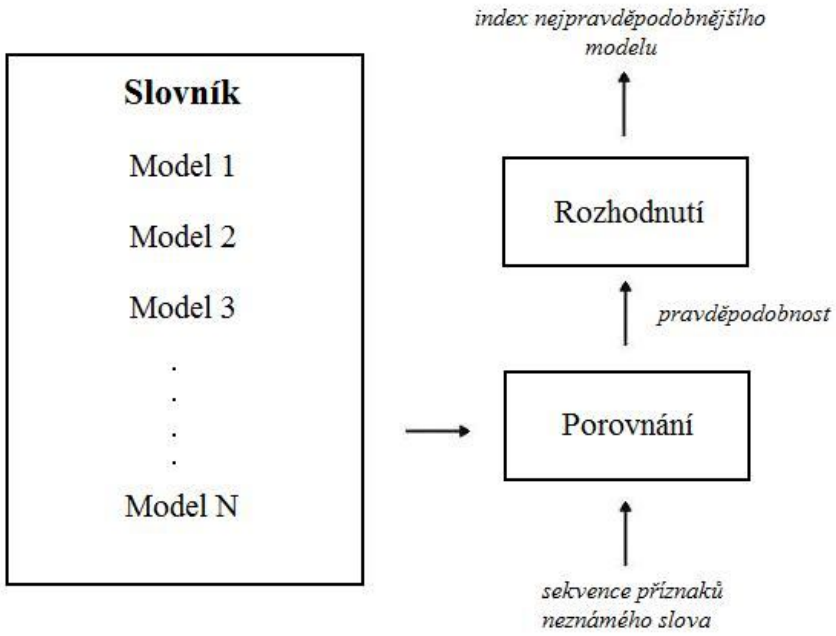
$$b_s(x) = \sum_{m=1}^{M} c_{sm} \frac{1}{(2\pi)^P \cdot \det \sum_{sm}} \exp\left[-\frac{1}{2} (x - \bar{x}_{sm})^T \sum_{sm}^{-1} (x - \bar{x}_{sm}) \right] \tag{3,4}$$

kde c je váhový koeficient a M je počet mixtur.

3.1.1 Trénování a rozpoznávání pomocí HMM

Ve skutečnosti při trénování není známo který frame patří ke kterému stavu, proto název „skryté“ Markovské modely. Trénování jednotlivých modelů a určování jejich parametrů je proto iterativní a dalo by se rozdělit do 4 kroků:

1) Inicializační (Framy všech nahrávek daného slova jsou rovnoměrně rozděleny jednotlivým stavům, jsou spočteny střední hodnoty, rozptyly a přechodové pravděpodobnosti.)
2) Přiřazovací (Pomocí Viterbiho algoritmu je nalezeno lepší, obvykle už ne rovnoměrné rozložení mezi framy a stavy.)
3) Reestimační (Pro nové rozložení jsou opět spočteny střední hodnoty, rozptyly a přechodové pravděpodobnosti.)
4) Opakování nebo Konec (Pokud se střední hodnoty a ostatní parametry liší od předchozích o více než stanovený práh, zopakuje se celý proces od kroku 2), jinak je trénování ukončeno.)



Obr. 7 –Schéma klasifikátoru HMM

Všechny natrénované modely jsou uloženy do slovníku (viz. obr. 7). Platí, že všechny modely mají stejný počet stavů a shodnou strukturu. Liší se pouze parametry výstupních funkcí a hodnotami přechodových pravděpodobností.

Klasifikace spočívá ve vyhodnocení míry pravděpodobnosti, že neznámé slovo, reprezentované posloupností příznaků, patří některému z modelů [2]. Pravděpodobnost lze popsat vztahem:

$$P(X,M) = \max_{f} \prod_{i=1}^{I} a_{f(i-1)f(i)} b_{f(i)} x(i) \tag{3,5}$$

kde slovo X je popsáno časovou posloupností x1,…,xi a model M určen S stavy s parametry a a b, f je přiřazovací funkce.

Vztah lze interpretovat tak, že pravděpodobnost modelu M pro slovo X se určí jako maximální z pravděpodobností vypočítaných přes všechna přípustná přiřazení stavů modelu a framů slova [2]. Přiřazovací funkce f přiřazuje framy slova ke stavům modelu. Omezení je dáno strukturou HMM, umožňuje pouze setrvat v aktuálním stavu, nebo přechod do následujícího. Častější je setrvání v současném stavu, to je dáno mnohem menším počtem stavů S než počtem framů slova I.

3.2 Rozpoznávání izolovaných slov

V této diplomové prácí jsou řešeny dva typy úloh, rozpoznávání izolovaných slov (Isolated word recognition – IWR) a fonémově orientované rozpoznávání slov reprezentovaných hláskovými modely, použitelné i pro rozpoznávání spojité řeči (Continuous speech recognition – CSR).

Úloha IWR má pouze omezené využití, jedná se o aplikace, kde ve slovníku stačí mít pouze omezené množství slov (řádově desítky, maximálně stovky) nebo krátkých slovních spojení a není nutné slovník často aktualizovat. Na Technické univerzitě v Liberci vznikly například aplikace pro hledání kontaktů v telefonním seznamu, linka InfoCity (informace o kultuře, sportu, dopravě apod. v Liberci) a další. Pro robustnější modely musí mít každé slovo nebo krátké slovní spojení („Do Liberce“) co nejvíce referenčních nahrávek od více mluvčích. Izolovaná slova jsou obvykle vyslovována

s větší pečlivostí a je snazší detekovat začátek a konec promluvy. V úloze IWR lze předpokládat, že v jedné detekované promluvě je právě jedno slovo (jedno slovní spojení), klasifikátor proto řeší úlohu, která z N položek ve slovníku to je. Úloha má tedy lineární složitost.

Protože nejmenší jednotkou ve slovníku je model slova, nelze tímto typem úlohy rozpoznávat spojitou řeč. Teoreticky ano, ale slovník by musel obsahovat všechna slova ve všech tvarech, což by znamenalo obrovské množství referenčních nahrávek a velmi rozsáhlé a pomalé rozpoznávání.

3.3 Fonémově orientované rozpoznávání

Cestu k neomezenému slovníku řeší použití modelů hlásek (fonémů). V mluvené řeči hrají fonémy obdobnou roli jako písmena v psaném jazyce. Fonémy jsou jazykově závislé a mají omezené množství, ve většině evropských jazyků se jejich počet pohybuje mezi 25 až 50. Některé fonémy jsou ve více jazycích, jiné jsou naopak unikátní („nosové“ hlásky ve francouzštině). V češtině máme 41 fonémů (viz. tab. 1), abeceda obsahuje například dva typy ř (znělé a neznělé), specifické hlásky jako „dž“ nebo „dz“. Každý foném je označen pouze jedním symbolem, ve fonetice se tedy neřeší rozdíl mezi y a i, dvoumístné ch je nahrazeno symbolem X, dva typy jedné hlásky jsou většinou rozlišeny malým a velkým písmenem (ř a Ř).

U fonémového rozpoznávání jsou běžně používány modely s menším počtem stavů (obvykle 3 stavy), než v úloze rozpoznávání izolovaných slov. Aby bylo možné dobře postihnout akustickou variabilitu, mívají výstupní funkce tvar vícesložkových Gaussovských rozložení (více mixtur), pracuje se s 16 i více mixturami [2]. U profesionálních programů je nutné modely natrénovat na záznamech řeči obsahujících tisíce realizací každé hlásky v různých slovech a slovních spojeních, od co největšího počtu osob. Pro jednu hlásku je možné natrénovat i více modelů, které se liší například kontextem, nebo okolím hlásky, jedná se o tzv. „trifóny“. Ke všem nahrávkám v trénovací databázi je nutné vytvořit i fonetický přepis.

Číslo	Foném vyjádřený českými hláskami	Foném dle PAC	Příklad	Číslo	Foném vyjádřený českými hláskami	Foném dle PAC	Příklad
1	„a“	a	táta	21	„m“	m	máma
2	„á“	á	táta	22	„m“	M	tramvaj
3	„b“	b	bába	23	„n“	n	víno
4	„c“	c	ocel	24	„n“	N	banka
5	„dz“	C	leckde	25	„ň“	ň	koně
6	„č“	č	čichá	26	„o“	o	kolo
7	„dž“	Č	rádža	27	„ó“	ó	óda
8	„d“	d	jeden	28	„p“	p	pupen
9	„ď“	ď	dělat	29	„r“	r	bere
10	„e“	e	lev	30	„ř“	ř	moře
11	„é“	é	méně	31	„ř“	Ř	keř
12	„f“	f	fauna	32	„s“	s	sud
13	„g“	g	guma	33	„š“	š	duše
14	„h“	h	aha	34	„t“	t	dutý
15	„ch“	X	chudý	35	„ť“	ť	kutil
16	„i“ nebo „y“	i	bil, byl	36	„u“	u	duše
17	„í“ nebo „ý“	í	vítr, lýko	37	„ú“ nebo „ů“	ú	růže
18	„j“	j	dojat	38	„v“	v	láva
19	„k“	k	kupec	39	„z“	z	koza
20	„l“	l	dělá	40	„ž“	ž	růže
				41	Neutrální samo hláska	E	*)

Tab. 1 – Nouza J. , Psutka J. , Uhlíř J. – Česká fonetická abeceda - PAC [9]

3.3.1 Vizémy

To co znamenají v akustické složce řečového signálu fonémy, jsou ve vizuální složce vizémy. Dalo by se tedy laicky říci, že se jedná o „polohu“ rtů při promluvě. Protože některé hlásky mají hodně podobnou vizuální odezvu, je vizémů méně než fonémů. V diplomové práci je použito rozdělení na 13 vizémových tříd, které je převzato z disertační práce Ing. Petra Císaře Ph.D. [4] (viz. tab. 2). Je vidět, že například hlásky b, m, p představují jeden vizém B, u těchto si lze snadno představit polohu rtů, která je skoro identická. Pro natrénování a rozpoznávání vizému je tedy nutná transkripce z fonetické podoby. Nejjednodušší je přímá transkripce, použitá i v diplomové práci, kterou se ovšem zavádí menší chyba. Akustická a vizuální složka totiž mohou být asynchronní.

Číslo	Fonémy	Vizém
1	a , á	A
2	b , m , M, p	B
3	c , C , s , z	C
4	č , Č , ř , Ř , š , ž	Č
5	d , n , N , t	D
6	ď , j , ň , ť	Ď
7	e , é	E
8	f , v	F
9	g , h , X , k	G
10	i , í	I
11	l , r	L
12	o , ó	O
13	u , ú	U

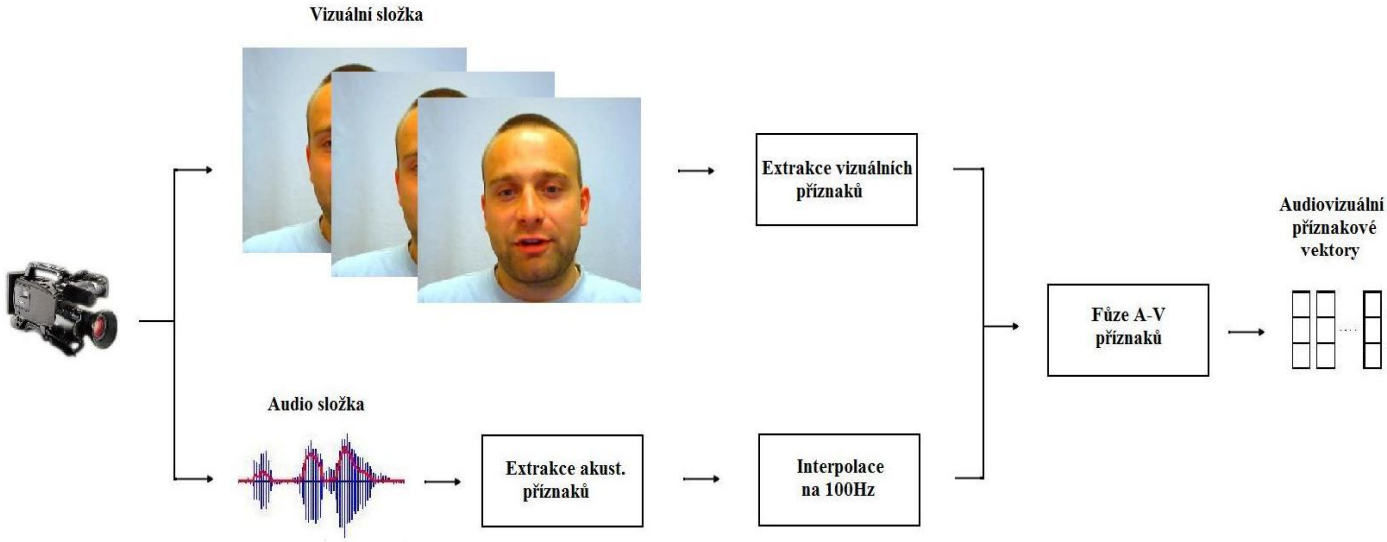
Tab. 2 – Tabulka transkripce z fonémů na vizémy [4]

4 Audiovizuální rozpoznávání řeči

V této kapitole je popsán princip rozpoznávání audiovizuálního signálu řeči. Proces se dá rozdělit do dvou kroků, parametrizace audiovizuálního signálu (viz. obr. 8) a samotné natrénování HMM modelů a rozpoznávání.

Video signál z kamery je rozdělen na akustickou a vizuální složku. Z akustického signálu jsou parametrizací extrahovány kepstrální příznaky (viz. kapitola 1). Vizuální složku reprezentují snímky mluvčího získané z video signálu. Na všech snímcích použitých pro trénování a rozpoznávání musí být nalezena oblast zájmu ROI a z ní jsou extrahovány vizuální DCT příznaky (viz. kapitola 2).

Dvě oddělené sady příznakových vektorů se spojí v jednu v kroku „fůze audiovizuálních příznaků“. Z tohoto souboru audiovizuálních příznakových vektorů jsou dále natrénovány skryté Markovské modely HMM (viz. kapitola 3) a může probíhat rozpoznávání.



Obr. 8 – Parametrizace audiovizuálního řečového signálu

4.1 Fůze audiovizuálních příznaků

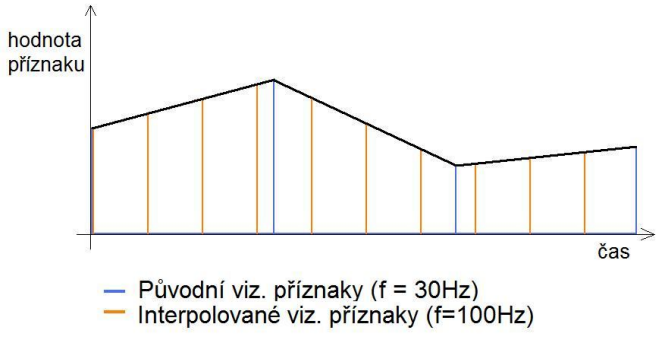
Jak již bylo uvedeno výše, parametrizace akustické a vizuální složky probíhá odděleně. Je proto nutné příznaky spojit dohromady. Existuje několik možností jak tuto fůzi a následné rozpoznávání provádět.

Jednou z nich je rozpoznávat akustický a vizuální řečový signál samostatně. Pro každý signál se odhaduje pravděpodobnost $P(X_i , x_a)$, $P(X_i , x_v)$ určující, nakolik odpovídá dané slovo X_i ze slovníku akustickému x_a nebo vizuálnímu x_v příznaku. Z těchto dvou pravděpodobností je na základě fůze určena výsledná pravděpodobnost $P(X_i , x_a\ x_v)$. Při fůzi se pracuje i s váhovými koeficienty [3].

V diplomové práci je využita metoda, kde se akustické a vizuální příznaky spojí v jeden příznakový vektor. Po jednotlivých framech se uloží nejprve akustické kepstrální příznaky a za ně vizuální DCT statické, dynamické a akcelerační příznaky.

4.1.1 Interpolace vizuálních příznaků

Akustické příznaky jsou získávány s frekvencí 100Hz, zatímco frekvence získávání vizuálních příznaků je závislá na snímkovací frekvenci digitální kamery, kterou jsou video nahrávky pořízeny. Tato frekvence bývá obvykle 25Hz nebo v našem případě 30Hz. U videokamer existuje i možnost pracovat v prokládaném režimu, ovšem na úkor snížení výsledného rozlišení. Aplikace audiovizuálního rozpoznávání řeči vyžadují synchronizaci akustického a vizuálního příznakového vektoru pro všechny nahrávky. Ne příliš používanou variantou je zopakování více snímků pro vyrovnání rozdílu (stejnou funkci má i interpolační metoda „nejbližšího souseda“). Potom jsou zde další dvě možnosti, získávat akustické příznaky také s frekvencí 30Hz (tato varianta se ale nepoužívá), nebo interpolovat vektor s vizuálními příznaky na 100Hz (viz. obr. 9). V diplomové práci je využívána lineární interpolace.



Obr. 9 –Lineární interpolace vizuálních příznaků

5 Audiovizuální databáze

V této kapitole je popsána audiovizuální databáze, kterou jsem dostal k dispozici pro účely rozpoznávání. Jsou rozebrány úpravy nutné pro natrénování HMM modelů, dobře reprezentujících nahrávky. Blíže je vysvětlena i metoda nalezení oblasti zájmu ROI ve vizuální složce řečového signálu. Od vedoucího práce mi byly pro tyto úpravy poskytnuty programy *Aligner* a *Marker*.

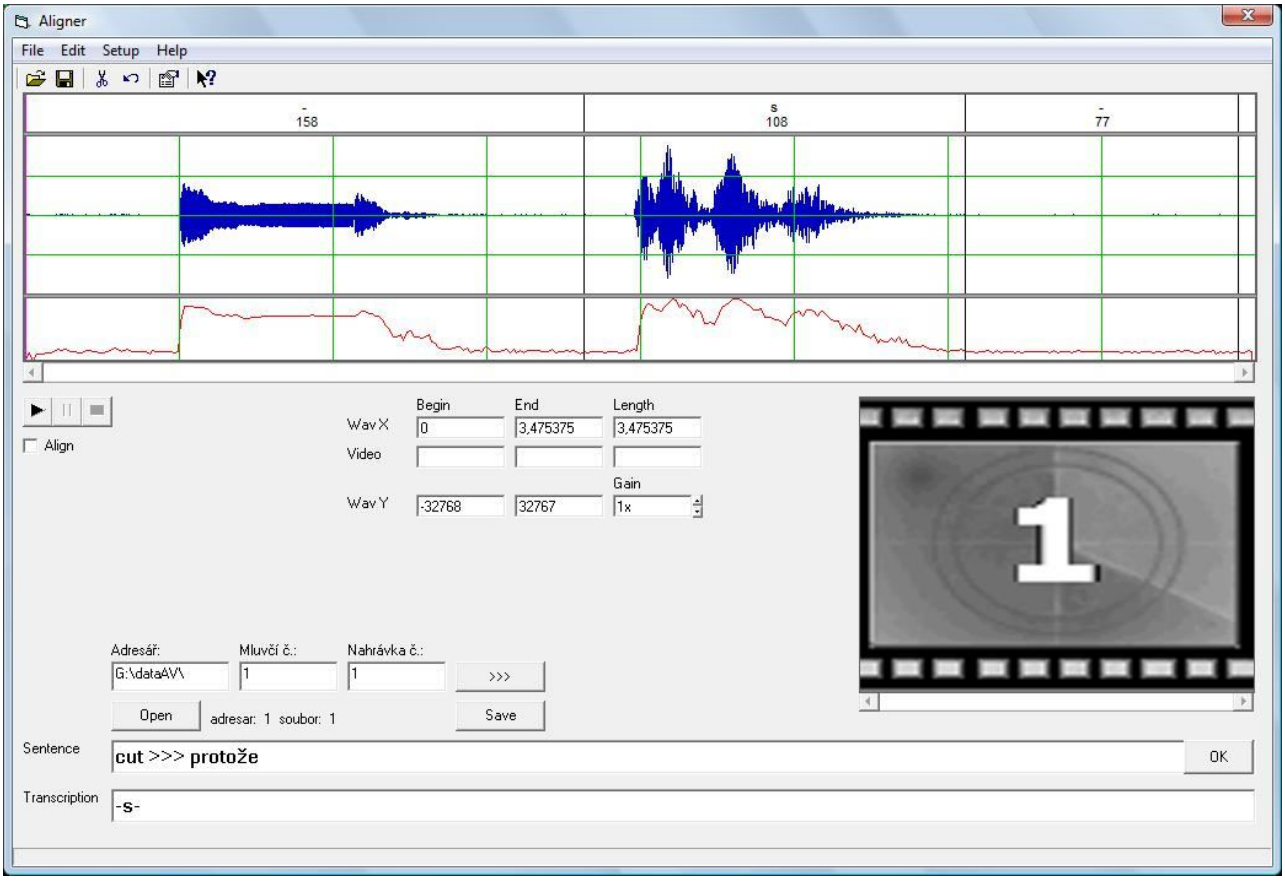
5.1 Popis AV databáze

Databáze vznikla na Technické univerzitě v Liberci v roce 2005, obsahuje nahrávky od 35 mluvčích, z čehož jsou 3 ženy a 32 mužů. Každý mluvčí poskytl 100 nahrávek. Slovník obsahuje 50 slov a 50 vět. Slova byla vybrána náhodně, věty byly vybrány z českých přísloví. Aby bylo možné dobře natrénovat i HMM modely hlásek, bylo 37 přísloví doplněno o 13 tzv. „foneticky bohatých“ vět. V českých příslovích se neobjevují hlásky jako „dž“, málo zastoupené jsou i hlásky „M“, „N“, nebo „ó“. Proto se ve slovníku objevily věty jako: „Virdžínie poslouchá ráda džes.“ apod.

Videonahrávky byly zaznamenány na digitální kameru se snímkovací frekvencí 30 snímků za sekundu a rozlišením 640x480 obrazových bodů. Dále byly rozděleny na vizuální a akustickou složku. Vzniklo tak 494780 snímků, uložených jako 24 bitové bitmapy. Zvukové stopy byly převzorkovány na frekvenci 8000 Hz s rozlišením 16 bitů. Databáze tak zabírá na disku přibližně 500GB.

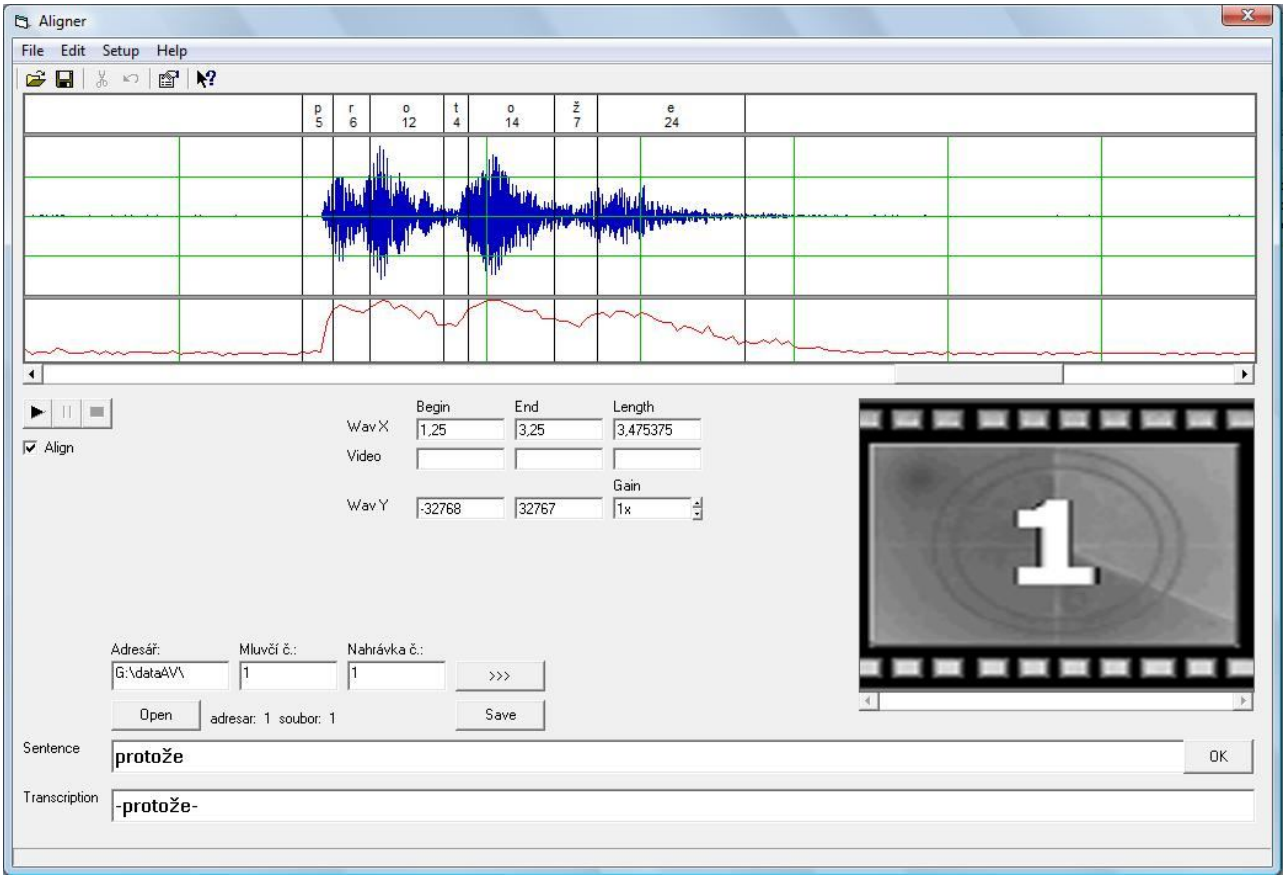
5.2 Úpravy akustické složky

Časově nejnáročnější, ale nutnou součástí rozpoznávání řeči je úprava dat pořízených do databáze určené pro trénování modelů. Zvukové stopy v databázi obsahovaly i oznamovací tón, který signalizoval mluvčímu, že může začít s promluvou. Bylo tedy potřeba všechny nahrávky sestříhat a uložit je (viz. obr. 10). Aby byla provázána akustická a vizuální složka, Aligner ukládá do informačního souboru *.nfo i informaci o prvním a posledním framu (snímku).



Obr. 10 – Ukázka z programu *Aligner* (krok stříhání)

Program *Aligner* dále umožňuje poloautomatické nalezení hranic jednotlivých hlásek v promluvě. Časové hranice hlásek, uložené do informačního souboru o nahrávce, jsou dále využity pro trénování modelů fonémů a vizémů. *Aligner* po načtení nahrávky nalezne přibližné hranice hlásek a umožňuje manuální dorovnání jednotlivých hranic. K větší přehlednosti a lepšímu nalezení hranic je zobrazen průběh zvukové stopy v čase, průběh energie signálu v čase a je možnost přehrát si jednotlivé úseky nahrávky (viz. obr. 11). Tento krok byl časově velice náročný, dorovnat hranice fonémů pro jednoho mluvčího (50 slov a hlavně 50 vět) zabralo průměrně 2 hodiny. Z odposlechu bylo nejprve velice složité poznat jednotlivé fonémy (časově se pohybujeme řádově v desítkách milisekund), postupem času se ale schopnost rozpoznat hlásku zlepšovala, hlavně díky využití informace o časovém průběhu energie.



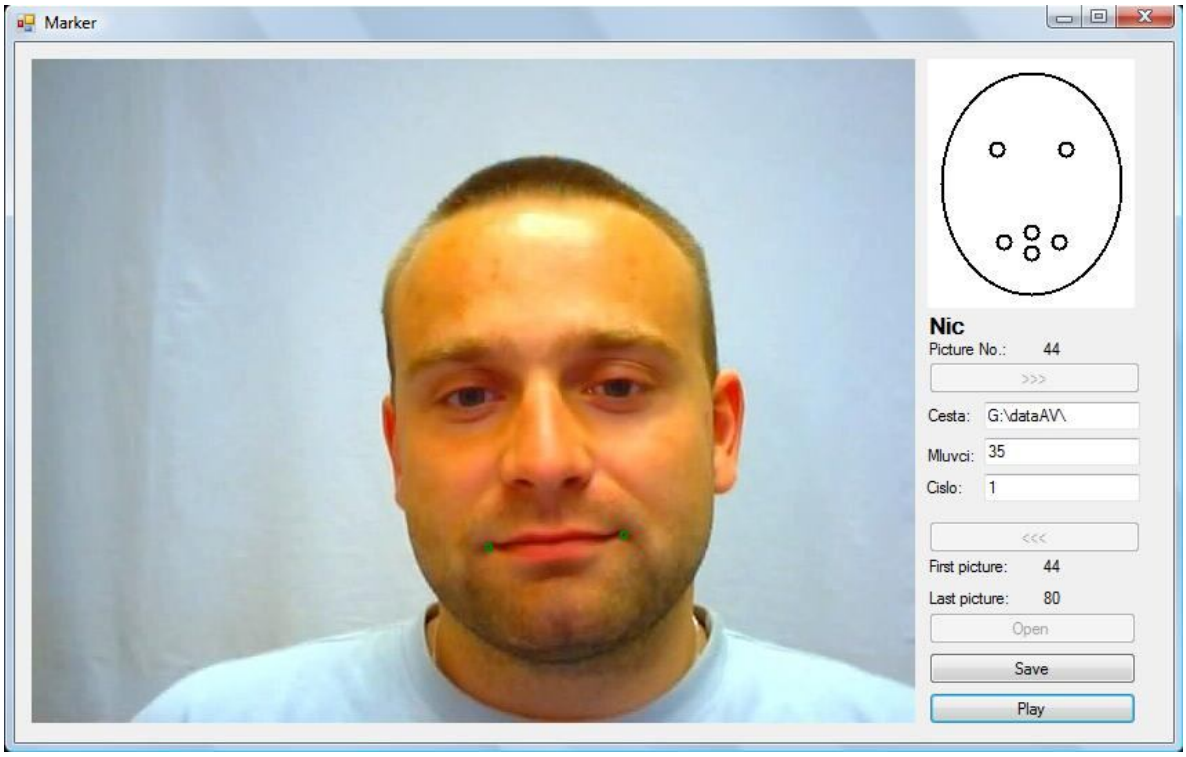
Obr. 11 – Ukázka z programu *Aligner* (krok zarovnávání časových hranic fonémů)

5.3 Úpravy vizuální složky

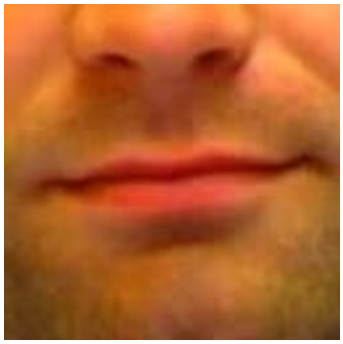
Pro vypočtení hodnotných vizuálních příznaků (viz. kapitola 2) musí nejprve vstupní obraz projít tzv. předzpracováním. Nejdůležitějším krokem je detekovat v obraze oblast zájmu ROI, ze které jsou příznaky extrahovány. U rozpoznávání řeči je zaměřena pozornost na obličej mluvčího, konkrétně potom na jeho rty. Detekci obličeje lze provádět automaticky například pomocí barevné a tvarové segmentace obrazu. Detekci rtů s využitím převodu do jiných barevných prostorů a metod pro segmentaci obrazu prahováním nebo segmentaci obrazu za pomocí různých barevných statistických modelů [3].

V diplomové práci byla detekce oblasti zájmu ROI, ve které jsou řečníkovi rty, prováděna poloautomaticky. V programu *Marker* (viz. obr. 12) byly vždy na prvním, prostředním a posledním snímku nahrávky označeny levý a pravý okraj úst mluvčího. Informace o poloze úst v obraze byly uloženy do souboru *.xls . Díky této informaci byla oblast zájmu ROI (výřez o velikosti 128x128 pixelů) nalezena u ostatních snímků

automaticky (viz. obr. 13). Poloautomatická detekce byla použita s očekáváním lepšího nalezení oblasti zájmu než u plně automatické detekce.



Obr. 12 – Ukázka z programu *Marker* (označení úst mluvčího na snímku)



Obr. 13 – Nalezená oblast zájmu ROI (128x128)

6 Testy prováděné na databázi

V kapitole jsou popsány jednotlivé testy a jejich výsledky. U každého testu je detailně popsána jeho specifikace (slovník, trénovací a testovací databáze, použité příznaky atd.), postup a přípravy při trénování a rozpoznávání. Řeč byla rozpoznávána pomocí akustické složky (fonémově orientované rozpoznávání), vizuální složky (DCT parametrizace, vliv použitých vizuálních příznaků) a nakonec bylo provedeno audiovizuální rozpoznávání s vyhodnocením vlivu okolního hluku na rozpoznávací skóre.

Trénování HMM modelů a rozpoznávání bylo prováděno pomocí programového balíku HTK Toolkit [5]. Programy pro ovládání HTK pomocí dávkových souborů, fonetické a vizémové transkripce, vyhodnocování rozpoznávacího skóre, fůzi audiovizuálních příznakových vektorů atd. byly vytvořeny v MS Visual C++ studio 2008 [6]. Pro DCT parametrizaci vizuální složky řeči byl využit program Matlab [7].

6.1 Rozpoznávání akustického signálu řeči

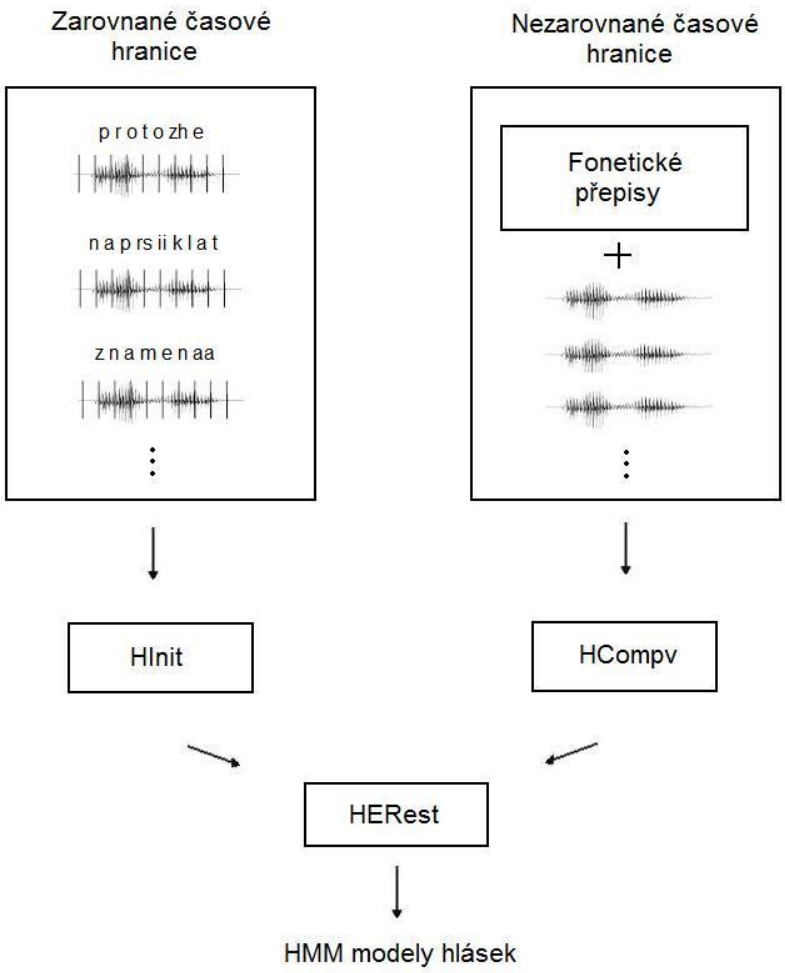
Specifikace testu	
Druh HMM modelů	Modely hlásek
Počet stavů HMM modelů	3
Délka framů	25ms (10ms překryv)
Druh příznaků	MFCC + dynamické a akcelerační
Počet příznaků	39 (13+13+13)
Slovník	50 slov
Trénovací databáze	50 vět od 35 mluvčích (1750 vět)
Testovací databáze	50 slov od 35 mluvčích (1750 slov)

V této úloze byl zjišťován vliv počtu mixtur v HMM modelech na výsledné rozpoznávací skóre. Rozpoznávána byla akustická složka řeči a bylo použito fonémově orientované rozpoznávání. Modely jednotlivých hlásek byly třístavové a měnil se počet mixtur. Bylo vytvořeno 41 modelů, 40 základních fonémů (viz. tab. 1) a 1 model pro krátkou pauzu (ticho). Pro fonémové rozpoznávání bylo nutné vytvořit ke všem

nahrávkám fonetické přepisy a tyto poté ještě upravit do podoby srozumitelné pro HTK. Program HTK vznikl na Cambridge University, proto se musí české hlásky rozepsat bez interpunkce (například místo *„á“* se použije *„aa“*). Existují dvě možnosti jak v HTK postupovat při trénování modelů hlásek, které se liší pouze v jednom kroku (viz. obr. 14). V prvním případě máme u fonetických přepisů uložené i časové hranice jednotlivých hlásek (zarovnávání hranic v programu *Aligner*), potom se použije program HInit, který nalezne všechny realizace každé hlásky a iterativně natrénuje jejich parametry. Ve druhém případě časové hranice nejsou určeny, použije se program HCompv. Ten provede tzv. plochý start („Flat Start“) a přes všechny nahrávky určí rozptyly a umístí je do modelů všech hlásek, tím dosáhne počáteční inicializace hodnot parametrů. Po jednom nebo druhém inicializačním kroku následuje v HTK krok reestimační (několik iterací programu HERest). Ten si na základě fonetického přepisu sestaví model celé nahrávky zřetězením všech dílčích hláskových modelů, pro každou nahrávku určí dílčí příspěvek k výpočtu parametrů a toto zopakuje pro všechny nahrávky. V testu bylo zvoleno 20 iterací programu HERest, při použití většího počtu iterací už hrozí tzv. přetrénování modelu. Byl zjišťován rozdíl v rozpoznávacím skóre pro modely s ručně zarovnanými časovými hranicemi (HInit) a modely natrénované s automatickým zarovnáváním (HCompv).

Ukázka fonetické transkripce:

Originál:	skutečně	
Fonetický přepis:	skutečňe	
Transkripce pro HTK:	0 300000	si
	300000 2300000	s
	2300000 3000000	k
	3000000 3700000	u
	3700000 4300000	t
	4300000 5300000	e
	5300000 6900000	ch
	6900000 7800000	nj
	7800000 9000000	e
	9000000 11999999	si

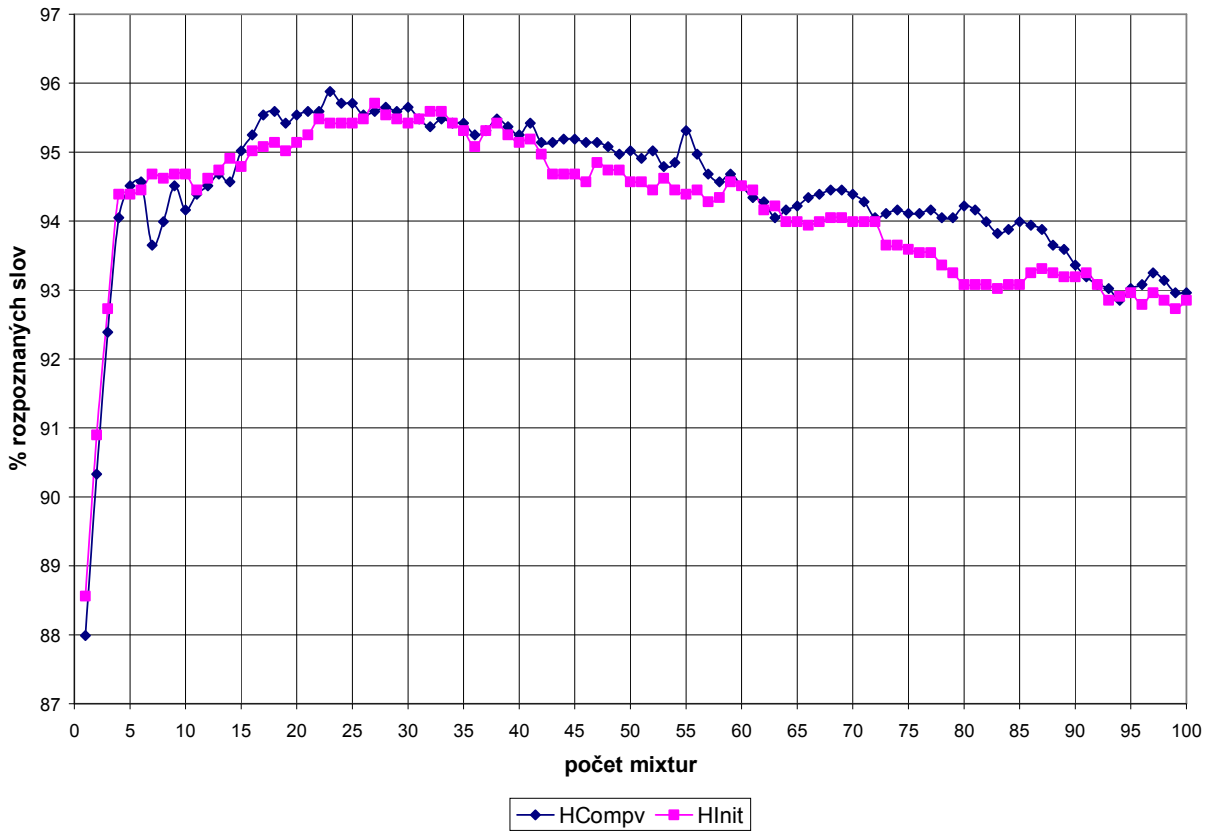


Obr. 14 – Schéma procesu trénování HMM modelů hlásek v HTK

Jak již bylo řečeno výše, v testu byl zjišťován i vliv počtu mixtur v HMM modelech na rozpoznávací skóre. Vícemixturové modely se v HTK trénují pomocí programu HHEd, který vezme stávající modely a rozdělí každou mixturu na zadaný počet mixtur nových [5]. Následně byl opět využit program HERest pro dotrénování parametrů. Počet mixtur byl volen od 1 do 100. U intervalu 1-80 mixtur bylo použito 5 iterací programu HERest, v intervalu 81-100 byly iterace jen 2. Rozpoznávací skóre už se nezlepšovalo a trénování bylo zbytečně velmi časově náročné.

Pro rozpoznávání bylo nutné vytvořit soubory se slovníkem (slovník obsahuje všechna slova i s fonetickým přepisem) a jazykovým modelem (vyjádřen jednoduchou gramatikou typu slovo1 nebo slovo2 nebo slovo3 atd.). V HTK byl pro rozpoznávání použit program HVite (název podle autora klasifikačního algoritmu – Viterbiho algoritmus). Pro vyhodnocování rozpoznávání byl využit program HResults, který

potřeboval vytvořit i tzv. referenční MLF soubor (Master Label File), kde byly uloženy referenční přepisy nahrávek. HResults porovnává tento referenční soubor s výstupním MLF souborem od programu HVite a určuje výsledné rozpoznávací skóre. Zároveň umí pro sekvenci slov určit počet rozpoznaných slov, vynechaných slov, atd. Tyto funkce nebyly využity, protože slovník obsahoval pouze jednotlivá slova.



Obr. 15 – Vliv počtu mixtur na rozpoznávací skóre u ručně zarovnaných časových hranic (HInit) a u automatického zarovnání (HCompv)

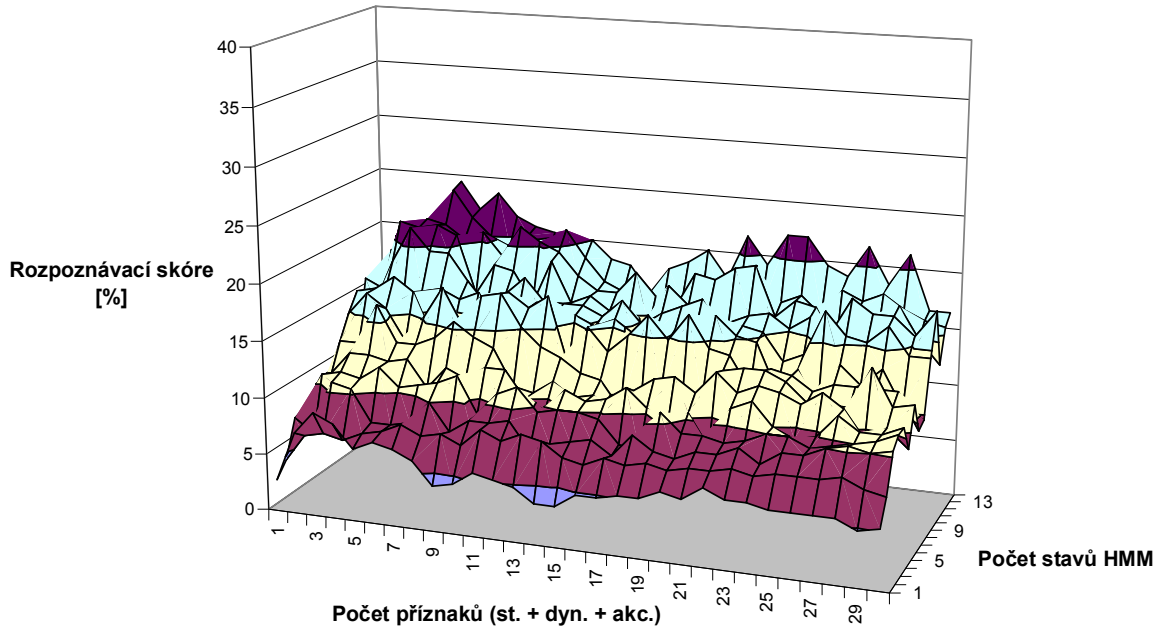
Z grafu je vidět, že nejlepšího skóre se dosáhlo použitím modelů s počtem mixtur přibližně od 20 do 30. Konkrétně potom pro program HCompv bylo nejlepší skóre 95,88% dosaženo pro modely s 23 mixturami a pro program HInit 95,71% pro modely s 27 mixturami. Obě křivky mají velmi podobný průběh, mírně lepších výsledků se dosáhlo použitím automatického zarovnávání. To zřejmě způsobila nedostatečná schopnost zarovnat manuálně časové hranice jednotlivých hlásek pomocí odposlechu. Použití většího počtu mixtur se ukázalo u takto malé databáze jako zbytečné, u větší databáze by zřejmě dosáhlo většího uplatnění.

6.2 Rozpoznávání vizuálního signálu řeči

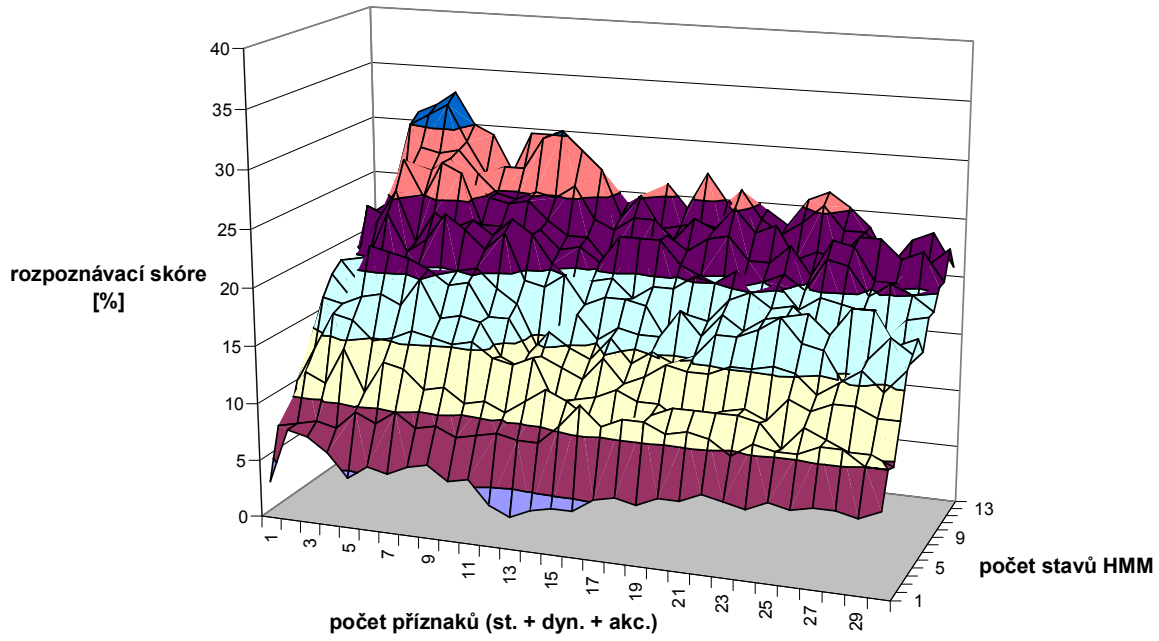
Specifikace testu	
Druh HMM modelů	Celoslovní modely
Počet stavů HMM modelů	1 - 14
Délka framů	33ms
Druh příznaků	DCT + dynamické a akcelerační
Počet příznaků	3 – 90 (1+1+1 , 2+2+2 , …)
Slovník	50 slov
Trénovací databáze	30 mluvčích (1500 slov)
Testovací databáze	5 mluvčích (250 slov)

V tomto testu byla hledána ideální kombinace vizuálních DCT příznaků vedoucí k nejlepšímu rozpoznávacímu skóre. Zjišťován byl vhodný počet stavů HMM modelů a počet příznaků v příznakovém vektoru. Trénovány byly celoslovní modely.

DCT parametrizace (viz. kapitola 2) byla naprogramována v prostředí Matlab, kde byla možnost využít funkci pro spočtení 2D diskrétní kosinové transformace (funkce *dct2*), převodu barevných RGB obrazů s oblastí zájmu ROI na šedotónové (funkce *rgb2gray*) a další operace s obrazem a maticemi. Bylo zjištěno i optimální pořadí jednotlivých kroků v parametrizaci. Tím bylo: vypočtení statických DCT příznaků, jejich normalizace a potom teprve počítání dynamických a akceleračních příznaků. Pro počítání dyn. a akc. příznaků bylo potřeba zvětšit matici statických příznakových vektorů o 2 na začátku. Jako vhodnější se ukázalo volit pro logaritmování energie přirozený logaritmus, desítkový logaritmus lepší skóre nepřinesl. Zlepšení rozpoznávacího skóre nepřinesla ani ekvalizace histogramu snímků s oblastí zájmu (funkce *histeq*). Program dále ukládal příznakové vektory do souborů ve formátu HTK, kde se do hlavičky ukládá například počet framů nahrávky, počet příznaků, druh parametrizace (USER), délka jednoho framu apod. Za hlavičkou se ukládají jednotlivé příznakové vektory frame po framu. Znatelné zlepšení potom přineslo opětovné oříznutí nahrávek, kde nebylo ponecháno ticho před a po skončení promluvy (viz. obr. 16, 17). Testování probíhalo pouze do 14 stavů, pro větší počet již nešlo natrénovat použitelné modely.



Obr. 16 – Vliv počtu stavů HMM a počtu příznaků na rozpoznávací skóre (1. střih)



Obr. 17 – Vliv počtu stavů HMM a počtu příznaků na rozpoznávací skóre (2. střih)

Zlepšení při druhém oříznutí způsobilo zaměření pouze na promluvu. U parametrizace akustické složky probíhá automaticky krok detekce začátku a konce řeči, proto ticho před a po promluvě nevadí (u úlohy rozpoznávání izolovaných slov je naopak žádoucí). U vizuální složky tato prodleva ovšem znamená zbytečné zařazení snímků, kde mluvčí nijak nemění polohu úst. V obou případech se jako nejvhodnější ukázala kombinace 5+5+5 (statické, dynamické a akcelerační DCT příznaky) a HMM modely se 14 stavy. V první části bylo maximální dosažené skóre 24,4% a v druhé 32,8%.

Pokud je k dispozici pouze menší databáze mluvčích, používá se pro dosažení objektivnějších výsledků rotace trénovací a testovací databáze. V diplomové práci potom konkrétně 35 rotací (30 mluvčích pro trénování, 5 pro rozpoznávání). Byly provedeny testy jednotlivých kombinací statických, dynamických a akceleračních DCT příznaků. Použito bylo optimální nastavení, které vzešlo z testu výše (5 maximálních příznaků, 14 stavové HMM modely, nově oříznuté nahrávky).

Kombinace DCT příznaků	**Rozpoznávací skóre po rotaci [%]**
Statické	23,2
Dynamické	20,59
Akcelerační	17,28
Statické + Dynamické	31,86
Statické + Akcelerační	30,33
Dynamické + Akcelerační	25,57
Stat. + Dyn. + Akc.	**33,14**

Tab. 3 – Výsledné rozpoznávací skóre [%] po „rotačních“ testech

Nejlepší skóre měla kombinace všech tří příznakových vektorů, naopak nejmenší dosažené rozpoznávací skóre měl samostatně použitý vektor akceleračních příznaků.

Specifikace testu	
Druh HMM modelů	Modely vizémů
Počet stavů HMM modelů	3
Délka framů	33ms
Druh příznaků	DCT + dynamické a akcelerační
Počet příznaků	5 + 5 + 5
Slovník	50 slov
Trénovací databáze	50 vět od 35 mluvčích (1750 vět)
Testovací databáze	50 slov od 35 mluvčích (1750 slov)

Další úloha se zaměřila na rozpoznávání řeči pomocí vizémů. Zjišťoval se vliv počtu mixtur HMM modelů na rozpoznávací skóre. Pro natrénování HMM modelů se využila obdobná technika jako u fonémového rozpoznávání. Použity byly ručně zarovnané časové hranice uložené ve fonetických přepisech a program HInit. Dále bylo nutno provést transkripci z fonémů na vizémy (viz. tab. 2). Jak již bylo řečeno výše (kapitola 3), touto transkripcí se do časových hranic zavádí menší chyba, protože akustická a vizuální složka nemusí být synchronní. Většinou je jako první zaznamenán zvuk a teprve potom člověk změní polohu rtů. Počet mixtur v modelech byly volen po pěti do 50.

Počet mixtur	1	5	10	15	20	25	30	35	40	45	50
Roz. skóre [%]	7,78	11,9	12,13	11,84	12,3	**12,47**	12,13	11,16	10,81	10,7	10,58

Tab. 4 – Vliv počtu mixtur na rozpoznávací skóre [%] v úloze vizémového rozpoznávání řeči

Nejlepší skóre bylo dosaženo pro modely s 25 mixturami a to 12,47%. Rozpoznávání řeči pouze na základě vizuální složky nedosahuje zdaleka výsledků v rozpoznávání akustického signálu řeči. Hlavní význam má vizuální složka při audiovizuálním rozpoznávání v hlučných podmínkách, jak bude vidět v další kapitole.

6.3 Audiovizuální rozpoznávání signálu řeči

Specifikace testu	
Druh HMM modelů	Celoslovní modely
Počet stavů HMM modelů	14
Délka framů	10ms
Druh příznaků	Kombinace MFCC + DCT
Počet příznaků	54 (39MFCC + 15DCT)
Slovník	50 slov
Trénovací databáze	30 mluvčích (1500 slov)
Testovací databáze	5 mluvčích (250 slov)

V testu bylo provedeno audiovizuální rozpoznávání řeči. Trénovány byly celoslovní modely se 14 stavy. Audiovizuální příznaky vznikly fůzí akustických kepstrálních MFCC příznaků (13 statických, 13 delta a 13 delta delta) a vizuálních DCT příznaků (kombinace 5 statických, 5 dynamických a 5 akceleračních). Bylo provedeno rozpoznávání v hlučných podmínkách a pozorován vliv přidání vizuální složky na rozpoznávací skóre.

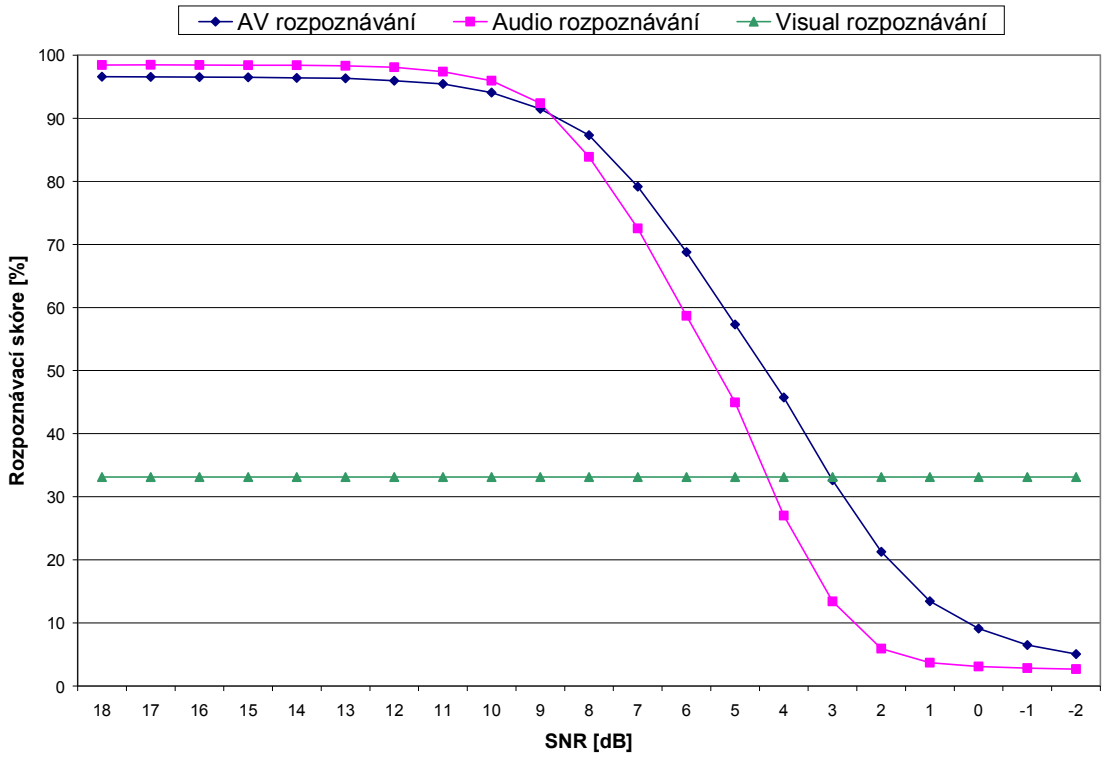
Jak již bylo řečeno v kapitole 4, akustické příznaky jsou získávány s frekvencí 100Hz, zatímco vizuální s frekvencí 30Hz. Prvním krokem tedy byla interpolace vizuálních příznaků na 100Hz, a tím synchronizace s akustickými příznaky. Interpolace byla naprogramována jako součást parametrizace vizuální složky v Matlabu (funkce *interp1*), použita byla lineární interpolace. Fůze příznaků, uložení souborů a vyhodnocení rozpoznávacího skóre bylo naprogramováno v MS Visual C++ studio 2008. Trénovací a testovací databáze se opět v každém kroku nechaly rotovat. Zjišťován byl vliv okolního hluku na rozpoznávací skóre, tento hluk byl simulován přidáváním dvou druhů šumu do akustických nahrávek. Prvním šumem byl tzv. „white“ šum (bílý šum), ten má rovnoměrnou výkonovou spektrální hustotu, signál má tedy stejný výkon v jakémkoli pásmu shodné šířky. Druhým použitým šumem byl tzv. „babble“ šum (řečový šum), ten simuluje ruch přidáním řeči dalších osob, nemá takový stupeň maskování jako white šum. Odstup signálu a šumu SNR (Signal to noise ratio) byl měněn od 18dB do -2dB. Z toho plyne i velká časová náročnost tohoto testu, trénování a

rozpoznávání bylo prováděno v 21 krocích pro SNR a v každém kroku byla ještě 35 krát rotována trénovací a testovací databáze. Pro představu, jedna tabulka (jedna křivka v grafu) byla počítána cca 30 hodin.

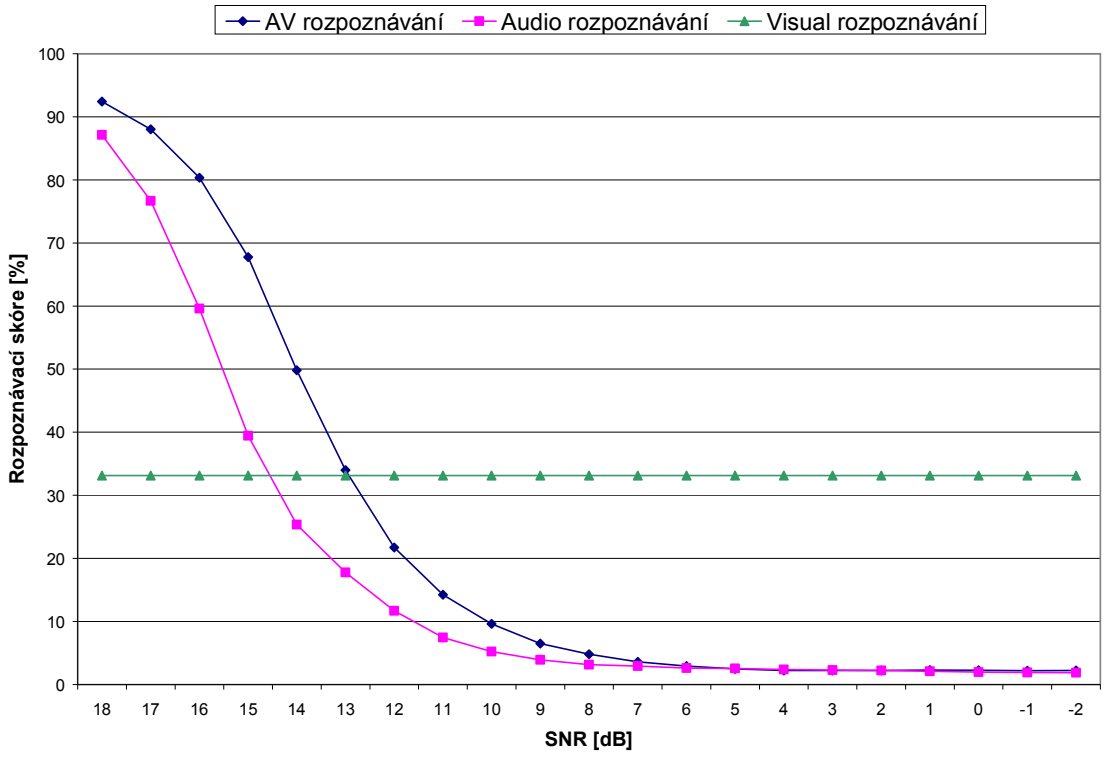
V programu byly konkrétně v každém kroku nejprve aktualizovány slovníky pro parametrizaci akustického signálu a slovníky pro trénování a rozpoznávání, se kterými pracuje balík HTK. Každý úkon HTK (parametrizace akustického signálu, trénování, rozpoznávání) vykonává jeden dávkový soubor, pro které jsou v programu vytvořeny jednotlivé procesy. U audiovizuálního rozpoznávání jsou tedy kroky následující:

- vytvoření souborů pro HTK
- proces parametrizace akustického signálu (HTK)
- fůze akustických a vizuálních příznaků
- proces trénování (HTK)
- proces rozpoznávání (HTK)
- vyhodnocení, uložení do tabulky a zapsání do souboru

Pro srovnání byly do grafů vloženy i výsledky rozpoznávání pouze akustické a vizuální složky řečového signálu (na tu samozřejmě okolní hluk vliv nemá). Z výsledku je patrný přínos audiovizuálního rozpoznávání řeči, přidáním vizuální složky se zmenšuje vliv okolního ruchu na výsledné rozpoznávací skóre. Dle předpokladu se ukázalo, že babble šum, který je při rozpoznávání řeči asi nejpřirozenější, nemá tak silnou schopnost maskovat původní signál. Největší zlepšení přineslo přidání vizuální složky pro SNR 3dB a sice 19,2% (viz. obr. 18). U testu, kde byly nahrávky zarušeny bílým šumem, se vliv šumu ukazoval již od začátku a křivka o dost rychleji klesala směrem k nule. Největší rozdíl mezi rozpoznáváním akustického signálu a audiovizuálním rozpoznáváním řeči lze pozorovat již při SNR 15dB a to velice slušných 28,31% (viz. obr. 19). Číselné hodnoty jednotlivých křivek a spočtené rozdíly jsou zaznamenány v tabulce č. 13.



Obr. 18 – Vliv okolního ruchu na rozpoznávací skóre, přidáván je „babble“ šum



Obr. 19 – Vliv okolního ruchu na rozpoznávací skóre, přidáván je „white“ šum

Závěr

Cílem práce bylo prostudovat metody audiovizuálního rozpoznávání řeči a používané akustické a vizuální příznaky řečového signálu. Dalším krokem byla úprava audiovizuální databáze pro účely rozpoznávání a vytvoření sady programů pro parametrizaci vizuálního signálu, trénování a rozpoznávání HMM modelů řeči. Provedena byla i simulace testů v hlučných podmínkách, řešená přidáním aditivního šumu do akustických nahrávek.

Přes velkou časovou náročnost (nejdéle trvala úprava audiovizuální databáze a testy v hlučných podmínkách) se nakonec podařilo všechny úkoly práce splnit. Nahrávky v databázi byly manuálně zkráceny o oznamovací tón, byly nalezeny i hranice jednotlivých fonémů a vizémů a je zde tedy možnost tohoto využít i do budoucna. Vznikla sada programů v prostředí Matlab pro DCT parametrizaci vizuální složky řečového signálu a v prostředí MS Visual C++ studio 2008 ve spolupráci s balíkem HTK pro audiovizuální rozpoznávání řeči. Rozpoznávání akustické složky signálu řeči je dnes již na velice dobré úrovni, v testu se dosáhlo rozpoznávacího skóre 95,88% pro fonémově orientované rozpoznávání a 98,47% pro rozpoznávání izolovaných slov (viz. kapitola 6.1 a 6.3). Rozpoznávání pouze na základě vizuální složky se příliš nepoužívá, v testech se dosáhlo skóre 12,47% pro vizémově orientované rozpoznávání a 33,14% pro rozpoznávání izolovaných slov (viz. kapitola 6.2). Hlavní přínos vizuální složky je vidět u audiovizuálního rozpoznávání v hlučných podmínkách. Při použití okolního babble šumu (hluk vytvořený z překrývající řeči, který se může standardně v místnosti vyskytnout) se skóre oproti akustickému rozpoznávání zlepšilo až o 19,2%. U použití bílého šumu (šum, který má rovnoměrnou výkonovou spektrální hustotu) byl pozorovaný rozdíl až velice slušných 28,31% (viz. kapitola 6.3).

Použitá literatura

[1] KOLEKTIV AUTORŮ, EDITOR NOUZA J., KOLDOVSKÝ Z., VÍCH R. : Sborník článků – principy hlasové komunikace, úlohy, metody a aplikace. Ve vydavatelství Technické univerzity v Liberci, Česká republika, Liberec, 2009, ISBN 978-80-7372-548-8

[2] 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

[3] CHALOUPKA J. : Rozpoznávání akustického signálu řeči s podporou vizuální informace, Disertační práce, Technická univerzita v Liberci 2005

[4] CÍSAŘ P. : Využití metod odezírání ze rtů pro podporu rozpoznávání řeči, Disertační práce, Západočeská univerzita v Plzni, fakulta aplikovaných věd, 2006

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

[6] Microsoft Developer Network, *http://msdn.microsoft.com*

[7] The MathWorks, MATLAB Documentation, *http://www.mathworks.com*

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

Příloha č.1 – Tabulka ke kapitole 6.1 (Tab. 5)

	Rozpoznávací skóre [%]			Rozpoznávací skóre [%]	
Počet mixtur	Hcompv	Hinit	**Počet mixtur**	Hcompv	Hinit
1	87,99	88,56	**51**	94,91	94,57
2	90,33	90,9	**52**	95,02	94,45
3	92,39	92,73	**53**	94,79	94,62
4	94,05	94,39	**54**	94,85	94,45
5	94,51	94,39	**55**	95,31	94,39
6	94,57	94,45	**56**	94,97	94,45
7	93,65	94,68	**57**	94,68	94,28
8	93,99	94,62	**58**	94,57	94,34
9	94,51	94,68	**59**	94,68	94,57
10	94,16	94,68	**60**	94,51	94,51
11	94,39	94,45	**61**	94,34	94,45
12	94,51	94,62	**62**	94,28	94,16
13	94,68	94,74	**63**	94,05	94,22
14	94,57	94,91	**64**	94,16	93,99
15	95,02	94,79	**65**	94,22	93,99
16	95,25	95,02	**66**	94,34	93,94
17	95,54	95,08	**67**	94,39	93,99
18	95,59	95,14	**68**	94,45	94,05
19	95,42	95,02	**69**	94,45	94,05
20	95,54	95,14	**70**	94,39	93,99
21	95,59	95,25	**71**	94,28	93,99
22	95,59	95,48	**72**	94,05	93,99
23	95,88	95,42	**73**	94,11	93,65
24	95,71	95,42	**74**	94,16	93,65
25	95,71	95,42	**75**	94,11	93,59
26	95,54	95,48	**76**	94,11	93,54
27	95,59	95,71	**77**	94,16	93,54
28	95,65	95,54	**78**	94,05	93,36
29	95,59	95,48	**79**	94,05	93,25
30	95,65	95,42	**80**	94,22	93,08
31	95,48	95,48	**81**	94,16	93,08
32	95,37	95,59	**82**	93,99	93,08
33	95,48	95,59	**83**	93,82	93,02
34	95,42	95,42	**84**	93,88	93,08
35	95,42	95,31	**85**	93,99	93,08
36	95,25	95,08	**86**	93,94	93,25
37	95,31	95,31	**87**	93,88	93,31
38	95,48	95,42	**88**	93,65	93,25
39	95,37	95,25	**89**	93,59	93,19
40	95,25	95,14	**90**	93,36	93,19
41	95,42	95,19	**91**	93,19	93,25

42	95,14	94,97	**92**	93,08	93,08
43	95,14	94,68	**93**	93,02	92,85
44	95,19	94,68	**94**	92,85	92,91
45	95,19	94,68	**95**	93,02	92,96
46	95,14	94,57	**96**	93,08	92,79
47	95,14	94,85	**97**	93,25	92,96
48	95,08	94,74	**98**	93,14	92,85
49	94,97	94,74	**99**	92,96	92,73
50	95,02	94,57	**100**	92,96	92,85

Tab. 5 – Vliv počtu mixtur na rozpoznávací skóre u ručně zarovnaných časových hranic (HInit) a u automatického zarovnání (HCompv) v úloze fonémově orientovaného rozpoznávání akustického signálu řeči

Příloha č.2 – Tabulky ke kapitole 6.2 (Tab. 6, Tab. 7 a Tab. 8)

Počet příznaků	Počet stavů HMM modelů													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
3	2,8	4	7,2	7,6	7,2	7,6	8	7,2	9,2	10,4	16	12,8	12,4	13,6
6	7,2	6,4	6	10,8	10,4	12,8	14,8	16,4	13,2	14,4	15,2	18,4	16	16,4
9	9,2	6,8	10,8	11,2	12,4	12,8	16,8	16,8	16	18,4	21,6	19,2	19,6	20
12	8,4	6,4	10,4	10,8	12,4	16,4	15,6	18	15,6	21,6	21,2	21,6	19,2	20,4
15	6,4	7,2	7,6	12	12	13,2	12,8	16,8	16,4	19,6	21,6	21,2	24	24,4
18	7,2	7,6	7,6	10,4	12,4	13,2	15,2	12,8	16,8	19,6	20,8	20	21,6	22
21	6,8	8,4	8	10,8	10	14	18	13,2	16	17,2	15,6	20,8	20	23,6
24	6	7,6	10	11,2	12	15,6	14,8	13,2	16,8	15,2	18,4	18,8	19,2	21,6
27	4	8	7,6	12	11,2	14,4	17,6	15,6	14	16,4	18	21,6	20,4	20,8
30	4,4	7,2	8,8	8,8	12,4	13,6	18	17,6	14,8	18	15,2	20,4	17,6	20,4
33	5,6	8,8	8,8	10,8	12,4	10,4	15,2	16,4	16,4	16	19,2	19,6	21,2	19,2
36	5,2	8	9,2	11,6	10	10	13,6	17,6	15,6	16	15,6	17,6	18,8	20
39	4,8	8,8	6,8	9,6	9,2	11,2	12,8	11,6	15,6	16	16,4	16,4	16,8	18,4
42	3,6	8,4	7,6	10	11,2	10	11,6	15,2	14,8	15,2	15,2	16,4	16	18
45	3,6	7,2	8	10	10,4	10,4	11,2	13,2	15,6	16,8	15,2	18,4	15,2	15,6
48	4,8	6,8	8,4	9,6	10,8	9,2	8,8	14,8	14,4	14,8	14	16,4	14,8	18
51	4,8	7,6	7,2	7,6	12	11,2	9,2	16,8	14,8	16	18	15,6	15,6	18,8
54	5,2	7,2	9,2	7,2	10,4	12,4	10,8	14,8	15,2	16,8	16,4	18,8	15,6	20
57	5,2	7,6	8	10,8	9,2	12,8	9,2	17,2	15,2	17,6	16	18,4	15,2	17,6
60	6	7,2	7,6	9,2	10,4	11,6	10,4	14,4	16,8	14,8	15,2	19,6	18	21,6
63	5,6	7,6	8,4	9,2	7,6	13,6	13,2	14	15,2	15,6	16	20	15,6	19,2
66	6,8	7,6	8,4	8,8	11,2	12	13,6	11,2	14,8	15,2	16,4	15,6	16,4	22
69	6	8	9,2	9,2	12	12,8	13,6	13,2	10,8	12,8	15,2	17,2	17,2	22
72	6	7,6	8	10	10	12,4	12,8	12,8	12,8	13,6	16,4	16,8	15,6	19,6
75	5,6	8	9,2	10	9,2	11,6	11,2	11,6	9,6	17,6	15,2	13,6	18	18,8
78	5,6	8	9,2	9,2	9,6	12,8	11,2	12,4	13,6	15,6	18,4	16,8	17,2	21,6
81	5,6	8	10	9,6	11,6	11,6	11,2	12,4	12	15,2	17,2	14,8	18	18,4
84	5,6	8,4	10,4	10	10,4	9,2	14,8	12,4	13,6	16,8	17,2	17,6	16,8	21,2
87	4,8	7,6	10	10,4	10,4	10,4	11,6	12,4	10,4	14,4	15,6	14,8	16	16,4
90	5,2	7,2	10,4	11,2	11,2	8,8	12	9,2	9,6	14,4	18	12,8	15,6	16,4

Tab. 6 – Vliv počtu stavů HMM a počtu DCT příznaků na rozpoznávací skóre [%] v úloze rozpoznávání vizuálního signálu řeči (1. střih nahrávek)

Počet příznaků	Počet stavů HMM modelů													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
3	3,2	7,6	7,2	10,4	12	9,6	9,6	12,8	14	13,6	18,8	21,6	20	20,8
6	8	8	12	13,6	17,6	19,2	20	19,2	19,6	23,6	22,8	21,6	23,6	24,8
9	7,6	8,4	11,2	16,8	17,2	17,2	18	20	22,8	22,8	24	24	30	30,8
12	6,4	8	14,4	16,4	18	22	19,6	24	25,6	28	27,2	28,4	31	31,6
15	4,4	9,6	10,8	14,8	18,4	21,6	18,4	20,8	21,2	24,8	26,8	28	32	32,8
18	5,6	8,8	13,6	15,2	17,6	21,2	20	22	24	28	28	28,4	28,8	30
21	5,2	10	13,2	16	20	20,4	17,6	24,4	22,8	27,2	24	26,4	26	29,2
24	6	8,8	11,6	15,6	16,8	19,2	20	23,2	23,6	22,8	24,8	24,8	26,8	26
27	6,4	9,2	12	15,2	17,2	19,6	20,4	21,6	22,4	22,8	24,4	24,4	26,8	29,6
30	5,2	8,4	10,8	15,6	16,8	16,8	18,8	23,2	24	21,2	24,8	25,2	29,6	29,6
33	5,6	8,8	11,2	14,4	14	16,8	17,6	19,2	22,8	22,8	23,2	24,8	30,4	28
36	3,6	9,2	10,4	14	13,2	14	14	18,8	20,4	21,2	22	26	28,8	24,4
39	2,8	8,8	11,6	14,8	14,4	16	14,4	20	18,8	21,6	22	23,2	27,2	22,8
42	3,6	8,8	12,8	11,2	13,6	14,8	17,2	17,6	19,2	20,4	20	22,4	23,2	24
45	4	10	12	11,6	12,8	14	15,2	18,4	19,6	22,8	19,6	23,6	25,2	24
48	4	12	10,8	12	14,8	16	16,4	18,4	18	22,8	18,4	23,6	22,4	26
51	5,2	11	9,2	12,8	12,8	14	16	18	19,2	22,8	22,8	22	22,8	24
54	5,6	12	9,2	12	13,6	15,2	16	15,2	19,2	20,8	21,6	24	23,2	27,2
57	5,2	12	11,6	12	14,8	14,4	18,4	17,6	20,4	20	22,8	23,2	22,8	24,8
60	6	11	12,4	10,8	14,4	15,2	16,4	17,2	20	17,6	23,2	25,2	26,4	23,6
63	6	11	12,4	10,8	14,4	17,2	18,4	19,2	21,2	19,6	21,2	23,2	24	24,4
66	6,8	11	12,4	12	15,2	17,2	17,2	20	18	20	21,2	24	22,8	23,2
69	6,4	9,6	12,4	12,4	15,2	18	18,4	18,8	20,4	20,8	22,8	21,2	24	25,6
72	6	11	12	11,6	15,6	19,2	18,8	19,2	18,4	19,6	20,4	22	25,6	26,4
75	6,8	10	10	13,6	15,2	17,2	19,2	17,6	19,6	22,4	23,6	21,6	23,6	25,2
78	6,4	10	12	12,8	13,6	16,8	17,6	16,8	19,2	22	23,2	23,6	23,6	23,6
81	6,8	11	11,2	13,6	14,8	14	20	16,8	18,8	19,6	20,4	23,6	22	20,4
84	6,8	11	10,4	13,6	16	16,4	20	16,8	16,4	20,8	20,4	22	20	22,8
87	6,4	10	10,4	12,8	14,4	18,4	17,2	14,4	18,4	20	18,8	22,8	21,2	23,6
90	7,2	10	9,6	12,8	16,4	17,2	18	16,4	18,8	18,8	20,4	20,4	22,8	20,8

Tab. 7 – Vliv počtu stavů HMM a počtu DCT příznaků na rozpoznávací skóre [%] v úloze rozpoznávání vizuálního signálu řeči (2. střih nahrávek)

Rotace číslo	Kombinace použitých DCT příznaků						
	S	D	A	S+D	S+A	D+A	S+D+A
1	20,8	17,6	12	27,2	27,6	20,8	32
2	19,6	17,2	16,4	28,8	29,2	23,2	30,4
3	20,8	21,6	20,8	33,2	30,4	26,8	35,2
4	22	25,6	21,2	36,4	35,2	27,6	39,6
5	26,4	23,2	23,6	41,2	33,2	28,8	40
6	26,4	22,4	24,8	39,2	39,6	32,4	40,8
7	30	25,6	21,2	37,2	34,4	34	42,8
8	32,4	24,4	22,8	39,2	38,8	34,8	42
9	24,8	18,8	21,2	34,4	33,6	30,8	34
10	21,6	17,6	15,6	30	28,8	26	29,2
11	18,8	20	13,6	26,8	24,8	21,6	27,6
12	15,6	18	12,8	25,2	25,6	21,6	27,6
13	14,8	12	10,4	20,8	18,8	19,6	24
14	14	12,4	6,8	22,8	19,6	16	22,8
15	16	12,8	10	20,8	18,8	15,2	22,8
16	13,6	12,4	10,8	23,2	18,8	17,2	22,4
17	20	16,8	12,8	21,6	22,4	18,8	25,2
18	20	16,8	14,4	26,4	26	20,8	28,4
19	20,8	14,8	17,6	28	28,8	23,6	28
20	20,4	23,2	18,8	34	33,2	28	38,4
21	27,6	26,8	22,8	40	40	34,4	41,2
22	28	22,4	20	37,6	39,2	28,8	41,2
23	25,6	23,6	20,8	32,8	34	28,4	36
24	26	25,6	18,4	35,6	35,2	34,8	38,8
25	28	24,8	21,2	39,2	38,8	28	38,4
26	26	21,2	20,4	34	33,6	23,2	36,4
27	22,8	20	17,2	31,6	28,4	23,6	32,4
28	26	22,8	15,2	33,2	28,4	24	32,8
29	25,6	20,8	16,4	34,4	30,8	24,8	31,2
30	26,8	22,4	18	33,2	30	28,4	31,2
31	24	24,8	19,6	32,8	30,8	26,4	32,4
32	29,2	27,6	18,8	36	35,2	29,2	37,6
33	28,8	24,8	16,8	34	32,8	27,2	34,4
34	23,6	18,8	19,2	32	30,4	22,8	31,6
35	25,2	21,2	12,4	32,4	26,4	23,2	31,2
Skóre po rotaci	23,2	20,59	17,28	31,86	30,33	25,27	33,14

Tab. 8 – Výsledné rozpoznávací skóre [%] u rotačního testu v úloze rozpoznávání vizuálního signálu řeči pro různé kombinace vizuálních DCT příznaků (S – statické, D – dynamické, A – akcelerační)

Příloha č.3 – Tabulky ke kapitole 6.3 (Tab. 9, 10, 11, 12, 13)

rotace číslo	odstup signál-šum SNR [dB]																				
	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	-1	-2
1	95,6	95,6	95,6	95,6	95,6	95,6	95,6	94,8	93,6	89,2	84	74	62,8	49,6	41,6	29,2	20,8	12,4	6,8	4,4	2,8
2	96,8	96,8	96,8	96,8	96,8	96,8	96,4	96,4	94	90	86,8	76,4	64,4	52	41,2	28,4	18,4	12,8	6,4	3,6	2,8
3	95,6	95,6	95,6	95,6	95,6	95,6	95,6	95,2	94,8	92	87,6	78,8	67,2	55,2	40,4	26,8	16,8	10	5,6	4	2,8
4	96	96	95,6	95,6	95,6	95,2	95,2	94	92	88,8	85,2	78	67,2	53,2	40,8	29,6	19,6	14,4	9,6	6,4	4,8
5	96	96	96	96	95,6	95,6	94,8	94,4	92,4	89,2	85,2	75,6	64,8	53,2	39,6	29,2	17,2	8,8	6,8	4	2,8
6	95,6	95,6	95,6	95,6	95,6	95,6	94,8	94	90,8	90	87,2	75,6	64,8	54	42	30	17,2	10	5,6	4	3,6
7	96,4	96,4	96,4	96,4	96,4	96,4	96	95,6	95,2	92,4	90,8	80,8	70,8	57,2	46,4	30,8	20,8	10,8	6	4,4	3,6
8	96,8	96,8	96,8	96,8	96,4	97,2	96,8	96,4	94,8	92,8	90,8	82,4	72,4	60	49,6	38,8	27,6	19,6	13,6	9,2	8
9	96,8	96,8	96,8	96,8	96,4	96,4	96,4	96	94,4	93,6	90,8	84,4	76,4	63,2	53,2	37,6	27,2	19,6	13,2	9,6	8
10	98,8	98,8	98,8	98,4	98,4	98,4	98,4	98	96,8	96	93,2	88	79,6	65,2	56	40,4	25,2	19,6	14,4	10,4	9,2
11	98	98	98	98	98	98	97,2	97,2	96,4	96,4	93,6	88,4	80,8	68	56	40,8	24	13,2	8,8	5,6	4,8
12	98	98	97,6	97,6	97,6	97,6	97,6	97,2	97,2	94	90	84,4	76,8	64	52	38,8	25,6	16	11,6	9,2	6,8
13	96,8	96,8	97,2	97,2	97,2	96,8	96,8	96,8	96	91,6	88	85,2	74	62,8	51,2	42,4	26	15,2	10	7,2	6,8
14	97,6	97,6	97,6	97,6	97,6	97,6	97,6	97,2	95,6	93,6	87,6	82,8	68,4	59,2	46,8	32,4	16,8	9,2	6,4	3,6	3,2
15	97,6	97,6	97,6	97,6	97,6	97,2	96,8	96	95,6	93,2	90	82,8	70,8	60	46,8	32,4	18,4	10	6	4,4	2,8
16	97,6	97,6	97,6	97,6	97,6	97,6	97,6	97,2	96,4	95,2	91,2	84,4	74	63,2	49,6	34,4	21,2	15,2	9,6	6,8	5,6
17	97,2	97,2	97,2	96,8	96,8	97,2	97,2	97,2	96,8	95,6	88,4	80,4	70	58	43,2	28	16,8	9,6	6,8	4,4	3,6
18	96,8	96,8	97,2	97,2	97,6	97,2	96,8	96,8	96,4	94,4	89,6	80,8	69,2	58	43,6	32	16,4	10,8	6,4	5,2	4,4
19	98,4	98,4	98,4	98,4	98,4	98,4	98	97,6	96	92,8	89,2	80,8	67,2	56,8	43,2	29,6	18	12	8,8	6,8	4,8
20	97,6	97,6	97,6	97,6	97,2	97,6	96,8	96	95,2	91,6	85,6	74,4	65,6	54,4	42,8	28,8	20,8	13,6	9,6	7,2	5,6
21	97,6	97,2	97,2	97,2	97,2	97,2	96,8	96	94,4	90,4	84,4	74	65,2	53,2	45,2	30	18,4	12	6,8	5,2	4,4
22	98,4	98,4	98,4	98,4	98,4	98,4	98	97,2	96,4	93,6	87,2	77,6	66	55,2	42,8	28,8	17,6	10,4	7,2	4,8	3,2
23	97,2	97,2	97,2	97,2	96,8	96,8	96	96	94,4	92,4	88,4	78,8	66,4	58	48	33,2	18,8	9,2	6,4	3,2	2,4
24	96,8	96,8	96,8	96,8	96,8	96,8	96,8	96,4	95,6	93,6	91,6	85,6	70,4	61,2	50,4	34,4	26,8	16	13,6	9,6	8,4
25	97,6	97,6	97,6	97,2	97,2	96,8	96,4	96	95,2	93,2	90,8	84,8	73,6	66,4	55,2	38,8	28,8	19,2	12,4	7,6	4,8
26	92,4	92,4	92,8	92,8	92,8	92,4	91,6	91,6	90,4	88	84,4	77,2	66	58,4	50	36	26,8	18,4	14,4	10	8,4
27	93,6	92,8	92,4	92,4	92,4	92	90,8	90,4	90	87,6	82	74	66,4	55,2	46,4	35,6	24,8	17,6	13,2	11,6	8,4
28	92,4	92,4	92	92	92	91,6	91,2	90,4	88,4	84,8	80,8	74,8	66,4	54,8	46,8	31,2	22,4	17,6	11,2	9,2	6,4
29	92,8	92,8	92,8	92,8	92,4	92,4	92	92	89,2	85,6	81,2	74	67,6	58	45,6	34,4	25,6	17,6	12,8	8,4	7,6
30	94,4	94,4	94,4	94	94	93,6	93,2	92	90,8	88	84	74	65,6	53,6	40,8	29,2	21,2	14,4	8,8	7,2	6
31	98,4	98,4	98	98	98,4	98,4	97,2	96	93,6	90,4	85,2	76,8	66	52	40	29,6	20	12	10	8,8	8
32	98,4	98	98	98	97,6	97,6	96,8	96	95,2	92	87,2	75,6	66,8	55,6	43,6	31,6	20,4	11,6	8,8	7,2	2,8
33	98	98	98	98	97,6	97,6	96,8	96	94,4	91,6	87,2	78,4	67,2	53,6	41,2	30	21,2	11,2	6,8	4,8	3,6
34	97,6	97,6	98	98	98	97,6	97,6	96	94,4	92	85,6	75,2	65,2	53,2	39,6	29,6	20,4	11,6	7,6	6	2,8
35	95,2	95,2	95,2	94,8	94,8	94,8	94,4	94	90	87,2	81,6	72	61,6	50,8	39,2	28,8	17,2	8,8	7,2	4	3,2
skóre	96,59	96,55	96,54	96,48	96,41	96,34	95,94	95,43	94,08	91,51	87,33	79,18	68,79	57,33	45,74	32,62	21,29	13,44	9,12	6,51	5,06

Tab. 9 – Rozpoznávací skóre [%] u rotačního testu v závislosti na SNR, jedná se o úlohu audiovizuálního rozpoznávání řeči, přidáván je „babble“ šum, na posledním řádku je výsledné zprůměrňované skóre

rotace číslo	odstup signál-šum SNR [dB]																				
	18	**17**	**16**	**15**	**14**	**13**	**12**	**11**	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**0**	**-1**	**-2**
1	92	85,6	77,2	62,4	42,8	31,2	20,4	14	12,8	8	5,6	4	2,4	2	2	2	2	2	2	2	2
2	92,8	88,8	79,6	64,8	50,8	37,2	24	16	10,8	6,8	5,2	4,4	3,2	2,4	2,8	2,4	2,4	2,4	2,4	2,8	2,8
3	93,6	88,4	79,6	64,8	48	34,8	22,4	16	11,2	6,8	5,6	4,4	3,2	2,8	2,8	2,4	2,4	2,4	2,4	2,4	2,8
4	88	82,8	74,8	64,4	49,6	31,2	20	15,2	10	6,8	4,8	4	4	4	3,6	2,8	2,8	2,8	2	2	2
5	90,8	83,6	76,8	65,2	48	32	20,8	13,6	9,2	6,4	6	3,6	3,2	3,2	3,2	3,2	3,2	3,2	3,2	3,2	3,2
6	87,6	81,6	75,2	62	43,6	31,2	19,2	13,2	10,4	8,4	7,2	6,8	4,4	4,8	4	4,8	4,4	3,2	2,8	2,8	2,4
7	91,2	85,2	77,6	64,4	47,2	26,8	18,8	12,4	9,2	6,8	6,4	5,2	3,2	2,4	2	2	2	2	2,8	2,8	2,8
8	90,8	85,6	77,2	66	44	27,6	18,4	12,8	10	8,4	6,4	4,8	4,4	2,4	2	2	2	3,2	2,4	2,8	2,4
9	93,2	89,6	82,4	72,8	54,4	36	25,6	18,4	12,8	8,8	6,4	6	4,8	4	2,8	2,8	2,4	2	1,6	1,6	1,6
10	96	93,6	87,2	76,4	59,6	38,4	26,4	19,6	13,2	9,2	4,8	3,6	2,4	2,4	2,4	2,4	2	2	2	2	2
11	95,6	93,6	89,6	80,4	65,2	44	32,4	22,8	14,4	9,2	6	4,4	2,8	2	1,2	1,2	1,6	2	2	1,6	1,2
12	95,2	92	84,8	74,8	59,2	43,6	28,8	22,4	16	10,4	8	5,6	4	2	1,6	2,4	2	2,4	2,8	2,4	2,8
13	94,8	90,8	85,2	74,4	59,6	45,6	30	22	14,8	8,8	6	3,6	3,2	2	2	2	1,6	2	2,4	2	2,8
14	93,2	89,6	82,8	71,6	54	40	25,2	17,6	10,8	6,4	5,2	3,6	2,4	1,2	1,6	2	2	2,4	2,4	1,6	2
15	93,6	88	82,8	72	57,6	40,4	26,8	14,8	11,6	4,8	4	2,8	1,6	2	2,4	2,4	2	2	2	2	2
16	93,6	90	82,8	70,8	56,8	40	22	15,2	10	5,6	3,6	2,8	2,8	2,4	2	2	2	2	2	2	2
17	92,4	86	77,2	64	48	31,2	16	7,2	4,4	2,4	2	2	2	2	2	2	2	2	2	2	2
18	92	84,8	77,6	63,6	44,4	29,6	14,8	8,4	6	4,8	3,2	2,4	2,4	2,4	2	2	2	2	2	2	2
19	91,6	87,6	78,8	62,8	42,4	26,8	16,8	9,6	5,2	2	2	2	2	2	2	2	2	2	2	2	2
20	92	86,8	76	59,6	42,8	28	17,2	9,6	6,4	4,4	3,2	2	2	2	2	2	2	2	2	2	2
21	93,2	87,2	76,4	61,2	44	28,4	15,6	10,4	6,4	4,4	2,8	2	1,6	1,6	2	1,6	1,6	2	2	2	2
22	94,4	91,6	82,4	65,6	46,8	32	19,2	11,2	8,8	7,2	5,2	5,2	4,4	3,2	2,8	2,8	3,2	2,8	3,2	3,2	3,6
23	94,4	91,6	83,6	71,6	46,4	34	22	14,4	9,6	8	6	3,6	3,6	2,4	2	2	2	2	2	2	2
24	95,2	91,2	83,6	74	55,2	42,4	30,8	17,2	12,4	8	5,6	4	3,2	4,4	3,6	4,4	4,4	4,4	3,6	2,8	3,2
25	95,6	92,8	84	75,6	59,2	45,2	32	18,8	11,2	8	6	2,8	2	1,2	0,8	0,8	1,2	1,6	1,6	1,6	1,6
26	90,8	88	82	71,2	53,2	39,2	27,2	18	8	5,6	5,2	2,4	2,8	2,4	0,8	1,2	2	1,6	1,6	1,6	1,6
27	89,6	85,6	79,6	70	56,4	38,8	24	13,2	8,8	7,2	4,8	3,2	3,2	3,2	2,4	2	2,4	2,4	2,4	2,4	2,4
28	89,2	84	78	72	57,6	44	30,4	18,4	10,8	8	5,2	4,4	4	2,4	2	1,6	2	2	2,4	1,6	1,6
29	88,4	86	80,4	68,8	50,4	34,8	18	11,6	9,6	8	6	5,2	4,8	3,6	2,4	2,8	2,4	2,8	2,4	2,4	2,4
30	90,8	88	81,2	67,6	48,4	30	20,8	12,8	9,2	6,4	4,4	3,6	2,4	2,4	2,4	2,4	2,4	2,4	2,4	2,4	2
31	90,8	88,8	83,6	68,4	46,4	30	14,8	8,8	6,8	4,4	2,8	2,8	2	2	2	2	2	2	2	2	2
32	93,6	90,8	82,4	65,6	45,2	25,6	16	11,6	6,8	4,8	3,6	2,4	2	2	2	2	2	2	2	2	2,4
33	94,8	90	79,6	64,8	42	26,8	17,6	9,6	5,2	4	3,2	2,8	2	2	2	2	2	2	2	2	2
34	92,4	89,2	78,4	62	38,4	22,4	13,6	9,6	6,8	4,4	2,8	2	2	2	2	2	2	2	2	2	2,8
35	91,6	82,4	72	56	36	20,4	12,8	11,6	7,2	4,4	3,2	2	2	2	2	2	2	2,4	2,4	2,4	1,6
skóre	92,42	88,03	80,35	67,76	49,82	33,99	21,74	14,23	9,62	6,51	4,81	3,61	2,93	2,49	2,22	2,24	2,24	2,30	2,26	2,18	2,23

Tab. 10 – Rozpoznávací skóre [%] u rotačního testu v závislosti na SNR, jedná se o úlohu audiovizuálního rozpoznávání řeči, přidáván je „white“ šum, na posledním řádku je výsledné zprůměrňované skóre

rotace číslo	odstup signál-šum SNR [dB]																				
	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	-1	-2
1	99,6	99,2	99,6	99,2	99,2	99,2	99,2	97,6	96,4	92	81,6	68	52,8	40	21,6	10	3,6	3,2	2	2	2
2	99,6	99,6	99,6	99,6	99,6	99,2	98,8	98,8	97,6	93,2	84,8	66,8	54,8	39,6	22,8	15,6	8	4	3,2	2,8	2
3	98,4	98,4	98,4	98,4	98,4	98,4	98	97,2	96,4	93,2	83,6	70	57,2	42,4	28,4	18,4	10	4,8	4	2,8	2
4	96,4	96,4	96,4	96,4	96	95,6	94,8	93,2	91,6	87,2	76,4	62,8	50,4	36,4	23,6	14,4	8,8	4,8	4	3,6	2,8
5	94,4	94,4	94,4	94	93,6	92,8	92,8	91,2	88,8	84,4	74,4	65,6	52	38	27,2	16	6	4,4	3,6	2,8	2,4
6	94	94	94	93,6	94	94,4	94	92,4	90,8	86,8	74,4	64	50,8	36,8	22,8	13,6	7,6	4	2,8	2,8	2,4
7	94,8	95,2	95,2	96	96	96	95,2	92,8	90	85,6	75,6	65,6	52,8	37,6	23,2	12,8	6,4	5,2	4	3,6	2,8
8	94,4	94,8	94,8	94,8	95,6	94,8	94,4	92,8	90,8	88,8	79,2	70,8	55,6	43,6	26	11,6	4,8	3,2	3,6	3,6	2,4
9	98,8	98,8	98,4	98,4	98,4	98,4	98	98,4	98	94,8	90,4	80,4	66	51,2	36	18	9,6	6	4	4	2,8
10	100	100	100	100	100	100	100	100	98,8	96	91,6	81,2	67,6	57,6	35,2	16,4	7,2	2,8	2,8	2,4	2,8
11	100	100	100	100	100	100	100	100	98,8	96,8	92,4	85,2	69,2	58,4	43,2	20	7,6	4	4,4	3,2	2,4
12	99,6	99,6	99,6	99,6	99,6	100	100	100	98,8	97,6	90,4	81,6	68	57,6	41,6	19,2	7,2	4	3,2	2,8	2,4
13	99,2	99,2	99,2	99,2	99,2	99,2	99,6	99,6	98	96,4	90	81,6	66,8	54	36,8	18,4	7,2	3,2	2,4	2,8	3,6
14	98	98	98	98	97,6	98	98,4	98,8	96,8	94,8	87,2	76,8	61,2	48	28,8	10	4	2,4	2	2	2
15	98	98	98	98	98	97,6	98	98,8	98,4	97,2	90	79,6	62	49,2	33,6	16	8	4	2,8	2	2,8
16	98	98,4	98	98	98	98	98	98	97,6	96	89,2	78,8	62	48,8	29,2	16,4	4,8	2,8	2	2	2
17	98,4	98	98	98	98	98	98	97,2	95,2	93,2	84,8	72,8	60	45,6	30	16,8	8	6,4	4,4	4,4	4,4
18	98,8	99,2	98,8	98,8	98,8	98,4	98,4	97,6	96,8	92	83,2	72	61,2	42,4	28	16,4	8,4	5,2	4	3,6	4
19	99,6	99,6	99,6	99,6	99,6	99,2	99,2	98	96,8	92,8	84	75,2	60	44,4	28,8	16	7,2	5,6	4,4	4,4	3,6
20	98,4	98,4	98,4	98,4	98,4	98,4	98	96,8	94,8	90,8	80	67,2	52,8	37,6	20	7,2	2	2	2	2	2
21	98,4	98,4	98,4	98,4	98,8	98,4	98,4	96,8	94,4	90,4	78,4	66,4	50,4	36,8	17,6	6,8	2,4	2	2	2	2
22	98,8	98,8	98,8	98,8	98,4	98,4	98	96,8	94,8	91,6	81,2	68,4	52,4	40	18,8	6	2,4	2	2	2	2
23	98,4	98,4	98,4	98,4	98	98	97,6	95,2	93,6	90,4	81,6	69,6	55,6	42,4	25,6	8,4	3,2	2,4	2,4	2,4	2,4
24	97,6	97,6	97,6	98	98,4	98,4	97,6	97,2	95,2	92	83,2	72	58	45,6	25,6	12,4	3,6	2,4	2,4	2,4	2,8
25	98,8	98,8	98,8	98,8	98,8	99,2	99,6	99,2	97,2	92,4	88	76,4	61,2	51,6	29,2	13,2	4,4	2,8	2,4	2,4	3,2
26	98,8	98,8	98,8	98,8	99,2	99,2	99,2	98,8	97,6	93,2	86,4	76	65,6	51,6	30,4	16,8	6,8	3,2	2,8	3,6	2,8
27	99,2	99,2	99,2	99,2	99,2	99,2	98,8	98,4	97,6	93,6	86	75,6	63,2	49,6	28,4	16,4	6,8	2,8	2,4	2	2,8
28	99,2	99,2	99,2	99,6	99,6	99,6	99,6	99,6	98	92,8	86,8	74,8	64,8	50	30	20	12	8	6	5,2	4
29	99,2	99,6	99,6	99,6	99,6	99,6	99,6	99,2	98,8	94,8	87,6	74,8	63,6	48	27,6	14,8	7,6	4,8	4,8	4,4	4,4
30	99,2	99,2	99,2	99,2	99,2	99,6	99,6	99,6	98,4	94,4	85,6	76	62	43,2	23,2	8,4	3,2	2	2	2	2
31	99,6	99,6	99,6	99,2	99,2	99,2	98,8	98,8	96	93,2	83,6	71,6	58	41,2	23,2	10,8	5,2	4	3,6	3,6	3,6
32	100	100	100	99,6	99,6	99,6	99,2	98,4	97,2	92,4	84	72,8	57,6	45,2	25,2	12,8	6	4,4	3,2	2	2
33	100	100	100	100	100	99,6	98,4	98	97,2	92,8	84,4	69,2	54,4	41,2	20	6,8	2,8	2	2	2	2
34	100	100	100	100	99,6	98,8	98	96,8	96	90,4	80	65,6	52,8	40,4	18	6,8	2,4	2	2	2	2,4
35	99,6	99,6	99,6	99,2	99,2	98,8	98,4	96,8	94,8	90	76	63,6	51,2	37,6	16,4	6	2,4	2	2	2	2
skóre	98,43	98,47	98,45	98,42	98,42	98,32	98,10	97,39	95,94	92,40	83,89	72,54	58,69	44,96	27,03	13,42	5,93	3,68	3,07	2,85	2,69

Tab. 11 – Rozpoznávací skóre [%] u rotačního testu v závislosti na SNR, jedná se o úlohu rozpoznávání akustického signálu řeči, přidáván je „babble“ šum, na posledním řádku je výsledné zprůměrňované skóre

rotace číslo	odstup signál-šum SNR [dB]																				
	18	**17**	**16**	**15**	**14**	**13**	**12**	**11**	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**0**	**-1**	**-2**
1	84,4	75,6	53,6	34,4	21,2	15,6	11,2	8	4	2,8	2,4	2,4	2,4	2,8	2,4	2,4	1,6	1,6	2	2,4	2
2	83,6	72,8	55,2	38,4	29,6	21,6	13,6	8,8	7,6	5,6	4,4	2,4	2	1,6	1,6	2	2	2	1,6	1,6	1,6
3	84	73,6	53,6	38,8	26	16,4	8,8	6	4,4	2,8	4	4,4	3,6	4	2,8	2	2	2,4	2,4	2	2
4	78,4	65,2	47,2	31,6	22	16,4	9,2	5,6	5,6	5,6	4,4	4,8	3,2	2,8	2,4	2	2	2	2	2	2
5	76,4	66,4	52	32,8	24,8	19,2	10,8	6,8	6	5,6	3,6	3,2	2,8	2,8	2,8	2,4	2,4	2	2,4	2	2
6	74,4	60,4	44	28,8	24,8	17,6	11,2	5,2	4	2,8	2,4	2	2	2	2	2	2,4	2	2	2	2
7	75,2	61,2	46,8	30,8	20	14,4	10	6	4	3,2	3,6	3,6	3,2	2,4	2	2	2	2	2	2	2
8	78,8	67,2	52	36	22	16,4	11,2	5,6	4,4	3,6	4	3,6	2	2	2	2	2,4	2	2	2	2
9	91,6	81,6	67,6	50	35,6	24	17,6	10	4,8	3,2	2	3,2	2	2,4	2,8	2,8	2,8	2,8	2	2	1,2
10	95,2	84,8	68,8	49,6	30,4	20,8	14,4	8,8	8	4,4	3,2	2,4	2	2	2	2	2	2	1,6	1,6	2
11	97,2	90,8	76	56,4	36,4	26,4	17,6	10	6,8	5,2	4	2,8	2,4	2,4	3,2	3,2	2	2,4	3,2	2,8	2,4
12	97,2	90,4	81,2	63,2	45,6	30	18,8	12,8	6	3,6	1,6	1,6	1,6	2,8	2	2	2	2	1,6	1,6	1,6
13	96,4	91,6	79,2	61,2	44	30	18,4	9,6	6	3,6	3,6	3,6	3,2	2,8	2,8	2	2	2	2	2	2,4
14	92,8	83,6	71,6	51,6	34,8	21,6	14	8,8	5,6	4,4	3,6	3,2	2,8	3,6	2	2	2	2	2	2	2
15	93,2	86	73,6	56	40,8	27,2	16,8	9,2	5,6	3,6	3,2	2,8	2	2	2	2,4	2,8	2,8	2	2	2
16	91,6	84,4	71,2	50	37,2	23,6	16,8	10	6,8	4,4	2,4	2	3,2	2	2,4	2	2	2,4	2,4	2	2
17	86,8	74,4	56,4	37,2	27,2	17,6	12,4	8	4,4	2,8	1,2	1,2	1,6	2	2	2	2	2	2	2	2
18	84,4	70,4	49,2	34,8	25,2	17,6	12,8	8	5,2	3,2	1,2	0,8	0,8	2	2	2	2	2	2	2	2
19	88,8	71,2	50,8	34,4	20,8	15,2	9,6	8	5,2	3,2	2	2	2	2	2	2	2	2	2	2	2
20	88	69,6	45,6	27,6	16,8	11,6	8,4	5,6	4	3,2	2	2,4	2,4	3,2	2,4	2	2	2	2	1,6	2,4
21	88	68,8	46	25,2	14,4	11,6	8,4	4	2	1,6	2	2,8	2,4	2,4	3,2	2,8	2,8	2	2	2	2
22	91,2	76,4	53,6	30,4	16,8	12,4	9,6	6	4	4	2,8	1,6	2	2	2	2	2	2	2	2	2
23	92,8	79,2	56,4	33,2	19,6	14,4	10,8	8	6	4,4	3,2	2,8	2,8	2,4	2	2	2	2	3,6	3,2	3,2
24	93,2	80	64	41,2	25,2	17,2	11,2	7,2	3,2	2,4	2,4	2,4	2,4	2	2	2	2	1,6	2	2	2
25	95,6	85,6	71,6	47,6	29,6	21,6	14	9,2	4,4	2,8	2	2	1,6	2	1,6	2,4	2	1,2	0,4	0,8	0,8
26	94,4	87,2	72,4	47,6	27,2	18,4	11,2	8	6	5,2	3,2	3,2	3,2	2	2	2	2	2	2	2	2
27	94,8	88,8	77,6	52,4	31,6	19,6	13,2	7,6	5,2	3,2	2,8	3,2	2	2	2	2	2	2	2	2	2
28	94,8	90	77,6	49,2	24	18,4	12	6,4	4,8	2,8	2,4	2,4	4	3,6	2,4	2,8	2,8	2	1,6	2	2
29	93,6	88,4	71,2	39,6	17,6	11,2	7,2	4,8	4	4,8	4	3,6	3,2	2,8	2,4	2,4	2,8	2,8	2	2	2
30	92,4	86	67,6	38	16	11,6	8,8	6	5,2	5,2	4	4	2,8	3,2	2,4	2,4	2,4	2,4	2,4	2	2
31	92,4	82	62	36,4	19,2	13,6	6,8	5,2	4,8	3,2	3,2	2,4	2	1,6	2,8	2,8	2,8	2,8	2,4	2	2
32	93,2	84	60	31,6	18,8	14,8	9,6	6,8	6,4	4,4	4	2,8	2,8	2,4	2,4	2,4	2,4	2	2	2	2
33	91,2	78,4	52,8	28,4	17,2	12,4	6	6,4	6	4,8	4	3,2	2,8	2,8	3,2	3,2	2,4	2,8	2	2	2
34	88	76,4	54,8	34	18	13,2	9,2	4,8	4	3,2	3,2	3,6	3,6	3,2	3,2	2,8	3,2	2,8	2	2	2
35	84,4	71,2	46,4	26,8	18,4	12,8	7,6	5,6	3,6	2,8	3,2	4	3,2	3,2	3,2	2,8	2,8	2	2	2	2
skóre	87,12	76,68	59,60	39,45	25,36	17,76	11,70	7,44	5,22	3,91	3,14	2,93	2,61	2,56	2,40	2,31	2,24	2,11	1,99	1,91	1,88

Tab. 12 – Rozpoznávací skóre [%] u rotačního testu v závislosti na SNR, jedná se o úlohu rozpoznávání akustického signálu řeči, přidáván je „white“ šum, na posledním řádku je výsledné zprůměrňované skóre

SNR [dB]	Druh rozpoznávání a použitý šum					
	"babble" šum			*"white" šum*		
	AV	*Audio*	rozdíl	*AV*	*Audio*	*rozdíl*
18	96,59	98,43	-1,84	92,42	87,12	5,30
17	96,55	98,47	-1,92	88,03	76,68	11,35
16	96,54	98,45	-1,91	80,35	59,60	20,75
15	96,48	98,42	-1,94	67,76	39,45	28,31
14	96,41	98,42	-2,01	49,82	25,36	24,46
13	96,34	98,32	-1,98	33,99	17,76	16,23
12	95,94	98,10	-2,16	21,74	11,70	10,04
11	95,43	97,39	-1,97	14,23	7,44	6,79
10	94,08	95,94	-1,86	9,62	5,22	4,40
9	91,51	92,40	-0,89	6,51	3,91	2,61
8	87,33	83,89	3,44	4,81	3,14	1,67
7	79,18	72,54	6,64	3,61	2,93	0,68
6	68,79	58,69	10,10	2,93	2,61	0,31
5	57,33	44,96	12,37	2,49	2,56	-0,07
4	45,74	27,03	18,71	2,22	2,40	-0,18
3	32,62	13,42	19,20	2,24	2,31	-0,07
2	21,29	5,93	15,36	2,24	2,24	0,00
1	13,44	3,68	9,76	2,30	2,11	0,19
0	9,12	3,07	6,05	2,26	1,99	0,27
-1	6,51	2,85	3,67	2,18	1,91	0,28
-2	5,06	2,69	2,38	2,23	1,88	0,35

Tab. 13 – Shrnutí rotačních testů posuzujících závislost na okolním ruchu, v tabulce je zobrazeno rozpoznávací skóre [%] a rozdíl vyjadřující zlepšení přidáním vizuální složky